Chapter 4

Resources in Organizing Systems

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

4.1 Introduction

This chapter builds upon the foundational concepts previously introduced to explain more carefully what we mean by resource. In particular, it focuses on identity, what will be treated as a separate resource, and discusses the issues and principles to consider when identifying and naming resources.

4.1.1. What is a Resource?

As a resource is "anything of value that can support goal-oriented activity," it might seem that the question of identity, of what a single resource is, should not be hard to answer. However, even when resources are tangible things, how to organize them is not always obvious because people think of them in different ways. Which properties garner our attention, and which we use in organizing, depends on our experiences, purposes, and context.

Naming or describing a resource adds information to it, and the information added varies because the same resource can be named or described in many different ways. Depending on the context and purpose, any given resource may be one of many members of a broad category, one of the few members of a narrow category, or a unique instance of a category with only one member. For example, we might recognize something as a piece of clothing, a sock, or the specific sock with the hole in the heel. Nevertheless, 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."

While it is not always possible to separate the decisions we need to make about resource instances from those about resource classes and types, this chapter focuses on the former; the latter are covered in Chapter 7, Categorization: Describing Resource Classes and Types.

4.1.1.1 Resources with Parts

As tricky as it can be to decide what a resource is when dealing with single objects, it is even more challenging when the resources are objects or systems composed of parts. In these cases, we must treat both the entirety of the object or system and its constituent parts as resources. Additionally, we must consider the relationships between the parts and the whole.

The answer to the question "how many things is a car?" depends on your point of view. If you are buying or selling a car, it is one thing. When assembling a car, it consists of several dozen large parts like the frame, suspension, drive train, fuel system, engine, exhaust system, passenger compartment, and other pre-assembled components. In turn, each of those components is itself made up of many

parts. If you count all of the parts of all of the components, including the smallest screws and wires, you will have counted about thirty thousand parts in the average car.

This ambiguity also holds for information resources. For example, 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. Similarly, a web page can be treated as a single resource, but it can also be considered as a collection of parts, each of which can be separately identified as the source or anchor of a link. Alternatively, various information components can be combined to form a single resource such as a credit score which is a statistical index that combines information about outstanding loans, payment history, current income, and more.

4.1.1.2 Bibliographic Resources, Information Components, and “Smart Things” as Resources

Information resources 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 resource descriptions, such as bibliographic records, describe the knowledge the books contain, rather than descriptions of their physical properties.

Another question about identify that is especially critical for bibliographic resources, but relevant to all information resources, is: What set of resources should be treated as the same because they contain essentially similar intellectual or artistic content? For example, we may talk about Shakespeare's play Macbeth, but what is this thing we call "Macbeth"? Is it a particular string of words, whether saved in a computer file or handwritten on 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 creations 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? Patrick Wilson proposed a genealogical analogy, characterizing a work as "a group or family of texts," with the idea that a creation like Shakespeare's Macbeth is the "ancestor of later members of the family."

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 appears in a specific instance and context. From a logical perspective, an order form may contain 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 across documents.

Similar questions are posed by the emergence of ubiquitous or pervasive computing. Information processing capabilities and connectivity are now commonly embedded into physical objects and 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.

4.1.2 Identity, Identifiers, and Names

The answer to the question “What Is a Resource?” has two parts. The first, identity, asks “what thing are we treating as the resource?” The second, identification, is the process of differentiating between a particular resource and other resources like it.

These problems are closely related, and once you have 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 resource 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.

Choosing names and identifiers—for a person, a service, a place, a document, a concept, or anything else—is often challenging and 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 a text. This issue of multiple names for the same resource or concept is sometimes called the "vocabulary problem."

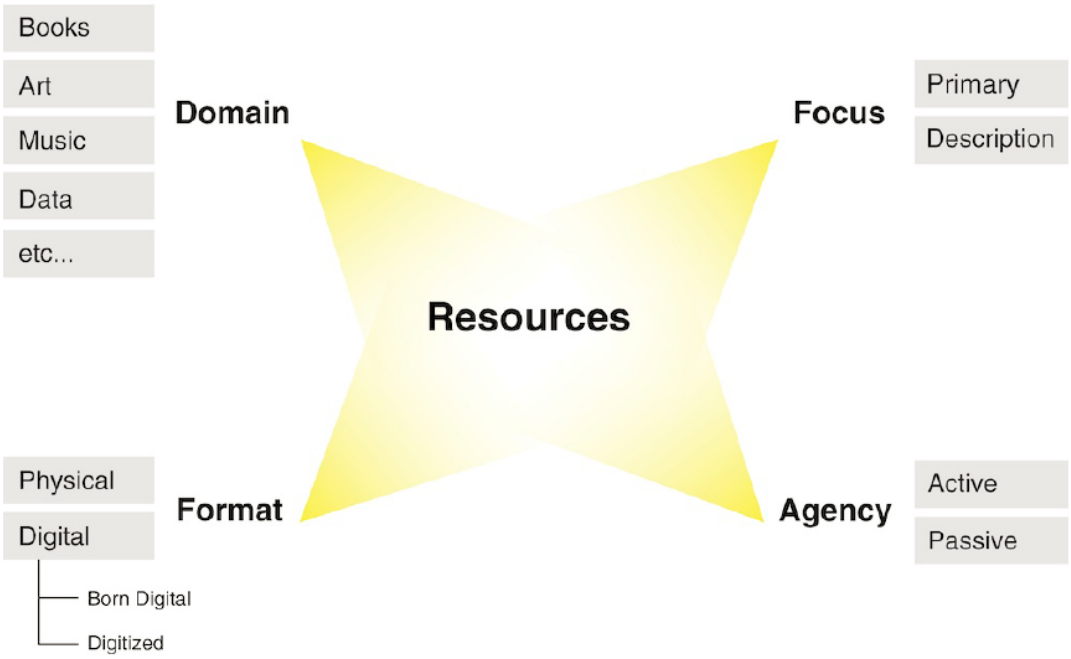
A partial solution to the vocabulary problem is to use a controlled vocabulary. We can impose rules that standardize the way in which names and labels for resources are assigned. However, vocabulary control cannot remove all ambiguity. Even if a passport or national identity system requires authoritative full names rather than nicknames, there will probably be more than one person named Robert John Smith or 张伟 Zhang Wei.

