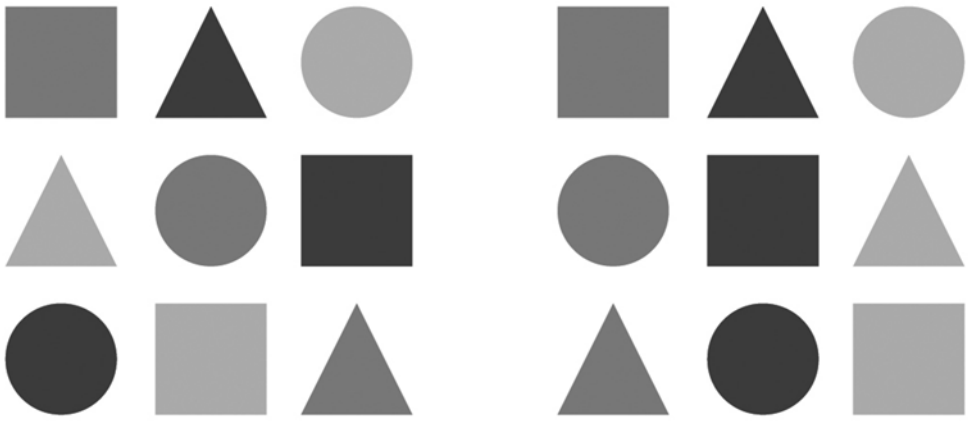
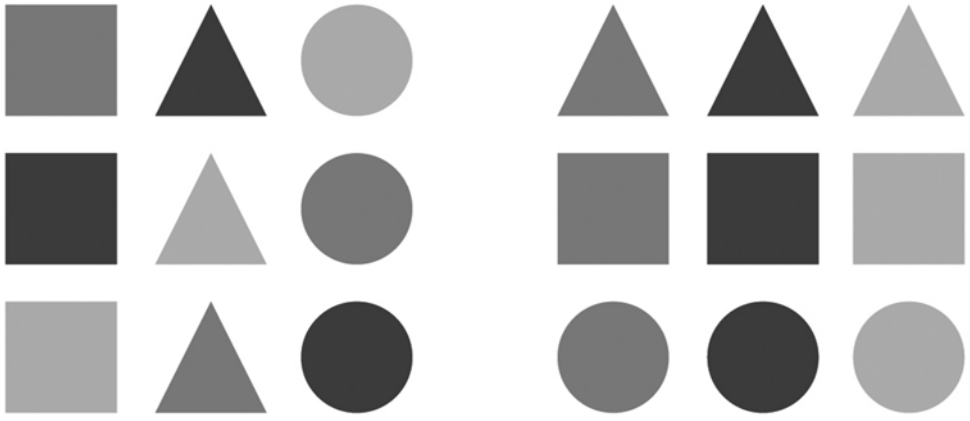


THE DISCIPLINE OF ORGANIZING



edited by ROBERT J. GLUSHKO

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 the shoes in our closet, the books on our book shelves, the spices in our kitchen, and the folders into which we file information for tax and other purposes. 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 things — whether they are physical resources like printed books and shoes or digital resources like web pages and MP3 files — 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.

This book evolved from a master's level university course on "Information Organization & Retrieval" I taught for several years at the *University of California, Berkeley's School of Information*. My goal was to synthesize insights from library science, information science, cognitive science, systems analysis, and computer science to provide my students with a richer understanding about information organization than any discipline alone could provide. I came to realize that information was just one of the many types of resources to organize and that it would be beneficial to think about the art and science of organizing in a more abstract way. This book is the product of countless discussions with students and faculty colleagues at Berkeley and other schools, and we are collaboratively developing a new discipline that unifies four types of organizing, 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.

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, e-books, MP3 and video files, appointments, and contacts— on our computers, smart phone, e-book 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 combined as "mashups" or choreographed to carry out information-intensive business models.

Let's take a closer look at these four different types or contexts of organizing. Are there clear, systematic and useful distinctions between them? 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 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 tool box or together on a work bench, 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 and specify the location of 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?

Geoff Nunberg has eloquently explained in *Farewell to the Information Age* that "Information" is "a collection of notions, rather than a single coherent concept." Most of its hundreds of definitions 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.

Michael Buckland's oft-cited essay *Information as Thing* rebuts the notion that information is inherently intangible and instead defines it more broadly and provocatively based on function. This makes the objects in museum or personal collections into information-as-thing resources because they can be learned from and serve as evidence.[1]

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 if you're lucky 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?

These differences and relationships between "**physical things**" and "**digital things**" have long been discussed and debated by philosophers, linguists, psychologists and others (See the sidebar, "What is Information?" (page 4)).

