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

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

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

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

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

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

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

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

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

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

**Designing resource-based interactions**
What are the most important and frequent queries that need to be pre-configured?
Maintaining

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

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

**Figure 3.1. Four Activities in all Organizing Systems.**

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

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

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

Domain-specific methods and vocabularies evolve over time to capture the complex and distinctive sets of experiences and practices of their respective disciplines. We can identify correspondences and overlapping meanings, but they are not synonyms or substitutes for each other. We propose more general terms like selection and maintenance, not as lowest common denominator replacements for these more specialized ones, but to facilitate communication and cooperation across the numerous disciplines that are concerned with organizing.
It might sound odd to describe the animals in a zoo as resources, to think of viewing a painting in a museum as an interaction, or to say that destroying information to comply with privacy regulations is maintenance. Taking a broader perspective on the activities in organizing systems so that we can identify best practices and patterns enables people with different backgrounds and working in different domains to understand and learn from each other.

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

### Navigating this chapter

In §1.3 The Concept of “Resource” (page 36) and §2.2 What Is Being Organized? (page 66) we briefly discussed the fundamental concept of a resource. In this chapter, we describe the four primary activities with resources, using examples from many different kinds of organizing systems.

- §3.2 Selecting Resources (page 96)
- §3.3 Organizing Resources (page 102)
- §3.4 Designing Resource-based Interactions (page 127)
- §3.5 Maintaining Resources (page 138)

We emphasize the activities of organizing and of designing resource-based interactions that make use of the organization imposed on the resources. We discuss selection and maintenance to create the context for the organizing activities and to highlight the interdependencies of organizing and these other activities. This broad survey enables us to compare and contrast the activities in different resource domains, setting the stage for a more thorough discussion of resources and resource description in Chapter 4 and Chapter 5.
3.2 Selecting Resources

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

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

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

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

3.2.1 Selection Criteria

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

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

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

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

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

For example, when a manufacturer of physical goods selects the materials or components that are transformed into its products, it carefully evaluates the candidate resources and their suppliers before making them part of its supply chain. The manufacturer would test the resources against required values of measurable characteristics like chemical purity, strength, capacity, and reliability. A business looking for transactional or demographic data to guide a business expansion strategy would specify different measurable characteristics; data files must be valid with respect to a schema, must contain no duplicates or personal-
Selection is an essential activity in creating organizing systems whose purpose is to combine separate web services or resources to create a composite service or application according to the business design philosophy of Service Oriented Architecture (SOA). When an information-intensive enterprise or application combines its internal services with ones provided by others via Application Programming Interfaces (APIs), the resources are selected to create a combined collection of services according to the “core competency” principle: resources are selected and combined to exploit the first party’s internal capabilities and those of its service partners better than any other combination of services could. For example, instead of writing millions of lines of code and collecting detailed maps to build an interactive map in an application, you can get access to the Google Maps organizing system with just a few lines of code. (See the sidebar, Selection of Web-based Resources (page 98))

Scientific and business data are ideally selected after assessments of their quality and their relevance to answering specific questions. But this is easy to say and hard to do. It is essential to assess the quality of individual data items to
find data entry problems such as misspellings and duplicate records, or data values that are illegal, statistical outliers, or otherwise suspicious. It is also essential to assess the quality of data as a collection to determine if there are problems in what data was collected, by whom or how it was collected and managed, the format and precision in which it is stored, whether the schema governing each instance is rigorous enough, and whether the collection is complete. In addition, copyright, licensing, consumer protection laws, competitive considerations, or simply the lack of incentives to share resources make it difficult to obtain the best or most appropriate resources. See the sidebar, Assessing and Addressing Data Quality (page 102)

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

Both libraries and museums typically formalize their selection principles in collection development policies that establish priorities for acquiring resources that reflect the people they serve and the services they provide to them. The diversity of user types in public libraries and many museums implies that narrowly-targeted criteria would produce a collection of resources that would fail to satisfy many of the users. As a result, libraries typically select resources on the basis of broader criteria like their utility and relevance to their user populations, and try to choose resources that add the most value to their existing collections, given the cost constraints that most libraries are currently facing. Museums often emphasize intrinsic value, scarcity, or uniqueness as selection criteria, even if the resources lack any contemporary use.

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

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

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

### 3.2.2 Looking “Upstream” and “Downstream” to Select Resources

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

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

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

The best outcome is to find a credible supplier of good quality resources. However, if an otherwise desirable supplier does not currently produce resources of sufficient quality, it is worth trying to improve the quality by changing the process using instruction or incentives. Advocates for open government have
succeeded in getting numerous US government entities to publish data for free in machine-readable formats, but it was partly as a result of somewhat subversive demonstration projects and shaming that the government finally created data.gov in 2009. A clear lesson from the “quality movement” and statistical process control is that interventions that fix quality problems at their source are almost always a better investment than repeated work to fix problems that were preventable (see endnote\textsuperscript{297} [Bus]). But if you cannot find the headwaters or you are not able to address quality problems at their source, you can sometimes transform the resources to give them the characteristics or quality they need.\textsuperscript{53[DS]} (See the sidebar, Assessing and Addressing Data Quality (page 102), and §10.3.2 Transforming Resources for Interactions (page 530).)

When you cannot obtain resources directly from their source, even if you have confidence in their quality at that point, it is important to analyze any evidence or records of their use or interactions as they flow downstream. (See §4.5) Physical resources are often associated with printed or digital documents that make claims about their origin and authenticity, and often have bar codes, RFID tags, or other technological mechanisms that enable them to be tracked from their headwaters to the places where they are used. Tracking is very important for data resources because they can often be added to, derived from, or otherwise changed without leaving visible traces. Just as the water from melted mountain snow becomes less pure as its flows downstream, a data resource can become “dirty” or “noisy” over time, reducing its quality from the perspective of another person or computational agent further downstream. Data often gets dirty when it is combined with other datasets that contain duplicate or seemingly-duplicate information. Data can also become dirty when the hardware or software that stores it changes. Subtle differences in representation formats, transaction management, enforcement of integrity constraints, and calculations of derived values can change the original data.

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

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

If an organizing system uses data acquired from some external source, it is essential to assess its quality as an “intake” process. Ideally, the data comes with a schema that explicitly specifies what is expected, including legal structures, data types, and values (See §9.2 Structuring Descriptions (page 469)). This intake process runs tests that find problems and then runs processes to fix the problems.

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

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

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

3.3 Organizing Resources

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

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

A simple principle for organizing resources is colocation — putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box. However, most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole. What properties are significant and how to think about them depends on the number of resources being organized, the purposes for which they are being organized, and on the experiences and implicit or explicit biases of the intended users of the organizing system. The implementation of the organizing system also shapes the need for, and the nature of, the resource properties.  

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

“Subject matter” organization involves the use of a classification system that provides categories and descriptive terms for indicating what a resource is about. Because they use aboutness properties that are not directly perceived, methods for assigning subject classifications are intellectually-intensive and in many cases require rigorous training to be performed consistently and appropriately.  

Nevertheless, the cost and time required for this human effort motivates the use of computational techniques for organizing resources.

