Chapter 3. Resources in Organizing Systems

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

Robert J. Glushko
Daniel D. Turner
Kimra McPherson
Jess Hemerly

3.1 Introduction to Resources

This chapter builds upon the foundational concepts introduced in Chapter 1 to explain more carefully what we mean by resource. In particular, we focus on the issue of identity – what will be treated as a separate resource – and discuss the issues and principles we need to consider when we give each resource a name or identifier.

In Section 3.2 we introduce four distinctions we can make when we discuss resources: **domain**, **format**, **agency**, and **focus**. In Section 3.3 we apply these distinctions as we discuss how resource identity is determined for physical resources, bibliographic resources, resources in information systems, as well as for active resources and "smart things." Section 3.4 then tackles the problems and principles for naming: once we have identified resources, how do we name and distinguish them? Finally, Section 3.5 considers issues that emerge with respect to resources over time.

3.1.1 What is a Resource?

Resources are what we organize.

We introduced the concept of "resource" in Section 1.2.1 with its ordinary sense of "anything of value that can support goal-oriented activity" and emphasized that a group of resources can be treated as a "collection" in an organizing system. And what do we mean by "anything of value," exactly? It might seem that the question of **identity**, of what a single resource is, shouldn't be hard to answer. After all, we live in a world of resources, and finding, selecting, describing, arranging, and referring to them are everyday activities.

Nevertheless, even when the resources we are dealing with are tangible things, how we go about organizing them is not always obvious, or at least not the same obvious to each of us at all times. Not everyone thinks of them in the same way. Recognizing something in the sense of perceiving it as a tangible thing is only the first step toward being able to organize it and other resources like it. Which properties garner our attention, and which we use in organizing depends on our experiences, purposes, and context.

We add information to a resource when we name or describe it; it then becomes more than "it." We can describe the same resource in many different ways. At various times we can consider any given resource to be one of many members of a broad category, as one of the few members of a narrow category, or as a unique instance of a category with only one member. For example, we might recognize something as a piece of clothing, as a sock, or as the specific dirty sock with the hole worn in the heel from yesterday's long hike. However, even after we categorize something, we might not be careful how we talk about it; we often refer to two objects as "the same thing" when what we mean is that they are "the same type of thing." Indeed, we could debate whether a category with only one possible member is really a category, because it blurs an important distinction between particular items or instances and the class or type to which they belong.

The issues that matter and the decisions we need to make about resource instances and resource classes and types are not completely separable. Nevertheless, we will strive to focus on the former ones in this chapter and the latter ones in Chapter 6, "Categories: Describing Resource Classes and Types."

3.1.1.1 Resources with Parts

As tricky as it can be to decide what a resource is when you are dealing with single objects, it is even more challenging when the resources are objects or systems composed of other parts. In these cases, we must focus on the entirety of the object or system and treat it as a resource, treat its constituent parts as resources, and deal with the relationships between the parts and the whole, as we do with engineering drawings and assembly procedures.

How many things is a car? If you're imagining the car being assembled you might think of several dozen large parts like the frame, suspension, drive train, gas tank, brakes, engine, exhaust system, passenger compartment, doors, and other pre-assembled components. Of course, each of those components is itself made up of many parts – think of the engine, or even just the radio. Some sources have counted ten or fifteen thousand parts in the average car, and but even at that precise granularity a lot of parts are still complex things. There are screws and wires and fasteners and on and on; really too many to count.

This ambiguity about the number of parts holds for information resources too; a newspaper can be considered a single resource but it might also consist of multiple sections, each of which contains separate stories, each of which has many paragraphs, and so on. From the typesetter's point of view, each character in a sentence can be taken as a distinct resource, selected from a font of similar resources.

3.1.1.2 Bibliographic Resources, Information Components, and "Smart Things" as Resources

Information resources generally pose additional challenges in their identification and description because their most important property is usually their content, which is not easily and consistently recognizable. Organizing systems for information resources in physical form, like those for libraries, have to juggle the duality of their tangible embodiment with what is inherently an abstract information resource; that is, the printed book versus the knowledge the book contains. Here the organizing system emphasizes description resources or surrogates like bibliographic records that describe the information content, rather than, their physical properties.

Another important question in libraries is what set of resources should be treated as the same work because they contain essentially similar intellectual or artistic content. We may talk about Shakespeare's play "Macbeth," but what is this thing we call "Macbeth"? Is it a particular string of words, saved in a computer file or handwritten upon a folio? Is it the collection of words printed with some predetermined font and pagination? Are all the editions and printings of these words the same "Macbeth"? How should we organize the numerous live and recorded performances of plays and movies that share the "Macbeth" name? What about works based on or inspired by "Macbeth" that do not share the title

"Macbeth," like the Kurosawa film "Kumonosu-jo" ("Throne of Blood") that transposes the plot to feudal Japan?

Information system designers and architects face analogous design challenges when they describe the "information components" in business or scientific organizing systems. Information content is intrinsically merged or confounded with structure and presentation whenever it is used in a specific instance and context. From a logical perspective, an order form contains information components for ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION, but the arrangement of these components, their type font and size, and other non-semantic properties can vary a great deal in different order forms and even across a single information system that repurposes these components for letters, delivery notices, mailing labels, and database entries.[1]

Similar questions about resource identity are posed by the emergence of ubiquitous or pervasive computing, in which information processing capability and connectivity are embedded into physical objects, in devices like smart phones, and in the surrounding environment. Equipped with sensors, radio-frequency identification (RFID) tags, GPS data, and user-contributed metadata, these "smart things" create a jumbled torrent of information about location and other properties that must be sorted into identified streams and then matched or associated with the original resource.

Section 3.3 discusses the issues and methods for determining "what is a resource?" for physical resources as well as for the bibliographic resources, information components and "smart things" discussed here in Section 3.1.1.1.

3.1.2 Identity, Identifiers, and Names

The answer to the question "What is a resource?" has two parts. The first part is identity: what thing are we treating as the resource? The second part is identification: differentiating between this single resource and other resources like it. These problems are closely related. Once you've decided what to treat as a resource, you create a name or an identifier so that you can refer to it reliably. A **name** is a label for a resource that is used to distinguish one from another. An **identifier** is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. For a digital resource, its identifier serves as the input to the system or function that determines its location so it can be retrieved, a process called **resolving** the identifier or **resolution**.

Choosing names and identifiers—be it for a person, a service, a place, a trend, a work, a document, a concept, etc.—is hardly straightforward. In fact, naming can often be challenging and is often highly contentious. Naming is made difficult by countless factors, including the audience that will need to access, share, and use the names, the limitations of language, institutional politics, and personal and cultural biases.

A common complication arises when a resource has more than one name or identifier. When something has more than one name each of the multiple names is a **synonym** or **alias**. A particular physical instance of a book might be called a hardcover or paperback or simply a text. George Furnas and his research collaborators called this issue of multiple names for the same resource or concept the "**vocabulary problem.**"[2]

Whether we call it a book or a text, the resource will usually have a Library of Congress catalog number as well as an ISBN as an identifier. When the book is in a carton of books being shipped from the publisher to a bookstore or library, that carton will have a bar-coded tracking number assigned by the delivery service, and a manifest or receipt document created by the publisher whose identifier associates the shipment with the customer. Each of these identifiers is unique with respect to some established scope or context.

A partial solution to the vocabulary problem is to use a **controlled vocabulary**. We can impose rules that standardize the way in which names and labels for resources are assigned in the first place. Alternatively, we can define mappings from terms used in our natural language to the authoritative or controlled terms. However, vocabulary control can't remove all ambiguity. Even if a passport or national identity system requires authoritative full names rather than nicknames, there could easily be more than one Robert John Smith in the system.

Controlling the language used for a particular purpose raises other questions: Who writes and enforces these rules? What happens when organizing systems that follow different rules get compared, combined, or otherwise brought together in contexts different from those for which they were originally intended?

3.2 Organizing How to Think About Resources

The nature of the resource is critical for the creation and maintenance of quality organizing systems. There are four distinctions we make in discussing resources: **domain**, **format**, **agency**, and **focus**.

3.2.1 Resource Domain

Every resource has some essence or type that distinguishes it from other resources, which we call the resource **domain**. Domain is an intuitive notion that we can help define by contrasting it with the alternative of ad hoc or arbitrary groupings of resources that just happen to be in the same place at some moment, rather than being based on natural or intrinsic characteristics.

For physical resources domains can be coarsely distinguished according to the type of matter of which they are made using properties that can be readily perceived. All languages and cultures make basic contrasts between animal, vegetable, or material substances and then make further distinctions to create a hierarchical system of domain categories. Many aspects of this system of domain categories are determined by natural constraints on category membership that are manifested in patterns of shared properties; once a resource is identified as a member of one category it must also be a member of another with which it shares some but not all properties. For example, a marble statue in a museum must also be a kind of material resource, and a fish in an aquarium must also be a kind of animal resource.

For information resources, easily perceived properties are less reliable and correlated, so we more often distinguish domains based on semantic properties; the definitions of the "encyclopedia," "novel," and "invoice" resource types distinguish them according to their typical subject matter, or the type of content, rather than according to the great variety of physical forms in which we might encounter them. Arranging books by color or size might be sensible for very small collections, or in a photo studio, but organizing according to physical properties would make it extremely impractical to find books in a large library.

