Chapter 1. Foundations for “Organizing Systems”

To appear in R. Glushko & C. Borgman (Eds.),
Intellectual Foundations for
Information Organization and Information Retrieval, 2011

Robert J. Glushko
glushko@berkeley.edu

Contents

Chapter 1. Foundations for “Organizing Systems” ..................................................................................... 1

1.1 Information Organization {and, or, vs.} Information Retrieval ......................................................... 1
  1.1.1 The Natural Coupling of IO and IR ............................................................................................... 2
  1.1.2 The Divergence of IO and IR ........................................................................................................... 2
  1.1.3 The Re-Convergence of IO and IR..................................................................................................... 3
  1.1.4 From “Collection Categories” to “Organizing Systems” ............................................................. 4
  1.1.5 “Organizing Systems” and the IO / IR Relationship................................................................. 7

1.2 The Organizing System ............................................................................................................................ 8
  1.2.1 What is Being Organized? ............................................................................................................. 12
  1.2.2 Why is it Being Organized? ............................................................................................................ 14
  1.2.3 How Much is it Being Organized? ................................................................................................. 17
  1.2.4 When is it Being Organized? ........................................................................................................ 20
  1.2.5 How (or by Whom) is it Organized? ............................................................................................... 22

1.3 Organizing this Book ............................................................................................................................. 23

References ..................................................................................................................................................... 25

1.1 Information Organization {and, or, vs.} Information Retrieval

Devising concepts, methods, and technologies for describing and organizing information has been an essential human activity for millennia, evolving both in response to human needs and to enable new ones. Concepts, methods, and technologies for finding and using information inherently co-evolve with those for describing, recording, and organizing it. Taken together these activities supported the development of civilization, from agriculture and commerce to government and warfare.
1.1.1 The Natural Coupling of IO and IR

For most of human history, the natural coupling of information and the container or medium in which it was represented made it unnecessary to distinguish “information organization” (IO) and “information retrieval” (IR) as separate tasks or concerns. Technologies and techniques for recording and organizing information were invented and improved by every developing civilization (Kilgour, 1998). In Mesopotamia, the oldest known documents are records of inventories, business transactions, and tax receipts, encoded using Cuneiform writing on clay tablets. Under the Pharaohs in Egypt, information was carved in stone or written in ink on papyrus made from reeds growing in Nile River marshes. The ancient Greeks invented parchment using animal skins. Its superior flexibility and durability enabled books in scroll form and ultimately led to the abandonment of papyrus as a writing medium. Paper made from the pulp mixture of macerated plant fiber was invented in China around 200 BCE but paper took almost a thousand years to reach the west, brought by Muslim traders.

When organizing information involves creating and arranging tangible artifacts—carving in stone, imprinting in clay, applying ink to paper by hand or with a printing press—and then putting the artifacts into boxes, bins, cabinets, or shelves, the affordances of the medium (Gibson, 1977) and the implications for information retrieval and use are immediately evident. Consider the codex, invented about 100 AD by binding thin wood writing tablets together. This ancestor of the book was a radical innovation for storing and retrieving information because it was vastly more efficient and usable than all other existing formats. Search and access for information in a scroll is inherently serial, and there are no externally visible markers or organizational devices. Unlike scrolls or tablets, books with paper pages could be easily handled and arranged on shelves, a first step toward the familiar look of the modern library.

How much effort will it take to organize some body of information? How many different access points or methods for finding information will this organizing effort create? The tradeoffs between the amount of work that goes into organizing information and the amount of work required to find it are inescapable when information is fixed in a physical artifact and can’t easily be “unfixed” for reuse in some other artifact or for some other purpose.

1.1.2 The Divergence of IO and IR

In the mid-20th century, research and technological developments in mathematics and computer science led to the mainframe computer and made it possible to organize large amounts of digital information. “Organizing” and “retrieving” activities diverged and their interdependencies became less visible. Computing and storage technology enabled a scale of information management not possible when dealing with traditional, tangible information artifacts, separating the work of creating and organizing information from the work of finding it and using it. In just a few decades, a huge gulf opened up between the disciplines concerned with information organization and those concerned with information retrieval.

Library and information science became increasingly systematized as disciplines that focused on the description and organization of institutional information resources, especially those created by professional and academic authors and collected in public-sector libraries. The overriding goal was to develop and apply methods by which people could organize information to enable unknown future users to find it later, often with questions different from those for which the information was created. Computer science and informatics emerged as disciplines
that formalized the design and implementation of the systems that store and process information, often in business or scientific contexts. The driving goal of the sub-field of information retrieval was the development and implementation of the logical models and algorithms needed by computers to index, search, and assess the relevance of information resources. A focal problem in IR is dealing with scale, because the effectiveness of organizational and retrieval methods sufficient for small collections of things or information decline precipitously as the collection grows. Over time, the academic curricula and textbooks targeted to library science and those aimed for computer science and informatics diverged even when they nominally involved common concerns of conceptual description and the creation and use of information models.

1.1.3 The Re-Convergence of IO and IR

Today the academic disciplines that study and teach information organization and retrieval remain divided, but within activities and design problems “in the real world,” issues and mechanisms for organization and retrieval are increasingly intermixed, thus re-converging. IO and IR are now so intertwined as the conceptual and practical foundations of our 21st century culture and economy that it makes little sense to us to force an artificial boundary between them.

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.” This digital information can be distributed, processed, reused, transformed, mixed, remixed and recombined into different formats for different purposes, applications, devices, or users in ways that are almost impossible to imagine when it is represented in a tangible medium like a book on a shelf or a box full of paper files. 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. Anyone can publish, annotate, “tag,” and evaluate information, bypassing traditional intermediaries between authors and readers. The massive scale of these bottom-up and distributed activities makes information organization emerge by “crowdsourcing” from the continuously aggregated actions and contributions of users (Howe, 2008). These informal and organic “folksonomies” (Trant, 2009b) create organization and authority through network effects, undermining traditional centralized mechanisms of organization and governance and threatening any business model in publishing, education, and entertainment that has relied on top-down control and organization of information.

Rather than fight these trends, many companies, universities, and groups of various kinds are adopting “wikis” modeled after the wildly popular Wikipedia, the “free encyclopedia that anyone can edit.” In a related vein, many governments as well as some progressive information providers and scientific researchers have begun to encourage the reuse and reorganization of their content by making geospatial, demographic, environmental, economic, and other datasets available in open formats or as web services rather than as “fixed” publications (Bizer, 2009; Robinson et al, 2009).

