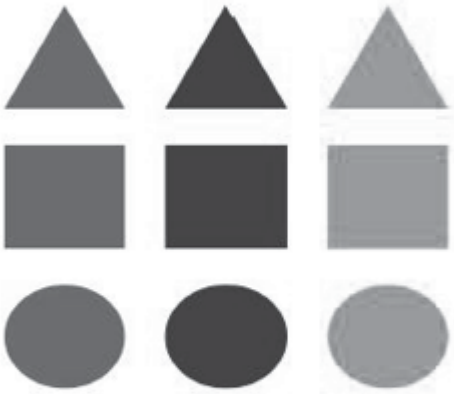
THE DISCIPLINE OF ORGANIZING

INFORMATICS EDITION



Edited by **ROBERT J. GLUSHKO**

The Discipline of Organizing

Informatics Edition

Edited by *Robert J. Glushko*

Principal Authors: Robert J. Glushko, Jess Hemerly, Murray Maloney, Kimra McPherson, Robyn Perry, Vivien Petras, Ryan Shaw, and Erik Wilde

Contributing Authors: Rachelle Annechino, Matt Earp, J.J.M. Ekaterin, Graham Freeman, Ryan Greenberg, Daniel Griffin, Carl Lagoze, Ian MacFarland, Michael Manoochehri, Sean Marimpietri, Matthew Mayernik, Karen Joy Nomorosa, Hyunwoo Park, Alberto Pepe, Jordan Shedlock, Isabelle Sperano, Daniel D. Turner, and Longhao Wang

O'Reilly Media • Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

The Discipline of Organizing

Editor: Robert J. Glushko
Principal Authors: Robert J. Glushko, Jess Hemerly, Murray Maloney, Kimra McPherson, Robyn Perry, Vivien Petras, Ryan Shaw, and Erik Wilde
Contributing Authors: Rachelle Annechino, Matt Earp, J.J.M. Ekaterin, Graham Freeman, Ryan Greenberg, Daniel Griffin, Carl Lagoze, Ian MacFarland, Michael Manoochehri, Sean Marimpietri, Matthew Mayernik, Karen Joy Nomorosa, Hyunwoo Park, Alberto Pepe, Jordan Shedlock, Isabelle Sperano, Daniel D. Turner, and Longhao Wang
Case Study Authors: Daniel Brenners, Gracen Brilmyer, Jason Danker, David Eicke, Colin Gerber, Shaun Giudici, Emilie Hardman, Hassan Jannah, Sandra Lee, Carlo Liquido, Ian MacFarland, Andrew McConachie, Emily Paul, Pratibha Rathore, Becca Stanger, and Suhaib Syed
Bibliography Editors: Lisa Jervis, Shohei Narron, and Anne Wootton
Cover Designer: Jen Wang
Illustrators: Divya Anand, Ajeeta Dhole, Robyn Perry, Christina Pham, and Raymon Sutedjo-The
CSS Page Designs: Nellie McKesson and Adam Witwer
Statistics Visualizations: Pablo Arvizu and Isabelle Sperano
DocBook Consultants: Bob Stayton and Jirka Kosek
Markup Editor: Murray Maloney

Published by Robert J. Glushko.

Produced in the United States of America.

Library of Congress Cataloging-in-Publication Data.

The Discipline of Organizing (4th ed.)/ edited by Robert J. Glushko.
p. cm
Includes bibliographical references and index.

ISBN 978-1-491-97065-2 (academic PDF ebook)

1. Information organization.
2. Information resources management.
3. Metadata.
I. Glushko, Robert J., editor of compilation.
Z666.5.D57 2013
025—dc23

10 9 8 7 6 5 4 3 2 1 0

To Aristotle, Plato, Linnaeus, Condorcet, Wittgenstein...
Panizzi, Cutter, Raganathan, Svenonius...
Gibson, Norman, Rosch, Barsalou...
Adam Smith, Coase, Williamson...
Simon, Salton, Miller, Dumais...
Bush, Engelbart, Nelson, Berners-Lee...
...and the countless others whose diverse perspectives
we have synthesized in the discipline of organizing.

Table of Contents

Chapter 1

Foundations for Organizing Systems

Robert J. Glushko

1.1 The Discipline of Organizing

To *organize* is to create capabilities by intentionally imposing order and structure.

Organizing is such a common *activity* that we often do it without thinking much about it. We organize shoes in our closet, books on our book shelves, spices in our kitchen, receipts and records in tax preparation folders, and people on business projects and sports teams. Quite a few of us have jobs that involve specific types of organizing tasks. We might even have been explicitly trained to perform them by following specialized disciplinary practices. We might learn to do these tasks very well, but even then we often do not reflect on the similarity of the organizing tasks we do and those done by others, or on the similarity of those we do at work and those we do at home. We take for granted and as givens the concepts and methods used in the Organizing System we work with most often.

The goal of this book is to help readers become more self-conscious about what it means to organize resources of any type and about the principles by which the resources are organized. In particular, this book introduces the concept of an *Organizing System*: an intentionally arranged collection of resources and the interactions they support. The book analyzes the design decisions that go into any systematic organization of resources and the design patterns for the interactions that make use of the resources, as follows:

We organize physical things. Each of us organizes many kinds of things in our lives—our books on bookshelves; printed financial records in folders and filing cabinets; clothes in dressers and closets; cooking and eating utensils in kitchen drawers and cabinets. Public libraries organize printed books, periodicals, maps, CDs, DVDs, and maybe some old record albums. Research libraries also organize rare manuscripts, pamphlets, musical scores, and many other kinds of printed information. Museums organize paintings, sculptures, and other artifacts of cultural, historical, or scientific value. Stores and suppliers organize their goods for sale to consumers and to each other. Sports leagues organize players into teams, and the teams organize players by position or role.

We organize information about physical things. Each of us organizes information about things: when we inventory the contents of our house for insurance purposes, when we sell our unwanted stuff on eBay, or when we rate a restaurant on Yelp. Library card catalogs, and their online replacements, tell us what books a library's collection contains and where to find them. Sensors and RFID tags track the movement of goods—even library books—through supply chains, and the movement (or lack of movement) of cars on highways.

We organize digital things. Each of us organizes personal digital information—email, documents, ebooks, MP3 and video files, appointments, contacts—on our computers, smartphone, ebook readers, or in "the cloud," —through information services that use Internet protocols. Large research libraries organize digital journals and books, computer programs, government and scientific datasets, databases, and many other kinds of digital information. Companies organize their digital business records and customer information in enterprise applications, content repositories, and databases. Hospitals and medical clinics maintain and exchange electronic health records and digital X-rays and scans.

We organize information about digital things. Digital library catalogs, web portals, and aggregation websites organize links to other digital resources. Web search engines use content and link analysis along with relevance ratings, to organize the billions of web pages competing for our attention. Web-based services, data feeds and other information resources can be interconnected and choreographed to carry out information-intensive business processes, or aggregated and analyzed to enable prediction and personalization of information services.

Let us take a closer look at these four different types or contexts of organizing. We contrasted "organizing things" with "organizing information." At first glance it might seem that organizing physical things like books, compact discs, machine parts, or cooking utensils has an entirely different character than organizing intangible digital things. We often arrange physical things according to their shapes, sizes, material of manufacture, or other intrinsic and visible properties: for example, we might arrange our shirts in the clothes closet by style and color, and we might organize our music collection by separating the old vinyl albums from the CDs. We might arrange books on bookshelves by their sizes, putting all the big, heavy picture books on the bottom shelf. Organization for clothes and information artifacts in tangible formats that is based on visible properties does not seem much like how you store and organize digital books on your Kindle or arrange digital music on your music player. Arranging, storing, and accessing X-rays printed on film might appear to have little in common with these activities when the X-rays are in digital form.

It is hardly surprising that organizing things and organizing information sometimes do not differ much when information is represented in a tangible way. The era of ubiquitous digital information of the last decade or two is just a blip in time compared with the more than ten thousand years of human experience with information carved in stone, etched in clay, or printed with ink on papyrus, parchment, or paper. These tangible information artifacts have deeply embedded the notion of information as a physical thing in culture, language, and methods of information design and organization. This perspective toward tangible information artifacts is especially prominent in rare book collections where books are revered as physical objects with a focus on their distinctive binding, calligraphy, and typesetting.

Nevertheless, at other times there are substantial differences in how we organize things and how we organize information, even when the latter is in physical form. We more often organize our "information things" according to what they are about rather than on the basis of their visible properties. At home we sort our CDs by artist or genre; we keep cookbooks separate from travel books, and fiction books apart from reference books. Libraries employ subject-based classification schemes that have a few hundred thousand distinct categories.

Likewise, there are times when we pay little attention to the visible properties of tangible things when we organize them and instead arrange them according to functional or task properties. We keep screwdrivers, pliers, a hammer, a saw, a drill, and a level in a toolbox or together on a workbench, even though they have few visual properties in common. We are not organizing them because of what we see about them, but because of what we know about to use them. The task-based organization of the tools has some similarity to the subject-based organization of the library.

We also contrasted "organizing things" with "organizing information about things." This difference seems clear if we consider the traditional library card catalog, whose printed cards describe the books on library shelves. When the things and the information about them are both in physical format, it is easy to see that the former is a primary resource and the latter a surrogate or associated resource that describes or relates to it.

What Is Information?

Most of the hundreds of definitions of information treat it as an idea that swirls around equally hard-to-define terms like "data," "knowledge," and "communication." Moreover, these intellectual and ideological perspectives on information coexist with more mundane uses of the term, as when we ask a station agent: "Can you give me some information about the train schedule?"

An abstract view of information as an intangible thing is the intellectual foundation for both modern information science and the information economy and society. Nevertheless, the abstract view of information often conflicts with the much older idea that information is a tangible thing that naturally arose when information was inextricably encoded in material formats. We often blur the sense of "information as content" with the sense of "information as container," and we too easily treat the number of stored bits on a computer or in "the cloud" as a measure of information content or value.

When it comes to "organizing information about digital things" the contrast is much less clear. When you search for a book using a search engine, first you get the catalog description of the book, and often the book itself is just a click away. When the things and the information about them are both digital, the contrast we posed is not as sharp as when one or both of them is in a physical format. And while we used X-rays—on film or in digital format—as examples of things we might organize, when a physician studies an X-ray, is it not being used as information about the subject of the X-ray, namely, the patient? And when businesspeople make marketing and pricing decisions by analyzing digital information about what and when people buy, we can think of this as organizing customers into categories, or as organizing customer information.

These differences and relationships between "physical things" and "digital things" have long been discussed and debated by philosophers, linguists, psychologists, and others. (See the sidebars, What Is Information? (page 28) and The Distinction between Data and Information (page 29).)

The distinctions among organizing physical things, organizing digital things, or organizing information about physical or digital things are challenging to describe because many of the words we might use are as overloaded with multiple

meanings as "information" itself. For example, the library science perspective often uses presentation or implementation properties in definitions of "document," using the term to refer only to traditional physical forms. In contrast, the informatics or computer science perspective takes an abstract view of "document" to refer to any self-contained unit of information, separating a document's content from its presentation or container.[2[Com]]

The Distinction between Data and Information

Astute readers might have noticed that we included sensor data as "information about physical things" and data feeds as "information about digital things." Many textbooks in the information science and knowledge management fields distinguish data and information in a more precise way. To them, data sits at the bottom of an Information Hierarchy, Knowledge Pyramid, or DIKW Hierarchy in which Data is transformed into Information, which is transformed into Knowledge, which is then transformed into Wisdom.

In this framework, data are raw or elementary observations about properties of objects, events, and their environment. Data becomes information when it is aggregated, processed, analyzed, formatted, and organized to add meaning and context so it can be used to answer questions. This processing can include calculation, inference, or refinement operations on the data. For example, measurements of temperature, precipitation, and wind speed are data. When combined and summarized, a set of data becomes statistical information about the weather on a particular day. When collected over a period of months or years, these datasets become information about the climate of the location where they were collected.

The Discipline of Organizing does not make this sharp contrast between data and information in the Hierarchy/Pyramid. People who read this book are likely to be aspiring or practicing professionals in information-intensive industries where information and data are often treated as synonyms to mean the content of a database or data-managing application. A distinction between data and information might be useful in theory, but not in these applied settings.

The distinction between data and information is also being blurred by the expansion in the scope of the definition of data in the emerging career field of *data science*. Indeed, a popular introductory text eliminates information entirely from the Hierarchy/Pyramid with its title, *Discovering knowledge in data: an introduction to data mining.*[4[DS]]

Similar definitional variation occurs with "author" or "creator." When we say that "Herman Melville is the author of *Moby Dick*" (Melville 1851) the meaning of "author" does not depend on whether we have a printed copy or an ebook in

mind, but what counts as authorship varies a great deal across academic disciplines. Furthermore, different standards for describing resources disagree in the precision with which they identify the person(s) or organization(s) primarily responsible for creating the intellectual content of the resource, People who are serious about music description rightly criticize streaming services and online stores that have only a single "artist" field because this fails to distinguish the composer, conductor, orchestra, and other people with distinct roles in creating the music.

If we allow the concept of information to be anything we can study—to be "anything that informs"—the concept becomes unbounded. Our goal in this book is to bridge the intellectual gulf that separates the many disciplines that share the goal of organizing but differ in what they organize. This requires us to focus on situations where information exists because of intentional acts to create or organize. (See the sidebar, The Discipline of Organizing (page 31))

Many of the foundational topics for a discipline of organizing have traditionally been presented from the perspective of the library sector and taught as "library and information science." These include bibliographic description, classification, naming, authority control, curation, and information standards. In recent decades these foundations have been built on and extended by computer science, cognitive science, informatics, and other new fields to include more private sector and non-bibliographic contexts, multimedia and social media, and new information-intensive applications and service systems enabled by mobile, pervasive, and scientific computing. The latest additions to the discipline of organizing are coming from *data science* and *machine learning*, introducing considerations of speed and scale that arise when massive computational power and new statistical techniques are harnessed to organize and act on information.

The new methods and tools of *data science* and *machine learning* let us organize more information, to do it faster, and to make predictions based on what people have clicked on, bought, or said. But this is not the first time that new ideas and technologies have challenged how people organized and interacted with resources. Fifty years ago, searchable online catalogs radically changed how people used libraries. The web, invented less than thirty years ago so that scientists could share technical reports, is now an essential part of many human activities. It is important not to view the latest new thing as changing everything, because new things will continue to come, and these technology breakthroughs still depend on and complement the organizing work done by people. Data science will not replace human organizers, any more than any other science has replaced humans. (See sidebar, Data Science and the Discipline of Organizing (page 32)).

The Discipline of Organizing

A *discipline* is an integrated field of study in which there is some level of agreement about the issues and problems that deserve study, how they are interrelated, how they should be studied, and how findings or theories about the issues and problems should be evaluated. A *framework* is a set of concepts that provide the basic structure for understanding a domain, enabling a common vocabulary for different explanatory theories.

Organizing is a fundamental issue in many disciplines, most notably library and information science, computer science, systems analysis, informatics, law, economics, and business. However, these disciplines have only limited agreement in how they approach problems of organizing and what they seek as their solutions. For example, library and information science has traditionally studied organizing from a public sector bibliographic perspective, paying careful attention to user requirements for access and preservation, and offering prescriptive methods and solutions. In contrast, computer science and informatics tend to study organizing in the context of information-intensive business applications with a focus on process efficiency, system architecture, and implementation. The disciplines of management and industrial organization deal with the organization of human, material, and information resources in contexts shaped by commercial, competitive, and regulatory forces.

This book presents a more abstract framework for issues and problems of organizing that emphasizes the common concepts and goals of the disciplines that study them. Our framework proposes that every system of organization involves a collection of resources, and we can treat physical things, digital things, and information about such things as resources. Every system of organization involves a choice of properties or principles used to describe and arrange the resources, and ways of supporting interactions with the resources. By comparing and contrasting how these activities take place in different contexts and domains, we can identify patterns of organizing and see that Organizing Systems often follow a common life cycle. We can create a discipline of organizing in a disciplined way.

This is why we need to take a transdisciplinary view that lets us emphasize what the different disciplines have in common and how they fit together rather than what distinguishes them. Resource selection, organizing, interaction design, and maintenance are taught in every discipline, but these concepts go by different names. A vocabulary for discussing common organizing challenges and issues that might be otherwise obscured by narrow disciplinary perspectives helps us understand existing systems of organizing better while also suggesting how to invent new ones by making different design choices.

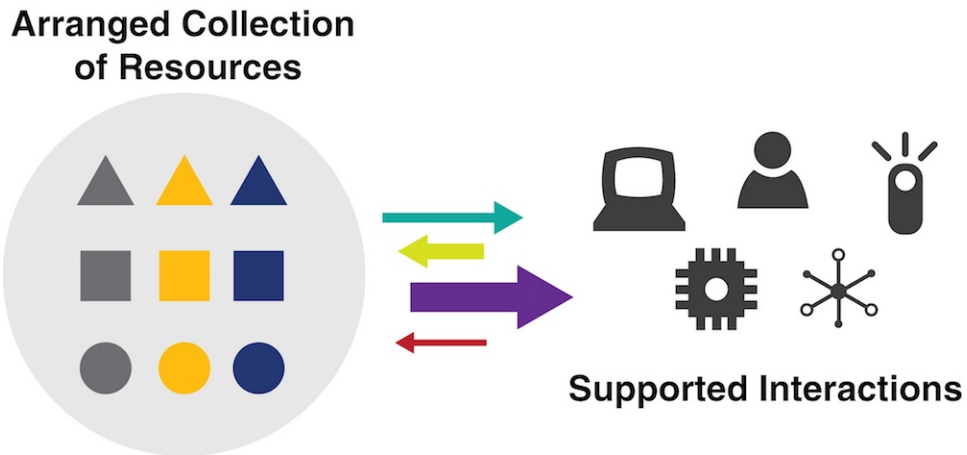
Data Science and the Discipline of Organizing

Advances in computing power and statistical techniques are making it possible to identify patterns in data and extract meaningful information at a scale never before possible. Many books and articles about *data science*, machine learning, and predictive analytics make bold predictions that these emerging fields will radically change the world. These claims are both provocative and promising, but at its core, data science is about how resources are selected, described, and organized; concepts with a long tradition in information and library science. Instead of organizing and describing the books in a library or the products in a warehouse, a data scientist might organize information about books or products into massive data tables, treating each resource as a row and its descriptive properties as the columns. After people might have organized books or products into categories, machine learning techniques might classify new books or products using those categories, or perhaps discover new categories based on access or purchasing behaviors. So while the techniques of data science are new, many of the challenges are not; data scientists need to select resources wisely and decide how best to describe them; they need to understand that resource description and categorization can be biased; they need to understand the tradeoffs and complements between people and computers; and, they need to test the discoveries that algorithms make with controlled experiments.

To make sense of the discussions around data science, one must understand the difference between kind and degree. A hundred years ago, a car's highway travel speed was about forty miles an hour. Today's cars travel twice as fast, but this is just a change in degree. However, an increase in speed to about 17,500 miles an hour achieves an "orbital velocity" that allows us to go into Earth orbit in space, travel that is different in kind.

What about data science? Some data science involves collections of data that are "tall," containing many millions or even billions of records that each have a relatively small number of variables. Being able to analyze "tall" data more rapidly than ever before is primarily a change in degree compared with traditional database techniques. Nevertheless, for collections of data that are "wide," where each record might contain hundreds or thousands of variables, data science techniques might allow us to see patterns that could not be seen at all, or could not be seen affordably and in quantity. Here, data science might be yielding changes in kind.[6[DS]]

Figure 1.1. An Organizing System.



An Organizing System is a collection of resources arranged in ways that enable people or computational agents to interact with them.

1.2 The "Organizing System" Concept

We propose to unify many perspectives about organizing and information with the concept of an *Organizing System*, *an intentionally arranged collection of resources and the interactions they support.* This definition brings together several essential ideas that we will briefly introduce in this chapter and then develop in detail in subsequent chapters.

Figure 1.1 depicts a conceptual model of an Organizing System that shows intentionally arranged resources, interactions (distinguished by different types of arrows), and the human and computational agents interacting with the resources in different contexts.

An Organizing System is an abstract characterization of how some collection of resources is described and arranged to enable human or computational agents to interact with the resources. The Organizing System is an architectural and conceptual view that is distinct from the physical arrangement of resources that might embody it, and also distinct from the person, enterprise, or institution that implements and operates it. These distinctions are sometimes hard to maintain in ordinary language; for example, we might describe some set of resource descriptions, organizing principles, and supported interactions as a "library" Organizing System. However, we also need at times to refer to a "library" as the

institution in which this Organizing System operates, and of course the idea of a "library" as a physical facility is deeply engrained in language and culture.

Our concept of the Organizing System was in part inspired by the concepts proposed in 2000 for bibliographic domains by Elaine Svenonius, in *The Intellectual Foundation of Information Organization.* She recognized that the traditional *information organization* activities of bibliographic description and cataloging were complemented, and partly compensated for, by automated text processing and indexing that were usually treated as part of a separate discipline of *information retrieval*. Svenonius proposed that decisions about organizing information and decisions about retrieving information were inherently linked by a tradeoff principle and thus needed to be viewed as an interconnected system: *"The effectiveness of a system for accessing information is a direct function of the intelligence put into organizing it" (p.ix)*. We celebrate and build upon her insights by beginning each of the sub-parts of Chapter 2 with a quote from her book.

A systems view of information organization and information retrieval captures and provides structure for the inherent tradeoffs obscured by the silos of traditional disciplinary and category perspectives: the more effort put into organizing information "on the way in" when it is created or added to a collection, the more effectively it can be retrieved, and the more effort put into retrieving information "on the way out," the less it needs to be organized first. Sometimes a collection of resources is highly organized, but because it was organized by someone else for different purposes that have in mind, we need to reorganize it "on the way in." This is especially common with digital text or datasets, where previously organized resources or their descriptions might be sorted, translated in format or language, combined, summarized, or otherwise transformed to fit into a new Organizing System. For example, to understand seasonal buying patterns, a retailer might combine shopping data with weather data and calendar data about commonly-watched sporting events (because bad weather and broadcast sports cause people to stay home), and all three datasets would need to describe "time" and "location" in the same way.

A systems view no longer contrasts information organization as a human activity and information retrieval as a machine activity, or information organization as a topic for library and information science and information retrieval as one for computer science. Instead, we readily see that computers now assist people in organizing and that people contribute much of the information used when computers analyze and organize resources. For example, many algorithms for computational classification use *supervised learning* approaches that start with items classified by people.

Finally, a systems view can be applied to Organizing Systems with any kind of resource, enabling more nuanced discussion of how economic, social, and cogni-

tive costs and benefits of organizing are allocated among different stakeholders. Every Organizing System is biased by the perspectives and experiences of the people who create it. Some of these biases are inescapable, a kind of automatic organizing, because they reflect innate human perceptual and cognitive capabilities. Our minds impose structure and find patterns, even when there aren't any, and we are not capable of acting perfectly rationally, so we simplify without realizing it. People are also not very good at thinking about future possibilities and revising their expectations given new evidence, and this mental inertia makes us preserve resources and interactions in Organizing Systems that are no longer needed. Other biases in Organizing Systems reflect more intentional choices that implicitly or explicitly create winners or losers, treat some interactions as preferred while deprecating others, or otherwise impose or overlay a set of values on the stakeholders of the system. For example, many Organizing Systems arrange people in groups or queues to make interactions more efficient, but when an airline gives boarding priority to customers who paid more for their tickets it might not seem fair to you if are in the last boarding group.

1.3 The Concept of "Resource"

Resource has an ordinary sense of anything of value that can support goal-oriented activity. This definition means that a resource can be a physical thing, a non-physical thing, information about physical things, information about non-physical things, or anything you want to organize. Other words that aim for this broad scope are *entity*, *object*, *item*, and *instance*. *Document* is often used for an information resource in either digital or physical format; *artifact* refers to resources created by people, and *asset* for resources with economic value.

Resource has specialized meaning in Internet architecture. It is conventional to describe web pages, images, videos, and so on as *resources*, and the protocol for accessing them, *Hypertext Transfer Protocol (HTTP)*, uses the *Uniform Resource Identifier (URI)*.[8][Web]

Treating as a *primary resource* anything that can be identified is an important generalization of the concept because it enables web-based services, data feeds, objects with RFID tags, sensors or other *smart devices,* or computational agents to be part of Organizing Systems.

Instead of emphasizing the differences between tangible and intangible resources, we consider it essential to determine whether the tangible resource has information content—whether it needs to be treated as being "about" or representing some other resource rather than being treated as a thing in itself. Whether a book is printed or digital, we focus on its information content, what it is about; its tangible properties become secondary. In contrast, the hangars in our closet and the measuring cups in our kitchen are not about anything more than their obvious utilitarian features, which makes their tangible properties

Concert Tickets

Tickets are physical artifacts that convey event-related metadata: including time, place, and seat number; price and terms of admission; and featured performers. For concert goers, tickets offer the promise of all that, and a memory of the ineffable quality of more.

(Photo by Murray Maloney.)

Concert Ticket

A concert ticket is a vehicle for conveying a package of assertions about an event, so it is a *description resource*, like a card in a library card catalog. A concert ticket is also a *resource* in its own right, with intrinsic value; it can be bought and sold, sometimes for a greater price than its *resource description* specifies. A ticket is a license to use a seat in a venue for a specified purpose at a specified time; after the event, the ticket loses its intrinsic value, but might acquire extrinsic value as an artifact in a collection like this one.

most important. (Of course, there is no sharp boundary here; you can buy "fashion hangers" that make a style statement, and the old measuring cup could be a family memento because it belonged to Grandma).

Many of the resources in Organizing Systems are *description resources* or *surrogate resources* that describe the *primary resources*; library catalog entries or the list of results in web search engines are familiar examples. In museums, information about the production, discovery, or history of ownership of a resource can be more important than the resource; a few shards of pottery are of little value without these associated information resources. Similarly, business or scientific data often cannot be understood or analyzed without additional information about the manner in which they were collected. Most web-based businesses exploit data about how users

interact with resources, such as the log files that record every web search you make, every link you click, and every web page you visit.

Resources that describe, or are associated with other resources are sometimes called *metadata*. However, when we look more broadly at Organizing Systems, it is often difficult to distinguish between the resource being described and any description of it or associated with it. One challenge is that when descriptions are embedded in resources, as metadata often is—in the title page of a book, the masthead of a newspaper, or the source of web pages—deciding which resources are primary is often arbitrary.

A second challenge is that what serves as metadata for one person or process can function as a primary resource or data for another one. Rather than being an inherent distinction, the difference between primary and associated resources is often just a decision about which resource we are focusing on in some situation. An animal specimen in a natural history museum might be a *primary resource* for museum visitors and scientists interested in anatomy, but information about where the specimen was collected is the *primary resource* for scientists interested in ecology or migration.

Organizing Systems can refer to people as resources, and we often use that term to avoid specifying the gender or specific role of an employee or worker, as in the management concept of the "human resources" department in a workplace. A business is defined by its intentional arrangement of human resources, and there is both variety and regularity in these arrangements (see the sidebar, Business Structures (page 296) in §6.5). [9[Phil]]

Human resources in Organizing Systems can be understood much the same way as inanimate physical or digital resources: they are selected, organized, and managed, and can create value individually or through their interactions with others inside and outside of the system.[10[Bus]] However, human beings are uniquely complicated resources, and any Organizing System that uses them must take into account their rights, motivations, and relationships. (See the sidebar, People as Resources (page 172).)

1.4 The Concept of "Collection"

A *collection* is a group of resources that have been selected for some purpose. Similar terms are *set* (mathematics), *aggregation* (data modeling), *dataset* (science and business), and *corpus* (linguistics and literary analysis).

We prefer *collection* because it has fewer specialized meanings. *Collection* is typically used to describe personal sets of physical resources (my stamp or record album collection) as well as digital ones (my collection of digital music). We distinguish law libraries from software libraries, knowledge management systems from data warehouses, and personal stamp collections from coin collec-

tions primarily because they contain different kinds of resources. Similarly, we distinguish document collections by resource type, contrasting narrative document types like novels and biographies with transactional ones like catalogs and invoices, with hybrid forms like textbooks and encyclopedias in between.

A collection can contain identifiers for resources along with or instead of the resources themselves, which enables a resource to be part of more than one collection, like songs in playlists.

A *collection* itself is also a *resource*. Like other resources, a collection can have description resources associated with it. An *index* is a *description resource* that contains information about the locations and frequencies of terms in a document *collection* to enable it to be searched efficiently.

Because *collections* are an important and frequently used kind of *resource*, it is important to distinguish them as a separate concept. In particular, the concept of *collection* has deep roots in libraries, museums and other institutions that select, assemble, arrange, and maintain resources. Organizing Systems in these domains can often be described as collections of collections that are variously organized according to resource type, author, creator, or collector of the resources in the collection, or any number of other principles or properties. In business contexts, the use of "collection" to describe a set of resources is much less common, but businesses organize many types of resources, including their employees, suppliers, customers, products, and the tangible and intangible assets used to create the products and run the business. Indeed, a business itself can sometimes be abstractly described as a collection of resources, especially when the resources are software components or services. (See endnote[46 [Com]].)

A type of resource and its conventional Organizing System are often the focal point of a discipline. Category labels such as library, museum, zoo, and data repository have core meanings and many associated experiences and practices. Specialized concepts and vocabularies often evolve to describe these. The richness that follows from this complex social and cultural construction makes it difficult to define category boundaries precisely.

We can always create new categories by stretching the conventional definitions of "library" or other familiar Organizing Systems and adding modifiers, as when Flickr is described as a web-based photo-sharing library. But whenever we define an Organizing System with respect to a familiar category, the typical or mainstream instances and characteristics of that category that are deeply embedded in language and culture are reinforced, and those that are atypical are marginalized. In the Flickr case, this means we suggest features that are not there (like authoritative classification) or omit the features that are distinctive (like tagging by users).

What Is a Library?

Most birds fly, but not all of them do. What characteristics are most important to us when we classify something as a bird? What characteristics are most important when we think of something as a library?

We might treat *circulation*, borrowing and returning the same item, as one of the interactions with resources that defines a library. In that case, an institution that lends items in its collection with the hope that the borrowers return something else that is better hardly seems like a library. But if the resources are the seeds of heirloom plants and the borrowers are expected to return seeds from the plants they grew from the borrowed seeds, perhaps "seed library" is an apt name for this novel Organizing System. Similarly, even though the resources in its collection are encyclopedia articles rather than living species, the Wikipedia open-source encyclopedia resembles the Seed Library by encouraging its users to "return" articles that are improvements of the current ones.

The photo-sharing website Flickr functions for most of its users as a personal photo archiving site. Flickr's billions of user-uploaded photos and the choice of many users to share them publicly transform it into a searchable shared collection, and many people also think of Flickr as a photo library. But Flickr lacks the authoritative description and standard classification that typify a library.

A similar categorization challenge arises with the Google Books digitization project. [11][Law]

More generally, a categorical view of Organizing Systems makes it matter greatly which category is used to anchor definitions or comparisons. The Google Books project makes out-of-print and scholarly works vastly more accessible, but when Google co-founder Sergei Brin described it as "a library to last forever" it upset many people with a more traditional sense of what the library category implies. We can readily identify design choices in Google Books that are more characteristic of the Organizing Systems in business domains, and the project might have been perceived more favorably had it been described as an online bookstore that offered many beneficial services for free.

1.5 The Concept of “Intentional Arrangement”

Intentional arrangement emphasizes explicit or implicit acts of organization by people, or by computational processes acting as proxies for, or as implementations of, human intentionality. Intentional arrangement is easiest to see in Organizing Systems created by individual people who can make all the necessary decisions about organizing their own resources. It is also easy to see in Organizing Systems created by institutions like libraries, museums, businesses, and governments where the responsibility and authority to organize is centralized and explicit in policies, laws, or regulations.

However, top-down intentionality is not always necessary to create an Organizing system. Organization can emerge over time via collective behavior in situations without central control when decisions made by individuals, each acting intentionally, create traces, records, or other information that accumulates over time. Organizing systems that use bottom-up rather than top-down mechanisms are sometimes called self-organizing, because they emerge from the aggregated interactions of actors with resources or with each other. *Self-organizing systems* can change their internal structure or their function in response to feedback or changed circumstances.

This definition is broad enough to include business and biological ecosystems, traffic patterns, and open-source software projects. Another good example of emergent organization involves path systems, where people (as well as ants and other animals) can follow and thereby reinforce the paths taken by their predecessors. When highly orderly and optimal arrangements emerge from local interactions among ants, bees, birds, fish, and other animal species, it is often called “swarm intelligence.” When this happens with human ratings for news stories, YouTube videos, restaurants, and other types of digital and physical resources we call it “crowdsourcing.” What the animal and human situations have in common is that information is being communicated between individuals. Sometimes this communication is direct, as when Amazon shows you the average rating for a book or what books have been bought by people like you. At other times the communication is indirect, achieved when the agents modify their environment (as they do when they create paths) and others can respond to these modifications. Adam Smith's “invisible hand” is another example where individuals collectively generate an outcome they did not directly intend but that arose from their separate self-interested actions as they respond to price signals in the marketplace. Likewise, even though there is no top-down organization, the web as a whole, with its more than a trillion unique pages, is a self-organizing system that at its core follows clear organizing principles.[12[Com] 13[Web] 14[IA]]

The Web as an Organizing System

Today's web barely resembles the system for distributing scientific and technical reports it was designed to be when physicist and computer scientist Tim Berners-Lee devised it in 1990 at the European Organization for Nuclear Research (CERN) lab near Geneva. However, as an *Organizing System* the web still follows the principles that Berners-Lee defined at its creation. These include standard data formats and interaction protocols; no need for centralized control of page creation or linking; remote access over the network from anywhere; and the ability to run on a large variety of computers and operating systems. This architecture makes the web open and extensible, but gives it no built-in mechanisms for authority or trust.[15][Web]

Because the web works without any central authority or authorship control, any person or organization can add to it. As a result, even though the web as a whole does not exhibit the centralized intentional arrangement of resources that characterizes many Organizing Systems, we can view it as consisting of millions of Organizing Systems that each embody a separate intentional arrangement of web pages. In addition, we most often interact with the web indirectly by using a search engine, which meets the definition of *Organizing System* because its indexing and retrieval algorithms are principled.

A great many Organizing Systems are implemented as collections of web pages. Some of these collections are created on the web as new pages, some are created by transforming existing collections of resources, and some combine new and existing resources.

The requirement for intentional arrangement excludes naturally occurring patterns created by physical or geological processes from being thought of as Organizing Systems. There is information in the piles of debris left after a tornado or tsunami and the strata of the Grand Canyon. But they are not Organizing Systems because the patterns of arrangement were created by deterministic natural forces rather than by agents following one or more organizing principles. On the other hand, collections of geological data like the measurements of chemical composition from different strata and locations in the Grand Canyon **are** Organizing Systems. Decisions about what to measure, how to combine and analyze the measurements, and any theories that are tested or created, reflect intentional arrangement of the data by the geologist.

Other patterns of resource arrangements are illusions or perceptions that require a particular vantage point. The best examples are patterns of stars as they appear to an observer on Earth. The three precisely aligned stars, often described as "Orion's belt," are hundreds of light years from Earth, and also from each other. The perceived arrangement of the stars is undeniable, but the stars

are not aligned in the universe. Astronomical constellations like Orion are *intentional arrangements* imposed on our perceived locations of the stars, and these perceived arrangements and the explanations for them that constellations provide, form an Organizing System that is deeply embedded in human culture and in the practice of celestial navigation over the seas.

Not an Intentional Arrangement

The composition and arrangement of the rock layers ("strata") in the Grand Canyon in the Southwest United States have been studied extensively by geologists. The composition of rock suggests the environment in which it was formed, and the absolute and relative arrangement of the rock layers reveals the timing of important geological events.

(Photo by B. Rosen. Creative Commons CC BY-ND 2.0 license.)

Taken together, the intentional arrangements of resources in an Organizing System are the result of decisions about what is organized, why it is organized, how much it is organized, when it is organized, and how or by whom it is organized (each of these will be discussed in greater detail in Chapter 2). An Organizing System is defined by the composite impact of the choices made on these design dimensions. Because these questions are interrelated their answers come together in an integrated way to define an Organizing System.

1.6 The Concept of "Organizing Principle"

The arrangements of resources in an Organizing System follow or embody one or more organizing principles that enable the Organizing System to achieve its purposes. *Organizing principles* are directives for the design or arrangement of a *collection* of resources that are ideally expressed in a way that does not assume any particular implementation or realization. We call this design philosophy "Architectural Thinking" (§11.5.2.)

Organizing Spices By Cuisine

An alternative to organizing spices alphabetically is to organize them according to cuisines or flavor profiles, which can be defined in terms of ingredients and spices that tend to be used together. Patricia Glushko organizes her spices into three groups: Indian (includes cayenne pepper, coriander, cumin, turmeric), Mediterranean / Middle Eastern (includes basil, dill, oregano, paprika, thyme), and seeds. Each group of spices is in a separate large container, which makes it convenient when cooking.

(Photo by R. Glushko.)

When we organize a bookshelf, home office, kitchen, or the MP3 files on our music player, the resources themselves might be new and modern but many of the principles that govern their organization are those that have influenced the design of Organizing Systems for thousands of years. For example, we organize many collections of resources using the properties that are easiest to perceive, or whose values vary the most among the items in the collection, because these principles make it easy to locate a particular resource. We also group together resources that we often use together, we make resources that we use often more accessible than those we use infrequently, and we put rare or unique resources where we can protect them. Very general and abstract organizing principles are sometimes called design heuristics (e.g., "make things easier to find"). More specific and commonly used organizing principles include *alphabetical ordering* (arranging resources according to their names) and *chronological ordering* (arranging resources according to the date of their creation or other important event in the lifetime of the resource). Some organizing principles sort resources into pre-defined categories and other organizing principles rely on novel combinations of resource properties to create new categories.

Because this book was motivated by the goal of broadening the study of information organization beyond its roots in library and information science, it emphasizes organizing principles with a specific functional purpose like identifying, selecting, retrieving, or preserving resources. However, for thousands of years people have systematically collected things, information about those things, and observations of all kinds, organizing them in an effort to understand how their world works; the Babylonians created inventories and star charts; ancient Egyptians tracked the annual Nile floods; and, Mesoamericans created astronomical calendars. The term *sensemaking* is often used to describe this generic and less specific purpose of organizing to derive meaning from experience by fitting new events or observations into what they already know.[16[CogSci]]

Expressing organizing principles in a way that separates design and implementation aligns well with the three-tier architecture familiar to software architects and designers: user interface (implementation of interactions), business logic (intentional arrangement), and data (resources). (See the sidebar, The Three Tiers of Organizing Systems (page 47).)

The logical separation between organizing principles and their implementation is easy to see with digital resources. In a digital library it does not matter to a user if the resources are stored locally or retrieved over a network. The essence of a library Organizing System emerges from the resources that it organizes and the interactions with the resources that it enables. Users typically care a lot about the interactions they can perform, like the kinds of searching and sorting allowed by the online library catalog. How the resources and interactions are implemented are typically of little concern.

The separation of organizing principles and their implementation is harder to recognize in an Organizing System that only contains physical resources, such as your kitchen or clothes closet, where you appear to have unmediated interactions with resources rather than accessing them through some kind of user interface or "presentation tier" that supports the principles specified in the "middle tier" and realized in the "storage tier." As a result, people can easily get distracted by presentation-tier concerns. Too often we waste time color-coding file folders and putting labels on storage containers, when it would have better to think more carefully about the logical organization of the folder and container contents. It does not help to use colors and labels to make the logical organization more salient if that is not well designed first.

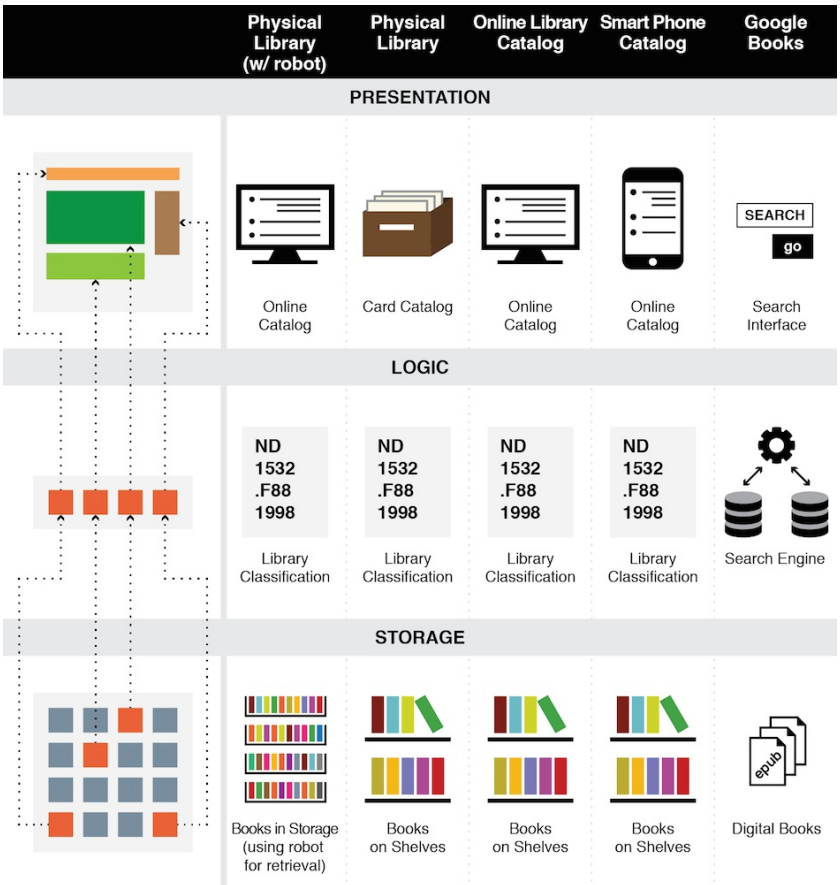
One place where you can easily appreciate these different tiers for physical resources is in the organization of spices in a kitchen. Different kitchens might all embody an *alphabetic order* organizing principle for arranging a collection of spices, but the exact locations and arrangement of the spices in any particular kitchen depends on the configuration of shelves and drawers, whether a spice rack or rotating tray is used, and other storage-tier considerations. Similarly, spices could be logically organized by cuisine, with Indian spices separated from Mexican spices, but this organizing principle does not imply anything about where they can be found in the kitchen.

Figure 1.2, Presentation, Logic and Storage Tiers. illustrates the separation of the presentation, logic, and storage tiers for four different types of library Organizing Systems and for Google Books. No two of them are the same in every tier. Note how a library that uses inventory robots to manage the storage of books does not reveal this in its higher tiers. (See the sidebar, Library Robot (page 127).)

Because tangible things can only be in one place at a time, many Organizing Systems, like those in the modern library with online catalogs and physical collections, resolve this constraint by creating digital proxies or surrogates to organize their tangible resources, or create parallel digital resources (e.g., digitized books).[17][Web] The implications for arranging, finding, using and reusing resources in any Organizing System directly reflect the mix of these two embodiments of information; in this way we can think of the modern library as a digital Organizing System that primarily relies on digital resources to organize a mixture of physical and digital ones.

The Organizing System for a small collection can sometimes use only the minimal or default organizing principle of *colocation*—putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box. If you do not cook much and have only a small number of spices in your kitchen, you do not need to alphabetize them because it is easy to find the one you want.[18][Com]

Figure 1.2. Presentation, Logic and Storage Tiers.



It is highly desirable when the design and implementation of an Organizing System separates the storage of the resources from the logic of their arrangement and the methods for interacting with them. This three-tier architect is familiar to designers of computerized Organizing Systems but it is also useful to think about Organizing Systems in this way even when it involves physical resources.

Some organization emerges implicitly through a *frequency of use* principle. In your kitchen or clothes closet, the resources you use most often migrate to the front because that is the easiest place to return them after using them. But as a collection grows in size, the time to arrange, locate, and retrieve a particular resource becomes more important. The collection must be explicitly organized to make these interactions efficient, and the organization must be preserved after the interaction takes place; i.e., resources are put back in the place they were

The Three Tiers of Organizing Systems

Software architects and designers agree that it is desirable to build applications that separate the storage of data, the business logic or functions that use the data, and the user interface or presentation components through which users or other applications interact with the data. This modular architecture allows each of the three tiers to be upgraded or reimplemented independently to satisfy changed requirements or to take advantage of new technologies. An analogous distinction is that between an algorithm as a logical description of a method for solving a computational problem and its implementation in a particular programming language like Java or Python.

These architectural distinctions are equally important to librarians and information scientists. Our new way of looking at Organizing Systems emphasizes the importance of identifying the desired interactions with resources, determining which organizing principles can enable the interactions, and then deciding how to store and manage the resources according to those principles. Applying architectural thinking to Organizing Systems makes it easier to compare and contrast existing ones and design new ones. Separating the organizing principles in the "middle tier" from their implications in the "data" and "presentation" tiers often makes it possible to implement the same logical Organizing System in different environments that support the same or equivalent interactions with the resources. For example, a new requirement to support searching through a library catalog on a smart phone would only affect the presentation tier.

found. As a result, most Organizing Systems employ organizing principles that make use of properties of the resources being organized (e.g., name, color, shape, date of creation, semantic or biological category), and multiple properties are often used simultaneously. For example, in your kitchen you might arrange your cooking pots and pans by size and shape so you can nest them and store them compactly, but you might also arrange things by cuisine or style and separate your grilling equipment from the wok and other items you use for making Chinese food.

Unlike those for physical resources, the most useful organizing properties for information resources are those that reflect their content and meaning, and these are not directly apparent when you look at a book, document, or collection of data. Significant intellectual effort or statistical computation is necessary to reveal these properties when assigning subject terms, creating an index, or using them as input features for machine learning and data analysis programs.

Separation Of Organizing Principle From Implementation

Whether spices are organized alphabetically by their names, by cuisines, by season, by frequency of use, or any other principle, this decision is logically distinct from the physical arrangement of the spices. There are many types of spice racks, shelves, circular "lazy susans," and other devices designed for arranging spices.

(Photo collage created by R. Glushko from various web catalogs.)

The most effective Organizing Systems for information resources often are based on statistical properties that emerge from analyzing the collection as a whole. For example, the relevance of documents to a search query is higher when they contain a higher than average frequency of the query terms compared to other documents in the collection, or when they are linked to relevant documents. Likewise, algorithms for classifying email messages continuously recalculate the probability that words like "beneficiary" or "Viagra" indicate whether a message is "spam" or "not spam" in the collection of messages processed.

1.7 The Concept of "Agent"

Many disciplines have specialized job titles to distinguish among the people who organize resources (for example: cataloger, archivist, indexer, curator, collections manager...). We use the more general word, *agent*, for any entity capable of autonomous and intentional organizing effort, because it treats organizing work done by people and organizing work done by computers as having common goals, despite obvious differences in methods.

We can analyze agents in Organizing Systems to understand how human and computational efforts to arrange resources complement and substitute for each other. We can determine the economic, social, and technological contexts in which each type of agent can best be employed. We can determine how the Organizing System allocates effort and costs among its creators, users, maintainers and other stakeholders.

A group of people can be an organizing *agent*, as when a group of people come together in a service club or standards body technical committee in which the members of the group subordinate their own individual agency to achieve a collective good.

We also use the term *agent* when we discuss interactions with Organizing Systems. The entities that most typically access the contents of libraries, museums, or other collections of physical resources are human agents—that is, people. In other Organizing Systems, such as business information systems or data repositories, interactions with resources are carried out by computational processes, robotic devices, or other entities that act autonomously on behalf of a person or group.

In some Organizing Systems, the resources themselves are capable of initiating interactions with other resources or with external agents. This is most obvious with human or other living resources, where a critical part of the design of any Organizing System with them is determining what kinds of interactions they should be encouraged or allowed to initiate. We will return to this issue after we discuss the design of interactions with ordinary resources that are passive, the situation in most Organizing Systems that involve physical resources.

Other resources that can initiate interactions are resources augmented with sensory, computational or communication capabilities that enable them to obtain information from their environment and then do something useful with it. You are probably familiar with RFID tags, which enable the precise identification and location of physical resources as they move through supply chains and stores, and with "smart" devices like Nest thermostats that learn how to program themselves.

1.8 The Concept of "Interactions"

An *interaction* is an action, function, service, or capability that makes use of the resources in a collection or the collection as a whole. The interaction of *access* is fundamental in any collection of resources, but many Organizing Systems provide additional functions to make access more efficient and to support additional interactions with the accessed resources. For example, libraries and similar Organizing Systems implement catalogs to enable interactions for *finding* a known resource, *identifying* any resource in the collection, and discriminating or *selecting* among similar resources.

Some of the interactions with resources in an Organizing System are inherently determined by the characteristics of the resource. Because many museum resources are unique or extremely valuable, visitors are allowed to view them but cannot borrow them, in contrast with most of the resources in libraries. A library might have multiple printed copies of *Moby Dick* but can never lend more of them than it possesses. After a printed book is checked out from the library, there are many types of interactions that might take place—reading, translating, summarizing, annotating, and so on—but these are not directly supported by the library Organizing System and are invisible to it.

For works not in the public domain, copyright law gives the copyright holder the right to prevent some uses, but at the same time "fair use" and similar copyright doctrines enable certain limited uses even for copyrighted works.[21][Law]

Digital resources enable a greater range of interactions than physical ones. Any number of people or processes can request a weather forecast from a web-based weather service because the forecast is not used up by the request and the marginal cost of allowing another access is nearly zero. Furthermore, with digital resources many new kinds of interactions can be enabled through application software, web services, or *application program interfaces (APIs)* in the Organizing System. In particular, translation, summarization, annotation, and keyword suggestion are highly useful services that are commonly supported by web search engines and other web applications. Similarly, an Organizing System with digital resources can implement a "keep everything up to date" interaction that automatically pushes current content to your browser.

But just as technology can enable interactions, it can prevent or constrain them. If your collection of digital resources (ebooks or music, for example) is not stored on your own computer or device, a continuous Internet connection is a requirement for access. In addition, access control policies and *digital rights management (DRM)* technology can limit the devices that can access the collection and prevent copying, annotation and other actions that might otherwise be enabled by the fair use doctrine.

Interaction design is especially crucial for managing resources that have the capability to initiate interactions with each other or with external agents. Consider the vast differences in how workers behave in businesses organized according to principles of scientific management and those that embody the Kaizen principles of continuous improvement. In the former, work is highly standardized and bureaucratic, giving workers little autonomy. In the latter, work is also standardized, but workers are motivated to analyze and improve work processes whenever possible, and they are given great discretion in how to do that.[22][Bus]

Just as with organizing principles, it is useful to think of interactions in an abstract or logical way that does not assume an implementation because it can encourage innovative designs for Organizing Systems.

1.9 The Concept of “Interaction Resource”

Interactions with physical resources sometimes leave traces or other evidence. Many of these traces are unintentional, like fingerprints, a coffee cup stain on a newspaper, or the erosion on a shortcut path across a lawn. Fans of *Sherlock Holmes* and *CSI* know that clever forensic investigators can use these residues of interactions to identify or vindicate suspects. Other interaction traces are intentional, like a student's yellow highlighting or notes in a textbook or spray-painted graffiti on a building. But not every interaction leaves a trace, traces fade over time, and different traces associated with the same resource lack consistency. This means that most traces are not of much use.

However, when Organizing Systems contain digital resources, or physical resources that have sensing, recording, or communication capabilities, interaction traces can be made predictable, persistent, and consistent. Each record of a user choice in accessing, browsing, buying, highlighting, linking, and other interactions then becomes an “interaction resource” that can be analyzed to reorganize the resource collection or otherwise influence subsequent interactions with the primary resources.

Interaction resources are often essential pieces of information that make Organizing Systems function. Most human toll-takers have been replaced by smart “toll tags” that broadcast their identity when the car they are in passes a radio receiver at a tolling location. Each interaction resource created identifies an account and credit card with which to pay the toll; taken together, the collection of these interaction resources can be used as the primary resources in other Organizing Systems that manage traffic congestion, or that support road design. Similarly, interaction resources created by search engines can be used to adjust the order of search hits, select ads, or personalize the content of web pages.

1.10 Organizing This Book

Devising concepts, methods, and technologies for describing and organizing resources have been essential human activities for millennia, evolving both in response to human needs and to enable new ones. Organizing Systems enabled the development of civilization, from agriculture and commerce to government and warfare. Today Organizing Systems are embedded in every domain of purposeful activity, including research, education, law, medicine, business, science, institutional memory, sociocultural memory, governance, public accountability, as well as in the ordinary acts of daily living.

With the World Wide Web and ubiquitous digital information, along with effectively unlimited processing, storage and communication capability, millions of people create and browse websites, blog, tag, tweet, and upload and download content of all media types without thinking "I am organizing now" or "I am retrieving now." Writing a book used to mean a long period of isolated work by an author followed by the publishing of a completed artifact, but today some books are continuously and iteratively written and published through the online interactions of authors and readers. When people use their smart phones to search the web or run applications, location information transmitted from their phone is used to filter and reorganize the information they retrieve. Arranging results to make them fit the user's location is a kind of computational curation, but because it takes place quickly and automatically we hardly notice it.

Likewise, almost every application that once seemed predominantly about information retrieval is now increasingly combined with activities and functions that most would consider to be information organization. Google, Microsoft, and other search engine operators have deployed millions of computers to analyze billions of web pages and millions of books and documents to enable the almost instantaneous retrieval of published or archival information. However, these firms increasingly augment this retrieval capability with information services that organize information in close to real-time. Further, the selection and presentation of search results, advertisements, and other information can be tailored for the person searching for information using his implicit or explicit preferences, location, or other *contextual information*.

Taken together, these innovations in technology and its application mean that the distinction between *information organization* and *information retrieval* that is often manifested in academic disciplines and curricula is much less important than it once was.

This book has few sharp divisions between *information organization* (IO) and *information retrieval* (IR) topics. Instead, it explains the key concepts and challenges in the design and deployment of Organizing Systems in a way that continuously emphasizes the relationships and tradeoffs between IO and IR. The concept of the Organizing System highlights the design dimensions and decisions that collectively determine the extent and nature of resource organization and the capabilities of the processes that compare, combine, transform and interact with the organized resources.

Navigating *The Discipline of Organizing*

Chapter 2, Design Decisions in Organizing Systems

This chapter introduces six broad design questions or dimensions whose intertwined answers define an Organizing System: What, why, how much, when, how, and where. This framework for describing and comparing Organizing Systems overcomes the biases and conservatism built into familiar categories like libraries and museums while enabling us to describe them as design patterns. We can then use these patterns to support inter-disciplinary work that cuts across categories and applies knowledge about familiar domains to unfamiliar ones.

Chapter 3, Activities in Organizing Systems

Developing a view that brings together how we organize as individuals with how libraries, museums, governments, research institutions, and businesses create Organizing Systems requires that we generalize the organizing concepts and methods from these different domains. Chapter 3 surveys a wide variety of Organizing Systems and describes four activities or functions shared by all of them: selecting resources, organizing resources, designing resource-based interactions and services, and maintaining resources over time.

Chapter 4, Resources in Organizing Systems

The design of an Organizing System is strongly shaped by what is being organized, the first of the six design decisions we introduced earlier in §2.2 What Is Being Organized? (page 64). To enable a broad perspective on this fundamental issue we use *resource* to refer to anything being organized, an abstraction that we can apply to physical things, digital things, information about either of them, or web-based services or objects. Chapter 4 discusses the challenges and methods for identifying the resources in an Organizing System in great detail and emphasizes how these decisions reflect the goals and interactions that must be supported—the "why" design decisions introduced in §2.3 Why Is It Being Organized? (page 66).

Chapter 5, Resource Description and Metadata

The principles by which resources are organized and the kinds of services and interactions that can be supported for them largely depend on the nature and explicitness of the resource descriptions. This "how much description" design question was introduced in §2.4 How Much Is It Being Organized? (page 70); Chapter 5 presents a systematic process for creating effective descriptions and analyzes how this general approach can be adapted for different types of Organizing Systems.

Chapter 6, Describing Relationships and Structures

An important aspect of organizing a collection of resources is describing the relationships between them. Chapter 6 introduces the specialized vocabulary used to describe semantic relationships between resources and between the concepts and words used in resource descriptions. It also discusses the structural relationships within multipart resources and between resources, like those expressed as citations or hypertext links.

Chapter 7, Categorization: Describing Resource Classes and Types

Groups or sets of resources with similar or identical descriptions can be treated as equivalent, making them members of an *equivalence class* or *category*. Identifying and using categories are essential human activities that take place automatically for perceptual categories like "red things" or "round things." Categorization is deeply ingrained in language and culture, and we use linguistic and cultural categories without realizing it, but categorization can also be a deeply analytic and cognitive process. Chapter 7 reviews theories of categorization from the point of view of how categories are created and used in Organizing Systems.

Chapter 8, Classification: Assigning Resources to Categories

The terms *categorization* and *classification* are often used interchangeably but they are not the same. Classification is applied categorization—the assignment of resources to a system of categories, called classes, using a predetermined set of principles. Chapter 8 discusses the broad range of how classifications are used in Organizing Systems. These include enumerative classification, faceted classification, activity-based classification, and computational classification. Because classification and standardization are closely related, the chapter also analyzes standards and standards-making as they apply to Organizing Systems.

Chapter 9, The Forms of Resource Descriptions

Chapter 9 complements the conceptual and methodological perspective on the creation of resource descriptions with an implementation perspective. Chapter 9 reviews a range of metamodels for structuring descriptions, with particular emphasis on XML, JSON, and RDF. It concludes by comparing and contrasting three "worlds of description" —document processing, the web, and the *Semantic Web*—where each of these three metamodels is most appropriate.

Chapter 10, Interactions with Resources

When Organizing Systems overlap, intersect, or are combined (temporarily or permanently), differences in resource descriptions can make it difficult or impossible to locate resources, access them, or otherwise impair their use. Chapter 10 reviews some of the great variety of concepts and techniques that different domains use when interacting with resources in Organizing Systems—integration, interoperability, data mapping, crosswalks, mash-ups, and so on. Interactions are characterized by the layers of resource properties they use: instance, collection-based, derived, or properties combined from different resources. Chapter 10 extends the idea of an information organization—information retrieval continuum, and describes information retrieval interactions (and others) in terms of information organization (i.e., resource description) requirements.

Chapter 11, The Organizing System Roadmap

Chapter 11 complements the descriptive perspective of chapters 2-10 with a more prescriptive one that analyzes the design choices and tradeoffs that must be made in different phases in an Organizing System's life cycle. System life cycle models exhibit great variety, but we use a generic four-phase model that distinguishes a domain identification and scoping phase, a requirements phase, a design and implementation phase, and an operational phase.

Chapter 12, Case Studies

In Chapter 12 we use the model described in Chapter 11 to guide the analysis of studies that span the range of Organizing Systems, and make reference to the principles, guidelines, vocabulary, and models discussed in the preceding chapters.

Endnotes for Chapter 1

[2][Com] (Glushko and McGrath 2005).

[4][DS] The DIKW hierarchy seems to have been inspired by *The Rock, A Pageant Play* (Eliot 1934) by the poet T S Eliot, whose opening chorus contains these lines:

Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

Most people credit Ackoff's *From Data to Wisdom* (Ackoff 1989) as the first articulation of the hierarchy in an information science and systems context. The hierarchy is mentioned in nearly twenty textbooks, but their close analysis by (Rowley 2007) reveals only partial agreement on the definitions and relationships among the four key concepts. The hierarchy has been criticized as lacking in philosophical rigor (Fricke 2009) and for ignoring the context-specificity of how knowledge is learned and applied (Jennex 2009). (Larose 2014)

[6][DS] Siegel's *Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die*" (Siegel 2013) is written for a non-technical audience and enthusiastically describes over 100 applications. *The Master Algorithm* (Domingos 2015) shares Siegel's enthusiasm but is far more technical; the book attempts to explain and compare the five "tribes" of machine learning: the symbolists, connectionists, evolutionaries, Bayesians, and analogizers. The title of Chris Anderson's provocative article in Wired Magazine (Anderson, 2008) is self-explanatory: "The end of theory: The data deluge makes the scientific method obsolete."

"Difference in kind or difference in degree" is an important issue in legal contexts and more generally arises whenever there is a disagreement about whether some difference or change is strict and categorical or whether it is incremental. We introduce it here so that readers can think critically about the socio-business-technical changes that might come about as a result of new methods and technologies for organizing and analyzing data. We believe that data science is on its way to becoming an important part of the organizing tool box. But everyone needs to remember that humans own the tool box, and that they design and build the tools..

[8][Web] The URI identifies a resource as an abstract entity that can have "multiple representations," which are the "things" that are actually exposed through applications or user interfaces. The HTTP protocol can transfer the representation that best satisfies the content properties specified by a web client, most often a browser. This means that interactions with web resources are always with their representations rather than directly with the resource *per se*. The representa-

tion of the resource might seem to be implied by the URI (as when it ends in `.htm` or `.html` to suggest text in *Hypertext Markup Language (HTML)* format), but the URI is not required to indicate anything about the "representation." A web resource can be a static web page, but it can also be dynamic content generated at the time of access by a program or service associated with the URI. Some resources like geolocations have "no representations at all;" the resource is simply some point or space and the interaction is "show me how to get there." The browser and web server can engage in "content negotiation" to determine which "representation" to retrieve, and this is particularly important when that format further requires an external application or "plug-in" in order for it to be rendered properly, as it does when the server returns a Power Point file or an other file format that is not built into the browser.

Internet architecture's definition of *resource* as a conceptual entity that is never directly interacted with is difficult for most people to apply when those resources are physical or tangible objects, because then it surely seems like we are interacting with something real. So we will most often talk about interactions with resources, and will mention "resource representations" only when it is necessary to align precisely with the narrower Internet architecture sense.

[9][Phil] In addition, groups of people have come together to form "intentional communities" for thousands of years in monasteries, communes, artist colonies, cooperative houses, and religious or ethnic enclaves so they can live with people who share their values and beliefs. A directory of intentional communities organized by type and location is managed by the Fellowship of Intentional Communities.

[10][Bus] The shift from a manufacturing to an information and services economy in the last few decades has resulted in greater emphasis on intellectual resources represented in skills and knowledge rather than on the natural resources of production materials and physical goods.

The intellectual resources of a firm are embodied in a firm's people, systems, management techniques, history of strategy and design decisions, customer relationships, and intellectual property like patents, copyrights, trademarks, and brands. Some of this knowledge is explicit, tangible, and traceable in the form of documents, databases, organization charts, and policy and procedure manuals. But much of it is tacit: informal and not systematized in tangible form because it is held in the minds and experiences of people; a synonym is "know-how." A more modern term is *Intellectual Capital,* a concept originated in a 1997 book with that title (Stewart 1997).

[11][Law] In 2004, Google began digitizing millions of books from several major research libraries with the goal of making them available through its search engine (Brin 2009). But many millions of these books are still in copyright, and in 2005 Google was sued for copyright infringement by several publishers and an

author's organization. In 2011 a US District Court judge rejected the proposed settlement the parties had negotiated in 2008 because many others objected to it, including the US Justice Department, several foreign governments, and numerous individuals (Samuelson 2011).

The major reason for the rejection was that the settlement was a "bridge too far" that went beyond the claims made against Google to address issues that were not in litigation. In particular, the judge objected to the treatment of the so-called "orphan works" that were still under copyright but out of print because money they generated went to the parties in the settlement and not to the rights holders who could not be located (why the books are "orphans") or to defray the costs of subscriptions to the digital book collection. The judge also was concerned that the settlement did not adequately address the concerns of academic authors—who wrote most of the books scanned from research libraries—who might prefer to make their books freely available rather than seek to maximize profits from them. Other concerns were that the settlement would have entrenched Google's monopoly in the search market and that there were inadequate controls for protecting the privacy of readers.

Google's plan would have dramatically increased access to out of print books, and the rejection of the proposed settlement has heightened calls for an open public digital library (Darnton 2011). A good start toward such a library was the digital copies that the research libraries received in return for giving Google books to scan, which were collected and organized by the Hathi Trust In 2010, the Alfred P. Sloan Foundation provided funding to launch the *Digital Public Library of America (DPLA)*: `http://dp.la/`. This non-proprietary goal might induce the US Congress and other governments to pass legislation that fixes the copyright problems for orphan works.

[12][Com] Self-organizing is also used to describe phenomena like climate, neural networks, and phase transitions and equilibrium states in physics and chemistry. But when systems involve collections of *inanimate* resources that are very large and open, with complex interactions among the resources, it seems less sensible to attribute intentional arrangement to the outcomes. The resource arrangements that emerge cannot always be interpreted as the result of intentional or deterministic principles and instead are more often described in probabilistic or statistical terms. And even though it involves animate resources, Charles Darwin's "natural selection" in evolutionary biology is a self-organizing mechanism where intentionality is hard to pinpoint or absent entirely.

The rules governing these local interactions can be simple and yet produce highly complex structures. For example, in flocks of birds or schools of fish the rules are: (1) follow things like you, (2) do not bump into each other, but stay close, and (3) move in the same direction as the rest of the group. With just these three rules computer models can create complex three-dimensional ar-

rangements that can make abrupt changes in shape and density while moving rapidly, just as live things do. (Friederici 2009)

The term "Crowdsourcing" was invented by Jeff Howe in a June 2006 article in Wired magazine, and the concept was developed further in a book published two years later (Howe 2006, 2008). "Folksonomy" was coined by Thomas Van der Wal at about the same time in 2004; see `http://vanderwal.net/folksonomy.html` and (Trant 2009).

(Goldstone and Gureckis 2009) present a cognitive science perspective on collective behavior, analyze important themes and controversies, and suggest areas for future research. (Moussaid et al. 2009) analyze self-organizing phenomena in animal swarms and human crowds in terms of information exchange among individuals.

Self-organizing behaviors in ants, bees, bats, cuckoos, fireflies and other animals have been analyzed to identify heuristics that can be applied to difficult optimization problems in network design, cryptography, and other domains where deterministic algorithms are infeasible. (Yang 2010)

(Smith 1776)

[13][Web] (Banzhaf 2009).

[14][IA] The concept of a web page is imprecise because many web pages, especially home pages designed as navigation gateways to an organized collection of pages, are constructed from heterogeneous blocks of content that could have been organized as separate pages.

[15][Web] The "plain web" (Wilde 2008a), whose evolution is managed by the *World Wide Web Consortium (W3C)*, is rigorously standardized, but unfortunately the larger ecosystem of technologies and formats in which the web exists is becoming less so. Web-based Organizing Systems often contain proprietary media formats and players (like Flash) or are implemented as closed environments that are intentionally isolated from the rest of the web (like Facebook or Apple's iTunes and other smart phone "app stores").

[16][CogSci] (Weick et al , 2005 p. 410).

[17][Web] Instead of thinking of a digital book as a "parallel resource" to a printed book, we could consider both of them as alternate representations of the same abstract resource that are linked together by an "alternative" relationship, just as we can use the HTML `ALT` tag to associate text with an image so its content and function can be understood by text-only readers.

[18][Com] For collections of non-trivial size the choice of searching or sorting algorithm in computer programs is a critical design decision because they differ greatly in the time they take to complete and the storage space they require.

For example, if the collection is arranged in an unorganized or random manner (as a "pile") and every resource must be examined, the time to find a particular item increases linearly with the collection size. If the collection is maintained in an ordered manner, a binary search algorithm can locate any item in a time proportional to the logarithm of the number of items. Analysis of algorithms is a fundamental topic in computer science; a popular textbook is *Introduction to Algorithms* by (Cormen et al. 2009).

[21][Law] Copyright law, license or contract agreements, terms of use and so on that shape interactions with resources are part of the Organizing System, but compliance with them might not be directly implemented as part of the system. With digital resources, digital rights management (DRM), passwords, and other security mechanisms can be built into the Organizing System to enforce compliance.

[22][Bus] Frederick Taylor developed "scientific management" to improve industrial efficiency and conducted detailed time and motion studies to devise what he thought were optimal ways to perform work tasks (Taylor 1914). The Kaizen principles of continuous improvement were introduced to Western audiences by Imai Masaaki and by numerous books about their application in the Toyota production system (Masaaki 1986).

Scientific management views a business as a machine, while Kaizen principles treat it as a brain that learns. These metaphors for business organization are among those described by (Morgan 1997) in a classic business textbook. Other metaphors discussed include organisms, cultures, political systems, and psychic prisons.

Chapter 2

Design Decisions in Organizing Systems

Robert J. Glushko

2.1 Introduction

A set of resources is transformed by an organizing system when the resources are described or arranged to enable interactions with them. Explicitly or by default, this requires many interdependent decisions about the identities of resources; their names, descriptions and other properties; the classes, relations, structures and collections in which they participate; and the people or technologies interacting with them.

One important contribution of the idea of the organizing system is that it moves beyond the debate about the definitions of "things," "documents," and "information," with the unifying concept of *"resource"* while acknowledging that "what is being organized" is just one of the questions or dimensions that need to be considered. These decisions are deeply intertwined, but it is easier to introduce them as if they were independent.

We introduce six groups of design questions, itemizing the most important dimensions in each group:

- **What is being organized?** What is the scope and scale of the domain? What is the mixture of physical things, digital things, and information about things in the organizing system? Is the organizing system being designed to create a new resource collection, catalog an existing and closed resource collection, or manage a collection in which resources are continually added or deleted? Are the resources unique, or are they interchangeable members of a category? Do they follow a predictable "life cycle" with a "useful life"? Does the organizing system use the interaction resources created through its use, or are these interaction resources extracted and aggregated for use by another organizing system? (§2.2)
- **Why is it being organized?** What interactions or services will be supported, and for whom? Are the uses and users known or unknown? Are the users primarily people or computational processes? Does the organizing system need to satisfy personal, social, or institutional goals? (§2.3)
- **How much is it being organized?** What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed? What organizing principles guide the organization? Are all resources organized to the same degree, or is the organization sparse and non-uniform? (§2.4)
- **When is it being organized?** Is the organization imposed on resources when they are created, when they become part of the collection, when interactions occur with them, just in case, just in time, all the time? Is any of this organizing mandated by law or shaped by industry practices or cultural tradition? (§2.5)
- **How or by whom, or by what computational processes, is it being organized?** Is the organization being performed by individuals, by informal groups, by formal groups, by professionals, by automated methods? Are the organizers also the users? Are there rules or roles that govern the organizing activities of different individuals or groups? (§2.6)
- **Where is it being organized?** Is the resource location constrained by design or by regulation? Are the resources positioned in a static location? Are the resources in transit or in motion? Does their location depend on other parameters, such as time? (§2.7)

How well these decisions coalesce in an organizing system depends on the requirements and goals of its human and computational users, and on understanding the constraints and tradeoffs that any set of requirements and goals impose. How and when these constraints and tradeoffs are handled can depend on the legal, business, and technological contexts in which the organizing system is designed and deployed; on the relationship between the designers and users of the organizing system (who may be the same people or different ones); on the

economic or emotional or societal purpose of the organizing system; and on numerous other design, deployment, and use factors.

Classifying organizing systems according to the kind of resources they contain is the most obvious and traditional approach. We can also classify organizing systems by their dominant purposes, by their intended user community, or other ways. No single fixed set of categories is sufficient by itself to capture the commonalities and contrasts between organizing systems.

We can augment the categorical view of organizing systems by thinking of them as existing in a multi-faceted or multi-dimensional design space in which we can consider many types of collections at the same time.

This *framework* for describing and comparing organizing systems overcomes some of the biases and conservatism built into familiar categories like libraries, museums, and archives, while enabling us to describe them as design patterns that embody characteristic configurations of design choices. We can then use these patterns to support inter-disciplinary work that cuts across categories and applies knowledge about familiar domains to unfamiliar ones. A dimensional perspective makes it easier to translate between category- and discipline-specific vocabularies so that people from different disciplines can have mutually intelligible discussions about their organizing activities. They might realize that they have much in common, and they might be working on similar or even the same problems.

A faceted or dimensional perspective acknowledges the diversity of instances of collection types and provides a generative, forward-looking framework for describing hybrid types that do not cleanly fit into the familiar categories. Even though it might differ from the conventional categories on some dimensions, an organizing system can be designed and understood by its *family resemblance* on the basis of its similarities on other dimensions to a familiar type of resource collection.

Thinking of organizing systems as points or regions in a design space makes it easier to invent new or more specialized types of collections and their associated interactions. If we think metaphorically of this design space as a map of organizing systems, the empty regions or “white space” between the densely-populated centers of the traditional categories represent organizing systems that do not yet exist. We can consider the properties of an organizing system that could occupy that white space and analyze the technology, process, or policy innovations that might be required to let us build it there. We can reason by analogy to identify and apply the principles used in one organizing system to understand or design others.

2.2 What Is Being Organized?

> *"What is difficult to identify is difficult to describe and therefore difficult to organize."*
>
> — (Svenonius 2000, p. 13)

Before we can begin to organize any resource we often need to identify it. It might seem straightforward to devise an organizing system around tangible resources, but we must be careful not to assume what a resource is. In different situations, the same "thing" can be treated as a unique item, one of many equivalent members of a broad category, or a component of an item rather than as an item on its own. For example, in a museum collection, a handmade, carved chess piece might be a separately identified item, identified as part of a set of carved chess pieces, or treated as one of the 33 unidentified components of an item identified as a chess set (including the board). When merchants assign a stock-keeping unit (SKU) to identify the things they sell, that SKU can be associated with a unique item, sets of items treated as equivalent for inventory or billing purposes, or intangible things like warranties.

You probably do not have explicit labels on the cabinets and drawers in your kitchen or clothes closet, but department stores and warehouses have signs in the aisles and on the shelves because of the larger number of things a store needs to organize. As a collection of resources grows, it often becomes necessary to identify each one explicitly; to create surrogates like bibliographic records or descriptions that distinguish one resource from another; and to create additional organizational mechanisms like shelf labels, store directories, library card catalogs and indexes that facilitate understanding the collection and locating the resources it contains. These organizational mechanisms often suggest or parallel the organizing principles used to organize the collection itself.

Organization mechanisms like aisle signs, store directories and library card catalogs are embedded in the same physical environment as the resources being organized. But when these mechanisms or surrogates are digitized, the new capabilities that they enable create design challenges. This is because a digital organizing system can be designed and operated according to more abstract and less constraining principles than an organizing system that only contains physical resources. A single physical resource can only be in one place at a time, and interactions with it are constrained by its size, location, and other properties. In contrast, digital copies and surrogates can exist in many places at once and enable searching, sorting, and other interactions with an efficiency and scale impossible for tangible things.

When the resources being organized consist of information content, deciding on the unit of organization is challenging because it might be necessary to look be-

yond physical properties and consider conceptual or intellectual equivalence. A high school student told to study Shakespeare's play *Macbeth* might treat any printed copy or web version as equivalent, and might even try to outwit the teacher by watching a film adaptation of the play. To the student, all versions of *Macbeth* seem to be the same resource, but librarians and scholars make much finer distinctions.

An increasing number of organizing systems handle resources that are born digital. Ideally, digital texts can be encoded with explicit markup that captures structural boundaries and content distinctions, which can be used to facilitate organization, retrieval, or both. In practice the digital representations of texts are often just image scans that do not support much processing or interaction. A similar situation exists for the digital representations of music, photographs, videos, and other non-text content like sensor data, where the digital formats are structurally and semantically opaque.

This book does not emphasize systems that organize people, but it would be remiss not to mention them. Businesses organize their employees, schools organize their faculties and students, sports leagues and teams organize their players, and governments organize their citizens and residents to enable them to vote, drive, attend schools, and receive medical care and other benefits. Data scientists in all of these fields increasingly predict how employees, students, athletes, voters, drivers - and other categories of people defined by intrinsic or derived characteristics - will behave, decide, live, or die. Once people die, it is no longer necessary to predict anything about them, but nonetheless cemeteries are highly organized.

We often think and talk about time as a resource, and time fits the definition of "anything of value that supports goal-oriented activity" from §1.3. Furthermore, we could think of the calendar and clock as organizing systems that define time at different levels of granularity to support different kinds of interactions. However, it is probably more useful to think of time as a constraint that influences how and how much to organize.

If you're sorting your own mail, you can question whether the time you spend on sorting is worth the time you save on searching. But at scale—imagine 10 million books in a library—the considerable effort required to organize resources saves vastly more time for the many users of the system over its lifetime. Note the inherent tradeoff between time spent on organizing versus retrieval; this will be a recurring theme throughout this book. In a personal context the tradeoff is a matter of individual need or preference, but in social or institutional contexts organization and retrieval are generally done by different people, and their time is likely valued in different ways by the system owner.

Computational Descriptions of People

Each of us is associated with a great many computational descriptions, some of which are used almost every day to make predictions about our behavior using a variety of statistical techniques that are collectively called "predictive analytics." Whenever you use a credit card, fraud detection algorithms use a model derived from your purchase history to decide, in fractions of a second, whether the transaction is being initiated by you, or by someone who has stolen your card. When you want to buy something expensive on credit, the seller consults your credit score—based on what you owe, your payment history, how long you have had credit, the kinds of credit you have, and other factors—to predict whether you are a good credit risk, and your credit score then gets adjusted if the seller decides to give you credit. Then, after you have bought that expensive item, the seller's predictive model can use that information to suggest other things you might want to buy.

Philosophers have long debated the extent to which observations of a person's behavior can yield an understanding of their true and unobservable nature. But whether or not computational descriptions capture a person's essence, there is no escaping them. If you want to get life or car insurance or a mortgage, models determine what you have to pay. Predictive models are being used to admit people to college, to hire them, to draft or trade them in professional sports, and to decide whether to monitor them closely because they might be planning a terrorist act. Some companies use "people analytics" software that analyzes every email, calendar item, and document created by employees to build a model of what they know, what they do, when they do it, and who they work with—the goal being to improve communication and collaboration within the firm and with customers.

2.3 Why Is It Being Organized?

> *"The central purpose of systems for organizing information [is] bringing like things together and differentiating among them."*
>
> — (Svenonius 2000 p. xi)

Almost by definition, the essential purpose of any organizing system is to describe or arrange resources so they can be located and accessed later. The organizing principles needed to achieve this goal depend on the types of resources or domains being organized, and in the personal, social, or institutional setting in which organization takes place.

Organizing systems can be distinguished by their dominant purposes or the priority of their common purposes. Libraries, museums, and archives are often

classified as *memory institutions* to emphasize their primary emphasis on resource preservation. In contrast, "management information systems" or "business systems" are categories that include the great variety of software applications that implement the organizing systems needed to carry out day-to-day business operations.

"Bringing like things together" is an informal organizing principle for many organizing systems. Almost as soon as libraries were invented over two thousand years ago, the earliest librarians saw the need to develop systematic methods for arranging and inventorying their collections. The invention of mechanized printing in the fifteenth century, which radically increased the number of books and periodicals, forced libraries to begin progressively more refined efforts to state the functional requirements for their organizing systems and to be explicit about how they met those requirements.

Today, any information-driven enterprise must have systematic processes and technologies in place that govern information creation or capture and then manage its entire life cycle. Commercial firms need processes for transacting with customers or other firms to carry out business operations, to support research and innovation, marketing, and to develop business strategy and tactics in compliance with laws and regulations for accounting, taxes, human resources, data retention, and so on. In large firms these functions are so highly specialized and complex that the different types of organizing systems have distinct names: *Enterprise Resource Planning (ERP)*, *Enterprise Content Management (ECM)*, *Enterprise Data Management (EDM) Supply Chain Management (SCM)*, *Records Management*, *Customer Relationship Management (CRM)*, *Business Intelligence (BI)*, *Knowledge Management (KM)*, and so on. And even though the most important functions in the organizing systems of large enterprises are those that manage the information resources needed for its business operation, these firms might also need to maintain corporate libraries and archives.

Preserving documents in their physical or original form is the primary purpose of archives and similar organizing systems that contain culturally, historically, or economically significant documents that have value as long-term evidence. Preservation is also an important motivation for the organizing systems of information- and knowledge-intensive firms, where information is primarily in digital formats. Businesses and governmental agencies are usually required by law to keep records of financial transactions, decision-making, personnel matters, and other information essential to business continuity, compliance with regulations and legal procedures, and transparency. As with archives, it is sometimes critical that these business knowledge or records management systems can retrieve the original documents, although digital copies that can be authenticated are increasingly being accepted as legally equivalent.

This discussion of the requirements for organizing resources in memory institutions and businesses might convey the impression that storing and retrieving resources efficiently are paramount goals, and indeed they are in many contexts. But there are many other reasons for organizing resources, as is easily seen when we look at personal organizing systems. And there are many other ways to compare organizing systems than just how efficiently they enable storing and retrieval functions.

An overarching goal when people are organizing their personal resources is to minimize the effort needed to find the resources. But unlike the finding task in institutional organizing systems, which is generally facilitated with external resource descriptions, finding aids, classifications, search engines, and orientation and navigation mechanisms, the finding task in personal organizing systems is primarily a cognitive one: you need to remember where the resources are and how they are arranged. Because each person has unique experiences and preferences, it is not surprising that people often organize the same types of resources in different ways to make the organization easier to perceive and remember. The resulting resource arrangements often emphasize aesthetic or emotional goals, as when books or clothes are arranged by color or preference, or behavioral goals, as when most frequently used condiments and spices are kept on the kitchen counter rather than stored in a pantry.

When individuals manage their papers, books, documents, record albums, compact discs, DVDs, and other information resources, their organizing systems can vary greatly. This is in part because the content of the resources being organized becomes a consideration. Furthermore, many of the organizing systems used by individuals are implemented by web applications, and this makes them more accessible than physical resources.[27][Web]

Put another way, an information resource inherently has more potential uses than resources like forks or frying pans, so it is not surprising that the organizing systems in offices are even more diverse than those in kitchens.

When the scale of the collection or the number of intended users increases, two things can happen. The first is that if the system can turn its interaction traces into interaction resources, additional value can be created by analyzing these resources to enhance the interactions, to suggest new ones, or make predictions about how individual users or groups of them will behave. Every business that has a high volume of customer transactions does this; for example, a fast-food restaurant would analyze time-stamped sales data, and might introduce a quick pickup line for items that sell the most, or create product bundles that increase sales while optimizing kitchen and counter work. Amazon.com and other retailers that can capture detailed browsing traces can augment the sales data they collect by treating items that were looked at but not purchased as potential

transactions, making them additional inputs to their sophisticated pricing and recommendation systems.

A second likely outcome of increased scale or use is that not everyone is likely to share the same goals and design preferences for the organizing system. If you share a kitchen with housemates, you might have to negotiate and compromise on some of the decisions about how the kitchen is organized so you can all get along. In more formal or institutional organizing systems conflicts between stakeholders can be much more severe, and the organizing principles and policies or permissions for the kinds of interactions available to different users might even be specified in commercial contracts or governed by laws or standards. For example, Bowker and Star note that physicians view the creation of patient records as central to diagnosis and treatment, insurance companies think of them as evidence needed for payment and reimbursement, and researchers think of them as primary data. These groups do not agree on the priority and quality requirements they assign to different information in the patient record, and physicians understandably resist doing work that has no direct benefit for them. Not surprisingly, policy making and regulations about patient records are highly contentious.

Once we acknowledge that stakeholders might not share the same goals, it is clear that efficiency is too narrow a measure for evaluating organizing systems. The ways that resources are organized and interacted with embody the priorities and values of those designing the organizing system, yielding arrangements and interactions designed to control or change the behaviors of the users. Put more bluntly, resources are always organized in ways that are designed to allocate value for some people (e.g., the owners of the resources, or the most frequent users of them) and not for others. From the perspective of the other types of user trying to interact with the system, this organization will likely seem unfair. In this way, organizing resources can often be seen as creating winners and losers, providing benefits to the former and imposing costs or constraints on the latter. For example, search engines analyze interaction resources to adjust search results and choose an ad that is related to your latest query. These are considered improved interactions from the perspective of the search engine, but you might consider it a violation of your privacy and a bit creepy to have the targeted ad follow you around the web until you click on it.

The emerging field of applied behavioral economics, popularized in books like *Freakonomics* and *Nudge,* explains how subtle differences in resource arrangement, the number and framing of choices, and default values can have substantial effects on the decisions people make. Consider the arrangement of salads, pasta dishes, bread, fish, meat, desserts and other types of food in a self-serve cafeteria buffet. In a school setting, the food might be organized and presented to encourage healthier eating, perhaps by making the fatty french fries and high-calorie desserts hard to reach or by providing smaller trays and plates. The

same foods would likely be organized differently in an all-you-can-eat restaurant, where the goal is to minimize food costs, with less expensive items like salads at the front of the line to ensure that trays and plates will already be full when the customer gets to the more expensive items at the end of the line.[29][Bus]

The organization of cafeteria buffets to shape user behavior might not seem sinister. However, organizing systems can control behavior in ways that create or perpetuate inequities among their users. This unfairness is a matter of degree: a person who does not own a computer who goes to the public library to check out a popular book loses out when the library enables patrons with computer access to check out books online and assumes that everyone has an equal shot at accessing books via the Internet.

Looking to a much more insidious organizing system, when the South African government adopted Apartheid policies to classify and segregate people by race, it systematized economic and political discrimination and great suffering for the nonwhite population. (See the sidebar, Power and Politics in Organizing (page 71).)

Chapter 8, *Classification: Assigning Resources to Categories* more fully explains the different purposes for organizing systems, the organizing principles they embody, and the methods for assigning resources to categories.

2.4 How Much Is It Being Organized?

> *"It is a general bibliographic truth that not all documents should be accorded the same degree of organization."*
>
> — (Svenonius 2000, p. 24)

Not all resources should be accorded the same degree of organization. In this section we will briefly unpack this notion of degree of organization into three important and related dimensions: the amount of description detail or organization applied to each resource, the amount of organization of resources into classes or categories, and the overall extent to which interactions in and between organizing systems are shaped by resource description and arrangement.

It is important to note that this section is **not** asking the question "how much stuff is being organized?" but rather to what degree is the stuff being organized. Another way to ask the same question is "how many organizing principles are at work?" in this organizing system. Your closet might be arranged only by body part covered and season; an online music store will organize resources by genre, artist name, band name, album name, popularity, date released, and maybe others. So we would say that the online music store is organized much more than the closet, because more organizing principles are at work.

Power and Politics in Organizing

It is tempting to think of organizing systems and the technologies used to implement them as neutral or objective in their goals and impacts, but it is impossible to argue that the use of racial classification in apartheid South Africa was not a conscious manifestation of prejudice. And even if making it hard for school kids to find the junk food in the cafeteria buffet has health benefits, it nevertheless reflects a paternalistic point of view that restricts individual choices.

Organizing systems and technology are not developed in a vacuum, unencumbered by politics or social context. As Langdon Winner underscores in *Do Artifacts Have Politics?*, systems and technologies can be conscious manifestations of the personal (and often political) biases of their creators. Because all people have different experiences and biases, even when they are not conscious of them they influence the design and implementation of organizing systems in ways that can create or perpetuate inequalities.[30][Phil]

Technology innovators whose expressed goals are to make something faster, smaller, or cheaper are ignoring the potential for their innovations and automation to render certain types of work less viable and discriminate against people who lack the technology or skills to use it. For example, Winner describes the inadvertent social and political consequences of the introduction of mechanical tomato harvesters in California agriculture in the 1960s. Their industry-wide adoption favored larger farms with more resources to buy the expensive machines, resulting in the disappearance of small tomato farms and large-scale changes to many rural communities whose economies had relied on them.

Some may argue that the mechanical tomato harvester created massive benefits by increasing productivity, but the determination that more efficient tomato production is worth its consequences could be debated. In any case, the debate cannot be answered with a definite yes or no, just as it cannot be with whether the Internet is bad because it has eroded the need for librarians, or whether Uber's clever technologies for matching drivers and riders unfairly avoid the regulations imposed on the taxi industry. Affirming the introduction of the mechanical tomato harvester, search engines, and Uber in the name of productivity, progress, and efficiency is a political point of view.

(See also §8.2.3 Classification Is Biased (page 408) and Chapter 11, *The Organizing System Roadmap*.)

(Chapter 5 and Chapter 7, more thoroughly address these questions about the nature and extent of description in organizing systems.)

Not all resources in a collection require the same degree of description for the simple reason we discussed in §2.3 Why Is It Being Organized? (page 66): Organizing systems exist for different purposes and to support different kinds of interactions or functions. Let us contrast two ends of the "degree of description" continuum. Many people use "current events awareness" or "news feed" applications that select news stories whose titles or abstracts contain one or more keywords (Google Alert is a good example). This exact match algorithm is easy to implement, but its all-or-none and one-item-at-a-time comparison misses any stories that use synonyms of the keyword, that are written in languages different from that of the keyword, or that are otherwise relevant but do not contain the exact keyword in the limited part of the document that is scanned. However, users with current events awareness goals do not need to see every news story about some event, and this limited amount of description for each story and the simple method of comparing descriptions are sufficient.

On the other hand, this simple organizing system is inadequate for the purpose of comprehensive retrieval of all documents that relate to some concept, event, or problem. This is a critical task for scholars, scientists, inventors, physicians, attorneys and similar professionals who might need to discover every relevant document in some domain. Instead, this type of organizing system needs rich bibliographic and semantic description of each document, most likely assigned by professional catalogers, and probably using terms from a *controlled vocabulary* to enforce consistency in what descriptions mean.

Similarly, different merchants or firms might make different decisions about the extent or granularity of description when they assign SKUs because of differences in suppliers, targeted customers, or other business strategies. If you take your car to the repair shop because windshield wiper fluid is leaking, you might be dismayed to find that the broken rubber seal that is causing the leak cannot be ordered separately and you have to pay to replace the "wiper fluid reservoir" for which the seal is a minor but vital part. Likewise, when two business applications try to exchange and merge customer information, integration problems arise if one describes a customer as a single "NAME" component while the other separates the customer's name into "TITLE," "FIRSTNAME," and "LASTNAME."

Even when faced with the same collection of resources, people differ in how much organization they prefer or how much disorganization they can tolerate. A classic study by Tom Malone of how people organize their office workspaces and desks contrasted the strategies and methods of "filers" and "pilers." Filers maintain clean desktops and systematically organize their papers into categories, while pilers have messy work areas and make few attempts at organization. This contrast has analogues in other organizing systems and we can easily imagine what happens if a "neat freak" and "slob" become roommates.[31][CogSci]

An equally wide range, from a little organization to a lot, can be seen in the organizing systems for businesses, armies, governments, or any other institutional organizing systems for people. Organizations with broad scope and many people usually have deep hierarchies and explicit reporting relationships with the CEO, general, or president at the top with numerous layers of vice presidents, directors, department heads, and managers (or colonels, majors, captains, lieutenants, and sergeants). Smaller organizations are more varied, with some embodying multi-layered management, and some embracing a flatter arrangement with fewer management levels, wider spans of authority, and more autonomy for individual workers. Many start-up firms try to grow without any management structure at all in the belief that it makes them more innovative and nimble, but evidence suggests that when no one is responsible for making decisions, the lack of accountability results in poor decisions, or in no decisions at all even when some were sorely needed.[32][Bus]

In any case, when people have to do it, describing and organizing resources is work. Stakeholders in an organizing system often have disagreements among about how much organization is necessary because of the implications for who performs the work and who derives the benefits, especially the economic ones. Physicians prefer narrative descriptions and broad classification systems because they make it easier to create patient notes. In contrast, insurance companies and researchers want fine-grained "form-filling" descriptions and detailed classifications that would make the physician's work more onerous.[33][Com]

The amount of resource description is always shaped by the currently available technology for capturing, storing, and making use of it. Nineteenth century geologists and paleontologists typically recorded only general information about the depth and surrounding geological features when they found fossils because they had no technology for making more precise measurements and everything they noted they had to record by hand. Today, vastly more detailed information is recorded by instruments and exploited by sophisticated techniques for carbon dating and 3D reconstruction.

Automatically generated descriptions are increasingly an alternative or complement to those created by people. "Smart" resources use sensors to capture information about themselves and their environments (see §4.3.4). Our own computers and phones record information about our keystrokes, clicks, communications, and locations. Business and government computers analyze and index most of the text and speech content that flows through and between our personal phones and computers. These indexes typically assign weights to the terms according to calculations that consider the frequency and distribution of the terms in both individual documents and in the collection as a whole to create a description of what the documents are about. These descriptions of the documents in the collection are more consistent than those created by human organizers. They allow for more complex query processing and comparison operations

by the retrieval functions in the organizing system. For example, query expansion mechanisms can automatically add synonyms and related terms to the search. Additionally, retrieved documents can be arranged by relevance, while "citing" and "cited-by" links can be analyzed to find related relevant documents.

It is important to recognize the potential downside to automated resource description. A detailed description produced by sensors or computers can seem more accurate or authoritative than a simpler one created by a human observer, even if the latter would be more useful for the intended purposes. Moreover, the more detailed the description, the greater the opportunity to use it for new purposes. This might be desirable, as when a company realizes that it can cross- and up-sell because it has been tracking every click in a web store to create a collection of interaction resources. But it could be undesirable, because detailed transaction data can be used to violate privacy and civil rights. It depends on who controls the collected information and their incentives for using it or not using it.

A second constraint on the degree of organization comes from the size of the collection within the scope of the organizing system. Organizing more resources requires more descriptions to distinguish any particular resource from the rest, and more constraining organizing principles. Similar resources need to be grouped or classified to emphasize the most important distinctions among the complete set of resources in the collection. A small neighborhood restaurant might have a short wine list with just ten wines, arranged in two categories for "red" and "white" and described only by the wine's name and price. In contrast, a gourmet restaurant might have hundreds of wines in its wine list, which would subdivide its "red" and "white" high-level categories into subcategories for country, region of origin, and grape varietal. The description for each wine might in addition include a specific vineyard from which the grapes were sourced, the vintage year, ratings of the wine, and tasting notes.

At some point a collection grows so large that it is not economically feasible for people to create bibliographic descriptions or to classify each separate resource, unless there are so many users of the collection that their aggregated effort is comparably large; this is organizing by "crowdsourcing." This leaves two approaches that can be done separately or in tandem.

- The simpler approach is to describe sets of resources or documents as a set or group.
- The second approach is to rely on automated and more general-purpose organizing technologies that organize resources through computational means. Search engines are familiar examples of computational organizing technology, and §8.6 Computational Classification (page 427) describes other common techniques in machine learning, clustering, and discriminant

Using "Information Theory" to Quantify Organization

We often hear news stories hyping "how much information" there is in the information society with breathless exuberance about the creation of peta-, exa-, whatever-bytes of content. A much more important and intellectually deeper question than absolute size in bytes is measuring how much information is encoded in the structure or organization of a system. For this we can turn to "Information Theory," a formal approach to understanding the theoretical maximum amount of information that can be carried by a communications system by using efficient coding, data compression, and error correction. It was developed by Claude Shannon, a researcher at Bell Laboratories, and first published as "a mathematical theory of communication" in 1948. We can apply it in the discipline of organizing to compare the amount of structure in different ways of organizing the same resources.[36[IA]]

Information theory quantifies the amount of organization in terms of the number of bits, binary decisions, or rules needed to describe some structure or pattern: the more complex or arbitrary a structure is, the more information it takes to describe it. For example, the organization of a company with a four-level hierarchy and a highly regular reporting structure where everyone supervises five people, can be described quite succinctly. In contrast, a company in which the number of direct reports at any management level is highly variable requires many more rules to describe.

Using measures from information theory to assess the amount of organization yields the somewhat counter-intuitive result that there is less information in the organization of a highly structured system than in a less structured one. It might help to flip this around and describe the amount of organization in terms of the reciprocal of the information measure. A system that is "highly organized" can be modeled or codified with relatively few rules or organizing principles, compared to a less organized system with many exceptions, corner cases, or one-off rules.

The "entropy" measure is often used to create predictive models of the "decision tree" variety, which is an algorithm that classifies or predicts by making a sequence of logical tests. Each test divides a collection of data into sets with less entropy (more predictability). (See §7.5)

analysis that can be used to create a system of categories and to assign resources to them.

Finally, we must acknowledge the ways in which information processing and telecommunications technologies have transformed and will continue to transform organizing systems in every sphere of economic and intellectual activity. A cen-

tury ago, when the telegraph and telephone enabled rapid communication and business coordination across large distances, these new technologies enabled the creation of massive vertically integrated industrial firms. In the 1920s, the Ford Motor Company owned coal and iron mines, rubber plantations, railroads, and steel mills so it could manage every resource needed in automobile production and reduce the costs and uncertainties of finding suppliers, negotiating with them, and ensuring their contractual compliance. Adam's Smith's invisible hand of the market as an organizing mechanism had been replaced by the visible hand of hierarchical management to control what Ronald Coase in 1937 termed "transaction costs" in *The Nature of the Firm.*

In recent decades, a new set of information and computing technologies enabled by Moore's law—unlimited computing power, effectively free bandwidth, and the Internet—have turned Coase upside down, leading to entirely new forms of industrial organization made possible as transaction costs plummet. When computation and coordination costs drop dramatically, it becomes possible for small firms and networks of services (provided by people or by computational processes) to out-compete large corporations through more efficient use of information resources and services, and through more effective information exchange with suppliers and customers, much of it automated. Herbert Simon, a pioneer in artificial intelligence, decision making, and human-computer interaction, recognized the similarities between the design of computing systems and human organizations and developed principles and mechanisms applicable to both.[37[Bus]]

Chapter 9, focuses on the representation of resource descriptions, taking a more technological or implementation perspective. Chapter 10, discusses how the nature and extent of descriptions determines the capabilities of the interactions that locate, compare, combine, or otherwise use resources in information-intensive domains.

2.5 When Is It Being Organized?

> *"Because bibliographic description, when manually performed, is expensive, it seems likely that the 'pre' organizing of information will continue to shift incrementally toward 'post' organizing."*
>
> — (Svenonius 2000, p. 194-195)

The organizing system framework recasts the traditional tradeoff between information organization and information retrieval as the decision about *when* the organization is imposed. We can contrast organization imposed on resources "on the way in" when they are created or made part of a collection with "on the way out" organization imposed when an interaction with resources takes place.

When an author writes a document, he or she gives it some internal organization via title, section headings, typographic conventions, page numbers, and other mechanisms that identify its parts and relationship to each other. The document could also have some external organization implied by the context of its publication, such as the name of its author and publisher, its web address, and citations or links to other documents or web pages.

Digital photos, videos, and documents are generally organized to some minimal degree when they are created because some descriptions, notably time and location, are assigned automatically to these types of resources by the technology used to create them. At a minimum, these descriptions include the resource's creation time, storage format, and chronologically ordered, auto-assigned filename (*IMG00001.JPG*, *IMG00002.JPG*, etc.), but often are much more detailed.[38][Com]

Digital resources created by automated processes generally exhibit a high degree of organization and structure because they are generated automatically in conformance with data or document schemas. These schemas implement the business rules and information models for the orders, invoices, payments, and the numerous other document types created and managed in business organizing systems.

Before a resource becomes part of a library collection, its author-created organization is often supplemented by additional information supplied by the publisher or other human intermediaries, such as an *International Standard Book Number (ISBN)* or *Library of Congress Call Number (LOC-CN)* or *Library of Congress Subject Headings (LOC-SH)*.

In contrast, Google and other search engines apply massive computational power to analyze the contents and associated structures (like links between web pages) to impose organization on resources that have already been published or made available so that they can be retrieved in response to a user's query "on the way out." Google makes use of existing organization within and between information resources when it can, but its unparalleled technological capabilities and scale yield competitive advantage in imposing organization on information that was not previously organized digitally.[39][Com] One reaction to the poor quality of some computational description has been the call for libraries to put their authoritative bibliographic resources on the open web, which would enable reuse of reliable information about books, authors, publishers, places, and subject classifications. This "linked data" movement is slowly gathering momentum.[40][Web]

Google makes almost all of its money through personalized ad placement, so much of the selection and ranking of search results is determined "on the way out" in the fraction of a second after the user submits a query by using information about the user's search history and current context. Of course, this "on the

way out" organization is only possible because of the more generic organization that Google's algorithms have imposed "on the way in."

In many organizing systems the nature and extent of organization changes over time as the resources are used. The arrangement of resources in a kitchen or office changes incrementally as frequently used things end up in the front of the pantry, drawer, shelf or filing cabinet or on the top of a pile of papers. Printed books or documents acquire margin notes, underlining, turned down pages or coffee cup stains that differentiate the most important or most frequently used parts. Digital documents do not take on coffee cup stains, but when they are edited, their new revision dates put them at the top of directory listings.

The scale of emergent organization of websites, photos on Flickr, blog posts, and other resources that can be accessed and used online dwarfs the incremental evolution of individual organizing systems. This organization is clearly visible in the pattern of links, tags, or ratings that are explicitly associated with these resources, but search engines and advertisers also exploit the less visible organization created over time by analyzing interaction resources, the recorded information about which resources were viewed and which links were followed.

The sort of organic or emergent change in organizing systems that takes place over time contrasts with the planned and systematic maintenance of organizing systems described as *curation* or *governance*, two related but distinct activities. *Curation* usually refers to the methods or systems that add value to and preserve resources, while the concept of *governance* more often emphasizes the institutions or organizations that carry out those activities. The former is most often used for libraries, museums, or archives and the latter for enterprise or inter-enterprise contexts. (For more discussion, see §3.5.4)

The organizing systems for businesses and industries often change because of the development of *de facto* or *de jure* standards, or because of regulations, court decisions, or other events or mandates.

We should always consider the extent to which people or technology in an organizing system are able to adapt when new resources, data, or people enter the picture. When and how much an organizing system can be changed depends on the extent of architectural thinking that went into its design (see The Three Tiers of Organizing Systems (page 47)), because it should be possible to make a change to a component without having to rethink the system entirely.

Sometimes what prevents adaptation are physical or technological constraints in the implementation of an organizing system, as with a desk or closet with fixed "pigeon holes," unmovable shelves, or with a music player with limited allowable formats and/or fixed storage capacity.

Machine learning algorithms use different techniques from those of human organizers; one of the important differences is that they're designed to adapt to

new inputs—which is why they're known to be "learning." In contrast, humans differ in how willing we are to re-organize to accommodate a different number or a different mix of resources. Without procedures in place to support or trigger adaptation, it may be quite difficult for us to change how we think or how we organize when our world changes, or even to realize that it has changed.

2.6 How (or by Whom) Is It Organized?

> *"The rise of the Internet is affecting the actual work of organizing information by shifting it from a relatively few professional indexers and catalogers to the populace at large. An important question today is whether the bibliographic universe can be organized both intelligently (that is, to meet the traditional bibliographic objectives) and automatically."*
>
> — (Svenonius 2000, p. 26)

In the preceding quote, Svenonius identifies three different ways for the "work of organizing information" to be performed: by professional indexers and catalogers, by the populace at large, and by automated (computerized) processes. Our notion of the organizing system is broader than her "bibliographic universe," making it necessary to extend her taxonomy. Authors are increasingly organizing the content they create, and it is important to distinguish users in informal and formal or institutional contexts. We have also introduced the concept of an organizing *agent* (§1.6) to unify organizing done by people and by computer algorithms.

Professional indexers and catalogers undergo extensive training to learn the concepts, controlled descriptive vocabularies, and standard classifications in the particular domains in which they work. Their goal is not only to describe individual resources, but to position them in the larger collection in which they reside. They can create and maintain organizing systems with consistent high quality, but their work often requires additional research, which is costly.

The class of professional organizers also includes the employees of commercial information services like Westlaw and LexisNexis, who add controlled and, often, proprietary metadata to legal and government documents and other news sources. Scientists and scholars with deep expertise in a domain often function as the professional organizers for data collections, scholarly publications and proceedings, and other specialized information resources in their respective disciplines. The *National Association of Professional Organizers (NAPO)* claims several thousand members who will organize your media collection, kitchen, closet, garage or entire house or help you downsize to a smaller living space.[42[Bus]]

Many of today's content creators are unlikely to be professional organizers, but presumably the author best understands why something was created and the

purposes for which it can be used. To the extent that authors want to help others find a resource, they will assign descriptions or classifications that they expect will be useful to those users. But unlike professional organizers, most authors are unfamiliar with controlled vocabularies and standard classifications, and as a result their descriptions will be more subjective and less consistent.

Similarly, most of us do not hire professionals to organize the resources we collect and use in our personal lives, and thus our organizing systems reflect our individual preferences and idiosyncrasies.

Non-author users in the "populace at large" are most often creating organization for their own benefit. These ordinary users are unlikely to use standard descriptors and classifications, and the organization they impose sometimes so closely reflects their own perspective and goals that it is not useful for others. Fortunately most users of "Web 2.0" or "community content" applications at least partly recognize that the organization of resources emerges from the aggregated contributions of all users, which provides incentive to use less egocentric *descriptors* and *classifications*. The staggering number of users and resources on the most popular applications inevitably leads to "tag convergence" simply because of the statistics of large sample sizes.

Finally, the vast size of the web and the even greater size of the "deep" or invisible web, composed of the information stores of business and proprietary information services, makes it impossible to imagine today that it could be organized by anything other than the massive computational power of search engine providers like Google and Microsoft. Likewise, data mining, predictive analytics, recommendation systems, and many other application areas that involve computational modeling and classification simply could not be done any other way.[43][Web]

Nevertheless, in the earliest days of the web, significant human effort was applied to organize it. Most notable is Yahoo!, founded by Jerry Yang and David Filo in 1994 as a directory of favorite websites. For many years the Yahoo! homepage was the best way to find relevant websites by browsing the extensive system of classification. Today's Yahoo! homepage emphasizes a search engine that makes it appear more like Google or Microsoft Bing, but the Yahoo! directory can still be found if you search for it.

2.7 Where is it being Organized?

> *"Bibliographic control requires fixing a document in the bibliographic universe by its space-time coordinates."*
>
> — (Svenonius 2000, p. 120)

Having identified the resources, reasoned about our motivations, limited the scope and scale, and determined when and by whom the organization will occur, we come finally to the question of where the resources are being organized.

In ordinary use, "Where" refers to a physical location. But the answer to "where?" often depends on whether we are asking about the current location, a past location, or an intended destination for resources that are in transit or in process. The answer to the question "where?" can take a lot of different forms. We can talk about an abstract space like "a library shelf" or we can talk about "the hidden compartment in Section XY at the Library of Congress," as depicted in the 2004 movie "National Treasure." We can answer "where?" with a description of a set of environmental conditions that best suit a class of wildlife, or a tire, or a sleeping bag. We can answer "where?" with "Renaissance Europe" or "Colonial Williamsburg." "Where?" can be a place in a mental construct, or even a place in an imagined location.

In the architectural design of an organizing system, its physical location is usually not a primary concern. In most organizing systems, the matter of where the organizing system and the resources are located can be abstracted away. So, in practice, resource location often is not as important as the other questions here. Physical constraints of the storage location should generally be relegated to an implementation concern rather than an architectural one. The construction of a special display structure for a valuable resource is not an independent design dimension; it is just the implementation of the user interface. (See §6.7 The Implementation Perspective (page 308))

Physical resources are often stored where it is convenient and efficient to do so, whether in ordinary warehouses, offices, storerooms, shelves, cabinets, and closets. It can be necessary to adapt an organizing system to characteristics of its physical environment, but this could undermine architectural thinking and make it harder to maintain the organization over time, as the collection evolves in scope and scale. (See §3.3.1 Organizing Physical Resources (page 100))

Digital resources, on the other hand, are increasingly organized and stored "in the cloud" and their actual locations are invisible, indeterminate, and generally irrelevant, except in situations where the servers and the information they hold may be subject to laws or practices of their physical location. For example, a controversy arose in Canada in 2013 when researchers discovered that Internet

service providers were, for various technical and business reasons, routinely routing trans-Canada web traffic through the United States. Because Canada has no jurisdiction over data traveling through cables and servers in another country, there was considerable outcry among Canadians who were concerned that their personal information was being subjected to the privacy laws and practices of another country without their knowledge or consent.

Sometimes location functions as an organizing principle in its own right, which in practice essentially collapses many of these architectural distinctions. This is frequently the case in our personal organizing systems, where we may exploit the innate human capability for spatial memory by always putting specific things like keys, eyeglasses, and cell phones in the same place, which makes them easy to find. But we can also see this happening in systems as complex and varied as: real estate information systems; wayfinding systems, such as road signage or mile markers; standardized international customs forms with position-specific data fields; geographic information systems; air, ground, sea, and space traffic control systems; and historic landmark preservation.

In §3.3.2 Organizing Places (page 103) we consider the organization of the land, built environments, and wayfinding systems. §6.5 The Structural Perspective (page 294) discusses the structural perspective on resource relationships, and in some systems, it may be very significant where resources are located in relation to one another. In The Barnes Collection (page 546), for example, works of art are physically grouped to enunciate common characteristics. Conversely, zoos do not mix the kangaroos with the wild dogs, and the military does not mix the ingredients for chemical weapons (at least, not until they plan to use them). There are also circumstances where resources can only exist in (or are particularly suited to) particular environments, such as the conditions required to grow wine grapes or mushrooms, or store spent nuclear fuel. UPS advises companies on where to put their warehouses and shipment centers. These are more substantial than questions of presentation, but it is debatable whether it falls under the storage or logic tier (you could have the principle of “keep the mushrooms somewhere moist” while not dictating where particularly).

Sometimes the location of an organizing system seems particularly salient, as in the design of cities where the street plan can be essential for orientation and navigation, and is embodied in zoning, voting, and other explicit organization, as well as in informal organization like neighborhood identity. But even here, it is really the people who live in the city who are being organized and whose interactions with the city and with each other are being encouraged or discouraged, not the physical location on which they live.

Indeed, in designing an organizing system you will often find that questions about location tumble naturally out of the other five design dimensions. For instance, questions about “when,” “what,” and “where” are often inseparable,

particularly when an organizing system is subject to outside regulations, which tend to have geographical jurisdictions. “Where” is also commonly bound up with “who” and “why,” when locational challenges or opportunities faced by a system's creators or users necessitate special design consideration. (See §4.5.2 Effectivity (page 200))

Location can be critically important to an organizing system—too important, in fact, to be considered alone. The question of “where?” is best considered in context of the other five design dimensions as a whole; a narrow focus on where the resources are being organized too often privileges past convention over architectural thinking and perpetuates legacy issues and poorly organized systems.

2.8 Key Points in Chapter Two

- A dimensional perspective makes it easier to translate between category- and discipline-specific vocabularies so that people from different disciplines can have mutually intelligible discussions about their organizing activities.

  (See §2.1 Introduction (page 61))
- In different situations, the same “thing” can be treated as a unique item, one of many equivalent members of a broad category, or a component of an item rather than as an item on its own.

  (See §2.2 What Is Being Organized? (page 64))
- A single physical resource can only be in one place at a time, and interactions with it are constrained by its size, location, and other properties. In contrast, digital copies and surrogates can exist in many places at once and enable searching, sorting, and other interactions with an efficiency and scale impossible for tangible things.

  (See §2.2 What Is Being Organized? (page 64))
- When the resources being organized consist of information content, deciding on the unit of organization is challenging because it might be necessary to look beyond physical properties and consider conceptual or intellectual equivalence.

  (See §2.2 What Is Being Organized? (page 64))
- Libraries, museums, and archives are often classified as *memory institutions* to emphasize their primary emphasis on resource preservation.

  (See §2.3 Why Is It Being Organized? (page 66))
- Businesses and governmental agencies are usually required by law to keep records of financial transactions, decision-making, personnel matters, and other information essential to business continuity, compliance with regulations and legal procedures, and transparency.

(See §2.3 Why Is It Being Organized? (page 66))

- If a system can turn its interaction traces into interaction resources, additional value can be created by analyzing these resources to enhance the interactions, to suggest new ones, or make predictions about how individual users or groups of them will behave.

  (See §2.3 Why Is It Being Organized? (page 66))

- Resources are always organized in ways that are designed to allocate value for some people (e.g., the owners of the resources, or the most frequent users of them) and not for others.

  (See §2.3 Why Is It Being Organized? (page 66))

- Subtle differences in resource arrangement, the number and framing of choices, and default values can have substantial effects on the decisions people make.

  (See §2.3 Why Is It Being Organized? (page 66))

- Different merchants or firms might make different decisions about the extent or granularity of description when they assign SKUs because of differences in suppliers, targeted customers, or other business strategies.

  (See §2.4 How Much Is It Being Organized? (page 70))

- A detailed description produced by sensors or computers can seem more accurate or authoritative than a simpler one created by a human observer, even if the latter would be more useful for the intended purposes. Detailed transaction data can be used to violate privacy and civil rights.

  (See §2.4 How Much Is It Being Organized? (page 70))

- Organizing more resources requires more descriptions to distinguish any particular resource from the rest, and more constraining organizing principles. Similar resources need to be grouped or classified to emphasize the most important distinctions among the complete set of resources in the collection.

  (See §2.4 How Much Is It Being Organized? (page 70))

- We can contrast organization imposed on resources “on the way in” when they are created or made part of a collection with “on the way out” organization imposed when an interaction with resources takes place.

  (See §2.5 When Is It Being Organized? (page 76))

- Digital resources created by automated processes generally exhibit a high degree of organization and structure because they are generated automatically in conformance with data or document schemas.

  (See §2.5 When Is It Being Organized? (page 76))

- The vast size of the web and the even greater size of the “deep” or invisible web makes it impossible to imagine today that it could be organized by anything other than the massive computational power of search engine providers like Google and Microsoft. Likewise, data mining, predictive analytics, recommendation systems, and many other application areas that involve computational modeling and classification simply could not be done any other way.

  (See §2.6 How (or by Whom) Is It Organized? (page 79))

Endnotes for Chapter 2

[27][Web] For example, many people manage their digital photos with Flickr, their home libraries with Library Thing, and their preferences for dining and shopping with Yelp. It is possible to use these “tagging” sites solely in support of individual goals, as tags like “my family,” “to read,” or “buy this” clearly demonstrate. But maintaining a personal organizing system with these web applications potentially augments the individual’s purpose with social goals like conveying information to others, developing a community, or promoting a reputation. Furthermore, because these community or collaborative applications aggregate and share the tags applied by individuals, they shape the individual organizing systems embedded within them when they suggest the most frequent tags for a particular resource.

[29][Bus] (Levitt 2005) and (Thaler 2008)

[30][Phil] (Winner 1980 p 121-136)

[31][CogSci] (Malone 1983) is the seminal research study, but individual differences in organizing preferences were the basis of Neil Simon’s Broadway play *The Odd Couple* in 1965, which then spawned numerous films and TV series.

[32][Bus] (Silverman 2013)

[33][Com] See Grudin’s classic work on non-technological barriers to the successful adoption of collaboration technology (Grudin 1994).

[36][IA] Information theory was developed to attack the technical problem of packing the maximum amount of data into the signal carrying telephone calls, but it quickly provided an essential statistical foundation in language analysis and computational linguistics. (Shannon 1948). Company organization and other examples applying information theory to the analysis of organizing systems can be found in (Levitin 2014, Chapter 7).

[37][Bus] Coase won the 1991 Nobel Prize in economics for his work on transaction costs, which he first published as a graduate student (Coase 1937). Berkeley

business professor Oliver Williamson received the prize in 2009 for work that extended Coase's framework to explain the shift from the hierarchical firm to the network firm (Williamson 1975, 1998). The notion of the "visible hand" comes from (Chandler 1977). Simon won the Nobel Prize in economics in 1978, but if there were Nobel Prizes in computer science or management theory he surely would have won them as well. Simon was the author or co-author of four books that have each been cited over 10,000 times, including (Simon 1997, 1996) and (Newell and Simon 1972).

[38][Com] Most digital cameras annotate each photo with detailed information about the camera and its settings in the *Exchangeable Image File Format (EXIF)*, and many mobile phones can associate their location along with any digital object they create.

[39][Com] Indeed, Geoff Nunberg criticized Google for ignoring or undervaluing the descriptive metadata and classifications previously assigned by people and replacing them with algorithmically assigned descriptors, many of which are incorrect or inappropriate. Calling Google's Book Search a "disaster for scholars" and a "metadata train wreck," he lists scores of errors in titles, publication dates, and classifications. For example, he reports that a search on "Internet" in books published before 1950 yields 527 results. The first 10 hits for Whitman's *Leaves of Grass* are variously classified as Poetry, Juvenile Nonfiction, Fiction, Literary Criticism, Biography & Autobiography, and Counterfeits and Counterfeiting. (Nunberg 2009)

[40][Web] (Byrne and Goddard 2010).

[42][Bus] NAPO: `http://www.napo.net` The name and scope of this organization seems a bit odd given how much professional organizing takes place in business, science, government, medicine, education, and other domains where closets and garages are not the most important focus.

[43][Web] (He et al. 2007) estimate that there are hundreds of thousands of websites and databases whose content is accessible only through query forms and web services, and there are over a million of those. The amount of content in this hidden web is many hundreds of times larger than that accessible in the surface or visible web.

See `http://www.worldwidewebsize.com/` for estimates of the size of the visible web calculated from comparisons of results from search engines.

Chapter 3

Activities in Organizing Systems

Robert J. Glushko
Erik Wilde
Jess Hemerly
Isabelle Sperano
Robyn Perry

3.1 Introduction

There are four *activities* that occur naturally in every *organizing system*; how explicit they are depend on the scope, the breadth or variety of the resources, and the scale, the number of resources that the organizing system encompasses. Consider the routine, everyday task of managing your wardrobe. When you organize your clothes closet, you are unlikely to write a formal *selection* policy that specifies what things go in the closet. You do not consciously itemize and prioritize the ways you expect to search for and locate things, and you are unlikely to consider explicitly the organizing principles that you use to arrange them. From time to time you will put things back in order and discard things you no longer wear, but you probably will not schedule this as a regular activity on your calendar.

Your clothes closet is an organizing system; defined as "an intentionally arranged collection of resources and the interactions they support." As such, it exposes these four highly interrelated and iterative *activities*:

Selecting
: Determining the scope of the organizing system by specifying which resources should be included. (*Should I hang up my sweaters in the clothes closet or put them in a dresser drawer in the bedroom?*)

Organizing
: Specifying the principles or rules that will be followed to arrange the resources. (*Should I sort my shirts by color, sleeve type, or season?)*

Designing resource-based interactions
: Designing and implementing the actions, functions or services that make use of the resources. (*Do I need storage places for clothes to be laundered? Should I have separate baskets for white and colors? Dry cleaning?)*

Maintaining
: Managing and adapting the resources and the organization imposed on them as needed to support the interactions. (*When is it time to straighten up the closet? What about mending? Should I toss out clothes based on wear and tear, how long I have owned them, or whether I am tired of them? What about excess hangers?)*

These activities are not entirely separable or sequential, and they can be informal for your clothes closet because its scope and scale are limited. In institutional organizing systems the activities and the inter-dependencies and iterations among them are more carefully managed and often highly formal.

For example, a data warehouse combines data from different sources like orders, sales, customers, inventory, and finance. Business analysts explore combinations and subsets of the data to find important patterns and relationships. The most important questions in the design and operation of the data warehouse can be arranged using the same activities as the clothes closet.

Selecting
: Which data sources should be included? How is their quality assessed? How much of the data is sampled? How are queries composed?

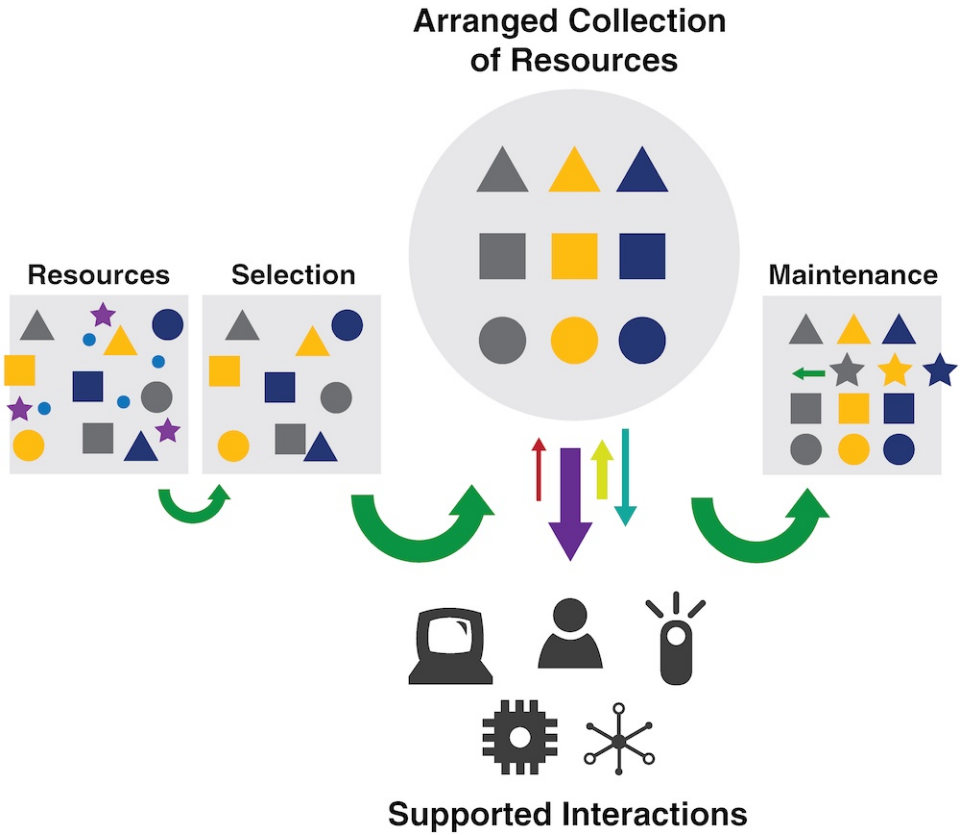
Organizing
: Which data formats and schemas will enable effective processing? Are needed transformations made at load time or query time?

Designing resource-based interactions
: What are the most important and frequent queries that need to be preconfigured?

Maintaining
: What governance policies and procedures are needed to satisfy retention, compliance, security, and privacy requirements?

Figure 3.1, Four Activities in all Organizing Systems. illustrates these four *activities* in all organizing systems, framing the depiction of the organizing and interaction design activities shown in Figure 1.1, An Organizing System. with the *selection* and *maintenance* activities that necessarily precede and follow them.

Figure 3.1. Four Activities in all Organizing Systems.



Four activities take place in all organizing systems: selection of resources for a collection; intentional organization of the resources; design and implementation of interactions with individual resources or with the collection, and; maintenance of the resources and the interactions over time.

These activities are deeply ingrained in academic curricula and professional practices, with domain-specific terms for their methods and results. Libraries and museums usually make their *selection* principles explicit in *collection development* policies. Adding a resource to a library collection is called *acquisition*, but adding to a museum collection is called *accessioning*. Documenting the contents of library and museum collections to organize them is called *cataloging*. *Circulation* is a central interaction in libraries, but because museum resources do not circulate the primary interactions for museum users are *viewing* or *visiting* the collection. *Maintenance* activities are usually described as *preservation* or *curation*.

In business information systems, *selection* of resources can involve *data generation*, *capture*, *sampling*, or *extraction*. Adding resources could involve *loading*, *integration*, or *insertion*. *Schema development* and *data transformation* are important organizing activities. Supported interactions could include *querying*, *reporting*, *analysis*, or *visualization*. *Maintenance* activities are often described as *deletion, purging, data cleansing*, *governance*, or *compliance*.

What about "Creating" Resources?

Our definition of organizing system as an intentionally arranged collection of resources might seem to imply that resources must exist before they are organized. This is often the case when we organize physical resources because the need for principled organization only arises when the collection gets too big for us to see everything in the collection at once. Similarly, many data analytics projects begin by bringing together data collected by others.

However, organizing systems for digital resources are often put in place as a prerequisite for creating them. This is always necessary when the resources are created by automated processes or data entry in business systems, and usually the case with professional writers in a technical publications context. We can think of database or document schemas (at the implementation tier) or data entry forms or word processor templates (in the user interface tier) as embodiments of the organizing principles in the data records or documents that are then created in conformance with them.

Domain-specific methods and vocabularies evolve over time to capture the complex and distinctive sets of experiences and practices of their respective disciplines. We can identify correspondences and overlapping meanings, but they are not synonyms or substitutes for each other. We propose more general terms like *selection* and *maintenance*, not as lowest common denominator replacements for these more specialized ones, but to facilitate communication and cooperation across the numerous disciplines that are concerned with organizing.

It might sound odd to describe the animals in a zoo as resources, to think of viewing a painting in a museum as an interaction, or to say that destroying information to comply with privacy regulations is maintenance. Taking a broader perspective on the activities in organizing systems so that we can identify best practices and patterns enables people with different backgrounds and working in different domains to understand and learn from each other.

Part of what a database administrator can learn from a museum curator follows from the rich associations the curator has accumulated around the concept of curation that are not available around the more general concept of maintenance. Without the shared concept of *maintenance* to bridge their disciplines, this learning could not take place.

Navigating this chapter

In §1.3 The Concept of "Resource" (page 35) and §2.2 What Is Being Organized? (page 64) we briefly discussed the fundamental concept of a *resource*. In this chapter, we describe the four primary *activities* with resources, using examples from many different kinds of organizing systems.

- §3.2 Selecting Resources (page 92)
- §3.3 Organizing Resources (page 98)
- §3.4 Designing Resource-based Interactions (page 122)
- §3.5 Maintaining Resources (page 133)

We emphasize the activities of organizing and of designing resource-based interactions that make use of the organization imposed on the resources. We discuss *selection* and *maintenance* to create the context for the organizing *activities* and to highlight the interdependencies of *organizing* and these other *activities*.

3.2 Selecting Resources

When we talk about organizing systems, we often do so in terms of the contents of their collections. This implies that the most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized. This decision is usually a constraint, not a choice; we acquire or encounter some resources that we need to interact with over time, and we need to organize them so we can do that effectively.

Selecting is the process by which resources are identified, evaluated, and then added to a collection in an organizing system. *Selection* is first shaped by the domain and then by the scope of the organizing system, which can be analyzed through six interrelated aspects:

1. the number and nature of users
2. the time span or lifetime over which the organizing system is expected to operate
3. the size of the collection
4. the expected changes to the collection
5. the physical or technological environment in which the organizing system is situated or implemented
6. the relationship of the organizing system to other ones that overlap with it in domain or scope.

(In Chapter 11, *The Organizing System Roadmap*, we discuss these six aspects in more detail.)

3.2.1 Selection Criteria

Selection must be an intentional process because, by definition, an organizing system contains resources whose selection and arrangement was determined by human or computational agents, not by natural processes. And given the broad definition of resource as "anything of value that can support goal-oriented activity" it follows that resources should be selected by an implicit or explicit assessment to determine whether they can be used to achieve those goals. So even though particular selection methods and criteria vary across resource domains, their common purpose is to determine how well the resource satisfies the specifications for the properties or capabilities that enable a person or nonhuman agent to perform the intended activities. "Fitness for use" is a common and concise way to summarize this idea, and while it highlights the need to have activities in mind before resources are selected to enable them, it also explains why precise selection criteria are harder to define for organizing systems that have

diverse sets of stakeholders or users with different goals, like those in public libraries.

Many resources are evaluated and selected one-at-a-time. This makes it impossible to specify in advance every property or criterion that might be considered in making a selection decision, especially for unique or rare resources like those being considered by a museum or private collector. In general, when resources are treated as instances, organizing activities typically occur after selection takes place, as in the closet organizing system with which we began this chapter.

When the resources being considered for a collection are more homogeneous and predictable, it is possible to treat them as a class or set, which enables selection criteria and organizing principles to be specified in advance. This makes selection and organizing into concurrent activities. This would be the case in the data warehouse organizing system, the other example at the beginning of this chapter, because each data source can be described by a schema whose structure is reflected in the organization of the data warehouse. Put another way, as long as subsequent datasets from a specific source do not differ in structure, only in temporal attributes like their creation or acquisition dates, the organization imposed on the initial dataset can be replicated for each subsequent one.

Well-run companies and organizations in every industry are highly systematic in selecting the resources that must be managed and the information needed to manage them. “Selecting the right resource for the job” is a clichéd way of saying this, but this slogan nonetheless applies broadly to raw materials, functional equipment, information resources and datasets, and to people, who are often called “human resources” in corporate-speak.

For some types of resources, the specifications that guide selection can be precise and measurable. Precise specifications are especially important when an organizing system will contain or make use of all resources of a particular type, or if all the resources produced from a particular source become part of the organizing system on some regular schedule. Selection specifications can also be shaped by laws, regulations or policies that require or prohibit the collection of certain kinds of objects or types of information.[44][Law]

For example, when a manufacturer of physical goods selects the materials or components that are transformed into its products, it carefully evaluates the candidate resources and their suppliers before making them part of its supply chain. The manufacturer would test the resources against required values of measurable characteristics like chemical purity, strength, capacity, and reliability. A business looking for transactional or demographic data to guide a business expansion strategy would specify different measurable characteristics; data files must be valid with respect to a schema, must contain no duplicates or personal-

ly identifiable information, and must be less than one month old when they are delivered. Similarly, employee selection has become highly data-intensive; employers hire people after assessing the match between their competencies and capabilities (expressed verbally or in a resume, or demonstrated in some qualification test) and what is needed to do the required activities.[45][DS]

Selection is an essential activity in creating organizing systems whose purpose is to combine separate web services or resources to create a composite service or application according to the business design philosophy of *Service Oriented Architecture (SOA)*.[46][Com] When an information-intensive enterprise or application combines its internal services with ones provided by others via Application Programming Interfaces (APIs), the resources are selected to create a combined collection of services according to the "core competency" principle: resources are selected and combined to exploit the first party's internal capabilities and those of its service partners better than any other combination of services could. For example, instead of writing millions of lines of code and collecting detailed maps to build an interactive map in an application, you can get access to the Google Maps organizing system with just a few lines of code.[47][Bus] (See the sidebar, Selection of Web-based Resources (page 94))

Selection of Web-based Resources

The nature and scale of the web changes how we collect resources and fundamentally challenges how we think of resources in the first place. Web-based resources cannot be selected for a collection by consulting a centralized authoritative directory, catalog, or index because one does not exist. ProgrammableWeb and other directories organize thousands of web-accessible APIs, and the dominant resource-organizing firms Amazon, Salesforce, Facebook, and Twitter offer hundreds of APIs to access massive amounts of information about products, people, and posts, but APIs enable access to only a fraction of the web's content. And although your favorite web search engine consults an index or directory of web resources when you enter a search query, you do not know where that index or directory came from or how it was assembled.[49][Web]

However, the web has universal scope and global reach, making most of the web irrelevant to most people most of the time. Researchers have attacked this problem by treating the web as a combination of a very large number of topic-based or domain-specific collections of resources, and then developing techniques for extracting these collections as digital libraries targeted for particular users and uses.[50][Web]

Scientific and business data are ideally selected after assessments of their quality and their relevance to answering specific questions. But this is easy to say and hard to do. It is essential to assess the quality of individual data items to

find data entry problems such as misspellings and duplicate records, or data values that are illegal, statistical outliers, or otherwise suspicious. It is also essential to assess the quality of data as a collection to determine if there are problems in what data was collected, by whom or how it was collected and managed, the format and precision in which it is stored, whether the schema governing each instance is rigorous enough, and whether the collection is complete. In addition, copyright, licensing, consumer protection laws, competitive considerations, or simply the lack of incentives to share resources make it difficult to obtain the best or most appropriate resources.[51][DS] (See the sidebar, Assessing and Addressing Data Quality (page 97))

In some domains, the nature of the resources or the goals they are intended to satisfy imply selection criteria that are inherently less quantifiable and more subjective. This is easy to see in personal collections, where selection criteria can be unconventional, idiosyncratic, or otherwise biased by the subjective perspective and experience of the collector. Most of the clothes and shoes you own have a reason for being in your closet, but could anyone else explain the contents of your closet and its organizing system, and why you bought that crazy-looking dress or shirt?

Even when selection criteria can be measured and evaluated in isolation, they are often incompatible or difficult to satisfy in combination. It would be desirable for data to be timely, accurate, complete, and consistent, but these criteria trade off against one other, and any prioritization that values one criterion over another is somewhat subjective. In addition, explicitly subjective perceptions of resource quality are hard to ignore; people are inclined to choose resources that come in attractive packages or that are sold and supported by attractive and friendly people.

Many of the examples in this section have involved selection principles whose purpose was to create a collection of desirable, rare, skilled, or otherwise distinctive resources. After all, no one would visit a museum whose artifacts were ordinary, and no one would watch a sports team made up of randomly chosen athletes because it could never win. However, choosing resources by randomly sampling from a large population is essential if your goal is to make inferences about it without having to study all its instances. Sampling is especially necessary with very large populations when timely decisions are required. A good sample for statistical purposes is one in which the selected resources are not different in any important way from the ones that were not selected.

Sampling is also important when large numbers of resources need to be selected to satisfy functional requirements. A manufacturer cannot test every part arriving at the factory, but might randomly test some of them from different shipments to ensure that parts satisfy their acceptance criteria.

3.2.2 Looking "Upstream" and "Downstream" to Select Resources

As we have seen, selection principles and activities differ across resource domains, and there is another important difference in selection that considers resources from the perspective of their history or the future.

In §3.2.1 we discussed the activity of selecting resources by assessing their conformance with specifications for required properties or capabilities. However, if you can determine where the resources come from, you can make better selection decisions by evaluating the people, processes, and organizing systems that create them. Using the analogy of a river, we can follow a resource "upstream" from us until we find the "headwaters." Physical resources might have their headwaters in a factory, farm, or artist's studio. Digital resources might have headwaters in a government agency, a scientist's laboratory, or a web-based commerce site.

When interaction resources (§1.9) are incorporated into the organizing system that creates them, as when records of a person's choices and behaviors are used to personalize subsequent information, the headwaters are obviously easy to find. However, even though finding the headwaters where resources come from is often not easy and sometimes not possible, that is where you are most likely to find the people best able to answer the questions, described in Chapter 2, that define any organizing system. The resource creators or producers will know the assumptions and tradeoffs they made that influence whether the resources will satisfy your requirements, and you can assess what they (or their documents that describe the resources) tell you and the credibility they have in telling it. You should also try to evaluate the processes or algorithms that produce the resources, and then decide if they are capable of yielding resources of acceptable quality.

The best outcome is to find a credible supplier of good quality resources. However, if an otherwise desirable supplier does not currently produce resources of sufficient quality, it is worth trying to improve the quality by changing the process using instruction or incentives. Advocates for open government have succeeded in getting numerous US government entities to publish data for free in machine-readable formats, but it was partly as a result of somewhat subversive demonstration projects and shaming that the government finally created data.gov in 2009. A clear lesson from the "quality movement" and statistical process control is that interventions that fix quality problems at their source are almost always a better investment than repeated work to fix problems that were preventable (see endnote[297 [Bus]]). But if you cannot find the headwaters or you are not able to address quality problems at their source, you can sometimes transform the resources to give them the characteristics or quality they need.[53[DS]] (See the sidebar, Assessing and Addressing Data Quality (page 97), and §10.3.2 Transforming Resources for Interactions (page 500).)

Assessing and Addressing Data Quality

If an organizing system uses data acquired from some external source, it is essential to assess its quality as an "intake" process. Ideally, the data comes with a schema that explicitly specifies what is expected, including legal structures, data types, and values (See §9.2 Structuring Descriptions (page 439)). This intake process runs tests that find problems and then runs processes to fix the problems.

There are a great many techniques for finding problems in numeric and string data. Some problems like missing data, duplicate records, spelling mistakes, and extreme or "outlier" values are easy to detect. A credit card charge for $10,000,000 is obviously a bad piece of data in a college student's account. Detecting anomalous and duplicate data is especially important because including them can produce misleading statistics and predictions, as well as creating the nuisance for consumers of receiving multiple copies of product catalogs, each with a different misspelling in a name or address.

Sometimes a dataset is valid with respect to its own specification but becomes problematical when it is combined with another dataset that has a different specification. Entities in the two datasets might not be described using the same units at the same point in time. So instead of analyzing and repairing resource instances, data cleaning must now be applied to every resource in a dataset, as when every "Zip Code" in a United States mailing directory is given the more universal "Postal Code" label, or when datasets using DD/MM/YY and MM/DD/YY formats for dates are combined.

Other data quality problems are harder to detect because they are contextual; a data value might be valid in some contexts but the same value might be invalid in others. For example, if you live in San Francisco and your credit card is used for transactions in Barcelona or Berlin, it could be fraud, or maybe you are on vacation. Similarly, high or low ratings for business establishments on sites like Yelp might be appropriate responses to excellent or poor service, but might also reflect "pay for rating" manipulation in the former case, and efforts by competitors to undermine rival businesses in the latter.[54][Com]

When you cannot obtain resources directly from their source, even if you have confidence in their quality at that point, it is important to analyze any evidence or records of their use or interactions as they flow downstream. (See §4.5) Physical resources are often associated with printed or digital documents that make claims about their origin and authenticity, and often have bar codes, RFID tags, or other technological mechanisms that enable them to be tracked from their headwaters to the places where they are used. Tracking is very important for

data resources because they can often be added to, derived from, or otherwise changed without leaving visible traces. Just as the water from melted mountain snow becomes less pure as its flows downstream, a data resource can become "dirty" or "noisy" over time, reducing its quality from the perspective of another person or computational agent further downstream. Data often gets dirty when it is combined with other datasets that contain duplicate or seemingly-duplicate information. Data can also become dirty when the hardware or software that stores it changes. Subtle differences in representation formats, transaction management, enforcement of integrity constraints, and calculations of derived values can change the original data.

In addition, a data resource can become inaccurate or obsolete simply because the world that the data describes has changed with the passage of time. People move, change jobs, get married or divorced, or die. Likewise, companies move, merge, get spun off, or go out of business. A poll taken a year before an election is often not a good predictor of the ultimate winner.[55][Com]

Other selection processes look "downstream" to select resources on the basis of predicted rather than current properties, capability, or suitability. Sports teams often sign promising athletes for their minor league teams, and businesses hire interns, train their employees, and run executive development programs to prepare promising low-level managers for executive roles. Businesses sometimes conduct experiments with variable product offers and pricing to collect data they will need in the future to power predictive models that will repay the investment in data acquisition many times over.

3.3 Organizing Resources

Organizing systems arrange resources according to many different principles. In libraries, museums, businesses, government agencies and other long-lived institutions, organizing principles are typically documented as cataloging rules, information management policies, or other explicit and systematic procedures so that different people can apply them consistently over time. In contrast, the principles for arranging resources in personal or small-scale organizing systems are usually informal and often inconsistent or conflicting.

For most types of resources, any number of principles could be used as the basis for their organization depending on the answers to the "why?" (§2.3), "how much?" (§2.4), and "how?" (§2.6) questions posed in Chapter 2.

A simple principle for organizing resources is *colocation —putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box.* However, most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole. What properties are significant and how to think about them

depends on the number of resources being organized, the purposes for which they are being organized, and on the experiences and implicit or explicit biases of the intended users of the organizing system. The implementation of the organizing system also shapes the need for, and the nature of, the resource properties.[57][DS]

Many resource collections acquire resources one at a time or in sets of related resources that can initially be treated the same way. Therefore, it is natural to arrange resources based on properties that can be assessed and interpreted when the resource becomes part of the collection.

"Subject matter" organization involves the use of a classification system that provides categories and descriptive terms for indicating what a resource is about. Because they use *aboutness* properties that are not directly perceived, methods for assigning subject classifications are intellectually-intensive and in many cases require rigorous training to be performed consistently and appropriately. Nevertheless, the cost and time required for this human effort motivates the use of computational techniques for organizing resources.

Property

In this book we use *property* in a generic and ordinary sense as a synonym for *feature* or "characteristic." Many cognitive and computer scientists are more precise in defining these terms and reserve *property* for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the *property* is called an *attribute,* "dimension," or "variable." *Feature* is used in *data science* and *machine learning* contexts for both "raw" or observable variables and "latent" ones, extracted or constructed from the original set.[56][CogSci]

As computing power steadily increases, the bias toward computational organization gets even stronger. However, an important concern arises when computational methods for organizing resources use so-called "black box" methods that create resource descriptions and organizing principles that are not inspectable or interpretable by people. In some applications more efficient information retrieval or question answering, more accurate predictions, or more personalized recommendations justify making the tradeoff. But comprehensibility is critical in many medical, military, financial, or scientific applications, where trusting a prediction can have life or death implications or cause substantial time or money to be spent.[59][DS]

3.3.1 Organizing Physical Resources

When the resources being arranged are physical or tangible things—such as books, paintings, animals, or cooking pots—any resource can be in only one place at a time in libraries, museums, zoos, or kitchens. Similarly, when organizing involves recording information in a physical medium—carving in stone, imprinting in clay, applying ink to paper by hand or with a printing press—how this information can be organized is subject to the intrinsic properties and constraints of physical things.

The inescapable tangibility of physical resources means that their organizing systems are often strongly influenced by the material or medium in which the resources are presented or represented. For example, museums generally collect original artifacts and their collections are commonly organized according to the type of thing being collected. There are art museums, sculpture museums, craft museums, toy museums, science museums, and so on.

Similarly, because they have different material manifestations, we usually organize our printed books in a different location than our record albums, which might be near but remain separate from our CDs and DVDs. This is partly because the storage environments for physical resources (shelves, cabinets, closets, and so on) have co-evolved with the physical resources they store.

3.3.1.1 Organizing with Properties of Physical Resources

Physical resources are often organized according to intrinsic physical properties like their size, color, or shape, because the human visual system automatically pays a lot of attention to them.

This inescapable aspect of visual perception was first formalized by German psychologists starting a century ago as the Gestalt principles (see the sidebar, Gestalt Principles (page 102)). Likewise, because people have limited attentional capacity, we ignore a lot of the ongoing complexity of visual (and auditory) stimulation, making us perceive our sensory world as simpler than it really is. Taken together, these two ideas explain why we automatically or “pre-attentively” organize separate things we see as groups or patterns based on their proximity and similarity. They also explain why arranging physical resources using these quickly perceived attributes can seem more aesthetic or satisfying than organizing them using properties that take more time to understand. Look at the cover of this book; the most organized arrangement of the colors and shapes just jumps out at you more than the others.

Physical resources are also commonly organized using intrinsically associated properties such as the place and time they were created or discovered. The shirts in your clothes closet might be arranged by color, by fabric, or style. We

Organizing People into Businesses

How people are organized into businesses is the essence of the discipline of management, and different aspects are taught in industrial organization and behavior, operations, entrepreneurship, and other courses. Organizing people in a business is often called "human resource management," and many of the principles for organizing physical resources and information resources apply to organizing people.

In addition, economics, strategy, and business culture are important considerations. There are a huge number of ways to organize people that differ in the extent of hierarchical structure, the flow of information up and down the hierarchy, the span of control for managers, and the discretion people have to deviate or innovate with respect to the work they have been assigned to do. For example, we can contrast law firms with a hierarchy of partners, associates, and paralegals with the self-management "holacracy" that companies like Zappos have experimented with, in which authority and decision-making are highly distributed among the employees.

Regardless of how the firm is organized, we can analyze it using economist Ronald Coase's idea of "transaction costs," which a business incurs in searching for and negotiating with suppliers, business partners, and customers, and in particular we can consider how new information and computing technologies reduce these costs to make the firm more efficient while remaining flexible.[61][Bus]

can view dress shirts, T-shirts, Hawaiian shirts and other styles as configurations of shirt properties that are so frequent and familiar that they have become linguistic and cultural categories. Other people might think about these same properties or categories differently, using a greater or lesser number of colors or ordering them differently, sorting the shirts by style first and then by color, or vice versa.

In addition to, or instead of, physical properties of your shirts, you might employ behavioral or usage-based properties to arrange them. You might separate your party and Hawaiian shirts from those you wear to the office. You might put the shirts you wear most often in the front of the closet so they are easy to locate. Unlike intrinsic properties of resources, which do not change, behavioral or usage-based properties are dynamic. You might move to Hawaii, where you can wear Hawaiian shirts to the office, or you could get tired of what were once your favorite shirts and stop wearing them as often as you used to.

Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users. Hazardous or flammable materials should not be stored where they might

Gestalt Principles

Psychologists Max Wertheimer, Wolfgang Kohler, and Kurt Koffka proposed several principles—proximity, similarity, continuity, connection, enclosure, and closure—that explain how our visual system imposes order on what it sees. There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

Koffka's pithy way of explaining the core idea of all the principles was that "The whole is other than the sum of the parts," which has been distorted over time to the cliché that "the whole is more than the sum of the parts."[62][CogSci]

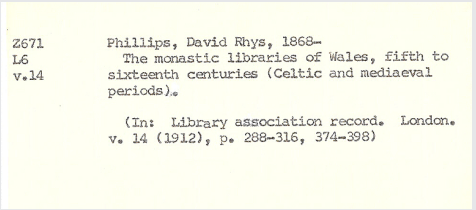
Designers of graphics and information visualizations rely on Gestalt rules because the automatic interpretations created by the human visual system enable their designs to be understood more quickly. This of course implies that designs that violate the Gestalt rules will be harder to understand. Camouflage—the use of disruptive coloration, colors and patterns that resemble backgrounds, countershading, shadow elimination, and similar techniques that make it difficult for the visual system to detect objects and edges—proves the power of Gestalt processing.[63][IA]

spill or ignite; lions and antelopes should not share the same zoo habitat or the former will eat the latter; adult books and movies should not be kept in a library where children might accidentally find them; and people who are confrontational, passive aggressive, or arrogant do not make good team members when tough decisions need to be made. For almost any resource, it seems possible to imagine a combination with another resource that might have unfortunate consequences. We have no shortage of professional certifications, building codes, MPAA movie ratings, and other types of laws and regulations designed to keep us safe from potentially dangerous resources.

3.3.1.2 Organizing with Descriptions of Physical Resources

To overcome the inherent constraints with organizing physical resources, organizing systems often use additional physical resources that describe the primary physical ones, with the library card catalog being the classic example. A specific physical resource might be in a particular place, but multiple description resources for it can be in many different places at the same time.

Card From Library Catalog


Z671
L6
v.14

Phillips, David Rhys, 1868–
The monastic libraries of Wales, fifth to sixteenth centuries (Celtic and mediaeval periods).

(In: Library association record. London. v. 14 (1912), p. 288-316, 374-398)

A catalog card from the library of the School of Library and Information Studies at the University of California, Berkeley. The card describes a book about the monastic libraries of Wales, which like the library in which this card came from are no longer in existence.

(Photo by R. Glushko.)

Card Catalog Cabinet

Library catalogs were managed as collections of printed cards for much of the 20th century, and the wooden cabinets that contained them were ubiquitous functional furniture in every library. Today such cabinets are often considered "retro" or antique treasures.

(Photo by R. Glushko.)

When the description resources are themselves digital, as when a printed library card catalog is put online, the additional layer of abstraction created enables additional organizing possibilities that can ignore physical properties of resources and many of the details about how they are stored.

In organizing systems that use additional resources to identify or describe primary ones, "adding to a collection" is a logical act that need not require any actual movement, copying, or reorganization of the primary resources. This virtual addition allows the same resources to be part of many collections at the same time; the same book can be listed in many bibliographies, the same web page can be in many lists of web bookmarks and have incoming links from many different pages, and a publisher's digital article repository can be licensed to any number of libraries.

3.3.2 Organizing Places

Places are physical resources, but unlike the previous two subsections where we treat the environment as given (the library or museum building, the card catalog or bookshelf) and discuss how we organize resources like books in that environment, we can take an alternative perspective and discuss how we design that physical environment. These environments could be any of the following:

- The land itself, as when we lay out city plans when organizing how people live together and interact in cities.
- A "built environment," a human-made space, particular building, or a set of connected spaces and buildings. A built environment could be a museum, airport, hospital, casino, department store, farm, road system, or any kind of building or space where resources are arranged and people interact with them.
- The orientation and navigation aids that enable users to understand and interact in built environments. These are resource descriptions that support the interaction requirements of the users.

These are not entirely separable contexts, but they are easier to discuss as if they are considered as such.

3.3.2.1 Organizing the Land

Cities naturally emerge in places that can support life and commerce. Almost all major cities are built on coasts or rivers because water provides sustenance, transportation and commercial links, and power to enable industry. Many very old cities have crowded and convoluted street plans that do not seem intentionally organized, but grid plans in cities also have a very long history. Cities in the Middle East were laid out in rough grids as far back as 2000+ BCE.

Because the United States, and especially the American West, was not heavily settled until much more recently compared to most of Europe and Asia, it was a place for people to experiment with new ideas in urban design. The natural human tendency to impose order on habitation location had ample room to do just that. The easiest and most efficient way to organize space is using a coordinate grid, with streets intersecting at perpendicular angles. Salt Lake City, Albuquerque, Phoenix, and Seattle are notable examples of grid cities. An interesting hybrid structure exists in Washington DC, which has radiating diagonal avenues overlaid on a grid.[64][IA]

3.3.2.2 Organizing Built Environments

Built environments influence the expectations, behaviors, and experiences of everyone who enters the space—employees, visitors, customers, and inhabitants are all subject to the design of the spaces they occupy. These environments can be designed to encourage or discourage interactions between people, to create a sense of freedom or confinement, to reward exploration or enforce efficiency, and of course, much much more. The arrangement of the resources in a built environment also encourages or discourages interactions, and sometimes the built environment is designed with a specific collection of resources in mind to enable and reinforce some particular interaction goals or policies.

If we contrast the built environments of museums, airports, and casinos, and the way in which each of them facilitates or constrains interactions are more obvious. Museums are often housed in buildings designed as architectural monuments that over time become symbols of national, civic, or cultural identity. Many old art museums mimic classical architecture, with grand stairs flanked by tall columns. They have large and dramatic entry halls that invite visitors inside. Modern museums are decidedly less traditional, and some people complain that the architecture of modern art museums can overshadow the art collection contained within because people are induced to pay more attention to the building than to its contents.

Some recently built airports have been designed with architectural flair, but airport design is more concerned with efficiency, walkability (maybe with the aid of moving walkways), navigability, and basic comfort for travelers getting in and out the airport. Wide walkways, multiple staircases, and people movers whose doors open in one direction at a time, all encourage people to move in certain directions, sometimes without the people even realizing they are being directed.

If you have ever been lost in a casino or had trouble finding the exit you can be sure you experienced a casino that achieved its main design goals: keeping people inside and making it easy for them to lose track of time because they lack both windows and clocks. As American architect Robert Venturi points out, “The intricate maze under the low ceiling never connects with outside light or outside space...This disorients the occupant in space and time... He loses track of where he is and when it is.”

If one accepts the premise that values and bias are at work in decisions about organizing systems, it is difficult not to see it in built environments. Consider queue design in banks, supermarkets, or boarding airplanes. Assuming that it is desirable to organize people efficiently to minimize wait times and crowding, how should the queue be designed? How many categories of people should there be? What is the basis for the categories?

It may be uncontroversial to include several express lanes in a supermarket checkout, because people can choose to buy fewer items if they do not want to wait. Similarly, it seems essential for hospital emergency rooms to have a triage policy that selects patients from the emergency room queue based on their likely benefit from immediate medical attention.

There are many other examples of how values and biases become part of built environments. In the mid-20th century the road systems of Long Island in New York were designed with low overpasses, which prevented public buses from passing under them, effectively segregating the beaches. The trend in college campus design after the student protests of the 1960s and 1970s was to create layouts that would prevent or frustrate large demonstrations.[67][Phil]

3.3.2.3 Orientation and Wayfinding Mechanisms

It is easy to move through an environment and stay oriented if the design is simple and consistent, but most built environments must include additional features or descriptions to assist people in these tasks. Distinctive architectural elements can create landmarks for orientation, and spaces can be differentiated with color, lighting, furnishings, or other means. More ubiquitous mechanisms include signs, room numbers, or directional arrows highlighting the way and distance to important destinations.

In airports, for example, there are many orientation signs and display terminals that help passengers find their departure gates, baggage, or ground transportation services. In contrast, casinos provide little orientation and navigation support because increased confusion leads to lengthier visits, and more gambling on the part of the casino's visitors.

A recent innovation in wayfinding and orientation mechanisms is to give them sensing and communication capabilities so they can identify people by their smartphones and then provide personalized directions or information. These so-called "beacon" systems have been deployed at numerous airports, including London's Gatwick, San Francisco, and Miami. [68][IA]

3.3.3 Organizing Digital Resources

Organizing systems that arrange digital resources like digital documents or information services have some important differences from those that organize physical resources. Because digital resources can be easily copied or interlinked, they are free from the "one place at a time" limitation.[69][Law] The actual storage locations for digital resources are no longer visible or very important. It hardly matters if a digital document or video resides on a computer in Berkeley or Bangalore if it can be located and accessed efficiently.[70][Web]

Moreover, because the functions and capabilities of digital resources are not directly manifested as physical properties, the constraints imposed on all *material* objects do not matter to digital content in many circumstances.[71][Com]

An organizing system for digital resources can also use digital description resources that are associated with them. Since the incremental costs of adding processing and storage capacity to digital organizing systems are small, collections of both primary digital resources and description resources can be arbitrarily large. Digital organizing systems can support collections and interactions at a scale that is impossible in organizing systems that are entirely physical, and they can implement services and functions that exploit the exponentially growing processing, storage and communication capabilities available today. This all sounds good, unless you are the small local business with limited onsite

inventory that cannot compete with global web retailers that offer many more choices from a network of warehouses.[73][Web]

Materiality

An emerging issue in the field of digital humanities is the requirement to recognize the *materiality* of the environment that enables people to create and interact with digital resources Even if the resources themselves are intangible, it can be necessary to study and preserve the technological and social context in which they exist to fully understand them.

There are inherently more arrangements of digital resources than there are for physical ones, but this difference emerges because of multiple implementation platforms for the organizing system as much as in the nature of the resources. Nevertheless, the organizing systems for digital books, music and video collections often maintain the distinctions embodied in the organizing system for physical resources because it enables their coexistence or simply because of legacy inertia. As a result, the organizing systems for collections of digital resources tend to be coarsely distinguished by media type (e.g., document management, digital music collection, digital video collection, digital photo collection, etc.).

Information resources in either physical or digital form are typically organized using intrinsic properties like author names, creation dates, publisher, or the set of words that they contain. Information resources can also be organized using assigned properties like subject classifications, names, or identifiers. Information resources can also be organized using behavioral or transactional properties collected about individuals or about groups of people with similar interaction histories. For example, Amazon and Netflix use browsing and purchasing behavior to make book and movie recommendations.[74][Com]

Complex organization and interactions are possible when organizing systems with digital resources are based on the data type or data model of the digital content (e.g., text, numeric, multimedia, statistical, geospatial, logical, scientific, personnel, and so on).

Interactions with numeric data can be further distinguished according to the *levels of measurement* embodied in the number, which determine how much quantitative processing makes sense:

- *Nominal* level data uses a number as an identifier for an instance or a category to distinguish it from other ones. Products in a catalog might have numbers associated with them, but the products have no intrinsic order, so no measurements using the numbers are meaningful other than the frequency with which they occur in the dataset. The most frequently occurring value is called the *mode*.

- *Ordinal* level data indicates a direction or ranking on some naturally ordered scale. We know that the first place finisher in a race came in ahead of the second place one, who finished ahead of the third place finisher, but this result conveys no information about the spacing among the racers at the finish line. The middle value in a sorted list is the *median*.
- *Interval* level data conveys order information, but in addition, the values that subdivide the scale are equally spaced. This makes it meaningful to calculate the distance between values, the *mean* or *average* value (the value for which the sum of its absolute distances to each other value is zero), the standard deviation, and other descriptive statistics about the data.
- *Ratio* level data is interval data with a fixed zero point, which makes assertions about proportions meaningful. $10,000 is twice as much as $5,000.

These distinctions are data type and levels of measurement are often strongly identifiable with business functions: operational, transactional, process control, and predictive analytics activities require the most fine-grained data and quantitative measurement scales, while strategic functions might rely on more qualitative analyses represented in narrative text formats.

Just as there are many laws and regulations that restrict the organization of physical resources, there are laws and regulations that constrain the arrangements of digital ones. Many information systems that generate or collect transactional data are prohibited from sharing any records that identify specific people. Banking, accounting, and legal organizing systems are made more homogeneous by compliance and reporting standards and rules.

3.3.3.1 Organizing Web-based Resources

The *Domain Name System (DNS)* is the most inherent scheme for organizing web resources. Top-level domains for countries (.us, .jp, .cn, etc.) and generic resource categories (.com, .edu. .org, gov, etc.) provide some clues about the resources organized by a website. These clues are most reliable for large established enterprises and publishers; we know what to expect at `ibm.com`, `Berkeley.edu`, and `sfgov.org`.[76][Web]

The network of hyperlinks among web resources challenges the notion of a collection, because it makes it impractical to define a precise boundary around any collection smaller than the complete web.[77][Web] Furthermore, authors are increasingly using “web-native” publication models, creating networks of articles that blur the notions of articles and journals. For example, scientific authors are interconnecting scientific findings with their underlying research data, to discipline-specific data repositories, or to software for analyzing, visualizing, simulation, or otherwise interacting with the information.[78][Web]

Organizing Mental Resources

Memories can be viewed either as physical (because at some level they are represented in the brain) or as digital (because they are retrieved as electrical impulses), but memory techniques like the method of loci and memory palaces reify this duality in an interesting way.

While physical resources must be stored in physical locations, our powerful spatial memory provides an opportunity for us to, in a sense, store mental resources in physical locations. Our hippocampus, the brain component dedicated to memory, is highly developed for storing and recalling memories of physical locations. The ancient Greeks relied on this capability and devised a mnemonic system—the method of loci—which involved attaching things to remember, the key ideas in a speech perhaps, to well-known physical locations. While giving the speech, then, all one must do is imagine walking through that physical location from idea to idea. Today, champion memorizers use this technique to associate items with places in vividly imagined "memory palaces." While you may not be interested in memorizing the order of a deck of cards, recognizing the power of our spatial memory may be worth considering when designing your organizing system or when analyzing the successes or failures of a system.[75][CogSci]

The conventional library is both a collection of books and the physical space in which the collection is managed. On the web, rich hyper linking and the fact that the actual storage location of web resources is unimportant to the end users fundamentally undermine the idea that organizing systems must collect resources and then arrange them under local control to be effective. The spectacular rise during the 1990s of the AOL "walled garden," created on the assumption that the open web was unreliable, insecure, and pernicious, was for a time a striking historical reminder and warning to designers of closed resource collections until its equally spectacular collapse in the following decade.[79][Web] But Facebook so far is succeeding by following a walled garden strategy.

3.3.3.2 "Information Architecture" and Organizing Systems

The discipline known as *information architecture* can be viewed as a specialized approach for designing the information models and their systematic manifestations in user experiences on websites and in other information-intensive organizing systems.[80][IA] Abstract patterns of information content or organization are sometimes called architectures, so it is straightforward from the perspective of the discipline of organizing to define the activity of *information architecture* as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use. Note how the

first part of this definition refers to intentional arrangement of resources, and the second to the interactions enabled by that arrangement.

Our definition of information architecture implies a methodology for the design of user interfaces and interactions that puts conceptual modeling at the foundation. Best practices in information architecture emphasize the use of systematic principles or design patterns for organizing the resources and interactions in user interfaces. The logical design is then translated into a graphical design that arranges windows, panes, menus, and other user interface components. The logical and graphical organization of a user interface together affect how people interact with it and the actions they take (or do not take).