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

3.3.1 Organizing Physical Resources

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

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

Similarly, because they have different material manifestations, we usually organize our printed books in a different location than our record albums, which might be near but remain separate from our CDs and DVDs. This is partly because the storage environments for physical resources (shelves, cabinets, closets, and so on) have co-evolved with the physical resources they store.60[LIS]

The resource collections of organizing systems in physical environments often grow to fit the size of the environment or place in which they are maintained—the bookshelf, closet, warehouse, library or museum building. Their scale can be large: the Smithsonian Institute in Washington, D.C., the world’s largest museum and research complex, consists of 19 museums, 9 research facilities, a zoo and a library with 1.5 million books. However, at some point, any physical space gets too crowded, and it is difficult and expensive to add new floors or galleries to an existing library or museum.

3.3.1.1 Organizing with Properties of Physical Resources

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

This inescapable aspect of visual perception was first formalized by German psychologists starting a century ago as the Gestalt principles (see the sidebar,
How people are organized into businesses is the essence of the discipline of management, and different aspects are taught in industrial organization and behavior, operations, entrepreneurship, and other courses. Organizing people in a business is often called “human resource management,” and many of the principles for organizing physical resources and information resources apply to organizing people.

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

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

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

Physical resources are also commonly organized using intrinsically associated properties such as the place and time they were created or discovered. The shirts in your clothes closet might be arranged by color, by fabric, or by style. We can view dress shirts, T-shirts, Hawaiian shirts and other styles as configurations of shirt properties that are so frequent and familiar that they have become linguistic and cultural categories. Other people might think about these same properties or categories differently, using a greater or lesser number of colors.
Psychologists Max Wertheimer, Wolfgang Kohler, and Kurt Koffka proposed several principles—proximity, similarity, continuity, connection, enclosure, and closure—that explain how our visual system imposes order on what it sees. There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

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

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

or ordering them differently, sorting the shirts by style first and then by color, or vice versa.

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

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

This simple example of a list of files illustrates the Gestalt similarity principle that elements that look similar are seen as being in the same group. The shape and visual design for the icons distinguishes files from folders, and file types from each other. These interpretations do not require you to be able to read the labels, although the visual similarity of the names suggests some similarity of content.

3.3.1.2 Organizing with Descriptions of Physical Resources

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

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

In organizing systems that use additional resources to identify or describe primary ones, “adding to a collection” is a logical act that need not require any actual movement, copying, or reorganization of the primary resources. This virtual addition allows the same resources to be part of many collections at the same time; the same book can be listed in many bibliographies, the same web page can be in many lists of web bookmarks and have incoming links from many different pages, and a publisher’s digital article repository can be licensed to any number of libraries.
Library catalogs were managed as collections of printed cards for much of the 20th century, and the wooden cabinets that contained them were ubiquitous functional furniture in every library. Today such cabinets are often considered “retro” or antique treasures.

(Photo by R. Glushko.)

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

(Photo by R. Glushko.)

3.3.2 Organizing Places

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

- The land itself, as when we lay out city plans when organizing how people live together and interact in cities.

- A “built environment,” a human-made space, particular building, or a set of connected spaces and buildings. A built environment could be a museum, airport, hospital, casino, department store, farm, road system, or any kind of building or space where resources are arranged and people interact with them.

- The orientation and navigation aids that enable users to understand and interact in built environments. These are resource descriptions that support the interaction requirements of the users.

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

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

Because the United States, and especially the American West, was not heavily settled until much more recently compared to most of Europe and Asia, it was a place for people to experiment with new ideas in urban design. The natural human tendency to impose order on habitation location had ample room to do just that. The easiest and most efficient way to organize space is using a coordinate grid, with streets intersecting at perpendicular angles. Salt Lake City, Albuquerque, Phoenix, and Seattle are notable examples of grid cities. An interesting hybrid structure exists in Washington DC, which has radiating diagonal avenues overlaid on a grid.641A

3.3.2.2 Organizing Built Environments

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

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

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

If you have ever been lost in a casino or had trouble finding the exit you can be
sure you experienced a casino that achieved its main design goals: keeping peo­
ple inside and making it easy for them to lose track of time because they lack
both windows and clocks. As American architect Robert Venturi points out, “The
intricate maze under the low ceiling never connects with outside light or out­
side space...This disorients the occupant in space and time... He loses track of
where he is and when it is.” 65[Mus]

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

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

However, consider the dynamic created by queue design at Disneyland to give
priority to people with physical limitations and disabilities. This seemingly so­
cially respectful decision was exploited by a devious collaboration between disa­
bled people and wealthy non-disabled people who hired them to pose as family
members, enabling the entire “family” to cut ahead of everyone else. In re­
response, Disney modified the policy favoring disabled patrons, causing numerous
complaints about Disney's insensitivity to their concerns. 66[IA]

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

3.3.2.3 Orientation and Wayfinding Mechanisms

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

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

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

3.3.3 Organizing Digital Resources

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

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

An organizing system for digital resources can also use digital description resources that are associated with them. Since the incremental costs of adding processing and storage capacity to digital organizing systems are small, collections of both primary digital resources and description resources can be arbitrarily large. Digital organizing systems can support collections and interactions at a scale that is impossible in organizing systems that are entirely physical, and they can implement services and functions that exploit the exponentially growing processing, storage and communication capabilities available today. This all sounds good, unless you are...
the small local business with limited onsite inventory that cannot compete with global web retailers that offer many more choices from a network of warehouses.\textsuperscript{73[Web]}

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

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

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

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

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

- \textit{Ordinal} level data indicates a direction or ranking on some naturally ordered scale. We know that the first place finisher in a race came in ahead of the second place one, who finished ahead of the third place finisher, but this result conveys no information about the spacing among the racers at the finish line. The middle value in a sorted list is the \textit{median}. 

\textit{The Discipline of Organizing}

Chapter 3 — Activities in Organizing Systems
Memories can be viewed either as physical (because at some level they are represented in the brain) or as digital (because they are retrieved as electrical impulses), but memory techniques like the method of loci and memory palaces reify this duality in an interesting way. While physical resources must be stored in physical locations, our powerful spatial memory provides an opportunity for us to, in a sense, store mental resources in physical locations. Our hippocampus, the brain component dedicated to memory, is highly developed for storing and recalling memories of physical locations. The ancient Greeks relied on this capability and devised a mnemonic system—the method of loci—which involved attaching things to remember, the key ideas in a speech perhaps, to well-known physical locations. While giving the speech, then, all one must do is imagine walking through that physical location from idea to idea. Today, champion memorizers use this technique to associate items with places in vividly imagined “memory palaces.” While you may not be interested in memorizing the order of a deck of cards, recognizing the power of our spatial memory may be worth considering when designing your organizing system or when analyzing the successes or failures of a system.

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

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

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

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

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

But Facebook so far is succeeding by following a walled garden strategy.

3.3.3.2 “Information Architecture” and Organizing Systems

The discipline known as *information architecture* can be viewed as a specialized approach for designing the information models and their systematic manifestations in user experiences on websites and in other information-intensive organizing systems. Abstract patterns of information content or organization are sometimes called architectures, so it is straightforward from the perspective of the discipline of organizing to define the activity of *information architecture* as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use. Note how the first part of this definition refers to intentional arrangement of resources, and the second to the interactions enabled by that arrangement.

Our definition of information architecture implies a methodology for the design of user interfaces and interactions that puts conceptual modeling at the foundation. Best practices in information architecture emphasize the use of systematic principles or design patterns for organizing the resources and interactions in...
The Activities of Information Architecture

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

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

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

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

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

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

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

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

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

However, interface designers can take advantage of this familiarity and employ design patterns in a less beneficial way to manipulate users, control their behaviors, or trick them into taking actions they do not intend. Patterns used this way are sometimes called Dark Patterns (page 117).