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?

4.2 Resource Domain, Format, Focus, and Agency

Considering 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, focus, and agency (see Figure 4.1).

Figure 4.1. Resource Domain, Format, Focus and Agency.



Four distinctions we can make when discussing resources concern their domain (their type of matter or content), format (physical or digital), agency (active or passive), and focus (primary or description).

4.2.1 Resource Domain

Resource domain is the notion that resources can be grouped according to the set of characteristics that distinguishes them from other resources.

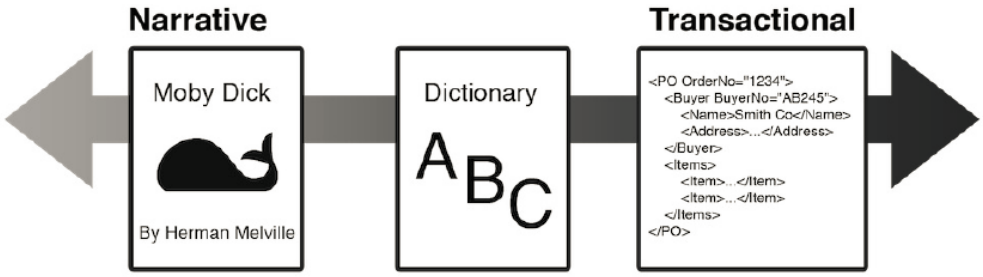
For physical resources, domains can be coarsely distinguished according to the type of matter they are made of using easily perceived properties. The top-level classification of all things into the animal, vegetable, and mineral kingdoms by Carl Linnaeus in 1735 is deeply embedded in most languages and cultures and creates a hierarchical system of domain categories. Many aspects of this system of domain categories are determined by natural constraints on category membership that exist as patterns of shared and correlated properties. A resource identified as a member of one category must also be a member of the categories above it in the hierarchy with which it shares some but not all properties. For example, a marble statue in a museum must also be a kind of mineral, and a fish in an aquarium must also be a kind of animal.

For information resources, we distinguish domains based on semantic properties. The definitions of "encyclopedia," "novel," and "invoice" distinguish them according to their typical subject matter, or the type of content, rather than according to the great variety of physical or digital forms in which we might encounter them.

While information resources can be arranged in a hierarchy, it is more useful to view domains of information resources on a continuum from weakly-structured narrative content to highly structured transactional content because the category boundaries are not sharp. 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.

Narrative types are authored by people, are heterogeneous in structure and content, and usually consist of just prose and graphic elements. 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 such as figures, data tables, and code examples.

Figure 4.2. Document Type Spectrum.



The Document Type Spectrum is a continuum of document types from narrative ones that are mostly text, like novels, to transactional ones with highly-structured information, like invoices. In between are hybrid types that contain both narrative and transactional content, like dictionaries and encyclopedias.

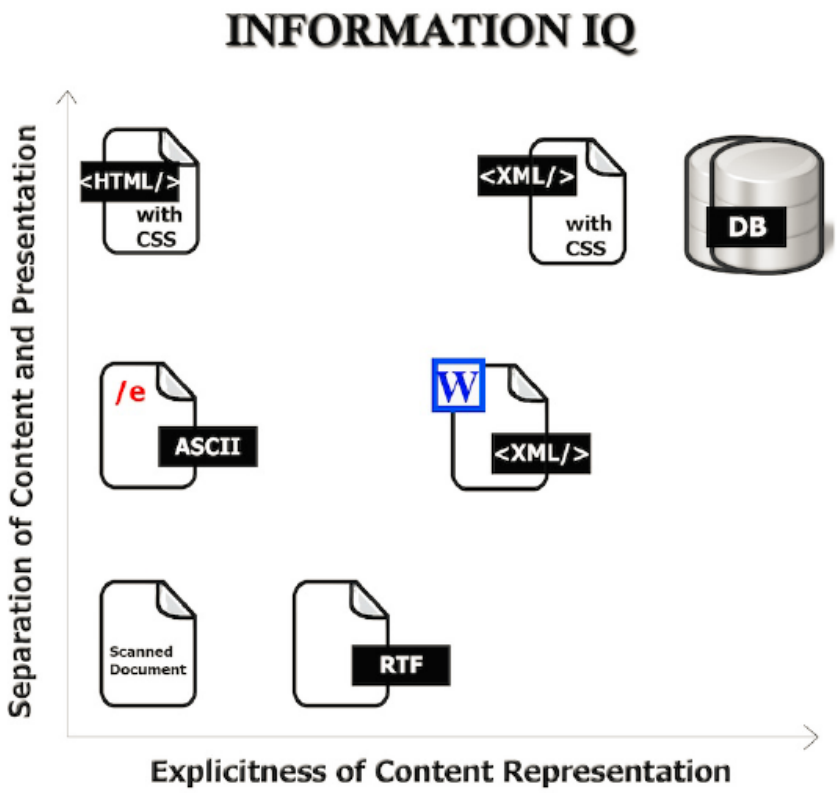
4.2.2 Resource Format

For information resources, the most basic format distinction is between physical and digital formats. This distinction is most important in the implementation of an organizing system, but less important at the logical level when designing the system and interactions. While the implementation of an organizing system for physical information resources will be constrained by their format, a similar system for digital information resources, or digital surrogates for physical ones, will not face the same constraints.

