Chapter 2. Activities in Organizing Systems

To appear in
***The Discipline of Organizing*,** 2012

Robert J. Glushko

Erik Wilde

Jess Hemerly

2.1 Four Activities in All Organizing Systems

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 to 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 in Chapter 1 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 can be informal for your clothes closet because its scope and scale are limited. In institutional organizing systems the selection, organizing, interaction design, and maintenance activities are often highly formal; they are deeply ingrained in academic curricula and professional practices, with domain-specific terms for their methods and results.

For example, libraries and museums usually make their selection principles explicit in **collection development** policies. Adding a resource to a library collection is called **acquisition**, but adding to a museum collection is called **accessioning.** Documenting the contents of library and museum collections to organize them is called **cataloguing**. **Circulation** is a central interaction in libraries, but because museum resources don't circulate the primary interactions for museum users are **viewing** or **visiting** the collection. Maintenance activities are usually described as **preservation** or **curation**.

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

These 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.

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.

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.

In Section 1.2.1 "The Concept of 'Resource'" and Section 1.3.2 "What Is Being Organized?" we briefly discussed the fundamental concept of resource. In this chapter, we describe the four primary activities with resources, using examples from many different kinds of organizing systems. 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 Chapters 3 and 4.

2.2 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) 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 10, "The Organizing System Road Map" we discuss these six aspects in more detail.

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. Selection methods and criteria vary across resource domains. Resource selection policies are often shaped by laws, regulations or policies that require or prohibit the collection of certain kinds of objects or types of information.[1]

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. Digitization is substantially changing how libraries select resources. Digital content can be delivered anywhere quickly and cheaply, making it easier for a group of cooperating libraries to share resources. For example, while each campus of the University of California system has its own libraries and library catalogs, system-wide catalogs and digital content delivery reduce the need for every campus to have any particular resource in its own collection.[2]

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.[3]

In the for-profit sector, well-run firms are similarly systematic in selecting the resources that must be managed and the information needed to manage them. The organizing systems for managing sales, orders, customers, inventory, personnel, and finance information are tailored to the specific information needed to run that part of the company's operations. Identifying this information is the job of business analysts and data modelers. Much of this operational data is combined in huge "data warehouses" to support the "business analytics" function in which novel combinations and relationships among data items are explored by selecting subsets of the data.[4]

Digitization has had extremely important impacts on the manner in which collections of information resources are created in information-intensive domains such as transportation, retailing, supply chain management, healthcare, energy management, and "big science" where a torrent of low-level information is captured from GPS devices, RFID tags, sensors and science labs. Businesses that once had to rely on limited historical data analysis and printed reports now have to deal with a constant stream of real-time information.

An analogous situation has evolved with personal collections of photographs. Less than two decades ago, before the digital camera became a consumer product, the time and expense of developing photographs induced people to take photos carefully and cautiously. Today the proliferation of digital cameras and photo-capable phones has made it so easy to take digital photos and videos that people are less selective and take many photos or videos of the same scene or event.

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).[5] When an information-intensive enterprise combines its internal services with outsourced ones provided by other firms, 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 enterprise's internal capabilities and those of its service partners better than any other combination of services could.[6]

The nature and scale of the web changes how we collect resources and fundamentally challenges how we think of resources in the first place. Web-based resources cannot be selected for a collection by consulting a centralized authoritative directory, catalog, or index because one does not exist. And although your favorite web search engine consults an index or directory of web resources when you enter a search query, you do not know where that index or directory came from or how it was assembled.[7]

The contents of a collection and how it is organized always reflect its intended users and uses. But the web has universal scope and global reach, making most of the web irrelevant to most people most of the time. Researchers have attacked this problem by treating the web as a combination of a very large number of topic-based or domain-specific collections of resources, and then developing techniques for extracting these collections as digital libraries targeted for particular users and uses.[8]

Even when the selection principles behind a collection are clear and consistent, they can be unconventional, idiosyncratic, or otherwise biased by the perspective and experience of the collector. This is sometimes the case in museum or library collections that began or grew opportunistically through the acquisition of private collections that reflect a highly individual point of view.

It is especially easy to see the collector's point of view in personal collections. 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?

2.3 Organizing Resources

Organizing systems arrange their resources according to many different principles. In libraries, museums, businesses, government agencies and other long-lived institutions these 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 not usually stated in any formal way and might even be 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?" (Section 1.3.3), "how much?" (Section 1.3.4), and "how?" (Section 1.3.6) questions posed in Chapter 1.

