Chapter 4. Resource Description and Metadata

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

Robert J. Glushko
Kimra McPherson
Ryan Greenberg
Matthew Mayernik

4.1 Introduction

Click. A professional photographer standing on a mountainside takes a picture with a digital camera. What information should be recorded and associated with the recorded image of the mountain scene? 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.[1] 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.

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. That does not matter, so long as their descriptions allow the boys to understand each other.[2]

Digital photos and Lego blocks are very different, yet for our purposes these scenarios are both about resource description. Together both scenarios raise important questions about describing resources that we 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?

We begin with an overview of resource description (Section 4.2), which we propose as a broad concept that includes the narrower concepts of bibliographic descriptions and metadata. Section 4.3 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 Section 4.4 we briefly discuss the issues that arise when describing museum and artistic resources, images, music, video, and contextual resources.

4.2 An Overview of Resource Description

We describe resources so that we can refer to them, distinguish among them, search for them, control access to them, and preserve them. Each purpose might require different resource descriptions. We use resource descriptions in every communication and conversation, and they are the enablers of organizing systems.

4.2.1 Naming {and, or, vs.} Describing

Chapter 3 discussed how to decide what things should be treated as resources and how names and identifiers distinguish one resource from another. Many names are literally resource descriptions, or once were. 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).[3] 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). In many cultures throughout the world, it has been very common for one spouse or the other to take on a name that describes their marital relationship. Historically, in many parts of the English-speaking world, married women have often referred to themselves using their husband's name; when Jane Smith married John Brown, her name became Jane Brown or Mrs. John Brown.[4]

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.[5]

Similarly, when we use a library catalog or search engine to locate a known resource, such as a particular book or document, we query for it using its name, or other specific information we know about it, to increase the likelihood that we can find it. In contrast, when we are looking for resources to satisfy an information need but don't 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.

4.2.2 "Description" as an Inclusive Term

Up to now we have used the concept of "description" without defining it because we have most often used it in its ordinary sense to mean 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. For example, 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 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.

All of these terms come from a relatively narrow semantic scope in which the purpose of description is to identify and characterize the essence, or aboutness, of a resource. This leaves out many kinds of information that can be associated with a resource to support additional purposes; for example, information that specifies access controls or possible uses of the resource, or information about actual uses. We will describe many of these purposes and the types of information needed to enable them in Section 4.3.2, and we use "resource description" in an expansive way to accommodate all of them.

Chapter 3 introduced the distinction of "Resource Focus" 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 Resource Description Framework (RDF) language used to make statements about Web resources and physical resources that can be identified on the Web.

4.2.2.1 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 resources and the "entities that populate the bibliographic universe," which include works, editions, authors, and subjects. Despite the "biblio-" root, bibliographic descriptions are applied to all of the resource types contained in libraries, not just books. Note also that this definition includes not just the information resources being described as distinct instances, but also as sets of related instances and the nature of those relationships.[6]

A bibliographic description of an information resource is typically realized as a structured record in a standardized format that describes a specific resource. The earliest bibliographic records in the 19th century were those in book catalogs, which organized for each author a list of his authored books, with separate entries for each edition and physical copy. Relationships between books by different authors were described using cross-references.

The nature and extent of bibliographic descriptions were highly constrained by the book catalog format, which also made the process of description a highly localized one because every library or collection of resources created its own catalog. The adoption of printed cards as the unit of organization for bibliographic descriptions around the turn of the 20th century made it easier to maintain the catalog, and also enabled the centralized creation of the records by the Library of Congress.

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 cataloguing rules designed for collections of books and other physical resources and intended only for use by people.

4.2.2.2 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, which makes it a superset of bibliographic 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. In addition, metadata has originally meant and still most often refers to descriptions of classes or collections of resources rather than descriptions of individual resources. The earliest metadata resources, 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 largely superseded beginning in 1997 by the eXtensible Markup Language (XML), whose initial purpose was to bring SGML to the web to make web content more structured and computer-processable.[7]

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 data sets, 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.[8]

The concept of metadata has more recently 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 web resource or physical resource with a web presence that can be annotated.

4.2.2.3 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.[9] 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 URI (Uniform Resource Identifier). URIs can take various forms, but you're probably most familiar with the URIs used to identify web pages, such as `http://springfield-elementary.edu/`. (You're 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 for relating a person to (the web page of) a school they attended. Values can be resources too, but they don't have to be: when a property takes simple values like numbers, dates, or text strings, these values do not have URIs and so aren't 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." The example below 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.[10] 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.

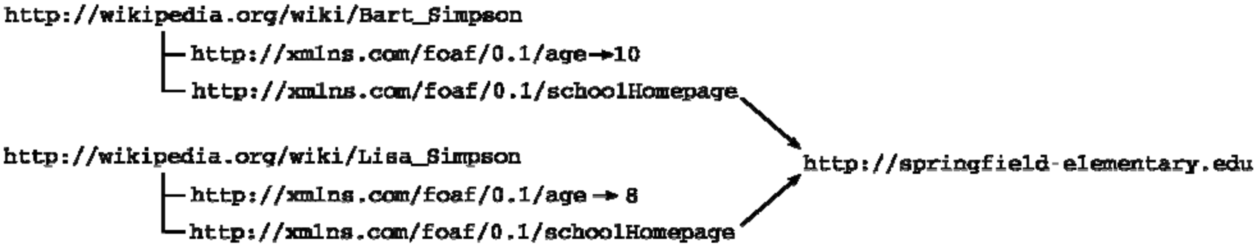
http://wikipedia.org/wiki/Bart_Simpson
http://xmlns.com/foaf/0.1/age
10
http://xmlns.com/foaf/0.1/schoolHomepage
http://wikipedia.org/wiki/Lisa_Simpson
http://xmlns.com/foaf/0.1/age
8
http://xmlns.com/foaf/0.1/schoolHomepage
http://springfield-elementary.edu

Figure 1. RDF triples arranged into a graph.