Today, many resources are born digital, created in word processors, graphics software, digital cameras, and the sensors in "smart things." Digital resources are also created when systems interact with barcodes, QR codes, RFID tags, or other mechanisms for tracking identity and location. Physical resources can be transformed into digital ones through digitization which allows them to be stored and manipulated by computers. Printed text, for example, can be digitized by scanning pages and using character recognition software or simply by re-typing it.

There are a vast number of digital formats that differ in many ways, but we can coarsely compare them on two dimensions: the degree to which they distinguish information content from presentation or rendering, and the explicitness with which content distinctions are represented. Taken together, these two dimensions comprise the "Information IQ" of a format, with the overarching principle being that "smarter" formats contain more information that can be processed by computers (see Figure 4.3).

Figure 4.3. Information IQ.



The notion of Information IQ captures the idea that document formats differ on two dimensions: the explicitness of content representation, and the separation of content and presentation. A scanned document is just a picture of a document with neither of these distinctions, so it is low on both dimensions. A database or XML document distinguishes explicitly between types of content and presentation is separately assigned, so they are high on both dimensions and have the highest Information IQ. An HTML document's content distinctions are usually presentational and, thus, it has lower IQ. Formats with high Information IQ facilitate computer processing.

Most document formats also explicitly encode a hierarchy of structural components, such as chapters, sections, or semantic components like descriptions or procedural steps. Additionally, some formats can encode the appearance of the rendered or printed form. An important distinction between formats is whether the information is encoded to be human readable as well as computer readable. For example, XML allows for layering of intentional coding or markup with "plain text" content. The most complex digital formats are those for multimedia resources and multidimensional data, where the data format is highly optimized for processing efficiency.

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. There may be multiple intended uses for a digitized resource that could require different digitization approaches and formats. Downstream users of digitized resources need to

know the format in which a 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. Sophisticated digital formats can enable interactions with annotated digital images or video, 3-D graphics, or embedded datasets. The Google Art Project contains extremely high-resolution photographs of famous paintings that make it possible to see details that are undetectable under normal viewing conditions in museums. That said, digital representations of physical resources can also lose information and capabilities. The distinctive sounds of hip-hop music produced by "scratching" vinyl records on turntables cannot be created from digital music files. Additionally, it is important to note that digital formats can intentionally limit interactions. For example, digital rights management (DRM) systems can prevent copying even if the copying is legally permissible.

4.2.3 Resource Agency

Agency is the extent to which a resource can initiate actions on its own. We can define a continuum from passive resources that cannot initiate any actions to active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources. To illustrate this, consider how the following three resources interact with the temperature of their environment: A book left out in the sun will grow warm, but it has no way of measuring its temperature meaning it is a passive resource. An ordinary mercury thermometer senses and displays the temperature, so it is an active resource, but it is not capable of communicating its reading. A digital wireless weather station that can both sense and communicate temperature information is also an active resource, and it has more agency than the mercury thermometer due to its ability to communicate information. In general, passive resources can be thought of as being like nouns that are acted upon, while active resources can be thought of as verbs that cause and carry out actions.

People as Resources

People organize themselves in innumerable ways to coexist, share knowledge, and accomplish more than they could as individuals. In human society, behaviors such as trust and reciprocity might be considered “organizing principles.” However, these organic relationships and interactions usually lack the intentional arrangement to be considered true Organizing Systems, except when the people are living in “intentional communities” like communes, monasteries, or ashrams where the members share social, political, or religious beliefs.

Nonetheless, human resources can be identified, categorized, described in terms of their attributes and relationships and take part in interactions to create value just like digital and physical resources. In businesses, people are organized to amplify their skills, knowledge, and agency. A company's organizational chart is often a formal hierarchy in which each worker's role is defined by his or her responsibilities and relationships to others in the company. Treating each employee abstractly as a resource with specific and predictable functions, inputs, and outputs enables employees or processes to depend upon each other without being distracted by the details of one another's work.

4.2.3.1 Passive Resources

Organizing systems containing passive resources are ubiquitous because we live in a world of physical resources that we identify and name. Passive resources are usually tangible and static, and their value comes as a result of some action or interaction with them.

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

4.2.3.2 Active Resources

Active resources can create effects or value on their own or when they initiate interactions with passive resources. Some examples of active resources are: people, other living resources, computational agents, active information sources, web-based services, self-driving cars, and robots. Additionally, otherwise ordinary objects like light bulbs, umbrellas, and shoes can be made "smarter" by exploiting computing capability, storage capacity, and communication bandwidth.

We can analyze active resources according to five capabilities that build on each other to give resources, and the organizing systems in which they participate, more ways to create value through interactions and information exchanges.