A simple principle for organizing resources is "co-location" – putting them in the same place. 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.[9]

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

This means that decisions about which resource properties will be used in organizing must often precede the creation or collection of the resources. This is especially critical for archeologists, naturalists, and scientists of every type. Without information about the context of creation or discovery, what might otherwise be important resources could be just a handful of pottery shards, a dead animal, or some random set of bits on a computer.

"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 properties like "aboutness" that are not directly perceived, methods for assigning subject classifications are intellectually-intensive and require rigorous training to be performed consistently and appropriately for the intended users. [10] The cost and time required for this human effort motivates many organizing systems for information resources to use computational approaches for arranging them.

2.3.1 Organizing Physical Resources

When the resources being arranged are physical or tangible things – such as books, paintings, animals, or cooking pots – any resource can only 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.[11]

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 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.

2.3.1.1 Organizing with Properties of Physical Resources

Physical resources are often organized according to intrinsic physical properties like their size, color or shape, or intrinsically associated properties such as the place and time they were created or discovered. The shirts in your clothes closet might be arranged by color, by fabric, or style. We can view dress shirts, T-shirts, Hawaiian shirts and other styles as configurations of shirt properties that are so frequent and familiar that they have become linguistic and cultural categories. Other people might think about these same properties or categories differently, using a greater or lesser number of colors or ordering them differently, sorting the shirts by style first and then by color, or vice versa.

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

Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users. Hazardous or flammable materials should not be stored where they might spill or ignite; lions and antelopes should not share the same zoo habitat or the former will eat the latter; and adult books and movies should not be kept in a library where children might accidentally find them. 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.

2.3.1.2 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 the 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.

2.3.2 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.[12] 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.[13]

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.[14] [15]

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.[16]

There are inherently more choices in the arrangement 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 extrinsic or behavioral properties like subject classifications, assigned names or identifiers, or even access frequency.[17]

Complex organization and interactions are possible when organizing systems with digital resources are based on the datatype or data model of the digital content (e.g., text, numeric, multimedia, statistical, geospatial, logical, scientific, or personnel data). These distinctions are often strongly identifiable with business functions: operational, transactional and process control activities require the most fine-grained data, while strategic functions might rely on more qualitative analyses represented in narrative text formats. Managerial and decision support functions might require a mixture of digital content types.

With digital resources we don't have to worry about hazardous resources blowing up or one resource eating another, although viruses, worms and other malevolent agents present threats to digital resources as dire as those faced by the zoo antelopes if lions are kept too close. Accordingly, 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.

2.3.2.1 Organizing Web-based Resources

The Domain Name System is the most inherent scheme for organizing web resources, and top-level domains for generic resource categories (.com, .edu. .org, gov, etc.) and countries provide some clues about the resources organized by a web site. These clues are most reliable for large established enterprises and publishers; we know what to expect at ibm.com, Berkeley.edu, and jstor.org.[18]

The network of hyperlinks among web resources challenges the notion of a collection, because it makes it impossible to define a precise boundary around any collection smaller than the complete web.[19] 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.[20]

The conventional library is both a collection of books and the physical space in which the collection is managed. On the web, rich hyperlinking 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 some kind of local control to be effective. The spectacular rise and fall 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.[21] But Facebook so far is succeeding by following a walled garden strategy.

2.3.3 Organizing with Multiple Resource Properties

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

If multiple resource properties are considered in a fixed order, the resulting arrangement forms a **logical hierarchy**. The top level categories of resources are created based on the values of the property evaluated first, and then each category is further subdivided using other properties until each resource is classified in only a single category. A typical example of hierarchical arrangement for digital resources is the system of directories or folders used by a professor to arrange his personal document collection in a computer file system; 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, an additional step of mapping categories to physical locations is required; for example, resources in the category "kitchen utensils" 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.

Another way to understand faceted classification is that it 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.

Chapter 7, "Classification," explains principles and methods for hierarchical and faceted classification in more detail.

2.4 Designing Resource-Based Interactions

There would be no point in selecting and organizing resources if they could not be accessed or interacted with in some way. Organizing systems vary a great deal in the types of resource-based interactions they enable and in the nature and extent of access they allow.

It is essential to distinguish the interactions that are designed into and directly supported by an organizing system 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 are directly designed into the library. 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. Note that some of these interactions might be explicitly supported in an organizing system containing digital books, as in Google's search engine where language translation is a supported service.

Users have direct access to original resources in a collection when they browse through library stacks or wander in museum galleries.[22] They have mediated or indirect access when they use catalogs or search engines, and sometimes they can only interact with copies or descriptions of the resources.