Some information design conventions have become design patterns. Documents use headings, boxes, white space, and horizontal rules to organize information by type and category. Large type signifies more important content than small type, red type indicates an advisory or warning, and italics or bold says “pay attention.”

Some patterns are general and apply to an entire website, page, or interface genre such as a government site, e-commerce site, blog, social network site, home page, “about us” page, and so on. Other patterns are more specific and affect a part of a site or a single component of a page (e.g., autocompletion of a text field, breadcrumb menu, slideshow).

In websites, different categories of content or interactions are typically arranged in different menus. The choices within each menu are then arranged to reflect typical workflows or ordered according to some commonly used property like size, percentage, or price.

All design patterns reflect and reinforce the user's past experiences with content and interface components, and this familiarity reduces the cognitive complexity of user interface interaction, requiring users to pay less attention.[82][IA]

The Activities of Information Architecture

IA is a relatively new field, but the ubiquity of the web and information-intensive applications that must implement many types of user interactions has inspired many conceptual and methodological innovations. Here are some of them.

Selecting Resources: To make good choices about what content to include in an information system or service, methods and tools for creating and organizing the information that is potentially available are important. Glushko and McGrath's method for creating a “Document Inventory” and Halvorson and Rach's “Information Inventory” both use a matrix or grid format to list information sources and various associated properties. Once the inventory is completed, the information must be evaluated with respect to the user and information requirements. This usually requires a more fine-grained analysis to choose the most reliable or reusable source when there are alternatives. This process is usually called content auditing, and tools or templates for organizing the work are easy to find on the web.

Organizing Resources: Tidwell proposes a set of design patterns for input forms, text and graphic editors, information graphics, calendars, and other common types of web applications that organize resources. Morville and Rosenfield classify design patterns as “organization schemes” and “organization structures,” reinforcing the idea that information architecture is a sub-specialty of the discipline of organizing.

Designing Interactions: Kalbach presents design patterns and implementations for navigation interactions. Resmini and Rosati discuss architectures and examples for information architectures that interconnect physical and digital channels. Marcotte introduces techniques for adapting user interfaces to the size and capabilities of different devices, collectively called responsive web design.

Information architects use a variety of tools for representing information and process models. Common ones include site maps, workflow and data-flow diagrams, and wireframe models. Brown's *Communicating Design* and Abel and Baillie's *The Language of Content Strategy* are concise sources.[81][IA]

However, interface designers can take advantage of this familiarity and employ design patterns in a less beneficial way to manipulate users, control their behaviors, or trick them into taking actions they do not intend. Patterns used this way are sometimes called Dark Patterns (page 112).

Many organizing systems need to support interactions to find, identify, and select resources. Some of these systems contain both physical and digital

Dark Patterns

Some websites and applications employ Dark Patterns, which rely on user familiarity with good design patterns to induce users to take actions or fail to take actions in ways counter to their best interests. For example, a website may exploit familiar patterns to induce users to click on an ad disguised as a news item, sign up for unwanted e-mails, disclose personal information, or ignore important terms and conditions because they are buried in tiny text or in unusual locations.

Darkpatterns.org collects and classifies dark patterns. The largest categories are "bait and switch" (suggest one action but cause another), "trick questions" (misleading phrasing of an option), and "misdirection" (focusing attention on one thing to distract from another). The website has numerous examples of interfaces that try to get users to install additional software or change their defaults to a company's product during installation. Other examples are from commerce sites that conceal the cheapest options, add additional fees at the very end of the purchase process, or make it difficult to accurately compare costs.

These practices are enough of a concern that some governments have begun to regulate the information that must be provided to consumers when purchasing digital products. The Directive on Consumer Rights published by the European Commission in June 2014 contains instructions about design choices that should be avoided, such as allowing additional purchases and payments without the consumer's consent. The Directive even includes a model set of patterns to help designers comply with it.[83][Law]

Dark patterns can be used to manipulate interactions with physical resources too. Gas pumps with three or four grades of gasoline invariably arrange the pumps in order of price, with the cheapest gas at the left and the most expensive on the right. Some gas stations put the cheapest gas in the middle, which causes inattentive customers who are relying on the usual pattern to buy more expensive gas than they intended.

resources, as in a bookstore with both web and physical channels, and many interactions are implemented across more than one device. Both the cross-channel and multiple-device situations create user expectations that interactions will be consistent across these different contexts. Starting with a conceptual model and separating content and structure from presentation, as we discussed in §1.6, gives organizing systems more implementation alternatives and makes them more robust in the face of technology diversity and change.

A model-based foundation is also essential in information visualization applications, which depict the structure and relationships in large data collections us-

ing spatial and graphical conventions to enable user interactions for exploration and analysis. By transforming data and applying color, texture, density, and other properties that are more directly perceptible, information visualization applications enable people to obtain more information than they can from text displays.[84][CogSci]

Some designers of information systems put less emphasis on conceptual modeling as an "inside-out" foundation for interaction design and more emphasis on an "outside-in" approach that highlights layout and other presentation-tier considerations with the goal of making interactions easy and enjoyable. This focus is typically called user experience design, and information architecture methods remain an important part of it, but not beginning with explicit organizing principles implies more heuristic methods and yields less predictable results.

3.3.4 Organizing With Descriptive Statistics

Descriptive statistics, about a collection or dataset, summarize it concisely and can identify the properties that might be most useful as organizing principles. The simplest statistical description of a collection is how big it is; how many resources or observations does it contain?

Descriptive statistics summarize a collection of resources or dataset with two types of measures:

- Measures of *central tendency*: Mean, median, and mode; which measure is appropriate depends on the level of measurement represented in the numbers being described (these measures and the concept of levels of measurements are defined in §3.3.3 Organizing Digital Resources (page 106)).
- Measures of *variability*: Range (the difference between the maximum and minimum values), and standard deviation (a measure of the spread of values around the mean).

Statistical descriptions can be created for any resource property, with the simplest being the number of resources that have the property or some particular value of it, such as the number of times a particular word occurs in a document or the number of copies a book has sold. Comparing summary statistics about a collection with the values for individual resources helps you understand how typical or representative that resource is. If you can compare your height of 6 feet, ½ inch with that of the average adult male, which is 5 feet, 10 inches, the difference is two and a half inches, but what does this mean? It is more informative to make this comparison using the standard deviation, which is three inches, because this tells you that 68% of adult men have heights between 5 feet, 7 inches and 6 feet, 1 inch. When measurements are normally distributed in the familiar bell-shaped curve around the mean, the standard deviation makes it easy to identify statistical outliers.

No matter how measurements are distributed, it can be useful to employ descriptive statistics to organize resources or observations into categories or quantiles that have the same number of them. Quartiles (4 categories), deciles (10), and percentiles (100) are commonly used partitions.

Alternatively, resources or observations can be organized by visualizing them in a histogram, which divides the range of values into units with equal intervals. Because values tend to vary around some central tendency, the intervals are unlikely to contain the same number of observations. Descriptive statistics and associated visualizations can suggest which properties make good organizing principles because they exhibit enough variation to distinguish resources in their most useful interactions. For example, it probably isn't useful to organize books according to their weight because almost all books weigh between ½ and 2 pounds, unless you are in the business of shipping books and paying according to how much they weigh.

3.3.4.1 Exploratory Analysis to Understand Data

Many experts recommend that data analysts should undertake some exploratory analysis with descriptive statistics and simple information visualizations to understand their data before applying sophisticated computational techniques to the dataset. In particular, because the human visual system quickly perceives shapes and patterns, analyzing and graphing the values of data attributes and other resource descriptions can suggest which properties might be useful and comprehensible organizing principles. In addition, data visualization makes it easy to recognize values that are typical or that are outliers. Some of this analysis might form part of data quality assessment during resource selection, but if not done then, it should be done as part of the organizing process.

A dataset whose fields or attributes lack information about data types and units of measure has little use because the data lacks meaning. When some, but not all parts of the data are named or annotated, avoid over-interpreting these descriptions' meanings. (See §4.4 Naming Resources (page 188).)

We will do some exploratory analysis to understand what an example dataset contains and how we might use it. For our example, we consider a collection of a few hundred records from a healthcare study, whose first eight records and first five data fields in each record are shown in Figure 3.2a, Example Dataset.

Figure 3.2a. Example Dataset

ID	Sex	Temp	Age	Weight	...	...	...	...	...
1	1	97.6	32	135					
2	0	97.6	19	118					
3	0	97.6	23	128					
4	1	98.7	34	140					
5	1	98.5	52	162					
6	1	98.7	60	160					
7	0	98.3	36	148					
8	0	98.3	38	155					
...	...								
260	1	99.0	23	123					

The “ID” column contains numeric data, but every value is a different integer, and the values are contiguous. The field label “ID” suggests that this is the resource identifier for the participants in the healthcare study. Further examination of other tables will reveal that this is a key value that points into a different dataset containing the resource names.

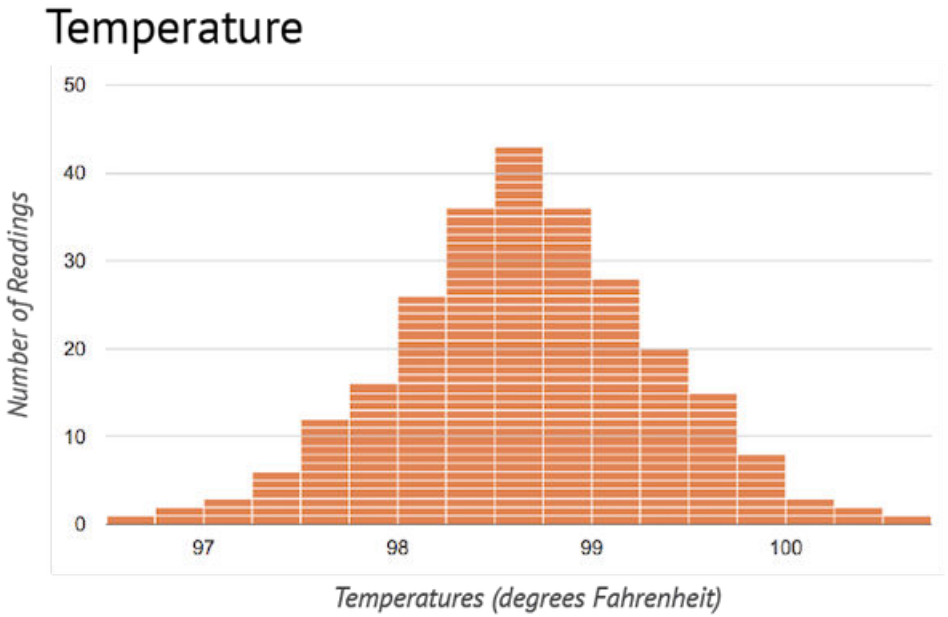
The “Sex” column is also numeric, but there are only two different values, 0 and 1, and in the complete dataset they are approximately equal in frequency. This attribute seems to be categorical or Boolean data. This makes sense for a “Sex” categorization, and it is likely to prove useful in understanding the dataset.

The “Temp” column contains several hundred different numeric values in the complete dataset, ranging from 96.8 to 100.6, with a mean of 98.6. These values are sensible if the label “Temp” means the under-the-tongue body temperature in degrees Fahrenheit of the study participant when the other measures were obtained. This type of data is usefully viewed as a histogram to get a sense of the spread and shape, shown in Figure 3.2b, Temperature.

Histogram

A histogram is the simplest visualization of one-dimensional data. It is a bar graph that takes the full range of values, organizes them into a set of intervals of equal size on one axis, and then counts the number of values in each interval on the other axis.

Figure 3.2b. Temperature



The data values of the "Temp" column follow the familiar normal or bell-shaped distribution, for which simple and useful descriptive statistics are the mean and the standard deviation. The mean (or average) is at the center of the distribution, and the standard deviation captures the width of the bell shape. In this dataset, the very narrow range of data values here suggests that this attribute is not useful as an organizing principle, since it does not distinguish the resources in any significant way. In a larger sample, however, there might be a few very low or very high temperatures, and it would be useful to investigate these "hypothermic" or "hyperthermic" outliers.

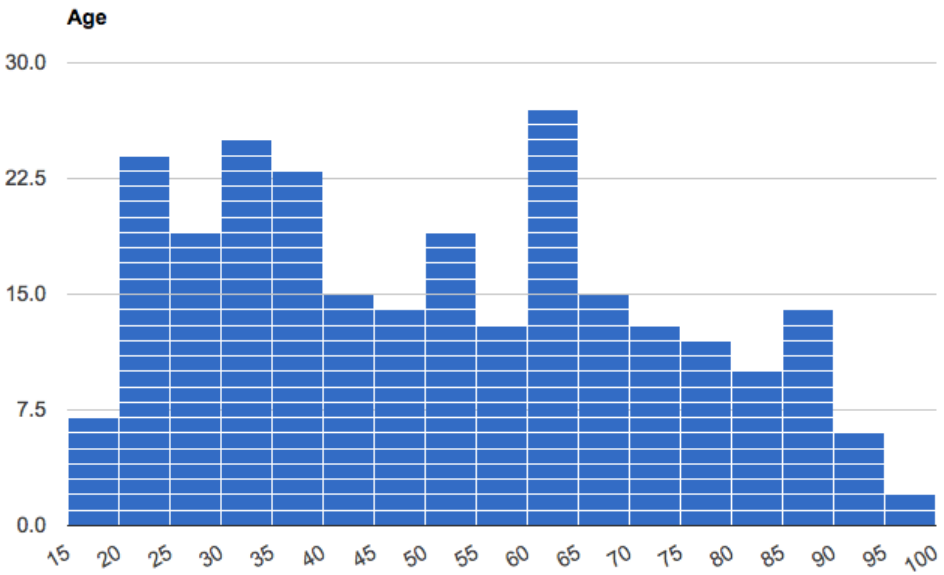
The data values of the "Age" column range from 18 to 97, and are spread broadly across the entire range; this is the age, in years, of the study participants. When a distribution is very broad and flat, or highly skewed with many values at one end or another, the mean value is less useful as a descriptive statistic. Instead of the mean, it is better to use the median or middle value as a summary of the data; the median value for "Age" in the complete dataset is 39.

Figure 3.2c. Age

Median versus Average

If ten people are in a bar, all of whom make $50,000 a year, when a movie star who made $25,000,000 this year walks in, the average income is now $2.3 million. The median income is still $50,000.

The End of Average tells the story of how the U.S. military designed aircraft cockpits beginning in 1926 on the basis of the average dimensions of a 1926 pilot. In 1950, researchers measured over four thousand pilots only to discover that no actual pilot had average values on all the measures, and recommended adjustable seats and controls in cockpit design.[85][DS]



The "Weight" column has about 220 different numeric values, from 82 to 300, and judging from this range we can infer that the weights are measured in pounds. The data follows an uneven distribution with peaks around 160 and 200, and a small peak at 300. This odd shape appears in the histogram of Figure 3.2d, Weight. The two peaks in this so-called multi-modal histogram suggest that this measure is mixing two different kinds of resources, and indeed it is because weights of men and women follow different distributions. It would thus be useful to use the categorical "Sex" data to separate these populations, and Figure 3.2e, Sex and Weight: Female shows how analyzing weight for women and men as different populations is much more informative as an organizing principle than combining them.

What about the odd peak in the distribution at 300? End of range anomalies like this generally reflect a limitation in the device or system that created the data. In this case, the weight scale must have an upper limit of 300 pounds, so the peak represents the people whose weight is 300 or greater.

Figure 3.2d. Weight

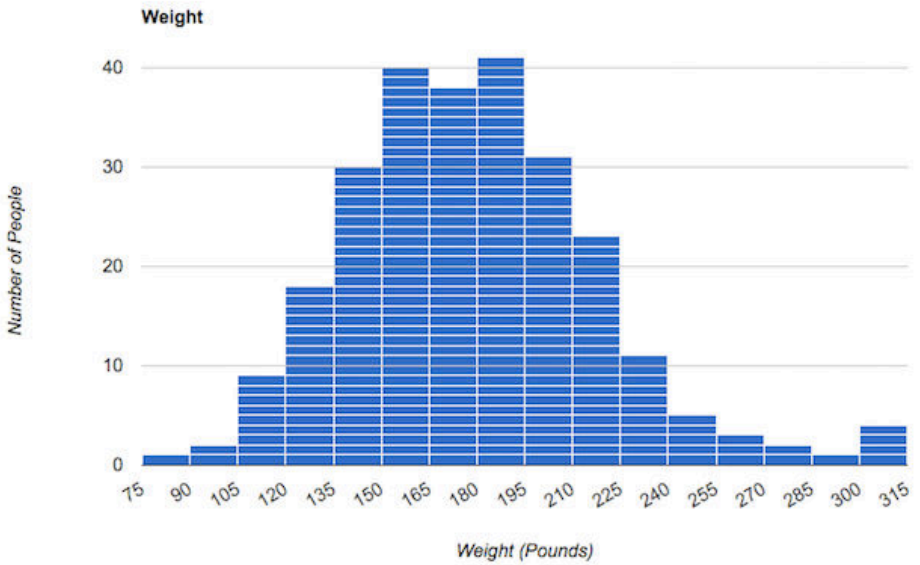


Figure 3.2e. Sex and Weight: Female

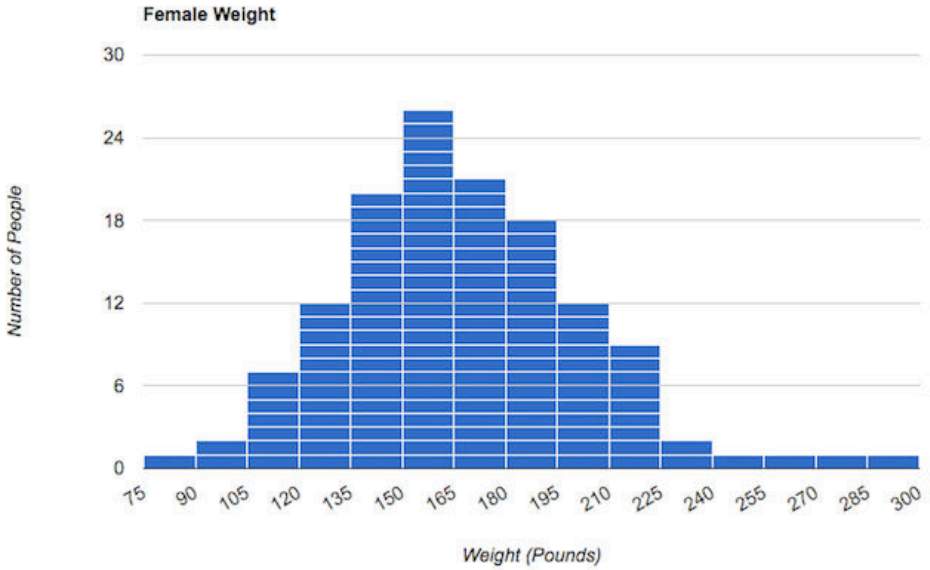
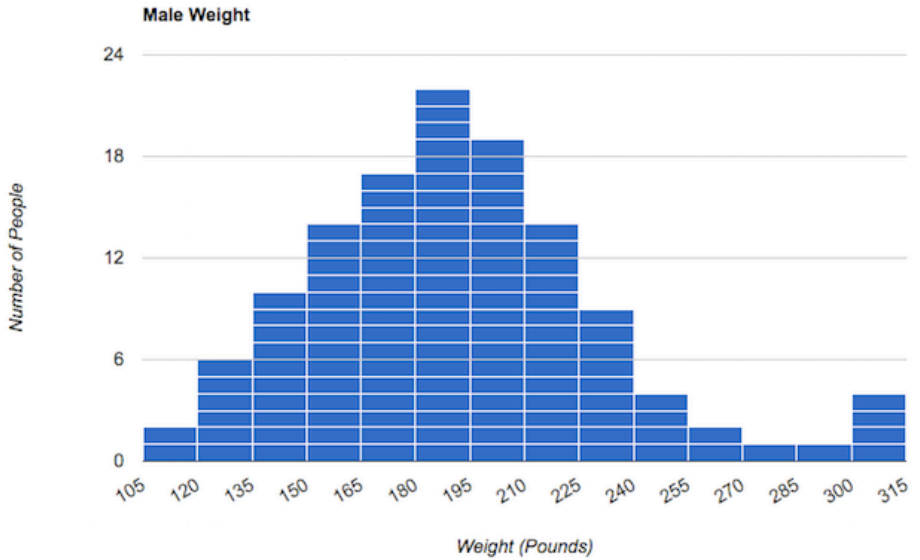


Figure 3.2f. Sex and Weight: Male



3.3.4.2 Detecting Errors and Fraud in Data

There are numerous techniques for evaluating individual data items or datasets to ensure that they have not been changed or corrupted during transmission, storage, or copying. These include parity bits, check digits, check sums, and cryptographic hash functions. They share the idea that a calculation will yield some particular value or match a stored result when the original data has not been changed. Another basic technique for detecting errors is to look for data values that are different or anomalous because they do not fall into expected ranges or categories.

More interesting challenges arise when the data might have been changed by intentional actions to commit fraud, launder money, or carry out some other crime. In these situations, the person tampering with data or creating fake data will try to make the data look normal or expected.

Forensic accountants and statisticians use many techniques for detecting possibly fraudulent data in these adversarial contexts. Some are quite simple:

- If expenses are reimbursed up to some maximum allowed value, look for data items with that exact value.
- When any value exceeding some threshold triggers more careful analysis, look for other data items just below that threshold.
- When invoices or claims are paid on receipt, and only a sample are subsequently audited, look for duplicate submissions.
- Calculate the ratio of the maximum to the minimum value for purchases in some category (such as the unit price paid for items from suppliers); items with large ratios might indicate fraud where the supplier "kicks back" some of the money to the purchaser.

Benford's Law, the observation that the leading digits in data sets are distributed in a non-uniform manner, is an effective technique for detecting fraudulent data because it is based on a counter-intuitive fact not known to most fraudsters, who often make up data to look random. You might think that the number 1 would occur 11% of the time as the first digit (since there are 9 possibilities), but for data sets whose values span several orders of magnitude, the number 1 is the first digit about 30% of the time, and 7, 8, and 9 occur around 5%.

Because of the very high transaction rate and the relatively small probability of fraud, credit card fraud is detected using machine learning algorithms. The classifier is trained with known good and bad transactions using properties like average amount, frequency, and location to develop a model of each cardholder's "data behavior" so that a transaction can quickly be assigned a probability that it is fraudulent. (More about this kind of computational classification in Chapter 7, *Categorization: Describing Resource Classes and Types*.)[86][DS]

3.3.5 Organizing with Multiple Resource Properties

Multiple properties of the resources, the person organizing or intending to use them, and the social and technological environment in which they are being organized can collectively shape their organization. For example, the way you organize your home kitchen is influenced by the physical layout of counters, cabinets, and drawers; the dishes you cook most often; your skills as a cook, which may influence the number of cookbooks, specialized appliances and tools you own and how you use them; the sizes and shapes of the packages in the pantry and refrigerator; and even your height.

If multiple resource properties are considered in a fixed order, the resulting arrangement forms a *logical hierarchy*. The top level categories of resources are created based on the values of the property evaluated first, and then each category is further subdivided using other properties until each resource is classified in only a single category. Consider the hierarchical system of folders used by a professor to arrange the digital resources on his computer; the first level distinguishes personal documents from work-related documents; work is then subdivided into teaching and research, teaching is subdivided by year, and year divided by course.

For physical resources, mapping categories to physical locations is another required step; for example, resources in the "kitchen utensils" category might all be arranged in drawers near a workspace, with "silverware" arranged more precisely to separate knives, forks, and spoons.

An alternative to hierarchical organization that is often used in digital organizing systems is *faceted classification*, in which the different properties for the resources can be evaluated in any order. For example, you can select wines from the wine.com store catalog by type of grape, cost, or region and consider these property facets in any order. Three people might each end up choosing the same moderately-priced Kendall Jackson California Chardonnay, but one of them might have started the search based on price, one based on the grape varietal, and the third with the region. This kind of interaction in effect generates a different logical hierarchy for every different combination of property values, and each user made his final selection from a different set of wines.

Faceted classification allows a collection of description resources to be dynamically re-organized into as many categories as there are combinations of values on the descriptive facets, depending on the priority or point of view the user applies to the facets. Of course this only works because the physical resources are not themselves being rearranged, only their digital descriptions.

Applications that organize large collections of digital information, including those for search, natural language processing, image classification, personal-

ized recommendation, and other computationally intensive domains, often use huge numbers of resource properties (which are often called "features" or "dimensions"). For example, in document collections each unique word might initially be treated as a feature by machine learning algorithms, so there might be tens of thousands of features.

Chapter 8, *Classification: Assigning Resources to Categories* explains principles and methods for hierarchical and faceted classification in more detail.

3.4 Designing Resource-based Interactions

We need to focus on the interactions that are enabled because of the intentional acts of description or arrangement that transform a collection of resources into an organizing system. With physical resources, it is easy to distinguish the interactions that are designed into and directly supported by an organizing system because of intentional acts of description or arrangement from those that can take place with resources after they have been accessed. For example, when a book is checked out of a library it might be read, translated, summarized, criticized, or otherwise used—but none of these interactions would be considered a capability of the book that had been designed into the library. Some physical resources can initiate interactions, as surely "human resources" and "smart" objects with sensors and other capabilities can, but most physical resources are passive. We will discuss this idea of resource *agency* in §4.2.3.

In contrast, in organizing systems that contain digital resources the logical boundary between the resources and their interactions is less clear because what you can do with a digital resource is often not apparent. Furthermore, some of the interactions that are outside of the boundary with physical resources can be inside of it with digital ones. For example, when you check a printed book out of the library, it is no longer in the library when you translate it. But a digital book in the Google Books library is not removed when you start reading it, and a language translation service runs "inside" of it.

Additional issues in the design of interactions with resources are whether users have direct or mediated access to the resources, and whether they interact with the resources themselves or only with copies or descriptions of them. For example, users have direct access to original resources in a collection when they browse through library stacks or wander in museum galleries. Users have mediated or indirect access when they use catalogs or search engines. Because digital resources can be easily reproduced, it can be difficult to distinguish a copy from the original, which raises questions of authenticity we will discuss in §4.5.3.

3.4.1 Affordance and Capability

The concept of *affordance*, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor's capabilities and cognition, what can be done with the resource.[88][CogSci]

Including capabilities and cognition brings accessibility considerations into the definition of affordance. A resource is only accessible when it supports interactions, and it is ineffective design to implement interactions with resources that some people are unable to perform. A person who cannot see text cannot read it, or if they are confined to a wheelchair they cannot select a book from a tall library shelf. Describing or transforming resources to ensure their accessibility is discussed in greater detail in §3.4.2.3 Accessibility (page 129).

When organizing resources involves arranging physical resources using boxes, bins, cabinets, or shelves, the affordances and the implications for access and use can be easily perceived. Resources of a certain size and weight can be picked up and carried away. Books on the lower shelves of bookcases are easy to reach, but those stored ten feet from the ground cannot be easily accessed.

We can analyze the organizing systems with physical resources to identify the affordances and the possible interactions they imply. We can compare the affordances or overall interaction *capability* enabled by different organizing systems for some type of physical resources, and we often do this without thinking about it. The tradeoffs between the amount of work that goes into organizing a collection of resources and the amount of work required to find and use them are inescapable when the resources are physical objects or information resources are in physical form. We can immediately see that storing information on scrolls does not enable the random access capability that is possible with books.

What and how to count to compare the capabilities of organizing systems becomes more challenging the further we get from collections of static physical resources, like books or shoes, where it is usually easy to perceive and understand the possible interactions. With computers, information systems, and digital resources in general, considerations about affordances and capabilities are not as straightforward.

First, the affordances we can perceive might not be tied to any useful interaction. Donald Norman joked that every computer screen within reaching distance affords touching, but unless the display is touch-sensitive, this affordance only benefits companies that sell screen-cleaning materials.[89][CogSci]

Second, most of the interactions that are supported by digital resources are not apparent when you encounter them. You cannot tell from their names, but you probably know from past experience what interactions are possible with files of

types ".doc" and ".pdf." You probably do not know what interactions take place with ".xpi" and ".mobi" files.[90][Com]

A similar difficulty exists when we look at resource descriptions and data collections, where we often cannot tell just by examining their values what kinds of interactions and operations with them are sensible. Think of all the different kinds of information that might be associated with a collection of people like the students in a university. A database might contain student names, student IDs, gender, birth dates, addresses, a numeric code for academic major, course units completed, grade point average, and other information. These pieces of information differ in their data type; some are integers, some are real numbers, some are Boolean, and some are just text strings. The numeric data also differs in the level of measurement it represents. Student IDs and the academic major codes are nominal data, the house or apartment number in the address is ordinal data, and the course units and grade point average are interval data. Data type and level of measurement influence the kind of interactions that are meaningful; we can create an alphabetical list of students using their last names, count up the number of students with the same academic major, and calculate the average GPA or units completed. But it makes no sense to use the numeric codes for academic major to compute an average major.

Once you have discovered it, the *capability* of digital resources and information systems can be assessed by counting the number of functions, services, or application program interfaces. However, this very coarse measure does not take into account differences in the capability or generality of a particular interaction. For example, two organizing systems might both have a search function, but differences in the operators they allow, the sophistication of pre-processing of the content to create index terms, or their usability can make them vastly differ in power, precision, and effectiveness.[91][Com]

An analogous measure of *functional capability* for a system with dynamic or living resources is the *behavioral repertoire*, the number of different activities, or range of actions, that can be initiated.

We should not assume that supporting more types of interactions necessarily makes a system better or more capable; what matters is how much value is created or invoked in each interaction. A smartphone cluttered with features and apps you never use enables a great many interactions, but most of them add little value. Doors that open automatically when their sensors detect an approaching person do not need handles or require explicit interactions. Organizing systems can use stored or computed information about user preferences or past interactions to anticipate user needs or personalize recommendations. This has the effect of substituting information for interaction to make interactions unnecessary or simpler.

For example, a “smart travel agent” service can use a user’s appointment calendar, past travel history, and information sources like airline and hotel reservation services to transform a minimal interaction like “book a business trip to New York for next week’s meeting” into numerous hidden queries that would have otherwise required separate interactions. These queries are interconnected by logical or causal dependencies that are represented by information that overlaps between them. For example, all travel-related services (airlines, hotels, ground transportation) need the traveler’s identity and the time and location of his travel. A New York trip might involve all of these services, and they need to fit together in time and location for the trip to make sense. The hotel reservation needs to begin the day the flight arrives in the destination city, the limousine service needs to meet the traveler shortly after the plane lands, and the restaurant reservation should be convenient in time and location to the hotel.[92][DS]

3.4.2 Interaction and Value Creation

A useful way to distinguish types of interactions with resources is according to the way in which they create value, using a classification proposed by Apte and Mason. They noted that interactions differ not just in their overall intensity but in the absolute and relative amounts of physical manipulation, interpersonal or empathetic contact, and symbolic manipulation or information exchange involved in the interaction.

Furthermore, Apte and Mason recognized that the proportions of these three types of value creating activities can be treated as design parameters, especially where the value created by retrieving or computing information could be completely separated from the value created by physical actions and person-to-person encounters. This configuration of value creation enables automated self-service, in which the human service provider can be replaced by technology, and outsourcing, in which the human provider is separated in space or time from the customer.[93][Bus]

3.4.2.1 Value Creation with Physical Resources

Physical manipulation is often the intrinsic type of interaction with collections of physical resources. The resource might have to be handled or directly perceived in order to interact with it, and often the experience of interacting with the resource is satisfying or entertaining, making it a goal in its own right. People often visit museums, galleries, zoos, animal theme parks or other institutions that contain physical resources because they value the direct, perceptual, or otherwise unmediated interaction that these organizing systems support.

Physical manipulation and interpersonal contact might be required to interact with information resources in physical form like the printed books in libraries.

However, for most people the primary purpose of interacting with a library is to access the information contained in its resources. Many people prefer accessing digital documents or books to accessing the original physical resource because the incidental physical and interpersonal interactions are unnecessary. In addition, many library searches are for known items, which is easily supported by digital search.

In some organizing systems robotic devices, computational processes, or other entities that can act autonomously with no need for a human agent carry out interactions with physical resources. Robots have profoundly increased efficiency in materials management, "picking and packing" in warehouse fulfillment, office mail delivery, and in many other domains where human agents once located, retrieved, and delivered physical resources. A "library robot" system that can locate books and grasp them from the shelves can manage seven times as many books in the same space used by conventional open stacks.

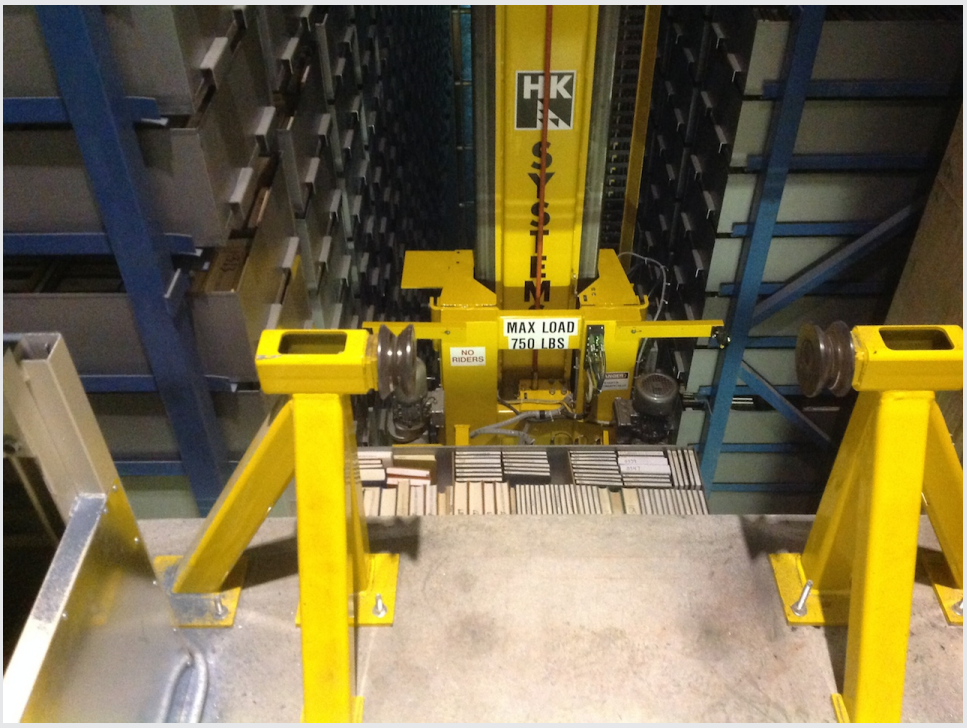
Interactions with physical resources often have highly tangible results; in the preceding examples of fulfillment and delivery interactions, resources move from one location to another. However, an abstract or architectural perspective on interaction design and value creation can create more flexibility in carrying out the interactions while still producing the expected value for the user. In general, more abstract descriptions of interactions and services allow for transparent substitution of the implementation, potentially enabling a computational process to be a substitute for one carried out by a person, or vice versa.

For example, a user buying from an internet-based store need not know and probably does not care which service delivers the package from the warehouse. Presenting the interaction to the shopper as the "delivery service" rather than as a "FedEx" or "UPS" service allows the retailer to choose the best service provider for each delivery. Going even further, if you need printed documents at a conference, sales meeting, or anywhere other than your current location, the interaction you desire is "provide me with documents" and not "deliver my documents." It does not matter that FedEx will print your documents at their destination rather than shipping them there.

3.4.2.2 Value Creation with Digital Resources

With digital resources, neither physical manipulation nor interpersonal contact is required, and the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource. Put another way, by replacing interactions that involve people and physical resources with symbolic ones, organizing systems can lower costs without reducing user satisfaction. This is why so many businesses have automated their information-intensive processes with self-service technology.

Library Robot



An automated robot library system at San Francisco State University.

The automated robot library system installed by the Dematic Group stores books in bins stacked on three-story-tall metal racks in five long aisles. Instead of using a library classification scheme, books are stored according to their sizes in one-foot deep metal bins, which contain about one hundred books each. Given an online catalog request for a book, the system looks up the bin where it was last stored, and then directs a robot to bring that bin to the circulation desk. Human librarians then find the requested book in the bin and scan its barcode, which notifies the requester that the book can be picked up. To store a book, the librarian scans its barcode, and it is then stored in the closest bin with available space.

(Photo by Scott Abel. Used with permission.)

Similarly, web search engines eliminate the physical effort required to visit a library and enable users to consult more readily accessible digital resources. A search engine returns a list of the page titles of resources that can be directly accessed with just another click, so it takes little effort to go from the query results to the primary resource.[98][Web]

The ease of use and speed of search engines in finding web resources creates the expectation that any resource worth looking at can be found on the web. This is certainly false, or Google would never have begun its ambitious and audacious project to digitize millions of books from research libraries. While research libraries strive to provide access to authoritative and specialized resources, the web is undeniably good enough for answering most of the questions ordinary users put to search engines, which largely deal with everyday life, popular culture, personalities, and news of the day.

Libraries recognize that they need to do a better job integrating their collections into the "web spaces" and web-based activities of their users if they hope to change the provably suboptimal strategies of "information foraging" most people have adopted that rely too much on the web and too little on the library.[99][Web] Some libraries are experimenting with *Semantic Web* and "Linked Data" technologies that would integrate their extensive bibliographic resources with resources on the open web.[100][Web]

Museums have aggressively embraced the web to provide access to their collections. While few museum visitors would prefer viewing a digital image over experiencing an original painting, sculpture, or other physical artifact, the alternative is often no access at all. Most museum collections are far larger than the space available to display them, so the web makes it possible to provide access to otherwise hidden resources.

The variety and functions of interactions with digital resources are determined by the amount of structure and semantics represented in their digital encoding, in the descriptions associated with the resources, or by the intelligence of the computational processes applied to them. Digital resources can support enhanced interactions of searching, copying, zooming, and other transformations. Digital or "ebooks" demonstrate how access to content can be enhanced once it is no longer tied to the container of the printed book, but ebook readers vary substantially in their interaction repertoires; the baseline they all share is "page turning," resizing, and full-text search.[102][Com]

To augment digital resources with text structures, multimedia, animation, interactive 3-D graphics, mathematical functions, and other richer content types requires much more sophisticated representation formats that tend to require a great deal of "hand-crafting." An alternative to hand-crafted resource description is sophisticated computer processing guided by human inputs. For example, Facebook and many web-based photo organizing systems implement face recognition analysis that detects faces in photos, compares features of detected faces to features of previously identified faces, and encourages people to tag photos to make the recognition more accurate. Some online services use similar image classification techniques to bring together shoes, jewelry, or other items that look alike.

Richer interactions with digital text resources are possible when they are encoded in an application or presentation-independent format. Automated content reuse and "single-source publishing" is most efficiently accomplished when text is encoded in XML, but much of this XML is produced by transforming text originally created in word processing formats. Once it is in XML, digital information can be distributed, processed, reused, transformed, mixed, remixed, and recombined into different formats for different purposes, applications, devices, or users in ways that are almost impossible to imagine when it is represented in a tangible (and therefore static) medium like a book on a shelf or a box full of paper files.[103][Com]

Businesses that create or own their information resources can readily take advantage of the enhanced interactions that digital formats enable. For libraries, however, copyright is often a barrier to digitization, both as a matter of law and because digitization enables copyright enforcement to a degree not possible with physical resources.

As a result, digital books are somewhat controversial and problematic for libraries, whose access models were created based on the economics of print publication and the social contract of the copyright *first sale* doctrine that allowed libraries to lend printed books.[104][Law]

Software-based agents do analogous work to robots in "moving information around" after accessing digital resources such as web services or physical resources with sensors attached that produce digital information. Agents can control or choreograph a set of interactions with digital resources to carry out complex business processes.

3.4.2.3 Accessibility

The United Nations Convention on the Rights of Persons with Disabilities recognizes accessibility to information and communications technologies as a basic human right. There is also a strong business case for accessibility: studies show that accessible websites are used more often, are easier to maintain, and produce better search results.[105][Phil]

Many of the techniques for making a resource accessible involve transforming the resource or its description into a different form so someone who could not perceive it or interact with it in its original form can now do so. The most common operating systems all come with general-purpose accessibility features such as reading text aloud, recognizing speech, magnifying text, increasing cursor size, signaling with flashing lights instead of with sounds, lights to signal keyboard shortcuts for selecting and navigating, and connecting to devices for displaying Braille. Google Translate converts text in one language to another, and many people use it to create a rough draft that is finished by a human translator.[106][Com]

Other techniques are not generic and automatic, and instead require investment by authors or designers to make information accessible. Websites are more accessible when images or other non-text content types have straightforward titles, captions, and "alt text" that describes what they are about. Consistent placement and appearance of navigation controls and interaction widgets is essential; for example, in a shopping site "My Cart" might always be found at the top right corner of the page.[107][Web]

If authors apply semantic and structural markup to the text and use formats that distinguish it from presentation instructions, page outlines and summaries can be generated to enhance navigation, and search can be made more precise by limiting it to particular sections or content types. As the "Information IQ" of the source format increases, more can be done to make it more accessible (see §4.2.2 Resource Format (page 169) and Figure 4.3, Information IQ.).[108][Web]

The Smithsonian Museum in Washington, DC invites visitors to record audio descriptions on mobile devices of the nearly 137 million objects in its collection, and then makes these available to everyone. This is just a small part of its efforts to make its exhibits more accessible. A company called D-Scriptive enables blind people to enjoy Broadway shows more by recording hundreds of audio descriptions that are synchronized with dialog spoken by the actors.

Transforming recorded spoken language to text to make it accessible and searchable is called transcription. At times transcription is necessary to comply with accessibility requirements, but is often done simply to add organization to content, as when a script is created to separate the multiple voices in a radio or television interview or story.

Transcriptions created by skilled people are highly accurate but labor-intensive to produce, so speech-to-text software is increasingly being used to transcribe speech using pre-trained acoustic and language models. Training these models is computationally intensive, and there are many clever techniques to acquire the "labeled" inputs. However, most of them are conceptually simple; they take the huge amount of data collected by voice search applications and analyze what the searcher does with the results to assess the accuracy of the transcription. Transcription accuracy can be improved when models can be specialized by industry or application. For example, speech-to-text software for doctors is trained to recognize medical terminology, while software for use by generic voice recognition services like Apple's is trained to understand dictation and commands or questions one would ask of a smartphone.

Since text transcripts are machine-readable, unlike audio or video files, adding text transcripts makes it possible for search engines to index audio and video in ways that were previously impossible. Pop Up Archive, an audio search company in Oakland, California, works with speech-to-text software specially trained for news media and spoken word content to make radio, podcasts, and archival

audio searchable. A challenge for audio search is that even though a transcription with a few mistakes works just fine for search engines, people often expect transcriptions to be perfect.[110][Com]

When the speech is in a language that is not understood, it needs to be translated as well. Perhaps you have watched a movie on an international flight and were able to choose from subtitles in many different languages. Creating subtitles for a foreign film is an asynchronous task that is substantially easier task than doing a real-time translation, and the demand for skilled translators for speeches and other synchronous situations (and interpreters, who translate speech to sign language for people with hearing disabilities) remains high.

3.4.3 Access Policies

Different levels of interactions or access can apply to different resources in a collection or to different categories of users. For example, library collections can range from completely open and public, to allowing limited access, to wholly private and restricted.

Because of their commercial and competitive purposes, organizing systems in business domains are more likely to enforce a granular level of access control that distinguishes people according to their roles and the nature of their interactions with resources. For example, administrative assistants in a company's Human Resources department are not allowed to see salaries; HR employees in a benefits administration role can see salaries but not change them; management-level employees in HR can change the salaries. Some firms limit access to specific times from authorized computers or IP addresses.[111][Bus]

A noteworthy situation arises when the person accessing the organizing system is the one who designed and implemented it. In this case, the person will have qualitatively better knowledge of the resources and the supported interactions. This situation most often arises in the organizing systems in kitchens, home closets, and other highly personal domains but can also occur in knowledge-intensive business and professional domains like consulting, customer relationship management, and scientific research.

Many of the organizing systems used by individuals are embedded in physical contexts where access controls are applied in a coarse manner. We need a key to get into the house, but we do not need additional permissions or passwords to enter our closets or to take a book from a bookshelf. In our online lives, however, we readily accept and impose more granular access controls. For example, we might allow or block individual "friend" requests on Facebook or mark photos on Flickr as public, private, or viewable only by named groups or individuals.

We can further contrast access policies based on their origins or motivations.

Designed resource access policies are established by the designer or operator of an organizing system to satisfy internally generated requirements. Examples of designed access policies are:

- giving more access to "inside" users (e.g., residents of a community, students or faculty members at a university, or employees of a company) than to anonymous or "outside" users;
- giving more access to paying users than to users who do not pay;
- giving more access to users with capabilities or competencies that can add value to the organizing system (e.g., material culture researchers like archaeologists or anthropologists, who often work with resources in museum collections that are not on display).

Imposed Policies are mandated by an external entity and the organizing system must comply with them. For example, an organizing system might have to follow information privacy, security, or other regulations that restrict access to resources or the interactions that can be made with them.

University libraries typically complement or replace parts of their print collections with networked access to digital content licensed from publishers. Typical licensing terms then require them to restrict access to users that are associated with the university, either by being on campus or by using virtual private network (VPN) software that controls remote access to the library network.[112][Law] Copyright law limits the uses of a substantial majority of the books in the collections of major libraries, prohibiting them from being made fully available in digital formats. Museums often prohibit photography because they do not own the rights to modern works they display.

Whether an access policy is designed or imposed is not always clear. Policies that were originally designed for a particular organizing system may over time become best practices or industry standards, which regulators or industry groups not satisfied with "self-regulation" later impose. Museums might aggressively enforce a ban on photography not just to comply with copyright law, but also to enhance the revenue they get from selling posters and reproductions.

3.5 Maintaining Resources

Maintaining resources is an important activity in every organizing system because resources must be available at the time they are needed. Beyond these basic shared motivations are substantial differences in *maintenance* goals and methods depending on the domain of the organizing system.

However, different domains sometimes use the same terms to describe different *maintenance* activities and different terms for similar activities. Common maintenance activities are *storage*, *preservation*, *curation*, and *governance*. Storage is most often used when referring to physical or technological aspects of maintaining resources; backup (for short-term storage), archiving (for long-term storage), and migration (moving stored resources from one storage device to another) are similar in this respect. The other three terms generally refer to activities or methods that more closely overlap in meaning; we will distinguish them in §3.5.2 through §3.5.4.

Selection and maintenance are interdependent. Selection is based on an initial set of rules that determine which resources enter the organizing system. Maintenance includes the work to preserve the resources, the processes for evaluating and revising the original selection criteria, and the removal of resources from the system when they no longer need to be preserved. More stringent rules for selecting resources generally imply a maintenance plan that carefully enforces the same constraints that limit selection. This is just common sense whether the resource is a piece of art, an automobile, a software package, or a star basketball player; if you worked hard to find or paid a lot to acquire a resource, you are going to take care of it and will not soon be buying another one.

Ideally, *maintenance* requirements for resources should be anticipated when organizing principles are defined and implemented. Resource descriptions to support preservation of digital resources are especially important.

3.5.1 Motivations for Maintaining Resources

The concept of *memory institution* broadly applies to a great many organizing systems that share the goal of preserving knowledge and cultural heritage. The primary resources in libraries, museums, data archives or other *memory institutions* are fixed cultural, historic, or scientific artifacts that are maintained because they are unique and original items with future value. This is why the Musée du Louvre preserves the portrait of the *Mona Lisa* and the United States National Archives preserves the *Declaration of Independence*.

In contrast, in businesses organizing systems, many of the resources that are collected and managed have limited intrinsic value. The motivation for preservation and maintenance is economic; resources are maintained because they are

essential in running the business. For example, businesses collect and preserve information about employees, inventory, orders, invoices, etc., because it ensures internal goals of efficiency, revenue generation, and competitive advantage. The same resources (e.g., customer information) are often used by more than one part of the business.[116][Bus] Maintaining the accuracy and consistency of changing resources is a major challenge in business organizing systems.[117][DS]

Many business organizing systems preserve information needed to satisfy externally imposed regulatory or compliance policies and serve largely to avoid possible catastrophic costs from penalties and lawsuits. In all these cases, resources are maintained as one of the means employed to preserve the business as an ongoing enterprise, not as an end in itself.

Unlike libraries, archives, and museums, indefinite preservation is not the central goal of most business organizing systems. These organizing systems mostly manage information needed to carry out day-to-day operations or relatively recent historical information used in decision support and strategic planning. In addition to these internal mandates, businesses have to conform to securities, taxation, and compliance regulations that impose requirements for long-term information preservation.[118][Law]

In between these contrasting purposes of preservation and maintenance are the motives in personal collections, which occasionally are created because of the inherent value of the items but more typically because of their value in supporting personal activities. Some people treasure old photos or collectibles that belonged to their parents or grandparents and imagine their own children or grandchildren enjoying them, but many old collections seem to end up as offerings on eBay. In addition, many personal organizing systems are task-oriented, so their contents need not be preserved after the task is completed.[120][CogSci]

3.5.2 Preservation

At the most basic level, *preservation* of resources means maintaining them in conditions that protect them from physical damage or deterioration. Libraries, museums, and archives aim for stable temperatures and low humidity. Permanently or temporarily out-of-service aircraft are parked in deserts where dry conditions reduce corrosion. Risk-aware businesses create continuity plans that involve off-site storage of the data and documents needed to stay in business in the event of a natural disaster or other disruption.

When the goal is indefinite preservation, other *maintenance* issues arise if resources deteriorate or are damaged. How much of an artifact’s worth is locked in with the medium used to express it? How much restoration should be attempted? How much of an artifact's essence is retained when digitized?

3.5.2.1 Digitization and Preserving Resources

Preservation is often a key motive for *digitization*, but digitization alone is not preservation. Digitization creates preservation challenges because technological obsolescence of computer software and hardware require ongoing efforts to ensure the digitized resources can be accessed.

Technological obsolescence is the major challenge in maintaining digital resources. The most visible one is a result of the relentless evolution of the physical media and environments used to store digital information in both institutional or business and personal organizing systems. Computer data began to be stored on magnetic tape and hard disk drives six decades ago, on floppy disks four decades ago, on CDs three decades ago, on DVDs two decades ago, on solid-state drives half a decade ago, and in "cloud-based" or "virtual" storage environments in the last decade. As the capacity of storage technologies grows, economic and efficiency considerations often make the case to adopt new technology to store newly acquired digital resources and raise questions about what to do with the existing ones.[121][Com]

The second challenge might seem paradoxical. Even though digital storage capacity increases at a staggering pace, the expected useful lifetimes of the physical storage media are measured in years or at best in decades. Colloquial terms for this problem are *data rot* or "bit rot." In contrast, books printed on acid-free paper can last for centuries. The contrast is striking; books on library shelves do not disappear if no one uses them, but digital data can be lost if no one wants access to it within a year or two after its creation.[122][Com]

However, limits to the physical lifetime of digital storage media are much less significant than the third challenge, the fact that the software and its associated computing environment used to parse and interpret the resource at the time of preservation might no longer be available when the resource needs to be accessed. Twenty-five years ago most digital documents were created using the Word Perfect word processor, but today the vast majority is created using Microsoft Word and few people use Word Perfect today. Software and services that convert documents from old formats to new ones are widely available, but they are only useful if the old file can be read from its legacy storage medium.

Because almost every digital device has storage associated with it, problems posed by multiple storage environments can arise at all scales of organizing systems. Only a few years ago people often struggled with migrating files from their old computer, music player or phone when they got new ones. Web-based email and applications and web-based storage services like Dropbox, Amazon Cloud Drive, and Apple iCloud eliminate some data storage and migration problems by making them someone else's responsibility, but in doing so introduce privacy and reliability concerns.

It is easy to say that the solutions to the problems of digital preservation are regular recopying of the digital resources onto new storage media and then migrating them to new formats when significantly better ones come along. In practice, however, how libraries, businesses, government agencies or other enterprises deal with these problems depends on their budgets and on their technical sophistication. In addition, not every resource should or can always be migrated, and the co-existence of multiple storage technologies makes an organizing system more complex because different storage formats and devices can be collectively incompatible.

(Interoperability and integration are discussed in Chapter 10, *Interactions with Resources*.)

3.5.2.2 Preserving the Web

Preservation of web resources is inherently problematic. Unlike libraries, museums, archives, and many other kinds of organizing systems that contain collections of unchanging resources, organizing systems on the web often contain resources that are highly dynamic. Some websites change by adding content, and others change by editing or removing it.[124][Web]

Longitudinal studies have shown that hundreds of millions of web pages change at least once a week, even though most web pages never change or change infrequently.[125][Web] Nevertheless, the continued existence of a particular web page is hardly sufficient to preserve it if it is not popular and relevant enough to show up in the first few pages of search results. Persistent access requires preservation, but preservation is not meaningful if there is no realistic probability of future access.

Comprehensive web search engines like Google and Bing use crawlers to continually update their indexed collections of web pages and their search results link to the current version, so preservation of older versions is explicitly not a goal. Furthermore, search engines do not reveal any details about how frequently they update their collections of indexed pages.[126][Web]

3.5.2.3 Preserving Resource Instances

A focus on preserving particular resource instances is most clear in museums and archives, where collections typically consist of unique and original items. There are many copies and derivative works of the *Mona Lisa,* but if the original *Mona Lisa* were destroyed none of them would be acceptable as a replacement.

Zoos often give a distinctive or attractive animal a name and then market it as a special or unique instance. For example, the Berlin Zoo successfully marketed a polar bear named Knut to become a world famous celebrity, and the zoo made millions of dollars a year through increased visits and sales of branded

The Internet Archive and the "Wayback Machine"

The Internet Archive (Archive.org), founded by Brewster Kahle, makes preservation of the web its first and foremost activity, and when you enter a URI into its "Wayback Machine" you can see what a site looked like at different moments in time. For example, `www.berkeley.edu` was archived about 2500 times between October 1996 and January 2013, including about twice a week on average during all of 2012. Even so, since a large site like berkeley.edu often changes many times a day, the Wayback Machine's preservation of berkeley.edu is incomplete, and it only preserves a fraction of the web's sites. Since 2006 the Internet Archive has hosted the "Archive-It" service to enable hundreds of schools, libraries, historical societies, and other institutions to archive collections of digital resources.

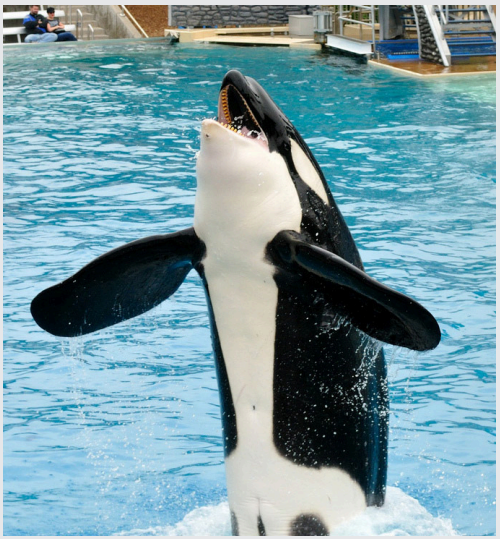
merchandise. Merchandise sales have continued even though Knut died unexpectedly in March 2011, which suggests that the zoo was less interested in preserving that particular polar bear than in preserving the revenue stream based on that resource.

Most business organizing systems, especially those that "run the business" by supporting day-to-day operations, are designed to preserve instances. These include systems for order management, customer relationship management, inventory management, digital asset management, record management, email archiving, and more general-purpose document management. In all of these domains, it is often necessary to retrieve specific information resources to serve customers or to meet compliance or traceability goals.

Recent developments in sensor technology enable very extensive data collection about the state and performance of machines, engines, equipment, and other types of physical resources, including human ones. (Are you wearing an activity tracker right now?) When combined with historical information about maintenance activity, predictive analytics techniques can use this data to determine normal operating ranges and indicators of coming performance degradation or failures. Predictive maintenance can maximize resource lifetimes while minimizing maintenance and inventory costs. These techniques have recently been used to predict when professional basketball players are at risk of an injury, potentially enabling NBA teams to identify the best time to rest their star players without impairing their competitive strategy.[131][DS]

3.5.2.4 Preserving Resource Types

"Shamu" the Killer Whale



This photo of "Shamu" was taken at one of the three Sea World marine parks in the US, but it does not matter which one because each of them has a killer whale (orca) performing there called Shamu. Similarly, it does not matter when this photo was taken because if a particular orca dies, it is replaced by another that also performs using Shamu as a stage name.

(Photo by Mike Saechang. Creative Commons CC BY-SA 2.0 license.)

Some business organizing systems are designed to preserve types or classes of resources rather than resource instances. In particular, systems for content management typically organize a repository of reusable or "source" information resources from which specific "product" resources are then generated. For example, content management systems might contain modular information about a company's products that are assembled and delivered in sales or product catalogs, installation guides, operating guides, or repair manuals.[132][Com]

Businesses strive to preserve the collective knowledge embodied in the company's people, systems, management techniques, past decisions, customer relationships, and intellectual property. Much of this knowledge is "know how"—knowing how to get things done or knowing how things work—that is tacit or informal. *Knowledge management systems (KMS)* are a type of business organizing system whose goal is to capture and systematize these information resources.[133][Bus] As with content management, the focus of knowledge management is the reuse of "knowledge as type," putting the focus on the knowledge rather than the specifics of how it found its way into the organizing system.

Libraries have a similar emphasis on preserving resource types rather than instances. The bulk of most library collections, especially public libraries, is made up of books that have many equivalent copies in other collections. When a library has a copy of *Moby Dick* it is preserving the abstract *work* rather than the particular physical *instance*—unless the copy of *Moby Dick* is a rare first edition signed by Melville.

Even when zoos give their popular animals individual names, it seems logical that the zoo's goal is to preserve animal species rather than instances because any particular animal has a finite lifespan and cannot be preserved forever.[134][Bus]

3.5.2.5 Preserving Resource Collections

In some organizing systems any specific resource might be of little interest or importance in its own right but is valuable because of its membership in a collection of essentially identical items. This is the situation in the data warehouses used by businesses to identify trends in customer or transaction data or in the huge data collections created by scientists. These collections are typically analyzed as complete sets. A scientist does not borrow a single data point when she accesses a data collection; she borrows the complete dataset consisting of millions or billions of data points. This requirement raises difficult questions about what additional software or equipment need to be preserved in an organizing system along with the data to ensure that it can be reanalyzed.[135][DS]

3.5.3 Curation

For almost a century *curation* has referred to the processes by which a resource in a collection is maintained over time, which may include actions to improve access or to restore or transform its representation or presentation.

Curation takes place in all organizing systems—at a personal scale when we rearrange a bookshelf to accommodate new books or create new file folders for this year's health insurance claims, at an institutional scale when a museum designs a new exhibit or a zoo creates a new habitat, and at web scale when people select photos to upload to Flickr or Facebook and then tag or "Like" those uploaded by others.

An individual, company, or any other creator of a website can make decisions and employ technology that maintains the contents, quality and character of the site over time. In that respect website curation and governance practices are little different than those for the organizing systems in memory institutions or business enterprises. The key to curation is having clear policies for collecting resources and maintaining them over time that enable people and automated processes to ensure that resource descriptions or data are authoritative, accurate, complete, consistent, and non-redundant.

3.5.3.1 Institutional Curation

Curation is most necessary and explicit in institutional organizing systems where the large number of resources or their heterogeneity requires choices to be made about which ones should be most accessible, how they should be organized to ensure this access, and which ones need most to be preserved to ensure continued accessibility over time. Curation might be thought of as an ongoing or deferred selection activity because curation decisions must often be made on an item-by-item basis.

3.5.3.2 Individual Curation

Curation by individuals has been studied a great deal in the research discipline of Personal Information Management (PIM).[140][CogSci] Much of this work has been influenced for decades by a seminal article written by Vannevar Bush titled "As We May Think." Bush envisioned the Memex, "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility." Bush's most influential idea was his proposal for organizing sets of related resources as "trails" connected by associative links, the ancestor of the *hypertext links* that define today's web.[141][Com]

3.5.3.3 Social and Web Curation

Many individuals spend a great amount of time curating their own websites, but when a site can attract large numbers of users, it often allows users to annotate, "tag," "like," "+1," and otherwise evaluate its resources. The concept of curation has recently been adapted to refer to these volunteer efforts of individuals to create, maintain, and evaluate web resources.[142][Web] The massive scale of these bottom-up and distributed activities is curation by "crowdsourcing," the continuously aggregated actions and contributions of users.[143][Web]

The informal and organic "folksonomies" that result from their aggregated effort create organization and authority through network effects.[144][Web] This undermines traditional centralized mechanisms of organization and governance and threatens any business model in publishing, education, and entertainment that has relied on top-down control and professional curation.[145][Bus] Professional curators are not pleased to have the *ad hoc* work of untrained people working on websites described as curation.

Most websites are not curated in a systematic way, and the decentralized nature of the web and its easy extensibility means that the web as a whole defies curation. It is easy to find many copies of the same document, image, music file, or

video and not easy to determine which is the original, authoritative or authorized version. Broken links return "Error 404 Not Found" messages.[146][Web]

Problems that result from lazy or careless webmastering are minor compared to those that result from deliberate misclassification, falsification, or malice. An entirely new vocabulary has emerged to describe these web resources with bad intent: "spam," "phishing," "malware," "fakeware," "spyware," "keyword stuffing," "spamdexing," "META tag abuse," "link farms," "cybersquatters," "phantom sites," and many more.[147][Com] Internet service providers, security software firms, email services, and search engines are engaged in a constant war against these kinds of malicious resources and techniques.[148][Com]

Since we cannot prevent these deceptions by controlling what web resources are created in the first place, we have to defend ourselves from them after the fact. "Defensive curation" techniques include filters and firewalls that block access to particular sites or resource types, but whether this is curation or censorship is often debated, and from the perspective of the government or organization doing the censorship it is certainly curation. Nevertheless, the decentralized nature of the web and its open protocols can sometimes enable these controls to be bypassed.

3.5.3.4 Computational Curation

Search engines continuously curate the web because the algorithms they use for determining relevance and ranking determine what resources people are likely to access. At a smaller scale, there are many kinds of tools for managing the quality of a website, such as ensuring that HTML content is valid, that links work, and that the site is being crawled completely. Another familiar example is the spam and content filtering that takes place in our email systems that automatically classifies incoming messages and sorts them into appropriate folders.

One might think that computational curation is always more reliable than any curation carried out by people. Certainly, it seems that we should always be able to trust any assertion created by context-aware resources like temperature or location sensors. But can we trust the accuracy of web content? Search engines use the popularity of web pages and the structure of links between them to compute relevance. But popularity and relevance do not always ensure accuracy. We can easily find popular pages that prove the existence of UFOs or claim to validate wacky conspiracy theories.

Computational curation is more predictable than curation done by people, but search engines have long been accused of bias built into their algorithms. For example, Google's search engine has been criticized for giving too much credibility to websites with `.edu` domain names, to sites that have been around for a long time, or that are owned by or that partner with the company, like Google Maps or YouTube.[149][DS]

In organizing systems that contain data, there are numerous tools for *name matching,* the task of determining when two different text strings denote the same person, object, or other named entity. This problem of eliminating duplicates and establishing a controlled or authoritative version of the data item arises in numerous application areas but familiar ones include law-enforcement and counter-terrorism. Done incorrectly, it might mean that you end up on a "watch list" and experience difficulties every time you want to fly commercially.

An extremely promising new approach to computational curation involves using scientific measuring equipment to analyze damaged physical resources and then building software models of the resources that can be manipulated to restore the resources or otherwise improve access to their content. For example, the first sound recordings were made using rotating wax cylinders; sounds caused a diaphragm to vibrate, the pattern of vibration was transferred into a connected stylus, which then cut a groove into the wax. When the cylinder was rotated past a passive stylus, it would vibrate according to the groove pattern, and the amplified vibrations could be heard as the replayed sound. Unfortunately, wax cylinders from the 19th century are now so fragile that they would fall apart if they were played. This dilemma was resolved by Carl Haber, an experimental physicist at the Lawrence Berkeley Laboratory. Haber used image processing techniques to convert microscope-detailed scans of the grooves in the wax cylinders. Measurements of the grooves could then be transformed to reproduce the sounds captured in the grooves.

A second example of computational curation applied to digital preservation is work done by a research team led by Melissa Terras and Tim Weyrich at University College London to build a 3-dimensional model of a 17th-century "Great Parchment Book" damaged in an 18th-century fire. The parchment was singed, shriveled, creased, folded, and nearly impossible to read (see website). After traditional document restoration techniques (e.g., illustrated in photos in §3.5.2) went as far as they could, the researchers used digital image capture and modeling techniques to create a software model of the parchment that could stretch and flatten the digital document to discover text hidden by the damage.

3.5.3.5 Discarding, Removing, and Not Keeping

So far, we have discussed maintenance as activities involved in preserving and protecting resources in an organizing system over time. An essential part of maintenance is the phasing out of resources that are damaged or unusable, expired or past their effectivity dates, or no longer relevant to any interaction.

Many organizations admit to a distinct lack of strategy in the removal aspect of maintenance. A firm with outdated storage technology might have to discard older data simply to make room for new data, and might do so without considering that keeping some summary statistics would be valuable for historical analy-

sis. Other firms might be biased towards keeping information just because they went to the trouble of collecting or acquiring it. Some amount of "intelligent" removal is an essential ingredient in any maintenance regime, and a popular book argues forcefully for continually discarding resources from personal organizing systems as a method of focusing on the resources that really matter.

In memory institutions, common terms for getting rid of resources include discarding, de-accession, de-selection, and weeding.

Other domains have other mechanisms and terms for removing resources. Employess are removed by firing, layoff, or retirement . Athletes are cut or waived or sent down from a sports team if their performance deteriorates.

Keeping an organizing system current often involves some amount of elimination of older resources in order to make space for the new: in fashion retail, the floor is constantly restocked with the latest styles. Software development teams will halt active support and documentation efforts of legacy versions.

Information resources are often discarded to comply with laws about retaining sensitive data. Governments and office holders sometimes destroy documents that might prove damaging or embarrassing if they are discovered through Freedom of Information requests or by opposing political parties.

More positively, the "right to be forgotten" movement and intentional destruction of information records about prior bankruptcy, credit problems, or juvenile arrests after a certain period of time has passed can be seen as a policy of "social forgetfulness" that gives people a chance to get on with their lives.[152][Law]

Some people have difficulty in discarding things, regardless of their actual value. This behavior is called hoarding, and is now regarded as a kind of obsessive-compulsive disorder that requires treatment because it can cause emotional, physical, social, and even legal problems for the hoarder and family members. It seems unsympathetic that many TV shows and stories have been produced about especially compulsive hoarding. A famous example is that of the Collyer brothers in New York, who shut themselves off from the world for years, and when they were found dead inside their home in 1947 it contained 140 tons of collected items, including 25,000 books, fourteen pianos, thousands of bottles and tin cans, hundreds of yards of fabrics, and even a Model T car chassis.

3.5.4 Governance

Governance overlaps with *curation* in meaning, but typically has more of policy focus (what should be done), rather than a process focus (how to do it). Governance is also more frequently used to describe curation in business and scientific organizing systems rather than in libraries, archives, and museums. Governance has a broader scope than curation because it extends beyond the resources in a collection and also applies to the software, computing, and networking environments needed to use them. This broader scope also means that governance must specify the rights and responsibilities for the people who might interact with the resources, the circumstances under which that might take place, and the methods they would be allowed to use.

Corporate governance is a common term applied to the ongoing maintenance and management of the relationship between operating practices and long-term strategic goals.

Data governance policies are often shaped by laws, regulations or policies that prohibit the collection of certain kinds of objects or types of information. Privacy laws prohibit the collection or misuse of personally identifiable information about healthcare, education, telecommunications, video rental, and in some countries restrict the information collected during web browsing.[155][Law]

3.5.4.1 Governance in Business Organizing Systems

Governance is essential to deal with the frequent changes in business organizing systems and the associated activities of data quality management, access control to ensure security and privacy, compliance, deletion, and archiving. For many of these activities, effective governance involves the design and implementation of standard services to ensure that the activities are performed in an effective and consistent manner.[156][Bus]

Today's information-intensive businesses capture and create large amounts of digital data. The concept of "business intelligence" emphasizes the value of data in identifying strategic directions and the tactics to implement them in marketing, customer relationship management, supply chain management and other information-intensive parts of the business.[157][Bus] A management aspect of governance in this domain is determining which resources and information will potentially provide economic or competitive advantages and determining which will not. A conceptual and technological aspect of governance is determining how best to organize the useful resources and information in business operations and information systems to secure the potential advantages.

Business intelligence is only as good as the data it is based on, which makes business data governance a critical concern that has rapidly developed its own

specialized techniques and vocabulary. The most fundamental governance activity in information-driven businesses is identifying the "master data" about customers, employees, materials, products, suppliers, etc., that is reused by different business functions and is thus central to business operations.[158][DS]

Because digital data can be easily copied, data governance policies might require that all sensitive data be anonymized or encrypted to reduce the risk of privacy breaches. To identify the source of a data breach or to facilitate the assertion of a copyright infringement claim a digital watermark can be embedded in digital resources.[159][Com]

Stop and Think: Business Data Governance

Ebay, Target, and other large companies have had tens of millions of passwords, credit card numbers, and other sensitive personal information breached by hackers or security lapses. Consider a data breach you have heard of or experienced. What secure information was leaked? How might the business's governance policies and practices have affected the severity of the breach? What changes could the businesses make to protect people's data better?

3.5.4.2 Governance in Scientific Organizing Systems

Scientific data poses special *governance* problems because of its enormous scale, which dwarfs the datasets managed in most business organizing systems. A scientific data collection might contain tens of millions of files and many petabytes of data. Furthermore, because scientific data is often created using specialized equipment or computers and undergoes complex workflows, it can be necessary to curate the technology and processing context along with data in order to preserve it. An additional barrier to effective scientific data curation is the lack of incentives in scientific culture and publication norms to invest in data retention for reuse by others. [160][Law]

The Long Tail of Dark Data

Almost all scientists admit that they are holding "dark data," data that has never been made available to the rest of the scientific community. There may only be a few scientists worldwide that would want to see a particular dataset, but there are many thousands of these datasets. Other dark data comes from research that fails to find effects; because these negative findings are less likely to be published, literature reviews can be skewed by their omission. Just as Netflix makes the long tail of movies available, perhaps dark data would become more accessible if it could be could easily uploaded to a Netflix for Science. (Heidorn 2008)

3.6 Key Points in Chapter Three

- Selection, organizing, interaction design, and maintenance activities occur in every organizing system.

  (See §3.1 Introduction (page 87))

- These activities are not identical in every domain, but the general terms enable communication and learning about domain-specific methods and vocabularies.

  (See §3.1 Introduction (page 87))

- The most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized.

  (See §3.2 Selecting Resources (page 92))

- Memory institutions select rare and distinctive resources, but in scientific research, a sample must contain representative instances.

  (See §3.2 Selecting Resources (page 92))

- Even when the selection principles behind a collection are clear and consistent, they can be unconventional, idiosyncratic, or otherwise biased.

  (See §3.2.1 Selection Criteria (page 92))

- If you can determine where the resources come from, you can make better selection decisions by evaluating the people, processes, and organizing systems that create them.

  (See §3.2.2 Looking “Upstream” and “Downstream” to Select Resources (page 96))

- In this book we use *property* in a generic and ordinary sense as a synonym for *feature* or “characteristic.” Many cognitive and computer scientists are more precise in defining these terms and reserve *property* for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the *property* is called an *attribute*, “dimension,” or “variable.”

  (See §3.3 Organizing Resources (page 98))

- Most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole.

  (See §3.3 Organizing Resources (page 98))

- There are a huge number of ways to organize people that differ in the extent of hierarchical structure, the flow of information up and down the hierarchy, the span of control for managers, and the discretion people have to deviate or innovate with respect to the work they have been assigned to do.

(See the sidebar, Organizing People into Businesses (page 101))

- Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users.

  (See §3.3.1.1 Organizing with Properties of Physical Resources (page 100))

- There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

  (See the sidebar, Gestalt Principles (page 102))

- Built environments can be designed to encourage or discourage interactions between people, to create a sense of freedom or confinement, to reward exploration or enforce efficiency.

  (See §3.3.2.2 Organizing Built Environments (page 104))

- It is straightforward from the perspective of the discipline of organizing to define the activity of information architecture as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use.

  (See §3.3.3.2 “Information Architecture” and Organizing Systems (page 109))

- The level of measurement (nominal, ordinal, interval, or ratio) of data determines how much quantitative organization of your data will be sensible.

  (See §3.3.4 Organizing With Descriptive Statistics (page 113))

- Statistical descriptions summarize a set of resources, and reveal other details that enable comparison of instances with the collection as a whole (such as identifying outliers).

  (See §3.3.4 Organizing With Descriptive Statistics (page 113))

- Multiple properties of the resources, the person organizing or intending to use them, and the social and technological environment in which they are being organized can collectively shape their organization.

  (See §3.3.5 Organizing with Multiple Resource Properties (page 121))

- The tradeoff between the amount of work that goes into organizing a collection of resources and the amount of work required to find and use them is inescapable when the resources are physical objects or information resources are in physical form.

  (See §3.4.1 Affordance and Capability (page 123))

- The concept of *affordance*, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor's capabilities and cognition, what can be done with the resource.

  (See §3.4.1 Affordance and Capability (page 123))

- We should not assume that supporting more types of interactions necessarily makes a system better or more capable; what matters is how much value is created or invoked in each interaction.

  (See §3.4.1 Affordance and Capability (page 123))

- A resource is only accessible when it supports interactions, and it is ineffective design to implement interactions with resources that some people are unable to perform.

  (See §3.4.1 Affordance and Capability (page 123))

- Many of the techniques for making a resource accessible involve transforming the resource or its description into a different form so someone who could not perceive it or interact with it in its original form can now do so.

  (See §3.4.1 Affordance and Capability (page 123))

- With digital resources, the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.

  (See §3.4.2.2 Value Creation with Digital Resources (page 126))

- The variety and functions of interactions with digital resources are determined by the amount of structure and semantics represented in their digital encoding, in the descriptions associated with the resources, or by the intelligence of the computational processes applied to them.

  (See §3.4.2.2 Value Creation with Digital Resources (page 126))

- Preservation of resources means maintaining them in conditions that protect them from physical damage or deterioration.

  (See §3.5.2 Preservation (page 134))

- Preservation is often a key motive for digitization, but digitization alone is not preservation.

  (See §3.5.2.1 Digitization and Preserving Resources (page 135))

- The essence of curation and governance is having clear policies for collecting resources and maintaining them over time that enable people and automated processes to ensure that resource descriptions or data are authoritative, accurate, complete, consistent, and non-redundant.

  (See §3.5.3 Curation (page 139) and §3.5.4 Governance (page 144))

- Data cleaning algorithms can eliminate duplicate data, search engines can improve the relevance of results using selection and navigation behavior, and sensor data can predict when machines need servicing.

  (See §3.5.3.4 Computational Curation (page 141))
- An essential part of maintenance is the phasing out of resources that are damaged or unusable, expired or past their effectivity dates, or no longer relevant to any interaction.

  (See §3.5.3.5 Discarding, Removing, and Not Keeping (page 142))
- Governance is essential to deal with frequent changes in business organizing systems, data quality management, access control to ensure security and privacy, compliance, deletion, and archiving.

  (See §3.5.4.1 Governance in Business Organizing Systems (page 144))
- Scientific data poses special governance problems because of its scale.

  (See §3.5.4.2 Governance in Scientific Organizing Systems (page 145))

Endnotes for Chapter 3

[44][Law] Some governments attempt to preserve and prevent misappropriation of cultural property by enforcing import or export controls on antiquities that might be stolen from archaeological sites (Merryman 2006). For digital resources, privacy laws prohibit the collection or misuse of personally identifiable information about healthcare, education, telecommunications, video rental, and might soon restrict the information collected during web browsing.

[45][DS] The popular LinkedIn site, which has hundreds of millions of resumes that it data mines to find statistically superior job candidates, is literally a gold mine for the company because it makes money by referring those candidates to potential employers. Data-intensive hiring practices in baseball are entertainingly presented in the book entitled *Moneyball* book (Lewis 2003) or the 2011 movie starring Brad Pitt. Pro football teams have begun to assess college football players by comparing them statistically with the best pro players (Robbins, 2016).

Many examples of business strategies that required significant investment to acquire data assets with no current value are reported in (Provost and Fawcett 2013).

[46][Com] See (Cherbakov et al. 2005), (Erl 2005a). The essence of SOA is to treat business services or functions as components that can be combined as needed. An SOA enables a business to quickly and cost-effectively change how it does business and whom it does business with (suppliers, business partners, or customers). SOA is generally implemented using web services that exchange *Exten-*

sible Markup Language (XML) documents in real-time information flows to interconnect the business service components. If the business service components are described abstractly it can be possible for one service provider to be transparently substituted for another—a kind of real-time resource selection—to maintain the desired quality of service. For example, a web retailer might send a Shipping Request to many delivery services, one of which is selected to provide the service. It probably does not matter to the customer which delivery service handles his package, and it might not even matter to the retailer.

[47][Bus] The idea that a firm's long term success can depend on just a handful of critical capabilities that cut across current technologies and organizational boundaries makes a firm's core competency a very abstract conceptual model of how it is organized. This concept was first proposed by (Pralahad and Hamel 1990), and since then there have been literally hundreds of business books that all say essentially the same thing: you cannot be good at everything; choose what you need to be good at and focus on getting better at them; let someone else do things that you do not need to be good at doing.

[49][Web] (Arasu et al. 2001), (Manning et al. 2008). The web is a graph, so all web crawlers use graph traversal algorithms to find URIs of web resources and then add any hyperlink they find to the list of URIs they visit. The sheer size of the web makes crawling its pages a bandwidth- and computation intensive process, and since some pages change frequently and others not at all, an effective crawler must be smart at how it prioritizes the pages it collects and how it recrawls pages. A web crawler for a search engine can determine the most relevant, popular, and credible pages from query logs and visit them more often. For other sites, a crawler adjusts its "revisit frequency" based on the "change frequency" (Cho and Garcia-Molina 2000).

[50][Web] Web resources are typically discovered by computerized "web crawlers" that find them by following links in a methodical automated manner. Web crawlers can be used to create topic-based or domain-specific collections of web resources by changing the "breadth-first" policy of generic crawlers to a "best-first" approach. Such "focused crawlers" only visit pages that have a high probability of being relevant to the topic or domain, which can be estimated by analyzing the similarity of the text of the linking and linked pages, terms in the linked page's URI, or locating explicit semantic annotation that describes their content or their interfaces if they are invokable services (Bergmark et al. 2002), (Ding et al. 2004).

[51][DS] FTC Fair Information Practice Principles say that consumer data collected for one purpose cannot be used for other purposes without the consumer's consent. Sometimes called the consumer privacy bill of rights.

See also (Zhu et al., 2014) and (Marchioni et al., 2012)

[53][DS] See (Tauberer 2014) for a history of the "civic hacking" and the open data movement.

The Sunlight Foundation (`http://sunlightfoundation.com/`) and Code For America (`https://www.codeforamerica.org/`) are good sources for keeping up with open government issues and initiatives.

[54][Com] On data modeling: see (Kent 2012), (Silverston 2000), (Glushko and McGrath 2005). For data warehouses see (Turban et al. 2010).

For a classification and review of data cleaning problems and methods, see (Rahm and Do, 2000). A recent and popular analysis that describes data cleaning as "data wrangling, data munging, and data janitor work" is (Lohr 2014). For a survey of anomaly detection see (Chandola 2009).

[55][Com] (Kim et al, 2003).

[56][CogSci] See (Barsalou and Hale 1983) for a rigorous contrast between feature lists and other representational formalisms in models of human categories.

[57][DS] For example, a personal or small organizing system would typically use properties that are easy to identify and understand. In contrast, an organizing system for very large collections of resources, or data about them, would choose properties that are statistically optimal, even if they are not interpretable by people, because of the greater need for operational efficiency and predictive accuracy.

[59][DS] (Freitas 2014) and (Burrell 2015).

[61][Bus] (Robertson 2015) and (Coase 1937).

[62][CogSci] The Gestalt principles are a staple in every introductory psychology textbook, but the classic text (Koffka 1935) has recently been reprinted. A group of distinguished contemporary researchers in visual perception (Wagemans et al, 2012) recently reviewed the history and impact of Gestalt psychology on their hundredth birthday.

[63][IA] Texts that ground graphic design and information visualization in Gestalt principles include (Cairo 2012) and (Few 2004). (Johnson 2013) explains them within the broader scope of user interface design.

[64][IA] Salt Lake City takes the use of a grid to an extreme because the central area is extremely flat. Streets are named by numbers and letters, so you might find yourself at the intersection of "North A Street" and "3rd Avenue N," or at the intersection of "W 100 S" and "S 200 W." It is a little creepy to think that your street address is a pinpoint location in the big grid.

In contrast, Seattle imposes the grid in an abstract way, ignoring the fact that there are many lakes, rivers, and hills that break up the grid. Streets keep the

same names even though they are not connected, and the grid stretches for many miles out from its origin in Seattle. You can be up in the mountains at the corner of “294th Avenue SE” and “472nd Street SE,” giving you precise information about your location and nearly 50 mile distance from downtown Seattle.

(See also Pierre Charles L'Enfant's plan for DC at http://en.wikipedia.org/wiki/Pierre_Charles_L%27Enfant)

This is not to say that imposing arbitrary grids on top of a physical environment to create a simple and easily understood organization is always desirable. It is essential that any organization imposed on a region be sensitive to any social, cultural, linguistic, ethnic, or religious organizing systems already in place. Much of the recent conflict and instability in the Middle East can be attributed to the implausibly straight line borders drawn by the French and British to carve up the defeated Ottoman Empire a century ago. Because the newly-created countries of Syria and Iraq lacked ethnic and religious cohesion, they could only be held together by dictatorships. (Trofimov 2015)

[67][Phil] The designer of the road system, Robert Moses, heralded as the master builder of mid-20th century New York City, built roads to enforce his idea of who should frequent Long Island (affluent whites). The overpasses were intentionally designed with clearances (often around nine feet) that were too low for public buses. Consequently, low-income bus riders (largely people of color) had no way to get to beaches. See (Winner 1980).

[68][IA] (Arthur and Passini 1992) (McCartney 2015)McCartney, Scott. Technology will speed you through the airport of the future. Wall Street Journal, July 15 2015.

[69][Law] In principle, it is easy to make perfect copies of digital resources. In practice, however, many industries employ a wide range of technologies including digital rights management, watermarking, and license servers to prevent copying of documents, music or video files, and other digital resources. The degree of copying allowed in digital organizing systems is a design choice that is shaped by law.

[70][Web] Web-based or “cloud” services are invoked through URIs, and good design practice makes them permanent even if the implementation or location of the resource they identify changes (Berners-Lee 1998). Digital resources are often replicated in content delivery networks to improve performance, reliability, scalability, and security (Pathan et al. 2008); the web pages served by a busy site might actually be delivered from different parts of the world, depending on where the accessing user is located.

[71][Com] Whether a digital resource seems intangible or tangible depends on the scale of the digital collection and whether we focus on individual resources or the entire collection. An email message is an identified digital resource in a

standard format, RFC 2822 (Resnick 2001). We can compare different email systems according to the kinds of interactions they support and how easy it is to carry them out, but how email resources are represented does not matter to us and they surely seem intangible. Similarly, the organizing system we use to manage email might employ a complex hierarchy of folders or just a single searchable in-box, but whether that organization is implemented in the computer or smart phone we use for email or exists somewhere "in the cloud" for web-based email does not much matter to us either. An email message is tangible when we print it on paper, but all that matters then is that there is well-defined mapping between the different representations of the abstract email resource.

On the other hand, at the scale at which Google and Microsoft handle billions of email messages in their Gmail and Hotmail services the implementation of the email organizing system is extremely relevant and involves many tangible considerations. The location and design of data centers, the configuration of processors and storage devices, the network capacity for delivering messages, whether messages and folder structures are server or client based, and numerous other considerations contribute to the quality of service that we experience when we interact with the email organizing system.

[73][Web] For example, a car dealer might be able to keep track of a few dozen new and used cars on his lot even without a computerized inventory system, but web-based AutoTrader.com offered more than 2,000,000 cars in 2012. The cars are physical resources where they are located in the world, but they are represented in the AutoTrader.com organizing system as digital resources, and cars can be searched for using any combination of the many resource properties in the car listings: price, body style, make, model, year, mileage, color, location, and even specific car features like sunroofs or heated seats.

[74][Com] Even when organizing principles such as *alphabetical*, *chronological*, or numerical ordering do not explicitly consider physical properties, how the resources are arranged in the "storage tier" of the organizing system can still be constrained by their physical properties and by the physical characteristics of the environments in which they are arranged. Books can only be stacked so high whether they are arranged alphabetically or by frequency of use, and large picture books often end up on the taller bottom shelf of bookcases because that is the only shelf they fit. Nevertheless, it is important to treat these idiosyncratic outcomes in physical storage as exceptions and not let them distort the choice of the organizing principles in the "logic tier."

[75][CogSci] (Spence 1985) This memory technique has continued to be used since, and in addition to being found in tips for studying and public speaking, is applied in memorization competitions. For example, journalist and author Joshua Foer, in his book on memory and his journey from beginner to winning the 2006 U.S. Memory Championship (Foer 2011), wrote that Scott Hagwood, a four-time

winner of the same competition, used locations in *Architectural Digest* to place his memories.

[76][Web] The Domain Name System (DNS) (Mockapetris 1987) is the hierarchical naming system that enables the assignment of meaningful domain names to groups of Internet resources. The responsibility for assigning names is delegated in a distributed way by the *Internet Corporation for Assigned Names and Numbers (ICANN)* (http://www.icann.org). DNS is an essential part of the Web's organizing system but predates it by almost twenty years.

[77][Web] HTML5 defines a "manifest" mechanism for making the boundary around a collection of web resources explicit even if somewhat arbitrary to support an "offline" mode of interaction in which all needed resources are continually downloaded (http://www.w3.org/TR/html5/browsers.html#offline), but many people consider it unreliable and subject to strange side effects.

[78][Web] (Aalbersberg and Kahler 2011).

[79][Web] (Munk 2004).

[80][IA] This definition of *information architecture* combines those in a Wikipedia article (http://en.wikipedia.org/wiki/Information_architecture) and in a popular book with the words in its title (Morville and Rosenfield 2006). Given the abstract elegance of "information" and "architecture" any definition of "information architecture" can seem a little feeble.

See (Resmini and Rosati 2011) for a history of information architecture.

[81][IA] See (Halvorson and Rach 2012), (Tidwell 2008), (Morville and Rosenfield 2006), (Kalbach 2007), (Resmini and Rosati 2011), (Marcotte 2011), (Brown 2010), (Abel and Baillie 2014)

[82][IA] Some popular collections of design patterns are (Van Duyne et. al, 2006), (Tidwell 2010), and http://ui-patterns.com/

[83][Law] The Directives can be found at http://ec.europa.eu/consumers/consumer_rights/rights-contracts/directive/index_en.htm

[84][CogSci] The classic text about information visualization is *The Visual Display of Quantitative Information* (Tufte 1983). More recent texts include (Few 2012) and (Yau 2011).

[85][DS] (Rose 2016)

[86][DS] See https://chapters.theiia.org/ottawa/Documents/Digital_Analysis.pdf for a short introduction to data analysis for fraud detection. See (Durtschi et al 2004) for the use of Benford's Law in forensic accounting.

[88][CogSci] (Gibson 1977), (Norman 1988). See also (Norman 1999) for a short and simple explanation of Norman's (re-)interpretation of Gibson.

[89][CogSci] (Norman 1999, p. 39).

[90][Com] The ".xpi" file type is used for Mozilla/Firefox browser extensions, small computer programs that can be installed in the browser to provide some additional user interface functionality or interaction. The ".mobi" file type was originally developed to enable better document display and interactions on devices with small screens. Today its primary use is as the base ebook format for the Amazon Kindle, except that the Kindle version is more highly compressed and locked down with digital rights management.

[91][Com] See (Hearst 2009), (Buettcher et al. 2010).

[92][DS] (Glushko and Nomorosa 2013).

[93][Bus] (Apte and Mason 1995) introduced this framework to analyze services rather than interactions *per se*.

[98][Web] It also erodes the authority and privilege that apply to resources because they are inside the library when a web search engine can search the "holdings" of the web faster and more comprehensively than you can search a library's collection through its online catalog.

[99][Web] (Pirolli 2007).

[100][Web] (Byrne and Goddard 2010).

[102][Com] To augment digital resources with text structures, multimedia, animation, interactive 3-D graphics, mathematical functions, and other richer content types requires much more sophisticated representation formats that tend to require a great deal of "hand-crafting."

An alternative to hand-crafted resource description is sophisticated computer processing guided by human inputs. For example, Facebook and many web-based photo organizing systems implement face recognition analysis that detects faces in photos, compares features of detected faces to features of previously identified faces, and encourages people to tag photos to make the recognition more accurate. Some online services use similar image classification techniques to bring together shoes, jewelry, or other items that look alike.

[103][Com] Even sophisticated text representation formats such as XML have inherent limitations: one important problem that arises in complex management scenarios, humanities scholarship, and bioinformatics is that XML markup cannot easily represent overlapping substructures in the same resource (Schmidt 2009).

[104][Law] Digital books change the economics and *first sale* is not as well-established for digital works, which are licensed rather than sold (Aufderheide and Jaszi 2011). To protect their business models, many publishers are limiting the number of times ebooks can be lent before they "self-destruct." Some librar-

ians have called for boycotts of publishers in response (http://boycottharpercollins.com).

In contrast to these new access restrictions imposed by publishers on digital works, many governments as well as some progressive information providers and scientific researchers have begun to encourage the reuse and reorganization of their content by making geospatial, demographic, environmental, economic, and other datasets available in open formats, as web services, or as data feeds rather than as "fixed" publications (Bizer 2009a), (Robinson et al. 2008). And we have made this book available as an open content repository so that it can be collaboratively maintained and customized.

[105][Phil] We cannot explain this any better than the UN does: "The Convention follows decades of work by the United Nations to change attitudes and approaches to persons with disabilities. It takes to a new height the movement from viewing persons with disabilities as 'objects' of charity, medical treatment and social protection towards viewing persons with disabilities as 'subjects' with rights, who are capable of claiming those rights and making decisions for their lives based on their free and informed consent as well as being active members of society." See https://www.un.org/disabilities/default.asp?navid=12&pid=150

[106][Com] See Microsoft Windows Accessibility, Apple Accessibility, and Android Accessibility Features.

[107][Web] The Web Accessibility Initiative works to make the Web accessible to people with visual, auditory, speech, cognitive, neurological, and physical disabilities.

[108][Web] Accessible Rich Internet Applications (ARIA)

[110][Com] For a recent historical and highly technical review of speech recognition written by some of the most prominent researchers in the field, see (Huang, Baker, and Reddy 2014) An easier to read story about Apple's Siri voice recognition program is (Geller 2012). Popup archive is https://www.popuparchive.org/ and its audio search service is https://www.audiosear.ch/

[111][Bus] These access controls to the organizing system or its host computer are enforced using passwords and more sophisticated software and hardware techniques. Some access control policies are mandated by regulations to ensure privacy of personal data, and policies differ from industry to industry and from country to country. Access controls can improve the credibility of information by identifying who created or changed it, especially important when traceability is required (e.g., financial accounting).

An important difference between interactions with physical resources and those with digital resources is how they use resource descriptions for access control.

Resources sometimes have associated security classifications like "Top Secret" that restrict who can learn about their existence or obtain them. Nonetheless, if you get your hands on a top secret printed document, nothing can prevent you from reading it. Similarly, printed resources often have "All rights reserved" copyright notices that say that you cannot copy them, but nothing can prevent you from making copies with a copy machine. On the other hand, learning of the existence of a digital resource might be of little value if copyright or licensing restrictions prevent you from obtaining it. Moreover, obtaining a digital resource might be of no value if its content is only available using a password, decryption key, or other resource description that enforces access control directly rather than indirectly like the security classifications.

[112][Law] In response to this trend, however, many libraries are supporting "open access" initiatives that strive to make scholarly publications available without restriction (Bailey 2007). Libraries and ebook vendors are engaged in a tussle about the extent to which the "first sale" rule that allows libraries to lend physical books without restrictions also applies to ebooks (Howard 2011).

[116][Bus] Customer information drives day-to-day operations, but is also used in decision support and strategic planning.

[117][DS] For businesses "in the world," a "customer" is usually an actual person whose identity was learned in a transaction, but for many web-based businesses and search engines a customer is a computational model extracted from browser access and click logs that is a kind of "theoretical customer" whose actual identity is often unknown. These computational customers are the targets of the computational advertising in search engines.

[118][Law] The *Sarbanes-Oxley Act* in the United States and similar legislation in other countries require firms to preserve transactional and accounting records and any document that relates to "internal controls," which arguably includes any information in any format created by any employee (Langevoort 2006). Civil procedure rules that permit discovery of evidence in lawsuits have long required firms to retain documents, and the proliferation of digital document types like email, voice mail, shared calendars and instant messages imposes new storage requirements and challenges (Levy and Casey 2006). However, if a company has a data retention policy that includes the systematic deletion of documents when they are no longer needed, courts have noted that this is not willful destruction of evidence.

[120][CogSci] For example. students writing a term paper usually organize the printed and digital resources they rely on; the former are probably kept in folders or in piles on the desk, and the latter in a computer file system. This organizing system is not likely to be preserved after the term paper is finished. An exception that proves the rule is the task of paying income taxes for which (in the USA) taxpayers are legally required to keep evidence for up to seven years after

filing a tax return (http://www.irs.gov/Businesses/Small-Businesses-&-Self-Employed/How-long-should-I-keep-records%3F).

[121][Com] (Rothenberg 1999).

[122][Com] (Pogue 2009).

[124][Web] This is tautologically true for sites that publish news, weather, product catalogs with inventory information, stock prices, and similar continually updated content because many of their pages are automatically revised when events happen or as information arrives from other sources. It is also true for blogs, wikis, Facebook, Flickr, YouTube, Yelp and the great many other "Web 2.0" sites whose content changes as they incorporate a steady stream of user-generated content.

In some cases changes to web pages are attempts to rewrite history and prevent preservation by removing all traces of information that later turned out to be embarrassing, contradictory, or politically incorrect. When pages cannot be changed, like the archives of newspapers published on the web, only the search engine can remove them from search results, and in 2014 the European Court ruled that people could ask Google to do that.

[125][Web] (Fetterly et al. 2003).

Most people understand that web pages can change, but most changed web pages do not highlight the changes. A "diff" tool from Microsoft reveals them. http://research.microsoft.com/en-us/projects/DiffIE/default.aspx

[126][Web] However, when a website disappears its first page can often be found in the search engine's index "cache" rather than by following what would be a broken link.

[131][DS] (Talukder 2016)

[132][Com] The set of content modules and their assembly structure for each kind of generated document conforms to a template or pattern that is called the document type model when it is expressed in XML.

[133][Bus] Company intranets, wikis, and blogs are often used as knowledge management technologies; Lotus Notes and Microsoft SharePoint are popular commercial systems. (See the case study in §12.2 Knowledge Management for a Small Consulting Firm (page 559).)

[134][Bus] In addition, the line between "preserving species" and "preserving marketing brands" is a fine one for zoos with celebrity animals, and in animal theme parks like Sea World, it seems to have been crossed. "Shamu" was the first killer whale (orca) to survive long in captivity and performed for several years at SeaWorld San Diego. Shamu died in 1971 but over forty years later all three US-based SeaWorld parks have Shamu shows and Shamu webcams.

[135][DS] (Manyika et al. 2011).

[140][CogSci] Because personal collections are strongly biased by the experiences and goals of the organizer, they are highly idiosyncratic, but still often embody well-thought-out and carefully executed curation activities (Kirsh 2000), (Jones 2007), (Marshall 2007),(Marshall 2008).

[141][Com] (Bush 1945). Bush imagined that Memex users could share these packages of trails and that a profession of trailbuilders would emerge. However, he did not envision that the Memexes themselves could be interconnected, nor did he imagine that their contents could be searched computationally.

[142][Web] (Howe 2008).

[143][Web] The most salient example of this so called "community curation" activity is the work to maintain the Wikipedia open-source encyclopedia according to a curation system of roles and functions that governs how and under what conditions contributors can add, revise, or delete articles; receive notifications of changes to articles; and resolve editing disputes (Lovink and Tkacz 2011). Some museums and scientific data repositories also encourage voluntary curation to analyze and classify specimens or photographs (Wright 2010).

[144][Web] (Trant 2009).

[145][Bus] Some popular "community content" sites like Yelp where people rate local businesses have been criticized for allowing positive rating manipulation. Yelp has also been criticized for allowing negative manipulation of ratings when competitors slam their rivals.

[146][Web] The resource might have been put someplace else when the site was reorganized or a new web server was installed. It is no longer the same resource because it will have another URI, even if its content did not change.

[147][Com] All of these terms refer to types of web resources or techniques whose purpose is to mislead people into doing things or letting things be done to their computers that will cost them their money, time, privacy, reputation, or worse. We know too well what spam is. Phishing is a type of spam that directs recipients to a fake website designed to look like a legitimate one to trick them into entering account numbers, passwords, or other sensitive personal information. Malware, fakeware, or spyware sites offer tempting downloadable content that installs software designed to steal information from or take control of the visiting computer. Keyword stuffing, spamdexing, and META tag abuse are techniques that try to mislead search engines about the content of a resource by annotating it with false descriptions. Link farms or scraper sites contain little useful or original content and exist solely for the purpose of manipulating search engine rankings to increase advertising revenue. Similarly, cybersquatters register

domain names with the hope of profiting from the goodwill of a trademark they do not own.

[148][Com] (Brown 2009).

[149][DS] (Diaz 2005), (Grimmelmann 2009).

[152][Law] (Blanchette and Johnson 2002)

[155][Law] Data governance decisions are also often shaped by the need to conform to information or process model standards, or to standards for IT service management like the *Information Technology Infrastructure Library (ITIL)*. See http://www.itil-officialsite.com/.

[156][Bus] In this context, these management and maintenance activities are often described as "IT governance" (Weill and Ross 2004). Data classification is an essential IT governance activity because the confidentiality, competitive value, or currency of information are factors that determine who has access to it, how long it should be preserved, and where it should be stored at different points in its lifecycle.

[157][Bus] (Turban et al. 2010).

[158][DS] This master data must be continually "cleansed" to remove errors or inconsistencies, and "de-duplication" techniques are applied to ensure an authoritative source of data and to prevent the redundant storage of many copies of the same resource. Redundant storage can result in wasted time searching for the most recent or authoritative version, cause problems if an outdated version is used, and increase the risk of important data being lost or stolen. (Loshin 2008).

[159][Com] (Cox et al. 2007).

[160][Law] Recently imposed requirements by the National Science Foundation (NSF), National Institute of Health (NIH) and other research granting agencies for researchers to submit "data management plans" as part of their proposals should make digital data curation a much more important concern (Borgman 2011). (NSF Data Management Plan Requirements: http://www.nsf.gov/eng/general/dmp.jsp).

Chapter 4

Resources in Organizing Systems

Robert J. Glushko
Daniel D. Turner
Kimra McPherson
Jess Hemerly

4.1 Introduction

This chapter builds upon the foundational concepts introduced in Chapter 1 to explain more carefully what we mean by *resource*. In particular, we focus on the issue of identity—what will be treated as a separate resource—and discuss the issues and principles we need to consider when we give each resource a name or identifier.

Navigating This Chapter

In §4.2 Four Distinctions about Resources (page 166) we introduce four distinctions we can make when we discuss resources: *domain*, *format*, *agency*, and *focus*. In §4.3 Resource Identity (page 182) we apply these distinctions as we discuss how resource identity is determined for physical resources, bibliographic resources, resources in information systems, as well as for *active resources* and *smart things*. §4.4 Naming Resources (page 188) then tackles the problems and principles for naming: once we have identified resources, how do we name and distinguish them? Finally, §4.5 Resources over Time (page 198) considers issues that emerge with respect to resources over time.

4.1.1 What Is a Resource?

Resources are what we *organize*.

We introduced the concept of *resource* in §1.3 The Concept of "Resource" (page 35) with its ordinary sense of "anything of value that can support goal-oriented activity" and emphasized that a group of resources can be treated as a *collection* in an organizing system. And what do we mean by "anything of value," exactly? It might seem that the question of *identity*, of what a single resource is, should not be hard to answer. After all, we live in a world of resources, and finding, selecting, describing, arranging, and referring to them are everyday activities. And while human resources are not a primary focus of this book, it would be remiss not to explain why it makes sense to think of people that way. See the sidebar, People as Resources (page 172).

Nevertheless, even when the resources we are dealing with are tangible things, how we go about organizing them is not always obvious, or at least not obvious to each of us in the same way at all times. Not everyone thinks of them in the same way. Recognizing something in the sense of perceiving it as a tangible thing is only the first step toward being able to organize it and other resources like it. Which properties garner our attention, and which we use in organizing depends on our experiences, purposes, and context.

We add information to a resource when we name it or describe it; it then becomes more than "it." We can describe the same resource in many different ways. At various times we can consider any given resource to be one of many members of a broad category, as one of the few members of a narrow category, or as a unique instance of a category with only one member. For example, we might recognize something as a piece of clothing, as a sock, or as the specific dirty sock with the hole worn in the heel from yesterday's long hike. However,

even after we categorize something, we might not be careful how we talk about it; we often refer to two objects as “the same thing” when what we mean is that they are “the same type of thing.” Indeed, we could debate whether a category with only one possible member is really a category, because it blurs an important distinction between particular items or instances and the class or type to which they belong.

The issues that matter and the decisions we need to make about resource instances and resource classes and types are not completely separable. Nevertheless, we will strive to focus on the former ones in this chapter and the latter ones in Chapter 7, *Categorization: Describing Resource Classes and Types*.

4.1.1.1 Resources with Parts

As tricky as it can be to decide what a resource is when you are dealing with single objects, it is even more challenging when the resources are objects or systems composed of other parts. In these cases, we must focus on the entirety of the object or system and treat it as a resource, treat its constituent parts as resources, and deal with the relationships between the parts and the whole, as we do with engineering drawings and assembly procedures.

How many things is a car? If you are imagining the car being assembled you might think of several dozen large parts like the frame, suspension, drive train, gas tank, brakes, engine, exhaust system, passenger compartment, doors, and other pre-assembled components. Of course, each of those components is itself made up of many parts—think of the engine, or even just the radio. Some sources have counted ten or fifteen thousand parts in the average car, but even at that precise granularity a lot of parts are still complex things. There are screws and wires and fasteners and on and on; really too many to count.

Ambiguity about the number of parts in the whole holds for information resources too; a newspaper can be considered a single resource but it might also consist of multiple sections, each of which contains separate stories, each of which has many paragraphs, and so on. Similarly, while a web page can be treated as a single resource, it can also be considered as a collection of more granular parts, each of which can be separately identified as the source or anchor of a link. Likewise, a bank's credit card application might ask about outstanding loans, payment history, current income, and other information, or the bank might just look up your credit score, which is a statistical index that combines this financial information into a single number.

How Many Things is a Chess Set?

A chess set exemplifies the many different ways to decide what to count as a separate resource. Is this a chess set, two sets of chess pieces, six types of chess pieces (1 king, 1 queen, 2 rooks, 2 bishops, 2 knights, 8 pawns for each side), or 33 separate things (the 32 pieces and a board on which to play the game)?

(Photo by Emma Jane Hogbin Westby. Creative Commons CC-BY-2.0 license.)

4.1.1.2 Bibliographic Resources, Information Components, and "Smart Things" as Resources

Information resources generally pose additional challenges in their identification and description because their most important property is usually their content, which is not easily and consistently recognizable. Organizing systems for information resources in physical form, like those for libraries, have to juggle the duality of their tangible embodiment with what is inherently an abstract information resource; that is, the printed book versus the knowledge the book contains. Here, the organizing system emphasizes description resources or surrogates, like bibliographic records that describe the information content, rather than their physical properties.

Another question about resource that is especially critical in libraries is: What set of resources should be treated as the same *work* because they contain essentially similar intellectual or artistic content? We may talk about Shakespeare's play *Macbeth,* but what is this thing we call "Macbeth"? Is it a particular string of words, saved in a computer file or handwritten upon a folio? Is it the collection of words printed with some predetermined font and pagination? Are all the editions and printings of these words the same *Macbeth*? How should we organize the numerous live and recorded performances of plays and movies that share the *Macbeth* name? What about creations based on or inspired by *Macbeth* that do not share the title "Macbeth," like the Kurosawa film "*Kumonosu-jo*" (*Throne of Blood*) that transposes the plot to feudal Japan? Patrick Wilson proposed a genealogical analogy, characterizing a *work* as "a group or family of texts," with the idea that a creation like Shakespeare's *Macbeth* is the "ancestor of later members of the family."

Information system designers and architects face analogous design challenges when they describe the *information components* in business or scientific organizing systems. Information content is intrinsically merged or confounded with structure and presentation whenever it is used in a specific instance and

context. From a logical perspective, an order form contains *information components* for ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION, but the arrangement of these components, their type font and size, and other non-semantic properties can vary a great deal in different order forms and even across a single information system that re-purposes these components for letters, delivery notices, mailing labels, and database entries.[162][Bus]

Similar questions about resource identity are posed by the emergence of ubiquitous or pervasive computing, in which information processing capability and connectivity are embedded into physical objects, in devices like smart phones, and in the surrounding environment. Equipped with sensors, radio-frequency identification (RFID) tags, GPS data, and user-contributed metadata, these *smart things* create a jumbled torrent of information about location and other properties that must be sorted into identified streams and then matched or associated with the original resource.

§4.3 Resource Identity (page 182) discusses the issues and methods for determining "what is a *resource*?" for physical resources, as well as for the *bibliographic resources*, *information components* and *smart things* discussed here, in §4.1.1.1 Resources with Parts (page 163).

4.1.2 Identity, Identifiers, and Names

The answer to the question posed in §4.1.1 What Is a Resource? (page 162) has two parts.

- The first part is *identity*: what thing are we treating as the resource?
- The second part is *identification*: differentiating between this single resource and other resources like it.

These problems are closely related. Once you have decided what to treat as a resource, you create a name or an identifier so that you can refer to it reliably. A *name* is a label for a resource that is used to distinguish one from another. An *identifier* is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. For a digital resource, its identifier serves as the input to the system or function that determines its location so it can be retrieved, a process called *resolving* the identifier or *resolution*.

Choosing names and identifiers—be it for a person, a service, a place, a trend, a work, a document, a concept, etc.—is often challenging and highly contentious. Naming is made difficult by countless factors, including the audience that will need to access, share, and use the names, the limitations of language, institutional politics, and personal and cultural biases.

A common complication arises when a resource has more than one name or identifier. When something has more than one name each of the multiple names is a *synonym* or *alias*. A particular physical instance of a book might be called a hardcover or paperback or simply a text. George Furnas and his research collaborators called this issue of multiple names for the same resource or concept the *vocabulary problem*.[163][CogSci]

Whether we call it a book or a text, the resource will usually have a Library of Congress catalog number as well as an ISBN as an *identifier*. When the book is in a carton of books being shipped from the publisher to a bookstore or library, that carton will have a bar-coded tracking number assigned by the delivery service, and a manifest or receipt document created by the publisher whose identifier associates the shipment with the customer. Each of these identifiers is unique with respect to some established scope or context.

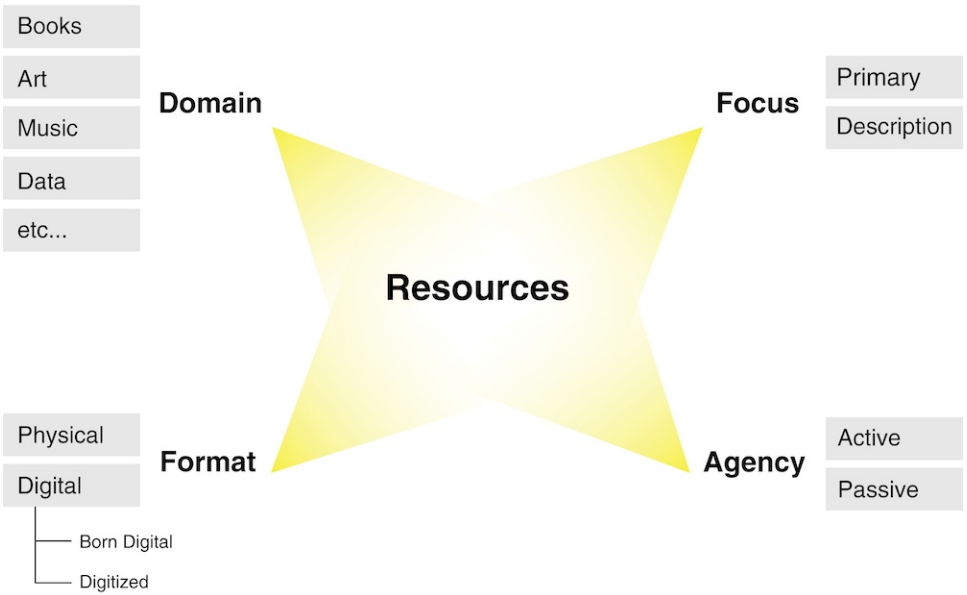
A partial solution to the *vocabulary problem* is to use a *controlled vocabulary*. We can impose rules that standardize the way in which names and labels for resources are assigned in the first place. Alternatively, we can define mappings from terms used in our natural language to the authoritative or controlled terms. However, vocabulary control cannot remove all ambiguity. Even if a passport or national identity system requires authoritative full names rather than nicknames, there could easily be more than one Robert John Smith in the system.

Controlling the language used for a particular purpose raises other questions: Who writes and enforces these rules? What happens when organizing systems that follow different rules get compared, combined, or otherwise brought together in contexts different from those for which they were originally intended?

4.2 Four Distinctions about Resources

The nature of the resource is critical for the creation and maintenance of quality organizing systems. There are four distinctions we make in discussing resources: *domain*, *format*, *agency*, and *focus*. Figure 4.1, Resource Domain, Format, Focus and Agency. depicts these four distinctions, perspectives or points of view on resources; because they are not independent, we cannot portray these distinctions as categories of resources.

Figure 4.1. Resource Domain, Format, Focus and Agency.



Four distinctions we can make when discussing resources concern their domain (their type of matter or content), format (physical or digital), agency (active or passive), and focus (primary or description).

4.2.1 Resource Domain

Resource domain is an intuitive notion that groups resources according to the set of natural or intuitive characteristics that distinguishes them from other resources. It contrasts with the idea of ad hoc or arbitrary groupings of resources that happen to be in the same place at some time.

For physical resources, domains can be coarsely distinguished according to the type of matter they are made of using easily perceived properties. The top-level classification of all things into the animal, vegetable, and mineral kingdoms by Carl Linnaeus in 1735 is deeply embedded in most languages and cultures to create a hierarchical system of domain categories. [164][Phil] Many aspects of this system of domain categories are determined by natural constraints on category membership that exist as patterns of shared and correlated properties; a resource identified as a member of one category must also be a member of another with which it shares some but not all properties. For example, a marble statue in a museum must also be a kind of material, and a fish in an aquarium must also be a kind of animal.

The Document Type Spectrum

Different domains or types of documents can be distinguished according to the extent to which their content is semantically prescribed, by the amount of internal structure, and by the correlations of their presentation and formatting to their content and structure. These three characteristics of content, structure, and presentation vary systematically from narrative document types like novels to transactional document types like invoices.

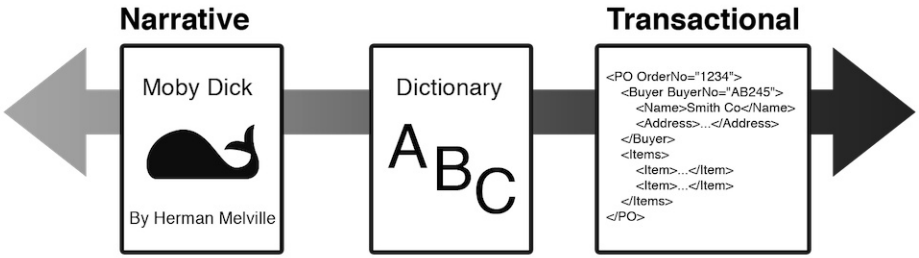
Narrative types are authored by people and are heterogeneous in structure and content, and their content is usually just prose and graphic elements. Their presentational characteristics carefully reinforce their structure and semantics; for example, the text of titles or major headings is large because the content is important, in contrast to the small text of footnotes. Transactional document types are usually created mechanically and, as a result, are homogeneous in structure and content; their content is largely "data" — strongly typed content with precise semantics that can be processed by computers.

In the middle of the spectrum are hybrid document types like textbooks, encyclopedias, and technical manuals that contain a mixture of narrative text and structured content in figures, data tables, code examples, and so on.

For information resources, easily perceived properties like a book's color or size are less reliably correlated with resource domain, so we more often distinguish domains based on semantic properties; the definitions of the "encyclopedia," "novel," and "invoice" resource types distinguish them according to their typical subject matter, or the type of content, rather than according to the great variety of physical forms in which we might encounter them. Arranging books by color or size might be sensible for very small collections, or in a photo studio, but organizing according to physical properties would make it extremely impractical to find books in a large library.

We can arrange types of information resources in a hierarchy. However, because the category boundaries are not sharp it is more useful to view domains of information resources on a continuum from weakly-structured narrative content to highly structured transactional content. This *framework*, called the *Document Type Spectrum* by Glushko and McGrath, captures the idea that the boundaries between resource domains, like those between colors in the rainbow, are easy to see for colors far apart in the spectrum but hard to see for adjacent ones.[165][Com] (See the sidebar, The Document Type Spectrum (page 168), and its corresponding depiction as Figure 4.2, Document Type Spectrum.)

Figure 4.2. Document Type Spectrum.



The Document Type Spectrum is a continuum of document types from narrative ones that are mostly text, like novels, to transactional ones with highly-structured information, like invoices. In between are hybrid types that contain both narrative and transactional content, like dictionaries and encyclopedias.

4.2.2 Resource Format

Information resources can exist in numerous formats with the most basic distinction between physical and digital ones. This distinction is most important in the implementation of a resource storage or preservation system because that is where physical properties are usually considerations, and very possibly constraints. This distinction is less important at the logical level when we design interactions with resources because digital surrogates for the physical resources can overcome the constraints posed by physical properties. When we search for cars or appliances in an online store it does not matter where the actual cars or appliances are located or how they are organized. (See the sidebar, The Three Tiers of Organizing Systems (page 47)).

Many digital representations can be associated with either physical or digital resources, but it is important to know which one is the original or primary resource, especially for unique or valuable ones.

Today many resources in organizing systems are *born digital*, created in word processors, digital cameras, audio and video recorders. Other digital resources are by sensors in “smart things” and by the systems that create digital resources when they interact with barcodes, QR codes, RFID tags, or other mechanisms for tracking identity and location.[166][Com]

Other digital resources are created by *digitization*, the process for transforming an artifact whose original format is physical so it can be stored and manipulated by a computer. We digitize the printed word, photographs, blueprints, and record albums. Printed text, for example, can be digitized by scanning pages and using character recognition software or simply re-typing it.

There are a vast number of digital formats that differ in many ways, but we can coarsely compare them on two dimensions: the degree to which they distinguish information content from presentation or rendering, and the explicitness with which content distinctions are represented. Taken together, these two dimensions allow us to compare formats on their overall "Information IQ" —with the overarching principle being that "smarter" formats contain more computer-processable information, as illustrated in Figure 4.3, Information IQ.

Simple digital formats for "plain text" documents contain only the characters that you see on your computer keyboard. *ASCII* is the most commonly used simple format, but ASCII is inadequate for most languages, which have larger character sets, and it also cannot handle mathematical characters.[168][Com] The Unicode standard was designed to overcome these limitations.[169][Com] (ASCII and Unicode are discussed in great detail in §9.3.1 Notations (page 462).)

Most document formats also explicitly encode a hierarchy of structural components, such as chapters, sections or semantic components like descriptions or procedural steps, and sometimes the appearance of the rendered or printed form.[170][Com] Another important distinction to note is whether the information is encoded as a sequence of text characters so that it is human readable as well as computer readable. Encoding character content with XML, for example, allows for layering of intentional coding or *markup* interwoven with the "plain text" content. Because XML processors are required to support Unicode, any character can appear in an XML document. The most complex digital formats are those for multimedia resources and multidimensional data, where the data format is highly optimized for specialized analysis or applications.[171][Com]

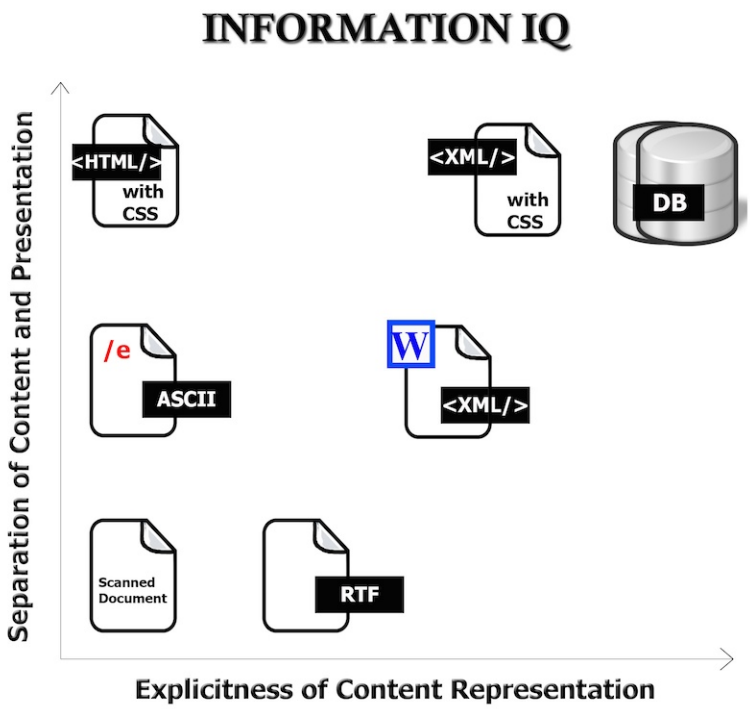
Digitization of non-text resources such as film photography, drawings, and analog audio and visual recordings raises a complicated set of choices about pixel density, color depth, sampling rate, frequency filtering, compression, and numerous other technical issues that determine the digital representation.[172][Com]

There may be multiple intended uses and devices for a digitized resource that could require different digitization approaches and formats. Downstream users of digitized resources need to know the format in which a digital artifact has been created, so they can reuse it as is, or process it in other ways.

Some digital formats support interactions that are qualitatively different and more powerful than those possible with physical resources. Museums are using virtual world technology to create interactive exhibits in which visitors can fly through the solar system, scan their own bodies, and change gravity so they can bounce off walls. Sophisticated digital document formats can enable interactions with annotated digital images or video, 3-D graphics or embedded datasets. The Google Art Project contains extremely high resolution photographs of famous paintings that make it possible to see details that are undetectable under the normal viewing conditions in museums.

Nevertheless, digital representations of physical resources can also lose important information and capabilities. The distinctive sounds of hip hop music produced by “scratching” vinyl records on turntables cannot be produced from digital MP3 music files.[174][Com]

Figure 4.3. Information IQ.



The notion of Information IQ captures the idea that document formats differ on two dimensions: the explicitness of content representation, and the separation of content and presentation. A scanned document is just a picture of a document with neither of these distinctions, so it is low on both dimensions. A database or XML document distinguishes explicitly between types of content and presentation is separately assigned, so they are high on both dimensions and have the highest Information IQ. An HTML document's content distinctions are usually presentational and, thus, it has lower IQ. Formats with high Information IQ facilitate computer processing.

Copyright often presents a barrier to digitization, both as a matter of law and because digitization itself enables copyright enforcement to a degree not possible prior to the advent of digitization, by eliminating common forms of access and interactions that are inherently possible with physical printed books like the ability to give or sell them to someone else.[175][Law]

4.2.3 Resource Agency

People as Resources

Earlier editions of this book sidestepped the question of people as resources to avoid complicating the Discipline of Organizing. People organize themselves in innumerable ways to coexist, share knowledge, and accomplish more than they could as individuals, and behaviors such as trust and reciprocity might be considered "organizing principles" for human society. But these organic relationships and interactions usually lack the intentional arrangement to be considered true Organizing Systems, except when the people are living in "intentional communities."[9 [Phil]]

However, people do qualify as resources in Organizing Systems under our definition: just like digital and physical resources, human resources can be identified, categorized, described in terms of their attributes and relationships, and take part in interactions to create value. In businesses, people are organized to amplify their skills, knowledge, and agency. A company's organizational chart is often a formal hierarchy in which each worker's role is defined through his or her responsibilities and relationships to others in the company. Treating an employee abstractly as a resource with specific and predictable functions, inputs, and outputs enables employees or processes to depend upon each other without being distracted by the details of one another's work. This so-called "black boxing" can encourage specialization and allow an organization to function more efficiently.

Nevertheless, while organizational charts are typically presented as neat hierarchies, human relationships cut across the hierarchy to create a network, and may be complicated by differing values and motivations. Conflicting incentives and lack of communication between people may cause breakdowns in the system. People are more than the sum of properties used to organize them: understanding and defining employees' roles too narrowly could exclude the aspects of the job that they find rewarding or consider part of their professional identity, while black boxing people's labor and treating them as "remote person calls" also risks dehumanizing them and ignoring their working conditions.

Agency is the extent to which a resource can initiate actions on its own. We can define a continuum between completely passive resources that cannot initiate

any actions and active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources. A book being read at the beach will grow warm from absorbing the sun's energy but it has no way of measuring its temperature and is a completely passive resource. An ordinary mercury thermometer senses and displays the temperature but is not capable of communicating its own reading, whereas a digital wireless thermometer or weather station can.

Passive resources serve as nouns or operands that are acted upon, while active resources serve as verbs or operants that cause and carry out actions. We need a concept of agency to bring resources that are active information sources, or computational in character, into the organizing system framework. This concept also lets us include living resources, or more specifically, humans, into discussions about organizing systems in a more general way that emphasizes their agency.[176][Bus]

4.2.3.1 Passive or Operand Resources

Organizing systems that contain passive or operand resources are ubiquitous for the simple reason that we live in a world of physical resources that we identify and name in order to interact with them. *Passive resources* are usually tangible and static and thus they become valuable only as a result of some action or interaction with them.

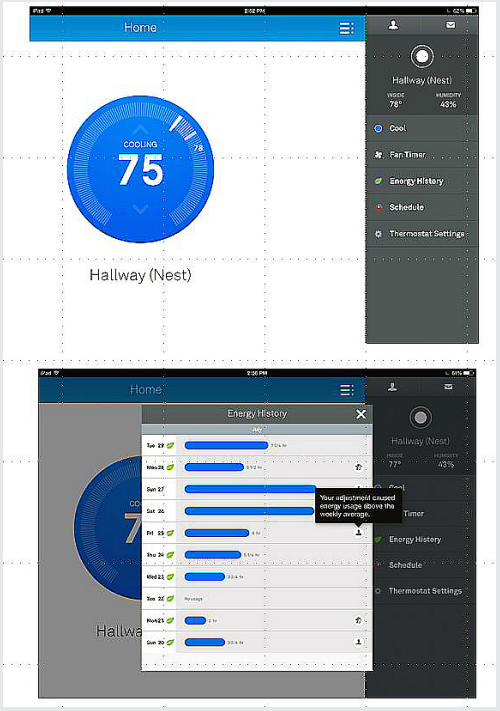
Most organizing systems with physical resources or those that contain resources that are digitized equivalents treat those resources as passive. A printed book on a library shelf, a digital book in an ebook reader, a statue in a museum gallery, or a case of beer in a supermarket refrigerator only create value when they are checked out, viewed, or consumed. None of these resources exhibits any agency and cannot initiate any actions to create value on their own.

4.2.3.2 Active or Operant Resources

Active resources create effects or value on their own, sometimes when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, web-based services, self-driving cars, robots, appliances, machines or otherwise ordinary objects like light bulbs, umbrellas, and shoes that have been made "smarter." We can exploit computing capability, storage capacity, and communication bandwidth to create active resources that can do things and support interactions that are impossible for ordinary physical passive resources.

We can analyze active resources according to five capabilities that progressively increase their agency. These capabilities build on each other to give resources and the organizing systems in which they participate more ways to create value through interactions and information exchanges.

Active Resources: The Nest Thermostat "Ecosystem"



These two screenshots of the Nest iPad app show the thermostat control panel and an energy history report with a pop-up note explaining that resetting the temperature resulted in higher than average energy use on that day. The Nest thermostat serves as a hub device, communicating with lights, appliances, smoke alarms, your car, your wearable fitness sensor, and other active resources (https://nest.com/works-with-nest/).

(Screenshots by Andrea Angquist. Used with permission.)

Sensing or awareness

The minimal capability for a resource to have some agency is for it to be able to sense or be aware of some aspect of its environment or its interactions with other resources. A thermometer measures temperature, a photodetector measures light, a gauge measures the fuel left in a car's gas tank, a GPS device computes its location after detecting and analyzing signals from satellites, a wearable fitness sensor tracks your heartbeat and how far you walk. But sensing something in itself does not create any value in an organizing system. Something needs to be done.

Actuation

A resource has the capability to actuate when it can create effects or value by initiating some action as a result of the information it senses; "actuator" is often used to describe a resource that can move or control a physical mechanism or system, while "effector" is used when the resource is a biological one. Resources can actuate by turning on lights, speakers, cameras, motors, switches, by sending a message about the state or value of a sensor, or by moving themselves around (as with robots).

A potential or latent actuation is created when a resource can display or broadcast some aspect of its state, but value is only created if another resource (possibly human) happens to see the display or hear the broadcast and then acts upon it.

For example, RFID chips, which are essentially bar codes with built-in radio transponders, can be attached to otherwise passive resources to make them active. RFID chips begin transmitting when they detect the presence of a

RFID reading device. This enables automated location tracking and context sensing. RFID receivers are built into assembly lines, loading docks, parking lots, toll booths, or store shelves to detect when some RFID-tagged resource is at some meaningful location. RFID tags can be made more useful by having them record and transmit information from sensors that detect temperature, humidity, acceleration, and even biological contamination.[177][DS]

Connectivity

For an active resource to do useful work it must be connected in some way to the actuation mechanism that manipulates or controls some other resource. This connection might be a direct and permanent one between the resource and the thing it actuates, like that of a thermostat whose temperature sensing capability has a fixed connection to a heating or cooling system that it turns off or on depending on the temperature.

An important innovation in the design of active resources is "wrapping" physical resources with software so they can be given IP addresses and make connections with Internet protocols, which allows them to send information to an application with more capability to act on it. Such resources are said to be part of the "Internet of Things."

Smart phones are *active resources* that can identify and share their own location, orientation, acceleration and a growing number of other *contextual parameters* to enable personalization of information services. Smart phones can also run the applications that receive messages from and send messages to other smart resources to monitor and optimize how they work.

Computation or programmability

Simple active resources operate in a deterministic manner: given this sensor reading, do this. Other active resources have computational capabilities that enable them to analyze the current and historical information from their sensors, identify significant data values or patterns in these interaction resources, and then adapt their behavior accordingly.

Many thermostats are programmable, but most people do not bother to program them so they miss out on potential energy savings. Nest Labs makes a *learning thermostat* that programs itself. The Nest thermostat uses sensors for temperature, humidity, motion, and light to figure out whether people are at home, and a Wi-fi connection to get local weather data.

The Roomba vacuum cleaning robot navigates around furniture, power cords, stairs, and optimizes its cleaning paths to go over particularly dirty places. But vacuuming is all it does. More sophisticated robots are designed to be versatile and adaptable so they can repetitively perform whatever task is needed for some manufacturing process, and their capabilities can be continually upgraded by software updates, just like the apps on your smart

phone. A new generation of robots typified by one called Baxter can be trained by example; a person moves Baxter's arms and hands to show him what to do, and when Baxter has programmed himself to repeat it, he nods.

Composability and cooperating

The "smartest" active resources can do more than analyze the information they collect and adapt what they do. In addition, they expose what they know and can do to other resources using standard or non-proprietary formats and protocols. This means that active resources that were independently designed and implemented can work together to create value.

Many organizing systems on the web consist of collections or configurations of active digital resources. Interactions among these active resources often implement information-intensive business models where value is created by exchanging, manipulating, transforming, or otherwise processing information, rather than by manipulating, transforming, or otherwise processing physical resources.

We are beginning to see the same principles of modularity and composability applied to physical resources, with open source software libraries for using sensors and micro-controllers and easy to use APIs. In essence, we are using software and physical resources in much the same way as functional building blocks, and standards will be critically important.

Service Oriented Architecture (SOA) is an emerging design discipline for organizing active software resources as functional business components that can be combined in different ways. SOA is generally implemented using web services that exchange XML documents in real-time information flows to interconnect the business service components.

A familiar design pattern for an organizing system composed from active digital resources is the "online store." The store can be analyzed as a composition or choreography in which some web pages display catalog items, others serve as "shopping carts" to assemble the order, and then a "checkout" page collects the buyer's payment and delivery information that gets passed on to other service providers who process payments and deliver the goods.

Design patterns for composing organizing systems from "smart" physical resources are emerging in work on the "smart home," "smart office building," or "smart city." Many experiments are underway and new products emerging that are trying out different combinations of hardware and software to understand the design tradeoffs between them to best determine where the "smarts" should go. For example, we can compare a "smart home" built around a super-intelligent hub device that communicates and coordinates with many other "not so smart" devices from the same manufacturer to one in which all of the devices are equally smart and come from different makers.

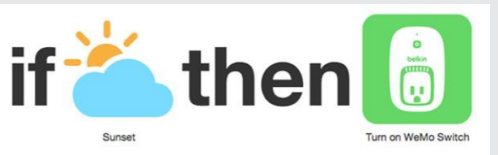
At more complex scales, a truly smart building will not just have programmable thermostats to control heating and cooling systems. It will take in weather forecasts, travel calendars, information about the cost of electricity from different sources, and other relevant information as inputs to a model of how the building heats and cools to optimize energy use and cost while keeping the rooms at appropriate temperatures.

Standard application interfaces enable active resources to interact with people to get information that might otherwise come from sensors or that enhances the value of the sensor information. A programmable thermostat that can record time-based preferences of the people who use the space controlled by the thermostat is more capable than one with just a single temperature threshold. A standard Internet protocol for communicating with the thermostat would enable it to be controlled remotely.

Open and standard data formats and communication protocols enable the aggregation and analysis of information from many instances of the same type of active resource. For example, smart phones running the Google Maps application transmit information about their speed and location. Machine learning and sophisticated optimization techniques of this dataset can yield collective intelligence that can then be given to the resources from which it was derived. In this case, Google can identify traffic jams and generate alternative routes for the drivers stuck in traffic.

If This, Then That

IFTTT is a visual programming system that lets non-programmers connect and control active resources in the physical and digital worlds. IFTTT programs, called recipes, can take information from a growing library of Internet services (date/time, calendar, weather, news, email, social media, and many others) and use this information with simple control logic to trigger actions in other services or resources. Example recipes can copy an Instagram photo to Google Drive, add daily Fitbit data to a spreadsheet, or control lights based on time, weather, or sunset.



"If sunset then electrical outlet on." The icon on the left is the trigger; the icon on the right is the action.

(Photo by R. Glushko)

But not everything can be done best by computers. The web has enabled the use of people as *active resources* to carry out tasks of short duration that can be precisely described but which cannot be done reliably by computers. These tasks often require aesthetic or subjective judgment. The people doing these web-based tasks are often called "Mechanical Turks" by analogy to a fake chess playing machine from the 18th century that had a human hidden inside who secretly moved the pieces.[178][Web]

Stop and Think: The Internet of Things

There is a great deal of hype about the Internet of Things, but there is also a great deal of innovation. If you search for the phrase "Internet of Things" along with almost any physical resource, chances are you will find something, Try "baby," "dog," "fork," "lettuce," "pajamas," "streetlamp," and then a few of your own.

4.2.4 Resource Focus

A fourth contrast between types of resources distinguishes original or *primary resources* from resources that describe them. Any primary resource can have one or more *description resources* associated with it to facilitate finding, interacting with, or interpreting the primary one. *Description resources* are essential in organizing systems where the primary resources are not under its control and can only be accessed or interacted with through the description. *Description resources* are often called *metadata*.

The distinction between *primary resources* and *description resources*, or *metadata*, is deeply embedded in library science and traditional organizing systems whose collections are predominantly text resources like books, articles, or other documents. In these contexts description resources are commonly called *bibliographic resources* or catalogs, and each primary resource is typically associated with one or more description resources.

In business enterprises, the organizing systems for digital information resources, such as business documents, or data records created by transactions or automated processes, almost always employ resources that describe, or are associated with, large sets or classes of primary resources.[179][Com]

The contrast between *primary resources* and *description resources* is very useful in many contexts, but when we look more broadly at organizing systems, it is often difficult to distinguish them, and determining which resources are primary and which are *metadata* is often just a decision about which resource is currently the *focus* of our attention.[180][Law]

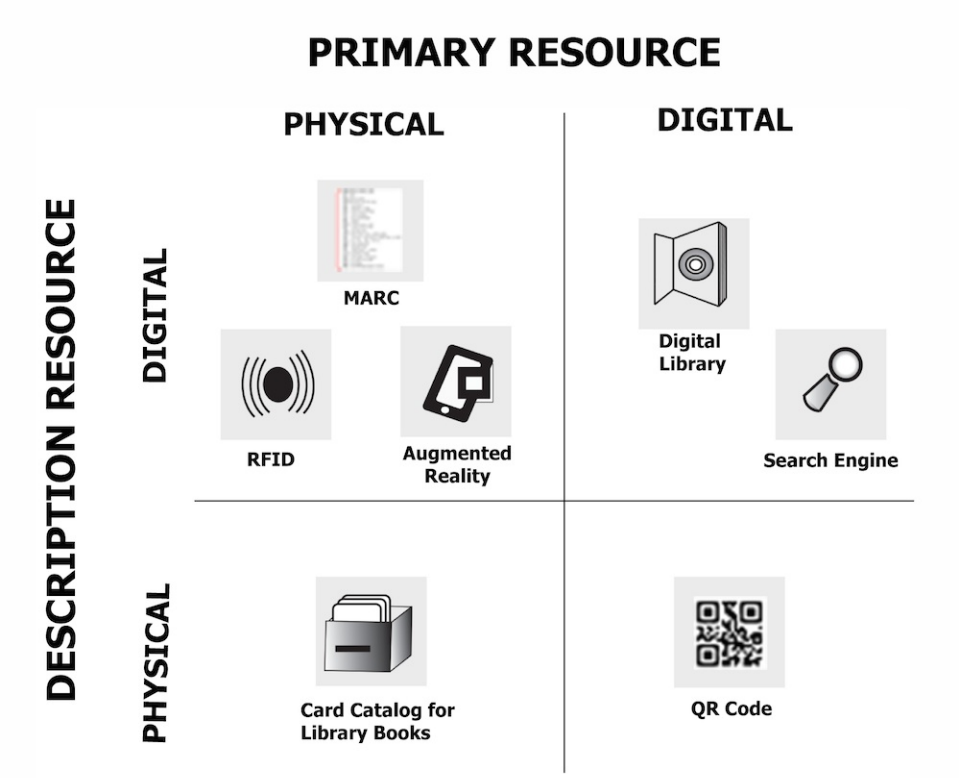
For example, many Twitter users treat the 140-character message body as the primary resource, while the associated metadata about the message and sender (is it a forward, reply, link, etc.) is less important. However, for firms that use Twitter metadata to measure sender and brand impact, or identify social networks and trends, the focus is the metadata, not the content.[181][DS]

As another example, players on professional sports teams are human resources, but millions of people participate in fantasy sports leagues where teams consist of resources based on the statistics generated by the actual human players. Put another way, the associated resources in the actual sports are treated as the primary ones in the fantasy leagues.[182][Bus]

4.2.5 Resource Format x Focus

Applying the format contrast between physical and digital resources to the focus distinction between primary and descriptive resources yields a useful *framework* with four categories of resources (Figure 4.4, Resource Format x Focus.).

Figure 4.4. Resource Format x Focus.



The distinctions of resource format and resource focus combine to distinguish four categories of resources: physical resources, digital resources, physical descriptions , and digital descriptions.

4.2.5.1 Physical Description of a Primary Physical Resource

The oldest relationship between descriptive resources and physical resources is when descriptions or other information about physical resources are themselves encoded in a physical form. Nearly ten thousand years ago in Mesopotamia small clay tokens kept in clay containers served as inventory information to count units of goods or livestock. It took 5000 years for the idea of stored tokens to evolve into Cuneiform writing in which marks in clay stood for the tokens and made both the tokens and containers unnecessary.

A Cuneiform Document at the Pergamon

The Pergamon Museum in Berlin contains a very large collection of Babylonian, Persian, and Assyrian artifacts that are nearly three thousand years old. including numerous cuneiform clay tablets like this one.

(Photo by R. Glushko.)

Printed cards served as physical description resources for books in libraries for nearly two centuries.

4.2.5.2 Digital Description of a Primary Physical Resource

Here, the digital resource describes a physical resource. The most familiar example of this relationship is the online library catalog used to find the shelf location of physical library resources, which beginning in the 1960s replaced the physical cards with database records. The online catalogs for museums usually contain a digital photograph of the painting, item of sculpture, or other museum object that each catalog entry describes.

Bar Code Shopping in A Virtual Supermarket

Woolworth's Australia created a"virtual supermarket" with product photos and bar codes. Scanning places an order, which is delivered from the customer's local supermarket.

(Photo by R. Glushko.)

Digital description resources for primary physical resources are essential in supply chain management, logistics retailing, transportation, and every business model that depends on having timely and accurate information about where things are or about their current states. This digital description resource is created as a result of an interaction with a primary physical resource like a temperature sensor or with some secondary physical resource that is already associated with the primary physical resource like an RFID tag or barcode.

Augmented reality systems combine a layer of real-time digital information about some physical object to a digital view or representation of it. The yellow "first down" lines superimposed in broadcasts of football games are a familiar example. Augmented reality techniques that superimpose identifying or descriptive metadata are used in displays to support the operation or maintenance of complex equipment, in smart phone navigation and tourist guides, in advertising, and in other domains where users might otherwise need to consult a separate information source. Advanced airplane cockpit technology includes heads-up displays that present critical data based on available instrumentation, including augmented reality runway lights when visibility is poor because of clouds or fog.

Augmented reality displays have recently been incorporated into wearable technology like Google Glass, which mounts on eyeglass frames to display information obtained from the Internet after being requested by voice commands. Some luxury car brands have incorporated similar technology to project dashboard data, traffic conditions, and directions on the driver's windshield.

4.2.5.3 Digital Description of a Primary Digital Resource

A digital resource describes a digital resource. This is the relationship in a digital library or any web-based organizing system, making it possible to access a primary digital resource directly from the digital secondary resource.

4.2.5.4 Physical Description of a Primary Digital Resource

This is the relationship implemented when we encounter an embedded QR barcode in newspaper or magazine advertisements, on billboards, sidewalks, t-shirts, or on store shelves. Scanning the QR code with a mobile phone camera can launch a website that contains information about a product or service, place an order for one unit of the pointed-to- item in a web catalog, dial a phone number, or initiate another application or service identified by the QR code.

4.3 Resource Identity

Determining the identity of resources that belong in a domain, deciding which properties are important or relevant to the people or systems operating in that domain, and then specifying the principles by which those properties encapsulate or define the relationships among the resources are the essential tasks when building any organizing system. In organizing systems used by individuals or with small scope, the methods for doing these tasks are often *ad hoc* and unsystematic, and the organizing systems are therefore idiosyncratic and do not scale well. At the other extreme, organizing systems designed for institutional or industry-wide use, especially in information-intensive domains, require systematic design methods to determine which resources will have separate identities and how they are related to each other. These resources and their relationships are then described in conceptual models which guide the implementation of the systems that manage the resources and support interactions with them.[186][Com]

4.3.1 Identity and Physical Resources

Our human visual and cognitive systems do a remarkable job at picking out objects from their backgrounds and distinguishing them from each other. In fact, we have little difficulty recognizing an object or a person even if we are seeing them from a novel distance and viewing angle or with different lighting, shading, and so on. When we watch a football game, we do not have any trouble perceiving the players moving around the field, and their contrasting uniform colors allow us to see that there are two different teams.

The perceptual mechanisms that make us see things as permanent objects with contrasting visible properties are just the prerequisite for the organizing tasks of identifying the specific object, determining the categories of objects to which it belongs, and deciding which of those categories is appropriate to emphasize. Most of the time we carry out these tasks in an automatic, unconscious way; at other times we make conscious decisions about them. For some purposes we consider a sports team as a single resource, as a collection of separate players for others, as offense and defense, as starters and reserves, and so on.[187][CogSci]

Although we have many choices about how we can organize football players, all of them will include the concept of a single player as the smallest identifiable resource. We are never going to think of a football player as an intentional collection of separately identified leg, arm, head, and body resources because there are no other ways to "assemble" a human from body parts. Put more generally, there are some natural constraints on the organization of matter into parts or collections based on sizes, shapes, materials, and other properties that make us identify some things as indivisible resources in some domain.

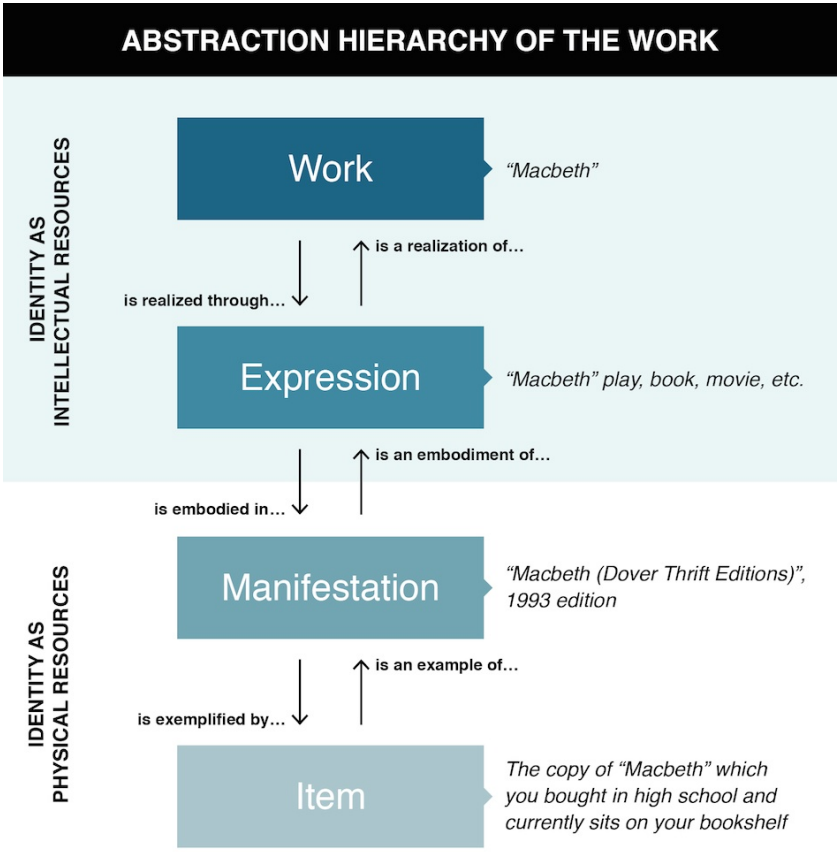
4.3.2 Identity and Bibliographic Resources

Pondering the question of *identity* is something relatively recent in the world of librarians and catalogers. Libraries have been around for about 4000 years, but until the last few hundred years librarians created "bins" of headings and topics to organize resources without bothering to give each individual item a separate identifier or name. This meant searchers first had to make an educated guess as to which bin might house their desired information—"Histories"? "Medical and Chemical Philosophy"?—then scour everything in the category in a quest for their desired item. The choices were *ad hoc* and always local—that is, each cataloger decided the bins and groupings for each catalog.

The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see Figure 4.5, The FRBR Abstraction Hierarchy.) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*.

If we revisit the question "What is this thing we call *Macbeth*?" we can see how different ways of answering fit into this abstraction hierarchy. The most specific answer is that " *Macbeth*" is a specific *item*, a very particular and individual resource, like that dog-eared paperback with yellow marked pages that you owned when you read *Macbeth* in high school. A more abstract answer is that *Macbeth* is an idealization called a *work*, a category that includes all the plays, movies, ballets, or other intellectual creations that share a recognizable amount of the plot and meaning from the original Shakespeare play.

Figure 4.5. The FRBR Abstraction Hierarchy.



The abstraction hierarchy for identifying resources yields four different answers about the identity of an information resource.

4.3.3 Identity and Information Components

In information-intensive domains, documents, databases, software applications, or other explicit repositories or sources of information are ubiquitous and essential to the creation of value for the user, reader, consumer, or customer. Value is created through the comparison, compilation, coordination or transformation of information in some chain or choreography of processes operating on information flowing from one information source or process to another. These processes are employed in accounting, financial services, procurement, logistics, supply chain management, insurance underwriting and claims processing, legal and professional services, customer support, computer programming, and energy management.

The processes that create value in information-intensive domains are "glued together" by shared *information components* that are exchanged in documents, records, messages, or resource descriptions of some kind. Information components are the primitive and abstract resources in information-intensive domains. They are the units of meaning that serve as building blocks of composite descriptions and other information artifacts.

The value creation processes in information-intensive domains work best when their component parts come from a common *controlled vocabulary* for components, or when each uses a vocabulary with a granularity and semantic precision compatible with the others. For example, the value created by a personal health record emerges when information from doctors, clinics, hospitals, and insurance companies can be combined because they all share the same "patient" component as a logical piece of information.

This abstract definition of information components does not help identify them, so we will introduce some heuristic criteria: An *information component* can be: (1) Any piece of information that has a unique label or identifier or (2) Any piece of information that is self-contained and comprehensible on its own.[192][Com]

These two criteria for determining the identity of information components are often easy to satisfy through observations, interviews, and task analysis because people naturally use many different types of information and talk easily about specific components and the documents that contain them. Some common components (e.g., person, location, date, item) and familiar document types (e.g., report, catalog, calendar, receipt) can be identified in almost any domain. Other components need to be more precisely defined to meet the more specific semantic requirements of narrower domains. These smaller or more fine-grained components might be viewed as refined or qualified versions of the generic components and document types, like course grade and semester components in academic transcripts, airport codes and flight numbers in travel itineraries and tickets, and drug names and dosages in prescriptions.

Decades of practical and theoretical effort in conceptual modeling, relational theory, and database design have resulted in rigorous methods for identifying *information components* when requirements and business rules for information can be precisely specified. For example, in the domain of business transactions, required information like item numbers, quantities, prices, payment information, and so on must be encoded as a particular type of data—integer, decimal, Unicode string, etc.— with clearly defined possible values and that follows clear occurrence rules.[193][Com]

Identifying components can seem superficially easy at the transactional end of the Document Type Spectrum (see the sidebar in §4.2.1 Resource Domain (page 167)), with orders or invoices, forms requiring data entry, or other highly-structured document types like product catalogs, where pieces of information are typically labeled and delimited by boxes, lines, white space or other presentation features that encode the distinctions between types of content. For example, the presence of ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION labels on the fields of an online order form suggests these pieces of information are semantically distinct components in a retail application. In addition, these labels might have analogues in variable names in the source code that implements the order form, or as tags in a XML document created by the ordering application; `<CustName>John Smith</CustName>` and `<Item>A-19</Item>` in the order document can be easily identified when it is sent to the other services by the order management application.

But the theoretically grounded methods for identifying components like those of relational theory and normalization that work for structured data do not strictly apply when information requirements are more qualitative and less precise at the narrative end of the Document Type Spectrum. These information requirements are typical of narrative, unstructured and semi-structured types of documents, and information sources like those often found in law, education, and professional services. Narrative documents include technical publications, reports, policies, procedures and other less structured information, where semantic components are rarely labeled explicitly and are often surrounded by text that is more generic. Unlike transactional documents that depend on precise semantics because they are used by computers, narrative documents are used by people, who can ask if they are not sure what something means, so there is less need to explicitly define the meaning of the *information components*. Occasional exceptions, such as where components in narrative documents are identified with explicit labels like NOTE and WARNING, only prove the rule.

4.3.4 Identity and Active Resources

Active resources (§4.4.3.2) initiate effects or create value on their own. In many cases an inherently passive physical resource like a product package or shipping pallet is transformed into an active one when associated with an RFID tag or bar code. Mobile phones contain device or subscriber IDs so that any information they communicate can be associated both with the phone and often, through indirect reference, with a particular person. If the resource has an IP address, it is said to be part of the "Internet of Things."[194][Com]

Organizing systems that create value from active resources often co-exist with or complement organizing systems that treat its resources as passive. In a traditional library, books sat passively on shelves and required users to read their spines to identify them. Today, some library books contain active RFID tags that make them dynamic information sources that self-identify by publishing their own locations. Similarly, a supermarket or department store might organize its goods as physical resources on shelves, treating them as passive resources; superimposed on that traditional organizing system is one that uses point-of-sale transaction information created when items are scanned at checkout counters to automatically re-order goods and replenish the inventory at the store where they were sold. In some stores the shelves contain sensors that continually "talk to the goods" and the information they gather can maintain inventory levels and even help prevent theft of valuable merchandise by tracking goods through a store or warehouse. The inventory becomes a collection of active resources; each item eager to announce its own location and ready to conduct its own sale. Another category of inanimate objects that are active resources are those that use Twitter to communicate their status or sensor measurements. These include bridges, rivers, and the Curiosity Rover on Mars.

The extent to which an active resource is "smart" depends on how much computing capability it has available to refine the data it collects and communicates. A large collection of sensors can transmit a torrent of captured data that requires substantial processing to distinguish significant events from those that reflect normal operation, and also from those that are statistical outliers with strange values caused by random noise. This challenge gets qualitatively more difficult as the amount of data grows to *big data* size, because a one in million event might be a statistical outlier that can be ignored, but if there are a thousand similar outliers in a billion sensor readings, this cluster of data probably reveals something important. On the other hand, giving every sensor the computing capability to refine its data so that it only communicates significant information might make the sensors too expensive to deploy.[195][DS]

Big Data Makes "Smart" Soccer Players

The German World Cup soccer team, which won the 2014 World Cup, took advantage of sophisticated data collection and analysis to optimize player skill and strategy training. German software firm SAP analyzed video data from on-field cameras that captured thousands of data points per second about player position and movement to identify improvements in passing and ball handling for German players and detect weaknesses in opponents. German sports equipment firm Adidas designed cleats with sensors that track mileage, field position, and movements. (Norton 2014) and (Reynolds 2014).

4.4 Naming Resources

Determining the identity of the thing, document, information component, or data item we need is not always enough. We often need to give that resource a name, a label that will help us understand and talk about what it is. But naming is not just the simple task of assigning a sequence of characters. In this section, we will discuss why we name, some of the problems with naming, and the principles that help us name things in useful ways.

4.4.1 What's in a Name?

When a child is born, its parents give it a name, often a very stressful and contentious decision. Names serve to distinguish one person from another, although names might not be unique—there are thousands of people named James Smith and Maria Garcia. Names also, intentionally or unintentionally, suggest characteristics or aspirations. The name given to us at birth is just one of the names we will be identified with during our lifetimes. We have nicknames, names we use professionally, names we use with friends, and names we use online. Our banks, our schools, and our governments will know who we are because of numbers they associate with our names. As long as it serves its purpose to identify you, your name could be anything.[196][Law]

Resources other than people need names so we can find them, describe them, reuse them, refer or link to them, record who owns them, and otherwise interact with them. In many domains the names assigned to resources are also influenced or constrained by rules, industry practice, or technology considerations.

4.4.2 The Problems of Naming

Giving names to anything, from a business to a concept to an action, can be a difficult process and it is possible to do it well or do it poorly. The following section details some of the major challenges in assigning a name to a resource.

4.4.2.1 The Vocabulary Problem

Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept. The words people choose to name or describe things are embodied in their experiences and context, so people will often disagree in the words they use. Moreover, people are often a bit surprised when it happens, because what seems like the natural or obvious name to one person is not natural or obvious to another. One way to avoid surprises is to have people cooperate in choosing names for resources, and information architects often use participatory design techniques of card sorting or free listing for this purpose.[197][Ling] [198][IA]

Back in the 1980s in the early days of computer user interface design, George Furnas and his colleagues at Bell Labs conducted a set of experiments to measure how much people would agree when they named some resource or function. The short answer: very little. Left to our own devices, we come up with a shockingly large number of names for a single common thing.

In one experiment, a thousand pairs of people were asked to "write the name you would give to a program that tells about interesting activities occurring in some major metropolitan area." Less than 12 pairs of people agreed on a name. Furnas called this phenomenon *the vocabulary problem,* concluding that no single word could ever be considered the "best" name.[199][CogSci]

4.4.2.2 Homonymy, Polysemy, and False Cognates

Sometimes the same word can refer to different resources—a "bank" can be a financial institution or the side of a river. When two words are spelled the same but have different meanings they are *homographs*; if they are also pronounced the same they are *homonyms*. If the different meanings of the homographs are related, they are *polysemes*.

Resources with homonymous and polysemous names are sometimes incorrectly identified, especially by an automated process that cannot use common sense or context to determine the correct referent. Polysemy can cause more trouble than simple homography because the overlapping meaning might obscure the misinterpretation. If one person thinks of a "shipping container" as being a cardboard box and orders some of them, while another person thinks of a "shipping container" as the large box carried by semi-trailers and stacked on cargo ships, their disagreement might not be discovered until the wrong kinds of containers arrive.[200][CogSci]

Many words in different languages have common roots, and as a result are often spelled the same or nearly the same. This is especially true for technology words; for example, "computer" has been borrowed by many languages. The existence of these *cognates* and borrowed words makes us vulnerable to false

Unreliable Names: Knockin' On Heaven's Door

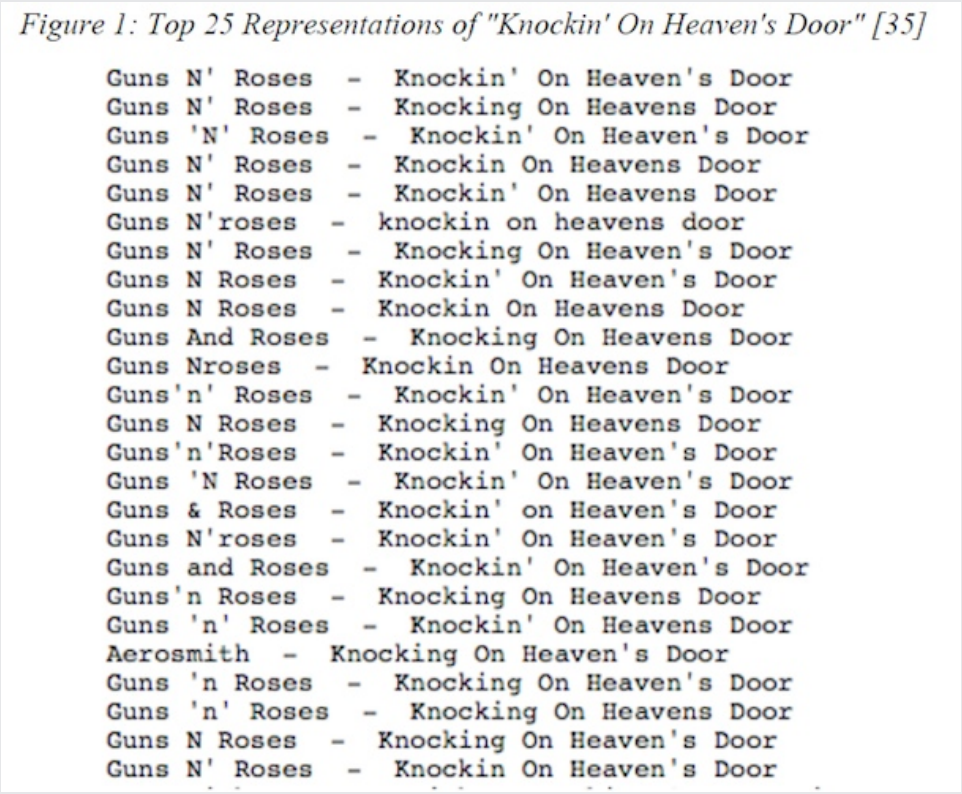


Figure 1: Top 25 Representations of "Knockin' On Heaven's Door" [35]

Guns N' Roses - Knockin' On Heaven's Door
Guns N' Roses - Knocking On Heavens Door
Guns 'N' Roses - Knockin' On Heaven's Door
Guns N' Roses - Knockin On Heavens Door
Guns N' Roses - Knockin' On Heavens Door
Guns N'roses - knockin on heavens door
Guns N' Roses - Knocking On Heaven's Door
Guns N Roses - Knockin' On Heaven's Door
Guns N Roses - Knockin On Heavens Door
Guns And Roses - Knocking On Heavens Door
Guns Nroses - Knockin On Heavens Door
Guns'n' Roses - Knockin' On Heaven's Door
Guns N Roses - Knocking On Heavens Door
Guns'n'Roses - Knockin' On Heaven's Door
Guns 'N Roses - Knockin' On Heaven's Door
Guns & Roses - Knockin' on Heaven's Door
Guns N'roses - Knockin' On Heaven's Door
Guns and Roses - Knockin' On Heaven's Door
Guns'n Roses - Knocking On Heavens Door
Guns 'n' Roses - Knockin' On Heavens Door
Aerosmith - Knocking On Heaven's Door
Guns 'n Roses - Knocking On Heaven's Door
Guns 'n' Roses - Knocking On Heavens Door
Guns N Roses - Knocking On Heaven's Door
Guns N' Roses - Knockin On Heaven's Door

In 2008, Music recommendation service Last.fm employee Richard Jones compiled a list of the 100 most descriptions of the Guns N' Roses recording of Bob Dylan's song "Knockin' on Heaven's Door." The 21st most common description of the song incorrectly attributes the recording to Aerosmith.

Reprinted in Figure 1 of (Hemerly 2011). Used by permission here.

cognates. When a word in one language has a different meaning and refers to different resources in another, the results can be embarrassing or disastrous. "*Gift*" is poison in German; "*pain*" is bread in French.

4.4.2.3 Names with Undesirable Associations

False cognates are a special category of words that make poor names, and there are many stories relating product marketing mistakes, where a product name or description translates poorly, into other languages or cultures, with undesirable associations.[201][Ling] Furthermore, these undesirable associations differ across cultures. For example, even though floor numbers have the straightforward purpose to identify floors from lowest to highest levels, most buildings in Western cultures skip the 13th floor because many people think 13 is an unlucky number. In many East and Southeast Asian buildings, the 4th floor is skipped. In China the number 4 is dreaded because it sounds like the word for "death," while 8 is prized because it sounds like the word for "wealth."