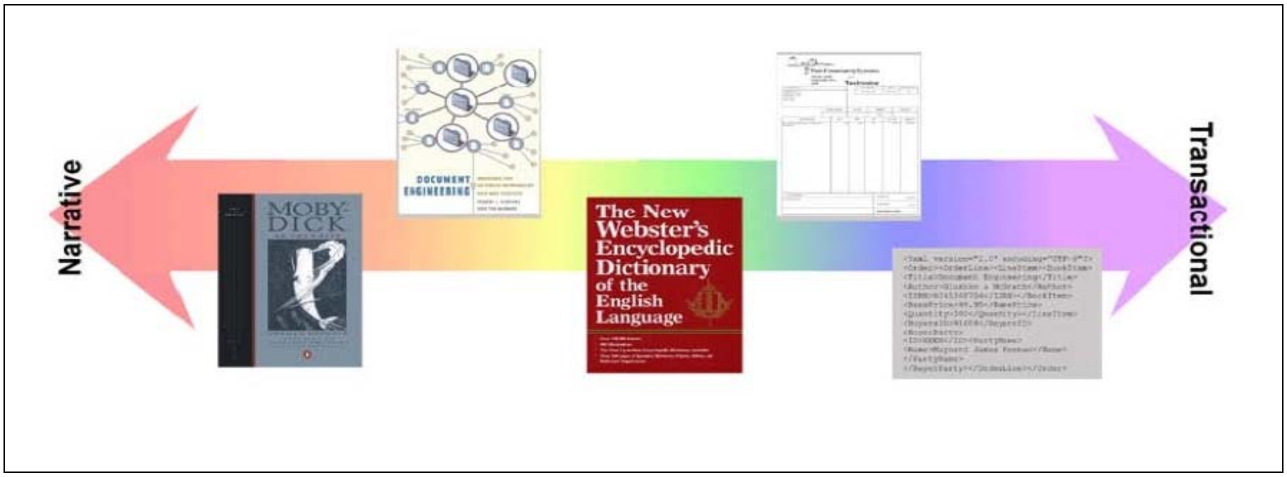
We can arrange types of information resources in a hierarchy but because the category boundaries are not sharp it is more useful to view domains of information resources on a continuum from weakly-structured narrative content to highly structured transactional content. This framework, called the Document Type Spectrum by Glushko and McGrath, captures the idea that the boundaries between resource domains, like those between colors in the rainbow, are easy to see for colors far apart in the spectrum but hard to see for adjacent ones.[3] See the Sidebar, "The Document Type Spectrum," and its corresponding figure.

THE "DOCUMENT TYPE SPECTRUM"

Different domains or types of documents can be distinguished according to the extent to which their content is semantically prescribed, by the amount of internal structure, and by the correlations of their presentation and formatting to their content and structure. These three characteristics of content, structure, and presentation vary systematically from narrative document types like novels to transactional document types like invoices.

Narrative types are authored by people and are heterogeneous in structure and content, and their content is usually just prose and graphic elements. Their presentational characteristics carefully reinforce their structure and semantics; for example, the text of titles or major headings is large because the content is important, in contrast to the small text of footnotes. Transactional document types are usually created mechanically and, as a result, are homogeneous in structure and content; their content is largely "data" -- strongly typed content with precise semantics that can be processed by computers.

In the middle of the spectrum are hybrid document types like textbooks, encyclopedias, and technical manuals that contain a mixture of narrative text and structured content in figures, data tables, code examples, and so on.



3.2.2 Resource Format

Information resources can exist in numerous formats with the most basic format distinction being whether the resource is physical or digital. This distinction is most important when it comes to the implementation of a resource storage or preservation system because that is where physical properties are usually considerations, and very possibly constraints. This distinction is less important at the logical level when we design interactions with resources because it is often possible to use digital surrogates for the physical resources to overcome the constraints posed by their physical properties. When we search for cars or appliances in an online store it doesn't matter where the actual cars or appliances are located or how they are physically organized.

Many digital representations can be associated with either physical or digital resources, but it is important to know which one is the original or primary resource, especially for unique or valuable ones.

Today a great many resources in organizing systems are **born digital**. They are created in word processors and digital cameras, or by audio and video recorders. Other resources are produced in digital form by the many types of sensors in "smart things" and by the systems that create digital resources when they interact with barcodes, QR ("quick response") codes, RFID tags, or other mechanisms for tracking identity and location.[4]

Other digital resources are created by **digitization**, the process for transforming an artifact whose original format is physical so that it can be stored and manipulated by a computer. We can digitize the printed word, photographs, blueprints and record albums. Printed text, for example, can be digitized by scanning the pages and employing character recognition software or simply by re-typing it.[5]

There are a vast number of digital formats. The simplest digital format for "plain text" documents typically consists of only the characters that you see on your computer keyboard; the alphanumerics and symbols with which we are all by now familiar. Most document formats also explicitly encode a hierarchy of structural components, such as chapters, sections or semantic components like descriptions or procedural steps, and sometimes the appearance of the rendered or printed form.[6] Another important distinction to note is whether the information is encoded as a sequence of text characters so that it is human readable as well as computer readable. Text formats such as EBCDIC, ASCII and UniCode offer progressively modern character encoding formats in common use today. Encoding character content with XML, for example, allows for layering of intentional coding or **markup** interwoven with the "plain text" content. The most complex digital formats are those for multimedia resources and multidimensional data, where the data format is highly optimized for specialized analysis or applications.[7]

Digitization of non-text resources such as film photography, drawings, and analog audio and visual recordings raises a complicated set of choices about pixel density, color depth, sampling rate, frequency filtering, compression, and numerous other technical issues that determine the digital representation.[8] There may be multiple intended uses and devices for the digitized resource that might require different digitization approaches and formats. Moreover, downstream users of digitized resources often need to know the format in which the digital artifact has been created so they can reuse it as is or process it in other ways.

Some digital formats support interactions that are qualitatively different and more powerful than those possible with physical resources. Museums are using virtual world technology to create interactive exhibits in which visitors can fly through the solar system, scan their own bodies, and change gravity so they can bounce off walls. Sophisticated digital document formats can enable interactions with annotated digital images or video, 3-D graphics or embedded data sets. The Google Art Project contains extremely high resolution photographs of famous paintings that make it possible to see details that are undetectable under the normal viewing conditions in museums.[9]

Nevertheless, digital representations of physical resources can also lose important information and capabilities. The distinctive sounds of hip hop music produced by "scratching" vinyl records on turntables cannot be produced from digital MP3 music files.

Copyright often presents a barrier to digitization, both as a matter of law and because digitization itself enables copyright enforcement to a degree not possible prior to the advent of digitization, by eliminating common forms of access and interactions that are inherently possible with physical printed books like the ability to give or sell them to someone else.[10]

3.2.3 Resource Agency

Agency, the extent to which a resource can initiate actions on its own is the third distinction we make about a resource. Another way to express this contrast is between passive resources that are acted upon and active resources that can initiate actions. Telephone answering and fax machines are agents because they are capable of independently responding to an outside stimulus, accepting and managing messages. An ordinary mercury thermometer is not capable of communicating its own reading, but a digital wireless thermometer or "weather station" can. Passive resources serve as nouns or operands, while active resources serve as verbs or operants.[11]

3.2.4.1 Passive or Operand Resources

Organizing systems that contain passive or operand resources are ubiquitous for the simple reason that we live in a world of physical resources that we identify and name in order to interact with them. Passive resources are usually tangible and static and thus they become valuable only as a result of some action or interaction with them.

Most organizing systems with physical resources or those that contain resources that are digitized equivalents treat those resources as passive. A printed book on a library shelf, a digital book in an e-book reader, a statue in a museum gallery, or a case of beer in a supermarket refrigerator only create value when they are checked out, viewed, or consumed. None of these resources exhibits any agency and cannot initiate any actions to create value on their own.

3.2.4.2 Active or Operant Resources

Active resources create effects or value on their own, sometimes when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, or web-based services. We can exploit computing capability, storage capacity and communication bandwidth to create active resources that can do things and support interactions that are impossible for ordinary physical passive resources.

Objects become active resources when they contain sensing or communication capabilities. RFID chips, which are essentially bar codes with built-in radio transponders, enable automated location tracking and context sensing. RFID receivers are built into store shelves, loading docks, parking lots, and toll booths to detect when some RFID-tagged resource is at some meaningful location. RFID tags can be made "smarter" by having them record and transmit information from sensors that detect temperature, humidity, acceleration, and even biological contamination.[12]

Smart phones are also active resources that can identify and share their own location, orientation, acceleration and a growing number of other contextual parameters to enable personalization of information services. Self-regulating appliances are active resources when they communicate with each other in a "smart building" to minimize energy consumption.

Many organizing systems on the web consist of collections or configurations of active digital resources. Interactions among these active resources often implement information-intensive business models where value is created by exchanging, manipulating, transforming, or otherwise processing information, rather than by manipulating, transforming, or otherwise processing physical resources.

"Service Oriented Architecture" (SOA) is an emerging design discipline for organizing active resources as functional business components that can be combined in different ways. SOA is generally implemented using web services that exchange XML documents in real-time information flows to interconnect the business service components.

A familiar design pattern for an organizing system composed from active digital resources is the "online store." The store can be analyzed as a composition or choreography in which some web pages display catalog items, others serve as "shopping carts" to assemble the order, and then a "checkout" page collects the buyer's payment and delivery information that gets passed on to other service providers who process payments and deliver the goods.

