

Chapter 5
**Resource Description and
Metadata**

*Robert J. Glushko
Kimra McPherson
Ryan Greenberg
Robyn Perry
Matthew Mayernik
Graham Freeman
Carl Lagoze*

5.1. Introduction 215
5.2. An Overview of Resource Description 219
5.3. The Process of Describing Resources 227
5.4. Describing Non-text Resources 257
5.5. Key Points in Chapter Five 262

5.1 Introduction

This chapter is a turning point in the book. The earlier chapters have discussed the key ideas of the discipline of organizing: identifying and selecting the resources to organize, and then organizing and maintaining them and their organizing system. We have emphasized that finding things later is the most important reason for organizing them. This can be surprisingly hard to do. People know things by different names or remember different aspects of them.

“Arrangement in Grey and Black No. 1”



“Arrangement in Grey and Black No. 1” (1871). Oil on canvas, by James Abbott McNeill Whistler. Alternative titles: “Portrait of the Artist's Mother” and “Whistler's Mother.” The painting is in Musée d'Orsay in Paris.

(Photo by Jean-Gilles Berizzi Source: [Wikimedia Commons](#).)

The famous painting here by the 19th century American painter James Whistler is exhibited in the Musée d'Orsay in Paris, and has been described as a Victorian-era *Mona Lisa*. What name do you know it by? How should it be described?

Resource descriptions for art usually contain the name of the artist, the medium, the year of its completion, and, of course, its title. Most of these map fairly obviously to the properties they describe; the title, owing to its prominence and expressive power, is often an exception.

Most often, a painting's title describes its subject. If you recognize the previous painting, you most likely know it by its colloquial name, *Whistler's Mother*. While it is a portrait of Anna McNeill Whistler, mother of painter James Abbott McNeill Whistler, the artist gave it a radically different title, *Arrangement in Grey and Black, No. 1*, because he believed the most important property of a painting

was not the subject it depicted, but its purely aesthetic properties and their effect on the viewer. So Whistler named his paintings, which were mostly landscapes and portraits, in the manner of musical compositions: *Nocturne in Black and Gold*; *Symphony in White*; *Arrangement in Pink, Red, and Purple*; and so on.

If Whistler's title surprises you, because you would have described it as a portrait of an elderly woman, this helps reinforce how wildly different names of the same resource can be. Resource descriptions and metadata provide meaning, but to whom? What is salient about a resource can depend on the context in which it is experienced, and thus may change over time. Descriptions that make sense to some people might not make sense to others. People searching on the “wrong descriptions” or the “wrong metadata” will not find what they are looking for.

Mt. St. Helens, in the southwest corner of Washington State, was usually just described as a mountain until 1980. Then, the deadliest and most economically destructive volcanic event in the history of the United States blew away the top of the mountain, killing 57 people, and leaving a mile-wide crater. Today almost every description of Mt. St. Helens mentions the volcanic eruption.

It would seem impossible to search using the wrong description if the descriptions of a resource were kept current to include all the latest information, but search engines are already too powerful, usually producing too much information. Technology improvements in search and retrieval do not eliminate the cognitive effort to remember what things are, how they are best described, and where they might be found. The design of resource descriptions and metadata depends on why we need to find the information later. This chapter is about how and why.

Mt. St. Helens Before and After



Before 1980, Mt. St. Helens was a “postcard-like” snow-covered mountain. Afterward, the mile-wide crater where its mountaintop once was reminds us of its violent volcanic eruption.

(Credit: Public domain images from US Forest Service and USGS.)

Stop and Think: These Places Have Their Moments

Our description of Mt. St. Helens forever changed after its volcanic eruption. Surely there are times and places that you remember differently because of their part in an important event. A family wedding? The Olympic Games? A natural disaster? The Twin Towers?

It is easy to find before and after images of Mt. St. Helens doing a web search. What information might be associated with these images? Modern cameras assign an identifier to the stored photograph and they also capture the technical description of the image’s production: the type of camera, lens, shutter speed, light sensitivity, aperture, and other settings.^{228[Com]} Many modern cameras also record information about the geographic and temporal circumstances

surrounding the image’s creation: the date, time and location on Earth where the photograph is taken. When the image is transferred out of the camera and is published for all to see, it might be useful to record biographical information about the photographer to help viewers relate to the photographer and better understand the photograph’s context. There may also be different licenses and copyright information to associate with the picture—who owns it and how it can be used.

Consider a completely different context. Four 7-year old boys are selecting Lego blocks to complete their latest construction. The first boy is looking for “cylinder one-ers,” another for “coke bottles,” the third for “golder wipers,” and the final boy is looking for “round one-bricks”? It turns out, they are all the same thing; each boy has devised his own set of descriptive terms for the tiny building blocks. Some of their many descriptions are based on color alone (“redder”), some on color and shape (“blue tunnel”), some on role (“connector”), some on common cultural touchstones (“light saber”). Others, like “jail snail” and “slug,” seem unidentifiable—unless, of course, you happen to be inside the mind of a particular 7-year-old kid. It doesn't matter if the boys use different description vocabularies when they play by themselves, but they will have to agree if they play together.²²⁹[CogSci]

Paintings, digital photos, and Lego blocks are all very different, but together these scenarios raise important questions about describing resources that we attempt to answer in this chapter:

- What is the purpose of resource description?
- What resource properties should be described?
- How are resource descriptions created?
- What makes a good resource description?

Navigating This Chapter

We begin with an overview of *resource description* (§5.2), which we propose as a broad concept that includes the narrower concepts of *bibliographic descriptions* and *metadata*. §5.3 *The Process of Describing Resources* (page 227) describes a 7-step process of describing resources that includes determining scope, focus and purposes, identifying resource properties, designing the description vocabulary, designing the description form and implementation, and creating and evaluating the descriptions. Because many principles and methods for resource description were developed for describing text resources in physical formats, in §5.4 *Describing Non-text Resources* (page 257) we briefly discuss the issues that arise when describing museum and artistic resources, images, music, video, and contextual resources.

5.2 An Overview of Resource Description

We describe resources so that we can refer to them, distinguish among them, search for them, manage access to them, preserve them, and make predictions about what might happen to them or what they might do. Each purpose may require different *resource descriptions*. We use *resource descriptions* in every communication and conversation; they are the enablers of organizing systems.

5.2.1 Naming {and, or, vs.} Describing

Chapter 4 discussed how to decide what things should be treated as resources and how names and identifiers distinguish one resource from another. Names can suggest the properties and principles an organizing system uses to arrange its resources. We can see how societies organize their people by noting that among the most common surnames in English are descriptions of occupations (Smith, Miller, Taylor), descriptions of kinship relations (Johnson, Wilson, Anderson), and descriptions of appearance (Brown, White).^{230[Ling]}

In many cultures, one spouse or the other takes a name that describes their marital relationship. In many parts of the English-speaking world, married women have often referred to themselves using their husband's name.^{231[Ling]}

Similarly, many other kinds of resources have names that are property descriptions, including buildings (Pentagon, White House), geographical locations (North America, Red Sea), and cities (Grand Forks, Baton Rouge).

Every resource can be given a name or identifier. Identifiers are especially efficient *resource descriptions* because, by definition, identifiers are unique over some domain or collection of resources. Names and identifiers do not typically describe the resource in any ordinary sense because they are usually assigned to the resource rather than recording a property of it.

However, the arbitrariness of names and identifiers means that they do not serve to distinguish resources for people who do not already know them. This is why we use what linguists call referring expressions or definite descriptions, like “the small black dog” rather than the more efficient “Blackie,” when we are talking to someone who does not know that is the dog's name.^{232[CogSci]}

Similarly, when we use a library catalog or search engine to locate a known resource, we query for it using its name, or some specific information we know about it, to make it easier to find. In contrast, when we look for resources to satisfy an information need but do not have specific resources in mind, we query for them using descriptions of their content or other properties. In general, information retrieval can be characterized as comparing the description of a user's needs with descriptions of the resources that might satisfy them.

5.2.2 “Description” as an Inclusive Term

Up to now we have used the concept of “description” in its ordinary sense to mean the labeling or explaining of the visible or important features that characterize or represent something. However, the concept is sometimes used more precisely in the context of organizing systems, where *resource description* is often more formal, systematic, and institutional. In the library science context of *bibliographic description*, a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources. In the contexts of conceptual modeling and information systems design, the terms in *resource descriptions* are also called “keywords,” “index terms,” *attributes*, *attribute values*, *elements*, “data elements,” “data values,” or “the vocabulary.” In business intelligence, predictive analytics or other data science contexts these are called “variables,” “features,” *properties*, or “measurements.” In contexts where descriptions are less formal or more personal the description terms are often called “labels” or “tags.” Rather than attempt to make fine distinctions among these synonyms or near-synonyms, we will use “description” as an inclusive term except where conventional usage overwhelmingly favors one of the other terms.

Many of these terms come from a narrow semantic scope in which the purpose of description is to identify and characterize the essence, or *aboutness*, of a resource. However, as it becomes trivial to associate computationally generated information with resources, many additional kinds of information beyond strict “aboutness” can support additional interactions. We describe many of these purposes and the types of information needed to enable them in §5.3.2 **Determining the Purposes** (page 234). We apply *resource description* in an expansive way to accommodate all of them.

Chapter 4 introduced the distinction of §4.2.4 **Resource Focus** (page 178) to contrast primary resources with resources that describe them, which we called Description Resources. We chose this term as a more inclusive and more easily understood alternative to two terms that are well established in organizing systems for information resources: *bibliographic descriptions* and *metadata*. We will also distinguish resource description as a general concept from the narrower senses of statistical description, tagging of web resources, and the *Resource Description Framework (RDF)* language used to make statements about web resources and physical resources that can be identified on the Web.

5.2.2.1 Bibliographic Descriptions

The purposes and nature of bibliographic description are the foundation of library and information science and have been debated and systematized for nearly two centuries. *Bibliographic descriptions* characterize information re-

sources and the entities that populate the bibliographic universe, which include works, editions, authors, and subjects.

A bibliographic description of an information resource is typically realized as a structured record in a standardized format that describes a specific resource.

The computerization of bibliographic records made them easier to use as aids for finding resources. However, digitizing legacy printed card-oriented descriptions for online use was not a straightforward task because the descriptions had been created according to cataloging rules designed for collections of books and other physical resources and intended only for use by people.

5.2.2.2 Metadata


Metadata is often defined as “data about data,” a definition that is nearly as ubiquitous as it is unhelpful. A more content-full definition of metadata is that it is structured description for information resources of any kind. Metadata is more useful when supported by a metadata schema that defines the elements in the structured description.

The concept of metadata originated in information systems and database design in the 1970s, so it is much newer than that of bibliographic description. The earliest metadata schemas, called data dictionaries, documented the arrangement and content of data fields in the records used by transactional applications on mainframe computers. A more sophisticated type of metadata emerged as the documentation of the data models in database management systems, called database schemas, which described the structure of relational tables, attribute names, and legal data types and values for content.

In 1986, the *Standard Generalized Markup Language (SGML)* formalized the *Document Type Definition (DTD)* as a metadata form for describing the structure and content elements in hierarchical and hypertextual document models. SGML was superseded in 1997 by eXtensible Markup Language (XML), whose purpose was structured and computer-processable web content.^{235[Com]}

Today, XML schemas and other web- and compute-friendly formats for resource description have broadened the idea of *resource description* far beyond that of bibliographic description to include the description of software components, business and scientific datasets, web services, and computational objects in both physical and digital formats. The resource descriptions themselves serve to enable discovery, reuse, access control, and the invocation of other resources needed for people or computational agents to effectively interact with the primary ones described by the metadata.^{236[Com]}

5.2.2.3 Tagging of Web-based Resources



Tags on Last.fm

lost.fm Music Search Music Listen Events Charts Originals Join Login

Female vocalists Artists Albums Tracks Videos Wiki

This tag describes any musical artist with a female singer as the centerpiece of the vocals. This tag is not limited to any one musical genre or era, and includes artists ranging from Billie Holiday to Lady Gaga. [View history](#)

Music tagged "female vocalists" [Play female vocalists Tag](#)

Related tags: [single-singer](#) [r&b](#) [pop](#) [female vocalists](#) [rock](#) [country](#) [search](#)

Top Artists

Tori Amos Kate Bush Lana Del Rey PJ Harvey

Recently Added

Tuesday Night Music Club
Released: 17 Nov 2009 (13 tracks)

Fallen
Released: 11 Sep 2009 (13 tracks)

No Need to Argue
The Cranberries
Released: 14 Jul 2009 (13 tracks)

Backferry
Released: 2 Feb 2009 (13 tracks)

Free Music Downloads

Wayland (140)
Larkland

Twice (Little Dragon Cover) (8:31)
Malmgren

Keeper of Secrets (1:23)
Theatre Des Vampires

Hyped Artists

Last.fm analyzes tags and other metadata to create rich multimedia “discovery” pages that bring together artist catalogs, new songs, free downloads, and music videos that its algorithm predicts will satisfy a user’s taste. This allows users to browse for new music in a more intuitive manner than searching by artist or genre.

(Screenshot by Ian MacFarland.)

The concept of metadata has been extended to include the tags, ratings, bookmarks or other types of descriptions that individuals apply to individual photos, blog or news items, or any other resource with a web presence. The practice of tagging has emerged as a way to apply labels to content in order to describe and identify it. Sets of tags are useful in managing one’s collection of websites or digital media, in sharing them with others, and enabling new types of interactions and services.^{237[IA]} For example, users of Last.fm tag music with labels that describe its nature, era, mood, or genre, and Last.fm uses these tags to generate radio stations that play music similar to that tag and related tags.