2.4.1 Affordance and Capability

The concept of **affordance**, introduced by J.J. Gibson and then extended and popularized by Don 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.[23]

When organizing resources involves arranging physical resources using boxes, bins, cabinets, or shelves, the affordances and the implications for access and use are immediately evident. 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. Overhead and end-of-aisle signs support navigation and orientation in libraries and stores, and the information on book spines or product packages help us select a specific resource.

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. When you implement the organizing system for your clothes closet, you implicitly consider the tradeoff between extensive and minimal organization and the implications for the amount of interaction effort required to put away and find clothes in each case.

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. For information systems, capability can be assessed by counting their 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.[24]

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. Doors that open automatically when their sensors detect an approaching person do not need handles. 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 current awareness service that automatically informs you about relevant news from many sources makes it unnecessary to search any of them separately. Similarly, 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.[25]

2.4.2 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 or disaggregated from the value created by physical actions and person-to-person encounters. [26]

2.4.1.1 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. 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 examining the primary resource.[27]

However, for most people the primary purpose of interacting with a library is to access the information contained in its resources.. For most people access in a digital library to copies of printed documents or books is equivalent to or even better than access to the original physical resource because the incidental physical and interpersonal interactions have been eliminated. [28]

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 "librarian robot" that can locate books and grasp them from the shelves shows promise.[29]

2.4.1.2 Value Creation with Digital Resources

With digital resources, neither physical manipulation nor interpersonal contact is required for interactions, and the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.[30] Put another way, by replacing interactions that involve people and physical resources with symbolic ones, organizing systems can lower their costs without reducing user satisfaction. This is why so many businesses have automated their information-intensive processes with self-service technology like ATMs, websites, or smartphone apps.

Similarly, web search engines eliminate the physical effort required to visit a library and enables 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 descriptions that libraries have always been known for because it enables rapid evaluation and iterative query refinement based on inspection of the primary resources.[31]

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. But 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.[32] Some libraries are experimenting with Semantic Web and "Linked Data" technologies that would integrate their extensive bibliographic resources with resources on the open web. But there is insufficient agreement about exactly how libraries should expose their collections some ambivalence about whether to do it at all.[33]

There seems to be less ambivalence for museums, which 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.[34]

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 "e-books" demonstrate how access to content can be enhanced once it is no longer tied to the container of the printed book, but some e-book formats have a limited interaction repertoire: typically only "page turning," resizing, and full-text search. [35]

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 (Extensible Markup Language), 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.[36]

Businesses that create or own their information resources can readily take advantage of the enhanced interactions that digital formats enable. For libraries, however, copyright is often a barrier to digitization, both as a matter of law and because digitization enables copyright enforcement to a degree not possible with physical resources. As a result, digital books are somewhat controversial and problematic for libraries, whose access models were created based on the economics of print publication and the social contract of the copyright first sale doctrine that allowed libraries to lend printed books.[37]

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

2.4.3 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 the rare documents in a special collection might be accessible only 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 that further distinguishes them according to the nature of their interactions with resources. For example, administrative assistants in a company's Human Resources department are not allowed to see employee salaries; HR employees in a benefits administration role can see the 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.[38]

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 the 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 kitchens or to take a book from a bookshelf. In our online lives, however, we readily accept and impose more granular access controls on our personal computers and in the applications we use, as when we 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: (1) 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; (2) giving more access to paying users than to users who don't pay; (3) 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, 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 or security 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 VPN software that controls remote access to the library network.[39] 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 often 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.

2.5 Maintaining Resources

Maintaining resources is an important activity in every organizing system regardless of the nature of its collection because resources or surrogates for them 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.

Different domains sometimes use the same terms to describe different maintenance activities and different terms for similar activities. The most common terms 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 and more closely overlap in meaning; we will distinguish them in section 2.5.2-2.5.4.

Ideally, maintenance requirements for resources should be anticipated when organizing principles are defined and implemented. In particular, resource descriptions to support long-term preservation of digital resources are important. [40]

2.5.1 Motivations for Maintaining Resources

The concept of "memory institution" broadly applies to a great many organizing systems that share the goal of preserving knowledge and cultural heritage. The primary resources in libraries, museums, data archives or other "memory institutions" are fixed cultural, historic, or scientific artifacts that are maintained because they are unique and original items with future value. This is why the Louvre preserves the portrait of the Mona Lisa and the United States National Archives preserves the Declaration of Independence.[41]

In contrast, in the organizing systems used by businesses 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 (such as information about a customer) are often used by more than one part of the business.[42] Maintaining the accuracy and consistency of changing resources is a major challenge in business organizing systems.[43]

Other 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 library, 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.[44]

Of course, libraries, museums, and archives also confront economic issues as they seek to preserve and maintain their collections and themselves as memory institutions. They view their collections as intrinsically valuable in ways that firms generally do not. 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.[45]

2.5.2 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 offsite 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 the essence of an artifact is retained if it is converted to a digital format?