The distinctions among organizing physical things, organizing digital things, or organizing information about physical things 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, some people use the term "document" to refer only to traditional physical forms, while others use it more abstractly to refer to any self-contained unit of information independent of its instantiation in physical or digital form. The most abstract definition, presented in *What is a Document?* is when Buckland provocatively asserts that an antelope is both "information as thing" and also a "document" when it is in a zoo, even though it is just an animal when it is running wild on the plains of Africa. Similar definitional variation occurs with "author" or "creator."[2]

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 that differ in what they organize. This requires us to focus on situations where information exists because of intentional acts to create or organize.

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.

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 in 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.[3] 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.

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. A ***framework*** is a set of concepts that provide the basic structure for understanding a domain, enabling a common vocabulary for different explanatory theories. 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.

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, "An Organizing System" 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 and generalizes to physical and web-based resource domains the concepts proposed in 2000 for bibliographic domains by Elaine Svenonius in *The Intellectual Foundation of In-*

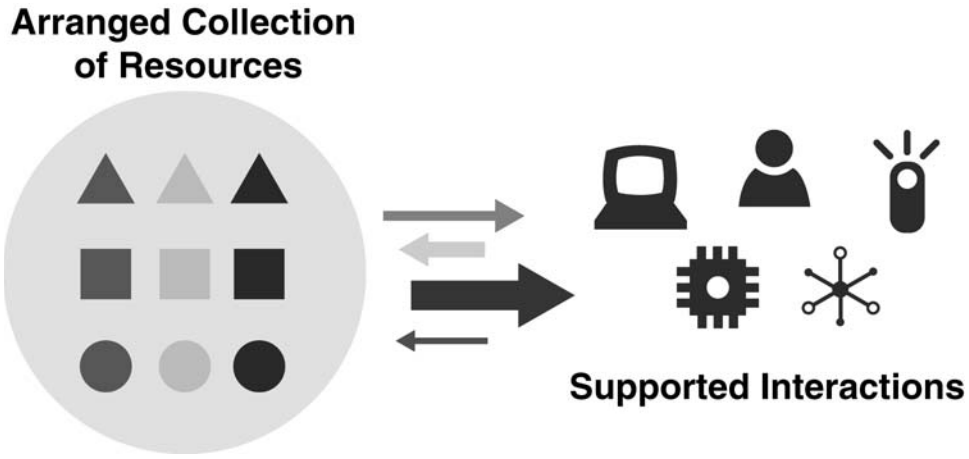


Figure 1.1. An Organizing System

formation 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. She 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 §1.3 with a quote from her book.[4]

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, the more effectively it can be retrieved, and the more effort put into retrieving information, the less it needs to be organized first. 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 by computers to enable retrieval.

1.2.1 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, product catalogs, and so on as resources and the protocol for accessing them, *Hypertext Transfer Protocol (HTTP)*, uses the *Uniform Resource Identifier (URI)*.[5]

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 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.

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, in

the masthead of a newspaper, or in the source of web pages, deciding which resources are primary is often arbitrary. A second challenge is that what serves as a 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" or HR department in a firm. 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.[6] It is important to consider the capabilities and motivations of the people who create and participate in Organizing Systems. We might discuss how human resources are selected, organized, and managed over time just as we might discuss these activities with respect to library resources. Nevertheless, these topics are much more appropriate for texts on human resources management and industrial organization so we will not consider them much further in this book.

1.2.2 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). 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 *collection*s 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 se-

lect, 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.

1.2.3 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 excludes naturally-occurring patterns created by physical, geological, biological or genetic processes. 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 an identifiable agent following one or more organizing principles selected by a human agent.

Self-organizing systems can change their internal structure or their function in response to feedback or changed circumstances. Requiring arrangement to be intentional also excludes those systems from our definition of Organizing System. These self-organizing systems have been used in physics, chemistry, and mathematics to explain phase transitions and equilibrium states. Self-organizing is also used to describe numerous natural and man-made phenomena like climate, communication networks, business and biological ecosystems, traffic and habitation patterns, neural networks, and online communities. All of these systems involve collections of resources that are very large and open, with complex interactions among the resources. 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. Adam Smith's "invisible hand" in economic markets and "Charles Darwin's natural selection" in evolutionary biology are classic examples of self-organizing mechanisms. The web as a whole with its more than a trillion unique pages is in many ways self-organizing, but at its core it follows clear organizing principles (See the Sidebar, "The Web as an Organizing System" (page 10)).[7]

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.[8]

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.

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 §1.3, "Design Decisions in Organizing Systems" (page 18)). 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.

1.2.3.1 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.