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

A model-based foundation is also essential in information visualization applications, which depict the structure and relationships in large data collections using spatial and graphical conventions to enable user interactions for exploration and analysis. By transforming data and applying color, texture, density, and other properties that are more directly perceptible, information visualization applications enable people to obtain more information than they can from text displays.

Some designers of information systems put less emphasis on conceptual modeling as an “inside-out” foundation for interaction design and more emphasis on an “outside-in” approach that highlights layout and other presentation-tier considerations with the goal of making interactions easy and enjoyable. This focus is typically called user experience design, and information architecture methods
Dark Patterns

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

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

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

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

remain an important part of it, but not beginning with explicit organizing principles implies more heuristic methods and yields less predictable results.
3.3.4 Organizing With Descriptive Statistics

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

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

- Measures of central tendency: Mean, median, and mode; which measure is appropriate depends on the level of measurement represented in the numbers being described (these measures and the concept of levels of measurements are defined in §3.3.3 Organizing Digital Resources (page 111)).

- Measures of variability: Range (the difference between the maximum and minimum values), and standard deviation (a measure of the spread of values around the mean).

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

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

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

### 3.3.4.1 Exploratory Analysis to Understand Data

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

A dataset whose fields or attributes lack information about data types and units of measure has little use because the data lacks meaning. When some, but not all parts of the data are named or annotated, avoid over-interpreting these descriptions’ meanings. (See §4.4 Naming Resources (page 203).)

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

**Figure 3.2a. Example Dataset**

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

The “ID” column contains numeric data, but every value is a different integer, and the values are contiguous. The field label “ID” suggests that this is the re-
A histogram is the simplest visualization of one-dimensional data. It is a bar graph that takes the full range of values, organizes them into a set of intervals of equal size on one axis, and then counts the number of values in each interval on the other axis.

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

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

**Figure 3.2b. Temperature**

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

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

**Median versus Average**

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

*The End of Average* tells the story of how the U.S. military designed aircraft cockpits beginning in 1926 on the basis of the average dimensions of a 1926 pilot. In 1950, researchers measured over four thousand pilots only to discover that no actual pilot had average values on all the measures, and recommended adjustable seats and controls in cockpit design.85[DS]
The “Weight” column has about 220 different numeric values, from 82 to 300, and judging from this range we can infer that the weights are measured in pounds. The data follows an uneven distribution with peaks around 160 and 200, and a small peak at 300. This odd shape appears in the histogram of Figure 3.2d, Weight. The two peaks in this so-called multi-modal histogram suggest that this measure is mixing two different kinds of resources, and indeed it is because weights of men and women follow different distributions. It would thus be useful to use the categorical “Sex” data to separate these populations, and Figure 3.2e, Sex and Weight: Female shows how analyzing weight for women and men as different populations is much more informative as an organizing principle than combining them.

What about the odd peak in the distribution at 300? End of range anomalies like this generally reflect a limitation in the device or system that created the data. In this case, the weight scale must have an upper limit of 300 pounds, so the peak represents the people whose weight is 300 or greater.
Figure 3.2d. Weight

Figure 3.2e. Sex and Weight: Female
3.3.4.2 Detecting Errors and Fraud in Data

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

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

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

- If expenses are reimbursed up to some maximum allowed value, look for data items with that exact value.
- When any value exceeding some threshold triggers more careful analysis, look for other data items just below that threshold.
• When invoices or claims are paid on receipt, and only a sample are sub-
sequently audited, look for duplicate submissions.

• Calculate the ratio of the maximum to the minimum value for purchases in
some category (such as the unit price paid for items from suppliers); items
with large ratios might indicate fraud where the supplier “kicks back” some
of the money to the purchaser.

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

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

3.3.5 Organizing with Multiple Resource Properties

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

If multiple resource properties are considered in a fixed order, the resulting ar-
rangement forms a logical hierarchy. The top level categories of resources are
created based on the values of the property evaluated first, and then each cate-
gory is further subdivided using other properties until each resource is classi-
fied in only a single category. Consider the hierarchical system of folders used
by a professor to arrange the digital resources on his computer; the first level
distinguishes personal documents from work-related documents; work is then
subdivided into teaching and research, teaching is subdivided by year, and year
divided by course.
For physical resources, mapping categories to physical locations is another required step; for example, resources in the “kitchen utensils” category might all be arranged in drawers near a workspace, with “silverware” arranged more precisely to separate knives, forks, and spoons.

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

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

Applications that organize large collections of digital information, including those for search, natural language processing, image classification, personalized recommendation, and other computationally intensive domains, often use huge numbers of resource properties (which are often called “features” or “dimensions”). For example, in document collections each unique word might initially be treated as a feature by machine learning algorithms, so there might be tens of thousands of features.

Chapter 8, *Classification: Assigning Resources to Categories* explains principles and methods for hierarchical and faceted classification in more detail.
3.4 Designing Resource-based Interactions

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

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

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

3.4.1 Affordance and Capability

The concept of affordance, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor’s capabilities and cognition, what can be done with the resource. Including capabilities and cognition brings accessibility considerations into the definition of affordance. A resource is only accessible when it supports interac-
tions, and it is ineffective design to implement interactions with resources that some people are unable to perform. A person who cannot see text cannot read it, or if they are confined to a wheelchair they cannot select a book from a tall library shelf. Describing or transforming resources to ensure their accessibility is discussed in greater detail in §3.4.2.3 Accessibility (page 134).

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

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

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

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

Second, most of the interactions that are supported by digital resources are not apparent when you encounter them. You cannot tell from their names, but you probably know from past experience what interactions are possible with files of types “.doc” and “.pdf.” You probably do not know what interactions take place with “.xpi” and “.mobi” files.\textsuperscript{90}[Com]

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

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

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

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

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

3.4.2 Interaction and Value Creation

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

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

3.4.2.1 Value Creation with Physical Resources

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

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

A large university library contains millions of books and academic journals, and access to those resources can require a long walk deep into the library stacks after a consultation with a reference librarian or a search in a library catalog. For decades library users searched through description resources—first printed library cards, and then online catalogs and databases of bibliographic citations—to locate the primary resources they wanted to access. The surrogate descriptions of the resources needed to be detailed so that users could assess the relevance of the resource without expending the significant effort of obtaining and examining the primary resource. 94[LIS]
However, for most people the primary purpose of interacting with a library is to access the information contained in its resources. Many people prefer accessing digital documents or books to accessing the original physical resource because the incidental physical and interpersonal interactions are unnecessary. In addition, many library searches are for known items, which is easily supported by digital search.\textsuperscript{95}[Mus]

In some organizing systems robotic devices, computational processes, or other entities that can act autonomously with no need for a human agent carry out interactions with physical resources. Robots have profoundly increased efficiency in materials management, “picking and packing” in warehouse fulfillment, office mail delivery, and in many other domains where human agents once located, retrieved, and delivered physical resources. A “library robot” system that can locate books and grasp them from the shelves can manage seven times as many books in the same space used by conventional open stacks.\textsuperscript{96}[LIS]

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

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

3.4.2.2 Value Creation with Digital Resources

With digital resources, neither physical manipulation nor interpersonal contact is required, and the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.\textsuperscript{97}[LIS] Put another way, by replacing interactions that involve people and physical resources with symbolic ones, organizing systems can lower costs without reducing user satisfaction. This is why so many businesses have automated their information-intensive processes with self-service technology.
An automated robot library system at San Francisco State University.