2.5.2.1 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 from kilobytes to megabytes to gigabytes to terabytes to petabytes, 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.[46]

The second challenge might seem paradoxical. Even as the capacities of digital storage technologies increase 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 between printed and digital resources is striking; books on library shelves don't disappear if no one uses them, but digital data can be lost just because no one wants access to it within a year or two after its creation.[47]

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.[48]

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. Dealing with interoperability and integration problems will be discussed further in Chapter 9, "Interactions in Organizing Systems."

2.5.2.2 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 web sites change by adding content, and others change by editing or removing it. [49]

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.[50] Nevertheless, the continued existence of a particular web page is hardly sufficient to preserve it if it not popular and relevant enough to show up in the first few pages of search results. Persistent access requires preservation, but preservation isn't 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 don't reveal any details about how frequently they update their collections of indexed pages.[51]

THE INTERNET ARCHIVE AND THE "WAYBACK MACHINE"

The Internet Archive (archive.org) makes preservation of the web its first and foremost activity, and when you enter a URL into its "Wayback Machine" you can see what a site looked like at different moments in time. For example, www.berkeley.edu was archived over 1400 times between October 1996 and June 2011, including about once a week on average during all of 2010. 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. The Internet Archive has recently launched the "Archive-It" service to enable schools, libraries and other public institutions to archive collections of digital resources.[52]

2.5.2.2 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.[53]

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 eighteenth 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.[54]

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.[55]

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.

2.5.2.3 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.[56]

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** are a type of business organizing system whose goal is to capture and systematize these information resources.[57] 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.

When businesses implement information-intensive processes employing web-based services, it is highly desirable to organize them as a collection of service types rather than service instances because this makes them more robust and maintainable. An abstract description of services or resources allows one service provider to transparently substitute for another. For example, the user of the organizing system that implements an Internet-based retail business model need not know and probably doesn't care which delivery service carries out a request to deliver a package from a warehouse. Similarly, an abstract service description might allow a computational process to substitute for one carried out by a person, or vice versa. For example, a credit card terminal in a restaurant offers the customer the capability to specify no tip, a specific amount, or calculating a percentage of the total.

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

Even when zoos give their popular animals individual names, it seems logical that the zoo's goal is to preserve animal species rather than instances because any particular animal has a finite lifespan and cannot be preserved forever.[58]

2.5.2.4 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 data set 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.[59]

At other times 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.[60]

2.5.3 Curation

For almost a century “curation” has been used to describe 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.[61] 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.[62]

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 web site can make decisions and employ technology that maintains the contents, quality and character of the site over time. In that respect web site curation and governance practices are little different than those for the organizing systems in memory institutions or business enterprises. The key 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.

2.5.3.1 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.[63]

Resource descriptions are more important in company Intranets than in the open web because the contents of the former lack the links that are critical in the latter.

2.5.3.2 Individual Curation

Curation by individuals has been studied a great deal in the research discipline of Personal Information Management.[64] 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.[65]

2.5.3.3 Social and Web Curation

Many individuals spend a great time of time curating their own web sites, 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.[66] The massive scale of these bottom-up and distributed activities is curation by "crowdsourcing" the continuously aggregated actions and contributions of users.[67]

The informal and organic "folksonomies" that result from their aggregated effort create organization and authority through network effects.[68] 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.[69] In addition, professional curators are not pleased to have the ad hoc work of untrained people working on web sites described as curation.

Most web sites 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.[70]

Nevertheless, problems like these 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 (See Sidebar).

WEB RESOURCES WITH BAD INTENT

"Spam," "phishing," "malware," "fakeware," "spyware," "keyword stuffing," "spamdexing," "META tag abuse,""link farms", "cybersquatters," "phantom sites"… Unfortunately 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 don't own.

ISPs, security software firms, email services, and search engines are engaged in a constant war against these kinds of malicious resources and techniques.[71]

Since we cannot prevent these deceptions by controlling what web resources are created in the first place, we have to respond to them after the fact with "defensive curation" techniques. These 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.