But tagging has a downside. The tendency for users to tag intuitively and spontaneously revives the vocabulary problem (§4.4) because one photographer’s “tree” is another’s “oak.” Likewise, unsystematic word choice leads to morphological inconsistency (§6.4.3 Relationships among Word

Forms (page 293)); the same photo might be tagged with “burning” and “trees” and also with “burn” and “tree” by another. This disparity in the descriptors people use to categorize the same or similar resources can turn systems that use tagging into a “tag soup” lacking in structure.^{238[IA]}

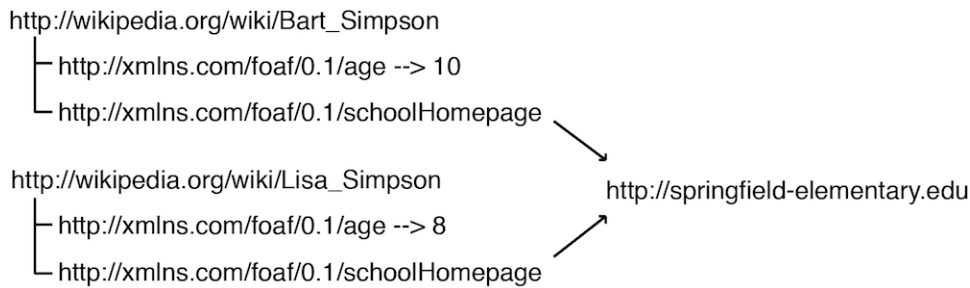
Some social media sites have incorporated mechanisms to make the tagging activity more systematic and to reduce *vocabulary problems*. For example, on Facebook, users can indicate that a specific person is in an uploaded picture by clicking on the faces of people in photographs, typing the person’s name, and then selecting the person from a list of Facebook friends whose names are formatted the way they appear on the friend’s profile. Some social media systems suggest the most popular tags, perform morphological normalization, or allow users to arrange tags in bundles or hierarchies.^{239[Web]}

5.2.2.4 Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a standard model for making computer-processable statements about web resources; it is the foundation for the vision of the *Semantic Web*.^{240[Web]} We have been using the word “resource” to refer to anything that is being organized. In the context of RDF and the web, however, “*resource*” means something more specific: a resource is anything that has been given a Uniform Resource Identifier (URI). URIs can take various forms, but you are probably most familiar with the URIs used to identify web pages, such as <http://springfield-elementary.edu/>. (You are probably also used to calling these URLs instead of URIs.) The key idea behind RDF is that we can use URIs to identify not only things “on” the web, like web pages, but also things “off” the web like people or countries. For example, we might use the URI <http://springfield-elementary.edu/> to refer to Springfield Elementary itself, and not just the school’s web page.

RDF models all descriptions as sets of “triples,” where each *triple* consists of the resource being described (identified by a URI), a property, and a value. Properties are resources too, meaning they are identified by URIs. For example, the URI <http://xmlns.com/foaf/0.1/schoolHomepage> identifies a property defined by the *Friend of a Friend (FOAF)* project for relating a person to (the web page of) a school they attended. Values can be resources too, but they do not have to be: when a property takes simple values like numbers, dates, or text strings, these values do not have URIs and so are not resources.

Because RDF uses URIs to identify described resources, their properties, and (some) property values, the triples in a description can be connected into a network or *graph*. **Figure 5.1, RDF Triples Arranged as a Graph**, shows four triples that have been connected into a graph. Two of the triples describe Bart Simpson, who is identified using the URI of his Wikipedia page.^{241[Web]} The other two describe Lisa Simpson. Two of the triples use the property `age`, which takes a simple number value. The other two use the property `schoolHomepage`, which takes a resource value, and in this case they happen to have the same resource (Springfield Elementary’s home page) as their value.

Figure 5.1. RDF Triples Arranged as a Graph.

Two RDF triples can be connected to form a graph when they have a resource, property, or value in common. In this example RDF triples that make a statement about the home page of the elementary school attended by Bart Simpson and Lisa Simpson can be connected because they have the same value, namely the URI for Springfield Elementary.

Using URIs as identifiers for resources and properties allows descriptions modeled as RDF to be interconnected into a network of “linked data,” in the same way that the web enabled information to be interconnected into a massive network of “linked documents.” Proponents of RDF claim that this will greatly benefit knowledge discovery and inference.^{242[Web]} But the benefits of RDF’s highly prescriptive description form must be weighed against the costs; turning existing descriptions into RDF can be labor-intensive.

5.2.2.5 Aggregated Information Objects

In the pre-digital age, information objects came with explicit tangible boundaries. Books consisted of pages bound within a cover, a vinyl record album physically bound together a set of songs (you could even see the groove pattern separating the songs), a movie was delivered on a strip of film spooled onto a reel, and a collection was (usually) demarcated as a designated shelf or room in a library.

Boundaries of information objects in the digital realm are neither tangible nor obvious. Consider the simple notion of a web page. Our cognitive notion of that which is rendered in our browser window (e.g., some formatted text with an associated image) is actually, in web architecture terms (Jacobs & Walsh, 2004), three information objects (aka resources); the HTML encoding the text, CSS that defines the formatting rules, and the JPEG that encodes the image. All three have URLs and can independently be retrieved and linked. The situation is even more ambiguous for the common notion of a web site, the boundaries of which are not defined technically and are cognitively difficult to express.

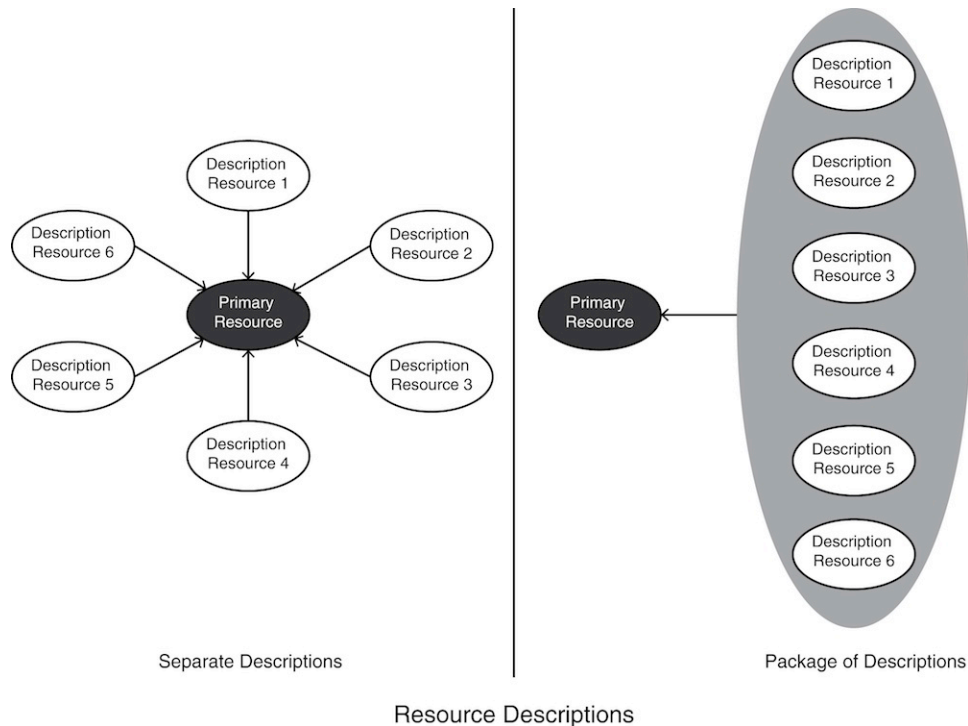
Aggregations can be convenience methods for simplifying dissemination or organization, but they can also be transformative; resources can derive nearly all their value from their inclusion in an aggregation. On a web page, the CSS file is virtually useless on its own, since its role is to style the HTML file. In iTunes, the playback and organization functions are optimized for pop music, where individual songs can usually stand on their own when separated from the rest of an album. Classical music fans often struggle with this, because the individual “tracks” of a recording, split up to reduce file size and facilitate navigation through long works, are not separable; pieces are meant to be listened to in their entirety, and it can be difficult to ensure that they are aggregated together and have the proper metadata assigned to their aggregations. In other words: you can't listen to symphonies on shuffle.^{244[1A]}

The problem here is how to architecturally and technically express the notion of an aggregation, a set of information objects that, when considered together, compose another named information object. Aggregations are prevalent all over our digital information space: the web page and site mentioned above; a scholarly publication consisting of text, figures, and data; a dataset that is the composition of multiple data files. Notably the notion is both recursive and non-exclusive. An object that is itself an aggregation may be aggregated into another object. Information objects included in one aggregation may also be included in other aggregations, allowing reuse and re-factoring of existing information objects. A solution to this problem is a critical aspect of organizing digital information because, without well-defined boundaries we cannot deterministically identify, reference, or describe information objects.

5.2.3 Frameworks for Resource Description

The broad scope of resources to which descriptions can be applied and the different communities that describe them means that many *frameworks* and classifications have been proposed to help make sense of resource description.

Figure 5.2. Architectures for Resource Description.



Two contrasting architectures for resource descriptions are separate descriptions versus packaged descriptions, which were dominant in library catalogs with printed cards containing descriptions about a resource.

The dominant historical view treats resource descriptions as a package of statements; this view is embodied in the printed library card catalog and its computerized analog in the MARC21 format (an exchange format for library catalog records), which contains many fields about the bibliographic characteristics of an object like author, title, publication year, publisher, and pagination. An alternate architecture for resource description focuses on each individual description or assertion about a single resource, as the RDF and linked data approaches do. These two alternatives are contrasted in **Figure 5.2, Architectures for Resource Description**.

In either case, these common ways of thinking about resource description emphasize—or perhaps even overemphasize—two implementation decisions:

- The first is whether to combine multiple resource descriptions into a structural package or to keep them as separate descriptive statements.
- The second is the choice of syntax in which the descriptions are encoded.

Both of these implementation decisions have important implications, but are secondary to the questions about the purposes of resource description, how resource properties are selected as the basis for description, how they are best created, and other logical or design considerations. In keeping with a fundamental idea of the discipline of organizing (introduced in §1.6 *The Concept of “Organizing Principle”* (page 43)), it is imperative to distinguish design principles from implementation choices. We treat the set of implementation decisions about character notations, syntax, and structure as the *form* of resource description and we will defer them as much as we can until *Chapter 9, The Forms of Resource Descriptions*.

Resource description is not an end in itself. Its many purposes are all means for enabling and using an organizing system for some collection of resources. As a result, our framework for resource descriptions aligns with the activities of organizing systems we discussed in *Chapter 3*: selecting, organizing, interacting with, and maintaining resources.

5.3 The Process of Describing Resources

We prefer the general concept of resource description over the more specialized ones of bibliographic description and metadata because it makes it easier to see the issues that cut across the domains where those terms dominate. In addition, it enables us to propose more standard process that we can apply broadly to the use of resource descriptions in organizing systems. A shared vocabulary enables the sharing of lessons and best practices.

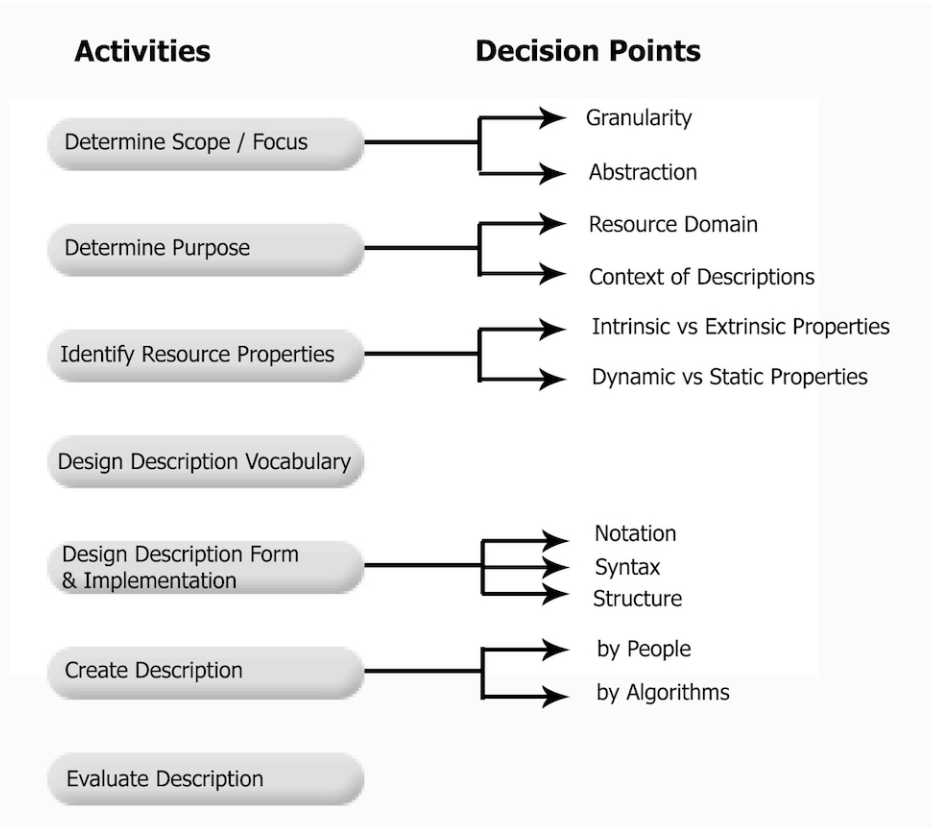
The process of describing resources involves seven interdependent and iterative steps. We begin with a generic summary of the process to set the stage for a detailed step-by-step discussion.

1. Identifying resources to describe is the first step; this topic is covered in detail in §4.3 *Resource Identity* (page 182). The resource *domain* and *scope* circumscribe the describable properties and the possible purposes that descriptions might serve. The resource *focus* determines which are primary information resources and which ones are treated as the corresponding resource descriptions. Two important decisions at this stage are *granularity* of description—are we describing individual resources or collections of them?

—and the *abstraction* level—are we describing resource instances, parts of them, or resource types?