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

(Photo by Scott Abel. Used with permission.)

Similarly, web search engines eliminate the physical effort required to visit a library and enable users to consult more readily accessible digital resources. A search engine returns a list of the page titles of resources that can be directly accessed with just another click, so it takes little effort to go from the query results to the primary resource. This reduces the need for the rich surrogate de-
scriptions that libraries have always been known for because it enables rapid evaluation and iterative query refinement.\[98\Web]

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

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

Museums have aggressively embraced the web to provide access to their collections. While few museum visitors would prefer viewing a digital image over experiencing an original painting, sculpture, or other physical artifact, the alternative is often no access at all. Most museum collections are far larger than the space available to display them, so the web makes it possible to provide access to otherwise hidden resources.\[101\Mus]

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

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

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

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

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

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

3.4.2.3 Accessibility

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

\textit{Chapter 3 — Activities in Organizing Systems}
Many of the techniques for making a resource accessible involve transforming the resource or its description into a different form so someone who could not perceive it or interact with it in its original form can now do so. The most common operating systems all come with general-purpose accessibility features such as reading text aloud, recognizing speech, magnifying text, increasing cursor size, signaling with flashing lights instead of with sounds, lights to signal keyboard shortcuts for selecting and navigating, and connecting to devices for displaying Braille. Google Translate converts text in one language to another, and many people use it to create a rough draft that is finished by a human translator. Other techniques are not generic and automatic, and instead require investment by authors or designers to make information accessible. Websites are more accessible when images or other non-text content types have straightforward titles, captions, and “alt text” that describes what they are about. Consistent placement and appearance of navigation controls and interaction widgets is essential; for example, in a shopping site “My Cart” might always be found at the top right corner of the page.

If authors apply semantic and structural markup to the text and use formats that distinguish it from presentation instructions, page outlines and summaries can be generated to enhance navigation, and search can be made more precise by limiting it to particular sections or content types. As the “Information IQ” of the source format increases, more can be done to make it more accessible (see §4.2.2 Resource Format (page 183) and Figure 4.3, Information IQ.).

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

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

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

Since text transcripts are machine-readable, unlike audio or video files, adding text transcripts makes it possible for search engines to index audio and video in ways that were previously impossible. Pop Up Archive, an audio search company in Oakland, California, works with speech-to-text software specially trained for news media and spoken word content to make radio, podcasts, and archival audio searchable. A challenge for audio search is that even though a transcription with a few mistakes works just fine for search engines, people often expect transcriptions to be perfect.\textsuperscript{110[Com]}

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

3.4.3 Access Policies

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

The library stacks might be open to anyone, but rare documents in a special collection are only accessible to authorized researchers. The same is true of museums, which typically have only a fraction of their collections on public display.

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

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

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

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

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

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

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

University libraries typically complement or replace parts of their print collections with networked access to digital content licensed from publishers. Typical licensing terms then require them to restrict access to users that are associated with the university, either by being on campus or by using virtual private network (VPN) software that controls remote access to the library network. Copyright law limits the uses of a substantial majority of the books in the collections of major libraries, prohibiting them from being made fully available in digital formats. Museums often prohibit photography because they do not own the rights to modern works they display.
Whether an access policy is designed or imposed is not always clear. Policies that were originally designed for a particular organizing system may over time become best practices or industry standards, which regulators or industry groups not satisfied with “self-regulation” later impose. Museums might aggressively enforce a ban on photography not just to comply with copyright law, but also to enhance the revenue they get from selling posters and reproductions.

### 3.5 Maintaining Resources

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

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

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

Ideally, maintenance requirements for resources should be anticipated when organizing principles are defined and implemented. Resource descriptions to support preservation of digital resources are especially important.\(^{113}\)\([\text{Arc}]\)

#### 3.5.1 Motivations for Maintaining Resources

The concept of memory institution broadly applies to a great many organizing systems that share the goal of preserving knowledge and cultural heritage.\(^{114}\)\([\text{LIS}]\)

The primary resources in libraries, museums, data archives or other memory institutions are fixed cultural, historic, or scientific artifacts that are maintained...
because they are unique and original items with future value. This is why the Musée du Louvre preserves the portrait of the *Mona Lisa* and the United States National Archives preserves the *Declaration of Independence*.\[^{115}\text{Arc}\]

In contrast, in businesses organizing systems, many of the resources that are collected and managed have limited intrinsic value. The motivation for preservation and maintenance is economic; resources are maintained because they are essential in running the business. For example, businesses collect and preserve information about employees, inventory, orders, invoices, etc., because it ensures internal goals of efficiency, revenue generation, and competitive advantage. The same resources (e.g., customer information) are often used by more than one part of the business.\[^{116}\text{Bus}\] Maintaining the accuracy and consistency of changing resources is a major challenge in business organizing systems.\[^{117}\text{DS}\]

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

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

Of course, libraries, museums, and archives also confront economic issues as they seek to preserve and maintain their collections and themselves as *memory institutions*.\[^{119}\text{LIS}\] They view their collections as intrinsically valuable in ways that firms generally do not. Because of this, extensive energy goes into preservation, protection, and storage of resources in memory institutions, and it is more rare that resources may be discarded or de-accessioned. Art galleries are an interesting hybrid because they organize and preserve collections that are valuable, but if they do not manage to sell some things, they will not stay in business.

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

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

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

3.5.2.1 Digitization and Preserving Resources

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

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

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

However, limits to the physical lifetime of digital storage media are much less significant than the third challenge, the fact that the software and its associated computing environment used to parse and interpret the resource at the time of preservation might no longer be available when the resource needs to be accessed. Twenty-five years ago most digital documents were created using the Word Perfect word processor, but today the vast majority is created using Microsoft Word and few people use Word Perfect today. Software and services that convert documents from old formats to new ones are widely available, but they are only useful if the old file can be read from its legacy storage medium.123[Arc]

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

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

3.5 Maintaining Resources 141
The Hathi Trust Digital Library

The Hathi Trust is a worldwide partnership of several dozen major research institutions and libraries dedicated to “collecting, organizing, preserving, communicating, and sharing the record of human knowledge.” The Hathi Trust was established in 2008 to coordinate the efforts of libraries in managing the digital copies of the books they received in return for providing books to Google for its book digitization projects. Since then the Hathi Trust has broadened its scope to include the public domain books collected by the Internet Archive and numerous other digital collections, and today its digital library has over ten million volumes. The costs of running the Hathi Trust and its digital library are shared in a transparent manner by the institutions that contributed digital collections or that want access to them, which reduces the costs for everyone compared to a “go it alone” strategy. The Hathi Trust Digital Library has separate modes for catalog search and full-text search of the library contents, unlike commercial search engines that do not distinguish them. A second important difference between the Hathi Trust Digital Library and commercial search engines is the absence of display advertising and “sponsored search” results.