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, like news feeds, traffic congestion map overlays, blogs and tweets. 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. Indeed, search engines are being driven to improve the timeliness and extent of information personalization because people are increasingly querying their social networks instead of search engines because they think the former will have more current and personally relevant answers.

Similarly, many business models and processes in retailing, supply chain management, healthcare, energy management, customer relationship management, and other information-intensive domains that once relied on analysis of historical information about transactions and customers increasingly benefit by collecting, analyzing, and acting on information in close to real-time. Coupons printed on supermarket checkout receipts are uniquely customized by business intelligence applications that use the billions of transactions tied to loyalty cards. When a customer logs onto Amazon.com or similar online retail site, the generic catalog seen by anonymous customers is dynamically replaced with a personalized one whose content and layout reflects shopping history, previous search queries, abandoned shopping carts, and even real-time browsing behavior (Shafer, Konstan, and Reidl, 2001). The slogan on the ubiquitous UPS brown delivery trucks doesn’t say “We Deliver”; it says “Synchronizing the World of Commerce,” emphasizing the business value UPS creates by always knowing the locations of packages. This information is essential to efficient logistics and meeting customer expectations in online retailing.

1.1.4 From “Collection Categories” to “Organizing Systems”

There is a rich set of categories for describing collections of information, things, or information about things. The traditional categories of public sector or nonprofit “memory institutions” – libraries, museums, and archives – have been evolving for thousands of years. We make distinctions between public libraries, university libraries, software libraries, presidential libraries, tool libraries, and countless other variants. We classify collections of art and other physical artifacts differently than those of animals and fish when we call them art museums, natural history museums, zoos, and aquaria. Business enterprises have always retained business information and technical data, and today they use different software applications for records management, content management, knowledge management, customer relationship management, data warehousing and business intelligence, e-mail archiving, and other subcategories of collections. Individuals organize their personal information, often in idiosyncratic ways and often using multiple devices or applications that impose different systems of organization (Karger and Jones, 2006; Marshall 2008).

Some of the proliferation of collection categories reflects technological innovation and especially digitization and “webification.” Music can be collected and organized on vinyl records, tapes in various formats, compact discs, and digital MP3 files. Photographs are hung on walls, arranged in photo albums, and, increasingly, collected as digital objects. Books can be collected in traditional print forms or downloadable digital books. These books then need to be further organized according to which e-book reader, tablet computer, smart phone or other device on which they can be read. The encyclopedia category, which once implied a collection of articles organized into handsomely printed volumes, now includes Wikipedia, a peer-produced
online encyclopedia that can be edited by anyone. Furthermore, the wiki genre itself has been subdivided into a taxonomy that includes single-contributor wikis, group or project wikis used as team collaboration tools, and “pedias”, enterprise-wide Wikipedia-like knowledge repositories (Poole and Grudin, 2010).

But as the Web and digital technologies blur the boundaries between IO and IR, we are seeing IO and IR coming together in ways that don’t fit existing categories and instead drive the development of new ones. Consider photo-sharing website Flickr, which functions for most of its users as a 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 library. Flickr lacks the authoritative description and standard classification that define traditional libraries. But Flickr offers application programming interfaces (APIs) and data feeds that enable clever software developers to invent new IO and IR services that surpass those of traditional libraries. Many of these involve searching for photos taken at specified places and times with the results integrated into maps and timelines; see the Flickr “App Garden” for many more examples (Flickr, 2011).

Two decades ago, when publishers first adopted digital submission of articles, CD-ROM and online access and storage, many aspects of scholarly research and communication were greatly improved even though the scientific article and journal per se changed relatively little. But more recently “web-native” publication models have emerged that embrace more modular and heavily hyperlinked architectures, creating networks of articles that blur the traditional notions of articles and journals and that interconnect scientific findings with their underlying research data or to discipline-specific data repositories (Aalbersberg and Kahler, 2011).

The Amazon Kindle and other e-book readers let users annotate or highlight passages they find particularly meaningful as a way of personalizing the books in their digital libraries. This is something people often do with books they own, but marking up a borrowed book violates library norms. But the Kindle also sends these individually highlighted passages back to Amazon servers to determine the “popular highlights,” which Kindle then displays as underlined passages in Kindle books, even for readers who don’t use the highlighting feature themselves (Amazon, 2011). While many readers may find this helpful, others concerned with privacy point out that libraries do not track what passages people read. For authors and publishers, this feature could be viewed as a key part of a market research or business intelligence information system.

A characteristic feature of museums is that they provide highly mediated access to the objects in their collections, whose descriptions and classification are created by curators or professional catalogers. How then should we classify museums and scientific projects that enlist amateur volunteers to describe and classify objects, as herbariaunited.org does for botanical specimens in British museum and university collections and as galaxyzoo.org does for astronomical images (Wright, 2010)?

Google has characterized its ambitious project to put tens of millions of books from research libraries online as “a library to last forever” (Brin, 2009). But the Google Books project has been widely criticized as not being true to library principles. First, Google classifies books using the bookstore-oriented BISAC categories rather than using the Library of Congress or Dewey Decimal systems used by academic or public libraries (Nunberg, 2009). Second, Google collects detailed records of user behavior, which conflicts with the cherished library goal to enable anonymous reading (Jones and Janes, 2010). Finally, Google uses these search records to serve targeted ads against book search results, just as it does for regular searches of ordinary web content. According to legal scholar Pamela Samuelson, “Anyone aspiring to create a modern
equivalent of the Alexandrian library would not have designed it to transform research libraries into shopping malls” (Samuelson, 2009).

Flickr, Google, Amazon, and any other applications that users access “in the cloud” as a hosted web service rather than as locally-installed software create value by aggregating and analyzing implicit and explicit IO and IR activities. “Library” and “museum” are inappropriate and inadequate descriptions for the hybrid types of collections and associated services enabled as a result. We can stretch the traditional definitions and add modifiers, as when we describe Flickr as a web-based photo-sharing library, but in doing so we suggest features that aren’t there (like authoritative classification) and omit the features that are distinctive (like tagging by users). 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 inflames many people with a more traditional sense of what the library category implies. To them, Google Books is simply another way for Google to make money.