While it can be tempting to dismiss unfamiliar biases and beliefs about names and identifiers as harmless superstitions and practices, their implications are ubiquitous and far from benign. *Alphabetical ordering* might seem like a fair and non-discriminatory arrangement of resources, but because it is easy to choose the name at the top of an alphabetical list, many firms in service businesses select names that begin with "A," "AA," or even "AAA" (look in any printed service directory). A consequence of this bias is that people or resources with names that begin with letters late in the alphabet are systematically discriminated against because they are often not considered, or because they are evaluated in the context created by resources earlier in the alphabet rather than on their own merit.[202][Bus]

4.4.2.4 Names that Assume Impermanent Attributes

Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.

Web resources are often referred to using URLs that contain the domain name of the server on which the resource is located, followed by the directory path and file name on the computer running the server. This treats the current location of the resource as its name, so the name will change if the resource is moved. It also means that resources that are identical in content, like those at an archive or mirror website, will have different names than the original even though they are exact copies. An analogous problem is faced by restaurants or businesses with street names or numbers in their names if they lose their leases or want to expand.[203][Bus]

From 'Kentucky Fried Chicken' to 'KFC'

"Kentucky Fried Chicken" was founded in 1930 by Harland Sanders as a tiny restaurant in a gas station storeroom in Corbin, Kentucky. It was one of the first fast-food chains to go international, and in 1987 was the first Western restaurant chain to open in China. It changed its name to "KFC" a few years later, no doubt in part because in Beijing, Moscow, London and other locations not anywhere near Kentucky many people have probably never heard of the place.

(Photo by Kyle Taylor. CC-BY-2.0 license.)

Some dynamic web resources that are generated by programs have URIs that contain information about the server technology used to create them. When the technology changes, the URIs will no longer work.[204][Web]

Some resources have names that include page numbers, which disappear or change when the resource is accessed in a digital form. For example, the standard citation format for legal opinions uses the page number from the printed volume issued by West Publishing, which has a virtual monopoly on the publishing of court opinions and other types of legal documents.[205][Law]

Some resources have names that contain dates, years or other time indicators, most often to point to the future. The film studio named "20th Century Fox" took on that name in the 1930s to give it a progressive, forward-looking identity, but today a name with "20th Century" in it does the opposite.[206][Ling]

4.4.2.5 The Semantic Gap

The *semantic gap* is the difference in perspective in naming and description when resources are described by automated processes rather than by people.[207][Com]

The *semantic gap* is largest when computer programs or sensors obtain and name some information in a format optimized for efficient capture, storage, *decoding*, or other technical criteria. The names—like *IMG20268.jpg* on a digital photo—might make sense for the camera as it stores consecutively taken photos but they are not good names for people. We may prefer names that describe the content of the picture, like *GoldenGateBridge.jpg*.

When we try to examine the content of computer-created or sensor-captured resources, like a clip of music or a compiled software program, a text rendering of the content simply looks like nonsense. It was designed to be interpreted by a computer program, not by a person.

Semantic Gap: Name This Tune

The format of this MP3 recording is designed to be read by a music player, not by people.

(Screenshot by R. Glushko.)

4.4.3 Choosing Good Names and Identifiers

If someone tells you they are having dinner with their best friend, a cousin, someone with whom they play basketball, and their professional mentor from work, how many places at the table will be set? Anywhere from two to five; it is possible all those relational descriptions refer to a single person, or to four different people, and because "friend," "cousin," "basketball teammate" and "mentor" do not name specific people you will have to guess who is coming to dinner.

If instead of descriptions you are told that the dinner guests are Bob, Carol, Ted, and Alice, you can count four names and you know how many people are having dinner. But you still cannot be sure exactly which four people are involved because there are many people with those names.

The uncertainty is eliminated if we use identifiers rather than names. *Identifiers* are names that refer unambiguously to a specific person, place, or resource because they are assigned in a controlled way. Identifiers are often strings of numbers or letters rather than words to avoid the biases and associations that words can convey. For example, a professor might grade exams that are identified by student numbers rather than names.

Names {and, or, vs} Identifiers

People change their names for many reasons: when they get married or divorced, because their name is often mispronounced or misspelled, to make a political or ethnic statement, or because they want to stand out. A few years a football player with a large ego named Chad Johnson, which is the second most common surname in the US, decided to change his name to his player number of 85, becoming Chad "Ochocinco." He had an ochocinco.com website and used the ochocinco name on Facebook and Twitter. In a bit of irony, when Ochocinco wanted to put Ocho Cinco on the back of his football jersey, the football league would not let him because his legal name does not have a space in it. That surely contributed to his decision to change his name back to Chad Johnson in 2012.

When you go to coffee shops, you are often asked your name, which the cashier writes on the empty cup so that your drink can be identified after the *barista* makes it. They do not actually need your name; just as some establishments use a receipt number to distinguish orders, what they need is an identifier. So even if your name is Joe, you can tell them it is Thor, Wotan, Mercurio, El Greco, Clark Kent, or any other name that is likely to be a unique identifier for the minute it takes to make your beverage.[208][Bus]

The distinction between *names* and *identifiers* for people is often not appreciated. (See the sidebar, Names {and, or, vs} Identifiers (page 194).)

4.4.3.1 Make Names Informative

The most basic principle of naming is to choose names that are informative, which makes them easier to understand and remember. It is easier to tell what a computer program or XML document is doing if it uses names like “ItemCost” and “TotalCost” rather than just “I” or “T.” People will enter more consistent and reusable address information if a form asks explicitly for “Street,” “City,” and “PostalCode” instead of “Line1” and “Line2.”

Identifiers can be designed with internal structure and semantics that conveys information beyond the basic aspect of pointing to a specific resource. An International Standard Book Number (ISBN) like “978-0-262-07261-8” identifies a resource (07261=“Document Engineering”) and also reveals that the resource is a book (978), in English (0), and published by The MIT Press (262).[209][Com]

The navigation points that mark intersections of radial signals from ground beacons or satellites that are crucial to aircraft pilots used to be meaningless five-letter codes that were changed to make them suggest their locations; semantic landmark names made pilots less likely to enter the wrong names into navigation systems, For example, some of the navigation points near Orlando, Florida —the home of Disney World—are MICKI, MINEE, and GOOFY.[210][Ling]

4.4.3.2 Use Controlled Vocabularies

One way to encourage good names for a given resource domain or task is to establish a *controlled vocabulary*. A *controlled vocabulary* is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use.

A *controlled vocabulary* is not simply a set of allowed words; it also includes their definitions and often specifies rules by which the vocabulary terms can be used and combined. Different domains can create specific controlled vocabularies for their own purposes, but the important thing is that the vocabulary be used consistently throughout that domain.

For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as *authority control*.

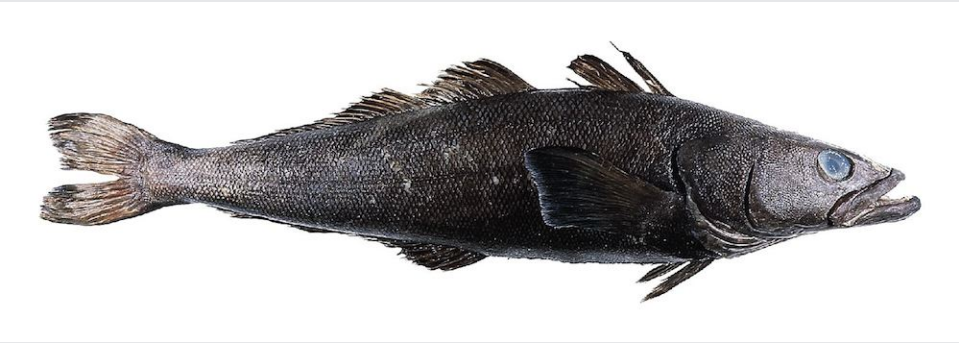
Official authority files are maintained for many resource domains: a gazetteer associates names and locations and tells us whether we should be referring to Bombay or Mumbai; the Domain Name System (DNS) maps human-oriented do-

main and host names to their IP addresses; the Chemical Abstracts Service Registry assigns unique identifiers to every chemical described in the open scientific literature; numerous institutions assign unique identifiers to different categories of animal species.

In some cases, authority files are created or maintained by a community, as in the case of MusicBrainz, an "open music encyclopedia" to which users contribute information about artists, releases, tracks, and other aspects of music. Music metadata is notoriously unreliable; one study found over 100 variations in the description of the *Knockin' on Heaven's Door* song (written by Bob Dylan) as recorded by Guns N' Roses.[214][Web]

4.4.3.3 Allow Aliasing

Aliasing: Bad for this Fish



A fish once known as the Patagonian Toothfish because of its large and unattractive teeth became popular in American restaurants when a fish wholesaler began marketing it as the Chilean Sea Bass even though it is usually found farther south in cold Antarctic waters and it is not a sea bass. Unfortunately for the fish, this alias was so successful that it led to overfishing, threatening the survival of the species. Some environmentally-oriented chefs, restauranteurs, and seafood distributors organized a boycott to save the fish. (Fabricant 2002)

(Photo published by the United States Government. Not protectable by copyright (17 USC Sec. 105).)

A *controlled vocabulary* is extremely useful to people who use it, but if you are designing an organizing system for other people who do not or cannot use it, you need to accommodate the variety of words they will actually use when they seek or describe resources. The authoritative name of a certain fish species is *Amphiprion ocellaris*, but most people would search for it as "clownfish," "anemone fish," or even by its more familiar film name of *Nemo.*

Furnas suggests “unlimited aliasing” to connect the uncontrolled or natural vocabularies that people use with the controlled one employed by the organizing system. By this he means that there must be many alternate access routes to each word or function that a user is trying to find. For example, the birth name of the 42nd President of the United States of America is “William Jefferson Clinton,” but web pages that refer to him as “Bill Clinton” are vastly more common, and searches for the former are redirected to the latter. A related mechanism used by search engines is spelling correction, essentially treating all the incorrect spellings as aliases of the correct one (“did you mean California?” when you typed “Claifornia”).

4.4.3.4 Make Identifiers Unique or Qualified

Even though an identifier refers to a single resource, this does not mean that no two identifiers are identical. One military inventory system might use stock number 99 000 1111 to identify a 24-hour, cold-climate ration pack, while another inventory system could use the same number to identify an electronic radio valve. Each identifier is unique in its inventory system, but if a supply request gets sent to the wrong warehouse hungry soldiers could be sent radio valves instead of rations.[215][Law]

We can prevent or reduce identifier collisions by adding information about the *namespace*, the domain from which the names or identifiers are selected, thus creating what are often called *qualified names*. There are several dozen US cities named “Springfield” and “Washington,” but adding state codes to mail addresses distinguishes them. Likewise, we can add prefixes to XML element names when we create documents that reuse components from multiple document types, distinguishing `<book:Title>` from `<legal:Title>`.

We can fix problems like these by qualifying or extending the identifier, or by creating a *globally unique identifier (GUID)*, one that will never be the same as another identifier in any organizing system anywhere else. One easy method to create a GUID is to use a URL you control and append a string to it, the same approach that gives every web page a unique address. GUIDs are often used to identify software objects, the resources in distributed systems, or data collections.[217][Com]

Because they are not created by an algorithm whose results are provably unique, we do not consider fingerprints, or other biometric information, to be globally unique identifiers for people, but for all practical purposes they are.[218][Com]

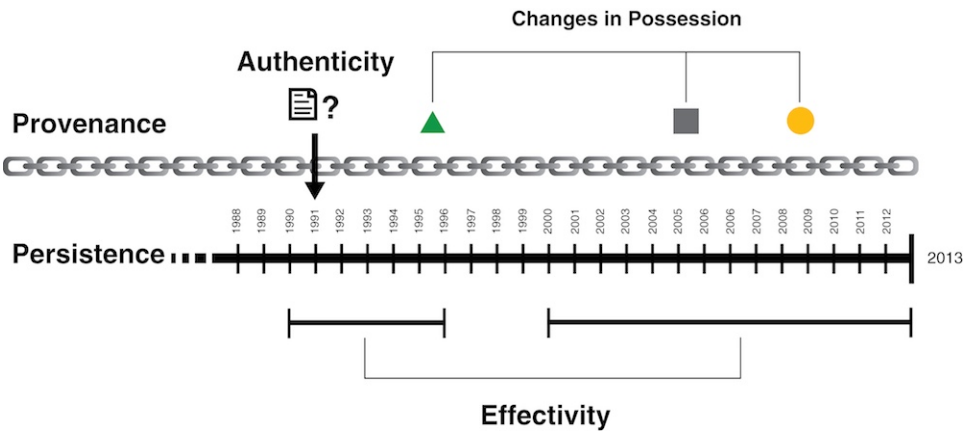
4.4.3.5 Distinguish Identifying and Resolving

When the identifier does not contain information about resource location, it must be“resolved” to determine the location. With physical resources, *resolution* takes place with the aid of signs, maps, or other associated resources that describe the resource arrangement in some physical environment; for example, “You are here” maps associate each resource identifier with a coordinate or other means of finding it on the map. With digital resources, the resolver is a directory system or service that interprets an identifier and looks up its location or directly initiates resource retrieval.

4.5 Resources over Time

Problems of “what is the resource?” and “how do we identify it?” are complex and often require ongoing work to ensure they are properly answered as an organizing system evolves. We might need to know how a resource does or does not change over time (its *persistence*), whether its state and content come into play at a specified point in time (its *effectivity*), whether the resource is what it is said to be (its *authenticity*), and sometimes who has certified its authenticity over time (its *provenance*). A resource might have persistence, but only the provenance provided by an documented chain of custody enables questions about authenticity to be answered with authority. Effectivity describes the limits of a resource's lifespan on the time line.

Figure 4.6. Resources over Time.



Four considerations that arise with respect to the maintenance of resources over time are their persistence, provenance, authenticity, and effectivity.

4.5.1 Persistence

Even if you have reached an agreement as to the meaning of "a thing" in your organizing system, you still face the question of the identity of the resource over time, or its *persistence*.

4.5.1.1 Persistent Identifiers

How long must an identifier last? Coyle gives the conventional, if unsatisfying, answer: "As long as it is needed." In some cases, the time frame is relatively short. When you order a specialty coffee and the *barista* asks for your name, this identifier only needs to last until you pick up your order at the end of the counter. But other time frames are much longer. For libraries and repositories of scientific, economic, census, or other data the time frame might be "forever."

The Great Sphinx at Giza

The Great Sphinx has persisted for over four thousand years. It has survived acts of vandalism, target practice by Napoleon's artillery, shoulder-deep burial in desert sandstorms, and eventual excavation in the early 20th century.

(Photo by R. Glushko.)

The design of a scheme for persistent identifiers must consider both the required time frame and the number of resources to be identified. When the Internet Protocol (IP) was designed in 1980, it contained a 32-bit address scheme, sufficient for over 4 billion unique addresses. But the enormous growth of the Internet and the application of IP addresses to resources of unexpected types have required a new addressing scheme with 128 bits.[220][Com]

Recognition that URIs are often not persistent as identifiers for web-based resources led the *Association of American Publishers (AAP)* to develop the *Digital Object Identifier (DOI)* system. The location and owner of a digital resource can change, but its DOI is permanent.[221][Com]

4.5.1.2 Persistent Resources

Even though persistence often has a technology dimension, it is more important to view it as a commitment by an institution or organization to perform activities over time to ensure that a resource is available when it is needed. Put another way, preservation (§3.5.2) and governance (§3.5.4) are activities carried out to ensure the outcome of persistence.

The subtle relationship between preservation and persistence raises some interesting questions about what it means for a resource to stay the same over time. One way to think of persistence is that a persistent resource is never changed. However, physical resources often require maintenance, repair, or restoration to keep them accessible and usable, and we might question whether at some point these activities have transformed them into different resources.[222][Phil] Likewise, digital resources require regular backup and migration to keep them available and this might include changing their digital format.

Many resources like online newspapers or blog feeds continually change their content but still have persistent identifiers. This suggests we should think of persistence more abstractly, and consider as persistent resources any that remain functionally the same to support the same interactions at any point in their lifetimes, even if their physical properties or information values change.

Active resources implemented as computational agents or web services might be re-implemented numerous times, but as long as they do not change their interfaces they can be deemed to be persistent from the perspective of other resources that use them. Similarly, the dataset that defines a user or customer model in a recommendation system should be treated as a persistent resource; it includes information like name and date of birth that is persistent in the traditional sense; but it might also include "last purchase" and "current location," which must change frequently to maintain the accuracy and usefulness of the customer model.

4.5.2 Effectivity

Many resources, or their properties, also have locative or temporal *effectivity*, meaning that they come into effect at a particular time and/or place; will almost certainly cease to be effective at some future date, and may cease to be effective in different places.

Temporal effectivity, sometimes known as "time-to-live," is generally expressed as a range of two dates. It consists of a date on which the resource is effective, and optionally a date on which the resource ceases to be effective, or becomes stale. For some types of resources, the *effective date* is the moment they are created, but for others, the effective date can be a time different from the moment of creation. For example, a law passed in November may take effect on January 1 of the following year, and credit cards first need to be activated and then can no longer be used after their expiration date. An "effective date" is the counterpart of the "Best Before" date on perishable goods. That date indicates when a product goes bad, whereas an item's effective date is when it "goes good" and the resource that it supersedes needs to be disposed of or archived.

Locative effectivity considers borders, security, roadways, altitude, depth and other geographic factors. Some types of resources, including people, are restricted as to where they may or may not be transported and/or used, such as hazardous cargo, explosives, narcotics, pharmaceuticals, alcohol, and cannabinoids. Jurisdictional issues concern borders, transportation corridors, weather stations, and geographic surveys. Parachutes are altitude-sensitive and scuba diving cylinders are depth-sensitive.

Effectivity concerns sometimes intersect with authority control for names and places. Name changes for resources often are tied to particular dates, events, and locations. Laws and regulations differ across organizational and geopolitical boundaries, and those boundaries often change. Some places that have been the site of civil unrest, foreign occupation, and other political disruptions have had many different names over time, and even at the same time.[224][Law]

In Which Country Do You Live?

Even if you always live in the same place, the answer to "what country do you live in?" can depend on when it is asked. Consider the case of an elderly woman born in 1929 in Zemum, a district in the eastern European city of Belgrade, who has never moved. The place she lives has been part of seven different countries during her lifetime: Kingdom of Yugoslavia (1929-1941); Independent State of Croatia (1941-1945); Federal People's Republic of Yugoslavia (1945-1963); Socialist Federal Republic of Yugoslavia (1963-1992); Federal Republic of Yugoslavia (1992-2003); State Union of Serbia and Montenegro (2003-2006); Republic of Serbia (2007—present).

Today these disputed borders cause a problem for Google Maps when it displays certain international borders. Because Google is subject to the laws of the country where its servers are located, it must present disputed borders to conform with the point of view of the host country when a country-specific Google site is used to access the map.[225][Com]

In most cases effectivity implies persistence requirements because it is important to be able to determine and reconstruct the configuration of resources that was in effect at some prior time. A new tax might go into effect on January 1, but if the government audits your tax returns what matters is whether you followed the law that was in effect when you filed your returns.[226][Bus]

4.5.3 Authenticity

Do You Trust This?

Ustar.net sells photos autographed by celebrities, and each comes with a Certificate of Authenticity that includes a replica of the photo and a signature from a Ustar employee to guarantee that the autograph is an actual hand-signed one. But Ustar does not provide a certificate to guarantee that the employee signature is an authorized one.

(Screenshot by R. Glushko. Source: ustar.net.)

In ordinary use we say that something is *authentic* if it can be shown to be, or has come to be accepted as what it claims to be. The importance and nuance of questions about authenticity can be seen in the many words we have to describe the relationship between "the real thing" (the "original") and something else: copy, reproduction, replica, fake, phony, forgery, counterfeit, pretender, imposter, ringer, and so on.

It is easy to think of examples where authenticity of a resource matters: a signed legal contract, a work of art, a historical artifact, even a person's signature.

The creator or operator of an organizing system, whether human or machine, can authenticate a newly created resource. A third party can also serve as proof of authenticity. Many professional careers are based on figuring out if a resource is authentic.[227][Law]

There is a large body of techniques for establishing the identity of a person or physical resource. We often use judgments about the physical integrity of recorded information when we consider the integrity of its contents.

Digital authenticity is more difficult to establish. Digital resources can be reproduced at almost no cost, exist in multiple locations, carry different names on identical documents or identical names on different documents, and bring about other complications that do not arise with physical items. Technological solutions for ensuring digital authenticity include time stamps, watermarking, encryption, and digital signatures. However, while scholars generally trust technological methods, technologists are more skeptical of them because they can

imagine ways for them to be circumvented or counterfeited. Even when a technologically sophisticated system for establishing authenticity is in place, we can still only assume the constancy of identity as far back as this system reaches in the "chain of custody" of the document.

4.5.4 Provenance

In §3.2.2 we recommended that you analyze any evidence or records about the use of resources as they made their way to you from their headwaters to ensure they have maintained their quality over time. The concept of provenance transforms the passive question of "what has happened to this resource?" into actions that can be taken to ensure that nothing bad can happen to a resource or to enable it to be detected.

The idea that important documents must be created in a manner that can be authenticated and then preserved, with an unbroken chain of custody, goes back to ancient Rome. Notaries witnessed the creation of important documents, which were then protected to maintain their integrity or value as evidence. In organizing systems like museums and archives that preserve rare or culturally important objects or documents this concern is expressed as the principle of *provenance*. This is the history of the ownership of a collection or the resources in it, where they have been and who has had access to the resources.

A uniquely Chinese technique in organizing systems is the imprinting of elaborate red seals on documents, books, and paintings that collectively record the provenance of ownership and the review and approval of the artifact by emperors or important officials.

However, it is not only art historians and custodians of critical documents that need to be concerned with provenance. If you are planning to buy a used car, it is wise to check the vehicle history (using the Vehicle Identification Number, the car's persistent identifier) to make sure it hasn't been wrecked, flooded, or stolen.

4.6 Key Points in Chapter Four

- We can consider a resource to be one of many members of a very broad category, as the unique instance of a category with only one member, or anywhere in between.

  (See §4.1.1 What Is a Resource? (page 162))

- The size of the category—the number of resources that are treated as equivalent—is determined by the properties or characteristics we consider when we examine the resource.

  (See §4.1.1 What Is a Resource? (page 162))

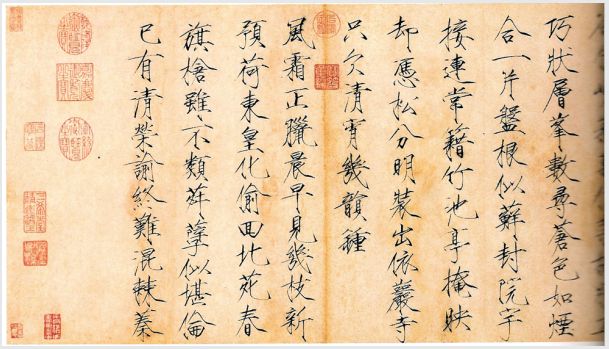
- Organizing systems for physical information resources emphasize description resources or surrogates like bibliographic records that describe the information content rather than their physical properties.

  (See §4.1.1.2 Bibliographic Resources, Information Components, and "Smart Things" as Resources (page 164))

- An identifier is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. The design of a scheme for persistent identifiers must consider both the required time frame and the number of resources to be identified.

  (See §4.1.2 Identity, Identifiers, and Names (page 165))

Chinese Manuscript With Provenance Seals



This beautiful manuscript, preserved in the National Palace Museum in Taipei, was created by Zhao Ji (赵佶), Emperor Huizong, the 8th Emperor of the Chinese Song Dynasty about a thousand years ago. He was famous for his skills in poetry, painting, and calligraphy. There are two poems here; the one on the right describes the techniques for Chinese landscape paintings, while the left one expresses the Emperor's appreciation of plum blossoms, which signal the onset of spring.

The red seals are those of several Ching Dynasty emperors over many generations, with the oldest being at least five hundred years after Huizong created the poems. Stamping your personalized red seal on a resource is analogous to but vastly more elegant and informative than "Liking" a web page today.

(Photo by R. Glushko.)

- *Active resources* create effects or value on their own, sometimes when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, web-based services, self-driving cars, robots, appliances, machines or otherwise ordinary objects like light bulbs, umbrellas, and shoes that have been made "smarter."

  (See §4.2.3.2 Active or Operant Resources (page 173))

- More fine-grained organization reduces *recall*, the number of resources you find or retrieve in response to a query, but increases the *precision* of the recalled set, the proportion of recalled items that are relevant.

  (See §4.3.3 Identity and Information Components (page 185))

- *Agency* is the extent to which a resource can initiate actions on its own. We can define a continuum between completely passive resources that cannot initiate any actions and active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources.

  (See §4.2.3 Resource Agency (page 172))

- Resources become active resources when they contain sensing and communication capabilities.

  (See §4.2.3 Resource Agency (page 172))

- Which resources are primary and which are metadata is often just a decision about which resource is the *focus* of our attention.

  (See §4.2.4 Resource Focus (page 178))

- It can be useful to view domains of information resources on the Document Type Spectrum from weakly-structured narrative content to highly structured transactional content.

  (See the sidebar, The Document Type Spectrum (page 168))

- The concept of identity for bibliographic resources has evolved into a four-level abstraction hierarchy between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*.

  (See §4.3.2 Identity and Bibliographic Resources (page 183) and Figure 4.5, The FRBR Abstraction Hierarchy.)

- If the resource has an IP address, it is part of the "Internet of Things."

  (See §4.3.4 Identity and Active Resources (page 187).)

- Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept.

(See §4.4.2 The Problems of Naming (page 188))

- Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.

  (See §4.4.2.4 Names that Assume Impermanent Attributes (page 191))

- The *semantic gap* is the difference in perspective in naming and description when resources are described by automated processes rather than by people.

  (See §4.4.2.5 The Semantic Gap (page 193))

- The most basic principle of naming is to choose names that are informative.

  (See §4.4.3.1 Make Names Informative (page 195))

- One way to encourage good names for a given resource domain or task is to establish a *controlled vocabulary*. A *controlled vocabulary* is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use.

  (See §4.4.3.2 Use Controlled Vocabularies (page 195))

- For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as *authority control*.

  (See §4.4.3.2 Use Controlled Vocabularies (page 195))

- Preservation and governance are activities carried out to ensure that resources will last as long as they are needed.

  (See §4.5.1 Persistence (page 199))

- Many resources, or their properties, also have locative or temporal *effectivity*, meaning that they come into effect at a particular time and/or place; will almost certainly cease to be effective at some future date, and may cease to be effective in different places.

  (See §4.5.2 Effectivity (page 200))

- The only guarantee of a resource's authenticity is having total oversight over the "chain of custody" from its creation to the present.

  (See §4.5.3 Authenticity (page 202) and §4.5.4 Provenance (page 203))

Endnotes for Chapter 4

[162][Bus] Separating information content from its structure and presentation is essential to re-purposing it for different scenarios, applications, devices, or users. The global information economy is increasingly driven by automated information exchange between business processes. When information flows efficiently from one type of document to another in this chain of related documents, the overlapping content components act as the "glue" that connects the information systems or web services that produce and consume the documents. (Glushko and McGrath 2005).

[163][CogSci] (Furnas, Landauer, Gomez, and Dumais 1987).

[164][Phil] (Linnaeus 1735). Linnaeus is sometimes called the father of modern taxonomy (which is unfair to Aristotle) but he certainly deserves enormous credit for the systematic approach to biological classification that he proposed in *Systema Naturae,* published in 1735. This seminal work contains the familiar kingdom, class, order, family, genus, species hierarchy.

[165][Com] (Glushko and McGrath 2005).

[166][Com] (Kuniavsky 2010).

[168][Com] The ASCII scheme was standardized in the 1960s when computer memory was expensive and most computing was in English-speaking countries, so it is minimal and distinguishes only 128 characters. (Cerf1969) *American Standard Code for Information Interchange (ASCII)* is an ANSI specification. (See `http://en.wikipedia.org/wiki/ASCII`.)

[169][Com] Unicode 6.0 (`http://www.unicode.org/`) has room to encode 109,449 characters for all the *writing systems* in the world, so a single standard can represent the characters of every existing language, even "dead" ones like Sumerian and Hittite. Unicode encodes the scripts used in languages, rather than languages *per se*, so there only needs to one representation of the Latin, Cyrillic, Arabic, etc scripts that are used for writing multiple language. Unicode also distinguishes *characters* from *glyphs*, the different forms for the same character—enabling different *fonts* to be identified as the same character.

[170][Com] Encoding of structure in documents is valuable because titles, sections, links and other structural elements can be leveraged to enhance the user interface and navigational interactions with the digital document and enable more precise information retrieval. Some uses of documents require formats that preserve their printed appearance. "Presentational fidelity" is essential if we imagine a banker or customs inspector carefully comparing a printed document with a computer-generated one to ensure they are identical.

[171][Com] Text encoding specs are well-documented; see (http://www.wotsit.org/list.asp?fc=10).

[172][Com] (Chapman and Chapman 2009).

[174][Com] However, scratching can be simulated using a smart phone or tablet app called djay. See http://www.algoriddim.com/djay.

[175][Law] As a result, digital books are somewhat controversial and problematic for libraries, whose access models were created based on the economics of print publication and the social contract of the copyright first sale doctrine that allowed libraries to lend printed books. Digital books change the economics and first sale is not as well-established for digital works, which are licensed rather than sold (Aufderheide and Jaszi 2011). To protect their business models, many publishers are limiting the number of times ebooks can be lent before they "self-destruct." Some librarians have called for boycotts of publishers in response (http://boycottharpercollins.com).

[176][Bus] The opposing categories of operands and operants have their roots in debates in political economics about the nature of work and the creation of value (Vargo and Lusch 2004) and have more recently played a central role in the development of modern thinking about service design (Constantin and Lusch 1994), (Maglio et al. 2009).

[177][DS] See (Allmendinger and Lombreglia 2005), (Want 2006). (Crawford and Johnson, 2012)

[178][Web] Luis Von Ahn (von Ahn 2004) was the first to use the web to get people to perform "microwork" or "human computation" tasks when he released what he called "the ESP game" that randomly paired people trying to agree on labeling an image. Not long afterward Amazon created the MTurk platform (http://www.mturk.com) that lets people propose microwork and others sign up to do it, and today there are both hundreds of thousands of tasks offered and hundreds of thousands of people offering to be paid to do them.

[179][Com] For semi-structured or more narrative documents these descriptions might be authoring templates used in word processors or other office applications, document schemas in XML applications, style sheets, or other kinds of transformations that change one resource representation into another one. Primary resources that are highly and regularly structured are invariably organized in databases or enterprise information management systems in which a *data schema* specifies the arrangement and type of data contained in each field or component of the resource.

[180][Law] Describing information as "metadata" suggests that it is of secondary importance, not as essential or informative as the resource being described. This is surely the reason why the US National Security Agency and those of oth-

er governments, whose unauthorized surveillance of global communications were revealed in 2013 by Edward Snowden, often stressed that they were only collecting message metadata, not its content. Of course, information about who you communicate with and when you do so defines your social network, information that is potentially very valuable, and the NSA knows this just as Facebook and Twitter do.

[181][DS] There are a large number of third-party Twitter apps. See http://twitter.pbworks.com/w/page/1779726/Apps. For a scholarly analysis see (Efron 2011).

[182][Bus] The basic idea behind fantasy sports is quite simple. You select a team of existing players in any sport, and then compare their statistical performance against other teams similarly selected by other people. Fantasy sports appeal mostly to die-hard fans who study player statistics carefully before "drafting" their players. The global fantasy sports business for companies who organize and operate fantasy leagues is estimated as between 1 and 2 billion US dollars annually (Montague 2010).

[186][Com] These methods go by different names in different disciplines, including "data modeling," "systems analysis," and "document engineering" (e.g., (Kent 2012), (Silverston 2000), (Glushko and McGrath 2005). What they have in common is that they produce conceptual models of a domain that specify their components or parts and the relationships among these components or parts. These conceptual models are called "schemas" or "domain ontologies" in some modeling approaches, and are typically implemented in models that are optimized for particular technologies or applications.

[187][CogSci] Specifically, an NFL football team needs to be considered a single resource for games through the season and in playoffs, and 53 individual players for other situations, like the NFL draft or play-calling. The team and the team's roster can be thought of as resources, and the team's individual players are also resources that make up the whole team.

[192][Com] This kind of advice can be found in many data or conceptual modeling texts, but this particular statement comes from (Glushko, Weaver, Coonan, and Lincoln 1988). Similar advice can also be found in the information science literature: "A unit of information...would have to be...correctly interpretable outside any context" (Wilson 1968, p. 18).

[193][Com] A group of techniques collectively called "normalization" produces a set of tightly defined *information components* that have minimal redundancy and ambiguity. Imagine that a business keeps information about customer orders using a "spreadsheet" style of organization in which a row contains cells that record the date, order number, customer name, customer address, item ID, item description, quantity, unit price, and total price. If an order contains multiple products, these would be recorded on additional rows, as would subsequent or-

ders from the same customer. All of this information is important to the business, but this way of organizing it has a great deal of redundancy and inefficiency. For example, the customer address recurs in every order, and the customer address field merges street, city, state and zip code into a large unstructured field rather than separating them as atomic components of different types of information with potentially varying uses. Similar redundancy exists for the products and prices. Canceling an order might result in the business deleting all the information it has about a particular customer or product.

Normalization divides this large body of information into four separate tables, one for customers, one for customer orders, one for the items contained in each order, and one for item information. This normalized information model encodes all of the information in the "spreadsheet style" model, but eliminates the redundancy and avoids the data integrity problems that are inherent in it.

Normalization is taught in every database design course. The concept and methods were proposed by (Codd 1970), who invented the relational data model, and has been taught to students in numerous database design textbooks like (Date 2003).

[194][Com] The "Internet of Things" concept spread very quickly after it was proposed in 1999 by Kevin Ashton, who co-founded the Auto-ID center at MIT that year to standardize RFID and sensor information. For a popular introduction, see (Gershenfeld, Krikorian, and Cohen 2004). For a recent technical survey and a taxonomy of application domains and scenarios see (Atzori, Iera, and Morabito 2010).

[195][DS] Pattern analysis can help escape this dilemma by enabling predictive modeling to make optimal use of the data. In designing *smart things* and devices for people, it is helpful to create a smart model in order to predict the kinds of patterns and locations relevant to the data collected or monitored. These allow designers to develop a set of dimensions and principles that will act as smart guides for the development of smart things. Modeling helps to enable automation, security, or energy efficiency, and baseline models can be used to detect anomalies. As for location, exact locations are unnecessary; use of a "symbolic space" to represent each "sensing zone"—e.g., rooms in a house—and an individual's movement history as a string of symbols—e.g., abcdegia—works sufficiently as a model of prediction. See (Das et al. 2002).

[196][Law] Well, maybe not anything. Books list traditional meanings of various names, charts rank names by popularity in different eras, and dozens of websites tout themselves as the place to find a special and unique name. See `http://www.ssa.gov/oact/babynames/` for historical trends about baby names in the US with an interactive visualization at `http://www.babynamewizard.com/voyager#`.

Different countries have rules about characters or words that may be used in names. In Germany, for example, the government regulates the names parents can give to their children; there's even a book, the *International Handbook of Forenames,* to guide them (Kulish 2009). In Portugal, the Ministry of Justice publishes lists of prohibited names (BBC News, 2007a). Meanwhile, in 2007, Swedish tax officials rejected a family's attempt to name their daughter Metallica (http://news.bbc.co.uk/2/hi/6525475.stm).

We can also change our names. Whether a woman takes on her husband's surname after marriage or, like the California man who changed his name to "Trout Fishing," we just find something that better suits us than our given name.

[197][Ling] While you may think that certain terms are more obviously "good" than others, studies show that "there is no one good access term for most objects. The idea of an 'obvious,' 'self-evident,' or 'natural' term is a myth!" (Furnas et al. 1987, p. 967).

[198][IA] (Spencer 2009). Free listing (see http://boxesandarrows.com/beyond-cardsorting-free-listing-methods-to-explore-user-categorizations/)

[199][CogSci] The most common names for this service were activities, calendar and events, but in all over a hundred different names were suggested, including cityevents, whatup, sparetime, funtime, weekender, and nightout, "People use a surprisingly great variety of words to refer to the same thing," Furnas wrote, "If everyone always agreed on what to call things, the user's word would be the designer's word would be the system's word. ... Unfortunately, people often disagree on the words they use for things" (Furnas et al. 1987, p. 964).

[200][CogSci] This example comes from (Farish 2002), who analyzes "What's in a Name?" and suggests that multiple names for the same thing might be a good idea because non-technical business users, data analysts, and system implementers need to see things differently and no one standard for assigning names will work for all three audiences.

[201][Ling] See, for example, *Handbook of Cross-Cultural Marketing*, (Kaynak 1997).

[202][Bus] See "As easy as YZX," http://www.economist.com/node/760345. In addition, the convention to list the co-authors of scientific publications in alphabetic order has been shown to affect reputation and employment by giving undeserved advantages to people whose names start with letters that come early in the alphabet. This bias might also affect admission to selective schools. (Efthyvoulou 2008).

[203][Bus] The Kentucky Fried Chicken franchise solved this problem by changing its name to KFC, which you can now find in Beijing, Moscow, London and other

locations not anywhere near Kentucky and where many people have probably never heard of the place.

Why is the professional basketball team in Los Angeles called the "Lakers" when there are few natural lakes there? The team was originally located in Minneapolis, Minnesota, a state nicknamed "The Land of 10,000 Lakes."

[204][Web] Tim Berners-Lee, the founder of the web, famously argued that "Cool URIs Don't Change" (Berners-Lee 1998).

[205][Law] Any online citation to one of the West printed court reports will use the West format. However, when Mead Data wanted to use the West page numbers in its LEXIS online service to link to specific pages, West sued for copyright infringement. The citation for the West Publishing vs. Mead Data Central case is 799 F.2d 1219 (8th Cir 1986), which means that the case begins on page 1219 of volume 799 in the set of opinions from the 8th Circuit Court of Appeals that West published in print form. West won the case and Mead Data had to pay substantial royalties. Fortunately, this logic behind this decision was repudiated by the US Supreme Court a few years later in a case that West published as *Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991),* and West can no longer claim copyright on page numbers.

[206][Ling] When George Orwell gave the title "*1984*" to a novel he wrote in 1949, he intended it as a warning about a totalitarian future as the Cold War took hold in a divided Europe, but today 1984 is decades in the past and the title does not have the same impact.

[207][Com] (Dorai and Venkatesh 2002).

[208][Bus] (Queenan 2011).

Most common US surnames; http://names.mongabay.com/most_common_surnames.htm.

Chad Ochocinco story: http://en.wikipedia.org/wiki/Chad_Ochocinco.

Fake names at Starbucks: http://online.wsj.com/article/SB10001424053111904106704576582834147448392.html.

Twitter on sports jerseys: http://www.forbes.com/sites/alexknapp/2011/12/30/pro-lacrosse-team-replaces-names-with-twitter-handles-on-jerseys/?partner=technology_newsletter.

[209][Com] Identifiers with meaningful internal structure are said to be structured or intelligent. Those that contain no additional information are sometimes said to be unstructured, opaque, or dumb. The 8 in the ISBN example is a check digit, not technically part of the identifier, that is algorithmically derived from the other digits to detect errors in entering the ISBN.

[210][Ling] (McCartney 2006).

[214][Web] (Hemerly 2011).

[215][Law] This rations / radio confusion is described in (Wheatley 2004). In 2008 a similar mistake in managing inventory at a US military warehouse led to missile launch fuses being sent to Taiwan instead of helicopter batteries, causing a high-level diplomatic furor when the Chinese government objected to this as a treaty violation (Hoffman 2008).

[217][Com] A more general technique is to use the *Universally Unique Identifier (UUID)* standard, which standardizes some algorithms that generate 128-bit tokens that, for all practical purposes, will be unique for hundreds, if not thousands, of years.

[218][Com] (OASIS 2003). The *Organization for the Advancement of Structured Information Systems (OASIS) XML Common Biometric Format (XCBF)* was developed to standardize the use of biometric data like DNA, fingerprints, iris scans, and hand geometry to verify identity (`https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xcbf`).

[220][Com] IP v6 for Internet addresses. The threat of exhaustion was the motivation for remedial technologies, such as classful networks, *Classless Inter-Domain Routing (CIDR)* methods, and *Network Address Translation (NAT)* that extend the usable address space.

[221][Com] Digital Object Identifier (DOI) system (`http://www.doi.org`). However, DOI has its issues too. It is a highly political, publisher-controlled system, not a universal solution to persistence.

[222][Phil] This is called the *Paradox of Theseus,* a philosophical debate since ancient times. Every day that Theseus's ship is in the harbor, a single plank gets replaced, until after a few years the ship is completely rebuilt: not a single original plank remains. Is it still the ship of Theseus? And suppose, meanwhile, the shipbuilders have been building a new ship out of the replaced planks? Is that the ship of Theseus? (Furner 2008, p. 6).

[224][Law] See `http://www.nationsonline.org/oneworld/hist_country_names.htm` for a list of formerly used country names and their respective effectivity.

[225][Com] See (Gravois 2010). One specific example of this effect of international geopolitics on an organizing system involves the northern border of the Crimean Peninsula. When running a query for "Ukraine" via `google.com/maps` (USA), the border appears as a dotted line, which reflects a "neutral" perspective in the aftermath of recent political and military conflicts. Alternatively, when submitting the same query via `google.com.ua/maps` (Ukraine), there is no border at all, which is a reflection of a Ukrainian perspective that the Crimean Peninsula is part of Ukraine. Lastly, when the query is submitted via `google.ru/maps` (Rus-

sia), the border is represented as a solid line, which reflects a Russian perspective that the territory is part of Russia. A 2014 study of Google Maps found 32 situations where the answer to "what country is that on the map?" depended on where it was asked (Yanovsky 2014)

[226][Bus] Effectivity in the tax code is simple compared to that relating to documents in complex systems, like commercial aircraft. Because of their long lifetimes—the Boeing 737 has been flying since the 1960s—and continual upgrading of parts like engines and computers, each airplane has its own operating and maintenance manual that reflects changes made to the plane over time. Every change to the plane requires an update to the repair manual, making the old version obsolete. And while an aircraft mechanic might refer to "the 737 maintenance manual," each 737 aircraft actually has its own unique manual.

[227][Law] A notary public is used to verify that a signature on an important document, such as a mortgage or other contract, is authentic, much as signet rings and sealing wax once proved that no one has tampered with a document since it was sealed.

Chapter 5

Resource Description and Metadata

Robert J. Glushko
Kimra McPherson
Ryan Greenberg
Robyn Perry
Matthew Mayernik
Graham Freeman
Carl Lagoze

5.1 Introduction

This chapter is a turning point in the book. The earlier chapters have discussed the key ideas of the discipline of organizing: identifying and selecting the resources to organize, and then organizing and maintaining them and their organizing system. We have emphasized that finding things later is the most important reason for organizing them. This can be surprisingly hard to do. People know things by different names or remember different aspects of them.

"Arrangement in Grey and Black No. 1"

"Arrangement in Grey and Black No. 1" (1871). Oil on canvas, by James Abbott McNeill Whistler. Alternative titles: "Portrait of the Artist's Mother" and "Whistler's Mother." The painting is in Musée d'Orsay in Paris.

(Photo by Jean-Gilles Berizzi Source: Wikimedia Commons.)

The famous painting here by the 19th century American painter James Whistler is exhibited in the Musée d'Orsay in Paris, and has been described as a Victorian-era *Mona Lisa.* What name do you know it by? How should it be described?

Resource descriptions for art usually contain the name of the artist, the medium, the year of its completion, and, of course, its title. Most of these map fairly obviously to the properties they describe; the title, owing to its prominence and expressive power, is often an exception.

Most often, a painting's title describes its subject. If you recognize the previous painting, you most likely know it by its colloquial name, *Whistler's Mother.* While it is a portrait of Anna McNeill Whistler, mother of painter James Abbott McNeill Whistler, the artist gave it a radically different title, *Arrangement in Grey and Black, No. 1,* because he believed the most important property of a painting was not the subject it depicted, but its purely aesthetic properties and their effect on the viewer. So Whistler named his paintings, which were mostly landscapes and portraits, in the manner of musical compositions: *Nocturne in Black and Gold*; *Symphony in White*; *Arrangement in Pink, Red, and Purple*; and so on.

If Whistler's title surprises you, because you would have described it as a portrait of an elderly woman, this helps reinforce how wildly different names of the same resource can be. Resource descriptions and metadata provide meaning, but to whom? What is salient about a resource can depend on the context in which it is experienced, and thus may change over time. Descriptions that make sense to some people might not make sense to others. People searching on the "wrong descriptions" or the "wrong metadata" will not find what they are looking for.

Mt. St. Helens Before and After

Before 1980, Mt. St. Helens was a "postcard-like" snow-covered mountain. Afterward, the mile-wide crater where its mountaintop once was reminds us of its violent volcanic eruption.

(Credit: Public domain images from US Forest Service and USGS.)

Mt. St. Helens, in the southwest corner of Washington State, was usually just described as a mountain until 1980. Then, the deadliest and most economically destructive volcanic event in the history of the United States blew away the top of the mountain, killing 57 people, and leaving a mile-wide crater. Today almost every description of Mt. St. Helens mentions the volcanic eruption.

It would seem impossible to search using the wrong description if the descriptions of a resource were kept current to include all the latest information, but search engines are already too powerful, usually producing too much information. Technology improvements in search and retrieval do not eliminate the cognitive effort to remember what things are, how they are best described, and where they might be found. The design of resource descriptions and metadata depends on why we need to find the information later. This chapter is about how and why.

Stop and Think: These Places Have Their Moments

Our description of Mt. St. Helens forever changed after its volcanic eruption. Surely there are times and places that you remember differently because of their part in an important event. A family wedding? The Olympic Games? A natural disaster? The Twin Towers?

It is easy to find before and after images of Mt. St. Helens doing a web search. What information might be associated with these images? Modern cameras assign an identifier to the stored photograph and they also capture the technical description of the image's production: the type of camera, lens, shutter speed, light sensitivity, aperture, and other settings.[228][Com] Many modern cameras also record information about the geographic and temporal circumstances surrounding the image's creation: the date, time and location on Earth where the photograph is taken. When the image is transferred out of the camera and is published for all to see, it might be useful to record biographical information about the photographer to help viewers relate to the photographer and better understand the photograph's context. There may also be different licenses and copyright information to associate with the picture—who owns it and how it can be used.

Consider a completely different context. Four 7-year old boys are selecting Lego blocks to complete their latest construction. The first boy is looking for "cylinder one-ers," another for "coke bottles," the third for "golder wipers," and the final boy is looking for "round one-bricks"? It turns out, they are all the same thing; each boy has devised his own set of descriptive terms for the tiny building blocks. Some of their many descriptions are based on color alone ("redder"), some on color and shape ("blue tunnel"), some on role ("connector"), some on common cultural touchstones ("light saber"). Others, like "jail snail" and "slug," seem unidentifiable—unless, of course, you happen to be inside the mind of a particular 7-year-old kid. It doesn't matter if the boys use different description vocabularies when they play by themselves, but they will have to agree if they play together.[229][CogSci]

Paintings, digital photos, and Lego blocks are all very different, but together these scenarios raise important questions about describing resources that we attempt to answer in this chapter:

- What is the purpose of resource description?
- What resource properties should be described?
- How are resource descriptions created?
- What makes a good resource description?

Navigating This Chapter

We begin with an overview of *resource description* (§5.2), which we propose as a broad concept that includes the narrower concepts of *bibliographic descriptions* and *metadata*. §5.3 The Process of Describing Resources (page 227) describes a 7-step process of describing resources that includes determining scope, focus and purposes, identifying resource properties, designing the description vocabulary, designing the description form and implementation, and creating and evaluating the descriptions. Because many principles and methods for resource description were developed for describing text resources in physical formats, in §5.4 Describing Non-text Resources (page 257) we briefly discuss the issues that arise when describing museum and artistic resources, images, music, video, and contextual resources.

5.2 An Overview of Resource Description

We describe resources so that we can refer to them, distinguish among them, search for them, manage access to them, preserve them, and make predictions about what might happen to them or what they might do. Each purpose may require different *resource descriptions*. We use *resource descriptions* in every communication and conversation; they are the enablers of organizing systems.

5.2.1 Naming {and, or, vs.} Describing

Chapter 4 discussed how to decide what things should be treated as resources and how names and identifiers distinguish one resource from another. Names can suggest the properties and principles an organizing system uses to arrange its resources. We can see how societies organize their people by noting that among the most common surnames in English are descriptions of occupations (Smith, Miller, Taylor), descriptions of kinship relations (Johnson, Wilson, Anderson), and descriptions of appearance (Brown, White).[230][Ling]

In many cultures, one spouse or the other takes a name that describes their marital relationship. In many parts of the English-speaking world, married women have often referred to themselves using their husband's name.[231][Ling]

Similarly, many other kinds of resources have names that are property descriptions, including buildings (Pentagon, White House), geographical locations (North America, Red Sea), and cities (Grand Forks, Baton Rouge).

Every resource can be given a name or identifier. Identifiers are especially efficient *resource descriptions* because, by definition, identifiers are unique over some domain or collection of resources. Names and identifiers do not typically describe the resource in any ordinary sense because they are usually assigned to the resource rather than recording a property of it.

However, the arbitrariness of names and identifiers means that they do not serve to distinguish resources for people who do not already know them. This is why we use what linguists call referring expressions or definite descriptions, like "the small black dog" rather than the more efficient "Blackie," when we are talking to someone who does not know that is the dog's name.[232][CogSci]

Similarly, when we use a library catalog or search engine to locate a known resource, we query for it using its name, or some specific information we know about it, to make it easier to find. In contrast, when we look for resources to satisfy an information need but do not have specific resources in mind, we query for them using descriptions of their content or other properties. In general, information retrieval can be characterized as comparing the description of a user's needs with descriptions of the resources that might satisfy them.

5.2.2 "Description" as an Inclusive Term

Up to now we have used the concept of "description" in its ordinary sense to mean the labeling or explaining of the visible or important features that characterize or represent something. However, the concept is sometimes used more precisely in the context of organizing systems, where *resource description* is often more formal, systematic, and institutional. In the library science context of *bibliographic description,* a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources. In the contexts of conceptual modeling and information systems design, the terms in *resource descriptions* are also called "keywords," "index terms," *attributes*, *attribute values*, *elements*, "data elements," "data values," or "the vocabulary." In business intelligence, predictive analytics or other data science contexts these are called "variables," "features," *properties*, or "measurements." In contexts where descriptions are less formal or more personal the description terms are often called "labels" or "tags." Rather than attempt to make fine distinctions among these synonyms or near-synonyms, we will use "description" as an inclusive term except where conventional usage overwhelmingly favors one of the other terms.

Many of these terms come from a narrow semantic scope in which the purpose of description is to identify and characterize the essence, or *aboutness*, of a resource. However, as it becomes trivial to associate computationally generated information with resources, many additional kinds of information beyond strict "aboutness" can support additional interactions. We describe many of these purposes and the types of information needed to enable them in §5.3.2 Determining the Purposes (page 234). We apply *resource description* in an expansive way to accommodate all of them.

Chapter 4 introduced the distinction of §4.2.4 Resource Focus (page 178) to contrast primary resources with resources that describe them, which we called Description Resources. We chose this term as a more inclusive and more easily understood alternative to two terms that are well established in organizing systems for information resources: *bibliographic descriptions* and *metadata*. We will also distinguish resource description as a general concept from the narrower senses of statistical description, tagging of web resources, and the *Resource Description Framework (RDF)* language used to make statements about web resources and physical resources that can be identified on the Web.

5.2.2.1 Bibliographic Descriptions

The purposes and nature of bibliographic description are the foundation of library and information science and have been debated and systematized for nearly two centuries. *Bibliographic descriptions* characterize information re-

sources and the entities that populate the bibliographic universe, which include works, editions, authors, and subjects.

A bibliographic description of an information resource is typically realized as a structured record in a standardized format that describes a specific resource.

The computerization of bibliographic records made them easier to use as aids for finding resources. However, digitizing legacy printed card-oriented descriptions for online use was not a straightforward task because the descriptions had been created according to cataloging rules designed for collections of books and other physical resources and intended only for use by people.

5.2.2.2 Metadata

Metadata is often defined as "data about data," a definition that is nearly as ubiquitous as it is unhelpful. A more content-full definition of metadata is that it is structured description for information resources of any kind. Metadata is more useful when supported by a metadata schema that defines the elements in the structured description.

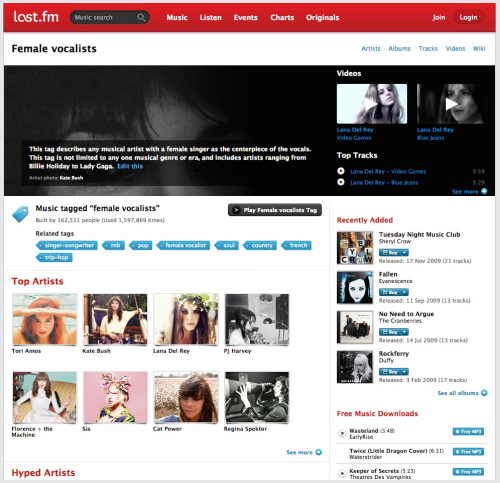
The concept of metadata originated in information systems and database design in the 1970s, so it is much newer than that of bibliographic description. The earliest metadata schemas, called data dictionaries, documented the arrangement and content of data fields in the records used by transactional applications on mainframe computers. A more sophisticated type of metadata emerged as the documentation of the data models in database management systems, called database schemas, which described the structure of relational tables, attribute names, and legal data types and values for content.

In 1986, the *Standard Generalized Markup Language (SGML)* formalized the *Document Type Definition (DTD)* as a metadata form for describing the structure and content elements in hierarchical and hypertextual document models. SGML was superseded in 1997 by eXtensible Markup Language (XML), whose purpose was structured and computer-processable web content.[235[Com]]

Today, XML schemas and other web- and compute-friendly formats for resource description have broadened the idea of *resource description* far beyond that of bibliographic description to include the description of software components, business and scientific datasets, web services, and computational objects in both physical and digital formats. The resource descriptions themselves serve to enable discovery, reuse, access control, and the invocation of other resources needed for people or computational agents to effectively interact with the primary ones described by the metadata.[236[Com]]

5.2.2.3 Tagging of Web-based Resources

Tags on Last.fm



Last.fm analyzes tags and other metadata to create rich multimedia "discovery" pages that bring together artist catalogs, new songs, free downloads, and music videos that its algorithm predicts will satisfy a user's taste. This allows users to browse for new music in a more intuitive manner than searching by artist or genre.

(Screenshot by Ian MacFarland.)

The concept of metadata has been extended to include the tags, ratings, bookmarks or other types of descriptions that individuals apply to individual photos, blog or news items, or any other resource with a web presence. The practice of tagging has emerged as a way to apply labels to content in order to describe and identify it. Sets of tags are useful in managing one's collection of websites or digital media, in sharing them with others, and enabling new types of interactions and services.[237][IA] For example, users of Last.fm tag music with labels that describe its nature, era, mood, or genre, and Last.fm uses these tags to generate radio stations that play music similar to that tag and related tags.

But tagging has a downside. The tendency for users to tag intuitively and spontaneously revives the vocabulary problem (§4.4) because one photographer's "tree" is another's "oak." Likewise, unsystematic word choice leads to morphological inconsistency (§6.4.3 Relationships among Word Forms (page 293)); the same photo might be tagged with "burning" and "trees" and also with "burn" and "tree" by another. This disparity in the descriptors people use to categorize the same or similar resources can turn systems that use tagging into a "tag soup" lacking in structure.[238][IA]

Some social media sites have incorporated mechanisms to make the tagging activity more systematic and to reduce *vocabulary problems*. For example, on Facebook, users can indicate that a specific person is in an uploaded picture by clicking on the faces of people in photographs, typing the person's name, and then selecting the person from a list of Facebook friends whose names are formatted the way they appear on the friend's profile. Some social media systems suggest the most popular tags, perform morphological normalization, or allow users to arrange tags in bundles or hierarchies.[239][Web]

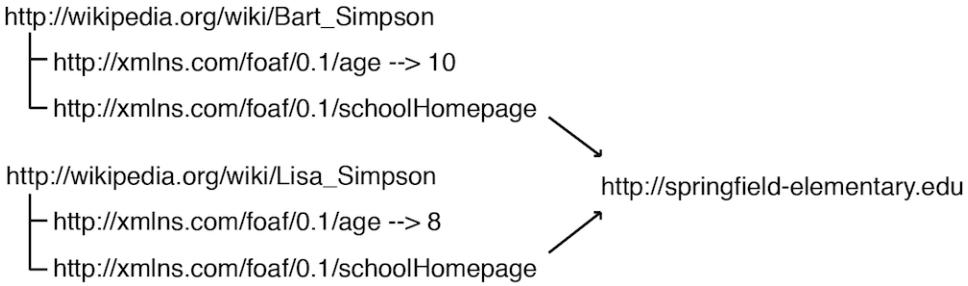
5.2.2.4 Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a standard model for making computer-processable statements about web resources; it is the foundation for the vision of the *Semantic Web*.[240][Web] We have been using the word "resource" to refer to anything that is being organized. In the context of RDF and the web, however, "*resource*" means something more specific: a resource is anything that has been given a Uniform Resource Identifier (URI). URIs can take various forms, but you are probably most familiar with the URIs used to identify web pages, such as `http://springfield-elementary.edu/`. (You are probably also used to calling these URLs instead of URIs.) The key idea behind RDF is that we can use URIs to identify not only things "on" the web, like web pages, but also things "off" the web like people or countries. For example, we might use the URI `http://springfield-elementary.edu/` to refer to Springfield Elementary itself, and not just the school's web page.

RDF models all descriptions as sets of "triples," where each *triple* consists of the resource being described (identified by a URI), a property, and a value. Properties are resources too, meaning they are identified by URIs. For example, the URI `http://xmlns.com/foaf/0.1/schoolHomepage` identifies a property defined by the *Friend of a Friend (FOAF)* project for relating a person to (the web page of) a school they attended. Values can be resources too, but they do not have to be: when a property takes simple values like numbers, dates, or text strings, these values do not have URIs and so are not resources.

Because RDF uses URIs to identify described resources, their properties, and (some) property values, the triples in a description can be connected into a network or *graph.* Figure 5.1, RDF Triples Arranged as a Graph. shows four triples that have been connected into a graph. Two of the triples describe Bart Simpson, who is identified using the URI of his Wikipedia page.[241][Web] The other two describe Lisa Simpson. Two of the triples use the property `age`, which takes a simple number value. The other two use the property `schoolHomepage`, which takes a resource value, and in this case they happen to have the same resource (Springfield Elementary's home page) as their value.

Figure 5.1. RDF Triples Arranged as a Graph.



Two RDF triples can be connected to form a graph when they have a resource, property, or value in common. In this example RDF triples that make a statement about the home page of the elementary school attended by Bart Simpson and Lisa Simpson can be connected because they have the same value, namely the URI for Springfield Elementary.

Using URIs as identifiers for resources and properties allows descriptions modeled as RDF to be interconnected into a network of "linked data," in the same way that the web enabled information to be interconnected into a massive network of "linked documents." Proponents of RDF claim that this will greatly benefit knowledge discovery and inference.[242][Web] But the benefits of RDF's highly prescriptive description form must be weighed against the costs; turning existing descriptions into RDF can be labor-intensive.

5.2.2.5 Aggregated Information Objects

In the pre-digital age, information objects came with explicit tangible boundaries. Books consisted of pages bound within a cover, a vinyl record album physically bound together a set of songs (you could even see the groove pattern separating the songs), a movie was delivered on a strip of film spooled onto a reel, and a collection was (usually) demarcated as a designated shelf or room in a library.

Boundaries of information objects in the digital realm are neither tangible nor obvious. Consider the simple notion of a web page. Our cognitive notion of that which is rendered in our browser window (e.g., some formatted text with an associated image) is actually, in web architecture terms (Jacobs & Walsh, 2004), three information objects (aka resources); the HTML encoding the text, CSS that defines the formatting rules, and the JPEG that encodes the image. All three have URLs and can independently be retrieved and linked. The situation is even more ambiguous for the common notion of a web site, the boundaries of which are not defined technically and are cognitively difficult to express.

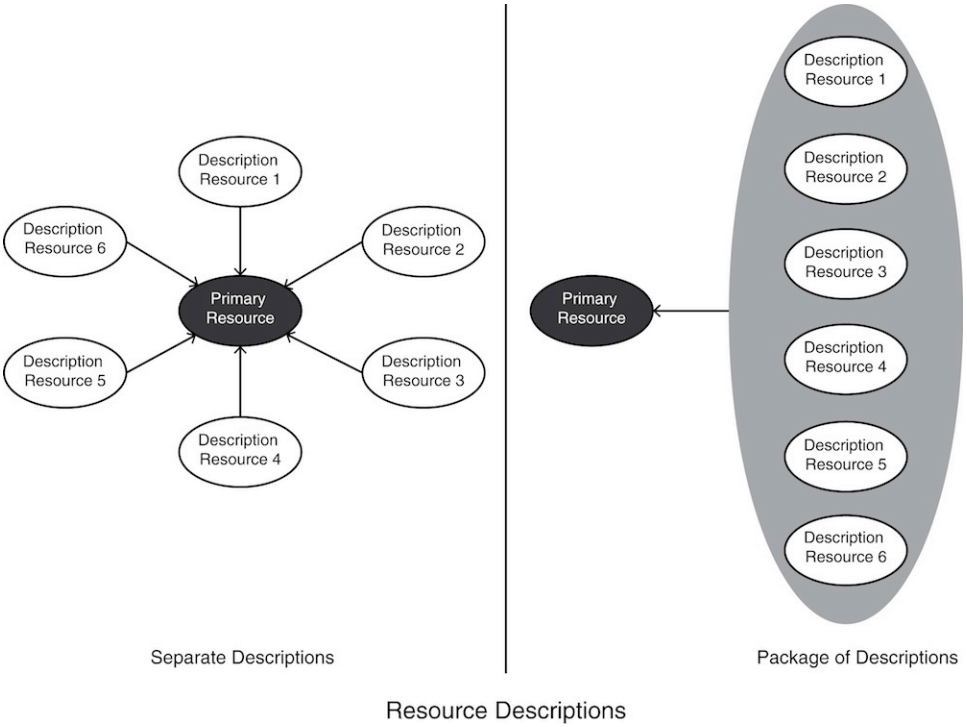
Aggregations can be convenience methods for simplifying dissemination or organization, but they can also be transformative; resources can derive nearly all their value from their inclusion in an aggregation. On a web page, the CSS file is virtually useless on its own, since its role is to style the HTML file. In iTunes, the playback and organization functions are optimized for pop music, where individual songs can usually stand on their own when separated from the rest of an album. Classical music fans often struggle with this, because the individual "tracks" of a recording, split up to reduce file size and facilitate navigation through long works, are not separable; pieces are meant to be listened to in their entirety, and it can be difficult to ensure that they are aggregated together and have the proper metadata assigned to their aggregations. In other words: you can't listen to symphonies on shuffle.[244][IA]

The problem here is how to architecturally and technically express the notion of an aggregation, a set of information objects that, when considered together, compose another named information object. Aggregations are prevalent all over our digital information space: the web page and site mentioned above; a scholarly publication consisting of text, figures, and data; a dataset that is the composition of multiple data files. Notably the notion is both recursive and non-exclusive. An object that is itself an aggregation may be aggregated into another object. Information objects included in one aggregation may also be included in other aggregations, allowing reuse and re-factoring of existing information objects. A solution to this problem is a critical aspect of organizing digital information because, without well-defined boundaries we cannot deterministically identify, reference, or describe information objects.

5.2.3 Frameworks for Resource Description

The broad scope of resources to which descriptions can be applied and the different communities that describe them means that many *frameworks* and classifications have been proposed to help make sense of resource description.

Figure 5.2. Architectures for Resource Description.



Two contrasting architectures for resource descriptions are separate descriptions versus packaged descriptions, which were dominant in library catalogs with printed cards containing descriptions about a resource.

The dominant historical view treats resource descriptions as a package of statements; this view is embodied in the printed library card catalog and its computerized analog in the MARC21 format (an exchange format for library catalog records), which contains many fields about the bibliographic characteristics of an object like author, title, publication year, publisher, and pagination. An alternate architecture for resource description focuses on each individual description or assertion about a single resource, as the RDF and linked data approaches do. These two alternatives are contrasted in Figure 5.2, Architectures for Resource Description.

In either case, these common ways of thinking about resource description emphasize—or perhaps even overemphasize—two implementation decisions:

- The first is whether to combine multiple resource descriptions into a structural package or to keep them as separate descriptive statements.
- The second is the choice of syntax in which the descriptions are encoded.

Both of these implementation decisions have important implications, but are secondary to the questions about the purposes of resource description, how resource properties are selected as the basis for description, how they are best created, and other logical or design considerations. In keeping with a fundamental idea of the discipline of organizing (introduced in §1.6 The Concept of "Organizing Principle" (page 43)), it is imperative to distinguish design principles from implementation choices. We treat the set of implementation decisions about character notations, syntax, and structure as the *form* of resource description and we will defer them as much as we can until Chapter 9, *The Forms of Resource Descriptions*.

Resource description is not an end in itself. Its many purposes are all means for enabling and using an organizing system for some collection of resources. As a result, our framework for resource descriptions aligns with the activities of organizing systems we discussed in Chapter 3: selecting, organizing, interacting with, and maintaining resources.

5.3 The Process of Describing Resources

We prefer the general concept of resource description over the more specialized ones of bibliographic description and metadata because it makes it easier to see the issues that cut across the domains where those terms dominate. In addition, it enables us to propose more standard process that we can apply broadly to the use of resource descriptions in organizing systems. A shared vocabulary enables the sharing of lessons and best practices.

The process of describing resources involves seven interdependent and iterative steps. We begin with a generic summary of the process to set the stage for a detailed step-by-step discussion.

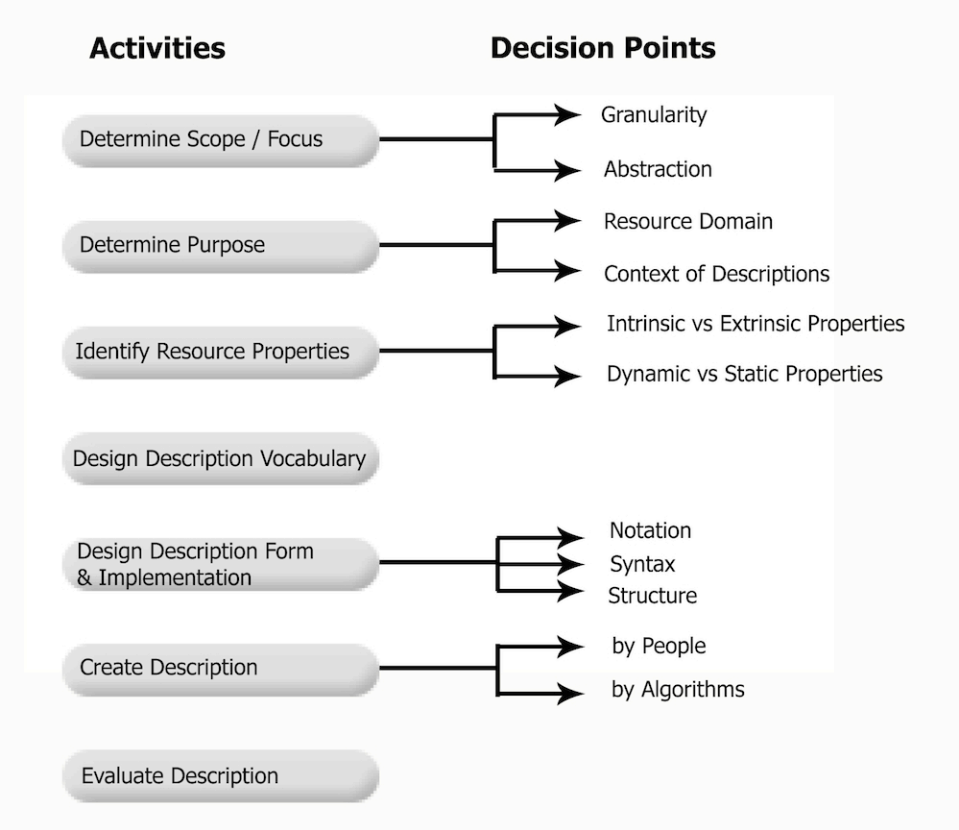
1. Identifying resources to describe is the first step; this topic is covered in detail in §4.3 Resource Identity (page 182). The resource *domain* and *scope* circumscribe the describable properties and the possible purposes that descriptions might serve. The resource *focus* determines which are primary information resources and which ones are treated as the corresponding resource descriptions. Two important decisions at this stage are *granularity* of description—are we describing individual resources or collections of them?

—and the *abstraction* level—are we describing resource instances, parts of them, or resource types?

2. Generally, the purpose of resource description is to support the activities common to all organizing systems: selecting, organizing, interacting with, and maintaining resources, as we saw in Chapter 3. The particular resource domain and the context in which descriptions are created and used imposes more specific requirements and constraints on the purposes that resource description can serve.
3. Once the purposes of description in terms of activities and interactions have been determined, the specific properties of the resources that are needed to enable them can be identified. The goal of description is not to be exhaustive; there are always more possible properties than can be reasonably described. Instead, the challenge is to use the properties that are most robust and reliable for supporting the desired interactions.
4. This step includes several logical and semantic decisions about how the resource properties will be described. What terms or element names should be used to identify the resource properties we have chosen to describe? Are there rules or constraints on the types of data or values that the property descriptions can assume? When dealing with numerical descriptions, their data types and levels of measurement constrain the kinds of processing to which they may submit. Nominal, ordinal, interval, and ratio data each are limited to particular transformations based on what they represent. A good description vocabulary will be easy to assign when creating resource descriptions and easy to understand when using them.
5. The logical and semantic decisions about the description vocabulary are reified by decisions about the notation, syntax and structure of the descriptions. Taken together, these decisions collectively determine what we call the *form* or *encoding* of the resource descriptions. The implementation of the descriptions involves decisions about how and where they are stored and the technology used to create, edit, store, and retrieve them.
6. Resource descriptions are created by individuals, by informal or formal groups of people, or by automated or computational means. Some types of descriptions can only be created by people, some types of descriptions can only be created by automated or algorithmic techniques, and some can be created in either manner.
7. The resource descriptions must be evaluated with respect to their intended purposes. The results of this evaluation will help determine which or the preceding steps need to be redone.

The next seven sub-sections discuss each of these steps in detail. A quick reference guide is Figure 5.3, The Process of Describing Resources.

Figure 5.3. The Process of Describing Resources.



The process of describing resources consists of seven steps: Determining the scope and focus, determining the purpose, identifying resource properties, designing the description vocabulary, designing the description form and implementation, creating the descriptions, and evaluating the descriptions.

How explicit and systematic each step needs to be depends on the resource domain and scope, and especially on the intended users of the organizing system. If we look carefully, we can see most of these steps taking place even in very informal contexts, like the kids playing with Lego blocks with which we started this chapter. The goal of building things with the blocks leads the boys to identify which properties are most useful to analyze. They develop descriptions of the blocks that capture the specific values of the relevant properties. Finally, they evaluate their descriptions by using them when they play together; it becomes

immediately obvious that a description is not serving its purpose when one boy hands a block to another that was not the one he thought he had asked for.

In contrast, a picture-taking scenario involves a much more explicit and systematic process of resource description. The resource properties, description vocabulary, and description form used automatically by a digital camera were chosen by an industry association and published as a technical specification implemented by camera and mobile phone manufacturers worldwide.

The resource descriptions used by libraries, archives, and museums are typically created in an even more explicit and systematic manner. Like the descriptions of the digital photo, the properties, vocabulary, and form of the descriptions used by their organizing systems are governed by standards. However, there is no equivalent to the digital camera that can create these descriptions automatically. Instead, highly trained professionals create them meticulously.

A great many resources and their associated descriptions in business and scientific organizing systems are created by automated or computational processes, so the process of describing individual resources is not at all like that in libraries and other memory institutions. However, the process for designing the data models or schemas for the class of resources that will be generated is equally systematic and is typically performed by highly skilled data analysts and data modelers.

5.3.1 Determining the Scope and Focus

Which resources do we want to describe? As we saw in Chapter 4, determining what will be treated as a separate resource is not always easy, especially for resources with component parts and for information resources where the most important property is their content, which is not directly perceivable. Identifying the thing you want to describe as precisely as practical is the first step to creating a useful description.

In §4.2.4 Resource Focus (page 178), we introduced the contrast between primary resources and description resources, which we called resource focus. Determining the resource focus goes hand in hand with determining which resources we intend to describe; these often arbitrary decisions then make a huge difference in the nature and extent of resource description. One person's metadata is another person's data.

- For a librarian, the price of a book might be just one more attribute that is part of the book's record.
- For an accountant at a bookstore, the price of that book—both the cost to buy the book and the price at which it is then sold to customers—is critical information for staying in business.

- In a medical records context, a patient's insurance provider isn't of much concern to the doctor, but to the person responsible for billing, it is central. For the nurse, the patient's current vital signs may be of most importance, while for the doctor, it may be most important to understand how those data in aggregate serve to indicate a longer-term prognosis of the patient's health.
- A scientist studying comparative anatomy preserves animal specimens and records detailed physical descriptions about them, but a scientist studying ecology or migration discards the specimens and focuses on describing the context in which the specimen was located.

5.3.1.1 Describing Instances or Describing Collections

It is simplest to think of a resource description as being associated with another individual resource. As we discussed in Chapter 4, it is challenging to determine what to treat as an individual resource when resources are themselves objects or systems composed of other parts or resources. For example, we sometimes describe a football team as a single resource and at other times we focus on each individual player. However, after deciding on resource granularity, the question remains whether each resource needs a separate description.

Libraries and museums specialize in curating resource descriptions about the instances in their collections. Resource descriptions are also applied to classes or collections of resources, because a collection is also a resource (§1.4 The Concept of "Collection" (page 37)). Archives and special collections of maps are typically assigned resource descriptions, but each document or map contained in the collection does not necessarily have its own bibliographic description. Similarly, business and scientific datasets are invariably described at collection-level granularity because they are often analyzed in their entirety.

Furthermore, the granularity of description for a collection of resources tends to differ for different users or purposes. An investor who owns many different stocks focuses on their individual prices, while other investors put their money in index funds that combine all the separate prices into a single value.

Many web pages, especially e-commerce product catalogs and news sites, are dynamically assembled and personalized from a large number of information resources and services that are separately identified and described in content management and content delivery systems. However, a highly complex collection of resources that comes together in a single page is treated as a single resource when that page appears in a list of search engine results. Moreover, all of the separately generated pages can be given a single description when a user creates a bookmark to make it easy to return to the home page of the site.

5.3.1.2 Abstraction in Resource Description

We can also associate resource descriptions with an entire type or domain of resources. (See §3.5.2.4 Preserving Resource Types (page 138) and §4.2.1 Resource Domain (page 167).) A collection of resource descriptions is vastly more useful when every resource is described using common description elements or terms that apply to every resource. A *schema* (or model, or metadata standard) specifies the set of descriptions that apply to an entire resource type. Sometimes this schema, model, or standard is inferred from or imposed on a collection of existing resources to ensure more consistent definitions, but more often, it is used as a specification when the resources are created or generated in the first place. (See What about "Creating" Resources? (page 90) in §3.1 Introduction (page 87).)

A relational database, for example, is easily conceptualized as a collection of records organized as one or more tables, with each record in its own row having a number of fields or attributes that contain some prescribed type of content. Each record or row in the database table is a description of a resource—an employee, a product, anything—and the individual attribute values, organized by the columns and rows of the table, are distinct parts of the description for some particular resource instance, like employee 24 or product 8012C.[251][Com]

The information resources that we commonly call documents are, by their nature, less homogeneous in content and structure than those that can be managed in databases. Document schemas, commonly represented in SGML or XML, usually allow for a mixture of data-like and textual descriptive elements.

XML schema languages have improved on SGML and XML by expressing the description of the document schema in XML itself, making it easy to create resources using the metadata as a template or pattern. XML schemas are often used as the specifications for XML resources created and used by information-intensive applications; in this context, they are often called XML vocabularies. XML schemas can be used to define web forms that capture resource instances (each filled-out form). XML schemas are also used to describe the interfaces to web services and other computational resources.[252][Com]

5.3.1.3 Scope, Scale, and Resource Description

If we only had one thing to describe, we could use a single word to describe it: "it." We would not need to distinguish it from anything else. A second resource implies at least one more term in the description language: "not it." However, as a collection grows, descriptions must become more complex to distinguish not only between, but also among resources.

Every element or term in a description language creates a dimension, or axis, along which resources can be distinguished, or it defines a set of questions

about resources. Distinctions and questions that arise frequently need to be easy to address, such as:

- What is the name of the resource?
- Who created it?
- What type of content or matter does it contain?

Therefore, as a collection grows, the language for describing resources must become more rigorous, and descriptions created when the collection was small often require revision because they are no longer adequate for their intended purposes.

Because the task of library resource description has been standardized at national and international levels, cataloging work is distributed among many describers whose results are shared. The principle of standardization has been the basis of centralized bibliographic description for a century.

Centralized resource description by skilled professionals works for libraries, but even in the earliest days of the web many library scientists and web authoring futurists recognized that this approach would not scale for describing web resources. In 1995, the *Dublin Core (DC)* metadata element set with only 15 elements was proposed as a vastly simpler description vocabulary that people not trained as professional catalogers could use. Since then, the Dublin Core initiative has been highly influential in inspiring numerous other communities to create minimalist description vocabularies, often by simplifying vocabularies that had been devised by professionals for use by non-professionals.

Of course, a simpler description vocabulary makes fewer distinctions than a complex one; replacing "author," "artist," "composer" and many other descriptions of the person or non-human resource responsible for the intellectual content of a resource with just "creator" (as Dublin Core does) results in a substantial loss of precision when the description is created and can cause misunderstanding when the descriptions are reused.[256][CogSci]

The negative impacts of growing scope and scale on resource description can sometimes be avoided if the ultimate scope and scale of the organizing system is contemplated when it is being created. It would not be smart for a business with customers in six US states to create an address field in its customer database that only handled those six states; a more extensible design would allow for any state or province and include a country code. In general, however, just as there are problems in adapting a simple vocabulary as scope and scale increase, designing and applying resource descriptions that will work for a large and continuously growing collection might seem like too much work when the collection at hand is small.

The challenges that arise with large description vocabularies are transformed when resource descriptions are created and assigned by computer algorithms. A large dataset might contain many thousands of descriptions for each resource, but clearly the computer does not have cognitive difficulty generating or using them. However, computer models with this many features can be hard for people to understand and trust.

5.3.2 Determining the Purposes

Resource description serves many purposes, and the mix of purposes and the resulting kinds of descriptions in any particular organizing system depends on the scope and scale of the resources being organized. We can identify and classify the most common purposes using the four activities that occur in every organizing system: selecting, organizing, interacting with, and maintaining resources (see Chapter 3). Resource description also has a more open-ended purpose in *sensemaking* and science (see §5.3.2.5); we observe and describe the world to make sense of our experiences and to predict future observations.

5.3.2.1 Resource Description to Support Selection

Defining selection as the process by which resources are identified, evaluated, and then added to a collection in an organizing system emphasizes resource descriptions created by someone other than the person who is using them. We can distinguish several different ways in which resource description supports selection:

Discovery
: What available resources might be added to a collection? New resources are often listed in directories, registries, or catalogs. Some types of resources are selected and acquired automatically through subscriptions or contracts.

Capability and Compatibility
: Will the resource meet functional or interoperability requirements? Technology-intensive resources often have numerous specialized types of descriptions that specify their functions, performance, reliability, and other "-ilities" that determine if they fit in with other resources in an organizing system. 257[Com] Some services have qualities of service levels, terms and conditions, or interfaces documented in resource descriptions that affect their compatibility and interoperability. Some resources have licensing or usage restrictions that might prevent the resources from being used effectively for the intended purposes. Decisions about "people selection" are becoming more data-driven, and sports teams, business employers,and dating sites now rely on predictive statistics to find the best person.

Authentication
: Is the resource what it claims to be? (§4.5.3 Authenticity (page 202)) Resource descriptions that can support authentication include technological ones like time stamps, watermarking, encryption, checksums, and digital signatures. The history of ownership or custody of a resource, called its provenance (§4.5.4 Provenance (page 203)), is often established through association with sales or tax records. Import and export certificates associated with the resource might be required to comply with laws designed to prevent the theft of antiquities or the transfer of technology or information with national security or foreign policy implications.

Appraisal
: What is the value of this resource? What is its cost? At what rate does it depreciate? Does it have a shelf life? Does it have any associated ratings, rankings, or quality measures? Moreover, what is the quality of those ratings, rankings and measures?

We also consider the perspective of the person creating the resource description and his or her primary purpose, which is often to encourage the selection of the resource by someone else. Product marketing is about devising names and descriptions to make a resource distinctive and attractive compared to alternatives. For many years prunes were promoted as a dietary supplement that people (especially old ones) need to "maintain regularity." But after the California Prune Board (the world's biggest supplier) re-branded them as "dried plums" and started marketing them as a snack food (and simultaneously renaming itself as the California Dried Plum Board) sales increased significantly.[258][Bus]

Many countries require that imported goods are labeled with their country or origin. Consumers often use this property in resource descriptions as an indicator of high quality, as they might with Swiss watches, French or Italian fashions, or Canadian bacon. Alternatively, consumers might want to buy domestic or locally-sourced goods out of economic patriotism or to comply with procurement regulations. Not surprisingly, when consumers view origin in a positive light, this information is conspicuous and easy to read. In contrast, when consumers view origin less positively, perhaps as a warning of low quality goods, the supplier is likely to make the origin information as inconspicuous as legally possible, or might even misrepresent the goods as domestic ones.[259][Bus]

This misrepresentation is also ubiquitous in online dating, though the amount of misrepresentation must be balanced with goals of the relationship and chances of the deception being discovered.[260][CogSci]

5.3.2.2 Resource Description to Support Organizing

We have defined *organizing* as specifying the principles for describing and arranging resources to create the capabilities upon which interactions are based. This definition treats the creation of resource descriptions and their use to organize resources for interactions as separate and sequential activities. This is easiest to see when people assign keywords and classifications to documents, or when sensors produce data, and these resource descriptions are later used to enable document retrieval or data analysis. A department store clerk might sort dress shirts on a display table using labels that describe their brands, sizes, and other properties. Rules governing the collection, integration, and analysis of personal information are also resource descriptions that influence the organization of information resources.

However, even if resource description and resource organization are logically separable, at times they are intertwined. When you arrange your own clothes, you don't use explicit resource descriptions and instead rely on implicit ones about easily perceived properties like color, shape, and material of composition. When algorithms rather than people analyze texts to identify descriptive features for applications like information retrieval, spam classification, and sentiment analysis, resource descriptions and resource organization co-evolve, often continuously as the algorithm adapts and learns with each new resource it describes. This tight connection between resource description and resource organization is also exploited in organizing systems that use usage records from session logs, browsing, or downloading activities as interaction resources, tying them to payments for using the resources or analyzing them to influence the selection and organizing of resources in future personalized interactions. (See §1.9 The Concept of "Interaction Resource" (page 51))

5.3.2.3 Resource Description to Support Interactions

Most discussions of the purposes of resource descriptions and metadata emphasize the interactions that are based on resource descriptions that have been intentionally and explicitly assigned. The Functional Requirements for Bibliographic Records (FRBR), defined by library scientists, specifies the four interactions of Finding, Identifying, Selecting, and Obtaining resources, but these apply generically to organizing systems, not just those in libraries.

Finding

> What resources are available that "correspond to the user's stated search criteria" and thus can satisfy an information need? Modern users accept that computerized indexing makes search possible over not only the entire description resource, but often over the entire content of the primary resource. Businesses search directories for descriptions of company capabilities to

find potential partners, and they also search for descriptions of application interfaces (APIs) that enable them to exchange information in an automated manner.

Identifying
: Another purpose of resource description is to enable a user to confirm the identity of a specific resource or to distinguish among several that have some overlapping descriptions. Computer processable resource descriptions like bar codes, QR codes, or RFID tags are also used to identify resources. In Semantic Web contexts, URIs serve this purpose. Color can be used as resource descriptions when physical resources need to be identified quickly.[262][CogSci]

Selecting
: Selecting in this context means the user activity of using resource descriptions to support a choice of resource from a collection, not the institutional activity of selecting resources for the collection in the first place. Search engines typically use a short "text snippet" with the query terms highlighted as resource descriptions to support selection. People often select resources with the least restrictions on uses as described in a Creative Commons license.[263][Law] A business might select a supplier or distributor that uses the same standard or industry reference model to describe its products or business processes because it is almost certain to reduce the cost of doing business with that business partner.[264][Bus]

Obtaining
: Physical resources often require significant effort to obtain after they have been selected. Catching a bus or plane involves coordinating your current location and time with the time and location the resource is available. With information resources in physical form, obtaining a selected resource usually meant a walk through the library stacks. With digital information resources, a search engine returns a list of the identifiers of resources that can be accessed with just another click, so it takes little effort to go from selecting among the query results to obtaining the corresponding primary resource.[265][Web]

Elaine Svenonius proposed that a fifth task called Navigation be added to the FRBR list, and in 2016 that happened but it was renamed as "Explore":

Navigation or Explore
: If users are not able to specify their information needs in a way that the *finding* functionality requires, they should be able to use relational and structural descriptions among the resources to navigate from any resource to other ones that might be better. Svenonius emphasizes generalization, aggregation, and derivational relationships. But in principle, any relationship or property could serve as the navigation "highway" between resources.

What some authors call “structural metadata” can be used to support the related tasks of moving within multi-part digital resources like electronic books, where each page might have associated information about previous, next, and other related pages. Documents described using XML models can use *Extensible Stylesheet Language Transformations (XSLT)* and *XPath* to address and select data elements, sub-trees, or other structural parts of the document.[268][Com]

5.3.2.4 Resource Description to Support Maintenance

Many types of resource descriptions that support selection (§5.3.2.1 Resource Description to Support Selection (page 234)) are also useful over time to support maintenance of specific resource and the collection to which they belong. In particular, technical information about resource formats and technology (software, computers, or other) needed to use the resources, and information needed to ensure resource integrity is often called “preservation metadata” in a maintenance context.

Resource descriptions that are more exclusively associated with maintenance activities include version information and effectivity, or useful life information. Equipment maintenance schedules are typically related to the number of miles driven (indicated by a car’s odometer), number of hours operated (stored by many engines), number of pages printed, or other easily recorded information about resource use or interactions. With smart resources now capable of capturing, analyzing, and communicating more data about real-time performance, more sophisticated prediction and scheduling of maintenance work is now possible. It is also easier to identify resources that are not being used as much as expected, which might imply that they are no longer needed and can thus be safely archived or discarded.

5.3.2.5 Resource Description for Sensemaking and Science

Up to now in §5.3.2, we have discussed how resource descriptions are used to perform well-defined tasks within an existing organizing system. However, there is a broader and less well-defined purpose of resource description that is older and more fundamental: the use of resource descriptions as the raw material for making sense of the world.

For thousands of years, even before the invention of written language, people have systematically collected things, information about those things, and observations of all kinds to understand how their world works. Paleolithic humans made cave paintings depicting the results of hunts and animal migrations; ancient Egyptians recorded the annual floods of the Nile River in stone carvings; and Babylonian, Egyptian, Chinese, and Mesoamerican astronomers organized

lunar, solar, and planetary observations as calendars starting about five thousand years ago.

These diverse efforts to impose meaning on experience by recording, analyzing, organizing, and reorganizing observations can be collectively described as *sensemaking*. (See the sidebar, Sensemaking and Organizing (page 240).)

Some aspects of sensemaking are hard-wired by evolution, which has given our brains powerful mechanisms that automatically simplify and organize the perceptual data we obtain from the world (see the sidebar Gestalt Principles (page 102)). But this automatic sensemaking is dominated and amplified by intentional sensemaking.

Intentional sensemaking takes place when systematic statistical, experimental, and scientific methods are consciously followed to extract and organize knowledge from collections of samples, observations, or measurements. It is critical to recognize here that the contents of these collections represent choices made about what to collect, because most things and most phenomena have a great many descriptions or properties that could be recorded about them.

After things or data have been collected, statistical methods summarize the values of properties in a collection or dataset and the relationships among them. Making sense of a single collection or dataset by determining the properties that contrast and classify the instances is the start toward the more important goal of understanding the larger set or population from which the initial collection is just a sample. There is no better example of this than the periodic table of elements developed by Mendeleev in 1869, who organized known elements on the basis of their common chemical properties and then successfully predicted some properties of yet undiscovered ones.

Computational models developed from the initial dataset can predict future observations. Classification models assign a new instance to a category (e.g., spam or not spam message, Madison or Hamilton as author, outdoor or indoor scene); regression models predict a specific value of some measurement (given a description of a new movie, how much money will it make?); ordinal regression models predict values for non-metric measures (how much will you like the movie?). Experimental methods for hypothesis testing help develop and refine models of any type by systematically varying the conditions under which observations are made to discover how the results change in different situations.

A fundamental challenge in sensemaking and modeling is finding a balance between the competing goals of understanding a particular collection or dataset and being able to apply that understanding to new instances. Models can differ in the number of resource descriptions they use as parameters, and it is easy and tempting to overfit a model by using more parameters that capture random

Sensemaking and Organizing

> People organize to make sense of equivocal inputs and enact this sense back into the world to make it more orderly.
>
> — (Weick 2005)

Sensemaking and *organizing* are intertwined. Ancient cultures recorded time-based observations and analyzed patterns among crop cycles, commodity prices, weather conditions, and astronomical sightings. Think back to the early astronomers, who oriented temple buildings to align with astronomical events and who decorated temple walls with zodiac imagery.

- Which of the planets and stars in the night sky should they observe and how should they record the details of those observations?
- What mathematical and statistical techniques should be used to analyze and describe these observations?
- What subset of observations are most useful in predicting the onset of the Nile River floods, caused by unobserved rainfall thousands of miles away?

Every choice about what to observe and how to describe it reflects a set of assumptions and potential hypotheses that are often implicit and unstated. Choices that increase understanding are built upon, and those that fail to provide insight are abandoned, but there is no guarantee that the iterative process of choosing what to observe and describe will lead to a correct understanding.

The principle that an accurate or comprehensive dataset is insufficient on its own to yield a correct model is exemplified in the interlocking efforts of Tycho Brahe and Johannes Kepler. Brahe was a 16th-century Danish nobleman astronomer who spent decades collecting data about the positions of hundreds of stars and the planets. However, because of prevailing religious and scientific biases, Brahe accepted the incorrect assumptions that the sun and planets revolved around the earth in circular orbits. After Brahe died in 1601, Kepler spent a decade analyzing Brahe's data, and then rejected the idea of earth-centric and circular planetary orbits in favor of elliptical ones with the sun at one focus. These new organizing principles for Brahe’s data made the model of the solar system vastly simpler, and Kepler was able to discover laws of planetary motion that are part of the foundation of modern astronomy and physics.[270][DS]

variations in observations. Overfitting produces spurious accuracy in reproducing the original observations, but it makes models less generalizable.

The highest level of sensemaking is the creation of scientific models or theories that propose interpretable and causal mechanisms for the observations. And just as automatic sensemaking creates simple explanations, scientists generally prefer simpler theories, a heuristic known as Occam's Razor, or the law of parsimony. Even though complex theories can sometimes be more accurate, simpler theories produce more testable predictions, making it easier to verify or refine the theory. Occam's famous principle, expressed eight centuries ago, is to prefer models that make the fewest assumptions, often measured in terms of the number of parameters or variables needed to make a prediction.[271][Bus] [272][Phil] [273][Com]

5.3.3 Identifying Properties

Once the purposes of description have been established, we need to identify the specific properties of the resources that can satisfy those purposes. There are four reasons why this task is more difficult than it initially appears.

- First, any particular resource might need many resource descriptions, all of which relate to different properties, depending on the interactions to be supported and the context in which they take place. Selecting people for a basketball team focuses on their physical properties such as height, strength, leaping ability, and coordination. Selections for a debate team will be more concerned with their verbal and intellectual properties.
- Second, different types of resources need to incorporate different properties in their descriptions. For resources in a museum, these might include materials and dimensions of pieces of art; for files and services managed by a network administrator, these include access control permissions; for electronic books or DVDs, they would include the digital rights management (DRM) code that expresses what you can and cannot do with the resource.
- Third, as we briefly touched on in §5.3.1.3, which properties participate in resource descriptions depends on who is doing the describing. It makes little sense to expect fine-grained distinctions and interpretations about properties from people who lack training in the discipline of organizing. We will return to this tradeoff in §5.3.6 and again in §5.4.1.
- Fourth, what might seem to be the same property at a conceptual level might be very different at an implementation level. Many resources have a resource description that is a surrogate or summary of the primary resource. For photos, paintings, and other resources whose appearance is their essence, an appropriate summary description can be a smaller, reduced resolution photo of the original. This surrogate is simple to create and easy for users to relate to the primary resource. On the other hand, distilling a text down to a short summary or abstract is a skill unto itself. Time-based resources provide greater challenges for summary. Should the summary of a

movie be a textual summary of the plot, a significant clip from the movie, a video summary, or something else altogether?

This implementation gap is often very large for properties about people because people are not as easy to measure as most types of resources. Businesses need to quantify a person's interest in their products to predict what price they would be willing to pay, but "interest" cannot be measured directly. Instead, predictions rely on proxy measures for "interest" like how long the customer looked at the product web page and whether they also looked at a competitor's web page.

Two important dimensions for understanding and contrasting resource properties used in descriptions and organizing principles are: property essence—whether the properties are intrinsically or extrinsically associated with the resource, and; property persistence—whether the properties are static or dynamic. Taken together these two dimensions yield four categories of properties, as illustrated in Figure 5.4, Property Essence x Persistence: Four Categories of Properties. These four categories provide a useful framework for thinking about resource properties, even if, at times, the classification of properties is debatable.[274[CogSci]]

5.3.3.1 Intrinsic Static Properties

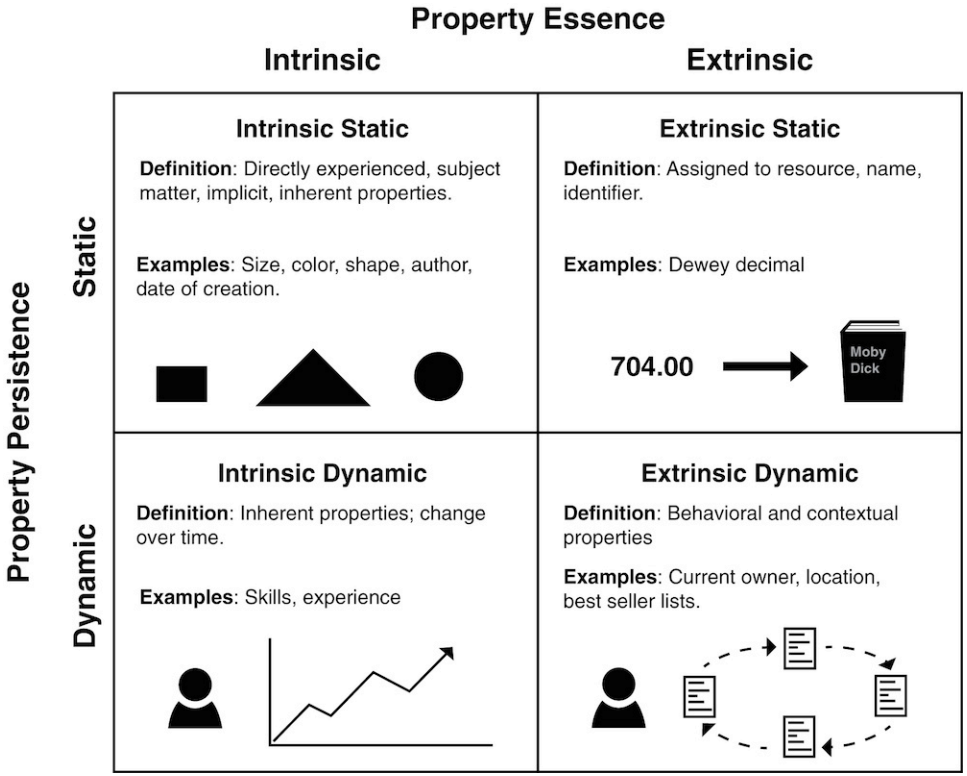
Intrinsic or implicit properties are inherent in the resource and can often be directly perceived or experienced. Static properties do not change their values over time. The species of an animal, the material of composition of a wooden chair, and the diameter of a wheel are all static properties that do not change their values over time. Static properties like color or shape are often used to describe and organize physical resources.

Intrinsic physical properties are usually just part of resource descriptions. In many cases, physical properties describe only the surface layer of a resource, revealing little about what something is or its original intended purpose, what it means, or when and why it was created. The author of a song and the context of its creation are other examples of intrinsic and static resource properties that are not directly perceivable.

Intrinsic descriptions are often extracted or calculated by computational processes. For example, a computer program might calculate the frequency and distribution of words in some particular document. Similarly, visual signatures or audio fingerprints are intrinsic descriptions (§5.4 Describing Non-text Resources (page 257)).

Some relationships among resources are intrinsic and static, like the parent-child relationship or the sibling relationship between two children with the same parents. Part-whole or compositional relationships for resources with

Figure 5.4. Property Essence x Persistence: Four Categories of Properties.



The distinctions of property persistence and property essence combine to distinguish four categories of properties: intrinsic static, extrinsic static, intrinsic dynamic, and extrinsic dynamic properties.

parts are also intrinsic static properties often used in resource descriptions. However, it is better to avoid treating resource relationships as properties, and instead express them as relations. Chapter 6, *Describing Relationships and Structures* discusses part-whole and other semantic relationships in great detail.

Intrinsic Static Properties Define a Dalmatian

The spots on a Dalmatian dog are intrinsic static properties that appear shortly after birth, and they are so distinctive that it is impossible to describe the breed without acknowledging the spots.

This particular Dalmatian is the "greeter" at the Viader Winery in Deer Park, California. The dog is nice and the wines are excellent.

(Photo by R. Glushko.)

5.3.3.2 Extrinsic Static Properties

Extrinsic or explicit properties are assigned to a resource rather than being inherent in it. The name or identifier of a resource is often arbitrary but once assigned does not usually change. Arranging resources according to the alphabetical or numerical order of their descriptive identifiers is a common organizing principle. Classification numbers and subject headings assigned to bibliographic resources are extrinsic static properties, as are the serial numbers stamped on or attached to manufactured products.

For information resources that have a digital form, the properties of their printed or rendered versions might not be intrinsic. Some text formats completely separate content from presentation, and as a result, style sheets can radically change the appearance of a printed document or web page without altering the primary resource in any way. For example, were a different style applied to this paragraph to highlight it in bold or cast in 24-point font, its content would remain the same.

5.3.3.3 Intrinsic Dynamic Properties

Intrinsic dynamic properties change over time. Developmental personal characteristics like a person's height and weight, skill proficiency, or intellectual capacity, for example. Because these properties are not static, they are usually employed only to organize resources whose membership in the collection is of limited duration. Sports programs or leagues that segregate participants by age or years of experience are using intrinsic dynamic properties to describe and organize the resources.

5.3.3.4 Extrinsic Dynamic Properties

Extrinsic dynamic properties are in many ways arbitrary and can change because they are based on usage, behavior, or context. The current owner or location of a resource, its frequency of access, the joint frequency of access with other resources, its current popularity or cultural salience, or its competitive advantage over alternative resources are typical extrinsic and dynamic properties that are used in resource descriptions. A topical book described as a best seller one year might be found in the discount sales bin a few years later. A student's grade point average is an extrinsic dynamic property.

Extrinsic dynamic properties are useful features for data scientists making prediction or classification models. Your current location, the thing you just bought, and the place you bought it can be viewed as manifestations of unobservable preferences and values. Fingerprints found on a doorknob at a crime scene are an extrinsic dynamic property associated with the door, and clever detectives would analyze them along with other interaction resources they discovered with the goal of identifying the person for whom the fingerprints are intrinsic static properties.

Many relationships between resources are extrinsic and dynamic properties, like that of best friend.

Contextual properties are those related to the situation or context in which a resource is described. Dey defines *context* as "any information that characterizes a situation related to the interactions between users, applications, and the surrounding environment." [275][Com] This open-ended definition implies a large number of contextual properties that might be used in a description; crisper definitions of context might be "location + activity" or "who, when, where, why." Since context changes, context-based descriptors might be appropriate when assigned but can have limited persistence and effectivity (§4.5 Resources over Time (page 198)); the description of a document as "receipt of a recent purchase" will not be useful for very long.

Citations of one information resource by another are extrinsic static descriptions when they are in print form, but when they are published in digital libraries it is usually the case that "cited by" is a dynamic resource description. Similarly, any particular link from one web page to another is an extrinsic static description, but because many web pages themselves are highly dynamic, we can also consider links as dynamic as well. Citations and web links are discussed in more detail in Chapter 6.

Resources are often described with *cultural properties* that derive from conventional language or culture, often by analogy, because they can be highly evocative and memorable. [277][Ling]

Why are Ottoman Carpets Named After a German Painter?

An example of a cultural category that has far outlasted its motivation is that of the Holbein carpet. A particular type of geometrically patterned Ottoman rug came to be known as a "Holbein carpet" after the German Renaissance painter Hans Holbein, who often depicted the rugs in his work (probably to show off his extremely meticulous technique). Holbein was famous in his time, and his commissioned paintings of the English King Henry VIII have Henry standing on such rugs. This painting, called "The Ambassadors," was painted in 1533 and now hangs in The National Gallery, London.

(Source: Google Art Project)

Sometimes a cultural description outlives its salience, losing its power to evoke anything other than puzzlement about what it might mean.[278][Ling]

Latent Feature Creation and Netflix Recommendations

Recent advances in computing technology and *data science* techniques are making it possible to discover or create resource properties that are called "latent" because they are inferred rather than observed. Many such features are used by businesses to segment customers or make recommendations to them based on their recent behavior, so these features are also extrinsic and dynamic.

Your own movie preferences prove that easy to identify properties like sex and age do not differentiate movie watchers enough to make good recommendations, even if you combine them to create a category like "single male students between 18 and 25." Netflix found that it was necessary to combine demographic properties, viewing history, and browsing behavior with very detailed ratings of dozens of movie properties to make good recommendations. It takes enormous computing power to discover a category of Netflix users who typically like action movies, yet consistently hover their mouse over romance movies, and to use this latent feature to recommend a sub-genre of western movies (one of nearly 100,000) that it calls "Romantic Action Adventure Movies."[276][Com]

For the Lego boys, current with the latest *Star Wars* movies, “light saber” was just the obvious description for a long, neon tube with a handle. However, someone unfamiliar with the *Star Wars* franchise might not understand“light saber,” and would describe the piece some other way.

5.3.4 Designing the Description Vocabulary

After we have determined the properties to use in resource descriptions, we need to design the description vocabulary: the set of words or values that represent the properties. §4.4 Naming Resources (page 188) discussed the problems of naming and proposed principles for good names, and since names are a very important resource description, much of what we said there applies generally to the design of the description vocabulary.

However, because the description vocabulary as a whole is much more than just the resource name, we need to propose additional principles or guidelines for this step. In addition, some new design questions arise when we consider all the resource descriptions as a set whose separate descriptions are created by many people over some period of time.

5.3.4.1 Principles of Good Description

In *The Intellectual Foundation of Information Organization,* Svenonius proposes a set of principles or “directives for design” of a description language. Her principles, framed in the narrow context of bibliographic descriptions, generally apply to the broad range of resource types we consider in this book.

User Convenience
: Choose description terms with the user in mind; these are likely to be terms in common usage among the target audience.

Representation
: Use descriptions that reflect how the resources describe themselves; assume that self-descriptions are accurate.

Sufficiency and Necessity
: Descriptions should have enough information to serve their purposes and not contain information that is not necessary for some purpose; this might imply excluding some aspects of self-descriptions that are insignificant.

Standardization
: Standardize descriptions to the extent practical, but also use aliasing to allow for commonly used terms.

Integration
: Prefer the same properties and terms for all types of resources.

Any set of general design principles faces two challenges.

- The first is that implementing any principle requires many additional and specific context-dependent choices for which the general principle offers little guidance. For example, how does the principle of Standardization apply if multiple standards already exist in some resource domain? Which of the competing standards should be adopted, and why?
- The second challenge is that the general principles can sometimes lead to conflicting advice. The User Convenience recommendation to choose description terms in common use fails if the user community includes both ordinary people and scientists who use different terms for the same resources; whose "common usage" should prevail?

5.3.4.2 Who Uses the Descriptions?

Focus on the user of the descriptions. This is a core idea that we cannot overemphasize because it is implicit in every step of the process of resource description. All of the design principles in the previous section share the idea that the design of the description vocabulary should focus on the user of the descriptions. Are the resources being organized personal ones, for personal and mostly private purposes? In that case, the description properties and terms can be highly personal or idiosyncratic and still follow the design principles.

Similarly, when resource users share relevant knowledge, or are in a context where they can communicate and negotiate, if necessary, to identify the resources, their resource descriptions can afford to be less precise and rigorous than they might otherwise need to be. This helps explain the curious descriptions in the Lego story with which we began this chapter. The boys playing with the blocks were talking to each other with the Legos in front of them. If they had not been able to see the blocks the others were talking about, or if they had to describe their toys to someone who had never played with Legos before, their descriptions would have been quite different.

More often, however, resource descriptions can not assume this degree of shared context and must be designed for user categories rather than individual users: library users searching for books, business employees or customers using part and product catalogs, scientists analyzing the datasets from experiments or simulations. In each of these situations resource descriptions will need to be understood by people who did not create them, so the design of the description vocabulary needs to be more deliberate and systematic to ensure that its terms are unambiguous and sufficient to ensure reliable context-free interpretation. A single individual seldom has the breadth of domain knowledge and experience with users needed to devise a description vocabulary that can satisfy diverse users with diverse purposes. Instead, many people working together typically

develop the required description vocabulary. We call the results institutional vocabularies, to contrast them with individual or cultural ones. (We will discuss this contrast more fully in Chapter 7, *Categorization: Describing Resource Classes and Types*)

Some resource descriptions are designed for use by machines, which seemingly reduces the importance of design principles that consider user preferences or common uses. However, even if resources are described and organized by algorithms, when people need to explain the classifications and predictions that the algorithms produce, resource descriptions that are comprehensible and easily communicated are preferable to statistically optimal ones. Moreover, standardization and integration principles become more important for inter-machine communication to enable efficient processing, reuse of data and software, and increased interoperability among organizing systems.[280][Com]

Stop and Think: Description and Expertise

Everyone knows something about trees, but some people know more than others, and their particular experience and perspective influences how they describe trees. What kind of properties and descriptions would be used by university students? By research botanists? By landscape designers? By park maintenance workers? By indigenous people who live in tropical rain forests?

5.3.4.3 Controlled Vocabularies and Content Rules

As we defined in §4.4.3.2, a *controlled vocabulary* is a fixed or closed set of description terms in some domain with precise definitions that is used instead of the vocabulary that people would otherwise use. For example, instead of the popular terms for descriptions of diseases or symptoms, medical researchers and teaching hospitals can use the National Library of Medicine's *Medical Subject Headings (MeSH)* controlled vocabulary.

We can distinguish a progression of vocabulary control: a glossary is a set of allowed terms; a thesaurus is a set of terms arranged in a hierarchy and annotated to indicate terms that are preferred, broader than, or narrower than other terms; an ontology expresses the conceptual relationships among the terms in a formal logic-based language so they can be processed by computers. We will say more about ontologies in Chapter 6.

Content rules are similar to controlled vocabularies because they also limit the possible values that can be used in descriptions. Instead of specifying a fixed set of values, content rules typically restrict descriptions by requiring them to be of a particular data type (integer, Boolean, Date, and so on). Possible values are constrained by logical expressions (e.g., a value must be between 0 and 99) or

regular expressions (e.g., must be a string of length 5 that must begin with a number). Content rules like these are used to ensure valid descriptions when people enter them in web forms or other applications.

5.3.4.4 Vocabulary Control as Dimensionality Reduction

In most cases, a controlled vocabulary is a subset of the natural or uncontrolled vocabulary, but sometimes it is a new set of invented terms. This might sound odd until we consider that the goal of a controlled vocabulary is to reduce the number of descriptive terms assignable to a resource. Stated this way the problem is one of *dimensionality reduction*, transforming a high-dimensional space into a lower-dimensional one. Reducing the number of components in a multidimensional description can be accomplished by many different statistical techniques that go by names like "feature extraction," "principle components analysis," "orthogonal decomposition," "latent semantic analysis," "multidimensional scaling," and "factor analysis." [282[DS]]

These techniques might sound imposing and they are computationally complex, but they all have the same simple concept at their core, that the features or properties that describe some resource are often highly correlated. For example, a document that contains the word "car" is more likely to contain the words "driver" and "traffic" than a document that does not. Similar correlations exist among the visual features used to describe images and the acoustic features that describe music. Dimensionality reduction techniques analyze the correlations between resource descriptions to transform a large set of descriptions into a much smaller set of uncorrelated ones. In a way this implements the principle of Sufficiency and Necessity we mentioned in §5.3.4.1 Principles of Good Description (page 247) because it eliminates description dimensions or properties that do not contribute much to distinguishing the resources.

Here is an oversimplified example that illustrates the idea. Suppose we have a collection of resources, and every resource described as "big" is also described as "red," and every "small" resource is also "green." This perfect correlation between color and size means that either of these properties is sufficient to distinguish "big red" things from "small green" ones, and we do not need clever algorithms to figure that out. But if we have thousands of properties and the correlations are only partial, we need the sophisticated statistical approaches to choose the optimal set of description properties and terms, and in some techniques the dimensions that remain are called "latent" or "synthetic" ones because they are statistically optimal but do not map directly to resource properties.

5.3.5 Designing the Description Form

By this step in the process of resource description we have made numerous important decisions about which resources to describe, the purposes for which we are describing, them, and the properties and terms we will use in the descriptions. As much as possible we have described the steps at a conceptual level and postponed discussion of implementation considerations about the notation, syntax, and deployment of the resource descriptions separately or in packages. Separating design from implementation concerns is an idealization of the process of resource description, but is easier to learn and think about resource description and organizing systems if we do. We discuss these implementation issues in Chapter 9, *The Forms of Resource Descriptions*.

Sometimes we have to confront legacy technology, existing or potential business relationships, regulations, standards conformance, performance requirements, or other factors that have implications for how resource descriptions must or should be implemented, stored, and managed. We will take this more pragmatic perspective in Chapter 11, *The Organizing System Roadmap,* but until then, we will continue to focus on design issues and defer discussion of the implementation choices.

5.3.6 Creating Resource Descriptions

Resource descriptions can be created by professionals, by the authors or creators of resources, by users, or by computational or automated means.

Professionally-created resource descriptions, author- or user-created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs. A natural solution is to try to combine desirable aspects from each in hybrid approaches. For example, the vocabulary for a new resource domain may arise from tagging by end users but then be refined by professionals, lay classifiers may create descriptions with help from software tools that suggest possible terms, or software that creates descriptions can be improved by training it with human-generated descriptions, a form of *supervised learning* (see §7.5.3.3).

Often existing resource descriptions can or must be transformed or enhanced to meet the ongoing needs of an organizing system, and sometimes these processes can be automated. We will defer further discussion of those situations to Chapter 10, *Interactions with Resources*. In the discussion that follows we focus on the creation of new resource descriptions where none yet exist.

5.3.6.1 Resource Description by Professionals

Before the web made it possible for almost anyone to create, publish, and describe their own resources and to describe those created and published by others, resource description was generally done by professionals in institutional contexts. Professional indexers and catalogers described bibliographic and museum resources after having been trained to learn the concepts, controlled descriptive vocabularies, and the relevant standards. In information systems domains professional data and process analysts, technical writers, and others created similarly rigorous descriptions after receiving analogous training. We have called these types of resource descriptions institutional ones to highlight the contrast between those created according to standards and those created informally in *ad hoc* ways, especially by untrained or undisciplined individuals.[285][Bus]

5.3.6.2 Resource Description by Authors or Creators

The author or creator of a resource can be presumed to understand the reasons why and the purposes for which the resource can be used. And, presumably, most authors want to be read, so they will describe their resources in ways that will appeal to and be useful to their intended users. However, these descriptions are unlikely to use the controlled vocabularies and standards that professional catalogers would use.

5.3.6.3 Resource Description by Users

Today's web contains a staggering number of resources, most of which are primary information resources published as web content, but many others are resources that stand for "in the world" physical resources. Most of these resources are being described by their users rather than by professionals or by their authors. These "at large" users are most often creating descriptions for their own benefit when they assign tags or ratings to web resources, and they are unlikely to use standard or controlled descriptors when they do so. The resulting variability can be a problem if creating the description requires judgment on the tagger's part. Most people can agree on the length of a particular music file but they may differ wildly when it comes to determining to which musical genre that file belongs. Fortunately most web users implicitly recognize that the potential value in these "Web 2.0" or "user-generated content" applications will be greater if they avoid egocentric descriptions. In addition, the statistics of large sample sizes inevitably leads to some agreement in descriptions on the most popular applications because idiosyncratic descriptions are dominated in the frequency distribution by the more conventional ones.[287][Web]

We are not suggesting that professional descriptions are always of high quality and utility, and socially produced ones are always of low quality and utility.[288][CogSci] Rather, it is important to understand the limitations and

qualifications of descriptions produced in each way. Tagging lowers the barrier to entry for description, making organizing more accessible and creating descriptions that reflects a variety of viewpoints. However, when many tags are associated with a resource, it increases *recall* while decreasing *precision.* (See §5.3.6.3 Resource Description by Users (page 252))

Metacrap

In an often-cited essay (Doctorow 2001) provocatively titled "Metacrap: Putting the torch to seven straw-men of the meta-utopia," Cory Doctorow argues that much human-created metadata *is* of low quality because "people lie, people are lazy, people are stupid, mission impossible—know thyself, schemas are not neutral, metrics influence results, (and) there is more than one way to describe something."

5.3.6.4 Automated and Computational Resource Description

A picture's EXIF file created by a digital camera records properties associated with the camera and its settings, as well as some properties of the photo-taking context. (See Figure 5.6, Contrasting Descriptions for a Work of Art. for an example.) Creating this highly detailed description by hand would be nearly impossible. The downside, however, is that the automated description does not capture the meaning of the photo; an automated picture description captures the time and place, but not that it is a picture of a honeymoon vacation. The difference between automated and human description is called the *semantic gap* (§4.4.2.5).

Any resource that is smart enough to collect data about its state or environment is creating resource descriptions automatically (See §4.2.3.2). Resources with computational capabilities can process the raw sensor data to identify important events and create more interpretable descriptions.

Some computational approaches create resource descriptions that are similar in purpose to those created by human describers. Text mining and summarization systems for customer comments about products can reduce thousands of comments to a list of the most important features.[289][Com] People shopping for books at Amazon.com get insights about a book's content and distinctiveness from the statistically improbable phrases that it has identified by comparing all the books for which it has the complete text.[290][Com]

Computational descriptions can use any observable or latent variable (see the sidebar, Latent Feature Creation and Netflix Recommendations (page 246)) except some that are prohibited by law, such as race, religion, national origin, and marital status, to prevent discrimination. In practice, however, this prohibition is easily circumvented because these properties can usually be predicted using other ones. For example, race can often be reliably predicted using residence address and surname.[291][DS]

Of course, all information retrieval systems compare a description of a user's needs with descriptions of the resources that might satisfy them. IR systems differ in the resource properties they emphasize; word frequencies and distributions for documents in digital libraries, links and navigation behavior for web pages, acoustics for music, and so on. These different property descriptions determine the comparison algorithms and the way in which relevance or similarity of descriptions is determined. We say a lot more about this in §5.4 Describing Non-text Resources (page 257) and in Chapter 10.

5.3.7 Evaluating Resource Descriptions

Evaluation is implicit in many of the activities of organizing systems we described in Chapter 3, *Activities in Organizing Systems* and is explicit when we maintain a collection of resources over time. In this section, we focus on the narrower problem of evaluating resource descriptions.

Evaluating means determining *quality* with respect to some criteria or dimensions. Many different sets of criteria have been proposed; for repositories of digital resources, the most commonly used ones are accuracy, completeness, and consistency. Other typical criteria are timeliness, interoperability, and usability. It is easy to imagine these criteria in conflict; efforts to achieve accuracy and completeness might jeopardize timeliness; enforcing consistency might preclude modifications and personalizations that would enhance usability.

Stop and Think: Defining Quality

What characteristics or criteria would you use to determine the quality of a car? Of food? Of clothing? Of a place to live? Which of these criteria are domain-specific, and which ones apply more generally to many types of resources?

The *quality* of the outcome of the multi-step process proposed in this chapter is a composite of the *quality* created or squandered at each step. A scope that is too granular or abstract, overly ambitious or vague intended purposes, a description vocabulary that is hard to use, or giving people inadequate time to create good descriptions can all cause quality problems, but none of these decisions is visible at the end of the process where users interact with resource descriptions.

5.3.7.1 Evaluating the Creation of Resource Descriptions

When professionals create resource descriptions in a centralized manner, which has long been the standard practice for many resources in libraries, there is a natural focus on *quality* at the point of creation to ensure that the appropriate controlled vocabularies and standards have been used. However, the need for resource description generalizes to resource domains outside of the traditional bibliographic one, and other *quality* considerations emerge in those contexts.

Resource descriptions in private sector firms are essential to running the business and in interacting efficiently with suppliers, partners, and customers. Compared to the public sector, there is much greater emphasis on the economics and strategy of resource description.[293[Bus]] What is the value of resource description? Who will bear the costs of producing them? Which of the competing industry standards will be followed? Some of these decisions are not free choices as much as they are constraints imposed as a condition of doing business with a dominant economic partner, which is sometimes a governmental entity.

For example, a firm like Wal-Mart with enormous market power can dictate terms and standards to its suppliers because the long-term benefits of a Wal-Mart contract usually make the initial accommodation worthwhile. Likewise, governments often require their suppliers to conform to open standards to avoid lock-in to proprietary technologies. [294[Bus]]

In both the public and private sectors there is increased use of computational techniques for creating resource descriptions because the number of resources to be described is simply too great to allow for professional description. A great deal of work in text data mining, web page classification, semantic enrichment, and other similar research areas is already under way and is significantly lowering the cost of producing useful resource descriptions. Some museums have embraced approaches that automatically create user-oriented resource descriptions and new user interfaces for searching and browsing by transforming the professional descriptions in their internal collections management systems. Google's ambitious project to digitize millions of books has been criticized for the quality of its algorithmically extracted resource descriptions, but we can expect that computer scientists will put the Google book corpus to good use as a research test bed to improve the techniques.[296[Com]]

Web 2.0 applications that derive their value from the aggregation and interpretation of user-generated content can be viewed as voluntarily ceding their authority to describe and organize resources to their users, who then tag or rate them as they see fit. In this context the consistency of resource description, or the lack of it, becomes an important issue, and many sites are using technology or incentives to guide users to create better descriptions.

5.3.7.2 Evaluating the Use of Resource Descriptions

Regardless of, or in addition to, any quality criteria applied to the creation and selection of resource descriptions, at some point the resource descriptions meet their intended users. The most important quality criterion at that point is whether the resource descriptions satisfy their intended purposes in a usable way. In many ways, the answer is a disappointing no.

For example, in one of the earliest revisions to the original HTML specification, a <META> tag was added to allow creators of web resources to define a set of key terms to describe a website or web page. This well-motivated resource description was to be used by search engines to improve the relevance of retrieved pages. However, it soon became obvious that it was possible to "game" the META tag by adding popular terms even though they did not accurately describe the page. Today search engines ignore the `<META>` tag for ranking pages, but many other techniques that use false *resource* descriptions continue to plague web users. (See §3.5.3.3.)

The design of a description vocabulary circumscribes what can be said about a resource, so it is important to recognize that it implicitly determines what cannot be said as well, with unintended negative consequences for users. The *resource* description schema implemented in a physician's patient management system defines certain types of recordable information about a patient's visit—the date of the visit, any tests that were ordered, a diagnosis that was made, a referral to a specialist. The schema, and its associated workflow, impose constraints that affect the kinds of information medical professionals can record and the amount of space they can use for those descriptions. Moreover, such a schema might also eliminate vital unstructured space that paper records can provide, where doctors communicate their rationale for a diagnosis or decision without having to fit it into any particular box.

However, when resource descriptions are the data used to train models for prediction or classification, the focus of evaluation is not on the descriptions, which are often assumed to be accurate observations about the world. Instead, evaluation focuses on the model, and "model selection" is the task of choosing which of several competing models best fit the original data while also generalizing well to new data. In any event, any quality problems or selection biases with the original data will undermine the value of whatever model is selected.

5.3.7.3 The Importance of Iterative Evaluation

The inevitable conflicts between *quality* goals mean that there will be compromises among the *quality* criteria. Furthermore, increasing *scale* in an organizing system and the steady improvements of computational techniques for resource description imply that the nature of the compromise will change over time. As a result, a single evaluation of resource descriptions at one moment in time will not suffice.

This makes usage records, navigation history, and transactional data extremely important kinds of resource descriptions because they enable you to focus efforts on improving *quality* where they are most needed. Furthermore, for organizing systems with many types of resources and user communities, this information can enable the tailoring of the nature and extent of resource description to find the right balance between "rich and comprehensive" and "simple and efficient" approaches. Each combination of resource type and user community might have a different solution.

The idea that *quality* is a property of an end-to-end process is embodied in the "quality movement" and statistical process control for industrial processes but it applies equally well to resource description. The central idea is that *quality* cannot be tested in by inspecting the final products. Instead, *quality* is achieved through process control—measuring and removing the variability of every process needed to create the products.[297][Bus] Explicit feedback from users or implicit feedback from the records of their resource interactions needs are essential as we iterate through the design process and revisit the decisions made there.

5.4 Describing Non-text Resources

Many of the principles and methods for resource description were developed for describing text resources in physical formats. Those principles have had to evolve to deal with different types of resources that people want to describe and organize, from paintings and statues to MP3s, JPEGs, and MPEGs.

Some descriptions for non-text resources are text-based, and are most often assigned by people. Other descriptions are in non-text formats are extracted algorithmically from the content of the non-text resource. These latter content-based resource descriptions capture intrinsic technical properties and in some domains are able to describe *aboutness* with some accuracy, thanks to breakthroughs in machine learning.

5.4.1 Describing Museum and Artistic Resources

The problems associated with describing multimedia resources are not all new. Museum curators have been grappling with them since they first started to collect, store, and describe artifacts hundreds of years ago. Many artifacts may represent the same work (think about shards of pottery that may once have been part of the same vase). The materials and forms do not convey semantics on their own. Without additional research and description, we know nothing about the vase; it does not come with any sort of title page or tag that connects it with a 9th-century Mayan settlement. Since museums can acquire large batches of artifacts all at once, they have to make decisions about which resources they can afford to describe and how much they can describe them.

German art historian Erwin Panofsky first codified one approach to these problems of description. In his classic *Studies in Iconology,* he defined three levels of description that can be applied to an artistic work or museum artifact. Figure 5.6, Contrasting Descriptions for a Work of Art. contrasts these three levels in the descriptions of a marble statue. It also shows the striking differences between the EXIF description in a digital photo of the statue and those created by people.

5.4.2 Describing Images

Digital cameras, including those in cell phones, take millions of photos each day. Unlike the images in museums and galleries, most of these images receive few descriptions beyond those created by the device that made them. Nevertheless, a great many of them end up with some limited descriptions in Facebook, Instagram, Flickr, Picasa, DeviantArt, or others of the numerous places where people share images, or in professional image applications like Light Room. All of these sites provide some facilities for users to assign tags to images or arrange them in named groups.

Many different computational approaches have been used to describe or classify images. One approach uses the visual signature of an image extracted from low-level features like color, shape, texture, and luminosity, which are then used to distinguish significant regions and objects. Image similarity is computed to create categories of images that contain the same kinds of colors, objects, or settings, which makes it easy to find duplicate or modified images.[300][Com]

For computers to identify specific objects or people in images, it is logically necessary to train them with images that are already identified. In 2005 Luis van Ahn devised a clever way to collect large amounts of labeled images with a web-based game called ESP that randomly paired people to suggest labels or tags for an image. The obvious choices were removed from contention, so a photo of a bird against a blue sky might already strike “bird” and “sky” from the set of

Figure 5.6. Contrasting Descriptions for a Work of Art.

EXIF Summary

Make	NIKON CORPORATION
Model	NIKON D90
Aperture	9
Exposure Time	1/320 (0.003125 sec)
Lens	ID AF-S DX VR Zoom-Nikkor 18-105mm f/3.5-5.6G ED
Focal Length	21.0 mm
Flash	Auto, Did not fire
File Size	4.7 MB
File Type	JPEG
Image Height	4288
Image Width	2848
Date & Time	2012:12:03 10:31:14

3 Levels

Primary
Marble statue of nude woman standing on a seashell.

Secondary
Statue made in 2005 by Lucio Carusi of Carrara, Italy, titled "Venus", made of local marble.

Interpretive
This is a 3d transformation of the 1486 painting by Italian painter Sondro Botticelli, titled "The Birth of Venus", now in the Uffizi Gallery in Florence. Carusi's Venus is substantially slimmer in proportions than Botticelli's because of changing notions of female beauty.

Descriptions for works of art can contrast a great deal, especially between those captured by a device like a digital camera and those created by people. Furthermore, the descriptions created by people differ according to the expertise of the creator and the amount of subjective interpretation applied in the description.

(Photo by R. Glushko. The statue, titled "Venus," was made by Lucio Carusi, of Carrara, Italy, and is currently part of a private collection.)

acceptable words, leaving users to suggest words such as "flying" and "cloudless." Van Ahn also invented the reCAPTCHA technique that presents images of text from old books being digitized, which improves the accuracy of the digitization while verifying that the user of a web site is a person and not a robot program.[301][Web]

However, if short text descriptions or low-level image properties are the only features available to train an image, otherwise irrelevant variations in the position, orientation, or illumination of objects in images will make it very difficult to distinguish objects that look similar, like a white wolf and the wolf-like white dog called a Samoyed. This problem can be addressed by using deep neural networks, which exploit the idea that low-level image features can be combined into many layers of higher-level ones; edges combine to form motifs or patterns, patterns combine to form parts of familiar objects, and parts combine to form complete objects. This hierarchical composition enables the highest-level representations to become insensitive to the lower-level variations that plague the other approaches.

In 2012, when deep learning techniques were applied to a dataset of about a million images that contained a thousand different object categories, they reduced the error rate by half. This spectacular breakthrough, and the fact that the deep learning techniques that derive layers of features from the input data are completely general, rapidly caused deep learning to be applied to many other domains with high-dimensional data. Facebook uses deep learning to identify people in photos, Google uses it for speech recognition and language translation, and rapid captioning for images and video are on the horizon. Wearable computers might use it to layer useful information onto people's views of the world, creating real-time augmented reality.[302][Com]

5.4.3 Describing Music

Some parts of describing a song are not that different from describing text: You might want to pull out the name of the singer and/or the songwriter, the length of the song, or the name of the album on which it appears. But what if you wanted to describe the actual content of the song? You could write out the lyrics, but describing the music itself requires a different approach.

Describing music presents challenges quite different from those involved in describing texts or images. Poems and paintings are tangible things that we can look at and contemplate, while the aural nature of music means that it is a fleeting phenomenon that can only be experienced in the performative moment. Even musical scores and recordings, while as much tangible things as paintings and poems, are merely containers that hold the potential for musical experience and not the music itself. Most contemporary popular music is in the form of songs, in which texts are set to a melody and supported by instrumental harmonies. If we want to categorize or describe such music by its lyrical content, we can still rely on methods for describing texts. But if we want to describe the music itself, we need to take a somewhat different approach.

Several people and companies working in multimedia have explored different processes for how songs are described. On the heavily technological side, software applications such as Shazam and Midomi can create a content-based *audio fingerprint* from a snippet of music. *Audio fingerprinting* renders a digital description of a piece of music, which a computer can then interpret and compare to other digital descriptions in a library.[303][Com]

On the face of it, contemporary music streaming services represent the apex of music classification and description. Pandora, for example, employs trained musicologists to listen to the music and then categorize the genres and musical materials according to a highly controlled musical vocabulary. The resulting algorithm, the "Music Genome," can essentially learn to define a listener's musical tastes by means of this musical tagging, and can then use that information to suggest other music with similar characteristics.[304][Com]

A DJ Describes and Organizes Music

Casual music fans might describe their music using the names of the songs or performers and might organize it according to genres like "Pop," "Rock," or "Classical." A professional DJ, however, emphasizes different properties, especially the beats per minute of the music.

This annotated photo shows a portion of the music collection of noted DJ "Kid Kameleon" (http://kidkameleon.com/).

(Photo and annotation by Matt Earp. Used with permission.)

But musicians have been thinking about how to describe music for centuries, and while the Music Genome certainly brims with complexity, it pales in comparison to the sophistication of the much older "pen-and-paper" methods from which it derives. Ethnomusicology (loosely defined as the study of global musical practices in their social contexts) has arguably made greater strides towards comprehensive descriptions of musical resources than any other field of musicological study. Since the late 19th century, ethnomusicologists have created complex methods of notation and stylistic taxonomies to capture and categorize the music of both Western and non-Western cultures.

On a more granular level, musicians are endlessly innovative in finding ways to categorize, describe, and analyze not simply large-scale musical genres, but the notes themselves. In the accompanying photo showing the record collection of professional DJ "Kid Kameleon," we see that the records are arranged not simply by genre, but also by beats-per-minute (BPM). For Kid Kameleon, these re-

cords represent the resources of his musical creative process, and arranging them by BPM allows him to pull exactly the correct musical material he needs to keep the music flowing during a performance. His classification system is therefore a taxonomy that moves from the broad strokes of genre down to the fine grains of specific arrangements of notes and rhythms. This photo is not simply a picture of a record collection: it is a visual representation of an artist's creative process.

5.4.4 Describing Video

Video is yet another resource domain where work to create resource descriptions to make search more effective is ongoing. Video analytics techniques can segment a video into shorter clips described according to their color, direction of motion, size of objects, and other characteristics. Identifying anomalous events and faces of people in video has obvious applications in security and surveillance.[307][Com] Identifying specific content details about a video currently takes a significant amount of human intervention, though it is possible that image signature-matching algorithms will take over in the future because they would enable automated ad placement in videos and television.[308][Bus]

5.5 Key Points in Chapter Five

- Information retrieval is characterized as comparing a description of a user's needs with descriptions of the resources that might satisfy them. Different property descriptions determine the comparison algorithms and the way in which relevance or similarity of descriptions is determined.

  (See §5.2.1 Naming {and, or, vs.} Describing (page 219))

- In different contexts, the terms in resource descriptions are called keywords, index terms, attributes, attribute values, elements, data elements, data values, or "the vocabulary," labels, or tags.

  (See §5.2.2 "Description" as an Inclusive Term (page 220))

- In the library science context of *bibliographic description,* a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources.

  (See §5.2.2 "Description" as an Inclusive Term (page 220))

- A bibliographic description of an information resource is most commonly realized as a structured record in a standard format that describes a specific resource.

  (See §5.2.2.1 Bibliographic Descriptions (page 220))

- Metadata is structured description for information resources of any kind, which makes it a superset of bibliographic description.

  (See §5.2.2.2 Metadata (page 221))

- A relational database schema is designed to restrict resource descriptions to be simple and completely regular sets of attribute-value pairs.

  (See §5.2.2.2 Metadata (page 221))

- The Resource Description Framework (RDF) is a language for making computer-processable statements about web resources that is the foundation for the vision of the Semantic Web.

  (See §5.2.2.4 Resource Description Framework (RDF) (page 223))

- An aggregation is a set of information objects that, when considered together, compose another named information object.

  (See §5.2.2.4 Resource Description Framework (RDF) (page 223))

- The dominant historical view treats resource descriptions as a package of statements, an alternate framework focuses on each individual description or assertion about a single resource.

  (See §5.2.3 Frameworks for Resource Description (page 226))

- Design of the description vocabulary should focus on the user of the descriptions. Svenonius proposes five principles for a description vocabulary: user convenience, representation, sufficiency and necessity, standardization, and integration.

  (See §5.3 The Process of Describing Resources (page 227))

- The process of describing resources involves several interdependent and iterative steps, including determining scope, focus and purposes, identifying resource properties, designing the description vocabulary, designing the description form and implementation, and creating and evaluating the descriptions.

  (See §5.3 The Process of Describing Resources (page 227) and Figure 5.3, The Process of Describing Resources.)

- A collection of resource descriptions is vastly more useful when every resource is described using common description elements or terms that apply to every resource; this specification is most often called a schema or model.

  (See §5.3.1.2 Abstraction in Resource Description (page 232))

- XML schemas are often used to define web forms that capture resource instances, and are also used to describe the interfaces to web services and other computational resources.

  (See §5.3.1.2 Abstraction in Resource Description (page 232))

- When the task of resource description is standardized, the work can be distributed among many describers whose results are shared. This is the principle on which centralized bibliographic description has been based for a century.

  (See §5.3.1.3 Scope, Scale, and Resource Description (page 232))

- Resource description can facilitate the discovery of resources, specify their capabilities and compatibility, authenticate them, and indicate their appraised value.

  (See §5.3.2.1 Resource Description to Support Selection (page 234))

- The Functional Requirements for Bibliographic Records (FRBR) presents four purposes that apply generically: Finding, Identifying, Selecting, and Obtaining resources.

  (See §5.3.2.3 Resource Description to Support Interactions (page 236))

- The variety and functions of the interactions with digital resources depends on the richness of their structural, semantic, and format description.

  (See §5.3.2.3 Resource Description to Support Interactions (page 236))

- *Sensemaking* is the foundation of organizing, as it is the basic human activity of making sense of the world. Sensemaking encompasses the range of organizing activities from the very informal and personal to systematic scientific processes.

  (See §5.3.2.5 Resource Description for Sensemaking and Science (page 238))

- Any particular resource might need many resource descriptions, all of which relate to different properties, depending on the interactions that need to be supported and the context in which they take place.

  (See §5.3.3 Identifying Properties (page 241))

- Two important dimensions for understanding and contrasting resource properties are whether the properties are intrinsically or extrinsically associated with the resource, and whether the properties are static or dynamic.

  (See §5.3.3 Identifying Properties (page 241))

- Recent advances in computing technology and *data science* techniques are making it possible to discover or create resource properties that are called "latent" because they are inferred rather than observed.

  (See the sidebar, Latent Feature Creation and Netflix Recommendations (page 246))

- A *controlled vocabulary* is a fixed or closed set of description terms in some domain with precise definitions that is used instead of the vocabulary that people would otherwise use. A controlled vocabulary reduces synonymy and homonymy.

- Professionally created resource descriptions, author or user created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs.

  (See §5.3.6 Creating Resource Descriptions (page 251))

- The most commonly used criteria for evaluating resource descriptions are accuracy, completeness, and consistency. Other typical criteria are timeliness, interoperability, and usability.

  (See §5.3.7 Evaluating Resource Descriptions (page 254))

- Computational methods can describe and classify images, identify and classify sounds and music, and identify anomalous events in video.

  (See §5.4 Describing Non-text Resources (page 257))

Endnotes for Chapter 5

[228][Com] Most digital cameras use the Exchangeable Image File Format (EXIF). The best source of information about it looks like its Wikipedia entry. `http://en.wikipedia.org/wiki/Exchangeable_image_file_format`.

[229][CogSci] This is much more than just a "kids say the darnedest things" story (see `http://en.wikipedia.org/wiki/Kids_Say_the_Darndest_Things`). Giles Turnbull (Turnbull 2009) noticed that his kids never used the official names for Lego blocks (e.g., Brick 2x2). He then asked other kids what their names were for 32 types of Lego blocks. His survey showed that the kids mostly used different names, but each created names that followed some systematic principles. The most standard name was the "light saber," used by every kid in Turnbull's sample.

[230][Ling] (Reaney and Wilson 1997) classify surnames as local, surnames of relationship, surnames of occupation or office, and nicknames. The dominance of occupational names reflects the fact that there are fewer occupations than places. While there are only a handful of kinship relationships used in surnames (patronymic or father-based names are most common), because the surname includes the father's name there is more variation than for occupations.

[231][Ling] This odd convention is preserved today in wedding invitations, causing some feminist teeth gnashing (Geller 1999).

[232][CogSci] See (Donnellan 1966). A contemporary analysis from the perspective of cognitive science is (Heller, Gorman, and Tanenhaus 2012).

[235][Com] (Rubinsky and Maloney 1997) capture this transitional perspective. A more recent text on XML is (Goldberg 2008).

[236][Com] See (Sen 2004), (Laskey 2005).

[237][IA] See (Marlow, Naaman, Boyd, and Davis 2006). These authors propose a conceptual model of tagging that includes (1) tags assigned to a specific resource, (2) connections or links between resources, and (3) connections or links between users and explain how any two of these can be used to infer information about the other.

[238][IA] (Hammond, Hanney, Lund, and Scott 2004) coined the phrase "tag soup" in an review of social bookmarking tools written early in the tagging era that remains insightful today. Many of the specific tools are no longer around, but the reasons why people tag are still the same.

[239][Web] Making tagging more systematic leads to "tag convergence" in which the distribution of tags for a particular resource stabilizes over time (Golder and Huberman 2006). Consider three things a user might do if his tag does not match the suggested tags; (1) Change the tag to conform? (2) Keep the tag to influence the group norm? (3) Add the proposed tag but keep his tag as well?

[240][Web] (RDF Working Group 2004). The official source for all things RDF is the W3C RDF page at `http://www.w3.org/RDF/`.

[241][Web] Some argue that the resource being described is thus Bart Simpson's Wikipedia page, not Bart Simpson himself. Whether or not that is an important distinction is a controversial question among RDF architects and users.

[242][Web] (Heath and Bizer 2011) and `http://linkeddata.org` are excellent sources.

[244][IA] (Pancake 2012)

[251][Com] Because the relational database schema serves as a model for the creation of resource descriptions, it is designed to restrict the descriptions to be simple and completely regular sets of attribute-value pairs. The database schema specifies the overall structure of the tables and especially their columns, which will contain the attribute values that describe each resource. An employee table might have columns for the attributes of employee ID, hiring date, department, and salary. A date attribute will be restricted to a value that is a date, while an employee salary will be restricted according to salary ranges established by the human resources department. This makes the name of the attribute and the constraints on attribute values into resource descriptions that apply to the entire class of resources described by the table.

It is often necessary to associate some descriptions with individual resources that are specific to that instance and other kinds of descriptions that reflect the abstract class to which the instance belongs. When a typical car comes off the assembly line, it has only one instance-level description that differentiates it from its peers: its vehicle identification number (VIN). Specific cars have individualized interior and exterior colors and installed options, and they all have a

date and location of manufacture. Other description elements have values that are shared with many other cars of the same model and year, like suggested price and the additional option packages, or configurations that can be applied to it before it is delivered to a customer. Alternatively, any descriptive information that applies to multiple cars of the same model year could be part of a resource description at that level that is referred to rather than duplicated in instance descriptions.

[252][Com] Web services are generally implemented using XML documents as their inputs and outputs. The interfaces to web services are typically described using an XML vocabulary called *Web Services Description Language (WSDL)*. See (Erl 2005b), especially Ch. 3, *Introduction to Web Services Technologies.*

[256][CogSci] The semantic "bluntness" of a minimalist vocabulary is illustrated by the examples for use of the "creator" element in an official Dublin Core user guide (Hillmann 2005) that shows "Shakespeare, William" and "Hubble Telescope" as creators.

[257][Com] The Intel Core 2 Duo Processor has detailed specifications (http://www.intel.com/products/processor/core2duo/specifications.htm) and seven categories of technical documentation: application notes, datasheets, design guides, manuals, updates, support components, and white papers (http://www.intel.com/design/core2duo/documentation.htm).

[258][Bus] Real estate advertisements are notorious for their creative descriptions; a house "convenient to transportation" is most likely next to a noisy highway, and a house in a "secluded location" is in a remote and desolate part of town.

[259][Bus] In its early days. when US consumers were generally unaware that Sony was a Japanese company and the quality of Japanese products was viewed in a negative light, Sony would make the "Made in Japan" label as inconspicuous as it could get away with. (John 1999)

In the summer of 2015, the consumer advocacy organization Truth in Advertising reported finding on Walmart's website over 100 product descriptions inaccurately presenting the products as being made in the United States. (See https://www.truthinadvertising.org/walmart-made-in-usa/)

[260][CogSci] Findings from a study of four online dating services (Toma et al 2008) found that 81% of people lied about at least one characteristic. Men were more likely to lie about height, while women lied more about weight, and the further their actual heights and weights were from the mean, they more they lied. A later study (Hall et al 2010) confirmed the finding for women and weight, but also found that men are highly likely to misrepresent their personal assets.

[262][CogSci] In the very busy and dangerous environment of an aircraft carrier flight deck, the sailors wear vests and shirts that are color-coded to their jobs.

For example, red shirts handle munitions, purple shirts handle fuel, green shirts run the catapults and hooks that launch and land the jets, and yellow shirts manage the flights. Color makes it faster and takes less attention for people to see if the right people are where they are supposed to be

The official Navy color chart for aircraft carrier personnel is available at http://www.navy.mil/navydata/ships/carriers/rainbow.asp

A similar principle is used in some sports; goalies wear different color jerseys to make it easier to enforce position-specific rules, and football quarterbacks wear distinctive practice jerseys to remind defensive players not to tackle them and possibly injure them.

It is worth noting that color blindness affects approximately 7% of the population.

[263][Law] The Creative Commons nonprofit organization defines six kinds of copyright licenses that differ in the extent they allow commercial uses or modifications of an original resource (see http://creativecommons.org/licenses/). The Flickr photo sharing application is a good example of a site where a search for reusable resources can use the Creative Commons licenses to filter the results (http://www.flickr.com/creativecommons/).

[264][Bus] Using the same standards to describe products or to specify the execution of business processes can facilitate the implementation and operation of information-intensive business models because information can then flow between services or firms without human intervention. In turn this enables the business to become more demand or event-driven rather than forecast driven, making it a more “adaptive,” “agile,” or “on demand” enterprise. See (Glushko and McGrath 2005), especially Ch. 5, *How Models and Patterns Evolve.*

[265][Web] For new resources, the labor-intensive cost of traditional bibliographic description is less justifiable when you can follow a link from a resource description to the digital resource it describes and quickly decide its relevance. That is, web search engines demonstrate that algorithmic analysis of the content of information resources can make them self-describing to a significant degree, reducing the need for bibliographic description.

[268][Com] Ken Holman’s *Definitive XSLT and XPath* (Holman 2001) is the book to get started on with XPath, and no one has taught more people about XPath than Holman. The first five hours of a 24-hour video course on *Practical Transformation Using XSLT and XPath* is available for free at http://www.udemy.com/practical-transformation-using-xslt-and-xpath.

[270][DS] (Lockyer 1893) and (Bell 1970)

The joint story of Brahe’s data collecting and Kepler’s analysis and theorizing is told in an entertaining manner in (Ferguson 2002). An equally fascinating analy-

sis that interprets Kepler's conceptual shifts with a model of analogical reasoning is (Gentner et al., 1997).

[271][Bus] The concept of *sensemaking* originated from business school research in management and organizational theory (Weick 1995) but has been widely employed by ethnographers in many contexts including emergency rooms, classrooms with minority students, airline safety inspections, and crime investigation. See (Weick 2005) and (Chater 2016)

[272][Phil] Occam's Razor has a long tradition in scientific philosophy, but some people have argued that it is overrated as a heuristic for choosing among alternative explanations or theories, particularly because it depends on how you define simplicity.

[273][Com] One way to make simplicity more useful as a guide for choosing between mathematical models is to explicitly penalize those that are more complex by adding error to the predictions, a technique that computer scientists have given the non-intuitive name of regularization. This penalty requires complex models to be significantly better at explaining the data than simpler ones because they have to overcome the added errors.

[274][CogSci] For example, the composition of a chair is presented here as a static intrinsic property but in fact a wooden chair might deteriorate over time as a result of exposure to sunlight, heat, or biological agents that attack it. A skill can be considered intrinsic and dynamic, but it might also be highly dependent on context, making it extrinsic. The subject category assigned to a book is extrinsic and static, but if the classification system is revised the book might be reclassified. Finally, while the location of a resource can be extrinsic and dynamic, the location history of the resource at some specific point in time is a fact, an intrinsic and static property.

[275][Com] (Dey 2001) further defines the "environment" of *context* as places, people, and things, and for each of "entities" there are four categories of context information: location, identity, status (or activity), and time. This *framework* thus yields 12 dimensions for describing the context of an environment.

[276][Com] A fascinating story about the Netflix's design and use of tens of thousands of movie sub-genres in its recommendation system is (Madrigal 2014).

[277][Ling] Consider how many events are named by appending a "-gate" suffix to imply that there is something scandalous or unethical going on that is being covered up. This cultural description is not immediately meaningful to anyone who does not know about the break-in at the headquarters of the Democratic National Committee headquarters at the Watergate hotel and subsequent cover-up that led to the 1974 resignation of US President Richard Nixon. A list of "-gate" events is maintained at `http://en.wikipedia.org/wiki/List_of_scandals_with_%22-gate%22_suffix`.

[278][Ling] (http://en.wikipedia.org/wiki/Holbein_carpet).

[280][Com] (Laskey 2005).

[282][DS] We cannot cite all of mathematical statistics in one short endnote, but if you are inclined to learn more, (Mardia, Kent, and Bibby 1980) and (Lee and Verleysen 2007) are the kindest and gentlest resources. If we look very generously at "dimensionality reduction" we might even consider the indexing step of eliminating "stop words" to be a form of dimensionality reduction. Stop words appear with such high frequency that they have no discriminating power, so they are discarded from queries and not part of the description of the indexed documents.

[285][Bus] Many institutional organizing systems are subject to a single centralized or governmental authority that can impose principles for describing and arranging resources. Examples of organizing systems where resources are described using standard centralized principles are:

Companies that follow industry standards for information or process models, product classification or identification to be eligible for government business (Shah and Kesan 2006).

Legislative documents that conform to National or European Community standards for structure, naming, and description (Biasiotti 2008).

The Internet Corporation for Assigned Names and Numbers (ICANN) and its policies for operating the Domain Name System (DNS) make it possible for every website to be located using its logical name (like "berkeley.edu" rather than using an IP address like 169.229.131.81). (http://www.icann.org/)

In other domains multiple organizations or institutions have the authority to impose principles of resource description. Sometimes this authority derives from the voluntary collaboration of multiple autonomous parties who set and conform to standards because they benefit from being able to share resources or information about resources. Examples of organizing systems where resources are described using standardized decentralized principles are:

Firms that establish company-wide standards for their information resources, typically including the organization and management of source content, document type models, and a style guide that applies to print and web documents.

Firms that participate in the OASIS (http://www.oasis-open.org/) or the W3C (http://www.w3.org/) industry consortia to establish specifications or technical recommendations for their information systems or web services).

[287][Web] (Sen et al. 2006) analyze the effects of four tag selection algorithms used in sites that allow user tags on vocabulary evolution (more often called

"tag convergence" in the literature), tag utility, tag adoption, and user satisfaction.

[288][CogSci] But in an often-cited essay (Doctorow, 2001) provocatively titled "Metacrap: Putting the torch to seven straw-men of the meta-utopia," Cory Doctorow argues that much human-created metadata *is* of low quality because "people lie, people are lazy, people are stupid, mission impossible—know thyself, schemas are not neutral, metrics influence results, (and) there is more than one way to describe something."

[289][Com] (Hu and Lui 2004).

[290][Com] `http://www.amazon.com/gp/search-inside/sipshelp.html/`.

[291][DS] The title says it all: *Predictive analytics: The power to predict who will click, buy, lie, or die.* (Siegel 2013). The Bayesian Surname and Geocoding technique for predicting race is described by (Elliott et al. 2008).

[293][Bus] However, these concerns are rapidly becoming more important in the public sector. In particular, many public universities in the US are struggling with cuts in state and federal funding that are affecting library services and practices.

[294][Bus] More generally, economists use the concept of the "mode of exchange" in a business relationship to include the procedures and norms that govern routine behavior between business partners. An "exit" mode is one in which the buyer makes little long-term commitment to a supplier, and problems with a supplier cause the buyer to find a new one. In contrast, in "voice" mode there is much greater commitment and communication between the parties, usually leading to improved processes and designs. See (Helper and McDuffie 2003).

[296][Com] (Nunberg 2009) called the quality of Google's metadata "a disaster for scholars," but (Sag 2012) argues that the otherwise neglected "orphan works" in the Google corpus are "grist for the data mill."

[297][Bus] The modern "quality movement" grew out of the efforts of the US to rebuild Japan after the Second World War and its "Bible" was Juran's 1951 *Quality Control Handbook* (Juran 1951).

[300][Com] See (Datta et al. 2008). The company Idée is developing a variety of image search algorithms, which use image signatures and measures of visual similarity to return photos similar to those a user asks to see.

[301][Web] (von Ahn and Dabbish 2008).

[302][Com] The key idea that made deep learning possible is the use of "backpropagation" to adjust the weights on features by working backwards from the output (the object classification produced by the network) all the way back to the input. Mathematically-sophisticated readers can find a concise explanation and history

of deep learning in (LeCun, Bengio, and Hinton 2015). LeCun and Hinton were part of research teams that independently invented backpropagation in the mid 1980s. Today, LeCun heads Facebook's research group on artificial intelligence, and Hinton has a similar role at Google.

[303][Com] (Cano et al. 2005).

[304][Com] (Walker 2009).

[307][Com] (Regazzoni et al. 2010) introduce a special issue in IEEE *Signal Processing on visual analytics.*

[308][Bus] One organization that sees a future in assembling better descriptions of video content is the United States' National Football League (NFL), whose vast library of clips can not only be used to gather plays for highlight reels and specials but can also be monetized by pointing out when key advertisers' products appear on film. Currently, labeling the video requires a person to watch the scenes and tag elements of each frame, but once those tags have been created and sequenced along with the video, they can be more easily searched in computerized, automated ways (Buhrmester 2007).

Chapter 6

Describing Relationships and Structures

Robert J. Glushko
Matthew Mayernik
Alberto Pepe
Murray Maloney

6.1 Introduction

We consider a family to be a collection of people affiliated by some connections, such as common ancestors or a common residence. The Simpson family includes a man named Homer and a woman named Marge, the married parents of three sibling children, a boy named Bart and two girls, Lisa and Maggie. This magical family speaks many languages, but most often uses the language of the local television station. In the English-speaking Simpson family, the boy describes his parents as his father and mother and his two siblings as his sisters. In the Spanish speaking Simpson family he refers to his parents as *su padre y su madre* and his sisters are *las hermanas*. In the Chinese Simpson family the sisters refer to each other according to their relative ages; Lisa, the elder, as *jiě jie* and, Maggie, the younger, as *mèi mei*.[309][Bus]

Kinship relationships are ubiquitous and widely studied, and the names and significance of kinship relations like "is parent of" or "is sibling of" are familiar ones, making kinship a good starting point for understanding *relationships* in organizing systems.[310][CogSci] An organizing system can make use of existing relationships among resources, or it can create relationships by applying organizing principles to arrange the resources. Organizing systems for digital resources or digital description resources are the most likely to rely on explicit relationships to enable interactions with the resources.

Simpson Family Trees

Because the Simpson family is known throughout the world, the Simpson family tree is often used to teach kinship terms to language learners.

- A website for teaching Spanish
- A website for teaching French
- A website for teaching German

In a classic book called *Data and Reality,* William Kent defines a *relationship* as an association among several things, with that association having a particular significance.[311][Com] "The things being associated," the components of the relationship, are people in kinship relationships but more generally can be any type of resource (Chapter 4), when we relate one resource instance to another. When we describe a resource (Chapter 5), the components of the relationship are a primary resource and a description resource. If we specify sets of relationships that go together, we are using these common relationships to define resource types or classes, which more generally are called categories (Chapter 7). We can then use resource types as one or both the components of a relationship when we want to further describe the resource type or to assert how two resource types go together to facilitate our interactions with them.

We begin with a more complete definition of relationship and introduce five perspectives for analyzing them: semantic, lexical, structural, architectural, and implementation. We then discuss each perspective, introducing the issues that each emphasizes, and the specialized vocabulary needed to describe and analyze relationships from that point of view. We apply these perspectives and vocabulary to analyze the most important types of relationships in organizing systems.

6.2 Describing Relationships: An Overview

The concept of a relationship is pervasive in human societies in both informal and formal senses. Humans are inescapably related to generations of ancestors, and in most cases they also have social networks of friends, co-workers, and casual acquaintances to whom they are related in various ways. We often hear that our access to information, money, jobs, and political power is all about "who you know," so we strive to "network" with other people to build relationships that might help us expand our access. In information systems, relationships between resources embody the organization that enables finding, selection, retrieval, and other interactions.

Most organizing systems are based on many relationships to enable the system to satisfy some intentional purposes with individual resources or the collection as a whole. In the domain of information resources, common resources include web pages, journal articles, books, datasets, metadata records, and XML documents, among many others. Important relationships in the information domain that facilitate purposes like finding, identifying, and selecting resources include "is the author of," "is published by," "has publication date," "is derived from," "has subject keyword," "is related to," and many others.

When we talk about relationships we specify both the resources that are associated along with a name or statement about the reason for the association. Just identifying the resources involved is not enough because several different relationships can exist among the same resources; the same person can be your brother, your employer, and your landlord. Furthermore, for many relationships the *directionality* or ordering of the participants in a relationship statement matters; the person who is your employer gives a paycheck to you, not vice versa. Kent points out that when we describe a relationship we sometimes use whole phrases, such as "is-employed-by," if our language does not contain a single word that expresses the meaning of the relationship.

Navigating This Chapter

In this chapter, we analyze relationships from several different perspectives:

Semantic perspective
: The semantic perspective is the most essential one; it characterizes the meaning of the association between resources. (§6.3)

Lexical perspective
: The lexical perspective focuses on how the conceptual description of a relationship is expressed using words in a specific language. (§6.4)

Structural perspective
: The structural perspective analyzes the actual patterns of association, arrangement, proximity, or connection between resources. (§6.5)

Architectural perspective
: The architectural perspective emphasizes the number and abstraction level of the components of a relationship, which together characterize its complexity. (§6.6)

Implementation perspective
: The implementation perspective considers how the relationship is implemented in a particular notation and syntax and the manner in which relationships are arranged and stored in some technology environment. (§6.7)

6.3 The Semantic Perspective

To describe relationships among resources, we need to understand what the relations mean. This *semantic perspective* is the essence of relationships and explains why the resources are related, relying on information that is not directly available from perceiving the resources. In our Simpson family example, we noted that Homer and Marge are related by marriage, and also by their relationship as parents of Bart, Lisa, and Maggie, and none of these relationships are directly perceivable. This means that "Homer is married to Marge" is a semantic assertion, but "Homer is standing next to Marge" is not.[312 [CogSci]]

Semantic relationships are commonly expressed with a predicate with one or more arguments. A *predicate* is a verb phrase template for specifying properties of objects or a relationship among objects. In many relationships the predicate is an action or association that involves multiple participants that must be of

particular types, and the arguments define the different roles of the participants.[313][CogSci]

We can express the relationship between Homer and Marge Simpson using a *predicate(argument(s))* syntax as follows:

is-married-to (Homer Simpson, Marge Simpson)

The sequence, type, and role of the arguments are an essential part of the relationship expression. The sequence and role are explicitly distinguished when predicates that take two arguments are expressed using a *subject-predicate-object* syntax that is often called a *triple* because of its three parts:

Homer Simpson → is-married-to → Marge Simpson

However, we have not yet specified what the "is-married-to" relationship means. People can demonstrate their understanding of "is-married-to" by realizing that alternative and semantically equivalent expressions of the relationship between Homer and Marge might be:

Homer Simpson → is-married-to → Marge Simpson
Homer Simpson → is-the-husband-of → Marge Simpson
Marge Simpson → is-married-to → Homer Simpson
Marge Simpson → is-the-wife-of → Homer Simpson

Going one step further, we could say that people understand the equivalence of these different expressions of the relationship because they have semantic and linguistic knowledge that relates some representation of "married," "husband," "wife," and other words. None of that knowledge is visible in the expressions of the relationships so far, all of which specify concrete relationships about individuals and not abstract relationships between resource classes or concepts. We have simply pushed the problem of what it means to understand the expressions into the mind of the person doing the understanding.

We can be more rigorous and define the words used in these expressions so they are "in the world" rather than just "in the mind" of the person understanding them. We can write definitions about these resource classes:

- The conventional or traditional marriage relationship is a consensual lifetime association between a husband and a wife, which is sanctioned by law and often by religious ceremonies;
- A husband is a male lifetime partner considered in relation to his wife; and
- A wife is a female lifetime partner considered in relation to her husband.[314][Law]

Definitions like these help a person learn and make some sense of the relationship expressions involving Homer and Marge. However, these definitions are not in a form that would enable someone to completely understand the Homer and Marge expressions; they rely on other undefined terms (consensual, law, lifetime, etc.), and they do not state the relationships among the concepts in the definitions.[315][CogSci] Furthermore, for a computer to understand the expressions, it needs a computer-processable representation of the relationships among words and meanings that makes every important semantic assumption and property precise and explicit. We will see what this takes starting in the next section.

6.3.1 Types of Semantic Relationships

In this discussion we will use *entity type, class, concept,* and *resource type* as synonyms. *Entity type* and *class* are conventional terms in data modeling and database design, *concept* is the conventional term in computational or cognitive modeling, and we use *resource type* when we discuss organizing systems. Similarly, we will use *entity occurrence, instance,* and *resource instance* when we refer to one thing rather than to a class or type of them.

There is no real consensus on how to categorize semantic relationships, but these three broad categories are reasonable for our purposes:

Inclusion Relationship
: One entity type contains or is comprised of other entity types; often expressed using "is-a," "is-a-type-of," "is-part-of," or "is-in" predicates.

Attribution Relationship
: Asserting or assigning values to properties; the predicate depends on the property: "is-the-author-of," "is-married-to," "is-employed-by," etc.

Possession Relationship
: Asserting ownership or control of a resource; often expressed using a "has" predicate, such as "has-serial-number-plate."[316][CogSci]

All of these are fundamental in organizing systems, both for describing and arranging resources themselves, and for describing the relationships among resources and resource descriptions.

6.3.1.1 Inclusion

There are three different types of inclusion relationships: *class inclusion*, meronymic inclusion, and topological inclusion. All three are commonly used in organizing systems.

Class inclusion is the fundamental and familiar "**is-a,**" "**is-a-type-of**," or "**subset**" relationship between two entity types or classes where one is contained in and thus more specific than the other more generic one.