2. Generally, the purpose of resource description is to support the activities common to all organizing systems: selecting, organizing, interacting with, and maintaining resources, as we saw in **Chapter 3**. The particular resource domain and the context in which descriptions are created and used imposes more specific requirements and constraints on the purposes that resource description can serve.
3. Once the purposes of description in terms of activities and interactions have been determined, the specific properties of the resources that are needed to enable them can be identified. The goal of description is not to be exhaustive; there are always more possible properties than can be reasonably described. Instead, the challenge is to use the properties that are most robust and reliable for supporting the desired interactions.
4. This step includes several logical and semantic decisions about how the resource properties will be described. What terms or element names should be used to identify the resource properties we have chosen to describe? Are there rules or constraints on the types of data or values that the property descriptions can assume? When dealing with numerical descriptions, their data types and levels of measurement constrain the kinds of processing to which they may submit. Nominal, ordinal, interval, and ratio data each are limited to particular transformations based on what they represent. A good description vocabulary will be easy to assign when creating resource descriptions and easy to understand when using them.
5. The logical and semantic decisions about the description vocabulary are reified by decisions about the notation, syntax and structure of the descriptions. Taken together, these decisions collectively determine what we call the *form* or *encoding* of the resource descriptions. The implementation of the descriptions involves decisions about how and where they are stored and the technology used to create, edit, store, and retrieve them.
6. Resource descriptions are created by individuals, by informal or formal groups of people, or by automated or computational means. Some types of descriptions can only be created by people, some types of descriptions can only be created by automated or algorithmic techniques, and some can be created in either manner.
7. The resource descriptions must be evaluated with respect to their intended purposes. The results of this evaluation will help determine which or the preceding steps need to be redone.

The next seven sub-sections discuss each of these steps in detail. A quick reference guide is **Figure 5.3, The Process of Describing Resources**.

Figure 5.3. The Process of Describing Resources.

The process of describing resources consists of seven steps: Determining the scope and focus, determining the purpose, identifying resource properties, designing the description vocabulary, designing the description form and implementation, creating the descriptions, and evaluating the descriptions.

How explicit and systematic each step needs to be depends on the resource domain and scope, and especially on the intended users of the organizing system. If we look carefully, we can see most of these steps taking place even in very informal contexts, like the kids playing with Lego blocks with which we started this chapter. The goal of building things with the blocks leads the boys to identify which properties are most useful to analyze. They develop descriptions of the blocks that capture the specific values of the relevant properties. Finally, they evaluate their descriptions by using them when they play together; it becomes

immediately obvious that a description is not serving its purpose when one boy hands a block to another that was not the one he thought he had asked for.

In contrast, a picture-taking scenario involves a much more explicit and systematic process of resource description. The resource properties, description vocabulary, and description form used automatically by a digital camera were chosen by an industry association and published as a technical specification implemented by camera and mobile phone manufacturers worldwide.

The resource descriptions used by libraries, archives, and museums are typically created in an even more explicit and systematic manner. Like the descriptions of the digital photo, the properties, vocabulary, and form of the descriptions used by their organizing systems are governed by standards. However, there is no equivalent to the digital camera that can create these descriptions automatically. Instead, highly trained professionals create them meticulously.

A great many resources and their associated descriptions in business and scientific organizing systems are created by automated or computational processes, so the process of describing individual resources is not at all like that in libraries and other memory institutions. However, the process for designing the data models or schemas for the class of resources that will be generated is equally systematic and is typically performed by highly skilled data analysts and data modelers.

5.3.1 Determining the Scope and Focus

Which resources do we want to describe? As we saw in [Chapter 4](#), determining what will be treated as a separate resource is not always easy, especially for resources with component parts and for information resources where the most important property is their content, which is not directly perceivable. Identifying the thing you want to describe as precisely as practical is the first step to creating a useful description.

In [§4.2.4 Resource Focus \(page 178\)](#), we introduced the contrast between primary resources and description resources, which we called resource focus. Determining the resource focus goes hand in hand with determining which resources we intend to describe; these often arbitrary decisions then make a huge difference in the nature and extent of resource description. One person's metadata is another person's data.

- For a librarian, the price of a book might be just one more attribute that is part of the book's record.
- For an accountant at a bookstore, the price of that book—both the cost to buy the book and the price at which it is then sold to customers—is critical information for staying in business.

- In a medical records context, a patient’s insurance provider isn’t of much concern to the doctor, but to the person responsible for billing, it is central. For the nurse, the patient’s current vital signs may be of most importance, while for the doctor, it may be most important to understand how those data in aggregate serve to indicate a longer-term prognosis of the patient’s health.
- A scientist studying comparative anatomy preserves animal specimens and records detailed physical descriptions about them, but a scientist studying ecology or migration discards the specimens and focuses on describing the context in which the specimen was located.

5.3.1.1 Describing Instances or Describing Collections

It is simplest to think of a resource description as being associated with another individual resource. As we discussed in [Chapter 4](#), it is challenging to determine what to treat as an individual resource when resources are themselves objects or systems composed of other parts or resources. For example, we sometimes describe a football team as a single resource and at other times we focus on each individual player. However, after deciding on resource granularity, the question remains whether each resource needs a separate description.

Libraries and museums specialize in curating resource descriptions about the instances in their collections. Resource descriptions are also applied to classes or collections of resources, because a collection is also a resource (§1.4 [The Concept of “Collection” \(page 37\)](#)). Archives and special collections of maps are typically assigned resource descriptions, but each document or map contained in the collection does not necessarily have its own bibliographic description. Similarly, business and scientific datasets are invariably described at collection-level granularity because they are often analyzed in their entirety.

Furthermore, the granularity of description for a collection of resources tends to differ for different users or purposes. An investor who owns many different stocks focuses on their individual prices, while other investors put their money in index funds that combine all the separate prices into a single value.

Many web pages, especially e-commerce product catalogs and news sites, are dynamically assembled and personalized from a large number of information resources and services that are separately identified and described in content management and content delivery systems. However, a highly complex collection of resources that comes together in a single page is treated as a single resource when that page appears in a list of search engine results. Moreover, all of the separately generated pages can be given a single description when a user creates a bookmark to make it easy to return to the home page of the site.

5.3.1.2 Abstraction in Resource Description

We can also associate resource descriptions with an entire type or domain of resources. (See §3.5.2.4 **Preserving Resource Types** (page 138) and §4.2.1 **Resource Domain** (page 167).) A collection of resource descriptions is vastly more useful when every resource is described using common description elements or terms that apply to every resource. A *schema* (or model, or metadata standard) specifies the set of descriptions that apply to an entire resource type. Sometimes this schema, model, or standard is inferred from or imposed on a collection of existing resources to ensure more consistent definitions, but more often, it is used as a specification when the resources are created or generated in the first place. (See **What about “Creating” Resources?** (page 90) in §3.1 **Introduction** (page 87).)

A relational database, for example, is easily conceptualized as a collection of records organized as one or more tables, with each record in its own row having a number of fields or attributes that contain some prescribed type of content. Each record or row in the database table is a description of a resource—an employee, a product, anything—and the individual attribute values, organized by the columns and rows of the table, are distinct parts of the description for some particular resource instance, like employee 24 or product 8012C.^{251[Com]}

The information resources that we commonly call documents are, by their nature, less homogeneous in content and structure than those that can be managed in databases. Document schemas, commonly represented in SGML or XML, usually allow for a mixture of data-like and textual descriptive elements.

XML schema languages have improved on SGML and XML by expressing the description of the document schema in XML itself, making it easy to create resources using the metadata as a template or pattern. XML schemas are often used as the specifications for XML resources created and used by information-intensive applications; in this context, they are often called XML vocabularies. XML schemas can be used to define web forms that capture resource instances (each filled-out form). XML schemas are also used to describe the interfaces to web services and other computational resources.^{252[Com]}

5.3.1.3 Scope, Scale, and Resource Description

If we only had one thing to describe, we could use a single word to describe it: “it.” We would not need to distinguish it from anything else. A second resource implies at least one more term in the description language: “not it.” However, as a collection grows, descriptions must become more complex to distinguish not only between, but also among resources.

Every element or term in a description language creates a dimension, or axis, along which resources can be distinguished, or it defines a set of questions

about resources. Distinctions and questions that arise frequently need to be easy to address, such as:

- What is the name of the resource?
- Who created it?
- What type of content or matter does it contain?

Therefore, as a collection grows, the language for describing resources must become more rigorous, and descriptions created when the collection was small often require revision because they are no longer adequate for their intended purposes.

Because the task of library resource description has been standardized at national and international levels, cataloging work is distributed among many describers whose results are shared. The principle of standardization has been the basis of centralized bibliographic description for a century.

Centralized resource description by skilled professionals works for libraries, but even in the earliest days of the web many library scientists and web authoring futurists recognized that this approach would not scale for describing web resources. In 1995, the *Dublin Core (DC)* metadata element set with only 15 elements was proposed as a vastly simpler description vocabulary that people not trained as professional catalogers could use. Since then, the Dublin Core initiative has been highly influential in inspiring numerous other communities to create minimalist description vocabularies, often by simplifying vocabularies that had been devised by professionals for use by non-professionals.

Of course, a simpler description vocabulary makes fewer distinctions than a complex one; replacing “author,” “artist,” “composer” and many other descriptions of the person or non-human resource responsible for the intellectual content of a resource with just “creator” (as Dublin Core does) results in a substantial loss of precision when the description is created and can cause misunderstanding when the descriptions are reused.²⁵⁶[CogSci]

The negative impacts of growing scope and scale on resource description can sometimes be avoided if the ultimate scope and scale of the organizing system is contemplated when it is being created. It would not be smart for a business with customers in six US states to create an address field in its customer database that only handled those six states; a more extensible design would allow for any state or province and include a country code. In general, however, just as there are problems in adapting a simple vocabulary as scope and scale increase, designing and applying resource descriptions that will work for a large and continuously growing collection might seem like too much work when the collection at hand is small.

The challenges that arise with large description vocabularies are transformed when resource descriptions are created and assigned by computer algorithms. A large dataset might contain many thousands of descriptions for each resource, but clearly the computer does not have cognitive difficulty generating or using them. However, computer models with this many features can be hard for people to understand and trust.

5.3.2 Determining the Purposes

Resource description serves many purposes, and the mix of purposes and the resulting kinds of descriptions in any particular organizing system depends on the scope and scale of the resources being organized. We can identify and classify the most common purposes using the four activities that occur in every organizing system: selecting, organizing, interacting with, and maintaining resources (see [Chapter 3](#)). Resource description also has a more open-ended purpose in *sensemaking* and science (see [§5.3.2.5](#)); we observe and describe the world to make sense of our experiences and to predict future observations.

5.3.2.1 Resource Description to Support Selection

Defining selection as the process by which resources are identified, evaluated, and then added to a collection in an organizing system emphasizes resource descriptions created by someone other than the person who is using them. We can distinguish several different ways in which resource description supports selection:

Discovery

What available resources might be added to a collection? New resources are often listed in directories, registries, or catalogs. Some types of resources are selected and acquired automatically through subscriptions or contracts.

Capability and Compatibility

Will the resource meet functional or interoperability requirements? Technology-intensive resources often have numerous specialized types of descriptions that specify their functions, performance, reliability, and other “ilities” that determine if they fit in with other resources in an organizing system. ^{257[Com]} Some services have qualities of service levels, terms and conditions, or interfaces documented in resource descriptions that affect their compatibility and interoperability. Some resources have licensing or usage restrictions that might prevent the resources from being used effectively for the intended purposes. Decisions about “people selection” are becoming more data-driven, and sports teams, business employers, and dating sites now rely on predictive statistics to find the best person.

Authentication

Is the resource what it claims to be? (§4.5.3 Authenticity (page 202)) Resource descriptions that can support authentication include technological ones like time stamps, watermarking, encryption, checksums, and digital signatures. The history of ownership or custody of a resource, called its provenance (§4.5.4 Provenance (page 203)), is often established through association with sales or tax records. Import and export certificates associated with the resource might be required to comply with laws designed to prevent the theft of antiquities or the transfer of technology or information with national security or foreign policy implications.

Appraisal

What is the value of this resource? What is its cost? At what rate does it depreciate? Does it have a shelf life? Does it have any associated ratings, rankings, or quality measures? Moreover, what is the quality of those ratings, rankings and measures?

We also consider the perspective of the person creating the resource description and his or her primary purpose, which is often to encourage the selection of the resource by someone else. Product marketing is about devising names and descriptions to make a resource distinctive and attractive compared to alternatives. For many years prunes were promoted as a dietary supplement that people (especially old ones) need to “maintain regularity.” But after the California Prune Board (the world’s biggest supplier) re-branded them as “dried plums” and started marketing them as a snack food (and simultaneously renaming itself as the California Dried Plum Board) sales increased significantly.^{258[Bus]}

Many countries require that imported goods are labeled with their country or origin. Consumers often use this property in resource descriptions as an indicator of high quality, as they might with Swiss watches, French or Italian fashions, or Canadian bacon. Alternatively, consumers might want to buy domestic or locally-sourced goods out of economic patriotism or to comply with procurement regulations. Not surprisingly, when consumers view origin in a positive light, this information is conspicuous and easy to read. In contrast, when consumers view origin less positively, perhaps as a warning of low quality goods, the supplier is likely to make the origin information as inconspicuous as legally possible, or might even misrepresent the goods as domestic ones.^{259[Bus]}

This misrepresentation is also ubiquitous in online dating, though the amount of misrepresentation must be balanced with goals of the relationship and chances of the deception being discovered.^{260[CogSci]}

5.3.2.2 Resource Description to Support Organizing

We have defined *organizing* as specifying the principles for describing and arranging resources to create the capabilities upon which interactions are based. This definition treats the creation of resource descriptions and their use to organize resources for interactions as separate and sequential activities. This is easiest to see when people assign keywords and classifications to documents, or when sensors produce data, and these resource descriptions are later used to enable document retrieval or data analysis. A department store clerk might sort dress shirts on a display table using labels that describe their brands, sizes, and other properties. Rules governing the collection, integration, and analysis of personal information are also resource descriptions that influence the organization of information resources.