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

3.5.2.2 Preserving the Web

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

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

Comprehensive web search engines like Google and Bing use crawlers to continually update their indexed collections of web pages and their search results link to the current version, so preservation of older versions is explicitly not a goal. Furthermore, search engines do not reveal any details about how frequently they update their collections of indexed pages.\(^{126}\)\(^{\text{[Web]}}\)
The Internet Archive and the “Wayback Machine”

The Internet Archive (Archive.org), founded by Brewster Kahle, makes preservation of the web its first and foremost activity, and when you enter a URI into its “Wayback Machine” you can see what a site looked like at different moments in time. For example, www.berkeley.edu was archived about 2500 times between October 1996 and January 2013, including about twice a week on average during all of 2012. Even so, since a large site like berkeley.edu often changes many times a day, the Wayback Machine’s preservation of berkeley.edu is incomplete, and it only preserves a fraction of the web’s sites. Since 2006 the Internet Archive has hosted the “Archive-It” service to enable hundreds of schools, libraries, historical societies, and other institutions to archive collections of digital resources.127[Arc]

3.5.2.3 Preserving Resource Instances

A focus on preserving particular resource instances is most clear in museums and archives, where collections typically consist of unique and original items. There are many copies and derivative works of the Mona Lisa, but if the original Mona Lisa were destroyed none of them would be acceptable as a replacement.128[Mus]

Archivists and historians argue that it is essential to preserve original documents because they convey more information than just their textual content. Paul Duguid recounts how a medical historian used faint smells of vinegar in 18th-century letters to investigate a cholera epidemic because disinfecting letters with vinegar was thought to prevent the spread of the disease. Obviously, the vinegar smell would not have been part of a digitized letter.129[Arc]

Zoos often give a distinctive or attractive animal a name and then market it as a special or unique instance. For example, the Berlin Zoo successfully marketed a polar bear named Knut to become a world famous celebrity, and the zoo made millions of dollars a year through increased visits and sales of branded merchandise. Merchandise sales have continued even though Knut died unexpectedly in March 2011, which suggests that the zoo was less interested in preserving that particular polar bear than in preserving the revenue stream based on that resource.130[Mus]

Most business organizing systems, especially those that “run the business” by supporting day-to-day operations, are designed to preserve instances. These include systems for order management, customer relationship management, inventory management, digital asset management, record management, email archiving, and more general-purpose document management. In all of these domains, it is often necessary to retrieve specific information resources to serve customers or to meet compliance or traceability goals.
Recent developments in sensor technology enable very extensive data collection about the state and performance of machines, engines, equipment, and other types of physical resources, including human ones. (Are you wearing an activity tracker right now?) When combined with historical information about maintenance activity, predictive analytics techniques can use this data to determine normal operating ranges and indicators of coming performance degradation or failures. Predictive maintenance can maximize resource lifetimes while minimizing maintenance and inventory costs. These techniques have recently been used to predict when professional basketball players are at risk of an injury, potentially enabling NBA teams to identify the best time to rest their star players without impairing their competitive strategy.\textsuperscript{131[DS]}

3.5.2.4 Preserving Resource Types

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

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

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

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

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

Sometimes, specific items in a collection might have some value or interest on their own, but they acquire even greater significance and enhanced meaning because of the context created by other items in the collection that are related in some essential way. The odd collection of “things people swallow that they should not” at the Mütter Museum is a perfect example.\[Mus\]
3.5.3 Curation

For almost a century *curation* has referred to the processes by which a resource in a collection is maintained over time, which may include actions to improve access or to restore or transform its representation or presentation.\textsuperscript{137}[LIS]

Furthermore, especially in cultural heritage collections, curation also includes research to identify, describe, and authenticate resources in a collection. Resource descriptions are often updated to reflect new knowledge or interpretations about the primary resources.\textsuperscript{138}[Arc]

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

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

3.5.3.1 Institutional Curation

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

Curation in these institutional contexts requires extensive professional training. The institutional authority empowers individuals or groups to make curation decisions. No one questions whether a museum curator or a compliance manager should be doing what they do.\textsuperscript{139}[LIS]

Institutional curation may be supported by automated methods. An “approval plan” is often implemented for the acquisition of new books by libraries that involves an initial selection of certain criteria (such as “published by an American university press; costs less than $100; not a reissue of an earlier edition; classed
within a particular Library of Congress range”) that enable libraries to automatically purchase all books meeting the criteria. While the approval plan can certainly be considered a selection activity, we cite it in maintenance as an example of a strategy to maintain the currency and relevancy of a given collection.

3.5.3.2 Individual Curation

**The Life-Changing Magic of Tidying Up**

Many of the ever growing number of self-help books about organizing seem to approach it as an intellectual contest to devise more elaborate and optimized storage strategies. Marie Kondo’s wildly popular 2014 book *The Life-changing Magic of Tidying Up*, an international best-seller, has upended the conversation with an unapologetic dogma of removal that promises to yield a happier—and much more minimalist—life for individuals with their at-home organizing systems.

Kondo’s method mandates that only what brings one joy may be kept. Everything else must be tossed — unused gifts, books kept only for reference but never referenced, unworn clothing, and anything else that does not bring its owner joy. Kondo’s approach is designed for personal organizing systems, and would be difficult to implement in systems in systems used by multiple individuals, much less institutions. However, Kondo’s rejection of the concept that things should be saved for a rainy day might benefit organizations by making them more attentive to the costs of maintaining resources with no current use.

While people must make up their own minds about how they manage their possessions, there is compelling evidence from cognitive science and behavioral economics that decision-making throughout the day can be mentally exhausting. Kondo’s approach implicitly recognizes this limitation by requiring cognitive energy up front to reduce the total number of resources to the bare minimum necessary (by one’s own “joy standards”). This philosophy has people spend decision-making energy where it counts the most and makes it easier to make maintenance decisions over time.

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

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

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

Most websites are not curated in a systematic way, and the decentralized nature of the web and its easy extensibility means that the web as a whole defies curation. It is easy to find many copies of the same document, image, music file, or video and not easy to determine which is the original, authoritative or authorized version. Broken links return “Error 404 Not Found” messages.

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

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

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

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

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

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

An extremely promising new approach to computational curation involves using scientific measuring equipment to analyze damaged physical resources and then building software models of the resources that can be manipulated to restore the resources or otherwise improve access to their content. For example, the first sound recordings were made using rotating wax cylinders; sounds caused a diaphragm to vibrate, the pattern of vibration was transferred into a connected stylus, which then cut a groove into the wax. When the cylinder was rotated past a passive stylus, it would vibrate according to the groove pattern, and the amplified vibrations could be heard as the replayed sound. Unfortunately, wax cylinders from the 19th century are now so fragile that they would fall apart if they were played. This dilemma was resolved by Carl Haber, an experimental physicist at the Lawrence Berkeley Laboratory. Haber used image processing
techniques to convert microscope-detailed scans of the grooves in the wax cylin-
ders. Measurements of the grooves could then be transformed to reproduce the
sounds captured in the grooves.\footnote{150[Arc]}

A second example of computational curation applied to digital preservation is
work done by a research team led by Melissa Terras and Tim Weyrich at Univer-
sity College London to build a 3-dimensional model of a 17th-century “Great
Parchment Book” damaged in an 18th-century fire. The parchment was singed,
shriveled, creased, folded, and nearly impossible to read (see website). After
traditional document restoration techniques (e.g., illustrated in photos in \S3.5.2)
grew as far as they could, the researchers used digital image capture and mod-
eling techniques to create a software model of the parchment that could stretch
and flatten the digital document to discover text hidden by the damage.