The web has enabled the novel application of human resources as active resources to carry out tasks of short duration that can be precisely described but which can't be done reliably by computers. These tasks include image classification or annotation, spoken language transcription, and sentiment analysis. The people doing these tasks over the web are sometimes called "Mechanical Turks" by analogy to a fake chess playing machine from the 18th century that had a human hidden inside who was secretly moving the pieces. [13]

3.2.4 Resource Focus

A fourth contrast between types of resources distinguishes primary or original resources from resources that describe them. Any primary resource can have one or more description resources associated with it to facilitate finding, interacting with, or interpreting the primary one. Description resources are essential in organizing systems where the primary resources are not under its control and can only be accessed or interacted with through the description. Description resources are often called **metadata**.

The distinction between primary resources and description resources, or metadata, is deeply embedded in library science and traditional organizing systems whose collections are predominantly text resources like books, articles, or other documents. In these contexts description resources are commonly called bibliographic resources or catalogs, and each primary resource is typically associated with one or more description resources.

In business enterprises, the organizing systems for digital information resources, such as business documents, or data records created by transactions or automated processes, almost always employ resources that describe, or are associated with, large sets or classes of primary resources.[14]

The contrast between primary resources and description resources is very useful in many contexts, but when we look more broadly at organizing systems, it is often difficult to distinguish them, and determining which resources are primary and which are metadata is often just a decision about which resource is currently the **focus** of our attention.

For example, many people who use Twitter focus on the 140-character message body as the primary resource, while the associated metadata about the sender and the message (is it a forward, reply, link, and so on?) is less important to them. However, for firms in the growing ecosystem of services that use Twitter metadata to measure sender and brand impact, identify social networks, and assess trends, the focus is on the metadata, not the message content.[15]

As another example, the players on professional sports teams are human resources that we enjoy watching as they compete, but millions of people participate in fantasy sports leagues where teams consist of fantasy players that are simulated resources based on the statistics generated by the actual human players. Put another way, the associated resources in the actual sports are treated as the primary ones in the fantasy leagues.

3.2.5 Resource Format x Focus

Applying the format contrast between physical and digital resources to the focus distinction between primary and descriptive resources yields a useful framework with four categories of resources.

3.2.5.1 Physical description resource for a primary physical resource

The oldest relationship between descriptive resources and physical resources is when descriptions or other information about physical resources are themselves encoded in a physical form. Nearly ten thousand years ago in Mesopotamia small clay tokens kept in clay containers served as inventory information to count units of goods or livestock. It took 5000 years for the idea of stored tokens to evolve into Cuneiform writing in which marks in clay stood for the tokens and made both the tokens and containers unnecessary.[16]

3.2.5.2 Digital description resource for a primary physical resource

Here the digital resource describes a physical resource. The most familiar example of this relationship is the online library catalog used to find the shelf location of physical library resources, which beginning in the 1970s replaced the physical cards with database records. The online catalogs for museums usually contain a digital photograph of the painting, item of sculpture, or other museum object that each catalog entry describes.

Digital description resources for primary physical resources are essential in supply chain management, logistics, retailing, transportation, and every business model that depends on having timely and accurate information about where things are or about their current states. This digital description resource is created as a result of an interaction with a primary physical resource like a temperature sensor or with some secondary physical resource that is already associated with the primary physical resource like an RFID tag, barcode, or two-dimensional QR code.

Augmented reality systems combine a layer of real-time digital information about some physical object to a digital view or representation of it. The yellow “first down” lines superimposed in broadcasts of football games are a familiar example. Augmented reality techniques that superimpose identifying or descriptive metadata have been used in displays to support the operation or maintenance of complex equipment, in smartphone navigation and tourist guides, in advertising, and in other domains where users might otherwise need to consult a separate information source. Advanced airplane cockpit technology includes heads-up displays that present critical data based on available instrumentation, including augmented reality runway lights when visibility is poor because of clouds or fog.

3.2.5.3 Digital description resource for a primary digital resource
Here the digital resource describes a digital resource. This is the relationship in a digital library or any web-based organizing system and it makes it possible to access the primary digital resource directly from the digital secondary resource.

3.2.5.4 Physical description resource for a primary digital resource
This is the relationship implemented when we encounter an embedded QR barcode in newspaper or magazine advertisements, on billboards, sidewalks, t-shirts, or on store shelves. Scanning the QR code with a mobile phone camera can launch a web site that contains information about a product or service, place an order for one unit of the pointed-to- item in a web catalog, dial a phone number, or initiate any other application or service identified by the QR code.[17]

3.3 Resource Identity
Determining the identity of resources that belong in a domain, deciding which properties are important or relevant to the people or systems operating in that domain, and then specifying the principles by which those properties encapsulate or define the relationships among the resources are the essential tasks when building any organizing system. In organizing systems used by individuals or with small scope, the methods for doing these tasks are often ad hoc and unsystematic, and the organizing systems are therefore idiosyncratic and do not scale well. At the other extreme, organizing systems designed for institutional or industry-wide use, especially in information-intensive domains, require systematic design methods to determine which resources will have separate identities and how they are related to each other. These resources and their relationships are then described in conceptual models which then are used to guide the implementation of the systems that manage the resources and support interactions with them.[18]

3.3.1 Identity and Physical Resources
Our human visual and cognitive systems do a remarkable job at picking out objects from their backgrounds and distinguishing them from each other. In fact, we have little difficulty recognizing an object or a person even if we're seeing them from a novel distance and viewing angle or with different lighting, shading, and so on. When we watch a football game, we don't have any trouble perceiving the players moving around the field, and their contrasting uniform colors allow us to see that there are two different teams.

The perceptual mechanisms that make us see things as permanent objects with contrasting visible properties are just the prerequisite for the organizing tasks of identifying the specific object, determining the categories of objects to which it belongs, and deciding which of those categories is appropriate to emphasize. Most of the time we carry out these tasks in an automatic, unconscious way; at other times we make conscious decisions about them. For some purposes we consider a sports team as a single resource, as a collection of separate players for others, as offense and defense, as starters and reserves, and so on.[19]

Although we have many choices about how we can organize football players, all of them will include the concept of a single player as the smallest identifiable resource. We are never going to think of a football player as an intentional collection of separately identified leg, arm, head, and body resources because there are no other ways to "assemble" a human from body parts. Put more generally, there are some natural constraints on the organization of matter into parts or collections based on sizes, shapes, materials, and other properties that make us identify some things as indivisible resources in some domain.

3.3.2 Identity and Bibliographic Resources

Pondering the question of identity is something relatively recent in the world of librarians and catalogers. Libraries have been around for about 4000 years, but until the last few hundred years librarians created "bins" of headings and topics to organize resources without bothering to give each individual item a separate identifier or name. This meant searchers first had to make an educated guess as to which bin might house their desired information—"Histories"? "Medical and Chemical Philosophy"?—then scour everything in the category in a quest for their desired item. The choices were ad hoc and always local—that is, each cataloger decided the bins and groupings for each catalog.[20]

The first systematic approach to dealing with the concept of identity for bibliographic resources was developed by Antonio Panizzi at the British Museum in the mid-19th century. Panizzi wondered: How do we differentiate similar objects in a library catalog? His solution was a catalog organized by author name with an index of subjects, along with his newly concocted *Rules for the Compilation of the Catalogue*. This contained 91 rules about how to identify and arrange author names and titles and what to do with anonymous works. The Rules were meant to codify how to differentiate and describe each singular resource in his library. Taken together, the rules serve to group all the different editions and versions of a work together under a single identity.[21]

The concept of identity for bibliographic resources was refined in the 1950s by Lubetzky, who enlarged the concept of "the work" to make it a more abstract idea of an author's intellectual or artistic creation. According to Lubetzky's principle, an audio book, a video recording of a play, and an electronic book should be listed each as distinct items, yet still linked to the original because of their overlapping intellectual origin.[22]

The distinctions put forth by Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy between the abstract **work**, an **expression** in multiple formats or genres, a particular **manifestation** in one of those formats or genres, and a specific physical **item**. The broad scope from the abstract work to the specific item is essential because organizing systems in libraries must organize tangible artifacts while expressing the conceptual structure of the domains of knowledge represented in their collections.

If we revisit the question "What is this thing we call Macbeth?" we can see how different ways of answering fit into this abstraction hierarchy. The most specific answer is that "Macbeth" is a specific **item**, a very particular and individual resource, like that dog-eared paperback with yellow marked pages that you owned when you read "Macbeth" in high school. A more abstract answer is that "Macbeth" is an idealization called a **work**, a category that includes all the plays, movies, ballets, or other intellectual creations that share a recognizable amount of the plot and meaning from the original Shakespeare play.

This hierarchy is defined in the Functional Requirements for Bibliographical Records (FRBR), published as a standard by the International Federation of Library Associations and Institutions (IFLA).[23]

3.3.3 Identity and Information Components

In information-intensive domains, documents, databases, software applications, or other explicit repositories or sources of information are ubiquitous and essential to the creation of value for the user, reader, consumer, or customer. Value is created through the comparison, compilation, coordination or transformation of information in some chain or choreography of processes operating on information flowing from one information source or process to another. These processes are employed in accounting, financial services, procurement, logistics, supply chain management, insurance underwriting and claims processing, legal and professional services, customer support, computer programming, and energy management.