However, even if resource description and resource organization are logically separable, at times they are intertwined. When you arrange your own clothes, you don't use explicit resource descriptions and instead rely on implicit ones about easily perceived properties like color, shape, and material of composition. When algorithms rather than people analyze texts to identify descriptive features for applications like information retrieval, spam classification, and sentiment analysis, resource descriptions and resource organization co-evolve, often continuously as the algorithm adapts and learns with each new resource it describes. This tight connection between resource description and resource organization is also exploited in organizing systems that use usage records from session logs, browsing, or downloading activities as interaction resources, tying them to payments for using the resources or analyzing them to influence the selection and organizing of resources in future personalized interactions. (See §1.9 The Concept of "Interaction Resource" (page 51))

5.3.2.3 Resource Description to Support Interactions

Most discussions of the purposes of resource descriptions and metadata emphasize the interactions that are based on resource descriptions that have been intentionally and explicitly assigned. The Functional Requirements for Bibliographic Records (FRBR), defined by library scientists, specifies the four interactions of Finding, Identifying, Selecting, and Obtaining resources, but these apply generically to organizing systems, not just those in libraries.

Finding

What resources are available that "correspond to the user's stated search criteria" and thus can satisfy an information need? Modern users accept that computerized indexing makes search possible over not only the entire description resource, but often over the entire content of the primary resource. Businesses search directories for descriptions of company capabilities to

find potential partners, and they also search for descriptions of application interfaces (APIs) that enable them to exchange information in an automated manner.

Identifying

Another purpose of resource description is to enable a user to confirm the identity of a specific resource or to distinguish among several that have some overlapping descriptions. Computer processable resource descriptions like bar codes, QR codes, or RFID tags are also used to identify resources. In Semantic Web contexts, URIs serve this purpose. Color can be used as resource descriptions when physical resources need to be identified quickly.^{262[CogSci]}

Selecting

Selecting in this context means the user activity of using resource descriptions to support a choice of resource from a collection, not the institutional activity of selecting resources for the collection in the first place. Search engines typically use a short “text snippet” with the query terms highlighted as resource descriptions to support selection. People often select resources with the least restrictions on uses as described in a Creative Commons license.^{263[Law]} A business might select a supplier or distributor that uses the same standard or industry reference model to describe its products or business processes because it is almost certain to reduce the cost of doing business with that business partner.^{264[Bus]}

Obtaining

Physical resources often require significant effort to obtain after they have been selected. Catching a bus or plane involves coordinating your current location and time with the time and location the resource is available. With information resources in physical form, obtaining a selected resource usually meant a walk through the library stacks. With digital information resources, a search engine returns a list of the identifiers of resources that can be accessed with just another click, so it takes little effort to go from selecting among the query results to obtaining the corresponding primary resource.^{265[Web]}

Elaine Svenonius proposed that a fifth task called Navigation be added to the FRBR list, and in 2016 that happened but it was renamed as “Explore”:

Navigation or Explore

If users are not able to specify their information needs in a way that the *finding* functionality requires, they should be able to use relational and structural descriptions among the resources to navigate from any resource to other ones that might be better. Svenonius emphasizes generalization, aggregation, and derivational relationships. But in principle, any relationship or property could serve as the navigation “highway” between resources.

What some authors call “structural metadata” can be used to support the related tasks of moving within multi-part digital resources like electronic books, where each page might have associated information about previous, next, and other related pages. Documents described using XML models can use *Extensible Stylesheet Language Transformations (XSLT)* and *XPath* to address and select data elements, sub-trees, or other structural parts of the document.^{268[Com]}

5.3.2.4 Resource Description to Support Maintenance

Many types of resource descriptions that support selection (§5.3.2.1 **Resource Description to Support Selection (page 234)**) are also useful over time to support maintenance of specific resource and the collection to which they belong. In particular, technical information about resource formats and technology (software, computers, or other) needed to use the resources, and information needed to ensure resource integrity is often called “preservation metadata” in a maintenance context.

Resource descriptions that are more exclusively associated with maintenance activities include version information and effectivity, or useful life information. Equipment maintenance schedules are typically related to the number of miles driven (indicated by a car’s odometer), number of hours operated (stored by many engines), number of pages printed, or other easily recorded information about resource use or interactions. With smart resources now capable of capturing, analyzing, and communicating more data about real-time performance, more sophisticated prediction and scheduling of maintenance work is now possible. It is also easier to identify resources that are not being used as much as expected, which might imply that they are no longer needed and can thus be safely archived or discarded.

5.3.2.5 Resource Description for Sensemaking and Science

Up to now in §5.3.2, we have discussed how resource descriptions are used to perform well-defined tasks within an existing organizing system. However, there is a broader and less well-defined purpose of resource description that is older and more fundamental: the use of resource descriptions as the raw material for making sense of the world.

For thousands of years, even before the invention of written language, people have systematically collected things, information about those things, and observations of all kinds to understand how their world works. Paleolithic humans made cave paintings depicting the results of hunts and animal migrations; ancient Egyptians recorded the annual floods of the Nile River in stone carvings; and Babylonian, Egyptian, Chinese, and Mesoamerican astronomers organized

lunar, solar, and planetary observations as calendars starting about five thousand years ago.

These diverse efforts to impose meaning on experience by recording, analyzing, organizing, and reorganizing observations can be collectively described as *sensemaking*. (See the sidebar, *Sensemaking and Organizing* (page 240).)

Some aspects of sensemaking are hard-wired by evolution, which has given our brains powerful mechanisms that automatically simplify and organize the perceptual data we obtain from the world (see the sidebar *Gestalt Principles* (page 102)). But this automatic sensemaking is dominated and amplified by intentional sensemaking.

Intentional sensemaking takes place when systematic statistical, experimental, and scientific methods are consciously followed to extract and organize knowledge from collections of samples, observations, or measurements. It is critical to recognize here that the contents of these collections represent choices made about what to collect, because most things and most phenomena have a great many descriptions or properties that could be recorded about them.

After things or data have been collected, statistical methods summarize the values of properties in a collection or dataset and the relationships among them. Making sense of a single collection or dataset by determining the properties that contrast and classify the instances is the start toward the more important goal of understanding the larger set or population from which the initial collection is just a sample. There is no better example of this than the periodic table of elements developed by Mendeleev in 1869, who organized known elements on the basis of their common chemical properties and then successfully predicted some properties of yet undiscovered ones.

Computational models developed from the initial dataset can predict future observations. Classification models assign a new instance to a category (e.g., spam or not spam message, Madison or Hamilton as author, outdoor or indoor scene); regression models predict a specific value of some measurement (given a description of a new movie, how much money will it make?); ordinal regression models predict values for non-metric measures (how much will you like the movie?). Experimental methods for hypothesis testing help develop and refine models of any type by systematically varying the conditions under which observations are made to discover how the results change in different situations.

A fundamental challenge in sensemaking and modeling is finding a balance between the competing goals of understanding a particular collection or dataset and being able to apply that understanding to new instances. Models can differ in the number of resource descriptions they use as parameters, and it is easy and tempting to overfit a model by using more parameters that capture random

Sensemaking and Organizing

People organize to make sense of equivocal inputs and enact this sense back into the world to make it more orderly.

— (Weick 2005)

Sensemaking and *organizing* are intertwined. Ancient cultures recorded time-based observations and analyzed patterns among crop cycles, commodity prices, weather conditions, and astronomical sightings. Think back to the early astronomers, who oriented temple buildings to align with astronomical events and who decorated temple walls with zodiac imagery.

- Which of the planets and stars in the night sky should they observe and how should they record the details of those observations?
- What mathematical and statistical techniques should be used to analyze and describe these observations?
- What subset of observations are most useful in predicting the onset of the Nile River floods, caused by unobserved rainfall thousands of miles away?

Every choice about what to observe and how to describe it reflects a set of assumptions and potential hypotheses that are often implicit and unstated. Choices that increase understanding are built upon, and those that fail to provide insight are abandoned, but there is no guarantee that the iterative process of choosing what to observe and describe will lead to a correct understanding.

The principle that an accurate or comprehensive dataset is insufficient on its own to yield a correct model is exemplified in the interlocking efforts of Tycho Brahe and Johannes Kepler. Brahe was a 16th-century Danish nobleman astronomer who spent decades collecting data about the positions of hundreds of stars and the planets. However, because of prevailing religious and scientific biases, Brahe accepted the incorrect assumptions that the sun and planets revolved around the earth in circular orbits. After Brahe died in 1601, Kepler spent a decade analyzing Brahe's data, and then rejected the idea of earth-centric and circular planetary orbits in favor of elliptical ones with the sun at one focus. These new organizing principles for Brahe's data made the model of the solar system vastly simpler, and Kepler was able to discover laws of planetary motion that are part of the foundation of modern astronomy and physics.^{270[DS]}

variations in observations. Overfitting produces spurious accuracy in reproducing the original observations, but it makes models less generalizable.

The highest level of sensemaking is the creation of scientific models or theories that propose interpretable and causal mechanisms for the observations. And just as automatic sensemaking creates simple explanations, scientists generally prefer simpler theories, a heuristic known as Occam's Razor, or the law of parsimony. Even though complex theories can sometimes be more accurate, simpler theories produce more testable predictions, making it easier to verify or refine the theory. Occam's famous principle, expressed eight centuries ago, is to prefer models that make the fewest assumptions, often measured in terms of the number of parameters or variables needed to make a prediction.^{271[Bus] 272[Phil] 273[Com]}

5.3.3 Identifying Properties

Once the purposes of description have been established, we need to identify the specific properties of the resources that can satisfy those purposes. There are four reasons why this task is more difficult than it initially appears.

- First, any particular resource might need many resource descriptions, all of which relate to different properties, depending on the interactions to be supported and the context in which they take place. Selecting people for a basketball team focuses on their physical properties such as height, strength, leaping ability, and coordination. Selections for a debate team will be more concerned with their verbal and intellectual properties.
- Second, different types of resources need to incorporate different properties in their descriptions. For resources in a museum, these might include materials and dimensions of pieces of art; for files and services managed by a network administrator, these include access control permissions; for electronic books or DVDs, they would include the digital rights management (DRM) code that expresses what you can and cannot do with the resource.
- Third, as we briefly touched on in §5.3.1.3, which properties participate in resource descriptions depends on who is doing the describing. It makes little sense to expect fine-grained distinctions and interpretations about properties from people who lack training in the discipline of organizing. We will return to this tradeoff in §5.3.6 and again in §5.4.1.
- Fourth, what might seem to be the same property at a conceptual level might be very different at an implementation level. Many resources have a resource description that is a surrogate or summary of the primary resource. For photos, paintings, and other resources whose appearance is their essence, an appropriate summary description can be a smaller, reduced resolution photo of the original. This surrogate is simple to create and easy for users to relate to the primary resource. On the other hand, distilling a text down to a short summary or abstract is a skill unto itself. Time-based resources provide greater challenges for summary. Should the summary of a

movie be a textual summary of the plot, a significant clip from the movie, a video summary, or something else altogether?

This implementation gap is often very large for properties about people because people are not as easy to measure as most types of resources. Businesses need to quantify a person's interest in their products to predict what price they would be willing to pay, but "interest" cannot be measured directly. Instead, predictions rely on proxy measures for "interest" like how long the customer looked at the product web page and whether they also looked at a competitor's web page.

Two important dimensions for understanding and contrasting resource properties used in descriptions and organizing principles are: property essence—whether the properties are intrinsically or extrinsically associated with the resource, and; property persistence—whether the properties are static or dynamic. Taken together these two dimensions yield four categories of properties, as illustrated in **Figure 5.4, Property Essence x Persistence: Four Categories of Properties**. These four categories provide a useful framework for thinking about resource properties, even if, at times, the classification of properties is debatable.²⁷⁴[CogSci]

5.3.3.1 Intrinsic Static Properties

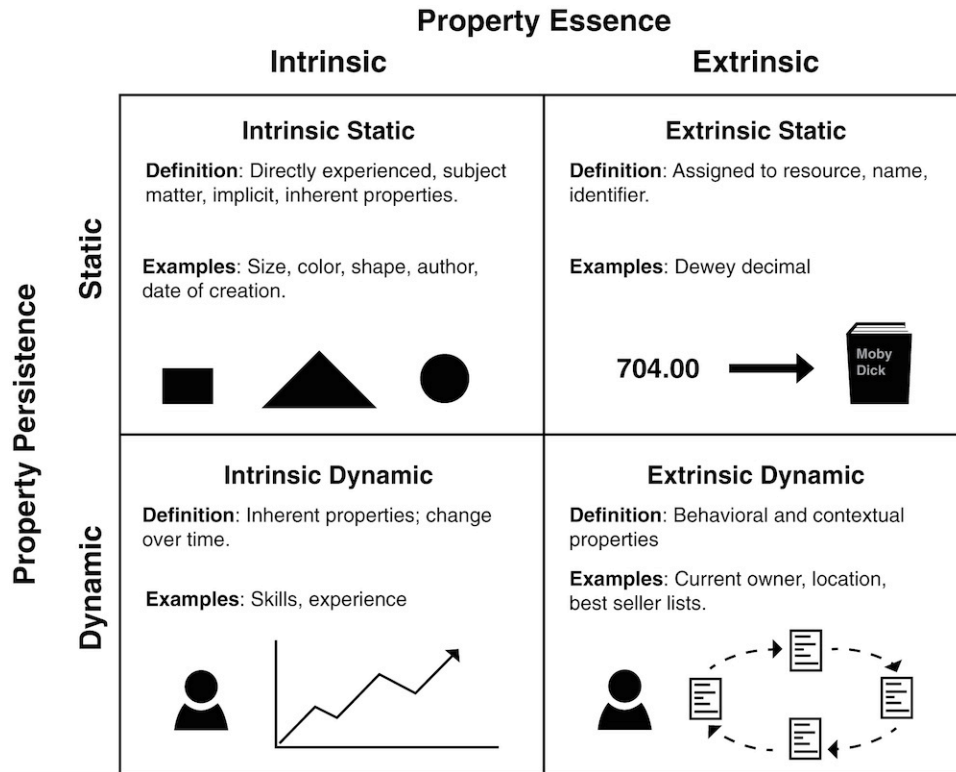
Intrinsic or implicit properties are inherent in the resource and can often be directly perceived or experienced. Static properties do not change their values over time. The species of an animal, the material of composition of a wooden chair, and the diameter of a wheel are all static properties that do not change their values over time. Static properties like color or shape are often used to describe and organize physical resources.

