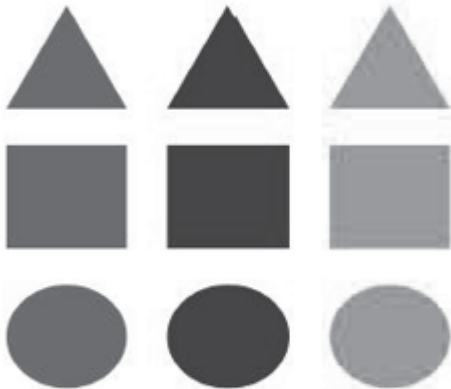


THE DISCIPLINE OF **ORGANIZING**

INFORMATICS EDITION



Edited by **ROBERT J. GLUSHKO**

This version of the 4th Edition (2016) is available under Creative Commons license CC BY-NC -
<https://creativecommons.org/licenses/by-nc/4.0/>

Chapter 2

Design Decisions in Organizing Systems

Robert J. Glushko

2.1.	Introduction	61
2.2.	What Is Being Organized?	64
2.3.	Why Is It Being Organized?	66
2.4.	How Much Is It Being Organized?	70
2.5.	When Is It Being Organized?	76
2.6.	How (or by Whom) Is It Organized?	79
2.7.	Where is it being Organized?	81
2.8.	Key Points in Chapter Two	83

2.1 Introduction

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

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

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

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

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

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

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

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

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

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

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

2.2 What Is Being Organized?

“What is difficult to identify is difficult to describe and therefore difficult to organize.”

— (Svenonius 2000, p. 13)

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

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

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

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

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

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

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

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

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

Computational Descriptions of People

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

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

2.3 Why Is It Being Organized?

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

— (Svenonius 2000 p. xi)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4 How Much Is It Being Organized?

“It is a general bibliographic truth that not all documents should be accorded the same degree of organization.”

— (Svenonius 2000, p. 24)

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

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

Power and Politics in Organizing

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Using “Information Theory” to Quantify Organization

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

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

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

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

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

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

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

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

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

2.5 When Is It Being Organized?

"Because bibliographic description, when manually performed, is expensive, it seems likely that the 'pre' organizing of information will continue to shift incrementally toward 'post' organizing."

— (Svenonius 2000, p. 194-195)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.6 How (or by Whom) Is It Organized?

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

— (Svenonius 2000, p. 26)

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

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

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

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

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

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

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

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

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

2.7 Where is it being Organized?

“Bibliographic control requires fixing a document in the bibliographic universe by its space-time coordinates.”

— (Svenonius 2000, p. 120)

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

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

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

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

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

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

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

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

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

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

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

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

2.8 Key Points in Chapter Two

- A dimensional perspective makes it easier to translate between category- and discipline-specific vocabularies so that people from different disciplines can have mutually intelligible discussions about their organizing activities.
(See §2.1 Introduction (page 61))
- In different situations, the same “thing” can be treated as a unique item, one of many equivalent members of a broad category, or a component of an item rather than as an item on its own.
(See §2.2 What Is Being Organized? (page 64))
- A single physical resource can only be in one place at a time, and interactions with it are constrained by its size, location, and other properties. In contrast, digital copies and surrogates can exist in many places at once and enable searching, sorting, and other interactions with an efficiency and scale impossible for tangible things.
(See §2.2 What Is Being Organized? (page 64))
- When the resources being organized consist of information content, deciding on the unit of organization is challenging because it might be necessary to look beyond physical properties and consider conceptual or intellectual equivalence.
(See §2.2 What Is Being Organized? (page 64))
- Libraries, museums, and archives are often classified as *memory institutions* to emphasize their primary emphasis on resource preservation.
(See §2.3 Why Is It Being Organized? (page 66))
- Businesses and governmental agencies are usually required by law to keep records of financial transactions, decision-making, personnel matters, and other information essential to business continuity, compliance with regulations and legal procedures, and transparency.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Endnotes for Chapter 2

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

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

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

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

[32][Bus] (Silverman 2013)

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

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

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

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

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

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

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

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

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

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