The processes that create value in information-intensive domains are "glued together" by shared information components that are exchanged in documents, records, messages, or resource descriptions of some kind. Information components are the primitive and abstract resources in information-intensive domains. They are the units of meaning that serve as building blocks of composite descriptions and other information artifacts.

The value creation processes in information-intensive domains work best when their component parts come from a common controlled vocabulary for components, or when each uses a vocabulary with a granularity and semantic precision compatible with the others. For example, the value created by a personal health record emerges when information from doctors, clinics, hospitals, and insurance companies can be combined because they all share the same "patient" component as a logical piece of information.

This abstract definition of information components doesn't help identify them, so we'll introduce some heuristic criteria: An "information component" can be (1) Any piece of information that has a unique label or identifier or (2) Any piece of information that is self-contained and comprehensible on its own.[24]

These two criteria for determining the identity of information components are often easy to satisfy through observations, interviews, and task analysis because people naturally use many different types of information and talk easily about specific components and the documents that contain them. Some common components (e.g., person, location, date, item) and familiar document types (e.g., report, catalog, calendar, receipt) can be identified in almost any domain. Other components need to be more precisely defined to meet the more specific semantic requirements of narrower domains. These smaller or more fine-grained components might be viewed as refined or qualified versions of the generic components and document types, like course grade and semester components in academic transcripts, airport codes and flight numbers in travel itineraries and tickets, and drug names and dosages in prescriptions.

Decades of practical and theoretical effort in conceptual modeling, relational theory, and database design have resulted in rigorous methods for identifying information components when requirements and business rules for information can be precisely specified. For example, in the domain of business transactions, required information like item numbers, quantities, prices, payment information, and so on must be encoded as a particular type of data—integer, decimal, Unicode string, etc.— with clearly defined possible values and that follows clear occurrence rules.[25]

Identifying components can seem superficially easy at the transactional end of the Document Type Spectrum (see Sidebar in Section 3.2.1), with orders or invoices, forms requiring data entry, or other highly-structured document types like product catalogs, where pieces of information are typically labeled and delimited by boxes, lines, white space or other presentation features that encode the distinctions between types of content. For example, the presence of ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION labels on the fields of an online order form suggests these pieces of information are semantically distinct components in a retail application. They follow the "self-contained and comprehensible" heuristic enough to interconnect the order management, payment, and delivery services that work together to carry out the transaction. In addition, these labels might have analogues in variable names in the source code that implements the order form, or as tags in a XML document created by the ordering application; <CustName>John Smith</CustName> and <Item>A-19</Item> in the order document can be easily identified when it is sent to the other services by the order management application.

But the theoretically grounded methods for identifying components like those of relational theory and normalization that work for structured data do not strictly apply when information requirements are more qualitative and less precise at the narrative end of the Document Type Spectrum. These information requirements are typical of narrative, unstructured and semi-structured types of documents, and information sources like those often found in law, education, and professional services. Narrative documents include technical publications, reports, policies, procedures and other less structured information, where semantic components are rarely labeled explicitly and are often surrounded by text that is more generic. Unlike transactional documents that depend on precise semantics because they are used by computers, narrative documents are used by people, who can ask if they aren't sure what something means, so there is less need to explicitly define the meaning of the information components. Occasional exceptions, such as where components in narrative documents are identified with explicit labels like NOTE and WARNING, only prove the rule.

3.3.4 Identity and Active Resources and "Smart Things"

Active resources (Section 3.4.3.2) initiate effects or create value on their own. In many cases an inherently passive physical resource like a product package or shipping pallet is transformed into an active one when it associated with an RFID tag or bar code. Mobile phones contain device or subscriber IDs so that any information they communicate can be associated both with the phone and often, through indirect reference, with a particular person. If the resource has an IP address, it is said to be part of the "Internet of Things." [26]

Organizing systems that create value from active resources often co-exist with or complement organizing systems that treat its resources as passive. In a traditional library, books sat passively on shelves and required users to read their spines to identify them. Today, some library books contain active RFID tags that make them dynamic information sources that self-identify by publishing their own locations. Similarly, a supermarket or department store might organize its goods as physical resources on shelves, treating them as passive resources; while superimposed on that traditional organizing system is one that uses point-of-sale transaction information created when items are scanned at checkout counters to automatically re-order goods and replenish the inventory at the store where they were sold. In some stores the shelves contain sensors that continually "talk to the goods" and the information they gather can maintain inventory levels and even help prevent theft of valuable merchandise by tracking goods through a store or warehouse. The inventory becomes a collection of active resources; each item eager to announce its own location and ready to conduct its own sale.

Blogjects—objects that blog—and Tweetjects—objects that post messages to Twitter—are neologisms for active resources that are plugged into the social web. Blogjects don't write editorial commentary about their experiences, but they use APIs and customized programs to harness the information captured by sensors and RFID that then appears on blogs in the form of human-readable maps, charts, and text. [27]

Tweetjects are sensors that send information about measurements or events to a Twitter account. For example, Sparkfun Electronics sells a kit consisting of a soil sensor that sends information about the water level in the soil through an Arduino circuit board, converting thresholds to Twitter messages like, "Please water me, I'm thirsty!" [28]

The extent to which an active resource is "smart" depends on how much computing capability it has available to refine the data it collects and communicates. A large collection of sensors can transmit a torrent of captured data that requires substantial processing to distinguish significant events from those that reflect normal operation, and also from those that are statistical outliers with strange values caused by random noise. This challenge gets qualitatively more difficult as the amount of data grows, because a one in million event might be a statistical outlier that can be ignored, but if there are a thousand similar outliers in a billion sensor readings, this cluster of data probably reveals something important. On the other hand, giving every sensor the computing capability to refine its data so that it only communicates significant information might make the sensors too expensive to deploy.[29]

3.4 Naming Resources

Determining the identity of the thing, document, information component, or data item we need isn't always enough. We often need to give that resource a name, a label that will help us understand and talk about what it is. But naming isn't just the simple task of assigning a sequence of characters. In this section, we'll discuss why we name, some of the problems with naming, and the principles that help us name things in useful ways.

3.4.1 What's In a Name?

When a child is born, its parents give it a name, often a very stressful and contentious decision. Names serve to distinguish one person from another. Names also, intentionally or unintentionally, suggest characteristics or aspirations. The name given to us at birth is just one of the names we will be identified with during our lifetimes. We have nicknames, names we use professionally, names we use with friends, and names we use online. Our banks, our schools, and our governments will know who we are because of numbers they associate with our names. As long as it serves its purpose to identify you, your name could be anything.[30]

Resources other than people need names so we can find them, describe them, reuse them, refer or link to them, record who owns them, and otherwise interact with them. In many domains the names assigned to resources are also influenced or constrained by rules, industry practice, or technology considerations.

3.4.2 The Problems of Naming

Giving names to anything, from a business to a concept to an action, can be a difficult process and it is possible to do it well or do it poorly. The following section details some of the major challenges in assigning a name to a resource.

3.4.2.1 The Vocabulary Problem

Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept. The words people choose to name or describe things are embodied in their experiences and context, so people will often disagree in the words they use. Moreover, people are often a bit surprised when it happens, because what seems like the natural or obvious name to one person isn't natural or obvious to another.[31]

Back in the 1980s in the early days of computer user interface design, George Furnas and his colleagues at Bell Labs conducted a set of experiments to measure how much people would agree when they named some resource or function. The short answer: very little. Left to our own devices, we come up with a shockingly large number of names for a single common thing.

In one experiment, a thousand pairs of people were asked to "write the name you would give to a program that tells about interesting activities occurring in some major metropolitan area." Less than 12 pairs of people agreed on a name. Furnas called this phenomenon "the vocabulary problem," concluding that no single word could ever be considered the "best" name. [32]

3.4.2.2 Homonymy, Polysemy, and False Cognates

Sometimes the same word can refer to different resources—a "bank" can be a financial institution or the side of a river. When two words are spelled the same but have different meanings they are **homographs;** if they are also pronounced the same they are **homonyms**. If the different meanings of the homographs are related, they are called **polysemes**.

Resources with homonymous and polysemous names are sometimes incorrectly identified, especially by an automated process that can't use common sense or context to determine the correct referent. Polysemy can cause more trouble than simple homography because the overlapping meaning might obscure the misinterpretation. If one person thinks of a "shipping container" as being a cardboard box and orders some of them, while another person thinks of a "shipping container" as the large box carried by semi-trailers and stacked on cargo ships, their disagreement might not be discovered until the wrong kinds of containers arrive.[33]

Many words in different languages have common roots, and as a result are often spelled the same or nearly the same. This is especially true for technology words; for example, "computer" has been borrowed by many languages. The existence of these cognates and borrowed words makes us vulnerable to false cognates. When a word in one language has a different meaning and refers to different resources in another, the results can be embarrassing or disastrous. "Gift" is poison in German; "pain" is bread in French.

3.4.2.3 Names with Undesirable Associations

False cognates are a special category of words that make poor names, and there are many stories relating product marketing mistakes, where a product name or description translates poorly, into other languages or cultures, with undesirable associations.[34] Furthermore, these undesirable associations differ across cultures. For example, even though floor numbers have the straightforward purpose to identify floors from lowest to highest levels, most buildings in Western cultures skip the 13th floor because many people think 13 is an unlucky number. In many East and Southeast Asian buildings, the 4th floor is skipped. In China the number 4 is dreaded because it sounds like the word for "death," while 8 is prized because it sounds like the word for "wealth."