Sensing or Awareness

> The minimal capability for a resource to have some agency is for it to be able to sense some aspect of its environment or its interactions with other resources. A thermometer measures temperature, a photodetector measures light, a fuel gauge measures the gas left in a car's tank, a GPS device computes its location by detecting and analyzing signals from satellites, a wearable fitness sensor tracks your heartbeat and how far you walk. However, sensing in itself does not create any value in an organizing system; value is created by responding to the sensed information.

Actuation

> A resource is an "actuator" when it uses the information it senses to move or control a physical mechanism or system. Resources can actuate by turning on lights, speakers, cameras, motors, switches, by sending a message about the state or value of a sensor, or by moving themselves around (as with robots).
>
> A potential or latent actuation is created when a resource can display or broadcast some aspect of its state. In this case, value is only created if another resource, possibly human, happens to see the display or hear the broadcast and then acts upon it.
>
> For example, RFID chips, which are essentially barcodes with built-in radio transponders, can be attached to otherwise passive resources to make them active. RFID chips transmit information when they are in the presence of an RFID reading device, and this transmission enables automated location tracking and context sensing. RFID reading devices built into assembly lines, loading docks, parking lots, toll booths, or store shelves can detect when an RFID-tagged resource is at, or has been moved from, a particular location.

Connectivity

> For an active resource to do useful work, it must be connected in some way to the actuation mechanism that manipulates or controls some other resource. This connection might be a direct and permanent one between the resource and the thing it actuates. For example, a

thermostat with a temperature sensing capability has a fixed connection to a heating or cooling system that it turns off or on depending on the temperature.

Additionally, physical resources can be "wrapped" with software and given an IP address to enable the use of Internet protocols to send information to an application that has more capability to act on it. Such resources are frequently said to be “smart” or part of the "Internet of Things." For example, smartphones can run applications that receive messages from, and send messages to, other resources to monitor and optimize how they work.

Computation or Programmability

Simple active resources operate in a deterministic manner: given this sensor reading, do this. Other active resources have computational capabilities that enable them to analyze current and historical information from their sensors, identify significant data values or patterns in their interaction resources, and adapt their behavior accordingly. For example, the Nest thermostat uses sensors for temperature, humidity, motion, and light to figure out whether people are at home, and then programs itself to optimize energy use.

Similarly, the Roomba vacuum cleaning robot navigates around furniture, power cords, stairs, and optimizes its cleaning path to go over particularly dirty places. More sophisticated robots are designed to be versatile and adaptable so they can repetitively perform whatever task is needed for some manufacturing process, and their capabilities can be continually upgraded by software updates. A new generation of robots, typified by one called Baxter, can be trained by example: a person moves Baxter's arms and hands to show it what to do, and when Baxter has programmed itself to repeat it, it nods.

Composability, Cooperation, and Standards

The "smartest" active resources can do even more than analyze information and adapt what they do. They can also expose what they know and can do to other resources using standard or non-proprietary formats and protocols. This means that active resources that were independently designed and implemented can work together to create additional value.

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.

The same principles of modularity and composability are being applied to physical resources that have been made “smart” through the use of open source software libraries and APIs for sensors and micro-controllers . As these resources serve as functional building blocks, standards will be critically important to ensure composability.

Standard application interfaces enable active resources to interact with people to get additional information or enhance the value of the sensor information. A programmable thermostat that can record time-based preferences of the people who use the space controlled by the thermostat is more capable than one with just a single temperature setting. A standard Internet protocol for communicating with the thermostat would enable it to be controlled remotely.

Open and standard data formats and communication protocols enable the aggregation and analysis of information from many instances of the same type of active resource. For example, smart phones running the Google Maps application transmit information about

their speed and location. This data can then be aggregated and analyzed using machine learning and other techniques to yield collective intelligence which is transmitted back to the resources from which it was derived. In this case, Google can identify traffic jams and generate alternative routes for the drivers stuck in traffic.

4.2.4 Resource Focus

Resource focus distinguishes original or primary 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 in the system's 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 is deeply embedded in library science and traditional organizing systems whose collections are predominantly text resources. 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.