Meat → is-a → Food

A set of interconnected class inclusion relationships creates a hierarchy, which is often called a *taxonomy*.

Meat → is-a → Food
Dairy Product → is-a → Food
Cereal → is-a → Food
Vegetable → is-a → Food
Beef → is-a → Meat
Pork → is-a → Meat
Chicken → is-a → Meat
Ground Beef → is-a → Beef
Steak → is-a → Beef
...

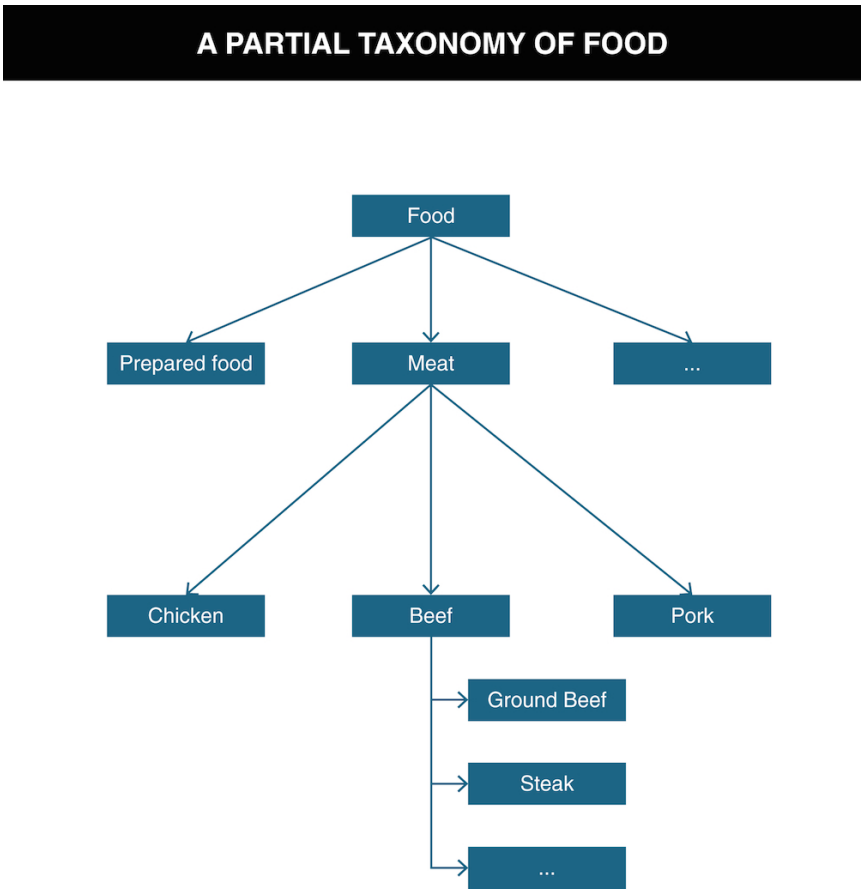
A visual depiction of the taxonomy makes the class hierarchy easier to perceive. See Figure 6.1, A Partial Taxonomy of Food.

Each level in a taxonomy subdivides the class above it into sub-classes, and each sub-class is further subdivided until the differences that remain among the members of each class no longer matter for the interactions the organizing system needs to support. We discuss the design of hierarchical organizing systems in §7.3, "Principles for Creating Categories."

All of the examples in the current section have expressed abstract relationships between classes, in contrast to the earlier concrete ones about Homer and Marge, which expressed relationships between specific people. Homer and Marge are instances of classes like "married people," "husbands," and "wives." When we make an assertion that a particular instance is a member of class, we are *classifying* the instance. *Classification* is a class inclusion relationship between an instance and a class, rather than between two classes. (We discuss Classification in detail in Chapter 8.)

Homer Simpson → is-a → Husband

Figure 6.1. A Partial Taxonomy of Food.



A partial taxonomy of food distinguishes the categories or prepared food from meat, distinguishes chicken, beef, and pork as subcategories of meat, and distinguishes ground beef and steak as subcategories of beef.

This is just the lowest level of the class hierarchy in which Homer is located at the very bottom; he is also a man, a human being, and a living organism (in cartoon land, at least).[317][CogSci] You might now remember the bibliographic class inclusion hierarchy we discussed in §4.3.2; a specific physical *item* like your dog-eared copy of *Macbeth* is also a particular *manifestation* in some format or genre, and this *expression* is one of many for the abstract *work*.

instance→ is-member-of → class

Part-whole inclusion or *meronymic inclusion* is a second type of inclusion relationship. It is usually expressed using "is-part-of," "is-partly," or with other similar predicate expressions. Winston, Chaffin, and Herrmann identified six distinct types of part-whole relationships. Their meaning subtly differs depending on whether the part is separately identifiable and whether the part is essential to the whole.[318[CogSci]]

- *Component-Object* is the relationship type when the part is a separate component that is arranged or assembled with other components to create a larger resource. In §4.1.1.1, "Resources with Parts," we used as an example the component-object relationship between an engine and a car:

  The Engine → is-part-of → the Car

  The components of this type of part-whole relationship need not be physical objects; "Germany is part of the European Union" expresses a component-object relationship. What matters is that the component is identifiable on its own as an integral entity and that the components follow some kind of patterned organization or structure when they form the whole. Together the parts form a composition, and the parts collectively form the whole. A car that lacks the engine part will not work.

- *Member-Collection* is the part-whole relationship type where "is-part-of" means "belongs-to," a weaker kind of association than component-object because there is no assumption that the component has a specific role or function in the whole.

  The Book → is-part-of → the Library

  The members of the collection exist independently of the whole; if the whole ceases to exist the individual resources still exist.

- *Portion-Mass* is the relationship type when all the parts are similar to each other and to the whole, unlike either of the previous types where engines are not tires or cars, and books are not like record albums or libraries.

  The Slice → is-part-of → the Pie

- *Stuff-Object* relationships are most often expressed using "is-partly" or "is-made-of" and are distinguishable from component-object ones because the stuff cannot be separated from the object without altering its identity. The stuff is not a separate ingredient that is used to make the object; it is a constituent of it once it is made.

  Wine → is-partly → Alcohol

- *Place-Area* relationships exist between areas and specific places or locations within them. Like members of collections, places have no particular functional contribution to the whole.

  The Everglades → are-part-of → Florida

- *Feature-Activity* is a relationship type in which the components are stages, phases, or sub activities that take place over time. This relationship is similar to component-object in that the components in the whole are arranged according to a structure or pattern.

  Overtime → is-part-of → a Football Game

A seventh type of part-whole relationship called *Phase-Activity* was proposed by Storey.[319][CogSci]

- *Phase-Activity* is similar to *feature-activity* except that the phases do not make sense as standalone activities without the context provided by the activity as a whole.

  Paying → is-part-of → Shopping

Topological, *Locative* and *Temporal Inclusion* is a third type of *inclusion relationship* between a container, area, or temporal duration and what it surrounds or contains. It is most often expressed using "is-in" as the relationship. However, the entity that is contained or surrounded is not a part of the including one, so this is not a *part-whole* relationship.

The Vatican City → is-in → Italy
The meeting → is-in → the afternoon

6.3.1.2 Attribution

In contrast to inclusion expressions that state relationships between resources, *attribution relationships* assert or assign values to properties for a particular resource. In Chapter 5 we used "attribute" to mean "an indivisible part of a resource description" and treated it as a synonym of "property." We now need to be more precise and carefully distinguish between the type of the *attribute* and the *value* that it has. For example, the color of any object is an *attribute* of the object, and the *value* of that attribute might be "green."

Some frameworks for semantic modeling define "attribute" very narrowly, restricting it to expressions with predicates with only one argument to assert properties of a single resource, distinguishing them from relationships between resources or resource types that require two arguments:[320][Bus]

Martin the Gecko → is-small
Martin the Gecko → is-green

However, it is always possible to express statements like these in ways that make them into relationships with two arguments:

Martin → has-size → small
Martin → has-skin-color → green

Dedre Gentner notes that this supposed distinction between one-predicate attributes and two-predicate relationships depends on context.[321][CogSci] For example, small can be viewed as an attribute, **X → is-small,** or as a relationship between X and some standard or reference Y, **X → is-smaller-than → Y.**

Another somewhat tricky aspect of attribution relationships is that from a semantic perspective, there are often many different ways of expressing equivalent attribute values.

Martin → has-size → 6 inches
Martin → has size → 152 mm

These two statements express the idea that Martin is small. However, many implementations of attribution relationships treat the attribute values literally. This means that unless we can process these two statements using another relationship that expresses the conversion of inches to mm, the two statements could be interpreted as saying different things about Martin's size.

Finally, we note that we can express attribution relationships about other relationships, like the date a relationship was established. Homer and Marge Simpson's wedding anniversary is an attribute of their "is-married-to" relationship.

The semantic distinctions between attributes and other types of relationships are not strong ones, but they can be made clearer by implementation choices. For example, XML attributes are tightly coupled to a containing element, and their literal values are limited to atomic items of information. In contrast, inclusion relationships are expressed by literal containment of one XML element by another.

6.3.1.3 Possession

A third distinct category of semantic relationships is that of possession. *Possession* relationships can seem superficially like part-whole ones:

Bob → has → a car
A car → has → wheels

However, in the second of these relationships "has" is an elliptical form of "has as a part," expressing a part-whole relationship rather that one of possession.

The concept of possession is especially important in institutional organizing systems, where questions of ownership, control, responsibility and transfers of ownership, control, and responsibility can be fundamental parts of the interactions they support. However, possession is a complex notion, inherently connected to societal norms and conventions about property and kinship, making it messier than institutional processes might like.

Possession relationships also imply duration or *persistence*, and are often difficult to distinguish from relationships based on habitual location or practice. Miller and Johnson-Laird illustrate the complex nature of possession relationships with this sentence, which expresses three different types of them:[322][Ling]

> **He owns an umbrella but she's borrowed it, though she doesn't have it with her.**

6.3.2 Properties of Semantic Relationships

Semantic relationships can have numerous special properties that help explain what they mean and especially how they relate to each other. In the following sections we briefly explain those that are most important in systems for organizing resources and resource descriptions.

6.3.2.1 Symmetry

In most relationships the order in which the subject and object arguments are expressed is central to the meaning of the relationship. If X has a relationship with Y, it is usually not the case that Y has the same relationship with X. For example, because "is-parent-of" is an *asymmetric relationship*, only the first of these relationships holds:

Homer Simpson → is-parent-of → Bart Simpson (`TRUE`)
Bart Simpson → is-parent-of → Homer Simpson (`NOT TRUE`)

In contrast, some relationships are *symmetric* or *bi-directional*, and reversing the order of the arguments of the relationship predicate does not change the meaning. As we noted earlier, these two statements are semantically equivalent because "is-married-to" is symmetric:

Homer Simpson → is-married-to → Marge Simpson
Marge Simpson → is-married-to → Homer Simpson

We can represent the *symmetric* and *bi-directional* nature of these relationships by using a double-headed arrow:

Homer Simpson ⇔ is-married-to ⇔ Marge Simpson

6.3.2.2 Transitivity

Transitivity is another property that can apply to semantic relationships. When a relationship is transitive, if X and Y have a relationship, and Y and Z have the same relationship, then X also has the relationship with Z. Any relationship based on ordering is transitive, which includes numerical, alphabetic, and chronological ones as well as those that imply qualitative or quantitative measurement. Because "is-taller-than" is transitive:

Homer Simpson → is-taller-than → Bart Simpson
Bart Simpson → is-taller-than → Maggie Simpson

implies that:

Homer Simpson → is-taller-than → Maggie Simpson

Inclusion relationships are inherently transitive, because just as "is-taller-than" is an assertion about relative physical size, "is-a-type of" and "is-part-of" are assertions about the relative sizes of abstract classes or categories. An example of transitivity in part-whole or meronymic relationships is: (1) the carburetor is part of the engine, (2) the engine is part of the car, (3) therefore, the carburetor is part of the car. 323[Ling]

Transitive relationships enable inferences about class membership or properties, and allow organizing systems to be more efficient in how they represent them since transitivity enables implicit relationships to be made explicit only when they are needed.

6.3.2.3 Equivalence

Any relationship that is both symmetric and transitive is an *equivalence relationship*; "is-equal-to" is obviously an equivalence relationship because if A=B then B=A and if A=B and B=C, then A=C. Other relationships can be equivalent without meaning "exactly equal," as is the relationship of "is-congruent-to" for all triangles.

We often need to assert that a particular class or property has the same meaning as another class or property or that it is generally substitutable for it. We make this explicit with an equivalence relationship.

Sister (English) ⇔ is-equivalent-to ⇔ Hermana (Spanish)

6.3.2.4 Inverse

For asymmetric relationships, it is often useful to be explicit about the meaning of the relationship when the order of the arguments in the relationship is reversed. The resulting relationship is called the *inverse* or the converse of the first relationship. If an organizing system explicitly represents that:

Is-child-of → is-the-inverse-of → Is-parent-of

We can then conclude that:

Bart Simpson → is-child-of → Homer Simpson

6.3.3 Ontologies

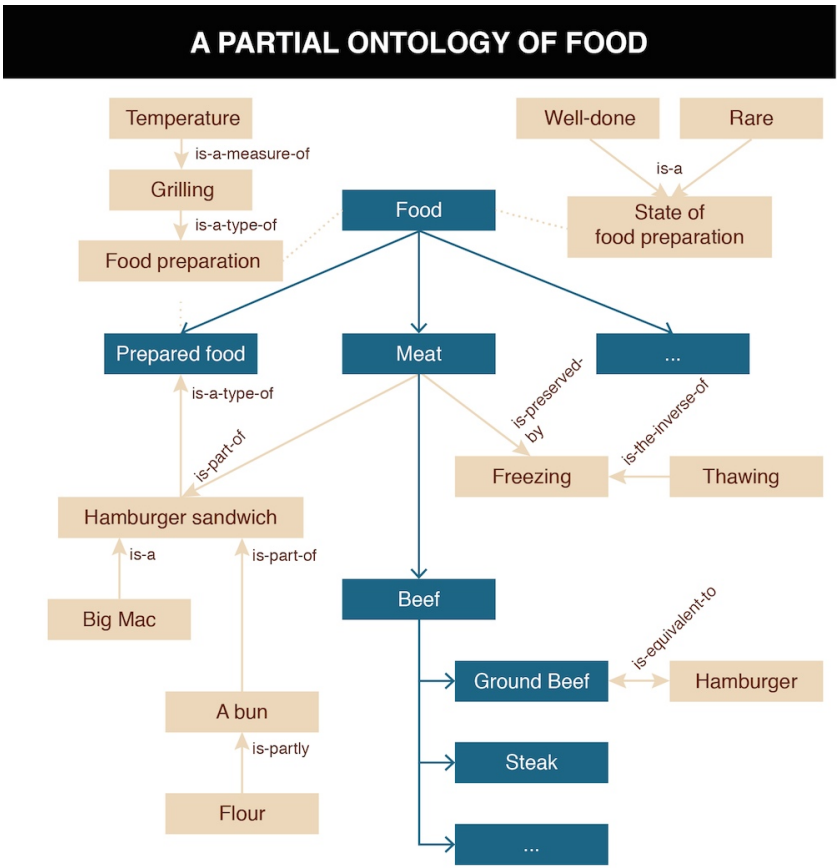
We now have described types and properties of semantic relationships in enough detail to return to the challenge we posed earlier: what information is required to fully understand relationships? This question has been asked and debated for decades and we will not pretend to answer it to any extent here. However, we can sketch out some of the basic parts of the solution.

Let us begin by recalling that a *taxonomy* captures a system of class inclusion relationships in some domain. But as we have seen, there are a great many kinds of relationships that are not about class inclusion. All of these other types of relationships represent knowledge about the domain that is potentially needed to understand statements about it and to make sense when more than one domain of resources or activities comes together.

For example, in the food domain whose partial taxonomy appears in Figure 6.2, A Partial Ontology of Food., we can assert relationships about properties of classes and instances, express equivalences about them, and otherwise enhance the representation of the food domain to create a complex network of relationships. In addition, the food domain intersects with food preparation, agriculture, commerce, and many other domains. We also need to express the relationships among these domains to fully understand any of them.

Grilling → is-a-type-of → Food Preparation
Temperature → is-a-measure-of → Grilling
Hamburger → is-equivalent-to → Ground Beef
Hamburger → is-prepared-by → Grilling
Hamburger Sandwich → is-a-type-of → Prepared Food
Rare → is-a → State of Food Preparation
Well-done → is-a → State of Food Preparation
Meat → is-preserved-by → Freezing
Thawing → is-the-inverse-of → Freezing
...

Figure 6.2. A Partial Ontology of Food.



A partial ontology of food overlays the taxonomy of food with statements that make assertions about categories, instances, and relationships in the food domain. Example statements might be that "Grilling is a type of food preparation," that"Meat is preserved by freezing," and that "Hamburger is equivalent to ground beef."

In this simple example we see that class inclusion relationships form a kind of backbone to which other kinds of relationships attach. We also see that there are many potentially relevant assertions that together represent the knowledge that just about everyone knows about food and related domains. A network of relationships like these creates a resource that is called an *ontology*.[324][Phil] A visual depiction of the ontology illustrates this idea that it has a taxonomy as its conceptual scaffold. (See Figure 6.2, A Partial Ontology of Food.)

There are numerous formats for expressing ontologies, but many of them have recently converged to or are based on the *Web Ontology Language (OWL)*, developed by the W3C. OWL ontologies use a formal logic-based language that builds on RDF (§5.2.2.4 Resource Description Framework (RDF) (page 223)) to define resource classes and assign properties to them in rigorous ways, arrange them in a class hierarchy, establish their equivalence, and specify the properties of relationships.[325][Web]

Ontologies are essential parts in some organizing systems, especially information-intensive ones where the scope and scale of the resources require an extensive and controlled description vocabulary. (See §5.3 The Process of Describing Resources (page 227).) The most extensive ontology ever created is Cyc, born in 1984 as an artificial intelligence research project. Three decades later, the latest version of the Cyc ontology contains several hundred thousand terms and millions of assertions that interrelate them.[326][Com]

6.4 The Lexical Perspective

The semantic perspective for analyzing relationships is the fundamental one, but it is intrinsically tied to the lexical one because a relationship is always expressed using words in a specific language. For example, we understand the relationships among the concepts or classes of "food," "meat," and "beef" by using the words "food," "meat," and "beef" to identify progressively smaller classes of edible things in a class hierarchy.

The connection between concept and words is not so simple. In the Simpson family example with which we began this chapter, we noted with "father" and "padre" that languages differ in the words they use to describe particular kinship relationships. Furthermore, we pointed out that cultures differ in which kinship relationships are conceptually distinct, so that languages like Chinese make distinctions about the relative ages of siblings that are not made in English.[327][Ling]

This is not to suggest that an English speaker cannot notice the difference between his older and younger sisters, only that this distinction is not lexicalized—captured in a single word—as it is in Chinese. This "missing word" in English from the perspective of Chinese is called a *lexical gap*. Exactly when a lexical gap exists is sometimes tricky, because it depends on how we define "word"—polar bear and sea horse are not lexicalized but they are a single meaning-bearing unit because we do not decompose and reassemble meaning from the two separate words. These "lexical gaps" differ from language to language, whereas "conceptual gaps"—the things we cannot think of or directly experience, like the pull of gravity— may be innate and universal. We revisit this issue as "linguistic relativity" in Chapter 7. [328][Ling]

Earlier in this book we discussed the naming of resources (§4.4.2 The Problems of Naming (page 188)) and the design of a vocabulary for resource description (§5.3.1.3 Scope, Scale, and Resource Description (page 232)), and we explained how increasing the scope and scale of an organizing system made it essential to be more systematic and precise in assigning names and descriptions. We need to be sure that the terms we use to organize resources capture the similarities and differences between them well enough to support our interactions with them. After our discussion about semantic relationships in this chapter, we now have a clearer sense of what is required to bring like things together, keep different things separate, and to satisfy any other goals for the organizing system.

For example, if we are organizing cars, buses, bicycles, and sleds, all of which are vehicles, there is an important distinction between vehicles that are motorized and those that are powered by human effort. It might also be useful to distinguish vehicles with wheels from those that lack them. Not making these distinctions leaves an unbalanced or uneven organizing system for describing the semantics of the vehicle domain. However, only the "motorized" concept is lexicalized in English, which is why we needed to invent the "wheeled vehicle" term in the second case.[329][Ling]

Simply put, we need to use words effectively in organizing systems. To do that, we need to be careful about how we talk about the relationships among words and how words relate to concepts. There are two different contexts for those relationships.

- First, we need to discuss relationships among the meanings of words. (§6.4.1) and the most commonly used tool for describing them (§6.4.2).
- Second, we need to discuss relationships among the form of words. (§6.4.3 Relationships among Word Forms (page 293))

6.4.1 Relationships among Word Meanings

There are several different types of relationships of word meanings. Not surprisingly, in most cases they parallel the types of relationships among concepts that we described in §6.3 The Semantic Perspective (page 276).

6.4.1.1 Hyponymy and Hyperonymy

When words encode the semantic distinctions expressed by class inclusion, the word for the more specific class in this relationship is called the *hyponym*, while the word for the more general class to which it belongs is called the *hypernym*. George Miller suggested an exemplary formula for defining a hyponym as its hypernym preceded by adjectives or followed by relative clauses that distinguish it from its *co-hyponyms*, mutually exclusive subtypes of the same hypernym.

hyponym = {adjective+} hypernym {distinguishing clause+}

For example, robin is a hyponym of bird, and could be defined as "a migratory bird that has a clear melodious song and a reddish breast with gray or black upper plumage." This definition does not describe every property of robins, but it is sufficient to differentiate robins from bluebirds or eagles.[330][Ling]

6.4.1.2 Metonymy

Part-whole or meronymic semantic relationships have lexical analogues in *metonomy*, when an entity is described by something that is contained in or otherwise part of it. A country's capital city or a building where its top leaders reside is often used as a metonym for the entire government: "The White House announced today..." Similarly, important concentrations of business activity are often metonyms for their entire industries: "Wall Street was bailed out again..."

6.4.1.3 Synonymy

Synonymy is the relationship between words that express the same semantic concept. The strictest definition is that *synonyms* "are words that can replace each other in some class of contexts with insignificant changes of the whole text's meaning."[331][Ling] This is an extremely hard test to pass, except for acronyms or compound terms like "USA," "United States," and "United States of America" that are completely substitutable.

Most synonyms are not *absolute synonyms*, and instead are considered *propositional synonyms*. *Propositional synonyms* are not identical in meaning, but they are equivalent enough that substituting one for the other will not change the truth value of the sentence. This weaker test lets us treat word as synonyms even though their meanings subtly differ. For example, if Lisa Simpson can play the violin, then because "violin" and "fiddle" are propositional synonyms, no one would disagree with an assertion that Lisa Simpson can play the fiddle.

An unordered set of synonyms is often called a *synset*, a term first used by the WordNet "semantic dictionary" project started in 1985 by George Miller at Princeton.[332][CogSci] Instead of using spelling as the primary organizing principle

for words, WordNet uses their semantic properties and relationships to create a network that captures the idea that words and concepts are an inseparable system. Synsets are interconnected by both semantic relationships and lexical ones, enabling navigation in either space.[333][Bus]

6.4.1.4 Polysemy

We introduced the lexical relationship of *polysemy*, when a word has several different meanings or senses, in the context of problems with names (§4.4.2.2 Homonymy, Polysemy, and False Cognates (page 189)). For example, the word "bank" can refer to a: river bank, money bank, bank shots in basketball and billiards, an aircraft maneuver, and other concepts.[334][Ling]

Polysemy is represented in WordNet by including a word in multiple synsets. This enables WordNet to be an extremely useful resource for sense disambiguation in natural language processing research and applications. When a polysemous word is encountered, it and the words that are nearby in the text are looked up in WordNet. By following the lexical relationships in the synset hierarchy, a "synset distance" can be calculated. The smallest semantic distance between the words, which identifies their most semantically specific hypernym, can be used to identify the correct sense. For example, in the sentence:

Put the money in the bank

Two of the three WordNet senses for "money" are:

1) the most common medium of exchange
2) the official currency issued by a government or national bank

and the first two of the ten WordNet senses for "bank" are:

1) a financial institution that accepts deposits
2) sloping land, especially the slope beside a body of water

The synset hierarchies for the two senses of "money" intersect after a very short path with the hierarchy for the first sense of "bank," but do not intersect with the second sense of "bank" until they reach very abstract concepts.[335][Com]

6.4.1.5 Antonymy

Antonymy is the lexical relationship between two words that have opposite meanings. *Antonymy* is a very salient lexical relationship, and for adjectives it is even more powerful than synonymy. In word association tests, when the probe word is a familiar adjective, the most common response is its antonym; a probe of "good" elicits "bad," and vice versa. Like synonymy, antonymy is sometimes exact and sometimes more graded.[336][Ling]

Contrasting or *binary antonyms* are used in mutually exclusive contexts where one or the other word can be used, but never both. For example, “alive” and “dead” can never be used at the same time to describe the state of some entity, because the meaning of one excludes or contradicts the meaning of the other.

Other antonymic relationships between word pairs are less semantically sharp because they can sometimes appear in the same context as a result of the broader semantic scope of one of the words. “Large” and “small,” or “old” and “young” generally suggest particular regions on size or age continua, but “how large is it?” or “how old is it?” can be asked about resources that are objectively small or young.[337][Ling]

6.4.2 Thesauri

The words that people naturally use when they describe resources reflect their unique experiences and perspectives, and this means that people often use different words for the same resource and the same words for different ones. Guiding people when they select description words from a *controlled vocabulary* is a partial solution to this *vocabulary problem* (§4.4.2.1 The Vocabulary Problem (page 189)) that becomes increasingly essential as the scope and scale of the organizing system grows. A *thesaurus* is a reference work that organizes words according to their semantic and lexical relationships. Thesauri are often used by professionals when they describe resources.

Thesauri have been created for many domains and subject areas. Some thesauri are very broad and contain words from many disciplines, like the Library of Congress Subject Headings (LOC-SH) used to classify any published content. Other commonly used thesauri are more focused, like the *Art and Architecture Thesaurus (AAT)* developed by the Getty Trust and the Legislative Indexing Vocabulary developed by the Library of Congress.

We can return to our simple food taxonomy to illustrate how a thesaurus annotates vocabulary terms with lexical and semantic relationships. The class inclusion relationships of *hypernomy* and *hyponymy* are usually encoded using `BT` (“broader term”) and `NT` (“narrower term”):

Food BT Meat
Beef NT Meat

The BT and NT relationships in a thesaurus create a hierarchical system of words, but a thesaurus is more than a lexical taxonomy for some domain because it also encodes additional lexical relationships for the most important words. Many thesauri emphasize the cluster of relationships for these key words and de-emphasize the overall lexical hierarchy.

Because the purpose of a thesaurus is to reduce synonymy, it distinguishes among synonyms or near-synonyms by indicating one of them as a preferred term using UF ("used for"):

Food UF Sustenance, Nourishment

A thesaurus might employ USE as the inverse of the UF relationship to refer from a less preferred or variant term to a preferred one:

Victuals USE Food

Thesauri also use RT ("related term" or "see also") to indicate terms that are not synonyms but which often occur in similar contexts:

Food RT Cooking, Dining, Cuisine

6.4.3 Relationships among Word Forms

The relationships among word meanings are critically important. Whenever we create, combine, or compare resource descriptions we also need to pay attention to relationships between word forms. These relationships begin with the idea that all natural languages create words and word forms from smaller units. The basic building blocks for words are called *morphemes* and can express semantic concepts (when they are called *root words*) or abstract concepts like "pastness" or "plural"). The analysis of the ways by which languages combine *morphemes* is called *morphology*.[339][Ling]

Simple examples illustrate this:

"dogs" = "dog" (root) + "s" (plural)
"uncertain" = "certain" (root) + "un" (negation)
"denied" = "deny" (root) + "ed" (past tense)

Morphological analysis of a language is heavily used in text processing to create indexes for information retrieval. For example, *stemming* (discussed in more detail in Chapter 10) is morphological processing which removes prefixes and suffixes to leave the root form of words. Similarly, simple text processing applications like hyphenation and spelling correction solve word form problems using roots and rules because it is more scalable and robust than solving them using word lists. Many misspellings of common words (e.g., "pain") are words of lower frequency (e.g., "pane"), so adding "pane" to a list of misspelled words would occasionally identify it incorrectly. In addition, because natural languages are generative and create new words all the time, a word list can never be complete; for example, when "flickr" occurs in text, is it a misspelling of "flicker" or the correct spelling of the popular photo-sharing site?

6.4.3.1 Derivational Morphology

Derivational morphology deals with how words are created by combining morphemes. *Compounding*, putting two "free morphemes" together as in "batman" or "catwoman," is an extremely powerful mechanism. The meaning of some compounds is easy to understand when the first morpheme qualifies or restricts the meaning of the second, as in "birdcage" and "tollbooth." However, many compounds take on new meanings that are not as literally derived from the meaning of their constituents, like "seahorse" and "batman."

Other types of derivations using "bound" morphemes follow more precise rules for combining them with "base" morphemes. The most common types of bound morphemes are prefixes and suffixes, which usually create a word of a different part-of-speech category when they are added. Familiar English prefixes include "a-," "ab-," "anti-," "co-," "de-," "pre-," and "un-." Among the most common English suffixes are "-able," "-ation," "-ify," "ing," "-ity," "-ize," "-ment," and "-ness." Compounding and adding prefixes or suffixes are simple mechanisms, but very complex words like "unimaginability" can be formed by using them in combination.

6.4.3.2 Inflectional Morphology

Inflectional mechanisms change the form of a word to represent tense, aspect, agreement, or other grammatical information. Unlike derivation, inflection never changes the part-of-speech of the base morpheme. The *inflectional morphology* of English is relatively simple compared with other languages.[341][Ling]

6.5 The Structural Perspective

The *structural perspective* analyzes the association, arrangement, proximity, or connection between resources without primary concern for their meaning or the origin of these relationships.[342][Com] We take a structural perspective when we define a family as "a collection of people" or when we say that a particular family like the Simpsons has five members. Sometimes all we know is that two resources are connected, as when we see a highlighted word or phrase that is pointing from the current web page to another. At other times we might know more about the reasons for the relationships within a set of resources, but we still focus on their structure, essentially merging or blurring all of the reasons for the associations into a single generic notion that the resources are connected.

Travers and Milgram conducted a now-famous study in the 1960s involving the delivery of written messages between people in the midwestern and eastern United States. If a person did not know the intended recipient, he was instructed to send the message to someone that he thought might know him. The study demonstrated what Travers and Milgram called the "small world problem," in which any two arbitrarily selected people were separated by an average of fewer than six links.

It is now common to analyze the number of "degrees of separation" between any pair of resources. For example, Markoff and Sengupta describe a 2011 study using Facebook data that computed the average "degree of separation" of any two people in the Facebook world to be 4.74.[343][Com]

Many types of resources have internal structure in addition to their structural relationships with other resources. Of course, we have to remember (as we discussed in §4.3 Resource Identity (page 182)) that we often face arbitrary choices about the abstraction and granularity with which we describe the parts that make up a resource and whether some combination of resource should also be identified as a resource. This is not easy when you are analyzing the structure of a car with its thousands of parts, and it is ever harder with information resources where there are many more ways to define parts and wholes. However, an advantage for information resources is that their internal structural descriptions are usually highly "computable," something we consider in depth in Chapter 10, *Interactions with Resources*.

Stop and Think: Kevin Bacon Numbers

See http://oracleofbacon.org/ for a web-based demonstration of "Kevin Bacon Numbers," which measure the average degrees of separation among more than 2.6 million actors in more than 1.9 million movies. Its name reflects the parlor game "Six Degrees of Kevin Bacon," a pun on "six degrees of separation" that is often associated with Travers and Milgram's work; the game relies on the remarkable variety of Bacon's roles, and hence the number of fellow actors in his movies (two actors in the same movie have one degree of separation). Bacon's Bacon Number is 2.994, but it turns out that more than 300 actors are closer to the center of the movie universe than Bacon. Try some famous actors and see if their Bacon Numbers are greater or smaller than Bacon's. (Hint: older actors have been in more movies.)

Business Structures

Management science is constantly reevaluating different structures for organizations. Many large businesses are organized similarly near the top, with a board of directors, a chief executive officer, and other executives who manage the vice presidents or directors of various business units. Within and across these business units, however, there are significant variations in how a business can organize its people.

Management strategies are built around the style of organization the business has chosen. These organizational choices reflect the CEO's management philosophy, the industry, regulatory requirements, operating scale, and other factors. Strict hierarchies are a traditional approach, with a tree structure leading from the lowest level worker directly up to the CEO. The strict management hierarchy at Foxconn is needed to enable close oversight of large numbers of low level employees in the manufacturing industry, with workers organized by physical location.

Other firms have a matrix structure in which an employee can be working on multiple projects, and reporting to a different manager for each one. A consulting firm's matrix structure might emphasize an employee's functional role (e.g., "process engineering consultant") and disassociate it from the employee's home location, which is why consultants spend so much time traveling on airplanes from project to project.

6.5.1 Intentional, Implicit, and Explicit Structure

In the discipline of organizing we emphasize "intentional structure" created by people or by computational processes rather than accidental or naturally-occurring structures created by physical and geological processes. We acknowledged in §1.5 that there is information in the piles of debris left after a tornado or tsunami and in the strata of the Grand Canyon. These structural patterns might be of interest to meteorologists, geologists, or others but because they were not created by an identifiable agent following one or more organizing principles, they are not our primary focus.

Some organizing principles impose very little structure. For a small collection of resources, co-locating them or arranging them near each other might be sufficient organization. We can impose two- or three-dimensional coordinate systems on this "implicit structure" and explicitly describe the location of a resource as precisely as we want, but we more naturally describe the structure of resource locations in relative terms. In English we have many ways to describe the structural relationship of one resource to another: "in," "on," "under," "behind," "above," "below," "near," "to the right of," "to the left of," "next to," and so on. Sometimes several resources are arranged or appear to be arranged in a

sequence or order and we can use positional descriptions of structure: a late 1990s TV show described the planet Earth as the "third rock from the Sun."[344][CogSci]

We pay most attention to intentional structures that are explicitly represented within and between resources because they embody the design or authoring choices about how much implicit or latent structure will be made explicit. Structures that can be reliably extracted by algorithms become especially important for very large collections of resources whose scope and scale defy structural analysis by people.

Stop and Think: Intentional, implicit/explicit structure

Find a map of the states (or provinces or other divisions) in your country. You probably think of some set of these as members of a collection. Other than their literal arrangement (e.g., "x is next to y, y is east of z"), how could you describe their relationships to each other within the collection? Are these relationships based on natural or unintentional properties or intentional ones? Example: in the United States, California, Oregon, and Washington are considered the "West Coast" and the Pacific Ocean determines their western boundaries. Some of the borders between the states are natural, determined by rivers, and other borders are more intentional and arbitrary.

6.5.2 Structural Relationships within a Resource

We almost always think of human and other animate resources as unitary entities. Likewise, many physical resources like paintings, sculptures, and manufactured goods have a material integrity that makes us usually consider them as indivisible. For an information resource, however, it is almost always the case that it has or might have had some internal structure or sub-division of its constituent data elements.

In fact, since all computer files are merely encodings of bits, bytes, characters and strings, all digital resources exhibit some internal structure, even if that structure is only discernible by software agents. Fortunately, the once inscrutable internal formats of word processing files are now much more interpretable after they were replaced by XML in the last decade.

When an author writes a document, he or she gives it some internal organization with its title, section headings, typographic conventions, page numbers, and other mechanisms that identify its parts and their significance or relationship to each other. The lowest level of this structural hierarchy, usually the paragraph, contains the text content of the document. Sometimes the author finds it useful to identify types of content like glossary terms or cross-references within the paragraph text. Document models that mix structural description with content "nuggets" in the text are said to contain *mixed content*.

Mixed Content

Mixed content distinguishes XML from other data representation languages. It is this structural feature, combined with the fact that child nodes in the XML Infoset (§9.2.2.2) are ordered, that makes it possible for XML documents to function both as human reader-oriented, textual documents and as structured data formats. It allows us to use natural language in writing descriptions while still enabling us to identify content by type by embedding markup to enclose "semantic nuggets" in otherwise undifferentiated text.[345][Com]

The *Guidelines for Electronic Text Encoding and Interchange*, produced by the *Text Encoding Initiative (TEI)*, for example, includes a set of elements and attributes for *Names, Dates, People and Places*.[346][Com]

In data-intensive or transactional domains, document instances tend to be homogeneous because they are produced by or for automated processes, and their information components will appear predictably in the same structural relationships with each other. These structures typically form a hierarchy expressed in an XML schema or word processing style template. XML documents describe their component parts using content-oriented elements like <ITEM>, <NAME>, and <ADDRESS>, that are themselves often aggregate structures or containers for more granular elements. The structures of resources maintained in databases are typically less hierarchical, but the structures are precisely captured in database schemas.

The internal parts of XML documents can be described, found and selected using the XPath language, which defines the structures and patterns used by XML forms, queries, and transformations. The key idea used by XPath is that the structure of XML documents is a tree of information items called nodes, whose locations are described in terms of the relationships between nodes. The relationships built into XPath, which it calls axes, include self, child, parent, following, and preceding, making it very easy to specify a structure-based query like "find all sections in Chapter 1 through Chapter 5 that have at least two levels of subsections."[347][Com] In addition, tools like Schematron take advantage of XPath's structural descriptions to test assertions about a document's structure and content. For example, a common editorial constraint might be that a numbered list must have at least three items.[348][Com]

In more qualitative, less information-intensive and more experience-intensive domains, we move toward the narrative end of the Document Type Spectrum, and document instances become more heterogeneous because they are produced by and for people. (See the sidebar, The Document Type Spectrum (page 168) in §4.2.1.) The information conveyed in the documents is conceptual or thematic rather than transactional, and the structural relationships between document parts are much weaker. Instead of precise structure and content rules,

there is usually just a shallow hierarchy marked up with Word processing or HTML tags like <HEAD>, <H1>, <H2>, and <LIST>.

Structural Metadata

Structural metadata, in the form of a schema for a database or document, describes a class of information resources, and may also prescribe grammatical details of inclusion and attribution relationships among the components. For example, the chapters of this book contain four levels of subsections. Each of those sections contains a title, some paragraphs and other text blocks, and subordinate sections. The textual content of the paragraphs includes highlighted terms and phrases that are defined *in situ* and referenced again in the glossary and index; there are also bibliographic citations that are reflected in the bibliography and index. We can discover these characteristics of a book through observation, but we could also examine its structural metadata, in its schema.

Structural metadata allows us to describe and prescribe relations among database tables, within the chapters of a book, or among parts in an inventory management system. The schema for HTML, for example, informs us that the <A> element can be used to signal a hypertext link-end; whether that link-end is an anchor or a target, or both, depends on the combination of values assigned to attributes. In HTML, the optional REL attribute may contain a value that signals the purpose of a hypertext link, and any HTML element may include a CLASS attribute value that may be used as a CSS selector for the purposes of formatting or dynamic interactions.

The usefulness of any given schema is often a function of the precision with which we may make useful statements based upon the descriptions and prescriptions it offers. Institutional schemas tend to be more prescriptive and restrictive, stressing professional orthodoxy and conformance to controlled vocabularies. Schemas for the information content in social and informal applications tend to be less prescriptive. Whether and how we use structural metadata is a tradeoff. Structural metadata is essential to enable quality control and maintenance in information collection and publishing processes, but someone has to do the work to create it.

The internal structural hierarchy in a resource is often extracted and made into a separate and familiar description resource called the “table of contents” to support finding and navigation interactions with the primary resource. In a printed media context, any given content resource is likely to only be presented once, and its page number is provided in the table of contents to allow the reader to locate the chapter, section or appendix in question. In a hypertext media context, a given resource may be a chapter in one book while being an appendix in another. Some tables of contents are created as a static structural de-

Stop and Think: Structural Metadata for a Course Syllabus

Analyze the structure of your syllabus for this course. What are its structural elements and some of the rules that specify how they are organized? Remember, think in terms of structural elements and not presentational elements. How does this structural schema compare to those of other course syllabi? What kinds of interactions would be enabled if all of your courses used the same syllabus schema?

scription, but others are dynamically generated from the internal structures whenever the resource is accessed. In addition, other types of entry points can be generated from the names or descriptions of content components, like selectable lists of tables, figures, maps, or code examples.

Identifying the components and their structural relationships in documents is easier when they follow consistent rules for structure (e.g., every non-text component must have a title and caption) and presentation (e.g., *hypertext links* in web pages are underlined and change cursor shapes when they are "moused over") that reinforce the distinctions between types of information components. Structural and presentation features are often ordered on some dimension (e.g., type size, line width, amount of white space) and used in a correlated manner to indicate the importance of a content component.[350[CogSci]]

DocBook Schema

The schema most commonly used for producing technical books is called DocBook; it describes every XML element and attribute and prescribes their grammatical forms. The schema lets us know that a formal paragraph must include a title, and that a title may contain emphasis. A schema can also describe and prescribe the lexical value space of a postal code, or require that every list must have at least three items. The DocBook schema is well-documented and has been production-tested in institutional publishing contexts for over twenty years.[349[Com]]

Many indexing algorithms treat documents as "bags of words" to compute statistics about the frequency and distribution of the words they contain while ignoring all semantics and structure. In Chapter 10, *Interactions with Resources*, we contrast this approach with algorithms that use internal structural descriptions to retrieve more specific parts of documents.

6.5.3 Structural Relationships between Resources

Many types of resources have "structural relationships" that interconnect them. Web pages are almost always linked to other pages. Sometimes the links among a set of pages remain mostly within those pages, as they are in an e-commerce

catalog site. More often, however, links connect to pages in other sites, creating a link network that cuts across and obscures the boundaries between sites.

The links between documents can be analyzed to infer connections between the authors of the documents. Using the pattern of links between documents to understand the structure of knowledge and of the intellectual community that creates it is not a new idea, but it has been energized as more of the information we exchange with other people is on the web or otherwise in digital formats. An important function in Google's search engine is the *page rank* algorithm that calculates the relevance of a page in part using the number of links that point to it while giving greater weight to pages that are themselves linked to often.[351][Web]

Web-based social networks enable people to express their connections with other people directly, bypassing the need to infer the connections from links in documents or other communications.

6.5.3.1 Hypertext Links

The concept of read-only or follow-only structures that connect one document to another is usually attributed to Vannevar Bush in his seminal 1945 essay titled "As We May Think." Bush called it *associative indexing,* defined as "a provision whereby any item may be caused at will to select immediately and automatically another."[352][Com] The "item" connected in this way was for Bush most often a book or a scientific article. However, the anchor and destination of a *hypertext link* can be a resource of any granularity, ranging from a single point or character, a paragraph, a document, or any part of the resource to which the ends of link are connected. The anchor and destination of a web link are its structural specification, but we often need to consider links from other perspectives. (See the sidebar, Perspectives on Hypertext Links (page 303)).

Theodor Holm Nelson, in a book intriguingly titled *Literary Machines,* renamed associative indexing as *hypertext* decades later, expanding the idea to make it a writing style as well as a reading style.[353][Com] Nelson urged writers to use hypertext to create non-sequential narratives that gave choices to readers, using a novel technique for which he coined the term *transclusion.*[354][Com]

At about the same time, and without knowing about Nelson's work, Douglas Engelbart's *Augmenting the Human Intellect,* described a future world in which professionals equipped with interactive computer displays utilize an information space consisting of a cross-linked resources.[355][Com]

In the 1960s, when computers lacked graphic displays and were primarily employed to solve complex mathematical and scientific problems that might take minutes, hours or even days to complete, Nelson's and Engelbart's visions of hypertext-based personal computing may have seemed far-fetched. In spite of

Transclusion

The inclusion, by hypertext reference, of a resource or part of a resource into another resource is called *transclusion*. Transclusion is normally performed automatically, without user intervention. The inclusion of images in web documents is an example of transclusion. Transclusion is a frequently used technique in business and legal document processing, where re-use of consistent and up-to-date content is essential to achieve efficiency and consistency.

this, by 1968, Engelbart and his team demonstrated human computer interface including the mouse, hypertext, and interactive media, along with a set of guiding principles.[356][Com]

Hypertext links are now familiar structural mechanisms in information applications because of the World Wide Web, proposed in 1989 by Tim Berners-Lee and Robert Cailliau.[357][Web] They invented the methods for encoding and following *hypertext links* using the now popular HyperText Markup Language (HTML).[358][Web] The resources connected by HTML's hypertext links are not limited to text or documents. Selecting a hypertext link can invoke a connected resource that might be a picture, video, or interactive application.[359][Com]

By 1993, personal computers, with a graphic display, speakers and a mouse pointer, had become ubiquitous. NCSA Mosaic is widely credited with popularizing the World Wide Web and HTML in 1993, by introducing inline graphics, audio and video media, rather than having to link to media segments in a separate window.[360][Web] The ability to transclude images and other media would transform the World Wide Web from a text-only viewer with links to a "networked landscape" with hypertext signposts to guide the way. On 12 November 1993, the first full release of NCSA Mosaic on the world's three most popular operating systems (X Windows, Microsoft Windows, and Apple Macintosh) enabled the general public to access the network with a graphical browser.[361][Web]

Since browsers made them familiar, hypertext links have been used in other computing applications as structure and navigation mechanisms.

Perspectives on Hypertext Links

A lexical perspective on hypertext links concerns the words that are used to signal the presence of a link or to encode its type. In web contexts, the words in which a structural link is embedded are called the *anchor text*. More generally, rhetorical structure theory analyzes how different conventions or signals in texts indicate relationships between texts or parts of them, like the subtle differences in polarity among "see," "see also," and "but see" as citation signals.[362][Ling]

Many hypertext links in web pages are purely structural because they lack explicit representation of the reason for the relationship. When it is evident, this semantic property of the link is called the *link type*.[363][Com]

An architectural perspective on links considers whether links are *one-way* or *bi-directional*. When a *bi-directional link* is created between an anchor and a destination, it is as though a one-way link that can be followed in the opposite direction is automatically created. Two one-way links serve the same purpose, but the return link is not automatically established when the first one is created. A second architectural consideration is whether to employ *binary links*, connecting one anchor to one destination, or *n-ary links*, connecting one anchor to multiple types of destinations.[364][Com] (See §6.6)

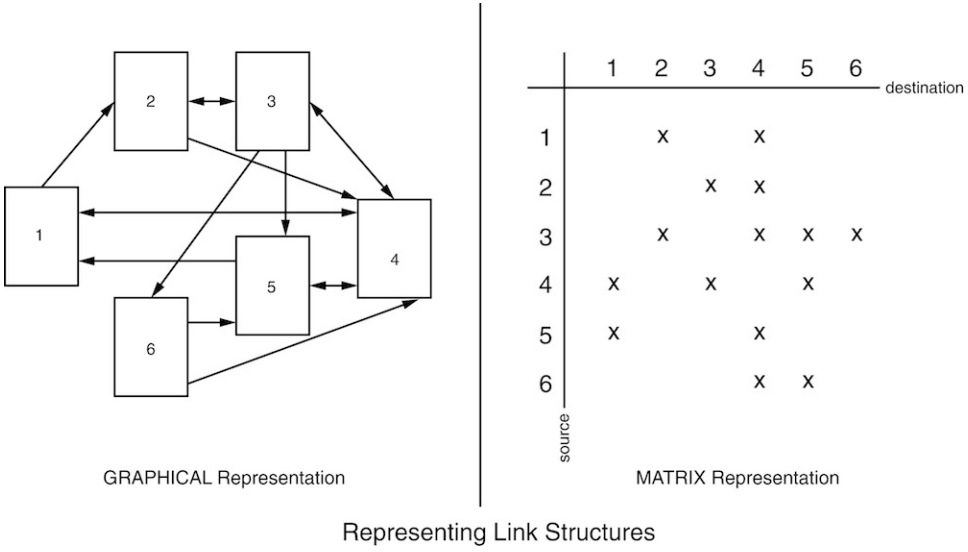
A "front end" or "surface" implementation perspective on hypertext links concerns how the presence of the link is indicated in a user interface; this is called the "link marker"; underlining or coloring of clickable text are conventional markers for web links.[365][Web] [366][IA] A "back end" implementation issue is whether links are contained or embedded in the resources they link or whether they are stored separately in a *link base*.[367][Web] (See §6.7)

6.5.3.2 Analyzing Link Structures

We can portray a set of links between resources graphically as a pattern of boxes and links. Because a link connection from one resource to another need not imply a link in the opposite direction, we distinguish one-way links from explicitly bi-directional ones.

A graphical representation of link structure is shown on the left panel of figure Figure 6.3, Representing Link Structures.. For a small network of links, a diagram like this one makes it easy to see that some resources have more incoming or outgoing links than other resources. However, for most purposes we leave the analysis of link structures to computer programs, and there it is much better to represent the link structures more abstractly in matrix form. In this matrix the resource identifiers on the row and column heads represent the source and destination of the link. This is a full matrix because not all of the links are symmetric; a link from resource 1 to resource 2 does not imply one from 2 to 1.

Figure 6.3. Representing Link Structures.



The structure of links between web resources can be represented graphically or in a matrix. The matrix representation is a more abstract one that can be analyzed by computers.

A matrix representation of the same link structure is shown on the right panel of Figure 6.3, Representing Link Structures.. This representation models the network as a directed graph in which the resources are the vertices and the relationships are the edges that connect them. We now can apply graph algorithms to determine many useful properties. A very important property is *reachability*, the "can you get there from here" property.[368][Com] Other useful properties include the average number of incoming or outgoing links, the average distance between any two resources, and the shortest path between them.

6.5.3.3 Bibliometrics, Shepardizing, Altmetrics, and Social Network Analysis

Information scientists began studying the structure of scientific citation, now called *bibliometrics*, nearly a century ago to identify influential scientists and publications. This analysis of the flow of ideas through publications can identify "invisible colleges" of scientists who rely on each other's research, and recognize the emergence of new scientific disciplines or research areas. Universities use bibliometrics to evaluate professors for promotion and tenure, and libraries use it to select resources for their collections.

The expression of citation relationships between documents is especially nuanced in legal contexts, where the use of legal cases as precedents makes it essential to distinguish precisely where a new ruling lies on the relational continuum between "Following" and "Overruling" with respect to a case it cites. The analysis of legal citations to determine whether a cited case is still good law is called *Shepardizing* because lists of cases annotated in this way were first published in the late 1800s by Frank Shepard, a salesman for a legal publishing company.[370][Law]

The links pointing to a web page might be thought of as citations to it, so it is tempting to make the analogy to consider Shepardizing the web. But unlike legal rulings, web pages aren't always persistent, and only courts have the authority to determine the value of cited cases as precedents, so Shepard-like metrics for web pages would be tricky to calculate and unreliable.

Nevertheless, the web's importance as a publishing and communication medium is undeniable, and many scholars, especially younger ones, now contribute to their fields by blogging, Tweeting, leaving comments on online publications, writing Wikipedia articles, giving MOOC lectures, and uploading papers, code, and datasets to open access repositories. Because the traditional bibliometrics pay no attention to this body of work, alternative metrics or "altmetrics" have been proposed to count these new venues for scholarly influence.

Facebook's valuation is based on its ability to exploit the structure of a person's social network to personalize advertisements for people and "friends" to whom they are connected. Many computer science researchers are working to determine the important characteristics of people and relationships that best identify the people whose activities or messages influence others to spend money.[372][Com]

6.6 The Architectural Perspective

The architectural perspective emphasizes the number and abstraction level of the components of a relationship, which together characterize the complexity of the relationship. We will briefly consider three architectural issues: degree (or arity), cardinality, and directionality.

These architectural concepts come from data modeling and they enable relationships to be described precisely and abstractly, which is essential for maintaining an organizing system that implements relationships among resources. Application and technology lifecycles have never been shorter, and vast amounts of new data are being created by increased tracking of online interactions and by all the active resources that are now part of the Internet of Things. Organizing systems built without clear architectural foundations cannot easily scale up in size and scope to handle these new requirements.

6.6.1 Degree

The *degree* or *arity* of a relationship is the number of entity types or categories of resources in the relationship. This is usually, though not always, the same as the number of arguments in the relationship expression.

Homer Simpson (husband) ⇔ is-married-to ⇔ Marge Simpson (wife)

is a relationship of degree 2, a **binary** relationship between two entity types, because the "is-married-to" relationship as we first defined it requires one of the arguments to be of entity type "husband" and one of them to be of type "wife."

Now suppose we change the definition of marriage to allow the two participants in a marriage to be any instance of the entity type "person." The relationship expression looks exactly the same, but its degree is now *unary* because only 1 entity type is needed to instantiate the two arguments:

Homer Simpson (person) ⇔ is-married-to ⇔ Marge Simpson (person)

Some relationships are best expressed as *ternary* ones that involve three different entity types. An example that appears in numerous data modeling books is one like this:

Supplier → provides → Part → assembled-in → Product

It is always possible to represent ternary relationships as a set of binary ones by creating a new entity type that relates to each of the others in turn. This new entity type is called a dummy in modeling practice.

Supplier → provides → DUMMY
Part → provided-for → DUMMY
DUMMY → assembled-in → Product

This transformation from a sensible ternary relationship to three binary ones involving a DUMMY entity type undoubtedly seems strange, but it enables all relationships to be binary while still preserving the meaning of the original ternary one. Making all relationships binary makes it easier to store relationships and combine them to discover new ones.

6.6.2 Cardinality

The *cardinality* of a relationship is the number of instances that can be associated with each entity type in a relationship. At first glance this might seem to be degree by another name, but it is not.

Cardinality is easiest to explain for binary relationships. If we return to Homer and Marge, the binary relationship that expresses that they are married husband and wife is a *one-to-one* relationship because a husband can only have one wife and a wife can only have one husband (at a time, in monogamous societies like the one in which the Simpsons live).

In contrast, the "is-parent-of" relationship is one-to-many, because the meaning of being a parent makes it correct to say that:

Homer Simpson → is-parent-of → Bart AND Lisa AND Maggie

As we did with the ternary relationship in §6.6.1 Degree (page 306), we can transform this more complex relationship architecture to a set of simpler ones by restricting expressions about being a parent to the one-to-one cardinality.

Homer Simpson → is-parent-of → Bart
Homer Simpson → is-parent-of → Lisa
Homer Simpson → is-parent-of → Maggie

The one-to-many expression brings all three of Homer's children together as arguments in the same relational expression, making it more obvious that they share the same relationship than in the set of separate and redundant one-to-one expressions.

6.6.3 Directionality

The *directionality* of a relationship defines the order in which the arguments of the relationship are connected. A *one-way* or *uni-directional* relationship can be followed in only one direction, whereas a *bi-directional* one can be followed in both directions.

All symmetric relationships are bi-directional, but not all bi-directional relationships are symmetric. (See §6.3.2.1 Symmetry (page 284).) A relationship between a manager and an employee that he manages is "employs," a different meaning than the "is-employed-by" relationship in the opposite direction. As in this example, the relationship is often lexicalized in only one direction.

6.7 The Implementation Perspective

Finally, the *implementation perspective* on relationships considers how a relationship is realized or encoded in a technology context. The implementation perspective contrasts strongly with the conceptual, structural, and architectural perspectives, which emphasize the meaning and abstract structure of relationships. The implementation perspective is a superset of the lexical perspective, because the choice of the language in which to express a relationship is an implementation decision. However, most people think of implementation as all of the decisions about technological form rather than just about the choice of words.

In this book we focus on the fundamental issues and challenges that apply to all organizing systems, and not just on information-intensive ones that rely extensively on technology. Even with this reduced scope, there are some critical implementation concerns about the notation, syntax, and deployment of the relationships and other descriptions about resources. We briefly introduce some of these issues here and then discuss them in detail in Chapter 9, *The Forms of Resource Descriptions*.

6.7.1 Choice of Implementation

The choice of implementation determines how easy it is to understand and process a set of relationships. For example, the second sentence of this chapter is a natural language implementation of a set of relationships in the Simpson family:

> **The Simpson family includes a man named Homer and a woman named Marge, the married parents of three sibling children, a boy named Bart and two girls, Lisa and Maggie.**

A subject-predicate-object syntax makes the relationships more explicit:

Example 6.1. Subject-predicate syntax

Homer Simpson → is-married-to → Marge Simpson
Homer Simpson → is-parent-of → Bart
Homer Simpson → is-parent-of → Lisa
Homer Simpson → is-parent-of → Maggie
Marge Simpson → is-married-to → Homer Simpson
Marge Simpson → is-parent-of → Bart
Marge Simpson → is-parent-of → Lisa
Marge Simpson → is-parent-of → Maggie
Bart Simpson → is-a → Boy
Lisa Simpson → is-a → Girl
Maggie Simpson → is-a → Girl

In the following example of a potential XML implementation syntax, we emphasize class inclusion relationships by using elements as containers, and the relationships among the members of the family are expressed explicitly through references, using XML's ID and IDREF attribute types:[373][Com]

Example 6.2. An XML implementation syntax

```
<Family name="Simpson">
   <Parents children="Bart Lisa Maggie">
      <Father name="Homer" spouse="Marge"  />
      <Mother name="Marge" spouse="Homer"  />
   </Parents>
   <Children parents="Homer Marge" >
      <Boy name="Bart" siblings="Lisa Maggie" />
      <Girl name="Lisa" siblings="Bart Maggie" />
      <Girl name="Maggie" siblings="Bart Lisa" />
   </Children>
</Family>
```

None of the models we have presented so far in this chapter represents the complexities of modern families that involve multiple marriages and children from more than one marriage, but they are sufficient for our limited demonstration purposes.

6.7.2 Syntax and Grammar

The *syntax* and *grammar* of a language consists of the rules that determine which combinations of its words are allowed and are thus grammatical or *well-formed*. Natural languages have substantial similarities by having nouns, verbs, adjectives and other parts of speech, but they differ greatly in how they arrange them to create sentences. Conformance to the rules for arranging these parts makes a sentence syntactically compliant but does not mean that an expression is semantically comprehensible; the classic example is Chomsky's anomalous sentence:

Colorless green ideas sleep furiously

Any meaning this sentence has is odd, difficult to visualize, and outside of readily accessible experience, but anyone who knows the English language can recognize that it follows its syntactic rules, as opposed to this sentence, which breaks them and seems completely meaningless:

Ideas colorless sleep furiously green[374][Ling]

6.7.3 Requirements for Implementation Syntax

The most basic requirement for implementation syntax is that it can represent all the expressions that it needs to express. For the examples in this chapter we have used an informal combination of English words and symbols (arrows and parentheses) that you could understand easily, but simple language is incapable of expressing most of what we readily say in English. But this benefit of natural language only accrues to people, and the more restrictive and formal syntax is easier to understand for computers.

A second consideration is that the implementation can be understood and used by its intended users. We can usually express a relationship in different languages while preserving its meaning, just as we can usually implement the same computing functionality in different programming languages. From a semantic perspective these three expressions are equivalent:

My name is Homer Simpson
Mon nom est Homer Simpson
Mein name ist Homer Simpson

However, whether these expressions are equivalent for someone reading them depends on which languages they understand.

An analogous situation occurs with the implementation of web pages. HTML was invented as a language for encoding how web pages look in a browser, and most of the tags in HTML represent the simple structure of an analogous print document. Representing paragraphs, list items and numbered headings with <P> and <LI> and <H*n*> makes using HTML so easy that school children can create web pages. However, the "web for eyes" implemented using HTML is of less efficient or practical for computers that want to treat content as product catalogs, orders, invoices, payments, and other business transactions and information that can be analyzed and processed. This "web for computers" is best implemented using domain-specific vocabularies in XML.

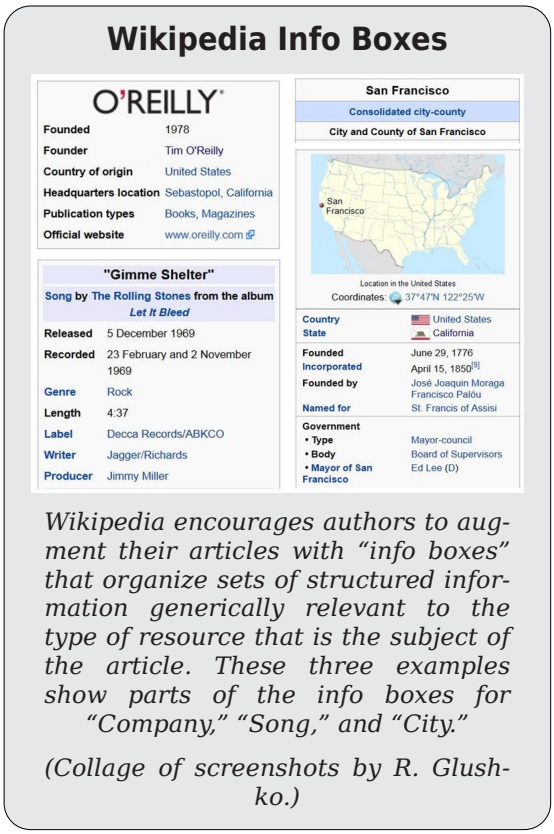
6.8 Relationships in Organizing Systems

In the previous sections as we surveyed the five perspectives on analyzing relationships we mentioned numerous examples where relationships had important roles in organizing systems. In this final section we examine three contexts for organizing systems where relationships are especially fundamental; the *Semantic Web* and Linked Data, bibliographic organizing systems, and situations involving system integration and interoperability.

6.8.1 The Semantic Web and Linked Data

In a classic 2001 paper, Tim Berners-Lee laid out a vision of a *Semantic Web* in which all information could be shared and processed by automated tools as well as by people.[375][Web] The essential technologies for making the web more semantic and relationships among web resources more explicit are applications of XML, including RDF (§5.2.2.4 Resource Description Framework (RDF) (page 223)), and OWL (§6.3.3 Ontologies (page 286)). Many tools have been developed to support more semantic encoding, but most still require substantial expertise in semantic technologies and web standards.[376][Com]

More likely to succeed are applications that aim lower, not trying to encode all the latent semantics in a document or web page. For example, some wiki and blogging tools contain templates for semantic annotation, and Wikipedia has thousands of templates and "info boxes" to encourage the creation of information in content-encoded formats.



Wikipedia encourages authors to augment their articles with "info boxes" that organize sets of structured information generically relevant to the type of resource that is the subject of the article. These three examples show parts of the info boxes for "Company," "Song," and "City."

(Collage of screenshots by R. Glushko.)

The "Linked Data" movement is an extension of the *Semantic Web* idea to reframe the basic principles of the web's architecture in more semantic terms. Instead of the limited role of links as simple untyped relationships between HTML documents, links between resources described by RDF can serve as the bridges between islands of semantic data, creating a Linked Data network or cloud.[377][Web]

6.8.2 Bibliographic Organizing Systems

Much of our thinking about relationships in organizing systems for information comes from the domain of bibliographic cataloging of library resources and the related areas of classification systems and descriptive thesauri. Bibliographic relationships provide an important means to build structure into library catalogs.

Bibliographic relationships are common among library resources. Smiraglia and Leazer found that approximately 30% of the works in the *Online Com-*

puter Library Center (OCLC) WorldCat union catalog have associated derivative works. Relationships among items within these bibliographic families differ, but the average family size for those works with derivative works was found to be 3.54 items. Moreover, "canonical" works that have strong cultural meaning and influence, such as "the plays of William Shakespeare" and *The Bible,* have very large and complex bibliographic families.

6.8.2.1 Tillett's Taxonomy

Barbara Tillett, in a study of 19th and 20th-century catalog rules, found that many different catalog rules have existed over time to describe bibliographic relationships. She developed a taxonomy of bibliographic relationships that includes equivalence, derivative, descriptive, whole-part, accompanying, sequential or chronological, and shared characteristic. These relationship types span the relationship perspectives defined in this chapter; equivalence, derivative, and description are semantic types; whole-part and accompanying are part semantic and part structural types; sequential or chronological are part lexical and part structural types; and shared characteristics are part semantic and part lexical types.

6.8.2.2 Resource Description and Access (RDA)

Many cataloging researchers have recognized that online catalogs do not do a very good job of encoding bibliographic relationships among items, both due to catalog display design and to the limitations of how information is organized within catalog records. Author name authority databases, for example, provide information for variant author names, which can be very important in finding all of the works by a single author, but this information is not held within a catalog record. Similarly, MARC records can be formatted and displayed in web library catalogs, but the data within the records are not available for re-use, re-purposing, or re-arranging by researchers, patrons, or librarians.

The Resource Description and Access (RDA) next-generation cataloging rules are attempting to bring together disconnected resource descriptions to provide more complete and interconnected data about works, authors, publications, publishers, and subjects.

RDA uses RDF to assert relationships among bibliographic materials.

6.8.2.3 RDA and the Semantic Web

The move in RDA to encode bibliographic data in RDF stems from the desire to make library catalog data more web-accessible. As web-based data mash-ups, application programming interfaces (APIs), and web searching are becoming ubiquitous and expected, library data are becoming increasingly isolated. The

developers of RDA see RDF as the means for making library data more widely available online.

In addition to simply making library data more web accessible, RDA seeks to leverage the distributed nature of the Semantic Web. Once rules for describing resources, and the relationships between them, are declared in RDF syntax and made publicly available, the rules themselves can be mixed and mashed up. Creators of information systems that use RDF can choose elements from any RDF schema. For example, we can use the Dublin Core metadata schema (which has been aligned with the RDF model) and the *Friend of a Friend (FOAF)* schema (a schema to describe people and the relationships between them) to create a set of metadata elements about a journal article that goes beyond the standard bibliographic information. RDA's process of moving to RDF is well underway.

6.8.3 Integration and Interoperability

Integration is the controlled sharing of information between two (or more) business systems, applications, or services within or between firms. *Integration* means that one party can extract or obtain information from another one, it does not imply that the recipient can make use of the information.

Interoperability goes beyond integration to mean that systems, applications, or services that exchange information can make sense of what they receive. *Interoperability* can involve identifying corresponding components and relationships in each system, transforming them syntactically to the same format, structurally to the same granularity, and semantically to the same meaning.

For example, an Internet shopping site might present customers with a product catalog whose items come from a variety of manufacturers who describe the same products in different ways. Likewise, the end-to-end process from customer ordering to delivery requires that customer, product and payment information pass through the information systems of different firms. Creating the necessary information mappings and transformations is tedious or even impossible if the components and relationships among them are not formally specified for each system.

In contrast, when these models exist as data or document schemas or as classes in programming languages, identifying and exploiting the relationships between the information in different systems to achieve *interoperability* or to merge different classification systems can often be completely automated. Because of the substantial economic benefits to governments, businesses, and their customers of more efficient information *integration* and exchange, efforts to standardize these information models are important in numerous industries. Chapter 10, *Interactions with Resources* will dive deeper into *interoperability* issues, especially those that arise in business contexts.

6.9 Key Points in Chapter Six

- A relationship is "an association among several things, with that association having a particular significance."

  (See §6.1 Introduction (page 273))
- Just identifying the resources involved is not enough because several different relationships can exist among the same resources.

  (See §6.2 Describing Relationships: An Overview (page 275))
- Most relationships between resources can be expressed using a subject-predicate-object model.

  (See §6.3 The Semantic Perspective (page 276) and §6.7.1 Choice of Implementation (page 308))
- For a computer to understand relational expressions, it needs a computer-processable representation of the relationships among words and meanings that makes every important semantic assumption and property precise and explicit.

  (See §6.3 The Semantic Perspective (page 276))
- Three broad categories of semantic relationships are inclusion, attribution, and possession.

  (See §6.3.1 Types of Semantic Relationships (page 278))
- A set of interconnected class inclusion relationships creates a hierarchy called a taxonomy.

  (See §6.3.1.1 Inclusion (page 279))
- Classification is a class inclusion relationship between an instance and a class.

  (See §6.3.1.1 Inclusion (page 279))
- Ordering and inclusion relationships are inherently transitive, enabling inferences about class membership and properties.

  (See §6.3.2.2 Transitivity (page 285))
- Class inclusion relationships form a framework to which other kinds of relationships attach, creating a network of relationships called an ontology.

  (See §6.3.3 Ontologies (page 286))
- When words encode the semantic distinctions expressed by class inclusion, the more specific class is called the hyponym; the more general class is the hypernym.

  (See §6.4.1.1 Hyponymy and Hyperonymy (page 290))

- Morphological analysis of how words in a language are created from smaller units is heavily used in text processing.

  (See §6.4.3 Relationships among Word Forms (page 293))

- Many types of resources have internal structure in addition to their structural relationships with other resources.

  (See §6.5.2 Structural Relationships within a Resource (page 297) and §6.5.2 Structural Relationships within a Resource (page 297))

- Using the pattern of links between documents to understand the structure of knowledge and the structure of the intellectual community that creates it is an idea that is nearly a century old.

  (See §6.5.3 Structural Relationships between Resources (page 300))

- Many hypertext links are purely structural because there is no explicit representation of the reason for the relationship.

  (See the sidebar, Perspectives on Hypertext Links (page 303))

- The architectural perspective on resources emphasizes the number and abstraction level of the components of a relationship; three important issues are degree, cardinality, and directionality.

  (See §6.6 The Architectural Perspective (page 305))

- The essential technologies for making the web more semantic and relationships among web resources more explicit are XML, RDF, and OWL.

  (See §6.8.1 The Semantic Web and Linked Data (page 311))

- Much of our thinking about relationships in organizing systems for information comes from the domain of bibliographic cataloging of library resources and the related areas of classification systems and descriptive thesauri.

  (See §6.8.2 Bibliographic Organizing Systems (page 311))

- The Resource Description and Access (RDA) next-generation cataloging rules are attempting to bring together disconnected resource descriptions.

  (See §6.8.2.2 Resource Description and Access (RDA) (page 312))

- Integration is the controlled sharing of information between two (or more) business systems, applications, or services within or between firms.

  (See §6.8.3 Integration and Interoperability (page 313))

- Interoperability goes beyond integration to mean that systems, applications, or services that exchange information can make sense of what they receive.

  (See §6.8.3 Integration and Interoperability (page 313))

Endnotes for Chapter 6

[309][Bus] *The Simpsons* TV show began in 1989 and is now the longest running scripted TV show ever. The official website is www.thesimpsons.com. The show is dubbed into French, Italian and Spanish for viewers in Quebec, France, Italy, Latin America and Spain. *The Simpson's Movie* has been dubbed into Mandarin Chinese and Cantonese. For more information about Mandarin kinship terms see http://mandarin.about.com/od/vocabularylists/tp/family.htm. (Yes, we know that Bart actually calls his father by his first name.)

[310][CogSci] Kinship can be studied from both anthropological and biological perspectives, which differ to the degree to which they emphasize social relationships and genetic ones. Kinship has been systematically studied since the nineteenth century: (Morgan 1871/ 1997) developed a system of kinship classification still taught today. A detailed interactive web tutorial developed by Brian Schwimer can be found at http://umanitoba.ca/faculties/arts/anthropology/kintitle.html.

[311][Com] Kent's *Data and Reality* was first published in 1978 with a second edition in 1998. Kent was a well-known and well-liked researcher in data modeling at IBM, and his book became a cult classic. In 2012, seven years after Kent's death, a third edition (Kent and Hoberman 2012) came out, slightly revised and annotated but containing essentially the same content as the book from 34 years earlier because its key issues about data modeling are timeless.

[312][CogSci] "Semantic" is usually defined as "relating to meaning or language" and that does not seem helpful here.

[313][CogSci] For decades important and vexing questions have been raised about the specificity of these predicate-argument associations and how or when the semantic constraints they embody combine with syntactic and contextual constraints during the process of comprehending language. Consider how "While in the operating room, the surgeon used a knife to cut the ____" generates a different expectancy from the same predicate and agent in "While at the fancy restaurant, the surgeon used a knife to cut the ____." See (Elman 2009).

[314][Law] This book is not the place for the debate over the definition of marriage. We are not bigots; we just do not need this discussion here. If these definitions upset you here, you will feel better in §6.6.1.

[315][CogSci] Typically, when people use language they operate on the assumption that everyone shares their model of the world, providing the common ground that enables them to communicate. As we saw in Chapter 4 and Chapter 5, (because of the *vocabulary problem* and different purposes for using resources and language) this assumption is often wrong, This paves the way for serious misun-

derstandings, since what is assumed to be shared knowledge may not really be shared or understood the same way.

[316][CogSci] See (Chaffin and Herrmann 1984), (Storey 1993).

[317][CogSci] Which of these classifications is most relevant depends on the context. In addition, there might be other Homer Simpsons who are not cartoon characters or who are not married, so we might have to disambiguate this homonymy to make sure we referring to the intended Homer Simpson.

[318][CogSci] (Winston, Chaffin, and Herman 1987).

[319][CogSci] (Storey 1993).

[320][Bus] Martin is the animated gecko who is the advertising spokesman for Geico Insurance (http://www.geico.com/). Martin's wit and cockney accent make him engaging and memorable, and a few years ago he was voted the favorite advertising icon in the US.

[321][CogSci] (Gentner 1983).

[322][Ling] (Miller and Johnson-Laird 1976, p 565).

[323][Ling] Some people have argued that meronomy is not transitive, but a closer look at their supposed counter-examples suggests otherwise. See Section 5 in (Winston, Chaffin, and Herman 1987).

[324][Phil] *ontology* is a branch of philosophy concerned with what exists in reality and the general features and relations of whatever that might be (Hofweber 2009). Computer science has adopted "ontology" to refer to any computer-processable resource that represents the relationships among words and meanings in some knowledge domain. See (Gruber 1993), (Guarino 1998).

[325][Web] Web Ontology Language (OWL) http://www.w3.org/2004/OWL/.

[326][Com] http://www.cyc.com/.

[327][Ling] Languages and cultures differ in how they distinguish and describe kinship, so Bart might find the system of family organization easier to master in some countries and cultures and more difficult in others.

[328][Ling] (Bentivogli and Pianta 2000).

[329][Ling] This example comes from (Fellbaum 2010, pages 236-237). German has a word *Kufenfahrzeug* for vehicle on runners.

[330][Ling] (Miller 1998).

[331][Ling] (Bolshakov and Gelbukh 2004), p, 314. The quote continues "The references to 'some class' and to 'insignificant change' make this definition rather vague, but we are not aware of any significantly stricter definition. Hence the

creation of synonymy dictionaries, which are known to be quite large, is rather a matter of art and insight."

[332][CogSci] George Miller made many important contributions to the study of mind and language during his long scientific career. His most famous article, "The Magical Number Seven, Plus or Minus Two" (Miller 1956), was seminal in its proposals about information organization in human memory, even though it is one of the most misquoted scientific papers of all time. Relatively late in his career Miller began the WordNet project to build a semantic dictionary, which is now an essential resource in natural language processing applications. See http://wordnet.princeton.edu/.

[333][Bus] This navigation is easiest to carry out using the commercial product called "The Visual Thesaurus" at http://www.visualthesaurus.com/.

[334][Ling] These contrasting meanings for "bank" are clear cases of polysemy, but there are often much subtler differences in meaning that arise from context. The verb "save" seems to mean something different in "The shopper saved..." versus "The lifeguard saved..." although they overlap in some ways. (Fillmore and Atkins 2000) and others have proposed definitions of polysemy, but there is no rigorous test for determining when word meanings diverge sufficiently to be called different senses.

[335][Com] Many techniques for using WordNet to calculate measures of semantic similarity have been proposed. See (Budanitsky and Hirst 2006).

[336][Ling] See (Gross and Miller, 1990).

[337][Ling] This type of "lexical asymmetry" is called "markedness." The broader or dominant term is the unmarked one and the narrower one is the marked one. See (Battistella 1996).

[339][Ling] Languages differ a great deal in morphological complexity and in the nature of their morphological mechanisms. Mandarin Chinese has relatively few morphemes and few grammatical inflections, which leads to a huge number of homophones. English is pretty average on this scale. A popular textbook on morphology is (Haspelmath and Sims 2010).

[341][Ling] English nouns have plural (book/books) and possessive forms (the professor's book), adjectives have comparatives and superlatives (big/bigger/biggest), and regular verbs have only four inflected forms (see http://cla.calpoly.edu/~jrubba/morph/morph.over.html). In contrast, in Classical Greek each noun can have 11 word forms, each adjective 30, and every regular verb over 300 (Anderson 2001).

[342][Com] Of the five perspectives on relationships in this chapter, the structural one comes closest to the meaning of "relation" in mathematics and computer science, where a relation is a set of ordered elements ("tuples") of equal degree

(§6.6.1 Degree (page 306)). A binary relation is a set of element pairs, a ternary relation is a set of 3-tuples, and so on. The elements in each tuple are "related" but they do not need to have any "significant association" or "relationship" among them.

[343][Com] See (Travers and Milgram 1969) and (Markoff and Sengupta 2011).

[344][CogSci] This seems like an homage to Jimi Hendrix based on the title from a 1967 song, *Third Stone from the Sun* http://en.wikipedia.org/wiki/Third_Stone_from_the_Sun.

[345][Com] The subfield of natural language processing called "named entity recognition" has as its goal the creation of mixed content by identifying people, companies, organizations, dates, trademarks, stock symbols, and so on in unstructured text.

[346][Com] Text Encoding Initiative13. Names, Dates, People, and Places.

[347][Com] See (Holman 2001) or (Tidwell 2008).

[348][Com] See (van der Vlist 2007) and schematron.org for overviews. See (Hamilton and Wood 2012) for a detailed case study.

[349][Com] (Walsh 2010).

[350][CogSci] These layout and typographic conventions are well known to graphic designers (Williams 2012) but are also fodder for more academic treatment in studies of visual language or semiotics (Crow 2010).

[351][Web] (Page, Brin, Motwani, and Winograd 1999) describes Page Rank when its inventors were computer science graduate students at Stanford. It is not a coincidence that the technique shares a name with one of its inventors, Google co-founder and CEO Larry Page. (Langville and Meyer 2012) is an excellent textbook. The ultimate authority about how page rank works is Google; see https://www.google.com/insidesearch/howsearchworks/thestory/.

[352][Com] (Bush 1945). "Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them..." See http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

[353][Com] (Nelson 1981).

[354][Com] See *Computer Lib/Dream Machines* (Nelson 1981) for an early example of Nelson's non-liner book style.

[355][Com] (Engelbart 1963) Douglas Engelbart credits Bush's "As We May Think" article as his direct inspiration. Engelbart was in the US Navy, living in a hut in the South Pacific during the last stages of WWII when he read *The Atlantic* monthly magazine in which Bush's article was published.

[356][Com] Doug Engelbart's demonstration has been called the "Mother of All Demos" and can be seen in its entirety at http://sloan.stanford.edu/MouseSite/1968Demo.html.

[362][Ling] See (Lorch 1989), (Mann and Thomson 1988). For example, an author might use "See" as in "See (Glushko et al. 2013)" when referring to this chapter if it is consistent with his point of view. On the other hand, that same author could use "but" as a contrasting citation signal, writing "But see (Glushko et al. 2013)" to express the relationship that the chapter disagrees with him.

[363][Com] Before the web, most hypertexts implementations were in stand-alone applications like CD-ROM encyclopedias or in personal information management systems that used "cards" or "notes" as metaphors for the information units that were linked together, typically using rich taxonomies of *link types*. See (Conklin 1987), (Conklin and Begeman 1988), and (DeRose 1989).

[364][Com] Many of the pre-web hypertext designs of the 1980s and 1990s allowed for n-ary links. The Dexter hypertext reference model (Halasz and Schwartz 1994) elegantly describes the typical architectures. However, there is some ambiguity in use of the term binary in hypertext link architectures. One-to-one vs. one-to-many is a cardinality distinction, and some people reserve binary to discussion about degree.

[365][Web] See (Weinreich, Obendorf, and Lamersdorf 2001).

[366][IA] Most designers use a variety of visual cues and conventions to distinguish hyperlinks (e.g., plain hyperlink, button, selectable menu, etc.) so that users can anticipate how they work and what they mean. A recent counter-trend called "flat design" —exemplified most notably by the user interfaces of Windows 8 and iOS 7— argues for a minimalist style with less variety in typography, color, and shading. Flat designs are easier to adapt across multiple devices, but convey less information.

[367][Web] See (Brailsford 1999), (Wilde and Lowe 2002).

[357][Web] (Gillies and Cailliau 2000).

[358][Web] Most web links are very simple in structure. The anchor text in the linking document is wrapped in <A> and </A> tags, with an HREF (hypertext reference) attribute that contains the URI of the link destination if it is in another page, or a reference to an ID attribute if the link is to a different part of the same page. HTML also has a <LINK> tag, which, along with <A> have REL (relationship) and REV (reverse relationship) attributes that enable the encoding of typed relationships in links. In a book context for example, link relationships and reverse relations include obvious candidates such as next, previous, parent, child, table of contents, bibliography, glossary and index.

[359][Com] Using hypertext links as interaction controls is the modern dynamic manifestation of cross references between textual commentary and illustrations in books, a mechanism that dates from the 1500s (Kilgour 1998). Hypertext links can be viewed as state transition controls in distributed collections of web-based resources; this design philosophy is known as *Representational State Transfer (REST)*. See (Wilde and Pautasso 2011).

[360][Web] Mosaic was developed in Joseph Hardin's lab at the National Center for Supercomputing Applications (NCSA), hosted by the University of Illinois, at Urbana/Champaign by Marc Andreesen, Eric Bina and a team of student programmers. Mosaic was initially developed on the Unix X Window System. See `http://www.ncsa.illinois.edu/Projects/mosaic.html`.

[361][Web] (Schatz and Hardin 1994).

[368][Com] Reachability is determined by calculating the transitive closure of the link matrix. A classic and well written explanation is (Agrawal, Borgida, and Jagadish 1989).

[370][Law] Shepard first put adhesive stickers into case books, then published lists of cases and their citations. Shepardizing is a big business for Lexis/Nexis and Westlaw (where the technique is called "KeyCite").

[372][Com] See (Watts 2004) for a detailed review of the theoretical foundations. See (Wu 2012) for applications in web-based social networks.

[373][Com] We are assuming a schema that establishes that the `name` attributes are of type ID and that the other attributes are of type IDREFS. This schema allows for polygamy, the possibility of multiple values for the `spouse` attribute. Restrictions on the number of spouses can be enforced with Schematron. (Also see the sidebar, Inclusions and References (page 456)).

[374][Ling] (Chomsky 1957) used these now famous sentences to motivate the distinction between syntax and semantics. He argued that since the probability in both cases that the words had previously occurred in this order was essentially zero, statistics of word occurrence could not be part of language knowledge. See. `http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously`.

[375][Web] (Berners-Lee, Hendler, and Lassila, 2001) is the classic paper, and (Shadbolt, Hall, and Berners-Lee 2006) is something of a revisionist history.

Ironically, the web was not semantic originally because Berners-Lee implemented web documents using a presentation-oriented HTML markup language. Designing HTML to be conceptually simple and easy to implement led to its rapid adoption. HTML documents can make assertions and describe relationships using `REL` and `REV` attributes, but browsers still do not provide useful interactions for link relations.

[376][Com] For example, Protégé a free, open-source platform with a suite of tools to construct domain models and knowledge-based applications with ontologies. (See http://protege.stanford.edu/)

[377][Web] See http://linkeddata.org/ and §9.3.3 Syntax (page 467).

Chapter 7

Categorization: Describing Resource Classes and Types

Robert J. Glushko
Rachelle Annechino
Jess Hemerly
Robyn Perry
Longhao Wang

7.1 Introduction

For nearly two decades, a TV game show called *Pyramid* aired in North America. The show featured two competing teams, each team consisting of two contestants: an ordinary civilian contestant and a celebrity. In the show's first round, both teams' members viewed a pyramid-shaped sign that displayed six category titles, some straightforward like "Where You Live" and others less conventional like "Things You Need to Feed." Each team then had an opportunity to compete for points in 30-second turns. The goal was for one team member to gain points by identifying a word or phrase related to the category from clues provided by the other team member. For example, a target phrase for the "Where You Live" category might be "zip code," and the clue might be "Mine is 94705." "Things you Need to Feed" might include both "screaming baby" and "parking meter."

The team that won the first round advanced to the "Winner's Circle," where the game was turned around. This time, only the clue giver was shown the category name and had to suggest concepts or instances belonging to that category so that the teammate could guess the category name. Clues like "alto," "soprano," and "tenor" would be given to prompt the teammate to guess "Singing Voices" or "Types of Singers."

As the game progressed, the categories became more challenging. It was interesting and entertaining to hear the clue receiver's initial guess and how subsequent guesses changed with more clues. The person giving clues would often become frustrated, because to them their clues seemed obvious and discriminating but would seem not to help the clue receivers in identifying the category. Viewers enjoyed sharing in these moments of vocabulary and category confusion.

The *Pyramid* TV game show developers created a textbook example for teaching about categories—groups or classes of things, people, processes, events or anything else that we treat as equivalent—and categorization—the process of assigning instances to categories. The game is a useful analog for us to illustrate many of the issues we discuss in this chapter. The Pyramid game was challenging, and sometimes comical, because people bring their own experiences and biases to understanding what a category means, and because not every instance of a category is equally typical or suggestive. How we organize reflects our thinking processes, which can inadvertently reveal personal characteristics that can be amusing in a social context. Hence, the popularity of the *Pyramid* franchise, which began on CBS in 1973 and has been produced in 20 countries.

Many texts in library science introduce categorization via cataloging rules, a set of highly prescriptive methods for assigning resources to categories that some describe and others satirize as "mark 'em and park 'em." Many texts in computer science discuss the process of defining the categories needed to create, process, and store information in terms of programming language constructs: "here's how to define an abstract type, and here's the data type system." Machine learning and *data science* texts explain how categories are created through statistical analysis of the correlations among the values of features in a collection or dataset. We take a very different approach in this chapter, but all of these different perspectives will find their place in it.[386][CogSci]

Navigating This Chapter

In the following sections, we discuss how and why we create categories, reviewing some important work in philosophy, linguistics, and cognitive psychology to better understand how categories are created and used in organizing systems. We discuss how the way we organize differs when we act as individuals or as members of social, cultural, or institutional groups (§7.2); later we share principles for creating categories(§7.3), design choices (§7.4), and implementation experience (§7.5). Throughout the chapter, we will compare how categories created by people compare with those created by computer algorithms. As usual, we close the chapter with a summary of the key points (§7.6).

7.2 The What and Why of Categories

Categories are *equivalence classes*, sets or groups of things or abstract entities that we treat the same. This does not mean that every instance of a category is identical, only that from some perspective, or for some purpose, we are treating them as equivalent based on what they have in common. When we consider something as a member of a category, we are making choices about which of its properties or roles we are focusing on and which ones we are ignoring. We do this automatically and unconsciously most of the time, but we can also do it in an explicit and self-aware way. When we create categories with conscious effort, we often say that we are creating a model, or just modeling. You should be familiar with the idea that a model is a set of simplified descriptions or a physical representation that removes some complexity to emphasize some features or characteristics and to de-emphasize others.[387][CogSci]

When we encounter objects or situations, recognizing them as members of a category helps us know how to interact with them. For example, when we enter an unfamiliar building we might need to open or pass through an entryway that we recognize as a door. We might never have seen that particular door before, but it has properties and affordances that we know that all doors have; it has a doorknob or a handle; it allows access to a larger space; it opens and closes. By mentally assigning this particular door to the "doors" category we distinguish it from "windows," a category that also contains objects that sometimes have handles and that open and close, but which we do not normally pass through to enter another space. Categorization judgments are therefore not just about what is included in a class, but also about what is excluded from a class. Nevertheless, the category boundaries are not sharp; a "Dutch door" is divided horizontally in half so that the bottom can be closed like a door while the top can stay open like a window.

Categories are *cognitive and linguistic models* for applying prior knowledge; creating and using categories are essential human activities. Categories enable us to relate things to each other in terms of similarity and dissimilarity and are involved whenever we perceive, communicate, analyze, predict, or classify. Without categories, we would perceive the world as an unorganized blur of things with no understandable or memorable relation to each other. Every wall-entry we encounter would be new to us, and we would have to discover its properties and supported interactions as though we had never before encountered a door. Of course, we still often need to identify something as a particular instance, but categories enable us to understand how it is equivalent to other instances. We can interchangeably relate to something as specific as "the wooden door to the main conference room" or more generally as "any door."

All human languages and cultures divide up the world into categories. How and why this takes place has long been debated by philosophers, psychologists and anthropologists. One explanation for this differentiation is that people recognize structure in the world, and then create categories of things that "go together" or are somehow similar. An alternative view says that human minds make sense of the world by imposing structure on it, and that what goes together or seems similar is the outcome rather than a cause of categorization. Bulmer framed the contrast in a memorable way by asking which came first, the chicken (the objective facts of nature) or the egghead (the role of the human intellect).[388][CogSci]

A secondary and more specialized debate going on for the last few decades among linguists, cognitive scientists, and computer scientists concerns the extent to which the cognitive mechanisms involved in category formation are specialized for that purpose rather than more general learning processes.[389][CogSci]

Even before they can talk, children behave in ways that suggest they have formed categories based on shape, color, and other properties they can directly perceive in physical objects.[390][CogSci] People almost effortlessly learn tens of thousands of categories embodied in the culture and language in which they grow up. People also rely on their own experiences, preferences, and goals to adapt these *cultural categories* or create entirely individual ones that they use to organize resources that they personally arrange. Later on, through situational training and formal education, people learn to apply systematic and logical thinking processes so that they can create and understand categories in engineering, logistics, transport, science, law, business, and other institutional contexts.

These three contexts of *cultural, individual,* and *institutional categorization* share some core ideas but they emphasize different processes and purposes for creating categories, so they are a useful distinction.[391][CogSci] Cultural categorization can be understood as a natural human cognitive ability that serves as a foundation for both informal and formal organizing systems. Individual categori-

zation tends to grow spontaneously out of our personal activities. Institutional categorization responds to the need for formal coordination and cooperation within and between companies, governments, and other goal-oriented enterprises.

In contrast to these three categorization contexts in which categories are created by people, *computational* categories are created by computer programs for information retrieval, machine learning, predictive analytics, and other applications. Computational categories are similar to those created by people in some ways but differ substantially in other ways.

7.2.1 Cultural Categories

Cultural categories are the archetypical form of categories upon which individual and institutional categories are usually based. Cultural categories tend to describe our everyday experiences of the world and our accumulated cultural knowledge. Such categories describe objects, events, settings, internal experiences, physical orientation, relationships between entities, and many other aspects of human experience. Cultural categories are learned primarily, with little explicit instruction, through normal exposure of children with their caregivers; they are associated with language acquisition and language use within particular cultural contexts.

Two thousand years ago Plato wrote that living species could be identified by "carving nature at its joints," the natural boundaries or discontinuities between types of things where the differences are the largest or most salient. Plato's metaphor is intuitively appealing because we can easily come up with examples of perceptible properties or behaviors of physical things that go together that make some ways of categorizing them seem more natural than others.[392][Phil]

Natural languages rely heavily on nouns to talk about categories of things because it is useful to have a shorthand way of referring to a set of properties that co-occur in predictable ways.[393][Ling] For example, in English (borrowed from Portuguese) we have a word for "banana" because a particular curved shape, greenish-yellow or yellow color, and a convenient size tend to co-occur in a familiar edible object, so it became useful to give it a name. The word "banana" brings together this configuration of highly interrelated perceptions into a unified concept so we do not have to refer to bananas by listing their properties.[394][CogSci]

Languages differ a great deal in the words they contain and also in more fundamental ways that they require speakers or writers to attend to details about the world or aspects of experience that another language allows them to ignore. This idea is often described as *linguistic relativity*. (See the sidebar, Linguistic Relativity (page 328).)

Linguistic Relativity

Linguistic diversity led Benjamin Whorf, in the mid-20th century, to propose an overly strong statement of the relationships among language, culture, and thought. Whorf argued that the particularities of one's native language determine how we think and what we can think about. Among his extreme ideas was the suggestion that, because some Native American languages lacked words or grammatical forms that refer to what we call "time" in English, they could not understand the concept. More careful language study showed both parts of the claim to be completely false.

Nevertheless, even though academic linguists have discredited strong versions of Whorf's ideas, less deterministic versions of *linguistic relativity* have become influential and help us understand cultural categorization. The more moderate position was crisply characterized by Roman Jakobson, who said that "languages differ essentially in what they *must* convey and not in what they *may* convey." In English one can say "I spent yesterday with a neighbor." In languages with grammatical gender, one must choose a word that identifies the neighbor as male or female.[395][Ling]

For example, speakers of the Australian aboriginal language, Guugu Yimithirr, do not use concepts of left and right, but rather use cardinal directions. Where in English we might say to a person facing north, "Take a step to your left," they would use their term for west. If the person faced south, we would change our instruction to "right," but they would still use their term for west. Imagine how difficult it would be for a speaker of Guugu Yimithirr and a speaker of English to collaborate in organizing a storage room or a closet.[396][CogSci]

It is not controversial to notice that different cultures and language communities have different experiences and activities that give them contrasting knowledge about particular domains. No one would doubt that university undergraduates in Chicago would think differently about animals than inhabitants of Guatemalan rain forests, or even that different types of "tree experts" (taxonomists, landscape workers, foresters, and tree maintenance personnel) would categorize trees differently.[397][CogSci]

On the other hand, despite the wide variation in the climates, environments, and cultures that produce them, at a high level "folk taxonomies" that describe natural phenomena are surprisingly consistent around the world. Half a century ago the sociologists Emile Durkheim and Marcel Mauss observed that the language and structure of folk taxonomies mirrors that of human family relationships (e.g., different types of trees might be "siblings," but animals would be part of another family entirely). They suggested that framing the world in terms of familiar human relationships allowed people to understand it more easily.[398][CogSci]

Anthropologist Brent Berlin, a more recent researcher, concurs with Durkheim and Mauss's observation that kinship relations and folk taxonomies are related, but argues that humans patterned their family structures after the natural world, not the other way around.[399][CogSci]

Invoking the Whorfian Hypothesis in a Clothing Ad

An advertisement for the "66 North" clothing brand invokes the Whorfian hypothesis to suggest that even though Icelanders have more than a hundred words for snow there is only one kind of winter clothing that matters to them; the kind that carries this brand name.

(Photo by R. Glushko. Taken in the Reykjavik airport.)

7.2.2 Individual Categories

Individual categories are created in an organizing system to satisfy the *ad hoc* requirements that arise from a person's unique experiences, preferences, and resource collections. Unlike cultural categories, which usually develop slowly and last a long time, individual categories are created by intentional activity, in response to a specific situation, or to solve an emerging organizational challenge. As a consequence, the categories in individual organizing systems generally have short lifetimes and rarely outlive the person who created them.

Individual categories draw from cultural categories but differ in two important ways. First, individual categories sometimes have an imaginative or metaphorical basis that is meaningful to the person who created them but which might distort or misinterpret cultural categories. Second, individual categories are often specialized or synthesized versions of cultural categories that capture particular experiences or personal history. For example, a person who has lived in China and Mexico, or lived with people from those places, might have highly individualized categories for foods they like and dislike that incorporate characteristics of both Chinese and Mexican cuisine.

Individual categories in organizing systems also reflect the idiosyncratic set of household goods, music, books, website bookmarks, or other resources that a person might have collected over time. The organizing systems for financial records, personal papers, or email messages often use highly specialized categories that are shaped by specific tasks to be performed, relationships with other people, events of personal history, and other highly individualized considerations. Put another way, individual categories are used to organize resource collections that are likely not representative samples of all resources of the type being collected. If everyone had the same collection of music, books, clothes, or toys the world would be a boring place.

Traditionally, *individual categorization* systems were usually not visible to, or shared with, others, whereas, this has become an increasingly common situation for people using web-based organizing system for pictures, music, or other personal resources. On websites like the popular Flickr, Instagram, and YouTube sites for photos and videos, people typically use existing cultural categories to tag their content as well as individual ones that they invent.[401][Ling]

7.2.3 Institutional Categories

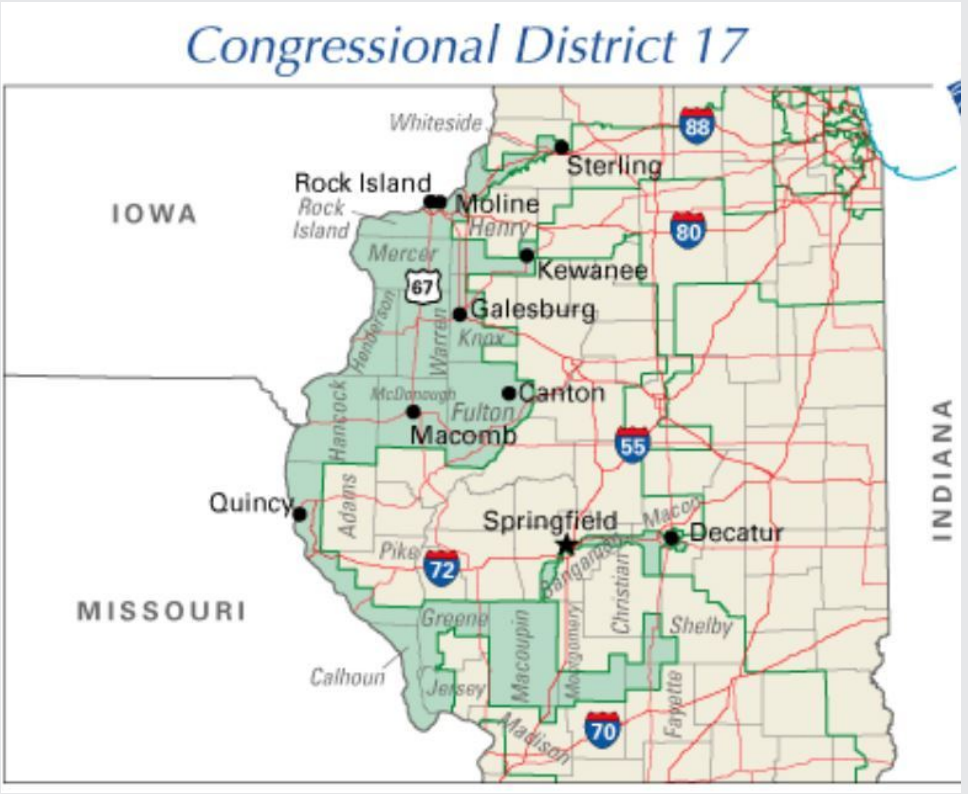
In contrast to cultural categories that are created and used implicitly, and to individual categories that are used by people acting alone, *institutional categories* are created and used explicitly, and most often by many people in coordination with each other. Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories are needed to regulate and systematize activity, to enable information sharing and reuse, and to reduce transaction costs. Furthermore, instead of describing the world as it is, institutional categories are usually defined to change or control the world by imposing semantic models that are more formal and arbitrary than those in cultural categories. Laws, regulations, and standards often specify institutional categories, along with decision rules for assigning resources to new categories, and behavior rules that prescribe how people must interact with them. The rigorous definition of institutional categories enables *classification:* the systematic assignment of resources to categories in an organizing system.[402][Law]

Creating institutional categories by more systematic processes than cultural or individual categories does not ensure that they will be used in systematic and rational ways, because the reasoning and rationale behind institutional categories might be unknown to, or ignored by, the people who use them. Likewise, this way of creating categories does not prevent them from being biased. Indeed, the goal of institutional categories is often to impose or incentivize biases in interpretation or behavior. There is no better example of this than the practice of gerrymandering, designing the boundaries of election districts to give one political party or ethnic group an advantage.[403][Ling](See the sidebar, Gerrymandering the Illinois 17th Congressional District (page 332).)

Institutional categorization stands apart from individual categorization primarily because it invariably requires significant efforts to reconcile mismatches between existing individual categories, where those categories embody useful working or *contextual knowledge* that is lost in the move to a formal institutional system.[404][Bus]

Institutional categorization efforts must also overcome the vagueness and inconsistency of cultural categories because the former must often conform to stricter logical standards to support inference and meet legal requirements. Furthermore, institutional categorization is usually a process that must be accounted for in a budget and staffing plans. While some kinds of institutional categories can be devised or discovered by computational processes, most of them are created through the collaboration of many individuals, typically from various parts of an organization or from different firms. For example, with the gerrymandering case we just discussed, it is important to emphasize that the inputs

Gerrymandering the Illinois 17th Congressional District



The 17th Congressional District in Illinois was dubbed "the rabbit on a skateboard" from 2003 through 2013 because of its highly contorted shape. The bizarre boundary was negotiated to create favorable voting constituencies for two incumbent legislators from opposing parties.

(Picture from nationatlas.gov. Not protectable by copyright (17 USC Sec. 105).)

to these programs and the decisions about districting are controlled by people, which is why the districts are institutional categories; the programs are simply tools that make the process more efficient. 405[Bus]

The different business or technical perspectives of the participants are often the essential ingredients in developing robust categories that can meet carefully identified requirements. And as requirements change over time, institutional categories must often change as well, implying version control, compliance testing, and other formal maintenance and governance processes.

Some institutional categories that initially had narrow or focused applicability have found their way into more popular use and are now considered cultural categories. A good example is the periodic table in chemistry, which Mendeleev developed in 1869 as a new system of categories for the chemical elements. The periodic table proved essential to scientists in understanding their properties and in predicting undiscovered ones. Today the periodic table is taught in elementary schools, and many things other than elements are commonly arranged using a graphical structure that resembles the periodic table of elements in chemistry, including sci-fi films and movies, desserts, and superheroes.[406][CogSci]

Stop and Think: Color

Think of the very broad category of "color." What are a few examples of a "cultural" category of color? How about an "individual" one? And an "institutional" one?

7.2.4 A "Categorization Continuum"

As we have seen, the concepts of cultural, individual, and institutional categorization usefully distinguish the primary processes and purposes when people create categories. However, these three kinds of categories can fuse, clash, and recombine with each other. Rather than viewing them as having precise boundaries, we might view them as regions on a continuum of categorization activities and methods.

Consider a few different perspectives on categorizing animals as an example. Scientific institutions categorize animals according to explicit, principled classification systems, such as the Linnaean taxonomy that assigns animals to a phylum, class, order, family, genus and species. Cultural categorization practices cannot be adequately described in terms of a master taxonomy, and are more fluid, converging with principled taxonomies sometimes, and diverging at other times. While human beings are classified within the animal kingdom in biological classification systems, people are usually not considered animals in most cultural contexts. Sometimes a scientific designation for human beings, *homo sapiens* is even applied to human beings in cultural contexts, since the genus-species taxonomic designation has influenced cultural conceptions of people and (other) animals over the years.

Animals are also often culturally categorized as pets or non-pets. The category "pets" commonly includes dogs, cats, and fish. A pet cat might be categorized at multiple levels that incorporate individual, cultural, and institutional perspectives on categorization—as an "animal" (cultural/institutional), as a "mammal" (institutional), as a "domestic short-hair" (institutional) as a "cat" (cultural), and as a "troublemaker" or a "favorite" (individual), among other possibilities, in addition to being identified individually by one or more pet names. Furthermore, not everyone experiences pets as just dogs, cats and fish. Some people have rel-

atively unusual pets, like pigs. For individuals who have pet pigs or who know people with pet pigs, “pigs” may be included in the “pets” category. If enough people have pet pigs, eventually “pigs” could be included in mainstream culture’s pet category.

Categorization skewed toward cultural perspectives incorporate relatively traditional categories, such as those learned implicitly from social interactions, like mainstream understandings of what kinds of animals are “pets,” while categorization skewed toward institutional perspectives emphasizes explicit, formal categories, like the categories employed in biological classification systems.

7.2.5 Computational Categories

Computational categories are created by computer programs when the number of resources, or when the number of descriptions or observations associated with each resource, are so large that people cannot think about them effectively. Computational categories are created for information retrieval, predictive analytics, and other applications where information scale or speed requirements are critical. The resulting categories are similar to those created by people in some ways but differ substantially in other ways.

The simplest kind of computational categories can be created using descriptive statistics (see §3.3.4). Descriptive statistics do not identify the categories they create by giving them familiar cultural or institutional labels. Instead, they create implicit categories of items according to how much they differ from the most typical or frequent ones. For example, in any dataset where the values follow the normal distribution, statistics of central tendency and dispersion serve as standard reference measures for any observation. These statistics identify categories of items that are very different or statistically unlikely outliers, which could be signals of measurement errors, poorly calibrated equipment, employees who are inadequately trained or committing fraud, or other problems. The “Six Sigma” methodology for process improvement and quality control rests on this idea that careful and consistent collection of statistics can make any measurable operation better.

Many text processing methods and applications use simple statistics to categorize words by their frequency in a language, in a collection of documents, or in individual documents, and these categories are exploited in many information retrieval applications (see §10.4.1 and §10.4.2).

Categories that people create and label also can be used more explicitly in computational algorithms and applications. In particular, a program that can assign an item or instance to one or more existing categories is called a classifier. The subfield of computer science known as *machine learning* is home to numerous techniques for creating classifiers by training them with already correctly categorized examples. This training is called *supervised learning*; it is supervised

CAFE Standards: Blurring the Lines Between Categorization Perspectives

The *Corporate Average Fuel Economy (CAFE)* standards sort vehicles into "passenger car" and "light truck" categories and impose higher minimum fuel efficiency requirements for cars because trucks have different typical uses.

When CAFE standards were introduced, the vehicles classified as light trucks were generally used for "light duty" farming and manufacturing purposes. "Light trucks" might be thought of as a "sort of" in-between category—a light truck is not really a car, but sufficiently unlike a prototypical truck to qualify the vehicle's categorization as "light." Formalizing this sense of in-between-ness by specifying features that define a "car" and a "light truck" is the only way to implement a consistent, transparent fuel efficiency policy that makes use of informal, graded distinctions between vehicles.

A manufacturer whose average fuel economy for all the vehicles it sells in a year falls below the CAFE standards has to pay penalties. This encourages them to produce "sport utility vehicles" (SUVs) that adhere to the CAFE definitions of light trucks but which most people use as passenger cars. Similarly, the PT Cruiser, a retro-styled hatchback produced by Chrysler from 2000-2010, strikes many people as a car. It looks like a car; we associate it with the transport of passengers rather than with farming; and in fact it is formally classified as a car under emissions standards. But like SUVs, in the CAFE classification system, the PT Cruiser is a light truck.

CAFE standards have evolved over time, becoming a theater for political clashes between holistic cultural categories and formal institutional categories, which plays out in competing pressures from industry, government, and political organizations. Furthermore, CAFE standards and manufacturers' response to them are influencing cultural categories, such that our cultural understanding of what a car looks like is changing over time as manufacturers design vehicles like the PT Cruiser with car functionality in unconventional shapes to take advantage of the CAFE light truck specifications.[407][Bus]

because it starts with instances labeled by category, and it involves learning because over time the classifier improves its performance by adjusting the weights for features that distinguish the categories. But strictly speaking, supervised learning techniques do not learn the categories; they implement and apply categories that they inherit or are given to them. We will further discuss the computational implementation of categories created by people in §7.5.

Supervised and Unsupervised Learning

Two subfields of *machine learning* that are relevant to organizing systems are *supervised* and *unsupervised* learning. In *supervised learning*, a machine learning program is trained with sample items or documents that are labeled by category, and the program learns to assign new items to the correct categories. In *unsupervised learning*, the program gets the same items but has to come up with the categories on its own by discovering the underlying correlations between the items; that is why unsupervised learning is sometimes called *statistical pattern recognition*.

In contrast, many computational techniques in machine learning can analyze a collection of resources to discover statistical regularities or correlations among the items, creating a set of categories without any labeled training data. This is called *unsupervised learning* or *statistical pattern recognition*. As we pointed out in §7.2.1 Cultural Categories (page 327), we learn most of our cultural categories without any explicit instruction about them, so it is not surprising that computational models of categorization developed by cognitive scientists often employ unsupervised statistical learning methods.

Many computational categories are like individual categories because they are tied to specific collections of resources or data and are designed to satisfy narrow goals. The individual categories you use to organize your email inbox or the files on your computer reflect your specific interests, activities, and personal network and are surely different than those of anyone else. Similarly, your credit card company analyzes your specific transactions to create computational categories of "likely good" and "likely fraudulent" that are different for every cardholder.

This focused scope is obvious when we consider how we might describe a computational category. "Fraudulent transaction for cardholder 4264123456780123" is not lexicalized with a one-word label as familiar cultural categories are. "Door" and "window" have broad scopes that are not tied to a single purpose. Put another way, the "door" and "window" cultural categories are highly reusable, as are institutional categories like those used to collect economic or health data that can be analyzed for many different purposes. The definitions of "door" and "window" might be a little fuzzy, but institutional categories are more precisely defined, often by law or regulation. Examples are the *North American Industry Classification System (NAICS)* from the US Census Bureau and the *United Nations Standard Products and Services Code (UNSPC)*.

A final contrast between categories created by people and those created computationally is that the former can almost always be inspected and reasoned about by other people, but only some of the latter can. A computational model that categorizes loan applicants as good or poor credit risks probably uses proper-

ties like age, income, home address, and marital status, so that a banker can understand and explain a credit decision. However, many other computational categories, especially those that created by clustering and deep learning techniques, are inseparable from the mathematical model that learned to use them, and as a result are uninterpretable by people.

A machine learning algorithm for classifying objects in images creates a complex multi-layer neural network whose features have no clear relationship to the categories, and this network has no other use. Put another way, machine learning programs are very general because they can be employed in any domain with high dimensional data, but what they learn cannot be applied in any other domain.

7.3 Principles for Creating Categories

§7.2 The What and Why of Categories (page 325) explained what categories are and the contrasting cultural, individual, and institutional contexts and purposes for which categories are created. In doing so, a number of different principles for creating categories were mentioned, mostly in passing.

We now take a systematic look at principles for creating categories, including: enumeration, single properties, multiple properties and hierarchy, probabilistic, similarity, and theory- and goal-based categorization. These ways of creating categories differ in the information and mechanisms they use to determine category membership.

7.3.1 Enumeration

The simplest principle for creating a category is *enumeration*; any resource in a finite or countable set can be deemed a category member by that fact alone. This principle is also known as *extensional definition*, and the members of the set are called the *extension*. Many institutional categories are defined by enumeration as a set of possible or legal values, like the 50 United States or the ISO currency codes (ISO 4217).

Enumerative categories enable membership to be unambiguously determined because a value like state name or currency code is either a member of the category or it is not. However, this clarity has a downside; it makes it hard to argue that something not explicitly mentioned in an enumeration should be considered a member of the category, which can make laws or regulations inflexible. Moreover, there comes a size when enumerative definition is impractical or inefficient, and the category either must be sub-divided or be given a definition based on principles other than enumeration.[408][Law]

For example, for millennia we earthlings have had a cultural category of “planet” as a “wandering” celestial object, and because we only knew of planets in

Too Many Planets to Enumerate: Keeping up with Kepler

Kepler is a space observatory launched by NASA in 2009 to search for Earth-like planets orbiting other stars in our own Milky Way galaxy. Kepler has already discovered and verified a few thousand new planets, and these results have led to estimates that there may be at least as many planets as there are stars, a few hundred billion in the Milky Way alone. Count fast.

our own solar system, the planet category was defined by enumeration: Mercury, Venus, Earth, Mars, Jupiter, and Saturn. When the outer planets of Uranus, Neptune, and Pluto were identified as planets in the 18th-20th centuries, they were added to this list of planets without any changes in the cultural category. But in the last couple of decades many heretofore unknown planets outside our solar system have been detected, making the set of planets unbounded, and definition by enumeration no longer works.

The *International Astronomical Union (IAU)* thought it solved this category crisis by proposing a definition of planet as "a celestial body that is (a) in orbit around a star, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighborhood around its orbit." Unfortunately, Pluto does not satisfy the third requirement, so it no longer is a member of the planet category, and instead is now called an "inferior planet."

Changing the definition of a significant cultural category generated a great deal of controversy and angst among ordinary non-scientific people. A typical headline was "Pluto's demotion has schools spinning," describing the outcry from elementary school students and teachers about the injustice done to Pluto and the disruption on the curriculum.

7.3.2 Single Properties

It is intuitive and useful to think in terms of properties when we identify instances and when we are describing instances (as we saw in §4.3 Resource Identity (page 182) and in Chapter 5, *Resource Description and Metadata*). Therefore, it should also be intuitive and useful to consider properties when we analyze more than one instance to compare and contrast them so we can determine which sets of instances can be treated as a category or *equivalence class*. Categories whose members are determined by one or more properties or rules follow the principle of *intensional definition*, and the defining properties are called the *intension*.

You might be thinking here that enumeration or extensional definition of a category is also a property test; is not "being a state" a property of California? But statehood is not a property precisely because "state" is defined by extension,

which means the only way to test California for statehood is to see if it is in the list of states.[410][Phil]

Any *single property* of a resource can be used to create categories, and the easiest ones to use are often the intrinsic static properties. As we discussed in Chapter 5, *Resource Description and Metadata*, intrinsic static properties are those inherent in a resource that never change. The material of composition of natural or manufactured objects is an intrinsic and static property that can be used to arrange physical resources. For example, an organizing system for a personal collection of music that is based on the intrinsic static property of physical format might use categories for CDs, DVDs, vinyl albums, 8-track cartridges, reel-to-reel tape and tape cassettes.[411][CogSci]

Using a single property is most natural to do when the properties can take on only a small set of discrete values like music formats, and especially when the property is closely related to how the resources are used, as they are with the music collection where each format requires different equipment to listen to the music. Each value then becomes a subcategory of the music category.

The author, date, and location of creation of an intellectual resource cannot be directly perceived but they are also intrinsic static properties. The subject matter or purpose of a resource, its "what it is about" or "what it was originally for," are also intrinsic static properties that are not directly perceivable, especially for information resources.

The name or identifier of a resource is often arbitrary but once assigned normally does not change, making it an extrinsic static property. Any collection of resources with alphabetic or numeric identifiers as an associated property can use sorting order as an organizing principle to arrange spices, books, personnel records, etc., in a completely reliable way. Some might argue whether this organizing principle creates a category system, or whether it simply exploits the ordering inherent in the identifier notation. For example, with alphabetic identifiers, we can think of alphabetic ordering as creating a recursive category system with 26 (A-Z) top-level categories, each containing the same number of second-level categories, and so on until every instance is assigned to its proper place.[412][CogSci]

Some resource properties are both extrinsic and dynamic because they are based on usage or behaviors that can be highly context-dependent. The current owner or location of a resource, its frequency of access, the joint frequency of access with other resources, or its current rating or preference with respect to alternative resources are typical extrinsic and dynamic properties that can be the basis for arranging resources and defining categories.

These properties can have a large number of values or are continuous measures, but as long as there are explicit rules for using property values to deter-

mine category assignment the resulting categories are still easy to understand and use. For example, we naturally categorize people we know on the basis of their current profession, the city where they live, their hobbies, or their age. Properties with a numerical dimension like "frequency of use" are often transformed into a small set of categories like "frequently used," "occasionally used," and "rarely used" based on the numerical property values.[413][CogSci]

While there are an infinite number of logically expressible properties for any resource, most of them would not lead to categories that would be interpretable and useful for people. If people are going to use the categories, it is important to base them on properties that are psychologically or pragmatically relevant for the resource domain being categorized. Whether something weighs more or less than 5000 pounds is a poor property to apply to things in general, because it puts cats and chairs in one category, and buses and elephants in another.[414][CogSci]

To summarize: The most useful single properties to use for creating categories for an organizing system used by people are those that are formally assigned, objectively measurable and orderable, or tied to well-established cultural categories, because the resulting categories will be easier to understand and describe.

If only a single property is used to distinguish among some set of resources and to create the categories in an organizing system, the choice of property is critical because different properties often lead to different categories. Using the age property, Bill Gates and Mark Zuckerberg are unlikely to end up in the same category of people. Using the wealth property, they most certainly would. Furthermore, if only one property is used to create a system of categories, any category with a large numbers of items in it will lack coherence because differences on other properties will be too apparent, and some category members will not fit as well as the others.

7.3.3 Multiple Properties

Organizing systems often use multiple properties to define categories. There are three different ways in which to do this that differ in the scope of the properties and how essential they are in defining the categories.

7.3.3.1 Multi-Level or Hierarchical Categories

If you have many shirts in your closet (and you are a bit compulsive or a "neat freak"), instead of just separating your shirts from your pants using a single property (the part of body on which the clothes are worn) you might arrange the shirts by style, and then by sleeve length, and finally by color. When all of the resources in an organizing system are arranged using the same sequence of re-

source properties, this creates a *logical hierarchy*, a multi-level category system.

If we treat all the shirts as the collection being organized, in the shirt organizing system the broad category of shirts is first divided by style into categories like "dress shirts," "work shirts," "party shirts," and "athletic or sweatshirts." Each of these style categories is further divided until the categories are very narrow ones, like the "white long-sleeve dress shirts" category. A particular shirt ends up in this last category only after passing a series of property tests along the way: it is a dress shirt, it has long sleeves, and it is white. Each test creates more precise categories in the intersections of the categories whose members passed the prior property tests.

Put another way, each subdivision of a category takes place when we identify or choose a property that differentiates the members of the category in a way that is important or useful for some intent or purpose. Shirts differ from pants in the value of the "part of body" property, and all the shirt subcategories share this "top part" value of that property. However, shirts differ on other properties that determine the subcategory to which they belong. Even as we pay attention to these differentiating properties, it is important to remember the other properties, the ones that members of a category at any level in the hierarchy have in common with the members of the categories that contain it. These properties are often described as "inherited" or "inferred" from the broader category.[415][Com] For example, just as every shirt shares the "worn on top part of body" property, every item of clothing shares the "can be worn on the body" property, and every resource in the "shirts" and "pants" category inherits that property.

Each differentiating property creates another level in the category hierarchy, which raises an obvious question: How many properties and levels do we need? In order to answer this question we must reflect upon the shirt categories in our closet. Our organizing system for shirts arranges them with the three properties of style, sleeve length, and color; some of the categories at the lowest level of the resulting hierarchy might have only one member, or no members at all. You might have yellow or red short-sleeved party shirts, but probably do not have yellow or red long-sleeved dress shirts, making them empty categories. Obviously, any category with only one member does not need any additional properties to tell the members apart, so a category hierarchy is logically complete if every resource is in a category by itself.

However, even when the lowest level categories of our shirt organizing system have more than one member, we might choose not to use additional properties to subdivide it because the differences that remain among the members do not matter to us for the interactions the organizing system needs to support. Suppose we have two long-sleeve white dress shirts from different shirt makers, but whenever we need to wear one of them, we ignore this property. Instead, we

just pick one or the other, treating the shirts as completely equivalent or substitutable. When the remaining differences between members of a category do not make a difference to the users of the category, we can say that the organizing system is pragmatically or practically complete even if it is not yet logically complete. That is to say, it is complete "for all intents and purposes." Indeed, we might argue that it is desirable to stop subdividing a system of categories while there are some small differences remaining among the items in each category because this leaves some flexibility or logical space in which to organize new items. This point might remind you of the concept of overfitting, where models with many parameters can very accurately fit their training data, but as a result generalize less well to new data. (See §5.3.2.5.)

On the other hand, consider the shirt section of a big department store. Shirts there might be organized by style, sleeve length, and color as they are in our home closet, but would certainly be further organized by shirt maker and by size to enable a shopper to find a Marc Jacobs long-sleeve blue dress shirt of size 15/35. The department store organizing system needs more properties and a deeper hierarchy for the shirt domain because it has a much larger number of shirt instances to organize and because it needs to support many shirt shoppers, not just one person whose shirts are all the same size.

7.3.3.2 Different Properties for Subsets of Resources

A different way to use multiple resource properties to create categories in an organizing system is to employ different properties for distinct subsets of the resources being organized. This contrasts with the strict multi-level approach in which every resource is evaluated with respect to every property. Alternatively, we could view this principle as a way of organizing multiple domains that are conceptually or physically adjacent, each of which has a separate set of categories based on properties of the resources in that domain. This principle is used for most folder structures in computer file systems and by many email applications; you can create as many folder categories as you want, but any resource can only be placed in one folder.

The contrasts between intrinsic and extrinsic properties, and between static and dynamic ones, are helpful in explaining this method of creating organizing categories. For example, you might organize all of your clothes using intrinsic static properties if you keep your shirts, socks, and sweaters in different drawers and arrange them by color; extrinsic static properties if you share your front hall closet with a roommate, so you each use only one side of that closet space; intrinsic dynamic properties if you arrange your clothes for ready access according to the season; and, extrinsic dynamic properties if you keep your most frequently used jacket and hat on a hook by the front door.[416][Bus]

Classifying Hawaiian "Boardshorts"

The swimsuits worn by surfers, called "boardshorts," have evolved from purely functional garments to symbols of extreme sports and the Hawaiian lifestyle. A 2012 exhibition at the Honolulu Museum of Art captured the diversity of boardshorts on three facets: their material, how they fastened around the surfer's fly and waist, and their length.

(Photo by R. Glushko.)

If we relax the requirement that different subsets of resources use different organizing properties and allow any property to be used to describe any resource, the loose organizing principle we now have is often called *tagging*. Using any property of a resource to create a description is an uncontrolled and often unprincipled principle for creating categories, but it is increasingly popular for organizing photos, web sites, email messages in gmail, or other web-based resources. We discuss tagging in more detail in §5.2.2.3 Tagging of Web-based Resources (page 222).

A Supermarket Map

A typical supermarket embodies a surprisingly complex classification system. Each section of the store employs a different set of properties to arrange its resources, and some properties such as perishability and onsite preparation are important in more than one section.

(Photo by R. Glushko.)

7.3.3.3 Necessary and Sufficient Properties

A large set of resources does not always require many properties and categories to organize it. Some types of categories can be defined precisely with just a few *essential* properties. For example, a prime number is a positive integer that has no divisors other than 1 and itself, and this category definition perfectly distinguishes prime and not-prime numbers no matter how many numbers are being categorized. "Positive integer" and "divisible only by 1 and itself" are *necessary* or *defining* properties for the prime number category; every prime number must satisfy these properties. These properties are also *sufficient* to establish membership in the prime number category; any number that satisfies the necessary properties is a prime number. Categories defined by necessary and sufficient properties are also called *monothetic*. They are also sometimes called *classical*

categories because they conform to Aristotle's theory of how categories are used in logical deduction using syllogisms.[417][Phil] (See the sidebar, The Classical View of Categories (page 345).)

Theories of categorization have evolved a great deal since Plato and Aristotle proposed them over two thousand years ago, but in many ways we still adhere to classical views of categories when we create organizing systems because they can be easier to implement and maintain that way.

An important implication of necessary and sufficient category definition is that every member of the category is an equally good member or example of the category; every prime number is equally prime. Institutional category systems often employ necessary and sufficient properties for their conceptual simplicity and straightforward implementation in *decision trees*, database *schemas*,and programming language *classes*.

The Classical View of Categories

The classical view is that categories are defined by necessary and sufficient properties. This theory has been enormously influential in Western thought, and is embodied in many organizing systems, especially those for information resources. However, as we will explain, we cannot rely on this principle to create categories in many domains and contexts because there are not necessary and sufficient properties. As a result, many psychologists, cognitive scientists, and computer scientists who think about categorization have criticized the classical theory.

We think this is unfair to Aristotle, who proposed what we now call the classical theory primarily to explain how categories underlie the logic of deductive reasoning: All men are mortal; Socrates is a man; Therefore, Socrates is mortal. People are wrong to turn Aristotle's thinking around and apply it to the problem of inductive reasoning, how categories are created in the first place. But this is not Aristotle's fault; he was not trying to explain how natural cultural categories arise.

Consider the definition of an address as requiring a street, city, governmental region, and postal code. Anything that has all of these *information components* is therefore considered to be a valid address, and anything that lacks any of them will not be considered to be a valid address. If we refine the properties of an address to require the governmental region to be a state, and specifically one of the United States Postal Service's list of official state and territory codes, we create a subcategory for US addresses that uses an enumerated category as part of its definition. Similarly, we could create a subcategory for Canadian addresses by exchanging the name "province" for state, and using an enumerated list of Canadian province and territory codes.

7.3.4 The Limits of Property-Based Categorization

Property-based categorization works tautologically well for categories like "prime number" where the category is defined by necessary and sufficient properties. Property-based categorization also works well when properties are conceptually distinct and the value of a property is easy to perceive and examine, as they are with man-made physical resources like shirts.

Historical experience with organizing systems that need to categorize information resources has shown that basing categories on easily perceived properties is often not effective. There might be indications "on the surface" that suggest the "joints" or boundaries between types of information resources, but these are often just presentation or packaging choices, That is to say, neither the size of a book nor the color of its cover are reliable cues for what it contains. Information resources have numerous descriptive properties like their title, author, and publisher that can be used more effectively to define categories, and these are certainly useful for some kinds of interactions, like finding all of the books written by a particular author or published by the same publisher. However, for practical purposes, the most useful property of an information resource is its *aboutness*, which may not be objectively perceivable and which is certainly hard to characterize. Any collection of information resources in a library or document filing system is likely to be about many subjects and topics, and when an individual resource is categorized according to a limited number of its content properties, it is at the same time not being categorized using the others.

When the web first started, there were many attempts to create categories of web sites, most notably by Yahoo! As the web grew, it became obvious that search engines would be vastly more useful because their near real-time text indexes obviate the need for *a priori* assignment of web pages to categories. Rather, web search engines represent each web page or document in a way that treats each word or term they contain as a separate property.

Considering every distinct word in a document stretches our notion of property to make it very different from the kinds of properties we have discussed so far, where properties were being explicitly used by people to make decisions about category membership and resource organization. It is just not possible for people to pay attention to more than a few properties at the same time even if they want to, because that is how human perceptual and cognitive machinery works. But computers have no such limitations, and algorithms for information retrieval and machine learning can use huge numbers of properties, as we will see later in this chapter and in Chapter 8 and Chapter 10.

Classifying the Web: Yahoo! in 1996

Their goal was to manually assign every web page to a category.

(Screenshot by R. Glushko. Source: Internet Archive wayback machine.)

7.3.5 Probabilistic Categories and "Family Resemblance"

As we have seen, some categories can be precisely defined using necessary and sufficient features, especially when the properties that determine category membership are easy to observe and evaluate. Something is either a prime number or it isn't. A person cannot be a registered student and not registered at the same time.

However, categorization based on explicit and logical consideration of properties is much less effective, and sometimes not even possible for domains where properties lack one or more of the characteristics of separability, perceptibility, and necessity. Instead, we need to categorize using properties in a probabilistic or statistical way to come up with some measure of resemblance or similarity between the resource to be categorized and the other members of the category.

Consider a familiar category like "bird." All birds have feathers, wings, beaks, and two legs. But there are thousands of types of birds, and they are distinguished by properties that some birds have that other birds lack: most birds can fly, most are active in the daytime, some swim, some swim underwater; some have webbed feet. These properties are correlated or clustered, a consequence of natural selection that conveys advantages to particular configurations of characteristics, and there are many different clusters; birds that live in trees have different wings and feet than those that swim, and birds that live in deserts have different colorations and metabolisms that those that live near water. So instead of being defined by a single set of properties that are both necessary and sufficient, the bird category is defined probabilistically, which means that decisions about category membership are made by accumulating evidence from the properties that are more or less characteristic of the category.

Categories of information resources often have the same probabilistic character. The category of spam messages is suggested by the presence of particular words (beneficiary, pharmaceutical) but these words also occur in messages that are not spam. A spam classifier uses the probabilities of each word in a message in spam and non-spam contexts to calculate an overall likelihood that the message is spam.

There are three related consequences for categories when their characteristic properties have a probabilistic distribution:

- The first is an effect of *typicality* or *centrality* that makes some members of the category better examples than others. Membership in probabilistic categories is not all or none, so even if they share many properties, an instance that has more of the characteristic properties will be judged as better or more typical.[419][CogSci] Try to define "bird" and then ask yourself if all of the things you classify as birds are equally good examples of the category (look

at the six birds in Family Resemblance and Typicality (page 350)). This effect is also described as *gradience* in category membership and reflects the extent to which the most characteristic properties are shared.

- A second consequence is that the sharing of some but not all properties creates what we call *family resemblances* among the category members; just as biological family members do not necessarily all share a single set of physical features but still are recognizable as members of the same family. This idea was first proposed by the 20th-century philosopher Ludwig Wittgenstein, who used "games" as an example of a category whose members resemble each other according to shifting property subsets.[420][Phil]
- The third consequence, when categories do not have necessary features for membership, is that the boundaries of the category are not fixed; the category can be stretched and new members assigned as long as they resemble incumbent members. Personal video games and multiplayer online games like World of Warcraft did not exist in Wittgenstein's time but we have no trouble recognizing them as games and neither would Wittgenstein, were he alive. Recall that in Chapter 1 we pointed out that the cultural category of "library" has been repeatedly extended by new properties, as when Flickr is described as a web-based photo-sharing library. Categories defined by family resemblance or multiple and shifting property sets are termed *polythetic*.

What Is a Game?

Ludwig Wittgenstein (1889-1951) was a philosopher who thought deeply about mathematics, the mind, and language. In 1999, his *Philosophical Investigations* was ranked as the most important book of 20th-century philosophy in a poll of philosophers.[421][Phil] In that book, Wittgenstein uses "game" to argue that many concepts have no defining properties, and that instead there is a "complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail." He contrasts board games, card games, ball games, games of skill, games of luck, games with competition, solitary games, and games for amusement. Wittgenstein notes that not all games are equally good examples of the category, and jokes about teaching children a gambling game with dice because he knows that this is not the kind of game that the parents were thinking of when they asked him to teach their children a game.[422][Phil]

We conclude that instead of using properties one at a time to assign category membership, we can use them in a composite or integrated way where together a co-occurring cluster of properties provides evidence that contributes to a *similarity* calculation. Something is categorized as an A and not a B if it is more similar to A's best or most typical member rather than it is to B's.[423][CogSci]

Family Resemblance and Typicality

These six animals have some physical features in common but not all of them, yet they resemble each other enough to be easily recognizable as birds. Most people consider a pigeon to be a more typical bird than a penguin.

A penguin, a pigeon, a swan, a stork, a flamingo, and a frigate bird. (Clockwise from top-left.)

(Photos by R. Glushko.)

7.3.6 Similarity

Similarity is a measure of the resemblance between two things that share some characteristics but are not identical. It is a very flexible notion whose meaning depends on the domain within which we apply it. Some people consider that the concept of similarity is itself meaningless because there must always be some basis, some unstated set of properties, for determining whether two things are similar. If we could identify those properties and how they are used, there would not be any work for a similarity mechanism to do.[424][CogSci]

To make similarity a useful mechanism for categorization we have to specify how the similarity measure is determined. There are four psychologically-motivated approaches that propose different functions for computing similarity: feature- or property-based, geometry-based, transformational, and alignment- or analogy-based. The big contrast here is between models that represent items as sets of properties or discrete conceptual features, and those that assume that properties vary on a continuous metric space.[425][CogSci]

7.3.6.1 Feature-based Models of Similarity

An influential model of feature-based similarity calculation is Amos Tversky's contrast model, which matches the features or properties of two things and computes a similarity measure according to three sets of features:

- those features they share,
- those features that the first has that the second lacks, and
- those features that the second has that the first lacks.

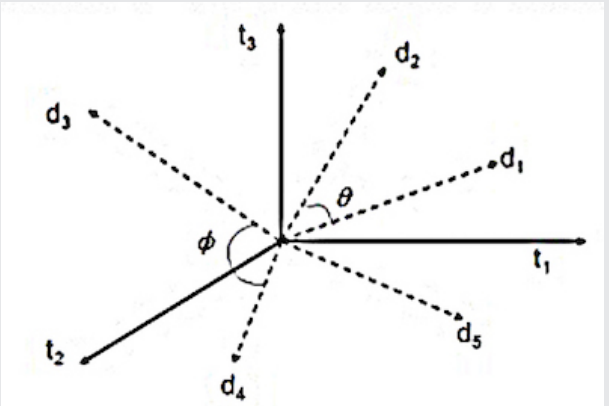
The similarity based on the shared features is reduced by the two sets of distinctive ones. The weights assigned to each set can be adjusted to explain judgments of category membership. Another commonly feature-based similarity measure is the Jaccard coefficient, the ratio of the common features to the total number of them. This simple calculation equals zero if there are no overlapping features and one if all features overlap. Jaccard's measure is often used to calculate document similarity by treating each word as a feature.[426][CogSci]

We often use a heuristic version of feature-based similarity calculation when we create multi-level or hierarchical category systems to ensure that the categories at each level are at the same level of abstraction or breadth. For example, if we were organizing a collection of musical instruments, it would not seem correct to have subcategories of "woodwind instruments," "violins," and "cellos" because the feature-based similarity among the categories is not the same for all pairwise comparisons among the categories; violins and cellos are simply too similar to each other to be separate categories given woodwinds as a category.

7.3.6.2 Geometric Models of Similarity

Geometric models are a type of similarity framework in which items whose property values are metric are represented as points in a multi-dimensional feature- or property-space. The property values are the coordinates, and similarity is calculated by measuring the distance between the items.

Document Similarity

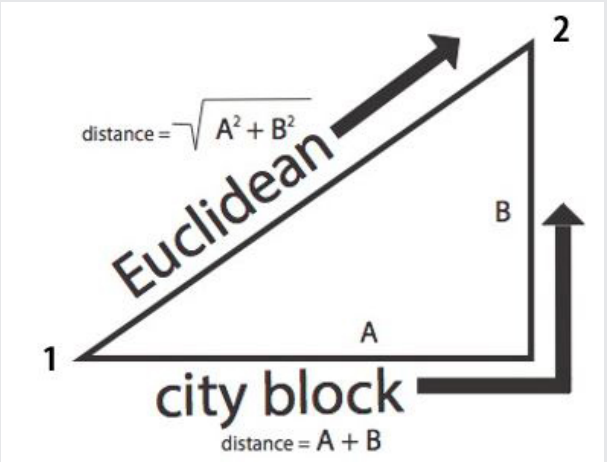


Documents represented as vectors in term space, with the angles between them as a measure of their similarity.

Geometric similarity functions are commonly used by search engines; if a query and document are each represented as a vector of search terms, relevance is determined by the distance between the vectors in the "term space." The simplified diagram in the sidebar, Document Similarity (page 352), depicts four documents whose locations in the term space are determined by how many of each of three terms they contain. The document vectors are normalized to length 1, which makes it possible to use the cosine of the angle between any two documents as a measure of their similarity. Documents d1 and d2 are more similar to each other than documents d3 and d4, because angle between the former pair (Θ) is smaller than the angle between the latter (Φ). We will discuss how this works in greater detail in Chapter 10, *Interactions with Resources*.

If the vectors that represent items in a multi-dimensional property space are of different lengths, instead of calculating similarity using cosines we need to calculate similarity in a way that more explicitly considers the differences on each dimension.

Geometric Distance Functions



The distance between points 1 and 2 depends on how the distance function combines the differences in values (A and B) on each dimension.

The diagram in the sidebar, Geometric Distance Functions (page 353) shows two different ways of calculating the distance between points 1 and 2 using the differences A and B. The Euclidean distance function takes the square root of the sum of the squared differences on each dimension; in two dimensions, this is the familiar Pythagorean Theorem to calculate the length of the hypotenuse of a right triangle, where the exponent applied to the differences is 2. In contrast, the City Block distance function, so-named because it is the natural way to measure distances in cities with "gridlike" street plans, simply adds up the differences on each dimension, which is equivalent to an exponent of 1.

We can interpret the exponent as a weighting function that determines the relative contribution of each property to the overall distance or similarity calculation. The choice of exponent depends on the type of properties that characterize a domain and how people make category judgments within it. The exponent of 1 in the City Block function ensures that each property contributes its full amount. As the exponent grows larger, it magnifies the impact of the properties on which differences are the largest.

The Chebyshev function takes this to the limit (where the exponent would be infinity) and defines the distance between two items as the difference of their values on the single property with the greatest difference. What this means in practice is that two items could have similar or even identical values on most properties, but if they differ much on just one property, they will be treated as very dissimilar. We can make an analogy to stereotyping or prejudice when a person is just like you in all ways except for the one property you view as negative, which then becomes the only one that matters to you.

At the other extreme, if the exponent is reduced to zero, this treats each property as binary, either present or absent, and the distance function becomes a count of the number of times that the value of the property for one item is different from the value for the other one. This is called the "Hamming distance."

7.3.6.3 Transformational Models of Similarity

Transformational models assume that the similarity between two things is inversely proportional to the complexity of the transformation required to turn one into the other. The simplest transformational model of similarity counts the number of properties that would need to change their values. More generally, one way to perform the *name matching* task of determining when two different strings denote the same person, object, or other named entity is to calculate the "edit distance" between them; the number of changes required to transform one into the other.

The simplest calculation just counts the number of insertion, deletion, and substitution operations and is called the Levenshtein distance; for example, the distance between "bob" and "book" is two: insert "o" and change the second "b" to "k". Two strings with a short edit distance might be variant spellings or misspellings of the same name, and transformational models that are sensitive to common typing errors like transposed or duplicated letters are very effective at spelling correction. Transformational models of similarity are also commonly used to detect plagiarism and duplicate web pages.[427][Com]

7.3.6.4 Alignment or Analogy Models of Similarity

None of the previous types of similarity models works very well when comparing things that have lots of internal or relational structure. In these cases, calculations based on matching features is insufficient; you need to compare features that align because they have the same role in structures or relationships. For example, a car with a green wheel and a truck with a green hood both share the feature green, but this matching feature does not increase their similarity much because the car's wheel does not align with the truck's hood. On the other hand, analogy lets us say that an atom is like the solar system. They have no common properties, but they share the relationship of having smaller objects revolving around a large one.

This kind of analogical comparison is especially important in problem solving. You might think that experts are good at solving problems in their domain of expertise because they have organized their knowledge and experience in ways that enable efficient search for and evaluation of possible solutions. For example, it is well known that chess masters search their memories of previous winning positions and the associated moves to decide what to play. However, top chess players also organize their knowledge and select moves on the basis of abstract similarities that cannot be explained in terms of specific positions of chess pieces. This idea that experts represent and solve problems at deeper levels than novices do by using more abstract principles or domain structure has been replicated in many areas. Novices tend to focus more on surface properties and rely more on literal similarity.[428][CogSci]

7.3.7 Goal-Derived Categories

Another psychological principle for creating categories is to organize resources that go together in order to satisfy a goal. Consider the category "Things to take from a burning house," an example that cognitive scientist Lawrence Barsalou termed an *ad hoc* or *goal-derived* category.[429][CogSci]

What things would you take from your house if a fire threatened it?? Possibly your cat, your wallet and checkbook, important papers like birth certificates and passports, and grandma's old photo album, and anything else you think is important, priceless, or irreplaceable—as long as you can carry it. These items have no discernible properties in common, except for being your most precious possessions. The category is derived or induced by a particular goal in some specified context.

Things Used at the Gym

A hand towel, a music player with headphones, and a bottle of water have no properties in common but they go together because they are members of the "things used at the gym when working out" category. This type of ad hoc or goal-derived category gave contestants trouble on the Pyramid game show.

(Photo by R. Glushko.)

7.3.8 Theory-Based Categories

A final psychological principle for creating categories is organizing things in ways that fit a theory or story that makes a particular categorization sensible. A *theory-based category* can win out even if probabilistic categorization, on the basis of *family resemblance* or *similarity* with respect to visible properties, would lead to a different category assignment. For example, a theory of phase change explains why liquid water, ice, and steam are all the same chemical compound even though they share few visible properties.

Theory-based categories based on origin or causation are especially important with highly inventive and computational resources because unlike natural kinds of physical resources, little or none of what they can do or how they behave is visible on the surface (see §3.4.1 Affordance and Capability (page 123)). Consider all of the different appearances and form factors of the resources that we categorize as "computers" —their essence is that they all compute, an invisible or theory-like principle that does not depend on their visible properties.[430][CogSci]

7.4 Category Design Issues and Implications

We have previously discussed the most important principles for creating categories: resource properties, similarity, and goals. When we use one or more of these principles to develop a system of categories, we must make decisions about its depth and breadth. Here, we examine the idea that some levels of abstraction in a system of categories are more basic or natural than others. We also consider how the choices we make affect how we create the organizing system in the first place, and how they shape our interactions when we need to find some resources that are categorized in it.

7.4.1 Category Abstraction and Granularity

We can identify any resource as a unique instance or as a member of a class of resources. The size of this class—the number of resources that are treated as equivalent—is determined by the properties or characteristics we consider when we examine the resources in some domain. The way we think of a resource domain depends on context and intent, so the same resource can be thought of abstractly in some situations and very concretely in others. As we discussed in Chapter 5, *Resource Description and Metadata*, this influences the nature and extent of resource description, and as we have seen in this chapter, it then influences the nature and extent of categories we can create.

Consider the regular chore of putting away clean clothes. We can consider any item of clothing as a member of a broad category whose members are any kind of garment that a person might wear. Using one category for all clothing, that is, failing to distinguish among the various items in any useful or practical way would likely mean that we would keep our clothes in a big unorganized pile.

However, we cannot wear any random combination of clothing items—we need a shirt, a pair of pants, socks, and so on. Clearly, our indiscriminate clothing category is too broad for most purposes. So instead, most people organize their clothes in more fine-grained categories that fit the normal pattern of how they wear clothes.

This tendency to use specific categories instead of broader ones is a general principle that reflects how people organize their experience when they see similar, but not identical, examples or events. This “size principle” for concept learning, as cognitive scientist Josh Tenenbaum describes it, is a preference for the most specific rules or descriptions that fit the observations. For example, if you visit a zoo and see many different species of animals, your conception of what you saw is different than if you visited a kennel that only contained dogs. You might say “I saw animals at the zoo,” but would be more likely to say “I saw dogs at the kennel” because using the broad “animal” category to describe your

kennel visit conveys less of what you learned from your observations there.[431][CogSci]

In §7.3.2 Single Properties (page 338) we described an organizing system for the shirts in our closet, so let us talk about socks instead. When it comes to socks, most people think that the basic unit is a pair because they always wear two socks at a time. If you are going to need to find socks in pairs, it seems sensible to organize them into pairs when you are putting them away. Some people might further separate their dress socks from athletic ones, and then sort these socks by color or material, creating a hierarchy of sock categories analogous to the shirt categories in our previous example.

Questions of resource abstraction and granularity also emerge whenever the information systems of different firms, or different parts of a firm, need to exchange information or be merged into a single system. All parties must define the identity of each thing in the same way, or in ways that can be related or mapped to each other either manually or electronically.

For example, how should a business system deal with a customer's address? Printed on an envelope, "an address" typically appears as a comprehensive, multi-line text object. Inside an information system, however, an address is best stored as a set of distinctly identifiable information components. This fine-grained organization makes it easier to sort customers by city or postal codes, for sales and marketing purposes. Incompatibilities in the abstraction and granularity of these information components, and the ways in which they are presented and reused in documents, will cause interoperability problems when businesses need to share information.[432][Com]

The *Universal Business Language (UBL)* (mentioned briefly in §8.1.5.2) is a library of information components designed to enable the creation of business document models that span a range of category abstraction. UBL comes equipped with XML schemas that define document categories like orders, invoices, payments, and receipts that many people are familiar with from their personal experiences of shopping and paying bills. However, UBL can also be used to design very specific or subordinate level transactional document types like "purchase order for industrial chemicals when buyer and seller are in different countries," or document types at the other end of the abstraction hierarchy like "fill-in-the-blank" legal forms for any kind of contract.

It might seem counterintuitive, but when a system of human-generated categories is too complex for people to interpret and apply reliably, computational classifiers that compute statistical similarity between new and already classified items can outperform people.[434][DS]

7.4.2 Basic or Natural Categories

Category abstraction is normally described in terms of a hierarchy of superordinate, basic, and subordinate category levels. “Clothing,” for example, is a superordinate category, “shirts” and “socks” are basic categories, and “white long-sleeve dress shirts” and “white wool hiking socks” are subordinate categories. Members of basic level categories like “shirts” and “socks” have many perceptual properties in common, and are more strongly associated with motor movements than members of superordinate categories. Members of subordinate categories have many common properties, but these properties are also shared by members of other subordinate categories at the same level of abstraction in the category hierarchy. That is, while we can identify many properties shared by all “white long-sleeve dress shirts,” many of them are also properties of “blue long-sleeve dress shirts” and “black long-sleeve pullover shirts.”

Psychological research suggests that some levels of abstraction in a system of categories are more basic or natural than others. Anthropologists have also observed that folk taxonomies invariably classify natural phenomena into a five- or six-level hierarchy, with one of the levels being the psychologically basic or “real” name (such as “cat” or “dog”), as opposed to more abstract names (e.g. “mammal”) that are used less in everyday life. An implication for organizing system design is that basic level categories are highly efficient in terms of the cognitive effort they take to create and use. A corollary is that classifications with many levels at different abstraction levels may be difficult for users to navigate effectively.[435[CogSci]]

7.4.3 The Recall / Precision Tradeoff

The abstraction level we choose determines how precisely we identify resources. When we want to make a general claim, or communicate that the scope of our interest is broad, we use superordinate categories, as when we ask, “How many animals are in the San Diego Zoo?” But we use precise subordinate categories when we need to be specific: “How many adult emus are in the San Diego Zoo today?”

If we return to our clothing example, finding a pair of white wool hiking socks is very easy if the organizing system for socks creates fine-grained categories. When resources are described or arranged with this level of detail, a similarly detailed specification of the resources you are looking for yields precisely what you want. When you get to the place where you keep white wool hiking socks, you find all of them and nothing else. On the other hand, if all your socks are tossed unsorted into a sock drawer, when you go sock hunting you might not be able to find the socks you want and you will encounter lots of socks you do not want. But you will not have put time into sorting them, which many people do

not enjoy doing; you can spend time sorting or searching depending on your preferences.

If we translate this example into the jargon of information retrieval, we say that more fine-grained organization reduces *recall*, the number of resources you find or retrieve in response to a query, but increases the *precision* of the recalled set, the proportion of recalled items that are relevant. Broader or coarse-grained categories increase recall, but lower precision. We are all too familiar with this hard bargain when we use a web search engine; a quick one-word query results in many pages of mostly irrelevant sites, whereas a carefully crafted multi-word query pinpoints sites with the information we seek. We will discuss recall, precision, and evaluation of information retrieval more extensively in Chapter 10, *Interactions with Resources*.

This mundane example illustrates the fundamental tradeoff between organization and retrieval. A tradeoff between the investment in organization and the investment in retrieval persists in nearly every organizing system. The more effort we put into organizing resources, the more effectively they can be retrieved. The more effort we are willing to put into retrieving resources, the less they need to be organized first. The allocation of costs and benefits between the organizer and retriever differs according to the relationship between them. Are they the same person? Who does the work and who gets the benefit?

7.4.4 Category Audience and Purpose

The ways in which people categorize depend on the goals of categorization, the breadth of the resources in the collection to be categorized, and the users of the organizing system. Suppose that we want to categorize languages. Our first step might be determining what constitutes a language, since there is no widespread agreement on what differentiates a language from a dialect, or even on whether such a distinction exists.

What we mean by “English” and “Chinese” as categories can change depending on the audience we are addressing and what our purpose is, however.[436[Ling]] A language learning school’s representation of “English” might depend on practical concerns such as how the school’s students are likely to use the language they learn, or which teachers are available. For the purposes of a school teaching global languages, and one of the standard varieties of English (i.e., those associated with political power), or an amalgamation of several standard varieties, might be thought of as a single instance (“English”) of the category “Languages.”

Similarly, the category structure in which “Chinese” is situated can vary with context. While some schools might not conceptualize “Chinese” as a category encompassing multiple linguistic varieties, but rather as a single instance within the “Languages” category, another school might teach its students Mandarin,

Wu, and Cantonese as dialects within the language category "Chinese," that are unified by a single standard *writing system*. In addition, a linguist might consider Mandarin, Wu, and Cantonese to be mutually unintelligible, making them separate languages within the broader category "Chinese" for the purpose of creating a principled language classification system.

If people could only categorize in a single way, the *Pyramid* game show, where contestants guess what category is illustrated by the example provided by a clue giver, would pose no challenge. The creative possibilities provided by categorization allow people to order the world and refer to interrelationships among conceptions through a kind of allusive shorthand. When we talk about the language of fashion, we suggest that in the context of our conversation, instances like "English," "Chinese," and "fashion" are alike in ways that distinguish them from other things that we would not categorize as languages.

7.5 Implementing Categories

Categories are conceptual constructs that we use in a mostly invisible way when we talk or think about them. When we organize our kitchens, closets, or file cabinets using shelves, drawers, and folders, these physical locations and containers are visible implementations of our personal category system, but they are not the categories. This distinction between category design and implementation is obvious when we follow signs and labels in libraries or grocery stores to find things, search a product catalog or company personnel directory, or analyze a set of economic data assembled by the government from income tax forms. These institutional categories were designed by people prior to the assignment of resources to them.

This separation between category creation and category implementation prompts us to ask how a system of categories can be implemented. We will not discuss the implementation of categories in the literal sense of building physical or software systems that organize resources. Instead, we will take a higher-level perspective that analyzes the implementation problem to be solved for the different types of categories discussed in §7.3, and then explain the logic followed to assign resources correctly to them.

7.5.1 Implementing Enumerated Categories

Categories defined by enumeration are easy to implement. The members or legal values in a set define the category, and testing an item for membership means looking in the set for it. Enumerated category definitions are familiar in drop-down menus and form-filling. You scroll through a list of all the countries in the world to search for the one you want in a shipping address, and whatever you select will be a valid country name, because the list is fixed until a new country is born. Enumerated categories can also be implemented with associative arrays (also known as hash tables or dictionaries). With these data structures, a test for set membership is even more efficient than searching, because it takes the same time for sets of any size (see §9.2.1 Kinds of Structures (page 442)).

7.5.2 Implementing Categories Defined by Properties

The most conceptually simple and straightforward implementation of categories defined by properties adopts the *classical view of categories* based on necessary and sufficient features. Because such categories are prescriptive with explicit and clear boundaries, classifying items into the categories is objective and deterministic, and supports a well-defined notion of *validation* to determine unambiguously whether some instance is a member of the category. Items are classified by testing them to determine if they have the required properties and property values. Tests can be expressed as rules:

- If instance X has property P, then X is in category Y.
- If a home mortgage loan in San Francisco exceeds $625,000, then it is classified as a “jumbo” loan by the US Office of Federal Housing Oversight.
- For a number to be classified as prime it must satisfy two rules: It must be greater than 1, and have no positive divisors other than 1 and itself.

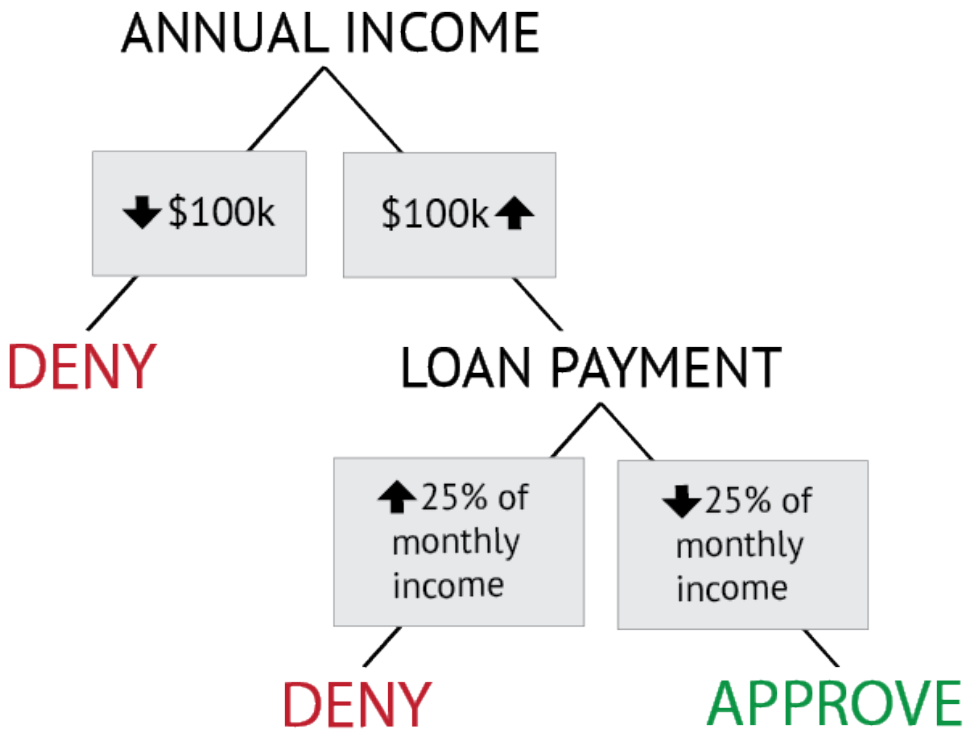
This doesn’t mean the property test is always easy; validation might require special equipment or calculations, and tests for the property might differ in their cost or efficiency. But given the test results, the answer is unambiguous. The item is either a member of the category or it isn’t.[437][Com]

A system of hierarchical categories is defined by a sequence of property tests in a particular order. The most natural way to implement multi-level category systems is with *decision trees*. A simple *decision tree* is an algorithm for determining a decision by making a sequence of logical or property tests. Suppose a bank used a sequential rule-based approach to decide whether to give someone a mortgage loan.

- If applicant's annual income exceeds $100,000, and if the monthly loan payment is less than 25% of monthly income, approve the mortgage application.
- Otherwise, deny the loan application.

This simple decision tree is depicted in Figure 7.1, Rule-based Decision Tree. The rules used by the bank to classify loan applications as "Approved" or "Denied" have a clear representation in the tree. The easy interpretation of decision trees makes them a common formalism for implementing classification models.

Figure 7.1. Rule-based Decision Tree



In this simple decision tree, a sequence of two tests for the borrower's annual income and the percentage of monthly income required to make the loan payment classify the applicants into the "deny" and "approve" categories.

Nevertheless, any implementation of a category is only interpretable to the extent that the properties and tests it uses in its definition and implementation can be understood. Because natural language is inherently ambiguous, it is not the optimal representational format for formally defined institutional categories. Categories defined using natural language can be incomplete, inconsistent, or

ambiguous because words often have multiple meanings. This implementation of the bank's procedure for evaluating loans would be hard to interpret reliably:

- If applicant is wealthy, and then if the monthly payment is an amount that the applicant can easily repay, then applicant is approved.

To ensure their interpretability, decision trees are sometimes specified using the controlled vocabularies and constrained syntax of "simplified writing" or "business rule" systems.

Artificial languages are a more ambitious way to enable precise specification of property-based categories. An artificial language expresses ideas concisely by introducing new terms or symbols that represent complex ideas along with syntactic mechanisms for combining and operating on them. Mathematical notation, programming languages, schema languages that define valid document instances (see §9.2.3.1), and regular expressions that define search and selection patterns (see §9.2.3.2) are familiar examples of artificial languages. It is certainly easier to explain and understand the Pythagorean Theorem when it is efficiently expressed as "$H^2 = A^2 + B^2$" than with a more verbose natural language expression: "In all triangles with an angle such that the sides forming the angle are perpendicular, the product of the length of the side opposite the angle such that the sides forming the angle are perpendicular with itself is equal to the sum of the products of the lengths of the other two sides, each with itself."[438][CogSci]

Artificial languages for defining categories have a long history in philosophy and science. (See the sidebar, Artificial Languages for Description and Classification (page 364)). However, the vast majority of institutional category systems are still specified with natural language, despite its ambiguities because people usually understand the languages they learned naturally better than artificial ones. Sometimes this is even intentional to allow institutional categories embodied in laws to evolve in the courts and to accommodate technological advances.[439][Law]

Data schemas that specify data entities, elements, identifiers, attributes, and relationships in databases and XML document types on the transactional end of the Document Type Spectrum (§4.2.1) are implementations of the categories needed for the design, development and maintenance of information organization systems. Data schemas tend to rigidly define categories of resources. [441][Com]

In object-oriented programming languages, *classes* are schemas that serve as templates for the creation of objects. A class in a programming language is analogous to a database schema that specifies the structure of its member instances, in that the class definition specifies how instances of the class are constructed in terms of data types and possible values. Programming classes may also specify whether data in a member object can be accessed, and if so, how.[442][Com]

Artificial Languages for Description and Classification

John Wilkins was one of the founders of the British Royal Society. In 1668 he published *An Essay towards a Real Character and a Philosophical Language* in which he proposed an artificial language for describing a universal taxonomy of knowledge that used symbol composition to specify a location in the category hierarchy. There were forty top level genus categories, which were further subdivided into differences within the genus, which were then subdivided into species. Each genus was a monosyllable of two letters; each difference added a consonant, and each species added a vowel.

This artificial language conveys the meaning of categories directly from the composition of the category name. For instance, *zi* indicates the genus of beasts, *zit* would be "rapacious beasts of the dog kind" whereas *zid* would be "cloven-footed beast." Adding for the fourth character an *a* for species, indicating the second species in the difference, would give *zita* for dog and *zida* for sheep.

In *The Analytical Language of John Wilkins,* Jorge Luis Borges remarks that Wilkins has many "ambiguities, redundancies and deficiencies" in the language and presents as a foil and parody an imagined "Celestial Empire of Benevolent Knowledge."

> In its remote pages it is written that the animals are divided into: (a) belonging to the emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camel hair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies.

Borges compliments Wilkins for inventing names that might signify in themselves some meaning to those who know the system, but notes that "it is clear that there is no classification of the Universe not being arbitrary and full of conjectures."[440[Ling]]

Unlike transactional document types, which can be prescriptively defined as *classical categories* because they are often produced and consumed by automated processes, narrative document types are usually descriptive in character. We do not classify something as a novel because it has some specific set of properties and content types. Instead, we have a notion of typical novels and their characteristic properties, and some things that are considered novels are far from typical in their structure and content.[443[CogSci]]

Nevertheless, categories like narrative document types can sometimes be implemented using document schemas that impose only a few constraints on structure and content. A schema for a purchase order is highly prescriptive; it uses

regular expressions, strongly data typed content, and enumerated code lists to validate the value of required elements that must occur in a particular order. In contrast, a schema for a narrative document type would have much optionality, be flexible about order, and expect only text in its sections, paragraphs and headings. Even very lax document schemas can be useful in making content management, reuse, and formatting more efficient.

7.5.3 Implementing Categories Defined by Probability and Similarity

Many categories cannot be defined in terms of required properties, and instead must be defined probabilistically, where category membership is determined by properties that resources are likely to share. Consider the category "friend." You probably consider many people to be your friends, but you have longtime friends, school friends, workplace friends, friends you see only at the gym, and friends of your parents. Each of these types of friends represents a different cluster of common properties. If someone is described to you as a potential friend or date, how accurately can you predict that the person will become a friend? (See the sidebar, Finding Friends and Dates: Lessons for Learning Categories (page 366))

Probabilistic categories can be challenging to define and use because it can be difficult to keep in mind the complex feature correlations and probabilities exhibited by different clusters of instances from some domain. Furthermore, when the category being learned is broad with a large number of members, the sample from which you learn strongly shapes what you learn. For example, people who grow up in high-density and diverse urban areas may have less predictable ideas of what an acceptable potential date looks like than someone in a remote rural area with a more homogeneous population.