While it can be tempting to dismiss unfamiliar biases and beliefs about names and identifiers as harmless superstitions and practices, their implications are ubiquitous and far from benign. Alphabetic ordering might seem like a fair and non-discriminatory arrangement of resources, but because it is easy to choose the name at the top of an alphabetical list, many firms in service businesses select names that begin with "A," "AA," or even "AAA" (look in any printed service directory). A consequence of this bias is that people or resources with names that begin with letters late in the alphabet are systematically discriminated against because they are often not considered, or because they are evaluated in the context created by resources earlier in the alphabet rather than on their own merit.[35]

3.4.2.4 Names that Assume Impermanent Attributes

Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.

Web resources are often referred to using URLs that contain the domain name of the server on which the resource is located, followed by the directory path and file name on the computer running the server. This treats the current location of the resource as its name, so the name will change if the resource is moved. It also means that resources that are identical in content, like those at an archive or mirror web site, will have different names than the original even though they are exact copies. An analogous problem is faced by restaurants or businesses with street names or numbers in their names if they lose their leases or want to expand.[36]

Some dynamic web resources that are generated by programs have URLs that contain information about the server technology used to create them. When the technology changes, the URLs will no longer work.[37]

Other resources have names that include page numbers, which disappear or change when the resource is accessed in a digital form. For example, the standard citation format for legal opinions uses the page number from the printed volume issued by West Publishing, which has a virtual monopoly on the publishing of court opinions and other types of legal documents.[38]

Some resources have names that contain dates, years or other time indicators, most often to point to the future. The film studio named "20th Century Fox" took on that name in the 1930s to give it a progressive identity, but today a name with "20th Century" in it does the opposite because it looks backward in time.[39]

3.4.2.5 The Semantic Gap

Another naming problem can arise when names are assigned by automated processes in ways that are conceptually different than how people do it. The difference in conceptual perspective in resource naming and description has been called the **semantic gap**.[40]

The semantic gap is largest when computer programs or sensors obtain and name some information in a format optimized for efficient capture, storage, decoding, or other technical criteria. The names — like IMG20268.jpg on a digital photo — might make sense for the camera as it stores consecutively taken photos but they are not good names for people. We may prefer names that describe the content of the picture, like "goldengatebridge.jpg."

And if we try to examine the content of computer-created or sensor-captured resources, like a clip of music or a compiled software program, a human-language text rendering of the content simply looks like nonsense. It was designed to be interpreted by a computer program, not by a person.

3.4.3 Choosing Good Names and Identifiers

If someone tells you they are having dinner with their best friend, a cousin, someone with whom they play basketball, and their professional mentor from work, how many places at the table will be set? Anywhere from two to five; it's possible all those relational descriptions refer to a single person, or to four different people, and because "friend," "cousin," "basketball teammate" and "mentor" don't name specific people you'll have to guess who is coming to dinner.

If instead of descriptions you're told that the dinner guests are Bob, Carol, Ted, and Alice, you can count four names and you know how many people are having dinner. But you still can't be sure exactly which four people are involved because there are many people with those names.

The uncertainty is completely eliminated only if we use identifiers for the people rather than names. **Identifiers** are names that refer unambiguously to a specific person, place, or resource because they are assigned in a controlled way Identifiers are often created as strings of numbers or letters rather than words to avoid the biases and associations that words can convey. For example, in some universities professors grade final exams that are identified with student numbers rather than names so that grades are assigned without the bias that could arise if the professor knows the student.

The distinction between names and identifiers for people is often not appreciated. See the Sidebar, "Names {and, or, vs} Identifiers."

NAMES {AND, OR, VS} IDENTIFIERS

People change their names for many reasons: when they get married or divorced, because their name is often mispronounced or misspelled, to make a political or ethnic statement, or because they want to stand out. A few years a football player with a large ego named Chad Johnson, which is the second most common surname in the US, decided to change his name to his player number of 85, becoming Chad "Ochocinco." He has an ochocinco.com web site and uses the ochocinco name on Facebook and Twitter. In a bit of irony, when Ochocinco wanted to put Ocho Cinco on the back of his football jersey, the football league would not let him because his legal name doesn't have a space in it.

A similar name change with an unintended consequence was that of the American singer and musician Price Rogers Nelson, who adopted the stage name of Prince and released numerous highly successful record albums under that name. But because Prince wasn't a legal name change, the record label trademarked it for marketing purposes, which led to disputes about control. In response, "The Artist Formerly Known as Prince" invented a graphical symbol that merged the symbols for male and female. Unfortunately, even though it is a unique identifier, this symbol isn't represented in any standard character set, so it can't be printed here and can't be searched for on the web.

Some minor league sports teams have replaced the player names on jerseys with twitter handles, which might be a good thing if their fans are into social media, but it must be strange for the announcers at the games when they say "@ifuentes4 just scored a goal."

When you go to coffee shops, you are often asked your name, which the cashier writes on the empty cup so that your drink can be identified after the barrista makes it. They don't actually need your name; just as some establishments use a receipt number to distinguish orders, what they need is an identifier. So even if your name is Joe, you can tell them it is Thor, Wotan, Mercurio, El Greco, Clark Kent, or any other name that is likely to be a unique identifier for the minute it takes to make your beverage.[41]

3.4.3.1 Make Names Informative

The most basic principle of naming is to choose names that are informative, which makes them easier to understand and remember. It is easier to tell what a computer program or XML document is doing if it uses names like "ItemCost" and "TotalCost" rather than just "I" or "T". People will enter more consistent and reusable address information if a form asks explicitly for "Street," "City," and "PostalCode" instead of "Line1" and "Line2."

Identifiers can be designed with internal structure and semantics that conveys information beyond the basic aspect of pointing to a specific resource. An International Standard Book Number like "ISBN 978-0-262-07261-8" identifies a resource (07261="Document Engineering") and also reveals that the resource is a book (978), in English (0), and published by MIT Press (262).[42]

The navigation points that mark intersections of radial signals from ground beacons or satellites that are crucial to aircraft pilots used to be meaningless five-letter codes. These identifiers were changed to make them suggest their locations, making them semantic landmarks that made pilots less likely to enter the wrong names into navigation systems, For example, some of the navpoints near Orlando, Florida - the home of Disney World – are MICKI, MINEE, and GOOFY.[43]

3.4.3.2 Reduce Synonymy and Homonymy with Controlled Vocabularies

One way to encourage good names for a given resource domain or task is to establish a **controlled vocabulary**. A controlled vocabulary can be thought of as a fixed or closed dictionary that includes all the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy and eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use. Controlled vocabularies are applied in many organizing systems, from bibliographic languages that determine the ways books are catalogued in a library to business languages that define the set of information components that can be used in transactional documents.

A controlled vocabulary isn't simply a set of allowed words; it also includes their definitions and often specifies rules by which the vocabulary terms can be used and combined. Different domains can create specific controlled vocabularies for their own purposes, but the important thing is that the vocabulary be used consistently throughout that domain.[44]

For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as **authority control**. When evaluating what name to use for an author, librarians typically look for the name form that's used most commonly across that author's body of work while conforming to rules for handling prefixes, suffixes and other name parts that often cause name variations. For example, a name like that of Johann Wolfgang von Goethe might be alphabetized as both a "G" name and a "V" name, but using "G" is the authoritative way. "See" and "see also" references then map the variations to the authoritative name. Similar rules are followed for identifying the authoritative form of titles when multiple translations and editions exist.[45]

Official authority files are maintained for many resource domains: a gazetteer associates names and locations and tells us whether we should be referring to Bombay or Mumbai; the Domain Name System maps human-oriented domain and host names to their IP addresses; the Chemical Abstracts Service Registry assigns unique identifiers to every chemical described in the open scientific literature; numerous institutions assign unique identifiers to different categories of animal species.[46]

In some cases, authority files are created or maintained by a community, as in the case of MusicBrainz, an "open music encyclopedia" to which users contribute information about artists, releases, tracks, and other aspects of music. Music metadata is notoriously unreliable; one study found over 100 variations in the description of the "Knockin' on Heaven's Door" song (written by Bob Dylan) as recorded by Guns N' Roses.[47]

3.4.3.3 Allow Aliasing

A controlled vocabulary is extremely useful to people who use it, but if you are designing an organizing system for other people who do not or cannot use it, you need to accommodate the variety of words they will actually use when they seek or describe resources. The authoritative name of a certain fish species is *Amphiprion ocellaris*, but most people would search for it as "clownfish," "anemone fish," or even by its more familiar film name of "Nemo."

Furnas suggests "unlimited aliasing" to connect the uncontrolled or natural vocabularies that people use with the controlled one employed by the organizing system. By this he means that there must be many alternate access routes to each word or function that a user is trying to find. For example, the birth name of the 42nd US President is "William Jefferson Clinton," but web pages that refer to him as "Bill Clinton" are vastly more common, and searches for the former are redirected to the latter. A related mechanism used by search engines is spelling correction, essentially treating all the incorrect spellings as aliases of the correct one ("did you mean California?" when you typed "Claifornia").