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 resources using easily perceived properties to make them easy to locate, we group together resources that we often use together, and we make resources that we use often more accessible than those we use infrequently. Very general and abstract organizing principles are sometimes called design heuristics (for exam-

ple, “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.

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 12).

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.

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. Similarly, many email applications have migrated to the web and the system of filters and folders that manage email messages is no longer implemented in a local network or on personal computers, but most people neither notice nor care.

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." Nevertheless, you can see these different tiers 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 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.

Because tangible things can only be in one place at a time, many Organizing Systems—like that 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 like digitized books.[9] 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 ***collocation*** — putting all the resources 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.[10]

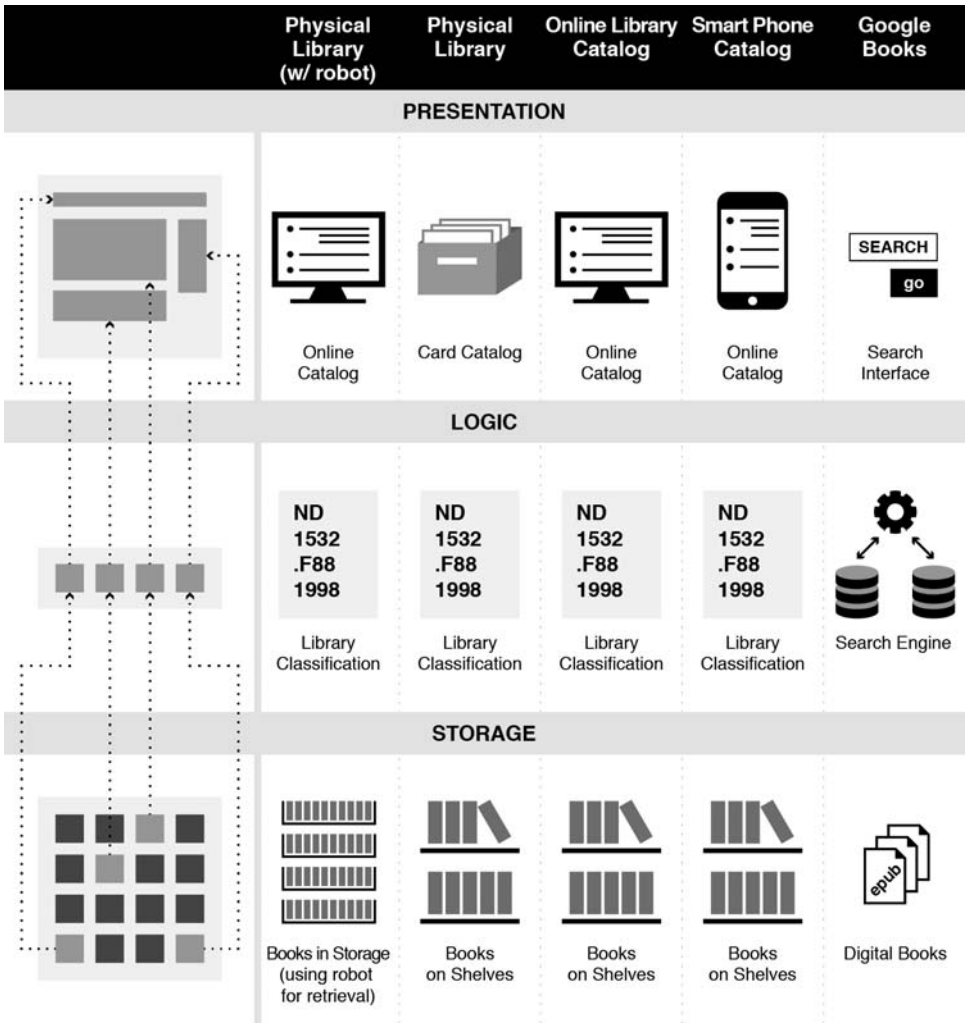


Figure 1.2. Presentation, Logic and Storage Tiers

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

found. As a result, most Organizing Systems employ organizing principles that make use of properties of the resources being organized (for example, 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 based on their content and meaning, and these are not directly apparent when you look at a book or document. Significant intellectual effort or computation is necessary to reveal these properties when assigning subject terms or creating an index. The most effective organizing systems for information resources often are based on 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.

1.2.3.2 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...).[11] 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 like 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 and is also the case with resources augmented with computational or communication capabilities. We are all familiar with RFID tags, which enable the precise identification and location of physical resources as they move through supply chains and stores.

1.2.4 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.[12]

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.[13]

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 Sys-

tem with digital resources can implement a "keep everything up to date" interaction that automatically pushes current content to your browser or computing device.