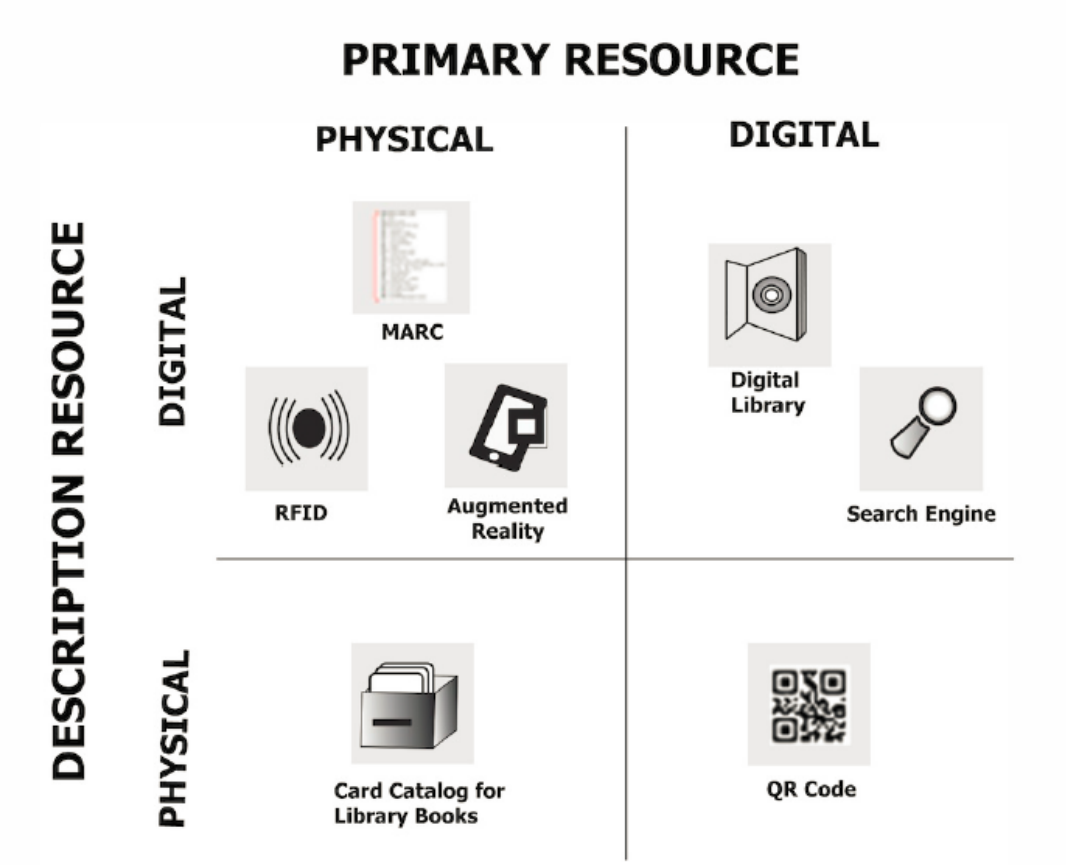
In business enterprises, the organizing systems for digital information resources, such as business documents, or data records created by transactions or automated processes, almost always employ resources that describe, or are associated with, large sets or classes of primary resources.

The contrast between primary resources and description resources is often useful, but looking more broadly at organizing systems, it can be difficult to distinguish between them. Which resources are primary or descriptive is just a decision about which resource is currently the focus of attention. For example, many Twitter users treat the 140-character message body as the primary resource, while the associated metadata about the message and sender (is it a forward, reply, link, etc.) is less important. However, for firms that use Twitter metadata to measure sender and brand impact, or identify social networks and trends, the focus is the metadata, not the content.

4.2.5 Resource Format x Focus

Crossing resource format with focus yields a useful framework with four categories of resources (see Figure 4.4).

Figure 4.4. Resource Format x Focus.



The distinctions of resource format and resource focus combine to distinguish four categories of resources: physical resources, digital resources, physical descriptions , and digital descriptions.

4.2.5.1 Physical Description of a Primary Physical Resource

The oldest relationships between descriptive resources and physical resources is when the descriptions 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 5,000 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. A more contemporary example is the printed cards that served as physical description resources for books in libraries for nearly two centuries.

4.2.5.2 Digital Description of a Primary Physical Resource

A common example of this relationship is the online library catalog used to find the shelf location of physical library resources, which beginning in the 1960s, began to replace physical cards with database records. Digital description resources for primary physical resources are essential for business models that depend on having timely and accurate information about where things are or their current state such as supply chain management, logistics retailing, and transportation.

Augmented reality systems combine a layer of real-time digital information about some physical object to a digital view or representation of it. A familiar example is the yellow “first down” line superimposed in broadcasts of football games. Augmented reality techniques that superimpose identifying or descriptive metadata are 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.

4.2.5.3 Digital Description of a Primary Digital Resource

When a digital resource describes a digital resource, it is often possible to access the primary resource directly from the descriptive resource. Digital libraries and web-based organizing system rely on this relationship.

4.2.5.4 Physical Description of a Primary Digital Resource

QR codes, whether in newspaper advertisements or on billboards, sidewalks, t-shirts, or store shelves, are physical descriptions of a primary digital resources. Scanning a QR code with a mobile phone camera can launch a website that contains information about a product or service, dial a phone number, or initiate another application or service identified by the QR code.

4.3 Resource Identity

Identifying the resources that will be organized is the essential task when building any organizing system. Once the resources are known, it is necessary to decide which resource properties are relevant to the people or systems operating in that domain. These properties are then used in conjunction with the organizing principles to define the relationships among the resources. In organizing systems used by individuals or with a small scope, the methods for doing these tasks are often ad hoc and unsystematic, while 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.

4.3.1 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. We have little difficulty recognizing an object or a person even if we see them from a novel distance and viewing angle or with different lighting and shading. When watching a football game, we do not have any trouble perceiving the players who are moving around the field, and their contrasting uniform colors allow us to see the different teams.

These perceptual mechanisms enable the organizing tasks of identifying some 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 and 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, or as offense and defense.