Intrinsic physical properties are usually just part of resource descriptions. In many cases, physical properties describe only the surface layer of a resource, revealing little about what something is or its original intended purpose, what it means, or when and why it was created. The author of a song and the context of its creation are other examples of intrinsic and static resource properties that are not directly perceivable.

Intrinsic descriptions are often extracted or calculated by computational processes. For example, a computer program might calculate the frequency and distribution of words in some particular document. Similarly, visual signatures or audio fingerprints are intrinsic descriptions (**§5.4 Describing Non-text Resources (page 257)**).

Some relationships among resources are intrinsic and static, like the parent-child relationship or the sibling relationship between two children with the same parents. Part-whole or compositional relationships for resources with

Figure 5.4. Property Essence x Persistence: Four Categories of Properties.



The distinctions of property persistence and property essence combine to distinguish four categories of properties: intrinsic static, extrinsic static, intrinsic dynamic, and extrinsic dynamic properties.

parts are also intrinsic static properties often used in resource descriptions. However, it is better to avoid treating resource relationships as properties, and instead express them as relations. **Chapter 6, Describing Relationships and Structures** discusses part-whole and other semantic relationships in great detail.

Intrinsic Static Properties Define a Dalmatian

The spots on a Dalmatian dog are intrinsic static properties that appear shortly after birth, and they are so distinctive that it is impossible to describe the breed without acknowledging the spots.



This particular Dalmatian is the “greeter” at the Viader Winery in Deer Park, California. The dog is nice and the wines are excellent.

(Photo by R. Glushko.)

5.3.3.2 Extrinsic Static Properties

Extrinsic or explicit properties are assigned to a resource rather than being inherent in it. The name or identifier of a resource is often arbitrary but once assigned does not usually change. Arranging resources according to the alphabetical or numerical order of their descriptive identifiers is a common organizing principle. Classification numbers and subject headings assigned to bibliographic resources are extrinsic static properties, as are the serial numbers stamped on or attached to manufactured products.

For information resources that have a digital form, the properties of their printed or rendered versions might not be intrinsic. Some text formats completely separate content from presentation, and as a result, style sheets can radically change the appearance of a printed document or web page without altering the primary resource in any way. For example, were a different style applied to this paragraph

to highlight it in bold or cast in 24-point font, its content would remain the same.

5.3.3.3 Intrinsic Dynamic Properties

Intrinsic dynamic properties change over time. Developmental personal characteristics like a person’s height and weight, skill proficiency, or intellectual capacity, for example. Because these properties are not static, they are usually employed only to organize resources whose membership in the collection is of limited duration. Sports programs or leagues that segregate participants by age or years of experience are using intrinsic dynamic properties to describe and organize the resources.

5.3.3.4 Extrinsic Dynamic Properties

Extrinsic dynamic properties are in many ways arbitrary and can change because they are based on usage, behavior, or context. The current owner or location of a resource, its frequency of access, the joint frequency of access with other resources, its current popularity or cultural salience, or its competitive advantage over alternative resources are typical extrinsic and dynamic properties that are used in resource descriptions. A topical book described as a best seller one year might be found in the discount sales bin a few years later. A student's grade point average is an extrinsic dynamic property.

Extrinsic dynamic properties are useful features for data scientists making prediction or classification models. Your current location, the thing you just bought, and the place you bought it can be viewed as manifestations of unobservable preferences and values. Fingerprints found on a doorknob at a crime scene are an extrinsic dynamic property associated with the door, and clever detectives would analyze them along with other interaction resources they discovered with the goal of identifying the person for whom the fingerprints are intrinsic static properties.

Many relationships between resources are extrinsic and dynamic properties, like that of best friend.

Contextual properties are those related to the situation or context in which a resource is described. Dey defines *context* as “any information that characterizes a situation related to the interactions between users, applications, and the surrounding environment.” ^{275[Com]} This open-ended definition implies a large number of contextual properties that might be used in a description; crisper definitions of context might be “location + activity” or “who, when, where, why.” Since context changes, context-based descriptors might be appropriate when assigned but can have limited persistence and effectivity (§4.5 *Resources over Time* (page 198)); the description of a document as “receipt of a recent purchase” will not be useful for very long.

Citations of one information resource by another are extrinsic static descriptions when they are in print form, but when they are published in digital libraries it is usually the case that “cited by” is a dynamic resource description. Similarly, any particular link from one web page to another is an extrinsic static description, but because many web pages themselves are highly dynamic, we can also consider links as dynamic as well. Citations and web links are discussed in more detail in [Chapter 6](#).

Resources are often described with *cultural properties* that derive from conventional language or culture, often by analogy, because they can be highly evocative and memorable. ^{277[Ling]}

Why are Ottoman Carpets Named After a German Painter?



An example of a cultural category that has far outlasted its motivation is that of the Holbein carpet. A particular type of geometrically patterned Ottoman rug came to be known as a “Holbein carpet” after the German Renaissance painter Hans Holbein, who often depicted the rugs in his work (probably to show off his extremely meticulous technique). Holbein was famous in his time, and his commissioned paintings of the English King Henry VIII have Henry standing on such rugs. This painting, called “The Ambassadors,” was painted in 1533 and now hangs in The National Gallery, London.

(Source: Google Art Project)

Sometimes a cultural description outlives its salience, losing its power to evoke anything other than puzzlement about what it might mean.^{278[Ling]}

Latent Feature Creation and Netflix Recommendations

Recent advances in computing technology and *data science* techniques are making it possible to discover or create resource properties that are called “latent” because they are inferred rather than observed. Many such features are used by businesses to segment customers or make recommendations to them based on their recent behavior, so these features are also extrinsic and dynamic.

Your own movie preferences prove that easy to identify properties like sex and age do not differentiate movie watchers enough to make good recommendations, even if you combine them to create a category like “single male students between 18 and 25.” Netflix found that it was necessary to combine demographic properties, viewing history, and browsing behavior with very detailed ratings of dozens of movie properties to make good recommendations. It takes enormous computing power to discover a category of Netflix users who typically like action movies, yet consistently hover their mouse over romance movies, and to use this latent feature to recommend a sub-genre of western movies (one of nearly 100,000) that it calls “Romantic Action Adventure Movies.”^{276[Com]}

For the Lego boys, current with the latest *Star Wars* movies, “light saber” was just the obvious description for a long, neon tube with a handle. However, someone unfamiliar with the *Star Wars* franchise might not understand “light saber,” and would describe the piece some other way.

5.3.4 Designing the Description Vocabulary

After we have determined the properties to use in resource descriptions, we need to design the description vocabulary: the set of words or values that represent the properties. §4.4 **Naming Resources** (page 188) discussed the problems of naming and proposed principles for good names, and since names are a very important resource description, much of what we said there applies generally to the design of the description vocabulary.

However, because the description vocabulary as a whole is much more than just the resource name, we need to propose additional principles or guidelines for this step. In addition, some new design questions arise when we consider all the resource descriptions as a set whose separate descriptions are created by many people over some period of time.

5.3.4.1 Principles of Good Description

In *The Intellectual Foundation of Information Organization*, Svenonius proposes a set of principles or “directives for design” of a description language. Her principles, framed in the narrow context of bibliographic descriptions, generally apply to the broad range of resource types we consider in this book.

User Convenience

Choose description terms with the user in mind; these are likely to be terms in common usage among the target audience.

Representation

Use descriptions that reflect how the resources describe themselves; assume that self-descriptions are accurate.

Sufficiency and Necessity

Descriptions should have enough information to serve their purposes and not contain information that is not necessary for some purpose; this might imply excluding some aspects of self-descriptions that are insignificant.

Standardization

Standardize descriptions to the extent practical, but also use aliasing to allow for commonly used terms.

Integration

Prefer the same properties and terms for all types of resources.

Any set of general design principles faces two challenges.

- The first is that implementing any principle requires many additional and specific context-dependent choices for which the general principle offers little guidance. For example, how does the principle of Standardization apply if multiple standards already exist in some resource domain? Which of the competing standards should be adopted, and why?
- The second challenge is that the general principles can sometimes lead to conflicting advice. The User Convenience recommendation to choose description terms in common use fails if the user community includes both ordinary people and scientists who use different terms for the same resources; whose “common usage” should prevail?

5.3.4.2 Who Uses the Descriptions?

Focus on the user of the descriptions. This is a core idea that we cannot overemphasize because it is implicit in every step of the process of resource description. All of the design principles in the previous section share the idea that the design of the description vocabulary should focus on the user of the descriptions. Are the resources being organized personal ones, for personal and mostly private purposes? In that case, the description properties and terms can be highly personal or idiosyncratic and still follow the design principles.

Similarly, when resource users share relevant knowledge, or are in a context where they can communicate and negotiate, if necessary, to identify the resources, their resource descriptions can afford to be less precise and rigorous than they might otherwise need to be. This helps explain the curious descriptions in the Lego story with which we began this chapter. The boys playing with the blocks were talking to each other with the Legos in front of them. If they had not been able to see the blocks the others were talking about, or if they had to describe their toys to someone who had never played with Legos before, their descriptions would have been quite different.

More often, however, resource descriptions can not assume this degree of shared context and must be designed for user categories rather than individual users: library users searching for books, business employees or customers using part and product catalogs, scientists analyzing the datasets from experiments or simulations. In each of these situations resource descriptions will need to be understood by people who did not create them, so the design of the description vocabulary needs to be more deliberate and systematic to ensure that its terms are unambiguous and sufficient to ensure reliable context-free interpretation. A single individual seldom has the breadth of domain knowledge and experience with users needed to devise a description vocabulary that can satisfy diverse users with diverse purposes. Instead, many people working together typically

develop the required description vocabulary. We call the results institutional vocabularies, to contrast them with individual or cultural ones. (We will discuss this contrast more fully in [Chapter 7, Categorization: Describing Resource Classes and Types](#))

Some resource descriptions are designed for use by machines, which seemingly reduces the importance of design principles that consider user preferences or common uses. However, even if resources are described and organized by algorithms, when people need to explain the classifications and predictions that the algorithms produce, resource descriptions that are comprehensible and easily communicated are preferable to statistically optimal ones. Moreover, standardization and integration principles become more important for inter-machine communication to enable efficient processing, reuse of data and software, and increased interoperability among organizing systems.^{280[Com]}

Stop and Think: Description and Expertise

Everyone knows something about trees, but some people know more than others, and their particular experience and perspective influences how they describe trees. What kind of properties and descriptions would be used by university students? By research botanists? By landscape designers? By park maintenance workers? By indigenous people who live in tropical rain forests?

5.3.4.3 Controlled Vocabularies and Content Rules

As we defined in [§4.4.3.2](#), a *controlled vocabulary* is a fixed or closed set of description terms in some domain with precise definitions that is used instead of the vocabulary that people would otherwise use. For example, instead of the popular terms for descriptions of diseases or symptoms, medical researchers and teaching hospitals can use the National Library of Medicine's *Medical Subject Headings (MeSH)* controlled vocabulary.

We can distinguish a progression of vocabulary control: a glossary is a set of allowed terms; a thesaurus is a set of terms arranged in a hierarchy and annotated to indicate terms that are preferred, broader than, or narrower than other terms; an ontology expresses the conceptual relationships among the terms in a formal logic-based language so they can be processed by computers. We will say more about ontologies in [Chapter 6](#).

Content rules are similar to controlled vocabularies because they also limit the possible values that can be used in descriptions. Instead of specifying a fixed set of values, content rules typically restrict descriptions by requiring them to be of a particular data type (integer, Boolean, Date, and so on). Possible values are constrained by logical expressions (e.g., a value must be between 0 and 99) or

regular expressions (e.g., must be a string of length 5 that must begin with a number). Content rules like these are used to ensure valid descriptions when people enter them in web forms or other applications.

5.3.4.4 Vocabulary Control as Dimensionality Reduction

In most cases, a controlled vocabulary is a subset of the natural or uncontrolled vocabulary, but sometimes it is a new set of invented terms. This might sound odd until we consider that the goal of a controlled vocabulary is to reduce the number of descriptive terms assignable to a resource. Stated this way the problem is one of *dimensionality reduction*, transforming a high-dimensional space into a lower-dimensional one. Reducing the number of components in a multidimensional description can be accomplished by many different statistical techniques that go by names like “feature extraction,” “principle components analysis,” “orthogonal decomposition,” “latent semantic analysis,” “multidimensional scaling,” and “factor analysis.” ^{282[DS]}

These techniques might sound imposing and they are computationally complex, but they all have the same simple concept at their core, that the features or properties that describe some resource are often highly correlated. For example, a document that contains the word “car” is more likely to contain the words “driver” and “traffic” than a document that does not. Similar correlations exist among the visual features used to describe images and the acoustic features that describe music. Dimensionality reduction techniques analyze the correlations between resource descriptions to transform a large set of descriptions into a much smaller set of uncorrelated ones. In a way this implements the principle of Sufficiency and Necessity we mentioned in §5.3.4.1 *Principles of Good Description* (page 247) because it eliminates description dimensions or properties that do not contribute much to distinguishing the resources.

Here is an oversimplified example that illustrates the idea. Suppose we have a collection of resources, and every resource described as “big” is also described as “red,” and every “small” resource is also “green.” This perfect correlation between color and size means that either of these properties is sufficient to distinguish “big red” things from “small green” ones, and we do not need clever algorithms to figure that out. But if we have thousands of properties and the correlations are only partial, we need the sophisticated statistical approaches to choose the optimal set of description properties and terms, and in some techniques the dimensions that remain are called “latent” or “synthetic” ones because they are statistically optimal but do not map directly to resource properties.