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. The common challenges of “going digital,” and the architectural/functional constraints imposed by web implementations are causing some convergence in the operation of libraries, museums, and archives (Trant, 2009a). Giving every physical object 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” (Gilliland-Swetland, 2000).

Taken together, these two trends have one profound implication. As the traditional categories for thinking about collections converge and become ambiguous, they are no longer capable of describing innovative collections and their associated services. Thus, we need a new approach, one that should:

- Apply the traditional categories where they are appropriate, but not impose them on new types of collections and services where they just don’t fit
- Apply comprehensively and consistently to collections of information, collections of things, collections of information about things, and any combination
- Explain and exploit the re-convergence of IO and IR.

We propose the idea of an Organizing System to complement the categorical view of collection types. Instead of treating collection types as mutually exclusive categories, we can describe collections as configurations of choices in a complex multidimensional space of organization design decisions. These decisions concern what is organized, why it is organized, how much it is organized, when it is organized, and how or by whom it is organized. An Organizing System embodies the composite impact of the choices made on these design dimensions. We can describe the familiar collection categories like libraries, museums, and archives as design patterns that embody characteristic configurations of design choices. But at the same time, a dimensional perspective acknowledges the diversity of instances of these collection types and provides a generative, forward-looking framework for inventing and describing new types of collections and their associated services. In addition, the increased nuance of multidimensional description makes it easier to trace the connections between specific requirements or constraints and particular aspects or functions in the Organizing System.
For example, an institution that lends items in its collection with the hope that the borrowers return something else hardly seems like a library if we treat “borrow and return the same item” as a defining feature of a library. But if the items 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 (Wang, 2010). Even though it differs from traditional libraries on some dimensions, its similarities on other dimensions of its Organizing System give a Seed Library more “family resemblance” to a traditional library than to any other type of collection. And because the dimensional perspective doesn’t impose sharp boundaries between categories, we can more readily notice that a Seed Library shares with museums the goal of preserving unique or rare items and with zoos the characteristic that the items in the collection are living species.

1.1.5 “Organizing Systems” and the IO / IR Relationship

The Organizing System framework captures and provides structure for the inherent tradeoffs obscured by the silos of traditional disciplinary and category perspectives:

- When we organize information to enable retrieval, the more effort put into organizing information, the more effectively it can be retrieved.
- The more effort put into retrieving information, the less it needs to be organized first.

An essential principle in the Organizing System framework is treating the work done by people and the work done by computers as having common goals, despite obvious differences in methods. Instead of a view that treats IO as a human activity and IR as a machine one, or IO as a topic for library and information science and IR as one for computer science, we can recognize that machines now assist people in IO and that people contribute much of the information used by the machines to enable IR. IO works hand-in-hand with IR, just as information created and analyzed by people works hand-in-hand with information created and analyzed by machines. We can look at different Organizing Systems to understand how human and computational efforts complement and substitute for each other and to determine the economic, social, and technological contexts in which each can best be employed. And we can determine how the Organizing System allocates effort and costs among its creators, users, maintainers and other stakeholders.

Another important concept that unifies IO and IR in the Organizing System framework is Description. Many of the design dimensions concern the nature and extent of the descriptions of the things being organized. These descriptions are later compared and combined when we search for, retrieve, and transform information. This is not a new idea; the process of information retrieval is often characterized as comparing the description of a user’s information need with descriptions of the information resources that might satisfy them. But the Organizing System framework highlights the factors that determine the nature and extent of these descriptions and how they determine the capabilities of the processes that locate, compare, combine, or otherwise use them. The Organizing System perspective, which views the traditional concerns of information organization and information retrieval in a more interconnected, systematic, and generative way, can readily adapt to new applications and technologies as they arise.
1.2 The Organizing System

We organize things. The local public library organizes printed books, periodicals, maps, CDs, DVDs, and maybe some old record albums. The large university research library also organizes dissertations, rare manuscripts, pamphlets, computer programs, government and scientific datasets, databases, musical scores, and every other kind of published information. Museums organize artifacts of cultural, historical, or scientific value. Stores and suppliers organize their goods for sale to consumers and to each other. And each of us organizes the many kinds of things in our lives—our books in our homes; financial records in folders and filing cabinets; clothes in dressers and closets; cooking and eating utensils in kitchen drawers and cabinets.

We organize information about things. Catalogs 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. And each of us organizes information about things, when we sell our unwanted stuff on eBay, when we tag our photos on Flickr or Facebook, or when we rate a restaurant on Yelp.

We organize 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. Web search engines use content and link analysis along with relevance ratings to discriminate among the billions of web pages competing for our attention. And each of us organizes digital information—email, documents, media, appointments, and contacts—on our personal computers or smart phones, or in “the cloud” through information services that use Internet protocols.

When we analyze these different contexts—organizing things, organizing information about things, and organizing digital information—we can be easily distracted by the specific information types, organizing principles, technology, functions or features, individuals or companies involved in any particular example. We can get lost trying to define “information” in ways that fit these different contexts because information is an abstract concept, and most of its hundreds of definitions treat it as an idea that swirls around equally hard to define terms like “data,” “knowledge,” and “communication” (Buckland, 1991; Bates, 2005). 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. Analogously, when analyzing “What is a Document?” Buckland rejects narrow definitions that emphasize traditional physical forms. Thus an antelope is “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 (Buckland, 1997).

These debates about what “things” and “information” have in common and how they contrast are essential if we are to bridge the intellectual gulf that separates the many disciplines that share the goal of organizing but that differ in what they organize. An abstract view of information as an intangible thing is the intellectual foundation for both modern information science and the information economy and society. But the abstract view of information must coexist and often conflict with a much older idea that information is a tangible thing. Five thousand years of human experience with tangible information artifacts has deeply embedded the notion of information as a physical thing in culture, language, and methods of information design and organization, and it always will, so long as we humans inhabit a physical world that contains
people and objects that we name, classify, and organize. And even intangible information must be stored and processed by tangible machines, whether massive racks of servers or a pocket-sized smartphone.

As a result, our concept of Organizing System has to confront head on the duality of information as an intangible thing versus information as a tangible thing that is more like the physical objects we organize on shelves and in cabinets and closets. When an Organizing System deals with tangible things, it follows different principles and must conform to different constraints than when it deals with intangible ones – the most obvious one being that 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 for their tangible things, thus accommodate both notions of information at the same time. The implications for arranging, finding, using and reusing things in any Organizing System directly reflect the mix of these two embodiments of information.