2.5.3.4 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 web site, 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.

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 are hassled every time you want to fly on a commercial plane.

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 a sensor that reports the temperature or current location. 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 in response to a query. But popularity and relevance don't always ensure accuracy. We can easily find popular pages that prove the existence of UFOs or claim to validate wacky conspiracy theories.

Furthermore, 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 web sites 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.[72]

2.5.4 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 the curation of the resources 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 different types of 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. 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.[73]

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 might soon restrict the information collected during web browsing.[74]

2.5.4.1 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 in the organizing system to ensure that the activities are performed in an effective and consistent manner.[75]

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.[76] A management aspect of governance in this domain is determining which resources and information will potentially provide economic or competitive advantages and determining which will not. A conceptual and technological aspect of governance is determining how best to organize the useful resources and information in business operations and information systems to secure the potential advantages.

Business intelligence is only as good as the data it is based on, which makes business data governance a critical concern that has rapidly developed its own specialized techniques and vocabulary. The most fundamental governance activity in information-driven businesses is identifying the "master data" about customers, employees, materials, products, suppliers, etc. that is reused by different business functions and is thus central to business operations.[77]

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 enable 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.[78]

2.5.4.2 Governance in Scientific Organizing Systems

Scientific data poses special governance problems because of its enormous scale, which dwarfs the data sets managed in most business organizing systems. A scientific data collection might contain tens of millions of files and 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.[79]

2.7 Key Points in Chapter Two

- Selection, organizing, interaction design, and maintenance activities occur in every organizing system.
- These activities are not identical in every domain, but the general terms enable communication and learning about domain-specific methods and vocabularies.
- The most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized.
- Even when the selection principles behind a collection are clear and consistent, they can be unconventional, idiosyncratic, or otherwise biased.
- Most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole.
- Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users.
- Digital organizing systems can support collections and interactions at a scale that is impossible in organizing systems that are entirely physical.
- 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.

- We focus on the interactions that are that are designed into and directly supported by an organizing system because of intentional acts of description or arrangement.
- 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 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.
- With digital resources, the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.
- 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.
- Preservation of resources means maintaining them in conditions that protect them from physical damage or deterioration.
- Preservation is often a key motive for digitization, but digitization alone is not preservation.
- Comprehensive web search engines 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.
- 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.
- Personal Information Management has been influenced for decades by a seminal article written by Vannevar Bush titled “As We May Think.”
- 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.
- Scientific data poses special governance problems because of its enormous scale.

---

[1] [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 archeological sites (Merryman, 2009). For digital resources, privacy laws prohibit the collection or misuse of personally identifiable information about healthcare, education, telecommunications, video rental, and might soon restrict the information collected during web browsing.

[2] [LIS] See Borgman (2000). 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 a perspective that emphasizes creation and preservation of physical collections to facilitating the use and creation of knowledge regardless of the medium of its representation and the physical or virtual location from which it is accessed.

[3] [LIS] 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 US much controversy exists because museums contain many human skeletal remains and artifacts that Native American groups want to be "repatriated." 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.

[4] [Citation] On data modeling: see Kent (1978/2000), Silverston (2000), Glushko & McGrath (2005). For data warehouses see Turban et al, (2010).

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

[6] [Business]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 can't 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 don't need to be good at doing.

[7] [Computing] (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).

[8] [Computing]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).

[9] [CogSci] 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., somethjng is red or not, round or not, and so on). If multiple values are possible, the "property" is called an "attribute," "dimension," or "variable." See Barsalou and Hale (1993) for a rigorous contrast between feature lists and other representational formalisms in models of human categories.

[10] [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 binominal (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.

[11] [Citation] Battles (2003).

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

[13] [Computing]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.

[14] [Computing]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, 2008). 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.

[15] [LIS] An emerging issue in the field of digital humanities (Schreibman, Siemens, and Unsworth, 2005) is the requirement to recognize the **materiality** of the environment that enables people to create and interact with digital resources (Leonardi, 2010). Even if the resources themselves are intangible, it can be necessary to study and preserve the technological and social context in which they exist to fully understand them. For example, a "Born-Digital Archives" program at Emory University is preserving a collection of the author Salmon Rushdie's work that includes his four personal computers and an external hard drive (Kirschenbaum, 2008; Kirschenbaum et al, 2009).

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

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

[18] [Computing]The Domain Name System or 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.

[19] [Computing][19] 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/offline.html), but many people consider it unreliable and subject to strange side effects.

[20] [Citation] (Aalbersberg and Kahler, 2011).