5.3.5 Designing the Description Form

By this step in the process of resource description we have made numerous important decisions about which resources to describe, the purposes for which we are describing, them, and the properties and terms we will use in the descriptions. As much as possible we have described the steps at a conceptual level and postponed discussion of implementation considerations about the notation, syntax, and deployment of the resource descriptions separately or in packages. Separating design from implementation concerns is an idealization of the process of resource description, but is easier to learn and think about resource description and organizing systems if we do. We discuss these implementation issues in *Chapter 9, The Forms of Resource Descriptions*.

Sometimes we have to confront legacy technology, existing or potential business relationships, regulations, standards conformance, performance requirements, or other factors that have implications for how resource descriptions must or should be implemented, stored, and managed. We will take this more pragmatic perspective in *Chapter 11, The Organizing System Roadmap*, but until then, we will continue to focus on design issues and defer discussion of the implementation choices.

5.3.6 Creating Resource Descriptions

Resource descriptions can be created by professionals, by the authors or creators of resources, by users, or by computational or automated means.

Professionally-created resource descriptions, author- or user-created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs. A natural solution is to try to combine desirable aspects from each in hybrid approaches. For example, the vocabulary for a new resource domain may arise from tagging by end users but then be refined by professionals, lay classifiers may create descriptions with help from software tools that suggest possible terms, or software that creates descriptions can be improved by training it with human-generated descriptions, a form of *supervised learning* (see §7.5.3.3).

Often existing resource descriptions can or must be transformed or enhanced to meet the ongoing needs of an organizing system, and sometimes these processes can be automated. We will defer further discussion of those situations to *Chapter 10, Interactions with Resources*. In the discussion that follows we focus on the creation of new resource descriptions where none yet exist.

5.3.6.1 Resource Description by Professionals

Before the web made it possible for almost anyone to create, publish, and describe their own resources and to describe those created and published by others, resource description was generally done by professionals in institutional contexts. Professional indexers and catalogers described bibliographic and museum resources after having been trained to learn the concepts, controlled descriptive vocabularies, and the relevant standards. In information systems domains professional data and process analysts, technical writers, and others created similarly rigorous descriptions after receiving analogous training. We have called these types of resource descriptions institutional ones to highlight the contrast between those created according to standards and those created informally in *ad hoc* ways, especially by untrained or undisciplined individuals.^{285[Bus]}

5.3.6.2 Resource Description by Authors or Creators

The author or creator of a resource can be presumed to understand the reasons why and the purposes for which the resource can be used. And, presumably, most authors want to be read, so they will describe their resources in ways that will appeal to and be useful to their intended users. However, these descriptions are unlikely to use the controlled vocabularies and standards that professional catalogers would use.

5.3.6.3 Resource Description by Users

Today's web contains a staggering number of resources, most of which are primary information resources published as web content, but many others are resources that stand for "in the world" physical resources. Most of these resources are being described by their users rather than by professionals or by their authors. These "at large" users are most often creating descriptions for their own benefit when they assign tags or ratings to web resources, and they are unlikely to use standard or controlled descriptors when they do so. The resulting variability can be a problem if creating the description requires judgment on the tagger's part. Most people can agree on the length of a particular music file but they may differ wildly when it comes to determining to which musical genre that file belongs. Fortunately most web users implicitly recognize that the potential value in these "Web 2.0" or "user-generated content" applications will be greater if they avoid egocentric descriptions. In addition, the statistics of large sample sizes inevitably leads to some agreement in descriptions on the most popular applications because idiosyncratic descriptions are dominated in the frequency distribution by the more conventional ones.^{287[Web]}

We are not suggesting that professional descriptions are always of high quality and utility, and socially produced ones are always of low quality and utility.^{288[CogSci]} Rather, it is important to understand the limitations and

qualifications of descriptions produced in each way. Tagging lowers the barrier to entry for description, making organizing more accessible and creating descriptions that reflects a variety of viewpoints. However, when many tags are associated with a resource, it increases *recall* while decreasing *precision*. (See §5.3.6.3 Resource Description by Users (page 252))

5.3.6.4 Automated and Computational Resource Description

A picture's EXIF file created by a digital camera records properties associated with the camera and its settings, as well as some properties of the photo-taking context. (See Figure 5.6, *Contrasting Descriptions for a Work of Art*, for an example.) Creating this highly detailed description by hand would be nearly impossible. The downside, however, is that the automated description does not capture the meaning of the photo; an automated picture description captures the time and place, but not that it is a picture of a honeymoon vacation. The difference between automated and human description is called the *semantic gap* (§4.4.2.5).

Any resource that is smart enough to collect data about its state or environment is creating resource descriptions automatically (See §4.2.3.2). Resources with computational capabilities can process the raw sensor data to identify important events and create more interpretable descriptions.

Some computational approaches create resource descriptions that are similar in purpose to those created by human describers. Text mining and summarization systems for customer comments about products can reduce thousands of comments to a list of the most important features.^{289[Com]} People shopping for books at Amazon.com get insights about a book's content and distinctiveness from the statistically improbable phrases that it has identified by comparing all the books for which it has the complete text.^{290[Com]}

Computational descriptions can use any observable or latent variable (see the sidebar, *Latent Feature Creation and Netflix Recommendations* (page 246)) except some that are prohibited by law, such as race, religion, national origin, and marital status, to prevent discrimination. In practice, however, this prohibition is easily circumvented because these properties can usually be predicted using other ones. For example, race can often be reliably predicted using residence address and surname.^{291[DS]}

Metacrap

In an often-cited essay (Doctorow 2001) provocatively titled “Metacrap: Putting the torch to seven straw-men of the meta-utopia,” Cory Doctorow argues that much human-created metadata is of low quality because “people lie, people are lazy, people are stupid, mission impossible—know thyself, schemas are not neutral, metrics influence results, (and) there is more than one way to describe something.”

Of course, all information retrieval systems compare a description of a user's needs with descriptions of the resources that might satisfy them. IR systems differ in the resource properties they emphasize; word frequencies and distributions for documents in digital libraries, links and navigation behavior for web pages, acoustics for music, and so on. These different property descriptions determine the comparison algorithms and the way in which relevance or similarity of descriptions is determined. We say a lot more about this in [§5.4 Describing Non-text Resources](#) (page 257) and in [Chapter 10](#).

5.3.7 Evaluating Resource Descriptions

Evaluation is implicit in many of the activities of organizing systems we described in [Chapter 3, Activities in Organizing Systems](#) and is explicit when we maintain a collection of resources over time. In this section, we focus on the narrower problem of evaluating resource descriptions.

Evaluating means determining *quality* with respect to some criteria or dimensions. Many different sets of criteria have been proposed; for repositories of digital resources, the most commonly used ones are accuracy, completeness, and consistency. Other typical criteria are timeliness, interoperability, and usability. It is easy to imagine these criteria in conflict; efforts to achieve accuracy and completeness might jeopardize timeliness; enforcing consistency might preclude modifications and personalizations that would enhance usability.

Stop and Think: Defining Quality

What characteristics or criteria would you use to determine the quality of a car? Of food? Of clothing? Of a place to live? Which of these criteria are domain-specific, and which ones apply more generally to many types of resources?

The *quality* of the outcome of the multi-step process proposed in this chapter is a composite of the *quality* created or squandered at each step. A scope that is too granular or abstract, overly ambitious or vague intended purposes, a description vocabulary that is hard to use, or giving people inadequate time to create good descriptions can all cause quality problems, but none of these decisions is visible at the end of the process

where users interact with resource descriptions.

5.3.7.1 Evaluating the Creation of Resource Descriptions

When professionals create resource descriptions in a centralized manner, which has long been the standard practice for many resources in libraries, there is a natural focus on *quality* at the point of creation to ensure that the appropriate controlled vocabularies and standards have been used. However, the need for resource description generalizes to resource domains outside of the traditional bibliographic one, and other *quality* considerations emerge in those contexts.

Resource descriptions in private sector firms are essential to running the business and in interacting efficiently with suppliers, partners, and customers. Compared to the public sector, there is much greater emphasis on the economics and strategy of resource description.^{293[Bus]} What is the value of resource description? Who will bear the costs of producing them? Which of the competing industry standards will be followed? Some of these decisions are not free choices as much as they are constraints imposed as a condition of doing business with a dominant economic partner, which is sometimes a governmental entity.

For example, a firm like Wal-Mart with enormous market power can dictate terms and standards to its suppliers because the long-term benefits of a Wal-Mart contract usually make the initial accommodation worthwhile. Likewise, governments often require their suppliers to conform to open standards to avoid lock-in to proprietary technologies.^{294[Bus]}

In both the public and private sectors there is increased use of computational techniques for creating resource descriptions because the number of resources to be described is simply too great to allow for professional description. A great deal of work in text data mining, web page classification, semantic enrichment, and other similar research areas is already under way and is significantly lowering the cost of producing useful resource descriptions. Some museums have embraced approaches that automatically create user-oriented resource descriptions and new user interfaces for searching and browsing by transforming the professional descriptions in their internal collections management systems. Google's ambitious project to digitize millions of books has been criticized for the quality of its algorithmically extracted resource descriptions, but we can expect that computer scientists will put the Google book corpus to good use as a research test bed to improve the techniques.^{296[Com]}

Web 2.0 applications that derive their value from the aggregation and interpretation of user-generated content can be viewed as voluntarily ceding their authority to describe and organize resources to their users, who then tag or rate them as they see fit. In this context the consistency of resource description, or the lack of it, becomes an important issue, and many sites are using technology or incentives to guide users to create better descriptions.

5.3.7.2 Evaluating the Use of Resource Descriptions

Regardless of, or in addition to, any quality criteria applied to the creation and selection of resource descriptions, at some point the resource descriptions meet their intended users. The most important quality criterion at that point is whether the resource descriptions satisfy their intended purposes in a usable way. In many ways, the answer is a disappointing no.

For example, in one of the earliest revisions to the original HTML specification, a `<META>` tag was added to allow creators of web resources to define a set of key terms to describe a website or web page. This well-motivated resource description was to be used by search engines to improve the relevance of retrieved pages. However, it soon became obvious that it was possible to “game” the META tag by adding popular terms even though they did not accurately describe the page. Today search engines ignore the `<META>` tag for ranking pages, but many other techniques that use false *resource* descriptions continue to plague web users. (See §3.5.3.3.)

The design of a description vocabulary circumscribes what can be said about a resource, so it is important to recognize that it implicitly determines what cannot be said as well, with unintended negative consequences for users. The *resource* description schema implemented in a physician’s patient management system defines certain types of recordable information about a patient’s visit—the date of the visit, any tests that were ordered, a diagnosis that was made, a referral to a specialist. The schema, and its associated workflow, impose constraints that affect the kinds of information medical professionals can record and the amount of space they can use for those descriptions. Moreover, such a schema might also eliminate vital unstructured space that paper records can provide, where doctors communicate their rationale for a diagnosis or decision without having to fit it into any particular box.

However, when resource descriptions are the data used to train models for prediction or classification, the focus of evaluation is not on the descriptions, which are often assumed to be accurate observations about the world. Instead, evaluation focuses on the model, and “model selection” is the task of choosing which of several competing models best fit the original data while also generalizing well to new data. In any event, any quality problems or selection biases with the original data will undermine the value of whatever model is selected.

5.3.7.3 The Importance of Iterative Evaluation

The inevitable conflicts between *quality* goals mean that there will be compromises among the *quality* criteria. Furthermore, increasing *scale* in an organizing system and the steady improvements of computational techniques for resource description imply that the nature of the compromise will change over time. As a result, a single evaluation of resource descriptions at one moment in time will not suffice.

This makes usage records, navigation history, and transactional data extremely important kinds of resource descriptions because they enable you to focus efforts on improving *quality* where they are most needed. Furthermore, for organizing systems with many types of resources and user communities, this information can enable the tailoring of the nature and extent of resource description to find the right balance between “rich and comprehensive” and “simple and efficient” approaches. Each combination of resource type and user community might have a different solution.

The idea that *quality* is a property of an end-to-end process is embodied in the “quality movement” and statistical process control for industrial processes but it applies equally well to resource description. The central idea is that *quality* cannot be tested in by inspecting the final products. Instead, *quality* is achieved through process control—measuring and removing the variability of every process needed to create the products.^{297[Bus]} Explicit feedback from users or implicit feedback from the records of their resource interactions needs are essential as we iterate through the design process and revisit the decisions made there.

5.4 Describing Non-text Resources

Many of the principles and methods for resource description were developed for describing text resources in physical formats. Those principles have had to evolve to deal with different types of resources that people want to describe and organize, from paintings and statues to MP3s, JPEGs, and MPEGs.

Some descriptions for non-text resources are text-based, and are most often assigned by people. Other descriptions in non-text formats are extracted algorithmically from the content of the non-text resource. These latter content-based resource descriptions capture intrinsic technical properties and in some domains are able to describe *aboutness* with some accuracy, thanks to breakthroughs in machine learning.

5.4.1 Describing Museum and Artistic Resources

The problems associated with describing multimedia resources are not all new. Museum curators have been grappling with them since they first started to collect, store, and describe artifacts hundreds of years ago. Many artifacts may represent the same work (think about shards of pottery that may once have been part of the same vase). The materials and forms do not convey semantics on their own. Without additional research and description, we know nothing about the vase; it does not come with any sort of title page or tag that connects it with a 9th-century Mayan settlement. Since museums can acquire large batches of artifacts all at once, they have to make decisions about which resources they can afford to describe and how much they can describe them.

German art historian Erwin Panofsky first codified one approach to these problems of description. In his classic *Studies in Iconology*, he defined three levels of description that can be applied to an artistic work or museum artifact. **Figure 5.6, Contrasting Descriptions for a Work of Art**, contrasts these three levels in the descriptions of a marble statue. It also shows the striking differences between the EXIF description in a digital photo of the statue and those created by people.