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.

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. See the Sidebar, "The Digital Zoo" (page 17).

The Digital Zoo

Consider the Organizing System of a zoo, which typically organizes the physical resources that we usually call animals according to principles of biological taxonomy or common habitat. The most important interaction supported by a modern zoo is enabling visitors to observe the behavior of animals in environments resembling their natural environments.

If all the animals are "stored" in a single location and we prefer unmediated interaction with them, the physical requirements for housing and showcasing the animals and our own mobility limits how big a zoo can be. However, this is simply a conventional manner of storing the animals and supporting interactions with them.

We can imagine a "digital zoo" in which all of the world's zoos are treated as a single digital collection. Instead of physical co-presence to enable viewing of the animals, "telepresence" quality video camera connections to zoo exhibits would enable us to view animals in the San Diego, Berlin, Singapore, Toronto and other great zoos from our living rooms almost as if we were there. Instead of a walk from exhibit to exhibit, a tour of a digital zoo would be implemented as a list of video connections, and different tours implemented as different sequences of cameras. Sensors in each animal habitat could detect activity and image recognition software could classify the behavior, enabling us to search the combined zoo for particular combinations of active animals to view (e.g., a mother and infant chimpanzee). No zoo would need to change how it "stores" the animals to enable these new principles of organization and our new kinds of interactions with them.

1.3 Design Decisions in Organizing Systems

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 who interact 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 five 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 enable a resource collection to be created, for an existing and closed resource collection, or for a collection in which resources are continually added or deleted? Are the resources unique, or are they interchangeable members of a class? Do they follow a predictable "life cycle" with a "useful life"?
- **Why it is 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?
- **How much is it being organized?** What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed? Is this description and structure imposed in a centralized or top-down manner or in a distributed or bottom-up manner? What organizing principles guide the organization? Are all resources organized to the same degree, or is the organization sparse and non-uniform?
- **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?
- **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?

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.

1.3.1 Organizing Systems in a “Design Space”

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 are at the same time.

1.3.1.1 Conventional Ways to Classify Organizing Systems

We distinguish law libraries from software libraries, knowledge management systems from data warehouses, and personal stamp collections from coin collections 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.

But there are three other conventional ways to classify Organizing Systems. A second way to distinguish Organizing Systems is by their dominant purposes or the priority of their common purposes. For example, 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.

A third conventional approach for classifying Organizing Systems is according to the nature or size of the intended user community. This size or scope can range from personal Organizing Systems created and used by a single person; to "community-based" Organizing Systems used by informal social groups; to those used by the employees, customers or stakeholders of an enterprise; to those used by an entire community or nation; to global ones potentially used by anyone in the world.

A fourth way to distinguish Organizing Systems is according to the technology used to implement them. Large businesses use different software applications for inventory management, records management, content management, knowledge management, customer relationship management, data warehousing and business intelligence, e-mail archiving, and other subcategories of collections.[14]

We can become overwhelmed by this proliferation of ways to classify collections of resources, especially when the classification is not clearly based on just one of these many approaches. For example, the list of "library types" used by the International Federation of Library Associations to organize its activities includes resource-based distinctions (e.g. art libraries, law libraries, social science libraries), purpose-based ones (e.g., academic and research libraries), and user-based distinctions (e.g., public libraries, school libraries, libraries serving persons with print disabilities).[15]

1.3.1.2 A Multifaceted or Multidimensional View

A type of resource and its conventional Organizing System are often the focal point of a discipline. Category labels like 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.

Consider Borgman's commonly accepted definition of libraries as institutions that "select, collect, organize, conserve, preserve, and provide access to information on behalf of a community of users." Many Organizing Systems are described as libraries, even though they differ from traditional libraries in important respects. See the Sidebar, "What is a Library?" (page 20)

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. Google co-founder Sergei Brin characterized its ambitious project to put tens of millions of books from research libraries online as "a library to last forever." But the Google Books project was widely criticized as not being true to library principles.[16]

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).

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 framing it in library terms to suggest it is a public good upsets 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.