More generally, if you are organizing a domain where the resources are active, change their state, or are measurements of properties that vary and co-occur probabilistically, the sample you choose strongly affects the accuracy of models for classification or prediction. In *The Signal and the Noise,* statistician Nate Silver explains how many notable predictions failed because of poor sampling techniques. One common sampling mistake is to use too short a historical window to assemble the training dataset; this is often a corollary of a second mistake, an over reliance on recent data because it is more available. For example, the collapse of housing prices and the resulting financial crisis of 2008 can be explained in part because the models that lenders used to predict mortgage foreclosures were based on data from 1980-2005, when house prices tended to grow higher. As a result, when mortgage foreclosures increased rapidly, the results were "out of sample" and were initially misinterpreted, delaying responses to the crisis.

Finding Friends and Dates: Lessons for Learning Categories

Online dating or matchmaking sites use many of the same features to describe people, but also have additional features to make more accurate matches for their targeted users. As the number of features grows, there are exponentially more combinations of shared properties. For example, the matchmaking site eHarmony employs 29 "Dimensions of Compatibility" and more than 200 questions to create a user profile. Even if the 29 dimensions were Boolean (would you describe yourself as x?) this yields 2^{29} or over 500,000,000 different combinations. Using these complex resource descriptions to predict the probability of a good match requires matchmaking sites to use proprietary machine learning algorithms to propose matches, which are ranked with unexplained measures and precision (what does an 80% match mean?). Not surprisingly, many people who try online dating give up after less success than they expected.

With such a large number of features in user profiles, any matching algorithm confronts what machine learning calls the curse of dimensionality. With high-dimensional data, there can never be enough instances to learn which features are really the most important. Neither you nor the online dating algorithm will ever meet enough different kinds of people to reliably predict the outcome of a possible match.

But all is not hopeless. Machine learning programs attack the curse of dimensionality using statistical techniques that use correlations among features to combine them or adjust the weights given to features to reflect their value in making predictions or classifications. For example, OKCupid asks people to rate how much importance they assign to match questions. You might prefer cats to dogs, and you might either never consider dating a dog lover or you might not care at all.

Another way to reduce the number of features needed to classify accurately is to reduce the scope of the category being learned. The matchmaking model for sites that target people with particular professions, religions, or political views would be less complex than the eHarmony one, because the former will have fewer relevant features, and hence fewer random correlations and noise that will undermine its accuracy. All other things being equal, the lower the variability in a set of examples, the better a model that learns from that data will perform.

Samples from dynamic and probabilistic domains result in models that capture this variability. Unfortunately, because many forecasters want to seem authoritative, and many people do not understand probability, classifications or predictions that are inherently imprecise are often presented with certainty and exactness even though they are probabilistic with a range of outcomes. Silver tells

the story of a disastrous 1997 flood caused when the Red River crested at 54 feet when the levees protecting the town of Grand Forks were at 51 feet. The weather service had predicted a crest between 40 and 58 feet, but emphasized the midpoint of the range, which was 49 feet. Unfortunately, most people interpreted this probabilistic prediction as if it were a binary classification, "flood" versus "no flood," ignored the range of the forecast, and failed to prepare for a flood that had about a 35% chance of occurring.[444][DS]

7.5.3.1 Probabilistic Decision Trees

In §7.5.2, we showed how a rule-based decision tree could be used to implement a strict property-based classification in which a bank uses tests for the properties of "annual income" and "monthly loan payment" to classify applicants as approved or denied. We can adapt that example to illustrate probabilistic decision trees, which are better suited for implementing categories in which category membership is probabilistic rather than absolute.

Banks that are more flexible about making loans can be more profitable because they can make loans to people that a stricter bank would reject but who still are able to make loan payments. Instead of enforcing conservative and fixed cutoffs on income and monthly payments, these banks consider more properties and look at applications in a more probabilistic way. These banks recognize that not every loan applicant who is likely to repay the loan looks exactly the same; "annual income" and "monthly loan payment" remain important properties, but other factors might also be useful predictors, and there is more than one configuration of values that an applicant could satisfy to be approved for a loan.

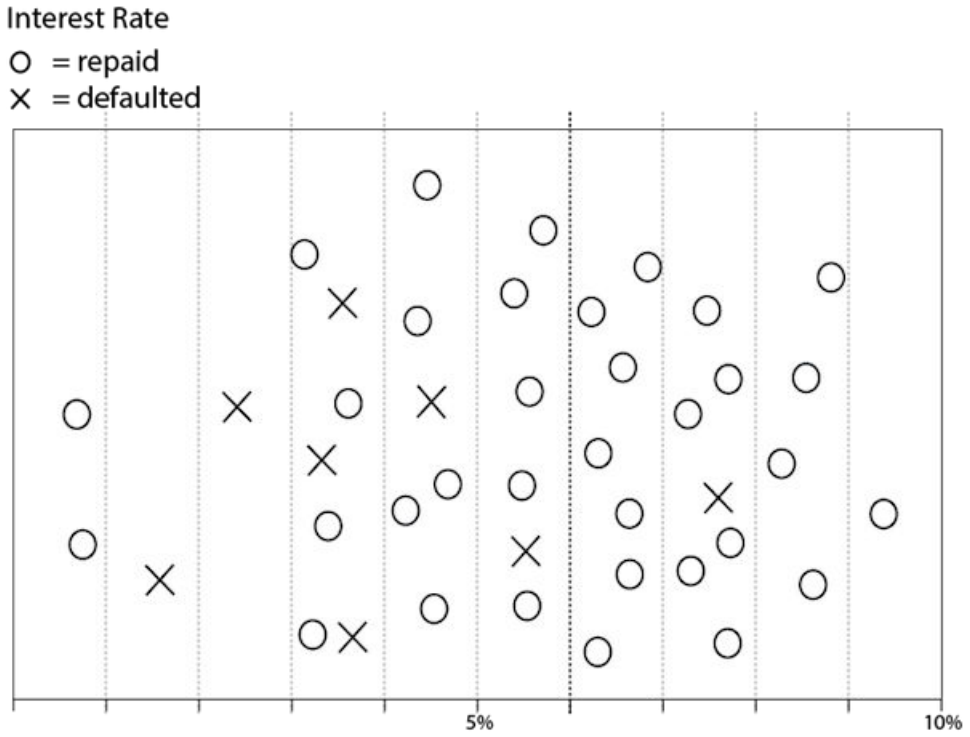
Which properties of applicants best predict whether they will repay the loan or default? A property that predicts each at 50% isn't helpful because the bank might as well flip a coin, but a property that splits the applicants into two sets, each with very different probabilities for repayment and defaulting, is very helpful in making a loan decision.

A data-driven bank relies upon historical data about loan repayment and defaults to train algorithms that create decision trees by repeatedly splitting the applicants into subsets that are most different in their predictions. Subsets of applicants with a high probability of repayment would be approved, and those with a high probability of default would be denied a loan. One method for selecting the property test for making each split is calculating the "information gain" (see the sidebar Using "Information Theory" to Quantify Organization (page 75)). This measure captures the degree to which each subset contains a "pure" group in which every applicant is classified the same, as likely repayers or likely defaulters.

For example, consider the chart in Figure 7.2, Historical Data: Loan Repayment Based on Interest Rate which is a simplified representation of the bank's histori-

cal data on loan defaults based on the initial interest rate. The chart represents loans that were repaid with "o" and those that defaulted with "x." Is there an interest rate that divides them into "pure" sets, one that contains only "o" loans and the other that contains only "x" loans?

Figure 7.2. Historical Data: Loan Repayment Based on Interest Rate



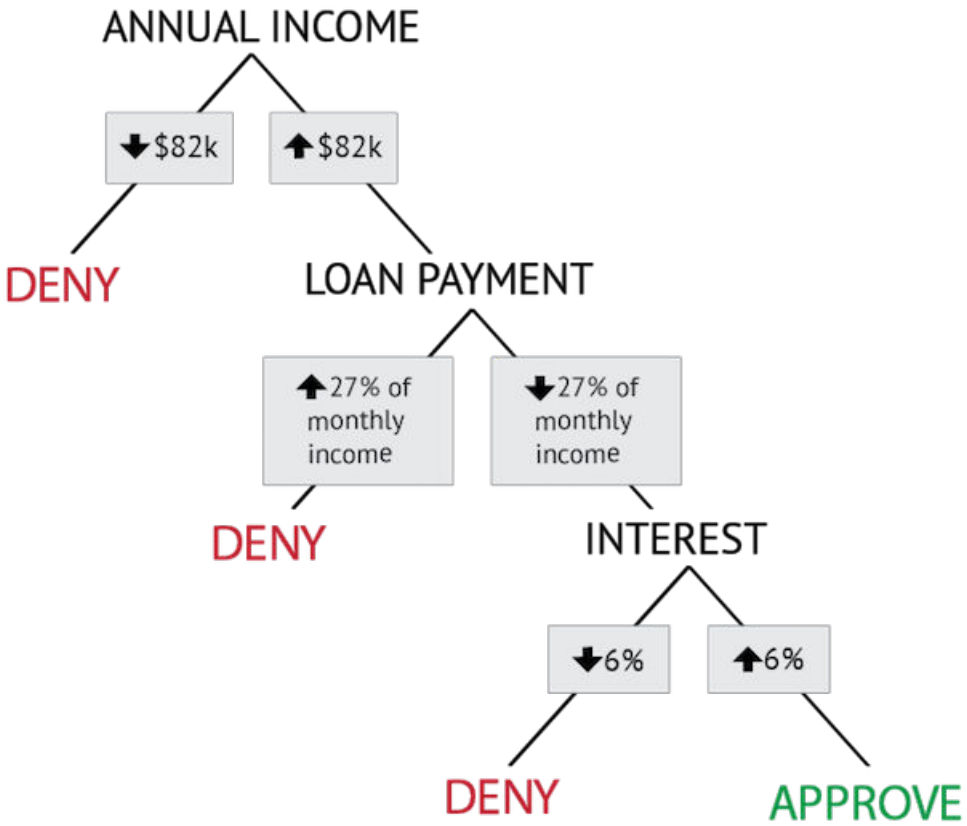
The "o" symbol represents loans that were repaid by the borrower; "x" represents loans on which the borrower defaulted. A 6% rate (darker vertical line) best divides the loans into subsets that differ in the payment outcome.

You can see that no interest rate divides these into pure sets. So the best that can be done is to find the interest rate that divides them so that the proportions of defaulters are most different on each side of the line.[445][DS]

This dividing line at the 6% interest rate best divides those who defaulted from those who repaid their loan. Most people who borrowed at 6% or greater repaid the loan, while those who took out loans at a lower rate were more likely to default. This might seem counter-intuitive until you learn that the lower-interest rate loans had adjustable rates that increased after a few years, causing the monthly payments to increase substantially. More prudent borrowers were will-

ing to pay higher interest rates that were fixed rather than adjustable to avoid radical increases in their monthly payments.

Figure 7.3. Probabilistic Decision Tree



In this probabilistic decision tree, the sequence of property tests and the threshold values in each test divide the loan applicants into categories that differ in how likely they are to repay the loan.

This calculation is carried out for each of the attributes in the historical data set to identify the one that best divides the applicants into the repaid and defaulted categories. The attributes and the value that defines the decision rule can then be ordered to create a decision tree similar to the rule-based one we saw in §7.5.2. In our hypothetical case, it turns out that the best order in which to test the properties is Income, Monthly Payment, and Interest Rate, as shown in Figure 7.3, Probabilistic Decision Tree. The end result is still a set of rules, but behind each decision in the tree are probabilities based on historical data that can more accurately predict whether an applicant will repay or default. Thus, in-

stead of the arbitrary cutoffs at $100,000 in income and 25% for monthly payment, the bank can offer loans to people with lower incomes and remain profitable doing so, because it knows from historical data that $82,000 and 27% are the optimal decision points. Using the interest rate in their decision process is an additional test to ensure that people can afford to make loan payments even if interest rates go up.[446[Bus]]

Because decision trees specify a sequence of rules that make property tests, they are highly interpretable, which makes them a very popular choice for data scientists building models much more complex than the simple loan example here. But they assume that every class is a conjunction of all the properties used to define them. This makes them susceptible to over-fitting because if they grow very deep with many property conjunctions, they capture exactly the properties that describe each member of the training set, effectively memorizing the training data. In other words, they capture both what is generally true beyond the set and what is particular to the training set only, when the goal is to build a model that captures only what is generally true. Overfitting in decision trees can be prevented by pruning back the tree after it has perfectly classified the training set, or by limiting the depth of the tree in advance, essentially pre-pruning it.

7.5.3.2 Naïve Bayes Classifiers

Another commonly used approach to implement a classifier for probabilistic categories is called Naïve Bayes. It employs Bayes' Theorem for learning the importance of a particular property for correct classification. There are some common sense ideas that are embodied in Bayes' Theorem:

- When you have a hypothesis or prior belief about the relationship between a property and a classification, new evidence consistent with that belief should increase your confidence.
- Contradictory evidence should reduce confidence in your belief.
- If the base rate for some kind of event is low, do not forget that when you make a prediction or classification for a new specific instance. It is easy to be overly influenced by recent information.

Now we can translate these ideas into calculations about how learning takes place. For property A and classification B, Bayes' Theorem says:

$$P(A \mid B) = P(B|A)\, P(A) / P(B)$$

The left hand side of the equation, P (A | B), is what we want to estimate but can't measure directly: the probability that A is the correct classification for an item or observation that has property B. This is called the conditional or posterior probability because it is estimated after seeing the evidence of property B.

P (B | A) is the probability that any item correctly classified as A has property B. This is called the likelihood function.

P (A) and P (B) are the independent or prior probabilities of A and B; what proportion of the items are classified as A? How often does property B occur in some set of items?

Using Bayes' Theorem to Calculate Conditional Probability

Your personal library contains 60% fiction and 40% nonfiction books. All of the fiction books are in ebook format, and half of the nonfiction books are ebooks and half are in print format. If you pick a book at random and it is in ebook format, what is the probability that it is nonfiction?

Bayes' Theorem tells us that:

P (nonfiction | ebook) = P (ebook |nonfiction) x P (nonfiction) / P (ebook).

We know: P (ebook | nonfiction) = .5 and P (nonfiction) = .4

We compute P (ebook) using the law of total probability to compute the combined probability of all the independent ways in which an ebook might be sampled. In this example there are two ways:

P (ebook) = P (ebook | nonfiction) x P (nonfiction)
+ P (ebook | fiction) x P (fiction)
= (.5 x .4) + (1 x .6) = .8

Therefore: P (nonfiction | ebook) = (.5 x .4) / .8 = .25

Now let's apply Bayes' Theorem to implement email spam filtering. Messages are classified as SPAM or HAM (i.e., non-SPAM); the former are sent to a SPAM folder, while the latter head to your inbox.

1. Select Properties. We start with a set of properties, some from the message metadata like the sender's email address or the number of recipients, and some from the message content. Every word that appears in messages can be treated as a separate property[447][Com]
2. Assemble Training Data. We assemble a set of email message that have been correctly assigned to the SPAM and HAM categories. These labeled instances make up the training set.
3. Analyze the Training Data. For each message, does it contain a particular property? For each message, is it classified as SPAM? If a message is classified as SPAM, does it contain a particular property? (These are the three probabilities on the right side of the Bayes equation).

4. Learn. The conditional probability (the left side of the Bayes equation) is recalculated, adjusting the predictive value of each property. Taken together, all of the properties are now able to correctly assign (most of) the messages into the categories they belonged to in the training set.
5. Classify. The trained classifier is now ready to classify uncategorized messages to the SPAM or HAM categories.
6. Improve. The classifier can improve its accuracy if the user gives it feedback by reclassifying SPAM messages as HAM ones or vice versa. The most efficient learning occurs when an algorithm uses "active learning" techniques to choose its own training data by soliciting user feedback only where it is uncertain about how to classify a message. For example, the algorithm might be confident that a message with "Cheap drugs" in the subject line is SPAM, but if the message comes from a longtime correspondent, the algorithm might ask the user to confirm that the classification.[448][Com]

7.5.3.3 Categories Created by Clustering

In the previous two sections we discussed how probabilistic decision trees and naïve Bayes classifiers implement categories that are defined by typically shared properties and similarity. Both are examples of supervised learning because they need correctly classified examples as training data, and they learn the categories they are taught.

In contrast, clustering techniques are unsupervised; they analyze a collection of uncategorized resources to discover statistical regularities or structure among the items, creating a set of categories without any labeled training data.

Clustering techniques share the goal of creating meaningful categories from a collection of items whose properties are hard to directly perceive and evaluate, which implies that category membership cannot easily be reduced to specific property tests and instead must be based on similarity. For example, with large sets of documents or behavioral data, clustering techniques can find categories of documents with the same topics, genre, or sentiment, or categories of people with similar habits and preferences.

Because clustering techniques are unsupervised, they create categories based on calculations of similarity between resources, maximizing the similarity of resources within a category and maximizing the differences between them. These statistically-learned categories are not always meaningful ones that can be named and used by people, and the choice of properties and methods for calculating similarity can result in very different numbers and types of categories. Some clustering techniques for text resources suggest names for the clusters based on the important words in documents at the center of each cluster. However, unless there is a labeled set of resources from the same domain that can

be used as a check to see if the clustering discovered the same categories, it is up to the data analyst or information scientist to make sense of the discovered clusters or topics.

There are many different distance-based clustering techniques, but they share three basic methods.

- The first shared method is that clustering techniques start with an initially uncategorized set of items or documents that are represented in ways that enable measures of inter-item similarity can be calculated. This representation is most often a vector of property values or the probabilities of different properties, so that items can be represented in a multidimensional space and similarity calculated using a distance function like those described in §7.3.6.2 Geometric Models of Similarity (page 352).[449][Com]
- The second shared method is that categories are created by putting items that are most similar into the same category. Hierarchical clustering approaches start with every item in its own category. Other approaches, notably one called "K-means clustering," start with a fixed number of K categories initialized with a randomly chosen item or document from the complete set.
- The third shared method is refining the system of categories by iterative similarity recalculation each time an item is added to a category. Approaches that start with every item in its own category create a hierarchical system of categories by merging the two most similar categories, recomputing the similarity between the new category and the remaining ones, and repeating this process until all the categories are merged into a single category at the root of a category tree. Techniques that start with a fixed number of categories do not create new ones but instead repeatedly recalculate the "centroid" of the category by adjusting its property representation to the average of all its members after a new member is added.[450][Com]

It makes sense that the algorithms that create clusters or categories of similar items can be later used as classifiers by using the same similarity measures to compare the unclassified items against items that are labeled by category. There are different choices about which items to compare with the unclassified one:

- The centroid: a prototypical or average item calculated on the properties of all the category members. However, the centroid might not correspond to any actual member (see the sidebar Median versus Average (page 117)), and this can make it hard to interpret the classification.
- Items that actually exist: Because the items in categories defined by similarity are not equally typical or good members, it is more robust to test against more than one exemplar. Classifiers that use this approach are called

nearest-neighbor techniques, and they essentially vote among themselves and the majority category is assigned to the new item.

- The edge cases: These are instances that are closest to the boundary between two categories, so there need to be at least two of them, one in each category. Because they are not typical members of the category, they are the hardest to classify initially, but using them in classifiers emphasizes the properties that are the most discriminating. This is the approach taken by support vector machines, which are not clustering algorithms but are somewhat like nearest-neighbor algorithms in that they calculate the similarity of an unclassified item to these edge cases. Their name makes more sense if you think of the vectors that represent the "edge cases" being used to "support" the category boundary, which falls between them.

7.5.3.4 Neural networks

Among the best performing classifiers for categorizing by similarity and probabilistic membership are those implemented using neural networks, and especially those employing deep learning techniques. Deep learning algorithms can learn categories from labeled training data or by using autoencoding, an unsupervised learning technique that trains a neural network to reconstruct its input data. However, instead of using the properties that are defined in the data, deep learning algorithms devise a very large number of features in hidden hierarchical layers, which makes them uninterpretable by people. The key idea that made deep learning possible is the use of "backpropagation" to adjust the weights on features by working backwards from the output (the object classification produced by the network) all the way back to the input. The use of deep learning to classify images was mentioned in §5.4.2.[451][DS]

7.5.4 Implementing Goal-Based Categories

Goal-based categories are highly individualized, and are often used just once in a very specific context. However, it is useful to consider that we could implement model goal-derived categories as rule-based decision trees by ordering the decisions to ensure that any sub-goals are satisfied according to their priority. We could understand the category "Things to take from a burning house" by first asking the question "Are there living things in the house?" because that might be the most important sub-goal. If the answer to that question is "yes," we might proceed along a different path than if the answer is "no." Similarly, we might put a higher priority on things that cannot be replaced (Grandma's photos) than those that can (passport).

7.5.5 Implementing Theory-Based Categories

Theory-based categories arise in domains in which the items to be categorized are characterized by abstract or complex relationships with their features and with each other. With this model an entity need not be understood as inherently possessing features shared in common with another entity. Rather, people project features from one thing to another in a search for congruities between things, much as clue receivers in the second round of the *Pyramid* game search for congruities between examples provided by the clue giver in order to guess the target category. For example, a clue like "screaming baby" can suggest many categories, as can "parking meter." But the likely intersection of the interactions one can have with babies and parking meters is that they are both "Things you need to feed."

Theory-based categories are created as cognitive constructs when we use analogies and classify, because things brought together by analogy have abstract rather than literal similarity. The most influential model of analogical processing is Structure Mapping, whose development and application has been guided by Dedre Gentner for over three decades.

The key insight in Structure Mapping is that an analogy "a T is like B" is created by matching relational structures and not properties between the base domain B and a target domain T. We take any two things, analyze the relational structures they contain, and align them to find correspondences between them. The properties of objects in the two domains need not match, and in fact, if too many properties match, analogy goes away and we have literal similarity:

- Analogy: The hydrogen atom is like our solar system
- Literal Similarity: The X12 star system in the Andromeda galaxy is like our solar system

Structure Mapping theory was implemented in the Structure-Mapping Engine (SME), which both formalized the theory and offered a computationally-tractable algorithm for carrying out the process of mapping structures and drawing inferences.[452][CogSci]

7.6 Key Points in Chapter Seven

- Categories are *equivalence classes*: sets or groups of things or abstract entities that we treat the same.

  (See §7.2 The What and Why of Categories (page 325))
- The size of the equivalence class is determined by the properties or characteristics we consider.

  (See §7.2 The What and Why of Categories (page 325))
- Cultural, individual, and institutional categorization share some core ideas but they emphasize different processes and purposes for creating categories.

  (See §7.2 The What and Why of Categories (page 325))
- Individual categories are created by intentional activity that usually takes place in response to a specific situation.

  (See §7.2.2 Individual Categories (page 330))
- Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories are needed.

  (See §7.2.3 Institutional Categories (page 331))
- The rigorous definition of institutional categories enables *classification,* the systematic assignment of resources to categories in an organizing system.

  (See §7.2.3 Institutional Categories (page 331))
- Computational categories are created by computer programs when the number of resources, or when the number of descriptions or observations associated with each resource, are so large that people cannot think about them effectively.

  (See §7.2.5 Computational Categories (page 334))
- In supervised learning, a machine learning program is trained by giving it sample items or documents that are labeled by category. In unsupervised learning, the program gets the samples but has to come up with the categories on its own.

  (See Supervised and Unsupervised Learning (page 336))
- Any collection of resources with sortable identifiers (alphabetic or numeric) as an associated property can benefit from using sorting order as an organizing principle.

  (See §7.3.2 Single Properties (page 338))

- If only a single property is used to distinguish among some set of resources and to create the categories in an organizing system, the choice of property is critical because different properties often lead to different categories.

  (See §7.3.2 Single Properties (page 338))

- A sequence of organizing decisions based on a fixed ordering of resource properties creates a *hierarchy*, a multi-level category system.

  (See §7.3.3.1 Multi-Level or Hierarchical Categories (page 340))

- An important implication of necessary and sufficient category definition is that every member of the category is an equally good member or example of the category.

  (See §7.3.3.3 Necessary and Sufficient Properties (page 344))

- For most purposes, the most useful property of information resources for categorizing them is their *aboutness*, which is not directly perceivable and which is hard to characterize.

  (See §7.3.4 The Limits of Property-Based Categorization (page 346))

- In domains where properties lack one or more of the characteristics of separability, perceptibility, and necessity, a probabilistic or statistical view of properties is needed to define categories.

  (See §7.3.5 Probabilistic Categories and “Family Resemblance” (page 348))

- Sharing some but not all properties is akin to *family resemblances* among the category members.

  (See §7.3.5 Probabilistic Categories and “Family Resemblance” (page 348))

- *Similarity* is a measure of the resemblance between two things that share some characteristics but are not identical.

  (See §7.3.6 Similarity (page 351))

- Feature- or property-based, geometry-based, transformational, and alignment- or analogy-based approaches are psychologically-motivated approaches that propose different functions for computing similarity.

  (See §7.3.6 Similarity (page 351))

- Classical categories can be defined precisely with just a few *necessary and sufficient* properties.

  (See §7.4.2 Basic or Natural Categories (page 358))

- Broader or coarse-grained categories increase *recall*, but lower *precision.*

  (See §7.4.3 The Recall / Precision Tradeoff (page 358))

- A simple *decision tree* is an algorithm for determining a decision by making a sequence of logical or property tests.

(See §7.5.2 Implementing Categories Defined by Properties (page 361))

- The most conceptually simple and straightforward implementation of categories in technologies for organizing systems adopts the classical view of categories based on necessary and sufficient features.

  (See §7.5.2 Implementing Categories Defined by Properties (page 361))

- An artificial language expresses ideas concisely by introducing new terms or symbols that represent complex ideas along with syntactic mechanisms for combining and operating on them.

  (See §7.5.2 Implementing Categories Defined by Properties (page 361))

- Naïve Bayes classifiers learn by revising the conditional probability of each property for making the correct classification after seeing the base rates of the class and property in the training data and how likely it is that a member of the class has the property.

  (See §7.5.3.2 Naïve Bayes Classifiers (page 370))

- Because clustering techniques are unsupervised, they create categories based on calculations of similarity between resources, maximizing the similarity of resources within a category and maximizing the differences between them.

  (See §7.5.3.3 Categories Created by Clustering (page 372))

Endnotes for Chapter 7

[386][CogSci] Cataloging and programming are important activities that need to be done well, and prescriptive advice is often essential. However, we believe that understanding how people create psychological and linguistic categories can help us appreciate that cataloging and information systems design are messier and more intellectually challenging activities than we might otherwise think.

[387][CogSci] Cognitive science mostly focuses on the automatic and unconscious mechanisms for creating and using categories. This disciplinary perspective emphasizes the activation of category knowledge for the purpose of making inferences and "going beyond the information given," to use Bruner's classic phrase (Bruner 1957). In contrast, the discipline of organizing focuses on the explicit and self-aware mechanisms for creating and using categories because by definition, organizing systems serve intentional and often highly explicit purposes. Organizing systems facilitate inferences about the resources they contain, but the more constrained purposes for which resources are described and arranged makes inference a secondary goal.

Cognitive science is also highly focused on understanding and creating computational models of the mechanisms for creating and using categories. These models blend data-driven or bottom-up processing with knowledge-driven or top-down processing to simulate the time course and results of categorization at both fine-grained scales (as in word or object recognition) and over developmental time frames (as in how children learn categories). The discipline of organizing can learn from these models about the types of properties and principles that organizing systems use, but these computational models are not a primary concern to us in this book.

[388][CogSci] However, even the way this debate has been framed is a bit controversial. Bulmer's chicken, the "categories are in the world" position, has been described as empirical, environment-driven, bottom-up, or objectivist, and these are not synonymous. Likewise, the "egghead" position that "categories are in the mind" has been called rational, constructive, top-down, experiential, and embodied—and they are also not synonyms. See (Bulmer 1970). See also (Lakoff 1990), (Malt 1995).

[389][CogSci] Is there a "universal grammar" or a "language faculty" that imposes strong constraints on human language and cognition? (Chomsky 1965) and (Jackendoff 1996) think so. Such proposals imply cognitive representations in which categories are explicit structures in memory with associated instances and properties. In contrast, generalized learning theories model category formation as the adjustment of the patterns and weighting of connections in neural processing networks that are not specialized for language in any way. Computational simulations of semantic networks can reproduce the experimental and behavioral results about language acquisition and semantic judgments that have been used as evidence for explicit category representations without needing anything like them. (Rogers and McClelland 2008) thoroughly review the explicit category models and then show how relatively simple learning models can do without them.

[390][CogSci] The debates about human category formation also extend to issues of how children learn categories and categorization methods. Most psychologists argue that category learning starts with general learning mechanisms that are very perceptually based, but they do not agree whether to characterize these changes as "stages" or as phases in a more complex dynamical system. Over time more specific learning techniques evolve that focus on correlations among perceptual properties (things with wings tend to have feathers), correlations among properties and roles (things with eyes tend to eat), and ultimately correlations among roles (things that eat tend to sleep). See (Smith and Thelen 2003).

[391][CogSci] These three contexts were proposed by (Glushko, Maglio, Matlock, and Barsalou 2008), who pointed out that cognitive science has focused on cultural

categorization and largely ignored individual and institutional contexts. They argue that taking a broader view of categorization highlights dimensions on which it varies that are not apparent when only cultural categories are considered. For example, institutional categories are usually designed and maintained using prescriptive methods that have no analogues with cultural categories. There is a difference between institutional categories created for people, and categories created in institutions by computers in the predictive analytics, data mining sense.

[392][Phil] This quote comes from Plato's *Phaedrus* dialogue, written around 370 BCE. Contemporary philosophers and cognitive scientists commonly invoke it in discussions about whether "natural kinds" exist. . For example, see (Campbell, O'Rourke, and Slater 2011), and (Hutchins 2010), (Atran 1987), and others have argued that the existence of perceptual discontinuities is not sufficient to account for category formation. Instead, people assume that members of a biological category must have an essence of co-occurring properties and these guide people to focus on the salient differences, thereby creating categories. Property clusters enable inferences about causality, which then builds a framework on which additional categories can be created and refined. For example, if "having wings" and "flying" are co-occurring properties that suggest a "bird" category, wings are then inferred as the causal basis of flying, and wings become more salient.

[393][Ling] Pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, particles, and numerals and other "parts of speech" are also grammatical categories, but nouns carry most of the semantic weight.

[394][CogSci] In contrast, the set of possible interactions with even a simple object like a banana is very large. We can pick, peel, slice, smash, eat, or throw a banana, so instead of capturing this complexity in the meaning of banana it gets parceled into the verbs that can act on the banana noun. Doing so requires languages to use verbs to capture a broader and more abstract type of meaning that is determined by the nouns with which they are combined. Familiar verbs like "set," "put," and "get" have dozens of different senses as a result because they go with so many different nouns. We set fires and we set tables, but fires and tables have little in common. The intangible character of verbs and the complexity of multiple meanings make it easier to focus instead on their associated nouns, which are often physical resources, and create organizing systems that emphasize the latter rather than the former. We create organizing systems that focus on verbs when we are categorizing actions, behaviors, or services where the resources that are involved are less visible or less directly involved in the supported interactions.

[395][Ling] Many languages have a system of grammatical gender in which all nouns must be identified as masculine or feminine using definite articles (*el* and

la in Spanish, *le* and *la* in French, and so on) and corresponding pronouns. Languages also contrast in how they describe time, spatial relationships, and in which things are treated as countable objects (one ox, two oxen) as opposed to substances or mass nouns that do not have distinct singular and plural forms (like water or dirt). (Deutscher 2011) carefully reviews and discredits the strong Whorfian view and makes the case for a more nuanced perspective on linguistic relativity. He also reviews much of Lera Boroditsky's important work in this area. George Lakoff's book with the title *Women, Fire, and Dangerous Things* (Lakoff 1990) provocatively points out differences in gender rules among languages; in an aboriginal language called Dyirbal many dangerous things, including fire have feminine gender, meanwhile "fire" is masculine in Spanish (*el feugo*) and French (*le feu*).

[396][CogSci] This analysis comes from (Haviland 1998). More recently, Lera Boroditsky has done many interesting studies and experiments about linguistic relativity. See (Boroditksy 2003) for an academic summary and (Boroditsky 2010, 2011) for more popular treatments.

[397][CogSci] (Medin et al. 1997).

[398][CogSci] This was ultimately reflected in complex mythological systems, such as Greek mythology, where genealogical relationships between gods represented category relationships among the phenomena with which they were associated. As human knowledge grew and the taxonomies became more comprehensive and complex, Durkheim and Mauss argued, they lay the groundwork for scientific classifications and shed their mythological roots. (Durkheim 1963).

[399][CogSci] (Berlin 2014)

[401][Ling] The typical syntactic constraint that tags are delimited by white space encourages the creation of new categories by combining existing category names using concatenation and camel case conventions; photos that could be categorized as "Berkeley" and "Student" are sometimes tagged as "BerkeleyStudent." Similar generative processes for creating individual category names are used with Twitter "hashtags" where tweets about events are often categorized with an ad hoc tag that combines an event name and a year identifier like "#NBAFinals16."

[402][Law] Consider how the cultural category of "killing a person" is refined by the legal system to distinguish manslaughter and different degrees of murder based on the amount of intentionality and planning involved (e.g., first and second degree murder) and the roles of people involved with the killing (accessory). In general, the purpose of laws is to replace coarse judgments of categorization based on overall similarity of facts with rule-based categorization based on specific dimensions or properties.

[403][Ling] The word was invented in 1812 in a newspaper article critical of Massachusetts governor Elbridge Gerry, who oversaw the creation of biased electoral districts. One such district was so contorted in shape, it was said to look like a salamander, and thus was called a Gerrymander. The practice remains widespread, but nowadays sophisticated computer programs can select voters on any number of characteristics and create boundaries that either "pack" them into a single district to concentrate their voting power or "crack" them into multiple districts to dilute it.

[404][Bus] The particularities or idiosyncrasies of individual categorization systems sometimes capture user expertise and knowledge that is not represented in the institutional categories that replace them. Many of the readers of this book are information professionals whose technological competence is central to their work and which helps them to be creative. But for a great many other people, information technology has enabled the routinization of work in offices, assembly lines, and in other jobs where new institutionalized job categories have "downskilled" or "deskilled" the nature of work, destroying competence and engendering a great deal of resistance from the affected workers.

[405][Bus] Similar technical concerns arise in within-company and multi-company standardization efforts, but the competitive and potentially anti-competitive character of the latter imposes greater complexity by introducing considerations of business strategy and politics. Credible standards-making in multi-company contexts depends on an explicit and transparent process for gathering and prioritizing requirements, negotiating specifications that satisfy them, and ensuring conformant implementations—without at any point giving any participating firm an advantage. See the OASIS Technical Committee Process for an example (`https://www.oasis-open.org/policies-guidelines/tc-process`) and (Rosenthal et al. 2004) for an analysis of best practices.

[406][CogSci] Unfortunately, in this transition from science to popular culture, many of these so-called periodic tables are just *ad hoc* collections that ignore the essential idea that the rows and columns capture explanatory principles about resource properties that vary in a periodic manner. A notable exception is Andrew Plotkin's Periodic Table of Dessert. See (Suehle 2012) and Plotkin's table at (Periodic Table of Dessert).

[407][Bus] The Corporate Average Fuel Economy (CAFE) standards have been developed by the United States National Highway Traffic Safety Administration (`http://www.nhtsa.gov/fuel-economy`) since 1975. For a careful and critical assessment of CAFE, including the politics of categorization for vehicles like the PT Cruiser, see the 2002 report from the Committee on the Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards, National Research Council.

[408][Law] Legal disputes often reflect different interpretations of category membership and whether a list of category members is exhaustive or merely illustrative. The legal principle of "implied exclusion"—*expressio unius est exclusio alterius* —says that if you "expressly name" or "designate" an enumeration of one or more things, any thing that is not named is excluded, by implication. However, prefacing the list with "such as," "including," or "like" implies that it is not a strict enumeration because there might be other members.

[410][Phil] The distinction between intension and extension was introduced by Gottlob Frege, a German philosopher and mathematician (Frege 1892).

[411][CogSci] The number of resources in each of these categories depends on the age of the collection and the collector. We could be more precise here and say "single atomic property" or otherwise more carefully define "property" in this context as a characteristic that is basic and not easily or naturally decomposable into other characteristics. It would be possible to analyze the physical format of a music resource as a composition of size, shape, weight, and material substance properties, but that is not how people normally think. Instead, they treat physical format as a single property as we do in this example.

[412][CogSci] We need to think of alphabetic ordering or any other organizing principle in a logical way that does not imply any particular physical implementation. Therefore, we do not need to consider which of these alphabetic categories exist as folders, files, or other tangible partitions.

[413][CogSci] Another example: rules for mailing packages might use either size or weight to calculate the shipping cost, and whether these rules are based on specific numerical values or ranges of values, the intent seems to be to create categories of packages.

[414][CogSci] If you try hard, you can come up with situations in which this property is important, as when the circus is coming to the island on a ferry or when you are loading an elevator with a capacity limit of 5000 pounds, but it just is not a useful or psychologically salient property in most contexts.

[415][Com] Many information systems, applications, and programming languages that work with hierarchical categories take advantage of this logical relationship to infer inherited properties when they are needed rather than storing them redundantly.

[416][Bus] Similarly, clothing stores use intrinsic static properties when they present merchandise arranged according to color and size; extrinsic static properties when they host branded displays of merchandise; intrinsic dynamic properties when they set aside a display for seasonal merchandise, from bathing suits to winter boots; and extrinsic dynamic properties when a display area is set aside for "Today's Special."

[417][Phil] Aristotle did not call them classical categories. That label was bestowed about 2300 years later by (Smith and Medin 1981).

[419][CogSci] Typicality and centrality effects were studied by Rosch and others in numerous highly influential experiments in the 1970s and 1980s (Rosch 1975). Good summaries can be found in (Mervis and Rosch 1981), (Rosch 1999), and in Chapter 1 of (Rogers and McClelland 2008).

[420][Phil] An easy to find source for Wittgenstein's discussion of "game" is (Wittgenstein 2002) in a collection of core readings for cognitive psychology (Levitin 2002).

[421][Phil] The philosopher's poll that ranked Wittgenstein's book #1 is reported by (Lackey 1999).

[422][Phil] It might be possible to define "game,"but it requires a great deal of abstraction that obscures the "necessary and sufficient" tests. "To play a game is to engage in activity directed toward bringing about a specific state of affairs, using only means permitted by specific rules, where the means permitted by the rules are more limited in scope than they would be in the absence of the rules, and where the sole reason for accepting such limitation is to make possible such activity." (Suits 1967)

[423][CogSci] The exact nature of the category representation to which the similarity comparison is made is a subject of ongoing debate in cognitive science. Is it a *prototype*, a central tendency or average of the properties shared by category members, or it one or more *exemplars,* particular members that typify the category. Or is it neither, as argued by connectionist modelers who view categories as patterns of network activation without any explicitly stored category representation? Fortunately, these distinctions do not matter for our discussion here. A recent review is (Rips, Smith, and Medin 2012).

[424][CogSci] Another situation where similarity has been described as a "mostly vacuous" explanation for categorization is with abstract categories or metaphors. Goldstone says "an unrewarding job and a relationship that cannot be ended may both be metaphorical prisons... and may seem similar in that both conjure up a feeling of being trapped... but this feature is almost as abstract as the category to be explained." (Goldstone 1994), p. 149.

[425][CogSci] (Medin, Goldstone, and Gentner 1993) and (Tenenbaum and Griffiths 2001).

[426][CogSci] Because Tversky's model separately considers the sets of non-overlapping features, it is possible to accurately capture similarity judgments when they are not symmetric, i.e., when A is judged more similar to B than B is to A. This framing effect is well-established in the psychological literature and

many machine learning algorithms now employ asymmetric measures. (Tversky 1974)

[427][Com] For a detailed explanation of distance and transformational models of similarity, see (Flach 2012), Chapter 9. There are many online calculators for Levenshein distance; http://www.let.rug.nl/kleiweg/lev/ also has a compelling visualization. The "strings" to be matched can themselves be transformations. The "soundex" function is very commonly used to determine if two words could be different spellings of the same name. It "hashes" the names into phonetic encodings that have fewer characters than the text versions. See (Christen 2006) and `http://www.searchforancestors.com/utility/soundex.html` to try it yourself.

[428][CogSci] This explanation for expert-novice differences in categorization and problem solving was proposed in (Chi et al 1981). See (Linhares 2007) for studies of abstract reasoning by chess experts.

[429][CogSci] (Barsalou 1983).

[430][CogSci] The emergence of theory-based categorization is an important event in cognitive development that has been characterized as a shift from "holistic" to "analytic" categories or from "surface properties" to "principles." See (Carey and Gelman 1991) (Rehder and Hastie 2004).

[431][CogSci] (Tenenbaum 2000) argues that this preference for the most specific hypothesis that fits the data is a general principle of Bayesian learning with random samples.

[432][Com] Consider what happens if two businesses model the concept of "address" in a customer database with different granularity. One may have a coarse "Address" field in the database, which stores a street address, city, state, and Zip code all in one block, while the other stores the components "StreetAddress," "City," and "PostalCode" In separate fields. The more granular model can be automatically transformed into the less granular one, but not vice versa (Glushko and McGrath 2005).

[434][DS] Statistician and baseball fan Nate Silver rejected a complex system that used twenty-six player categories for predicting baseball performance because "it required as much art as science to figure out what group a player belonged in." (Silver 2012, p, 83). His improved system used the technique of "nearest neighbor" analysis to identify current baseball players whose minor league statistics were most similar to the current minor league players being evaluated. (See §7.5.3.3 Categories Created by Clustering (page 372)).

Silver later became famous for his extremely accurate predictions of the 2008 US presidential elections. He is the founder and editor of the FiveThirtyEight blog, so named because there are 538 senators and representatives in the US Congress.

[435][CogSci] (Rosch 1999) calls this the principle of cognitive economy, that "what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible. [...] It is to the organism's advantage not to differentiate one stimulus from another when that differentiation is irrelevant to the purposes at hand." (Pages 3-4.)

[436][Ling] For example, some linguists think of "English" as a broad category encompassing multiple languages or dialects, such as "Standard British English," "Standard American English," and "Appalachian English."

If we are concerned with linguistic diversity and the survival of minority languages, we might categorize some languages as endangered in order to mobilize language preservation efforts. We could also categorize languages in terms of shared linguistic ancestors ("Romance languages," for example), in terms of what kinds of sounds they make use of, by how well we speak them, by regions they are commonly spoken in, whether they are signed or unsigned, and so on. We could also expand our definition of the languages category to include artificial computer languages, or body language, or languages shared by people and their pets—or thinking more metaphorically, we might include the language of fashion.

[437][Com] For example, you can test whether a number is prime by dividing it by every number smaller than its square root, but this algorithm is ridiculously impractical for any useful application. Many cryptographic systems multiply prime numbers to create encryption keys, counting on the difficulty of factoring them to protect the keys; so, proving that ever larger numbers are prime is very important. See (Crandall and Pomerance 2006).

If you are wondering why prime numbers aren't considered an enumerative category given that every number that is prime already exists, it is because we have not found all of them yet, and we need to test through to infinity.

[438][CogSci] This example comes from (Perlman 1984), who introduced the idea of "natural artificial languages" as those designed to be easy to learn and use because they employ mnemonic symbols, suggestive names, and consistent syntax.

[439][Law] When the US Congress revised copyright law in 1976 it codified a "fair use" provision to allow for some limited uses of copyrighted works, but fair use in the digital era is vastly different today; website caching to improve performance and links that return thumbnail versions of images are fair uses that were not conceivable when the law was written. A law that precisely defined fair uses using contemporary technology would have quickly become obsolete, but one written more qualitatively to enable interpretation by the courts has remained viable. See (Samuelson 2009).

[440][Ling] (Wilkins 1668) and (Borges 1952)

[441][Com] "Rigid" might sound negative, but a rigidly defined resource is also precisely defined. Precise definition is essential when creating, capturing, and retrieving data and when information about resources in different organizing systems needs to be combined or compared. For example, in a traditional relational database, each table contains a field, or combination of fields, known as a primary key, which is used to define and restrict membership in the table. A table of email messages in a database might define an email message as a unique combination of sender address, recipient address, and date/time when the message was sent, by enforcing a primary key on a combination of these fields. Similar to category membership based on a single, monothetic set of properties, membership in this email message table is based on a single set of required criteria. An item without a recipient address cannot be admitted to the table. In categorization terms, the item is not a member of the "email message" class because it does not have all the properties necessary for membership.

[442][Com] Like *data schemas*, programming classes specify and enforce rules in the construction and manipulation of data. However, programming classes, like other implementations that are characterized by specificity and rule enforcement, can vary widely in the degree to which rules are specified and enforced. While some class definitions are very rigid, others are more flexible. Some languages have abstract types that have no instances but serve to provide a common ancestor for specific implemented types.

[443][CogSci] The existence of chapters might suggest that an item is a novel; however, a lack of chapters need not automatically indicate that an item is not a novel. Some novels are hypertexts that encourage readers to take alternative paths. Many of the writings by James Joyce and Samuel Beckett are "stream of consciousness" works that lack a coherent plot, yet they are widely regarded as novels.

[444][DS] See (Silver 2012). Over reliance on data that is readily available is a decision-making heuristic proposed by (Tversky and Kahneman 1974), who developed the psychological foundations for behavioral economics. (See the sidebar, Behavioral Economics (page 495).)

[445][DS] To be precise, this "difference of proportions" calculation uses an algorithm that also uses the logarithm of the proportions to calculate entropy, a measure of the uncertainty in a probability distribution. An entropy of zero means that the outcome can be perfectly predicted, and entropy increases as outcomes are less predictable. The information gain for an attribute is how much it reduces entropy after it is used to subdivide a dataset.

[446][Bus] Unfortunately, this rational data-driven process for classifying loan applications as "Approved" or "Denied" was abandoned during the "housing bubble" of the early 2000s. Because lending banks could quickly sell their mortgages to investment banks who bundled them into mortgage-backed securities, appli-

cants were approved without any income verification for "subprime" loans that initially had very low adjustable interest rates. Of course, when the rates increased substantially a few years later, defaults and foreclosures skyrocketed. This sad story is told in an informative, entertaining, but depressing manner in *"The Big Short"* (Lewis, 2010) and in a 2015 movie with the same name.

[447][Com] Machine learning algorithms differ in which properties they use in how they select them. A straightforward method is to run the algorithms using different sets of properties, and select the set that yields the best result. However, it can be very computationally expensive to run algorithms multiple times, especially when the number of properties is large. A faster alternative is to select or filter features based on how well they predict the classification. The information gain calculation discussed in §7.5.3.1 Probabilistic Decision Trees (page 367) is an example of a filter method.

Naïve Bayes classifiers make the simplifying assumption that the properties are independent, an assumption that is rarely correct, which is why the approach is called naïve. For example, a document that contains the word "insurance" is also likely to contain "beneficiary," so their presence in messages is not independent.

Nevertheless, even though the independence assumption is usually violated, Naive Bayes classifiers often perform very well. Furthermore, treating properties as independent means that the classifier needs much less data to train than if we had to calculate the conditional probabilities of all combination of properties. Instead, we just have to count separately the number of times each property occurs with each of the two classification outcomes.

[448][Com] See (Blanzieri and Bryl 2009) for a review of the spam problem and the policy and technology methods for fighting it. (Upsana and Chakravarty 2010) is somewhat more recent and more narrowly focused on text classification techniques.

A very thorough yet highly readable introduction to Active Learning is (Settles 2012).

[449][Com] In particular, documents are usually represented as vectors of frequency-weighted terms. Other approaches start more directly with the similarity measure, obtained either by direct judgments of the similarity of each pair of items or by indirect measures like the accuracy in deciding whether two sounds, colors, or images are the same or different. The assumption is that the confusability of two items reflects how similar they are.

[450][Com] Unlike hierarchical clustering methods that have a clear stopping rule when they create the root category, k-means clustering methods run until the centroids of the categorize stabilize. Furthermore, because the k-means algorithm is basically just hill-climbing, and the initial category "seed" items are

random, it can easily get stuck in a local optimum. So it is desirable to try many different starting configurations for different choices of K.

[451][DS] In addition, the complex feature representations of neural networks compute very precise similarity measurements, which enable searches for specific images or that find duplicate ones.

[452][CogSci] Structure Mapping theory was proposed in (Gentner 1983), and the Structure Mapping Engine followed a few years later (Falkenhainer et al 1989). The SME was criticized for relying on hand-coded knowledge representations, a limitation overcome by (Turney 2008), who used text processing techniques to extract the semantic relationships used by Structure Mapping.

Chapter 8

Classification: Assigning Resources to Categories

Robert J. Glushko
Jess Hemerly
Vivien Petras
Michael Manoochehri
Longhao Wang
Jordan Shedlock
Daniel Griffin

8.1 Introduction

In Chapter 6 we discussed different types of semantic relationships and contrasted abstract relationships between categories that define a semantic hierarchy like

Meat → is-a → Food

with concrete relationships involving specific people like members of the Simpson family:

Homer Simpson → is-a → Husband

When we make an assertion that a particular instance like Homer Simpson is a member of class, we are *classifying* the instance.

Classification, the systematic assignment of resources to intentional categories, is the focus of this chapter. In Chapter 7, *Categorization: Describing Resource Classes and Types*, we described categories created by people as cognitive and linguistic models for applying prior knowledge and we discussed a set of principles for creating categories and category systems. We explained how cultural categories serve as the foundations upon which individual and institutional categories are based. Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories enable *classification* to be purposeful and principled. Computational categories inherited by supervised learning techniques are usually as interpretable as those created by people, but categories created by unsupervised machine learning techniques are statistical patterns that might or might not be interpretable.

A system of categories and its attendant rules or access methods is typically called a *classification scheme* or just the *classifications*. A system of categories captures the distinctions and relationships among its resources that are most important in a domain and for a particular context of use, creating a reference model or conceptual roadmap for its users. This classification creates the structure and support for the interactions that human or computational agents perform. For example, research libraries and bookstores do not use the same classifications to organize books, but the categories they each use are appropriate for their contrasting types of collections and the different kinds of browsing and searching activities that take place in each context. Likewise, the scientific classifications for animals used by biologists contrast with those used in pet stores because the latter have no need for the precise differentiation enabled by the former.

Navigating This Chapter

Most of the chapter is a survey of topics that span the broad range of how classifications are used in organizing systems. These include enumerative classification (§8.3), faceted classification (§8.4), activity-based classification (§8.5), and computational classification (§8.6). Because classification and standardization are closely related, we also analyze standards and standards making as they apply to organizing systems. Throughout, we observe how personal, institutional, cultural, linguistic, political, religious, and even artistic biases can affect otherwise principled and purposeful classification schemes. We finish the chapter with §8.7 Key Points in Chapter Eight (page 429).

8.1.1 Classification vs. Categorization

Classification requires a system of categories, so not everyone distinguishes classification from categorization. Batley, for example, says classification is "imposing some sort of structure on our understanding of our environment," a vague definition that applies equally well to categorization.

In the discipline of organizing, the definition of classification is narrower and more formal. The contrasts among cultural, individual, and institutional categories in §7.2 The What and Why of Categories (page 325) yield a precise definition of classification: *The systematic assignment of resources to a system of intentional categories, often institutional ones.* This definition highlights the intentionality behind the system of categories, the systematic processes for using them, and implies the greater requirements for *governance* and *maintenance* that are absent for cultural categories and most individual ones.

8.1.2 Classification vs. Tagging

Precise and reliable classification is possible when the shared properties of a collection of resources are used in a principled and systematic manner. This method of classification is essential to satisfy institutional and commercial purposes. However, this degree of rigor might be excessive for personal classifications and for classifications of resources in social or informal contexts.

Instead, a weaker approach to organizing resources is to use any property of a resource and any vocabulary to describe it, regardless of how well it differentiates it from other resources to create a system of categories. This method of organizing resources is most often called *tagging* (§5.2.2.3), but it has also been called *social classification*.[454][Web]

Tagging is often used in personal organizing systems, but is social when it serves goals to convey information, develop a community, or manage reputation. Regardless of its name, however, tagging is popular for organizing and rating photos, websites, email messages, or other web-based resources or web-based descriptions of physical resources like stores and restaurants.

The distinction between classification and tagging was blurred when Thomas Vander Wal coined the term "folksonomy" —combining "folk" and "taxonomy" (which is a classification; see §6.3.1.1 Inclusion (page 279)) —to describe the collection of tags for a particular web site or application.[455][Web] Folksonomies are often displayed in the form of a *tag cloud*, where the frequency with which the tag is used throughout the site determines the size of the text in the tag cloud. The tag cloud emerges through the bottom-up aggregation of user tags and is a statistical construct, rather than a semantic one.[456][IA]

Tagging seems insufficiently principled to be considered classification. Tagging a photo as "red" or "car" is an act of resource description, not classification, be-

cause the other tags that would serve as the alternative classifications are unspecified. Furthermore, when tagging principles are followed at all, they are likely to be idiosyncratic ones that were not pre-determined or arrived at through an analysis of goals and requirements.

Noticeably, some uses of tags treat them as category labels, turning tagging into classification. Many websites and resources encourage users to assign "Like" or "+1" tags to them, and because these tags are pre-defined, they are category choices in an implied classification system; for example, we can consider "Like" as an alternative to a "Not liked enough" category.

When users or communities establish sets of principles to govern their tagging practices, tagging is even more like classification. Such a tagging system can be called a *tagsonomy*, a neologism we have invented to describe more systematic tagging. For example, a tagsonomy could predetermine tags as categories to be assigned to particular contents of a blog post, or specify the level of abstraction and granularity for assigning tags without predetermining them (§7.4 Category Design Issues and Implications (page 356)). Some people use multiple user accounts for the same application to establish distinct personas or contexts (e.g., personal vs. business photo collections) as a way to make their tagsonomies more distinct.

Making these decisions about tagging content and form and applying them in the tagging process transforms an *ad hoc* set of tags into a principled tagsonomy. When tagging is introduced in a business setting, more pragmatic purposes and more systematic tagging—for example, by using tags from lists of departments or products—also tends to create tagsonomic classification.[457][Bus]

"Tagging documents by computer," or multi-label classification, is a glib way to describe topic modeling, an unsupervised learning technique for organizing and summarizing collections of unstructured documents by discovering patterns or clusters in the words they contain. The basic intuition behind topic modeling is that the words in a document are probabilistic indications of what the document is about; a document that contains words like "election, government, and candidate" is probably about the "politics" topic, while words like "adore, wedding, and marriage" are good indications of a "love" topic. Topic models are not quite tagging because the words they identify to describe documents are not atomic tags or labels explicitly assigned to individual documents. Instead, topics are more like themes that different documents are more or less likely to contain.

Topic models have been used to implement user interfaces for browsing large document collections because they let a user explore using themes instead of specific search terms. In digital humanities, topic models have been used to discover changes in "what's written about" by some author or resource (like a newspaper) over time. Web commerce companies use topic models to organize books or products for their recommendation engines.[458][DS]

8.1.3 Classification vs. Physical Arrangement

We have often stressed the principle in the discipline of organizing that logical issues must be separated from implementation issues. (See §1.6 The Concept of "Organizing Principle" (page 43), §5.3.5 Designing the Description Form (page 251), and §6.7 The Implementation Perspective (page 308)) With classification we separate the conceptual act of assigning a resource to a category from the subsequent but often incidental act of putting it in some physical or digital storage location. This focus on the logical essence of classification is elegantly expressed in a definition by Gruenberg: Classification is "a higher order thinking skill requiring the fusion of the naturalist's eye for relationships... with the logician's desire for structured order... the mathematician's compulsion to achieve consistent, predictable results... and the linguist's interest in explicit and tacit expressions of meaning."

Taking a conceptual or cognitive perspective on classification contrasts with much conventional usage in library science, where classification is mostly associated with arranging tangible items on shelves, emphasizing the "parking" function that realizes the "marking" function of identifying the category to which the resource belongs.

From a library science or collection curation perspective, it seems undeniable that when the resources being classified are physical or tangible things such as books, paintings, animals, or cooking pots, the end result of the classification activity is that some resource has been placed in some physical location. Moreover, the placement of physical resources can be influenced by the physical context in which they are organized. Once placed, the physical context often embodies some aspects of the organization when similar or related resources are arranged in nearby locations. In libraries and bookstores, this adjacency facilitates the serendipitous discovery of resources, as anyone well knows who has found an interesting book by browsing the shelves.

However, once we broaden the scope of organizing to include digital resources, it is clear that we rely on their logical classifications when we interact with them, not whether they reside on a computer in Berkeley or Bangalore. It is better to emphasize that a classification system is foremost a specification for the logical arrangement of resources because there are usually many possible and often arbitrary mappings of logical references to physical locations.

8.1.4 Classification Schemes

A classification scheme is a realization of one or more organizing principles. Physical resources are often classified according to their tangible or perceivable properties. As we discussed in §7.3.2 Single Properties (page 338) and §7.3.3 Multiple Properties (page 340), when properties take on only a small set of discrete values, a classification system naturally emerges in which each catego-

ry is defined by one property value or some particular combination of property values. Classification schemes in which all possible categories to which resources can be assigned are defined explicitly are *enumerative*. For example, the *enumerative classification* for a personal collection of music recorded on physical media might have categories for CDs, DVDs, vinyl albums, 8-track cartridges, reel-to-reel tape, and tape cassettes; every music resource fits into one and only one of these categories.

When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is *hierarchical* or *taxonomic*. (See §6.3.1.1 Inclusion (page 279).)

For information resources, their *aboutness* is usually more important than their physical properties. For example, a professor planning a new course might organize candidate articles for the syllabus in a fixed set of categories, one for each potential lecture topic. But it is more challenging to enumerate all the subjects or topics that a larger collection of resources might be about. The Library of Congress Classification (LCC) is a hierarchical and enumerative scheme with a very detailed set of subject categories because books can be about almost anything. We discuss the LCC more in §8.3 Bibliographic Classification (page 412).

In addition to or instead of their *aboutness,* information resources are sometimes organized using intrinsic properties like author names or creation dates. Our professor might primarily organize his collection of articles by author name, and when he plans a new course, he might put those he selects for the syllabus into a classification system with one category for every scheduled lecture.

Because names and dates can take on a great many values, an organizing principle like *alphabetical* or *chronological* ordering is unlikely to enumerate in advance an explicit category for each possible value. Instead, we can consider these organizing principles as creating an *implicit or latent* classification system in which the categories are generated only as needed. For example, the Q category only exists in an alphabetical scheme if there is a resource whose name starts with Q.

Many resource domains have multiple properties that might be used to define a classification scheme. For example, wine can be classified by type of grape (varietal), color, flavor, price, winemaker, region of origin (appellation), blending style, and other properties. Furthermore, people differ in their knowledge or preferences about these properties; some people choose wine based on its price and varietal, while others studiously compare winemakers and appellations. Each order of considering the properties creates a different hierarchical classification, and using all of them would create a very deep and unwieldy system. Moreover, many different hierarchies might be required to satisfy divergent preferences. An alternative classification scheme for domains like these is

faceted classification, a type of classification system that takes a set of resource properties and then generates only those categories for combinations that actually occur.

The most common types of facets are enumerative (mutually exclusive); Boolean (yes or no); hierarchical or taxonomic (logical containment); and spectrum (a range of numerical values). We discuss *faceted classification* in detail (in §8.4 Faceted Classification (page 416)) because it is very frequently used in online classifications. Faceted schemes enable easier search and browsing of large resource collections like those for retail sites and museums than hierarchical enumerative schemes. In library science a classification system that builds categories by combination of facets is sometimes also called *analytico-synthetic*.

8.1.5 Classification and Standardization

Classifications impose order on resources. Standards do the same by making distinctions, either implicitly or explicitly, between "standard" and "nonstandard" ways of creating, organizing, and using resources. Classification and standardization are not identical, but they are closely related. Some classifications become standards, and some standards define new classifications. Institutional categories (§7.2.3) are of two broad types.

8.1.5.1 Institutional Taxonomies

Institutional taxonomies are classifications designed to make it more likely that people or computational agents will organize and interact with resources in the same way. Among the thousands of standards published by the *International Organization for Standardization (ISO)* are many institutional taxonomies that govern the classification of resources and products in agriculture, aviation, construction, energy, healthcare, information technology, transportation, and almost every industry sector.[461][Bus]

Institutional taxonomies are especially important in libraries and knowledge management. The Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) enable different libraries to arrange books in the same categories, and the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* in clinical psychology enables different doctors to assign patients to the same diagnostic and insurance categories.

8.1.5.2 Institutional Semantics

Systems of *institutional semantics* offer precisely defined abstractions or *information components* (§4.3.3 Identity and Information Components (page 185)) needed to ensure that information can be efficiently exchanged and used. Organizing systems that use different information models often cannot share and combine information without tedious negotiation and excessive rework.

Automating transactions with suppliers and customers in a supply chain requires that all the parties use the same data format or formats that can be transformed to be interoperable. Retrofitting or replacing these applications to enable efficient interoperability is often possible, and it is usually desirable for the firm to develop or adopt enterprise standards for information exchange models rather than pay the recurring transaction costs to integrate or transform incompatible formats.

Standard semantics are especially important in industries or markets that have significant network effects where the value of a product depends on the number of interoperable or compatible products—these include much of the information and service economies.

An example of a system of institutional semantics is the Universal Business Language (UBL) a library of about 2000 semantic "building blocks" for common concepts like "Address," "Item," "Payment," and "Party" along with nearly 100 document types assembled from the standard components. UBL is widely used to facilitate the automated exchange of transactional documents in procurement, logistics, inventory management, collaborative planning and forecasting, and payment.[463][Bus]

8.1.5.3 Specifications vs. Standards

Implementing an organizing system of significant scope and complexity in a robust and maintainable fashion requires precise descriptions of the resources it contains, their formats, the classes, relations, structures and collections in which they participate, and the processes that ensure their efficient and effective use. Rigorous descriptions like these are often called "specifications" and there are well-established practices for developing good ones.

There is a subtle but critical distinction between "specifications" and "standards." Any person, firm, or *ad hoc* group of people or firms can create a specification and then use it or attempt to get others to use it.[464][Bus] In contrast, a standard is a published specification that is developed and maintained by consensus of all the relevant stakeholders in some domain by following a defined and transparent process, usually under the auspices of a recognized standards organization.[465][Bus] In addition, implementations of standards often are subject to conformance tests that establish the completeness and accuracy of the implementation. This means that users can decide either to implement the specification themselves or choose from other conforming implementations.

The additional rigor and transparency when specifications are developed and maintained through a standards process often makes them fairer and gives them more legitimacy. Governments often require or recommend these *de jure* standards, especially those that are "open" or "royalty free" because they are

typically supported by multiple vendors, minimizing the cost of adoption and maximizing their longevity.

For example, work on UBL has gone on for over a decade in a technical committee under the auspices of a standards development consortium called the Organization for the Advancement of Structured Information Standards (OASIS), which has developed scores of standards for web services and information-intensive industries.

Despite these important distinctions between "specifications" and "standards," however, in conventional usage "standard" is often simply a synonym for "dominant or widely-adopted specification." These *de facto* standards, in contrast with the *de jure* standards created by standards organizations, are typically created by the dominant firm or firms in an industry, by a new firm that is first to use a new technology or innovative method, or by a non-profit entity like a foundation that focuses on a particular domain.[466][Bus]

De facto standards and *ad hoc* standards often co-exist and compete in "standards wars," especially in information-intensive domains and industries with rapid innovation. Standards "wars" tend to occur when different firms or groups of firms develop two or more standards that tend to address the same needs. Not surprisingly, the competing standards are often incompatible on purpose. At first this lets each standard attract customers with features not enabled by the other, but it ends up locking them in by imposing switching costs. Current examples include Google vs. Apple on mobile phones and Kindle versus Apple on ebook readers.[467][Bus]

For example, the Dewey Decimal Classification (DDC) is the world's most widely used library classification system, and most people treat it as a standard. In fact, the DDC is proprietary and it is maintained and licensed for use by the Online Computer Library Center (OCLC). Similarly, the DSM is maintained and published by the *American Psychiatric Association (APA)* and it earns the APA many millions of dollars a year.

In contrast, *de jure* standards include the Library of Congress Classification (LCC), developed under the auspices of the US government, the familiar MARC record format used in online library catalogs (ISO 2709), and its American counterpart ANSI Z39.2.

As a result, even though it would be technically correct to argue that "while all standards are specifications, not all specifications are standards," this distinction is hard to maintain in practice.

8.1.5.4 Mandated Classifications

Standards are often imposed by governments to protect the interests of their citizens by coordinating or facilitating activities that might otherwise not be possible or safe. Some of them primarily concern public or product safety and are only tangentially relevant to systems for organizing information. Others are highly relevant, especially those that specify the formats and content of information exchange; many European governments require firms doing business with the government to adopt UBL.[469][Law]

Other government standards that are important in organizing systems are those that express requirements for classification and retention of auditing information for financial activities, such as the *Sarbanes-Oxley Act,* or for non-retention of personal information, such as HIPAA and FERPA.[470][Bus]

8.2 Understanding Classification

Classifications arrange resources to support discovery, selection, combination, integration, analysis, and other purposeful activity in every organizing system. A classification of diseases facilitates diagnosis and development of medical procedures, as well as accounting and billing. In addition, classifications facilitate understanding of a domain by highlighting the important resources and relationships in it, supporting the training of people who work in the domain and their acquisition of specialized skills for it.

We consider classification to be systematic when it follows principles that govern the structure of categories and their relationships. However, being systematic and principled does not necessarily ensure that a classification will be unbiased or satisfy all users' requirements. For example, the zoning, environmental, economic development, and political district classifications that overlay different parts of a city determine the present and future allocation of services and resources, and over time influence whether the city thrives or decays. These classifications reflect tradeoffs and negotiations among numerous participants, including businesses, lobbyists, incumbent politicians, donors to political parties, real estate developers, and others with strong self-interests.

8.2.1 Classification Is Purposeful

Categories often arise naturally, but by definition classifications do not because they are systems of categories that have been intentionally designed for some purpose. Every classification brings together resources that go together, and in doing so differentiates among them. However, bringing resources together would be pointless without reasons for finding, accessing, and interacting with them later.

8.2.1.1 Classifications Are Reference Models

A classification creates a semantic or conceptual roadmap to a domain by highlighting the properties and relationships that distinguish the resources in it. This reference model facilitates learning, comprehension, and use of organizing systems within the domain. Standard classifications like those used in libraries enable people to rely on one system that they can use to locate resources in many libraries. Standard business, job, and product classifications enable the reliable collection, analysis, and interchange of economic data and resources.

Another important use of standard classifications created by people is as a "gold standard" for comparison with unsupervised computational classifications carried out on the same collection of resources or in the same domain. Presumably no unsupervised classifier could exactly reproduce the classifications created by careful experts.

8.2.1.2 Classifications Support Interactions

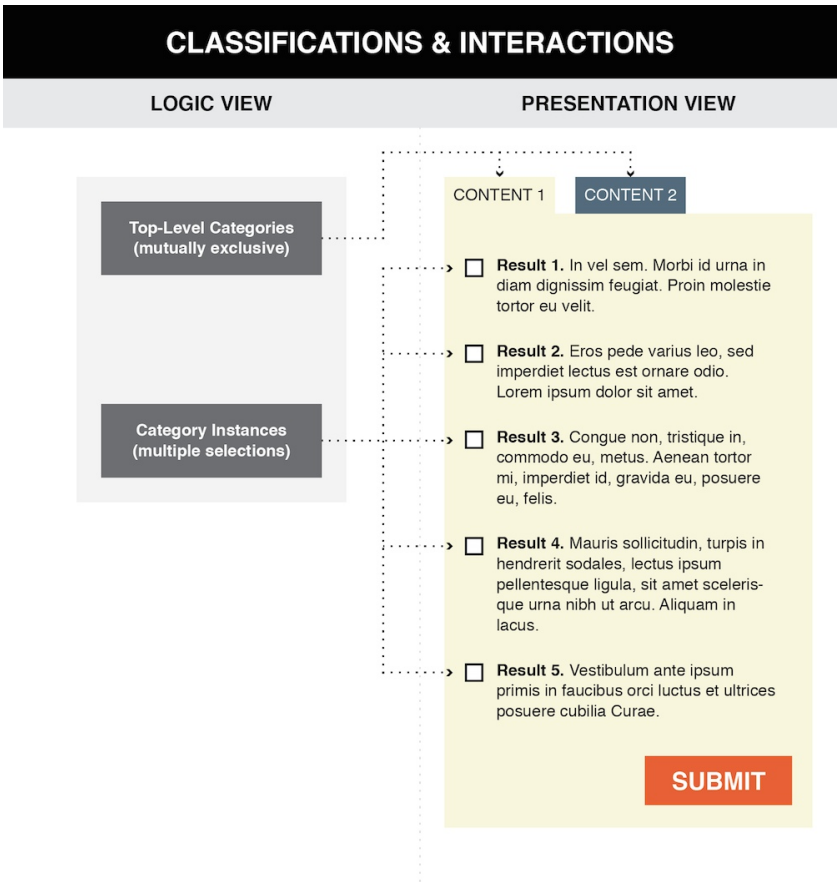
A classification creates structure in the organizing system that increases the variety and capability of the interactions it can support. With physical resources, classification increases useful co-location; in kitchens, for example, keeping resources that are used together near each other (e.g., baking ingredients) makes cooking and cleanup more efficient (see "activity-based" classification in §8.5).

Classification makes systems more usable when it is manifested in the arrangement of resource descriptions or controls in user interface components like list boxes, tabs, buttons, function menus, and structured lists of search results.[471][IA]

A typical mapping between the logic of a classification scheme and a user interface is illustrated in Figure 8.1, Classification and Interactions.

How a business classifies its product or service strongly influences whether a customer can find it; this is the essential task of marketing. The business of "search engine optimization" exists to help a firm with a web presence choose the categories and descriptive terms that will improve its ranking in search results and attract the number of types of customer it desires.[472][Web] How a customer interacts with a supplier is influenced by how the supplier classifies its offerings in its shopping aisles or catalogs; the "science of shopping" uses creative classifications and co-location of goods to shape browsing behavior and encourage impulse buying.[473][Bus] In business-to-business contexts, standard classifications for business processes and their application interfaces enable firms to more easily build and maintain supply chains and distribution networks that interconnect many business partners.[474][Bus]

Figure 8.1. Classification and Interactions.



Good user interface design creates a clear mapping between the logic of a classification scheme and the selection methods and arrangements presented to users. Categories that are mutually exclusive imply different tabs or other visualizations that imply a single selection, for example.

Classification In A Novel User Interface

The meat from animals used as food is classified into numerous "cuts" based on its origin. In the US, these classifications are standardized by the Department of Agriculture to ensure that meat is labeled correctly. The most natural way to convey the classification system is to label the parts of the animal in a diagram, because this binds each logical category to the "user interface."

(Photo by R. Glushko. Taken in 2011 at the Union Square Greenmarket in New York City.)

8.2.2 Classification Is Principled

§7.3 Principles for Creating Categories (page 337) explained principles for creating categories, including enumeration, single properties, multiple properties and hierarchy, probabilistic co-occurrence of properties, theory and goal-based categorization. It logically follows that the principles considered in designing categories are embodied in classifications that use those categories. However, when we say, "classification is principled," we are going further to say that the processes of assigning resources to categories and maintaining the classification scheme over time must also follow principles.

The design and use of a classification system involves many choices about its purposes, scope, scale, intended lifetime, extensibility, and other considerations. Principled classification means that once those design choices are made they should be systematically and consistently followed.

Principled does not necessarily equate to "good," because many of the choices can be arbitrary and others may involve tradeoffs that depend on the nature of the resources, the purposes of the classification, the amount of effort available, the complexity of the domain, and the capabilities of the people doing the classification and of the people using it (see §7.4 Category Design Issues and Implications (page 356)). Every classification system is biased in one way or another (see §8.3 Bibliographic Classification (page 412)).

Consider the classifications of resources in a highly-organized kitchen. (See §12.5 Organizing a Kitchen (page 566)). Tableware, dishes, pots and pans, spices and food provisions, and other resources have dedicated locations determined by a set of intersecting requirements and organizing principles. There is no written specification, and other people organize their kitchens differently.

On the other hand, complex institutional classification systems like those used in libraries or government agencies are implemented with detailed specifications, methods, protocols, and guidelines. The people who apply these methods in the field have studied the protocols in school or they have received extensive on-the-job training to ensure that they apply them correctly, consistently, and in accordance with the specifications and guidelines.

8.2.2.1 Principles Embodied in the Classification Scheme

Some of the most important principles that lead us to say that classification is principled are those that guide the design of the classification scheme in the first place. These principles are fundamental in the discipline of library science but they apply more broadly to other domains.

The *warrant* principle concerns the justification for the choice of categories and the names given to them. The principle of *literary warrant* holds that a classification must be based only on the specific resources that are being classified. In the library context, this *ad hoc* principle that builds a classification from a particular collection principle is often posed in opposition to a more philosophical or epistemological perspective, first articulated by Francis Bacon in the seventeenth century, that a classification should be universal and must handle all knowledge and all possible resources. The principle of *scientific warrant* argues that only the categories recognized by the scientists or experts in a domain should be used in a classification system, and it is often opposed by the principle of *use* or *user warrant*, which chooses categories and descriptive terms according to their frequency of use by everyone, not just experts. With classifications of physical resources like those in a kitchen, we see *object warrant*, where

similar objects are put together, but more frequently the justifying principle will be one of use warrant, where resources are organized based on how they are used.

A second principle embodied in a classification scheme concerns the breadth and depth of the category hierarchy. We discussed this in §7.4 Category Design Issues and Implications (page 356) but in the context of classification this principle has additional implications and is framed as the extent to which the scheme is *enumerative* (§8.1.3 Classification vs. Physical Arrangement (page 395)). The decision to classify broadly or precisely depends largely on the variety or heterogeneity of the resources that the system of categories has been designed to organize. Because of the diversity of resources for a sale in a department store, a broad classification is necessary to accommodate everything in the store. Kitchen goods will be grouped together in a few aisles on a single floor. But a specialty kitchen store or a wholesale kitchen supply store for restaurants would classify much more precisely because of the restricted resource domain and the greater expertise of those who want to buy things there. An entire section might be dedicated just to knives, organized by knife type, manufacturer, quality of steel, and other categories that are not used in the kitchen section of the department store.[477][CogSci]

Starbucks Coffee Sizes: "Anti-User" Warrant?

The Starbucks coffee chain seemingly goes out of its way to confuse its customers by calling the smallest (twelve ounces) of its three coffee sizes the "tall" size, calling its sixteen-ounce size a "*grande*," and calling its largest a "*venti*," which is Italian for twenty (ounces). Outside of Starbucks, something that is "tall" is never also considered "small." Ironically, despite having more than five thousand coffeehouses in over fifty countries, Starbucks has none in Italy where *venti* would be in the local language.

The precision or enumerativeness of a classification scheme increases the similarity of resources that are assigned to the same category and sharpens the distinctions between resources in different categories. However, when different classifications must be combined, mismatches in their precision or granularity can create challenges (see §10.3 Reorganizing Resources for Interactions (page 499)).

8.2.2.2 Principles for Assigning Resources to Categories

The *uniqueness principle* means the categories in a classification scheme are mutually exclusive. Thus, when a logical concept is assigned to a particular category, it cannot simultaneously be assigned to another category. Resources, however, can be assigned to several categories if they embody several concepts

represented by those different categories. This can present a challenge when a physical storage solution is based on storing resources according to its assigned category in a logical classification system. This is not a serious problem for resource types like technical equipment or tools, for which the properties used to classify them are highly salient, and that have very narrow and predictable contexts of use. It is also not a problem for highly-specialized information resources like scientific research reports or government economic data, which might end up in only one specialized class. However, many resources are inherently more difficult to classify because they have less salient properties or because they have many more possible uses.

We face this kind of problem all the time. For example, should we store a pair of scissors in the kitchen or in the office? One solution is to buy a second pair of scissors so that scissors can be kept in both locations where they are typically used, but this is not practical for many types of resources and this principle would be difficult to apply in a systematic manner.

Many books are about multiple subjects. A self-help book about coping with change in a business setting might reasonably be classified as either about applied psychology or about business. It is not helpful that book titles are often poor clues to their content; *Who Moved My Cheese?* is in fact a self-help book about coping with change in a business setting. Its Library of Congress Classification is BF 637, "Applied Psychology," and at UC Berkeley it is kept in the business school library.

The general solution to satisfying the *uniqueness principle* in library classifications when resources do not clearly fit in a single category is to invent and follow a detailed set of often arbitrary rules. Usually, the primary subject of the book is used for assigning a category, which will then determine the book's place on a shelf.

8.2.2.3 Principles for Maintaining the Classification over Time

Most personal classifications are created in response to a specific situation to solve an emerging organizational challenge. As a consequence, personal classification systems change in an *ad hoc* or opportunistic manner during their limited lifetimes. For example, the classification schemes in your kitchen or closet are deconstructed and disappear when you move and take your possessions to a different house or apartment. Your efforts to re-implement the classifications will be influenced by the configuration of shelves and cabinets in your new residence, so they will not be exactly the same.

In contrast, the institutional classification schemes for many library resources, culturally or scientifically-important artifacts, and much of the information created or collected by businesses, governments and researchers might have useful lives of decades or centuries. Classification systems like these can only be

changed incrementally to avoid disruption of the work flows of the organization. We described maintaining resources as an activity in all organizing systems (§3.5 Maintaining Resources (page 133)) and the issues of persistence, effectivity, authenticity, and provenance that emerge with resources over time (§4.5 Resources over Time (page 198)). Much of this previous discussion applies in a straightforward manner to maintaining classifications over time.

However, some additional issues arise with classifications over time. The warrant principle (§8.2.2.1) implicitly treats the justification for designing and naming categories as a one-time decision. This is reasonable if you are organizing a collection of bibliographic resources or common types of physical resources like printed books, clothing or butterflies. However, in domains where the resources are active, change their state or implementation, or otherwise have a probabilistic character it might be necessary to revisit warrant and the decisions based on it from time to time. Put another way, if the world that you are sampling from or describing has some randomness or change in it, the categories and descriptions you imposed on it probably need to change as well. It often happens that the meaning of an underlying category can change, along with its relative and absolute importance with respect to the other categories in the classification system. Categories sometimes change slowly, but they can also change quickly and radically as a result of technological, process, or geopolitical innovation or events. Entirely new types of resources and bodies of knowledge can appear in a short time. Consider what the categories of "travel," "entertainment," "computing," and "communication" mean today compared to just a decade or two ago.

Changes in the meaning of the categories in a classification threaten its *integrity*, the principle that categories should not move within the structure of the classification system. One way to maintain integrity while adapting to the dynamic and changing nature of knowledge is to define a new version of a classification system while allowing earlier ones to persist, which preserves resource assignments in the previous version of the classification system while allowing it to change in the new one. If we adopt a logical perspective on classification (§8.1.2 Classification vs. Tagging (page 393)) that dissociates the conceptual assignment of resources to categories from their physical arrangement, there is no reason why a resource cannot have contrasting category assignments in different versions of a classification.

However, the conventional library with collections of physical resources cannot easily abandon its requirement to use a classification to arrange books on shelves in specific places so they can be located, checked out, and returned to the same location.

A related principle about maintaining classifications over time is *flexibility*, the degree to which the classification can accommodate new categories. Computer

scientists typically describe this principle as *extensibility*, and library scientists sometimes describe it as *hospitality*. In any case the concern is the same and we are all familiar with it. When you buy a bookshelf, clothes wardrobe, file cabinet, or computer, it makes sense to buy one that has some extra space to accommodate the books, clothes, or files you will acquire over some future time frame. As with other choices that need to be made about organizing systems, how much extra space and "organizing room" you will acquire involves numerous tradeoffs.

Classification schemes can increase their flexibility by creating extra "logical space" when they are defined. Library classifications accomplish this by using naming or numbering schemes for classification that can be extended easily to create new subcategories. Classification schemes in information systems can also anticipate the evolution of document or database schemas.[481][Com]

8.2.3 Classification Is Biased

The discipline of organizing is fundamentally about choices of properties and principles for describing and arranging resources. We discussed choices about describing resources in §5.3 The Process of Describing Resources (page 227), choices for creating resource categories in §7.3 Principles for Creating Categories (page 337), and choices for creating classifications in this chapter. The choices made reflect the purposes, experiences, professions, politics, values, and other characteristics and preferences of the people making them. As a result, every system of classification is biased because it takes a point of view that is a composite of all of these influences.

But first we need to point out that there are at least two quite different senses of "bias" that people reading this book are likely to encounter. The colloquial sense of bias we discuss in this section reflects value-based decisions in organizing systems that implicitly or explicitly favor some interactions or users over others. In contrast, statistical bias is systematic error or distortion in a measurement. (See the sidebar, Statistical Bias and Variance (page 408).)

The claim that classification is biased might seem surprising, because many classification systems are formal and institutional, created by governments or firms participating in standards organizations. We expect these classifications to be impartial and objective. However, consider the classification of people as "employed" or "unemployed." Many people think that any employable person who is not currently employed would be counted as unemployed. But the US government's Department of Labor only counts someone as unemployed if they have actively looked for work in the past month, effectively removing anyone who has given up on finding work from the unemployed category by assigning them to a "discouraged worker" category. In 2012 this classification scheme allowed the government to report that unemployment was about 8% and falling,

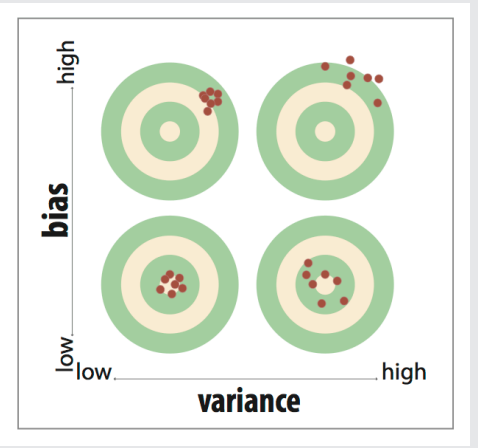
Statistical Bias and Variance

Statistical bias is the systematic error in measurements introduced by miscalibration of the measurement instrument, by ineffective measurement techniques, an algorithm that makes incorrect assumptions, or some environment interference, all of which distort the measured value in a predictable way. Measurement bias contrasts with the variability or variance of a measurement, the amount of dispersion around an average or expected value, most often due to random factors. Some variance arises because the property being measured is not the same for all instances, as we would expect for measurements of the weight of a random sample of people, or in the set of tags or topics assigned to a random sample of news articles by people or algorithms. By analyzing a large enough set of instances it is possible to determine the most likely values of the property and also to estimate the amount of random error.

High variance in the measurements for a sample of resources when we expect all of them to have more similar values can be a quality problem. High bias, on the other hand, might be less of a quality problem, because systematic sources of inaccuracy might be easier to correct.

Bias and Variance on Dartboards

Precise and accurate dart throws demonstrate low bias and low variance (lower left in the figure). Precise but inaccurate darts reflect high bias and low variance (upper left). Imprecise but accurate ones have low bias but high variance (lower right). Finally, a lack of accuracy and precision shows both high bias and high variance (upper right).



when in fact it was closer to 20% and rising. The political implications of this classification are substantial.[482][Law]

Classification bias is often intentionally or unintentionally shown in data visualizations, including choropleth maps, in which map regions are colored, patterned, or otherwise distinguished according to a statistical variable being displayed on the map. Choropleths are commonly used to display election results, with the districts or states won by each candidate shown in different colors; in the United States, the

convention is to show those won by Democratic Party candidates in blue, and those won by Republicans in Red. These election choropleths are often misleading because coloring an entire state in the winner's colors ignores population density and the regional concentrations of votes that differ from the majority. California voters are reliably "blue" as a whole, but as you can see in the nearby figure with election results divided by county, this majority is amassed in the large cities along the coast, and inland and rural counties are more reliably "red" in their voting.

California Election Map: A "Blue" State with "Red" Counties

A more subtle way in which choropleths encode bias reflects the decisions made to organize the data into the categories that are represented by different colors or patterns. Choropleth categories might present data divided into equal range intervals, into sets with the same number of observations, or into categories that reflect clusters or natural breaks in the observed data. Small changes in the data ranges or proportions that are then assigned to each category can communicate entirely different stories with the same data. To learn "how to lie with maps" or how to prevent being lied to, refer to the classic book with that title by Mark Monmonier.[483][IA]

Friedman and Nissenbaum's *Bias In Computer Systems* offers a framework for conceptualizing the various types of bias that may be present in technical systems. Friedman and Nissenbaum define bias as "a system that systematically and unfairly discriminates against individuals or groups of individuals in favor of others" Their taxonomy includes pre-existing, technical, and emergent bias.[484][Com]

Pre-existing bias is the type people are most familiar with: it occurs when an organizing system's design embodies personal or societal biases that exist at the time of its creation, either intentionally or inadvertently, and sometimes despite one's best intentions to prevent it.

Technical bias arises from limitations and constraints of technical systems that result in unfairness when the system is applied to the real world. Automated

decision-making is especially ripe for this sort of bias: alphabetical ordering, processes that rely on pseudo-random number generation, and other automated ways of sorting or grouping resources may systematically create different opportunities for different user groups (e.g., people or companies whose names begin with "A").

Emergent bias is related to the interplay between actual users and a technical system. Problems of this type arise when, due to the designer's incomplete understanding of the user population, or a change in that population over time, there is a mismatch between users and the system. User interfaces are especially susceptible to this form of bias, given their need to reflect the habits and capacities of intended users. Unfairness can emerge when an unexpected user group uses the system, or as new societal knowledge arises that the system is not able to incorporate or respond to.

Both pre-existing and emergent bias may be difficult to assess accurately; the former may be difficult for the biased to see or admit to, and the latter, arising due to unanticipated circumstances after implementation, is hard to predict.

Bowker and Star have written extensively about biases in classification systems but acknowledge that many people do not see them:

> *Information scientists work every day on the design, delegation and choice of classification systems and standards, yet few see them as artifacts embodying moral and aesthetic choices that in turn craft people's identities, aspirations and dignity.*
>
> — (Bowker and Star 2000)

Bowker and Star describe many examples where seemingly neutral and benign classifications implement controversial assumptions. A striking example is found in the ethnic classifications of the United States Census and the categories to which US residents are required to assign themselves. These categories have changed nearly every decade since the first census in 1790 and strongly reflect political goals, prevailing cultural sensitivities or lack thereof, and non-scientific considerations. Some recent changes included a "multi-racial" category, which some people viewed as empowering, but which was attacked by African-American and Hispanic civil rights groups as diluting their power.

A more positive way to think about bias in classification is that the choices made in an organizing system about resource selection, description, and arrangement come together to convey the values of the organizers. This makes a classification a rhetorical or communicative vehicle for establishing credibility and trust with those who interact with the resources in the classification. Seen in this light, an objective or neutral classification is not only unrealistic as a goal; it may also consume valuable time and energy when instead it might be more de-

sirable to seize the opportunity to interpret the resources in a creative way to communicate a particular message to a particular user group. Melanie Feinberg makes the point that "fair trade" or "green" supermarkets differentiate themselves by a relatively small proportion of the goods they offer compared with ordinary stores, but these particular items signal the values that their customers care most about.[487][Bus]

Bias is clearly evident in the most widely used bibliographic classifications, the Library of Congress and the Dewey Decimal, which we discuss next.

8.3 Bibliographic Classification

Much of our thinking about classification comes from the bibliographic domain. Libraries and the classification systems for the resources they contain have been evolving for millennia, shaped by the intellectual, social, and technological conditions of the societies that created them. As early as the third millennium BCE, there were enough written documents—papyrus scrolls or clay tablets—that the need arose to organize them. Some of the first attempts, by Mesopotamian scribes, were simple lists of documents in no particular order. The ancient Greeks, Romans, and Chinese created more principled systems, both sorting works by features such as language and alphabetical order, and placing them into semantically significant categories such as topic or genre. Medieval European libraries were tightly focused on Christian theology, but as secular books and readers proliferated thanks to new technologies and increased literacy, bibliographic classifications grew broader and more complex to accommodate them. Modern classification systems are highly nuanced systems designed to encompass all knowledge; however, they retain some of the same features and biases of their forebears.

We will briefly describe the most important systems for bibliographic classification, especially the Dewey Decimal Classification (DDC) and *Library of Congress Classification (LCC)* systems. However, there are several important ways in which bibliographic classification is distinctive and we will discuss those first:

Scale, Complexity, and Degree of Standardization:
: Department stores and supermarkets typically offer tens of thousands of different items (as measured by the number of "stock keeping units" or SKUs), and popular online commerce sites like Amazon.com and eBay are of similar scale. However, the standard product classification system for supermarkets has only about 300 categories.[489][Bus] The classifications for online stores are typically deeper than those for physical stores, but they are highly idiosyncratic and non-standard. In contrast, scores of university libraries have five million or more distinct items in their collections, and they almost all use the same standard bibliographic classification system that has about 300,000 distinct categories.

Legacy of Physical Arrangement, User Access, and Re-Shelving:
A corollary to the previous one that distinguishes bibliographic classification systems is that they have long been shaped and continue to be shaped by the legacy of physical arrangement, user access to the storage locations, and re-shelving that they support. These requirements constrain the evolution and extensibility of bibliographic classifications, making them less able to keep pace with changing concepts and new bodies of knowledge. Amazon classifies the products it sells in huge warehouses, but its customers do not have to pick out their purchases there, and most goods never return to the warehouse. Amazon can add new product categories and manage the resources in warehouses far more easily than libraries can.

With digital libraries, constraints of scale and physical arrangement are substantially eliminated, because the storage location is hidden from the user and the resources do not need to be returned and re-shelved. However, when users can search the entire content of the library, as they have learned to expect from the web, they are less likely to use the bibliographic classification systems that have painstakingly been applied to the library's resources.

8.3.1 The Dewey Decimal Classification

The Dewey Decimal Classification (DDC) is the world's most widely used bibliographic system, applied to books in over 200,000 libraries in 135 countries. It is a proprietary and *de facto* standard, and it must be licensed for use from the Online Computer Library Center (OCLC).

In 1876, Melvil Dewey invented the DDC when he was hired to manage the Amherst College library immediately after graduating. Dewey was inspired by Bacon's attempt to create a universal classification for all knowledge and considered the DDC as a numerical overlay on Bacon with 10 main classes, each divided into 10 more, and so on. Despite his explicit rejection of literary warrant, however, Dewey's classification was strongly influenced by the existing Amherst collection, which reflected Amherst's focus on the time on the "education of indigent young men of piety and talents for the Christian ministry."

The resulting nineteenth-century Western bias in the DDC's classification of religion seems almost startling today, where it persists in the 23rd revision (see Figure 8.2, "Religion" in Dewey Decimal Classification.). "Religion" is one of the 10 main classes, the 200 class, with nine subclasses, Six of these nine subclasses are topics with "Christian" in the name; one class is for the *Bible* alone; and another section is entitled "Natural theology." Everything else related to the world's many religions is lumped under 290, "Other religions."

The notational simplicity of a decimal system makes the DDC easy to use and easy to subdivide existing categories, So-called subdivision tables allow facets

Figure 8.2. "Religion" in Dewey Decimal Classification.

```
200 Religion
  210 Natural Theology
  220 Bible
  230 Christian theology
  240 Christian moral and devotional theology
  250 Christian orders and local church
  260 Christian social theology
  270 Christian church history
  280 Christian sects and denominations
  290 Other religions
```

for language, geography or format to be added to many classes, making the classification more specific. But the overall system is not very hospitable to new areas of knowledge.

8.3.2 The Library of Congress Classification

The US Library of Congress is the largest library in the world today, but it got off to a bad start after being established in 1800. In 1814, during the War of 1812, British troops burned down the US Capitol building where the library was located and the 3000 books in the collection went up in flames.[493[CogSci]] The library was restarted a year later when Congress purchased the personal library of former president Thomas Jefferson, which was over twice the size of the collection that the British burned. Jefferson was a deeply intellectual person, and unlike the narrow historical and legal collection of the original library, Jefferson's library reflected his "comprehensive interests in philosophy, history, geography, science, and literature, as well as political and legal treatises."

Restarting the Library of Congress around Jefferson's personal collection and classification had an interesting implication. When Herbert Putnam formally created the Library of Congress Classification (LCC) in 1897, he meant it not as a way to organize all the world's knowledge, but to provide a practical way to organize and later locate items within the Library of Congress's collection. However, despite Putnam's commitment to literary warrant, the breadth of Jefferson's collection made the LCC more intellectually ambitious than it might otherwise had been, and probably contributed to its dominant adoption in university libraries.

The LCC has 21 top-level categories, identified by letters instead of using numbers like the DDC (see Figure 8.3, Top Level Categories in the Library of Congress Classification.). Each top-level category is divided into about 10-20 subclasses, each of which is further subdivided. The complete LCC and supporting information takes up 41 printed volumes.

Figure 8.3. Top Level Categories in the Library of Congress Classification.

```
A – GENERAL WORKS
B – PHILOSOPHY. PSYCHOLOGY. RELIGION
C – AUXILLARY SCIENCES OF HISTORY (GENERAL)
D – WORLD HISTORY (EXCEPT AMERICAN HISTORY)
E – HISTORY: AMERICA
F – HISTORY: AMERICA
G – GEOGRAPHY. ANTHROPOLOGY. RECREATION
H – SOCIAL SCIENCE
J – POLITICAL SCIENCE
K – LAW
L – EDUCATION
M – MUSIC
N – FINE ARTS
P – LANGUAGE AND LITERATURE
Q – SCIENCE
R – MEDICINE
S – AGRICULTURE
T – TECHNOLOGY
U – MILITARY SCIENCE
V – NAVAL SCIENCE
Z – BIBLIOGRAPHY. LIBRARY SCIENCE
```

Bias is apparent in the LCC as it is in the DDC, but is somewhat more subtle. A library for the US emphasizes its own history. “Naval science” was vastly more important in the 1800s when it was given its own top level category, separated from other resources about “Military science” (which had a subclass for “Cavalry”).

The LCC is highly enumerative, and along with the uniqueness principle, this creates distortions over time and sometimes requires contortions to incorporate new disciplines. For example, it might seem odd today that a discipline as broad and important as computer science does not have its own second level category under the Q category of science, but because computer science was first taught in math departments, the LCC has it as the QA76 subclass of mathematics, which is QA.[496[CogSci]]

8.3.3 The BISAC Classification

A very different approach to bibliographic classification is represented in the *Book Industry Standards Advisory Committee classification (BISAC)*. BISAC is developed by the *Book Industry Study Group (BISG)*, a non-profit industry association that “develops, maintains, and promotes standards and best practices

that enable the book industry to conduct business more efficiently." The BISAC classification system is used by many of the major businesses within the North American book industry, including Amazon, Baker & Taylor, Barnes & Noble, Bookscan, Booksense, Bowker, Indigo, Ingram and most major publishers.[497][Bus]

The BISAC classifications are used by publishers to suggest to booksellers how a book should be classified in physical and online bookstores. Because of its commercial and consumer focus, BISAC follows a principle of use warrant, and its categories are biased toward common language usage and popular culture. Some top-level BISAC categories, including Law, Medicine, Music, and Philosophy, are also top-level categories in the LCC. However, BISAC also has top-level categories for Comics & Graphic Novels. Cooking, Pets, and True Crime.

The differences between BISAC and the LCC are understandable because they are used for completely different purposes and generally have little need to come into contact. This changed in 2004, when Google began its ambitious project to digitize the majority of the world's books. (See the sidebar, What Is a Library? (page 39)). To the dismay of many people in the library and academic community, Google initially classified books using BISAC rather than the LCC.[498][Bus]

In addition, some new public libraries have adopted BISAC rather than the DDC because they feel the former makes the library friendlier to its users. Some librarians believe that their online catalogs need to be more like web search engines, so a less precise classification that uses more familiar category terms seems like a good choice.

8.4 Faceted Classification

We have noted several times that strictly enumerative classifications constrain how resources are assigned to categories and how the classification can evolve over time. *Faceted classifications* are an alternative that overcome some of these limitations. In a *faceted classification* system, each resource is described using properties from multiple facets, but a person searching for resources does not need to consider all of the properties (and consequently the facets) and does not need to consider them in a fixed order, which an enumerative hierarchical classification requires.

Faceted classifications are especially useful in web user interfaces for online shopping or for browsing a large and heterogeneous museum collection. The process of considering facets in any order and ignoring those that are not relevant implies a dynamic organizational structure that makes selection both flexible and efficient. We can best illustrate these advantages with a shopping example in a domain that we are familiar with from §7.3.3.

If a department store offers shirts in various styles, colors, sizes, brands, and prices, shoppers might want to search and sort through them using properties from these facets in any order. However, in a physical store, this is not possible because the shirts must be arranged in actual locations in the store, with dress shirts in one area, work shirts in another, and so on.

Assume that the shirt store has shirts in four styles: dress shirts, work shirts, party shirts, and athletic shirts. The dress shirts come in white and blue, the work shirts in white and brown, and the party and athletic shirts come in white, blue, brown, and red. White dress shirts come in large and medium sizes.

Suppose we are looking for a white dress shirt in a large size. We can think of this desired shirt in two equivalent ways, either as a member of a category of "large white dress shirts" or a shirt with "dress," "white," and "large" values on style, color, and size facets. Because of the way the shirts are arranged in the physical store, our search process has to follow a hierarchical structure of categories. We go to the dress shirt section, find white shirts, and then look for a large one. This process corresponds to the hierarchy shown in Figure 8.4, Enumerative Classification with Style Facet Followed by Color Facet.

Although unlikely, a store might choose to organize its shirts by color. In our search for a "white dress shirt in a large size," if we consider the color first, because shirts come in four colors, there are four color categories to choose from. When we choose the white shirts, there is no category for work shirts because there are no work shirts that come in white. We then choose the dress shirts, and then finally find the large one. (Figure 8.5, Enumerative Classification with Color Facet Followed by Style Facet.)

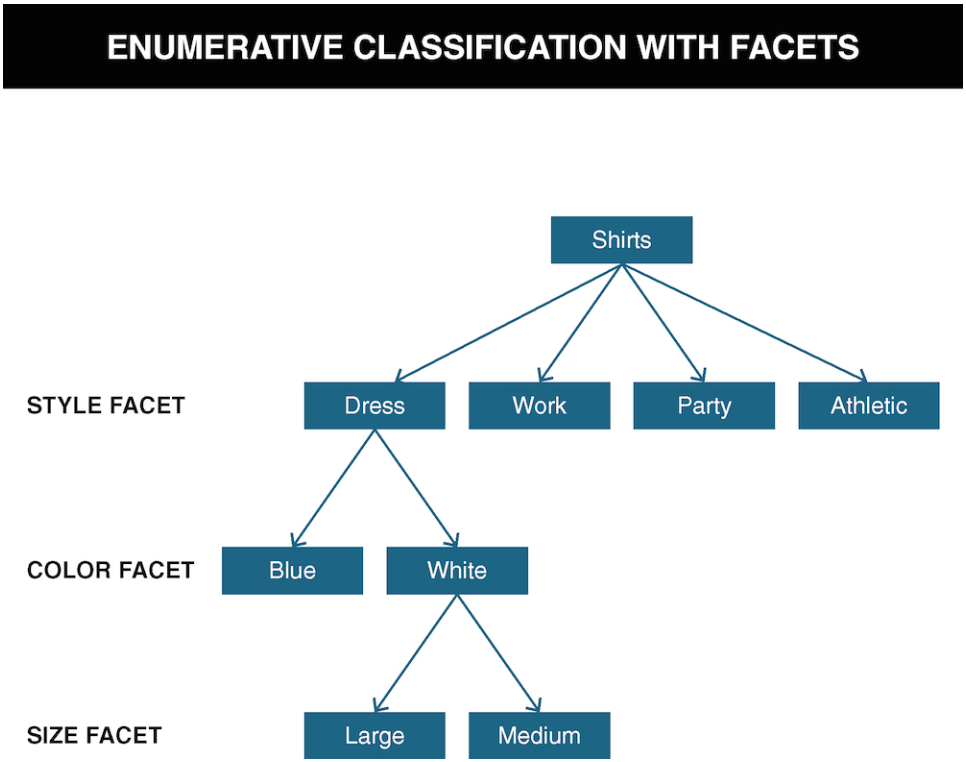
This department store example shows that for a physical organization, one property facet guides the localization of resources; all other facets are subordinated under the primary organizing property. In hierarchical enumerative classifications, this means that the primary organizing facet determines the primary form of access. The shirts are either organized by style and then color, or by color then style, which enforces an inflexible query strategy (style first or color first).

In an online store, however, descriptions of the shirts are being searched and sorted instead of the real shirts, and different organizations are possible. When the shirts are described using a faceted classification system, we treat all facets independently (i.e., they can all be the primary facet).

We can enumerate all the properties needed to assign resources appropriately, but we create the categories (i.e., union of properties from different facets) only as needed to sort resources with a particular combination of properties.

An additional aspect of the flexibility of faceted classification is that a facet can be left out of a resource description if it is not needed or appropriate. For example, because party shirts are often multi-colored with exotic patterns, it is not

Figure 8.4. Enumerative Classification with Style Facet Followed by Color Facet.

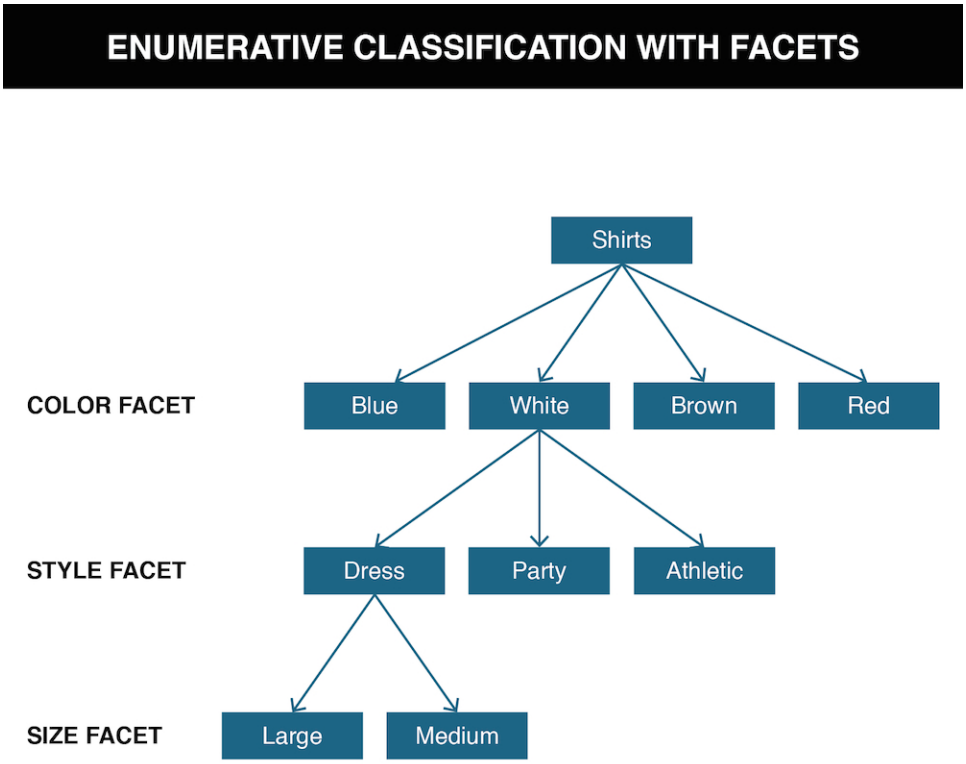


In an enumerative classification system the order of the facets determines the classification hierarchy. For example, a store might classify shirts first using a style facet, next with a color facet, and finally with a size facet. This ordering could result in two piles of dress shirts, one blue and one white, in which each pile contains shirts of large and medium sizes.

that useful to describe their color. Likewise, certain types of athletic shirts might be very loose-fitting, and as a result not be given a size description, but their color is important because it is tied to a particular team. Figure 8.6, Faceted Classification. shows how these two resource types can be classified with the faceted Shirt classification. Resource 1 describes a party shirt in medium; resource 2 describes an athletic shirt in blue without information about size.

A faceted classification scheme like that shown in Figure 8.6, Faceted Classification. eliminates the requirement for predetermining a combination and ordering of facets like those in Figure 8.4, Enumerative Classification with Style Facet

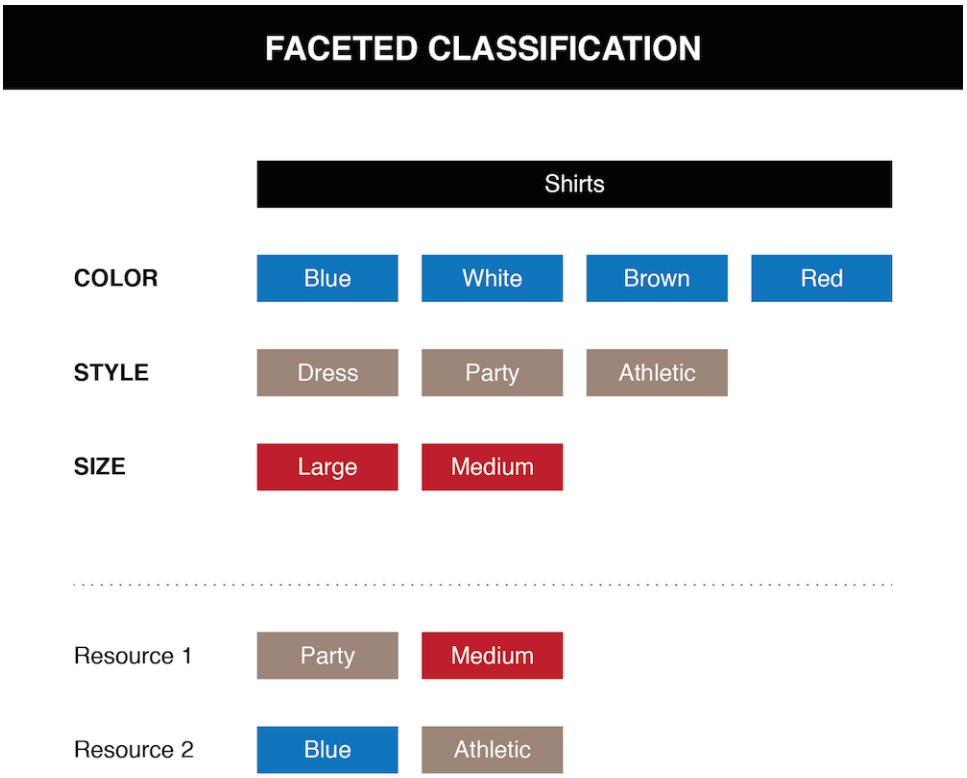
Figure 8.5. Enumerative Classification with Color Facet Followed by Style Facet.



An alternative ordering of the same shirt facets changes the classification hierarchy. If the first facet considered is color, style is next, and finally size, this ordering could result in two piles of white shirts, one for dress shirts and one for athletic shirts, in which each pile contains shirts of large and medium sizes.

Followed by Color Facet. and Figure 8.5, Enumerative Classification with Color Facet Followed by Style Facet. Instead, imagine a shirt store where you decide when you begin shopping which facets are important to you ("show me all the medium party shirts," "show me the blue athletic shirts") instead of having to adhere to whatever predetermined (pre-combined) enumerative classification the store invented. In a digital organizing system, faceted classification enables highly flexible access because prioritizing different facets can dynamically reorganize how the collection is presented.

Figure 8.6. Faceted Classification.



In a pure faceted classification, not every facet needs to apply to every resource, and there is no requirement for a predetermined order in which the facets are considered.

8.4.1 Foundations for Faceted Classification

In library and information science texts it is common to credit the idea of faceted classification to S.R. Ranganathan, a Hindu mathematician working as a librarian. Ranganathan had an almost mystical motivation to classify everything in the universe with a single classification system and notation, considering it his dharma (the closest translation in English would be "fundamental duty" or "destiny"). Facing the limitations of Dewey's system, where an item's essence had to first be identified and then the item assigned to a category based on that essence, Ranganathan believed that all bibliographic resources could be organized around a more abstract variety of aspects.

In 1933 Ranganathan proposed that a set of five facets applied to all knowledge:

***P**ersonality*
: The type of thing.

***M**atter*
: The constituent material of the thing.

***E**nergy*
: The action or activity of the thing.

***S**pace*
: Where the thing occurs.

***T**ime*
: When the thing occurs.

This classification system is known as colon classification (or PMEST) because the notation used for resource identifiers uses a colon to separate the values on each facet. These values come from tables of categories and subcategories, making the call number very compact. Colon classification is most commonly used in libraries in India.

Ranganathan deserves credit for implementing the first faceted classification system, but people other than librarians generally credit the idea to Nicolas de Condorcet, a French mathematician and philosopher. About 140 years before Ranganathan, Condorcet was concerned that "systems of classification that imposed a given interpretation upon Nature... represented an insufferable obstacle to... scientific advance." Condorcet thus proposed a flexible classification scheme for "arranging a large number of subjects in a system so that we may straightway grasp their relations, quickly perceive their combinations, and readily form new combinations."[502][Bus]

Faceted classification is most commonly used in narrow domains, each with its own specific facets. This makes intuitive sense because even if resources can be distinguished with a general classification, doing so requires lengthy notations, and it is much harder to add to a general classification than to a classification created specifically for a single subject area. We could probably describe shirts using the PMEST facets, but style, color, and size seem more natural.

8.4.2 Faceted Classification in Description

Elaine Svenonius defines facets as "groupings of terms obtained by the first division of a subject discipline into homogeneous or semantically cohesive categories." The relationships between these facets results in a controlled vocabulary (§4.1.2) governing the resources we are organizing. From this controlled vocabulary we can generate many descriptions that are complex but formally structured, enabling us to describe things for which terms do not yet exist.

Getty's Art & Architecture Thesaurus (AAT) is a robust and widely used *controlled vocabulary* consisting of generic terms to describe artifacts, objects, places and concepts in the domains of "art, architecture, and material culture."

AAT is a thesaurus with a faceted hierarchical structure. The AAT's facets are "conceptually organized in a scheme that proceeds from abstract concepts to concrete, physical artifacts:"

Associated Concepts
: Concepts, philosophical and critical theory, and phenomena, such as "love" and "nihilism."

Physical Attributes
: Material characteristics that can be measured and perceived, like "height" and "flexibility."

Styles and Periods
: Artistic and architectural eras and stylistic groupings, such as "Renaissance" and "Dada."

Agents
: Basically, people and the various groups and organizations with which they identify, whether based on physical, mental, socio-economic, or political characteristics—e.g., "stonemasons" or "socialists."

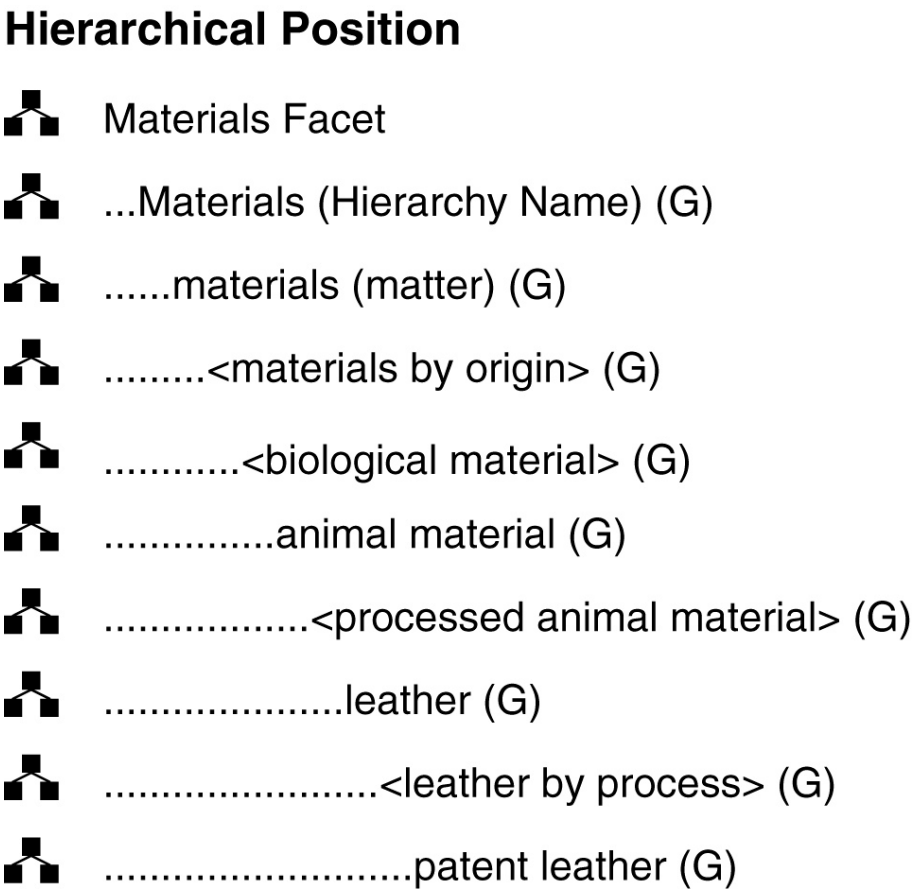
Activities
: Actions, processes, and occurrences, such as "body painting" and "drawing." These are different from the "Objects" facet, which may also contain "body painting," in terms of the actual work itself, not the creation process.

Materials
: Concerned with the actual substance of which a work is made, like "metal" or "bleach." "Materials" differ from "Physical Attributes" in that the latter is more abstract than the former.

Objects
: The largest facet, *objects* contains the actual works, like "sandcastles" and "screen prints."

Within each facet is a strict hierarchical structure drilling down from broad term to very specific instance.

Figure 8.7. "Patent Leather" in the Art & Architecture Thesaurus.



The Art and Architecture Thesaurus has a faceted hierarchical structure.

Figure 8.7, "Patent Leather" in the Art & Architecture Thesaurus. shows how a particular instance may be described on a number of dimensions for the purpose of organizing the item and retrieving information about it. And by using a standard *controlled vocabulary*, catalogers and indexers make it easier for users to understand and adapt to the way things are organized for the purpose of finding them.

8.4.3 A Classification for Facets

There are four major types of facets.

Enumerative facets
: Have mutually exclusive possible values. In our online shirt store, "Style" is an enumerative facet whose values are "dress," "work," "party," and "athletic."

Boolean facets
: Take on one of two values, yes (true) or no (false) along some dimension or property. On a sportswear website, "Waterproof" would be a Boolean facet because an item of clothing is either waterproof or it is not.

Hierarchical facets
: Organize resources by logical inclusion (§6.3.1.1). At Williams-Sonoma's website, the top-level facet includes "Cookware," "Cooks' Tools," and "Cutlery." At wine.com the "Region" facet has values for "US," "Old World," and "New World," each of which is further divided geographically. [508][CogSci] Also see *taxonomic facets*.

Spectrum facets
: Assume a range of numerical values with a defined minimum and maximum. Price and date are common spectrum facets. The ranges are often modeled as mutually exclusive regions (potential price facet values might include "$0—$49," "$50—$99," and "$100—$149").

8.4.4 Designing a Faceted Classification System

It is important to be systematic and principled when designing a faceted classification. In some respects the process and design concerns overlap with those for describing resources, and much of the advice in §5.3 The Process of Describing Resources (page 227) is relevant here.

8.4.4.1 Design Process for Faceted Classification

We advocate a five step process for designing a faceted classification system.

1. Define the purposes of the classification (§5.3.2 Determining the Purposes (page 234), §8.2.1 Classification Is Purposeful (page 400)) and specify the collection of concepts or resources to be classified.
2. For each facet, determine its logical type (§8.4.3 A Classification for Facets (page 424)) and possible values. Specify the order of the values for each facet so that they make sense to users; useful orderings are alphabetical, chronological, procedural, size, most popular to least popular, simple to complex, and geographical or topological.

3. Analyze and describe a representative sample of resource instances to identify properties or dimensions as candidate facets (See §5.3.3 Identifying Properties (page 241)).
4. Examine the relationships between the facets to create sub-facets if necessary. Determine how the facets will be combined to generate the classifications.
5. Test the classification on new instances, and revise the facets, facet values, and facet grammar as needed.

8.4.4.2 Design Principles and Pragmatics

Here is some more specific advice about selecting and designing facets and facet values:

Orthogonality
: Facets should be independent dimensions, so a resource can have values of all of them while only having one value on each of them. In an online kitchen store, one facet might be “Product” and another might be “Brand.” A particular item might be classified as a “Saucepan” in the “Product” facet and as “Calphalon” in the “Brand” one. Other saucepans might have other brands, and other Calphalon products might not be saucepans, because Product and Brand are orthogonal.

Semantic Balance
: Top-level facets should be the properties that best differentiate the resources in the classification domain. The values should be of equal semantic scope so that resources are distributed among the subcategories. Subfacets of “Cookware” like “Sauciers and Saucepans” and “Roasters and Brasiers” are semantically balanced as they are both named and grouped by cooking activity.[510][Ling]

Coverage
: The values of a facet should be able of classifying all instances within the intended scope.

Scalability
: Facet values must accommodate potential additions to the set of instances. Including an “Other” value is an easy way to ensure that a facet is flexible and hospitable to new instances, but it not desirable if all new instances will be assigned that value.

Objectivity
: Although every classification has an explicit or implicit bias (§8.2.3 Classification Is Biased (page 408)), facets and facet values should be as unambiguous and concrete as possible to enable reliable classification of instances.

Normativity
: To make a faceted classification as useful by as many people as possible, the terms used for facets and facet values should not be idiosyncratic, metaphorical, or require special knowledge to interpret.[511][CogSci]

As we will see in §8.6 Computational Classification (page 427), classification can sometimes be done by computers rather than by people. Computer algorithms can analyze resource properties and descriptions to identify dimensions on which resources differ and the most frequent descriptive terms, which can then be used to design a faceted classification scheme. Resources can then be assigned to the appropriate categories, either without human intervention or in collaboration with a human who trains the algorithm with classified instances.

8.5 Classification by Activity Structure

Institutional classification systems are often strongly hierarchical and taxonomic because their many users come to them for diverse purposes, making a context-free or semantic organization the most appropriate. However, in narrow domains that offer a more limited variety of uses it can be much more effective to classify resources according to the tasks or activities they support. A task or activity-based classification system is called a *taskonomy*, a term invented by anthropologists Janet Dougherty and Charles Keller after their ethnographic study of how blacksmiths organized their tools. Instead of keeping things together according to their semantic relationships in what Donald Norman called "hardware store organization," the blacksmiths arranged tools in locations where they were used— "fire tools," "stump tools," "drill press rack tools," and so on.[512][CogSci]

Personal organizing systems are often taskonomic. Think about the way you cook when you are following a recipe. Do you first retrieve all the ingredients from their storage places, and arrange them in activity-based groups in the preparation area?[513][CogSci]

Looking at the relationship between tasks and tools in this way can help a cook determine the best way to organize tools in a kitchen. Cutting items would necessarily be kept together near a prep area; having to run across the kitchen to another area where a poultry knife is kept with, say, chicken broth would be detrimental to the cook's workflow. It would make far more sense to have all of the items for the task of cutting in a single area.

The intentional arrangement of tools in a working kitchen might look something like Table 8.1:

Table 8.1. A cook's taskonomy

Prep	Oven	Stove
Poultry knife	Oven mitts	Pots and pans
Paring knife	Baking sheets	Wooden spoons
Vegetable knife	Aluminum foil	Wok
Cutting board	Parchment paper	
	Roasting pan	

Stop and Think: Office Taskonomy

Think about your personal office space. It may be an interesting hybrid space—it probably contains documents that could be classified in a hierarchical system, but it is also a work space that could lend itself to "taskonomy" organization. Which does it more closely resemble? How have any conflicts between hierarchy and "taskonomy" been resolved?

8.6 Computational Classification

Because of its importance, ubiquity, and ease of processing by computers, it should not be surprising that a great many computational classification problems involve text. Some of these problems are relatively simple, like identifying the language in which a text is written, which is solved by comparing the probability of one, two, and three character-long contiguous strings in the text against their probabilities in different languages. For example, in English the most likely strings are "the", "and", "to", "of", "a", "in", and so on. But if the most likely strings are "der", "die", "und", and "den" the text is German and if they are "de", "la", "que", "el", and "en" the text is Spanish.

More challenging text classification problems arise when more features are required to describe each instance being classified and where the features are less predictable. The unknown author of a document can sometimes be identified by analyzing other documents known to be written by him to identify a set of features like word frequency, phrase structure, and sentence length that create a "writeprint" analogous to a fingerprint that uniquely identifies him. This kind of analysis was used in 2013 to determine that *Harry Potter* author J. K. Rowling had written a crime fiction novel entitled *The Cuckoo's Calling* under the pseudonym Robert Galbraith.[514][Com]

Another challenging text classification problem is sentiment analysis, determining whether a text has a positive or negative opinion about some topic. Much

academic and commercial research has been conducted to understand the sentiment of Twitter tweets, Facebook posts, email sent to customer support applications, and other similar contexts. Sentiment analysis is hard because messages are often short so there is not much to analyze, and because and because sarcasm, slang, clichés, and cultural norms obscure the content needed to make the classification.

A crucial consideration whenever supervised learning is used to train a classifier is ensuring that the training set is appropriate. If we were training a classifier to detect spam messages using email from the year 2000, the topics of the emails, the words they contain, and perhaps even the language they are written in would be substantially different than messages from this year. Up to date training data is especially important for the classification algorithms used by Twitter, Facebook, YouTube, and similar social sites that classify and recommend content based on popularity trends.

When the relevant training data is constantly changing and there is a great deal of it, there is a risk that by the time a model can learn to classify correctly it is already out of date. This challenge has led to the development of streaming algorithms that operate on data as it comes in, using it as a live data source rather than as a static training set. Streaming algorithms are essential for tackling datasets that are too large to store or for models that must operate under intense time pressure. Streaming approaches complement rather than replace those that work with historical datasets because they make different tradeoffs between accuracy and speed. The streaming system might provide real-time alerting and recommendations, while historical analyses are made on the batch-oriented system that works with the entire data collection.[515][Com]

How a computational classifier "learns" depends on the specific machine learning algorithm. Decision trees, Naive Bayes, support vector machines, and neural net approaches were briefly described in §7.5 Implementing Categories (page 360).

Stop and Think: Sentiment Analysis

Sometimes, a text message might seem complimentary, but really is not. Is the customer happy if he tweets "Nice job, United. You only lost one of my bags this time." Think of some other short messages where sarcasm or slang makes sentiment analysis difficult. How would you write a product or service review that is unambiguously positive, negative, or neutral? How would you write a review whose sentiment is difficult to determine?

8.7 Key Points in Chapter Eight

- Classification is the systematic assignment of resources to a system of intentional categories, often institutional ones.

  (See §8.1 Introduction (page 391))

- A classification system is foremost a specification for the logical arrangement of resources because there are usually many possible and often arbitrary mappings of logical locations to physical ones.

  (See §8.1.3 Classification vs. Physical Arrangement (page 395))

- A classification creates structure in the organizing system that increases the variety and capability of the interactions it can support.

  (See §8.2.1.2 Classifications Support Interactions (page 401))

- Classifications are always biased by the purposes, experiences, professions, politics, values, and other characteristics and preferences of the people making them.

  (See §8.2.3 Classification Is Biased (page 408))

- Three types of bias in technical systems are pre-existing, technical, and emergent bias.

  (See §8.2.3 Classification Is Biased (page 408))

- Classification schemes in which all possible categories to which resources can be assigned are defined explicitly are called *enumerative.*

  (See §8.1.4 Classification Schemes (page 395))

- When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is *hierarchical* or *taxonomic.*

  (See §8.1.4 Classification Schemes (page 395))

- Classification and standardization are not identical, but they are closely related. Some classifications become standards, and some standards define new classifications.

  (See §8.1.5 Classification and Standardization (page 397))

- A standard is a published specification that is developed and maintained by consensus of all the relevant stakeholders in some domain by following a defined and transparent process.

  (See §8.1.5.3 Specifications vs. Standards (page 398))

- Standard semantics are especially important in industries or markets that have significant network effects where the value of a product depends on the number of interoperable or compatible products.

  (See §8.1.5.2 Institutional Semantics (page 397))

- The principle of *literary warrant* holds that a classification must be based only on the specific resources that are being classified.

  (See §8.2.2.1 Principles Embodied in the Classification Scheme (page 404))

- The *uniqueness principle* means the categories in a classification scheme are mutually exclusive. Thus, when a logical concept is assigned to a particular category, it cannot simultaneously be assigned to another category.

  (See §8.2.2.2 Principles for Assigning Resources to Categories (page 405))

- The general solution to satisfying the uniqueness principle in library classifications when resources do not clearly fit in a single category is to invent and follow a detailed set of often-arbitrary rules.

  (See §8.2.2.2 Principles for Assigning Resources to Categories (page 405))

- Categories sometimes change slowly, but they can also change quickly and radically as a result of technological, process, or geopolitical innovation or events.

  (See §8.2.2.3 Principles for Maintaining the Classification over Time (page 406))

- *Flexibility*, *extensibility*, and *hospitality* are synonyms for the degree to which the classification can accommodate new resources.

  (See §8.2.2.3 Principles for Maintaining the Classification over Time (page 406))

- Bibliographic classification is distinctive because of a legacy of physical arrangement and its scale and complexity.

  (See §8.3 Bibliographic Classification (page 412))

- *Faceted* classification systems enumerate all the categories needed to assign resources appropriately, but instead of combining them in advance in a fixed hierarchy, they are applied only if they are needed to sort resources with a particular combination of properties.

  (See §8.4 Faceted Classification (page 416))

- Facets should be independent dimensions, so a resource can have values of all of them while only having one value on each of them.

  (See §8.4.4.2 Design Principles and Pragmatics (page 425))

- Top-level facets should be the properties that best differentiate the resources in the classification domain. The values should be of equal semantic

scope so that resources are distributed among the subcategories. Subfacets of “Cookware” like “Sauciers and Saucepans” and “Roasters and Brasiers” are semantically balanced as they are both named and grouped by cooking activity.

(See §8.4.4.2 Design Principles and Pragmatics (page 425))

- Facet values must accommodate potential additions to the set of instances. Including an “Other” value is an easy way to ensure that a facet is flexible and hospitable to new instances, but it not desirable if all new instances will be assigned that value.

  (See §8.4.4.2 Design Principles and Pragmatics (page 425))

- Most tagging seems insufficiently principled to be considered classification, except when tags are treated as category labels or when decisions that make tagging more systematic turn a set of tags into a *tagsonomy*.

  (See §8.1.2 Classification vs. Tagging (page 393))

- A task or activity-based classification system is called a *taskonomy*.

  (See §8.5 Classification by Activity Structure (page 426))

- *Supervised* learning techniques start with a designed classification scheme and then train computers to assign new resources to the categories.

  (See §8.6 Computational Classification (page 427))

Endnotes for Chapter 8

[454][Web] (Hammond et al. 2004) note that the “unstructured (or better, free structured) approach to classification with users assigning their own labels is variously referred to as a folksonomy, folk classification, ethnoclassification, distributed classification, or social classification.”

[455][Web] Thomas Vander Wal invented the term “folksonomy” in 2004, and the term quickly gained traction. His personal account of the creation and dispersion of the term is (Vander Wal 2007).

[456][IA] See (Halvey and Keane 2007), (Sinclair and Cardrew-Hall 2007)) for analyses of the usability of different presentations, and (Kaser and Lemire 2007) for algorithms for drawing tag clouds.

[457][Bus] See (Millen, Feinberg, and Kerr 2006), (John and Seligmann 2006).

[458][DS] The statistical techniques used in topic models are intimidating; to vastly oversimplify, topic models start with a document x term matrix and extract topics by reducing the dimensionality through linear algebra techniques. (Blei 2012) is a relatively easy introduction.

[461][Bus] The most "standard" of all standards organization is the International Organization for Standardization (ISO), whose members are themselves national standards organizations, which as a result gives the nearly 20,000 ISO standards the broadest and most global coverage. See http://ISO.org. In addition, there are scores of other national and industry-specific standards bodies whose work is potentially relevant to organizing systems of the sorts discussed in this book. We encounter these kinds of standards every day in codes for countries, currencies, and airports, in file formats, in product barcodes, and in many other contexts.

[463][Bus] (OASIS 2006). All the finished work of OASIS is freely available at https://www.oasis-open.org; the UBL committee is at https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ubl.

[464][Bus] A small number of people can often informally agree on an organizing system that meets the needs of each participant. But each new person often brings new requirements and it is not feasible to resolve every disagreement between every pair of participants. Instead, for a large-scale organizing system, decisions are usually made by entities that have the authority to coordinate actions and prevent conflicts by imposing a single solution on all the participants. (Rosenthal, Seligman, and Renner 2004) call this the "person-concept" tradeoff, which we can paraphrase as "a few people can agree on a lot, but a lot of people can only agree on a little."

This authority can come from many different sources, but they can be roughly categorized as "authority from power" and "authority from consensus." Often the economic dominance of a firm allows it to control how business gets done in its industry. One key part of that is establishing specifications for data formats and classification schemes in organizing systems, which usually means requiring other firms to use the ones developed by the dominant firm for its own use. This ensures the continued efficiency of their own business processes while making it harder for other firms to challenge their market power.

In contrast, consensus is the authority mechanism embodied in the workings of the open source community, where the freedom to view and change data formats and code that uses them encourages cooperation and adoption. Consensus also underlies the authority of voluntary standards activities, where firms work together under the auspices of a standards body and agree to follow its procedures for creating, ratifying, and implementing standards.

[465][Bus] International and national standards bodies derive their authority from the authority of the governments that created them. But standards organizations arguably derive most of their authority from the collective power of their members, because many influential standards organizations like OASIS, W3C, OMG, and IETF are not chartered or sponsored by governments. In addition, firms often create *ad hoc* "quasi-standards" organizations or "communities of in-

terest" to facilitate relatively short-term cooperative standards-making activities that in the former case would otherwise be prohibited by anti-trust considerations. Finally, at the extreme "lightweight" end of the standards-making continuum, the codification of simple and commonly used information models as "microformats" depends on authority that emerges from the collaboration of individuals rather than firms.

[466][Bus] Often a standard evolves from an existing specification submitted to a standards organization by the firm that created it. In other cases, the specifications used by a dominant firm becomes a *de facto* standard by other firms in its industry, and it is never submitted to a formal standards-making process.

[467][Bus] See (Shapiro and Varian, 1998).

[469][Law] Governments have inherently long time horizons for their actions, they need to serve all citizens fairly and without discrimination, and they (should seek to) minimize cost to taxpayers. Each of these principles is an independent argument for standards and taken together they make a very strong one. Indeed, one the founding goals in the US Constitution is to protect the public interest, and this is enabled in Article I, Section 8 by granting Congress the power to set standards "of Weights and Measures" to facilitate commerce. Setting standards is a key role of the National Institute of Standards and Technology (NIST), part of the Department of Commerce, and other departments have similar standards-setting responsibilities and agencies, like the *Food and Drug Administration (FDA)* in the Department of Health and Social Services. In addition, independent government agencies like the *Federal Communications Commission (FCC)* and *Federal Trade Commission (FTC)* set numerous standards that are relevant to information organizing systems. And of course, the Library of Congress (LOC) maintains procedures and standards needed "to sustain and preserve a universal collection of knowledge... for future generations" (LOC.gov/about).

[470][Bus] The *Sarbanes-Oxley Act* is *US Public Law 107-204,* http://www.sec.gov/about/laws/soa2002.pdf.

The definitive source for the *Health Insurance Portability and Accountability Act (HIPAA)* is the US Department of Health & Human Services, http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.

The definitive source for the *Family Educational Rights and Privacy Act (FERPA)* is the US Department of Education, http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

Complying with government regulations like these can be expensive and difficult, and many companies, especially smaller ones, complain about the cost. On the other hand, the argument can be made that investing in a rigorous system

for organizing information can provide competitive advantages, turning the compliance burden into a competitive weapon (Taylor 2006).

[471][IA] The application of classification and organizing principles more generally to the design of user interfaces to facilitate information access, navigation, and use is often called "Information Architecture." See (Morville and Rosenfeld 2006).

[472][Web] (Grappone and Couzin 2011) is a search engine optimization "cookbook" for do-it-yourselfers. See (Malaga 2008) for a critique of typical SEO practices.

[473][Bus] See (Gladwell 1996), (Schwartz 2005), (Underhill 2008).

[474][Bus] The RosettaNet standards are used by thousands of firms as specifications and implementations of business-to-business processes in several industries, especially component manufacturing and electronics. The specifications are defined using a three-level hierarchy of process clusters, segments, and partner interface processes (PIPs) to enable firms to find a level of process abstraction that works best for them. See http://RosettaNet.org.

[477][CogSci] Very detailed classification of knives are at http://www2.knifecenter.com/knifecenter/kitchen/ and http://kitchenknives.com/.

[481][Com] (Rahm and Bernstein 2006) provide a crisp introduction to the challenges and approaches for changing deployed schemas in databases, conceptual models, ontologies, XML schemas, and software application interfaces. They operate an online bibliography on schema evolution that contains several hundred sources. See http://se-pubs.dbs.uni-leipzig.de/.

[482][Law] See *How the Government Measures Unemployment,* http://www.bls.gov/cps/cps_htgm.htm from the Department of Labor's Bureau of Labor Statistics, and a critical commentary about the measurement scheme titled *Making 9 Million Jobless Vanish: How the Government Manipulates Unemployment Statistics* at http://danielamerman.com/articles/2012/WorkC.html.

[483][IA] (Monmonier 1996) is a highly-readable treatment of intentional and inadvertent bias in mapmaking. A web search for "lying with maps" yields a large number of examples. See also *When Maps Lie*" by Wiseman

[484][Com] (Friedman and Nissenbaum 1996)

[487][Bus] (Feinberg 2012).

[489][Bus] Supermarkets typically carry anywhere from 15,000 to 60,000 SKUs (depending on the size of the store), and may offer a service deli, a service bakery, and/or a pharmacy. 300 standard product categories (http://www.fmi.org/research-resources/supermarket-facts).

[493][CogSci] That was not a typo. The “War of 1812” lasted well into 1815. The persistence of an inaccurate name for this war reflects its unique characteristics. Wars (in the English language) are generally named for the location of the fighting or the enemy being fought (the Mexican-American War, the Korean War, the Vietnam War, the Iraq War), or for a particular ideal or ambition (the Revolutionary War, the Civil War). The War of 1812 does not satisfy any of these naming conventions; the war was fought across a huge range of geography from eastern Canada to Louisiana, between a diverse range of groups from Canadians and Native American tribes, with national armies getting involved very late in the war. While nominally fought over freedom the seas, the war quickly morphed into one about territorial ambition in North America. Of course, if the world were a place where people could agree on naming standards for wars, it is likely we would no longer have wars. See http://en.wikipedia.org/wiki/List_of_wars_involving_the_United_States.

[496][CogSci] Cognitive Science has an even harder time finding its proper place in the LCC because it emerged as the intersection of psychology, linguistics, computer science, and other disciplines. Cognitive science books can be found scattered throughout the LCC, with concentrations in BF, P, and QA.

[497][Bus] The Book Industry Study Group (BISG) first and foremost is focused on resource description and classification as means to business ends; this purpose contrasts with goals of DDC or LOC. BISG classifications are used for barcodes and shipping labels to support supply chain and inventory management, marketing, and promotion activities. See http://www.bisg.org/.

[498][Bus] See (Pope and Holley 2011), (Samuelson 2010).

[502][Bus] (Baker 1962). The first quote is on page 104; the second one is on page 100. This article contains Condorcet’s 1805 essay in French, but fortunately for us Baker’s analysis is in English, This motivation of Condorcet’s classification scheme sounds like the description of a data warehouse or business intelligence system in which transactional data can be “sliced and diced” into new combinations to answer questions in support of strategic decision-making. See (Watson and Wixon 2007).

[508][CogSci] You might have thought that the US was in the new world, but according to wine.com, the new world of wine includes Australia, New Zealand, Argentina, Chile, and South Africa. The geography under the US facet is equally distorted by the uneven distribution of quality wine making regions, so the values of that facet are California. Oregon, Washington, and Other US.

[510][Ling] Should remind you of issues of lexical gap in §6.4 The Lexical Perspective (page 288).

[511][CogSci] Semantic balance is a bit hard to define, but you can often tell when facet values are not balanced. A cookware facet whose values include sauce-

pans, frying pans, stock pots, and pizza pans will not evenly distribute resources across the facets.

[512][CogSci] See (Dougherty and Keller 1985) for the ethnography of blacksmithing, and also (Norman 2006), who extends the taskonomy idea to the design of user interfaces for cell phones and other computing devices. You probably have not worked as blacksmith, but you have certainly used taskonomic classification. For example, a student writing a term paper or doing a course project checks out books from the library's taxonomic classification system (or prints them out from the web) and then organizes them in piles on a desk or on the floor according to the plan for the paper or project. Some of the original classification might persist, but the emphasis clearly shifts toward getting work done. When the task is completed the books go back to the library and are put back into the context-free taxonomy.

[513][CogSci] See (Kirsh 1995) for theoretical motivation and a classification scheme for the "intelligent use of space," and (de Leon 2003) for an example of cooking ethnography.

[514][Com] (SeeLi, Zheng, and Chen 2006), (Juola 2014), (Rowling 1997-2007), (Rowling as "Galbraith" 2013),

[515][Com] (Ellis 2014). A compelling demonstration of the need to sample big data streams to ensure against bias is (Morstatter et al 2013).

Chapter 9

The Forms of Resource Descriptions

Ryan Shaw
Murray Maloney

9.1 Introduction

Throughout this book, we have emphasized the importance of separately considering fundamental organizing principles, application-specific concepts, and details of implementation. The three-tier architecture we introduced in §1.6 is one way to conceptualize this separation. In §6.7, we contrasted the implementation-focused perspective for analyzing relationships with other perspectives that focus on the meaning and abstract structure of relationships. In this chapter, we present this contrast between conceptualization and implementation in terms of separating the *content* and *form* of resource descriptions.

In the previous chapters, we have considered principles and concepts of organizing in many different contexts, ranging from personal organizing systems to cultural and institutional ones. We have noted that some organizing systems have limited scope and expected lifetime, such as a task-oriented personal organizing system like a shopping list. Other organizing systems support broad uses that rely on standard categories developed through rigorous processes, like a product catalog.

By this point you should have a good sense of the various conceptual issues you need to consider when deciding how to describe a resource in order to meet the

goals of your organizing system. Considering those issues will give you some sense of what the content of your descriptions should be. In order to focus on the conceptual issues, we have deferred discussion of specific implementation issues. Implementation involves choosing the specific form of your descriptions, and that is the topic of this chapter.

We can approach the problem of how to form resource descriptions from two perspectives: structuring and writing. From one perspective, resource descriptions are things that are *used* by both people and computational agents. From this perspective, choosing the form of resource descriptions is a kind of design. This is easy to see for certain kinds of resource descriptions, notably signs and maps found in physical environments like airport terminals, public libraries, and malls. In these spaces, resource descriptions are quite literally designed to help people orient themselves and find their way. But any kind of resource description, not just those embedded in the built environment, can be viewed as a designed object. Designing an object involves making decisions about how it should be structured so that it can best be used for its intended purpose. From a design perspective, choosing the form of a resource description means making decisions about its *structure*.

In §6.5, we took a structural perspective on resources and the relationships among them. In this chapter, we will take a structural perspective on resource *descriptions*. The difference is subtle but important. A structural perspective on resource *relationships* focuses on how people or computational processes associate, arrange, and connect those resources. A structural perspective on resource *descriptions* focuses on how those associations, arrangements, and connections are explicitly represented or implemented in the descriptions we create. Mismatches between the structure imposed on the resources being organized and the structure of the descriptions used to implement that organization could result in an organizing system that is complex, inefficient, and difficult to maintain, as you will see in our first example (Example 9.1, Description structured as a dictionary).

The structures of resource descriptions enable or inhibit particular ways of interacting with those descriptions, just as the descriptions themselves enable or inhibit particular ways of interacting with the described resources. (See §3.4 Designing Resource-based Interactions (page 122), and Chapter 10, *Interactions with Resources*) Keep in mind that resource descriptions are themselves information resources, so much of what we will say in this chapter is applicable to the structures and forms of information resources in general. Put another way, the structure and form of information resources informs the design of resource descriptions.

From another perspective, creating resource descriptions is a kind of writing. I may describe something to you orally, but such a description might not be very

useful to an organizing system unless it were transcribed. Organizing systems need persistent descriptions, and that means they need to be written. In that sense, choosing the form of a resource description means making decisions about *notation* and *syntax*.

Modern Western culture tends to make a sharp distinction between designing and writing, but there are areas where this distinction breaks down, and the creation of resource descriptions in organizing systems is one of them. In the following sections, we will use designing and writing as two lenses for looking at the problem of how to choose the form of resource descriptions. Specifically, we will examine the spectrum of options we have for structuring descriptions, and the kinds of syntaxes we have for writing those descriptions.

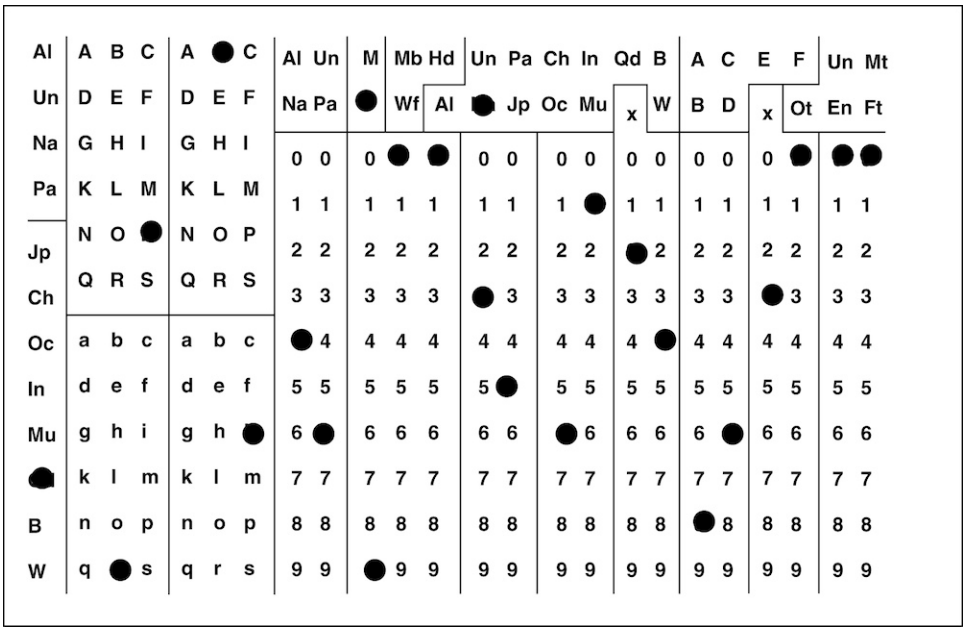
9.2 Structuring Descriptions

Choosing how to structure resource descriptions is a matter of making principled and purposeful design decisions in order to solve specific problems, serve specific purposes, or bring about some desirable property in the descriptions. Most of these decisions are specific to a *domain*: the particular context of application for the organizing system being designed and the kinds of interactions with resources it will enable. Making these kinds of context-specific decisions results in a model of that domain. (See §5.3.1.2 Abstraction in Resource Description (page 232).)

Over time, many people have built similar kinds of descriptions. They have had similar purposes, desired similar properties, and faced similar problems. Unsurprisingly, they have converged on some of the same decisions. When common sets of design decisions can be identified that are not specific to any one domain, they often become systematized in textbooks and in design practices, and may eventually be designed into standard formats and architectures for creating organizing systems. These formally recognized sets of design decisions are known as *abstract models* or *metamodels*. *Metamodels* describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain. While any designer of an organizing system will usually create a model of her specific domain, she usually will not create an entirely new metamodel but will instead make choices from among the metamodels that have been formally recognized and incorporated into existing standards. The resulting model is sometimes called a “domain-specific language.” Reusing standard metamodels can bring great economical advantages, as developers can reuse tools designed for and knowledge about these metamodels, rather than having to start from scratch.

In the following sections, we examine some common kinds of structures used as the basis for metamodels. But first, we consider a concrete example of how the structure of resource descriptions supports or inhibits particular uses. As we

Figure 9.1. A Batten Card.



An example of a punch card used by Batten to describe a particular patent in a patent collection. Each card represented an individual description term, and each punch position on a card represented a particular patent.

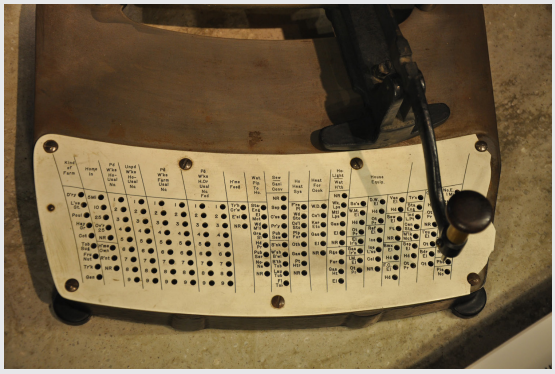
explained in Chapter 1, the concept of a resource de-emphasizes the differences between physical and digital things in favor of focusing on how things, in general, are used to support goal-oriented activity. Different kinds of books can be treated as information resources regardless of the particular mix of tangible and intangible properties they may have. Since resource descriptions are also information resources, we can similarly consider how their structures support particular uses, independent of whether they are physical, digital, or a mix of both.

During World War II, a British chemist named W. E. Batten developed a system for organizing patents.[516][Com] The system consisted of a language for describing the product, process, use, and apparatus of a patent, and a way of using punched cards to record these descriptions. Batten used cards printed with matrices of 800 positions (see Figure 9.1). Each card represented a specific value from the vocabulary of the description language, and each position corresponded to a particular patent. To describe patent #256 as covering *extrusion of polythene to produce cable coverings*, one would first select the cards for the val-

ues *polythene*, *extrusion*, and *cable coverings*, and then punch each card at the 256th position. The description of patent #256 would thus extend over these three cards.

The advantage of this structure is that to find patents covering *extrusion of polythene* (for any purpose), one needs only to select the two cards corresponding to those values, lay one on top of the other, and hold them up to a light. Light will shine through wherever there is a position corresponding to a patent described using those values. Patents meeting a certain description are easily found due to the structure of the cards designed to describe the patents.

Punchcard Machine



Punchcards were an important information input and storage medium for decades, even before the invention of computers. The Hollerith keyboard punch was used to transcribe the information collected in the 1890 US census. The template being used in this photo is for recording information about a farm. The punch cards were tabulated by electromechanical machines. A merger of four tabulating machine companies in 1911 created a company whose current name is IBM.

This keyboard punch machine is in the collection of the Computer History Museum in Mountain View, California.

(Photo by R. Glushko.)

Of course, this system has clear disadvantages as well. Finding the concepts associated with a particular patent is tedious, because every card must be inspected. Adding a new patent is relatively easy as long as there is an index that allows the cards for specific concepts to be located quickly. However, once the cards run out of space for punching holes, the whole set of cards must be duplicated to accommodate more patents: a very expensive operation. Adding new concepts is potentially easy: simply add a new card. But if we want to be able to find existing patents using the new concept, all the existing patents would have to be reexamined to determine whether their positions on the new card should be punched: also an expensive operation.

The structure of Batten's cards supported rapid selection of resources given a partial description. The kinds of structures we will examine in the following sections are not quite so elaborate as Batten's cards. But like the cards, each kind of structure supports more efficient mechanical execution of certain operations, at the cost of less efficient execution of others.

9.2.1 Kinds of Structures

Sets, lists, dictionaries, trees, and graphs are kinds of structures that can be used to form resource descriptions. As we shall see, each of these kinds is actually a family of related structures. These structures are *abstractions*: they describe formal structural properties in a general way, rather than specifying an exact physical or textual form. Abstractions are useful because they help us to see common properties shared by different specific ways of organizing information. By focusing on these common properties, we can more easily reason about the operations that different forms support and the affordances that they provide, without being distracted by less relevant details.

9.2.1.1 Blobs

The simplest kind of structure is no structure at all. Consider the following description of a book: *Sebald's novel uses a walking tour in East Anglia to meditate on links between past and present, East and West*.[517][Ling] This description is an unstructured text expression with no clearly defined internal parts, and we can consider it to be a *blob*. Or, more precisely, it has structure, but that structure is the underlying grammatical structure of the English language, and none of that grammatical structure is explicitly represented in a surface structure when the sentence is expressed. As readers of English we can interpret the sentence as a description of the subject of the book, but to do this mechanically is difficult.[518][Ling] On the other hand, such a written description is relatively easy to create, as the describer can simply use natural language.

A blob need not be a blob of text. It could be a photograph of a resource, or a recording of a spoken description of a resource. Like blobs of text, blobs of pixels or sound have underlying structure that any person with normal vision or hearing can understand easily.[519][CogSci] But we can treat these blobs as unstructured, because none of the underlying structure in the visual or auditory input is explicit, and we are concerned with the ways that the structures of resource descriptions support or inhibit mechanical or computational operations.[520][Com]

9.2.1.2 Sets

The simplest way to structure a description is to give it parts and treat them as a *set*. For example, the description of Sebald's novel might be reformulated as a set of terms: *Sebald*, *novel, East Anglia, walking, history*. Doing this has lost much of the meaning, but something has been gained: we now can easily distinguish *Sebald* and *walking* as separate items in the description. This makes it easier to find, for example, all the descriptions that include the term *walking*. (Note that this is different from simply searching through blob-of-text descriptions for the word *walking*. When treated as a set, the description *Fiji, fire walk-*

ing, memoir does not include the term *walking*, though it does include the term *fire walking*.)

Sets make it easy to find intersections among descriptions. *Sets* are also easy to create. In §8.1.2 we looked at "folksonomies," organizing systems in which non-professional users create resource descriptions. In these systems, descriptions are structured as *sets* of "tags." To find resources, users can specify a *set* of tags to obtain resources having descriptions that intersect at those tags. This is more valuable if the tags come from a *controlled vocabulary*, making intersections more likely. But enforcing vocabulary control adds complexity to the description process, so a balance must be struck between maximizing potential intersections and making description as simple as practical.[522][Com]

A *set* is a type or class of structure. We can refine the definition of different kinds of sets by introducing *constraints*. For example, we might introduce the constraint that a given set has a maximum number of items. Or we might constrain a set to always have the same number of items, giving us a fixed-size set. We can also remove constraints. Sets do not contain duplicate items (think of a tagging system in which it does not make sense to assign the same tag more than once to the same resource). If we remove this *uniqueness* constraint, we have a different structure known as a "bag" or "multiset."

9.2.1.3 Lists

Constraints are what distinguish lists from sets. A *list*, like a *set*, is a collection of items with an additional constraint: their items are ordered. If we were designing a tagging system in which it was important that the order of the tags be maintained, we would want to use lists, not sets. Unlike sets, *lists* may contain duplicate items. In a *list*, two items that are otherwise the same can be distinguished by their position in the ordering, but in a set this is not possible. For example, we might want to organize the tags assigned to a resource, listing the most used tag first, the least frequently used last, and the rest according to their frequency of use.

Again, we can introduce constraints to refine the definition of different kinds of *lists*, such as fixed-length lists. If we constrain a list to contain only items that are themselves lists, and further specify that these contained lists do not themselves contains lists, then we have a *table* (a list of lists of items). A spreadsheet is a list of lists.

9.2.1.4 Dictionaries

One major limitation of *lists* and *sets* is that, although items can be individually addressed, there is no way to distinguish the items except by comparing their values (or, in a list, their positions in the ordering). In a set of terms like *Sebald, novel, East Anglia, walking, history*, for example, one cannot easily tell that *Sebald* refers to the author of the book while *East Anglia* and *walking* refer to what it is about. One way of addressing this problem is to break each item in a set into two parts: a *property* and a *value*. So, for example, our simple set of tags might become *author: Sebald, type: novel, subject: East Anglia, subject: walking, subject: history*. Now we can say that *author*, *type*, and *subject* are the properties, and the original items in the set are the values.

author
: Sebald

type
: novel

subject1
: East Anglia

subject2
: walking

subject3
: history

This kind of structure is called a *dictionary*, a *map* or an *associative array*. A *dictionary* is a set of property-value pairs or entries. It is a set of entries, not a list of entries, because the pairs are not ordered and because each entry must have a unique key.[523][Com] Note that this specialized meaning of *dictionary* is different from the more common meaning of "dictionary" as an alphabetized list of terms accompanied by sentences that define them. The two meanings are related, however. Like a "real" dictionary, a *dictionary* structure allows us to easily find the value (such as a definition) associated with a particular property or *key* (such as a word). But unlike a real dictionary, which orders its keys alphabetically, a *dictionary* structure does not specify an order for its keys.

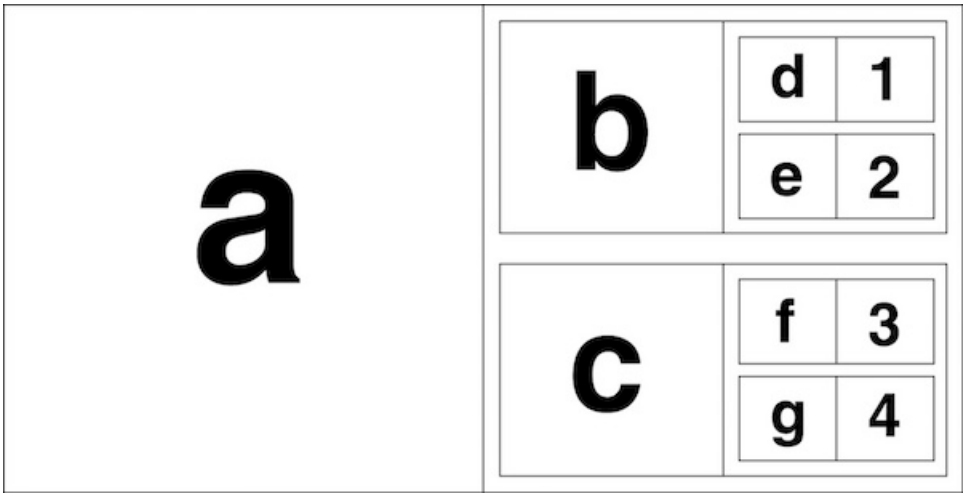
Dictionaries are ubiquitous in resource descriptions. Structured descriptions entered using a form are easily represented as dictionaries, where the form items' labels are the properties and the data entered are the values. Tabular data with a "header row" can be thought of as a set of dictionaries, where column headers are the properties for each dictionary, and each row is a set of corresponding values. Dictionaries are also a basic type of data structure found in nearly all programming languages (referred to as associative arrays).

Again, we can introduce or remove constraints to define specialized types of dictionaries. A sorted dictionary adds an ordering over entries; in other words, it is a list of entries rather than a set. A *multimap* is a dictionary in which multiple entries may have the same key.

9.2.1.5 Trees

In dictionaries as they are commonly understood, properties are terms and values are their corresponding definitions. The terms and values are usually words, phrases, or other expressions that can be ordered alphabetically. But if generalize the notion of a dictionary as abstract sets of property-value pairs, the values can be anything at all. In particular, the values can themselves be dictionaries. When a dictionary structure has values that are themselves dictionaries, we say that the dictionaries are *nested*. Nesting is very useful for resource descriptions that need more structure than what a (non-nested) dictionary can provide.

Figure 9.2. Four Nested Dictionaries.



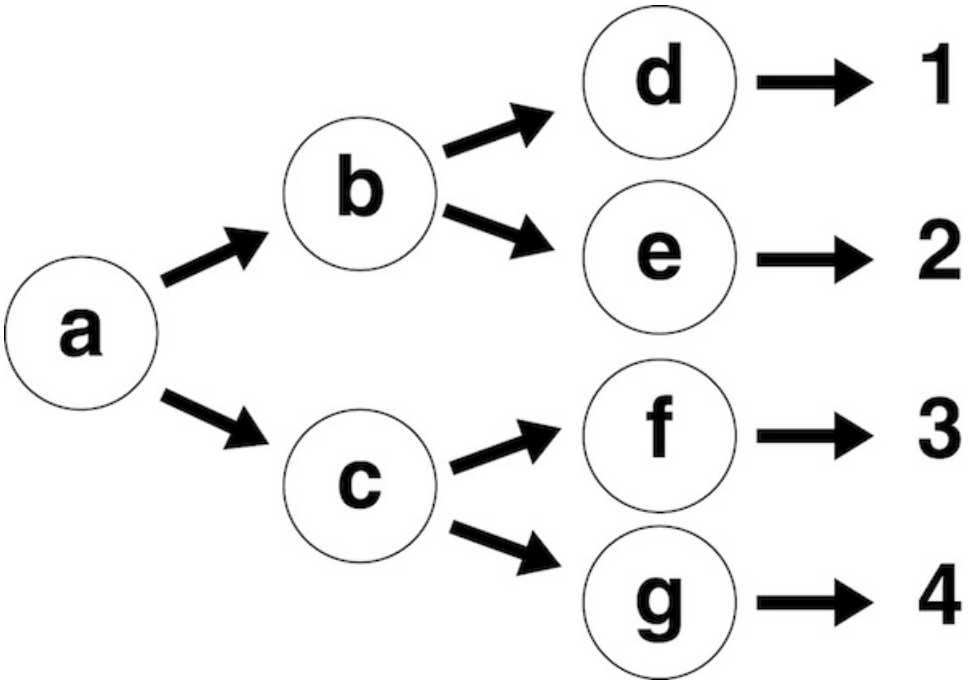
When a dictionary contains other dictionaries, they are said to be nested.

Figure 9.2, Four Nested Dictionaries. presents an example of nested dictionaries. At the top level there is one dictionary with a single entry having the property *a*. The value associated with *a* is a dictionary consisting of two entries, the first having property *b* and the second having property *c*. The values associated with *b* and with *c* are also dictionaries.

If we nest dictionaries like this, and our "top" dictionary (the one that contains all the others) has only one entry, then we have a kind of *tree* structure. Figure 9.3, A Tree of Properties and Values. shows the same properties and values

as Figure 9.2, this time arranged to make the tree structure more visible. *Trees* consist of *nodes* (the letters and numbers in Figure 9.3) joined by *edges* (the arrows). Each node in the tree with a circle around it is a property, and the value of each property consists of the nodes below (to the right of) it in the tree. A node is referred to as the *parent* of the nodes below it, which in turn are referred to as the *children* of that node. The edges show these "parent of" relationships between the nodes. The node with no parent is called the *root* of the tree. Nodes with no children are called *leaf* nodes.

Figure 9.3. A Tree of Properties and Values.



An alternative representation of nested dictionaries is as a tree. The lowest level or leaf nodes of the tree contain property values.

As with the other types of structures we have considered, we can define different kinds of trees by introducing different types of constraints. For example, the predominant metamodel for XML is documents is a kind of tree called the *XML Information Set* or Infoset. [525][Com]

The *XML Information Set* defines a specific kind of tree structure by adding very specific constraints, including ordering of child nodes, to the basic definition of a tree. The addition of an ordering constraint distinguishes XML trees from nested dictionaries, in which child nodes do not have any order (because

dictionary entries do not have an ordering). Ordering is an important constraint for resource descriptions, since without ordering it is impossible to, for example, list multiple authors while guaranteeing that the order of authors will be maintained. Figure 9.3 depicts a kind of tree with a different set of constraints: all non-leaf nodes are properties, and all leafs are values. We could also define a tree in which every node has both a property and a value. Trees exist in a large variety of flavors, but they all share a common topology: the edges between nodes are directed (one node is the parent and the other is the child), and every node except the root has exactly one parent.