3.5.3.5 Discarding, Removing, and Not Keeping

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

Many organizations admit to a distinct lack of strategy in the removal aspect of
maintenance. A firm with outdated storage technology might have to discard
older data simply to make room for new data, and might do so without consider-
ing that keeping some summary statistics would be valuable for historical analy-
sis. Other firms might be biased towards keeping information just because they
went to the trouble of collecting or acquiring it. Some amount of “intelligent”
removal is an essential ingredient in any maintenance regime, and a popular
book argues forcefully for continually discarding resources from personal or-
ganizing systems as a method of focusing on the resources that really matter.
(See the sidebar, The Life-Changing Magic of Tidying Up (page 147).)

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

Efforts by libraries to automate the discarding of books that have not circulated
for several years might seem like the obvious counterpart to their automated ac-
quision, but such efforts often produce passionate complaints from library pa-
trons.\footnote{151[LIS]}

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

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

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

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

It is worth noting that the ability to discard without having to reuse is relatively recent. Historically, the urge and need to discard has clashed with the availability of resources. In the Middle Ages, liturgical texts or music would be phased out, perhaps when the music had gone out of style or when entire sections of the liturgy were phased out by decree. When this happened, they would reuse the parchment or vellum, either by scraping it down or by flipping it over, pasting it in a book, and using the other side. The former of these solutions often created a palimpsest, a document or other resource in which the remnants of older content remain visible under the new.

Some people have difficulty in discarding things, regardless of their actual value. This behavior is called hoarding, and is now regarded as a kind of obsessive-compulsive disorder that requires treatment because it can cause emotional, physical, social, and even legal problems for the hoarder and family members. It seems unsympathetic that many TV shows and stories have been produced about especially compulsive hoarding. A famous example is that of the Collyer brothers in New York, who shut themselves off from the world for years, and when they were found dead inside their home in 1947 it contained 140 tons of collected items, including 25,000 books, fourteen pianos, thousands of bottles and tin cans, hundreds of yards of fabrics, and even a Model T car chassis.153[Arc]

3.5.4 Governance

Governance overlaps with curation in meaning, but typically has more of policy focus (what should be done), rather than a process focus (how to do it). Governance is also more frequently used to describe curation in business and scientific organizing systems rather than in libraries, archives, and museums. Governance has a broader scope than curation because it extends beyond the resources in a collection and also applies to the software, computing, and networking environments needed to use them. This broader scope also means that governance must specify the rights and responsibilities for the people who might interact with the resources, the circumstances under which that might take place, and the methods they would be allowed to use.
Corporate governance is a common term applied to the ongoing maintenance and management of the relationship between operating practices and long-term strategic goals.\textsuperscript{154} [LIS]

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

3.5.4.1 Governance in Business Organizing Systems

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

Stop and Think: Business Data Governance

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

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

Business intelligence is only as good as the data it is based on, which makes business data governance a critical concern that has rapidly developed its own specialized techniques and vocabulary. The most fundamental governance activity in information-driven businesses is identifying the “master data” about cus-
customers, employees, materials, products, suppliers, etc., that is reused by different business functions and is thus central to business operations.\footnote{DS}{158}

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

3.5.4.2 Governance in Scientific Organizing Systems

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

The Long Tail of Dark Data

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

- Selection, organizing, interaction design, and maintenance activities occur in every organizing system.
  (See §3.1 Introduction (page 91))
- These activities are not identical in every domain, but the general terms enable communication and learning about domain-specific methods and vocabularies.
  (See §3.1 Introduction (page 91))
- The most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized.
  (See §3.2 Selecting Resources (page 96))
- Memory institutions select rare and distinctive resources, but in scientific research, a sample must contain representative instances.
  (See §3.2 Selecting Resources (page 96))
- Even when the selection principles behind a collection are clear and consistent, they can be unconventional, idiosyncratic, or otherwise biased.
  (See §3.2.1 Selection Criteria (page 96))
- If you can determine where the resources come from, you can make better selection decisions by evaluating the people, processes, and organizing systems that create them.
  (See §3.2.2 Looking “Upstream” and “Downstream” to Select Resources (page 100))
- In this book we use property in a generic and ordinary sense as a synonym for feature or “characteristic.” Many cognitive and computer scientists are more precise in defining these terms and reserve property for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the property is called an attribute, “dimension,” or “variable.”
  (See §3.3 Organizing Resources (page 102))
- Most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole.
  (See §3.3 Organizing Resources (page 102))
- Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users.
  (See §3.3.1.1 Organizing with Properties of Physical Resources (page 104))
• There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

(See the sidebar, Gestalt Principles (page 106))

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

(See §3.3.2.2 Organizing Built Environments (page 109))

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

(See §3.3.3.2 “Information Architecture” and Organizing Systems (page 114))

• An emerging issue in the field of digital humanities is the requirement to recognize the materiality of the environment that enables people to create and interact with digital resources.

(See §3.3.3 Organizing Digital Resources (page 111))

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

(See §3.3.4 Organizing With Descriptive Statistics (page 118))

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

(See §3.3.4 Organizing With Descriptive Statistics (page 118))

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

(See §3.3.5 Organizing with Multiple Resource Properties (page 125))

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

(See §3.4.1 Affordance and Capability (page 127))

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

(See §3.4.1 Affordance and Capability (page 127))

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

(See §3.4.1 Affordance and Capability (page 127))

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

(See §3.4.1 Affordance and Capability (page 127))

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

(See §3.4.2.2 Value Creation with Digital Resources (page 131))

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

(See §3.4.2.2 Value Creation with Digital Resources (page 131))

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

(See §3.5.2 Preservation (page 140))

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

(See §3.5.2.1 Digitization and Preserving Resources (page 140))

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

(See §3.5.3 Curation (page 146) and §3.5.4 Governance (page 151))

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

(See §3.5.3.4 Computational Curation (page 149))
• An essential part of maintenance is the phasing out of resources that are
damaged or unusable, expired or past their effectivity dates, or no longer
relevant to any interaction.

(See §3.5.3.5 Discarding, Removing, and Not Keeping (page 150))

• Governance is essential to deal with frequent changes in business organiz­ing
systems, data quality management, access control to ensure security and
privacy, compliance, deletion, and archiving.

(See §3.5.4.1 Governance in Business Organizing Systems (page 152))

• Scientific data poses special governance problems because of its scale.

(See §3.5.4.2 Governance in Scientific Organizing Systems (page 153))

Endnotes for Chapter 3

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

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

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

[46][Com] See (Cherbakov et al. 2005), (Erl 2005a). The essence of SOA is to treat
business services or functions as components that can be combined as needed.
An SOA enables a business to quickly and cost-effectively change how it does
business and whom it does business with (suppliers, business partners, or cus­
tomers). SOA is generally implemented using web services that exchange Exten­sible Markup Language (XML) documents in real-time information flows to in­terconnect the business service components. If the business service components
are described abstractly it can be possible for one service provider to be trans­
parently substituted for another—a kind of real-time resource selection—to
maintain the desired quality of service. For example, a web retailer might send
a Shipping Request to many delivery services, one of which is selected to provide the service. It probably does not matter to the customer which delivery service handles his package, and it might not even matter to the retailer.

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

[48][LIS] See (Borgman 2000) on digitization and libraries. But while shared collections benefit users and reduce acquisition costs, if a library has defined itself as a physical place and emphasizes its holdings— the resources it directly controls —it might resist anything that reduces the importance of its physical reification, the size of its holdings, or the control it has over resources (Sandler 2006). A challenge facing conventional libraries today is to make the transition from emphasizing creation and preservation of physical collections to facilitating the use and creation of knowledge regardless of its medium and the location from which it is accessed.