Explicitly or by default, an Organizing System makes many interdependent decisions about the identities of entities and information components; their names, descriptions and other properties; the classes, relations, structures and collections in which they participate; and the people or technologies who create, transform, combine, compare and use 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 by recognizing that “what is being organized” is just one of the questions or dimensions we need to analyze.

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 mixture of things, information about things, and information in the Organizing System? Is the information in the collection narrative or transactional? Do all non-digital things have digital proxies or surrogates? What is the scope and scale of the domain? Are the things unique? Do they change? Do they follow a predictable “life cycle” with a “useful life”—or are they living things? Is the collection stable, or are items continually added or deleted?

- **Why is it being organized?** What functions or capabilities will be enabled, and for whom? Are the uses and users known or unknown? Are the users primarily people or computational processes? How are users given access to items in the Organizing System? Can the items be viewed, borrowed, “checked out,” or otherwise accessed or used outside of the collection? Does the Organizing System need to satisfy personal, social, or institutional goals?

- **How much is it being organized?** What is the extent 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 principles guide and are embodied in the organization? Do the descriptions and classifications conform to standards or are they ad hoc or proprietary?

- **When is it being organized?** Is the organization imposed at design time, or at runtime, just in case, just in time, all the time? Is any of this organizing, governance or curation activity mandated by law or shaped by industry or business practices?

- **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 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.

Because any of these factors can change, Organizing Systems change in response to them. In a large library or museum, items are continually added or edited, but the Organizing System changes only incrementally, perhaps by adding or deleting items from descriptive or classification vocabularies. In contrast, in Organizing Systems used in business or in technology-intensive domains, new competitors or technologies can have a disruptive impact, forcing major changes in description and classification to enable competitive positioning, refined customer segmentation, or faster finding, retrieval, or delivery of informational or physical goods. For example, the Association for Computing Machinery professional society created a classification system (ACM, 2011) in 1964 to organize articles in its many publications, but relentless change in the computing field has required the ACM to significantly revise the system almost every decade. Even the society’s name seems outdated.

New laws or regulations can disrupt Organizing Systems by mandating specific changes in information organization or processing that must take place at a specified time. Examples in the United States include the Health Insurance Portability and Accountability Act (HIPAA), which specifies controls on personally identifiable health information, and the Sarbanes-Oxley Act, passed in the wake of accounting scandals to require more control and transparency. Both of these policies imposed substantial compliance obligations.

The Organizing Systems created by individuals change, too. The set of clothes in a college student’s closet and the categories and principles for arranging them will change somewhat if he graduates and takes a job in a downtown office building. He might even need an entirely new wardrobe if the job is in a new city with a different climate.

The same domain or body of information can have more than one Organizing System, and one Organizing System can contain multiple others. The Organizing System for books in a library arranges books about cooking according to the Library of Congress or Dewey Decimal classifications and bookstores use the BISAC ones, mostly using cuisine as the primary factor. In turn, cookbooks employ an Organizing System for their recipes that arranges them by type of dish, main ingredient, or method of preparation. Within a cookbook, recipes might follow an Organizing System that standardizes the order of their component parts like the description, ingredients, and preparation steps.

Sometimes these multiple Organizing Systems are designed in coordination so they can function as a single hierarchical, or nested, Organizing System in which it is possible to emphasize different levels depending on the user’s task or application. Most books and many documents have an internal structure with chapters and hierarchical headings that enable readers to understand smaller units of content in the context of larger ones. Similarly, a collection of songs can be treated as an album and organized using that level of abstraction for the item, but
each of those songs can also be treated as the unit of organization, especially when they are embodied in separate digital files.

Unfortunately, much of the time that multiple Organizing Systems exist in the same domain, they don’t have the same answers for the what/why/how much/when/how questions, resulting in inefficiency or incompatibility. Many Organizing Systems that use the web as an implementation platform don’t work well together because of conflicting choices platform designers make about formats for describing content. The most significant decisions that web-based Organizing Systems make about content description involve the tradeoff between simplicity and expressive power. The web rapidly became ubiquitous in no small part because HTML, the markup language used to describe the structure and appearance of web pages, was initially so simple that almost anyone could create a web page, and most web content is still represented in some version of HTML. But using HTML tags like “<li>” to identify a list item or “<h2>” to render some text as a second-level heading creates a “web for eyes only” whose content can’t be processed by business systems or other computer applications. A web page might use a text label to indicate that the content tagged as “list item” is a price, and “<li>PRICE: $24.99</li>” provides additional information for human readers. But turning the “web for eyes” into a “web for computers” requires computer-interpretable content. The most common way to make content computable is to use content-oriented XML tags like “<price>” or “<quantity>” whose arrangement and possible values can be validated against document type models and code lists expressed using XML schema languages.

Even when a web-based Organizing System uses content-based encoding of information, several other important design decisions remain. Should the Organizing System use domain-specific XML vocabularies like those defined by the OASIS standards body, or generic vocabularies for semantic description like RDF or RDF-a? Is it necessary for transparency or traceability to follow rigorous and formal modeling methods to create the descriptive vocabulary? Or can informal bottom-up approaches be used that assume that useful semantic agreement can emerge as “microformats” that systematize the description of small “semantic components” like “address,” “contact,” “calendar event,” and “catalog item” (Allsopp, 2007)?

And in a non-web and non-business context, consider that roommates or spouses sometimes argue about the “right” way to organize items in the kitchen, in the refrigerator, or in some other shared space. To a person who arranges spices alphabetically and condiment jars by size, arranging them according to cuisine or frequency of use makes no sense. Thus, Organizing Systems are highly subjective, designed on a use-and-user basis.

Organizing Systems overlap and intersect. People and enterprises routinely interact with many different Organizing Systems because what they do requires them to use information and things in ways that cut across context, device, or application boundaries. Just consider how many different Organizing Systems we use for managing personal information like contacts, appointments, and messages. Now consider this at an institutional scale in the interactions among the Organizing Systems of physicians, hospitals, medical labs, insurance companies, government agencies, and other parties involved in healthcare. Finally, think about Web “mashups” and composite services. We have come to expect that the boundaries between Organizing Systems are often arbitrary and that we should be able to merge or combine them when that would create additional value. The variety and complexity of Organizing Systems would be overwhelming without a framework like the one we’re proposing to understand them.