A complementary perspective on Organizing Systems is that they exist in a multi-faceted or multi-dimensional design space. 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 multidisciplinary 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. Try this analogy test: Google Books is to Library as ? is to Natural History Museum.[17]

But even though digital technology is radically subdividing the traditional categories of collections by supporting new kinds of specialized information-intensive applications, an opposite and somewhat paradoxical trend has emerged. Jennifer Trant argues that the common challenges of "going digital," and the architectural and functional constraints imposed by web implementations, are causing some convergence in the operation of libraries, museums, and archives. Similarly, Anne Gilliland suggests that giving every physical resource in a collection a digital surrogate or proxy that is searchable and viewable in a web browser is "erasing the distinctions between custodians of information and custodians of things."[18]

Taken together, these two trends have one profound implication. If the traditional categories for thinking about collections are splintering in some respects and converging in others, they are less useful in describing innovative collec-

tions and their associated interactions. Thus, we need a new concept — the Organizing System — that:

- Applies comprehensively and consistently to collections of resources of any type
- Reuses familiar categories where they are appropriate, but does not impose them on new types of collections and services where they do not fit well
- Makes it easier to trace the connections between specific requirements or constraints and particular functions or implementation choices.

1.3.2 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, as one of many equivalent members of a broad category, or as 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, an SKU can be associated with a unique item, to sets of items treated as equivalent for inventory or billing purposes, or to 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 ca-

pabilities 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 beyond 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.[19]

Archival Organizing Systems implement a distinctive answer to the question of what is being organized. Archives are a type of collection that focuses on resources created by a particular person, organization, or institution, often during a particular time period. This means that archives have themselves been previously organized as a result of the processes that created and used them. The "original order" of the resources in an archive embodies the implicit or explicit Organizing System of the person or entity that created the documents and it is treated as an essential part of the meaning of the collection. As a result, the unit of organization for archival collections is the ***fonds***—the original arrangement or grouping, preserving any hierarchy of boxes, folders, envelopes, and individual documents—and thus they are not re-organized according to other (perhaps more systematic) classifications.[20]

Some Organizing Systems contain legal, business or scientific documents or data that are the digital descendants of paper reports or records of transactions or observations. These Organizing Systems might need to deal with legacy information that still exists in paper form or in electronic formats like image scans that are different from the structural digital format in which more recent information is likely to be preserved. When legacy conversions from printed information artifacts are complete or unnecessary, an Organizing System no longer deals with any of the traditional tangible artifacts. Digital libraries dispense with these artifacts, replacing them with the capability to print copies if needed. This enables libraries of digital documents or data collections to be vastly larger and more accessible across space and time than any library that stores tangible, physical items could ever be.

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.

1.3.3 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. "Bringing like things together" is an informal organizing principle for many Organizing Systems. But there will likely be a number of more precise requirements or constraints to satisfy.

Organizing Systems involving physical resources are more likely to emphasize aesthetic or emotional goals than those for information resources, which more often are dominated by functional goals like efficiency of storage and access. This contrast is often magnified by the tendency for major library and museum collections to be housed in buildings designed as architectural monuments that over time become symbols of national or cultural identity.

The fine distinctions between Organizing Systems that have many characteristics in common reflect subtle differences in the priority of their shared goals. For example, many Organizing Systems create collections and enable interactions with the goals of supporting scientific research, public education, and entertainment. We can contrast zoos, animal theme parks, and wild animal preserves in terms of the absolute and relative importance of these three goals with respect to animal resources.[21]

When individuals manage their papers, books, documents, record albums, compact discs, DVDs, and other information resources, their Organizing Systems 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 because their resources can be accessed from anywhere with a web browser.[22]

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, 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 might even be specified in commercial contracts or governed by law. 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. Not surprisingly, policymaking and regulations about patient records are highly contentious.[23]

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.[24] 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, the Organizing Systems in a large academic research library must also support many functions and services other than those that directly support search and location of resources in their collections. In these respects, the Organizing Systems in non-profit libraries have much in common with those in corporate information repositories and business applications. See the Sidebar, "Library {and, or, vs.} Business Organizing Systems."

Library {and, or, vs.} Business Organizing Systems

Any information-driven enterprise must have processes and technologies in place that govern information creation or capture and then manage its entire life cycle. In addition to the Organizing Systems that manage and provide access to their collections, large libraries also need business Organizing Systems to support acquisition, billing, interlibrary loan record routing and systems, licenses of digital resources from publishers, course material websites, and the library's own web presence. Commercial firms need processes for transacting with customers or other firms to carry out business

operations, to support research and innovation, 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), 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. 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.

Chapter 7, "*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.

1.3.4 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 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. Chapter 4

and Chapter 6, 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 §1.3.3, "Why is it Being Organized?" (page 25): Organizing Systems exist for different purposes and to support different kinds of interactions or functions. Let's 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. 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 will 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 organiza-

tion. This contrast has analogues in other Organizing Systems and we can easily imagine what happens if a “neat freak” and “slob” become roommates.[25]

Different preferences and disagreements between stakeholders in an Organizing System about how much organization is necessary often result because of the implications for who does the work and who gets 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.[26]

The cost-effectiveness of creating systematic and comprehensive descriptions of the resources in an information collection has been debated for nearly two centuries and in the last half century the scope of the debate grew to consider the role of computer-generated resource descriptions.[27]

An alternative and complement to man-made descriptions for each resource are computer-generated indexes of their textual contents. 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 or thesauri 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.