3.4.4.4 Make Identifiers Unique or Qualified

Even though an identifier refers to a single resource, this doesn't mean that no two identifiers are identical. One military inventory system might use stock number 99 000 1111 to identify a 24-hour, cold-climate ration pack, while another inventory system, the same number could be used to identify an electronic radio valve. Each identifier is unique in its inventory system, but if a supply request gets sent to the wrong warehouse hungry soldiers could be sent radio valves instead of rations. [48] [49]

We can prevent or reduce identifier collisions by adding information about the **namespace**, the domain from which the names or identifiers are selected, thus creating what are often called **qualified names**. There are several dozen US cities named "Springfield" and "Washington," but adding state codes to mail addresses distinguishes them. Likewise, we can add prefixes to XML element names when we create documents that reuse components from multiple document types, distinguishing <book:Title> from <legal:Title>.

We can fix problems like these by qualifying or extending the identifier, or by creating a **globally unique identifier** (or GUID), one that will never be the same as another identifier in any organizing system anywhere else. One easy method to create a GUID is to use a URL you control and append a string to it, the same approach that gives every web page a unique address. GUIDs are often used to identify software objects, the resources in distributed systems, or data collections. [50]

Because they aren't created by an algorithm whose results are provably unique, we do not consider fingerprints, or other biometric information, to be globally unique identifiers for people, but for all practical purposes they are.[51]

3.4.4.5 Distinguish Identifying and Resolving

Library call numbers are identifiers that do not contain any information about where the resource can be found in the library stacks on in a digital repository. This separation enables this identification system to work when there are multiple copies in different locations, in contrast to URLs that serve as both identifiers and locations much of the time. When the identifier does not contain information about resource location, we need a way to interpret or **resolve** it to determine the location. With physical resources, resolution takes place with the aid of signs, maps, or other associated resources that describe the arrangement of resources in some physical environment; for example, "you are here" maps have a list of its buildings and associate each with a coordinate or other means of finding it on the map.. With digital resources, the resolver is a directory system or service that interprets an identifier and looks up its location or directly initiates the retrieval of the resource.

3.5 Resources Over Time

Problems of "what is the resource?" and "how do we identify it?" are complex and often require ongoing work to ensure they are properly answered as the content and context of an organizing system evolves. As a result, we might need to know how a resource does or does not change over time (its **persistence**), whether its state and content come into play at a specified point in time (its **effectivity**), whether the resource is what it is said to be (its **authenticity**), and sometimes who has certified its authenticity over time (its **provenance**).

3.5.1 Persistence

Even if you have reached an agreement as to the meaning of "a thing" in your organizing system, you still face the question of the identity of the resource over time, or its **persistence**.

3.5.1.1 Persistent Identifiers

How long must an identifier last? Coyle gives the conventional, if unsatisfying, answer: "As long as it's needed".[52] In some cases, the time frame is relatively short. When you order a specialty coffee and the barrista asks for your name, this identifier only needs to last until you pick up your order at the end of the counter. But other time frames are much longer. For libraries and repositories of scientific, economic, census, or other data the timeframe might be "forever."

The design of a scheme for persistent identifiers must consider both the required time frame and the number of resources to be identified. When the Internet Protocol was designed in 1980, it contained a 32-bit address scheme, sufficient for over 4 billion unique addresses. But the enormous growth of the Internet and the application of IP addresses to resources of expected types have required a new addressing scheme with 12 bits.[53]

Recognition that URLs are often not persistent as identifiers for web-based resources led the Association of American Publishers (AAP) to develop the Digital Object Identifier (DOI) system. The location and owner of a digital resource can change, but its DOI is permanent.[54]

3.5.1.2 Persistent Resources

Even though persistence often has a technology dimension, it is more important to view it as a commitment by an institution or organization to perform activities over time to ensure that a resource is available when it is needed. Put another way, preservation (section 2.3.2) and governance (section 2.3.3) are activities carried out to ensure the outcome of persistence.

The subtle relationship between preservation and persistence raises some interesting questions about what it means for a resource to stay the same over time. One way to think of persistence is that a persistent resource is never changed. However, physical resources often require maintenance, repair, or restoration to keep them accessible and usable, and we might question whether at some point these activities have transformed them into different resources.[55] Likewise, digital resources require regular backup and migration to keep them available and this might include changing their digital format.

We might instead think of persistence more abstractly, and expect that persistent resources need only to remain functionally the same to support the same interactions at any point in their lifetimes even if their physical properties change. Active resources implemented as computational agents or web services might be re-implemented numerous times, but as long as they don't change their interfaces they can be deemed to be persistent from the perspective of any other resource that uses them. Similarly, many resources like online newspapers or blog feeds continually change their content but still could have persistent identifiers.

Some organizing systems closely monitor their resources and every interaction with them to prevent or detect tampering with them or other unauthorized changes. Some organizing systems, like those for software or legal documents, explicitly maintain every changed version to satisfy expectations of persistence because different users might not be relying on the same version. With digital resources determining whether two resources are the same or determining how they are related or derived from one another are very challenging problems. [56]

3.5.2 Effectivity

Many resources also have **effectivity**, meaning that they come into effect, or being, at a particular time and may cease to be effective at some future date. Effectivity is sometimes known as time-to-live. It consists of a date on which the resource is effective, and optionally a date on which the resource ceases to be effective, or becomes stale. For some types of resources, the effective date can be the moment when they're created, but for others, the effective date can be a time different from the moment of creation. For example, a law can be passed in November but not take effect until January 1 of the following year. An effectivity date is the counterpart of the "Best Before" date on perishable goods. That date indicates when a product goes bad, whereas an item's effectivity date is when it "goes good" and the resource that it supersedes needs to be disposed of or archived.

In most cases effectivity implies persistence requirements because it is important to be able to determine and reconstruct the configuration of resources that was in effect at some prior time. A new tax might go into effect on January 1, but if the government audits your tax returns what matters is whether you followed the law that was in effect when you filed your returns.[57]

3.5.3 Authenticity

In ordinary use we say that something is authentic if it can be shown to be, or has come to be accepted as what it claims to be. It is easy to think of examples where authenticity of a resource matters: a signed legal contract, a work of art, a historical artifact, even a person's signature. The importance and nuance of questions about authenticity can be seen in the many words we have to describe the relationship between "the real thing" (the "original") and something else: copy, reproduction, replica, fake, phony, forgery, counterfeit, pretender, imposter, ringer, and so on.

The creator or operator of an organizing system, whether human or machine, can authenticate a newly created resource. A third party can also serve as proof of authenticity. Many professional careers are based on figuring out if a resource is authentic. [58]

There is large body of techniques for establishing the identity of a person or physical resource. We often use judgments about the physical integrity of recorded information when we consider the integrity of its contents.

Digital authenticity is more difficult to establish. Digital resources can be reproduced at almost no cost, exist in multiple locations, carry different names on identical documents or identical names on different documents, and bring about other complications that do not arise with physical items. Technological solutions for ensuring digital authenticity include time stamps, watermarking, encryption, and digital signatures. However, while scholars generally trust technological methods, technologists are more skeptical of them because they can imagine ways for them to be circumvented or counterfeited. Even when a technologically sophisticated system for establishing authenticity is in place, we can still only assume the constancy of identity as far back as this system reaches in the "chain of custody" of the document.

3.5.4 Provenance

The idea that important documents must be created in an authenticatable manner and then preserved with an unbroken chain of custody goes back to ancient Rome. Notaries witnessed the creation of important documents, which were then protected to maintain their integrity or value as evidence. In Organizing Systems like museums and archives that preserve rare or culturally important objects or documents this concern is expressed as the principle of **provenance**. This is the history of the ownership of a collection or the resources in it, where they have been and who has had it.

A uniquely Chinese technique in Organizing Systems is the imprinting of elaborate red seals on documents, books, and paintings that collectively record the provenance of ownership and the review and approval of the artifact by emperors or important officials.

3.6 Key Points in Chapter Three

- We can consider a resource to be one of many members of a very broad category, as the unique instance of a category with only one member, or anywhere in between.
- The size of the category - the number of resources that are treated as equivalent - is determined by the properties or characteristics we consider when we examine the resource.
- More fine-grained organization reduces **recall**, the number of resources you find or retrieve in response to a query, but increases the **precision** of the recalled set, the proportion of recalled items that are relevant.
- Organizing systems for physical information resources emphasize description resources or surrogates like bibliographic records that describe the information content rather than their physical properties.
- Which resources are primary and which are metadata is often just a decision about which resource is the **focus** of our attention.

- It can be useful to view domains of information resources on the Document Type Spectrum from weakly-structured narrative content to highly structured transactional content.
- Organizing systems designed for institutional or industry-wide use require systematic design methods to determine which resources will have separate identities and how they are related to each other.
- The concept of identity for bibliographic resources has evolved into a four-step abstraction hierarchy between the abstract **work**, an **expression** in multiple formats or genres, a particular **manifestation** in one of those formats or genres, and a specific physical **item**.
- Resources become active resources when they contain sensing and communication capabilities.
- Organizing systems that create value from active resources often co-exist with or complement organizing systems that treat its resources as passive.
- If the resource has an IP address, it is said to be part of the "Internet of Things."
- The most basic principle of naming is to choose names that are informative.
- A controlled vocabulary can be thought of as a fixed or closed dictionary that includes all the terms that can be used unambiguously in a particular domain
- Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.
- An identifier is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. The design of a scheme for persistent identifiers must consider both the required time frame and the number of resources to be identified.
- Preservation and governance are activities carried out to ensure the outcome of persistence.