Our concept of the Organizing System was in part inspired by and generalizes to non-bibliographic domains the concepts proposed by Elaine Svenonius in “The Intellectual
Foundation of Information Organization” (Svenonius, 2000). She recognized that the traditional information organization activities of bibliographic description and cataloging were complemented, and partly compensated for, by automated text processing and indexing that were usually treated as part of a separate discipline of information retrieval. Svenonius—who strongly identifies with a LIS perspective—unequivocally expresses the tradeoff principle in the Organizing 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 following sub-sections with a quote from her book.

1.2.1 What is Being Organized?

“What is difficult to identify is difficult to describe and therefore difficult to organize” (Svenonius, p. 13).

Before we can begin to organize anything we need to identify it. It might seem straightforward to devise an Organizing System around tangible things, but we must be careful not to beg the question of what a thing 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 a sub-part 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 sub-parts of an item identified as a chess set (including the board).

When merchants assign stock-keeping units to identify the things they sell, a 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. Different merchants or firms might make different decisions about abstraction and granularity when they assign SKUs because of differences in suppliers, targeted customers, or marketing strategies. Similarly, 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 breaks the description into “TITLE”, “FIRSTNAME,” and “LASTNAME.”

For most types of things, any number of properties or characteristics could be used in an Organizing System. You might arrange your shirts in your clothes closet by style (long sleeve, short sleeve, pullover), by color, by fabric, or by formality (party and Hawaiian shirts are separated from those you would wear to the office). For any choice of arrangement, you make decisions about what properties are significant and how abstractly you think about those properties. This depends on the number of things being organized and the purposes for which things are being organized. If you only have one cooking pan, you’ll probably decide where to store it in your kitchen differently than if you have many different kinds of pans that are designed for different kinds of cooking.

You probably don’t have 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 things grows, it becomes necessary to explicitly identify each thing; to create surrogates and descriptions like bibliographic records or metadata that distinguish one thing from another; and to create additional organizational mechanisms like store directories, library card catalogs and indexes that facilitate understanding the collection and locating the things.

Organization mechanisms like store directories and library card catalogs are embedded in the same physical environment as the things being organized. But when these mechanisms or surrogates are computerized, the new capabilities that they enable create design challenges. This
is because such an Organizing System can be designed and operated according to different principles than the Organizing System for the physical things. A single physical thing 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 efficiency and scale impossible for tangible, physical things.

When the things being organized contain or 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 version of Macbeth seem to be the same thing, and to some degree she has a point: Organizing Systems that follow the rules set forth in the Functional Requirements for Bibliographic Records (Tillett, 2004) treat all the Macbeths as the same “work.” However, they also enforce an hierarchical set of distinctions for finer-grained organization. FRBR views books and movies as different “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.

Preserving documents in their physical or original form is the primary purpose of many Organizing Systems that contain culturally, historically, or economically significant documents that have value as long-term evidence. Archives are a type of collection that focus on items created by a particular person, organization, or institution, often during a particular time period. Typical examples of archives might be national or government document collections or the specialized Julia Morgan archive at the University of California, Berkeley (Online Archive of California, 2011), 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.

Archival Organizing Systems implement a distinctive answer to the question of what is being organized. Archives are sets of documents that have themselves been previously organized as a result of the processes that created and used them. This "original order" 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—and thus they are not re-organized according to other (perhaps more systematic) classifications. This fundamental principle of archival Organizing Systems was first defined by 19th century French archivists and is often identified as “respect pour les fonds.”

The systematic appraisal and curation processes to identify and preserve documents that are characteristic of archives are also essential in 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.
Thus, 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 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 information that has digital origins. 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 contrast, the digital representations of music, photographs, videos, and other non-text content like sensor data are structurally and semantically opaque. As a result, they are generally organized and retrieved using text surrogates or with descriptions extracted by computational processing of the content.

Chapter 4, “Identity,” discusses the challenges and methods for identifying the things in an Organizing System in great detail.

1.2.2 Why is it Being Organized?

“The central purpose of systems for organizing information [is] bringing like things together and differentiating among them” (Svenonius p. xi).

Almost by definition, the essential function of any Organizing System is to enable its users to arrange things so they can be found later. How this general purpose of “bringing like things together” is achieved depends on the types of things or domains being organized, and in the personal, social, or institutional setting in which organization takes place. Chapter 8, “Classification,” more fully explains the different purposes for Organizing Systems, the organizing principles they embody, and the methods for assigning items to classifications. Chapter 3, “Domains and Design Patterns for Organizing Systems,” explains how these purposes and techniques apply in libraries, museums and archives—the traditional domains of library and information science—and extends the analysis to commercial applications of content, knowledge, and data management.

“Bringing like things together” is an informal specification of the goals for Organizing Systems. But because there will likely be a number of more precise requirements or constraints to satisfy, it is only a first approximation even for those created by individuals with small collections of things being organized. For example, the way you organize your home kitchen is influenced by the physical layout of counters, cabinets, and drawers; the dishes you cook most often; your skills as a cook, which may influence the number of cookbooks, specialized appliances and tools you own and how you use them; the sizes and shapes of the packages in the pantry and refrigerator; and even your height. But if other people also share the kitchen, you might have to negotiate and compromise on some of the decisions about what goes where because all users’ requirements or preferences differ. The more people, companies, or stakeholders are involved in an Organizing System, the more likely this decision making will be explicit, contractual, or even governed by law.
When individuals manage their papers, books, documents, record albums, compact discs, DVDs, and other information sources or information artifacts, their Organizing Systems vary greatly. This is in part because the content of the things being organized becomes a consideration. Put another way, an information object has inherently more potential uses than objects like forks or frying pans, so it isn’t surprising that the Organizing Systems in offices are even more diverse than those in kitchens. A classic study of personal information management in offices contrasted the strategies of “filers” and “pilers” (Malone, 1983). The Organizing Systems of filers require significant effort to define and maintain a systematic classification, and their offices and desktops are clean and free of clutter. In contrast, pilers don’t invest in explicit classification and instead put items in piles that are occasionally task or project-oriented, so their offices are messy. Pilers rely on spatial memory to find documents and use the document on the top of each pile as an informal label of the pile’s contents.