4.3.2 Identity and Bibliographic Resources

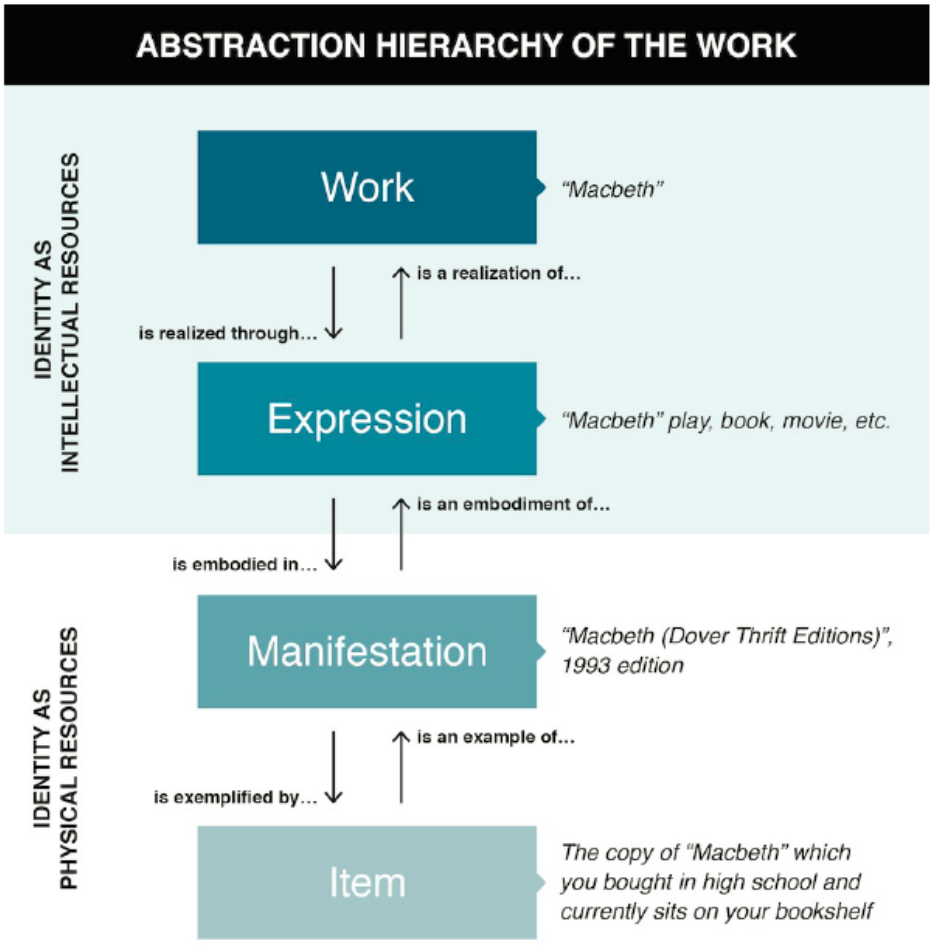
The question of identity is relatively recent in the world of librarians and catalogers. Libraries have been around for approximately 4,000 years, but until the last few hundred years, librarians created "bins" of headings and topics to organize resources without bothering to give each item a separate

identifier or name. Additionally, these choices were ad hoc and each cataloger independently decided the bins and groupings for each catalog.

However, contemporary thinking about resource identity in library science is much more sophisticated and systematic. A four-step abstraction hierarchy has been proposed that distinguishes the abstract work, its expression in multiple formats or genres, a particular manifestation in one of those formats or genres, and a specific physical item (see Figure 4.5).

Revisiting 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 particular and individual resource, such as the paperback with yellow marked pages that you owned when you read Macbeth in 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.

Figure 4.5. The FRBR Abstraction Hierarchy.



The abstraction hierarchy for identifying resources yields four different answers about the identity of an information resource.

4.3.3 Identity and Information Components

In information-intensive domains, value is created through the comparison, compilation, coordination, or transformation of information as it flows from one information source or process to another. These processes are “glued together” by shared information components that are exchanged in documents, records, messages, or resource descriptions. Information components are the primitive and abstract resources in information-intensive domains. They are the units of meaning that serve as building blocks for composite descriptions and other information artifacts. Given this, an information component can be any piece of information that has a unique label or identifier or any piece of information that is self-contained and comprehensible on its own.

The value creation processes in information-intensive domains work best when the information components use a common controlled vocabulary or 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.

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, or payment information must be encoded as a particular type of data—integer, decimal, or Unicode string—with clearly defined possible values.

However, the theoretically grounded methods for identifying components in structured data do not work as well when information is more qualitative and less precise such as at the narrative end of the Document Type Spectrum. Narrative documents include technical publications, reports, policies, procedures and other less structured information in which semantic components are rarely labeled explicitly. Unlike transactional documents that depend on precise semantics, narrative documents are used by people, who can ask if they are not sure what something means, so there is less need to precisely define the meaning of the information components.

4.4 Naming Resources

Determining the identity of the thing, document, information component, or data item we need is not always enough. We often need to give that resource a name that will help us understand and talk about it. As such, naming is not just the task of assigning a sequence of characters to a resource.

4.4.1 What’s in a Name?

When a child is born, its parents give her or him a name which serves to distinguish that child from all others even though the name might not be unique—there are thousands of people named James Smith and 王丽 Wang Li. Beyond this, names, intentionally or unintentionally, suggest characteristics or aspirations. Furthermore, the name given 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, schools, and 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.

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.

4.4.2 The Problems of Naming

4.4.2.1 The Vocabulary Problem

Every natural language offers more than one way to express any thought, and 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 sometimes surprised that these disagreements because what seems like the natural or obvious name to one person is not natural or obvious to another. One way to avoid surprises is to have people cooperate when choosing names for resources, and participatory design techniques can be used for this purpose.

4.4.2.2 Homonymy, Polysemy, and False Friends

Sometimes the same word can refer to different resources—in English, 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 polysemes.

Resources with homonymous and polysemous names are sometimes incorrectly identified if common sense or context are not used 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 places an order, while another person thinks of a “shipping container” as the large metal box carried by semi-trailers and stacked on cargo ships, their disagreement might not be discovered until the wrong kinds of containers arrive.

False friends 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 due to undesirable associations. For example, "gift" means "a present" in English but "poison" in German.

4.4.2.3 Names with Undesirable Associations

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. Alphabetical ordering might seem like a fair and non-discriminatory arrangement of resources, but 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 merit.