Trees provide a way to group statements describing different but related resources. For example, consider the description structured as a dictionary here:

Example 9.1. Description structured as a dictionary

author given names → Winfried Georg
author surname → Sebald
title → Die Ringe des Saturn
pages → 371

The dictionary groups together four property-value pairs describing a particular book. (The arrows are simply a schematic way to indicate property-value relations. Later in the chapter we look at ways to "write" these relations using some specific syntax.)

But really the first two entries are not describing the book; they are describing the book's author. So, it would be better to group those two statements somehow. We can do this by nesting the entries describing the author within the book description, creating a tree structure:

Example 9.2. Nesting an author description within a book description

author →
given names → Winfried Georg
surname → Sebald
title → Die Ringe des Saturn
pages → 371

Using a tree works well in this case because we can treat the book as the primary resource being described, making it the root of our tree, and adding on the author description as a "branch."

We also could have chosen to make the author the primary resource, giving us a tree like the one in Example 9.3.

Example 9.3. Nesting book descriptions within an author description

given names → Winfried Georg
surname → Sebald
books authored →
1. title → Die Ringe des Saturn
pages → 371
2. title → Austerlitz
pages → 416

Note that in this dictionary, the value of the *books authored* property is a *list* of dictionaries. Making the author the primary or root resource allows us to include multiple book descriptions in the tree (but makes it more difficult to describe books having multiple authors). A tree is a good choice for structuring descriptions as long as we can clearly identify a primary resource. In some cases, however, we want to connect descriptions of related resources without having to designate one as primary. In these cases, we need a more flexible data structure.

9.2.1.6 Graphs

Suppose we were describing two books, where the author of one book is the subject of the other, as in Example 9.4, Two related descriptions:

Example 9.4. Two related descriptions

1. author → Mark Richard McCulloch
title → Understanding W. G. Sebald
subject → Winfried Georg Sebald
2. author → Winfried Georg Sebald
title → Die Ringe des Saturn

By looking at these descriptions, we can guess the relationship between the two books, but that relationship is not explicitly represented in the structure: we just have two separate dictionaries and have inferred the relationship by matching property values. It is possible that this inference could be wrong: there might be two people named *Winfried Georg Sebald*. How can we structure these descriptions to explicitly represent the fact that the *Winfried Georg Sebald* that is the subject of the first book is the same *Winfried Georg Sebald* who authored the second?

One possibility would be to make *Winfried Georg Sebald* the root of a tree, similar to the approach taken in Example 9.3, Nesting book descriptions within an author description, adding a *book about* property alongside the *books authored*

one. This solution would work fine if people were our primary resources, and it thus made sense to structure our descriptions around them. But suppose that we had decided that our descriptions should be structured around books, and that we were using a vocabulary that took this perspective (with properties such as *author* and *subject* rather than *books authored* and *books about*). We should not let a particular structure limit the organizational perspective we can take, as Batten's cards did. Instead, we should consciously choose structures to suit our organizational perspective. How can we do this?

If we treat our two book descriptions as trees, we can join the two branches (subject and author) that share a value. When we do this, we no longer have a tree, because we now have a node with more than one parent (Figure 9.4, Descriptions Linked into a Graph.). The structure in Figure 9.4, Descriptions Linked into a Graph. is a *graph*. Like a *tree*, a *graph* consists of a set of nodes connected by edges. These edges may or may not have a direction (§6.6.3 Directionality (page 307)). If they do, the *graph* is referred to as a "directed graph." If a *graph* is directed, it may be possible to start at a node and follow edges in a path that leads back to the starting node. Such a path is called a "cycle." If a directed graph has no cycles, it is referred to as an "acyclic graph."

A tree is just a more constrained kind of *graph*. Trees are *directed* graphs because the "parent of" relationship between nodes is asymmetric: the edges are arrows that point in a certain direction. (See §6.3.2.1 Symmetry (page 284).) Furthermore, trees are *acyclic* graphs, because if you follow the directed edges from one node to another, you can never encounter the same node twice. Finally, trees have the constraint that every node (except the root) must have exactly one parent.[526][Com]

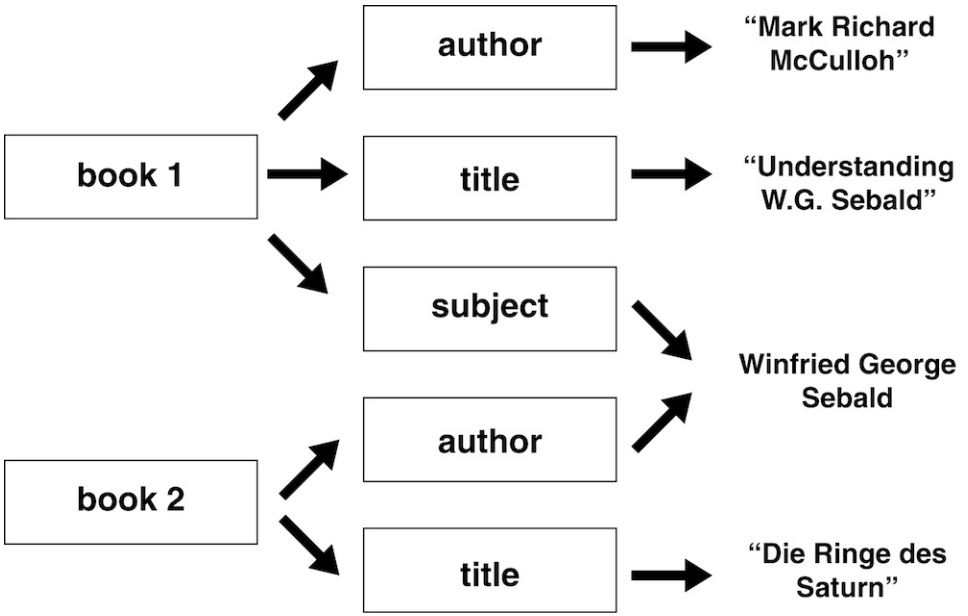
In Figure 9.4, Descriptions Linked into a Graph. we have violated this constraint by joining our two book trees. The graph that results is still directed and acyclic, but because the *Winfried George Sebald* node now has two parents, it is no longer a tree.

Graphs are very general and flexible structures. Many kinds of systems can be conceived of as nodes connected by edges: stations connected by subway lines, people connected by friendships, decisions connected by dependencies, and so on. Relationships can be modeled in different ways using different kinds of *graph*s. For example, if we assume that friendship is symmetric (see §6.3.2.1 Symmetry (page 284)), we

Stop and Think: Social Network Properties

Compare the concept of "friend" in Facebook with that of "follower" in Twitter, in terms of the semantic properties discussed in §6.3.2 Properties of Semantic Relationships (page 284) and the graph properties discussed in this section.

Figure 9.4. Descriptions Linked into a Graph.



Descriptions can be linked to form a graph when the value assigned to two different properties is the same.

would use an undirected *graph* to model the relationship. However, in web-based social networks friendship is often asymmetric (you might "friend" someone who does not reciprocate), so a directed *graph* is more appropriate.

Often it is useful to treat a *graph* as a set of pairs of nodes, where each pair may or may not be directly connected by an edge. Many approaches to characterizing structural relationships among resources (see §6.5.3 Structural Relationships between Resources (page 300)) are based on modeling the related resources as a set of pairs of nodes, and then analyzing patterns of connectedness among them. As we will see, being able to break down a *graph* into pairs is also useful when we structure resource descriptions as *graph*s.

In §9.4.1 we will use XML to model the graph shown in Figure 9.4, Descriptions Linked into a Graph. by using "references" to connect a book to its title, authors and subject. This will allow us to develop sophisticated graphs of knowledge within a single XML document instance. (See also the sidebar, Inclusions and References (page 456))[527][Com]

9.2.2 Comparing Metamodels: JSON, XML and RDF

Now that we are familiar with the various kinds of metamodels used to structure resource descriptions, we can take a closer look at some specific metamodels. A detailed comparison of the affordances of different metamodels is beyond the scope of this chapter. Here we will simply take a brief look at three popular metamodels—JSON, XML, and RDF—in order to see how they further specify and constrain the more general kinds of metamodels introduced above.

9.2.2.1 JSON

JavaScript Object Notation (JSON)

> *JavaScript Object Notation (JSON)* is a textual format for exchanging data that borrows its metamodel from the JavaScript programming language. Specifically, the JSON metamodel consists of two kinds of structures found in JavaScript: lists (called "arrays" in JavaScript) and dictionaries (called "objects" in JavaScript). Lists and dictionaries contain values, which may be strings of text, numbers, Booleans (true or false), or the null (empty) value. Again, these types of values are taken directly from JavaScript. Lists and dictionaries can be values too, meaning lists and dictionaries can be nested within one another to produce more complex structures such as tables and trees.

Lists, dictionaries, and a basic set of value types constitute the JSON metamodel. Because this metamodel is a subset of JavaScript, the JSON metamodel is very easy to work with in JavaScript. Since JavaScript is the only programming language that is available in all web browsers, JSON has become a popular choice for developers who need to work with data and resource descriptions on the web. (See §9.3.2 Writing Systems (page 464) later in this chapter.) Furthermore, many modern programming languages provide data structures and value types equivalent to those provided by JavaScript. So, data represented as JSON is easy to work with in many programming languages, not just JavaScript.

9.2.2.2 XML Information Set

The *XML Information Set* metamodel is derived from data structures used for document markup. (See §5.2.2.2.) These markup structures—*elements* and *attributes*—are well suited for programmatically manipulating the structure of documents and data together.[528][Com]

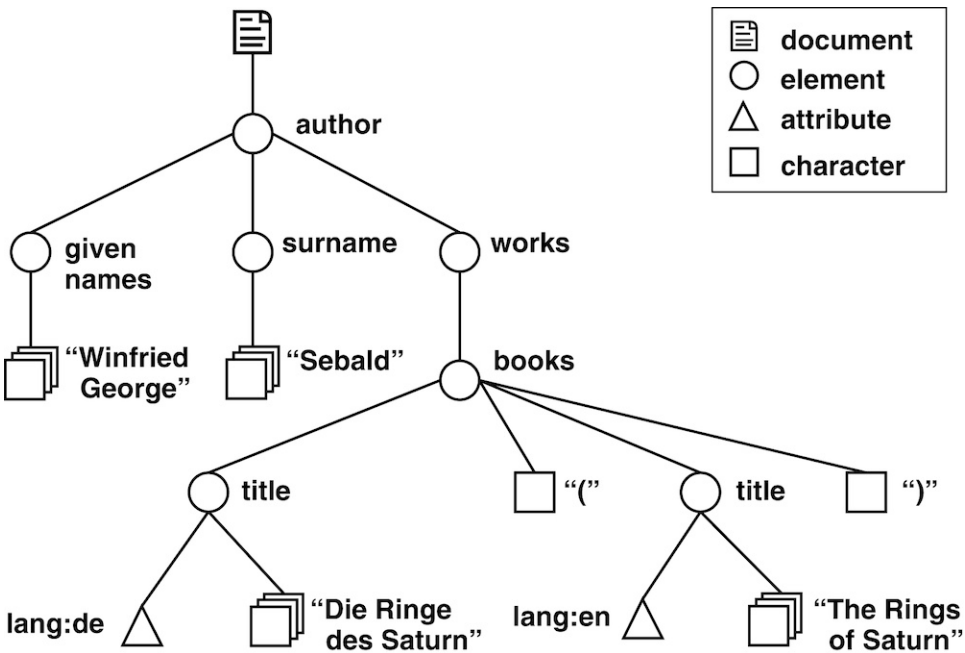
XML Infoset

> The *XML Infoset* is a tree structure, where each node of the tree is defined to be an "information item" of a particular type. Each information item has a set of type-specific properties associated with it. At the root of the tree is a "document item," which has exactly one "element item" as its child. An *element* item has a set of *attribute* items, and a list of child nodes. These child

nodes may include other element items, or they may be character items. (See §9.2.1 Kinds of Structures (page 442) below for more on characters.) *Attribute* items may contain character items, or they may contain typed data, such as name tokens, identifiers and references. Element identifiers and references (ID/IDREF) may be used to connect nodes, transforming a tree into a graph. (See the sidebar, Inclusions and References (page 456))[529][Com]

Figure 9.5, A Description Structure. is a graphical representation of how an XML document might be used to structure part of a description of an author and his works. This example demonstrates how we might use element items to model the domain of the description, by giving them names such as author and title. The character items that are the children of these elements hold the content of the description: author names, book titles, and so on. Attribute items are used to hold auxiliary information about this content, such as its language.

Figure 9.5. A Description Structure.



An XML document can be described as a tree in which elements are nodes that can contain character content directly or attributes that contain character content.

This example also demonstrates how the XML Infoset supports mixed content by allowing element items and character items to be "siblings" of the same parent element. In this case, the Infoset structure allows us to specify that the

book description can be displayed as a line of text consisting of the original title and the translated title in parentheses. The elements and attributes are used to indicate that this line of text consists of two titles written in different languages, not a single title containing parentheses.

If not for mixed content, we could not write narrative text with *hypertext links* embedded in the middle of a sentence. It gives us the ability to identify the subcomponents of a sentence, so that we could distinguish the terms “Sebald,” “walking” and “East Anglia” as an author and two subjects.

Using schemas to define data representation formats is a good practice that facilitates shared understanding and contributes to long-term maintainability in institutional or business contexts. An XML schema represents a contract among the parties subscribing to its definitions, whereas JSON depends on out-of-band communication among programmers. The notion that “the code is the documentation” may be fashionable among programmers, but modelers prefer to design at a higher level of abstraction and then implement.

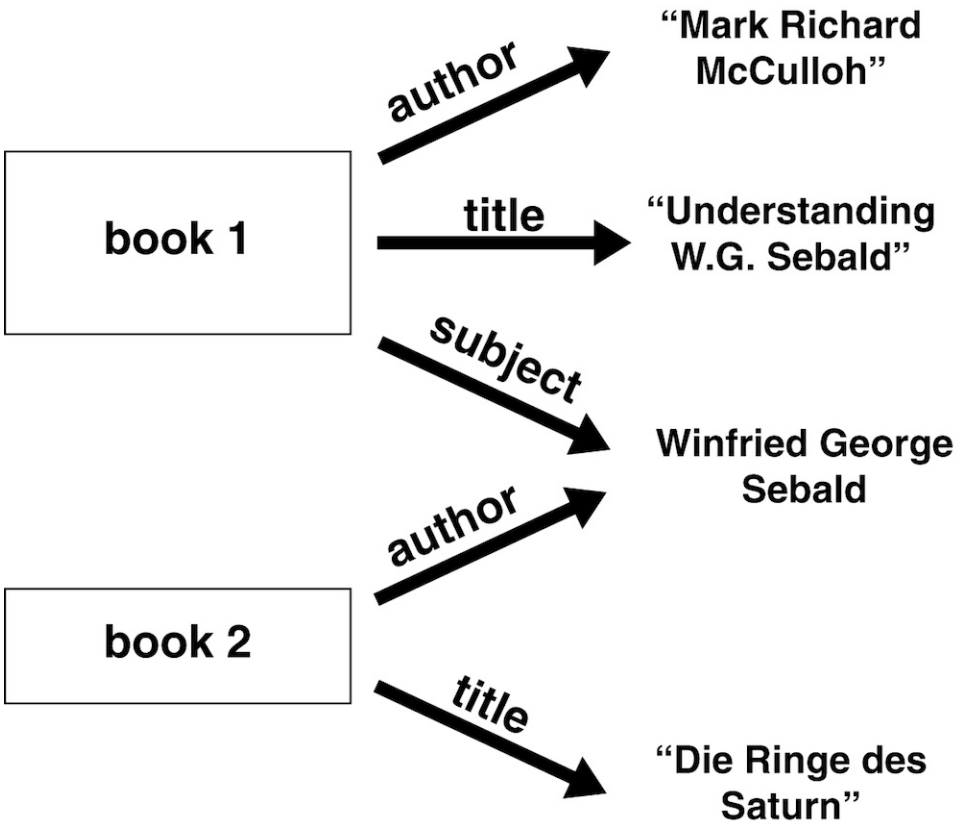
The XML Infoset presents a strong contrast to JSON and does not always map in a straightforward way to the data structures used in popular web scripting languages. Whereas JSON’s structures make it easier for object-oriented programmers to readily exchange data, they lack any formal schema language and cannot easily handle mixed content.

9.2.2.3 RDF

In Figure 9.4, Descriptions Linked into a Graph., we structured our resource description as a graph by treating resources, properties, and values as nodes, with edges reflecting their combination into descriptive statements. However, a more common approach is to treat resources and values as nodes, and properties as the edges that connect them. Figure 9.6, Treating Properties as Edges Rather Than Nodes. shows the same description as Figure 9.4, Descriptions Linked into a Graph., this time with properties treated as edges. This roughly corresponds to the particular kind of graph metamodel defined by *RDF*. (§5.2.2.4 Resource Description Framework (RDF) (page 223))

We have noted that we can treat a graph as a set of pairs of nodes, where each pair may be connected by an edge. Similarly, we can treat each component of the description in Figure 9.6, Treating Properties as Edges Rather Than Nodes. as a pair of nodes (a resource and a value) with an edge (the property) linking them. In the RDF metamodel, a pair of nodes and its edge is called a *triple*, because it consists of three parts (two nodes and one edge). The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple.

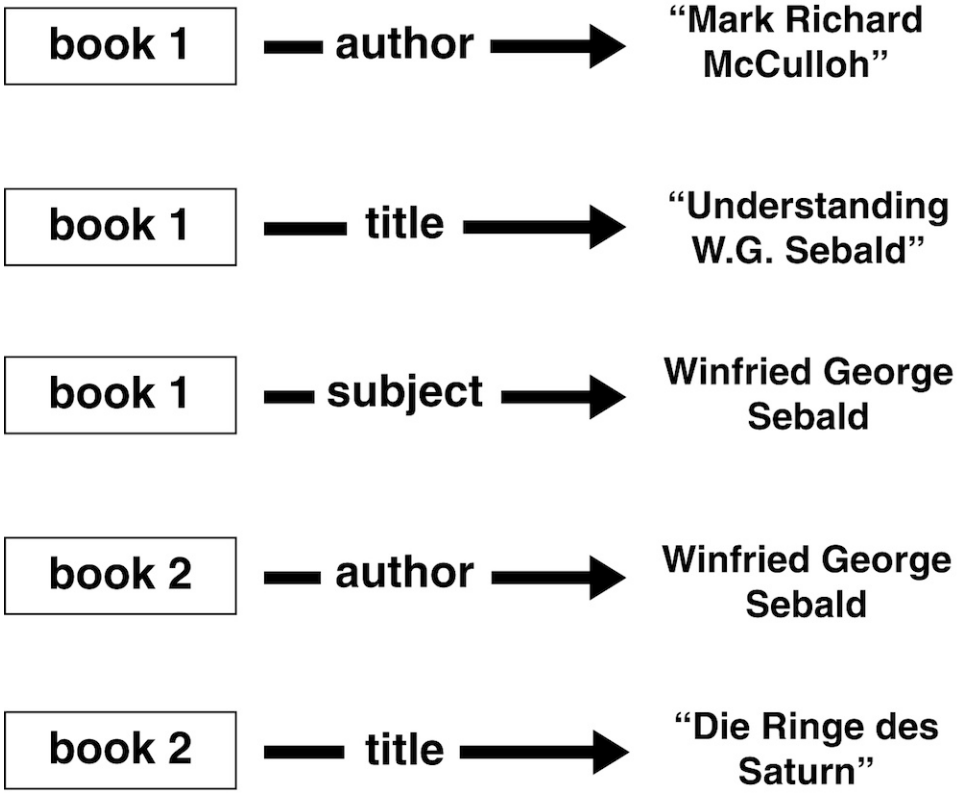
Figure 9.6. Treating Properties as Edges Rather Than Nodes.



We can treat each component of a description as a pair of nodes (a resource and a value) with an edge (the property) linking them. Here, we have two book resources that are related to four values through five properties. The single value node, "Winfried George Sebald" is the subject of one book while being the author of the second book. The books are depicted as boxes, the edges as labeled arrows and the values as text strings.

Figure 9.7, Listing Triples Individually. lists separately all the triples in Figure 9.6 However, there is something missing in Figure 9.7. Figure 9.6 clearly indicates that the *Winfried George Sebald* who is the subject of book 1 is the same *Winfried George Sebald* who is the author of book 2. In Figure 9.7, Listing Triples Individually. this relationship is not clear. How can we tell if the *Winfried George Sebald* of the third triple is the same as the *Winfried George Sebald* of the triple statement? For that matter, how can we tell if the first three triples all involve the same book 1? This is easy to show in a diagram of the entire description

Figure 9.7. Listing Triples Individually.



Lists each of the triples individually. Here, each statement relates one resource to one value through an edge. Thus, we have two distinct “Winfried George Sebald” value nodes. The books are depicted as boxes, the edges as labeled arrows and the values as text strings.

graph, where we can have multiple edges attached to a node. But when we disaggregate that graph into triples, we need some way of uniquely referring to nodes. We need identifiers (§4.4.3 Choosing Good Names and Identifiers (page 194)). When two triples have nodes with the same identifier, we can know that it is the same node. RDF achieves this by associating URIs with nodes. (See §5.2.2.4 Resource Description Framework (RDF) (page 223))

The need to identify nodes when we break down an RDF graph into triples becomes important when we want to “write” RDF graphs—create textual representations of them instead of depicting them—so that they can be exchanged as data. Tree structures do not necessarily have this problem, because it is

Inclusions and References

An XML Infoset is typically the result of processing a well-formed XML document instance.[530][Com] Schemas associated with XML document instances "inform" the corresponding XML Infoset. Thus, the "truth value" of any XML Infoset is dependent upon its related schemas.[531][Com] Traditionally, any documentation that is related to the schema is considered to be part of the schema definition and, at least notionally, informs human understanding and interpretation of corresponding documents.[532][Com]

The XML family offers several mechanisms to create inclusion relationships: by employing element references; by way of entity definition and reference; by using *XML Inclusions (XInclude)* or XLink. These inclusions and references can also inform the XML Infoset, if they are processed.

Any XML node may refer to another node simply by referencing it by its assigned ID. Assuming attributes are declared, the Infoset exposes this information as a *references* property as an ordered list of element information items. That is to say that an element may contain other element nodes by subordination, or by reference.[533][Com]

XInclude "specifies a processing model and syntax for general purpose inclusion. Inclusion is accomplished by merging a number of XML information sets into a single composite infoset." XInclude offers the most versatile mechanism for addressing whole documents, specific information items, ranges of information items, and even parts of information items, which has led to its widespread adoption in document processing.[534][Com]

XLink "allows elements to be inserted into XML documents in order to create and describe links between resources. It uses XML syntax to create structures that can describe links similar to the simple unidirectional hyperlinks of today's HTML, as well as more sophisticated links."[535][Com]

Entities are similar to macros found in many programming languages; a value is assigned to a token, the token is referenced wherever the value is needed, and macro expansion happens when the XML document instance is read into the Infoset.[536][Com] Entities are a handy feature, but since they are expanded on their way in, entities do not survive as information items in the XML Infoset. The ID/IDREF feature is more popular than the use of entities because it carries more information into the XML Infoset.

possible to textually represent a tree structure without having to mention any node more than once. Thus, one price paid for the generality and flexibility of graph structures is the added complexity of recording, representing or writing those structures.

9.2.2.4 Choosing Your Constraints

This tradeoff between flexibility and complexity illustrates a more general point about constraints. In the context of managing and interacting with resource descriptions, constraints are a good thing. As discussed above, a tree is a graph with very specific constraints. These constraints allow you to do things with trees that are not possible with graphs in general, such as representing them textually without repeating yourself, or uniquely identifying nodes by the path from the root of the tree to that node. This can make managing descriptions and the resources they describe easier and more efficient—*if* a tree structure is a good fit to the requirements of the organizing system. For example, an ordered tree structure is a good fit for the hierarchical structure of the content of a book or book-like document, such as an aircraft service manual or an SEC filing. On the other hand, the network of relationships among the people and organizations that collaborated to produce a book might be better represented using a graph structure. XML is most often used to represent hierarchies, but is also capable of representing network structures.

9.2.3 Modeling within Constraints

A metamodel imposes certain constraints on the structure of our resource descriptions. But in organizing systems, we usually need to further specify the content and composition of descriptions of the specific types of resources being organized. For example, when designing a system for organizing books, it is not sufficient to say that a book's description is structured using XML, because the XML metamodel constrains structure and not the content of descriptions. We need also to specify that a book description includes a list of contributors, each entry of which provides a name and indicates the role of that contributor. This kind of specification is a *model* to which our descriptions of books are expected to conform. (See §5.3.1.2 Abstraction in Resource Description (page 232).)

When designing an organizing system we may choose to reuse a standard model. For example, ONIX for Books is a standard model (conforming to the XML metamodel) developed by the publishing industry for describing books.[537][Bus]

If no such standard exists, or existing standards do not suit our needs, we may create a new model for our specific domain. But we will not usually create a new metamodel: instead we will make choices from among the metamodels, such as JSON, XML, or RDF, that have been formally recognized and incorporated into existing standards. Once we have selected a metamodel, we know the constraints we have to work with when modeling the resources and collections in our specific domain.[538][Com]

9.2.3.1 Specifying Vocabularies and Schemas

Creating a model for descriptions of resources in a particular domain involves specifying the common elements of those descriptions, and giving those elements standard names. (See §5.3 The Process of Describing Resources (page 227)) The model may also specify how these elements are arranged into larger structures, for example, how they are ordered into lists nested into trees. Metamodels vary in the tools they provide for specifying the structure and composition of domain-specific models, and in the maturity and robustness of the methods for designing them.[539][Com] RDF and XML each provide different, metamodel-specific tools to define a model for a specific domain. But not every metamodel provides such tools.

In XML, models are defined in separate documents known as *schemas*. An XML schema defining a domain model provides a vocabulary of terms that can be used as element and attribute names in XML documents that adhere to that model. For example, Onix for Books schema specifies that an author of a book should be called a `Contributor`, and that the page count should be called an `Extent`. An XML schema also defines rules for how those elements, attributes, and their content can be arranged into higher-level structures. For example, the Onix for Books specifies that the description of a book must include a list of `Contributor` elements, that this list must have at least one element in it, and that each `Contributor` element must have a `ContributorRole` child element.

If an XML schema is given an identifier, XML documents can use that identifier to indicate that they use terms and rules from that schema. An XML document may use vocabularies from more than one XML schema.[540][Com] Associating a schema with an XML instance enables *validation*: automatically checking that vocabulary terms are being used correctly.[541][Com]

If two descriptions share the same XML schema and use only that schema, then combining them is straightforward. If not, it can be problematic, unless someone has figured out exactly how the two schemas should "map" to one another. Finding such a mapping is not a trivial problem, as XML schemas may differ semantically, lexically, structurally, or architecturally despite sharing a common implementation form. (See Chapter 6, *Describing Relationships and Structures*.)

Tree structures can vary considerably while still conforming to the XML Infoset metamodel. Users of XML often specify rules for checking whether certain patterns appear in an XML document (document-level validation). This is less often done with RDF, because graphs that conform to the RDF metamodel all have the same structure: they are all sets of triples. This shared structure makes it simple to combine different RDF descriptions without worrying about checking structure at the document level. However, sometimes it is desirable to check descriptions at the document level, as when part of a description is required. As with XML, if consumers of those descriptions want to assert that they expect

those descriptions to have a certain structure (such as a required property), they must check them at the document level.

Because the RDF metamodel already defines structure, defining a domain-specific model in RDF mainly involves specifying URIs and names for predicates. A set of RDF predicate names and URIs is known as an *RDF vocabulary*. Publication of vocabularies on the web and the use of URIs to identify and refer to predicate definitions are key principles of Linked Data and the *Semantic Web*. (Also see §6.8.1, as well as later in this chapter.)[542][Web]

For example, the Resource Description and Access (RDA) standard for cataloging library resources includes a set of RDF vocabularies defining predicates usable in cataloging descriptions. One such predicate is:

```
<http://rdvocab.info/Elements/extentOfText>
```

which is defined as "the number and type of units and/or subunits making up a resource consisting of text, with or without accompanying illustrations." The vocabulary further specifies that this predicate is a refinement of a more general predicate:

```
<http://rdvocab.info/Elements/extent>
```

which can be used to indicate, "the number and type of units and/or subunits making up a resource" regardless of whether it is textual or not.

JSON lacks any standardized way to define which terms can be used. That does not mean one cannot use a standard vocabulary when creating descriptions using JSON, only that there is no agreed-upon way to use JSON to communicate which vocabulary is being used, and no way to automatically check that it is being used correctly.

9.2.3.2 Controlling Values

So far, we have focused on how models specify vocabularies of terms and how those terms can be used in descriptions. But models may also constrain the values or content of descriptions. Sometimes, a single model will define both the terms that can be used for property names and the terms that can be used for property values. For example, an XML schema may enumerate a list of valid terms for an attribute value.[543][Com]

Often, however, there are separate, specialized vocabularies of terms intended for use as property values in resource descriptions. Typically these vocabularies provide values for use within statements that describe what a resource is about. Examples of such subject vocabularies include the Library of Congress Subject Headings (LOC-SH) and the *Medical Subject Headings (MeSH)*. Other vocabularies may provide authoritative names for people, corporations, or places. Classi-

fication schemes are yet another kind of vocabulary, providing the category names for use as the values in descriptive statements that classify resources.

Because different metamodels take different approaches to specifying vocabularies, there will usually be different versions of these vocabularies for use with different metamodels. For example the LCSH are available both as XML conforming to the *Metadata Authority Description Standard (MADS)* schema, and as RDF using the *Simple Knowledge Organization System (SKOS)* vocabulary.

Specifying a vocabulary is just one way models can control what values can be assigned to properties. Another strategy is to specify what types of values can be assigned. For example, a model for book descriptions may specify that the value of a *pages* property must be a positive integer. Or it could be more specific; a course catalog might give each course an identifier that contains a two-letter department code followed by a 1-3 digit course number. Specifying a data type like this with a *regular expression* narrows down the set of possible values for the property without having to enumerate every possible value. (See the sidebar.)

In addition to or in lieu of specifying a type, a model may specify an encoding scheme for values. An *encoding scheme* is a specialized *writing system* or syntax for particular types of values. For example, a model like Atom for describing syndicated web content requires a publication date. But there are many different ways to write dates: `9/2/76`, `2 Sept. 1976`, `September 2nd 1976`, etc. Atom also specifies an encoding scheme for date values. The encoding scheme is RFC3339, a standard for writing dates. When using RFC3339, one always writes a date using the same form: `1976-09-02`.[545][Com]

Encoding schemes are often defined in conjunction with standardized identifiers. (See §4.4.3.1 Make Names Informative (page 195).) For example, International Standard Book Numbers (ISBN) are not just sequences of Arabic numerals: they are values written using the ISBN encoding scheme. This scheme specifies how to separate the sequence of numerals into parts, and how each of these parts should be interpreted. The ISBN `978-3-8218-4448-0` has five parts, the first three of which indicate that the resource with this identifier is 1) a product of the book publishing industry, 2) published in a German-speaking country, and 3) published by the publishing house Eichborn.

Encoding schemes can be viewed as very specialized models of particular kinds of information, such as dates or book identifiers. But because they specify not only the structure of this information, but also how it should be written, we can also view them as specialized *writing systems*. That is, encoding schemes specify how to *textually represent* information.

In the second half of this chapter, we will focus on the issues involved in textually representing resource descriptions—writing them. Graphs, trees, dictio-

Regular Expressions

Regular expressions have been used to describe patterns in text documents since the early days of computing and came into widespread use when Ken Thompson incorporated them into early UNIX text processing tools, such as *ed* and *grep*. There are too many variations of regular expression syntax for us to detail them here, but it is worthwhile to consider them briefly while we are on the subject of controlling values.[546][Com]

Regular expressions are employed by modern text processing tools for selection and retrieval purposes. In search and replace applications, one might search for the string "`Chapter [1-5]`" to express your intent to select chapters 1 through 5, or "`it['']?s`" to locate every use of "it's" and "its" in a manuscript; this capability is highly valued by anyone who has had to edit a book. Programmers and data modelers use regular expressions to describe expected encoding schemes when they design documents, data elements, databases, and encoding schemes. You experience regular expression processing when you enter a phone number or postal code into a Web-based form. Many data modeling, programming and XML schema languages employ regular expressions to control data entry and validation of values. In the context of controlling values, we can use regular expressions to describe data values as varied as identifiers, names, dates, telephone numbers, and postal codes. We can, likewise, define rules for white space handling and punctuation within a data value.

naries, lists, and sets are general types of structures found in different metamodels. Thinking about these broad types and how they fit or do not fit the ways we want to model our resource descriptions can help us select a specific metamodel. Specific metamodels such as the XML Infoset or RDF are formalized and standardized definitions of the more general types of structures discussed above. Once we have selected a metamodel, we know the constraints we have to work with when modeling the *resources* and collections in our specific domain. But because metamodels are abstract and exist only on a conceptual level, they can only take us so far. If we want to create, store, and exchange individual resource descriptions, we need to make the structures defined by our abstract metamodels concrete. We need to write them.

9.3 Writing Descriptions

Suppose that I am organizing books, and I have decided that it is important for the purposes of this organizing to know the title of each book and how many pages it has. Before me I have a book, which I examine to determine that its title is *Die Ringe des Saturn* and it has 371 pages. Example 9.5, Basic ways of writing part of a book description. lists a few of the ways to write this description. Let us examine these various forms of writing to see what they have in common and where they differ.

Example 9.5. Basic ways of writing part of a book description.

The title is *Die Ringe des Saturn* and it has 371 pages.
`{ book: {"title":"Die Ringe des Saturn","pages":371} }`
`<book pages="371"> <title>Die Ringe des Saturn</title> </book>`

`<div class="book">The title is`
`<span class="title">Die Ringe des Saturn</span>`
`and it has <span class="pages">371 pages.</span>`
`</div>`

`<http://lccn.loc.gov/96103072>`
`<http://rdvocab.info/Elements/title> "Die Ringe des Saturn"@de ;`
`<http://rdvocab.info/Elements/extentOfText> "371 p." .`

We examine the notations, writing systems and syntax of each of these description forms, and others, in the following sections.

9.3.1 Notations

First, let us look at the actual marks on the page. To write you must make marks or—more likely—select from a menu of marks using a keyboard. In either case, you are using a *notation*: a set of characters with distinct forms.[547][Com] The Latin alphabet is a notation, as are Arabic numerals. Some more exotic notations include the symbols used for editorial markup and alchemical symbols.[548][Com] The characters in a notation usually have an ordering. Arabic numerals are ordered *1 2 3* and so on. English-speaking children usually learn the ordering of the Latin alphabet in the form of an alphabet song.[549][CogSci]

A character may belong to more than one notation. The examples in Example 9.5, Basic ways of writing part of a book description. use characters from a few different notations: the letters of the Latin alphabet, Arabic numerals, and a handful of auxiliary marks: `. { } " :< > / $` Collectively, all of these characters —alphabet, numerals, and auxiliary marks—also belong to a notation called the American Standard Code for Information Interchange (ASCII).[550][Com]

ASCII is an example of a notation that has been codified and standardized for use in a digital environment. A traditional notation like the Latin alphabet can withstand a certain degree of variation in the form of a particular mark. Two people might write the letter *A* rather differently, but as long as they can mutually recognize each other's marks as an "A," they can successfully share a notation. Computers, however, cannot easily accommodate such variation. Each character must be strictly defined. In the case of ASCII, each character is given a number from 0 to 127, so that there are 128 ASCII characters.[551][Com] When using a computer to type ASCII characters, each key you press selects a character from this "menu" of 128 characters. A notation that has had numbers assigned to its characters is called a *character encoding*.

Table 9.1. ASCII

	0	1	2	3	4	5	6	7
0	NUL	DLE	space	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

The most ambitious character coding in existence is Unicode, which as of version 6.0 assigns numbers to 109,449 characters.[552][Com] Unicode makes the important distinction between *characters* and *glyphs*. A *character* is the smallest meaningful unit of a written language. In alphabet-based languages like English, *characters* are letters; in languages like Chinese, characters are ideographs. Unicode treats all of these *characters* as abstract ideas (*Latin capital A*) rather than specific marks (A *A* A **A**). A specific mark that can be used to depict a *character* is a *glyph*. A *font* is a collection of glyphs used to depict some set of *characters*. A Unicode font explicitly associates each glyph with a particular

number in the Unicode character encoding. The inability of computers to use contextual understanding to bridge the gap between various glyphs and the abstract *character* depicted by those glyphs turns out to have important consequences for organizing systems.

Different notations may include very similar marks. For example, modern music notation includes marks for indicating the pitch of note, known as accidentals. One of these music notation marks is ♯ ("sharp"). The sharp sign looks very much like the symbol used in English as an abbreviation for the word *number*, as in *We're #1!*[553][Ling] If you were to write a sharp sign and a number sign by hand, they would probably look identical. In a non-digital environment, we would rely on context to understand whether the written mark was being used as part of music notation, or mathematical notation, or as an English abbreviation.

Computers, however, have no such intuitive understanding of context. Unicode encodes the number sign and the sharp sign as two different characters. As far as a computer using Unicode is concerned, ♯ and # are completely different, and the fact that they have similar-looking glyphs is irrelevant. That is a problem if, for example, a cataloger has carefully described a piece of music by correctly using the sharp sign, but a person looking for that piece of music searches for descriptions using the number sign (since that is what you get when you press the keyboard button with the symbol that most closely resembles a sharp sign).[554][Com]

9.3.2 Writing Systems

A *writing system* employs one or more notations, and adds a set of rules for using them. Most writing systems assume knowledge of a particular human language. These writing systems are known as *glottic* writing systems. But there are many writing systems, such as mathematical and musical ones, that are not tied to human languages in this way. Many of the writing systems used for describing resources belong to this latter group, meaning that (at least in principle) they can be used with equal facility by speakers of any language.

Glottic writing systems, being grounded in natural human languages, are difficult to describe precisely and comprehensively. Non-glottic writing systems, on the other hand, can be described precisely and comprehensively using an abstract model. That is the connection between the structural perspective taken in the previous section, and the textual perspective taken in this section. A non-glottic writing system is described by a particular metamodel, and structures that fit within the constraints of a given metamodel can be textually represented using one or more writing systems that are described by that metamodel.

Some writing systems are closely identified with specific metamodels. For example, XML and JSON are *both* 1) metamodels for structuring information *and* 2)

writing systems for textually representing information. In other words, they specify both the abstract structure of a description and how to write it down. It is possible to conceive of other ways to textually represent the structure of these metamodels, but for each of these metamodels just one writing system has been standardized.[555][Com]

RDF, on the other hand, is *only* a metamodel, not a writing system. RDF only defines an abstract structure, not how to write that structure. So how do we write information that is structured as RDF? It turns out that we have many choices. Unlike XML and JSON, several different writing systems for the RDF metamodel have been standardized, including N-Triples, Turtle, RDFa, and RDF/XML.[556][Com] Each of these is a writing system that is abstractly described by the RDF metamodel.

Writing systems provide rules for arranging characters from a notation into meaningful structures. A character in a notation has no inherent meaning. Characters in a notation only take on meaning in the context of a writing system that uses that notation. For example: what does the letter *I* from the Latin alphabet mean? That question can only be answered by looking at how it is being used in a particular writing system. If the writing system is American English, then whether *I* has a meaning depends on whether it is grouped with other letters or whether it stands alone. Only in the latter case does it have an assignable meaning. However in the arithmetic writing system of ancient Rome, which also uses as a notation the letters of the Latin alphabet, *I* has a different meaning: *one*.

This example also serves to illustrate how the ordering of a notation can differ from the ordering of a writing system that uses that notation. According to the ordering of the Latin alphabet, the twelfth letter *L* comes before the twenty-second letter *V*. But in the Roman numeric writing system, *V* (the number 5) comes before *L* (the number 50). Unless we know which ordering we are using, we cannot arrange *L* and *V* "in order."[557][Ling]

Table 9.2. Roman Numerals

Roman Number	Arabic Number
I	1
V	5
X	10
L	50
C	100
D	500
M	1000

This kind of difference in ordering can arise in more subtle ways as well. When we alphabetically order names, we first compare the first character of each name, and arrange them according to the ordering of the writing system. The first known use of alphabetical ordering was in the Library of Alexandria about two thousand years ago, when Zenodotus arranged the collection according to the first letter of resource names.[558][Ling] If the first characters of two names are the same, we compare the second character, and so on. We can also apply this same kind of ordering procedure to sequences of numerals. If we do, then *334* will come before *67*, because *3* (the first character of the first sequence) comes before *6* (the first character of the second sequence) according to the ordering of our notation (Arabic numerals). However, it is more common when ordering sequences of numerals to treat them as decimal numbers, and thus to use the ordering imposed by the decimal system. In the decimal writing system, *67* precedes *334*, since the latter is a greater number.

This difference is important for organizing systems. Computers will sort values differently depending on whether they are treating sequences of numerals as numbers or just as sequences. Some organizing systems mix multiple ways of ordering the same characters. For example, Library of Congress call numbers have four parts, and sequences of Arabic numerals can appear in three of them. In the second part, indicating a narrow subject area, and fourth part, indicating year of publication, sequences of numerals are treated as numbers and ordered according to the decimal system. In the third part, however, sequences of numerals are treated as sequences and ordered "notationally" as in the example above (*334* before *67*).

Differences in ordering demonstrate just one way that multiple writing systems may use the same notation differently. For example, the American English and British English writing systems both use the same Latin alphabet, but impose slightly different spelling rules.[559][Ling] The Japanese writing system employs a number of notations, including traditional Chinese characters (*kanji*) as well as the Latin alphabet (*rōmaji*). Often, writing systems do not share the same exact notation but have mostly overlapping notations. Many European languages, for example, extend the Latin alphabet with characters such as *Å* and *Ü* that add additional marks, known as diacritics, to the basic characters.[560][Com]

In organizing systems it is often necessary to represent values from one writing system in another writing system that uses a different notation, a process known as *transliteration*. For example, early computer systems only supported the ASCII notation, so text from writing systems that extend the Latin alphabet had to be converted to ASCII, usually by removing (or sometimes transliterating) diacritics. This made the non-ASCII text usable in an ASCII-based computerized organizing system, at the expense of information loss.

Even in modern computer systems that support Unicode, however, transliteration is often needed to support organizing activities by users who cannot read text written using its original system. The Library of Congress and the American Library Association provide standard procedures for transliterating text from over sixty different writing systems into the (extended) Latin alphabet.

9.3.3 Syntax

The examples in Example 9.5, Basic ways of writing part of a book description. express the same information using different writing systems. The examples use the same notation (ASCII) but differ in their *syntax*: the rules that define how characters can be combined into words and how words can be combined into higher-level structures.[561][Com]

- Consider the first entry: *The title is* Die Ringe des Saturn *and it has 371 pages*. The leading capital letter and the period ending this sequence of characters indicate to us that this is a sentence. This sentence is one way we might use the English writing system to express two statements about the book we are describing. A *statement* is one distinct fact or piece of information. In glottic writing systems like English, there is usually more than one sentence we could write to express the same statement. For example, instead of *it has 371 pages* we might have written *the number of pages is 371*. English writing also enables us to construct complex sentences that express more than one statement.[562][Ling]

  In contrast, when we create descriptions of resources in an organizing system, we generally use non-glottic writing systems in which each sentence only expresses a single statement, and there is just one way to write a sentence that expresses a given statement.[563][Com] These restrictions make these writing systems less expressive, but simplify their use. In particular, since there is a one-to-one correspondence between sentences and statements, we can drop the distinction and just talk about the statements of a description.

  Now we return to our example and look at the structure of the statement, *The title is* Die Ringe des Saturn *and it has 371 pages*. Spaces are used to separate the text into words, and English syntax defines the functions of those words. The verb *is* in this statement functions to link the word *title* to the phrase *Die Ringe des Saturn*. This is typical of the kind of statements found in a resource description. Each statement identifies and describes some aspect of the resource. In this case, the statement attributes the value *Die Ringe des Saturn* to the property *title*.

  As we saw when we looked at description structures, we can analyze descriptions as involving properties of resources and their corresponding values or content. In a writing system like English, it is not always so straightforward to determine which words refer to properties and which refer to val-

ues. (This is why blobs are not ideal description structures.) Writing systems designed for expressing resource descriptions, on the other hand, usually define syntax that makes this determination easier. In our dictionary examples above, we used an arrow character → to indicate the relationship between properties and values.

This ease of distinguishing properties and values comes at a price, however. The syntax of English is forgiving: we can read a sentence with somewhat garbled syntax such as *371 pages it has* and often still make out its meaning.[564][Ling] This is usually not the case with writing systems intended for expressing resource descriptions. These systems strictly define their rules for how characters can be combined into higher-level structures. Structures that follow the rules are *well formed* according to that system.

- Take for example the second entry in Example 9.5, Basic ways of writing part of a book description..

  ```
  { book: {"title":"Die Ringe des Saturn","pages":371} }
  ```

  This fragment is written in JSON. As explained earlier in this chapter, JSON is a metamodel for structuring information using lists and dictionaries. But JSON is also a writing system, which borrows its syntax from JavaScript. The JSON syntax uses brackets to textually represent lists `[1,2,3]` and braces to textually represent dictionaries `{title:"Die Ringe des Saturn", "pages": 371}`. Within braces, the colon character `:` is used to link properties with their values, much as is was used in the previous example. So `"pages":371` is a statement assigning the value 371 to the property `pages`.

- The third fragment is written in XML.

  ```
  <book pages="371"> <title>Die Ringe des Saturn</title> </book>
  ```

  Like JSON, XML is a metamodel and also a writing system. Here we have XML elements and attributes. XML elements are textually represented as *tags* that are marked using the special characters `<`, `>` and `/`. So, this fragment of XML consists of a `book` element with a child element, `title`, and a `pages` attribute, each of which has some text content. In this case, `pages="371"` is a statement assigning the value 371 to the property `pages`. The difference is syntax is subtle; quotation marks surround the value and equal sign = is used to assign the property to its value.

- The fourth is a fragment of HTML.

  ```
  <div class="book">The title is
  <span class="title">Die Ringe des Saturn</span>
  and it has <span class="pages">371 pages.</span>
  </div>
  ```

The writing system that HTML employs is close enough to XML to ignore any differences in syntax. In this example, the CLASS attribute contains the property name and the property value is the element content.

- The fifth entry is a fragment of Turtle, one of the writing systems for RDF.

```
<http://lccn.loc.gov/96103072>
<http://rdvocab.info/Elements/title> "Die Ringe des Saturn"@de ;
<http://rdvocab.info/Elements/extentOfText> "371 p." .
```

  Turtle provides a syntax for writing down RDF *triples*. Each triple consists of a subject, predicate, and object separated by spaces. Recall that RDF uses URIs to identify subjects, predicates, and some objects; these URIs are written in Turtle by enclosing them in angle brackets < >. Triples are separated by period . characters, but triples that share the same subject can be written more compactly by writing the subject only once, and then writing the predicate and object of each triple, separated by a semicolon ; character. This is what we see in Example 9.2, Nesting an author description within a book description: two triples that share a subject.

The two fragments in Example 9.6, Writing part of a book description in Semantic XML. demonstrate namespaces, terms from the Dublin Core and DocBook namespaces, and the facility with which XML embraces semantic encoding of description resources.

Example 9.6. Writing part of a book description in Semantic XML.

```
<book xmlns:dc="http://purl.org/dc/terms/" dc:extent="371 p.">
<dc:title>Die Ringe des Saturn</title>
...
</book>
```

```
<book xmlns:db="http://www.docbook.org/xml/4.5/docbookx.dtd">
<bookinfo>
<title>Die Ringe des Saturn</title>
<pagenums>371 p.</pagenums>...</bookinfo>
...
</book>
```

- The first example extends the third fragment from Example 9.5, Basic ways of writing part of a book description.; the `xmlns:dc="..."` segment is a namespace declaration, which is associating `dc` with the quoted URI, which happens to be the Dublin Core Metadata Initiative (DCMI); the child `<dc:title>` element and the attached `dc:extent="371"` tell us that the corresponding values are attributable to the title and extent properties, respectively, from the Dublin Core namespace.
- The next fragment employs DocBook DTD namespace; we now have a `<pagenums>` element for which the meaning is contextually obvious; the title is still

a title; an extra layer of markup reflects the fact that it could be metadata in the source file of a book that is being edited, is in production or is on your favorite tablet right now.[565][Com]

Microformats, RDFa and Microdata

When Tim Berners-Lee deployed HTML, its syntax contained the basic elements and attributes needed to make formal statements about the document as a whole by using `<LINK/>`, or about specific parts of the document by using the `<A>` element. Each of these elements have four attributes in common: the famous HREF attribute contains a URI that names an object resource; the NAME attribute allows the element to be the target end of a link; the REL and REV attributes contain descriptions of the link relations.

Microformats, RDFa and Microdata are the latest generation of metadata extensions to HTML. Each approach is widely used on the web and by search engines. As such, they are potential targets when transforming into HTML from richer semantic formats.

Microformats are the simplest of the three. It uses controlled vocabularies of terms in REL/REV, and in the CLASS attribute, to declare high-level information types.

RDFa is RDF in Attributes. That is, RDFa is a formal specification for writing RDF expressions by using attributes in XML and HTML documents. It uses an ABOUT attribute to name the subject of the relation; the REL and REV attributes; HREF is joined by SRC and RESOURCE to name the object of the link; a TYPEOF attribute declares a type; PROPERTY and CONTENT attributes are used to *attribute* a value to an object's property.

Microdata is similar, inasmuch as it uses attributes extensively. The presence of an ITEMSCOPE attribute identifies an item while the ITEMTYPE attribute value identifies its type; ITEMID declares an items name or unique identifier; ITEMPROP is a name value pair, and; ITEMREF relates this item to other elements that are outside of the scope of the container element.

The two fragments in Example 9.7, Writing part of a book description in RDFa or microdata. demonstrate RDFa and microdata formats, which each rely upon specific attributes to establish the type of the property values contained by the HTML elements. In each example, the book title is contained by a `<span>` element. Whereas RDFa relies upon the `property` attribute, the microdata example employs the `itemprop` attribute to specify that the contents of the element is, effectively, a "title" in exactly the same sense as we know that the contents of `<dc:title>` is a "title."

Example 9.7. Writing part of a book description in RDFa or microdata.

```
<div class="book">The title is
<span property="http://purl.org/dc/terms/title">Die Ringe des Saturn</span>
and it has <span property="http://purl.org/dc/terms/extent">371 p.</span></div>
```

```
<div itemscope itemtype="book">The title is
<span itemprop="http://purl.org/dc/terms/title">Die Ringe des Saturn</span>
and it has <span itemprop="http://purl.org/dc/terms/extent">371 p.</span></div>
```

9.4 Worlds of Description

In the previous two sections we have considered descriptions as designed objects with particular structures and as written documents with particular syntaxes. As we have seen, there are many possible choices of structure and syntax. But these choices are never made in isolation. Just as an architect or designer must work within the constraints of the existing built environment, and just as any author must work with existing writing systems, descriptions are always created as part of a pre-existing "world" over which any one of us has little control.

In the final part of this chapter, we will consider how choices of structure and syntax have converged historically into broad patterns of usage. For lack of a better term, we call these broad patterns "worlds." "World" is not a technical term and should not be taken too literally: the broad areas of application sketched here have considerable overlap, and there are many other ways one might identify patterns of description structure and syntax. That said, the three worlds described here do reflect real patterns of description form that influence tool and technology choices. In your own work creating and managing resource descriptions, it is likely that you will need to think about how your descriptions fit into one or more of these worlds.

9.4.1 The Document Processing World

The first world we will consider is concerned primarily with the creation, processing and management of hybrid narrative-transactional documents such as instruction manuals, textbooks, or annotated medieval manuscripts. (See The Document Type Spectrum (page 168)). These are quite different kinds of documents, but they all contain a mixture of narrative text and structured data, and they all can be usefully modeled as tree structures. Because of these shared qualities, tools as different as publishing software, supply-chain management software, and scholarly editing software have all converged on common XML-based solutions. ("The XML world" would be another appropriate name for the document-processing world.)

This convergence was no accident, because XML was designed specifically to address the problem of how to add structure and data to documents by "marking them up." XML is the descendant of Standard Generalized Markup Language (SGML), which in turn descended from International Business Machines (IBM)'s Generalized Markup Language, which was invented to enable the production and management of large-scale technical documentation. The explicitness of markup makes it well-suited for representing structure and content type distinctions in institutional contexts, where the scope, scale, and expected lifetime of organizing systems for information implies reuse by unknown people for unanticipated purposes.

The abstract data model underlying XML is called the XML Information Set or Infoset. The Infoset defines a document as a partially ordered tree of "information items." Every XML document can thus be understood as a specific kind of tree, although not every tree structure is expressible as an XML document.[566][Com]

As we discussed in Inclusions and References (page 456), XML has the ability to describe graphs by incorporating the use of ID and IDREF attribute types to create references among element information items within the same document. This modest form of hypertext linking allows us to present the following document fragment that approximates the graph we saw modeled in Figure 9.4, Descriptions Linked into a Graph.

Example 9.8. XML implementation of a biblio-graph

```
<person id="WG.Sebald">Winfried George Sebald</person>
<person id="MR.McCulloch>Mark Richard McCulloch</person>

<book>
    <title>Understanding W.G. Sebald</title>
    <subject idref="WG.Sebald"/>
    <author idref="WG.Sebald"/>
    <author idref="MR.McCulloch"/>
</book>

<book pages="371">
    <title lang="de">Die Ringe des Saturne</title>
    <title lang="en">The Rings of Saturn</title>
    <author idref="WG.Sebald"/>
</book>

<book pages="416">
    <title lang="de">Austerlitz</title>
    <author idref="WG.Sebald"/>
</book>
```

As one might expect, tools and technologies in the document-processing world are optimized for manipulating and combining tree structures. A "toolchain" is set of tools intended to be used together to achieve some goal.

The XML Toolchain

The XML toolchain is quite comprehensive. It consists of tools for creating XML documents (XML editors), tools for expressing logical document and data models (DTD, XML Schema, *REgular LAnguage for XML Next Generation (RELAX NG)*, Schematron), tools for transforming XML documents (XSLT), tools for describing document processing "pipelines" (*XProc: An XML Pipeline Language*), and tools for storing and querying collections of XML documents (XML databases, queried using XML Query Language (XQuery)). Used together, these tools provide very powerful means of working with tree-structured documents. XML editors incorporate knowledge of DTDs, schemas, transformations, style sheets, queries, databases and pipelines. Pipelines choreograph the plumbing and inter-dependencies involved in processing a complex dataset and publishing a useful result in one or more output formats.

For programmers who do not to use the XML toolchain, other programming languages also provide libraries for working with XML. This fact has led some to propose, and others to believe, that XML is a kind of universal format for exchanging data among systems. However, programmers have observed that a random XML Infoset does not map easily to the data structures commonly found in many programming languages. "Working with XML" frequently means translating from XML tree structures to data structures native to another language, usually meaning lists and dictionaries. This translation can be problematic and often means giving up many of the strengths of XML. By the same token, there are decades more practical experience working with markup languages and institutional publishing than there is with JSON and RDF.

XML is not a universal solution for every possible problem. That does not mean that it is not the best solution for a wide variety of problems, including yours. To gauge whether your resource descriptions are, or ought to be, part of the document-processing world, ask yourself the following questions:

- Do my resource descriptions contain mixtures of narrative text, hypertext, structured data and a variety of media formats?
- Can my descriptions easily be modeled using tree structures, hypertext links, and transclusion?
- Are the vocabularies I need or want to use made available using XML technologies?

- Do I need to work with a body of existing descriptions already encoded as XML?
- Do I need to interoperate with processes or partners that utilize the XML toolchain?
- Do I need to publish my resource descriptions in multiple formats from a single source?

If the answer to one or more of these questions is "yes," then chances are good that you are working within the document processing world, and you will need to become familiar with conceptualizing your descriptions as trees and working with them using XML tools.

9.4.2 The Web World

The second "world" emerged in the early 1990s with the creation of the World Wide Web. The web was developed to address a need for simple and rapid sharing of scientific data. Of course, it has grown far beyond that initial use case, and is now a ubiquitous infrastructure for all varieties of information and communication services. ("The browser world" would be another appropriate name for what we are calling the Web World.)

Documents, data, and services on the web are conceptualized as resources, identified using Uniform Resource Identifiers (URI), and accessible through *representations* transferred via Hypertext Transfer Protocol (HTTP). Representations are sequences of bytes, and could be HTML pages, JPEG images, tabular data, or practically anything else transferable via HTTP. No matter what they are, representations transferred over the web include descriptions of themselves. These descriptions take the form of property-value pairs, known as "HTTP headers." The HTTP headers of web representations are structured as dictionaries.

Dictionary structures appear many other places in web infrastructure. URIs may include a *query* component beginning with a ? character. This component is used for purposes such as providing query parameters to search services. The query component is commonly structured as a dictionary, consisting of a series of property-value pairs separated by the & character. For example, the following URI:

```
https://www.google.com/search?q=sebald&tbs=qdr:m
```

includes the query component `q=sebald&tbs=qdr:m`. This is a dictionary with the properties `q` and `tbs`, respectively specifying the search term and temporal constraints on the search.

Data entered into an HTML form is also structured as a dictionary. When an HTML form is submitted, the entered data is used either to compose the query

component of a URI, or to create a new representation to be transferred to a web server. In either case, the data is structured as a set of properties and their corresponding values.

HTML documents are structured as trees, but descriptions embedded within HTML documents can also be structured as dictionaries. HTML documents may include a dictionary of metadata elements, each of which specifies a property and its value. Recently support for *microdata* was added to HTML, which is another method of adding dictionaries of property-value pairs to documents. Using *microdata*, authors can annotate web content with additional information, making it easier to automatically extract structured descriptions of that content.[567][Web] *Microformats* are another method for doing this by mapping existing HTML attributes and values to (nested) dictionary structures.[568][Web]

Dictionary structures are easy to work with in any programming language, and they pervade various popular frameworks for programming the Web. In the programming languages used to implement web services, HTTP headers and query parameters are easily mapped to dictionary data structures native to those languages. On the client side, there is only one programming language that runs within all web browsers: JavaScript. The dictionary is the fundamental data structure within JavaScript as well.

Thus it is unsurprising that JSON, a dictionary-structured, JavaScript-based syntax, has become the *de facto* standard for application-to-application interchange of data on the web in contexts that do not involve business transactions. Web services providing structured data intended for programmatic use can make that data available as JSON, which is well-suited for use either by JavaScript programs running within browsers, or by programs written in other languages running outside of browsers (e.g., smart phone applications).

It is now commonly accepted that there are useful differences of approach between the document-processing world and the Web World. This does not mean that the two worlds do not have significant overlaps. Some very important web representation types are XML-based, such as the Atom syndication format. Trees will continue to be the structure of choice for web representations that consist primarily of narrative rather than transactional data. But for structured descriptions that are intended to be accessed and manipulated on the Web, dictionary structures currently rule.

To gauge whether your resource descriptions are or ought to be part of the Web world, ask yourself the following questions:

- Is the web the primary platform upon which I will be making my descriptions available?
- Are my resource descriptions primarily structured, transaction-oriented data?

- Can my descriptions easily be modeled as lists of properties and values (dictionaries)?
- Are the vocabularies I need or want to use made available primarily using HTML technologies such as microdata or microformats?
- Do I need to make my descriptions easily usable for use within a wide array of programming languages?

If the answer to one or more of these questions is "yes," then chances are good that you are working within the Web World, and you will need to become familiar with conceptualizing your descriptions as dictionaries and working with them using programming languages such as JavaScript.

9.4.3 The Semantic Web World

The last world we consider is still somewhat of a possible world, at least in comparison with the previous two. While the document processing world and the web world are well-established, the Semantic Web world is only starting to emerge, despite having been envisioned over a decade ago.

The vision of a *Semantic Web* world builds upon the web world, but adds some further prescriptions and constraints for how to structure descriptions. The Semantic Web world unifies the concept of a resource as it has been developed in this book, with the web notion of a resource as anything with a URI. On the Semantic Web, anything being described must have a URI. Furthermore, the descriptions must be structured as graphs, adhering to the RDF metamodel and relating resources to one another via their URIs. Advocates of Linked Data further prescribe that those descriptions must be made available as representations transferred over HTTP.[569][Web]

This is a departure from the web world. The web world is also structured around URIs, but it does not require that every resource being described have a URI. For example, in the web world a list of bibliographic descriptions of books by W.G. Sebald might be published at a specific URI, but the individual books themselves might not have URIs. In the Semantic Web world, in addition to the list having a URIs, each book would have a URI too, in addition to whatever other identifiers it might have.[570][Web]

Making an HTTP request to an individual book URI may return a graph-structured description of that book, if best practices for Linked Data are being followed. This, too, is a departure from the web world, which is agnostic about the form representations or descriptions of resources should take (although as we have seen, dictionary structures are often favored on the web when the clients consuming those descriptions are computer programs). On the Semantic Web, all descriptions are structured as RDF graphs. Each description graph links to other description graphs by referring to these related resources using

their URIs. Thus, at least in theory, all description graphs on the Semantic Web are linked into a single massive graph structure. In practice, however, it is far from clear that this is an achievable, or even a desirable, goal.

Although the Semantic Web is in its infancy, a significant number of resource descriptions have already been made available in accordance with the principles outlined above. Descriptions published according to these principles are often referred to as "Linked Data." Prominent examples include: DBpedia, a graph of descriptions of subjects of Wikipedia articles; the *Virtual International Authority File (VIAF)*, a graph of descriptions of names collected from various national libraries' name authority files; GeoNames, a graph of descriptions of places; and Data.gov.uk, a graph of descriptions of public data made available by the UK government.[571][Web]

Despite the growing amount of Linked Data, tools for working with graph-structured data are still immature in comparison to the XML toolchain and Web programming languages. Although there is an XML syntax for RDF, using the XML toolchain to work with graph-structured data is generally a bad idea. And just as most programming languages do not support natively working with tree structures, most do not support natively working with graph structures either. Storing and querying graph-structured data efficiently requires a graph database or *triple store*.

Still, the Semantic Web world has much to recommend it. Having a common way of identifying resources (the URI) and a single shared metamodel (RDF) for all resource descriptions makes it much easier to combine descriptions from different sources. To gauge whether your resource descriptions are or ought to be part of the Semantic Web world, ask yourself the following questions:

- Is the web the primary platform upon which I will be making my descriptions available?
- Is it important that I be able to easily and freely aggregate the elements of my descriptions in different ways and to combine them with descriptions created by others?
- Are my descriptions best modeled as graph structures?
- Have the vocabularies I need or want to use been created using RDF?
- Do I need to work with a body of existing descriptions that have been published as Linked Data?

If the answer to one or more of these questions is "yes," then chances are good that you should be working within the Semantic Web world, and you ought to become familiar with conceptualizing your descriptions as graphs and working with them using Semantic Web tools.

9.5 Key Points in Chapter Nine

- We can approach the problem of how to form resource descriptions from two perspectives: structuring and writing.

  (See §9.1 Introduction (page 437))
- Metamodels describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain.

  (See §9.2 Structuring Descriptions (page 439))
- Blobs, sets, lists, dictionaries, trees, and graphs are all kinds of structures that can be used to form resource descriptions.

  (See §9.2.1 Kinds of Structures (page 442))
- A *list*, like a set, is a collection of items with an additional constraint: their items are *ordered*.

  (See §9.2.1.3 Lists (page 443))
- A *dictionary*, also known as a *map* or an *associative array*, is a set of property-value pairs or *entries*.

  (See §9.2.1.4 Dictionaries (page 444))
- Nested dictionaries form a *tree*.

  (See §9.2.1.4 Dictionaries (page 444))
- Trees consist of *nodes* joined by *edges*.

  (See §9.2.1.5 Trees (page 445))
- JSON consists of two kinds of structures: lists (called *arrays* in JavaScript) and dictionaries (called *objects* in JavaScript).

  (See §9.2.2.1 JSON (page 451))
- The XML Infoset is a tree structure, where each node of the tree is defined to be an *information item* of a particular type.

  (See §9.2.2.2 XML Information Set (page 451))
- Using schemas to define data representation formats is a good practice that facilitates shared understanding and contributes to long-term maintainability.

  (See §9.2.2.2 XML Information Set (page 451))
- The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple.

  (See §9.2.2.3 RDF (page 453))

- An “encoding scheme” is a specialized writing system or syntax for particular types of values. Encoding schemes specify how to *textually represent* information.

  (See §9.3.1 Notations (page 462))

- A *writing system* employs notations, and adds a set of rules for using them.

  (See §9.3.2 Writing Systems (page 464))

- Differences in ordering demonstrate just one way that multiple writing systems may use the same notation differently.

  (See §9.3.2 Writing Systems (page 464))

- Syntax is the rules that define how characters can be combined into words and how words can be combined into higher-level structures.

  (See §9.3.3 Syntax (page 467))

- The document processing world is concerned primarily with the creation, processing and management of hybrid narrative-transactional documents.

  (See §9.4.1 The Document Processing World (page 471))

- In the web world, documents, data, and services are conceptualized as resources, identified using Uniform Resource Identifiers (URI), and accessible through *representations* transferred via the Hypertext Transfer Protocol (HTTP).

  (See §9.4.2 The Web World (page 474))

- The Semantic Web world unifies the concept of a resource as it has been developed in this book, with the web notion of a resource as anything with a URI. Descriptions must be structured as graphs, adhering to the RDF metamodel and relating *resources* to one another via their URIs.

  (See §9.4.3 The Semantic Web World (page 476))

Endnotes for Chapter 9

[516][Com] This discussion of Batten’s cards is based on (Lancaster 1968, pages 28-32). Batten’s own explanation is in (Batten 1951).

[517][Ling] (Silman 1998). (Sebald 1995).

[518][Ling] The technique of diagramming sentences was invented in the mid-19th century by Stephen W. Clark, a New York schoolmaster; (Clark2010) is an exact reprinting of a nearly 100 year old edition of his book *A Practical Grammar.* A recent tribute to Clark is (Florey 2012).

[519][CogSci] It is easy to underestimate the incredible power of the human perceptual and cognitive systems to apply neural computation and knowledge to enable vision and hearing to seem automatic. Computers are getting better at extracting features from visual and auditory signals to identify and classify inputs, but our point here is that none of these features are explicitly represented in the input "blob" or "stream."

[520][Com] As we commented earlier, an oral description of a resource may not be especially useful in an organizing system because computers cannot easily understand it. On the other hand, there are many contexts in which an oral description would be especially useful, such as in a guided tour of a museum where visitors can use audio headsets.

[522][Com] It is rarely practical to make things as simple as possible. According to Einstein, we should endeavor to "Make everything as simple as possible, but not simpler."

[523][Com] This structural metamodel only allows one value for each property, which means it would not work for books with multiple authors or that discuss multiple subjects.

[525][Com] The XML Information Set (Cowan2004)

RDF/XML is one example where meta models meet. In *Document Design Matters,* (Wilde and Glushko 2008b) point out that "If the designer of an exchange format uses a non-XML conceptual metamodel because it seems to be a better fit for the data model, XML is only used as the physical layer for the exchange model. The logical layer in this case defines the mapping between the non-XML conceptual model, and any reconstruction of the exchange model data requires the consumer to be fully aware of this mapping. In such a case, it is good practice to make users of the API aware of the fact that it is using a non-XML metamodel. Otherwise they might be tempted to base their implementation on a too small set of examples, creating implementations which are brittle and will fail at some point in time."

[526][Com] Technically, what is described here is referred to as "rooted tree" by mathematicians, who define trees more generally. Since trees used as data structures are always rooted trees, we do not make the distinction here.

[527][Com] This feature relies upon the existence of an XML schema. An XML schema can declare that certain attributes are of type ID, IDREF or IDREFS. Whether an XML DTD or one of the many schema languages that have been developed under the auspices of the W3C or ISO.

[528][Com] `http://www.w3.org/TR/xml-infoset/`.

[529][Com] The *XML Infoset* is one of many metamodels for XML, including the DOM and XPath. Typically, an XML Infoset is created as a by-product of parsing

a well-formed XML document instance. An XML document may also be informed by its DTD or schema with information about the types of attribute values, and their default values. Attributes of type ID, IDREF and IDREFs provide a mechanism for intra-document hypertext linking and transclusion. An XML document instance may contain entity definitions and references that get expanded when the document is parsed, thereby offering another form of transclusion.

[530][Com] A well-formed XML document instance, when processed, will yield an XML Information Set, as described here. Information sets may also be constructed by other means, such as transforming from another information set. See the section on *Synthetic Infosets* at `http://www.w3.org/TR/xml-infoset/#intro.synthetic` for details.

[531][Com] The Infoset contains knowledge of whether all related declarations have been read and processed, the base URI of the document instance, information about attribute types, comments, processing instructions, unparsed entities and notations, and more.

A well-formed XML document instance for which there are associated schemas, such as a DTD, may contribute information to the Infoset. Notably, schemas may associate data types with element and attribute information items, and it may also specify default or fixed values for attributes. A DTD may define entities that are referenced in the document instance and are expanded in-place when processed. These contributions can affect the truth value of the document.

[532][Com] The SGML standard explicitly stated that documentation describing or explaining a DTD is part of the document type definition. The implication being that a schema is not just about defining syntax, but also semantics. Moreover, since DTDs do not make possible to describe all possible constraints, such as co-occurrence constraints, the documentation could serve as human-consumable guidance for implementers as well as content creators and consumers.

[533][Com] Attribute types may be declared in an XML DTD or schema. Attributes whose type is ID must have a valid XML name value that is unique within that XML document; an attribute of type IDREF whose value corresponds to a unique ID has a “references” property whose value is the element node that corresponds to the element with that ID. An attribute of type IDREFS whose value corresponds to a list of unique ID has a “references” property whose value is a list of element node(s) that corresponds to the element(s) with matching IDs.

[534][Com] XML Inclusions (XInclude) is (Marsh, Orchard, and Veillard 2006).

[535][Com] XML Linking Language (XLink) is (DeRose, Maler, Orchard, and Walsh 2010).

[536][Com] Within the document’s DTD, one simply declares the entity and its corresponding value, which could be anything from an entire document to a phrase

and then it may be referenced in place within the XML document instance. The entity reference is replaced by the entity value in the XML Infoset. Entities, as nameable wrappers, effectively disappear on their way into the XML Infoset.

[537][Bus] Online Information Exchange (ONIX) is the international standard for representing and communicating book industry product information in electronic form: http://www.editeur.org/11/Books/.

[538][Com] Do not take on the task of creating a new XML model lightly. Literally thousands of XML vocabularies have been created, and some represent hundreds or thousands of hours of effort. See (Bray 2005) for advice on how to reduce the risk of vocabulary design if you cannot find an existing one that satisfies your requirements.

[539][Com] See (Glushko and McGrath 2005) for a synthesis of best practices for creating domain-specific languages in technical publishing and business-to-business document exchange contexts. You need best practices for big problems, while small ones can be attacked with *ad hoc* methods.

[540][Com] Unless an XML instance is associated with a schema, it is fair to say that it does not have any model at all because there is no way to understand the content and structure of the information it contains. The assignment of a schema to an XML instance requires a "Document Type Declaration." If some of the same vocabulary terms occur in more than one XML schema, with different meanings in each, using elements from more than one schema in the same instance requires that they be distinguished using *namespaces.* For example, if an element named "title" means the "title of the book" in one schema and "the honorific associated with a person" in another, instances might have elements with namespace prefixes like <book:title>The Discipline of Organizing</book:title> and <hon:title>Professor</hon:title>. Namespaces are a common source of frustration in XML, because they seem like an overly complicated solution to a simple problem. But in addition to avoiding naming collisions, they are important in schema composition and organization.

[541][Com] What "correctly" means depends on the schema language used to encode the conceptual model of the document type. The XML family of standards includes several schema languages that differ in how completely they can encode a document type's conceptual model. The Document Type Definition (DTD) has its origins in publishing and enforces structural constraints well; it expresses strong data typing through associated documentation resources. *XML Schema Definition Language (XSD)* is better for representing transactional document types but its added expressive power tends to make it more complex.

[542][Web] For example, see Linked Open Vocabularies at http://lov.okfn.org/dataset/lov/index.html.

[543][Com] Attribute values can be constrained in a schema by specifying a data type, a default value, and a list of potential values. Data types allow us to specify whether a value is supposed to be a name, a number, a date, a token or a string of text. Having established the data type, we can further constrain the value of an attribute by specifying a range of values, for a number or a date, for example. We can also use *regular expression* patterns to describe a data type such as a postal code, telephone number or ISBN number. Specifying default values and lists of legal values for attributes simplifies content creation and quality assurance processes. In Schematron, a rule-based XML schema language for making test assertions about XML documents, we can express constraints between elements and attributes in ways that other XML schema languages cannot. For example, we can express the constraint that if two `<title>` elements are provided, then each must contain a unique string value and different `language` attribute values.

[545][Com] The Atom Publishing Protocol is IETF RFC 5023, (https://tools.ietf.org/html/rfc5023); a good introduction is (Sayre 2005). IETF RFC is http://www.ietf.org/rfc/rfc3339.txt.

[546][Com] There is no single authority on the subject of regular expressions or their syntax. A good starting point is the Wikipedia article on the subject: http://en.wikipedia.org/wiki/Regular_expression.

[547][Com] The terminology here and in the following sections comes from (Harris 1996).

[548][Com] See http://unicode.org/charts/PDF/U1F700.pdf.

[549][CogSci] Entitled “The ABC,” the song was copyrighted in 1835 by Boston music publisher Charles Bradlee. It is sung to a tune that was originally developed by Wolfgang Amadeus Mozart, and is commonly recognizable as *Twinkle, Twinkle, Little Star.*

[550][Com] http://tools.ietf.org/html/rfc20.

[551][Com] Only 95 of these characters are actually “marks” in the sense of being visible and printable. The other 33 ASCII characters are “control codes” that indicate things like horizontal and vertical tabs, the ends of printed lines, form feeds, and transmission control. We can think of many of these as special auxiliary marks, similar to the kind of symbols editors and proofreaders use to annotate texts.

[552][Com] The Unicode standard is maintained by a global non-profit organization. Everything you need to know is at http://www.unicode.org/.

[553][Ling] The Chinese character 井 (water well) looks like the # character too. The # symbol was historically used to denote pounds, the Imperial unit of

weight, as in 10# of potatoes. In the United Kingdom, the # character is called“hash.” We could go on, but we will leave it to you to discover more.

[554][Com] To add to the confusion, while the American standard (ASCII) places the # character at position 23, the British equivalent (BS 4730) places the currency symbol £ at the same position. As a result, improperly configured computers sometimes display # in place of £ and vice versa.

[555][Com] Recently, an alternative writing system for XML-structured data has been standardized: Efficient XML Interchange (EXI). However it is not yet widely used.

[556][Com] RDF/XML is a bit confusing; it is a writing system that uses XML syntax to textually represent RDF structure. This means that while XML tools can read and write RDF/XML, they cannot manipulate the graph structures it represents, because they were designed to work with XML’s tree structures.

[557][Ling] Although we use alphabetic characters today to represent Roman numerals, originally they were represented by unique symbols.

[558][Ling] It took a few hundred years before alphabetization became recursive and applied to letters other than the first (Casson 2002, p. 37). Alphabetization relies on the ordering of the writing system, not the notation. For example, Swedish and German are two writing systems that assign different orderings to the same notation.

[559][Ling] For example, the American spelling of the words “center” and “color” contrasts slightly with the English spelling of “centre” and “colour.” There are too many examples to include here. Wikipedia has a comprehensive analysis of American and British spelling differences at `http://en.wikipedia.org/wiki/American_and_British_English_spelling_differences`.

[560][Com] ASCII’s 128 characters are insufficient to represent these more complex character sets, so a new family of character encodings was created, ISO-8859, in which each encoding enumerates 256 characters. Each encoding thus has more space to accommodate the additional characters of regionally-specific notations. ISO 8859-5, for example, has extensions to support the Cyrillic alphabet.

[561][Com] In discussions of glottic writing systems, “syntax” usually refers only to the rules for combining words into sentences. In discussions of programming languages, “syntax” has the broader sense we use here.

[562][Ling] Compound sentences contain two independent clauses joined by a conjunction, such as “and,” “or,” “nor,” “but.” For example: I went to the store and I bought a book.” Complex sentences contain an independent clause joined by one or more dependent clauses. For example: “I read the book that I bought at the store.”

[563][Com] In truth, even non-glottic writing systems designed to encode resource descriptions unambiguously can have variant forms of the same statement. For example, XML permits some variation in the way the same Infoset may be textually represented. Often these variations involve the treatment of content that may under some circumstances be treated as optional, such as white space. The difference is that in writing systems designed for resource description, these variations can be precisely enumerated and rules developed to reconcile them, while this is not generally true for glottic writing systems.

[564][Ling] Fortunately for Yoda. There are many web services for converting English to Yoda-speak; an example is `http://www.yodaspeak.co.uk/`.

[565][Com] DocBook (Walsh 2010) is widely used to publish academic, commercial, industrial book, scientific, and computing book, papers and articles. The book that you are reading is encoded with DocBook markup; complete bibliographic information for the book is contained within the source files, ready to be extracted on the way into one of the latest ebook formats.

[566][Com] It should be noted that the content of the Infoset for a given document may be affected by knowledge of any related DTDs or schemas. That is to say that, upon examination of a given XML document instance, its Infoset may be augmented with some useful information, such as default attribute values and attribute types. (See Inclusions and References (page 456).)

[567][Web] Microdata is an invention of WHATWG and exists and part of what they call a “living standard.” It was supported by Google, so it was widely used and there exist numerous controlled vocabularies, including those for creative works, persons, events and organizations. Support for microdata has since been withdrawn from Apple Safari and Google Chrome browsers.

[568][Web] Microformats is a non-standard that emerged from the community and has been sponsored by CommerceNet and Microformats.org.

[569][Web] (Bizer, Heath, and Berners-Lee 2009).

[570][Web] It is worth noting that URIs are not required to have anything at their endpoints. Resolvability of URIs is evangelized as a best practice for Linked Data but not a requirement within the broader Semantic Web paradigm. Merely asserting that a URI is associated with a book is enough. If the URI can return a description or a resource, so much the better, but if not, at least you can talk about the book by referring to the same URI.

[571][Web] Many more available datasets are listed at linkeddata.org.

Chapter 10

Interactions with Resources

Vivien Petras
Robert J. Glushko
Ian MacFarland
Karen Joy Nomorosa
J.J.M. Ekaterin
Hyunwoo Park
Robyn Perry
Sean Marimpietri

10.1 Introduction

Picture a dim room in the basement of a Detroit police station, lined with metal shelves: the shelves contain boxes and boxes of cold case files, evidence meticulously logged and categorized for no one to look at, documenting murders that will never be solved. Or the library of a small-town historical society in New Jersey: struggling with budget cuts, the board of directors has been forced to close its doors, locking its treasures inside, carefully curated and preserved but inaccessible to the public. Or a valuable data store encoded in an orphaned storage format: business records in a legacy database system that will not run on modern computers, census data on proprietary magnetic tape reels from the 1970s, your unfinished novel on a series of eight-inch floppy disks. You know the data is there, but you cannot interact with it.

An organizing system without interactions is a sad one indeed.

Interactions are the answer to two of the fundamental questions we posed back in Chapter 1: why and when are the resources organized?

The question of "why?" has been in the background (and often the foreground) of every chapter in this book thus far; whenever we select a resource for inclusion in an organizing system, describe it, or arrange it according to an organizing principle, we have an interaction in mind. We include a resource in our system because our users will need it; we assign a resource to one or more categories to help our users find it, understand it, and connect it with other resources in a meaningful way.

In this chapter we will pivot from design for interactions to the design of interactions—and to do this we must pause to consider the question of "when?" In §2.5, we contrasted organization done "on the way in" with that done "on the way out," but this distinction is not always a particularly relevant one. Consider a bookshelf: if you do not organize its resources on the way in (i.e., when you put a book on the shelf), you cannot really organize them on the way out; you just have a disorganized bookshelf. When the time comes to retrieve a book, you'll have to employ a brute-force linear search algorithm—reading every spine until you find the one you want, and it will not make the remaining books on the shelf any more organized.

But digital resources and networked organizing systems are an entirely different story. In fact, we argue that they blur the traditional boundary between the academic disciplines of "information organization" and "information retrieval"; with the World Wide Web, ubiquitous digital information, and effectively unlimited processing, storage, and communication capability driven by cloud computing architecture and Moore's law, billions of people create and browse websites, blog, tag, tweet, and upload and download content of all media types without thinking "I am organizing now" or "I am retrieving now." When people use their smartphones to search the web or run applications, location information transmitted from their phone is used to filter and reorganize the information they retrieve. Arranging results to make them fit the user's location is a kind of computational curation, but because it takes place quickly and automatically we hardly notice it. Likewise, almost every application that once seemed predominantly about information retrieval is now increasingly combined with activities and functions that most would consider to be information organization.

Thus we come to the question of when a system's resources are organized: we may apply the techniques of computational information retrieval to a set of resources that simply are not organized the way we need them to be in order to support our desired interaction. Maybe the system was designed poorly or for a different purpose than the one we are pursuing; maybe we are attempting to collect or aggregate resources from multiple organizing systems, each of which

has its own separate purposes and design flaws. Regardless of the reasons, what we are essentially doing is reorganizing these resources on the fly, or "on the way out," following many of the same principles and procedures we've covered in the preceding eight chapters of this book.

Most Common Museum Interaction

Because museums often contain extremely rare or valuable resources that do not circulate, their most popular items are mobbed by visitors. The crowding often makes it impossible to get a good look at the rare item. This ironic situation is typified by the crowd control cordon that creates a 20-foot barrier around La Gioconda (aka "The Mona Lisa") at Musée du Louvre in Paris.

(Photo by R. Glushko.)

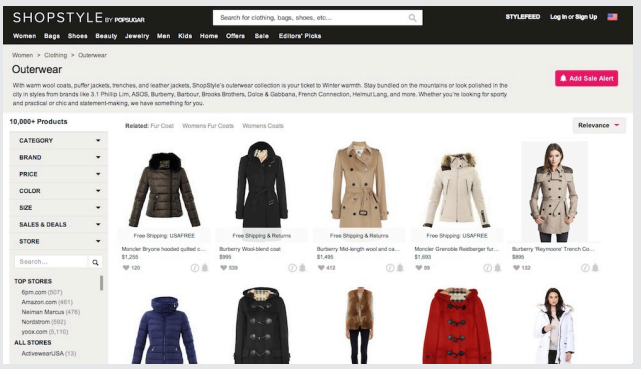
The fundamental interaction of any organizing system is accessing resources or resource descriptions, whether physically or digitally. Sometimes we must combine or merge resources or resource descriptions to access them effectively; this poses numerous strategy, design, and implementation challenges, as producers often use different identifiers, description or cataloging formats, and practices for similar resources. Different service providers use different technol-

ogies, have different information policies, and follow different processes developed in their separate organizing systems.

Some organizing systems have the power to determine the description standards that others must use. Walmart, the largest retailer in the United States, has devised an organizing system for its supply chain that supports access and movement of physical goods with maximal efficiency and effectiveness. This system saves the corporation money on inventory management and distribution, but to maximize savings, Walmart requires its suppliers to employ the same data model, follow company-set standards, and adopt new technologies such as bar codes and RFID tags that support the highly efficient interactions it requires.[572][Bus]

Other organizing systems must adapt to whatever their counterparts develop. Online retailer Shopstyle.com presents a typical ecommerce interface, allowing shoppers to browse a multitude of fashion and beauty products organized into familiar categories. But behind the scenes, Shopstyle is aggregating the catalogs of more than 250 online stores and providing a seamless access interaction for all their merchandise. It does not actually sell anything: it directs shoppers to those third-party stores to make their purchases. Rather than moving physical resources like Walmart, Shopstyle's most important interactions involve moving and combining digital resource descriptions.

Browsing Merchandise Catalogs



Shopstyle.com provides a transparent interface to the catalogs of hundreds of other online clothing retailers, aggregating their listings to allow users to browse them all from a single page.

(Screenshot by Ian MacFarland.)

Still others choose to abide by what a standard-setting body decides, or participate in laborious, democratic processes to align their organizing practices and interactions.[573][Bus] Libraries and museums are the classic examples of this. The most important interaction in a library, of course, is borrowing: checking out a book to use it off the premises, and checking it back in when you're done. Patrons search descriptions in a catalog to find books on a certain topic, by a certain author, or with a certain title, and access them by fetching them from the stacks or asking a librarian to retrieve them. As institutions that serve the public interest, libraries adhere to standards and democratic processes to ensure consistent and familiar user experiences for patrons, but also to enable powerful search interactions such as union catalogs, where resource descriptions from multiple libraries are merged before they are offered for search. Union catalogs allow patrons to find out with a single search whether a resource is available from any library that is accessible to them.

Museums serve the public interest as well, and employ standards and democratic procedures for similar reasons as libraries, but their visitors generally look at their resources rather than borrowing them. Museums enable people to discover or experience resources by exhibiting artifacts in creative contexts, and when they implement this interaction digitally, as in a website, they vastly increase the opportunity for public access. Virtual collections are accessible to remote patrons who are unable to visit the physical museum, and they allow access to resources that are not currently on view.

The digitization of museum resources also allows visitors to experience them from a perspective that might not be possible in a physical museum. For example, in Google's Art Project, users can zoom in to view fine details of digitized paintings. Museums are starting to leverage technology and the popularity of Web 2.0 features such as tagging and social networking to attract new audiences.

Implemented in 2004, the MuseumFinland project aims to provide a portal for publishing heterogeneous museum collections on the Semantic Web. Institutions such as the Getty Information Institute and the International Committee for Documentation of the International Council of Museums have worked on standards that ensure worldwide consistency in how museums manage information about their collections.

How can these differences be handled in order to provide seamless interactions within and across organizing systems? Which requirements have to be met in order to provide the interactions that are desired? How are different interaction types implemented? Finally, how can the quality of interactions be evaluated with respect to their requirements? These are the main questions for interactions that we will try to answer.

Navigating This Chapter

This chapter concentrates on the processes that develop interactions based on leveraging the resources of organizing systems to provide valuable services to their users (human or computational agents). It will discuss the determination of the appropriate interactions (§10.2), the organization of resources for interactions (§10.3), the implementation of interactions (§10.4), and their evaluation and adaptation (§10.5). Although the fundamental questions pertain to all types of organizing systems, this chapter focuses on systems that use computers to satisfy their goals.

10.2 Determining Interactions

Creating a strategy for successfully implementing interactions involves an intricate balance between the resources, the organizing system that arranges and manages them, its producers, and its intended users or consumers. The design of interactions is driven by user requirements and their impact on the choices made in the implementation process. It is constrained by resource and technical system properties and by social and legal requirements. Determining the scope and scale of interactions requires a careful analysis of these individual factors, their combination, and the consequences thereof.

Stop and Think: Constraint vs Flexibility

Think of an information organization project you were involved in. Can you recall ways in which you were constrained in representing an idea by the organizing system the project was implemented with? In what ways was the project negatively affected by the implementation? In what ways might the constraint have had a positive effect?

It is useful to distinguish decisions that involve choices, where multiple feasible alternatives exist, from decisions that involve constraints, where design choices have been eliminated or rendered infeasible by previous ones. The goal when creating an organizing system is to make design decisions that preserve subsequent choices or that create constraints that impose design decisions that would have been preferred anyway.

10.2.1 User Requirements

Users (human or computational agents) search or navigate resources in organizing systems not just to identify them, but also to obtain and further use the selected resources (e.g., read, cluster, annotate, buy, copy, distribute, adapt, etc.). How resources are used and by whom affects how much of the resource or its

description is exposed, across which channels it is offered, and the *precision* and *accuracy* of the interaction.

An organizing system should enable interactions that allow users to achieve their goals. The more abstract and intermediated the interaction between a user and an organizing system becomes, the more precisely the requirements must be expressed. User requirements can be stated or implied, depending on the sophistication and functional capabilities of the system.

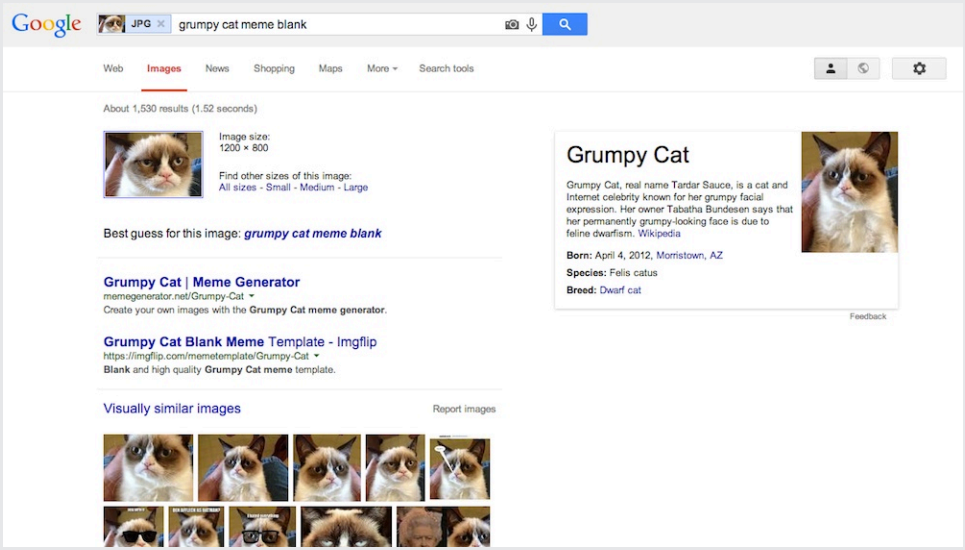
In a closet, which is a personal organizing system for physical resources, the person searching with an intent to find a particular shirt might think, "Where is my yellow Hawaiian shirt?" but does not need to communicate the search criteria to anyone else in an explicit way. In a business or institutional organizing system, however, the user needs to describe the desired resource and interact with the system to select from candidate resources. This interaction might involve a human intermediary like a salesperson or reference librarian, or a computational one like a search engine.

A user's information need usually determines the kind and content of resources required. User information needs are most often expressed in search queries (whatever is typed into a search box) or manifest themselves in the selection of one or more of the system categories that are offered for browsing. Queries can be as simple as a few keywords or very complex and specialized, employing different search fields or operators; they may even be expressed in a query language by expert users. Techniques such as spelling correction, query expansion, and suggestion assist users in formulating queries. Techniques like breadcrumb navigation and faceted filtering assist users in browsing an organizing system's category system. Some systems allow the query to be expressed in natural language and then transform it into a description that is easier for the system to process. Queries for non-textual information like photos or videos are typically expressed as text, but some systems compute descriptions from non-textual queries such as images or audio files. For example, a user can hum a tune or draw or drag an image into an image query box.

Information needs of computational agents are determined by rules and criteria set by the creators of the agents (i.e., the function or goal of the agent). When a computational agent interacts with another computational agent or service by using its API, in the ideal case its output precisely satisfies those information needs.

While search queries are explicitly stated user information needs, organizing systems increasingly attempt to solicit the user's context or larger work task in order to provide more suitable or precise interactions. Factors such as level of education, physical disabilities, location, time, or deadline pressure often specify and constrain the types of resources needed as well as the types of interactions the user is willing or able to engage in. Implicit information can be col-

Google Image Search



Google's Image Search tool can accept an image file as an input rather than text, and will find visually similar images as well as making its best guess of the image's subject matter.

(Screenshot by Ian MacFarland.)

lected from user behavior, for example, search or buying history, current user location or language, and social or collaborative behavior (other people with the same context). Methods for explicitly soliciting user requirements include observation, surveys, focus groups, interviews, work task analysis and many more.[579][IA]

Designers of organizing systems must recognize that people are not perfectly capable and rational decision makers. Limited memory and attention capacities prevent people from remembering everything and make them unable to consider more than a few things or choices at once. As a result of these fundamental limitations, people consciously and unconsciously reduce the cognitive effort they make when faced with decisions.

One important way in which this affects how people behave demonstrates what Barry Schwartz calls *The Paradox of Choice*. You might think that people would prefer many options rather than just a few because that would better enable them to select a resource that best meets their requirements. In fact, because considering more choices requires more mental effort, this can cause stress and indecision and might cause people to give up. For example, when there were 24 different types of jam offered at an upscale market, more people stopped to

Behavioral Economics

Classical economics assumes that humans are perfectly rational goal-oriented actors who act to achieve maximal satisfaction or utility. In contrast, behavioral economics recognizes the cognitive and emotional constraints on human behavior and assumes that people are biased and flawed decision makers.

Daniel Kahneman and Amos Tversky systematized the psychological foundations for behavioral economics, building on the work of Herbert Simon, who first proposed to understand people as "boundedly rational." Kahneman and Tversky identified the systematic biases that prevent people from making optimal decisions and the heuristics they use to save cognitive effort. Kahneman contrasts classical and behavioral economics as follows:

> Psychological theories of intuitive thinking cannot match the elegance and precision of formal normative models of belief and choice, but this is just another way of saying that rational models are psychologically unrealistic.
>
> — (Kahneman 2003, p 1449)

Sunstein and Thaler popularized the application of behavioral economics as "libertarian paternalism," with the goal of encouraging the design of organizing systems and policies that maintain or increase freedom of choice but which at the same time influence people to make choices that they would judge as good ones. This perspective is nicely captured by the title of their best-selling book, *Nudge.* Many government agencies and businesses in the US and elsewhere are building "nudging" principles into policies and products in the areas of social services, healthcare, and financial services because of the complexity of their offerings.

Behavioral economics complements the discipline of organizing by offering insights into the thinking and behavior of typical users that can lead to classifications and choices that make them more effective and satisfied. However, the principles of behavioral economics can be used to design organizing systems that manipulate people into taking actions and making choices that they might not intend or that are not in their best interests. (See Dark Patterns (page 112).)[582[CogSci]]

taste than when only 6 choices were offered, but a greater percentage of people who were presented a smaller number of options actually made a purchase.[580[CogSci]]

We see the same phenomenon when we compare libraries and bookstores. A rational book seeker should prefer the detailed classification system used in libraries over the very coarse BISAC system used in bookstores. However, many peo-

ple say that the detailed system makes them work too hard, leading to calls that new libraries adopt the bookstore organizing system. (See §8.3.3)

People can avoid making choices if a system proposes or pre-selects an option for them that becomes a default choice if they do nothing. Often people will make a cursory assessment about how well the option satisfies a requirement and if it is good enough they will not consider any other alternatives.[581][Law]

The study of the limits to human rationality in decision-making is the centerpiece of the discipline known as Behavioral Economics (page 495).

Organizing systems should plan for interactions based on non-purposeful user behavior. A user who does not have a particular resource need in mind might interact with an organizing system to see what it contains or to be entertained or educated. Imagine a user going to a museum to avoid the heat outside. Their requirement is to be out of the heat and—possibly—to see interesting things. A visitor to a zoo might go there to view a specific animal, but most of the time, visitors follow a more or less random path among the zoo's resources. Similarly, web surfing is random, non-information-need-driven behavior. This type of requirement cannot be satisfied by providing search capabilities alone; other interaction types (e.g., browsing, suggestions) must be provided as well.

Lastly, not all users are human beings, typing in search queries or browsing through catalogs. An organizing system should plan for interaction scenarios where computational agents access the system via APIs (application programming interfaces), which require heavily standardized access procedures and resource descriptions in order to enable interactions.

10.2.2 Socio-Political and Organizational Constraints

An important constraint for interaction design choices is the access policies imposed by the producers of organizing systems, as already described in §3.4.3 Access Policies (page 131). If resources or their descriptions are restricted, interactions may not be able to use certain properties and therefore cannot be supported.

Inter-organizational or socio-political constraints are imposed when certain parties in an interaction, or even producers of an organizing system, can exert power over other parties and therefore control the nature of the interaction (or even the nature of the resource descriptions). We can distinguish different types of constraints:

Information and economic power asymmetry
: Some organizations are able to impose their requirements for interactions and their resource description formats upon their clients or customers. For example, Google and Apple each have the power to control the extent of interoperability attainable in products, services, or applications that utilize

their numerous platforms through mandated APIs and the process by which third-party applications are approved. The asymmetry between these dominant players and the myriad of smaller entities providing peripheral support, services, or components can result in *de facto* standards that may pose significant burden for small businesses and reduce overall competition.

Standards
: Industry-wide or community standards can be essential in enabling interoperability between systems, applications, and devices. A standard interface describes the data formats and protocols to which systems should conform.[583][Com] Failure to adhere to standards complicates the merging of resources from different organizing systems. Challenges to standardization include organizational inertia; closed policies, processes, or development groups; intellectual property; credentialing; lack of specifications; competing standards; high implementation costs; lack of conformance metrics; lack of clarity or awareness; and abuse of standards as trade barriers.

Public policy
: Beyond businesses and standards-setting organizations, the government sector wields substantial influence over the implementation and success of possible interactions in organizing systems. As institutions with large and inalienable constituents, governments and governmental entities have similar influences as large businesses due to their size and substantial impact over society at large. Different forms of government around the world, ranging from centrally planned autocracy to loosely organized nation-states, can have far-reaching consequences in terms of how resource description policies are designed. Laws and regulations regarding data privacy prevent organizing systems from recording certain user data, therefore prohibiting interactions based on this information.[584][Bus]

Even within the same firm or organization, constraints on interaction design may result from contradictory policies for organizing systems or even require the implementation of separate, disjoint systems that cannot be integrated without additional investment. Siloed business functions may be resistant to the merging of resources or resource descriptions in order to gain competitive advantage or command resources over other business functions.

Stop and Think: Standards

It is easy to take standards for granted, but without them our lives would run less smoothly because many products and services would not work very well or even be dangerous to use. If you search for the phrase "ISO Standard" along with almost anything, there is a good chance that you will find something. Try "currency," "food," "sunglasses," "tea," "water," "wine," and then a few of your own.

Often characterized by different kinds of value contribution, different policies, processes, and practices, organizational units must clearly define and prioritize different interaction goals, align and coordinate processes, and build collaboration capabilities to achieve a high level of interoperability within the organizing system or between different organizing systems in the organization.

In addition to information exchange, organizational interoperability also aims to provide services that are widely available, easily identifiable, and accessible across the enterprise.

Nevertheless, inter-organizational constraints are inherently less deterministic than intra-organizational ones, because it is possible that a decision-maker with broad authority can decide that some interaction is important enough to warrant the change of institutional policies, formats, or even category systems. (See §7.2.3 Institutional Categories (page 331).)

Regulatory Constraints: Right to be Forgotten

A controversial idea known as the "right to be forgotten" gained the force of law in the European Union in May 2014 after the EU's highest court ruled that people could ask search engines such as Google, which dominates the European market, to remove certain kinds of personal information from their search results.

The ruling had its foundations in the EU's 1995 Data Protection Directive, a data retention policy crafted in a time before the dominance of the Internet and search engines. While many privacy advocates hailed it as a victory, others in the technology and media firms have decried it as censorship. Either way, it has highlighted the need for the European Commission to update and modernize its data policy; a proposal has been before the European Parliament since 2012, and plans for its adoption were underway as of summer 2014. (Source: EC fact sheet on the "right to be forgotten" ruling.)

10.3 Reorganizing Resources for Interactions

Once the scope and range of interactions is defined according to requirements and constraints, the resources and the technology of the organizing system have to be arranged to enable the implementation of the desired interactions.

Commonly, interactions are determined at the beginning of a development process of the organizing system. It follows that most required resource descriptions (which properties of a resource are documented in an organizing system) need to be clarified at the beginning of the development process as well; that is, resource descriptions are determined based on the desired interactions that an organizing system should support. Most of these processes have been described in detail in Chapter 5, Chapter 6 and Chapter 9.

Resources from different organizing systems are often aggregated to be accessed within one larger organizing system (warehouses, portals, search engines, union catalogs, cross-brand retailers), which requires resources and resource descriptions to be transformed in order to adapt to the new organizing system with its extended interaction requirements. Elsewhere, legacy systems often need to be updated to accommodate new standards, technologies, and interactions (e.g, mobile interfaces for digital libraries). That means that the necessary resources and resource descriptions for an interaction need to be identified, and, if necessary, changes have to be made in the description of the resources. Sometimes, resources are merged or transformed in order to perform new interactions.

10.3.1 Identifying and Describing Resources for Interactions

Individual and collection resource descriptions need to be carefully considered in order to record the necessary information for the designed interactions. (See Chapter 9.) The type of interaction determines whether new properties need to be derived or computed with the help of external factors and whether these properties will be represented permanently in the organizing system (e.g., an extended topical description added due to a user comment) or created on the fly whenever a transaction is executed (e.g., a frequency count).

Determining which resources or resource descriptions will be used in an interaction is simple when all resources are included (e.g., in a simple search interaction over all resources in a data warehouse). Sometimes resources need to be identified according to more selective criteria such as resources exhibiting a certain property (e.g., all restaurants in your neighborhood with four stars on Yelp in an advanced search interaction).

10.3.2 Transforming Resources for Interactions

When an organizing system and its interactions are designed with resources or resource descriptions from legacy systems with outdated formats or from multiple organizing systems or when the new organizing systems has a different purpose and requires different resource properties, resources and their descriptions need to be transformed. The processing and transformation steps required to produce the expected modification can be applied at different layers:

Infrastructure or notation transformation
: When resources are aggregated, the organizing systems must have a common basic infrastructure to communicate with one another and speak the same language. This means that participating systems must have a common set of communication protocols and an agreed upon way of representing information in digital formats, i.e., a notation (§9.3.1), such as the Unicode encoding scheme.[586][Com]

Writing system transformation
: During a writing system transformation (Chapter 9), the syntax or vocabulary—also called the data exchange format—of the resource description will be changed to conform to another model, e.g., when library records are mapped from the MARC21 standard to the Dublin Core format in order to be aggregated, or when information in a business information system is transformed into an EDI or XML format so that it can be sent to another firm.[587][Bus] Sometimes customized vocabularies are used to represent certain types of properties. These vocabularies were probably introduced to reduce errors or ambiguity or abbreviate common organizational resource properties. These customized vocabularies need to be explained and agreed upon by organizations combining resources to prevent interoperability problems.

Semantic transformation
: Agreeing on a category or classification system (Chapter 7 & Chapter 8) is crucial so that organizing systems agree semantically—that is, so that resource properties and descriptions share not only technology but also meaning. For example, because the US Census has often changed its system of race categories, it is difficult to compare data from different censuses without some semantic transformation to align the categories.[588][Com]

Resource or resource description transformation
: Resources or resource descriptions are often directly transformed, as when they are converted to another file format. In computer-based interactions like search engines, text resources are often pre-processed to remove some of the ambiguity inherent in natural language. These steps, collectively called *text processing*, include decoding, filtering, normalization, stopword elimination, and stemming. (See the sidebar, Text Processing (page 501))

Text Processing

Decoding
A digital resource is first a sequence of bits. Decoding transforms those bits into characters according to the encoding scheme used, extracting the text from its stored form. (See §9.3.1 Notations (page 462).)

Filtering
If a text is encapsulated by formatting or non-semantic markup, these characters are removed because this information is rarely used as the basis of further interactions.

Tokenization
Segments the stream of characters (in an encoding scheme, a space is also a character) into textual components, usually words. In English, a simple rule-based system can separate words using spaces. However, punctuation makes things more complicated. For example, periods at the end of sentences should be removed, but periods in numbers should not. Other languages introduce other problems for tokenization; in Chinese, a space does not mark the divisions between individual concepts.

Normalization
Normalization removes superficial differences in character sequences, for example, by transforming all capitalized characters into lower-case. More complicated normalization operations include the removal of accents, hyphens, or diacritics and merging different forms of acronyms (e.g., U.N. and UN are both normalized to UN).

Stopword elimination
Stopwords are those words in a language that occur very frequently and are not very semantically expressive. Stopwords are usually articles, pronouns, prepositions, or conjunctions. Since they occur in every text, they can be removed because they cannot distinguish them. Of course, in some cases, removing stopwords might remove semantically important phrases (e.g., “To be or not to be”).

Stemming
These processing steps normalize inflectional and derivational variations in terms, e.g., by removing the “-ed” from verbs in the past tense. This homogenization can be done by following rules (*stemming*) or by using dictionaries (*lemmatization*). Rule-based stemming algorithms are easy to implement, but can result in wrongly normalized word groups, for example when “university” and “universe” are both stemmed to “univers.”

10.3.2.1 Transforming Resources from Multiple or Legacy Organizing Systems

The traditional approach to enabling heterogeneous organizing systems to be accessed together has been to fully integrate them, which has allowed the "unrestricted sharing of data and business processes among any connected applications and data sources" in the organization.[589][Bus] This can be a strategic approach to improving the management of resources, resource descriptions, and organizing systems as a whole, especially when organizations have disparate systems and redundant information spread across different groups and departments. However, it can also be a costly approach, as integration points may be numerous, with vastly different technologies needed to get one system to integrate with another. Maintenance also becomes an issue, as changes in one system may entail changes in all systems integrating with it.[590][Bus]

Planning the transformation of resources from different organizing systems to be merged in an aggregation is called *data mapping* or alignment. In this process, aspects of the description layers (most often writing system or semantics) are compared and matched between two or more organizing systems. The relationship between each component may be unidirectional or bidirectional. In addition, resource properties and values that are semantically equivalent might have different names (the *vocabulary problem* of §4.4.2.1). The purpose of mapping may vary from allowing simple exchanges of resource descriptions, to enabling access to longitudinal data, to facilitating standardized reporting.[592][Com] The preservation of version histories of resource description elements and relations in both systems is vital for verifying the validity of the data map.

Similar to mapping, a straightforward approach to transformation is the use of *crosswalks*, which are equivalence tables that relate resource description elements, semantics, and writing systems from one organizing system to those of another.[593][Com] Crosswalks not only enable systems with different resource descriptions to interchange information in real-time, but are also used by third-party systems, such as harvesters and search engines to generate union catalogs and perform queries on multiple systems as if they were one consolidated system.[594][Com]

As the number of organizing systems increases, crosswalks and mappings become increasingly impractical if each pair of organizing systems requires a separate crosswalk. A more efficient approach would be the use of one vocabulary or format as a switching mechanism (also called a pivot or *hub language*) for all other vocabularies to map towards. Another possibility, which is often used in asymmetric power relationships between organizing systems, is to force all systems to adhere to the format that is used by the most powerful party.

10.3.2.2 Modes of Transformation

The conceptual relationships between different descriptions can be mapped out manually when creating simple maps. This, however, becomes more difficult as maps become more complex, due to the number of properties being mapped or when there are more structural or granularity issues to consider.

The use of automatic tools to create these alignments become vital in ensuring their accuracy and robustness. Graphical mapping tools provide users with a graphical user interface to connect description elements from source to target by drawing a line from one to the other.[600][Com] Other tools perform automatic mappings based on predetermined rules and criteria.[601][Com]

We often perform manual run-time transformations for decisions that require consulting more than one organizing system in our daily lives. For example, when planning a vacation, we use a variety of systems to negotiate a wide set of *ad hoc* requirements such as our resources and time, our fellow travelers and their availability, and the bookings for hotel and transportation, as well as desirable destinations and their various offerings. We somehow reconcile the different descriptions used in each of the systems and match these against each other so that the relevant information can be combined and compared. Even though the systems use different formats, vocabularies and structures, they are targeted toward human users and are relatively easy to interpret. For automatic run-time transformations, which need to be handled computationally, designers face the challenge of creating more structured processes for merging information from different systems.[602][Com]

The time of the transformation—at design time when organizing system resources are merged, or at run time when a certain interaction is performed— depends on the nature of the collaboration between organizing systems. Design-time transformations depend on highly cooperative environments where specific design requirements (like mapping rules and criteria) can be negotiated ahead of the system implementation. In cases where high-flexibility, *ad hoc* or real-time transformations would not be possible due to a lack of cooperation (such as the ShopStyle.com), run-time transformation processes may provide appropriate alternatives. Some low-level incompatibilities between organizing systems, such as the presence of syntactical, encoding, and particular structural and content issues, can also be rectified by implementing run-time transformation techniques, creating more loosely-coupled interoperating systems.

10.3.2.3 Granularity and Abstraction

Within writing system and semantic transformations, issues of *granularity* and level of abstraction (§5.3.1 Determining the Scope and Focus (page 230) and §7.4.1 Category Abstraction and Granularity (page 356)) pose the most challenges to cross-organizing system interoperability.[603][Com] *Granularity* refers to the

level of detail or precision for a specific information resource property. For instance, the postal address of a particular location might be represented as several different data items, including the number, street name, city, state, country and postal code (a high-granularity model). It might also be represented in one single line including all of the information above (a low-granularity model). While it is easy to create the complete address by aggregating the different *information components* from the high-granularity model, it is not as easy to decompose the low-granularity model into more specific information components.

This does not mean, however, that a high-granularity model is always the best choice, especially if the context of use does not require it, as there are corresponding tradeoffs in terms of efficiency and speed in assembling and processing the resource information. (See the sidebar, AccuWeather Request Granularity (page 504))

The level of abstraction is the degree to which a resource description is abstracted from the concrete use case in order to fit a wider range of resources. For example, many countries have an address field called *state*, but in some countries, a similar regional division is called *province*. In order to accommodate both concepts, we can abstract from the original concrete concepts and establish a more abstract description of *administrative region*. Granularity and abstraction differences can occur at every resource property layer when resources need to be transformed; therefore, they need to be recognized and analyzed at every layer.

AccuWeather Request Granularity

Requests for AccuWeather data have exploded in the last years, due to automated requests from mobile devices to keep weather apps updated. The company has dealt with this challenge by truncating the GPS coordinates sent by the mobile device when it requests weather data (a transformation to lower granularity). If the request with the truncated coordinates is identical to one recently made, a cached version of the content is served, resulting in 300 million to 500 million fewer requests a day.[604][Com]

10.3.2.4 Accuracy of Transformations

Automatic mapping tools can only be as accurate as the specifications and criteria that are included in the mapping guidelines. Intellectual checks and tests performed by humans are almost always necessary to validate the *accuracy* of the transformation. Because description systems vary in expressive power and complexity, challenges to transformations may arise from differences in semantic definitions, rules regarding whether an element is required or requires multiple values, hierarchical or value constraints, and controlled vocabularies. As a result of these complexities, absolute transformations that ensure exact map-

pings will result in a loss of precision if the source description system is substantially richer than the target system.

In practice, relative crosswalks where all elements in a source description are mapped to at least one target, regardless of semantic equivalence, are often implemented. This lowers the quality and *accuracy* of the mapping and can result in "down translation" or "dumbing down" of the system for resource description. As a result of mapping compromises due to different granularity or abstraction levels, transformations from different organizing systems usually result in less granular or specific resource descriptions. Consequently, whereas some interactions are now enabled (e.g., cross-organizing system search), others that were once possible can no longer be supported. For example, conflating geographical and person subject fields from one system (e.g., geographical subject = Alberta, person subject = Virginia) to a joint subject field (e.g. subject = Alberta, Virginia) to transform to the resource description of another system does not allow for searches that distinguish between these specific categories anymore.

10.4 Implementing Interactions

Stop and Think: Dumbing Down

Can you think of an example where resource description elements from one system are available for interaction in another due to a transformation, where the target system does not retain all the details of the descriptions in the source?

The next sections describe some common interactions in digital organizing systems. One way to distinguish among them is to consider the source of the algorithms that are used in order to perform them. We can mostly distinguish *information retrieval* interactions (e.g., search and browse), *machine learning* interactions (e.g., cluster, classify, extract) or *natural language processing* interactions (e.g., named entity recognition, summarization, sentiment analysis, anaphoric resolution). Another way to distinguish among interactions is to note whether resources are changed during the interaction (e.g., annotate, tag, rate, comment) or unchanged (search, cluster). Yet another way would be to distinguish interactions based on their absolute and relative complexity, i.e., on the progression of actions or steps that are needed to complete the interaction. Here, we will distinguish interactions based on the different resource description layers they act upon.

Chapter 3, *Activities in Organizing Systems*, introduced the concept of *affordance* or behavioral repertoire—the inherent actionable properties that determine what can be done with resources. We will now look at affordances (and constraints) that resource properties pose for interaction design. The interactions that an individual resource can support depend on the nature and extent

of its inherent and described properties and internal structure. However, the interactions that can be designed into an organizing system can be extended by utilizing collection properties, derived properties, and any combination thereof. These three types of resource properties can be thought of as creating layers because they build on each other.

The further an organizing system moves up the layers, the more functional capabilities are enabled and more interactions can be designed. The degree of possible interactions is determined by the extent of the properties that are organized, described, and created in an organizing system. This marks a correlation between the extent of organization and the range of possible interactions: *The more extensive the organization and the number of identifiable resource properties, the larger the universe of "affordable" interactions.*

Interactions can be distinguished by four layers:

Interactions based on properties of individual resources
: Resource properties have been described extensively in Chapter 4 and Chapter 5. Any information or property that describes the resource itself can be used to design an interaction. If a property is not described in an organizing system or does not pertain to certain resources, an interaction that needs this information cannot be implemented. For example, a retail site like Shopstyle cannot offer to reliably search by color of clothing if this property is not contained in the resource description.

Interactions based on collection properties
: Collection-based properties are created when resources are aggregated. (See Chapter 1.) An interaction that compares individual resources to a collection average (e.g., average age of publications in a library or average price of goods in a retail store) can only be implemented if the collection average is calculated.

Interactions based on derived or computed properties
: Derived or computed properties are not inherent in the resources or collections but need to be computed with the help of external information or tools. The popularity of a digital resource can be computed based on the frequency of its use, for example. This computed property could then be used to design an access interaction that searches resources based on their popularity. An important use case for derived properties is the analysis of non-textual resources like images or audio files. For these content-based interactions, intrinsic properties of the resources like color distributions are computationally derived and stored as resource properties. A search can then be performed on color distributions (e.g., a search for outdoor nature images could return resources that have a high concentration of blue in the upper half and a high concentration of green on the bottom: a meadow on a sunny day).

Interactions based on combining resources
: Combining resources and their individual, collection or derived properties can be used to design interactions based on joint properties that a single organizing system and its resources do not contain. This can lead to interactions that individual organizing systems with their particular purposes and resource descriptions cannot offer.

Whether a desired interaction can be implemented depends on the layers of resource properties that have been incorporated into the organizing system. How an interaction is implemented (especially in digital organizing systems) depends also on the algorithms and technologies available to access the resources or resource descriptions.

In our examples, we write primarily about textual resources or resource descriptions. Information retrieval of physical goods (e.g., finding a favorite cookie brand in the supermarket) or non-textual multimedia digital resources (e.g., finding images of the UC Berkeley logo) involves similar interactions, but with different algorithms and different resource properties.

10.4.1 Interactions Based on Instance Properties

Interactions in this category depend only on the properties of individual resource instances. Often, using resource properties on this lower layer coincides with basic action combinations in the interaction.

10.4.1.1 Boolean Retrieval

In a Boolean search, a query is specified by stating the information need and using operators from Boolean logic (AND, OR, NOT) to combine the components. The query is compared to individual resource properties (most often terms), where the result of the comparison is either TRUE or FALSE. The TRUE results are returned as a result of the query, and all other results are ignored. A Boolean search does not compare or rank resources so every returned resource is considered equally relevant. The advantage of the Boolean search is that the results are predictable and easy to explain. However, because the results of the Boolean model are not ranked by relevance, users have to sift through all the returned resource descriptions in order to find the most useful results.[606[Com] 607[Com]]

10.4.1.2 Tag / Annotate

A tagging or annotation interaction allows a user (either a human or a computational agent) to add information to the resource itself or the resource descriptions. A typical tagging or annotation interaction locates a resource or resource description and lets the user add their chosen resource property. The resulting changes are stored in the organizing system and can be made available for oth-

er interactions (e.g., when additional tags are used to improve the search). An interaction that adds information from users can also enhance the quality of the system and improve its usability.

10.4.2 Interactions Based on Collection Properties

Interactions in this category utilize collection-level properties in order to improve the interaction, for example, to improve the ranking in a search or to enable comparison to collection averages.

10.4.2.1 Ranked Retrieval with Vector Space or Probabilistic Models

Ranked retrieval sorts the results of a search according to their relevance with respect to the information need expressed in a query. The Vector Space and Probabilistic approaches introduced here use individual resource properties like term occurrence or term frequency in a resource and collection averages of terms and their frequencies to calculate the rank of a resource for a query.[609][Com]

The simplicity of the Boolean model makes it easy to understand and implement, but its binary notion of relevance does not fit our intuition that terms differ in how much they suggest what a document is about. Gerard Salton invented the vector space model of information retrieval to enable a continuous measure of relevance.[610][Com] In the vector space model, each resource and query in an organizing system is represented as a vector of terms. Resources and queries are compared by comparing the directions of vectors in an n-dimensional space (as many dimensions as terms in the collection), with the assumption is that "closeness in space" means "closeness in meaning."

In contrast to the vector space model, the underlying idea of the probabilistic model is that given a query and a resource or resource description (most often a text), probability theory is used to estimate how likely it is that a resource is relevant to an information need. A probabilistic model returns a list of resources that are ranked by their estimated *probability of relevance* with respect to the information need so that the resource with the highest probability to be relevant is ranked highest. In the vector space model, by comparison, the resource whose term vector is most similar to a query term vector (based on frequency counts) is ranked highest.[611][Com]

Both models utilize an intrinsic resource property called the *term frequency (tf)*. For each term, term frequency (tf) measures how many times the term appears in a resource. It is intuitive that *term frequency* itself has an ability to summarize a resource. If a term such as "automobile" appears frequently in a resource, we can assume that one of the topics discussed in the resource is automobiles and that a query for "automobile" should retrieve this resource. Another problem with the term frequency measure occurs when resource descriptions have

different lengths (a very common occurrence in organizing systems). In order to compensate for different resource description lengths that would bias the term frequency count and the calculated relevance towards longer documents, the length of the term vectors are normalized as a percentage of the description length rather than a raw count.

Relying solely on term frequency to determine the relevance of a resource for a query has a drawback: if a term occurs in all resources in a collection it cannot distinguish resources. For example, if every resource discusses automobiles, all resources are potentially relevant for an "automobile" query. Hence, there should be an additional mechanism that penalize a term appearing in too many resources. This is done with *inverse document frequency*, which signals how often a term or property occurs in a collection.

Inverse document frequency (idf) is a collection-level property. The *document frequency* (df) is the number of resources containing a particular term. The inverse document frequency (idf) for a term is defined as $idf_t = \log(N/df_t)$, where N is the total number of documents. The inverse document frequency of a term decreases the more documents contain the term, providing a discriminating factor for the importance of terms in a query. For example, in a collection containing resources about automobiles, an information retrieval interaction can handle a query for "automobile accident" by lowering the importance of "automobile" and increasing the importance of "accident" in the resources that are selected as result set.

As a first step of a search, resource descriptions are compared with the terms in the query. In the vector space model, a metric for calculating similarities between resource description and query vectors combining the term frequency and the inverse document frequency is used to rank resources according to their relevance with respect to the query.[612][Com]

The probability ranking principle is mathematically and theoretically better motivated than the vector space ranking principle. However, multiple methods have been proposed to estimate the probability of relevance. Well-known probabilistic retrieval methods are Okapi BM25, *language models (LM)* and *divergence from randomness models (DFR)*.[613][Com] Although these models vary in their estimations of the probability of relevance for a given resource and differ in their mathematical complexity, intrinsic properties of resources like term frequency and collection-level properties like inverse document frequency and others are used for these calculations.

10.4.2.2 Synonym Expansion with Latent Semantic Indexing

Latent semantic indexing is a variation of the vector space model where a mathematical technique known as singular value decomposition is used to combine similar term vectors into a smaller number of vectors that describe their "statis-

tical center." [614][Com] This method is based mostly on collection-level properties like co-occurrence of terms in a collection. Based on the terms that occur in all resources in a collection, the method calculates which terms might be synonyms of each other or otherwise related. Put another way, latent semantic indexing groups terms into topics. Let us say the terms "roses" and "flowers" often occur together in the resources of a particular collection. The latent semantic indexing methodology recognizes statistically that these terms are related, and replaces the representations of the "roses" and "flower" terms with a computed "latent semantic" term that captures the fact that they are related, reducing the dimensionality of resource description (see §5.3.4.4 Vocabulary Control as Dimensionality Reduction (page 250)). Since queries are translated into the same set of components, a query for "roses" will also retrieve resources that mention "flower." This increases the chance of a resource being found relevant to a query even if the query terms do not match the resource description terms exactly; the technique can therefore improve the quality of search.

Latent semantic indexing has been shown to be a practical technique for estimating the substitutability or semantic equivalence of words in larger text segments, which makes it effective in information retrieval, text categorization, and other NLP applications like question answering. In addition, some people view it as a model of the computational processes and representations underlying substantial portions of how knowledge is acquired and used, because latent semantic analysis techniques produces measures of word-word, word-passage, and passage-passage relations that correlate well with human cognitive judgments and phenomena involving association or semantic similarity. These situations include vocabulary tests, rate of vocabulary learning, essay tests, prose recall, and analogical reasoning.[615][CogSci]

Another approach for increasing the quality of search is to add similar terms or properties to a query from a controlled vocabulary or classification system. When a query can be mapped to terms in the controlled vocabulary or classes in the classification, the inherent semantic structure of the vocabulary or classification can suggest additional terms (broader, narrower, synonymous) whose occurrence in resources can signal their relevance for a query.

10.4.2.3 Structure-Based Retrieval

When the internal structure of a resource is represented in its resource description a search interaction can use the structure to retrieve more specific parts of a resource. This enables parametric or zone searching, where a particular component or resource property can be searched while all other properties are disregarded.[616][Com] For example, a search for "Shakespeare" in the title field in a bibliographic organizing system will only retrieve books with Shakespeare in the title, not as an author. Because all resources use the same structure, this structure is a collection-level property.

A common structure-based retrieval technique is the search in relational databases with Structured Query Language (SQL). With the help of tools to facilitate selection and transformation, particular tables and fields in tables and in many combination or with various constraints can be applied to yield highly precise results.

A format like XML enables structured resource descriptions and is therefore very suitable for search and for structured navigation and retrieval. XPath (see §6.5.2) describes how individual parts of XML documents can be reached within the internal structure. *XML Query Language (XQuery)*, a structure-based retrieval language for XML, executes queries that can fulfill both topical and structural constraints in XML documents. For example, a query can be expressed for documents containing the word “apple” in text, and where “apple” is also mentioned in a title or subtitle, or in a glossary of terms.

10.4.2.4 Clustering / Classification

Clustering (§7.5.3.3) and computational classification (§8.7) are both interactions that use individual and collection-level resource properties to execute their operation. During clustering (unsupervised learning), all resources are compared and grouped with respect to their similarity to each other. During computational classification (supervised learning), an individual resource or a group of resources is compared to a given classification or *controlled vocabulary* in an organizing system and the resource is assigned to the most similar class or descriptor. Another example for a classification interaction is spam detection. (See §8.7.) Author identification or characterization algorithms attempt to determine the author of a given work (a classification interaction) or to characterize the type of author that has or should write a work (a clustering interaction).

10.4.3 Interactions Based on Derived Properties

Interactions in this category derive or compute properties or features that are not inherent to the resources themselves or the collection. External data sources, services, and tools are employed to support these interactions. Building interactions with conditionality based on externally derived properties usually increases the quality of the interactions by increasing the system’s context awareness.

10.4.3.1 Popularity-Based Retrieval

Google’s PageRank (see §6.5.3) is the most well-known popularity measure for websites.[617][Web] The basic idea of PageRank is that a website is as popular as the number of links referencing the website. The actual calculation of a website’s PageRank involves more sophisticated mathematics than counting the number

Retail Store Activity Tracking

Retail analytics companies use cameras and other sensors to analyze shopper activity in retail stores and generate heatmaps of which areas see the most foot traffic and which items customers interact with most.

(Photo by Flickr user m01229. Creative Commons license. Illustration of heatmap by Ian MacFarland.)

of in-links, because the source of links is also important. Links that come from quality websites contribute more to a website's PageRank than other links, and links to qualitatively low websites will hurt a website's PageRank.

An information retrieval model for web pages can now use PageRank to determine the value of a web page with respect to a query. Google and other web search engines use many different ranking features to determine the final rank of a web page for any search, PageRank as a popularity measure is only one of them.

Other popularity measures can be used to rank resources. For example, frequency of use, buying frequency for retail goods, the number of laundry cycles a particular piece of clothing has gone through, and even whether it is due for a laundry cycle right now.

10.4.3.2 Citation-Based Retrieval

Citation-based retrieval is a sophisticated and highly effective technique employed within bibliographic information systems. Bibliographic resources are linked to each other by citations, that is, when one publication cites another. When a bibliographic resource is referenced by another resource, those two resources are probably thematically related. The idea of citation-based search is to use a known resource as the information need and retrieve other resources that are related by citation.

Citation-based search can be implemented by directly following citations from the original resource or to find resources that cite the original resource. Another comparison technique is the principle of bibliographic coupling, where the information retrieval system looks for other resources that cite the same resources as the original resource. Citation-based search results can also be ranked,

for example, by the number of in-citations a publication has received (the PageRank popularity measure actually derives from this principle).

10.4.3.3 Translation

During translation, resources are transformed into another language, with varying degrees of success. In contrast to the transformations that are performed in order to merge resources from different organizing systems to prepare them for further interactions, a translation transforms the resource after it has been retrieved or located. Dictionaries or parallel corpora are external resources that drive a translation.

During a dictionary-based translation, every individual term (sometimes phrases) in a resource description is looked up in a dictionary and replaced with the most likely translation. This is a simple translation, as it cannot take grammatical sentence structures or context into account. Context can have an important impact on the most likely translation: the French word *avocat* should be translated into *lawyer* in most organizing systems, but probably not in a cookbook collection, in which it is the *avocado* fruit.

Parallel corpora are a way to overcome many of these challenges. Parallel corpora are the same or similar texts in different languages. The Bible or the protocols of United Nations (UN) meetings are popular examples because they exist in parallel in many different languages. A machine learning algorithm can learn from these corpora to derive which phrases and other grammatical structures can be translated in which contexts. This knowledge can then be applied to further resource translation interactions.

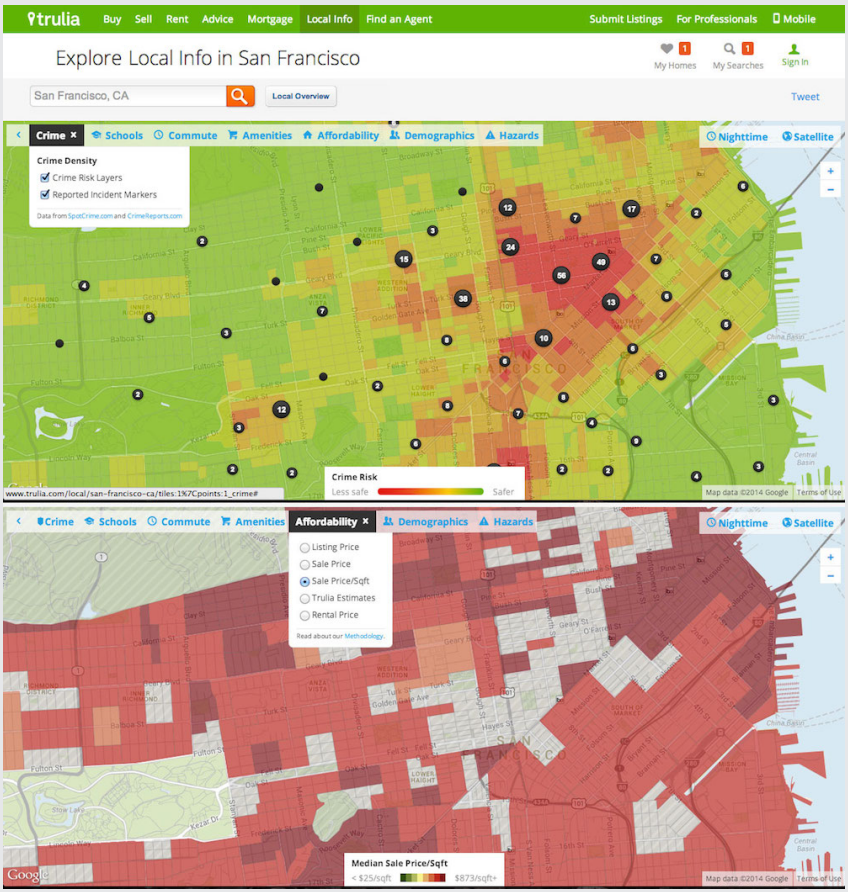
10.4.4 Interactions Based on Combining Resources

Interactions in this category combine resources mostly from different organizing systems to provide services that a single organizing system could not enable. Sometimes different organizing systems with related resources are created on purpose in order to protect the privacy of personal information or to protect business interests. Releasing organizing systems to the public can have unwanted consequences when clever developers detect the potential of connecting previously unrelated data sources.

10.4.4.1 Mash-Ups

A mash-up combines data from several resources, which enables an interaction to present new information that arises from the combination.[618][Web] For example, housing advertisements have been combined with crime statistics on maps to graphically identify rentals that are available in relatively safe neighborhoods.

Mash-up of Housing and Crime Stats



The "Local Info" map on real-estate website Trulia mashes up data on crime, schools, housing prices, commute times, and other factors relevant to people searching for a new place to live.

(Screenshots by Ian MacFarland.)

Mash-ups are usually *ad-hoc* combinations at the resource level and therefore do not impact the "mashed-up" organizing systems' internal structures or vocabularies; they can be an efficient instrument for rapid prototyping on the web. On the other hand, that makes them not very reliable or robust, because a mash-up can fail in its operation as soon as the underlying organizing systems change.

10.4.4.2 Linked Data Retrieval and Resource Discovery

In §9.4.3 The Semantic Web World (page 476), linked data relates resources among different organizing system technologies via standardized and unique identifiers (URIs). This simple approach connects resources from different systems with each other so that a cross-system search is possible.[619][Web] For example, two different online retailers selling a Martha Stewart bedspread can link to a website describing the bedspread on the Martha Stewart website. Both retailers use the same unique identifier for the bedspread, which leads back to the Martha Stewart site.

Resource discovery or linked data retrieval are search interactions that traverse the network (or semantic web graph) via connecting links in order to discover semantically related resources. A search interaction could therefore use the link from one retailer to the Martha Stewart website to discover the other retailer, which might have a cheaper or more convenient offer.

10.5 Evaluating Interactions

Managing the quality of an interaction with respect to its intent or goal is a crucial part of every step from design through implementation and especially during operation. Evaluating the quality of interactions at different times in the design process (design concept, prototype, implementation, and operation) reveals both strengths and weaknesses to the designers or operators of the organizing system.

During the design and implementation stages, interactions need to be tested against the original goals of the interaction and the constraints that are imposed by the organizing system, its resources and external conditions. It is very common for processes in interactions to be tweaked or tuned to better comply with the original goals and intentions for the interaction. Evaluation during these stages often attempts to provide a calculable way to measure this compliance and supports the fine-tuning process. It should be an integral part of an iterative design process.

During the later implementation and operation stages, interactions are evaluated with respect to the dynamically changing conditions of the organizing system and its environment. User expectations as well as environmental conditions or constraints can change and need to be checked periodically. A systematic evaluation of interactions ensures that changes that affect an interaction are observed early and can be integrated in order to adjust and even improve the interaction. At these stages, more subjective evaluation aspects like satisfaction, experience, reputation, or “feel” also play a role in fine-tuning the interactions. This subjective part of the evaluation process is as important as the quantitative, objective part. Many factors during the design and implementation pro-

cesses need to be considered and made to work together. Ongoing quality evaluation and feedback ensures that interactions work as intended.

Evaluation aspects can be distinguished in numerous ways: by the effort and time to perform them (both data collection and analysis); by how quantifiable they are or how comparable they are with measures in other organization systems; by what component of the interaction or organizing system they focus on; or by the discipline, expertise, or methodologies that are used for the evaluation.

A common and important distinction is the difference between efficiency, effectiveness, and satisfaction. An interaction is efficient when it performs its actions in a timely and economical manner, effective when it performs its actions correctly and completely, satisfactory when it performs as expected. Satisfaction is the least quantifiable of the evaluation aspects because it is highly dependent on individual tastes and experiences.

Let us assume that Shopstyle.com develops a new interaction that lets you compare coat lengths from the offerings of their various retailers. Once the interaction is designed, an evaluation takes place in order to determine whether all coats and their lengths are integrated in the interaction and whether the coat lengths are measured and compared correctly. The designers would not only want to know whether the coat lengths are represented correctly but also whether the interaction performs efficiently. When the interaction is ready to be released (usually first in beta or test status), users and retailers will be asked whether the interaction improves their shopping experience, whether the comparison performs as they expected, and what they would change. These evaluation styles work hand in hand in order to improve the interaction.

10.5.1 Efficiency

When evaluating the efficiency of an organizing system, we focus on the time, energy and economic resources needed in order to achieve the interaction goals of the system. Commonly, the fewer resources are needed for achieving a successful interaction, the more efficient the interaction.

Efficiency measures are usually related to engineering aspects such as the time to perform an action, number of steps to perform an interaction, or amount of computing resources used. Efficiency with respect to the human costs of memory load, attention, and cognitive processing is also important if there is to be a seamless user experience where users can interact with the system in a timely manner.

For a lot of organizing system interactions, however, effectiveness is the more important aspect, particularly for those interactions that we have looked at so far. If search results are not correct, then users will not be satisfied by even the

most usable interface. Many interactions are evaluated with respect to their ability to return relevant resources. Why and how this is evaluated is the focus of the remainder of this section.

10.5.2 Effectiveness

Effectiveness evaluates the correct output or results of an interaction. An effective interaction achieves relevant, intended or expected results. The concept of *relevance* and its relationship to effectiveness is pivotal in information retrieval and machine learning interactions. (§10.5.2.1) Effectiveness measures are often developed in the fields that developed the algorithm for the interaction, information retrieval, or machine learning. *Precision* and *recall* are the fundamental measures of *relevance* or effectiveness in information retrieval or machine learning interactions. (§10.5.2.2)

10.5.2.1 Relevance

Relevance is widely regarded as the fundamental concept of information retrieval, and by extension, all of information science. Despite being one of the more intuitive concepts in human communication, relevance is notoriously difficult to define and has been the subject of much debate over the past century.

Historically, relevance has been addressed in logic and philosophy since the notion of inference was codified (to infer B from A, A must be relevant to B). Other fields have attempted to deal with relevance as well: sociology, linguistics, and psychology in particular. The subject knowledge view, subject literature view, logical view, system's view, destination's view, pertinence view, pragmatic view and the utility-theoretic interpretation are different perspectives on the question of when something is relevant. In 1997, Mizzaro surveyed 160 research articles on the topic of relevance and arrived at this definition: "relevance can be seen as a point in a four-dimensional space, the values of each of the four dimensions being: (*i*) Surrogate, document, information; (*ii*) query, request, information need, problem; (*iii*) topic, context, and each combination of them; and (*iv*) the various time instants from the arising of problem until its solution." This rather abstract definition points to the terminological ambiguity surrounding the concept.

For the purpose of organizing systems, relevance is a concept for evaluating effectiveness that describes whether a stated or implicit information need is satisfied in a particular user context and at a particular time. One of the challenges for the evaluation of relevance in organizing systems is the gap between a user's information need (often not directly stated), and an expression of that information need (a query). This gap might result in ambiguous results in the interaction. For example, suppose somebody speaks the word "Paris" (query) into a smart phone application seeking advice on how to travel to Paris, France. The

response includes offers for the Paris Hotel in Las Vegas. Does the result satisfy the information need? What if the searcher receives advice on Paris but has already seen every one of the resources the organizing system offers? What is the correct decision on relevance here?

The key to calculating effectiveness is to be aware of what is being measured. If the information need as expressed in the query is measured, the topical relevance or topicality—a system-side perspective is analyzed. If the information need as in a person's mind is measured, the pertinence, utility, or situational relevance—a subjective, personal perspective is analyzed. This juxtaposition is the point of much research and contention in the field of information retrieval, because topical relevance is objectively measurable, but subjective relevance is the real goal. In order to evaluate relevance in any interaction, an essential prerequisite is deciding which of these notions of relevance to consider.

10.5.2.2 The Recall / Precision Tradeoff

Precision measures the *accuracy* of a result set, that is, how many of the retrieved resources for a query are relevant. *Recall* measures the completeness of the result set, that is, how many of the relevant resources in a collection were retrieved. Let us assume that a collection contains 20 relevant resources for a query. A retrieval interaction retrieves 10 resources in a result set, 5 of the retrieved resources are relevant. The precision of this interaction is 50% (5 out of 10 retrieved resources are relevant); the recall is 25% (5 out of 20 relevant resources were retrieved).[622][Com]

It is in the nature of information retrieval interactions that recall and precision trade off with each other. To find all relevant resources in a collection, the interaction has to cast a wide net and will not be very precise. In order to be very precise and return only relevant resources to the searcher, an interaction has to be very discriminating and will probably not find all relevant resources. When a collection is very large and contains many relevant resources for any given query, the priority is usually to increase precision. However, when a collection is small or the information need also requires finding all relevant documents (e.g., in case law, patent searches, or medical diagnosis support), then the priority is put on increasing recall.

The completeness and granularity of the organizing principles in an organizing system have a large impact on the trade-off between *recall* and *precision*. (See Chapter 4.) When resources are organized in fine-grained category systems and many different resource properties are described, high-precision searches are possible because a desired resource can be searched as precisely as the description or organization of the system allows. However, very specialized description and organization may preclude certain resources from being found; consequently, recall might be sacrificed. If the organization is superficial—like

your sock drawer, for example—you can find all the socks you want (high recall) but you have to sort through a lot of socks to find the right pair (low precision). The trade-off between recall and precision is closely associated with the extent of the organization.

10.5.3 Satisfaction

Satisfaction evaluates the opinion, experience or attitude of a user towards an interaction. Because satisfaction depends on individual user opinions, it is difficult to quantify. Satisfaction measures arise out of the user's experience with the interaction—they are mostly aspects of user interfaces, usability, or subjective and aesthetic impressions.

Usability measures whether the interaction and the user interface designed for it correspond with the user's expectations of how they should function. It particularly focuses on the usefulness of the interaction. Usability analyzes ease-of-use, learnability, and cognitive effort to measure how well users can use an interaction to achieve their task.

Although efficiency, effectiveness, and satisfaction are measured differently and affect different components of the interaction, they are equally important for the success of an interaction. Even if an interaction is fast, it is not very useful if it arrives at incorrect results. Even if an interaction works correctly, user satisfaction is not guaranteed. One of the challenges in designing interactions is that these factors invariably involve tradeoffs. A fast system cannot be as precise as one that prioritizes the use of contextual information. An effective interaction might require a lot of effort from the user, which does not make it very easy to use, so the user satisfaction might decrease. The priorities of the organizing system and its designers will determine which properties to optimize.

Let us continue our Shopstyle coat-length comparison interaction example. When the coat length calculation is performed in an acceptable amount of time and does not consume a lot of the organizing systems resources, the interaction is efficient. When all coat lengths are correctly measured and compared, the interaction is effective. When the interaction is seamlessly integrated into the shopping process, visually supported in the interface, and not cognitively exhausting, is it probably satisfactory for a user, as it provides a useful service (especially for someone with irregular body dimensions). What aspect should Shopstyle prioritize? It will probably weigh the consequences of effectiveness versus efficiency and satisfaction. For a retail- and consumer-oriented organizing system, satisfaction is probably one of the more important aspects, so it is highly likely that efficiency and effectiveness might be sacrificed (in moderation) in favor of satisfaction.

10.6 Key Points in Chapter Ten

- Interactions arise naturally from the affordances of resources or are purposefully designed into organizing systems.

  (See §10.1 Introduction (page 487))

- Accessing and merging resources are fundamental interactions that occur in almost every organizing system.

  (See §10.1 Introduction (page 487))

- User requirements, which layer of resource properties is used, and the legal, social and organizational environment can distinguish interactions.

  (See §10.2 Determining Interactions (page 492))

- Limited memory and attention capacities prevent people from remembering everything and make them unable to consider more than a few things or choices at once.

  (See §10.2.1 User Requirements (page 492))

- The principles of behavioral economics can be used to design organizing systems that manipulate people into taking actions and making choices that they might not intend or that are not in their best interests.

  (See the sidebar, Behavioral Economics (page 495))

- In order to enable interactions, it is necessary to identify, describe, and sometimes transform the resources in an organizing system.

  (See §10.3.1 Identifying and Describing Resources for Interactions (page 499))

- Similar to mapping, a straightforward approach to transformation is the use of *crosswalks*, which are equivalence tables that relate resource description elements, semantics, and writing systems from one organizing system to those of another.

  (See §10.3.2.2 Modes of Transformation (page 503))

- Merging transformations can be distinguished by type (mapping or crosswalk), time (design time or run time) and mode (manual or automatic).

  (See §10.3.2.3 Granularity and Abstraction (page 503))

- Implementations can be distinguished by the source of the algorithm (information retrieval, machine learning, natural language processing), by their complexity (number of actions needed), by whether resources are changed, or by the resource description layers they are based on.

  (See §10.4 Implementing Interactions (page 505))

- Important aspects for the evaluation of interactions are efficiency (timeliness and cost-effectiveness), effectiveness (accuracy and relevance) and satisfaction (positive attitude of the user).

  (See §10.5 Evaluating Interactions (page 515))

- The concept of *relevance* and its relationship to effectiveness is pivotal in information retrieval and machine learning interactions.

  (See §10.5.2.1 Relevance (page 517))

- The trade-off between recall and precision decides whether a search finds all relevant documents (high recall) or only relevant documents (high precision).

  (See §10.5.2.2 The Recall / Precision Tradeoff (page 518))

- The extent of the organization principles also impacts recall and precision: more fine-grained organization allows for more precise interactions.

  (See §10.5.2.2 The Recall / Precision Tradeoff (page 518))

Endnotes for Chapter 10

[572][Bus] Walmart uses its market power to impose technology and process decisions on its suppliers and partners. See (Fishman 2003), (Grean and Shaw 2005), (Wilbert 2006). Walmart's website for suppliers is http://walmartstores.com/Suppliers/248.aspx/

[573][Bus] In order to more easily use and reuse content, as well as have the ability to integrate different learning tools into a single *Learning Management System (LMS)*, Global Learning Consortium, an organization composed of 140 members from leading educational institutions and education-related companies, has released specifications to make this possible. Called *Common Cartridge and Learning Tools Interoperability* (http://www.imsglobal.org/commoncartridge.html), the specifications provide a common format and guidelines to construct tools and create content that can be easily imported into learning management systems. Common Cartridge (CC) specifications give detailed descriptions of the directory structure, metadata and information models associated with a particular learning object. For example, a learning package from a provider from McGraw-Hill may contain content from a book, some interactive quizzes, and some multimedia to support the text. CC specifies how files would be organized within a directory, how links would be represented, how the package would communicate with a backend server, how to describe each of the components, and the like. This would enable a professor or a student using any capable learning management system to import a "cartridge" or learning material and have it appear in a consistent manner with all other learning materials within the LMS. This means

that content providers need not maintain multiple versions of the same content just to conform to the formats of different systems, allowing them to focus their resources on creating more content as opposed to maintaining the ones they already have. Looking at this in the context of the interoperability framework, we see that while information from providers are in a structured digital form, the main problem was that users were consuming the content using competing systems that had their own data formats by which to accept content. Huge publishers, wanting to increase distribution of their product, offered their content in all these different formats. While the specifications that the LMS created refer to the technical considerations in creating content and tools, the process of getting to that point involved a lot of organizational and political discussions. Internally, content and LMS providers needed to set aside the necessary resources to refactor their products to conform to the standards. Externally, competing providers had to collaborate with one another to create the specifications.

[579][IA] A conceptual framework for analyzing users and their work tasks for design requirements is (Fidel 2012). A general survey of design methods is given in (Hanington and Martin 2012). Designing particularly for successful interactions (services) is discussed in (Polaine, Løvlie and Reason). (Resmini and Rosati 2011) describe designing for engaged users using cross-channel, cross-media information architecture.

[582][CogSci] (Simon 1982), (Tversky and Kahneman 1974), (Kahneman and Tversky 1979), (Kahneman 2003), and (Thaler 2008).

[580][CogSci] (Schwartz 2005) and (Iyengar 2000)

[581][Law] A comprehensive review of the power of defaults in software and other technical systems from the perspectives of law, computer science, and behavioral economics is (Kesan and Shah 2006)

[583][Com] (von Riegen 2006).

[584][Bus] A good example for the importance of standards and interoperability rules is E-government. E-government refers to the ability to deliver government services through electronic means. These services can range from government-to-citizen, government-to-business, government-to-employees, government-to-government, and vice-versa (Guijarro 2007), (Scholl 2007). This could range from a government unit providing a portal where citizens can apply for a driver's license or file their taxes, to more complex implementations such as allowing different government agencies to share certain pertinent information with one another, such as providing information on driver's license holders to the police. Because the government interacts with heterogeneous entities and their various systems, e-government planners must consider how to integrate and interoperate with different systems and data models. Countries belonging to the

Organization for Economic Cooperation and Development (OECD) have continuously refined their strategies for e-government.

An example of a highly successful implementation of a business-to-government implementation is the use of the Universal Business Language (UBL) by the Government of Denmark. UBL is a "royalty-free library of standard electronic XML business documents such as purchase orders and invoices" [oasis-open.org]. The Government of Denmark localized these standards, and mandated all organizations wanting to do business with the government to use these formats for invoicing. By automating the matching process between an electronic order and an electronic invoice, the government expects total potential savings of about 160 million Euros per year [UBL case study], thus highlighting the need for a standard format by which businesses can send in orders and invoices electronically.

Recognizing that its position as government entails that all types of suppliers, big or small, must have equal opportunity to sell its products and services, the Government of Denmark not only set data format standards, it also gave several options by which information can be exchanged. Paper-based invoices would be sent to scanning agencies that would scan and create electronic versions to be submitted to the government. This highlights the different organizational and consumption issues that the government of Denmark had to consider when designing the system.

[586][Com] While data encoding describes how information is represented, and data exchange formats describe how information is structured, communication protocols refer to how information is exchanged between systems. These protocols dictate how these documents are enclosed within messages, and how these messages are transmitted across the network. Things such as message format, error detection and reporting, security and encryption are described and considered. Nowadays, there are a number of communication protocols that are used over networks, including *File Transfer Protocol (FTP)*, Hypertext Transfer Protocol (HTTP) commonly used in the Internet, *Post Office Protocol (POP)* commonly used for e-mail, and other protocols under the *Transmission Control Protocol/Internet Protocol (TCP/IP)* suite. Different product manufacturers normally also have more proprietary protocols that they employ, including Apple Computer Protocols Suite and Cisco Protocols. In addition, different types of networks would also have corresponding protocols, including Mobile Wireless Protocols and such.

[587][Bus] Electronic Data Interchange (EDI), is used to exchange formatted messages between computers or systems. Organizations use this format to conduct business transactions electronically without human intervention, such as in sending and receiving purchase orders or exchange invoice information and such. There are four main standards that have been developed for EDI, includ-

ing the *UN/EDIFACT* standard recommended by the *United Nations (UN)*, *ANSI ASC X12* standard widely used in the US, TRADACOMS standard that is widely used in the UK, and the ODETTE standard used in the European automotive industry. These standards include formats for a wide range of business activities, such as shipping notices, fund transfers, and the like. EDI messages are highly formatted, with the meaning of the information being transmitted being highly dependent on its position in the document. For instance, a line in an EDI document with `BEG*00*NE*MOG009364501**950910*CSW11096^` corresponds to a line in the X12 standard for Purchase Orders (standard 850). "BEG" specifies the start of a Purchase Order Transaction Set. The asterisk (*) symbol delineates between items in the line, with each value corresponding to a particular field or information component described in the standard. "NE," for example, corresponds to the Purchase Order Type Code, which in this instance is "New Order." As can be seen in the example, the description of the information being transmitted is not readily available within the document. Instead, parties exchanging the information must agree on these formats beforehand, and need to ensure that the information instance is at the right position within the document so that the receiving party can correctly interpret it.

*EDI samples come from `http://miscouncil.org`.

American National Standards Association (ANSI) can be found at `http://www.ansi.org`.

[588][Com] This and more examples for difficult categorizations can be found in: (Bowker and Star 2000).

[589][Bus] (Linthicum 1999).

[590][Bus] Allowing unrestricted access to data and business processes also becomes a problem when working across organizations. Fully integrating systems between two companies, for instance, may entail the exposure of business intelligence and information that should be kept private. This type of exposure is too much for most businesses, regardless of whether the relationship with the other business is collaborative rather than competitive. There are security issues to be considered, as collaborating organizations would need to access private networks and secure servers. The heterogeneity in supporting organizing systems along with the need to quickly evolve with the rapid changes in an organization's competitive and collaborative environment has pushed organizations to shift from more vertical, isolated structures to a more loosely coupled, ecosystem paradigm This has led to more componentized and modularized systems that need only to exchange information or transform resources when an interaction requires it.

The emerging paradigm then is to enable independent systems to interoperate, or to have "the ability of two or more systems or components to exchange infor-

mation and to use the information that has been exchanged." Because the focus is in the exchange of resources or resource descriptions, independent systems need not necessarily know other systems' underlying logic or implementation, for example, how they store resources. What is important is knowing what kind of resource is expected and in what format (notation, writing system, semantics), and what kind of information is returned for a particular. This is a strategic approach to exchanging resources, as systems can remain highly independent of each other. Changes in one system need not necessarily affect how other systems work as long as the information that is sent and received through an interface stays the same. This allows greater adaptability, as changes to system logic or business processes can be done in self-contained modules without necessarily affecting others. The transformation then happens in an intermediate space where the agreements on resource descriptions are fixed.

[592][Com] (McBride et al. 2006).

[593][Com] (NISO 2004).

[594][Com] http://journal.code4lib.org/articles/54 (Section 1.), http://www.dlib.org/dlib/june06/chan/06chan.html.

[600][Com] More commonly, graphical data mapping tools are included in an extract, transform, and load (ETL) database suite that provides additional powerful data transformation capabilities. Whereas data mapping is the first step in capturing the relationships between different systems, data transformation entails code generation that uses the resulting maps to produce an executable transformational program that converts the source data into target format. ETL databases *extract* the information needed from the outside sources, *transform* these into information that can be used by the target system using the necessary data mappings, and then *loads* it into the end system.

[601][Com] Languages such as XSLT and *Turing eXtender Language (TXL)* facilitate the ease of data transformation while various commercial data warehousing tools provide varying functionalities such as single/multiple source acquisition, data cleansing, and statistical and analytical capabilities. Based on XML, XSLT is a declarative language designed for transforming XML documents into other documents. For example, XSLT can be used to convert XML data into HTML documents for web display or PDF for print or screen display. XSLT processing entails taking an input document in XML format and one or more XSLT style sheets through a template-processing engine to produce a new document.

[602][Com] (Carney et al. 2005).

[603][Com] For an in-depth discussion of interoperability challenges, see Chapter 6 of (Glushko and McGrath 2005).

[604][Com] (AT&T 2011).

[606][Com] Each of the four information retrieval models discussed in the chapter has different combinations of the comparing, ranking, and location activities. Boolean and vector space models compare the description of the information need with the description of the information resource. Vector space and probabilistic models rank the information resource in the order that the resource can satisfy the user's query. Structure-based search locates information using internal or external structure of the information resource.

[607][Com] Our discussion of information retrieval models in this chapter does not attempt to address information retrieval at the level of theoretical and technical detail that informs work and research in this field (Manning et al. 2008), (Croft et al. 2009). Instead, our goal is to introduce IR from a more conceptual perspective, highlighting its core topics and problems using the vocabulary and principles of IO as much as possible.

[609][Com] (Manning et al. 2008), Ch. 1.

[610][Com] Salton was generally viewed as the leading researcher in information retrieval for the last part of the 20th century until he died in 1995. The vector model was first described in (Salton, Wong, and Yang 1975).

[611][Com] (Manning et al. 2008), p.221.

[612][Com] See (Manning et al. 2008), Ch. 6 for more explanations and references on the vector space model.

[613][Com] See (Robertson 2005), (Manning et al. 2008), Ch. 12 for more explanations and references.

[614][Com] (Deerwester, Dumais et al. 1990).

[615][CogSci] See (Dumais 2003).

[616][Com] (Manning et al. 2008), Section 6.1.

[617][Web] (Page et al. 1999).

[618][Web] (Yee 2008).

[619][Web] (Bizer, Heath, and Berners-Lee 2009).

[622][Com] Recall and precision are only the foundation of measures that have been developed in information retrieval to evaluate the effectiveness of search algorithms. See (Baeza-Yates and Ribeiro 2011), (Manning et al. 2008) Ch. 8; (Demartini and Mizzaro 2006).

Chapter 11

The Organizing System Roadmap

Robert J. Glushko

11.1 Introduction

Chapter 1 defined an organizing system as "an intentionally arranged collection of resources and the interactions they support." An organizing system emerges as the result of decisions about **what** is organized, **why** it is organized, **how much** it is organized, **when** it is organized, and **how or by whom** it is organized. These decisions and the tradeoffs they embody are manifested in the four common activities of organizing systems—selecting resources, organizing them, designing and supporting interactions with them, and maintaining them—which we described in Chapter 2. Chapters 4-10 progressively explained each of the parts of the organizing system: resources, resource descriptions, resource categories and collections, and interactions with resources—introducing additional concepts and methods associated with each of these parts.

Along the way we described many types of organizing systems. Sometimes we discussed broad categories of organizing systems, like those for libraries, museums, business information systems, and compositions of web-based services. At other times we described specific instances of organizing systems, like those in the Seed Library, the Flickr photo sharing site, Amazon's drop shipment store, and your home kitchen or closet.

We can now build on the foundation created by Chapters 1-10 to create a "roadmap" that organizes and summarizes the design issues and choices that emerge

during an organizing system's lifecycle. These design choices follow patterns that help us understand existing organizing systems better, while also suggesting how to invent new ones by making a different set of design choices.

The roadmap is extensively annotated with references to the preceding chapters where the issues and choices mentioned in the roadmap were introduced and discussed in detail. We will use this roadmap to analyze a variety of case study examples in Chapter 12, and to explore the "design neighborhood" around each of them. The design questions from Chapter 1 serve as a template to give each case study the same structure, which we hope enables instructors, students, and others who read this book to add to this collection of case studies by contributing their own at `DisciplineOfOrganizing.org`.

Navigating This Chapter

We begin with a look at §11.2 The Organizing System Lifecycle (page 529) which proposes four phases, each of which is discussed in its own section. The first phase, which largely determines the extent and complexity of the resource descriptions needed for organizing and interactions, is §11.3 Defining and Scoping the Organizing System Domain (page 530). The second phase, which is highly shaped by economic factors and technology constraints or choices, is §11.4 Identifying Requirements for an Organizing System (page 536). §11.5 Designing and Implementing an Organizing System (page 542), emphasizes the need for clearly separating requirements from implementation, a principle we call architectural thinking. The final phase is discussed in §11.6 Operating and Maintaining an Organizing System (page 546), where we distinguish the maintenance of specific resources and descriptions from the maintenance of the system as a whole.

11.2 The Organizing System Lifecycle

System lifecycle models exhibit great variety; for our purposes it suffices to use a generic four-phase model that distinguishes a domain definition and scoping phase, a requirements phase, a design and implementation phase, and an operational and maintenance phase. These phases are brief and mostly inseparable for some simple organizing systems, more sequential for others, and more systematic and iterative for complex organizing systems.

Most of the specific decisions that must be made for an organizing system are strongly shaped by the initial decisions about its domain, scope (the breadth or variety of the resources), and scale (the number of resource instances). In organizing systems with limited scope and scale, most of these decisions are made in an informal, unanalyzed, and holistic manner. For example, when we arrange our bookshelves or closets it is not necessary to think explicitly about scoping, requirements, design, implementation, and operational phases. For complex organizing systems, however, especially those in information-intensive domains, it is important to follow a more systematic methodology.

Initial decisions about scope and requirements can create lasting technology and process legacies that impact operational efficiency and flexibility. They can also have profound and unforeseen ramifications for the users of the system and other people affected by the work the system enables. A rigorous, well-documented planning process can help organizers minimize unfair and ineffective outcomes, justify their difficult tradeoffs and decisions, and figure out what went wrong so they can learn from their mistakes.

The consequences of releasing technical systems and tools into the world always include social, business, political, and legal dimensions in addition to technical ones. Some of these implications are due to the context in which the system will operate (§10.4 Implementing Interactions (page 505)). Others are due to the fact that the work of organizing system designers, architects, and developers is shaped by their experiences, values, beliefs, and circumstances—the often hidden constraints and influences of their social position, education, cultural context, and mental models of the world. Inevitably, the work of information professionals involves “carving up the world at its joints,” creating classifications, models, and architectures that support interactions with resources. In practice this often translates to creating artificial “joints” where none truly exist, which will always favor some and injure others. No modeling is ever completely faithful to reality for all people with all experiences (nor is it intended to be), so those people not considered target users for a system, or who have unique circumstances, may end up feeling slighted or ignored and may actually suffer as a result.

11.3 Defining and Scoping the Organizing System Domain

The most fundamental decision for an organizing system is defining its *domain*, the set or type of resources that are being organized. This is why "What is Being Organized?" (§2.2) was the first of the design decisions we introduced in Chapter 1.

We refine how we think about an organizing system domain by breaking it down into five interrelated aspects:

1. the scope and scale of the collection
2. the number and nature of users
3. the time span or lifetime over which the organizing system will operate
4. the physical or technological environment in which the organizing system is situated
5. the relationship of the organizing system to other ones that overlap with it in domain or scope

Addressing these issues is a prerequisite for prioritizing requirements for the organizing system, proposing the principles of its design, and implementing the organizing system.

11.3.1 Scope and Scale of the Collection

The *scope* of a collection is the dominant factor in the design of an organizing system, because it largely determines the extent and complexity of the resource descriptions needed by organizing principles and interactions (§5.3.1.3). The impact of broad scope arises more from the heterogeneity of the resources in a collection than its absolute scale. It takes more effort to manage a broad and large collection than a narrow and small one; it takes less effort to manage a large collection if it has a narrow scope. A cattle ranch can get by with just one worker for every thousand cows, unlike zoos, which typically have a small number of instances of many types of animals. A zoo needs many more workers because each animal type and sometimes even individual animals can have distinct requirements for their arrangement and care.