Kirsh (2000) has studied many different kinds of “information-intensive” work environments and has contributed to a theory and methodology that is called “information ethnography” or “distributed cognition” for understanding how and why people think about and value information in different ways (Hollan, Hutchins, and Kirsh, 2000). Kirsh analyzes work environments and strategies for information management and use in terms of the cognitive “cost structure” or “computational complexity” that their Organizing Systems impose. For example, this kind of analysis of “filing” and “piling” shows that filing is an objectively good strategy with predictably large costs because work is required “just in case” it will be necessary. In contrast, piling is objectively suboptimal but imposes only small and incremental costs because it organizes information “just in time” for it to be used. Kirsh’s approach also highlights the ways in which Organizing Systems can be made more effective with small amounts of information added to the environment, like Post-it notes or the markings on the cups in Starbucks coffee shops.

Many of the Organizing Systems used by individuals are implemented by web applications, and this makes them more useful because they can be accessed from anywhere with a web browser. For example, many people manage their digital photos with Flickr, their home libraries with LibraryThing, 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 “myfamily,” “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. In Chapter 8 on “Classification,” section 8.4 discusses this “Social/Distributed Classification” in more detail.

“Tag convergence” in web applications like Flickr or del.icio.us is a simple case of the negotiation of meaning that takes place in the interactions and conflicts between Organizing Systems in social or institutional settings. In “Sorting Things Out: Classification and its Consequences,” Bowker and Star (1999) present case studies in which differences in perspectives, roles, status or power can shape and in turn be shaped by Organizing Systems. For example, in studies of medical diagnoses and hospital procedures, they showed that classifications of procedures were biased by who performed them, with “nurse” work being viewed as requiring less expertise than “physician” work (and being reimbursed at lower rates by insurance companies) even when they objectively were the same tasks.
Bowker, Star, and other researchers in the “sociotechnical systems” tradition caution us against the uncritical assumption that Organizing Systems embody fair and objective solutions to problems of information organization, management, and use. Even when they are developed and promoted through institutional processes by governments, standards organizations, industry trade groups, and other authoritative bodies, Organizing Systems reflect the biases of their creators, and might be designed to achieve or subvert goals like inclusion, transparency, or political correctness even though their stated purposes might be otherwise.

When the scale of the collection or the number of intended users increases, it becomes more important to be explicit about the goals of an Organizing System and in the description of the mechanisms or processes by which these goals are met. We can look back to the invention of mechanized printing in the seventeenth century, which radically increased the number of books and periodicals, as the motivation for libraries to develop systematic methods for cataloging and classifying what they owned and to view themselves as doing more than just preserving a collection. Libraries began progressively more refined efforts to state the functional requirements for their Organizing Systems and to be explicit about how they met those requirements.

For example, in 1841 Panizzi argued that a simple list of the items in the British Library was inadequate, and he published 91 cataloging rules that defined authoritative forms for titles and author names (Panizzi, 1841). In 1876, Cutter proposed a set of rules for bibliographic descriptions three times larger than Panizzi’s, expanding their scope to cover cross references between entries so that related items could be found—i.e., to confront the question of “What is a Macbeth?” that we raised in the previous section. Cutter justified this extra cataloging effort as necessary to satisfy a library user’s “finding,” “collocating,” and “choice” objectives, arguing that “the convenience of the public is always to be set before the ease of the cataloguer” (Cutter, 1876). Cutter’s framework was widely adopted and, in slightly revised form, became the basis of international cataloging standards. The most recent statement of the goals of a library catalog is the Functional Requirements for Bibliographic Records (FRBR), released in 1998.

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 items in their collections. These include billing systems, interlibrary loan record routing and systems, course reserves, licenses of digital resources from publishers, course material websites, and the library’s own web presence. In these respects, the Organizing Systems in non-profit libraries have much in common with those in corporate information repositories and business applications.

Any firm with an information-driven business model must have processes and technologies in place that govern information creation or capture and then manage its entire life cycle. These processes are diverse and complex: supporting transactions 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, Records Management, Customer Relationship Management (CRM), Business Intelligence (BI), Knowledge Management, and so on.
1.2.3 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, p. 24).

We’ve broadened the notion of an Organizing System to include more than just documents, but it remains a general truth that not all things should be accorded the same degree of organization. In this section we will briefly unpack this notion of degree of organization into its two more important and related dimensions: the amount of description assigned to each item and the amount of organization of items into classes or categories. Chapter 5, “Instances,” Chapter 6, “Metadata,” and Chapter 7, “Classes and Types,” more thoroughly address these questions about the nature and extent of description in Organizing Systems.

Not all items in a collection require the same degree of description for the simple reason we discussed in Section 1.2.2: Organizing Systems exist for different purposes and to support different tasks or functions. Let’s contrast two ends of the “degree of description” continuum. Many people use “current awareness” or “news feed” applications that simply select items 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 items that use synonyms of the keyword; that are written in languages different from that of the keyword; or that are otherwise relevant but don’t contain the exact keyword in the limited part of the document that is scanned. However, users with current awareness goals don’t need to see every news story about some event, and this limited amount of description for each item and the simple method of comparing descriptions are sufficient. In addition, tasks like current awareness can’t employ any organizing technique that takes any significant time, especially in domains where “time is money.”

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, 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 cataloguers, and probably using terms from a controlled vocabulary to enforce consistency in what descriptions mean.

When Panizzi published his 91 cataloging rules for the British Library that defined authoritative forms for titles and author names, he ignited a debate about the cost-effectiveness of creating systematic and comprehensive descriptions of the items in an information collection that continues to the present day (Anderson and Perez-Carballo 2001a, 2001b). The effort to comply with standards for bibliographic description is essential if items are to be shared between libraries.

An alternative and complement to man-made descriptions for each item 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 capture more precisely what the documents are about. Naturally, these more sophisticated descriptions of the documents in the collection allow for more sophisticated 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.