5.4.2 Describing Images

Digital cameras, including those in cell phones, take millions of photos each day. Unlike the images in museums and galleries, most of these images receive few descriptions beyond those created by the device that made them. Nevertheless, a great many of them end up with some limited descriptions in Facebook, Instagram, Flickr, Picasa, DeviantArt, or others of the numerous places where people share images, or in professional image applications like Light Room. All of these sites provide some facilities for users to assign tags to images or arrange them in named groups.

Many different computational approaches have been used to describe or classify images. One approach uses the visual signature of an image extracted from low-level features like color, shape, texture, and luminosity, which are then used to distinguish significant regions and objects. Image similarity is computed to create categories of images that contain the same kinds of colors, objects, or settings, which makes it easy to find duplicate or modified images.^{300[Com]}

For computers to identify specific objects or people in images, it is logically necessary to train them with images that are already identified. In 2005 Luis van Ahn devised a clever way to collect large amounts of labeled images with a web-based game called ESP that randomly paired people to suggest labels or tags for an image. The obvious choices were removed from contention, so a photo of a bird against a blue sky might already strike “bird” and “sky” from the set of

Figure 5.6. Contrasting Descriptions for a Work of Art.



EXIF Summary

Make	NIKON CORPORATION
Model	NIKON D90
Aperture	9
Exposure Time	1/320 (0.003125 sec)
Lens	ID AF-S DX VR Zoom-Nikkor 18-105mm f/3.5-5.6G ED
Focal Length	21.0 mm
Flash	Auto, Did not fire
File Size	4.7 MB
File Type	JPEG
Image Height	4288
Image Width	2848
Date & Time	2012:12:03 10:31:14

3 Levels

Primary

Marble statue of nude woman standing on a seashell.

Secondary

Statue made in 2005 by Lucio Carusi of Carrara, Italy, titled "Venus", made of local marble.

Interpretive

This is a 3d transformation of the 1486 painting by Italian painter Sandro Botticelli, titled "The Birth of Venus", now in the Uffizi Gallery in Florence. Carusi's Venus is substantially slimmer in proportions than Botticelli's because of changing notions of female beauty.

Descriptions for works of art can contrast a great deal, especially between those captured by a device like a digital camera and those created by people. Furthermore, the descriptions created by people differ according to the expertise of the creator and the amount of subjective interpretation applied in the description.

(Photo by R. Glushko. The statue, titled "Venus," was made by Lucio Carusi, of Carrara, Italy, and is currently part of a private collection.)

acceptable words, leaving users to suggest words such as “flying” and “cloudless.” Van Ahn also invented the reCAPTCHA technique that presents images of text from old books being digitized, which improves the accuracy of the digitization while verifying that the user of a web site is a person and not a robot program.^{301[Web]}

However, if short text descriptions or low-level image properties are the only features available to train an image, otherwise irrelevant variations in the position, orientation, or illumination of objects in images will make it very difficult to distinguish objects that look similar, like a white wolf and the wolf-like white dog called a Samoyed. This problem can be addressed by using deep neural networks, which exploit the idea that low-level image features can be combined into many layers of higher-level ones; edges combine to form motifs or patterns, patterns combine to form parts of familiar objects, and parts combine to form complete objects. This hierarchical composition enables the highest-level representations to become insensitive to the lower-level variations that plague the other approaches.

In 2012, when deep learning techniques were applied to a dataset of about a million images that contained a thousand different object categories, they reduced the error rate by half. This spectacular breakthrough, and the fact that the deep learning techniques that derive layers of features from the input data are completely general, rapidly caused deep learning to be applied to many other domains with high-dimensional data. Facebook uses deep learning to identify people in photos, Google uses it for speech recognition and language translation, and rapid captioning for images and video are on the horizon. Wearable computers might use it to layer useful information onto people's views of the world, creating real-time augmented reality.^{302[Com]}

5.4.3 Describing Music

Some parts of describing a song are not that different from describing text: You might want to pull out the name of the singer and/or the songwriter, the length of the song, or the name of the album on which it appears. But what if you wanted to describe the actual content of the song? You could write out the lyrics, but describing the music itself requires a different approach.

Describing music presents challenges quite different from those involved in describing texts or images. Poems and paintings are tangible things that we can look at and contemplate, while the aural nature of music means that it is a fleeting phenomenon that can only be experienced in the performative moment. Even musical scores and recordings, while as much tangible things as paintings and poems, are merely containers that hold the potential for musical experience and not the music itself. Most contemporary popular music is in the form of songs, in which texts are set to a melody and supported by instrumental harmonies. If we want to categorize or describe such music by its lyrical content, we can still rely on methods for describing texts. But if we want to describe the music itself, we need to take a somewhat different approach.

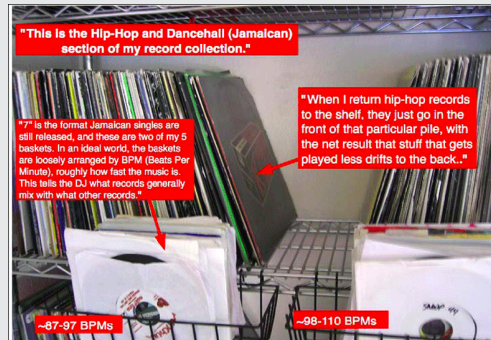
Several people and companies working in multimedia have explored different processes for how songs are described. On the heavily technological side, software applications such as Shazam and Midomi can create a content-based *audio fingerprint* from a snippet of music. *Audio fingerprinting* renders a digital description of a piece of music, which a computer can then interpret and compare to other digital descriptions in a library.^{303[Com]}

On the face of it, contemporary music streaming services represent the apex of music classification and description. Pandora, for example, employs trained musicologists to listen to the music and then categorize the genres and musical materials according to a highly controlled musical vocabulary. The resulting algorithm, the “Music Genome,” can essentially learn to define a listener’s musical tastes by means of this musical tagging, and can then use that information to suggest other music with similar characteristics.^{304[Com]}

But musicians have been thinking about how to describe music for centuries, and while the Music Genome certainly brims with complexity, it pales in comparison to the sophistication of the much older “pen-and-paper” methods from which it derives. Ethnomusicology (loosely defined as the study of global musical practices in their social contexts) has arguably made greater strides towards comprehensive descriptions of musical resources than any other field of musicological study. Since the late 19th century, ethnomusicologists have created complex methods of notation and stylistic taxonomies to capture and categorize the music of both Western and non-Western cultures.

On a more granular level, musicians are endlessly innovative in finding ways to categorize, describe, and analyze not simply large-scale musical genres, but the notes themselves. In the accompanying photo showing the record collection of professional DJ “Kid Kameleon,” we see that the records are arranged not simply by genre, but also by beats-per-minute (BPM). For Kid Kameleon, these re-

A DJ Describes and Organizes Music



Casual music fans might describe their music using the names of the songs or performers and might organize it according to genres like “Pop,” “Rock,” or “Classical.” A professional DJ, however, emphasizes different properties, especially the beats per minute of the music.

This annotated photo shows a portion of the music collection of noted DJ “Kid Kameleon” (<http://kidkameleon.com/>).

(Photo and annotation by Matt Earp. Used with permission.)

cords represent the resources of his musical creative process, and arranging them by BPM allows him to pull exactly the correct musical material he needs to keep the music flowing during a performance. His classification system is therefore a taxonomy that moves from the broad strokes of genre down to the fine grains of specific arrangements of notes and rhythms. This photo is not simply a picture of a record collection: it is a visual representation of an artist's creative process.

5.4.4 Describing Video

Video is yet another resource domain where work to create resource descriptions to make search more effective is ongoing. Video analytics techniques can segment a video into shorter clips described according to their color, direction of motion, size of objects, and other characteristics. Identifying anomalous events and faces of people in video has obvious applications in security and surveillance.^{307[Com]} Identifying specific content details about a video currently takes a significant amount of human intervention, though it is possible that image signature-matching algorithms will take over in the future because they would enable automated ad placement in videos and television.^{308[Bus]}

5.5 Key Points in Chapter Five

- Information retrieval is characterized as comparing a description of a user's needs with descriptions of the resources that might satisfy them. Different property descriptions determine the comparison algorithms and the way in which relevance or similarity of descriptions is determined.

(See §5.2.1 Naming {and, or, vs.} Describing (page 219))

- In different contexts, the terms in resource descriptions are called keywords, index terms, attributes, attribute values, elements, data elements, data values, or “the vocabulary,” labels, or tags.

(See §5.2.2 “Description” as an Inclusive Term (page 220))

- In the library science context of *bibliographic description*, a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources.

(See §5.2.2 “Description” as an Inclusive Term (page 220))

- A bibliographic description of an information resource is most commonly realized as a structured record in a standard format that describes a specific resource.

(See §5.2.2.1 Bibliographic Descriptions (page 220))

- Metadata is structured description for information resources of any kind, which makes it a superset of bibliographic description.
(See §5.2.2.2 Metadata (page 221))
- A relational database schema is designed to restrict resource descriptions to be simple and completely regular sets of attribute-value pairs.
(See §5.2.2.2 Metadata (page 221))
- The Resource Description Framework (RDF) is a language for making computer-processable statements about web resources that is the foundation for the vision of the Semantic Web.
(See §5.2.2.4 Resource Description Framework (RDF) (page 223))
- An aggregation is a set of information objects that, when considered together, compose another named information object.
(See §5.2.2.4 Resource Description Framework (RDF) (page 223))
- The dominant historical view treats resource descriptions as a package of statements, an alternate framework focuses on each individual description or assertion about a single resource.
(See §5.2.3 Frameworks for Resource Description (page 226))
- Design of the description vocabulary should focus on the user of the descriptions. Svenonius proposes five principles for a description vocabulary: user convenience, representation, sufficiency and necessity, standardization, and integration.
(See §5.3 The Process of Describing Resources (page 227))
- The process of describing resources involves several interdependent and iterative steps, including determining scope, focus and purposes, identifying resource properties, designing the description vocabulary, designing the description form and implementation, and creating and evaluating the descriptions.
(See §5.3 The Process of Describing Resources (page 227) and Figure 5.3, The Process of Describing Resources.)
- A collection of resource descriptions is vastly more useful when every resource is described using common description elements or terms that apply to every resource; this specification is most often called a schema or model.
(See §5.3.1.2 Abstraction in Resource Description (page 232))
- XML schemas are often used to define web forms that capture resource instances, and are also used to describe the interfaces to web services and other computational resources.
(See §5.3.1.2 Abstraction in Resource Description (page 232))

- When the task of resource description is standardized, the work can be distributed among many describers whose results are shared. This is the principle on which centralized bibliographic description has been based for a century.

(See §5.3.1.3 Scope, Scale, and Resource Description (page 232))

- Resource description can facilitate the discovery of resources, specify their capabilities and compatibility, authenticate them, and indicate their appraised value.

(See §5.3.2.1 Resource Description to Support Selection (page 234))

- The Functional Requirements for Bibliographic Records (FRBR) presents four purposes that apply generically: Finding, Identifying, Selecting, and Obtaining resources.

(See §5.3.2.3 Resource Description to Support Interactions (page 236))

- The variety and functions of the interactions with digital resources depends on the richness of their structural, semantic, and format description.

(See §5.3.2.3 Resource Description to Support Interactions (page 236))

- *Sensemaking* is the foundation of organizing, as it is the basic human activity of making sense of the world. Sensemaking encompasses the range of organizing activities from the very informal and personal to systematic scientific processes.

(See §5.3.2.5 Resource Description for Sensemaking and Science (page 238))

- Any particular resource might need many resource descriptions, all of which relate to different properties, depending on the interactions that need to be supported and the context in which they take place.

(See §5.3.3 Identifying Properties (page 241))

- Two important dimensions for understanding and contrasting resource properties are whether the properties are intrinsically or extrinsically associated with the resource, and whether the properties are static or dynamic.

(See §5.3.3 Identifying Properties (page 241))

- Recent advances in computing technology and *data science* techniques are making it possible to discover or create resource properties that are called “latent” because they are inferred rather than observed.

(See the sidebar, Latent Feature Creation and Netflix Recommendations (page 246))

- A *controlled vocabulary* is a fixed or closed set of description terms in some domain with precise definitions that is used instead of the vocabulary that people would otherwise use. A controlled vocabulary reduces synonymy and homonymy.

- Professionally created resource descriptions, author or user created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs.

(See §5.3.6 Creating Resource Descriptions (page 251))

- The most commonly used criteria for evaluating resource descriptions are accuracy, completeness, and consistency. Other typical criteria are timeliness, interoperability, and usability.

(See §5.3.7 Evaluating Resource Descriptions (page 254))

- Computational methods can describe and classify images, identify and classify sounds and music, and identify anomalous events in video.

(See §5.4 Describing Non-text Resources (page 257))

Endnotes for Chapter 5

[228][Com] Most digital cameras use the Exchangeable Image File Format (EXIF). The best source of information about it looks like its Wikipedia entry. http://en.wikipedia.org/wiki/Exchangeable_image_file_format.

[229][CogSci] This is much more than just a “kids say the darnedest things” story (see http://en.wikipedia.org/wiki/Kids_Say_the_Darndest_Things). Giles Turnbull (Turnbull 2009) noticed that his kids never used the official names for Lego blocks (e.g., Brick 2x2). He then asked other kids what their names were for 32 types of Lego blocks. His survey showed that the kids mostly used different names, but each created names that followed some systematic principles. The most standard name was the “light saber,” used by every kid in Turnbull’s sample.

[230][Ling] (Reaney and Wilson 1997) classify surnames as local, surnames of relationship, surnames of occupation or office, and nicknames. The dominance of occupational names reflects the fact that there are fewer occupations than places. While there are only a handful of kinship relationships used in surnames (patronymic or father-based names are most common), because the surname includes the father’s name there is more variation than for occupations.

[231][Ling] This odd convention is preserved today in wedding invitations, causing some feminist teeth gnashing (Geller 1999).

[232][CogSci] See (Donnellan 1966). A contemporary analysis from the perspective of cognitive science is (Heller, Gorman, and Tanenhaus 2012).