---

[1] [Business] Separating information content from its structure and presentation is essential to repurposing it for different scenarios, applications, devices, or users. The global information economy is increasingly driven by automated information exchange between business processes. When information flows efficiently from one type of document to another in this chain of related documents, the overlapping content components act as the "glue" that connects the information systems or web services that produce and consume the documents. Glushko and McGrath (2005).

[2] [Citation] Furnas, Landauer, Gomez, and Dumais (1987).

[3] [Citation] Glushko and McGrath (2005).

[4] [Citation] Kuniavsky (2010)

[5] [LIS] Project Gutenberg, begun in 1971, was the first large-scale effort to digitize books; its thousands of volunteers have created about 40,000 digital versions of classic printed works. Systematic research in digital libraries began in the 1990s when the US National Science Foundation, the Advanced Research Projects Agency, and NASA launched a Digital Library Initiative that emphasized the enabling technologies and

infrastructure. At about the same time numerous pragmatic efforts to digitize library collections began, characterized by some as a race against time as old books in libraries were literally disintegrating and turning into dust. The Internet Archive, started in 1996, now has a collection of over 3 million texts and has estimated the cost of digitizing to be about $30 for the average book. Multiply this by the scores of millions ofoks held in the world's research libraries and it is easy to why many libraries endorsed Google's offer to digitize their collections.

[6] [CogSci] Encoding of structure in documents is valuable because titles, sections, links and other structural elements can be leveraged to enhance the user interface and navigational interactions with the digital document and enable more precise information retrieval. Some uses of documents require formats that preserve their printed appearance. For example, "presentational fidelity" is essential if we imagine a banker or customs inspector carefully comparing a printed document with a computer-generated one to ensure they are identical.

[7] [Computing][Text encoding specs are well-documented; see (http://www.wotsit.org/list.asp?fc=10).

[8] [Citation] (Chapman and Chapman, 2009).

[9] [LIS] Numerous museums have created web collections, but a great many of them seem to have focused on the quantity of information they could put online rather than on the user experience they were creating. Perhaps not surprisingly, the ambitious use of virtual world technology to create novel forms of interaction described by Rothfarb and Doherty (2007) reflects the highly interactive character of its host museum, the Exploratorium in San Francisco (http://www.exploratorium.edu/). Similarly, the Google Art Project (googleartproject.com) is notable for its goal of complementing and extending, rather than merely imitating, the museum visitor's encounter with artwork (Proctor, 2011). A feature that let people create a "personal art collection" is very popular, enabling a fan of Van Gogh to bring together paintings that hang in different museums.

[10] [Law] 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.. 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).

[11] [Business] The opposing categories of operands and operants have their roots in debates in political economics about the nature of work and the creation of value (Vargo, Lusch, & Morgan 2006) and have more recently played a central role in the development of modern thinking about service design (Constantin & Lusch, 1994; Maglio et al 2009). The concept of agency or operant resources is needed to bring resources that are active information sources, or computational in character, into the organizing system framework. This concept also lets us include living resources, or more specifically, humans, into discussions about organizing systems in a more general way that emphasizes their agency and de-emphasizes other characteristics that could otherwise be distracting,

[12] [Citation] See Allmendinger and Lombreglia (2005), Want (2006).

[13] [CogSci] Luis Von Ahn (Von Ahn, 2004) was the first to use the web to get people to perform "microwork" or "human computation" tasks when he released what he called "the ESP game" that randomly paired people trying to agree on labeling an image. Not long afterwards Amazon created the MTurk platform (www.mturk.com) that lets people propose microwork and others sign up to do it, and today there are both hundreds of thousands of tasks offered and hundreds of thousands of people offering to be paid to do them.

14 [Computing] For semi-structured or more narrative documents these descriptions might be authoring templates used in word processors or other office applications, document schemas in XML applications, style sheets, or other kinds of transformations that change one resource representation into another one. Primary resources that are highly and regularly structured are invariably organized in databases or enterprise information management systems in which a data schema specifies the arrangement and type of data contained in each field or component of the resource.

15 [Computing] There are a large number of 3rd party Twitter apps. See http://twitter.pbworks.com/w/page/1779726/Apps. For a scholarly analysis see Efron (2011).

16 [Citation] (Schmandt-Besserat, 1997)

17 [LIS] We treat resource format and resource focus as distinct dimensions, so there are four categories here. This contrasts with David Weinberger's three "orders of order" that he proposes in the first chapter of a book called *Everything is Miscellaneous* (Weinberger, 2007). Weinberger starts with the assumption that physical resources are inherently the primary ones, so the first "order of order" emerges when physical resources are arranged. The second "order of order" emerges when physical description resources are arranged, and the third "order of order" emerges when digital description resources for physical resources are arranged. Later in the book Weinberger mentions the use of bar codes associated with web sites, a physical description of a digital resource, but because he started with the assumption that physical resources define the "first order" this example doesn't fit into his orders of order.

18 [Computing] These methods go by different names in different disciplines, including "data modeling," "systems analysis," and "document engineering" (e.g., Kent, 1978/2000; Silverston, 2001; Glushko & McGrath, 2005). What they have in common is that they produce conceptual models of a domain that specify their components or parts and the relationships among these components or parts. These conceptual models are called "schemas" or "domain ontologies" in some modeling approaches, and are typically implemented in models that are optimized for particular technologies or applications.

19 [CogSci] Specifically, an NFL football team needs to be considered a single resource for games through the season and in playoffs, and 53 individual players for other situations, like the NFL draft or play-calling. The team and the team's roster can be thought of as resources, and the team's individual players are also resources that make up the whole team.

20 [LIS] Denton (2007) is a highly readable retelling of the history of cataloguing that follows four themes – the use of axioms, user requirements, the "work," and standardization and internationalization – culminating with their synthesis in the Functional Requirements for Bibliographic Records (FRBR).

21 [LIS] This was a surprisingly controversial activity. Many people opposed Panizzi's efforts as a waste of time of effort because they assumed that "building a catalog was a simple matter of writing down a list of titles" (Denton 2007, p. 38).

22 [LIS] Lubetzsky worked for the US Library of Congress from 1943-1960 where he tirelessly sought to simplify the proliferating mass of special case cataloguing rules proposed by the American Library Association , because at the time the LOC had the task of applying those rules and making the catalog cards other libraries used. Lubetsky's book on Cataloguing Rules and Principles (Lubetsky, 1953) bluntly asks "Is this rule necessary?" and was a turning point in cataloguing.

23 [LIS] In between the abstraction of the WORK and the specific single ITEM are two additional levels in the FRBR abstraction hierarchy. An EXPRESSION denotes the multiple the multiple realizations of a work in some particular medium or notation, where it can actually be perceived. There are many editions and translations of Macbeth, but they are all the same expression, and they are a different expression from all of

the film adaptations of Macbeth. A MANIFESTATION is the set of physical artifacts with the same expression. All of the copies of the Folger Library print edition of Macbeth are the same manifestation.

[24] [Computing] This kind of advice can be found in many data or conceptual modeling texts, but this particular statement comes from Glushko, Weaver, Coonan, and Lincoln (1988).

[25] [Computing] A group of techniques collectively called normalization produces a set of tightly defined information components that have minimal redundancy and ambiguity. Imagine that a business keeps information about customer orders using a "spreadsheet" style of organization in which a row contains cells that record the date, order number, customer name, customer address, item ID, item description, quantity, unit price, and total price. If an order contains multiple products, these would be recorded on additional rows, as would subsequent orders from the same customer. All of this information is important to the business, but this way of organizing it has a great deal of redundancy and inefficiency. For example, the customer address recurs in every order, and the customer address field merges street, city, state and zip code into a large unstructured field rather than separating them as atomic components of different types of information with potentially varying uses. Similar redundancy exists for the products and prices. Cancelling an order might result in the business deleting all the information it has about a particular customer or product.

Normalization divides this large body of information into four separate tables, one for customers, one for customer orders, one for the items contained in each order, and one for item information. This normalized information model encodes all of the information in the "spreadsheet style" model, but eliminates the redundancy and avoids the data integrity problems that are inherent in it.

Normalization is taught in every database design course. The concept and methods were proposed by Codd (1970), who invented the relational data model, and has been taught to countless students in numerous database design textbooks like Date (2003).

[26] [Computing] The "Internet of Things" concept spread very quickly after it was proposed in 1999 by Kevin Ashton, who co-founded the Auto-ID center at MIT that year to standardize RFID and sensor information (Ashton, 1999). For a popular introduction, see (Gershenfeld, Krikorian, & Cohen, 2004). For a recent technical survey and a taxonomy of application domains and scenarios see (Atzori, Iera, & Morabito, 2010).

[27] [Computing] University of Southern California professor Julian Bleecker (2006) coined the term "Blogjects" to describe objects that blog (p. 2). Bleecker's early example of a Blogject is Beatriz da Costa's Pigeon Blog. Da Costa, a Los Angeles–based artist working at the intersection of life sciences, politics, and technology, armed urban pigeons with pollution sensors and locative tracking devices, released them, and created a web interface—in this case Pigeon Blog—to display their flight patterns on Google Maps alongside the pollution levels in the air as they flew. "Whereas once the pigeon was an urban varmint whose value as a participant in the larger social collective was practically nil or worse, the Pigeon that Blogs now attains first-class citizen status" (Bleecker, 2006, p. 5).