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 re-

source, unless there are so many users of the collection that their aggregated effort is comparably large; this is organizing by "crowdsourcing" (See the Sidebar on "Web 2.0" in §1.3.6). 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, which is especially sensible for archives with its emphasis on the *fonds* (see §1.3.2, "What is Being Organized?" (page 23)). 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 §7.6, "Computational Classification" (page 310) describes other common techniques in machine learning, clustering, and discriminant 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 century 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*.

But in recent decades a new set of information and computing technologies enabled by Moore's Law — unlimited computer 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 outcompete 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 that could apply to both.[28]

Chapter 8, "*The Forms of Resource Descriptions*", focuses on the representation of resource descriptions, taking a more technological or implementation per-

spective. Chapter 9, "*Interactions with Resources*", 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.

1.3.5 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 their significance or relationship to each other. The document could also have some external organization implied by the context of its publication, like the name of its author and publisher, its web address if it is online or has a website, 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 like time and location are assigned automatically to these types of resources by the technology used to create them.[29]

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 types of document resources that are 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 pa-

ges) 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. 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.[30] 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.[31]

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, but that only reminds us of how much the traditional distinction between "information organization" and "information retrieval" is no longer defensible.

In many Organizing Systems the nature and extent of organization changes over time as the resources governed by the Organizing System are used. The arrangement of resources in a kitchen or in an 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 web sites, 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 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. (See §2.5.4, "Governance" (page 77) for more discussion).

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 from entities with the authority to impose them.

1.3.6 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* (§1.2.3.1) 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.[32] 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 will help you downsize to a smaller living space.[33]

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, many authors will be 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. Not only are these ordinary users unlikely to use standard descriptors and classifications, the organization they impose sometimes so closely reflects their own perspective and goals that it is not useful or accurate for others. Fortunately most users of "Web 2.0" or "community content" applications at least partly recognize that in these applications 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.[34] 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 web sites. 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.

Web 2.0, Enterprise 2.0, Library 2.0, Museum 2.0, Science 2.0, Gov 2.0, ...

The Web was invented as a publishing and document distribution medium, and later became a platform for business transactions. But after the bursting of the "dot com bubble" in 2000-2001 it was clear that moving a transactional business model to the web was not enough. In 2005 Tim O'Reilly and Dale Dougherty proposed the concept of "Web 2.0" for firms whose applications literally get better the more people use them because they "harness the collective intelligence" of their users.

Google, Amazon.com, eBay, Wikipedia, Facebook, Twitter, and YouTube are familiar examples today of web-based applications and services where value is based on aggregating, interpreting, and responding to enormous amounts of user-generated data and content. Web sites and resources that attract many visitors collect user interactions implicitly and also allow users to annotate, "tag," and evaluate them explicitly. These bottom-up and distributed activities have been called "folksonomies" and "crowdsourcing."

Tagging, bookmarking, and rating mechanisms are increasingly being adapted for use inside companies as techniques for knowledge management, a trend named "Enterprise 2.0" by Andrew McAfee to emphasize its similarity with "Web 2.0" while pointing out how it differs. Because every user is authenticated to their real identities, and organizational norms and incentives restrict and shape the purposes and nature of user contributions, Enterprise 2.0 applications have been successful at capturing expertise and institutional knowledge.

The core Web 2.0 design principle of empowering users to contribute information to help organize some collection of resources is rapidly being generalized to many other domains of Organizing Systems. Some libraries are now discussing how a "Library 2.0" could provide personalized catalogs and information services and enable patrons to interact online with people of similar interests. Similarly, some museums, scientific repositories, and governments are conducting "open access" or "citizen participation" experiments by allowing users access to identify and annotate items, analyze raw data, or create "mashups" or applications that reuse and transform information that formerly was available only in summary form or in finished documents.[35]

1.4 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.

Many of the foundational topics for a discipline of organizing have traditionally been presented from the perspective of the public sector library and taught as "library and information science." These include bibliographic description, classification, naming, authority control, and information standards. We need to update and extend the coverage of these topics to include more private sector and non-bibliographic contexts, multi- and social media, and new information-intensive applications and service systems enabled by mobile, pervasive, and scientific computing. In so doing we can reframe the foundational concepts to make them equally compatible with the disciplinary perspectives of informatics, data and process modeling, and document engineering.