4.4.2.4 Names that Assume Impermanent Attributes

If a resource is given a name based on attributes that can change in value or interpretation, it may later cause problems. 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 rule treats the current location of the resource as its name, so the name will change if the resource is moved. Additionally, some dynamic web resources that are generated by programs have URIs that contain information about the server technology used to create them; when the technology changes, the URIs will no longer work. An analogous problem is

faced by restaurants or other businesses with street names or numbers in their names if they lose their leases or want to expand.

4.4.2.5 The Semantic Gap

The semantic gap is the difference in perspective in naming and description when resources are described by automated processes rather than by people. The 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. Such names—like IMG20268.jpg for a digital photo—make sense for the camera as it stores consecutively taken photos, but they are not useful names for people who may prefer names that describe the content of the picture, like GoldenGateBridge.jpg.

4.4.3 Choosing Good Names and Identifiers

If someone tells you they are having dinner with their best friend, cousin, someone with whom they play basketball, and their professional mentor from work, how many people will be at the dinner? Anywhere from two to five. It is possible all those relational descriptions refer to a single person, or to four different people, and because "friend," "cousin," "basketball teammate," and "mentor" do not name specific people, you will have to guess who is coming to dinner. If instead of descriptions you are 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. However, you still cannot be sure exactly which four people are involved because there are many people with those names.

Uncertainty can be eliminated by using identifiers rather than standard 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 strings of numbers or letters rather than words to avoid the biases and associations that words can convey. For example, a professor might grade exams that are identified by student numbers rather than names.

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.

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 barista makes it. They do not need your name; they need an identifier. So even if your name is Joe, you can tell them it is Thor, Wotan, 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.

4.4.3.1 Make Names Informative

The most basic principle of naming is to choose names that are informative as doing so 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." Similarly, 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 pointing to a specific resource. For example, an International Standard Book Number (ISBN) like

“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 The MIT Press (262).

4.4.3.2 Use Controlled Vocabularies

One way to encourage good names is to establish a controlled vocabulary. A controlled vocabulary is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use.

A controlled vocabulary is not simply a set of allowed words; it also includes their definitions and often specific rules by which the vocabulary terms can be used and combined. Different domains can create separate controlled vocabularies, but it is important that the vocabulary is used consistently throughout that domain. 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.

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 (DNS) 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.

4.4.3.3 Allow Aliasing

A controlled vocabulary is extremely useful for people able to use it, but if you are designing an organizing system for people who do not or cannot use it, you need to accommodate the variety of words they will 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 familiar film name of “Nemo.”

The technique of "unlimited aliasing" connects the uncontrolled or natural vocabularies that people use with the controlled one employed by the organizing system. For example, the birth name of the 42nd President of the United States of America is "William Jefferson Clinton," but web pages that refer to him as "Bill Clinton" are vastly more common, and searches that use the former name are redirected to the latter. A related mechanism used by search engines is spelling correction which essentially treats all the incorrect spellings as aliases of the correct one ("did you mean California?" when you typed "Claifornia”).

4.4.3.4 Make Identifiers Unique or Qualified

Even though an identifier refers to a single resource, this does not 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 could use the same number to identify an electronic radio valve. While each identifier is unique in its inventory system, if a supply request gets sent to the wrong warehouse, hungry soldiers could be sent radio valves instead of rations.

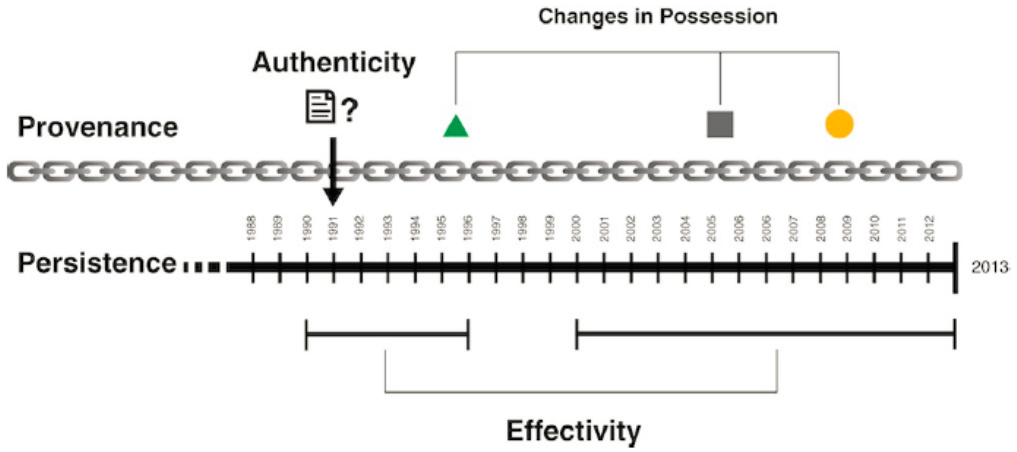
We can prevent or reduce identifier collisions by adding information about the namespace, the domain from which the names or identifiers are selected, thus creating what are called qualified

names. There are several dozen U.S. 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 to distinguish between them such as with <book:Title> and <legal:Title>.

4.5 Resources over Time

The questions 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 an organizing system evolves. We might need to know how a resource does or does not change over time (its persistence), or if it might not be available or relevant until a specified point in time (its effectivity). Additionally, we need to be able to determine whether the resource is what it is said to be (its authenticity), and sometimes who has certified its authenticity over time (its provenance). A resource might have persistence, but only the provenance provided by a documented chain of custody enables questions about authenticity to be answered with authority, and effectivity describes the limits of a resource's lifespan (see Figure 4.6).