[21] [Citation] (Munk, 2004).

[22] [LIS] 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.

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

[24] [Citation] See Hearst (2009), Buttcher et al(2010).

[25] [Citation] Glushko and Nomorosa (2012)

[26] [Business]Apte and Mason (1995) introduced this framework to analyze services rather than interactions per se. They paid special attention to services where the value created by symbolic manipulation or information exchange could be completely separated or disaggregated from the value created by person-to-person interactions. 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.

[27] [LIS] 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.

[28] [LIS] In addition, many of the interactions in libraries are searches for known items, and this function is easily supported by digital search. 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.

29 [Citation] (Viswanadham, 2002; Madrigal 2009). (Prats et al 2008).

30 [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.

31 [LIS] 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.

32 [Citation] (Pirolli, 2007).

33 [Citation] (Byrne and Goddard, 2010).

34 [Citation] See (Simon, 2011). 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 UC 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 CollectionSpace (http://www.collectionspace.org/), an open source web collections management system for museum collections, collaboratively being developed by UC Berkeley, Cambridge University, Ontario Academy of Art and Design, and numerous museums.

35 [Computing] 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 the features of detected faces to the features of previously identified faces, and encourages people to tag photos to make the recognition more accurate. Some online use similar image classification techniques to bring together shoes, jewelry, or other items that look alike.

36 [Computing]However, 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).

37 [Law] Digital books change the economics and first sale is not as well-established for digital works, which are licensed rather than sold (Aufderheide and Jaszi, 2011). To protect their business models, many publishers are limiting the number of times e-books 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, 2009; Robinson et al, 2009). And we have made this book available as an open content repository so that it can be collaboratively maintained and customized.

[38] [Business]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).

[39] [LIS] 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 e-book 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 e-books (Howard, 2011).

[40] [Citation] (Guenther and Wolfe, 2009).

[41] [LIS] Today the United States National Archives displays the Declaration of Independence, Bill of Rights, and 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.

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

[43] [Computing]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.

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

[45] [CogSci] For example. students writing a term paper usually organize the printed and digital resources they rely on; the former are probably kept in folders or in piles on the desk, and the latter in a computer file system. This organizing system is not likely to be preserved after the term paper is finished. An exception that proves the rule is the task of paying income taxes for which (in the US) taxpayers are legally required to keep evidence for up to seven years after filing a tax return (Internal Revenue Service, 2011).

[46] [Citation] (Rothenberg, 1995).

[47][Citation] (Pogue, 2009). (Smith, 2000, p. 74).

[48] [Computing] 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 application 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 e-book reader, web browser, or other software used to access the resource has capabilities that rely on later versions of the standards the "old data" won't have taken advantage of them.

[49] [Computing]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 the changes 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.

[50] [Citation] (Fretterly et al, 2003)

[51] [Computing] However, when a web site 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.

[52] [Computing]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, 2011).

[53] [Computing] But people might still enjoy the many Mona Lisa parodies and recreations. See http://www.megamonalisa.com, http://www.oddee.com/item_96790.aspx, http://www.chilloutpoint.com/art_and_design/the-best-mona-lisa-parodies.html

[54] [Citation] (Brown and Duguid, 2000).

[55] [Citation] (Savodnik, 2011).

[56] [Computing]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.

[57] [Business]Company intranets, wikis, and blogs are often used as knowledge management technologies; Lotus Notes and Microsoft SharePoint are popular commercial systems.

[58] [Business] 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.

[59] [Citation] (Manyika et al, 2011). (Gable et al, 2008).

[60] [LIS] 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 include 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.

[61][LIS] 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, 2009). 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.

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

[63] [LIS] 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.

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

[65] [Citation] Bush 1945.

[66] [Citation] (Howe, 2008).

[67] [LIS] 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 (**CITE**). Some museums and scientific data repositories also encourage voluntary curation to analyze and classify specimens or photographs (Wright, 2010).

[68] [Citation] Trant 2009b

[69] [Business] 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.

[70] [Computing] 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.

[71] [Citation] (Brown, 2009).

[72] [Citation] (Diaz, 2008; Grimmelmann, 2009).

[73] [Citation] (Kim, Nofsinger, and Mohr, 2009).

[74] [Computing] 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, 2011).

[75][Business] 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.

[76][Citation] (Turban et al, 2010)

[77] [Computing]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).

[78] [Citation](Cox et al, 2007).

[79] [Law] Recently imposed requirements by the National Science Foundation, National Institute of Health 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).