Consider a business information system being designed to contain millions of highly structured and similar instances of a small number of related resource types, such as purchase orders and their corresponding invoices.[623][Com] The analysis to determine the appropriate properties and principles for resource description and organization is straightforward, and any order or invoice is an equally good instance to study.[624][Com]

Contrast this large but very narrow collection with a small but very broad one that contains a thousand highly variable instances of dozens of different resource types. This heterogeneity makes it difficult to determine if an instance is representative of its resource type, and every resource might need to be analyzed. This variability implies a large and diverse set of resource descriptions where individual resource instances might not be described with much precision because it costs too much to do it manually (§5.3.6.1). We can extrapolate to understand why organizing systems whose resource collections are both broad and deep, like those of Amazon or eBay, have come to rely on machine learning techniques to identify description properties and construct resource taxonomies (§5.3.6.4, §7.5.3.3).[625][Com]

A partial remedy or compromise when the resource instances are highly dissimilar is to define resource types more broadly or abstractly, reducing the overall number of types. We illustrated this approach in §8.2.2.1 when we contrasted how kitchen goods might be categorized broadly in a department store but much more precisely in a wholesale kitchen supply store. The broader categories in the department store blur many of the differences between instances, but in doing so yield a small set of common properties that can be used to describe them. Because these common properties will be at a higher level of abstraction, using them to describe resources will require less expertise and probably less effort (§5.3.1.3, §7.4.1). However, this comes at a cost: Poets, painters, composers, sculptors, technical writers, and programmers all create resources, but describing all of them with a "creator" property, as the *Dublin Core* requires, loses a great deal of precision.

Challenges caused by the scale of a collection are often related to constraints imposed by the physical or technological environment in which the collection exists that limit how large the collection can be or how it can be organized. (See §3.3.1) Only a few dozen books can fit on a small bookshelf but thousands of books can fit in your two-car garage, which is a typical size because most people and families do not have more than two cars. On the other hand, if you are a Hollywood mogul, superstar athlete, or sultan with a collection of hundreds of cars, a two-car garage is orders of magnitudes too small to store your collection.[626][Bus] Even collections of digital things can be limited in size by their technological environment, which you might have discovered when you ran out of space for your songs and photos on your portable media player.

Estimating the ultimate size of a collection at the beginning of an organizing system's lifecycle can reduce scaling issues related to storage space for the resources or for their descriptions. Other problems of scale are more fundamental. Larger collections need more people to organize and maintain them, creating communication and coordination problems that grow much faster than the collection, especially when the collection is distributed in different locations.

The best way to prevent problems of scope and scale is through standardization. Standardization of resources can take place if they are created by automated means so that every instance conforms to a schema or model (§7.5.2).[627][Com] Standards for describing bibliographic resources enable libraries to centralize and share much resource description, and using the same standards for resources of diverse types helps address the challenge of broad scope by reducing the need for close monitoring and coordination. Analogous standards for describing information resources, services, or economic activities business, governmental, or scientific information systems to systematically manage hundreds of millions or even billions of transactional records or pieces of data (§5.3.1.3).

11.3.2 Number and Nature of Users

An organizing system might have only one user, as when an individual creates and operates an organizing system for a clothes closet, a home bookcase or file cabinet, or for digital files and applications on a personal computer or smart phone. Collections of personal resources are often organized for highly individualized interactions using *ad hoc* categories that are hard to understand for any other user (§7.2.2). Personal collections or collections used by only a small number of people typically contain resources that they themselves selected, which makes the most typical interaction with the organizing system searching for a familiar known resource (§10.2.1).

At the other extreme, an organizing system can have national or even global scope and have millions or more users like the Library of Congress classification system, the United Nations Standard Products and Services Code, or the Internet Domain Name System. These organizing systems employ systems of institutional categories (§7.2.3) that are designed to support systematically specified and purposeful interactions, often to search for previously unknown resources. In between these extremes are the many kinds of organizing systems created by informal and formal groups, by firms of every size, and by sets of cooperating enterprises like those that carry out supply chains and other information-intensive business processes.

The nature and number of users strongly shapes the contents of an organizing system and the interactions it must be designed to support. (See §10.2.1) Some generic categories of users that apply in many domains are customers, clients, visitors, operators, and managers. We can adapt the generic interactions supported by most organizing systems (§5.3.2) to satisfy these generic user types. For example, while most organizing systems allow any type of user to browse or search the collection to discover its content, only operators or managers are likely to have access to information about the browsing and searching activities of customers, clients, or visitors.

Once we have identified the organizing system's domain more precisely we can refine these generic user categories, classifying users and interactions with more precision. For example, the customers of university libraries are mostly professors and students, while the customers of online stores are mostly shoppers seeking to find something to purchase. Library customers borrow and return resources, often according to different policies for professors and students, whereas online stores might only allow resources to be returned for refunds or exchanges under limited circumstances.

Just as it is with collection scope, the heterogeneity of the user base is more critical than its absolute size. An airport bookstore typically has a narrowly focused collection and treats its customers as generic travelers browsing imprecisely for something to fill their time in the terminal or on the airplane. In contrast, the local public library will have a much broader collection because it has to meet the needs of a more diverse user base than the airport bookstore, and it will support a range of interactions and services targeted to children learning to read, school students, local businesses, retirees, and other categories of users. A company library will focus its collection on its industry segment, making it narrower in coverage than a local or university research library, but it might provide specialized services for marketing, engineering, research, legal, or other departments of the firm.

Each category of users, and indeed each individual user, brings different experiences, goals, and biases into interactions with the organizing system. As a result, organizing systems in the same domain and with nominally the same scope can differ substantially in the resources they contain and the interactions they support for different categories of users. The library for the Centers for Disease Control and the WebMD website both contain information about diseases and symptoms, but the former is primarily organized to support research in public health and the latter is organized for consumers trying to figure out why they are sick and how to get well. These contrasting purposes and targeted users are manifested in different classification systems and descriptive vocabularies.

The designers of these systems do not necessarily share the same biases as their users, and more importantly, they may not always understand them completely or correctly. This is precisely why good design is iterative: successive cycles of evaluation and revision can shape crude, provisional, and misguided ideas into wildly successful ones. But such nimbleness is not always feasible in highly complex, political, or bureaucratic institutional contexts. Even then, as Bowker and Star conclude, transparency is the best corrective for these sorts of design failures. Designers who recognize that their systems have real consequences for real people should commit to an ongoing process of negotiation that enables those affected by the technology to voice and push back against any detrimental effect it might have on them and their communities. This helps set the stage for effective operation and maintenance of the system (§11.6.2).

11.3.3 Expected Lifetime

The scope and scale of a collection and the size of its user population are often correlated with the expected lifetime of its organizing system. Because small personal organizing systems are often created in response to a specific situation or to accomplish a specific task, they generally have short lifetimes (§7.2.2).

The expected lifetime of the organizing system is not the same as the expected lifetime of the resources it contains because motivations for maintaining resources differ a great deal. (See §3.5.1) As we have just noted, some organizing systems created by individuals are tied to specific short-term tasks, and when the task is completed or changes, the organizing system is no longer needed or must be superseded by a new one. At the other extreme are libraries, museums, archives, and other memory institutions designed to last indefinitely because they exist to preserve valuable and often irreplaceable resources.

However, most business organizing systems contain relatively short-lived resources that arise from and support day-to-day operations, in which case the organizing system has a long expected lifetime with impermanent resources. Finally, just to complete our 2 x 2 matrix, the auction catalog that organizes valuable paintings or other collectibles is a short-lived single-purpose organizing system whose contents are descriptions of resources with long expected lifetimes.

11.3.4 Physical or Technological Environment

An organizing system is often tied to a particular physical or technological environment. A kitchen, closet, card cabinet, airplane cockpit, handheld computer or smartphone, and any other physical environment in which resources are organized provides affordances to be taken advantage of and constraints that must be accommodated by an organizing system (§3.4.1).[629][Bus]

The extent of these physical and technological constraints affects the lifetime of an organizing system because they make it more difficult to adapt to changes in the set of resources being organized or the reasons for their organization. A desk or cabinet with fixed “pigeon holes” or drawers affords less flexible organization than a file cabinet or open shelves. A building with hard-walled offices constrains how people interact and collaborate more than an open floor plan with modular cubicles does. Business processes implemented in a monolithic enterprise software application are tightly coupled; those implemented as a choreography of loosely-coupled web services can often transparently substitute one service provider for another.

11.3.5 Relationship to Other Organizing Systems

The same domain or set of resources can have more than one organizing system, and one organizing system can contain multiple others. The organizing system for books in a library arranges books about cooking according to the Library of Congress or Dewey Decimal classifications and bookstores use the BISAC ones, mostly using cuisine as the primary factor (§8.3). In turn, cookbooks employ an organizing system for their recipes that arranges them by type of dish, main ingredient, or method of preparation. Within a cookbook, recipes might follow an organizing system that standardizes the order of their component parts like the description, ingredients, and preparation steps.

Sometimes these multiple organizing systems can be designed in coordination so they can function as a single hierarchical, or nested, organizing system in which it is possible to emphasize different levels depending on the user's task or application. Most books and many documents have an internal structure with chapters and hierarchical headings that enable readers to understand smaller units of content in the context of larger ones (§6.5.2). Similarly, a collection of songs can be treated as an album and organized using that level of abstraction for the item, but each of those songs can also be treated as the unit of organization, especially when they are embodied in separate digital files.

Organizing systems overlap and intersect. People and enterprises routinely interact with many different organizing systems because what they do requires them to use resources in ways that cut across context, device, or application boundaries. Just consider how many different organizing systems we use as individuals for managing personal information like contacts, appointments, and messages. As company employees we create and organize information in email, document repositories, spreadsheets, and CRM and ERP systems. Now consider this at an institutional scale in the inter-enterprise interactions among the organizing systems of physicians, hospitals, medical labs, insurance companies, government agencies, and other parties involved in healthcare. Consider how many of these are "mash-ups" and composite services that combine information and resources from independently designed systems.

We have come to expect that the boundaries between organizing systems are often arbitrary and that we should be able to merge or combine them when that would create additional value. It is surely impossible to anticipate all of these *ad hoc* or dynamic intersections of organizing systems, but it is surely necessary to recognize their inevitability, especially when the organizing systems contain digital information and are implemented using web architectures.

11.4 Identifying Requirements for an Organizing System

The two parts of the definition of an organizing system explicitly suggest two categories of requirements, those that specify the intentional arrangement of the resources and those that specify the interactions with the resources. These categories of requirements both depend on resource descriptions, which are implied by but not explicitly called out in the definition of an organizing system.

Because description, arrangement, and interaction are interrelated it is impossible to describe them separately without some redundancy. Nevertheless, in this book we have done that on purpose because taking different perspectives on organizing systems in Chapters 2-10 has enabled us to introduce a broad range of concepts, issues, and methods:

- Every organizing system must enable users to interact with its collection of resources (Chapters 3 and 10);
- The possible interactions depend primarily on the nature and extent of the descriptions associated with the resources (Chapters 4, 5 and 6);
- Intentional arrangement emerges when one or more resource descriptions are used by organizing principles (Chapters 7 and 8);
- Different implementations of the same organizing principle can determine the efficiency or effectiveness of the interactions it enables. (Chapter 10).

If you are creating a personal organizing system or otherwise small-scale one with only a small number of users, you might think there is little reason to think explicitly about requirements. However, any project benefits from the discipline of being more systematic about its purposes and their priorities. In addition, being explicit about requirements enables traceability and impact analysis. Traceability means being able to relate an interaction or feature of a system to the requirement it satisfies; impact analysis runs the causal link between requirements and features in the opposite direction to assess what or who will be affected if requirements change.[630][Bus]

11.4.1 Requirements for Interactions

When we describe interactions in a generic or broad way as we did in Chapter 3, *Activities in Organizing Systems* we see that all organizing systems have some common interactions, but most of the time we want to pay attention to the more specific interactions that are designed to create value in a particular organizing system because of the kind of resources it contains (§3.4.2). The domain, scope, and scale of the organizing system determines which interactions are possible and which ones must be explicitly supported, but the priorities of

different interactions are more often determined by decisions about intended users. (See §11.3.2.)

For most organizing systems other than personal ones, the set of interactions that are implemented in an organizing system is strongly determined by business model considerations, funding levels, or other economic factors. For-profit firms often differentiate themselves by the number and quality of the interactions they support with their resources, some by supporting many of them and some by supporting a minimal number. This differentiation is strongly shaped by and also shapes user preferences; some people prefer self-service or unmediated interactions, while others prefer full service and mediated interactions.[631][Bus] Non-profit institutions like public libraries and museums are also subject to these constraints, but unfortunately they have fewer options for adjusting service levels or changing their targeted user populations when their funding is reduced.

Some requirements for interactions come along with technology requirements, to have resources in a particular format, to conform to a particular specification or standard in order to operate in some technology environment, or to interoperate with other parties or their organizing systems.

An essential requirement in every organizing system is ensuring that the supported interactions can be discovered and invoked by their intended users. In organizing systems with physical resources, good designers enhance the inherent affordances of resources with navigation and orientation aids that direct users to points of interactions (§3.4.1). With digital resources and information-intensive organizing systems, interactions are not immediately perceivable, and poor design can create overly complicated user interfaces in which many interactions are never discovered and thus never used.

It is tricky to compare the overall capabilities of organizing systems in terms of the number or variety of their interactions because what matters more is how much value they create. Organizing systems with active resources can create value on their own without an explicit user interaction (§4.2.3.2). Other organizing systems exploit stored, computed, or *contextual information* to create value by eliminating the need for user interactions, such as location-based smartphone apps that push information to you when you are near some particular location or some person you know (§3.4.1).[633][Bus]

11.4.2 About the Nature and Extent of Resource Description

Interactions with resources within an organizing system often depend on descriptions of individual resources or descriptions of the collections that contain them. In the bibliographic domain, generic or common interactions make use of descriptions that can be associated with almost any type of resource, such as the name, creator, and creation date.

For example, any resource with a sortable name or identifier can be arranged alphabetically to enable it to be easily found, and any resource with a creation date can be discovered by a "what's new" query to a resource collection.

Different types of resources must have differentiating properties, otherwise there would be no reason to distinguish them as different types. These resource properties can be recorded in the terms of a description language to support one or more interactions or to answer one or more questions. Simply put, choices about the nature and extent of resource description depend on which interactions or questions are most frequent or important (§5.3.1.1). If a particular property of a resource has no interactions that depend on it, there is no need to describe it. However, if an interaction depends on a description of a particular resource property, a missing description or one of inadequate precision and granularity means that the interaction will be impossible or inefficient to carry out because the resource will need to be further analyzed to create or extract the required description. An ISBN is a sufficient description to find a book in a directory, but if the ISBN is the only description associated with the book you will not be able to tell who wrote it. The tradeoffs imposed by the extent and timing of resource description have been a recurring theme in this book, with the tradeoff between recall and precision being the most salient (§2.5, §3.4.1, §7.4.1, §7.4.3).

The properties of resources that are easiest to describe are not always the most useful ones, especially for information resources. Anyone can determine the number of pages in a book, but often only a skilled cataloger can accurately describe what the book is about, a far more important property. (§5.2.2 "Description" as an Inclusive Term (page 220) and §7.3.4 The Limits of Property-Based Categorization (page 346)) For non-text information resources this problem is magnified because the content is often in a semantically opaque format that it optimized for the devices that creates and processes it but which cannot usefully be analyzed by people. (§4.4.2.5 The Semantic Gap (page 193) and §5.4 Describing Non-text Resources (page 257))

Business strategy and economics strongly influence the extent of resource description. In many museums and archives there are not enough trained people and time to describe every pottery fragment or document, and many resources are described only at an aggregate level. In contrast, some people argue that the explosion of content in physical and digital form mandates significant investment in descriptions that facilitate resource discovery in a crowded marketplace.

Automated and computerized processes can create the resource descriptions in an organizing system and their use is primarily driven by scale (§5.3.6.4). Search engines index web pages and analyze their link structures because it would be impossible to treat the web as a traditional library and organize it by

human effort. The benefits of digital cameras, video recorders, and similar devices would be far fewer if people had to manually identify and describe each resource when creating it. Instead, these devices can automatically assign some contextual metadata. Similarly, competitive pressures on vendors to provide real-time and context-sensitive information services mandate automated collection of contextual information like location from mobile phones, portable book readers and tablet computers.

Color Coded Library

Because he is presumably familiar with the contents of all of his books, interaction designer Juhan Sonin organized his library according to their spine colors. This organizing principle is a highly individual and aesthetic one, but it would probably not appeal to people unfamiliar with the collection and would bring chaos to a library of larger size.

(Photo by See-ming Lee. Creative Commons CC-BY-SA-2.0 license.)

We might seek some optimal degree of description given some set of requirements or purposes for an organizing system and some estimate of the organizing effort that could be applied; in practice this is elusive for two reasons, both relating to scope and scale. First, as the number of users of an organizing sys-

tem increases, it becomes more difficult to identify and anticipate all its possible purposes and constraints it must satisfy. Even if most users share the goals for the organizing system, any particular user might have some additional specialized use for some attributes or relationships that would require more description to satisfy.[636][Bus]

Second, even if it were possible to implement some optimal degree of description in a particular organizing system, we would still encounter problems when multiple organizing systems exist in the same domain or in domains that intersect across context, device, or application boundaries. Since organizing systems are designed and evolve to satisfy the specific requirements of their particular context, companies will often describe the same resources differently, which creates integration and interoperability problems when companies need to exchange and combine information resources (§10.3.2).

11.4.3 About Intentional Arrangement

Organizing principles depend on resource descriptions, so requirements for the former are always intertwined with those for the latter. Specifying requirements for the intentional arrangement of resources is analogous to specifying why and how resource categories can be created (§7.3). In turn, the creation of resource categories often becomes a question about the number and kind of resource properties that might be analyzed and exploited by organizing principles.

We noted that there is a continuum of category formation that ranges from minimal use of resource properties to more rigorous use of multiple properties, and finally to statistical or composite use of multiple properties, some of which are induced or inferred rather than explicit. The simplest principle for defining a category is by enumeration, just putting the resources into a set without any specification of any properties they might share. The enumerated resources might very well have common properties, but the principle of enumeration ignores them; the only property that matters for that principle is that the resources are in the same set. This corresponds to the simplest principle of intentional arrangement, that of *collocation*, just putting the resources in the same location without any additional organization.

Collocated resources often acquire some additional arrangement as a result of their use; consider how the books, papers, or other resources gathered for some writing project often end up in piles in your office or on your desk close to your work area. For a small collection, the proximity-to-use organizing principle is the easiest way to satisfy a requirement to minimize the time to find frequently used resources.

As we have often seen, the scope and scale of the organizing system is a dominant design consideration and it applies to principles of resource arrangement too. The *collocation* principle of arrangement is sufficient for small resource col-

lections because it is not necessary to define the optimal organization if the time to find any particular resource is short even for an inefficient search method of scanning the entire collection. Using the extrinsic property of frequency of use makes search slightly more efficient, but only in organizing systems where the user population is small or interacts with the resources in similar ways. Otherwise, arranging resources to facilitate their frequent access for some users would hinder other users who never use them. Imagine if you shared your office desk with someone who works all night on other writing projects and leaves his frequently used resources in piles close to his work area—which becomes your work area in the morning.

Larger resource collections usually require multiple organizing principles to manage the complexity that emerges when more users and more varied interactions must be supported. A valet parking lot might organize cars only by size to make optimal use of limited space when parking and fetching them, but when cars are organized for sale they would be organized by price, performance, seating capacity, manufacturer, and many other properties. It is essential to establish the priority of users and interactions because these requirements determine the order in which the principles are applied to arrange the resources. This ordering creates a logical resource hierarchy that affects the efficiency of interactions and the maintenance of the organizing system over time.

Information resources are invariably challenging to arrange because their aboutness is not an easily perceived property and because of the open-ended purposes they can serve. Information collections with broad scope most often use a standard system of bibliographic classification (§8.3). In contrast, special libraries have narrower collections that need to support domain-specific interactions for a relatively small set of users, and as a result they require more specialized organizational schemes. The principles for resource arrangement in large firms of every type are often required to conform with laws and regulations for accounting, taxes, human resources, data retention and non retention, access control, and other functions. (See §3.5.4.1, §8.1.5.4)

11.4.4 Dealing with Conflicting Requirements

Any individual, group, or enterprise can create an organizing system that meets their specific requirements, but once this organizing system involves two or more parties with different requirements, there is a potential for conflict. Roommates or spouses sometimes argue about how to organize items in the kitchen, in the refrigerator, or in some other shared space. To a person who arranges spices alphabetically and condiment jars by size, arranging them according to cuisine or frequency of use makes no sense. Similarly, if you are the sole user of a Dropbox or other cloud storage account, you can organize it any way you want. You can use any number of folders that need only make sense to you, or you can leave everything unorganized in a single folder. However, if you share

the Dropbox account with another person, they are likely to have different organizational needs or preferences. Perhaps you tend to organize resources by file type, while they prefer to organize resources by topic or project.

A small number of people can often agree on an organizing system that meets the needs of each participant through informal negotiations. The potential for conflict increases when more people are involved, and "bottom-up" *ad hoc* negotiations to resolve every disagreement between every pair of participants just are not feasible. In many domains conflicts are avoided or suppressed because the parties have developed or agreed to conform to standards (§8.1.5). Nevertheless, conflicts in organizing principles for large-scale organizing systems are often resolved by parties with the legal authority or economic power to impose a solution on all the participants in a "top-down" manner.[638][Bus]

11.5 Designing and Implementing an Organizing System

Requirements define what must be done but NOT how to do it; that's the role of the design and implementation phases. Being explicit about requirements and the intended scope and scale of an organizing system before moving onto these phases in an organizing system's lifecycle avoids two problems. The first is taking a narrow and short-term focus on the initial resources in a collection, which might not be representative of the collection when it reaches its planned scope and scale. This can result in overly customized and inflexible resource descriptions or arrangements that cannot easily accommodate the future growth of the collection. A second problem, often a corollary of the first, is not separating design principles from their implementation in some specific environment or technology.

11.5.1 Choosing Scope- and Scale-Appropriate Technology

A simple organizing system to satisfy personal record keeping or some short-lived information management requirements can be implemented using folders and files on a personal computer or by using "off the shelf" generic software such as web forms, spreadsheets, databases, and wikis. Other simple organizing systems run as applications on smart phones. Some small amount of configuration, scripting, structuring or programming might be involved, but in many cases this work can be done in an *ad hoc* manner. The low initial cost to get started with these kinds of applications must be weighed against the possible cost of having to redo a lot of the work later because the resources and the resource descriptions might not be easily exported to new ones.

More capable organizing systems that enable the persistent storage and efficient retrieval of large amounts of structured information resources generally require additional design and implementation efforts. Flat word processing files and spreadsheets are not adequate. Instead, XML document models and data-

base schemas often must be developed to ensure more control of and validation of the information content and its descriptions. Software for version and configuration management, security and access control, query and transformation, and for other functions and services must also be developed to implement the organizing system.

Technology for organizing systems will always evolve to enable new capabilities. For example, cloud computing and storage are radically changing the scale of organizing systems and the accessibility of the information they contain. It might be possible to implement these capabilities and services to an organizing system in an incremental fashion with informal design and implementation methods. If information models, processing logic, business rules and other constraints are encoded in the software without explicit traceability to requirements and design decisions the organizing system will be difficult to maintain if the context, scope or requirements change. This is why we have repeatedly emphasized the importance of architectural thinking about organizing systems, beginning in §1.6 where we proposed that organizing principles should ideally be expressed in a way that did not assume how they would be implemented. (See also §3.3.3.2, §8.1.3, and §9.1)

11.5.2 Architectural Thinking

Much of the advice about designing and implementing an organizing system can be summarized as "architectural thinking," introduced in §1.6. The overall purpose of architectural thinking is to separate design issues from implementation ones to make a system more robust and flexible. Architectural thinking leads to more modularity and abstraction in design, making it easier to change an implementation to satisfy new requirements or to take advantage of new technologies or procedures. It is also important to think architecturally about the design of the vocabularies and schemas for resource description and of classification systems to leave room for expansion to accommodate new resource types (§7.5 and §8.2.2.3). Doing so is easier if the descriptions are logically and physically distinct from the resources they describe. A checklist the brings together useful principles and processes for architectural thinking from all parts of this book is in the nearby sidebar.

Nevertheless, architectural thinking requires more careful analysis of resources and implementation alternatives, and most people do not think this way, especially for personal and informal organizing systems. You can imagine that someone might arrange a collection of paperback books in a small bookcase whose shelf height and width were perfectly suited for the paperbacks they currently own. However, this organizing system would not work at all for large format books, and a paperback could not be added to the collection unless one was purged from the collection. It would be more sensible to start with a bigger

Principles and Processes for Architectural Thinking

- Explicitly define the purposes and users of the organizing system, recognizing that users might not agree on purposes or their priorities (§2.3).
- Select resources (§3.2.1) and design interactions (§3.4.2) to support the primary or highest priority users.
- Specify information and interaction requirements in a conceptual and technology-neutral way (§1.6) that conforms as much as possible to domain standards, schemas, or vocabularies (§8.1.5).
- Implement user interactions with design patterns to make them more discoverable, usable, and effective (§3.3.3.2).
- Follow principles for good names (§4.4.3) and good resource descriptions (§5.3.4.1).
- Make an informed decision about the tradeoff between flexibility and complexity; a simpler system might be easier to adapt (§8.2.2.3).
- Make design and technology decisions consistent with the expected life of the organizing system (§11.6).

bookcase with adjustable shelves so that the organizing system would have a longer lifetime.

You might think that large institutional organizing systems would avoid these problems caused by tying a collection too tightly to the physical environment in which it is initially organized, but sometimes they do not. A famous example involves the art collection of the Barnes Foundation, which had to keep its paintings in the exact same crowded arrangements when the museum made a controversial move from a small building to a larger one because the donor had mandated that the paintings never be moved from their original settings. (See the sidebar, The Barnes Collection (page 546)).

For digital resources, inexpensive storage and high bandwidth have largely eliminated capacity as a constraint for organizing systems, with an exception for *big data*, which is defined as a collection of data that is too big to be managed by typical database software and hardware architectures.[639[Com]] Even so, big data collections are often large but homogeneous, so their scale is not their most important challenge from an organizing system perspective (§11.3.1).

11.5.3 Distinguishing Access from Control

Because large resource collections are often used for multiple purposes by many different people or projects, they illustrate another important architectural issue for collections of digital resources. A requirement for access to resources does not imply a need to directly own or control them, and information-intensive and web-based businesses have increasingly adopted organizing

Stop and Think: What is a Library?

The word "library" has several meanings that differ in how much architectural thinking they embody. When you tell someone you will meet them at the library for a cup of coffee and a study session it is a specific physical place. At other times you might use a more abstract notion of a library as an organizing system with a predictable type of collection, resource arrangement, and supported interactions. Both meanings are important for a city creating a new library. How would you ensure that both are considered in an effective way?

system designs that involve storage of digital resources in the cloud, licensing of globally distributed resources, and outsourcing of information services. Designs that use these architectural concepts can realize functional and quality improvements because the location and identity of the service provider is hidden by an abstraction layer (§3.4.2.1, §4.4.3.5). However, separating access from ownership has been a cultural challenge for some libraries and museums whose institutional identities emphasize the resources they directly control and the physical buildings in which they control them.

11.5.4 Standardization and Legacy Considerations

As we noted with the Barnes Collection, a building becomes old and outdated over time. The technology used in digital organizing systems becomes obsolete faster than physical buildings do. The best way to slow the inevitable transformation of today's cutting edge technology to tomorrow's legacy technology is to design with standard data formats, description vocabularies and schemas, and classification systems unless you have specific requirements that preclude these choices.

Even a requirement to interoperate with an organizing system that uses proprietary or non-standard specifications can usually be satisfied by transforming from a standard format (§8.1.5.2, §10.4). Similarly, it is better to design the APIs and data feeds of an organizing system in a generic or standard way that abstracts from their hidden implementation. This design principle makes it easier for external users to understand the supported interactions, and also prevents disclosure of any aspects of resource description or organization that provide competitive advantage. For example, the way in which a business classifies products, suppliers, customers, or employees can be competitively important.

Two important design questions that arise with data transformation or conversion, whether it is required by a technology upgrade or an interoperability requirement, are when to do it and where to do it. The job of converting all the resources in a collection can typically be outsourced to a firm that specializes in format conversion or resource description, and a batch or pre-emptive conversion of an entire collection enables an upgraded or new organizing system to

The Barnes Collection

Albert Barnes was a chemist who made a fortune inventing a preventive treatment for gonorrhea and who then amassed perhaps the greatest private art collection ever, one that contained over 800 paintings by artists like Picasso, Renoir, Matisse, van Gogh, and Cezanne. In 1922 Barnes built a museum for his collection in his residential neighborhood in Merion, PA, a suburb of Philadelphia. Barnes did not open his collection to the public and in his will mandated that the collection never be moved, loaned, or sold.

In the decades after Barnes died in 1951 the Merion museum needed extensive repairs and security upgrades, and some people suggested that its remote location and access restrictions jeopardized its financial viability. However, a proposal to relocate the collection to Philadelphia seemingly violated the terms of the Barnes will.

A legal fight dragged on for decades. Finally in 2004 a judge ruled that the collection could be moved to Philadelphia, but only if the new museum contained exact copies of the gallery rooms of the original museum and arranged the paintings exactly as they were in Merion. The new museum building, opened in 2012, is ten times larger than the old one, but the collection takes up the same space as it did in Merion. The other 90% of the building is occupied by an auditorium, offices, classrooms, a gift shop, and other space that contains none of the collection.

operate more efficiently when it is not distracted by ongoing conversion activity. On the other hand, if resources vary greatly in their frequency of use, a "do-it-yourself on-demand" method is probably more cost effective as long as the conversion does not impact the interactions that need to be supported.

11.6 Operating and Maintaining an Organizing System

After the organizing system has been designed and implemented it can be put into its operation and maintenance phases. We will look at these from two perspectives, first from the point of view of individual resources, and then from the point of view of the organizing system's design and implementation. These two perspectives are not always clearly distinguished. Curation, for example, is often used to describe actions taken to maintain individual resources as well as those that result in new arrangements of them.

11.6.1 Resource Perspective

Sometimes an organizing system is implemented with its organizing structures and relationships waiting to be populated by resources as they are acquired and described. The scope and scale of the organizing system shapes how the de-

scriptions are created and how the descriptions are then used to assign resources to the logical or physical containers of the organizing system. The most important decisions to be made at this point involve determining an appropriate mix of methods for creating the resource descriptions, because their cost, quality, consistency, completeness, and semantic richness depends on which human or computational agents do the work (§5.3.6).

For web-based and consumer-focused organizing systems, it is tempting to rely on users to assign descriptions, tags, or ratings to resources (§5.2.2.3). Some of these systems attempt to improve the quality and precision of these descriptions by providing forms, controlled vocabularies, or suggestions. Finding a balance is tricky; too much direction and control is demotivating to uncompensated volunteer describers, and too little of it results in the proverbial "tag soup."

An essential operational and maintenance activity is evaluation of resource descriptions, first with respect to the time and process by which they are created, and second with respect to how and when they support the designed interactions (§5.3.7).

Some organizing systems are initiated with a fixed set of resources that will not change in any way. For example, in an archive as most narrowly defined, neither the individual resources nor the organization of the collection as a whole will change. If an archive of Abraham Lincoln letters is established, we know that Lincoln will never revise any letters or write any new ones, and any new classifications or descriptions devised by people studying the archive will not be used to rearrange the letters.

Most organizing systems, however, need to support ongoing interactions with a collection that changes over time as new resources enter the collection and old ones leave. These selection and collection management processes are explicit in libraries, museums and similar institutions that maintain collections to satisfy the changing needs and preferences of their user communities (§3.2.2).

11.6.2 Properties, Principles and Technology Perspective

It is useful to consider how an organizing system as a whole is operated and maintained over time. We can analyze how the system's organizing properties, principles and technology might change, and we can roughly order different types of change according to their overall impact.

The most predictable maintenance activities for an organizing system with an expected long lifetime (§11.3.3) are incremental changes in description vocabularies and classification schemes (§8.2.2.3). These need to evolve when new instances or contexts require additional properties to maintain the distinctions between types of resources, but the basic principles embodied in the organizing systems are not affected.

Incremental category maintenance takes place even in personal organizing systems where the categories are not always explicit. The collection of clothes in a college student's closet and the categories and properties for arranging them will change somewhat when he graduates and takes a job in a downtown office building where he needs to dress more formally than he did as a student. He will learn that despite the common term in the category name, "student casual" and "business casual" do not contain the same sets of resources.

Category maintenance is an ongoing activity in institutional organizing systems. The most commonly used bibliographic classification systems all have numbering and naming schemes that allow for subdivision and extension to create new subcategories to accommodate resources about new fields of knowledge and technology.

As another example, the *Association for Computing Machinery (ACM)* professional society created a keyword classification in 1964 to organize articles in its many publications, but relentless change in the computing field driven by Moore's Law has required the ACM to significantly revise the system almost every decade.[642[Com]]

In contrast, changes in business organizing systems are more likely to be driven by economic factors. Resource properties for managing collections of resource and the information that describes them often change over time as a result of new products and services, mergers and acquisitions, or refined customer segmentation. More substantial changes in business organizing systems reflect the need to comply with laws and regulations that impose new requirements for tracing money flows or transactions. These mandated classifications and processes might require new organizing principles, not just incremental properties (§8.1.5.4).

The choice of implementation technology influences how easy it is to handle these types of changes in organizing systems. In databases this problem is known as "schema migration." With XML implementations, schemas can be designed with "extension" or "codelist" elements to enable changes that will not invalidate existing information. Business processes that are driven by "executable specifications" like the *Business Process Execution Language (BPEL)* can be easily modified because the BPEL XML instance is used to configure the software that carries out the process it describes.[643[Com]]

Another very predictable type of activity over time with organizing systems is a technology upgrade that improves its quality or capabilities without affecting the organizing principles. A student might replace his handwritten lecture notes with typed notes on a laptop or tablet computer but not significantly change the way the notes are organized.

Institutional organizing systems are adopting tiered storage systems that automatically move resources between different types of storage media to meet performance, availability and recovery requirements. For example, firms with high financial impact of downtime like banks run critical organizing systems with copies in “failsafe” or “hot” modes that are synchronized with the production environments to prevent any interruptions in information access if the latter are disrupted. On the other hand, resources needed for regulatory compliance can be kept on lower cost disk storage.

The most challenging kinds of maintenance activities for organizing systems involve changes to the principles for arranging resources along with changes in the implementing technology. An example is the ambitious effort to introduce semantic web and linked data concepts in bibliographic organizing systems (§6.8.2, §9.4.3). And change comes faster to businesses than to libraries and museums. New technologies can have a disruptive impact on business organizing system, forcing major changes to enable strategy changes that involve faster finding, retrieval, or delivery of informational or physical goods.[644][Com]

Sometimes major changes to organizing principles and technologies can be introduced incrementally, with changes “rolled out” to different sets of resources or user groups during a transition period. However, sometimes the changes are inherently ”all or none” because it is impossible to have two conflicting organizing systems operating in the same context. An easy to understand example of an organizing system that changed radically is the system governing which side of the road you drive on, which was changed in Samoa in 2009. (See the sidebar, Driving in Samoa (page 550)).

11.7 Key Points in Chapter Eleven

- Most of the specific decisions that must be made for an organizing system are shaped by the initial decisions about its domain, scope, and scale.

  (See §11.2 The Organizing System Lifecycle (page 529))
- The impact of broad scope arises more from the heterogeneity of the resources and users in a collection rather than from their absolute number.

  (See §11.3.1 Scope and Scale of the Collection (page 530))
- Larger collections need more people to organize and maintain them, creating communication and coordination problems that grow much faster than the collection.

  (See §11.3.1 Scope and Scale of the Collection (page 530))
- Standardization is the best way to prevent problems of scope and scale.

  (See §11.3.1 Scope and Scale of the Collection (page 530))

Driving in Samoa

Whether you travel by bus, car, or bicycle, you always keep to one side of the road. The convention of driving on either the right side or the left side is a legal standard that everyone takes for granted. However, you must follow it to ensure safe driving and avoid collisions and crashes.

This standard of which side of the road you drive on was not decided arbitrarily, but rather, it was adopted as a result of history, convention, and the need for organization. If you were the only person to use the road, you could choose to travel on any side you wanted, even travel right down the middle. As soon as more than one person needs to use the same road, the risk of collisions compels the creation of a coordinating standard.

In 2009, the government of Samoa took the rare step of changing the side of the road standard from driving on the right to driving on the left. The original standard reflected the influence of German colonization in the early 1900s. However, Samoa is both geographically close to and economically intertwined with Australia and New Zealand, former British colonies that follow the British convention of driving on the left side. This proximity gives Samoa access to a nearby source of used cars that would be attractive to Samoa's relatively poor population. So, the Samoan government decided to use its authority to change the driving standard so that more of its people could afford to buy cars.

As one could imagine, this decision was not implemented without controversy and opposition. While the decision benefited people currently without cars, it negatively affected those who already owned them. After a switch like this, what happens to the current market value of the thousands of cars designed to drive on the right? Opponents also claimed that the switch would cause unprecedented safety hazards. If even a small fraction of drivers were not able to immediately get the hang of driving on the other side, the accident rate could increase tremendously. Imagine the current pool of buses designed with doors that open on the right hand side—would they now let passengers out in the middle of the street? Who would pay to have the buses modified to put doors on the left hand side?[645][Bus]

- Organizing systems in the same domain and with nominally the same scope can differ substantially in the resources they contain and the interactions they support if they have different categories of users.

  (See §11.3.2 Number and Nature of Users (page 532))

- Designers who recognize that their systems have real consequences for people should commit to measures of transparency and an ongoing process of

negotiation that enables those affected to voice concerns related to any detrimental effects the technology might have on them and their communities.

(See §11.3.2 Number and Nature of Users (page 532))

- For most organizing systems other than personal ones, the set of interactions that are implemented in an organizing system is strongly determined by economic factors.

  (See §11.4.1 Requirements for Interactions (page 536))

- An essential requirement in every organizing system is ensuring that the supported interactions can be discovered and invoked by their intended users.

  (See §11.4.1 Requirements for Interactions (page 536))

- Automated and computerized processes can create the resource descriptions in an organizing system and their use is primarily driven by scale.

  (See §11.4.2 About the Nature and Extent of Resource Description (page 537))

- Organizing principles depend on resource descriptions, so requirements for the former are always intertwined with those for the latter.

  (See §11.4.3 About Intentional Arrangement (page 540))

- Overly customized and inflexible resource descriptions or arrangements cannot easily accommodate the future growth of the collection.

  (See §11.5 Designing and Implementing an Organizing System (page 542))

- Architectural thinking leads to more modularity and abstraction in design, making it easier to change an implementation to satisfy new requirements or to take advantage of new technologies or procedures.

  (See §11.5.2 Architectural Thinking (page 543))

- For digital resources, inexpensive storage and high bandwidth have largely eliminated capacity as a constraint for organizing systems, with an exception for *big data*, which is defined as a collection of data that is too big to be managed by typical database software and hardware architectures.

  (See §11.5.2 Architectural Thinking (page 543))

- The most predictable maintenance activities for an organizing system with an expected long lifetime are incremental changes in description vocabularies and classification schemes.

  Another very predictable type of activity over time with organizing systems is a technology upgrade that improves its quality or capabilities without affecting the organizing principles.

  (See §11.6.2 Properties, Principles and Technology Perspective (page 547))

- The most challenging kinds of maintenance activities for organizing systems involve changes to the principles for arranging resources along with changes in the implementing technology.

  (See §11.6.2 Properties, Principles and Technology Perspective (page 547))

- What resources are being organized? Why are the resources being organized? Who does the organizing? When are the resources organized? Where are the resources organized? How much are the resources organized?

  (See the case studies presented in Chapter 12, *Case Studies*)

Endnotes for Chapter 11

[623][Com] For some kinds of resources with highly regular structure, the distinction between the resource and its description is a bit arbitrary. A transactional document like a payment contains at its core a specification of the amount paid, which we could consider the payment resource. Information about the payer, the payee, the reason for the payment, and other essential information might be viewed as descriptions of the payment resource. In a payment or financial management system, the entire document might be treated as the resource.

[624][Com] The results of this analysis can be represented in a conceptual model or document /database schema that can guide the automated creation of the resource instances and their descriptions (§5.3.1.2). Furthermore, these models or schemas can also be used in "model-based" or "model-driven" architectures to generate much of the software that implements the functionality to store the instances and interchange them with other information systems; "imagine if the construction worker could take his blueprint, crank it through a machine, and have the foundation for the building simply appear." Quote comes from (Miller and Mukerji 2003). See also (Kleppe, Warmer, and Bast 2003).

[625][Com] See (Chen, Li, Liang, and Zhang 2010), (Pohs 2013).

[626][Bus] See `http://autos.ca.msn.com/editors-picks/the-worlds-biggest-car-collectors`.

[627][Com] Model-driven software generation can be simple—an XFORM specification that creates an input form on a web page. Or it can be complex—a detailed architectural specification in UML sufficient to generate a complete application.

[629][Bus] Service design, architecture, and user interaction design are the primary disciplines that care about the influence of layout and spatial arrangement on user interaction behavior and satisfaction. One type of physical framework is the "Servicescape" (Bitner 1992), the man-made physical context in which services are delivered. For example, the arrangement of waiting lines in banks, su-

permarkets, and post offices or the use of centrally-visible "take a number" systems strongly influence the encounters in service systems (Zhou and Soman 2003). Related concepts for describing the use of features and orienting mechanisms in "the built environment" come from the "Wayfinding" (Arthur and Passini 1992) literature in urban planning and architecture.

[630][Bus] An easy to remember framework for prioritizing requirements is MoSCoW, which classifies them as Must, Should, Could, and Won't (Desoky 2010). (Winkler and Pilgrim 2010) is a comprehensive review of academic research and best practices for requirements traceability.

[631][Bus] "Customer segments" or "customer models" are well-established constructs in product and service marketing and operations (Batt 2000) (Zeithami, Rust, and Lemon 2001). They are key parts of strategies for acquiring customers, increasing market share, and retaining customers. Customer segments can be identified using numerous overlapping criteria, including demographic variables, product or behavior choices, and preferred interaction locations or channels. For example, an airline might segment its customers according to their age, gender, home airport, ticketing class, and travel frequency.

[633][Bus] Organizing systems differ in the extent they can initiate interactions or use information to make them unnecessary. In libraries the organizing systems are typically designed not to preserve user activity records longer than absolutely necessary; in commercial organizing systems, user activity records are the basis of business processes that create highly detailed user models (called "microsegments" or "microcategories") that enable personalized product and service offerings See (Taylor and Raden 2007), (Rosen 2012).

[636][Bus] For example, we have often used a home kitchen as a setting for organizing systems. Suppose the home kitchen is to be used as the set for a cooking show and the designers want to arrange cookbooks to make the background visually pleasing. The designers would like to search for cookbooks on the basis of size and spine color, but these descriptive elements would be of little value to other users.

[638][Bus] Some people call this the "Wal-Mart approach" to standardization. A firm with dominant market power does not need to negotiate standards because it can impose whatever standards it chooses on its partners as a condition of doing business with them. When there are conflicting requirements, different relationships within the set of participants trying to reach agreement, and different extents to which they are subject to the authority behind the desired agreement, it is not surprising that approaches "that require perfect coordination and altruism are of no practical interest" (Rosenthal et al. 2004, p 47).

[639][Com] Note that this definition does not include any specific size threshold, such as some number of terabytes (thousands of gigabytes). This allows the

threshold size that makes a collection a big data one to increase as storage technology advances. It also recognizes that different industries or domains have different thresholds (Manyika et al 2011).

[642][Com] No classification scheme ever devised is as unstable as the ACM's because new computing concepts, technologies, and application areas are constantly emerging. Even the society's name seems outdated.

[643][Com] For a formal computer science treatment of BPEL see (Fu, Bultan, and Su 2004); for a commercial perspective see `http://www.oracle.com/technetwork/middleware/bpel/overview/index.html`.

[644][Com] This means that the organizing systems used by business applications more often employ configuration management, version control, model-based code generation, and other computing techniques that robustly support the need for qualitative changes in the organizing systems.

[645][Bus] (Barta 2009).

Chapter 12
Case Studies

Applying the Roadmap

Robert J. Glushko

Overview

We now fulfill the promise of this book, with a set of case study examples that apply the concepts and phases of the roadmap. (The first four case studies appeared in the first print edition of the book. All the others have been contributed by students or other readers of the book and edited for consistency.——Ed.)

Case Study Template

For the sake of consistency, we employ the questions posed in Chapter 2, *Design Decisions in Organizing Systems* as a template for the case studies. We remind you of six groups of design decisions, itemizing the most important dimensions in each group:

- **What is being organized?** What is the scope and scale of the domain? What is the mixture of physical things, digital things, and information about things in the organizing system? Is the organizing system being designed to create a new resource collection, catalog an existing and closed resource collection, or manage a collection in which resources are continually added or deleted? Are the resources unique, or are they interchangeable members of a category? Do they follow a predictable "life cycle" with a "useful life"? Does the organizing system use the interaction resources created through its use, or are these interaction resources extracted and aggregated for use by another organizing system? (§2.2)
- **Why is it being organized?** What interactions or services will be supported, and for whom? Are the uses and users known or unknown? Are the users primarily people or computational processes? Does the organizing system need to satisfy personal, social, or institutional goals? (§2.3)
- **How much is it being organized?** What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed? What organizing principles guide the organization? Are all resources organized to the same degree, or is the organization sparse and non-uniform? (§2.4)
- **When is it being organized?** Is the organization imposed on resources when they are created, when they become part of the collection, when interactions occur with them, just in case, just in time, all the time? Is any of this organizing mandated by law or shaped by industry practices or cultural tradition? (§2.5)
- **How or by whom, or by what computational processes, is it being organized?** Is the organization being performed by individuals, by informal groups, by formal groups, by professionals, by automated methods? Are the organizers also the users? Are there rules or roles that govern the organizing activities of different individuals or groups? (§2.6)
- **Where is it being organized?** Is the resource location constrained by design or by regulation? Are the resources positioned in a static location? Are the resources in transit or in motion? Does their location depend on other parameters, such as time? (§2.7)

As we discussed in §2.7 Where is it being Organized? (page 81), when location is a constraint, it will typically be identified as such in the other questions. As result, we will only examine "Where?" as distinct design dimension in cases where it is warranted.

12.1 A Multi-generational Photo Collection

Overview. Your grandfather has died, at age 91, and under his bed is a suitcase containing several photo albums with a few hundred photos. Some of them have captions, but many do not. What do you do with them?

Your first thought was to create a digital photo archive of Grandpa's collection so that you and all your relatives could see them, and you would also want to generate accurate captions where none exist. Since you have an extensive digital photo collection of your own in a web-based application, perhaps you can combine the two collections to create a multi-generational photo organizing system.

This project involves digitization, archiving, social media issues, and negotiations with and collecting information from other family members who might have different views about what to do.

What is being organized? It is easy to find advice about how to digitize old photos, but there are more choices than you might think. What resolution and format should you use? Should you do the work yourself or send Grandpa's precious photos to a service and take the risk that they might get lost? Should you do any restoration or enhancement of the photos as part of the digitization process?[646][CogSci]

More fundamental design questions concern the scope and scale of the organizing system. If you are digitizing Grandpa's photos and combining them with yours, you are skipping a generation. Should not you also include photos from your parents and the rest of Grandpa's children? That generation has both printed photos and digital ones, but it is not as comfortable with computers as you are, and their digital photos are stored less systematically on a variety of CD-ROM, DVDs, flash memory sticks, and SD photo cards, making the digitizing and organizing work more complicated. Do these differences in storage media reflect an intentional arrangement that needs to be preserved? And what about that box full of Super 8 cartridges and VHS tapes with family videos on them, and the audio cassettes with recordings made at long-ago family gatherings?

A family history management system that includes many different resource types is a much bigger project than the one you contemplated when you first opened Grandpa's suitcase. It is easier to consider using separate but related organizing systems for each media type, because there are many web-based applications you could use. In fact, there are far too many choices of web applica-

tions for you to consider. You might compare some for their functionality and usability, but given the long expected lifetime of your organizing system there are more critical considerations: whether the site is likely to last as long as your collection and, if it does not, how easily you can export your resources and resource descriptions.[647][Web]

Why is it being organized? The overall goal of preserving Grandpa's photos needs no justification, but is preservation the primary goal? Or, rather, is to enable access to the images for far-flung family members? Or is it to create a repository for family photos as they continue to be produced? Alternatively, is it less about the images themselves and perhaps more about collecting family history information contained in the photos, thus making the collection of metadata (accurate information about when and where the photo was taken, who is in it, etc.) most important?

These decisions determine requirements for the interactions that the photo organizing system must support, but the repertoire of interactions is mostly determined by the choice of photo storage and sharing application. Some applications combine photo storage in a cloud-based repository tied to a very powerful set of digital photography tools, but this functionality comes with complexity that would overwhelm your less technology-savvy relatives. They would be happy just to be able to browse and search for photos.

How much is it being organized? Because you realize that a carefully designed set of categories and a controlled tagging vocabulary will enable precise browsing and search, you chose an application that supports grouping and tagging. But not everyone should be allowed to group or tag photos, and maybe some of the more distant relatives can view photos but not add any.

Will your categories and tags include all of those that Grandpa used when he arranged pictures in albums and made notes on the back of many of them? Do you want to allow annotations? Maybe this is a picture from a vacation; if you go back to the same place, do you want to create an association between the pictures?

Do not forget to keep Grandpa's original albums in a safe place, not under a bed somewhere.

When is it being organized? Once you create your categories and tags, you can require people to use them when they add new photos to the collection. Perhaps the existing resource descriptions can be completed or enhanced as a collective activity at a family reunion. Do not put this off too long—the people who can identify Grandpa's sister Gladys, her second husband, and his sister in an uncaptioned photo are getting on in years.

How or by whom is it being organized? You have taken on the role of the editor and curator, but you cannot do everything and having a group of people in-

volved will probably result in more robust organizing. A group can also better handle sticky situations like what to do if people get divorced or have a falling out with other family members; do pictures taken of or by them get deleted?

Other considerations. Maintenance of this collection for an indefinite time raises the important issue or a succession plan for the curator. If only one name is on the account and only that person knows the password, you run the risk of losing access to the photos if that person dies. One of Grandpa's mistakes was dying without clearly specifying his intentions for his photo collection, so whatever you decide you should document carefully and include a continuity plan when you are no longer the curator.[648][Law]

12.2 Knowledge Management for a Small Consulting Firm

Overview. A senior professor who has done part-time consulting for many years is very pleased when his latest book becomes a best-seller and he is inundated with new consulting opportunities. He decides to take a two-year leave of absence from his university to start a small consulting firm with several of his current and former graduate students as his junior consulting partners.

An organizing system for knowledge management is required, but what gets designed will depend on the scoping decision. Is the goal of the system to support the consulting business, or also to support ongoing and future research projects that sooner or later will generate the consulting opportunities?

What is being organized? The professor concludes that since his consulting is based on his research, he needs to include in the new knowledge management system his research articles and the raw and analyzed data that is discussed in the articles. These resources are already organized to a great extent according to the research project that led to their creation. These have been kept in the professor's university office.

The professor also has a separate collection of consulting proposals, client reports, and presentations that he has made at client firms. Because of restrictive university rules about faculty consulting, the professor has always kept these resources in his home office rather than on campus.[649][Law]

In addition to these existing resource types, it will be necessary to create new ones that make systematic and explicit information that the professor has managed in an informal and largely tacit manner. This includes consulting inquiries, information about prospects, and information about specific people in client firms.

Why is it being organized? The professor has usually just done one consulting project at a time, very opportunistically. He has often turned down projects that involved more work than he could do himself. He now sees the opportunity to do

much more consulting and to take on more significant projects if he can leverage his expertise in a more efficient way.

The professor can take on the "rainmaker" role to secure new consulting engagements and make the important decisions, and he is confident that he can train and support his new staff of current and former students to do much of the actual consulting work.

The knowledge management system must enable everyone in the firm to access and contribute to project repositories that contain proposals, plans, work in progress, and project deliverables. Much of this work can be reused from one project to another, increasing the productivity of the firm and the quality of its deliverables.[650][Bus]

Just as it is essential that the professor's knowledge is systematized and made available via a knowledge management system, so must the knowledge created by the new staff of consultants. The professor cannot expect that all of the students will work for him forever, so any knowledge that they acquire and create in the course of their work will be lost to the firm unless it is captured along with the professor's.

The consulting firm probably will not have an indefinite lifetime. After his leave of absence, the professor might return to his university duties, perhaps on a part-time basis. The knowledge management system will enable him to leave the firm in someone else's hands while enabling him to keep tabs on and possibly contribute to ongoing projects. Alternatively, if the firm is doing very well, perhaps the professor will resign his university position and take on the role of growing the firm. A larger consultancy might want to acquire the professor's firm, and the firm's valuation will in part be determined be the extent to which the firm's capabilities and resources are documented in the knowledge management system.

How much is it being organized? A small firm has neither the money nor the people to invest in complex technology and a rigorous process for knowledge management, but appropriate technology is readily available and affordable. Decisions about organizing principles must be made that reflect the mix of consulting projects; resources might be organized in a shared file system by customer type, project type, the lead consultant, or all of these ways using a faceted classification approach.

Standard document templates and style sheets for the resource types created by consultants can be integrated into word processors and spreadsheets. Contact and customer management functionality can be licensed as a hosted application.

Many small teams make good use of wikis for knowledge management because they are very flexible in the amount of structure they impose.[651][Bus]

When is it being organized? The professor's decision to take a leave of absence reflects his belief that getting the firm started quickly is essential if he is to capitalize on his recent bestselling book to generate consulting business. This makes managing the prospect pipeline and the proposal-writing process the highest priority targets for knowledge management.

Much of the other organizing work can emerge as adjuncts to consulting projects if some effort is made to coordinate the organizing across projects.

How or by whom is it being organized? Because many of the early organizing decisions have implications for the types of customers and projects that the firm can take on, only the professor is capable of making most of them. The principal goal of the knowledge management system is to enable the professor to delegate work to his consulting staff, so he needs to enlist them in the design of the organizing system to ensure it is effective.

Other considerations. As the consulting firm grows, it is inevitable that some consultants will be better than others at creating and using knowledge to create customer value, and they will expect to be compensated accordingly. It is essential for the ongoing success of the firm not to let this create disincentives for knowledge capture and sharing between consultants. The solution is to develop a company culture that promotes and rewards them.[652][Bus]

12.3 Smarter Farming in Japan

Overview. Unlike the first two case studies, this is an actual case rather than a hypothetical or composite one. It shares with the first two cases a focus on preserving valuable resources but in the radically different domain of farming.

This case concerns an initiative by Fujitsu, a Japanese technology firm, to apply "smart computing" and lean manufacturing techniques to the agricultural sector, which lags in technology use. Fujitsu is testing a "farm work management system" at six Japanese farms. In this case study we will focus on the farm highlighted in a 2011 Wall Street Journal story.[653][Bus]

This test farm is located in southern Japan. It has 60 different crops spread over 100 hectares (about 250 acres), an area slightly larger than the central campus of the University of California at Berkeley.

What is being organized? Sensors are deployed in each of 300 different farm plots to collect readings on temperature, soil, and moisture levels. Video cameras also monitor each plot.

The 72 relatively unskilled workers on the farm are also managed resources. Each of them carries a mobile phone for communication, transmission of pictures, and GPS tracking of their location.

Why is it being organized? The highest-level goal for Fujitsu is to expand its reach as a technology firm by applying the concepts of lean manufacturing, statistical process control, and continual improvement to new domains. Farming is an obvious choice in Japan because it is a relatively unproductive sector where the average age is over sixty. It is essential that farms use more computing capability to increase efficiency and to capture and reuse the scarce knowledge possessed by aging workers.

The Fujitsu farm work management system supports numerous types of interactions to achieve these goals. For example, workers can send pictures of infected crops for diagnosis by an expert farmer in the farm's office, who can then investigate further by studying recorded video from the affected plot.

As more farms deploy the Fujitsu system, the aggregated knowledge and sensor information can be analyzed to enable economies of scale that will allow separate and widely distributed farms to function as if they were all part of a single large firm with centralized management.[654][Bus]

How much is it being organized? The current design of the system treats farm workers as relatively passive resources that are managed very closely. The system generates a daily schedule of planting, maintenance, harvesting, and other activities for each worker. At a daily wrap-up meeting the farm manager reviews each worker's performance based on GPS and sensor readings.

The sensor data is analyzed and organized extensively by Fujitsu computers to make recommendations, both agricultural ones (e.g., what crop grows best in each plot and the work schedule that optimizes quality and yield) and business ones (the profitability of growing this crop on this plot of land).

When is it being organized? The farm work management system is continually organizing and reorganizing what it knows about the farm as it analyzes sensor and production information. In contrast, the information created by the workers is captured but its analysis is deferred to an expert.

It is conceivable that as the farm workers become more expert as a result of the guidance and instruction they receive from the system that they can be more autonomous and do more analysis and interpretation on their own. It is also likely that the inexorable forces of Moore's law will enable more data collection and more processing of the sensor data at its time of collection, which might result in increased real-time information exchange with the workers.

How or by whom is it being organized? The physical organization of the farm, with 300 small plots of land with 60 different fruits and vegetables, is the legacy arrangement of the farm before the Fujitsu trial began. Because of the sizable investment that Fujitsu has made in the farm to deploy the system, it is likely that the farm manager defers to recommendations made by the system to change crop arrangements. So it is reasonable to conclude that most of the de-

cisions about the organizing system are made by computational processes rather than by people.

Other considerations. Fujitsu built this system for managing farms, but there are several other resource domains with similar challenges about capturing and reusing operational knowledge: vineyards, forests, and fish farms come to mind.[655][Com] It will be interesting to see if the farm work management system can be made more abstract and configurable so that the same system can be used in all of these domains.

Farm crops, vineyards, trees, and fish pens do not move around, so a more challenging application of sensor technologies arises with cattle herd management. Nevertheless, sensors inserted in the genitals of a female dairy cow can trigger a text message to a herd manager's cell phone when the cow is in heat, preventing the economic loss of missing a reproductive cycle.[656][Ling]

Somewhat more remote domains for potential application of systems that combine sensor networks with workforce management include sales, field support, and logistics.

12.4 Single-Source Textbook Publishing

Overview. The fourth case is also an actual case—a self-referential one. It is a case study about the organizing system involved in the creation, production, and distribution of *The Discipline of Organizing*. See (Glushko 2015).

We have known since the beginning of this project that this book should not just be a conventional text. A printed book is an intellectual snapshot that is already dated in many respects the day it is published. In addition, the pedagogical goal of *The Discipline of Organizing* as a textbook for information schools and similar programs is made more difficult by the relentless growth of computing capability and the resulting technology innovation in our information-intensive economy and culture. We think that the emergence of ebook publishing opens up innovative possibilities as long as we can use a single set of source files to produce and update the print and digital versions of this book.

What is being organized? The content of this book began in early 2010 as more than 1000 slides and associated instructor notes for a graduate course "Information Organizing and Retrieval" that Robert J. Glushko, the primary author and editor of *The Discipline of Organizing*, was teaching at the University of California, Berkeley. These slides and notes were created in XML and transformed to HTML for presentation in a web browser.[657][Com]

The first decision to be made about resource organization led to the iterative sorting of the slides from 26 lectures into the 10 chapters in the initial outline for the book. The second decision concerned the granularity of the new content

resources being created for the book. The team of authors was organized by chapters, which made chapters the natural granularity for file management and version control. Because authors were widely dispersed we relied on the Dropbox cloud storage service to synchronize work. Nevertheless, the broad and deep topical coverage of the book meant that chapters had substantial internal structure (four levels of headings in some places), and many of these subsections became separately identified resources that moved from chapter to chapter until they found their natural home.

In addition to the text content and illustrations that make up the printed text, we needed to organize short videos, interactive examples, and other applications to incorporate in digital versions of the book.

Finally, it has been essential to view the software that transforms, assembles, formats, and assigns styles when turning source files into deliverable artifacts as resources that must be managed. For the first and second editions of the book, we were fortunate to get much of the software required to build both print and ebooks from O'Reilly and Associates, an innovative technology publisher that has been developing a single-source publishing system called Atlas. Because we have recently been experimenting with including richer interactivity and navigation capability, reader-controlled personalization, and other features that go beyond what Atlas enables, we now use our own custom-built single-source publishing system.

Why is it being organized? Publishing print and ebook versions of a text from the same source files is the only way to produce both in a cost-effective and maintainable fashion. Approaches that require any "hand-crafting" would make it impossible to revise the book on a timely schedule. Furthermore, a survey of Berkeley students in the summer of 2012 revealed a great diversity of preferred platforms for reading digital books that included laptop computers, Apple and Android tablets, and seven different dedicated ebook readers. Only an automated single-source publishing strategy could produce all these outputs.

The highly granular structure for the content resources that comprise this book makes cross-referencing vastly more precise, making it easier to use the book as a textbook and job aid. It will also make it easier to maintain and adapt the text for use in online courses. (The emerging best practice for online courses is to break up lectures and study content into smaller units than used in traditional classroom lectures.)

How much is it being organized? The nature and extent of resource organization for this book reflects its purpose of bringing together multiple disciplines that recognize organizing as a fundamental issue but from different perspectives. The book contains many specialized topics and domain-specific examples that might overwhelm the shared concepts. Our solution was to write a lean core text and to move much of the disciplinary and domain-specific content into tag-

ged endnotes. These categories of endnotes are somewhat arbitrary, but the authoring task of identifying content to go into endnotes is a non-trivial one.

The extent of resource organization is also affected by the choice of XML vocabulary, and we carefully considered whether to choose DITA or DocBook. DITA has the benefit of having more native support for modular authoring and transparent customization and updating, but DocBook is much older and hence has better toolkits. We eventually chose DocBook.[658][Com]

When is it being organized? Despite the fact that the lecture notes with which the book began were in XML, we decided to author the book using Microsoft Word. Many of the authors had little experience with XML editors, and the highly developed commenting and revision management facilities in Word proved very useful. This tradeoff imposed the burden of converting files to XML during the production process, but only two of the authors were still working on the book at that stage, and both have decades of experience with hypertext markup languages.

How or by whom is it being organized? The chapter authors used Word style sheets in a careful manner, tagging text with styles rather than using formatting overrides. This enabled a conversion vendor to convert most of the book from Word to XML semi-automatically. Some cleanup of the markup is inevitable because of the ambiguity created when the source markup with Word styles is less granular than the target markup in XML. We do not know whether the amount of work left for us was atypical.

Nevertheless, waiting until the book was substantially finished to convert to XML meant that we were also deferring the effort to mark up the text with cross references, citations, glossary terms, and index entries, because these types of content were not included in the Word authoring templates and style sheets. As a result, a substantial amount of effort has been required of our copy and markup editor that could have been done by chapter editors if they had authored natively in XML. However, having a single markup editor has given this book a more consistent and complete bibliography, glossary, and index than would be have possible with multiple authors.

Other considerations. Because every bit of content in the book is tagged as either “core” or discipline-specific, our source files collectively represent a “family of books” with 2048 different members, any one of which we can build by filtering the content to include any combination from zero to eleven disciplines. It is impractical to publish this many editions, but we hope to use this flexibility to enable instructors to tailor the text for a wide range of courses in many different academic disciplines and customize the text for both graduate and undergraduate students. Better still would be an approach that defers the generation of a particular version of an ebook from “publishing time” to “reading time.” The same algorithms apply, but now the reader decides when and how to apply

them, enabling the dynamic configuration of the book's content. This radical capability is experimental as of August 2015, but we expect it to generally available before too long.

This design for a book challenges conventional definitions of book editions and forces us to imagine new ways to acknowledge collaborative authorship. But asking "What is *The Discipline of Organizing*?," given these new authoring and publishing models, is a similar question to the one asked in Chapter 4, "What is *Macbeth*?"

12.5 Organizing a Kitchen

By Emilie Hardman, April 2013.

Overview. Just about everyone has a kitchen in their home or apartment, and most kitchens contain many of the same resources. These include pots and pans, dishes, bowls, drinking glasses, silverware, and cooking tools of various kinds. Kitchens are also often the location for organizing food items, cooking ingredients, spices, wine, and other beverages. Kitchens also invariably contain refrigerators and freezers for storing prepared and preserved food.

The organizing system for a kitchen is highly influenced by the size, shape, and arrangement of the counters, cabinets, shelves, and other parts of the physical environment of the kitchen. A person building a new home might be able to design this kitchen environment, but most people treat this as a given and work within its affordances, often because there are limits to how much the physical environment can be easily changed.

What is being organized? Our wine, wine glasses, cocktail glasses and ingredients, as well as tea and coffee stuff were stored in the cabinet by the fridge, close to the center worktable so people could have easy access to them. Because of space limitations, this meant that our water glasses had to be somewhere else, but as we would usually put out water for dinner parties or have a pitcher and glasses on a tray when people came over, we thought this was reasonable, since the things people would most often be looking for and need easy access to for themselves would be these more social drinking glasses.

We also bought a freestanding worktable with a butcher's block and stainless steel for pastry and chocolate work, as well as extra counter space in general. It worked as a prep space and as an area to lay out finished dishes or drinks for people to serve themselves when we had parties.

Some kitchen tools were kept with the food items to which they applied: for example, the coffee and the coffee grinder, or the cutting board, toaster, bread knife and bread all together. Other tools were kept with like tools: potato peelers, julienne tools, knives, etc. This was probably because of the kind of flexibili-

Kitchen Organizing System

My kitchen. I did my annual deep kitchen clean and it deserved a picture.

(Photo by Emilie Hardman. Creative Commons CC-BY-SA-2.0 license.)

ty something like a potato peeler would have versus a coffee grinder; it also made more sense to put lots of these little things together in a drawer rather than leave them strewn out around the apples or potatoes.

Pots and pans had their own spaces and were stacked within one another; same with dishes. Most frequently used things were given preference over specialty tools.

Other things that were organized around the social dimensions of the kitchen were some food items and serving elements. For example, we used bowls to organize chocolate bars and treats that might easily be grabbed to set out and serve. Similarly, we kept stacks of serving bowls easily at hand so we could empty pretzels or tortilla chips, olives, etc., quickly and casually.

Why is it being organized? We wanted to emphasize a feeling of comfort and openness in our kitchen, so people would feel free to get what they wanted

when they needed it. It also had to work on a practical level to be an efficient work area in a small space, so those concerns had to be balanced as well.

When is it being organized? We ended up moving silverware at one point because friends would consistently open a particular drawer in our center work island to look for silverware. Initially, I had specialized tools in that drawer because they were what I would reach for when I was working on something like making chocolates, but because of the continuous confusion, we moved those tools to another drawer and put the silverware where people seemed to expect it.

The fridge and freezer was organized by type of food for orderliness, ease of access, and immediacy of knowing when we had. We have a pull-out freezer, so things could get a little hidden, but assuming no one had compromised the system, you would know it was frozen fruit all of the way down in one segment and flours in another.

Some food items demanded different placement or storage based on their ripeness, the season, etc. In August we might be overrun with tomatoes, for example, and the window sills would fill up with them, whereas we would usually put them in a bowl if there were just a few.

How or by whom is it being organized? I think one thing to sum up would be to say that my partner is a librarian and I am trained as an archivist. We both care about classification and public service, so as people who also entertain a lot, I think these very practical and intuitive systems of grouping things is a motivation.

My father, an engineer who in his retirement does a lot of woodworking, built two cabinets that would *just* fit into the space and provide more storage than the two upper cabinets and three base cabinets provided in the kitchen.

Other considerations. The whole kitchen was not organized around guests, though. We also arranged things to be practical for cooking and for space saving. Food in the cabinets was organized by general function: for example, there was a shelf of dried beans in jars, another of dried chilies and spices—things that give flavor. Spices were organized within that by general type in rows and then alphabetically within those rows. This was because the rows helped group things which might be likely used together (e.g., cinnamon, cloves, mace, nutmeg) and alphabetically because so many of them look the same from the outside; knowing that the oregano would necessarily be shelved before the thyme was useful. Beans, though, because they are more immediately identifiable, less used, and certainly not as often used in concert (as one would with spices), I was a little more loose with and sometimes just arranged to a general aesthetic preference; if we had heirloom money beans, I might have preferred to see them over the standard red lentils, for example.

12.6 Earth Orbiting Satellites

By Daniel Brenners, December 2014.

Overview. Twenty two thousand miles above our heads, a global race for orbital real estate is underway. A single circular orbit around the Earth, called the geostationary Earth orbit (GEO), is the only area in space that allows a satellite to remain in a fixed point in the sky above Earth's surface while it rotates.[1] This prime location allows for satellites to have consistent communication with the ground below. Satellite television, a $100 billion industry, relies on satellites within the GEO to broadcast signals to homes across the world. Global positioning systems (GPS) and military applications also depend on satellites within this thin ring around the Earth. Unfortunately, space is severely limited in the GEO, and tension is growing over who gets to send their satellites to this valuable parking lot in the sky. The principles used to organize which satellites get to be placed in the GEO have many unforeseen legal and sociopolitical complications. As room becomes limited, it becomes increasingly important to find a solution to the problem of multiple organizing agents competing to organize this system to support varying interactions.

What is being organized? The scope of resources being organized are the satellites being deployed to the GEO. These are physical objects that have been launched into orbit. The satellites are each unique and are able to provide a variety of interactions. The only unifying attribute that they share is that they are computers that are able to send and receive radio signals to and from Earth. To stay in orbit, they are also able to adjust their position with propulsion systems.

This organizing system is designed to manage a collection in which resources are continually added and removed. The International Telecommunications Union (ITU) records which portions of the orbit are already occupied.[2] Satellites cannot stay in the orbit forever, as they expend lots of energy performing computational processes and maintaining orbit, and eventually run out of power. The resources follow a lifecycle that is unique to each individual resource, but the timescale is typically one to fifteen years.[3]

Why is it being organized? Satellites are being organized in the GEO to support several interactions. The GEO allows satellites to move at the same rate as the Earth, giving it a stationary view of more than 40 percent of the Earth's surface. Such a view is ideal for broadcasting signals to large regions and performing remote sensing, such as weather forecasting. They also serve as crucial relay points to transfer telecommunications across the globe. Other interactions that these satellites provide include surveillance, scientific research, global positioning, navigation, and military reconnaissance.[3] Longitudinal positioning along the GEO shapes which interactions can occur and which users can interact with the

satellite. For instance, a satellite directly over the Atlantic Ocean may not be well suited to broadcast a television signal, but may be positioned to relay signals from North America to Europe.

The users are practically everyone on Earth. Civilians use geostationary satellites directly when they use GPS or need to have a call relayed to distant regions of the world. Commercial organizations, such as television providers, use these satellites to broadcast signals down to viewers. Geostationary satellites are also particularly useful for early warning systems used by the military to detect ballistic events around the globe.

How much is it being organized? If resources are able to be placed in the GEO, they are placed in a vacant slot that the applicant chooses, based on what types of interactions they want to support and what users they want interacting with the satellite. To prevent signal interference and collision, satellites need to be placed very far apart, leaving only 2,000 total orbital slots where satellites can be placed in the GEO.[4] The ITU uses a first-come, first-served organizing principle to decide which resources are placed into orbital slots, provided the applicant completes the lengthy application process.

The organization applying for the slot chooses where to place its satellite. The ITU catalogs these slots as degrees longitude, and includes other resource descriptions such as the name of the satellite, country of operator, types of users, mass, expected lifetime, and contractor.[3] Organizations choose to place satellites around the longitude of the Earth that the satellite is supposed to interact with. Since the latitude is fixed at zero degrees, countries with the same longitude but different latitudes (countries directly north or south of each other) must vie for the same slots.

When is it being organized? Satellites are added as soon as they can be approved by the ITU and launched into orbit. At the end of their life cycle, the Federal Communications Commission mandates that U.S. satellites are pushed into what is called the graveyard orbit, which is a few hundred kilometers outside of the GEO.[5] At this point, another satellite can be added to the vacant slot via the ITU application process.

How or by whom is it being organized? Many organizing agents are competing with each other to organize this system according to their own needs. Applications to occupy the GEO come from countries, scientific organizations, companies, and civilians. Satellite TV companies such as DirecTV, Dish Network, and Intelsat own a large number of the slots across the western hemisphere. Countries such as the United States, Russia, and the United Kingdom own a majority of the military satellites, and multinational European organizations own a large share of orbital slots as well.[3]

Other considerations. Although the ITU serves as an authoritative entity for this organizing system, the reality is that the ambiguous legality of ownership in outer space means that anyone can attempt to organize this system. The ITU is in place to perform the useful task of cataloging occupied slots and facilitating the filling of vacancies, but it has no way of enforcing these guidelines.

This organizing system is interesting because many agents are attempting to organize the same system. There are also interesting social implications that stem from the system's principles of organization. The first-come, first-served system of the ITU has the effect of allowing only technologically advanced organizations to manage the collection. It does not take into consideration that by the time many countries are finally ready to use this type of technology, there will be no more room in the GEO belt.

Ironically, the only legal claim to sovereignty that has been made of this organizing system has been from countries that, generally, have no means of organizing it themselves. In 1976 eight equatorial countries, which lie directly below the GEO belt, stated that they had exclusive rights over these slots in a document known as the Bogotá Declaration.[6] The tenuous claim was that the orbit is not a part of outer space, because its existence is solely dependent on Earth's gravity, and that the earth within the borders of the equatorial countries creates GEO with its gravitational pull. Many experts disagree, stating that the gravitational pull from the moon and other celestial bodies defines the GEO, and state that the orbit does indeed lie in outer space because it is further than 100 kilometers from Earth. This demarcation, known as the Kármán line, is a widely accepted definition of when space begins.[7] This would then make the GEO fall within the 1967 Outer Space Treaty, effectively leaving no possibility for ownership of the orbit.

Finding a dividing line between space and Earth's atmosphere is an interesting topic, especially considering that ownership of valuable resources may be decided based on what is included in the category of space versus the category of atmosphere. In this case, the Kármán line roughly represents the altitude at which an aircraft would have to propel itself faster than the speed at which the Earth rotates to establish enough lift to keep itself up. While this is not intuitive (hardly carving nature at its joints), it does serve as a useful demarcation that is not completely arbitrary. It can be seen as a goal-based category, where the goal is using traditional means of traveling through the air using aeronautics. It makes sense that this is the line the Fédération Aéronautique Internationale uses to divide astronautics and aeronautics.

The limited availability of spots in the GEO, along with the relatively small number of countries able to launch satellites, has the potential to further divide countries. By the time most countries will be able to launch satellites, there will likely not be any room left. Although there are only around 400 satellites cur-

rently in geostationary orbit, there are already more filings for ITU applications than there are spots available.[4] Only a select few countries will be able to take advantage of the GEO, leaving others to depend on these countries for communication, scientific research, and surveillance. Furthermore, this could limit the interactions of these less developed countries to those interactions dictated by the countries with geostationary satellites. In particular, these developed countries can greatly influence the information that citizens in other countries can receive via satellite.

But even within the technologically advanced countries, competition for orbital slots may be heating up. In early 2014, the US unveiled its Geosynchronous Space Situational Awareness Program (GSSAP), which aims to create maneuverable satellites that monitor and protect the precious GEO belt.[8] This reveal comes only months after China was seen practicing its anti-satellite missile capabilities.[9] In Russia, $300 million is being spent to construct a craft that would act as a "space broom" to push satellites out of geostationary orbit. The US has a similar program, named the Phoenix project under DARPA, developing a robotic device that can help maintain satellites and possibly dismantle others without causing excess space debris.

Although this might simply be countries attempting to flex their military muscles, these technologies represent a newfound ability for countries to organize resources in the GEO to fit their own agenda. Years ago, the countries that were able to get satellites into orbit were the ones that could reap the benefits. Now, it seems that we may be entering an age where a country's ability to make room for itself, possibly by force, will determine if it can make use of precious interactions created by these limited resources.

Notes: The following notes relate to this case study.

1. NASA Jet Propulsion Laboratory Basics of Space Flight Section 1 Chapter 5: Planetary Orbits http://www2.jpl.nasa.gov/basics/bsf5-1.php
2. ITU Space Services Department (SSD) 2014 http://www.itu.int/ITU-R/go/space/en
3. Union of Concerned Scientists Satellite Database http://www.ucsusa.org/nuclear_weapons_and_global_security/solutions/space-weapons/ucs-satellite-database.html#.VJKNXmTF-5I
4. Posen M., Have We Got a Slot? RPC Telecommunications Ltd. World Space Forum Dubai March 2010 http://www.rpctelecom.com/files/Have We Got A Slot.pdf
5. De Selding P., FCC Enters Orbital Debris Debate. Space News, 28 Jun. 2004
6. Finch M., Limited Space: Allocating the Geostationary Orbit. Northwestern Journal of International Law Vol 7 Issue 4 Fall 1986

7. Haraszti G., Questions of International Law Volume 2. Akademiai Kiado Budapest 1981
8. Hsu J., Global Conflict Could Threaten Geostationary Satellites: China, Russia and the U.S. have the ability to destroy one another's eyes in the sky. Scientific American March 31, 2014 http://www.scientificamerican.com/article/global-conflict-could-threaten-geostationary-satellites/
9. Shalal-Esa A. U.S. sees China launch as test of anti-satellite muscle. Reuters May 2013 http://www.reuters.com/article/2013/05/15/us-china-launch-idUSBRE94E07D20130515

12.7 CalBug and its Search Interface Redesign

By Gracen Brilmyer, December 2014.

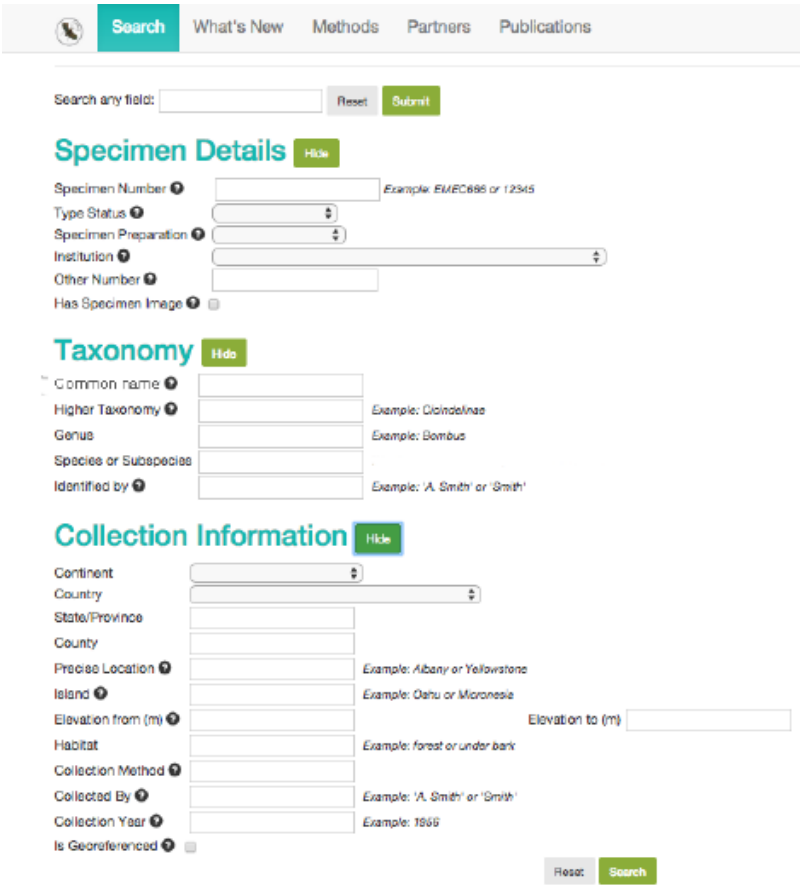
Overview. The CalBug project, housed out of the The Essig Museum of Entomology at the University of California, Berkeley, is a collaborative initiative between nine California institutions with a goal to digitize over a million specimens. Digitization involves imaging both specimens and their labels as well as storing their collection info in a database. The CalBug project also is attempting to georeference, or locate the original latitude and longitude coordinates, for these million specimens (some dating back to the 18th century) so that they can be better used for research. The project uses many student workers, graduate students, and volunteers to capture the images and data. Over the past few years, it has participated in the Notes from Nature project, which helps connect citizen scientists to scientific research. Through the images generated of the specimen labels by the team at the Essig Museum, citizen scientists digitally transcribe the data that can be read from the image. The Essig, after each label is transcribed by 24 citizen scientists, runs an R program to find the most accurate transcription and transfer it into the Essig's database. These combined efforts have accumulated in over 209,000 specimen records and over 400,000 images and counting. This project has a large scope and an ever-increasing scale.

What is being organized? The insect specimens in the CalBug project are digitized on an individual level, with unique identifying numbers, and new specimen records and their associated data are continually being added to the digital collection. Both the specimens and their data are being organized. Existing groups of specimens are prioritized for digitization and new physical specimens are accessioned into the collection and are databased upon arrival.

Why is it being organized? An individual specimen's associated data can be highly variable; however, as long as a specimen has the time and place of its collection (no matter how vague) associated with it, it is valuable research material. The physical specimens are organized to facilitate the collection manager's

use of the collection. When physical specimens need to be borrowed, they must be efficiently found, packaged, and sent out on loan, so fastidious organization is key when locating thousands of specimens. The digital organization of the collection also facilitates the duties of museum staff and the collection manager by allowing for expanded interaction with the collection by using the database. The digital collection's web interface, undergoing a redesign as of the time of this writing, makes the collection accessible for researchers and novices alike, as well as to foster data sharing to other data repositories. Since the specimen data follows digital curatorial standards, a web interface that allows these fields to be easily searchable and navigable can add to the use of the collection for a broader audience, which is a major impetus for the redesign.

Figure 12.1. CalBug search interface



CalBug's redesigned web search interface

How much is it being organized? As discussed in the previous section, the specimens and their information are subject to multiple levels of organization, and each level of organization supports a different type of user. The data of the CalBug Project is organized according to Darwin Core (DwC), a standard "designed to facilitate the exchange of information about the geographic occurrence of species and the existence of specimens in collections."[1] Certain specimen attributes have concrete institutional parameters, such as unique identifying numbers and taxonomic identification, while others have less strict parameters (e.g. a precise location of where a specimen is found), although they still must use specific DwC fields. Although there are institutional taxonomies in place for information associated with a specimen's collection and identification, the CalBug search interface design in Figure 12.1, CalBug search interface allows for an outward-facing reorganization of the existing fields.

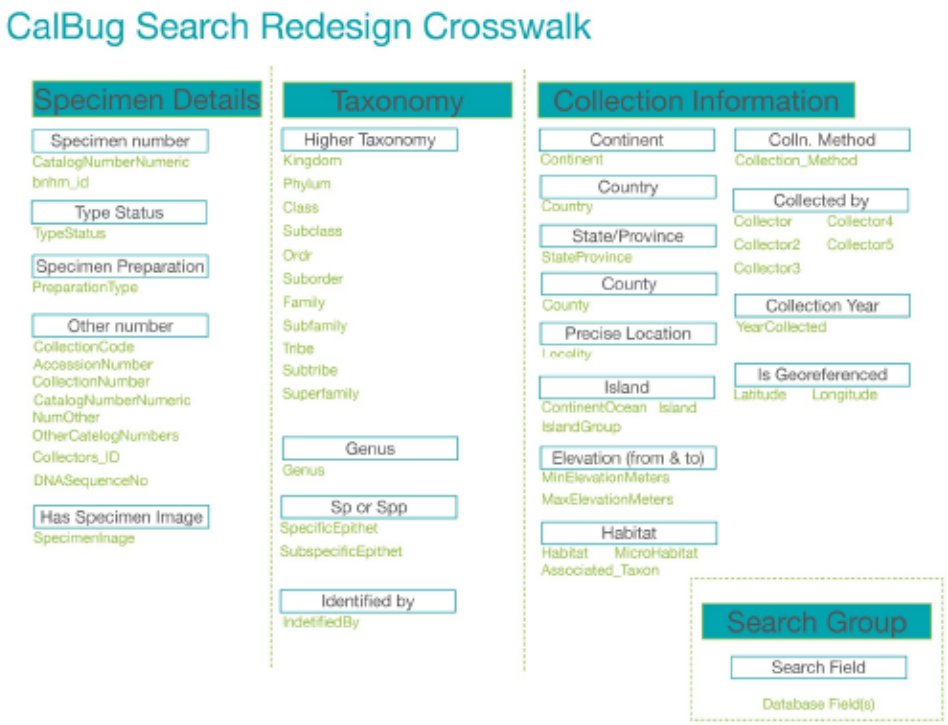
When is it being organized, and by whom? The categorization and organization happens at multiple times for one specimen. If identified, the specimen is already inserted into the taxonomic classification scheme—the hierarchy of how species are related. This scientific warrant is inherited and replicated in the physical curation of the collection, and specimens are further sorted (within a taxon) by geographic region. Aligning with taxonomic categories provides a clear hierarchy for sorting and locating physical specimens and, with changes in taxonomy having to be published, makes collection maintenance fairly consistent.

The specimens are organized a second time when they are databased, either by interns or through Notes from Nature. The data is stored in a MySQL database that uses mostly DwC fields, an institutional taxonomy for specimen data. The digitization of specimens, through utilizing DwC institutional semantics, makes collection maintenance, governance, and interaction easier, as the collection manager can search in a multifaceted manner, better understand the holdings of the museum, and track specimens for loans. The unique specimen numbers allow for individual tracking, and the other DwC fields provide multiple areas for accurate search and retrieval.

For the CalBug web search interface, the specimens retain their classification hierarchy within the database. However, the outward-facing search fields aim to serve a broader audience, not just the collection manager and museum staff. Thus the search application organizes the resources a third time "on the way out" of the database in response to a user query. As this design is optimized for researchers and students, the classification appears to focus more on taskonomy instead of the institutional taxonomy (see Figure 12.1, CalBug search interface). The 20 search fields provided in the search interface, while actually searching through the ~100 fields in the database, facilitate precise information retrieval. Although fewer search fields might yield lower accuracy, user testing

has shown that the new search design improves accuracy by not requiring users to know exactly which DwC field to query.

Figure 12.2. Crosswalk table



This crosswalk table maps the fields in the CalBug search interface to the underlying database columns.

The search is further expanded by having a 'Search any field' box, which literally looks in every DwC field for a term, as well as a "Common Name" field, to support novice searches, such as "beetle" and "butterfly" instead of "coleoptera" and "lepidoptera." The intrinsic properties of the specimens lend the results to simple (alphabetic and numeric) sorting as well as filtering (through the "Refine" option) on the list view of the results pages. Additional views of results, including a map view showing collection locations and a grid view that displays specimen photos, help users locate desired specimens and reorganize as needed to suit their needs.

Notes: 1. `http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome`

12.8 Weekly Newspaper

By Ian MacFarland, December 2013.

Overview. A weekly neighborhood newspaper in New York City now covers the entire borough of Queens. Rather than publish a single weekly edition for this highly diverse area of more than 2 million people, its owners have opted to produce 14 separate editions, each centered on a different neighborhood. All editions share a deadline, delivery schedule, and staff pool, but each has unique content tailored to its target readers.

What is being organized? The newspaper's resources—its content—consist mainly of articles and photos generated by staff and freelance contributors throughout the week. Often, newspapers will assign their reporters to beats based on subject matter (politics, education, "cops and courts," etc.), making them domain experts who cover stories on that beat throughout a wide geographical area. However, because of this paper's historical orientation toward "hyper-local" neighborhood news, it has given each of its seven full-time reporters a more granular geographical beat that corresponds to two of the 14 editions' coverage areas, within which they are responsible for general assignment reporting. Most reporters also have a specialty for covering news that is of more general interest throughout the borough, such as citywide government or transportation issues, and they will include coverage of these domains in their story budgets for the week as well. The staff maintains a centralized story list that includes a handful of resource descriptions for each story: its slug (an abbreviated, descriptive name, including tags for its relevant neighborhoods), its length, and whether it has "art."

Why is it being organized? The media market in New York is crowded and extremely competitive, and this newspaper believes its competitive edge lies in its laser-focus on individual neighborhoods. Furthermore, most of its readers are subscribers who receive the paper in the mail, not newsstand buyers. As a result, the paper generally eschews the familiar tabloid approach of splashing the most salacious story of the week across the front page and usually fronts two stories that are "small-bore" but extremely relevant to the neighborhood, such as the doings of local school or government officials, notable crimes, or human-interest stories featuring neighborhood residents. The deeper into the paper one goes, the less local its content becomes, and stories often appear in more than one edition, in different locations and even with different headlines, to tailor them to an appropriate level of localization.

On a more general level, of course, the paper must support the conventional interactions all readers expect from newspapers. Readers are rarely expected to progress through the paper from front to back, so it supports a wide variety of

reading styles; large headlines and photos and concise, compelling story "ledes" (opening paragraphs) facilitate skimming and scanning interactions, and dividing the paper into sections, such as "Opinion," "Sports," and "Arts & Entertainment," lets readers skip directly to their areas of interest after turning past page one.

How much is it being organized? The level of organization behind the scenes at this small, local newspaper is surprisingly complex. The primary organizing principle that determines a story's placement is its relevance, which is a function of location granularity (does it directly affect the people of this neighborhood? Did it happen here?), significance (will readers find it important?), and time (is it old news? Has anyone else reported it yet?). Counterbalancing that is the economic reality of the struggling newspaper industry, which results in often severely limited space for the news (because paper and press time are costly physical constraints) and manpower with which to produce all 14 editions before deadline. The result is a hierarchical system in which the 14 editions are categorized into three zones; in each zone, about two-thirds of the pages are common to all editions, and the remaining third (including, most crucially, pages one through three) are unique to each single edition. Thus, for instance, a general-interest story about transportation need not be laid out 14 separate times, but one about a fatal car accident can appear on page one for the neighborhoods where it occurred and where the victims were from, and further back (or not at all) for other neighborhoods.

When is it being organized? In a weekly news cycle, selection, creation, and organizing of editorial resources is largely concurrent. The story list is updated on a rolling basis throughout the week, and an article or photo's placement in the paper is often planned based on its intended subject matter well in advance of when the resource is actually created. However, organizing must be completed long before it reaches its intended users, because the final layouts must be printed, collated, and mailed to readers, which, due to logistical concerns, takes several days—so the paper is laid out on Tuesday (as late as possible to maximize the window for ad sales), printed on Wednesday, and delivered by the Postal Service on Thursday or Friday.

How or by whom is it being organized? Human agents—specifically, editors—are the newspaper's primary organizers. They rely heavily on the judgment of the reporters, who are most familiar with their beats, to determine a story's relevance and placement for each edition, as well as their own news judgment, assessment of the story's quality, and estimation of where the story will physically fit based on ad placements (which are decided first). The implementation of their organizing system is carried out by page layout designers, with some software automation on the part of the paper's content management system.

Other considerations. Part of the grind of a weekly news cycle is that the effectivity of the paper's resources is never guaranteed; when the next edition comes out, they all become yesterday's news, and one never knows when new developments will render a story irrelevant or incorrect; in fact, because of the latency between layout and delivery, a story's effectivity may even expire before its publication.

12.9 The CODIS DNA Database

By Becca Stanger, December 2013.

Overview. Operating on a local, state, and federal level, the Combined DNA Index System (CODIS) is the FBI DNA database. As of October 2013, the National DNA Index (NDIS), or the federal level of the CODIS, contained over 10,647,800 offender profiles, 1,677,100 arrestee profiles, and 522,200 forensic profiles. Designed to help solve crimes, this database has generated over 255,400 hits and has aided over 216,200 investigations. While this organizing system has played a crucial role in reducing crime by enabling more interactions in the law enforcement agency than ever before, it provokes numerous ethical questions worth exploring.

What is being organized? The CODIS database maintains digital records or "DNA profiles" for a wide range of people involved in criminal justice cases, including convicted offenders, arrestees, missing persons, and more. Specifically, these profiles are measurements of one or two alleles of 13 predetermined unique genetic sequence loci. These precise measurements provide enough granularity for the profiles to uniquely identify a single individual.

These resource descriptions are generated, often with polymerase chain reaction technology, from the original DNA specimen resources by accredited laboratories nationwide. Upon creation, the resources themselves—the specimens—are kept at the laboratories, while the resource descriptions—the digital profiles—are added to the CODIS database. No offender personal identifiers are assigned to the profiles; however, information on the submitting agency, specimen, and personnel is stored with the profile.

Rather than focusing on collecting resource descriptions, the FBI could have chosen to collect the original resources themselves. Presumably, though, this level of coordination of physical resources (e.g., shipping, storage, maintenance, etc.) would have placed an additional cost on the federal government and required legislative approval. Thus, it is understandable that the FBI would choose to minimize cost and effort by focusing on the resource descriptions alone.

Why is it being organized? In the past, law enforcement agencies were limited to solving crimes within their geographic region. A detective working on a mur-

der in California, for example, may never have heard of a related murder in New York. The CODIS database organizing system encourages that coordination between law enforcement agencies in an effort to solve crimes.

With 10,647,800 offender profiles in the NDIS alone, though, the massive CODIS database required an organizing system in order to prove useful to the law enforcement agencies involved. The successful creation and maintenance of this organizing system has offered newfound interactions to a wide variety of government officials. In addition to law enforcement agencies, judicial courts, criminal defense agencies, and population statistics agencies can access the CODIS organizing system, enabling them to perform a wide variety of functions, including identifying potential suspects in criminal investigations, identifying missing persons, collecting population statistics, and exonerating convicted criminals.

How much is it being organized? As mentioned previously, the high degree of resource description granularity in measuring 13 specific genetic sequence loci enables DNA profiles to uniquely identify each individual in the database. That being said, the DNA profiles are not simply heaped into one massive database.

Instead, the databases are maintained on both a state and federal level. A new profile might be checked against a smaller state database as well as the larger national one. In addition, the databases are divided into different indices dependent on the DNA source, including an offender index, arrestee index, forensic index, and missing persons index.

This division of the database into separate indices poses a tradeoff dilemma, though. If CODIS did not subdivide the database into federal, state and source indices, it is possible the algorithm would be able to find more obscure hits, since the search parameters would be broadened. This increase in hit frequency might result in more investigations aided.

The tradeoff, however, is that the broadened search parameters would also necessitate a more complex search algorithm and a longer search process. This delay would most likely lead to fewer hits overall. Thus, in government institutions where time and resources are limited, it is more important for the CODIS organizing system users to generate a larger number of hits with subdivided databases than more accurate hits in one collective database. Categories in the CODIS organizing system help simplify the interaction processes.

When is it being organized? DNA profiles enter the CODIS organizing system when participating, accredited local, state, and federal laboratories submit them. Thus, the laboratory technicians handling the resource and resource description decide on a case-by-case basis how a given profile should be categorized and which indices it should be added to and checked against.

That being said, the lab technicians are given strict standards on how a given DNA profile should be categorized. These standards vary state by state depending on state law.

How or by whom is it being organized? Beyond laboratory and state involvement in CODIS, the FBI ultimately maintains and oversees the CODIS database. It maintains the software and search algorithms, performs searches throughout the system, and oversees strict quality assurance standards for all participating laboratories.

To avoid the risk of bias or error amongst lab technicians, the FBI could potentially choose to instead perform the laboratory processing and categorization themselves. This alteration, however, would present new challenges, such as new federal costs related to maintaining and processing the resources mentioned previously. In addition, pulling together all resources into a FBI processing center would necessitate a meticulous record of the resource's originating state to ensure resource descriptions are categorized in accordance with state laws. The FBI's strict maintenance of standards and laws is the best option for addressing the risk of error and bias.

Other considerations. The CODIS organizing system presents a wide range of intriguing ethical questions surrounding race, gender, criminal justice, and privacy. Perhaps the most hotly debated issue surrounding DNA databases arose when the private DNA testing company 23andMe announced that it would discontinue the sale of its genetic tests in response to FDA demands, prompting more media questions than ever before on the maintenance and use of DNA databases.

Likewise, many have questioned the legitimacy of the CODIS maintenance of DNA profiles. The ACLU, for example, has noted the possibility of "function creep" in the maintenance of a government DNA database which could lead our country down a slippery slope towards a "brave new world" where private genetic information could be collected and used in abusive, discriminating manners.

With the commercial surveillance of 23andMe and government surveillance by the NSA at the forefront of media attention, it is possible we will see more attention turned to the legitimacy of the maintenance of the CODIS organizing system in the coming years.

12.10 Honolulu Rail Transit

By Carlo Liquido, December 2015

Overview. The Honolulu Rail Transit Project is an urban rail rapid transit system under construction in Honolulu on the island of O'ahu, Hawaii. Honolulu's notoriously bad traffic has plagued locals and tourists for decades, and for almost as long, proposals to address the traffic problems and pay for the solution have been very contentious and political. Construction began in 2011 and is expected to finish in 2019, but delays have been frequent.

What resources are being used? The new railway transit system under construction in O'ahu will run along the southwest region of the island spanning a total of 20 miles, from East Kapolei to Downtown Honolulu with a total of 21 stops strategically placed throughout. There are a number of ways in which one could scope this project. What are the cultural and political limitations? What are the environmental effects and resources that will be indirectly affected? What are the topographic constraints of a railway system in Hawaii? In terms of the scope of my analysis, however, the people—namely the things the organizing system is intended for—are the primary resources. The principle guiding the organizing system is to reduce traffic and make the traveling experience more efficient as a whole.

O'ahu Traffic

O'ahu traffic is usually congested, especially in and around Honolulu on the south and southeast sides of the island

(Hawaii Dep. of Economic Development and Tourism. CC-BY-2.0.)

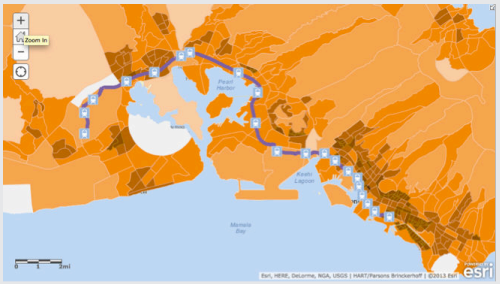
Why are the resources organized? The guiding principle behind the organizing system of a rail transit system is to reduce traffic and make commuting more efficient. According to the Department of Business, Economic Development and Tourism, the amount of traffic on almost every major highway on O'ahu has increased from 2012-2014. Moreover, the dearth of job creation on other parts of the island, namely the west side, has focused traffic into and out of downtown Honolulu, as shown in the first map.

This skewed traffic pattern, limited real estate, and inflexible road infrastructure has necessitated an above-ground railway system linking the west side of O'ahu with the burgeoning downtown area of Honolulu. This new organizing system seeks to rebalance the traffic system by reorganizing its resources, that is, by taking drivers and bus commuters

off the road and onto the rail. O'ahu has only three major freeways, the H1, H2, and H3. The freeway H2 bottlenecks from the west into H1. Drivers and bus commuters are organized in such a way that peak hours of traffic are unavoidable. The new transit will conceivably provide an additional layer of organization to the currently

How much are the resources organized? There are 21 planned stations that run along the 20-mile span of track. The train stations are arranged to serve as many people as possible by concentrating them in the most densely populated areas.

Population Density



Population in Honolulu area is highly concentrated in lowland areas.

(Hawaii Dep. of Economic Development and Tourism. CC-BY-2.0.)

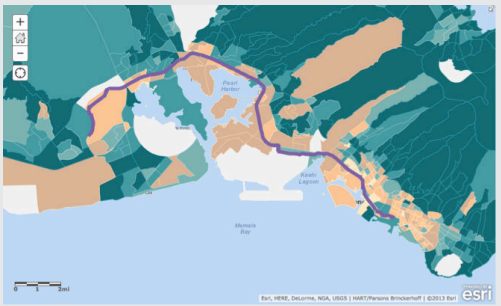
Darker areas represent high-density tracts while lighter areas represent low-density tracts. The densely-populated stretch from Keahi Lagoon to Honolulu Downtown, also has the highest density of traffic. It makes sense that this portion of the rail system constitutes almost half the number of total stops in just a quarter of the total mileage.

Income per household also plays a vital role in how these stops were selected. The rail transit system predominately runs along areas of low-income neighborhoods (tan and brown indicates low income per household, while green indicates high income per household). This design principle embodies an assumption that people with lower incomes are more likely to rely on public transit.

When are the resources organized? As with any construction project of this magnitude, the organizing system was planned in detail before construction—down to the number of pillars, the amount of concrete, the imported steel for rail cars, etc. However, after construction excavation revealed ancient burial sites, the Native Hawaiian community demanded many changes to the project. The number of stops has remained the same but the route has changed dramatically.

How or by whom is it being organized? There are a number of interested parties with varying degrees of power. At the forefront, the government—that is, the State of Hawaii—makes the final decision. However, the people of Hawaii directly influence their decisions.

Honolulu Area Income per Household



Household income is lowest in the most densely populated areas.

(Hawaii Dep. of Economic Development and Tourism. CC-BY-2.0.)

The protection of cultural resources, practices, and beliefs is important in Hawaii, both as a matter of law and of culture. Private archeology firms, state officials, and cultural descendants work together to reduce and mitigate impacts to archaeologically significant properties. The Oahu Island Burial Council, for instance, is a state council created to help protect *iwi kupuna* (ancestral bones). It stresses the importance of consulting recognized lineal descendants before any excavation for the rail project is carried out.

Where is it being organized? The "where" component of the organizing system is not as important for the scope of this analysis as other design questions. However, the physical nature of the project highly constrains how the system can be organized. The volcanic origin of O'ahu, does not allow for a below-ground rail system. The limited real estate, similarly, does not allow for a ground-level system. The sharp and steep volcanic ridges that cut across the island are barriers that limit where the rail system might go.

12.11 The Antikythera Mechanism

By Murray Maloney, 2 March 2014.

Overview. In 1900, a strange looking mechanical device was recovered from a shipwreck off of the island of Antikythera, Greece. Only in the 1970s was it determined that the device was an ancient mechanical computer that performed astronomical calculations; it had a manual crank control with a rate of one turn per day, forward or backward in time; its user interface presented calendrical, solar, lunar, and planetary positions.[1]

The Antikythera Mechanism persists through time as a collection of artifacts and a model of intellectual achievement. Thought to have been constructed by Archimedes at Syracuse or by Posidonius at Rhodes, the mechanism was recovered from a ship wreck near the Greek island of Antikythera in 1900-1. The significance of the find only began to become apparent in the 1970s when researchers applied modern scanning technology.[2]

What is being organized? The Antikythera Mechanism was an arrangement of resource descriptions that represented a classical Alexandrian sol-lunar calendar, complete with an almanac of the positions of the sun, moon, known planets, and specific stars over time. The resource descriptions are represented simply by the measurements of the gears, and the corresponding information that is displayed on the front and rear panels, based on the position of those gears. These resource descriptions accounted for the range of known astronomical phenomena at the time.[3]

The organization of the mechanism consists of a main solar gear connected to a hand crank and a collection of gear trains that ultimately control the rotation of pointers indicating the calendar, lunar position and phases, the position of the sun and of all the known planets, and the nearest eclipse. The mechanism was housed in a wood frame box with bronze panels whose physicality was obviously intrinsic to the use of the device; the panels the back door was inscribed with what seems to be a user's guide.

The Antikythera Mechanism

The Antikythera Mechanism exhibit at the National Archaeological Museum of Athens.

(Photo by Tilemahos Efthimiadis. CC-BY-2.0 license.)

The Antikythera Mechanism calculated the position of the moon by employing five gear trains to take into account the Saros, Metonic, Callipic, and Exeligmos cycles. Thus, it was able to predict the dates of solar and lunar eclipses.

Today, the Antikythera is a collection of the eighty-two fragments that have been recovered from the ship wreck and sea bed, twenty three of which are evidently inscribed. The fragments have been dated to about 70 BCE based on the coincident presence of some coins from Pergamum and Ephesus that were recovered in the 1970s.

Why is it being organized? From a purely pragmatic perspective, the Antikythera Mechanism was a relatively portable computational device. that would have been used to accurately reckon a very specific calendar system, and to predict the cycles of days, months, years, and saro, as well as lunations, eclipses, and Olympic games. It would be an invaluable tool for astronomers, mathematicians, civil engineers, and cartographers of the time.

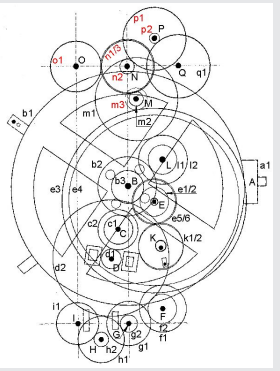
From a philosophical perspective, the Antikythera Mechanism was built to prove that it could be done. It represents a fulfillment of Aristotelian thought. Through the ages, the lure of scientific answers to the mathematical riddles presented among the patterns in the heavens has challenged our burgeoning intellects. The Antikythera Mechanism realized then-modern thinking on mathematics, engineering, astronomy and calendrical calculation in a portable mechanical computational device.[4]

How much is it being organized? Some of the major fragments are on display at the National Archaeological Museum of Athens; the others are stored.

The Antikythera Mechanism is reported to have had about thirty gears within a frame whose size was less than the volume of a large book. The level of miniaturization and the precision of fabrication was not thereafter seen until the next millennium. The engineering and machining would have required trial models, accurate plans, and custom tooling. There have been various modern attempts to re-create the Antikythera Mechanism, or at least to re-create the model it seems to have manifested.[5]

When is it being organized? The person who operates the mechanism turns a hand-operated crank to establish a date, or contra-wise confirms the current date by taking sightings and comparing with the dial settings. The front face offers a solar-lunar calendar dial, a tropical zodiac dial, and an almanac dial with rising and setting times of various stars. The rear panel offers dials representing the five lunar cycles.

The Gears



Gear arrangement.

(Wiki Commons.)

The organization of the engineering data required to build, operate, and maintain the Antikythera Mechanism is staggering to imagine, yet it pales in comparison to the organization required to collect and archive astronomical sightings on clay tablets for hundreds of years.[6] (See the sidebar, A Cuneiform Document at the Pergamon (page 180).)

The organization of the fragments of the Antikythera Mechanism is in the hands of the Bronze Collection of the National Archeological Museum in Athens.

How or by whom is it being organized? Ancient Chaldean, Greek, and Roman astronomers and engineers; modern divers, marine archaeologists, curators and researchers. In 1978, Jacques Cousteau led an expedition to the sea bed and returned some historical artifacts, that, while unrelated to the Antikythera Mechanism itself, provide additional historical context and may help date the discovery.

The Antikythera Mechanism Research Project is a collaboration of academic, industrial, and scientific researchers, who are applying some of the world's most advanced technology to study the capabilities and applications of the Antikythera Mechanism, as well as its historical context and significance.

Other considerations. From the perspective of one ship's unlucky captain and crew, the Antikythera Mechanism was likely just a piece of cargo, although it may have accompanied an equally unlucky passenger carrying the world's first computer to Caesar's court in Rome. It remains unknown how or why the device was aboard the ship or what fate befell it, but that is a story for researchers and historians to uncover in the fulness of time.

The Antikythera Mechanism

A recreation of the Antikythera Mechanism on display at the National Archaeological Museum of Athens.

(Photo by Tilemahos Efthimiadis. CC-BY-2.0 license.)

Notes: The following notes relate to this case study.

1. PBS aired *Ancient Computer* on April 3, 2013. The BBC aired Ancient Moon 'computer' revisited
2. The Antikythera Mechanism Research Project recently published *The Inscriptions of the Antikythera Mechanism. 2016.* Y. Bitsakis, M.G.Edmunds, A. Jones, *et alia* Almagest 7-1, May 2016
3. Cicero wrote about a similar device, created by Archimedes, in M. Tvlli Ciceronis de Republica Liber Primvs

Gears from the Greeks. The Antikythera Mechanism: A Calendar Computer from ca. 80 B. C. Derek de Solla Price Transactions of the American Philosophical Society New Series, Vol. 64, No. 7 (1974), pp. 1-70

4. Aristotle's work on the subject *On the Heavens* (c 350 BCE) avers to the mathematical symmetry and perfection in the travels of the spheres, envisioning cycles and epicycles in motion.

   In 343 BCE, Aristotle was head of the Macedonian Academy, where he tutored Alexander and his future general, Soter Ptolemy. Following Alexander's conquest of Babylon in 331 BCE, he ordered Kallisthenes to organize the translation of all historical astronomical observations, initiating the transfer of the world's greatest collection of astronomical observations, dating back to 747 BCE. Within a year, Callippus had developed a new calendar, designating the summer solstice of 330 BCE as an epoch for astronomers and calendrical calculation. The Callipic cycle of 76 years less a day, equates to 27,759 days, and 940 lunations, is represented in the gearing of the mechanism.

   Ptolemy established his capital at Alexandria and founded a museum, spawning the need for a library, in the Platonic style. His successors, through to Cleopatra, added to the papyrus rolls. Mathematicians, astronomers, mechanical engineers, scientists; the most famous thinkers of the ancient world studied in the halls of the Library at Alexandria. Notable to us in this context are Euclid, Archimedes, Eratosthenes, Hipparchus, Aristarchus, and Posidonius.

   According to Pliny, the calendar reform of Julius Caesar, was assisted by Cleopatra's astronomer, Sosigenes, of Alexandria, who "brought the separate years back into conformity with the course of the sun."

5. In 2010, Andrew Carol built a Lego model of the Antikythera Mechanism on a dare. John Pavlus wrote and directed a short film, Behind the Scenes: Lego Antikythera Mechanism.

   Hublot, the Swiss maker of luxury time pieces, created a special edition Antikythera Watch. Hublot is also sponsoring ongoing research. See Return to Antikythera: A project of the Hellenic Ministry of Culture and Sports with support from the Woods Hole Oceanographic Institution

   A simulation of the Antikythera Mechanism is available as an open source application on Github.

   The Antikythera Mechanism Research Project maintains a list of Solid Models of the Antikythera Mechanism.

6. In his *Almagest*, Claudius Ptolemy marks the beginning of an epoch in recorded time, 1 Thoth 1 Nabonassar, with the coincident occurrence of a solar eclipse and the ascension of the Chaldean, King. Nabonassar in 747 BCE.

(See the Almagest Ephemeris.) Nabonassar's calendar reform began a period of seven hundred years of meticulous record keeping, indexing, summarizing, and studying. The scientific study of astronomy based upon recorded observation is thought to have begun with Nabonassar. When we talk about the discipline of organizing, we can tip our hats to Nabonassar.

John M. Steele (2000). Observations and Predictions of Eclipse Times by Early Astronomers. Kluwer Academic Publications. pp. 43–45.

The British Museum stores the "Babylonian astronomical diaries," a highly systematized collection of ancient cuneiform texts that record periodic astronomical events, commodity prices and weather conditions over a period extending from 652 BCE to the 1st century BCE.

Aaboe, Asger. *The culture of Babylonia: Babylonian mathematics, astrology, and astronomy.* The Assyrian and Babylonian Empires and other States of the Near East, from the Eighth to the Sixth Centuries B.C.E Eds. John Boardman, I. E. S. Edwards, N. G. L. Hammond, E. Sollberger and C. B. F. Walker. Cambridge University Press, (1991)

Related Readings. See §4.5 Resources over Time (page 198)

12.12 Autonomous Cars

By Jason Danker, December 2015.

Overview. Automation in cars is nothing new. Automatic transmissions and cruise control have been around since 1939 and 1958 respectively, but these systems serve to aid, rather than replace, human drivers. What is new is a near future potential for fully autonomous cars, cars that are capable of full operation without an attending human driver.

While other vehicles, such as light rail and monorail trains, have been capable of fully automatic operation since 1967, these vehicles have the luxury of operating in closed environments and only need to be able to respond to a defined set of inputs. Autonomous cars do not have this luxury. In operating "in the wild," the systems guiding these cars may be forced to respond to any number of unanticipated situations. As the automation system cannot enumerate all possible situations, it must instead rely on continuous organization of its operating environment.

This is clearly a technical challenge, but it also raises ethical and legal issues. As autonomous cars act based on the organization of sensory inputs, the organizing systems are necessarily developed relative to ethical considerations, whether intentional or not. At the most basic level, the organizing system will direct the autonomous car in making decisions analogous to those posited in the trolley problem, a famous thought experiment in ethics that forces a choice be-

tween saving five endangered people or taking the life of an innocent person who had not been in danger. Beyond ethics, autonomous cars also raise legal questions: if an autonomous car crashes, who is liable for the damages?

What is being organized? An autonomous car will organize information about the car itself, the objects in its vicinity, and environmental conditions. The car must keep track of its movements, those of other objects, and the relative positions of itself and the other objects. It must organize this information within the environmental framework of lane markings, speed limits, road signs, traffic signals, weather and traffic conditions, and numerous other constraints. As autonomous cars become common, the cars will likely communicate with one another and this information will also need to be brought into the organizing system. The car will also need to organize, and likely prioritize, inputs from human occupants. Regardless of the exact implementation, the organizing system will necessarily limit what is worthy of organization: it is likely not possible, or desirable, to keep track of every insect in the vicinity of the car.

Why is it being organized? The car organizes its surroundings in order to safely navigate to a destination. While this is the primary interaction enabled by the organization, countless other interactions support this primary interaction. The supporting interactions fall into the two categories of prediction and reaction. The systems being developed by Google use the information that has been organized to predict what is most likely to happen next: "It predicts that the cyclist will ride by and the pedestrian will cross the street." The systems that have been launched by Tesla tend to be more reactionary: "Side Collision Warning further enhances Model S's active safety capabilities by sensing range and alerting drivers to objects, such as cars, that are too close to the side of Model S."

How much is it being organized? The extent of organization varies based on the implementation. While Google uses on-board sensors and extremely detailed street maps to implement self-driving functionality, Tesla's Autopilot relies on-board sensors and standard GPS data. While the exact extent of the organization is not publicly available information, Google has publicly stated "the system is engineered to work hardest to avoid vulnerable road users (think pedestrians and cyclists), then other vehicles on the road, and lastly avoid things that don't move." Given this, Google's categories, and their hierarchy, appear to be defined by their vulnerability.

When is it being organized? For information gathered by on-board sensors, organization takes place as objects enter and leave the vicinity of the autonomous car. The organization is ongoing as the car's surrounding and environment are constantly changing. In addition to the sensor data, autonomous cars also rely on map data which is organized in advance. Google's cars rely on specialized, highly detailed maps that are being developed as part of the self-driving car

project and, as such, are unable to drive on roads that have not yet been mapped to the necessary level of detail. While Tesla's Autopilot also relies on maps, it uses standard GPS maps and is not similarly restricted.

How or by whom is it being organized? The car's computational processes are responsible for the organization. That said, the car is restricted to organizing within the organizing system implemented by the manufacturer. While Google and Tesla are two of the main companies in this space, many traditional automotive companies are also developing autonomous systems.

Where is it being organized? Except for map data, the organization takes place within the car's onboard systems. The organization must take place in the car itself due to the potential catastrophic consequences of a lag in information flow. Additionally, ensuring all organization takes place within the car provides greater security: a self-contained car is less susceptible to attack than a network dependent one.

Other considerations. While it is likely that fully autonomous cars will be technologically feasible within a few years, the cars may still require human interactions for legal reasons. This is clearly seen in Tesla's press release for Autopilot: "The driver is still responsible for, and ultimately in control of, the car." This human-in-the-loop design principle creates a legal buffer for autonomous car manufacturers by treating the "driver" as a "liability sponge" or "moral crumple zone." As articulated by Madeleine Elish and Tim Hwang, "the human in an autonomous system may become simply a component—accidentally or intentionally—that is intended to bear the brunt of the moral and legal penalties when the overall system fails."

While these issues will ultimately play out through a combination of court rulings and policy decisions, it is interesting to note that there is legal precedent that could either blame, or exonerate, the "driver" of an autonomous car. Drawing parallels to aviation automation, precedent suggests that the human "driver" will be held responsible for liability claims arising from the operation of the car. On the other hand, product liability law offers recourse for consumers when a company's products fail. Many people have argued that this existing legal framework is sufficient to handle the liability issues brought up by autonomous vehicles.

Regardless of the legal complexities that will arise from specific incidents, autonomous cars have great potential to reduce car crashes and improve overall road safety. The promise of the autonomous technology, even for partially autonomous systems, is so great that the National Highway Traffic Safety Administration is proposing updates to its safety ratings that will penalize manufacturers that don't include autonomous technologies in their vehicles.

12.13 IP Addressing in the Global Internet

By Andrew McConachie, December 2013.

Overview. Most people take for granted that the Internet just works. They connect their computer to the Internet, it gets an IP address, and they are able to communicate with a computer with a different IP address on the other side of the planet. How did their computer get the correct IP address? How does any computer or router get the correct IP address? How did the routers and other computers on the Internet get their IP addresses? Who decides which computers and which routers get which IP addresses?

What is being organized? At their simplest, an IPv4 address is a 32-bit series of 0′s and 1′s. They are resources that are born-digital, as they have no canonical physical representation. Their digital canonical representation, with which we have all become familiar, is called the "dotted quad" format and is 4 numbers between 0-255 separated with dots. For example, 169.229.216.200 is the IPv4 address for `www.berkeley.edu`.

Not all IP addresses are of equivalent classes. There are unicast, multicast, broadcast, and experimental IPv4 addresses, and unicast addresses can be either public or private. There are also two different versions of IP addresses currently in use on the Internet, IPv4 and IPv6. We will focus on IPv4 unicast public IP addresses, since these are not only the most common, but also the most important. This is roughly the range of IP addresses from 1.0.0.0 to 223.255.255.255, with some breaks in the middle for private IP address space.

Why is it being organized? IP addresses are the foundation of network connectivity and the Internet; they identify each device on a computer network and also serve as its address, so that routers and other devices can locate and communicate with it. You cannot get online without one. IP addresses can be represented into blocks, or subnetworks, using a prefix and a mask. For example, 169.229.216.0/24 represents all IP addresses in the range of 169.229.216.1 – 169.229.216.255. Internet routers do not have enough memory to hold routes for every individual IP address on the Internet. So by organizing the Internet into subnetworks based loosely on a hierarchical model, routers are able to determine the correct path for every destination in the network without actually storing every address in their memory. If the organization of IP addresses is not handled properly, Internet routers would exhaust their memory space and parts of the Internet would become unreachable.

How much is it being organized? Currently there is too much granularity in the global Internet routing table. For a router it takes the same amount of memory to store a subnetwork with 255 IP addresses as it does to store a subnetwork with 65536 addresses. So if our main concern is to minimize memory usage in

Internet routers, thereby lowering operator costs and increasing stability, we want as little granularity as possible in the Internet routing table. The problem is that many organizations use non-contiguous IP subnetworks that cannot be aggregated into larger subnetworks. This results in routers having to store many small subnetworks instead of fewer larger subnetworks, which will eventually lead to memory exhaustion in older routers and possible reachability issues. Currently the full Internet routing table is approaching 500,000 routes. Most network engineers expect problems once the routing table grows past 512,000 entries, since router memory limitations are always at bit boundaries.

When is it being organized? IP addresses are organized once someone configures one on a device or sets up a Dynamic Host Configuration Protocol (DHCP) server. If an organization exhausts their supply of free IP addresses. it will have to make a request to the upstream provider or Regional Internet Registry (RIR) for more address space. In the early days of the Internet, large blocks of IP addresses were given to organizations, but this led to many of the addresses in these blocks not being used. We are now reaching a point where we no longer have new addresses to assign to organizations.

Markets are now emerging for organizations to buy and sell IP addresses, and the organizations who have held on to large amounts of unused addressing space stand to make significant revenue from selling their unused space. When these organizations sell their unused IP address space, they will break up large allocations into smaller subnetworks, thereby increasing granularity and further accelerating the growth of the Internet routing table.

How or by whom is it being organized? The Internet Corporation for Assigned Names and Numbers (ICANN) is currently responsible for initial allocation of IP addresses. They allocate 8 blocks of IP addresses to RIRs, who are then responsible for distributing allocations to organizations that request them. These organizations can then allocate IP addresses to smaller organizations, thus forming a loose hierarchy of organizations, where each level lower in the hierarchy receives a subset of the IP address space from the organization above it. ICANN no longer has any /8 blocks of IP addresses left to allocate to RIRs. Once all of the RIRs have exhausted their last allocations from ICANN, organizations will have to rely on secondary markets to increase their IP address space.

Other Considerations. The world of IP addressing will soon get a lot more interesting. The introduction of IPv6 as a replacement for IPv4 has been slow in coming and, while gathering momentum, continues at a snail's pace. As organizations start purchasing IP addresses from one another, we should expect increased granularity and decreased stability in the Internet routing infrastructure. Whether or not normal Internet users notice will ultimately be determined by how well equipment vendors and engineers expediently address the coming problems.

12.14 The Art Genome Project

By David Eicke, December 2014.

What is being organized? Artsy.net carries the ambitious mission of making "all the world's art" accessible to anyone with an Internet connection. This is not only challenging purely from a scale perspective, with the number of artworks in the world daunting even if it were not being incremented constantly, but it is also challenging in that "art" is a nebulous term. Creators of music and literature often refer to themselves and each other as "artists." The same goes for dancers and other performers. Will their works be included? The current collection seems to be mostly visual art, with some architecture and design objects included.

Artsy's mission is to be carried out by their Art Genome Project, which is the organizational engine that powers their search and interactions. The name was inspired by Pandora's project, as was their term for their organizing process: "genoming." Genoming is not yet automated and still costly, so Artsy selects the art that is to be "genomed" carefully. Their first priority is the works featured in galleries with whom Artsy has contracts. Galleries pay to have their work organized and searchable on the site. Those works, then, must be genomed quickly in order to keep the company running. Artsy's engine also takes in works from museums and other institutions who do not have contracts with them, but many of those institutions have image-rights concerns, and not all their artworks can be published. In other cases, the images of the works are simply too low-quality to be displayed.

Why is it being organized? Why organize art? The simplest answer is to educate. That said, art has been being organized into movements and -isms for a very long time. The Getty Foundation even created an authoritative art vocabulary called the Categories for the Description of Works of Art a few decades ago. At first glance, Artsy seems to be reinventing the wheel. However, the organizing system Artsy uses is unique in that it facilitates a special kind of interaction with its body of published works.

The way resources are organized on Artsy is a cross between a hierarchical structure and a graph structure. They have over 1,000 characteristics (which they call "genes") to describe their resources. These characteristics can have to do with art movements, formal qualities, techniques, subject, etc. The emphasis here, however, is on relationships between works of art. For example, one of the genes Artsy uses is "eye-contact," and if you have a photo taken last month where the subject is looking directly into the camera and an oil painting from hundreds of years ago where the subject's eyes are looking at the painter, those

two can be one click away from each other. No other organizing system could facilitate that sort of easy link between two such disparate works.

This free-flowing linkage between works enables the "berry-picking" model of knowledge seeking, where a user searching for something doesn't necessarily have to know what he or she is searching for. A user could begin her exploration with only a vague notion that she enjoys this long-legged rhinoceros sculpture by Salvador Dali. She may not know what she likes about it, but she will see his other work there. Maybe she finds a painting she likes in the "other works by Dali" section, and she clicks on it. Then the characteristics of this painting are listed in the interface, and she is free to click on any one of them. She might click on "Surrealism" and find more works from that movement. She may click on "waterscapes" and find other oceanic imagery. She is free to explore and discover art in a self-directed way and free to discover what she likes and why she likes it. The director of Artsy's Art Genome Project says the system was intended to parallel a professor who is adept at "riffing" on things.

How much is it being organized? As mentioned above, Artsy currently uses over 1,000 characteristics ("genes") to describe its resources. These characteristics can describe anything from the art's form to the art's subject to the technique used to create the art. Experts assign these genes to the artworks and then assign those genes a weight from 0 to 100, depending on the salience of the characteristic within the work. Aside from the genes, the art is described in terms of physical dimensions (how much space it takes up), whether it has been sold or not, its gallery, its price (if for sale), its creation date, and, of course, who created it. Having such a rich set of descriptions has allowed Artsy to create a public API for developers to use all of this information as they see fit.

When is it being organized? Description of Artsy's resources is an ongoing process. Their ingested collection of art is much larger than their published collection. Most of the artworks are waiting to be genomed, with some of them waiting for permissions or image-rights paperwork to process. Another factor in determining when something is organized is the signing of new contracts with galleries. Works from galleries with contracts have first priority, and Artsy experts genome those works as they come in.

While these experts are assigning genes on a rolling basis, they are also drawing upon hundreds of years of art history scholarship when assigning them. For example, the Arsty experts did not come up with Dadaism as an organizational concept. So, in a way, some of these works were organized long ago.

How or by whom is it being organized? Artsy has a team of art historians and experts working to describe the resources that Artsy has ingested (and those that it will ingest). They have done some experiments with image-recognition software, but its descriptions are simply not rich enough to facilitate the sorts of interactions the organization is trying to facilitate. The strategy of employing

experts has its obvious downsides, however. It does not scale well, and it is reminiscent of Yahoo's early strategy of employing librarians to describe web content. There will also be inevitable biases in human resource description.

Other considerations. With such a grand ambition, one thing that may stand in Artsy's way of becoming an authoritative organizing system in the art space is that they are for-profit. Even if they are able to avoid too much bias in the interest of revenue generation, the perception remains that they are less interested in classifying art for educational purposes and more interested in making money.

12.15 Making a Documentary Film

By Suhaib Syed, December 2013.

Overview. As part of a small crew, I was in pursuit of making a documentary film shedding light on the problems in the higher education system in India. We had traveled far and wide, capturing many thought-provoking stories, illuminating interviews, and shocking truths. Due to the relatively small crew and a tight schedule, we ended up with our raw footage being labeled in a generic format (MVI_1234 etc.). I, being the director, had the task of assisting the editor in renaming and reorganizing the files to make our lives easier, do justice to all the efforts that were put into capturing all the clips, and incorporate them in an impactful manner.

What is being organized? The primary resources being organized were the video clips (digital, shot on DSLRs) acquired during the shoot. In this context, they could be classified as passive resources having no real capability to produce any significant value on their own, and which had to be acted upon or interacted with to produce any effect. But the key problem here was to formulate usable resource descriptions based on the following resource properties:

Intrinsic static
: Date and time of creation, duration of the clip, type of external lighting used, camera used, lens used, exposure, ISO, white balance, frame rate, compression type

Extrinsic static
: Shot sequence number (assigned to each story element during story-boarding), shot movement type (dolly, follow focus, zoom, macro, etc.)

During this particular stage, the intrinsic and extrinsic dynamic properties did not play a large role in the resource descriptions.

We had done a lot of work on story-boarding and identified the right level of granularity so that we could capture each shot sequence separately, so we di-

rectly used the shot sequence number as an important part of the resource description. This helped us keeping our descriptions short and meaningful.

Additionally, we realized that the corresponding audio clips captured along with the video also had to be organized, but since the two were intricately linked to each other we decided to use the same name as the corresponding video clip, the only difference being the extension. We relied on the editing software to capture the intrinsic static properties of the audio files (e.g., bit rate and compression type).

Why is it being organized? Essentially, we were organizing these digital resources to find, identify, and select them so as to weave a powerful narrative enabling us to convey the truth in an impactful manner.

Hence, the interactions were directly with the primary resource.

The interactions that had to be supported by our organization scheme involved:

- Finding the clips related to a particular story-board section
- Selecting the best set of clips to be included in the film based on relevance to story, progression, continuation and several other inter-connected factors
- Manipulating the clip (i.e., color-correcting, white balancing, and stabilizing) to create an aesthetic effect
- Matching the video of a clip to corresponding audio recording
- Adding the right background score based on sentiment being portrayed in the clips and the progression of the story
- Providing subtitles in case of a foreign language or incoherent speech

How much is it being organized? Since the scope and size of our organizing system was relatively limited and all the resources were already available, we were able to make some bold decisions without causing a lot of problems. We formed a controlled, vertical vocabulary for resource description by deliberately choosing certain resource properties over others. Our main objective was to keep the description as short as possible and at the same time convey the most valuable information that would help us interact with the resources (i.e., the video clips).

We could have easily opted for a date- and time-stamp based id and every resource in a collection (i.e., clips specific to one camera) would have a unique identifier, but we realized that our cameras already attached this information to the file along with the technical details like frame rate, aperture, shutter speed, ISO, and white balance, which our operating system and editing software could easily capture, display, and search through, hence, we decided not to use these details.

We also decided not to include important lighting condition properties (kino-flo, LeikoLite, etc.) and location, because the first frame in most of our clips consisted of the clap-board which contained all of this information, and our editing software showed all the video files as thumbnails using first frame of the video.

Thus we leveraged all of these to form a controlled vocabulary that placed the shot sequence number first, followed by the take number followed by camera identifier (e.g., camA, camB, etc.). For instance: 2A_1_camB.

However, we did realize that these decisions were specific to our OS and video editing software and hence lacked interoperability.

When is it being organized? In our case, although we intended to organize the resources as soon as they were acquired, we failed and then came up with an organizing system after all the resources were acquired. We leveraged this fact to our benefit and formed a more specific description system.

How or by whom is it being organized? Ideally it is the role of the first assistant cinematographer (AC), even 2nd or 3rd AC (depending on the budget), to make sure all the file names are stored properly and all the cards properly backed up. But due to our limitations we (i.e., the director and cinematographer) collaborated to organize the set of raw footage.

Other considerations. One important consideration that we left out in the discussion was the need for certain people appearing in the documentary to have their identity hidden by means of facial blurring and voice modulation. Although we could not accommodate this interaction of identifying which clips had footage of people who did not want to reveal themselves, we could easily add the special effects over an entire sequence once all the clips were brought together.

12.16 The Dabbawalas of Mumbai

Indian Lunch Box System

By Pratibha Rathore, December 2014.

Overview. The Mumbai dabbawala tiffin service is the source of much fascination from around the world, and I am no different: I worked in Mumbai for two years and used the services of dabbawalas to get my lunch box (called a “dabba”) delivered from home to my office, which was about 44 miles away. Without the use of any technology or digital resources, this organizing system has been coordinating the delivery of home-cooked lunches to thousands of Indian office workers for over a century, charging just a small fee of $3-7 per month. The community of dabbawalas has been able to create value for its customers by optimizing and standardizing the principles of its operations and devising an organizing system that is down to earth and human-centric.

What is being organized? The primary resources in the dabbawala system are the dabbas that are delivered to respective customer's offices and organized using a simple but effective color-coding system. The secondary resource is the workforce, consisting of 5,000–6,000 people known as dabbawalas, who organize themselves and their supporting supply chain and logistics operations to deliver the dabbas to the right location and at the right time each day without failure. The dabbawala community, called the Mumbai Tiffin Box Suppliers Association (MTBSA), follows a flat organization structure, meaning the motivation to perform consistently is a matter of personal drive and accountability.

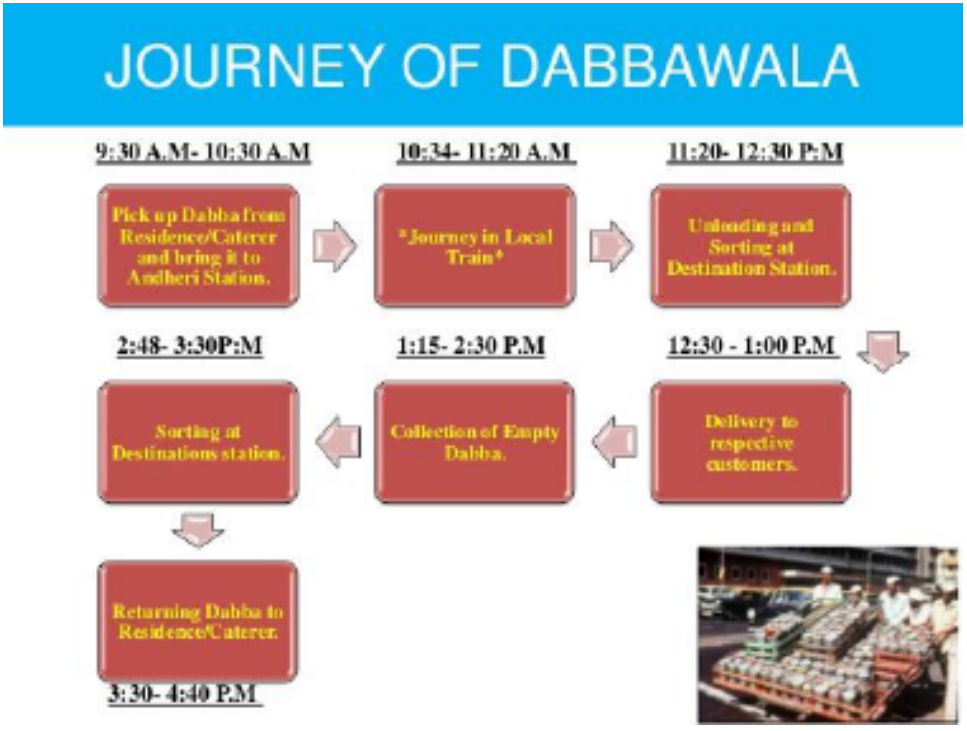
Why is it being organized? The primary reason people use the service of the dabbawalas is to eat a proper, home-prepared meal during lunch, a way to connect with their family while busy at work. The interactions supported by the dabbawala organizing system provide two significant benefits to the customers: managing their budgets while eating healthy, and leveraging time constraints. Most of the office-goers usually leave by 7 a.m. to commute from the suburbs of Mumbai, traveling south to the main commercial area of Mumbai and returning back home after 7 p.m. The railway network during the peak hours is jam-packed with commuters hanging onto the trains with one hand; therefore, carrying one's lunch at that time is not feasible. Most of the commuters cannot afford to eat takeout every day, and eating on the roadside is unhealthy and unhygienic. In addition, catering to the diverse food habits and taste needs of employees is very difficult for office canteens to manage. Thankfully, the dabbawala system solves all these problems with 100 percent customer satisfaction by delivering to each employee his lunch filled with food prepared at his home.

How much is it being organized? The Mumbai lunch box system is a successful and a socially sustainable enterprise. The number of dabbas delivered per day to offices and back home is around 300,000; that means 600,000 transactions per day. Although the number of transactions is very large, each person handles a small subset of transactions at a time. The scope of the organizing system and the scale of operations pretty much remain consistent, with the addition or deletion of few dabbas every month. Most interestingly, despite the lack of computers, mobile technology, or any automated processes, a dabba goes astray only once every two months, making less than one mistake in every 6 million deliveries. Now that's efficiency! The system is able to achieve consistency in its operations because of successful implementation of several organizing principles. Firstly, containers used to house the lunch boxes are of a standard shape and size. Second, the color coding done on the dabbas takes advantage of people's visual acuity, following a human-centric design approach. Third, the sequence of transactions to deliver each dabba from its source to destination and back to source is repeatable, predictable, systematic, and iterative in nature, enabling easy tracking and monitoring. Finally, governance within the community is ach-

ieved by instilling ethics, values, and principles in employees and by holding employees accountable at all times.

When is it being organized? The interactions between dabbawalas to deliver the dabbas follow a "hub and spoke" process model. During a dabba's journey from kitchen to consumer, it is handled by between three and twelve different deliverymen. The typical day for a dabbawala begins at 9:30 a.m., and he spends about an hour collecting all the 25–30 dabbas from the assigned set of homes in his designated area. The households are expected to have the lunch box ready when he arrives for collection. When he is done with collection, he goes to the local train station and gathers with the other dabbawalas of his area. Next, the dabbas are sorted in the order of stops on that rail line and handed off to the dabbawala who is responsible for that particular station for delivery to their final destination. At every departure station, the dabbas are passed out according to their next destinations. The same process is repeated when returning empty dabbas back to homes.

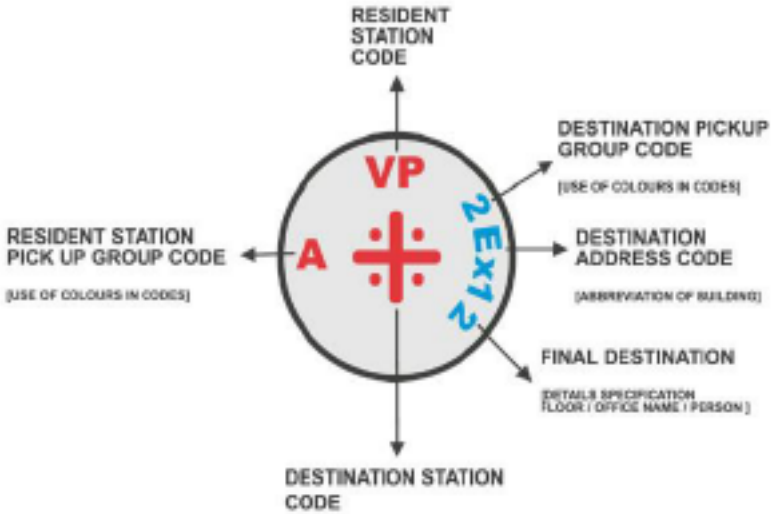
Figure 12.3. Dabbawalla Delivery Process



A model of the dabbawalla delivery process

How or by whom is it being organized? The key to this successful delivery management system is the color coding done on the dabbas. The dabbawalas use simple design measures such as signs, different colors, numbers, dashes, dots, letters, and simple symbols to indicate various parameters such as origination suburb, route to take, destination station, who is responsible, the street, building, floor, etc. As most of the dabbawals are illiterate, the choice of syntax for markings is done in such a way to ensure it is easy to understand and implement. The vocabulary used to implement and describe markings on the dabbas follows a standardized and self-descriptive process, thereby eliminating ambiguity and variability and making the organizing system more effective. Since only numbers and letters are used, the syntax for description of the primary resource (dabbas) is intentionally made to be independent of any local language, so that everyone can learn, understand, and process without any confusion, bias, or information overload.

Figure 12.4. Dabba Routing Codes



A breakdown of the coding system used to identify and route a dabba.

At each stage of the process, only one part of this code needs to be read, which works as a signal and thus allows picking up the right dabbas very quickly. It is also particularly efficient for traceability, since any dabbawala seeing a dabba knows which path it has to take. In case a dabba is lost or forbidden somewhere, any dabbawala is able to put it back on the right track. There is no need for the structure of color coding to be more granular than described above, as dabbawalas know the collection areas by heart. Furthermore, the process of adding a new resource to the organizing system is straightforward and struc-

tured. If a new resource—that is, a new customer—is added to the system, the dabbawala will do the complete journey to check the address of delivery and coordinate with other colleagues in the community to see who has a free place in his crate to add one more dabba. Once the sequence of delivery has been established and all the necessary stops for exchange decided, the address on the dabba is marked and it becomes part of the whole system.

Other considerations. It would be interesting to know if this delivery model could be used by other cities as the problem of longer commute and need for homemade food for lunch by office workers is always there in major cities. In my view, standardization of operations and understanding cultural and regional biases can provide opportunities for other cities to implement this model, at the same time providing jobs to many semi-skilled workforces.

12.17 Managing Information About Data Center Resources

By Hassan Jannah, December 2013.

Overview. Nowadays, there is an app for almost everything! Yet, we show little or no regard to what happens behind our shiny little screen until something breaks down and our lives descend to near chaos. That is the conundrum of IT guys. The truth is that IT solutions are, in many cases, fragile things that need constant care. This is no easy task. In fact, most of the cost and effort involved in IT solutions is maintenance. A million things could go wrong. Words like preventive maintenance, service monitoring, business continuity, and disaster recovery are examples of the different activities done to maximize availability, and expedite troubleshooting. Everyone involved with these activities needs access to resources. Above all, they all need access to information.

What is being organized? IT data centers have both physical and digital resources. Physical resources include the facility (i.e., building), utilities, computer hardware (e.g., network switches, cables, servers, storage, etc.), and, also people. Digital resources are much fuzzier to define. A simplistic approach could classify them into data and applications. Each category can be further subclassified into an entire ontology. The complexity increases when you consider the great number of potential resource types that can be created by combining physical and digital resources. Capturing, storing, and maintaining information about these resources is a big challenge. A lot of information can be retrieved from the resources themselves. Usually, each team responsible for supporting a certain group of resources would store information in spreadsheets and documents. More organized teams would use databases or knowledge management systems. More diligent organizations would have a central repository for everything.

What many fail to capture is the information about how all of these different clusters of resources are interconnected. That is often a much bigger and complex challenge. That information could be either buried deep in these systems (e.g., the user name used to run a certain service), or is stored in people's brains. The added value of an organizing system for data about data center resources can be multiplied if effectively organized information about their interactions.

Why is it being organized? Running an IT data center is complex, resource intensive, and risky. Customers require around the clock availability of services with no room for failure. The consequences of such failures go beyond financial loss and customer dissatisfaction. They could affect people's safety and, even, national security. Cyber threats have become a constant threat for IT service providers, especially those that host highly sensitive data or serve critical operations. People can survive if their emails were inaccessible for an hour. However, what are the ramifications of a total failure of the IT infrastructure of the New York Stock Exchange? What if the airport systems of Heathrow airport failed? These are some of the conditions that IT data center managers must work in. Furthermore, technology advances have created highly diverse, complex, and integrated solutions. New resources are introduced frequently as old resources are retired. These activities require careful planning and execution to prevent the intricate eco-system from crashing. Having all the information required to plan these activities would mitigate that risk.

Nevertheless, when something wrong does happen, having the required information is equally important to expedite fixing it. In fact, availability of information increases with the severity of the problem. How can you rebuild a system if you do not know how to connect its parts? How much are the resources organized? The granularity of the data required about data center resources varies between organizations and also between stakeholders of the same organization. The information can be classified into operational, and planning information.

How much is it being organized? Operational information is required for running day-to-day operations. These include information about resources and how they are interconnected. Many organizations put most of their focus on organizing operational information with high granularity. The granularity could be influenced by economic, political, an intellectual factors. Higher granularity means that more time and money are required to organize the information.

The level of granularity used to describe a resource type can be driven by the motives of the team leading the activity. For example, a hardware systems support team would invest more in building a robust organizing system for hardware systems and not focus on applications running on that hardware. Finally, the team's intellectual abilities and knowledge would influence the granularity of the system. As the boundaries between physical and digital resources fade,

system designers could face some challenging questions. For example, servers are, traditionally, considered hardware resources. However, many organization have switched to virtual servers running on big machines. In such a case, how would you define a server? Is it the big machine or the individual virtual servers? Is it a physical resource or a digital resource? If you have a standby clone of a virtual server, would you consider both to be the same entity or not?

Planning information is usually required to make business decisions and is usually less granular. This could include information about the purchase and maintenance costs, contracts, hardware life-times …etc. Managers and planners could use this information to better plan for business activities, manage operational and capital costs, and make strategic decisions about the services and products the data center offers.

When is it being organized? Many data centers start building an organizing system of data about their resources based on existing resources. In such cases, building the system is the easy part. The real challenge is maintaining the information up-to-date in an ever-changing environment. Clear information life-cycle and change management processes are required in parallel with work processes to ensure information is updated.

How or by whom is it being organized? Based on the scope and level of granularity of the system, the number of resources could potentially be gargantuan. The organization must try to maximize the amount of information collected automatically using auto discovery “agents” to keep update information. Inevitably, other information, especially information describing interdependencies, will require human entry. The organization must have a clear and comprehensive governance framework that details the roles and responsibilities of different parties in adding, and maintaining information.

Other considerations. Most big companies in the past operated their own corporate data centers. Their organizing system might have a smaller scope. The emergence of global cloud service providers has extended the commoditization of IT products and services across the entire technology landscape; from the consumers all the way back to the servers that provide them. These providers will have a bigger scope due to the diversity and dynamic provisioning of their services.

12.18 Neuroscience Lab

By Colin Gerber, December 2013.

Overview. A neuroscience lab is doing Parkinson's disease research in which they do experiments with rats. They use different types of rats, surgeries, and drugs for experiments and have to keep track of all this information for data analysis, publications, and lab inspectors.

The existing organizing system was developed before personal computers were prevalent and has slowly evolved over time. However, much of the underlying structure of the system still has its roots in pre-computer concepts. In order to update the system to incorporate more modern technologies what are the changes to the resources, their descriptions, and the systems structure that need to be made?

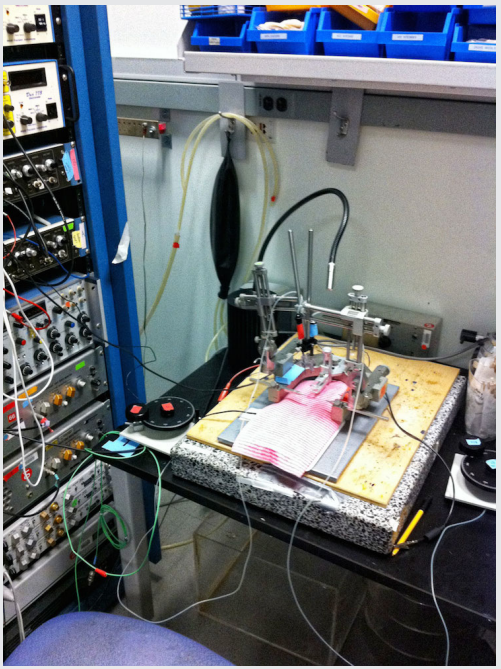
What is being organized? Resources in the current organizing system include rats, surgeries, experiments, drugs, and data recorded from the experiments. There are some other resources that could be incorporated into the organizing system.

One such new resource is surgery techniques. Surgery techniques have historically been passed down by the master apprentice method and information was largely tacit knowledge that was held by the researchers performing the surgeries and not explicitly in the system. This was done because it is inherently difficult to store the intricacies of surgery in text and even more difficult for a new researcher to learn how to perform the surgery from textual information. The ability to store and annotate multimedia changes this however. It is now possible to make instructional videos for each type of surgery, add resource descriptions to the video file and store it in the organizing system.

There is also a resource that is treated as one resource through its entire lifetime when it may actually be two. When rats are originally brought into the organizing system they are treated as a manifestation of the rat resource type. Meaning the rats are interchangeable, you can use any rat from that group in your surgery. Once the surgery has been performed the rat is modified into a new resource instance. The specific rat the surgery was performed on now has a new set of resource descriptions.

Why is it being organized? Is the main purpose of the organization system to make sure the correct rats are used in each different experiment? Or is it to make sure the records are kept up to date for the lab inspectors? It could also be making data analysis and paper writing more efficient. These decisions will affect how many different types of resource descriptions are required and the granularity needed for those descriptions.

Neuroscience Research Equipment



Physical resources in the author's lab are arranged to facilitate the precise accuracy of interactions required in medical research. In this photo, an array of amplifiers and filters for processing and recording rats' brainwave signals (left) is installed in a vertical rack that can be located close to the equipment used to perform surgeries.

(Photo by Colin Gerber. Used with permission.)

This system is just one of many organizing systems within a lab so deciding the scope and interactions it will have with the other organizing systems is very important. One important decision is if the system will support the training of new members of the lab or not. Having resources such as video recording of surgeries and experiments could enable teaching interactions for new researchers. But there are many other aspects of training a new researcher must go through, should these also be included in the organizing system? If so, it would make the system much more complex and expand the scope of the organizing system outside of surgeries and experiments but would keep all of the teaching resource in one system.

Another option would be to have a separate organizing system that is responsible for training material which is able to interact with the multimedia in the system that are relevant to training. This does not expand the scope of the system but would make the maintenance of it more difficult. Each time a surgery technique or experiment is changed two systems would have to be updated to take the changes into account.

How much is it being organized? The system is accessed by many types of users, each requiring a different type of interaction. The researchers need to search for the correct rat and surgery technique. The lab inspector needs to check for drug logs and make sure all the surgery methods and equipment are up to date. The principal investigator needs to see an overview of progress on projects.

Currently the system is organized in hierarchical categories where the top-level categories are surgery and experiments. This organization makes it easy to re-

trieve specific resources. However, the interactions normally performed with the system use resources from both sub-trees, which makes the hierarchical approach less than optimal.

A faceted classification approach could work well to enable these interactions. The facets would incorporate the original categories of surgery and experiments but also add facets for each common type of interaction. In this case different resource descriptions of the same resource will often be classified into different facets. These resource descriptions will often act as resources themselves. For example, a lab inspector is interested in retrieving the expiration date and times a drug was used in surgery, not the drug itself.

When is it being organized? In a neuroscience lab resource descriptions are often lost if they are not recorded at the time they are measured. For example, if a rat is weighed to calculate the correct dosage of a drug, both the dosage and the weight should be entered into the system. If the weight is not entered at the time of measurement it would be impossible to weigh the rat later and get the same result (as the rat changes weight over time.) This is a common problem, so as a rule all resources and descriptions should be entered into the system at the time they are acquired.

How or by whom is it being organized? The researchers working in the lab do all of the organizing. They are the ones creating new resources, descriptions and have the most knowledge about the resources and how they relate to each other.

Other considerations. Changing the system and entering all of the data at the time of measurement will initially cause more work for the researchers but will result in more accuracy for the interactions supported by the system and less retrieval work during data analysis and paper writing.

12.19 A Nonprofit Book Publisher

By Emily Paul, December 2014.

Overview. The New Press, a nonprofit book publisher with approximately 1,000 published titles, roughly 800 of which are actively in print and featured on the website, updated its book categories for use on thenewpress.com as part of a website redesign. Rather than fully adhering to an established book classification system, such as BISAC, which is commonly used in book retail, The New Press developed its own classification system. In addition to the standard goal of allowing readers to browse categories, this classification system is designed to represent the press's focus and mission. The New Press classification system employs a mix of principles and levels of granularity while incorporating some elements of the institutional categories from BISAC.

In order to gain some insight into how these dual goals affect usability, I ran user tests on a mockup of the website with the proposed categories. I conducted a think-aloud exercise in which the users verbalized their thoughts as they browsed through the categories and subcategories. I then asked the users to walk through where they would go for a particular book in response to a prompt from me that included the book's title, subtitle, and a brief description. Lastly, I asked the users about what their impressions were of The New Press after looking at the categories, whether they were confused by the categories, and which categories they would be interested in looking at if they visited the site.

What is being organized? The resource being organized is the digital presence of the books on thenewpress.com. The classification system is only used on The New Press website and is stored in a FileMaker database that pushes data to the website. There is already a dedicated website classification system that this new system builds on. It is worth noting that the book records in the database also contain BISAC categories. These are entered so that they can be sent out to distribution and bookseller feeds that require the industry-standard categories. The BISAC categories are institutional categories created by the Book Industry Standards Group. The BISAC system is designed to reflect the interests and understanding of general readers. As such, the BISAC categories are informed by cultural categories and also influence cultural categories because of their broad adoption in the book industry. In addition to using some institutional categories from BISAC and mainstream cultural categories, The New Press is using cultural categories from specific groups, namely academics and political progressives, to connect with specific readers.

Why is it being organized? The books are being categorized to facilitate browsing by readers and supporters on The New Press website. In addition to the primary browsing interaction, the categories are also being used as an opportunity to position The New Press and to convey a sense of its mission.

How much is it being organized? For the purposes of The New Press website, books can be placed in multiple categories and subcategories, but all books will have at least one category designation. Because The New Press is not concerned with the physical presentation of the resources, the books can be placed in as many categories as are relevant. In contrast, library and bookstore classifications need to satisfy the uniqueness principle, because the book can only be located in one physical location.

Most of the categories are based on the subject matter of the books. A book's subject matter is an intrinsic static property because it does not change once it is published. However, the categories used to describe this subject matter may change over time as new categories are added to the classification system and retroactively assigned to previously published books. The book subject categories can generally be thought of as extrinsic and static because the threshold for

changing them is higher than it is for more dynamic properties such as Current Season, Next Season, and Bestsellers. These categories are also included on the site in a separate section and are all extrinsic, dynamic properties because they are based either on time or sales, rather than intrinsic properties of the books.

The New Press classification system includes hierarchical categories, though only the subjects in which the press publishes more extensively have subcategories. In areas for which there are more books, the organization can be more granular without creating a subcategory that contains only one or a few books. Additionally, the greater institutional knowledge of the subject area enables the staff to make more specific distinctions within the broader subject category. One of the questions I explored in my user testing was whether these differentiations are necessary to support users' interactions with the books. If the users do not share the same level of knowledge in the subject it may not be useful, and may even diminish usability, to differentiate at the level of granularity provided by the subcategories.

Even at the top category level, there is a range of granularity and also a range of principles embodied in the categories. For example, History and Immigration are both top-level categories, but Immigration covers a more specific group of topics than History does. Most categories are based on the subject of the books, but there are several top-level categories based on other principles. These include Graphic Nonfiction, which refers to format; Primary Source Documents, which refers to the source material; and Biography, which refers to the genre of the book but does not express anything about its subject matter beyond the fact that it is about someone's life. Mixing category principles can be useful, particularly in a faceted system, which allows users to combine different categories to increase precision. In a faceted version of this system, a user could select Biography and Law in order to find biographies written about a judge or lawyer. Because books are assigned to all relevant categories in this system, this interaction is feasible at the logic level even though the current presentation does not allow it. If The New Press wanted to switch to a faceted presentation it would likely visually separate the categories into blocks based on the principles, so that users knew which facets they could pivot their searches on. This might include creating a genre section with Biography, Oral History, and Primary Source Documents as well as a geography section with the subcategories from World.

When is it being organized? Once the updated categories are finalized, all previously published books will be reviewed and assigned to new categories as necessary. Going forward, new books will be categorized on a seasonal basis and new categories may occasionally be assigned to previously published books on an ad hoc basis (this could be due to previous oversight in not assigning the category, or to the creation of a new category or subcategory). This system is flexible because books can be assigned to all relevant categories, so the introduction of a new category does not mean that all previous assignments will need to be

changed. The subcategories also allow for flexibility because if one of these categories becomes more important over time, it can be changed at the presentation layer to a top-level category with minimal effort.

How or by whom is it being organized? The sales, marketing, and inventory manager assigns the categories, with input from the editorial and marketing teams. From time to time other departments, such as fundraising or publicity, may suggest a new category or category assignment for consideration. The categories are assigned in a FileMaker database in which the categories can be selected from a list of existing categories and subcategories. The category assignments in the FileMaker database are pushed to the website along with other book data.

Other considerations. Creating a classification system that can be widely understood is difficult to do. In this case, simplifying the system would support The New Press's goal of reaching a broad audience of readers. User testing revealed that the current category system may be hindering this because of issues with semantics, granularity, and structure. The structural issues are the most important to address because the inconsistent use of subcategories generated significant confusion during the user testing. By removing the subcategories and instead allowing expert users or those who know exactly what they are looking for to use search, the press could maximize the categories' relevance for general readers. This could be strengthened by an emphasis on using relevant keywords in the book descriptions that support searching. Despite some initial surprise from the test users about certain unusual top-level categories, I would argue that after simplifying other aspects of the system, the press could successfully keep some of these in order to represent its publishing areas and connect with like-minded readers. For example, Immigration and Criminal Justice are not top-level BISAC categories, but are easily understood by general readers and serve to highlight these important areas for The New Press. Biases in classification systems are unavoidable. While this can be negative, particularly when the organizers are not aware of the biases, it can also be harnessed positively and used to communicate a sense of the organization and its values. This needs to be approached thoughtfully and carefully and tested on users to understand how people outside the organization will interact with the system.

Endnotes for Chapter 12

[646][CogSci] (Ctein 2010) and (Taylor 2010) are popular guides for photo digitization and restoration.

[647][Web] For example, http://web.appstorm.net/roundups/media-roundups/top-20-photo-storage-and-sharing-sites/ reviews 20 photo storage and sharing sites and http://photo-book-review.toptenreviews.com/ compares 10 sites for creating

printed albums from digital photos in case you want to "round trip" from Grandpa's photos and print photo books for family members.

[648][Law] (Herbst 2009) is a thoughtful legal primer on the novel property, jurisdiction, and terms of service complexities in gaining access to accounts of deceased people. A popular treatment about what has come to be called the "digital afterlife" is (Carroll and Romano 2011).

[649][Law] `http://www.spo.berkeley.edu/guide/consultquick.html` is an example of such a policy. Indeed, it is because of rules like these that the professor determined he needed to take a leave of absence from the university.

[650][Bus] For a high-level theoretical framework about capturing value from knowledge assets see (Teece 1998); for a detailed case study see (Goodwin et al. 2012).

[651][Bus] (Poole and Grudin 2010).

[652][Bus] (Hansen 2009).

[653][Bus] (Wakabayashi 2011).

[654][Bus] (Hori, Kawashima, and Yamazaki 2010). Fujitsu expects that the system will eventually integrate business management functions, production history, and operational support for best practices.

[655][Com] See(Burrell, Brooke, and Beckwith 2004) for a study of the use of sensor networks in Oregon vineyards.

[656][Ling] (Tagliabue 2012). We cannot resist describing this as "sexting" by cows.

[657][Com] (Wilde and Catin 2007). Looking back it seems ironic to start with a single-source XML publishing system, abandon it to author the book in Word, and then convert the files Word back to XML to enable single-source publishing.

[658][Com] (Kimber 2012) seems destined to become the definitive resource for DITA-based publishing. The definitive source for DocBook has long been (Walsh 2010).

Bibliography

Note

The bibliography presents an alphabetical listing of entries which detail the authors, publication dates, titles, publishers, issues, volumes, page numbers, and digital object identifiers associated with the works cited throughout this book.

Entries begin with an identifier based upon principal author and date of publication; where a principal author is associated with multiple works in a given year, a suffix is added to differentiate.—MM

A

[Aalbersberg2011] AALBERSBERG, Ijsbrand Jan. *"Supporting Science through the Interoperability of Data and Articles": D-Lib Magazine*. 2011. http://www.dlib.org/dlib/january11/aalbersberg/01aalbersberg.html.

[Abel2014] ABEL, Scott and BAILLIE, Rahel Anne. *The Language of Content Strategy*. XML Press, 2014.

[Ackoff1989] ACKOFF, Russell. *"From data to wisdom.": Journal of applied systems analysis*. 1989. pp. 3-9.

[Agrawal1989] AGRAWAL, Rakesh, BORGIDA, Alexander, and JAGADISH, H. V. *"Efficient Management of Transitive Relationships in Large Data and Knowledge Bases": SIGMOD '89: Proceedings of the 1989 ACM SIGMOD international conference on management of data*. 1989. pp. 253-262.

[Allmendinger2005] ALLMENDINGER, Glen and LOMBREGLIA, Ralph. *"Four Strategies for the Age of Smart Services": Harvard Business Review*. 2005. http://hbr.org/2005/10/four-strategies-for-the-age-of-smart-services/ar/1.

[Anderson2008] ANDERSON, Chris. *"The end of theory: The data deluge makes the scientific method obsolete": Wired*. 2008. http://www.wired.com/2008/06/pb-theory/ .

[Anderson2001c] ANDERSON, Stephen R. *"Morpholology": In The MIT Encyclopedia of the Cognitive Sciences*. A Bradford Book, 2001. 562-563.

[Apte1995] APTE, Uday M. and MASON, Richard O. *"Global Disaggregation of Information-Intensive Services": Management Science*. 1995. pp. 1250-1262. http://www.jstor.org/stable/2632780.

[Arasu2001] ARASU, Arvind, et al. *"Searching the Web": ACM Transactions on Internet Technology*. 2001. pp. 2-43.

[Aristotle350BC] ARISTOTLE. *On the heavens: (De Caelo et Mundo)*. 350 BC. https://archive.org/details/decaeloleofric00arisuoft.