With the Web and ubiquitous digital information, along with effectively unlimited processing, storage and communication capability, millions of people create and browse web sites, blog, tag, tweet, and upload and download content of all media types without thinking "I'm organizing now" or "I'm 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.

Chapter 2, "*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 2 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 3, "*Resources in Organizing Systems*". The design of an Organizing System is strongly shaped by what is being organized, the first of the five design decisions we introduced earlier in §1.3.2, "What is Being Organized?" (page 23). 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 3 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 §1.3.3, "Why is it Being Organized?" (page 25).

Chapter 4, "*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 §1.3.4, "How Much is it Being Organized?" (page 27); Chapter 4 presents a systematic process for creating effective descriptions and analyzes how this general approach can be adapted for different types of Organizing Systems.

Chapter 5, "*Describing Relationships and Structures*". An important aspect of organizing a collection of resources is describing the relationships between them. Chapter 5 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 6, "*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 6 reviews theories of categorization from the point of view of how categories are created and used in Organizing Systems.

Chapter 7, "*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 7 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 8, "*The Forms of Resource Descriptions*". Chapter 8 complements the conceptual and methodological perspective on the creation of resource descriptions with an implementation perspective. Chapter 8 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 9, "*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 9 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, mashups, and so on. Similarly, processes for information retrieval are often characterized as comparing the description of a user's needs

with descriptions of the resources that might satisfy them. Chapter 9 extends and more broadly applies this core idea to describe IR and related applications of natural language processing (NLP) in terms of locating, comparing, and ranking descriptions.

Chapter 10, *"The Organizing System Roadmap".* Chapter 10 complements the descriptive perspective of Chapter 2—Chapter 9 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. This model is then used to guide the analysis of four case studies that span the range of Organizing Systems.

Notes

1. [Citation] (Nunberg 1996, 2011). (Buckland 1991). See also (Bates 2005).

2. [LIS] (Buckland 1997); (Glushko and McGrath 2005) and others with an informatics or computer science perspective take an abstract view of "document" that separates its content from its presentation or container (see §3.3.3, "Identity and Information Components" (page 112)). In contrast, the library science perspective often uses presentation or implementation properties in definitions of "document." On authorship: when we say that "Herman Melville is the author of *Moby Dick*" the meaning of "author" does not depend on whether we have a printed copy or a e-book 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, which creates interoperability problems (see Chapter 9).

3. [LIS] We can continue the debate in the previous paragraphs and the "What is Information?" (page 4) sidebar by pointing out that in both common and professional usage, "bibliographic" activities involve describing and organizing information resources of the kinds that might be found in a library. But noted information scientist Patrick Wilson argued for a much broader expanse of the bibliographic universe, suggesting that "it includes manuscripts as well as printed books, bills of lading and street signs as well as personal letters, inscriptions on stone as well as phonograph recordings of speeches, and most notably, memorized texts in human heads and texts stored up in the memories of machines" (Wilson 1968, p. 12).

4. [Citation] (Svenonius 2000).

5. [Computing] 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 representation 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 PowerPoint 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.

6. [Business] 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).

7. [Citation] (Banzhaf 2009).

8. [Computing] 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 environ-

ments that are intentionally isolated from the rest of the web (like Facebook or Apple's iTunes and other smart phone "app stores").

9. [Computing] 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.

10. [Computing] 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).

11. [LIS] For precise distinctions, see the US Department of Labor, Bureau of Labor Statistics occupational outlook handbooks at *http://www.bls.gov/oco/ocos065.htm* and *http://www.bls.gov/oco/ocos068.htm* and *http://www.michellemach.com/jobtitles/realjobs.html*

12. [LIS] The four objectives listed in this paragraph as those proposed in 1997 by the *International Federation of Library Associations and Institutions (IFLA)*. The first statement of the objectives for a bibliographic system was made by (Cutter 1876), which (Svenonius 2000) says it is likely the most cited text in the bibliographic literature. Cutter called his three objectives "finding," "co-locating," and "choice."

13. [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.

14. [Computing] Sometimes many of these Organizing Systems and their associated applications are implemented using a unified storage foundation provided by an enterprise content management (ECM) or enterprise data management (EDM) system. An integrated storage tier can improve the integrity and quality of the information but is invisible to users of the applications.

15. [Citation] IFLA Library Types (*http://www.ifla.org/library-types*)

16. [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 (See the Sidebar, "The Hathi Trust Digital Library" (page 71). 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.

17. [Citation] Depending on which characteristics of Google Books and libraries you think about, you might complete this analogy with an animal theme park like Sea World (*http://www.seaworld.com/*) or a private hunting reserve that creates personalized "big game" hunts. Or maybe you can invent something completely new.