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

[50][Web] Web resources are typically discovered by computerized “web crawlers” that find them by following links in a methodical automated manner. Web crawlers can be used to create topic-based or domain-specific collections of web resources by changing the “breadth-first” policy of generic crawlers to a “best-first” approach. Such “focused crawlers” only visit pages that have a high probability of being relevant to the topic or domain, which can be estimated by analyzing the similarity of the text of the linking and linked pages, terms in the linked page’s URI, or locating explicit semantic annotation that describes their content or their interfaces if they are invokable services (Bergmark et al. 2002), (Ding et al. 2004).
FTC Fair Information Practice Principles say that consumer data collected for one purpose cannot be used for other purposes without the consumer’s consent. Sometimes called the consumer privacy bill of rights.

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

Large research libraries have historically viewed their collections as their intellectual capital and have policies that specify the subjects and sources that they intend to emphasize as they build their collections. See (Evans 2000). Museums are often wary of accepting items that might not have been legally acquired or that have claims on them from donor heirs or descendant groups; in the USA, much controversy exists because museums contain many human skeletal remains and artifacts that Native American groups want to be repatriated.

Adding a resource to a museum implies an obligation to preserve it forever, so many museums follow rigorous accessioning procedures before accepting it. Likewise, archives usually perform an additional appraisal step to determine the quality and value of materials offered to them.

In archives, common appraisal criteria include uniqueness, the credibility of the source, the extent of documentation, and the rights and potential for reuse. To oversimplify: libraries decide what to keep, museums decide what to accept, and archives decide what to throw away.

See (Tauberer 2014) for a history of the “civic hacking” and the open data movement.

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

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

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

(Kim et al, 2003).

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

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

[58][LIS] Libraries and bookstores use different classification systems. The kitchen in a restaurant is not organized like a home kitchen because professional cooks think of cooking differently than ordinary people do. Scientists use the Latin or binomial (genus + species) scheme for identifying and classifying living things to avoid the ambiguities and inconsistencies of common names, which differ across languages and often within different regions in a single language community.

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

[60][LIS] Many of the ancient libraries in Greece and Rome have been identified by archaeologists by characteristic architectural features (Casson 2002). See also (Battles 2003).

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

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

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

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

In contrast, Seattle imposes the grid in an abstract way, ignoring the fact that there are many lakes, rivers, and hills that break up the grid. Streets keep the same names even though they are not connected, and the grid stretches for many miles out from its origin in Seattle. You can be up in the mountains at the corner of “294th Avenue SE” and “472nd Street SE,” giving you precise information about your location and nearly 50 mile distance from downtown Seattle.

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

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

(Shiner 2007). The comparison of the organizing systems in casinos and airports comes from (Curran 2011). (Venturi 1972)

The number of queues, their locations and their layout (if spatial) is referred to as the “queue configuration.” The “queue discipline” is the policy for selecting the next customer from the queue. Most common discipline is “First come, first served.” Frequent, higher-paying, or some other customer segment might have their own queue with FCFS applied within it.

See the New York Post article at http://nypost.com/2013/05/14/rich-manhattan-moms-hire-handicapped-tour-guides-so-kids-can-cut-lines-at-disney-world/

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


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

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

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

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

(The Discipline of Organizing  Chapter 3 — Activities in Organizing Systems)
ic outcomes in physical storage as exceptions and not let them distort the choice of the organizing principles in the “logic tier.”

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

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

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


[79][Web] (Munk 2004).

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

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

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

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

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

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

Except when the resources on display are replicas of the originals, which is more common than you might suspect. Many nineteenth-century museums in the United States largely contained copies of pieces from European museums. Today, museums sometimes display replicas when the originals are too fragile or valuable to risk damage (Wallach 1998). Whether the “resource-based interaction” is identical for the replica and original is subjective and depends on how well the replica is implemented.

(Gibson 1977), (Norman 1988). See also (Norman 1999) for a short and simple explanation of Norman’s (re-)interpretation of Gibson.


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

See (Hearst 2009), (Buettcher et al. 2010).

(Glushko and Nomorosa 2013).

(Apte and Mason 1995) introduced this framework to analyze services rather than interactions per se.

Furthermore, many of the resources might not be available in the user’s own library and could only be obtained through inter-library loan, which could take days or weeks.

In contrast, far fewer interactions in museum collections are searches for known items, and serendipitous interactions with previously unknown resources are often the goal of museum visitors. As a result, few museum visitors would prefer an online visit to experiencing an original painting, sculpture, or other physical artifact. However, it is precisely because of the unique character of museum resources that museums allow access to them but do not allow visitors to borrow them, in clear contrast to libraries.

(Viswanadham 2002), (Madrigal 2009), (Prats et al. 2008). A video of a robot librarian in action at the University of Missouri, Kansas City is at http://www.youtube.com/watch?v=8wJIlTq7ts.

[97][LIS] Providing access to knowledge is a core mission of libraries, and it is worth pointing out that library users obtain knowledge both from the primary resources in the library collection and from the organizing system that manages the collection.

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

[99][Web] (Pirolli 2007).

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

[101][Mus] See (Simon 2010). An exemplary project to enhance museum access is Delphi (Schmitz and Black 2008), the collections browser for the Phoebe A. Hearst Museum of Anthropology at University of California, Berkeley. Delphi very cleverly uses natural language processing techniques to build an easy-to-use faceted browsing user interface that lets users view over 600,000 items stored in museum warehouses. Delphi is being integrated into Collection Space (http://www.collectionspace.org/), an open source web collections management system for museum collections, collaboratively being developed by University of California, Berkeley, Cambridge University, Ontario Academy of Art and Design (OCAD), and numerous museums.

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

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

[103][Com] Even sophisticated text representation formats such as XML have inherent limitations: one important problem that arises in complex management scenarios, humanities scholarship, and bioinformatics is that XML markup cannot easily represent overlapping substructures in the same resource (Schmidt 2009).
Digital books change the economics and *first sale* is not as well-established for digital works, which are licensed rather than sold (Aufderheide and Jaszi 2011). To protect their business models, many publishers are limiting the number of times ebooks can be lent before they “self-destruct.” Some librarians have called for boycotts of publishers in response (http://boycottharpercollins.com).

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

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

See Microsoft Windows Accessibility, Apple Accessibility, and Android Accessibility Features.

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

Accessible Rich Internet Applications (ARIA)

Smithsonian Guidelines for Accessible Exhibition Design

Because not every performance of a Broadway is exactly the same, the D-Scriptive audio snippets are tied to particular bits of dialog. The theater's stage manager triggers the D-Scriptive system to broadcast the corresponding visual explanations to audience members listening on earpieces. For example, in the Lion King a snippet might explain that “on the left are two giraffes and a cheetah.” (Giridharadas 2014)

In 2015 Netflix began a similar audio description service to accompany some of its original series. See http://blog.netflix.com/2015/04/netflix-begins-audio-description-for.html
For a recent historical and highly technical review of speech recognition written by some of the most prominent researchers in the field, see (Huang, Baker, and Reddy 2014). An easier to read story about Apple's Siri voice recognition program is (Geller 2012). Popup archive is https://www.popuparchive.org/ and its audio search service is https://www.audiosearch.org/.