Another reason why all of the items within the scope of an Organizing System should not necessarily have the same degree of description is that different types of items aren’t equally describable. Narrative or publication-style document types are heterogeneous in structure and content. They typically have titles and authors that are important elements in their description, and their hierarchical structure of chapters and sections suggests units to be described. If information needs to be reused in multiple document types, different output formats, or translated into several languages, a “single-source” publishing strategy can obtain significant benefits from fine-grained semantic specification of paragraph or section-level content units even though this approach imposes substantial costs on authors or production personnel (Rockley, 2002). But it is more common that the content of narrative documents is just text, and Organizing Systems rely on document-level metadata. This is especially evident when there are limits to the amount of “organizing intelligence” that can be applied by people to create content descriptions. In contrast, transactional document types—e.g., catalogs, resumes, orders, invoices, tax forms and the data sets that populate them—are more homogeneous in structure and content, primarily because they are often created by automated processes or by people as inputs to automated processes. These types of documents are inherently easier to describe and indeed have often been called “self-describing” when they are encoded using XML or other syntaxes that explicitly “mark up” the content as “item,” “customer,” “price,” “date” and so on. The fine-grained and regular component boundaries in transactional documents enable them to be organized automatically in content repositories and databases once some “organizing intelligence” has been applied to define the document or database schemas that will apply to each instance. Because it is easier to identify and describe the components of transactional document types, it is easier to combine them to create information repositories and databases, easier to specify information needs or queries that use them, and easier to carry out the comparison processes that find the information that satisfies those needs. This more granular description for transactional documents also enables more opportunity to reuse information and tailor user interfaces for different contexts and devices.

There is no clear boundary between narrative and transactional document types—indeed, Glushko and McGrath (2005) use the notion of a “Document Type Spectrum” by analogy with the continuous rainbow of the visible light spectrum—but there are nevertheless systematic differences in the degree of explicit structure and content markup that shapes how much they can be described. Because narrative document types contain a great variety of content, the set of appropriate descriptions is potentially open-ended. Having trained professionals use a controlled or standard vocabulary to yield more consistent descriptions is the traditional approach in library science to solve this problem, but doesn’t easily scale in the 21st century internet and information economy. It is certainly easy to argue that the vastness of the Web and its distributed authorship makes consistent description and authoritative organization impossible, making everything “miscellaneous” (Weinberger, 2007).

In principle, we might be able to define some degree of description that would yield optimal results given some set of requirements or purposes for an Organizing System. But in practice we usually can’t do this for two reasons, both relating to scope and scale. First, as the number of users of an Organizing System increases, it becomes more difficult to identify and anticipate all its possible purposes and constraints it must satisfy. Even if most users share the goals for the Organizing System, any particular user might have some additional specialized use
for some attributes or relationships that would require more description to satisfy, potentially “spawning description ad infinitum” (Svenonius, 2000, p. 23). If we go back to the Organizing System for the home kitchen we talked about in Section 1.2.2, suppose the home kitchen is to be used as the set for a cooking show and the designers want to arrange cookbooks to make the background visually pleasing. It seems unreasonable to expect that they could search a catalog of cookbooks on the basis of size and spine color because these additional descriptive elements are of little value to users who aren’t designers of sets for cooking shows. Nevertheless, in some cases, the exception proves the rule, and for specialized or unique items like archeological artifacts from Pompeii or Darwin’s notebooks and collections (Darwin, 2011), there is almost no limit to the amount of description that can and will be assigned by domain experts.

Even if it were possible to implement some optimal degree of description in a particular Organizing System, we would still encounter problems when multiple Organizing Systems exist in the same domain or in domains that intersect across context, device, or application boundaries. Since Organizing Systems are designed and evolve to satisfy the specific requirements of their particular context, companies will often describe the same things differently, which creates integration and interoperability problems when companies need to exchange and combine information. Chapter 11, “Interoperability,” discusses these challenges in the creation of an aggregated product catalog when different manufacturers describe their items using information models of varying degrees of structure. For example, one manufacturer’s catalog might describe a clothing item like a shirt by name, item number, and a text description, while another might decompose the shirt description into a set of explicit properties and possible values for size, color, and fabric. Some industries strive to reduce these problems caused by incompatible description by promoting reference models or standards, but developing and complying with them can take substantial effort, as we’ll see in Chapter 2, “The Context for Organizing Systems.”

A second constraint on the degree of organization comes from the absolute size of the collection within the scope of the Organizing System. Organizing more items requires more descriptions to distinguish any particular item from the rest. Similar items also need to be grouped or classified to emphasize the most important distinctions among the complete set of items in the collection. A small neighborhood restaurant might have a short wine list with just ten items, 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 item. This leaves two approaches that can be done separately or in tandem. The simpler approach is to describe sets of items or documents as a set or group, which is especially sensible for archives with its emphasis on the fonds (see Section 1.2.1). The second approach is to rely on automated and more general-purpose organizing technologies that organize items through computational means. Search engines like Google are familiar examples of computational organizing technology, and section 8.5, “Computational Classification,” describes other common techniques in machine learning, clustering, and discriminant analysis that can be used to create a system of categories and to assign items to them.
Chapter 6, “Metadata,” focuses on the representation and management of descriptions, taking a more technological or implementation perspective. Chapter 11, “Interoperability,” and Chapter 12, “From IO to IR,” discuss how the nature and extent of descriptions determines the capabilities of the processes that locate, compare, combine, or otherwise use them in information-intensive domains.

1.2.4 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, p. 194-195).

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. These explicit acts of organization by the author are often supplemented by additional organizing information supplied by the publisher or others, such as an ISBN or Library of Congress call number or subject headings. All of these pieces of data constitute metadata, or information about information.

Digital photos, videos, and documents are always organized to some minimal degree because descriptions are assigned automatically to them by the technology used to create them. At a minimum, these descriptions include the creation time and storage format for the item, or chronologically by the auto-assigned filename (IMG00001.JPG, IMG00002.JPG, etc.), but often are much more detailed. For example, most digital cameras annotate each photo with metadata 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.

The Organizing System framework recasts the traditional tradeoff between information organization and information retrieval as the decision about when the organization is imposed, “on the way in” or “on the way out.” This contrast is easiest to see if we compare traditional libraries with Google and other search engines. Before an item is made available to a user in a library, it is assigned a description according to a controlled vocabulary and embedded in standard classification systems. Not all items receive the same amount of description and classification, but some amount of human “organizing intelligence” is applied to every information source in the library “on the way in” when it is added to the library’s collection. And because libraries have always treated information about a reader’s activities and choices as private, they would not alter the description or organization of items to personalize them “on the way out.”

On the other hand, Google applies massive computational power to analyze the contents and associated structures (like links between web pages) to impose organization on items of information 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 items when it can, but its unparalleled technological capabilities and scale yield competitive advantage in imposing organization on information that wasn’t previously organized digitally. Google has even been criticized for ignoring or undervaluing the descriptive metadata and classifications previously assigned to items by people and replacing them with computationally assigned descriptors (Nunberg, 2009). 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 retrieval” and “information organization” is no longer defensible.