18. [Citation] (Trant 2009a), (Gilliland-Swetland 2000).

19. [LIS] Organizing Systems that follow the rules set forth in the Functional Requirements for Bibliographic Records (FRBR) (Tillett 2005) treat all instances of *Macbeth* as the same "work." However, they also enforce a hierarchical set of distinctions for finer-grained organization. FRBR views books and movies as dif-

ferent “expressions,” different print editions as “manifestations,” and each distinct physical thing in a collection as an “item.” This Organizing System thus encodes the degree of intellectual equivalence while enabling separate identities where the physical form is important, which is often the case for scholars.

20. [LIS] Typical examples of archives might be national or government document collections or the specialized Julia Morgan archive at the University of California, Berkeley (*http://www.oac.cdlib.org/findaid/ark:/13030/tf7b69n9k9/*), which houses documents by the famous architect who designed many of the university’s most notable buildings as well as the famous Hearst Castle along the central California coast. The “original order” organizing principle of archival Organizing Systems was first defined by 19th century French archivists and is often described as “*respect pour les fonds*.”

21. [CogSci] But seeking absolute boundaries between types of Organizing Systems is an impossible quest because how we define them varies with context or point of view. Zoos, animal theme parks, and wild animal parks all contain live animals, so we might conclude that they are more similar to each other than to a natural history musuem in which the animals are all dead. Colonial Williamsburg (*http://www.colonialwilliamsburg.com*) has people re-enacting 18th century Virginia and describes itself as a “living history museum,” but could it not be considered an animal theme park that has human animals? Is a cemetery in some ways a natural history museum?

22. [Computing] 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,” “toread,” or “buythis” 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.

23. [Citation] (Bowker and Star 2000).

24. [Citation] (Casson 2002).

25. [Citation] (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.

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

27. [LIS] Sir Anthony Panizzi is most often associated with the origins of modern library cataloging. In 1841 (Panizzi 1841) published 91 cataloging rules for the British Library that defined authoritative forms for titles and author names, but the complexity of the rules and the resulting resource descriptions were widely criticized. For example, the famous author and historian Thomas Carlyle argued that a library catalog should be nothing more than a list of the names of the books in it. Standards for bibliographic description are essential if resources are to be shared between libraries. See (Denton 2007), (Anderson and Perez-Carballo 2001a, 2001b).

28. [Business] 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).

29. [Computing] At a minimum, these descriptions include the creation time and storage format for the resource, or chronologically by the auto-assigned file-name (*IMG00001.JPG*, *IMG00002.JPG*, etc.), but often are much more detailed. 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. Nevertheless, these descriptions are not always correct. For example, Microsoft Office applications extract the author name from any template associated with a document, presentation, or spreadsheet and then embed it in the new documents. And if you have not set the time correctly in your digital camera any timestamp it associates with a photo will be wrong.

30. [LIS] (Nunberg 2009) calls 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.

31. [LIS] (Byrne and Goddard 2010).

32. [LIS] This is an important distinction in library science education and library practice. Individual resources are described ("formal" cataloging) using "bibliographic languages" and their classification in the larger collection is done using

"subject languages" (Svenonius 2000, Ch. 4 and Ch. 8, respectively). These two practices are generally taught in different library school courses because they use different languages, methods and rules and are generally carried out by different people in the library. In other organizations, the resource description (both formal and subject) is created in the same step and by the same person.

33. [Citation] NAPO: *http://www.napo.net*

34. [Computing] (He et al. 2007) estimate that there are hundreds of thousands of web sites 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.

35. [Citation] The "manifesto" for Web 2.0 is Tim O'Reilly's *What is Web 2.0?* (*http://oreilly.com/web2/archive/what-is-web-20.html*).

"Folksonomy" was coined by Thomas Van der Wal at about the same time in 2004; see *http://vanderwal.net/folksonomy.html* and (Trant 2009b).

The term "Crowdsourcing" was invented by Jeff Howe in a June 2006 article in Wired magazine, *http://www.wired.com/wired/archive/14.06/crowds.html*, and the concept was developed further in a book published two years later (Howe 2008).

(Millen et al. 2005) describe an enterprise application of social bookmarking at *IBM* called Dogear. The Library 2.0 idea is presented in (Maness 2006) and several more recent surveys of Web 2.0 features in university library web sites have been reported by (Xu et al. 2009) and (Harinarayana and Raju 2010).

Nina Simon's book, *The Participatory Museum*, is itself an example of Web 2.0 concepts, available online with reader comments (*http://www.participatorymuseum.org/read/*). For Science 2.0., see (Shneiderman 2008). For Government 2.0, see (Robinson et al. 2008) and (Drapeau 2010).