[28] [Computing] IBM's Andy Stanford-Clark has been credited with coining the term when he wired his house with sensors, enabling appliances to send information to the house's Twitter account, @andy_house (MacManus, 2009, para. 4). The house plant kit: http://www.sparkfun.com/products/10334

[29] [Computing] Pattern analysis can help escape this dilemma by enabling predictive modeling to make optimal use of the data. In designing smart things and devices for people, it is helpful to create a smart model in order to predict the kinds of patterns and locations relevant to the data collected or monitored. These allow designers to develop a set of dimensions and principles that will act as smart guides for the development of smart things. Modeling helps to enable automation, security, or energy efficiency, and baseline models can be used to detect anomalies. As for location, exact locations are unnecessary; use of a "symbolic space" to represent each "sensing zone"—e.g., rooms in a house—and an individual's movement history as a string of symbols—e.g., abcdegia—works sufficiently as a model of prediction. See (Das et al 2002).

[30] [Law] Well, maybe not anything. Books list traditional meanings of various names, charts rank names by popularity in different eras, and dozens of websites tout themselves as the place to find a special and unique name. See http://www.ssa.gov/oact/babynames/ for historical trends about baby names in the US with an interactive visualization at http://www.babynamewizard.com/voyager#

Different countries have rules about characters or words that may be used in names. In Germany, for example, the government regulates the names parents can give to their children; there's even a book, the International Handbook of Forenames, to guide them (Kulish, 2009). In Portugal, the Ministry of Justice publishes lists of prohibited names (Cornell, 2006). Meanwhile, in 2007, Swedish tax officials rejected a family's attempt to name their daughter Metallica (BBC, 2007).

We can also change our names. Whether a woman takes on her husband's surname after marriage or, like the California man who changed his name to "Trout Fishing," we just find something that better suits us than the name given by our parents. In California in the 1990s, a high school student made waves by changing his name to "Trout Fishing in America" (Associated Press, 1994).

[31] [CogSci] And while you may think that certain terms are more obviously "good" than others, studies show that "there is no one good access term for most objects. The idea of an 'obvious,' 'self-evident,' or 'natural' term is a myth!" (Furnas et al, 1987, p. 967).

[32] [CogSci] The most common names for this service were activities, calendar and events, but in all over a hundred different names were suggested, including cityevents, whatup, sparetime, funtime, weekender, nightout, and many more, "People use a surprisingly great variety of words to refer to the same thing," Furnas wrote. "If everyone always agreed on what to call things, the user's word would be the designer's word would be the system's word. . . . Unfortunately, people often disagree on the words they use for things" (Furnas, 1987, p. 964).

[33] [CogSci] This example comes from (Farish, 2002), who analyzes "What's in a Name?" and suggests that multiple names for the same thing might be a good idea because non-technical business users, data analysts, and system implementers need to see things differently and no one standard for assigning names will work for all three audiences.

[34] [CogSci] See, for example, *Handbook of Cross-Cultural Marketing*, (Herbig 1998). The Starbucks coffee chain seemingly goes out of its way to confuse its customers by calling the smallest of its three coffee sizes (12 ounces) the "tall" size, calling its 16-ounce size a "grande," and calling its largest a "venti," which is Italian for 20 (ounces). Outside of Starbucks, something that is "tall" is never also considered "small." Ironically, despite having about 20,000 stores in about 60 countries, Starbucks has none in Italy where "venti" would be in the local language.

[35] [Business] Economist, As easy as YZX, August 30 2001. For example, the convention to list the co-authors of scientific publications in alphabetic order has been shown to affect reputation and employment by giving undeserved advantages to people whose names start with letters that come early in the alphabet. This bias might also affect admission to selective schools. (Efthyvoulou, 2008).

[36] [Business] The Kentucky Fried Chicken franchise solved this problem by changing its name to KFC, which you can now find in Beijing, Moscow, London and other locations not anywhere near Kentucky and where many people have probably never heard of the place.

[37] [Computing] Tim Berners-Lee, the founder of the web, famously argued that "Cool URIs Don't Change" (Berners-Lee, 1998).

[38] [Law] Any online citation to one of the West printed court reports will use the West format. However, when Mead Data wanted to use the West page numbers in its LEXIS online service to link to specific pages, West sued for copyright infringement. The citation for the West Publishing vs. Mead Data Central case is 799

F.2d 1219 (8th Cir 1986), which means that the case begins on page 1219 of volume 799 in the set of opinions from the 8th circuit court of appeals that West published in print form. West won the case and Mead Data had to pay substantial royalties. Fortunately, this logic behind this decision was repudiated by the US Supreme Court a few years later in a case that West published as Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991), and West can no longer claim copyright on page numbers.

39 [CogSci] When George Orwell gave the title "1984" to a novel he wrote in 1949 he intended it as a warning about a totalitarian future as the Cold War took hold in a divided Europe, but today 1984 is decades in the past and the title doesn't have the same impact.

40 [Citation] (Dorai and Venkatesh, 2001).

41 [Citation] Most common US surnames; http://names.mongabay.com/most_common_surnames.htm; Chad Ochocinco story: http://en.wikipedia.org/wiki/Chad_Ochocinco. The Artist Formerly Known as Prince: http://en.wikipedia.org/wiki/Prince_%28musician%29. Fake names at Starkbucks: http://online.wsj.com/article/SB10001424053111904106704576582834147448392.html. Twitter on sports jerseys: : http://www.forbes.com/sites/alexknapp/2011/12/30/pro-lacrosse-team-replaces-names-with-twitter-handles-on-jerseys/?partner=technology_newsletter

42 [Computing] Identifiers with meaningful internal structure are said to be structured or intelligent. Those that contain no additional information are sometimes said to be unstructured, opaque, or dumb. The 8 in the ISBN example is a check digit, not technically part of the identifier, that is algorithmically derived from the other digits to detect errors in entering the ISBN.

43[Citation] (McCartney 2006).

44 [LIS] Svenonius (2000) calls vocabulary control "the sine qua non of information organization" (p. 89). "The imposition of vocabulary control creates an artificial language out of a natural language" (p. 89), leaving behind an official, normalized set of terms and their uses.

45 [LIS] This mapping is "the means by which the language of the user and that of a retrieval system are brought into sync" (Svenonius, 2000, p. 93) and allows an information-seeker to understand the relationship between, say, Samuel Clemens and Mark Twain. The Library of Congress maintains a list of standard, accepted names for authors, subjects, and titles called the Name Authority File. http://id.loc.gov/authorities/names.html

46 [Citation] PESI www.eu-nomen.eu/pesi; CBOL www.barcoding.si.edu/; http://services.natureserve.org/BrowseServices/getSpeciesData/getSpeciesListREST.jsp

47 [Citation] (Hemerly 2011).

48 [Law] This rations / radio confusion is described in (Wheatley 2004). In 2008 a similar mistake in managing inventory at a US military warehouse led to missile launch fuses being sent to Taiwan instead of helicopter batteries, causing a high-level diplomatic furor when the Chinese government objected to this as a treaty violation (Hoffman 2008).

49 [LIS] Organizing systems in libraries, museums, and businesses often give sequential accession numbers to resources when they are added to a collection, but these identifiers are of no use outside of the context in which they are assigned, as when a union catalog or merged database is created.

50 [Computing] A more general technique is to use the UUID standard, which standardizes some algorithms that generate 128-bit tokens that, for all practical purposes, will be unique for hundreds, if not thousands, of years

[51] [Computing] The OASIS XML Common Biometric Format (XCBF) was developed to standardize the use of biometric data like DNA, fingerprints, iris scans, and hand geometry to verify identity (OASIS 2003).

[52] [Citation] (Coyle, 2006, p. 429).

[53] [Computing] IP 6 for internet addresses. The threat of exhaustion was the motivation for remedial technologies, such as classful networks, Classless Inter-Domain Routing (CIDR) methods, and network address translation (NAT) that extend the usable address space.

[54] [Computing] Digital Object Identifier system (www.doi.org). However, DOI has its issues too. It's a highly political, publisher-controlled system, not a universal solution to persistence.

[55] [CogSci] This is called the Paradox of Theseus, a philosophical debate since ancient times. Every day that Theseus's ship is in the harbor, a single plank gets replaced, until after a few years the ship is completely rebuilt: not a single original plank remains. Is it still the ship of Theseus? And suppose, meanwhile, the shipbuilders have been building a new ship out of the replaced planks? Is that the ship of Theseus? (Furner, 2008, p. 6)

[56] [Citation] See (Renear and Dubin 2003), (Wynholds 2011).

[57] [Business] Effectivity in the tax code is simple compared to that relating to documents in complex systems, like commercial aircraft. Because of their long lifetimes—the Boeing 737 has been flying since the 1960s—and continual upgrading of parts like engines and computers, each airplane has its own operating and maintenance manual that reflects the changes made to the plane over time. Every change to the plane requires an update to the repair manual, making the old version obsolete. And while an aircraft mechanic might refer to "the 737 maintenance manual," each 737 aircraft actually has its own unique manual.

[58] [Law] Notary publics are used on a daily basis to verify that a signature on an important document such as a mortgage or other contract is authentic, much as signet rings and sealing wax once proved that no one has tampered with a document since it was sealed.