[Arthur1992] ARTHUR, Paul and PASSINI, Romedi. *Wayfinding: People, Signs and Architecture*. McGraw-Hill, 1992.

[Atran1987] ATRAN, Scott. *"Ordinary Constraints on the Semantics of Living Kinds: A Commonsense Alternative to Recent Treatments of Natural-Object Terms": Mind & Language*. 1987. pp. 27-63. http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0017.1987.tb00107.x/abstract.

[ATT2011] AT&T. *"An AccuWeather Cloudlet Answers a Hail of Data Requests"*. 2011. http://www.business.att.com/content/customertestimonial/Case_Study-_AccuWeather_4.7.11.pdf.

[Atzori2010] ATZORI, Luigi, IERA, Antonio, and MORABITO, Giacomo. *"The Internet of Things: A survey": Computer Networks*. 2010. pp. 2787-2805. http://www.sciencedirect.com/science/article/pii/S1389128610001568.

[Aufderheide2011] AUFDERHEIDE, Patricia and JASZI, Peter. *Reclaiming Fair Use: How to Put Balance Back in Copyright*. University of Chicago Press, 2011.

B

[Baeza-Yates2011] BAEZA-YATES, Ricardo and RIBEIRO-NETO, Berthier. *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison Wesley, 2011.

[Bailey2007] BAILEY, Charles W. *"Open Access and Libraries": Collection Management*. 2007. pp. 351-383.

[Baker1962] BAKER, Keith M. *"An unpublished essay of Condorcet on technical methods of classification": Annals of Science*. 1962. pp. 99-123.

[Banzhaf2009] BANZHAF, Wolfgang. *"Self-organizing Systems": Proceedings of Encyclopedia of Complexity and Systems Science*. 2009. pp. 8040-8050.

[Barsalou1983] BARSALOU, Lawrence W. *"Ad hoc categories": Memory & Cognition*. 1983. pp. 211-227. http://www.ncbi.nlm.nih.gov/pubmed/6621337.

[Barta2009] BARTA, Patrick. *"Shifting the Right of Way to the Left Leaves Some Samoans Feeling Wronged": The Wall Street Journal*. 2009. http://online.wsj.com/article/SB125086852452149513.html.

[Batt2000] BATT, Rosemary. *"Strategic Segmentation in Frontline Services: Matching Customers, Employees, and Human Resource Systems": International Journal of Human Resource Management*. 2000. pp. 540-561.

[Batten1951] BATTEN, W. E. *"Specialized Files for Patent Searching": In Punched Cards: Their Applications to Science and Industry*. Reinhold Publishing Corporation, 1951. 169-181.

[Battistella1996] BATTISTELLA, Edwin. *The Logic of Markedness*. Oxford University Press, 1996.

[Bell1970] BELL, Barbara. *"The oldest records of the Nile floods.": The Geographical Journal*. 1970. pp. 569-573. .

[Bentivogli2000] BENTIVOGLI, Luisa and PIANTA, Emanuele. *"Looking for lexical gaps": Proceedings of Euralex-2000 International Congress*. 2000.

[Bergmark2002] BERGMARK, Donna, LAGOZE, Carl, and SBITYAKOV, Alex. *"Focused Crawls, Tunneling, and Digital Libraries": Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. 2002. pp. 91-106.

[Berlin2014] BERLIN, Brent. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies.* Princeton University Press, 2014.

[Berners-Lee1998] BERNERS-LEE, Tim. *"Cool URIs don't change": World Wide Web Consortium (W3C)*. 1998. http://www.w3.org/Provider/Style/URI.html.

[Berners-Lee2001] BERNERS-LEE, Tim, HENDLER, James, and LASSILA, Ora. *"The Semantic Web": Scientific American*. 2001.

[Biasiotti2008] BIASIOTTI, Mariangela, et al. *"Legal informatics and management of legislative documents"*. Global Centre for ICT in Parliament, 2008.

[Bitner1992] BITNER, Mary Jo. *"Servicescapes: The impact of physical surroundings on customers and employees": Journal of Marketing*. 1992. pp. 57-71.

[Bizer2009a] BIZER, Christian. *"The Emerging Web of Linked Data": IEEE Intelligent Systems*. 2009. pp. 87-92. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5286174.

[Bizer2009b] BIZER, Christian, HEATH, Tom, and BERNERS-LEE, Tim. *"Linked Data—The Story So Far": International Journal on Semantic Web and Information Systems*. 2009. pp. 1-22.

[Blanchette2002] BLANCHETTE, Jean-François and JOHNSON, Deborah G. *"Data Retention and the Panoptic Society: The Social Benefits of Forgetfulness": The Information Society*. 2002.

[Blanzieri2009] BLANZIERI, Enrico and BRYL, Anton. *"A survey of learning-based techniques of email spam filtering": Artificial Intelligence Review*. 2009. pp. 63-92.

[Blei2012] BLEI, D. M. *"Probabilistic topic models": Communications of the ACM*. 2012. 77-84.

[Board2002] Board on Energy and Environmental Systems, Division on Engineering and Physical Sciences, and Transportation Research Board. *Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards*. The National Academies Press, 2002. http://www.nap.edu/openbook.php?isbn=0309076013.

[Bolshakov2004] BOLSHAKOV, Igor A. and GELBUKH, Alexander. *"Synonymous Paraphrasing Using WordNet and Internet": Proceedings of NLDB: International Conference on Applications of Natural Language to Information Systems*. 2004. pp. 312-323.

[Borges1952] BORGES, Jorge Luis. *"The Analytical Language of John Wilkins" (El idioma analitico de John Wilkins): In Otras Inquisiciones (1937–1952)*. 1952.

[Borgman2011] BORGMAN, Christine L. *"The Conundrum of Sharing Research Data": Journal of the American Society for Information Science and Technology*. 2011. pp. 1-40. http://papers.ssrn.com/abstract=1869155.

[Boroditsky2003] BORODITSKY, Lera. *"Linguistic Relativity": In Encyclopedia of Cognitive Science*. Wiley, 2003.

[Boroditsky2010] BORODITSKY, Lera. *"Lost in Translation": The Wall Street Journal*. 2010. http://online.wsj.com/article/SB10001424052748703467304575383131592767868.html.

[Boroditsky2011] BORODITSKY, Lera. *"How Language Shapes Thought": Scientific American*. 2011. http://www.sciamdigital.com/index.cfm?fa=Products.ViewIssuePreview&ARTICLEID_CHAR=94C85092-237D-9F22-E874366AD6B49809.

[Bowker2000] BOWKER, Geoffrey C. and STAR, Susan Leigh. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, 2000.

[Brailsford1999] BRAILSFORD, David F. *"Separable Hyperstructure and Delayed Link Binding": ACM Computing Surveys*. 1999.

[Bray2005] BRAY, Tim. *"On Language Creation": XML 2005*. 2005.

[Brin2009] BRIN, Sergey. *"A Library to Last Forever": The New York Times*. 2009. http://www.nytimes.com/2009/10/09/opinion/09brin.html.

[Brown2009] BROWN, Bruce C. *How to Stop E-Mail Spam, Spyware, Malware, Computer Viruses, and Hackers from Ruining Your Computer or Network: The Complete Guide for Your Home and Work*. Atlantic Publishing Group Inc, 2009.

[Brown2010] BROWN, Dan. *Communicating design: developing web site documentation for design and planning*. New Riders, 2010.

[Bruner1957] BRUNER, Jerome S. *"Going beyond the information given": In Contemporary approaches to cognition*. Harvard University Press, 1957. 41-69. http://www.jimdavies.org/summaries/bruner1957.html.

[Budanitsky2006] BUDANITSKY, Alexander and HIRST, Graeme. *"Evaluating WordNet-based Measures of Lexical Semantic Relatedness": Computational Linguistics*. 2006. pp. 13-47.

[Buettcher2010] BUETTCHER, Stefan, CLARKE, Charles L. A., and CORMACK, Gordon V. *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, 2010.

[Buhrmester2007] BUHRMESTER, Jason. *"NFL Films' Exhaustive Archive Is Rushing Into the Digital Age": Wired*. 2007. http://www.wired.com/culture/lifestyle/magazine/15-10/ps_nfl.

[Bulmer1970] BULMER, R. N. H. *"Which came first, the chicken or the egg-head?": In Échanges et communications: mélanges offerts à Claude Lévi-Strauss à l'occasion de son 60ème anniversaire* . Mouton & Co, 1970. pp. 1069-1091.

[Burke1993] BURKE, Colin. *Information and Secrecy: Vannevar Bush, Ultra, and the other Memex*. Scarecrow Press, 1993.

[Burrell2015] BURRELL, Jenna. *"How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms"*. September 15, 2015. http://ssrn.com/abstract=2660674 and http://dx.doi.org/10.2139/ssrn.2660674.

[Burrell2004] BURRELL, Jenna, BROOKE, Tim, and BECKWITH, Richard. *"Vineyard Computing: Sensor Networks in Agricultural Production": Pervasive Computing, IEEE*. 2004. pp. 38-45.

[Bush1945] BUSH, Vannevar. *"As We May Think": The Atlantic*. 1945.

[Buttcher2010] BÜTTCHER, Stefan, CLARKE, Charles, and CORMACK, Gordon V. *Information retrieval: Implementing and evaluating search engines*. The MIT Press, 2010.

[Byrne2010] BYRNE, Gillian. *"The Strongest Link: Libraries and Linked Data": D-Lib Magazine*. 2010. http://www.dlib.org/dlib/november10/byrne/11byrne.html.

C

[Cairo2012] CAIRO, Alberto. *The Functional Art: An introduction to information graphics and visualization*. New Riders, 2012.

[Campbell2011] CAMPBELL, Joseph Keim, O'ROURKE, Michael, and SLATER, Matthew H. *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science*. A Bradford Book, 2011.

[Cano2005] CANO, Pedro, et al. *"Audio Fingerprinting: Concepts And Applications": In Computational Intelligence for Modelling and Prediction*. Springer, 2005.

[Carey1991] CAREY, Susan and GELMAN, Rochel. *The Epigenesis of Mind: Essays on Biology and Cognition*. Psychology Press, 1991. http://www.amazon.co.uk/The-Epigenesis-Mind-Cognition-Symposia/dp/toc/0805804382.

[Carney2005] CARNEY, David, et al. *"Some Current Approaches to Interoperability": Carnegie Mellon Software Engineering Institute*. 2005.

[Carroll2010] CARROLL, Evan and ROMANO, John. *Your Digital Afterlife: When Facebook, Flickr, and Twitter Are Your Estate, What's Your Legacy?*. New Riders, 2010.

[Casson2002] CASSON, Lionel. *Libraries in the Ancient World*. Yale University Press, 2002.

[Cerf1969] CERF, Vint. *ASCII Format for Network Interchange*. 1969. http://tools.ietf.org/html/rfc20.

[Chaffin1984] CHAFFIN, Roger and HERRMANN, Douglas J. *"The similarity and diversity of semantic relations": Memory & Cognition*. 1984. pp. 134-41.

[Chandler1977] CHANDLER JR, Alfred Dupont. *The Visible Hand: The Managerial Revolution in American Business*. Belknap Press, 1977.

[Chandola2009] CHANDOLA, Varun, BANERJEE, Arindam, and KUMAR, Vipin. *"Anomaly detection: A survey.": ACM Computing Surveys (CSUR)*. 2009. pp. 15.

[Chapman2009] CHAPMAN, Nigel and CHAPMAN, Jenny. *Digital Multimedia*. Wiley, 2009.

[Chater2016] CHATER, N and LOEWENSTEIN, G. *"The under-appreciated drive for sense-making.": Journal of Economic Behavior & Organization*. 2016.

[Chen2010] CHEN, Donglin, et al. *"Research on the Theory of Customer-Oriented E-Catalog Ontology Automatic Construction": 2010 International Conference on E-Business and E-Government*. 2010. pp. 2961-2964. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5590430.

[Cherbakov2005] CHERBAKOV, Luba, et al. *"Impact of service orientation at the business level": IBM Systems Journal*. 2005. pp. 653-668. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5386696.

[Chi1981] CHI, M.T., FELTOVICH, P.J., and GLASER, R. *"Categorization and representation of physics problems by experts and novices": Cognitive Science*. 1981. pp. 121-152.

[Cho2000] CHO, Junghoo and GARCIA-MOLINA, Hector. *"The Evolution of the Web and Implications for an Incremental Crawler": Proceedings of the 26th International Conference on Very Large Data Bases*. 2000. pp. 200-209.

[Chomsky1957] CHOMSKY, Noam. *Syntactic Structures*. Mouton & Co, 1957.

[Chomsky1965] CHOMSKY, Noam. *Aspects of the Theory of Syntax*. The MIT Press, 1965. http://www.worldcat.org/title/aspects-of-the-theory-of-syntax/oclc/309976.

[Christen2006] CHRISTEN, Peter. *A Comparison of Personal Name Matching: Techniques and Practical Issues: The Australian National University*. 2006. p. 14.

[Clark1868] CLARK, Stephen Watkins. *A Practical Grammar: In Which Words, Phrases, and Sentences Are Classified According to Their Offices and Their Various Relations to One Another*. A.S. Barnes & Co, 1868. http://archive.org/details/practicalgrammar00clar.

[Clark2010] CLARK, Stephen Watkins. *A Practical Grammar: In Which Words, Phrases, and Sentences Are Classified According to Their Offices and Their Various Relations to One Another*. Nabu Press. Originally published in 1847 by A.S. Barnes & Co, 2010. http://archive.org/details/practicalgrammar00clar.

[Coase1937] COASE, Ronald H. *"The Nature of the Firm": Economica, New Series*. 1937. pp. 386-405.

[Codd1970] CODD, E. F. *"A relational model of data for large shared data banks": Communications*. 1970. pp. 377-387. http://www.ncbi.nlm.nih.gov/pubmed/9617087.

[Conklin1987] CONKLIN, Jeff. *"Hypertext : An Introduction and Survey": IEEE Computer*. 1987. pp. 17-41.

[Conklin1988] CONKLIN, Jeff and BEGEMAN, Michael L. *"glBIS : A Hypertext Tool for Exploratory Policy Discussion": ACM Transactions on Information Systems*. 1988. pp. 303-331.

[Constantin1994] CONSTANTIN, James A. and LUSCH, Robert F. *Understanding Resource Management: How to deploy your people, products, and processes for maximum productivity*. Irwin Professional, 1994.

[Cormen2009] CORMEN, Thomas H., et al. *Introduction to Algorithms*. The MIT Press, 2009.

[Cowan2004] COWAN, John and TOBIN, Richard. *XML Information Set: Recommendation of the World Wide Web Consortium (W3C)*. 2004. http://www.w3.org/TR/xml-infoset/.

[Cox2007] COX, Ingemar, et al. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2007.

[Coyle2006] COYLE, Karen. *"Identifiers: Unique, Persistent, Global": The Journal of Academic Librarianship*. 2006. pp. 428-431.

[Crandall2006] CRANDALL, R and POMERANCE, C. *"Prime numbers: a computational perspective": Springer Science & Business Media*. 2006.

[Crawford2012] CRAWFORD, Stephanie and JOHNSON, Bernadette. *"How the Nest Learning Thermostat Works"*. HowStuffWorks.com, 06 March 2012.

[Croft2009] CROFT, Bruce W., METZLER, Donald, and STROHMAN, Trevor. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

[Crow2010] CROW, David. *Visible Signs: An Introduction to Semiotics in the Visual Arts*. AVA Publishing, 2010.

[Ctein2010] CTEIN. *Digital Restoration from Start to Finish: How to Repair Old and Damaged Photographs*. Focal Press, 2010.

[Cutter1876] CUTTER, Charles. *Rules for a printed dictionary catalogue: Issued as part 2 of Special report on public libraries, by the United States Education Bureau*. Government Printing Office, 1876. http://www.openlibrary.org/books/OL24156277M.

D

[Darnton2011] DARNTON, Robert. *"Google's Loss: The Public's Gain": The New York Review of Books*. 2011. http://www.nybooks.com/articles/archives/2011/apr/28/googles-loss-publics-gain.

[Das2002] DAS, Sajal K., et al. *"The role of prediction algorithms in the MavHome smart home architecture": IEEE Wireless Communications*. 2002. pp. 77-84.

[Date2003] DATE, C.J. *An Introduction to Database Systems*. Addison Wesley, 2003.

[Datta2008] DATTA, Ritendra, et al. *"Image retrieval: ideas, influences, and trends of the new age": ACM Computing Surveys*. 2008.

[Deerwester1990] DEERWESTER, Scott, et al. *"Indexing by Latent Semantic Analysis": Journal of the American Society for Information Science*. 1990. pp. 391-407.

[deLeon2003] DE LÉON, David. *"Actions, Artefacts, and Cognition: An Enthnography of Cooking": Lund University Cognitive Studies*. 2003.

[Demartini2006] DEMARTINI, Gianluca and MIZZARO, Stefano. *"A Classification of IR Effectiveness Metrics IR Metrics: A Survey and a Classification": In Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006*. Springer Berlin Heidelberg, 2006. 488-491.

[DeRose1989] DEROSE, Steven J. *"Expanding the Notion of Links": Proceedings of the second annual ACM conference on Hypertext (HYPERTEXT '89)*. 1989. pp. 249-257.

[DeRose2010] DEROSE, Steven J., et al. *XML Linking Language (XLink): Recommendation of the World Wide Web Consortium (W3C)*. 2010. http://www.w3.org/TR/xlink11/.

[Desoky2010] DESOKY, Ashraf. *MoSCoW Prioritisation*. 2010. http://certifications.groupsite.com/beta/discussion/topics/310632/messages.

[Deutscher2011] DEUTSCHER, Guy. *Through the Language Glass: Why the World Looks Different in Other Languages*. Arrow Books, 2011.

[Dey2001] DEY, Anind K. *"Understanding and Using Context": Personal and Ubiquitous Computing*. 2001. pp. 4-7.

[Diaz2005] DIAZ, Alejandro M. *Through the Google Goggles: Sociopolitical Bias in Search Engine Design: Stanford University Program in Science, Technology and Society*. 2005.

[Ding2004] DING, Li, et al. *"Swoogle: a search and metadata engine for the semantic web": Proceedings of the 2004 Conference on Information and Knowledge Management*. 2004. pp. 652-659.

[Doctorow2001] DOCTOROW, Cory. *Metacrap*. 2001. http://www.well.com/~doctorow/metacrap.htm.

[Domingos2015] DOMINGOS, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books, 2015.

[Donnellan1966] DONNELLAN, Keith S. *"Reference and Definite Descriptions": The Philosophical Review*. 1966. pp. 281-304.

[Dorai2002] DORAI, Chitra and VENKATESH, Svetha. *"Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics": Media Computing*. 2002. pp. 1-9.

[Dougherty1985] DOUGHERTY, Janet W. D. and KELLER, Charles M. *"Taskonomy: A practical approach to knowledge structures": In Directions in cognitive anthropology*. University of Illinois Press, 1985. pp. 161-174.

[Drapeau2010] DRAPEAU, Mark. *"The Three Phases of Government 2.0": O'Reilly Radar*. 2010. http://radar.oreilly.com/2010/05/the-three-phases-of-government.html.

[Dumais2003] DUMAIS, Susan. *"Data-driven approaches to information access": Cognitive Science*. 2003. pp. 491-524. http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog2703_7/abstract.

[Durkheim1963] DURKHEIM, Emile and MAUSS, Marcel. *Primitive classification*. University of Chicago Press, 1963.

[Durtschi2004] DURTSCHI, C., HILLISON, W., and PACINI, C. *"The effective use of Benford's law to assist in detecting fraud in accounting data": Journal of Forensic Accounting*. 2004. 17-34.

E

[Efron2011] EFRON, Miles. *"Information Search and Retrieval in Microblogs": Journal of the American Society for Information Science and Technology*. 2011. pp. 996-1008. http://onlinelibrary.wiley.com/doi/10.1002/asi.21512/abstract.

[Efthyvoulou2008] EFTHYVOULOU, George. *"Alphabet Economics: The link between names and reputation": The Journal of Socio-Economics*. 2008. pp. 1266-1285. http://sheffield.academia.edu/GeorgiosEfthyvoulou/Papers/330894/Efthyvoulou_G._2008_._Alphabet_Economics_The_link_between_names_and_reputation._The_Journal_of_Socio-_Economics_37_3_1266-1285.

[Eliot1934] ELIOT, T.S. *"Choruses from the rock.": The Rock. A Pageant Play*. 1934.

[Elliott2008] ELLIOTT, M, et al. *"A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity": Health Services Research*. 2008. pp 1722–1736.

[Ellis2014] ELLIS, B. *Real-time analytics: Techniques to analyze and visualize streaming data*. John Wiley & Sons, 2014.

[Elman2009] ELMAN, Jeffrey L. *"On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon": Cognitive Science*. 2009. pp. 547-582. http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2009.01023.x/abstract.

[Engelbart1963] ENGELBART, Douglas. *"A Conceptual Framework for the Augmentation of Man's Intellect": In Vistas in Information Handling*. Spartan Books, 1963. 1-29.

[Erl2005a] ERL, Thomas. *Service-Oriented Architecture (SOA): Concepts, Technology, and Design*. Prentice Hall, 2005.

[Erl2005b] ERL, Thomas. *Service-Oriented Architecture: A Field Guide to Integrating XML and Web Services*. Prentice Hall, 2005.

F

[Fabricant2002] FABRICANT, Florence. *"Chilean Sea Bass: More Than an Identity Problem": The New York Times*. 2002.

[Falkenhainer1989] FALKENHAINER, Brian, FORBUS, Kenneth, and GENTNER, Dedre. *"The structure-mapping engine: Algorithm and examples.": Artificial intelligence*. 1989. pp. 1-63.

[Farish2002] FARISH, J. Brian. *What's in a Name?: Vertaasis*. 2002. http://www.vertaasis.com/articles/whats_in_a_name.htm.

[Feinberg2012] FEINBERG, Melanie. *"Synthetic Ethos: The Believability of Collections at the Intersection of Classification and Curation": The Information Society*. 2012. pp. 329-339. http://www.tandfonline.com/doi/abs/10.1080/01972243.2012.708709.

[Ferguson2002] FERGUSON, K. *Tycho and Kepler: the unlikely partnership that forever changed our understanding of the heavens.* Bloomsbury Publishing, 2002.

[Fetterly2003] FETTERLY, Dennis, et al. *"A Large-Scale Study of the Evolution of Web Pages": Proceedings of the Twelfth International World Wide Web Conference*. 2003.

[Few2004] FEW, Stephen. *Show me the numbers: Designing tables and graphs to enlighten*. Analytics Press, 2004.

[Few2012] FEW, Stephen. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. 2nd edition. 2012.

[Fidel2012] FIDEL, Raya. *Human Information Interaction: An Ecological Approach to Information Behavior*. The MIT Press, 2012.

[Fillmore2000] FILLMORE, Charles J. and ATKINS, B. T. S. *"Describing polysemy: The case of 'crawl'": In Polysemy: Theoretical and computational approaches*. Oxford University Press, 2000. 91-110.

[Fishman2003] FISHMAN, Charles. *"The Wal-Mart You Don't Know": Fast Company*. 2003. http://www.fastcompany.com/47593/wal-mart-you-dont-know.

[Flach2012] FLACH, P. *Machine learning: the art and science of algorithms that make sense of data.* . Cambridge University Press, 2012.

[Florey2012] FLOREY, Kitty Burns. *"A Picture of Language": The New York Times*. 2012.

[Foer2011] FOER, Joshua. *Moonwalking with Einstein: The art and science of remembering everything*. Penguin, 2011.

[Freeman2005] FREEMAN, Geoffrey T., et al. *Library as Place: Rethinking Roles, Rethinking Space*. Council on Library and Information Resources, 2005.

[Frege1892] FREGE, Gottlob. *"Uber Sinn und Bedeutung": In Zeitschrift fur Philosophie und philosophische Kritik 100*. Translated as "On sense and reference." In Translations from the Philosophical Writings of Gottlob Frege, edited by P.T. Geach and M. Black, 1952, pp. 56-78. Oxford: Basil Blackwell, 1892. pp. 25-50.

[Freitas2014] FREITAS, A.A. *"Comprehensible classification models: a position paper.": ACM SIGKDD Explorations Newsletter,* . 2014. pp. 1-10.

[Fricke2009] FRICKE, Martin. *"The knowledge pyramid: a critique of the DIKW hierarchy": Journal of Information Science*. 2009. pp. 131-142.

[Friederici2009] FRIEDERICI, Peter. *Explaining Bird Flocks*. Audubon, March-April 2009.

[Friedman1996] FRIEDMAN, Batya and NISSENBAUM, Helen. *"Bias in Computer Systems": ACM Trans. Inf. Syst.* 1996. p 330–47.

[Fu2004] FU, Xiang, BULTAN, Tevfik, and SU, Jianwen. *"Analysis of interacting BPEL web services": Proceedings of the 13th conference on World Wide Web—WWW '04*. 2004.

[Furnas1987] FURNAS, G W, et al. *"The Vocabulary Problem in Human-System Communication: an Analysis and a Solution": Communications of the ACM*. 1987. pp. 964-971.

[Furner2008] FURNER, Jonathan. *"Interrogating 'identity': A philosophical approach to an enduring issue in knowledge organization": Knowledge Organization*. 2008. pp. 3-16.

G

[Geller1999] GELLER, Jacklyn. *"The Contemporary Wedding Invitation: A Social Document in Crisis": Salmagundi*. 1999. pp. 175-187.

[Geller2012] GELLER, Tom. *"Talking to machines": Communications of the ACM*. 2012. pp 14-16.

[Gentner1983] GENTNER, Dedre. *"Structure-mapping: A theoretical framework for analogy": Cognitive Science*. 1983. pp. 155-170. http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog0702_3/abstract.

[Gentner1997] GENTNER, D., et al. *"Analogical reasoning and conceptual change: A case study of Johannes Kepler": The journal of the learning sciences*. 1997. pp. 3-40.

[Gershenfeld2004] GERSHENFELD, Neil, KRIKORIAN, Raffi, and COHEN, Danny. *"The Internet of Things": Scientific American*. 2004. http://www.scientificamerican.com/article.cfm?id=the-internet-of-things.

[Gibson1977] GIBSON, James J. *"The Theory of Affordances": In Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Lawrence Erlbaum Associates, 1977.

[Gillies2000] GILLIES, James and CAILLIAU, Robert. *How the Web Was Born: The story of the World Wide Web*. Oxford University Press, USA, 2000.

[Gladwell1996] GLADWELL, Malcolm. *"The Science of Shopping": The New Yorker*. 1996.

[Glushko1988] GLUSHKO, Robert J., et al. *"Hypertext engineering: practical methods for creating a compact disk encyclopedia": Proceedings of the ACM Conference on Document Processing Systems—DOCPROCS '88*. 1988. pp. 11-19.

[Glushko2005] GLUSHKO, Robert J. and MCGRATH, Tim. *Document Engineering: Analyzing and Designing Documents for Business Informatics and Web Services*. The MIT Press, 2005. http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=10476.

[Glushko2008] GLUSHKO, Robert J., et al. *"Categorization in the wild": Trends in cognitive sciences*. 2008. pp. 129-35. http://www.ncbi.nlm.nih.gov/pubmed/18343710.

[Glushko2013] GLUSHKO, Robert J. and NOMOROSA, Karen J. *"Substituting Information for Interaction: A Framework for Personalization in Service Encounters and Service Systems": Journal of Service Research*. 2013. pp. 21-38.

[Glushko2013b] GLUSHKO, Robert J. (ed.). *The Discipline of Organizing*. The MIT Press, 2013.

[Glushko2015] GLUSHKO, Robert J. *Collaborative Authoring, Evolution, and Personalization for a "Transdisciplinary" Textbook: OpenSym'15*. 2015.

[Godby2008] GODBY, Carol Jean, SMITH, Devon, and CHILDRESS, Eric. *"Toward element-level interoperability in bibliographic metadata": Code{4}lib*. 2008.

[Goldberg2008] GOLDBERG, Kevin Howard. *XML: Visual QuickStart Guide*. Peachpit Press, 2008.

[Goldstone2009] GOLDSTONE, R.L. and GURECKIS, T. M. *"Collective behavior.": Topics in cognitive science*. 2009. p 412-438.

[Golder2006] GOLDER, Scott A. and HUBERMAN, Bernardo A. *"Usage patterns of collaborative tagging systems.": Journal of Information Science*. 2006. 198-208.

[Goldstone1994] GOLDSTONE, R L. *"The role of similarity in categorization: providing a groundwork": Cognition*. 1994. pp. 125-57. http://www.ncbi.nlm.nih.gov/pubmed/7924201.

[Goodwin2012] GOODWIN, Richard, et al. *"Effective Content Reuse for Business Consulting Practices": 2012 Annual SRII Global Conference*. 2012. pp. 682-690. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6311054.

[Grappone2011] GRAPPONE, Jennifer and COUZIN, Gravida. *Search Engine Optimization (SEO): An Hour a Day*. Sybex, 2011.

[Gravois2010] GRAVOIS, John. *"The Agnostic Cartographer": Washington Monthly*. July/August 2010.

[Grean2005] GREAN, Michael and SHAW, Michael J. *"Supply-Chain Integration through Information Sharing: Channel Partnership between Wal-Mart and Procter & Gamble": Center for IT and e-Business Management, University of Illinois at Urbana-Champaign*. 2005.

[Grimmelmann2009] GRIMMELMANN, James. *"How to Fix the Google Book Search Settlement": Journal of Internet Law*. 2009. http://works.bepress.com/james_grimmelmann/23/.

[Gross1990] GROSS, Derek and MILLER, Katherine J. *"Adjectives in WordNet": International Journal of Lexicography*. 1990. pp. 265-277.

[Gruber1993] GRUBER, Thomas R. *"A Translation Approach to Portable Ontology Specifications": Knowledge Acquisition*. 1993. pp. 199-220.

[Grudin1994] GRUDIN, Jonathan. *"Groupware and social dynamics: eight challenges for developers": Communications of the ACM*. 1994. pp. 92-105.

[Guarino1998] GUARINO, Nicola. *"Formal Ontology and Information Systems": Formal ontology in information systems: proceedings of FOIS '98*. 1998. pp. 3-15.

[Guijarro2007] GUIJARRO, Luis. *"Interoperability frameworks and enterprise architectures in e-government initiatives in Europe and the United States": Government Information Quarterly*. 2007. pp. 89-101. http://linkinghub.elsevier.com/retrieve/pii/S0740624X06000864.

H

[Halasz1994] HALASZ, Frank and SCHWARTZ, Mayer. *"The Dexter Hypertext Reference Model": Communications of the ACM*. 1994. pp. 30-39.

[Hall2010] HALL, Jeffrey A., et al. *"Strategic misrepresentation in online dating: The effects of gender, self-monitoring, and personality traits": Journal of Social and Personal Relationships*. 2010. pp 117-135.

[Halvey2007] HALVEY, Martin and KEANE, Mark T. *"An Assessment of Tag Presentation Techniques": Proceedings of the 16th international conference on World Wide Web (WWW '07)*. 2007. pp. 1313-1314.

[Halvorson2012] HALVORSON, Kristina and RACH, Melissa. *Content strategy for the web*. New Riders, 2012.

[Hamilton2012] HAMILTON, Kate and WOOD, Lauren. *"Schematron in the Context of the Clinical Document Architecture (CDA)": Proceedings of Balisage: The Markup Conference 2012*. 2012.

[Hammond2004] HAMMOND, Tony, et al. *"Social Bookmarking Tools (I): A General Review": D-Lib Magazine*. 2004. http://www.dlib.org/dlib/april05/hammond/04hammond.html.

[Hanington2012] HANINGTON, B. and MARTIN, B. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers, 2012.

[Hansen2009] HANSEN, Morten. *Collaboration: How leaders avoid the traps, build common ground, and reap big results*. Harvard Business Press, 2009.

[Harris1996] HARRIS, Roy. *Signs of Writing*. Routledge, 1996.

[Haspelmath2010] HASPELMATH, Martin and SIMS, Andrea. *Understanding Morphology*. Routledge, 2010.

[Haviland1998] HAVILAND, John B. *"Guugu Yimithirr Cardinal Directions": Ethos*. 1998. pp. 25-47.

[He2007] HE, Bin, et al. *"Accessing the Deep Web: A Survey": Communications of the ACM*. 2007. pp. 94-101.

[Hearst2009] HEARST, Marti A. *Search User Interfaces*.

[Heath2011] HEATH, Tom and BIZER, Christian. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool Publishers, 2011.

[Heidorn2008] HEIDORN, P. Bryan. *Shedding Light on the Dark Data in the Long Tail of Science: Library Trends, Volume 57, Number 2, Fall 2008, pp. 280-299 (Article)*. The Johns Hopkins University Press, 2008.

[Heller2012] HELLER, Daphna, GORMAN, Kristen S., and TANENHAUS, Michael K. *"To name or to describe: shared knowledge affects referential form": Topics in cognitive science*. 2012. pp. 290-305. http://www.ncbi.nlm.nih.gov/pubmed/22389094.

[Helper2003] HELPER, Susan and MACDUFFIE, John Paul. *B2B and modes of exchange: evolutionary and transformative effects: The Global Internet Economy*. 2003. 331-380.

[Hemerly2011] HEMERLY, Jess. *"Making Metadata: The Case of MusicBrainz": SSRN Electronic Journal*. 2011. http://papers.ssrn.com/abstract=1982823.

[Herbst2009] HERBST, Charles. *"Death in cyberspace": Res Gestae*. 2009. pp. 16-25.

[Hillmann2005] HILLMANN, Diane. *Using Dublin Core—The Elements: Dublin Core Metadata Initiative*. 2005. http://dublincore.org/documents/usageguide/elements.shtml.

[Hoffman2008] HOFFMAN, Michael. *"Details emerging on how fuses got to Taiwan": Air Force Times*. 2008. http://www.airforcetimes.com/news/2008/03/airforce_loose_fuses3_032708w/.

[Hofweber2009] HOFWEBER, Thomas. *"Logic and Ontology": In The Stanford Encyclopedia of Philosophy*. 2009. http://plato.stanford.edu/archives/spr2009/entries/logic-ontology/.

[Holman2001] HOLMAN, G. Ken. *Definitive XSLT and XPath*. Prentice Hall, 2001.

[Hori2010] HORI, Mitsuyoshi, KAWASHIMA, Eiji, and YAMAZAKI, Tomihoro. *"Application of Cloud Computing to Agriculture and Prospects in Other Fields": Fujitsu Scientific and Technical Journal*. 2010. pp. 446-454.

[Howard2011] HOWARD, Jennifer. *"Librarians Puzzle Over E-Books They May Buy but Not Truly Own": The Chronicle of Higher Education*. 2011. http://chronicle.com/article/Hot-Type-Librarians-Puzzle/127538/.

[Howe2006] HOWE, Jeff. *"The Rise of Crowdsourcing": Wired*. 2006. http://www.wired.com/wired/archive/14.06/crowds.html.

[Howe2008] HOWE, Jeff. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Random House, 2008.

[Hu2004] HU, Minqing and LIU, Bing. *"Mining and summarizing customer reviews": Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004. pp. 168.

[Huang2014] HUANG, Xuedong, BAKER, James, and REDDY, Raj. *"A historical perspective of speech recognition": Communications of the ACM*. 2014. pp 94-103.

[Hutchins2010] HUTCHINS, Edwin. *"Cognitive Ecology": Topics in cognitive science*. 2010. pp. 705-715. http://philpapers.org/rec/EDWCE.

I

[Iyengar2000] IYENGAR, Sheena S. and LEPPER, Mark R. *"When choice is demotivating: Can one desire too much of a good thing?": Journal of personality and social psychology*. 2000. . 995.

J

[Jackendoff1996] JACKENDOFF, Ray S. *The Architecture of the Language Faculty*. The MIT Press, 1996.

[Jacobs2004] JACOBS, Ian and WALSH, Norman. *Architecture of the World Wide Web*. W3C, 2004. Retrieved from http://www.w3.org/TR/webarch/.

[Jennex2009] JENNEX, Murray. *"Re-visiting the knowledge pyramid": HICSS'09. IEEE*. 2009. pp. 1-7.

[John2006] JOHN, Ajita and SELIGMANN, Dorée. *"Collaborative Tagging and Expertise in the Enterprise": Proceedings of the Fifteenth International World Wide Web Conference*. 2006.

[John1999] JOHN, Nathan. *Sony: The private life*. Houghton Mifflin, 1999.

[Johnson2013] JOHNSON, Jeff. *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Guidelines*. Elsevier, 2013.

[Johnson1998] JOHNSON, Spencer. *Who Moved My Cheese?: An Amazing Way to Deal with Change in Your Work and in Your Life*. Putnam Adult, 2010.

[Jones2007] JONES, William. *Keeping Things Found: The Study and Practice of Personal Information Management*. Morgan Kaufmann, 2007.

[Juola2014] JUOLA, Patrick. *"Rowling and 'Galbraith': an authorial analysis": Language Log*. 2014-08-10. http://languagelog.ldc.upenn.edu/nll/?p=5315.

[Juran1951] . *Quality Control Handbook*. McGraw-Hill, 1951.

K

[Kahn1995] KAHN, R. and WILENSKY, R. *A Framework for Distributed Digital Object Services*. Corporation for National Research Initiatives, 1995. Retrieved from http://www.cnri.reston.va.us/k-w.html.

[Kahneman2003] KAHNEMAN, Daniel. *"Maps of bounded rationality: Psychology for behavioral economics": American economic review*. 2003. 1449-1475.

[Kahneman1979] KAHNEMAN, Daniel and TVERSKY, Amos. *"Prospect theory: An analysis of decision under risk": Econometrica: Journal of the Econometric Society*. 1979. 263-291.

[Kalbach2007] KALBACH, James. *Designing Web Navigation*. O'Reilly Media, Inc, 2007.

[Kaser2007] KASER, Owen and LEMIRE, Daniel. *"Tag-Cloud Drawing : Algorithms for Cloud Visualization": WWW'07 Workshop on Taggings and Metadata for Social Information Organization*. 2007.

[Kaynak1997] KAYNAK, Erdener and HERBIG, Paul. *Handbook of Cross-Cultural Marketing*. Routledge, 1997.

[Kent2012] KENT, William and HOBERMAN, Steve. *Data and Reality: A Timeless Perspective on Perceiving and Managing Information in Our Imprecise World*. Technics Publications, LLC, 2012.

[Kesan2006] KESAN, Jay P. and SHAH, Rajiv C. *"Setting software defaults: Perspectives from law, computer science and behavioral economics": Notre Dame L. Rev.* 2006. p. 583.

[Kilgour1998] KILGOUR, Frederick G. *The Evolution of the Book*. Oxford University Press, 1998.

[Kim2003] KIM, W., et al. *A taxonomy of dirty data.: Data mining and knowledge discovery*. 2003. 81-99.

[Kimber2012] KIMBER, Eliot. *DITA for Practitioners Volume 1: Architecture and Technology*. XML Press, 2012.

[Kirsh1995] KIRSH, David. *"The Intelligent Use of Space": Artificial Intelligence*. 1995. pp. 31-36.

[Kirsh2000] KIRSH, David. *"A Few Thoughts on Cognitive Overload": Intellectica*. 2000. pp. 19-51.

[Kleppe2003] KLEPPE, Anneke, WARMER, Jos, and BAST, Wim. *MDA Explained: The Model Driven Architecture: Practice and Promise*. Addison Wesley, 2003.

[Koffka1935] KOFFKA, Kurt. *Principles of Gestalt psychology*. Routledge, 1935.

[Kondo2014] KONDO, Marie. *The Life-changing Magic of Tidying Up: The Japanese Art of Decluttering and Organizing*. Ten Speed Press, 2014.

[Kulish2009] KULISH, Nicholas. *"High Court in Germany Pops Names That Balloon": The New York Times*. 2009. http://www.nytimes.com/2009/05/06/world/europe/06germany.html.

[Kuniavsky2010] KUNIAVSKY, Mike. *Smart Things: Ubiquitous Computing User Experience Design*. Morgan Kaufmann, 2010.

[Kurosawa1957] KUROSAWA, Akira. *Kumonosu-jo (蜘蛛巣城): (Throne of Blood)*. Toho Studios, 1957.

L

[Lackey1999] LACKEY, Douglas P. *"What are the modern classics? The Baruch poll of great philosophy in the twentieth century": Philosophical Forum*. 1999. pp. 329-346. http://philpapers.org/rec/LACWAT.

[Lagoze1996] LAGOZE, Carl. *A Container Architecture for Diverse Sets of Metadata: D-Lib Magazine*. 1996. Retrieved from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

[Lagoze2005] LAGOZE, Carl, et al. *Fedora: An architecture for complex objects and their relationships: International Journal of Digital Libraries*. 2005. 124–138.

[Lagoze2008] LAGOZE, Carl, et al. *Object Re-Use & Exchange: A Resource-Centric Approach (arXiv)*. 2008. Retrieved from http://arxiv.org/abs/0804.2273.

[Lakoff1990] LAKOFF, George. *Women, Fire, and Dangerous Things*. University of Chicago Press, 1990.

[Lancaster1968] LANCASTER, Frederick Wilfrid. *Information Retrieval Systems: Characteristics, Testing, and Evaluation*. John Wiley & Sons, 1968.

[Langevoort2006] LANGEVOORT, Donald. *"Internal Controls After Sarbanes-Oxley: Revisiting Corporate Law's 'Duty of Care as Responsibility for Systems'": Georgetown Law Faculty Publications and Other Works*. 2006. http://scholarship.law.georgetown.edu/facpub/144.

[Langville2012] LANGVILLE, Amy and MEYER, Carl. *Google's Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2012.

[Larose2014] LAROSE, Daniel T. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[Laskey2005] LASKEY, Kenneth J. *"Metadata Concepts to Support a Net-Centric Data Environment": In Net-Centric Approaches to Intelligence and National Security*. Springer, 2005. 29-54.

[LeCun2015] LECUN, Yann, BENGIO, Yoshua, and HINTON, Geoffrey. *"Deep learning.": Nature*. 2015. pp. 436-444.

[Lee2007] LEE, John A. and VERLEYSEN, Michel. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[Levitin2002] LEVITIN, Daniel J. *Foundations of Cognitive Psychology: Core Readings*. A Bradford Book, 2002.

[Levitin2014] LEVITIN, Daniel J. *The organized mind: Thinking straight in the age of information overload*. Penguin, 2014.

[Levitt2005] LEVITT, Steven D. and LUBNER, Stephen J. *Freakonomics*. William Morrow, 2005.

[Levy2006] LEVY, Robert L. and CASEY, Patricia L. *Electronic Evidence and the Large Document Case: Common Evidence Problems: Haynes and Boone, LLP*. 2006.

[Levy2010] LEVY, Steven. *"How Google's Algorithm Rules the Web": Wired*. 2010. http://www.wired.com/magazine/2010/02/ff_google_algorithm/.

[Lewis2003] LEWIS, Michael. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2003.

[Lewis2010] LEWIS, Michael. *The Big Short: Inside the Doomsday Machine.* WW Norton & Company, 2010.

[Linhares2007] LINHARES, M.T. and BRUM, P. *Understanding our understanding of strategic scenarios: What role do chunks play?: Cognitive Science*. 2007. pp. 989-1007.

[Linnaeus1735] LINNAEUS, C. von. *Systema Naturae 1*. Editio Decima, Reformata. (Holmiae, fasc reprint 1939) (1758), 1735.

[Linthicum1999] LINTHICUM, David S. *Enterprise Application Integration*. Addison Wesley, 1999.

[Liptak2014] LIPTAK, Adam. *Justices to weigh finance law as it was applied to little fish*. NY Times, April 28 2014. http://www.nytimes.com/2014/04/29/us/politics/justices-to-weigh-fishermans-conviction-under-federal-finance-law.html.

[Lockyer1893] LOCKYER, Norman. *The Dawn of astronomy: a study of the temple-worship and mythology of the ancient Egyptians*. Macmillan and Company, 1893.

[Lohr2014] LOHR, Steve. *"For big-data scientists, janitor work is key hurdle to insights."* The New York Times, 2014.

[Lorch1989] LORCH, Robert F. *"Text-signaling devices and their effects on reading and memory processes": Educational Psychology Review*. 1989. pp. 209-234.

[Loshin2008] LOSHIN, David. *Master Data Management*. Morgan Kaufmann, 2008.

[Lovink2011] LOVINK, Geert and TKACZ, Nathaniel. *Critical Point of View: A Wikipedia Reader*. Institute of Network Cultures, 2011. http://www.amazon.co.uk/CRITICAL-POINT-OF-VIEW-WIKIPEDIA/dp/9078146133.

[Lubetzky2001] LUBETZKY, Seymour. *Seymour Lubetzky: Writings on the Classical Art of Cataloging*. Libraries Unlimited, 1953.

M

[Madrigal2014] MADRIGAL, Alexis. *"How Netflix Reverse Engineered Hollywood"*. The Atlantic, 2014.

[Maglio2009] MAGLIO, Paul P., et al. *"The service system is the basic abstraction of service science.": Information Systems and e-business Management*. 2009. 395-406.

[Malaga2008] MALAGA, Ross A. *"Worst Practices in Search Engine Optimization": Communications of the ACM*. 2008. pp. 147-150.

[Malone1983] MALONE, Thomas W. *"How do people organize their desks?: Implications for the design of office information systems": ACM Transactions on Information Systems*. 1983. pp. 99-112.

[Malt1995] MALT, Barbara C. *"Category Coherence in Cross-Cultural Perspective": Cognitive Psychology*. 1995. pp. 85-148. http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ514283.

[Maness2006] MANESS, Jack M. *"Library 2.0 Theory: Web 2.0 and its Implications for Libraries": Webology*. 2006. http://www.webology.org/2006/v3n2/a25.html.

[Mann1988] MANN, William C. and THOMPSON, Sandra A. *"Rhetorical Structure Theory: A Theory of Text Organization": Text*. 1988. pp. 243-281.

[Manning2008] MANNING, Christopher D., RAGHAVAN, Prabhakar, and SCHÜTZE, Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Manyika2011] MANYIKA, James, et al. *"Big data: The next frontier for innovation, competition, and productivity": McKinsey Global Institute*. 2011. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.

[Marchioni2012] MARCHIONI, G., et al. *"Curating for quality: Ensuring data quality to enable new science": Final Report of NSF Workshop*. 2012. http://datacuration.web.unc.edu/files/2012/10/NSF_Data_Curation_Workshop_Report.pdf .

[Marcotte2011] MARCOTTE, Ethan. *Responsive web design*. Editions Eyrolles, 2011.

[Mardia1980] MARDIA, Kanti V., KENT, J. T., and BIBBY, J. M. *Multivariate Analysis*. Academic Press, 1980.

[Margolis1999] MARGOLIS, Eric and LAURENCE, Stephen. *Concepts: Core Readings*. A Bradford Book, 1999.

[Markoff2011] MARKOFF, John and SENGUPTA, Somini. *"Separating you and me? 4.74 degrees": The New York Times*. 2011.

[Marlow2006] MARLOW, Cameron, et al. *"HT06, tagging paper, taxonomy, Flickr, academic article, to read.": In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*. ACM, 2006. 31-40.

[Marsh2006] MARSH, Jonathan, ORCHARD, David, and VEILLARD, Daniel. *XML Inclusions (XInclude): Recommendation of the World Wide Web Consortium (W3C)*. 2006. http://www.w3.org/TR/xinclude/.

[Marshall2007] MARSHALL, Catherine C., MCCOWN, Frank, and NELSON, ML. *"Evaluating personal archiving strategies for Internet-based information": Proceedings of IS&T Archiving 2007*. 2007. http://arxiv.org/abs/0704.3647.

[Marshall2008] MARSHALL, Catherine C. *"Rethinking Personal Digital Archiving Part 1: Four Challenges from the Field": D-Lib Magazine*. 2008. http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html.

[Masaaki1986] MASAAKI, I. *Kaizen: The key to Japan's competitive success*. McGraw-Hill/Irwin, 1986.

[McBride2006] MCBRIDE, S., et al. *"Data Mapping": Journal of the American Health Information Management Association*. 2006. pp. 44-52.

[McCartney2006] MCCARTNEY, Scott. *"When Pilots Pass the BRBON, They Must Be in Kentucky": The Wall Street Journal*. 2006. http://online.wsj.com/article/SB114291174429403797.html.

[McCartney2015] MCCARTNEY, Scott. *"Technology will speed you through the airport of the future": Wall Street Journal*. 15 July 2015.

[McCulloch2003] MCCULLOCH, Mark. *Understanding W.G. Sebald*. University of South Carolina Press, 2003.

[McDonough2006] MCDONOUGH, J.P. *METS: Standardized encoding for digital library objects: International Journal of Digital Libraries*. 2006. pp. 148–158.

[McGrath2015] MCGRATH, Ben. *"Dream teams": The New Yorker*. 13 April 2015. http://www.newyorker.com/magazine/2015/04/13/dream-teams.

[Medin1993] MEDIN, Douglas L., GOLDSTONE, Robert L., and GENTNER, Dedre. *"Respects for Similarity": Psychological Review*. 1993. pp. 254-278.

[Medin1997] MEDIN, Douglas L., et al. *"Categorization and reasoning among tree experts: do all roads lead to Rome?": Cognitive psychology*. 1997. pp. 49-96. http://www.ncbi.nlm.nih.gov/pubmed/9038245.

[Melville1851] MELVILLE, Herman. *Moby Dick: or The Whale*. Richard Bentley, 1851.

[Merryman2006] MERRYMAN, John Henry. *Imperialism, Art and Restitution*. Cambridge University Press, 2006.

[Mervis1981] MERVIS, Carolyn B. and ROSCH, Eleanor. *"Categorization of natural objects": The Annual Review of Psychology*. 1981. pp. 89-115.

[Millen2006] MILLEN, David R., FEINBERG, Jonathan, and KERR, Bernard. *"Dogear: Social Bookmarking in the Enterprise": Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. 2006. pp. 111-120.

[Miller1956] MILLER, George. *"The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information": Psychological Review*. 1956. pp. 81-97.

[Miller1976] MILLER, George A. and JOHNSON-LAIRD, Philip. *Language and Perception*. Belknap Press, 1976.

[Miller1998] MILLER, George A. *"Nouns in WordNet": In WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[Miller2003] MILLER, Joaquin and MUKERJI, Jishnu. *MDA Guide Version 1.0.1: Object Management Group*. 2003. http://www.omg.org/cgi-bin/doc?omg/03-06-01.

[Mockapetris1987] MOCKAPETRIS, Paul. *RFC 1035: Domain Name System (DNS)*. 1987. http://tools.ietf.org/html/rfc1035.

[Monmonier1996] MONMONIER, M. *How to Lie with Maps (2nd Edition)*. University of Chicago Press, 1996.

[Montague2010] MONTAGUE, James. *The rise and fall of fantasy sports*. 2010. http://www.cnn.com/2010/SPORT/football/01/06/fantasy.football.moneyball.sabermetrics/index.html.

[Morgan1871] MORGAN, Lewis Henry. *Systems of Consanguinity and Affinity of the Human Family*. University of Nebraska Press, 1871.

[Morgan1997] MORGAN, Lewis Henry. *Systems of consanguinity and affinity of the human family: Reprint of Morgan(1871)*. Smithsonian Institution, 1997.

[GMorgan1997] GARETH, GREGORY, F., and ROACH, C. *Images of Organization*. John Willey Online, 1997.

[Morstatter2013] MORSTATTER, F.J., LIU, H, and CARLEY, K.M. *Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose: In Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM), July 8-10*. 2013.

[Morville2006] MORVILLE, Peter and ROSENFELD, Louis. *Information Architecture for the World Wide Web*. O'Reilly, 2006.

[Moussaid2009] MOUSSAID, M., et al. *Collective information processing and pattern formation in swarms, flocks, and crowds.: Topics in Cognitive Science*. 2009. p 469-497.

[Munk2004] MUNK, Nina. *Fools Rush In: Steve Case, Jerry Levin, and the Unmaking of AOL Time Warner*. HarperCollins, 2004.

N

[Nelson1974] NELSON, Theodor Holm. *Computer Lib*. Microsoft Press, 1974.

[Nelson1981] NELSON, Theodor H. *Literary Machines: The Report On, and Of, Project Xanadu Concerning Word Processing, Electronic Publishing, Hypertext, Thinkertoys, Tomorrow's Intellectual Revolution, and Certain Other Topics Including Knowledge, Education and Freedom*. 3rd edition. T. Nelson, 1981.

[Newell1972] NEWELL, Allen and SIMON, Herbert Alexander. *Human Problem Solving*. Prentice Hall, 1972.

[NISO2004] National Information Standards Organization. *Understanding Metadata*. NISO Press, 2004. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

[Norman1988] NORMAN, Donald A. *The psychology of everyday things*. Basic Books, 1988.

[Norman1999] NORMAN, Donald A. *"Affordance, conventions, and design": interactions*. 1999. pp. 38-43.

[Norman2006] NORMAN, Donald A. *"Logic Versus Usage: The Case for Activity-Centered Design": interactions*. 2006. pp. 45-ff.

[Norton2014] NORTON, Steven. *"Germany's 12th Man at the World Cup: Big Data"*. The Wall Street Journal, 2014-07-10. http://blogs.wsj.com/dailyfix/2014/07/10/germanys-12th-man-at-the-world-cup-big-data/.

[Nunberg1996] NUNBERG, Geoffrey. *"Farewell to the Information Age": In The Future of the Book*. University of California Press, 1996.

[Nunberg2009] NUNBERG, Geoffrey. *"Google's Book Search: A Disaster for Scholars": The Chronicle of Higher Education*. 2009. http://chronicle.com/article/Googles-Book-Search-A/48245/.

[Nunberg2011] NUNBERG, Geoffrey. *"James Gleick's History of Information": The New York Times*. 2011. http://www.nytimes.com/2011/03/20/books/review/book-review-the-information-by-james-gleick.html.

O

[OASIS2003] OASIS. *XML Common Biometric Format*. 2003. http://www.oasis-open.org/committees/download.php/3353/oasis-200305-xcbf-specification-1.1.doc.

[OASIS2006] OASIS. *Universal Business Language (UBL)*. 2006. http://docs.oasis-open.org/ubl/os-UBL-2.0/UBL-2.0.html.

[Orwell1949] ORWELL, George. *Nineteen Eighty-Four (1984)*. Secker and Warburg, 1949.

P

[Page1999] PAGE, Lawrence, et al. *"The PageRank Citation Ranking: Bringing Order to the Web": Stanford InfoLab*. 1999. http://ilpubs.stanford.edu:8090/422/.

[Pancake2012] PANCAKE, David. *Searching for Beethoven*. Wall Street Journal, 4 Jan 2012.

[Pandey2010] PANDEY, Upasana and CHAKRAVARTY, Shampa. *"A Survey on Text Classification Techniques for E-mail Filtering": 2010 Second International Conference on Machine Learning and Computing*. 2010. pp. 32-36.

[Panizzi1841] PANIZZI, Anthony. *Rules for the Compilation of the Catalogue*. 1841.

[Panofsky1972] PANOFSKY, Erwin and PANOFSKY, Gerda S. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Westview Press, 1972.

[Pathan2008] PATHAN, Mukaddim, BUYYA, Rajkumar, and VAKALI, Athena. *"Content Delivery Networks: State of the Art , Insights, and Imperatives"*. Springer, 2008. pp. 3-32.

[Perlman1984] PERLMAN, G. *Natural artificial languages: low level processes: International Journal of Man-Machine Studies*. 1984.

[Pirolli2007] PIROLLI, Peter L. T. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, 2007.

[Plato370BC] PLATO. *Phaedrus*. 370 BC.

[Pogue2009] POGUE, David. *"Should You Worry About Data Rot?": The New York Times*. 2009. http://pogue.blogs.nytimes.com/2009/03/26/should-you-worry-about-data-rot/.

[Pohs2013] POHS, Wendi. *"Building a taxonomy for auto-classification": Bulletin of the American Socieity for Information Science and Technology*. 2013. pp. 34-38.

[Polaine2013] POLAINE, A., LØVLIE, L, and REASON, B. *Service design: from insight to implementation*. Rosenfeld Media, 2013.

[Poole2010] POOLE, Erika Shehan and GRUDIN, Jonathan. *"A taxonomy of Wiki genres in enterprise settings": Proceedings of the 6th International Symposium on Wikis and Open Collaboration - WikiSym '10*. 2010. http://portal.acm.org/citation.cfm?doid=1832772.1832792.

[Pope2011] POPE, Julia T. and HOLLEY, Robert P. *"Google Book Search and Metadata": Cataloging & Classification Quarterly*. 2011. pp. 1-13.

[Pralahad1990] PRALAHAD, C. K. and HAMEL, Gary. *"The Core Competence of the Corporation": Harvard Business Review*. 1990. http://hbr.org/1990/05/the-core-competence-of-the-corporation/ar/1.

[Provost2013] PROVOST, Foster and FAWCETT, Tom. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, 2013.

Q

[Queenan2011] QUEENAN, Joe. *"Wotan, Your Double-Skim Latte Is Ready": The Wall Street Journal*. 2011. http://online.wsj.com/article/SB10001424053111904106704576582834147448392.html.

R

[Rahm2000] RAHM, Erhard and DO, Hong Hai. *"Data cleaning: Problems and current approaches": IEEE Data Eng. Bull*. 2000.

[Rahm2006] RAHM, Erhard and BERNSTEIN, Philip A. *"An Online Bibliography on Schema Evolution": ACM SIGMOD Record*. 2006. pp. 30-31.

[Ranganathan1967] RANGANATHAN, Shiyali Ramamrita. *Prologmena to Library Classification*. Asia Publishing House, 1967.

[RDFWorkingGroup2004] RDF Working Group. *Resource Description Framework: Recommendation of the World Wide Web Consortium (W3C)*. 2004. http://www.w3.org/RDF/.

[Reaney1997] REANEY, P. H. and WILSON, R. M. *A Dictionary of English Surnames*. Oxford University Press, 1997.

[Regazzoni2010] REGAZZONI, Carlo S., et al. *"Video Analytics for Surveillance: Theory and Practice": IEEE Signal Processing*. 2010. pp. 16-17.

[Rehder2004] REHDER, Bob and HASTIE, Reid. *"Category coherence and category-based property induction": Cognition*. 2004. pp. 113-153. http://linkinghub.elsevier.com/retrieve/pii/S0010027703001677.

[Renear2003] RENEAR, Allen and DUBIN, David. *"Towards Identity Conditions for Digital Documents": International Conference on Dublin Core and Metadata Applications*. 2003. pp. 181-189. http://dcpapers.dublincore.org/index.php/pubs/article/view/746.

[Resmini2011a] RESMINI, Andrea and ROSATI, Luca. *A Brief History of Information Architecture: Journal of Information Architecture*. 2011. 33-46.

[Resmini2011b] RESMINI, Andrea and ROSATI, Luca. *Pervasive Information Architecture: Designing Cross-Channel User Experiences*. Elsevier, 2011.

[Resnick2001] RESNICK, P. *"Internet Message Format": The Internet Society*. 2001.

[Reynolds2014] REYNOLDS, Gretchen. *"Train Like a German Soccer Star"*. The New York Times, 2014-07-16. http://nyti.ms/1rfy8rd.

[Rips2012] RIPS, Lance J., SMITH, Edward E., and MEDIN, Douglas L. *"Concepts and Categories: Memory, Meaning, and Metaphysics": In The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, USA, 2012. pp. 177-209.

[Robbins2016] ROBBINS, Lenn. *"A Data Scientist Dissects the 2016 NFL Draft": Wall Street Journal*. 27 April 2016. http://www.wsj.com/articles/a-data-scientist-dissects-the-2016-nfl-draft-1461793878 .

[Robertson2015] ROBERTSON, Brian J. *Holacracy: The New Management System for a Rapidly Changing World*. Henry Holt and Company, 2015.

[Robertson2005] ROBERTSON, Stephen. *"How Okapi came to TREC": In TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005. 287-299.

[Robinson2008] ROBINSON, David, et al. *"Government Data and the Invisible Hand": Yale Journal of Law and Technology*. 2008. pp. 160-176. http://papers.ssrn.com/abstract=1138083.

[Rogers2008] ROGERS, Timothy T. and MCCLELLAND, James L. *"Précis of Semantic Cognition: A Parallel Distributed Processing Approach": Behavioral and Brain Sciences*. 2008. pp. 689-749.

[Rosch1975] ROSCH, Eleanor. *"Cognitive representations of semantic categories": Journal of Experimental Psychology: General*. 1975. pp. 192-233. http://psycnet.apa.org/journals/xge/104/3/192/.

[Rosch1999] ROSCH, Eleanor. *"Principles of Categorization": In Concepts: Core Readings (Margolis 1999)*. The MIT Press, 1999. pp. 189-206.

[Rose2016] ROSE, Tom. *The End of Average: How We Succeed in a World of Sameness*. Harper Collins, 2016.

[Rosen2012] ROSEN, Jeffrey. *"Who Do Online Advertisers Think You Are?": The New York Times Magazine*. 2012. http://www.nytimes.com/2012/12/02/magazine/who-do-online-advertisers-think-you-are.html.

[Rosenthal2004] ROSENTHAL, Arnon, SELIGMAN, Len, and RENNER, Scott. *"From semantic integration to semantics management: case studies and a way forward": ACM SIGMOD Record*. 2004. pp. 44-50.

[Rothenberg1999] ROTHENBERG, Jeff. *"Ensuring the Longevity of Digital Information": RAND*. 1999.

[Rowley2007] ROWLEY, Jennifer. *"The wisdom hierarchy: representations of the DIKW hierarchy": Journal of Information Science*. 2007. pp. 163-180.

[Rowling1997-2007] ROWLING, J.K. *Harry Potter Series*. Bloomsbury Publishing, 1997-2007.

[Rowling2013] ROWLING, J.K. *The Cuckoo's Calling*. Sphere Books, 2013.

[Rubinsky1997] RUBINSKY, Yuri and MALONEY, Murray. *SGML on the Web: Small Steps Beyond HTML*. Prentice Hall, 1997.

S

[Sag2012] SAG, Matthew. *"Orphan Works as Grist for the Data Mill": SSRN Electronic Journal*. 2012. http://papers.ssrn.com/abstract=2038889.

[Salton1975] SALTON, Gerard, WONG, Anita, and YANG, Chung-Shu. *"A Vector Space Model for Automatic Indexing": Communications of the ACM*. 1975. pp. 613-320.

[Samuelson2009] SAMUELSON, Pamela. *Google Books Is Not a Library: Huffington Post*. 2009. http://www.huffingtonpost.com/pamela-samuelson/google-books-is-not-a-lib_b_317518.html.

[Samuelson2010] SAMUELSON, Pamela. *"Google Book Search and the Future of Books in Cyberspace": Miinnesota Law Review*. 2010. pp. 1308-1374.

[Samuelson2011] SAMUELSON, Pamela. *"The Google Book Settlement as Copyright Reform": Wisconsin Law Review*. 2011. pp. 479-562. http://papers.ssrn.com/abstract=1683589.

[Sayre2005] SAYRE, Robert. *"Atom: The Standard in Syndication": IEEE Internet Computing*. 2005. pp. 71-78.

[Schatz1994] SCHATZ, Bruce R. and HARDIN, Joseph B. *"NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet": Science*. 1994. pp. 895-901.

[Schmidt2009] SCHMIDT, Desmond. *"Merging Multi-Version Texts: A General Solution to the Overlap Problem": Proceedings of Balisage: The Markup Conference 2009*. 2009. http://www.balisage.net/Proceedings/vol3/print/Schmidt01/BalisageVol3-Schmidt01.html.

[Scholl2007] SCHOLL, Hans Jochen and KLISCHEWSKI, Ralf. *"International Journal of Public E-Government Integration and Interoperability: Framing the Research Agenda": International Journal of Public Administration*. 2007. pp. 889-920.

[Schwartz2005] SCHWARTZ, Barry. *The Paradox of Choice: Why More Is Less*. Harper Perennial, 2005.

[Sebald1995] SEBALD, W.G. *The Rings of Saturn. (English Ed.). (Die Ringe des Saturn. Eine englische Wallfahrt.)*. Harvill, 1995.

[SeeLi2006] SEELI, Jiexun, ZHENG, Rong, and CHEN, Hsinchun. *"From fingerprint to writeprint": Communications of the ACM*. 2006. pp. 76-82. dl.acm.org/citation.cfm?id=1121951 .

[Sen2004] SEN, Arun. *"Metadata management: past, present and future": Decision Support Systems*. 2004. pp. 151-173. http://dx.doi.org/10.1016/S0167-9236(02)00208-7.

[Sen2006] SEN, Shilad, et al. *"tagging, communities, vocabulary, evolution": CSCW '06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*. 2006. pp. 181-190.

[Settles2012] SETTLES, Burr. *"Active learning": Synthesis Lectures on Artificial Intelligence and Machine Learning*. 2012. p 1-114.

[Shadbolt2006] SHADBOLT, Nigel, BERNERS-LEE, Tim, and HALL, Wendy. *"The Semantic Web Revisited": IEEE Intelligent Systems*. 2006. pp. 96-101.

[Shah2006] SHAH, Rajiv C. and KESAN, Jay P. *"Open Standards and the Role of Politics": The Proceedings of the 8th Annual International Digital Government Research Conference*. 2006.

[Shannon1948] SHANNON, Claude E. *"A Mathematical Theory of Communication": Bell System Technical Journal*. July 1948. pp 379–423.

[Shakespeare1623] SHAKESPEARE, William. *The Tragedie of Macbeth: Folio*. 1623.

[Shapiro1998] SHAPIRO, Carl and VARIAN, Hal R. *Information rules: a strategic guide to the network economy*. Harvard Business Review Press, 1998.

[Siegel2013] SIEGAL, E. *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons, 2013.

[Silman1998] SILMAN, Roberta. *"In the Company of Ghosts": The New York Times*. 1998. http://www.nytimes.com/books/98/07/26/reviews/980726.26silmant.html.

[Silver2012] SILVER, Nate. *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. Penguin Press, 2012.

[Silverman2013] SILVERMAN, Rachel. *Some Tech Firms Ask: Who Needs Managers?*. Wall Street Journal, 6 August 2013.

[Silverston2000] SILVERSTON, Len. *The Data Model Resource Book, Vol. 2: A Library of Data Models for Specific Industries*. John Wiley & Sons, 2000.

[Simon1982] SIMON, Herbert Alexander. *Models of bounded rationality: Empirically grounded economic reason*. MIT Press, 1982.

[Simon1996] SIMON, Herbert Alexander. *The Sciences of the Artificial*. The MIT Press, 1996.

[Simon1997] SIMON, Herbert Alexander. *Administrative Behavior*. Free Press, 1997.

[Simon2010] SIMON, Nina. *The Participatory Museum*. Museum 2.0, 2010. http://www.participatorymuseum.org/read/.

[Sinclair2007] SINCLAIR, James and CARDEW-HALL, Michael. *"The folksonomy tag cloud: when is it useful?": Journal of Information Science*. 2007. pp. 15-29.

[Smith1776] SMITH, Adam. *Wealth of Nations*. W. Strahan and T. Cadell, 1776.

[Smith1981] SMITH, Edward E. and MEDIN, Douglas L. *Categories and Concepts*. Harvard University Press, 1981.

[Smith2003] SMITH, Linda B. and THELEN, Esther. *"Development as a dynamic system": Trends in Cognitive Sciences*. 2003. pp. 343-348. http://linkinghub.elsevier.com/retrieve/pii/S1364661303001566.

[Spence1985] SPENCE, Johnathan D. *The Memory Palace of Matteo Ricci*. Penguin Books, 3 September 1985.

[Spencer2009] SPENCER, D. *Card Sorting*. Rosenfeld Media, 2009.

[Stewart1997] STEWART, Thomas A. *Intellectual Capital: The New Wealth of Organizations*. Crown Business, 1997.

[Storey1993] STOREY, Veda C. *"Understanding Semantic Relationships": VDLB Journal*. 1993. pp. 455-488.

[Suehle2012] SUEHLE, Ruth. *"The Periodic Tables of Everything but Elements": Wired*. 2012. http://www.wired.com/geekmom/2012/03/the-periodic-tables-of-everything-but-elements/.

[Suits1967] SUITS, Bernard. *"What is a Game?": Philosophy of Science*. 1967. pp. 148-156.

[Svenonius2000] SVENONIUS, Elaine. *The Intellectual Foundation of Information Organization*. The MIT Press, 2000.

T

[Tagliabue2012] TAGLIABUE, John. *"Swiss Cows Send Texts to Announce They're in Heat": The New York Times*. 2012. http://www.nytimes.com/2012/10/02/world/europe/device-sends-message-to-swiss-farmer-when-cow-is-in-heat.html.

[Talukder2016] TALUKDER, Hisham. *"Preventing in--game injuries for NBA players": MIT Sports Analytics Conference*. 2016. http://www.sloansportsconference.com/wp-content/uploads/2016/02/1590-Preventing-in-game-injuries-for-NBA-players.pdf .

[Tauberer2014] TAUBERER, Joshua. *Open Government Data: The book*. 2014. https://opengovdata.io/2014/history-the-movement/ .

[Taylor2009] TAYLOR, Arlene G. and JOUDREY, Daniel N. *The Organization of Information*. Libraries Unlimited, 2009.

[Taylor1914] TAYLOR, F.W. *The principles of scientific management*. Harper, 1914.

[Taylor2006] TAYLOR, Hugh. *The Joy of SOX: Why Sarbanes-Oxley and Service-Oriented Architecture May Be the Best Thing That Ever Happened to You*. Wiley, 2006.

[Taylor2007] TAYLOR, James and RADEN, Neil. *Smart (Enough) Systems: How to Deliver Competitive Advantage by Automating Hidden Decisions*. Prentice Hall, 2007.

[Taylor2010] TAYLOR, Maureen A. *Preserving Your Family Photographs*. Picture Perfect Press, 2010.

[Teece1998] TEECE, David J. *"Capturing Value from Knowledge Assets: The New Economy, Markets for Know-how, and Intangible Assets": California Management Review*. 1998. pp. 55-79.

[Tenenbaum2000] TENENBAUM, J.B. *"Rules and similarity in concept learning": Advances in neural information processing systems*. 2000. pp. 59-65.

[Tenenbaum2001] TENENBAUM, J.B. and GRIFFITHS, T.L. *"Generalization, similarity, and Bayesian inference": Behavioral and brain sciences*. 2001. pp. 629-640.

[Thaler2008] THALER, Richard H. and SUNSTEIN, Cass R. *Nudge: Improving Decisions about Health, Wealth and Happiness*. Yale University Press, 8 April 2008.

[Tidwell2008] TIDWELL, Doug. *XSLT: Mastering XML Transformations*. O'Reilly, 2008.

[Tidwell2010] TIDWELL, Jennifer. *Designing Interfaces*. (2nd Edition). O'Reilly, 2010.

[Tillett1991] TILLETT, Barbara B. *"A Taxonomy of Bibliographic Relationships": Library Resources & Technical Services*. 1991. pp. 150-158.

[Tillett1992] TILLETT, Barbara B. *"Bibliographic relationships: An empirical study of the LC machine-readable records": Library Resources & Technical Services*. 1992. pp. 162-188.

[Tillett2001] TILLETT, Barbara B. *"Bibliographic Relationships": In Relationships in the Organization of Knowledge*. Kluwer, 2001. 19-36.

[Tillett2003] TILLETT, Barbara. *"What is FRBR? A Conceptual Model for the Bibliographic Universe": Technicalities*. 2003.

[Tillett2005] TILLETT, Barbara B. *"FRBR and Cataloging for the Future": Cataloging & Classification Quarterly*. 2005. pp. 197-205.

[Toma2008] TOMA, Catalina L., HANCOCK, Jeffrey T., and ELLISON, Nicole B. *"Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles": Personality and Social Psychology Bulletin*. 2008. pp 1023-1036.

[Trant2009] TRANT, Jennifer. *"Emerging convergence? Thoughts on museums, archives, libraries, and professional training": Museum Management and Curatorship*. 2009. pp. 1-24.

[Travers1969] TRAVERS, Jeffrey and MILGRAM, Stanley. *"An experimental study of the small world problem": Sociometry*. 1969. pp. 425-443.

[Trofimov2015] TROFIMOV, Y. *"Would new borders mean less conflict in the Middle East?": Wall Street Journal*. 10 April 2015. `http://on.wsj.com/1DsapZP`.

[Turban2010] TURBAN, Efraim, SHARDA, Ramesh, and DELEN, Dursun. *Decision Support and Business Intelligence Systems*. Prentice Hall, 2010.

[Tufte1983] TUFTE, Edward R. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[Turnbull2009] TURNBULL, Giles. *"A Common Nomenclature for Lego Families": The Morning News*. 2009. `http://www.themorningnews.org/article/a-common-nomenclature-for-lego-families`.

[Turney2008] TURNEY, Peter. *"The latent relation mapping engine: Algorithm and experiments.": Journal of Artificial Intelligence Research*. 2008. pp. 615-655. .

[Tversky1974] TVERSKY, Amos and KAHNEMAN, Daniel. *"Judgment under uncertainty: Heuristics and biases": Science*. no. 4157. 1974.

U

[Underhill2008] UNDERHILL, Paco. *Why We Buy: The Science of Shopping – Updated and Revised for the Internet, the Global Consumer, and Beyond*. Simon & Schuster, 2008.

V

[vanderVlist2007] VAN DER VLIST, Eric. *Schematron*. O'Reilly, 2007.

[VanderWal2007] VANDER WAL, Thomas. *Folksonomy: vanderwal.net*. 2007. http://vanderwal.net/folksonomy.html.

[VanDuyne2006] VANDUYNE, Douglas K., LANDAY, J.A., and HONG, J.I. *The Design of Sites: Patterns for Creating Winning Web Sites*. (2nd Edition). Prentice Hall, 2006.

[Vargo2004] VARGO, Stephen and LUSCH, Robert F. *"Evolving to a new dominant logic for marketing.": Journal of marketing*. 2004. 1-17.

[vonAhn2004] VON AHN, Luis and DABBISH, Laura. *"Labeling images with a computer game": Proceedings of the 2004 Conference on Human Factors in Computing Systems*. 2004. pp. 319-326.

[vonAhn2008] VON AHN, Luis and DABBISH, Laura. *"Designing games with a purpose": Communications of the ACM*. 2008. pp. 57-67.

[vonRiegen2006] VON RIEGEN, Claus. *How Standards Address Interoperability Needs: An Industry View: OASIS Symposium, May 10, 2006, San Francisco*. 2006.

W

[Wagemans2012] WAGEMANS, Johan, et al. *"A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization.": Psychological bulletin*. 2012. p 1172.

[Wakabayashi2011] WAKABAYASHI, Daisuke. *"Japanese Farms Look to the 'Cloud'": The Wall Street Journal*. 2011. http://online.wsj.com/article/SB10001424052748704029704576087910899748444.html.

[Walker2009] WALKER, Rob. *"The Song Decoders at Pandora": The New York Times*. 2009. http://www.nytimes.com/2009/10/18/magazine/18Pandora-t.html.

[Walsh2010] WALSH, Norman. *DocBook 5: The Definitive Guide*. O'Reilly, 2010.

[Want2006] WANT, Roy. *"An Introduction to RFID Technology": Pervasive Computing*. 2006. pp. 25-33.

[Watson2007] WATSON, Hugh J. and WIXOM, Barbara H. *"The Current State of Business Intelligence": Computer*. 2007. pp. 96-99.

[Watts2004] WATTS, Duncan J. *"The 'New' Science of Networks": Annual Review of Sociology*. 2004. pp. 243-270.

[Weick1995] WEICK, Karl E. *"Sensemaking in organizations": Sage*. 1985.

[Weick2005] WEICK, Karl, SUTCLIFFE, Kathleen, and OBSTFELD, David. *"Organizing and the process of sensemaking": Organization science*. 2005. pp. 409-42.

[Weill2004] WEILL, Peter and ROSS, Jeanne. *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Harvard Business Review Press, 2004.

[Weinreich2001] WEINREICH, Harald, OBENDORF, Hartmut, and LAMERSDORF, Winfried. *"The Look of the Link—Concepts for the User Interface of Extended Hyperlinks": Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (HYPERTEXT '01)*. 2001. pp. 19-28.

[Wheatley2004] WHEATLEY, Malcolm. *"Operation Clean Data": CIO*. 2004. http://www.cio.com.au/article/166533/operation_clean_data/.

[Whistler1890] WHISTLER, James. Project Gutenberg, 1890. http://www.gutenberg.org/ebooks/24650.

[Wilbert2006] WILBERT, Caroline. *How Wal-Mart Works: HowStuffWorks.com*. 2006. http://money.howstuffworks.com/wal-mart.htm.

[Wilde2002] WILDE, Erik and LOWE, David. *XPath, XLink, XPointer, and XML: A Practical Guide to Web Hyperlinking and Transclusion*. Addison Wesley, 2002.

[Wilde2007] WILDE, Erik and CATTIN, Philippe. *"Presenting in HTML": DocEng '07: Proceedings of the 2007 ACM Symposium on Document Engineering*. 2007.

[Wilde2008a] WILDE, Erik. *The Plain Web*. 2008. http://dret.net/netdret/docs/wilde-wsw2008/.

[Wilde2008b] WILDE, Erik and GLUSHKO, Robert J. *"Document design matters": Communications of the ACM*. 2008. pp. 43-49. http://portal.acm.org/citation.cfm?doid=1400181.1400195.

[Wilde2011] WILDE, Erik and PAUTASSO, Cesare. *REST: From Research to Practice*. Springer, 2011.

[Wilkins1668] WILKINS, John. *"Essay towards a Real Character and a Philosophical Language"*. The Royal Society, 1668.

[Williams2012] WILLIAMS, Robin. *The Non-Designer's Design Book*. Peachpit Press, 2012.

[Williamson1975] WILLIAMSON, Oliver E. *"Markets and hierarchies: analysis and antitrust implications: a study in the economics of internal organization": University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*. 1975. http://ssrn.com/abstract=1496220.

[Williamson1998] WILLIAMSON, Oliver E. *The Economic Institutions of Capitalism*. Free Press, 1998.

[Wilson1968] WILSON, Patrick. *Two Kinds of Power: An Essay on Bibliographical Control*. University of California Press, 1968.

[Winkler2010] WINKLER, Stefan and VON PILGRIM, Jens. *"A survey of traceability in requirements engineering and model-driven development": Software and Systems Modeling*. 2010. pp. 529-565.

[Winner1980] WINNER, Langdon. *"Do artifacts have politics?": Daedelus*. 1980.

[Wittgenstein2002] WITTGENSTEIN, Ludwig. *"Philosophical Investigations, Sections 65-78": In Foundations of Cognitive Psychology: Core Readings*. The MIT Press, 2002. pp. 271-276.

[Winston1987] WINSTON, Morton E., CHAFFIN, Roger, and HERRMANN, Douglas. *"A taxonomy of part-whole relations": Cognitive Science*. 1987. 417-444.

[Wright2010] WRIGHT, Alex. *"Managing Scientific Inquiry in a Laboratory the Size of the Web": The New York Times*. 2010. http://www.nytimes.com/2010/12/28/science/28citizen.html.

[Wu2012] WU, Michael. *The Science of Social: Beyond Hype, Likes, and Followers*. Lithium Technologies, 2012.

X

Y

[Yang2010] YANG, Xin-She. *Nature-Inspired Optimization Algorithms*. Luniver Press, 2010.

[Yanovsky2014] YANOVSKY, David. *Here are the 32 Countries Google Maps Won't Draw Borders Around*. Quartz, 10 June 2014.

[Yau2011] YAU, Nathan. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. Wiley, 2011.

[Yee2008] YEE, Raymond. *Pro Web 2.0 Mashups: Remixing Data and Web Services*. Apress, 2008.

Z

[Zeithami2001] ZEITHAMI, Valarie A., RUST, Roland T., and LEMON, Katherine N. *"Customer Pyramid: Creating and Serving Profitable Customers": California Management Review*. 2001. pp. 118-142.

[Zhou2003] ZHOU, Rongron and SOMAN, Dilip. *"Looking Back : Exploring the Psychology of Queuing and the Effect of the Number of People Behind": Journal of Consumer Research*. 2003. pp. 517-530.

[Zhu2014] ZHU, H., et al. *"Data and Information Quality Research: Its Evolution and Future"*. 2014.

Glossary

> **Note**
>
> The glossary presents an alphabetical listing of entries, each with a term and a corresponding meaning. As much as possible and wherever practical, the contents of the glossary definitions are transcluded from the chapters. In the case of abbreviations, the meaning transcludes the expanded form of the abbreviation. Where abbreviations relate to formal organizations or standards, we provide a reference URI to encourage discovery.—MM

A

AAP
: Association of American Publishers (AAP)

  (http://www.publishers.org/)

AAT
: Art and Architecture Thesaurus (AAT)

  (http://www.getty.edu/research/tools/vocabularies/aat/)

aboutness
: "Subject matter" organization involves the use of a classification system that provides categories and descriptive terms for indicating what a resource is about. Because they use *aboutness* properties that are not directly perceived, methods for assigning subject classifications are intellectually-intensive and in many cases require rigorous training to be performed consistently and appropriately. (From §3.3 Organizing Resources.)

absolute synonyms
: The strictest definition is that *synonyms* "are words that can replace each other in some class of contexts with insignificant changes of the whole text's meaning." (From §6.4.1.3 Synonymy.)

  See also synonym

abstract models
: **Abstract models** describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain. (From abstract models.)

accessioning
: Adding a resource to a library collection is called *acquisition*, but adding to a museum collection is called *accessioning*. (From §3.1 Introduction.)

See also acquisition, collection development

accuracy
See precision.

ACM
Association for Computing Machinery (ACM)

(http://www.acm.org/)

acquisition
Adding a resource to a library collection is called *acquisition*, but adding to a museum collection is called *accessioning*. (From §3.1 Introduction.)

See also accessioning, collection development

active resources
Active resources create effects or value on their own, sometimes when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, web-based services, self-driving cars, robots, appliances, machines or otherwise ordinary objects like light bulbs, umbrellas, and shoes that have been made "smarter." (From §4.2.3.2 Active or Operant Resources.)

activities
There are four *activities* that occur naturally in every *organizing system*; how explicit they are depend on the scope, the breadth or variety of the resources, and the scale, the number of resources that the organizing system encompasses. (From §3.1 Introduction.)

See also selecting, organizing, designing resource-based interactions, maintaining

ad hoc category
An **ad hoc category** or goal-derived category is a collection of resources that happen to go together to satisfy a goal. The resources might not have any discernible properties in common. (From ad hoc.)

affordance
The concept of *affordance*, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor's capabilities and cognition, what can be done with the resource. (From §3.4.1 Affordance and Capability.)

See also capability

agency
Agency is the extent to which a resource can initiate actions on its own. We can define a continuum between completely passive resources that cannot initiate any actions and active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources. (From §4.2.3 Resource Agency.)

agent
We use the more general word, *agent*, for any entity capable of autonomous and intentional organizing effort, because it treats organizing work done by people and organizing work done by computers as having common goals, despite obvious differences in methods. (From §1.7 The Concept of "Agent".)

agents
A facet in the hierarchical structure of the AAT thesaurus. Basically, people and the various groups and organizations with which they identify, whether based on physical, mental, socio-economic, or political characteristics—e.g., "stonemasons" or "socialists."

(From §8.4.2 Faceted Classification in Description (page 421).)

alias
: See synonym.

alphabetical ordering
: **Alphabetical ordering** is arranging resources according to their names (From §1.6 The Concept of "Organizing Principle".)

  See also chronological ordering

American Society for Information Science and Technology
: See ASIS&T.

analysis
: A common interaction with an organizing system.

analytico-synthetic classification
: In library science a classification system that builds categories by combination of facets is sometimes also called *analytico-synthetic*. (From §8.1.4 Classification Schemes.)

anchor text
: In web contexts, the words in which a structural link is embedded are called the *anchor text*. (From §6.5.3.1 Hypertext Links.)

  See also hypertext

ANSI
: American National Standards Association (ANSI)

  (http://www.ansi.org/)

antonymy
: *Antonymy* is the lexical relationship between two words that have opposite meanings. *Antonymy* is a very salient lexical relationship, and for adjectives it is even more powerful than synonymy. (From §6.4.1.5 Antonymy.)

APA
: American Psychiatric Association (APA)

  (http://www.psych.org)

API
: application program interfaces (APIs)

appraisal
: What is the value of this resource? What is its cost? At what rate does it depreciate? Does it have a shelf life? Does it have any associated ratings, rankings, or quality measures? Moreover, what is the quality of those ratings, rankings and measures? (From §5.3.2.1 Resource Description to Support Selection (page 234).)

architectural perspective
: The architectural perspective emphasizes the number and abstraction level of the components of a relationship, which together characterize its complexity. (From §6.2 Describing Relationships: An Overview (page 275).)

arity
: The *degree* or *arity* of a relationship is the number of entity types or categories of resources in the relationship. This is usually, though not always, the same as the number of arguments in the relationship expression. (From §6.6.1 Degree.)

  See also degree

ARPA
: Advanced Research Projects Agency

  (http://www.darpa.mil/)

artifact
: See resource.

ASCII
: American Standard Code for Information Interchange (ASCII)

  American National Standard for Information Systems—Coded Character Sets—7-Bit American National Standard Code for Information Interchange (7-Bit ASCII), ANSI X3.4-1986, Ameri-

can National Standards Institute, Inc., March 26, 1986

asset
: See resource.

ASIS&T
: American Society for Information Science and Technology (ASIS&T)

  (http://www.asis.org)

associated resource
: See description resources.

associative array
: See dictionary.

asymmetric relationships
: **Asymmetric relationships** express a one-way relationship from the subject to the object. For example, "is-parent-of."

  See also hypertext, directionality, and one-way link. From asymmetric relationship.)

attribute
: **Attribute** is a synonym for "*property*."

  To **attribute** is to assert or assign a value to a property. See attribution relationship

  An **attribute** is a syntactic component of XML elements and a conceptual component of the XML Infoset, consisting of a potentially qualified name and a value, whose type may influence its interpretation. The value of an attribute in an XML document is a Unicode string. The value of that attribute in the XML Infoset could be a simple string of text, a precisely-typed numeric or temporal value, a list of references to document nodes, a hypertext link, or a reference to a formal notation. (See also element item)

attribution relationship
: Asserting or assigning values to properties; the predicate depends on the property: "is-the-author-of," "is-married-to," "is-employed-by," etc. (From §6.3.1 Types of Semantic Relationships (page 278).)

authentication
: Is the resource what it claims to be? ((page 202)§4.5.3 Authenticity) Resource descriptions that can support authentication include technological ones like time stamps, watermarking, encryption, checksums, and digital signatures. (From §5.3.2 Determining the Purposes (page 234).)

authenticity
: In ordinary use we say that something is *authentic* if it can be shown to be, or has come to be accepted as what it claims to be. The importance and nuance of questions about authenticity can be seen in the many words we have to describe the relationship between "the real thing" (the "original") and something else: copy, reproduction, replica, fake, phony, forgery, counterfeit, pretender, imposter, ringer, and so on. (From §4.5.3 Authenticity.)

  See also provenance

authority control
: For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as *authority control*. (From §4.4.3.2 Use Controlled Vocabularies.)

B

BI
: Business Intelligence (BI)

bi-directional
: See symmetric relationships.

bi-directional links
: When a *bi-directional link* is created between an anchor and a destination, it is as though a one-way link that can be followed in the opposite direction is automatically created. Two one-way links serve the same purpose, but the return link is not automatically established when the first one is created. (From §6.5.3.1 Hypertext Links.)

  See also hypertext, directionality, one-way link

bibliographic description
: *Bibliographic descriptions* characterize information resources and the entities that populate the bibliographic universe, which include works, editions, authors, and subjects. (From §5.2.2.1 Bibliographic Descriptions.)

bibliography
: A **bibliography** is a description resource in the domain of library science. (Ed.)

bibliometrics
: Information scientists began studying the structure of scientific citation, now called *bibliometrics*, nearly a century ago to identify influential scientists and publications. (From §6.5.3.3 Bibliometrics, Shepardizing, Altmetrics, and Social Network Analysis.)

big data
: For digital resources, inexpensive storage and high bandwidth have largely eliminated capacity as a constraint for organizing systems, with an exception for *big data*, which is defined as a collection of data that is too big to be managed by typical database software and hardware architectures. (From §11.5.2 Architectural Thinking.)

binary antonyms
: Contrasting or *binary antonyms* are used in mutually exclusive contexts where one or the other word can be used, but never both. For example, “alive” and “dead” can never be used at the same time to describe the state of some entity, because the meaning of one excludes or contradicts the meaning of the other. (From §6.4.1.5 Antonymy.)

binary link
: A **binary link** connects one anchor to one destination. (From binary links.)

  See also hypertext

blob
: A **blob** is any resource whose internal structure is functionally opaque for the purpose at hand. (From blob.)

Boolean facet
: Take on one of two values, yes (true) or no (false) along some dimension or property. (From Boolean facets.)

  See also §8.4.3 A Classification for Facets (page 424)

born digital
: Resources in organizing systems that are created in digital format are **born digital**. These include resources created by word processors and digital cameras, or by audio and video recorders. Other resources are produced in digital form by “smart things” and by the systems that create digital resources when they interact with barcodes, QR (“quick response”) codes, RFID tags, or other mechanisms for tracking identity and location. (From born digital.)

BPEL
: Business Process Execution Language (BPEL)

  (https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsbpel)

C

CAFE
: Corporate Average Fuel Economy (CAFE)

  (http://www.nhtsa.gov/staticfiles/rulemaking/pdf/cafe/CAFE-GHG_MY_2012-2016_Final_Rule_FR.pdf)

capability
: **Capability** is a function of the affordances offered by an *organizing system* and the possible interactions they imply. (From capability.)

  See also affordance

capability and compatibility
: Will the resource meet functional or interoperability requirements? Technology-intensive resources often have numerous specialized types of descriptions that specify their functions, performance, reliability, and other "-ilities" that determine if they fit in with other resources in an organizing system. (From §5.3.2.1 Resource Description to Support Selection (page 234).)

cardinality
: The *cardinality* of a relationship is the number of instances that can be associated with each entity type in a relationship. (From §6.6.2 Cardinality.)

cataloging
: Documenting the contents of library and museum collections to organize them is called *cataloging* (From §3.1 Introduction.)

  See also collection development

categories
: *Categories* are *equivalence classes*, sets or groups of things or abstract entities that we treat the same. (From §7.2 The What and Why of Categories.)

  See also equivalence class

CBS
: CBS Corporation and CBS Broadcasting Inc.

  (http://www.cbs.com/)

CC
: *Common Cartridge and Learning Tools Interoperability*

  (http://www.imsglobal.org/commoncartridge.html)

centrality
: The **centrality** of a resource instance as a member of a category is a measure of how close it is to a mathematical average on some measures or property values that apply to all the members. (From §7.3.5 Probabilistic Categories and "Family Resemblance" (page 348).)

CERN
: European Organization for Nuclear Research (*Centre Européen de Research Nuclear*)

  (http://public.web.cern.ch/public/)

character
: Unicode makes the important distinction between *characters* and *glyphs*. A *character* is the smallest meaningful unit of a written language. In alphabet-based languages like English, *characters* are letters; in languages like Chinese, characters are ideographs. (From §9.3.1 Notations.)

character encoding
: A notation that has had numbers assigned to its characters is called a *character encoding*. (From §9.3.1 Notations.)

The most ambitious character coding in existence is Unicode, which as of version 6.0 assigns numbers to 109,449 characters. Unicode makes the important distinction between characters and glyphs.

chronological ordering

Chronological ordering is arranging resources according to the date of their creation or other important event in the lifetime of the resource (From §1.6 The Concept of "Organizing Principle".).

See also alphabetical ordering

CIDR

Classless Inter-Domain Routing (CIDR)

circulation

We might treat *circulation*, borrowing and returning the same item, as one of the interactions with resources that defines a library. (From §1.4 The Concept of "Collection".)

classes

In object-oriented programming languages, *classes* are schemas that serve as templates for the creation of objects. A class in a programming language is analogous to a database schema that specifies the structure of its member instances, in that the class definition specifies how instances of the class are constructed in terms of data types and possible values. Programming classes may also specify whether data in a member object can be accessed, and if so, how. (From §7.5.2 Implementing Categories Defined by Properties.)

classical categories

Categories defined by necessary and sufficient properties are also called *monothetic*. They are also sometimes called *classical categories* because they conform to Aristotle's theory of how categories are used in logical deduction using syllogisms. (From §7.3.3.3 Necessary and Sufficient Properties.)

classification

The systematic assignment of resources to a system of intentional categories, often institutional ones. (From §8.1.1 Classification vs. Categorization.)

Classification is applied categorization – the assignment of resources to a system of categories, called classes, using a predetermined set of principles.

See also inclusion

classification scheme

See classifications

classifications

A system of categories and its attendant rules or access methods is typically called a *classification scheme* or just the *classifications*. A system of categories captures the distinctions and relationships among its resources that are most important in a domain and for a particular context of use, creating a reference model or conceptual roadmap for its users. (From §8.1 Introduction.)

classifying

When we make an assertion that a particular instance is a member of class, we are *classifying* the instance. (From §6.3.1.1 Inclusion.)

class inclusion

Class inclusion is the fundamental and familiar "**is-a,**" "**is-a-type-of**," or "**subset**" relationship between two entity types or classes where one is contained in and thus more specific than the other more generic one. (From §6.3.1.1 Inclusion.)

See also inclusion

clustering
: *Clustering* techniques share the goal of creating meaningful categories from a collection of items whose properties are hard to directly perceive and evaluate, which implies that category membership cannot easily be reduced to specific property tests and instead must be based on similarity. (From §7.5.3.3 Categories Created by Clustering (page 372).)

  The end result of **clustering** is a statistically optimal set of categories in which the similarity of all the items within a category is larger than the similarity of items that belong to different categories.

cognates
: Many words in different languages have common roots, and as a result are often spelled the same or nearly the same. This is especially true for technology words; for example, "computer" has been borrowed by many languages. The existence of these *cognates* and borrowed words makes us vulnerable to false cognates. When a word in one language has a different meaning and refers to different resources in another, the results can be embarrassing or disastrous. "*Gift*" is poison in German; "*pain*" is bread in French. (From §4.4.2.2 Homonymy, Polysemy, and False Cognates.)

collection
: A *collection* is a group of resources that have been selected for some purpose. (From §1.4 The Concept of "Collection".)

collection development
: Libraries and museums usually make their *selection* principles explicit in *collection development* policies. Adding a resource to a library collection is called *acquisition*, but adding to a museum collection is called *accessioning*. Documenting the contents of library and museum collections to organize them is called *cataloging*. *Circulation* is a central interaction in libraries, but because museum resources do not circulate the primary interactions for museum users are *viewing* or *visiting* the collection. *Maintenance* activities are usually described as *preservation* or *curation*. (From §3.1 Introduction.)

collocation
: The Organizing System for a small collection can sometimes use only the minimal or default organizing principle of *colocation*—putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box. (From §1.6 The Concept of "Organizing Principle".)

compliance
: **Compliance** is a maintenance activity.

component-object inclusion
: *Component-Object* is the relationship type when the part is a separate component that is arranged or assembled with other components to create a larger resource. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

compounding
: Putting two "free morphemes" together. (From Compounding.)

constraint
: A limit or bound on a data type or structure, most usefully expressed in a schema or regular expression. Constraints on data types and structures can be expressed in a variety of natural, programming and schema languages with varying degrees of efficacy. (Ed.)

content rules
: *Content rules* are similar to controlled vocabularies because they also limit the possible values that can be used in descriptions. Instead of specifying a fixed set of values, content rules typically restrict descriptions by requiring them to be of a particular data type (integer, Boolean, Date, and so on). (From §5.3.4.3 Controlled Vocabularies and Content Rules.)

contextual properties
: *Contextual properties* are those related to the situation or context in which a resource is described. Dey defines *context* as "any information that characterizes a situation related to the interactions between users, applications, and the surrounding environment." (From §5.3.3.4 Extrinsic Dynamic Properties.)

controlled vocabulary
: One way to encourage good names for a given resource domain or task is to establish a *controlled vocabulary*. A *controlled vocabulary* is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use. (From §4.4.3.2 Use Controlled Vocabularies.)

coverage
: The values of a facet should be able of classifying all instances within the intended scope. (From §8.4.4.2 Design Principles and Pragmatics (page 425).)

CRM
: Customer Relationship Management (CRM)

crosswalk
: Similar to mapping, a straightforward approach to transformation is the use of *crosswalks*, which are equivalence tables that relate resource description elements, semantics, and writing systems from one organizing system to those of another. (From §10.3.2.2 Modes of Transformation (page 503).)

cultural categories
: *Cultural categories* are the archetypical form of categories upon which individual and institutional categories are usually based. Cultural categories tend to describe our everyday experiences of the world and our accumulated cultural knowledge. (From §7.2.1 Cultural Categories.)

cultural properties
: **Cultural properties** derive from conventional language or culture, often by analogy, because they can be highly evocative and memorable. (From §5.3.3.4 Extrinsic Dynamic Properties (page 245)

curation
: **Curation** is a maintenance activity.

  Curation usually refers to the methods or systems that add value to and preserve resources, while the concept of *governance* more often emphasizes the institutions or organizations that carry out those activities. The former is most often used for libraries, museums, or archives and the latter for enterprise or inter-enterprise contexts. (From §2.5 When Is It Being Organized?.)

D

data
: **Data** is a collection of one or more pieces if information. The singular noun form is "datum"; the plural forms

are “datums" and “data"; the collective noun form is also “data". For example: Starting with a single datum; many more datums are subsequently identified; those data are then intentionally arranged; and, finally the data is organized.

data activities
: **Data** capture, extraction and generation are resource selection activities.

  Data cleaning and cleansing are maintenance activities.

  Data insertion and integration add resources to a collection.

data rot
: **Data rot** is a colloquial term intended to convey the fact that the physical medium of a digital resource deteriorates over time.

data schema
: *Data schemas* that specify data entities, elements, identifiers, attributes, and relationships in databases and XML document types on the transactional end of the Document Type Spectrum (§4.2.1) are implementations of the categories needed for the design, development and maintenance of information organization systems. Data schemas tend to rigidly define categories of resources. (From §7.5.2 Implementing Categories Defined by Properties.)

data science
: Data science, actuarial science, statistics, probability, and predictive analytics. Predicting future outcomes by applying statistical analysis over many large datasets and calculating probabilities. Ancient roots in the fields of economics, insurance, cartography, astronomy, and meteorology.

DC
: Dublin Core (DC)

  (http://dublincore.org/documents/dcmi-terms/)

  See also [Hillmann2005]

decision tree
: A simple *decision tree* is an algorithm for determining a decision by making a sequence of logical or property tests. (From §7.5.2 Implementing Categories Defined by Properties.)

decoding
: A digital resource is first a sequence of bits. Decoding transforms those bits into characters according to the encoding scheme used, extracting the text from its stored form.

degree
: The *degree* or *arity* of a relationship is the number of entity types or categories of resources in the relationship. This is usually, though not always, the same as the number of arguments in the relationship expression. (From §6.6.1 Degree.)

derivational morphology
: *Derivational morphology* deals with how words are created by combining morphemes. (From §6.4.3.1 Derivational Morphology.)

description resources
: Any primary resource can have one or more *description resources* associated with it to facilitate finding, interacting with, or interpreting the primary one. *Description resources* are essential in organizing systems where the primary resources are not under its control and can only be accessed or interacted with through the description. *Description resources* are often called *metadata*. (From §4.2.4 Resource Focus.)

  Description resources, such as physical or online catalog records, de-

scribe the *primary resources* that comprise the collection.

descriptor
: In the library science context of *bibliographic description,* a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources. (From §5.2.2 "Description" as an Inclusive Term.)

designed resource access policies
: Designed resource access policies are established by the designer or operator of an organizing system to satisfy internally generated requirements. (From §3.1 Introduction (page 87).)

designing resource-based interactions
: Designing and implementing the actions, functions or services that make use of the resources. (From §3.1 Introduction (page 87).)

dictionary
: A *dictionary* is a set of property-value pairs or entries. It is a set of entries, not a list of entries, because the pairs are not ordered and because each entry must have a unique key.

  Note that this specialized meaning of *dictionary* is different from the more common meaning of "dictionary" as an alphabetized list of terms accompanied by sentences that define them. (From §9.2.1.4 Dictionaries.)

digitization
: Other digital resources are created by *digitization*, the process for transforming an artifact whose original format is physical so it can be stored and manipulated by a computer. (From §4.2.2 Resource Format.)

dimensionality reduction
: Dimensionality reduction implies transforming a high-dimensional space into a lower-dimensional one. Reducing the number of components in a multidimensional description can be accomplished by many different statistical techniques that go by names like "feature extraction," "principle components analysis," "orthogonal decomposition," "latent semantic analysis," "multidimensional scaling," and "factor analysis." (From §5.3.4.4 Vocabulary Control as Dimensionality Reduction.)

directionality
: The *directionality* of a relationship defines the order in which the arguments of the relationship are connected. A *one-way* or *uni-directional* relationship can be followed in only one direction, whereas a *bi-directional* one can be followed in both directions. (From §6.6.3 Directionality.)

  See also hypertext, directionality, one-way link, bi-directional

discipline
: A *discipline* is an integrated field of study in which there is some level of agreement about the issues and problems that deserve study, how they are interrelated, how they should be studied, and how findings or theories about the issues and problems should be evaluated. (From §1.1 The Discipline of Organizing.)

discovery
: What available resources might be added to a collection? New resources are often listed in directories, registries, or catalogs. Some types of resources are selected and acquired automatically through subscriptions or contracts. (From §5.3.2 Determining the Purposes (page 234).)

DNS
: Domain Name System (DNS)

  (http://tools.ietf.org/html/rfc1035)

document
: See resource.

document frequency
: *Inverse document frequency* (idf) is a collection-level property. The *document frequency* (df) is the number of resources containing a particular term. The inverse document frequency (idf) for a term is defined as $idf_t = \log(N/df_t)$, where N is the total number of documents. The inverse document frequency of a term decreases the more documents contain the term, providing a discriminating factor for the importance of terms in a query. (From §10.4.2.1 Ranked Retrieval with Vector Space or Probabilistic Models.)

DOI
: Digital Object Identifier (DOI)

  (http://www.doi.org)

domain
: *Resource domain* is an intuitive notion that groups resources according to the set of natural or intuitive characteristics that distinguishes them from other resources. It contrasts with the idea of ad hoc or arbitrary groupings of resources that happen to be in the same place at some time. (From §4.2.1 Resource Domain.)

DPLA
: Digital Public Library of America (DPLA)

  (http://dp.la/)

DRM
: digital rights management (DRM)

DSM
: Diagnostic and Statistical Manual of Mental Disorders (DSM)

  (http://www.dsm5.org/)

DTD
: Document Type Definition (DTD)

E

ECM
: Enterprise Content Management (ECM)

edge
: See tree.

EDI
: Electronic Data Exchange (EDI)

  Typically refers to one or all of the UN/EDIFACT, ANSI ASC X12, TRADCOMS and ODETTE standards.

EDM
: Enterprise Data Management (EDM)

effectivity
: Many resources, or their properties, also have locative or temporal *effectivity*, meaning that they come into effect at a particular time and/or place; will almost certainly cease to be effective at some future date, and may cease to be effective in different places. (From §4.5.2 Effectivity.)

element item
: An *element* item has a set of *attribute* items, and a list of child nodes. These child nodes may include other element items, or they may be character items. (From §9.2.2.2 XML Information Set (page 451).)

encoding scheme
: An *encoding scheme* is a specialized *writing system* or syntax for particular types of values. (From §9.2.3.2 Controlling Values.)

energy facet
: One of Ranganathan's universal facets in colon classification. The action or activity of the thing. (From

§8.4.1 Foundations for Faceted Classification (page 420).)

entity
: See resource.

entity type
: See classes

enumeration
: The simplest principle for creating a category is *enumeration*; any resource in a finite or countable set can be deemed a category member by that fact alone. (From §7.3.1 Enumeration.)

  See also extensional definition.

enumerative classification
: Classification schemes in which all possible categories to which resources can be assigned are defined explicitly are *enumerative*. (From §8.1.4 Classification Schemes.)

enumerative facets
: Have mutually exclusive possible values. (From §8.4.3 A Classification for Facets (page 424).)

equivalence class
: See categories

equivalence relationship
: Any relationship that is both symmetric and transitive is an *equivalence relationship*; "is-equal-to" is obviously an equivalence relationship because if A=B then B=A and if A=B and B=C, then A=C. Other relationships can be equivalent without meaning "exactly equal," as is the relationship of "is-congruent-to" for all triangles. (From §6.3.2.3 Equivalence.)

ERP
: Enterprise Resource Planning (ERP)

ETL
: Extract, Transform, and Load

EXIF
: Exchangeable Image File Format (EXIF)

  (`http://www.exif.org/`)

expression
: The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see Figure 4.5, The FRBR Abstraction Hierarchy.) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*. (From §4.3.2 Identity and Bibliographic Resources.)

extensibility of classification
: See flexibility

extension
: See extensional definition

extensional definition
: The simplest principle for creating a category is *enumeration*; any resource in a finite or countable set can be deemed a category member by that fact alone. This principle is also known as *extensional definition*, and the members of the set are called the *extension*. (From §7.3.1 Enumeration.)

F

faceted classification
: In a *faceted classification* system, each resource is described using properties from multiple facets, but a person searching for resources does not need to consider all of the properties (and consequently the facets) and does not need to consider them in a fixed order, which an enumerative hierarchical classification requires. (From §8.4 Faceted Classification.)

family resemblance
: A second consequence is that the sharing of some but not all properties creates what we call *family resemblances* among the category members; just as biological family members do not necessarily all share a single set of physical features but still are recognizable as members of the same family. (From §7.3.5 Probabilistic Categories and "Family Resemblance" (page 348).)

FCC
: Federal Communications Commission (FCC)

  (http://www.fcc.gov/)

FDA
: Food and Drug Administration (FDA)

  (http://www.fda.gov/)

feature
: *Feature* is used in *data science* and *machine learning* contexts for both "raw" or observable variables and "latent" ones, extracted or constructed from the original set. (From §3.3 Organizing Resources (page 98).)

  See also property

feature-activity inclusion
: *Feature-Activity* is a relationship type in which the components are stages, phases, or sub activities that take place over time. This relationship is similar to component-object in that the components in the whole are arranged according to a structure or pattern. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

FERPA
: Family Educational Rights and Privacy Act (FERPA)

  (http://www2.ed.gov/policy/gen/guid/fpco/ferpa/)

finding
: What resources are available that "correspond to the user's stated search criteria" and thus can satisfy an information need? Modern users accept that computerized indexing makes search possible over not only the entire description resource, but often over the entire content of the primary resource. (From §5.3.2.3 Resource Description to Support Interactions (page 236).)

flexibility of classification
: A related principle about maintaining classifications over time is *flexibility*, the degree to which the classification can accommodate new categories. Computer scientists typically describe this principle as *extensibility*, and library scientists sometimes describe it as *hospitality*. (From §8.2.2.3 Principles for Maintaining the Classification over Time.)

FOAF
: Friend of a Friend (FOAF)

  (http://www.foaf-project.org/)

focus
: The contrast between *primary resources* and *description resources* is very useful in many contexts, but when we look more broadly at organizing systems, it is often difficult to distinguish them, and determining which resources are primary and which are *metadata* is often just a decision about which resource is currently the *focus* of our attention. (From §4.2.4 Resource Focus.)

font
: A *font* is a collection of glyphs used to depict some set of *characters*. A Unicode font explicitly associates each glyph with a particular number in the Unicode character encoding. (From §9.3.1 Notations.)

form
: We treat the set of implementation decisions about character notations, syntax, and structure as the *form* of resource description (From §5.2.3 Frameworks for Resource Description.)

format
: Information resources can exist in numerous **formats** with the most basic format distinction being whether the resource is physical or digital.

framework
: A *framework* is a set of concepts that provide the basic structure for understanding a domain, enabling a common vocabulary for different explanatory theories. (From §1.1 The Discipline of Organizing.)

frequency of use principle
: Some organization emerges implicitly through a *frequency of use* principle. In your kitchen or clothes closet, the resources you use most often migrate to the front because that is the easiest place to return them after using them. (From §1.6 The Concept of "Organizing Principle".)

FTC
: Federal Trade Commission (FTC)

  (`http://www.ftc.gov/`)

FTP
: File Transfer Protocol (FTP)

  (`http://tools.ietf.org/html/rfc959`)

G

globally unique identifier (GUID)
: A **globally unique identifier** (or GUID), is an identifier that will never be the same as another identifier in any organizing system anywhere else. (Ed.)

glossary definition
: A glossary definition states the meaning of its corresponding term. There must be one and there may be more definitions for a given term. The most common presentation is a set of words or symbols that convey the semantic of the term, such as the expanded form of an abbreviation or acronym, or a paragraph of text. Definition by reference is often used for synonym terms.

  See also synonym

glyph
: A specific mark that can be used to depict a *character* is a *glyph*. (From §9.3.1 Notations.)

  See also character, font

governance
: *Curation* usually refers to the methods or systems that add value to and preserve resources, while the concept of *governance* more often emphasizes the institutions or organizations that carry out those activities. The former is most often used for libraries, museums, or archives and the latter for enterprise or inter-enterprise contexts. (From §2.5 When Is It Being Organized?.)

GPS
: Global Positioning System

  (`http://www.schriever.af.mil/GPS/`)

gradience
: When category members differ in their centrality or typicality to the category definition, this effect is called category gradience. (From §7.3.5 Probabilistic Categories and "Family Resemblance" (page 348).)

grammar
: The *syntax* and *grammar* of a language consists of the rules that determine which combinations of its words

are allowed and are thus grammatical or *well-formed*. Natural languages have substantial similarities by having nouns, verbs, adjectives and other parts of speech, but they differ greatly in how they arrange them to create sentences. (From §6.7.2 Syntax and Grammar.)

granularity
: *Granularity* refers to the level of detail or precision for a specific information resource property. For instance, the postal address of a particular location might be represented as several different data items, including the number, street name, city, state, country and postal code (a high-granularity model). It might also be represented in one single line including all of the information above (a low-granularity model). (From §10.3.2.3 Granularity and Abstraction.)

graph
: Like a *tree*, a *graph* consists of a set of nodes connected by edges. These edges may or may not have a direction ((page 307)§6.6.3 Directionality). If they do, the *graph* is referred to as a "directed graph." If a *graph* is directed, it may be possible to start at a node and follow edges in a path that leads back to the starting node. Such a path is called a "cycle." If a directed graph has no cycles, it is referred to as an "acyclic graph." (From §9.2.1.6 Graphs.)

GUID
: Globally Unique Identifier

H

hierarchical classification
: When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is *hierarchical* or *taxonomic*. (From §8.1.4 Classification Schemes.)

hierarchical facet
: Organize resources by logical inclusion (§6.3.1.1). (From §8.4.3 A Classification for Facets (page 424).)

  See also inclusion

HIPAA
: Health Insurance Portability and Accountability Act (HIPAA)

  (`http://www.hhs.gov/ocr/privacy/hipaa/understanding/`)

homographs
: When two words are spelled the same but have different meanings they are *homographs*; if they are also pronounced the same they are *homonyms*. If the different meanings of the homographs are related, they are *polysemes*. (From §4.4.2.2 Homonymy, Polysemy, and False Cognates.)

homonyms
: **Homonyms** are homographs that are pronounced the same. (From §4.4.2.2 Homonymy, Polysemy, and False Cognates.)

hospitality of classification
: See flexibility

HR
: Human Resources

HTML
: Hypertext Markup Language (HTML)

  (`http://www.w3.org/community/webed/wiki/HTML/Specifications`)

HTTP
: Hypertext Transfer Protocol (HTTP)

  (`http://www.w3.org/Protocols/`)

hypernym
: When words encode the semantic distinctions expressed by class inclusion, the word for the more specific class in

this relationship is called the *hyponym*, while the word for the more general class to which it belongs is called the *hypernym*. (From §6.4.1.1 Hyponymy and Hyperonymy.)

hypertext
: **Hypertext** expresses relationships among resources. Hypertext is "a provision whereby any item may be caused at will to select immediately and automatically another." It can be used to create non-sequential narratives that gives choices to readers. (See §6.5.3.1 Hypertext Links.)

hypertext link
: The concept of read-only or follow-only structures that connect one document to another is usually attributed to Vannevar Bush in his seminal 1945 essay titled "As We May Think." Bush called it *associative indexing,* defined as "a provision whereby any item may be caused at will to select immediately and automatically another." (From §6.5.3.1 Hypertext Links.)

hyponym
: When words encode the semantic distinctions expressed by class inclusion, the word for the more specific class in this relationship is called the *hyponym*, while the word for the more general class to which it belongs is called the *hypernym*. (From §6.4.1.1 Hyponymy and Hyperonymy.)

I

IAU
: International Astronomical Union (IAU)

  (http://www.iau.org/)

IBM
: International Business Machines (IBM)

  (http://www.ibm.com)

ICANN
: Internet Corporation for Assigned Names and Numbers (ICANN)

  (http://www.icann.org/)

identifier
: An *identifier* is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. (From §4.1.2 Identity, Identifiers, and Names.)

identifying
: Another purpose of resource description is to enable a user to confirm the identity of a specific resource or to distinguish among several that have some overlapping descriptions. Computer processable resource descriptions like bar codes, QR codes, or RFID tags are also used to identify resources. In Semantic Web contexts, URIs serve this purpose. (From §5.3.2.3 Resource Description to Support Interactions (page 236).)

identity
: When some thing or things are treated as a single resource this establishes an **identity**. (Ed.)

IEEE
: Institute of Electrical and Electronics Engineers

  (http://www.ieee.org/index.html)

IETF
: Internet Engineering Task Force

  (http://ietf.org)

implementation perspective
: The implementation perspective considers how the relationship is implemented in a particular notation and syntax and the manner in which relationships are arranged and stored in some technology environment. (From

§6.2 Describing Relationships: An Overview (page 275).)

implicit classification
: Because names and dates can take on a great many values, an organizing principle like *alphabetical* or *chronological* ordering is unlikely to enumerate in advance an explicit category for each possible value. Instead, we can consider these organizing principles as creating an *implicit or latent* classification system in which the categories are generated only as needed. For example, the Q category only exists in an alphabetical scheme if there is a resource whose name starts with Q. (From §8.1.4 Classification Schemes.)

imposed policies
: *Imposed Policies* are mandated by an external entity and the organizing system must comply with them. (From §3.4.3 Access Policies.)

inclusion relationship
: One entity type contains or is comprised of other entity types; often expressed using "is-a," "is-a-type-of," "is-part-of," or "is-in" predicates. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also component-object, feature-activity inclusion, locative, member-collection, meronymic, part-whole, phase-activity, place-area, portion-mass, stuff-object, temporal, topological, taxonomy and classification

index
: An *index* is a *description resource* that contains information about the locations and frequencies of terms in a document *collection* to enable it to be searched efficiently. (From §1.4 The Concept of "Collection".)

individual categorization
: *Individual categories* are created in an organizing system to satisfy the *ad hoc* requirements that arise from a person's unique experiences, preferences, and resource collections. Unlike cultural categories, which usually develop slowly and last a long time, individual categories are created by intentional activity, in response to a specific situation, or to solve an emerging organizational challenge. (From §7.2.2 Individual Categories.)

inflectional morphology
: Inflectional mechanisms change the form of a word to represent tense, aspect, agreement, or other grammatical information. Unlike derivation, inflection never changes the part-of-speech of the base morpheme. The *inflectional morphology* of English is relatively simple compared with other languages. (From §6.4.3.2 Inflectional Morphology.)

informatics
: Informatics is a broad academic category encompassing the science of information, including the automation of information processing. Computer science, information architecture and web architecture are among the related academic disciplines.

information architecture
: Abstract patterns of information content or organization are sometimes called architectures, so it is straightforward from the perspective of the discipline of organizing to define the activity of *information architecture* as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use. (From §3.3.3.2 "Information Architecture" and Organizing Systems.)

information component
: An *information component* can be: (1) Any piece of information that has a unique label or identifier or (2) Any

piece of information that is self-contained and comprehensible on its own. (From §4.3.3 Identity and Information Components.)

information organization
: Traditional information organization activities include bibliographic description and cataloging.

information retrieval
: Traditional information retrieval activities include automated text processing, indexing and search.

instance
: See resource.

institutional categorization
: In contrast to cultural categories that are created and used implicitly, and to individual categories that are used by people acting alone, *institutional categories* are created and used explicitly, and most often by many people in coordination with each other. Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories are needed to regulate and systematize activity, to enable information sharing and reuse, and to reduce transaction costs. (From §7.2.3 Institutional Categories.)

institutional semantics
: Systems of *institutional semantics* offer precisely defined abstractions or *information components* ((page 185) §4.3.3 Identity and Information Components) needed to ensure that information can be efficiently exchanged and used. (From §8.1.5.2 Institutional Semantics.)

institutional taxonomies
: *Institutional taxonomies* are classifications designed to make it more likely that people or computational agents will organize and interact with resources in the same way. (From §8.1.5.1 Institutional Taxonomies.)

integration
: *Integration* is the controlled sharing of information between two (or more) business systems, applications, or services within or between firms. *Integration* means that one party can extract or obtain information from another one, it does not imply that the recipient can make use of the information. (From §6.8.3 Integration and Interoperability.)

integrity of classification
: Changes in the meaning of the categories in a classification threaten its *integrity*, the principle that categories should not move within the structure of the classification system. (From §8.2.2.3 Principles for Maintaining the Classification over Time.)

intension
: Categories whose members are determined by one or more properties or rules follow the principle of *intensional definition*, and the defining properties are called the *intension*. (From §7.3.2 Single Properties.)

intensional definition
: Categories whose members are determined by one or more properties or rules follow the principle of *intensional definition*, and the defining properties are called the *intension*. (From §7.3.2 Single Properties.)

intentional arrangement
: *Intentional arrangement* emphasizes explicit or implicit acts of organization by people, or by computational processes acting as proxies for, or as implementations of, human intentionality. (From §1.5 The Concept of "Intentional Arrangement".)

interaction
: An *interaction* is an action, function, service, or capability that makes use of the resources in a collection or the collection as a whole. The interaction of *access* is fundamental in any collection of resources, but many Organizing Systems provide additional functions to make access more efficient and to support additional interactions with the accessed resources. (From §1.8 The Concept of "Interactions".)

interoperability
: *Interoperability* goes beyond integration to mean that systems, applications, or services that exchange information can make sense of what they receive. *Interoperability* can involve identifying corresponding components and relationships in each system, transforming them syntactically to the same format, structurally to the same granularity, and semantically to the same meaning. (From §6.8.3 Integration and Interoperability.)

inverse document frequency
: *Inverse document frequency* (idf) is a collection-level property. The *document frequency* (df) is the number of resources containing a particular term. The inverse document frequency (idf) for a term is defined as $idf_t = \log(N/df_t)$, where N is the total number of documents. The inverse document frequency of a term decreases the more documents contain the term, providing a discriminating factor for the importance of terms in a query. (From §10.4.2.1 Ranked Retrieval with Vector Space or Probabilistic Models.)

inverse relationship
: For asymmetric relationships, it is often useful to be explicit about the meaning of the relationship when the order of the arguments in the relationship is reversed. The resulting relationship is called the *inverse* or the converse of the first relationship. (From §6.3.2.4 Inverse.)

ISBN
: International Standard Book Number (ISBN)

  (`http://www.isbn.org/`)

ISO
: International Organization for Standardization (ISO)

  (`http://www.iso.org/iso/`)

item
: The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see Figure 4.5, The FRBR Abstraction Hierarchy.) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*. (From §4.3.2 Identity and Bibliographic Resources.)

  See also resource

J

JavaScript Object Notation (JSON)
: *JavaScript Object Notation (JSON)* is a textual format for exchanging data that borrows its metamodel from the JavaScript programming language. Specifically, the JSON metamodel consists of two kinds of structures found in JavaScript: lists (called "arrays" in JavaScript) and dictionaries (called "objects" in JavaScript). (From §9.2.2.1 JSON (page 451).)

  (`http://www.json.org/`)

JPEG
: Joint Photographic Experts Group

  (`http://www.jpeg.org/`)

K

KM
: Knowledge Management (KM)

KMS
: *Knowledge management systems (KMS)* are a type of business organizing system whose goal is to capture and systematize these information resources. (From §3.5.2.4 Preserving Resource Types.)

L

LCC
: Library of Congress Classification (LCC)

  (`http://www.loc.gov/catdir/cpso/lcc.html`)

learns
: See machine learning.

lexical gap
: A **lexical gap** in a language exists when it lacks a word for a concept that is expressed as a word in another language. (From lexical gap.)

lexical perspective
: The lexical perspective focuses on how the conceptual description of a relationship is expressed using words in a specific language. (From §6.2 Describing Relationships: An Overview (page 275).)

linguistic relativity
: Languages differ a great deal in the words they contain and also in more fundamental ways that they require speakers or writers to attend to details about the world or aspects of experience that another language allows them to ignore. This idea is often described as *linguistic relativity*. (From §7.2.1 Cultural Categories.)

link
: See hypertext link

link base
: A **link base** is a collection of links stored separately from the resources that they link. (Mentioned in the sidebar, Perspectives on Hypertext Links (page 303).)

link type
: When it is evident, this semantic property of the link is called the *link type*. (From §6.5.3.1 Hypertext Links.)

list
: A *list*, like a *set*, is a collection of items with an additional constraint: their items are ordered. (From §9.2.1.3 Lists.)

literary warrant
: The principle of *literary warrant* holds that a classification must be based only on the specific resources that are being classified. (From §8.2.2.1 Principles Embodied in the Classification Scheme.)

LM
: language models (LM)

LMS
: Learning Management System (LMS)

loading
: Adding resources to a collection.

LOC-CN
: Library of Congress Call Number (LOC-CN)

LOC-SH
: Library of Congress Subject Headings (LOC-SH)

locative inclusion
: **Locative inclusion** is a type of inclusion relationship between an area and what it surrounds or contains. It is most often expressed using “is-in” as the relationship. However, the entity that is contained or surrounded is not a part of the including one, so this is not a part-whole relationship.

See also §6.3.1.1 Inclusion (page 279)

logical hierarchy
: If multiple resource properties are considered in a fixed order, the resulting arrangement forms a *logical hierarchy*. (From §3.3.5 Organizing with Multiple Resource Properties.)

M

machine learning
: *machine learning* is home to numerous techniques for creating classifiers by training them with already correctly categorized examples. This training is called *supervised learning*; it is supervised because it starts with instances labeled by category, and it involves learning because over time the classifier improves its performance by adjusting the weights for features that distinguish the categories. But strictly speaking, supervised learning techniques do not learn the categories; they implement and apply categories that they inherit or are given to them. (From §7.2.5 Computational Categories.)

MADS
: Metadata Authority Description Standard (MADS)

  (`http://www.loc.gov/standards/mads/`)

maintaining
: Managing and adapting the resources and the organization imposed on them as needed to support the interactions. (From §3.1 Introduction (page 87).)

manifestation
: The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see Figure 4.5, The FRBR Abstraction Hierarchy.) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*. (From §4.3.2 Identity and Bibliographic Resources.)

map
: See dictionary

markup
: **Markup** is an encoding of character content with a layer of intentional coding, typically by surrounding the character text with "pointy brackets" or tags whose name suggests a content type, structural role, or formatting. (Ed.)

materiality
: It is the requirement to recognize the *materiality* of the environment that enables people to create and interact with digital resources (From §3.3.3 Organizing Digital Resources.)

materials facet
: Concerned with the actual substance of which a work is made, like "metal" or "bleach." "Materials" differ from "Physical Attributes" in that the latter is more abstract than the former. (From §8.4.2 Faceted Classification in Description (page 421).)

matter facet
: One of Ranganathan's universal facets in colon classification. The constituent material of the thing. (From §8.4.1 Foundations for Faceted Classification (page 420).)

member-collection inclusion
: *Member-Collection* is the part-whole relationship type where "is-part-of" means "belongs-to," a weaker kind of association than component-object because there is no assumption that the component has a specific role or function in the whole. (From §6.3.1 Types of Semantic Relationships (page 278).)

See also inclusion

memory institution
: The concept of *memory institution* broadly applies to a great many organizing systems that share the goal of preserving knowledge and cultural heritage. (From §3.5.1 Motivations for Maintaining Resources.)

meronymic inclusion
: See part-whole

  See also inclusion

MeSH
: Medical Subject Headings (MeSH)

  (http://www.nlm.nih.gov/mesh/)

metadata
: *Metadata* is often defined as "data about data," a definition that is nearly as ubiquitous as it is unhelpful. A more content-full definition of metadata is that it is structured description for information resources of any kind. (From §5.2.2.2 Metadata.)

  See also description resources

metamodels
: When common sets of design decisions can be identified that are not specific to any one domain, they often become systematized in textbooks and in design practices, and may eventually be designed into standard formats and architectures for creating organizing systems. These formally recognized sets of design decisions are known as *abstract models* or *metamodels*. *Metamodels* describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain. (From §9.2 Structuring Descriptions.)

metonymy
: Part-whole or meronymic semantic relationships have lexical analogues in *metonomy*, when an entity is described by something that is contained in or otherwise part of it. (From §6.4.1.2 Metonymy.)

monothetic categories
: **Monothetic categories** are defined by necessary and sufficient properties.

  See classical categories

morphemes
: See morphology

morphology
: The basic building blocks for words are called *morphemes* and can express semantic concepts (when they are called *root words*) or abstract concepts like "pastness" or "plural"). The analysis of the ways by which languages combine *morphemes* is called *morphology*. (From §6.4.3 Relationships among Word Forms.)

MPAA
: Motion Picture Association of America (MPAA)

  (http://www.mpaa.org/)

N

n-ary links
: **n-ary links** connect one anchor to multiple types of destinations. (Mentioned in n-ary links.)

NAICS
: North American Industry Classification System (NAICS)

  (http://www.census.gov/eos/www/naics/)

name
: A *name* is a label for a resource that is used to distinguish one from another. (From §4.1.2 Identity, Identifiers, and Names.)

name matching
: In organizing systems that contain data, there are numerous tools for *name matching,* the task of determining

when two different text strings denote the same person, object, or other named entity. (From §3.5.3.4 Computational Curation.)

namespace
: We can prevent or reduce identifier collisions by adding information about the *namespace*, the domain from which the names or identifiers are selected, thus creating what are often called *qualified names*. (From §4.4.3.4 Make Identifiers Unique or Qualified.)

NAPO
: National Association of Professional Organizers (NAPO)

  (http://www.napo.net/)

natural language processing
: Natural language processing

navigation
: If users are not able to specify their information needs in a way that the *finding* functionality requires, they should be able to use relational and structural descriptions among the resources to navigate from any resource to other ones that might be better. (From §5.3.2.2 Resource Description to Support Organizing (page 236).)

NCSA
: National Center for Supercomputing Applications (NCSA)

  (http://www.ncsa.illinois.edu/)

NFL
: National Football League (NFL)

  (http://www.nfl.com/)

NIH
: National Institute of Health (NIH)

  (http://www.nih.gov/)

NIST
: National Institute of Standards and Technology (NIST)

  (http://www.nist.gov/)

NLP
: Natural Language Processing (NLP)

node
: **Nodes** are objects in an entity-relationship system.

  In the RDF metamodel, a pair of nodes and its edge is called a *triple*, because it consists of three parts (two nodes and one edge). The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple. (From §9.2.2.3 RDF.)

notation
: A **notation** a set of characters with distinct forms. (From §9.3.1 Notations.)

  The Latin alphabet is a *notation*, as are Arabic numerals. Some more exotic *notations* include alchemical symbols and the symbols used for editorial markup.

NSF
: National Science Foundation (NSF)

  (http://www.nsf.gov/)

O

OASIS
: Organization for the Advancement of Structured Information Standards (OASIS)

  (https://www.oasis-open.org/)

object
: In the RDF metamodel, a pair of nodes and its edge is called a *triple*, because it consists of three parts (two nodes and one edge). The RDF metamodel is

a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple. (From §9.2.2.3 RDF.)

See also: resource

object warrant
: With classifications of physical resources like those in a kitchen, we see *object warrant*, where similar objects are put together, but more frequently the justifying principle will be one of use warrant, where resources are organized based on how they are used. (From §8.2.2.1 Principles Embodied in the Classification Scheme.)

objectivity
: Although every classification has an explicit or implicit bias ((page 408) §8.2.3 Classification Is Biased), facets and facet values should be as unambiguous and concrete as possible to enable reliable classification of instances. (From §8.4.4.2 Design Principles and Pragmatics (page 425).)

objects facet
: The largest facet, *objects* contains the actual works, like "sandcastles" and "screen prints." (From §8.4.2 Faceted Classification in Description (page 421).)

obtaining
: Physical resources often require significant effort to obtain after they have been selected. (From §5.3.2.2 Resource Description to Support Organizing (page 236).)

OCLC
: Online Computer Library Center (OCLC)

  (http://www.oclc.org/)

OECD
: Organization for Economic Cooperation and Development (OECD)

  (http://www.oecd.org/)

OMG
: Object Management Group (OMG)

  (http://www.omg.org/)

one-way
: Allowing physical or conceptual movement in one direction only. (Ed.)

one-way link
: A **one-way link** asserts a link from a resource to one or more resources. A one-way link does not imply a link in the return direction, or among the target resources. (From one-way.)

  See also hypertext

  See also directionality

ONIX
: Online Information Exchange (ONIX)

  (http://www.editeur.org/8/ONIX)

ontology
: **Ontology** is a branch of philosophy concerned with what exists in reality and the general features and relations of whatever that might be. Computer science has adopted *ontology* to refer to any computer-processable resource that represents the relationships among words and meanings in some knowledge domain. (See ontology

organize
: To *organize* is to create capabilities by intentionally imposing order and structure. (From §1.1 The Discipline of Organizing.)

organizing
: Specifying the principles or rules that will be followed to arrange the resources. (From §3.1 Introduction (page 87).)

organizing principles
: *Organizing principles* are directives for the design or arrangement of a *collection* of resources that are ideally expressed in a way that does not assume any particular implementation or realization. (From §1.6 The Concept of "Organizing Principle"

organizing system
: *Organizing System*: an intentionally arranged collection of resources and the interactions they support. (From §1.1 The Discipline of Organizing.)

orthogonality
: Facets should be independent dimensions, so a resource can have values of all of them while only having one value on each of them. (From §8.4.4.2 Design Principles and Pragmatics (page 425).)

OWL
: Web Ontology Language (OWL)

  (http://www.w3.org/TR/owl2-overview/)

P

part-whole inclusion
: *Part-whole inclusion* or *meronymic inclusion* is a second type of inclusion relationship. It is usually expressed using "is-part-of," "is-partly," or with other similar predicate expressions. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

passive resources
: *Passive resources* are usually tangible and static and thus they become valuable only as a result of some action or interaction with them. (From §4.2.3.1 Passive or Operand Resources.)

PDF
: Portable Document Format (PDF)

  (http://www.adobe.com/products/acrobat/adobepdf.html)

persistence
: **Persistence** is the quality of resisting change over time. See §4.5.1 Persistence (page 199) and §5.3.3 Identifying Properties (page 241).

personality facet
: One of Ranganathan's universal facets in colon classification. The type of thing. (From §8.4.1 Foundations for Faceted Classification (page 420).)

phase-activity inclusion
: *Phase-Activity* is similar to *feature-activity* except that the phases do not make sense as standalone activities without the context provided by the activity as a whole. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

physical attributes facet
: Material characteristics that can be measured and perceived, like "height" and "flexibility." (From §8.4.2 Faceted Classification in Description (page 421).)

PIM
: Personal Information Management (PIM)

place-area inclusion
: *Place-Area* relationships exist between areas and specific places or locations within them. Like members of collections, places have no particular functional contribution to the whole. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

polysemes
: If the different meanings of the homographs are related, they are called **polysemes**. (From §4.4.2.2 Homony-

my, Polysemy, and False Cognates (page 189).)

polysemy
: **Polysemy** is the linguistic term for words with multiple meanings or senses. (From §4.4.2.2 Homonymy, Polysemy, and False Cognates (page 189).)

polythetic
: Categories defined by family resemblance or multiple and shifting property sets are termed *polythetic*. (From §7.3.5 Probabilistic Categories and "Family Resemblance" (page 348).)

POP
: Post Office Protocol (POP)

  (`https://tools.ietf.org/html/rfc1939`)

portion-mass inclusion
: **Portion-Mass** is the relationship type when all the parts are similar to each other and to the whole. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

possession relationship
: Asserting ownership or control of a resource; often expressed using a "has" predicate, such as "has-serial-number-plate." (From §6.3.1 Types of Semantic Relationships (page 278).)

precision
: *Precision* measures the *accuracy* of a result set, that is, how many of the retrieved resources for a query are relevant. (From §10.5.2.2 The Recall / Precision Tradeoff.)

predicate
: A *predicate* is a verb phrase template for specifying properties of objects or a relationship among objects. (From §6.3 The Semantic Perspective.)

preservation
: **Preservation** is a maintenance activity.

preservation metadata
: **Preservation metadata** is technical information about resource formats and technology needed to ensure resource and collection integrity in a maintenance context. (From §5.3.2.4 Resource Description to Support Maintenance (page 238).)

primary resource
: Treating as a *primary resource* anything that can be identified is an important generalization of the concept because it enables web-based services, data feeds, objects with RFID tags, sensors or other *smart devices,* or computational agents to be part of Organizing Systems. (From §1.3 The Concept of "Resource".)

property
: In this book we use *property* in a generic and ordinary sense as a synonym for *feature* or "characteristic." Many cognitive and computer scientists are more precise in defining these terms and reserve *property* for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the *property* is called an *attribute,* "dimension," or "variable." (From §3.3 Organizing Resources.)

property-based categorization
: *Property-based categorization* works tautologically well for categories like "prime number" where the category is defined by necessary and sufficient properties. Property-based categorization also works well when properties are conceptually distinct and the value of a property is easy to perceive and examine, as they are with man-made physical resources like shirts. (From §7.3.4 The Limits of Property-Based Categorization.)

propositional synonyms
: *Propositional synonyms* are not identical in meaning, but they are equivalent enough that substituting one for the other will not change the truth value of the sentence. (From §6.4.1.3 Synonymy.)

provenance
: **Provenance** is the history of the ownership of a collection or the resources in it, where they have been and who has possessed them. In organizing systems like museums and archives that preserve rare or culturally important objects or documents, **provenance** describes a record of who has authenticated a resource over time. (From §4.5.4 Provenance (page 203)

Q

QR
: Quick Response (QR)

  (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43655)

qualified names
: **Qualified names** are identifiers which explicitly identify the domain, or namespace, from which they are drawn, thereby reducing identifier collision. (From §4.4.3.4 Make Identifiers Unique or Qualified (page 197).)

quality
: A **quality** is an attribute or property of a resource. A quality is logically ascribable by a subject. (Ed.)

  Quality is a measure of the fitness of purpose of a resource or service. It is the difference between what was planned or expected versus what was realized or manifest; it is as an assessment of the suitability of a resource or interaction. (Ed.)

querying
: Querying is a very common interaction in many organizing systems, including libraries, museums, archives, computer science, information architecture, data science, the Web, philosophy, cognitive sciences, linguistics, business, and law. Formulating a query in natural language is typically a precursor to application of more systematized techniques discussed throughout this book.

  (See especially Chapter 2, *Design Decisions in Organizing Systems*, Chapter 10, *Interactions with Resources*, and Chapter 12, *Case Studies*

R

RDF
: Resource Description Framework (RDF)

  (http://www.w3.org/RDF/)

RDF vocabulary
: A set of RDF predicate names and URIs is known as an *RDF vocabulary*. (From §9.2.3.1 Specifying Vocabularies and Schemas.)

reachability
: **Reachability** is the "can you get there from here" property between two resources in a directed graph. (From §6.5.3.2 Analyzing Link Structures (page 303).)

recall
: *Recall* measures the completeness of the result set, that is, how many of the relevant resources in a collection were retrieved. (From §10.5.2.2 The Recall / Precision Tradeoff.)

regular expressions
: Regular expressions are used in computing for matching text patterns. A regular expression is written in a for-

mal language, which may vary among implementations.

See the sidebar, Regular Expressions (page 461) in §9.2.3.2 Controlling Values (page 459).

relationship
: A **relationship** is an association among several things, with that association having a particular significance. (From §6.1 Introduction (page 273).)

RELAX-NG
: REgular LAnguage for XML Next Generation (RELAX NG)

(`http://relaxng.org/`)

relevance
: The concept of *relevance* and its relationship to effectiveness is pivotal in information retrieval and machine learning interactions. (From §10.5.2 Effectiveness.)

reporting
: A common interaction with an organizing system.

representation
: A principle of good descriptions: Use descriptions that reflect how the resources describe themselves; assume that self-descriptions are accurate. (From §5.3.4.1 Principles of Good Description (page 247).)

resolution
: For a digital resource, its identifier serves as the input to the system or function that determines its location so it can be retrieved, a process called *resolving* the identifier or *resolution*. (From §4.1.2 Identity, Identifiers, and Names.)

resource
: *Resource* has an ordinary sense of anything of value that can support goal-oriented activity. This definition means that a resource can be a physical thing, a non-physical thing, information about physical things, information about non-physical things, or anything you want to organize. Other words that aim for this broad scope are *entity*, *object*, *item*, and *instance*. *Document* is often used for an information resource in either digital or physical format; *artifact* refers to resources created by people, and *asset* for resources with economic value.

  Resource has specialized meaning in Internet architecture. It is conventional to describe web pages, images, videos, and so on as *resources*, and the protocol for accessing them, *Hypertext Transfer Protocol (HTTP)*, uses the *Uniform Resource Identifier (URI)*. (From §1.3 The Concept of "Resource".)

resource description
: We describe resources so that we can refer to them, distinguish among them, search for them, manage access to them, preserve them, and make predictions about what might happen to them or what they might do. Each purpose may require different *resource descriptions*. We use *resource descriptions* in every communication and conversation; they are the enablers of organizing systems.

Resource Description Framework (RDF)
: The Resource Description Framework (RDF) *metamodel* is a directed graph, so it identifies one node (the one from which the edge is pointing) as the subject of the triple, and the other node (the one to which the edge is pointing) as its object. The edge is referred to as the predicate of the triple. (From §9.2.2.3 RDF (page 453).)

REST
: Representational State Transfer (REST)

RFID
: Radio-frequency Identification (RFID)

  See US Patent 4,384,288

rich descriptions
: **Rich descriptions** are created by trained and disciplined professionals, often in institutional contexts.

  See §5.3.6 Creating Resource Descriptions (page 251)

root word
: The form of a word after all affixes are removed. (From §6.4.3 Relationships among Word Forms (page 293).)

S

scalability
: Facet values must accommodate potential additions to the set of instances. Including an "Other" value is an easy way to ensure that a facet is flexible and hospitable to new instances, but it not desirable if all new instances will be assigned that value. (From §8.4.4.2 Design Principles and Pragmatics (page 425).)

scale
: The number of resources and interactions that the collection entails. (Ed.)

schema
: A *schema* (or model, or metadata standard) specifies the set of descriptions that apply to an entire resource type. (From §5.3.1.2 Abstraction in Resource Description.)

scientific warrant
: The principle of *scientific warrant* argues that only the categories recognized by the scientists or experts in a domain should be used in a classification system, and it is often opposed by the principle of *use* or *user warrant*, which chooses categories and descriptive terms according to their frequency of use by everyone, not just experts. (From §8.2.2.1 Principles Embodied in the Classification Scheme.)

scope
: The resource *domain* and *scope* circumscribe the describable properties and the possible purposes that descriptions might serve. (From §5.3 The Process of Describing Resources.)

selecting
: Determining the scope of the organizing system by specifying which resources should be included.

  Selecting in this context means the user activity of using resource descriptions to support a choice of resource from a collection, not the institutional activity of selecting resources for the collection in the first place. (From §5.3.2.2 Resource Description to Support Organizing (page 236).)

self-organizing systems
: *Self-organizing systems* can change their internal structure or their function in response to feedback or changed circumstances. (From §1.5 The Concept of "Intentional Arrangement".)

semantic balance
: Top-level facets should be the properties that best differentiate the resources in the classification domain. The values should be of equal semantic scope so that resources are distributed among the subcategories. Subfacets of "Cookware" like "Sauciers and Saucepans" and "Roasters and Brasiers" are semantically balanced as they are both named and grouped by cooking activity. (From §8.4.4.2 Design Principles and Pragmatics (page 425).)

semantic gap
: The *semantic gap* is the difference in perspective in naming and description when resources are described by auto-

mated processes rather than by people. (From §4.4.2.5 The Semantic Gap.)

semantic perspective
: The **semantic perspective** characterizes the meaning of the association between resources. (From §6.2 Describing Relationships: An Overview (page 275).)

semantic web
: The vision of a *Semantic Web* world builds upon the web world, but adds some further prescriptions and constraints for how to structure descriptions. The Semantic Web world unifies the concept of a resource as it has been developed in this book, with the web notion of a resource as anything with a URI. On the Semantic Web, anything being described must have a URI. Furthermore, the descriptions must be structured as graphs, adhering to the RDF metamodel and relating resources to one another via their URIs. Advocates of Linked Data further prescribe that those descriptions must be made available as representations transferred over HTTP. (From §9.4.3 The Semantic Web World.)

sensemaking
: Sensemaking (or sense-making) is the set of processes used by humans to derive meaning from experience or to enhance our understanding. Philosophy, the cognitive sciences and linguistics are among the related academic disciplines.

SEO
: Search Engine Optimization (SEO)

set
: The simplest way to structure a description is to give it parts and treat them as a *set*. (From §9.2.1.2 Sets.)

SGML
: Standard Generalized Markup Language (SGML)

  (`http://www.w3.org/TR/html4/intro/sgmltut.html`)

Shepardizing
: The analysis of legal citations to determine whether a cited case is still good law is called **Shepardizing** because lists of cases annotated in this way were first published in the late 1800s by Frank Shepard, a salesman for a legal publishing company. (From §6.5.3.3 Bibliometrics, Shepardizing, Altmetrics, and Social Network Analysis (page 304).)

SKOS
: Simple Knowledge Organization System (SKOS)

  (`http://www.w3.org/2004/02/skos/`)

SKU
: Stock Keeping Unit (SKU)

similarity
: *Similarity* is a measure of the resemblance between two things that share some characteristics but are not identical. It is a very flexible notion whose meaning depends on the domain within which we apply it. (From §7.3.6 Similarity (page 351).)

smart things
: See active resources.

social classification
: Using any property of a resource to create a description is an uncontrolled and often unprincipled principle for creating categories is called **social classification** or **tagging**. (From §8.1.2 Classification vs. Tagging (page 393).)

SOA
: Service Oriented Architecture

space facet
: One of Ranganathan's universal facets in colon classification. Where the thing occurs. (From §8.4.1 Foundations for Faceted Classification (page 420).)

spectrum facets
: Assume a range of numerical values with a defined minimum and maximum. Price and date are common spectrum facets. The ranges are often modeled as mutually exclusive regions (potential price facet values might include "$0—$49," "$50—$99," and "$100—$149"). (From §8.4.3 A Classification for Facets (page 424).)

SQL
: Structured Query Language (SQL)

  ISO/IEC 9075:2011 "Information technology – Database languages – SQL"

standardization
: A principle of good description: Standardize descriptions to the extent practical, but also use aliasing to allow for commonly used terms. (From §5.3.4.1 Principles of Good Description (page 247).)

statistical pattern recognition
: See unsupervised learning

stemming
: These processing steps normalize inflectional and derivational variations in terms, e.g., by removing the "-ed" from verbs in the past tense. This homogenization can be done by following rules (*stemming*) or by using dictionaries (*lemmatization*). Rule-based stemming algorithms are easy to implement, but can result in wrongly normalized word groups, for example when "university" and "universe" are both stemmed to "univers." (From §10.3.2 Transforming Resources for Interactions (page 500).)

stopword elimination
: Stopwords are those words in a language that occur very frequently and are not very semantically expressive. Stopwords are usually articles, pronouns, prepositions, or conjunctions. Since they occur in every text, they can be removed because they cannot distinguish them. Of course, in some cases, removing stopwords might remove semantically important phrases (e.g., "To be or not to be"). (From §10.3.2 Transforming Resources for Interactions (page 500).)

storage
: **Storage** is a maintenance activity.

  See also preservation, curation

structural perspective
: The **structural perspective** analyzes the patterns of association, arrangement, proximity, or connection between resources without primary concern for their meaning or the origin of these relationships. (From §6.2 Describing Relationships: An Overview (page 275).)

structured descriptions
: See §5.3.6 Creating Resource Descriptions (page 251)

stuff-object inclusion
: *Stuff-Object* relationships are most often expressed using "is-partly" or "is-made-of" and are distinguishable from component-object ones because the stuff cannot be separated from the object without altering its identity. The stuff is not a separate ingredient that is used to make the object; it is a constituent of it once it is made. (From §6.3.1 Types of Semantic Relationships (page 278).)

  See also inclusion

styles and periods facet
: Artistic and architectural eras and stylistic groupings, such as "Renaissance" and "Dada." (From §8.4.2 Faceted Classification in Description (page 421).)

subject
: In the RDF metamodel, a pair of nodes and its edge is called a *triple*, because it consists of three parts (two nodes and one edge). The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple. (From §9.2.2.3 RDF.)

sufficiency and necessity
: Descriptions should have enough information to serve their purposes and not contain information that is not necessary for some purpose; this might imply excluding some aspects of self-descriptions that are insignificant. (From §5.3.4.1 Principles of Good Description (page 247).)

supervised learning
: In *supervised learning*, a machine learning program is trained with sample items or documents that are labeled by category, and the program learns to assign new items to the correct categories. (From §7.2.5 Computational Categories

surrogate resource
: See description resources.

SUV
: Sport Utility Vehicle (SUV)

SVM
: Support Vector Machine (SVM)

symmetric relationships
: **Symmetric relationships** are bidirectional; they express the same relationship from the subject to object as they do from the object to the subject. For example, "is-married-to."

synonym
: When something has more than one name, each of the multiple names is a **synonym** or **alias**. (From §6.4.1.3 Synonymy (page 290).)

synonymy
: **Synonymy** is the relationship between words that express the same semantic concept. (From §6.4.1.3 Synonymy (page 290).)

synset
: An unordered set of synonyms is often called a **synset**. Synsets are interconnected by both semantic relationships and lexical ones, enabling navigation in either space. (From §6.4.1.3 Synonymy (page 290).)

syntax
: The *syntax* and *grammar* of a language consists of the rules that determine which combinations of its words are allowed and are thus grammatical or *well-formed*. Natural languages have substantial similarities by having nouns, verbs, adjectives and other parts of speech, but they differ greatly in how they arrange them to create sentences. (From §6.7.2 Syntax and Grammar.)

T

tag cloud
: Folksonomies are often displayed in the form of a **tag cloud**,where the frequency with which the tag is used throughout the site determines the size of the text in the tag cloud. The tag cloud emerges through the bottom-up aggregation of user tags

and is a statistical construct, rather than a semantic one. (From §8.1.2 Classification vs. Tagging (page 393)

tagging
: Using any property of a resource to create a description is an uncontrolled and often unprincipled principle for creating categories is called **social classification** or **tagging**. (From §8.1.2 Classification vs. Tagging (page 393)

tagsonomy
: When users or communities establish sets of principles to govern their tagging practices, tagging is even more like classification. Such a tagging system can be called a *tagsonomy*, a neologism we have invented to describe more systematic tagging. (From §8.1.2 Classification vs. Tagging (page 393)

taskonomy
: A task or activity-based classification system is called a **taskonomy**. (From §8.5 Classification by Activity Structure (page 426)

taxonomic classification
: When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is **hierarchical** or **taxonomic**. (From §8.5 Classification by Activity Structure (page 426)

taxonomic facets
: **Taxonomic facets**, also known as hierarchical facets are based on logical containment. (From §8.4.3 A Classification for Facets (page 424)

taxonomy
: A **taxonomy** is a hierarchy that is created by a set of interconnected class inclusion relationships. (From §6.3.1.1 Inclusion (page 279)

  See also inclusion

TCP/IP
: Transmission Control Protocol/Internet Protocol (TCP/IP)

  (`https://tools.ietf.org/html/rfc1180`)

TEI
: Text Encoding Initiative (TEI)

  (`http://www.tei-c.org/index.xml`)

temporal inclusion
: **Temporal inclusion** is a type of inclusion relationship between a temporal duration and what it surrounds or contains. It is most often expressed using "is-in" as the relationship. However, the entity that is contained or surrounded is not a part of the including one, so this is not a part-whole relationship. (From §6.3.1.1 Inclusion (page 279).)

  See also inclusion

term frequency
: A vector space ranking utilizes an intrinsic resource property, the number of individual terms in a resource, called the **term frequency**. For each term, term frequency measures how many times the term appears in a resource. (From §10.4.2.1 Ranked Retrieval with Vector Space or Probabilistic Models (page 508)

theory-based category
: A final psychological principle for creating categories is organizing things in ways that fit a theory or story that makes a particular categorization sensible. A *theory-based category* can win out even if probabilistic categorization, on the basis of *family resemblance* or *similarity* with respect to visible properties, would lead to a different category assignment. (From §7.3.8 Theory-Based Categories.)

thesaurus
: A **thesaurus** is a reference work that organizes words according to their semantic and lexical relationships. Thesauri are often used by professionals when they describe resources. (From §6.4.2 Thesauri (page 292).)

time facet
: One of Ranganathan's universal facets in colon classification. When the thing occurs. (From §8.4.1 Foundations for Faceted Classification (page 420).)

tokenization
: Segments the stream of characters (in an encoding scheme, a space is also a character) into textual components, usually words. In English, a simple rule-based system can separate words using spaces. However, punctuation makes things more complicated. For example, periods at the end of sentences should be removed, but periods in numbers should not. Other languages introduce other problems for tokenization; in Chinese, a space does not mark the divisions between individual concepts. (From the sidebar Text Processing (page 501) in §10.3.2 Transforming Resources for Interactions (page 500).)

topological inclusion
: **Topological inclusion** is a type of inclusion relationship between a container and what it surrounds or contains. It is most often expressed using "is-in" as the relationship. However, the entity that is contained or surrounded is not a part of the including one, so this is not a part-whole relationship. (From §6.3.1.1 Inclusion (page 279).)

  See also inclusion

training set
: A **training set** for supervised learning is taken from the labeled instances. The remaining instances are used for validation. (From §8.6 Computational Classification (page 427).)

transclusion
: The inclusion, by hypertext reference, of a resource or part of a resource into another resource is called *transclusion*. Transclusion is normally performed automatically, without user intervention. The inclusion of images in web documents is an example of transclusion. Transclusion is a frequently used technique in business and legal document processing, where re-use of consistent and up-to-date content is essential to achieve efficiency and consistency. (From §6.5.3.1 Hypertext Links)

transformation
: **Transformation** is a very broad concept but in the context of organizing systems it typically means a change in a resource representation or description. The transformation can involve the selection, restructuring, or rearrangement of resources or parts of them. (See §10.3.2 Transforming Resources for Interactions (page 500).)

transitivity
: *Transitivity* is another property that can apply to semantic relationships. When a relationship is transitive, if X and Y have a relationship, and Y and Z have the same relationship, then X also has the relationship with Z. Any relationship based on ordering is transitive, which includes numerical, alphabetic, and chronological ones as well as those that imply qualitative or quantitative measurement. (From §6.3.2.2 Transitivity.)

tree
: *Trees* consist of nodes joined by edges, recursively nested. When a single, root dictionary is connected to child nodes that are themselves dic-

tionaries, we say that the dictionaries are *nested* into a kind of **tree** structure.

A *tree* is a constrained *graph*. Trees are *directed* graphs because the "parent of" relationship between nodes is asymmetric: the edges are arrows that point in a certain direction. Trees are *acyclic* graphs, because if you follow the directed edges from one node to another, you can never encounter the same node twice. Finally, trees have the constraint that every node (except the root) must have exactly one parent. (From §9.2.1.5 Trees (page 445).)

triple
: In the RDF metamodel, a pair of nodes and its edge is called a *triple*, because it consists of three parts (two nodes and one edge). The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the *subject* of the triple, and the other node (the one to which the edge is pointing) as its *object*. The edge is referred to as the *predicate* or (as we have been saying) *property* of the triple. (From §9.2.2.3 RDF.)

TXL
: Turing eXtender Language (TXL)

  (http://www.txl.ca/)

typicality
: **Typicality** or **centrality** considers some members of the category better examples than others, even if they share most properties. (From §7.3.5 Probabilistic Categories and "Family Resemblance" (page 348).)

U

UBL
: Universal Business Language (UBL)

  (https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ubl)

UK
: United Kingdom (UK)

  (https://www.gov.uk/)

UN
: United Nations (UN)

  (http://www.un.org/en/)

uniqueness principle
: The *uniqueness principle* means the categories in a classification scheme are mutually exclusive. Thus, when a logical concept is assigned to a particular category, it cannot simultaneously be assigned to another category. (From §8.2.2.2 Principles for Assigning Resources to Categories.)

UNSPC
: United Nations Standard Products and Services Code (UNSPC)

  (http://www.unspsc.org/)

unsupervised learning
: In *unsupervised learning*, the program gets the same items but has to come up with the categories on its own by discovering the underlying correlations between the items; that is why unsupervised learning is sometimes called *statistical pattern recognition*. (From: §7.2.5 Computational Categories

  See also: machine learning and supervised learning

URI
: Uniform Resource Identifier (URI)

  (http://www.w3.org/Addressing/)

URL
: Uniform Resource Locator (URL)

  (http://www.w3.org/TR/url/)

URN
: Uniform Resource Name (URN)

(http://www.w3.org/TR/uri-clarification/)

user convenience
: Choose description terms with the user in mind; these are likely to be terms in common usage among the target audience. (From §5.3.4.1 Principles of Good Description (page 247).)

user warrant
: The principle of *scientific warrant* argues that only the categories recognized by the scientists or experts in a domain should be used in a classification system, and it is often opposed by the principle of *use* or *user warrant*, which chooses categories and descriptive terms according to their frequency of use by everyone, not just experts. (From §8.2.2.1 Principles Embodied in the Classification Scheme.)

UUID
: Universally Unique Identifier (UUID)

(http://www.ietf.org/rfc/rfc4122.txt)

V

validation
: *Validation* is the process of verifying that a document or data structure conforms with its schema or schemas. Markup validation confirms the structure of the document. Type validation confirms that the content of leaf nodes conforms with the specification of data types. Content validation confirms that the values of the leaf nodes are appropriate. Link validation confirms the integrity of the links between nodes and between documents. Cross validation is the method commonly used for model selection. Business rule validation confirms compliance with business rules. (Discussed in §7.5.2 Implementing Categories Defined by Properties (page 361), §8.4.4.2 Design Principles and Pragmatics (page 425), §9.2.3.1 Specifying Vocabularies and Schemas (page 458)

value
: We distinguish between the type of the *attribute* and the *value* that it has. For example, the color of any object is an *attribute* of the object, and the *value* of that attribute might be "green." (From §6.3.1.2 Attribution (page 282).)

VIAF
: Virtual International Authority File (VIAF)

(http://viaf.org/)

viewing
: **Viewing** is a central interaction in museums and zoos.

See also collection development

VIN
: Vehicle Identification Number (VIN)

(ISO 3779:2009)

visiting
: **Visiting** is a central interaction in museums and zoos.

See also collection development

visualization
: A common interaction with an organizing system.

vocabulary problem
: Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept. (From §4.4.2.1 The Vocabulary Problem.)

VPN
: Virtual Private Network

W

W3C
: World Wide Web Consortium (W3C)

(http://www.w3.org/)

warrant principle
: The *warrant* principle concerns the justification for the choice of categories and the names given to them. (From §8.2.2.1 Principles Embodied in the Classification Scheme.)

  See also: literary warrant, scientific warrant, user warrant and object warrant

well-formed
: The *syntax* and *grammar* of a language consists of the rules that determine which combinations of its words are allowed and are thus grammatical or *well-formed*. Natural languages have substantial similarities by having nouns, verbs, adjectives and other parts of speech, but they differ greatly in how they arrange them to create sentences. (From §6.7.2 Syntax and Grammar.)

work
: An abstract idea of an author's intellectual or artistic creation.

  The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see Figure 4.5, The FRBR Abstraction Hierarchy.) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*.

writing system
: A *writing system* employs one or more notations, and adds a set of rules for using them. Most writing systems assume knowledge of a particular human language. These writing systems are known as *glottic* writing systems. But there are many writing systems, such as mathematical and musical ones, that are not tied to human languages in this way. Many of the writing systems used for describing resources belong to this latter group, meaning that (at least in principle) they can be used with equal facility by speakers of any language. (From §9.3.2 Writing Systems.)

  Some writing systems, such as XML and JSON, are closely identified with specific metamodels.

WSDL
: Web Services Description Language (WSDL)

  (http://www.w3.org/TR/wsdl)

X

XCBF
: XML Common Biometric Format (XCBF)

  (https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xcbf)

XInclude
: XML Inclusions (XInclude)

  (http://www.w3.org/TR/xinclude/)

XML
: Extensible Markup Language (XML)

  (http://www.w3.org/XML/)

XML Information Set
: The *XML Infoset* is a tree structure, where each node of the tree is defined to be an "information item" of a particular type. Each information item has a set of type-specific properties associated with it. At the root of the tree is a "document item," which has exactly one "element item" as its child. An *element* item has a set of *attribute* items, and a list of child nodes. These child nodes may include other element items, or they may be character items. (See (page 442)§9.2.1 Kinds of Structures below for more on char-

acters.) *Attribute* items may contain character items, or they may contain typed data, such as name tokens, identifiers and references. Element identifiers and references (ID/IDREF) may be used to connect nodes, transforming a tree into a graph. (From §9.2.2.2.)

XSD
: XML Schema Definition Language (XSD)

  (http://www.w3.org/XML/Schema.html)

XSLT
: Extensible Stylesheet Language Transformations (XSLT)

  Based on XML, XSLT is a declarative language designed for transforming XML documents into other documents. For example, XSLT can be used to convert XML data into HTML documents for web display or PDF for print or screen display. XSLT processing entails taking an input document in XML format and one or more XSLT style sheets through a template-processing engine to produce a new document.

  (http://www.w3.org/TR/xslt)

Z

zoo
: A **zoo** is an organizing system for living animals that arranges them according to principles of biological taxonomy or common habitat. (Ed.)

Index

Note
In this PDF-format hypertext ebook, the index is an organizing system that presents an alphabetical arrangement of terms to enable look-up and discovery of corresponding subject matter, selectable with page references that lead the reader to a term's first use, its definition or a featured mention within the content. Where a page number repeats, it indicates a distinct and noteworthy entry on that page; we have indexed down to the paragraph, footnote, and even phrase level. Our intent is for this Index to be a useful discovery interface. If the PDF reader interface frustrates our humble efforts, we apologize for your discomfort.—MM

B

C

D

E

F

G

H

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z