[235][Com] (Rubinsky and Maloney 1997) capture this transitional perspective. A more recent text on XML is (Goldberg 2008).

[236][Com] See (Sen 2004), (Laskey 2005).

[237][IA] See (Marlow, Naaman, Boyd, and Davis 2006). These authors propose a conceptual model of tagging that includes (1) tags assigned to a specific resource, (2) connections or links between resources, and (3) connections or links between users and explain how any two of these can be used to infer information about the other.

[238][IA] (Hammond, Hanney, Lund, and Scott 2004) coined the phrase “tag soup” in an review of social bookmarking tools written early in the tagging era that remains insightful today. Many of the specific tools are no longer around, but the reasons why people tag are still the same.

[239][Web] Making tagging more systematic leads to “tag convergence” in which the distribution of tags for a particular resource stabilizes over time (Golder and Huberman 2006). Consider three things a user might do if his tag does not match the suggested tags; (1) Change the tag to conform? (2) Keep the tag to influence the group norm? (3) Add the proposed tag but keep his tag as well?

[240][Web] (RDF Working Group 2004). The official source for all things RDF is the W3C RDF page at <http://www.w3.org/RDF/>.

[241][Web] Some argue that the resource being described is thus Bart Simpson’s Wikipedia page, not Bart Simpson himself. Whether or not that is an important distinction is a controversial question among RDF architects and users.

[242][Web] (Heath and Bizer 2011) and <http://linkeddata.org> are excellent sources.

[244][IA] (Pancake 2012)

[251][Com] Because the relational database schema serves as a model for the creation of resource descriptions, it is designed to restrict the descriptions to be simple and completely regular sets of attribute-value pairs. The database schema specifies the overall structure of the tables and especially their columns, which will contain the attribute values that describe each resource. An employee table might have columns for the attributes of employee ID, hiring date, department, and salary. A date attribute will be restricted to a value that is a date, while an employee salary will be restricted according to salary ranges established by the human resources department. This makes the name of the attribute and the constraints on attribute values into resource descriptions that apply to the entire class of resources described by the table.

It is often necessary to associate some descriptions with individual resources that are specific to that instance and other kinds of descriptions that reflect the abstract class to which the instance belongs. When a typical car comes off the assembly line, it has only one instance-level description that differentiates it from its peers: its vehicle identification number (VIN). Specific cars have individualized interior and exterior colors and installed options, and they all have a

date and location of manufacture. Other description elements have values that are shared with many other cars of the same model and year, like suggested price and the additional option packages, or configurations that can be applied to it before it is delivered to a customer. Alternatively, any descriptive information that applies to multiple cars of the same model year could be part of a resource description at that level that is referred to rather than duplicated in instance descriptions.

[252][Com] Web services are generally implemented using XML documents as their inputs and outputs. The interfaces to web services are typically described using an XML vocabulary called *Web Services Description Language (WSDL)*. See (Erl 2005b), especially Ch. 3, *Introduction to Web Services Technologies*.

[256][CogSci] The semantic “bluntness” of a minimalist vocabulary is illustrated by the examples for use of the “creator” element in an official Dublin Core user guide (Hillmann 2005) that shows “Shakespeare, William” and “Hubble Telescope” as creators.

[257][Com] The Intel Core 2 Duo Processor has detailed specifications (<http://www.intel.com/products/processor/core2duo/specifications.htm>) and seven categories of technical documentation: application notes, datasheets, design guides, manuals, updates, support components, and white papers (<http://www.intel.com/design/core2duo/documentation.htm>).

[258][Bus] Real estate advertisements are notorious for their creative descriptions; a house “convenient to transportation” is most likely next to a noisy highway, and a house in a “secluded location” is in a remote and desolate part of town.

[259][Bus] In its early days, when US consumers were generally unaware that Sony was a Japanese company and the quality of Japanese products was viewed in a negative light, Sony would make the “Made in Japan” label as inconspicuous as it could get away with. (John 1999)

In the summer of 2015, the consumer advocacy organization Truth in Advertising reported finding on Walmart’s website over 100 product descriptions inaccurately presenting the products as being made in the United States. (See <https://www.truthinadvertising.org/walmart-made-in-usa/>)

[260][CogSci] Findings from a study of four online dating services (Toma et al 2008) found that 81% of people lied about at least one characteristic. Men were more likely to lie about height, while women lied more about weight, and the further their actual heights and weights were from the mean, the more they lied. A later study (Hall et al 2010) confirmed the finding for women and weight, but also found that men are highly likely to misrepresent their personal assets.

[262][CogSci] In the very busy and dangerous environment of an aircraft carrier flight deck, the sailors wear vests and shirts that are color-coded to their jobs.

For example, red shirts handle munitions, purple shirts handle fuel, green shirts run the catapults and hooks that launch and land the jets, and yellow shirts manage the flights. Color makes it faster and takes less attention for people to see if the right people are where they are supposed to be

The official Navy color chart for aircraft carrier personnel is available at <http://www.navy.mil/navydata/ships/carriers/rainbow.asp>

A similar principle is used in some sports; goalies wear different color jerseys to make it easier to enforce position-specific rules, and football quarterbacks wear distinctive practice jerseys to remind defensive players not to tackle them and possibly injure them.

It is worth noting that color blindness affects approximately 7% of the population.

[263][Law] The Creative Commons nonprofit organization defines six kinds of copy-right licenses that differ in the extent they allow commercial uses or modifications of an original resource (see <http://creativecommons.org/licenses/>). The Flickr photo sharing application is a good example of a site where a search for reusable resources can use the Creative Commons licenses to filter the results (<http://www.flickr.com/creativecommons/>).

[264][Bus] Using the same standards to describe products or to specify the execution of business processes can facilitate the implementation and operation of information-intensive business models because information can then flow between services or firms without human intervention. In turn this enables the business to become more demand or event-driven rather than forecast driven, making it a more “adaptive,” “agile,” or “on demand” enterprise. See (Glushko and McGrath 2005), especially Ch. 5, *How Models and Patterns Evolve*.

[265][Web] For new resources, the labor-intensive cost of traditional bibliographic description is less justifiable when you can follow a link from a resource description to the digital resource it describes and quickly decide its relevance. That is, web search engines demonstrate that algorithmic analysis of the content of information resources can make them self-describing to a significant degree, reducing the need for bibliographic description.

[268][Com] Ken Holman’s *Definitive XSLT and XPath* (Holman 2001) is the book to get started on with XPath, and no one has taught more people about XPath than Holman. The first five hours of a 24-hour video course on *Practical Transformation Using XSLT and XPath* is available for free at <http://www.udemy.com/practical-transformation-using-xslt-and-xpath>.

[270][DS] (Lockyer 1893) and (Bell 1970)

The joint story of Brahe’s data collecting and Kepler’s analysis and theorizing is told in an entertaining manner in (Ferguson 2002). An equally fascinating analy-

sis that interprets Kepler's conceptual shifts with a model of analogical reasoning is (Gentner et al., 1997).

[271][Bus] The concept of *sensemaking* originated from business school research in management and organizational theory (Weick 1995) but has been widely employed by ethnographers in many contexts including emergency rooms, classrooms with minority students, airline safety inspections, and crime investigation. See (Weick 2005) and (Chater 2016)

[272][Phil] Occam's Razor has a long tradition in scientific philosophy, but some people have argued that it is overrated as a heuristic for choosing among alternative explanations or theories, particularly because it depends on how you define simplicity.

[273][Com] One way to make simplicity more useful as a guide for choosing between mathematical models is to explicitly penalize those that are more complex by adding error to the predictions, a technique that computer scientists have given the non-intuitive name of regularization. This penalty requires complex models to be significantly better at explaining the data than simpler ones because they have to overcome the added errors.

[274][CogSci] For example, the composition of a chair is presented here as a static intrinsic property but in fact a wooden chair might deteriorate over time as a result of exposure to sunlight, heat, or biological agents that attack it. A skill can be considered intrinsic and dynamic, but it might also be highly dependent on context, making it extrinsic. The subject category assigned to a book is extrinsic and static, but if the classification system is revised the book might be reclassified. Finally, while the location of a resource can be extrinsic and dynamic, the location history of the resource at some specific point in time is a fact, an intrinsic and static property.

[275][Com] (Dey 2001) further defines the "environment" of *context* as places, people, and things, and for each of "entities" there are four categories of context information: location, identity, status (or activity), and time. This *framework* thus yields 12 dimensions for describing the context of an environment.

[276][Com] A fascinating story about the Netflix's design and use of tens of thousands of movie sub-genres in its recommendation system is (Madrigal 2014).

[277][Ling] Consider how many events are named by appending a "-gate" suffix to imply that there is something scandalous or unethical going on that is being covered up. This cultural description is not immediately meaningful to anyone who does not know about the break-in at the headquarters of the Democratic National Committee headquarters at the Watergate hotel and subsequent cover-up that led to the 1974 resignation of US President Richard Nixon. A list of "-gate" events is maintained at http://en.wikipedia.org/wiki/List_of_scandals_with_%22-gate%22_suffix.

[278][Ling] (http://en.wikipedia.org/wiki/Holbein_carpet).

[280][Com] (Laskey 2005).

[282][DS] We cannot cite all of mathematical statistics in one short endnote, but if you are inclined to learn more, (Mardia, Kent, and Bibby 1980) and (Lee and Verleysen 2007) are the kindest and gentlest resources. If we look very generously at “dimensionality reduction” we might even consider the indexing step of eliminating “stop words” to be a form of dimensionality reduction. Stop words appear with such high frequency that they have no discriminating power, so they are discarded from queries and not part of the description of the indexed documents.

[285][Bus] Many institutional organizing systems are subject to a single centralized or governmental authority that can impose principles for describing and arranging resources. Examples of organizing systems where resources are described using standard centralized principles are:

Companies that follow industry standards for information or process models, product classification or identification to be eligible for government business (Shah and Kesan 2006).

Legislative documents that conform to National or European Community standards for structure, naming, and description (Biasiotti 2008).

The Internet Corporation for Assigned Names and Numbers (ICANN) and its policies for operating the Domain Name System (DNS) make it possible for every website to be located using its logical name (like “berkeley.edu” rather than using an IP address like 169.229.131.81). (<http://www.icann.org/>)

In other domains multiple organizations or institutions have the authority to impose principles of resource description. Sometimes this authority derives from the voluntary collaboration of multiple autonomous parties who set and conform to standards because they benefit from being able to share resources or information about resources. Examples of organizing systems where resources are described using standardized decentralized principles are:

Firms that establish company-wide standards for their information resources, typically including the organization and management of source content, document type models, and a style guide that applies to print and web documents.

Firms that participate in the OASIS (<http://www.oasis-open.org/>) or the W3C (<http://www.w3.org/>) industry consortia to establish specifications or technical recommendations for their information systems or web services).

[287][Web] (Sen et al. 2006) analyze the effects of four tag selection algorithms used in sites that allow user tags on vocabulary evolution (more often called

“tag convergence” in the literature), tag utility, tag adoption, and user satisfaction.

[288][CogSci] But in an often-cited essay (Doctorow, 2001) provocatively titled “*Metacrap: Putting the torch to seven straw-men of the meta-utopia*,” Cory Doctorow argues that much human-created metadata is of low quality because “people lie, people are lazy, people are stupid, mission impossible—know thyself, schemas are not neutral, metrics influence results, (and) there is more than one way to describe something.”

[289][Com] (Hu and Lui 2004).

[290][Com] <http://www.amazon.com/gp/search-inside/sipshelp.html/>.

[291][DS] The title says it all: *Predictive analytics: The power to predict who will click, buy, lie, or die*. (Siegel 2013). The Bayesian Surname and Geocoding technique for predicting race is described by (Elliott et al. 2008).

[293][Bus] However, these concerns are rapidly becoming more important in the public sector. In particular, many public universities in the US are struggling with cuts in state and federal funding that are affecting library services and practices.

[294][Bus] More generally, economists use the concept of the “mode of exchange” in a business relationship to include the procedures and norms that govern routine behavior between business partners. An “exit” mode is one in which the buyer makes little long-term commitment to a supplier, and problems with a supplier cause the buyer to find a new one. In contrast, in “voice” mode there is much greater commitment and communication between the parties, usually leading to improved processes and designs. See (Helper and McDuffie 2003).

[296][Com] (Nunberg 2009) called the quality of Google’s metadata “a disaster for scholars,” but (Sag 2012) argues that the otherwise neglected “orphan works” in the Google corpus are “grist for the data mill.”

[297][Bus] The modern “quality movement” grew out of the efforts of the US to rebuild Japan after the Second World War and its “Bible” was Juran’s 1951 *Quality Control Handbook* (Juran 1951).

[300][Com] See (Datta et al. 2008). The company Idée is developing a variety of image search algorithms, which use image signatures and measures of visual similarity to return photos similar to those a user asks to see.

[301][Web] (von Ahn and Dabbish 2008).

[302][Com] The key idea that made deep learning possible is the use of “backpropagation” to adjust the weights on features by working backwards from the output (the object classification produced by the network) all the way back to the input. Mathematically-sophisticated readers can find a concise explanation and history

of deep learning in (LeCun, Bengio, and Hinton 2015). LeCun and Hinton were part of research teams that independently invented backpropagation in the mid 1980s. Today, LeCun heads Facebook's research group on artificial intelligence, and Hinton has a similar role at Google.

[303][Com] (Cano et al. 2005).

[304][Com] (Walker 2009).

[307][Com] (Regazzoni et al. 2010) introduce a special issue in IEEE *Signal Processing on visual analytics*.

[308][Bus] One organization that sees a future in assembling better descriptions of video content is the United States' National Football League (NFL), whose vast library of clips can not only be used to gather plays for highlight reels and specials but can also be monetized by pointing out when key advertisers' products appear on film. Currently, labeling the video requires a person to watch the scenes and tag elements of each frame, but once those tags have been created and sequenced along with the video, they can be more easily searched in computerized, automated ways (Buhrmester 2007).