Figure 4.6. Resources over Time.



Four considerations that arise with respect to the maintenance of resources over time are their persistence, provenance, authenticity, and effectivity.

4.5.1 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.

4.5.1.1 Persistent Identifiers

How long does an identifier need to last? Coyle gives the conventional, if unsatisfying, answer: "As long as it is needed." In some cases, the time frame is relatively short such as when you order a coffee and the barista asks for your name. For libraries and repositories of scientific, economic, census, or other data, the time frame might be "forever."

The design of a scheme for persistent identifiers must consider both the required time frame and the anticipated number of resources to be identified. When the Internet Protocol (IP) was designed in 1980, it contained a 32-bit address scheme, sufficient for over 4 billion unique addresses. However, the enormous growth of the Internet and the application of IP addresses to resources of unexpected types have required a new addressing scheme with 128 bits.

4.5.1.2 Persistent Resources

Even though persistence often has a technological 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 and governance are activities carried out to ensure the outcome of persistence. The subtle relationship between preservation and persistence raises 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. At some point we might question whether these activities have transformed the resource into a different one. Likewise, digital resources require regular backup and migration to keep them available, and this might include changing their digital format.

The Paradox of Theseus

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?

Philosophical pondering aside, it is usually better to think of persistence more abstractly and consider resources that remain functionally the same, even if their physical properties or information values change, as persistent.

4.5.2 Effectivity

Many resources, or their properties, have locative or temporal effectivity, meaning that they come into effect at a particular time and place, and they almost certainly will cease to be effective at some future date, or they may cease to be effective in different places. Temporal effectivity, sometimes known as "time-to-live," is expressed as a range of two dates. It consists of a date on which the resource is effective, and a date on which the resource ceases to be effective or becomes stale. For some types of resources, the effective date is the moment they are created, but for others the effective date can be a time different from the moment of creation. For example, a law passed in November may take effect on January 1 of the following year, and credit cards first need to be activated and then can no longer be used after their expiration date. An "effective date" is the counterpart of the "Best Before" date on perishable goods. That date indicates when a product goes bad, whereas an item's effective date is when it "goes good," and the resource that it supersedes needs to be disposed of or archived.

In Which Country Do You Live?

Even if you always live in the same place, the answer to "what country do you live in?" can depend on when it is asked. Consider the case of an elderly woman born in 1929 in Zemun, a district in the eastern European city of Belgrade, who has never moved. The place she lives has been part of seven different countries during her lifetime: Kingdom of Yugoslavia (1929-1941); Independent State of Croatia (1941-1945); Federal People's Republic of Yugoslavia (1945-1963); Socialist Federal Republic of Yugoslavia (1963-1992); Federal Republic of Yugoslavia (1992-2003); State Union of Serbia and Montenegro (2003-2006); Republic of Serbia (2007-present).

Locative effectivity considers borders, security, roadways, altitude, depth, and other geographic factors. Some types of resources, including people, are restricted as to where they may or may not be transported and used, such as hazardous cargo, explosives, narcotics, pharmaceuticals, or alcohol.

Effectivity concerns sometimes intersect with authority control for names and places. Name changes for resources often are tied to particular dates, events, and locations. Laws and regulations differ across organizational and geopolitical boundaries, and those boundaries often change. Some places that have been the site of civil unrest, foreign occupation, and other political disruptions have had many different names over time and even at the same time. Today these disputed borders cause a problem for Google Maps when it displays certain international borders. As Google is subject to the laws of the country where its servers are located, it must present disputed borders to conform with the point of view of the host country when a country-specific Google site is used to access the map.

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.

4.5.3 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. The importance and nuance of questions about authenticity can be seen in the many words used to describe the relationship between "the real thing" or 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. Alternatively, a third party can also serve as proof of authenticity, and many professional careers are based on figuring out if a resource is authentic.

For people and physical resources, there is a large body of techniques for establishing identity. One technique is to assess the physical integrity of recorded information when considering the integrity of its contents.

For digital resources, authenticity is often 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. While scholars are likely to trust technological methods, technologists are more skeptical 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 resource.

4.5.4 Provenance

The concept of provenance transforms the passive question of “what has happened to this resource?” into actions that can be taken to ensure nothing bad can happen to a resource or to detect possible changes. The idea that important documents must be created in a manner that can be authenticated and 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 access to the resources.

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.

However, it is not only art historians and custodians of critical documents that need to be concerned with provenance. If you are planning to buy a used car, it is wise to check the vehicle history (using the Vehicle Identification Number, the car's persistent identifier) to make sure it hasn't been wrecked, flooded, or stolen.

Chinese Manuscript with Provenance Seals

This beautiful manuscript, preserved in the National Palace Museum in Taipei, was created by Zhao Ji (赵佶), Emperor Huizong, the 8th Emperor of the Chinese Song Dynasty approximately a thousand years ago. He was famous for his skills in poetry, painting, and calligraphy. There are two poems here; the one on the right describes the techniques for Chinese landscape paintings, while the left one expresses the Emperor's appreciation of plum blossoms, which signal the onset of spring. The red seals are those of several Ching Dynasty emperors over many generations, with the oldest being at least five hundred years after Huizong created the poems. Stamping your personalized red seal on a resource is analogous to, but vastly more elegant and informative than, "Liking" a web page today. (Photo by R. Glushko.)

4.6 Key Points in Chapter Four

TBD