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

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

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

This is the historical and dominant conception of the research library, but libraries are now fighting to prove that they are much more than just repositories because many of their users place greater value “on-the-fly access” of current materials. See (Teper 2005) for a sobering analysis of this dilemma.

Today the United States National Archives displays the Declaration of Independence, Bill of Rights, and the U.S. Constitution in sealed titanium cases filled with inert argon gas. Unfortunately, for over a century these documents were barely preserved at all; the Declaration hung on the wall at the United States Patent Office in direct sunlight for about 40 years.
Customer information drives day-to-day operations, but is also used in decision support and strategic planning.

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

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

Libraries are increasingly faced with the choice of providing access to digital resources through renewable licensing agreements, “pay per view” arrangements, or not at all. To some librarians, however, the failure to obtain permanent access rights “offends the traditional ideal of libraries” as memory institution (Carr 2010).

For example, students writing a term paper usually organize the printed and digital resources they rely on; the former are probably kept in folders or in piles on the desk, and the latter in a computer file system. This organizing system is not likely to be preserved after the term paper is finished. An exception that proves the rule is the task of paying income taxes for which (in the USA) taxpayers are legally required to keep evidence for up to seven years after filing a tax return (http://www.irs.gov/Businesses/Small-Businesses-&-Self-Employed/How-long-should-I-keep-records%3F).

(Arc) Many of those Word Perfect documents were stored on floppy disks because floppy disk drives were built into almost every personal computer for decades, but it would be hard to find such disk drives today. And even if someone with a collection of word processor documents stored of floppy disks in 1995 had copied those files to newer storage technologies, it is unlikely that the current version of the word processor would be able to read them. Software appli-
cation vendors usually preserve “backwards compatibility” for a few years with earlier versions to give users time to update their software, but few would support older versions indefinitely because to do so can make it difficult to implement new features.

Digital resources can be encoded using non-proprietary and standardized data formats to ensure “forward compatibility” in any software application that implements the version of the standard. However, if the ebook reader, web browser, or other software used to access the resource has capabilities that rely on later versions of the standards the “old data” will not have taken advantage of them.

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

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

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


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

[127][Arc] Brewster Kahle has been described as a computer engineer, Internet entrepreneur, Internet activist, advocate of universal access to knowledge, and digital librarian (http://en.wikipedia.org/wiki/Brewster_Kahle). In addition to websites, the Internet Archive preserves several million books, over a million pieces of video, 400,000 news programs from broadcast TV, over a million audio recordings, and over 100,000 live music concerts.

The Memento project has proposed a specification for using HTTP headers to perform “datetime negotiation” with the Wayback Machine and other archives of web pages, making it unnecessary for Memento to save anything on its own.
Memento is implemented as a browser plug-in to "browse backwards in time" whenever older versions of pages are available from archives that use its specification. (VandeSompel 2010).


(Brown and Duguid 2002).

(Savodnik 2011).

(Talukder 2016)

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

Company intranets, wikis, and blogs are often used as knowledge management technologies; Lotus Notes and Microsoft SharePoint are popular commercial systems. (See the case study in §12.2 Knowledge Management for a Small Consulting Firm (page 591).)

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

(The College of Physicians of Philadelphia’s Mütter Museum houses a novel collection of artifacts meant to "educate future doctors about anatomy and human medical anomalies." No museum in the world is like it; it contains display cases full of human skulls, abnormal fetuses in jars, preserved human bodies, a garden of medicinal herbs, and many other unique collections of resources.

However, one sub-collection best reflects the distinctive and idiosyncratic selection and arrangement of resources in the museum. Chevalier Jackson, a distinguished laryngologist, collected over 2,000 objects extracted from the throats of patients. Because of the peculiar focus and educational focus of this collection, and because there are few shared characteristics of "things people swallow that they should not," the characteristics and principles used to organize and describe the collection would be of little use in another organizing system. What other collection would include toys, bones, sewing needles, coins, shells, and dental material? It is hard to imagine that any other collection that would in-
clude all of these items plus fully annotated record of sex and approximate age of patient, the amount of time the extraction procedure took, the tool used, and whether or not the patient survived.

Curation is a very old concept whose Medieval meaning focused on the “preservation and cure of souls” by a pastor, priest, or “curate” (Simpson and Weiner 1989). A set of related and systematized curation practices for some class of resources is often called a curation system, especially when they are embodied in technology.

Information about which resources are most often interacted with in scientific or archival collections is essential in understanding resource value and quality.

In memory institutions, the most common job titles include “curator” or “conservator.” In for-profit contexts where “governance” is more common than “curation” job titles reflect that difference. In addition to “governance,” job titles often include “recordkeeping,” “compliance,” or “regulatory” prefixes to “officer,” “accountant,” or “analyst” job classifications.

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

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

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

Some popular “community content” sites like Yelp where people rate local businesses have been criticized for allowing positive rating manipulation. Yelp has also been criticized for allowing negative manipulation of ratings when competitors slam their rivals.
The resource might have been put someplace else when the site was reorganized or a new web server was installed. It is no longer the same resource because it will have another URI, even if its content did not change.

All of these terms refer to types of web resources or techniques whose purpose is to mislead people into doing things or letting things be done to their computers that will cost them their money, time, privacy, reputation, or worse. We know too well what spam is. Phishing is a type of spam that directs recipients to a fake website designed to look like a legitimate one to trick them into entering account numbers, passwords, or other sensitive personal information. Malware, fakeware, or spyware sites offer tempting downloadable content that installs software designed to steal information from or take control of the visiting computer. Keyword stuffing, spamdexing, and META tag abuse are techniques that try to mislead search engines about the content of a resource by annotating it with false descriptions. Link farms or scraper sites contain little useful or original content and exist solely for the purpose of manipulating search engine rankings to increase advertising revenue. Similarly, cybersquatters register domain names with the hope of profiting from the goodwill of a trademark they do not own.

(Brown 2009).

(Diaz 2005), (Grimmelmann 2009).

See video of Haber explaining how this works, Haber has recently been able to build a version of his scanning and image processing technology for use outside the laboratory that he calls Irene (Image, Reconstruct, Erase Noise, Etc.). (Cowen 2015) and (Wilkinson 2014)


For a cogent discussion of when and for what reasons weeding must take place in university libraries, see https://mrlibrarydude.wordpress.com/2014/03/12/why-we-weed-book-deselection-in-academic-libraries/.

A typical reaction when libraries discard books is described in (Jackman 2015)

(Blanchette and Johnson 2002)

(Neziroglu 2014) and (Lidz 2003).

Libraries and museums must also deal with long-term strategy, but the lesser visibility of library governance and museum governance might simply reflect the greater concerns about fraud and malfeasance in for-profit business contexts than in non-profit contexts and the greater number of standards or “best practices” for corporate governance. (Kim, Nofsinger, and Mohr 2009).
Data governance decisions are also often shaped by the need to conform to information or process model standards, or to standards for IT service management like the *Information Technology Infrastructure Library (ITIL)*. See [http://www.itil-officialsite.com/](http://www.itil-officialsite.com/).

In this context, these management and maintenance activities are often described as “IT governance” (Weill and Ross 2004). Data classification is an essential IT governance activity because the confidentiality, competitive value, or currency of information are factors that determine who has access to it, how long it should be preserved, and where it should be stored at different points in its lifecycle. (Turban et al. 2010).

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

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