However, it is an oversimplification to contrast “on the way in” organization performed by authors or librarians with “on the way out” organization performed by computers because the nature and extent of organization almost always changes over time as the items governed by the Organizing System are used. The arrangement of items in a kitchen or in an office changes incrementally as frequently used items 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 don’t 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 information items 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 things, but search engines and advertisers also exploit the less visible organization created over time by information about which items were viewed and which links were followed.

This 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. These two terms are roughly equivalent, but the former is most often used for libraries, museums, or archives and the latter for enterprise or inter-enterprise contexts. Curation and governance are activities whose goals are to preserve the investments in an Organizing System by anticipating and responding to changes in requirements and opportunities. 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. These topics are discussed in Chapter 2, “The Context for Organizing Systems.”

An important activity in collections of both physical objects and information items is ensuring their persistence. Physical objects deteriorate over time, usually in a gradual or continuous way that can be managed by regular inspection and maintenance. Digital objects, in contrast, can be lost in a moment, making periodic backup and offsite storage essential to enable business continuity in the event of technology failure. But storage of the digital content alone is insufficient. To ensure meaningful future access, it might be necessary to preserve or continually upgrade any software or computers that implement the Organizing System. The rapid obsolescence of computing and digital storage technology has encouraged the migration of many Organizing Systems to implementation “in the cloud,” but this simply shifts the technology concern to the cloud operator and introduces other concerns like data privacy.

A related concern to ensuring that the items in a collection persist over time is ensuring that they do so in a way that preserves their authenticity. In Organizing Systems like museums and archives that preserve rare or culturally important objects or documents this concern is expressed as the principle of provenance. This is the history of the ownership of a collection or the items in it. Since ancient Rome notaries have authenticated the creation of important
documents, which then must be preserved with an unbroken “chain of custody” to demonstrate that the items have maintained their integrity or value as evidence. A uniquely Chinese technique in Organizing Systems is the imprinting of elaborate red seals on documents, books, and paintings that collectively record the provenance of ownership and the review and approval of the artifact by emperors or important officials.

1.2.5 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, 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. Because our notion of the Organizing System is broader than the “bibliographic universe,” it is necessary to extend her taxonomy. In particular, we identify authors as a subset of the non-professional user population, and further distinguish users in informal and formal/institutional contexts.

Professional organizers undergo extensive training to learn the concepts, controlled descriptive vocabularies, and standard classifications in the particular domains in which they work. They can create and maintain Organizing Systems with consistent high quality, but “any task that requires an organizing intelligence to engage in research is costly” (Svenonius, 2000, p. 27). Expanding the scope of Organizing Systems beyond the “bibliographic universe” expands the class of professional organizers to include 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.

Authors 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 the item, they will assign descriptions or classifications that they expect will be useful to those users.

On the other hand, 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 isn’t useful or accurate for others. For example, many pictures in Flickr are tagged with “mywife” or “mybaby” or “mytruck” (or “my” concatenated with almost anything), and many web sites in del.icio.us are tagged with “readthis” or “buythis”—highly subjective tags that clearly wouldn’t mean the same thing to other users.

Fortunately most users of these sorts of “Web 2.0” or “community content” applications at least partly recognize that organization emerges from the aggregated contributions of all users, which provides incentive to use less egocentric descriptors and classifications. The staggering number of users and items on the most popular sites inevitably leads to “tag convergence” simply because of the statistics of large sample sizes.

The creation of Organizing Systems from collective contributions of loosely coupled participants in social networks or Web 2.0 applications seems to work best in domains where the
economic costs and benefits are small. In contrast, where business efficiency or competitive advantage depends on an Organizing System, more formal processes like those of standards or quasi-standards organizations are typically used to define descriptive vocabularies and other kinds of information models. Even so, for simple information models, the “microformats” movement can be seen as an attempt to supplant formally created Organizing Systems with crowdsourced ones. Getting from merely “collected intelligence” to truly “collective intelligence” isn’t trivial, but incorporating templates in wikis, blogs, and other socially oriented web applications encourages people to provide more structured descriptions and “a little semantics goes a long way” when combined with machine learning and other computational techniques for identifying patterns (Gruber, 2008).

Tagging, bookmarking, and rating mechanisms are increasingly being adapted for use inside companies as techniques for knowledge management, a trend named “Enterprise 2.0” to contrast with “Web 2.0” (McAfee, 2009). Because every user is authenticated to their real identities, and organizational norms and incentives restrict and shape the purposes and nature of the descriptions that users provide, these applications seem to capture expertise and interests implicitly and at lower cost than traditional knowledge management applications. Similarly, a class of Organizing Systems targeted for academic researchers, including CiteULike and Connotea, has had some success combining authoritative or professional metadata about books and publications with user-generated tags. The former metadata is reliable for retrieving the specific item, and the latter is more useful in support of exploratory browsing to find related information.

Automated and computerized processes can create the descriptions and classifications in an Organizing System “on the way in” or “on the way out,” but in both cases, their use is primarily driven by scale. The benefits of digital cameras, videorecorders, and similar devices would be far fewer if people had to manually identify each item when creating it. Instead, as we discussed earlier in this section, the devices can automatically assign some contextual metadata. Similarly, competitive pressures on vendors to provide real-time and context-sensitive information services mandates automated collection of contextual information like location from mobile phones, portable book readers and tablet computers.

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 (He, et al 2007)—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. 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.

1.3 Organizing this Book

The concept of the Organizing System highlights the dimensions that collectively determine the extent and nature of information organization and the capabilities of the processes that compare, combine, and transform the organized information. So this book, unlike other ones that nominally deal with the same topics, has no need to be divided into separate “information
important topics in information organization, such as conceptual modeling, semantic representation, vocabulary and metadata design, classification, naming, authority control, and standardization have traditionally been framed by a public sector bibliographic perspective. This book will update and broaden the coverage of these topics to include more private sector and non-bibliographic public sector contexts, multi- and social media, and new information-intensive applications and service systems enabled by mobile, pervasive computing, and scientific computing. The book will reinvigorate the foundations and principles of traditional library and information sciences and extend them to these new contexts.
References