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." roponents of RDF claim that this will greatly benefit knowledge discovery and inference.[11] But the benefits of RDFs highly prescriptive description form must be weighed against the costs; turning existing descriptions as RDF can be labor-intensive.

RDF can be used for bibliographic description, and some libraries are exploring whether RDF transformations of their legacy bibliographic records can be exposed and integrated with resource descriptions on the open web. This activity has raised technical concerns about whether the RDF model of description is sophisticated enough and more fundamental concerns about the desirability of losing control over library resources.[12]

4.2.3 Frameworks and Classifications 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.

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 framework for resource description focuses on each individual description or assertion about a single resource. This "statement at a time" view of resource is more typical when descriptions are assigned to web objects or resources, and much of the discussion about this view is framed in terms of the particular syntactic forms being used in RDF.

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 the fundamental idea of the discipline of organizing (introduced in Section 1.2.3.1), 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 8.

In library and information science, it is very common to discuss resource descriptions using a classification proposed by Arlene Taylor, which distinguishes administrative, structural, and descriptive metadata.[13] A similar typology proposed by Gilliland breaks metadata down into five types: administrative, descriptive, preservation, use, and technical.[14] Both of these classifications imply a narrow notion of descriptive metadata that reflects the historical emphasis on bibliographic description, in contrast to our view that treats resource description as a more inclusive category. In addition, these classifications do not always distinguish between intrinsic and extrinsic properties (as we will see in Section 4.3.3), and they often mix and match design and implementation considerations.

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 2: selecting, organizing, interacting with, and maintaining resources.

4.3 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 also enables us to propose more standard vocabulary 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 several interdependent and iterative steps. We begin with a generic summary of the process to set the stage for a detailed step-by-step discussion.

- **Determining Scope and Focus**: Identifying resources to describe is the first step; this topic is covered in detail in Section 3.3. The resource scope and domain 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 or resource types?

- **Determining Purposes**: 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 2. 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;

- **Identifying Resource Properties**: 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 contrasts between intrinsic and extrinsic properties, and between static and dynamic properties, are useful to identify appropriate resource properties;

- **Designing the Description Vocabulary**: 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? A good description vocabulary will be easy to assign when creating resource descriptions and easy to understand when using them;

- **Designing the Description Form and Implementation**: 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;

- **Creating the Descriptions**: 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;

- **Evaluating the Descriptions**: 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.

How explicit and systematic each step needs to be depends on the resource scope and domain, 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, the 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 the digital camera were chosen by an industry association and published as a technical specification implemented by camera and mobile phone manufacturers worldwide. If a professional photographer is taking the photo for commercial purposes, many of the other descriptions assigned to the image to identify ownership, rights management and syndication are likely to conform to formal specifications and be managed in institutional information systems.

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.

4.3.1 Determining the Scope and Focus of Resource Description

Which resources do we want to describe? As we saw in Chapter 3, 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 Section 3.2.4, 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. 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.

4.3.1.1 Granularity – Describing Instances or Describing Collections

It is simplest to think of a resource description as being associated with another individual resource. However, as discussed in Chapter 3, it can be challenging to determine what to treat as an individual resource when resources are themselves objects or systems that are 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 we have decided 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; see Section 1.2.2). 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 data sets 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. Consider the information systems that commodity traders use to access descriptions of real resources in markets all over the world: some traders are concerned with weekly production in their region, while others are monitoring real-time global flows of precious metals and petroleum products. 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 thousands of 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.

4.3.1.2 Abstraction in Resource Description

We can also associate resource descriptions with an entire type or domain of resources (see Sections 2.5.2.3 and 3.2.1). 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. The specification of the set of descriptions that apply to an entire resource type is called a schema, model, or metadata standard. 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?" in Section 2.1).

A relational database, for example, is easily conceptualized as a collection of records organized as a table, 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.

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 table and especially its 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.

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 greatly improved on SGML and XML by expressing the description of the document schema in XML itself, using the same syntax as the resource it describes, making it easy to create resources using the metadata as a template or pattern. As a result, 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 are often used to define web forms that capture resource instances (each filled-out form). XML schemas as also used to describe the interfaces to web services and other computational resources.[15]

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.

4.3.1.3 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, such as “what is the name of the resource?”, “who created it?”, or “what type of content or matter does it contain?”, need to be easy to address. 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. The description language typically evolves from a simple list of descriptive terms to a glossary with definitions, to a highly controlled vocabulary with content rules for allowable values, and, finally, to a thesaurus in which each term is also defined with respect to its semantic relationships to other terms that are broader, narrower, or otherwise associated with it.[16]

This co-evolution of descriptive scope and description complexity is easy to see in the highly complex bibliographic descriptions created by professional cataloguers. The commonly used AACR2 cataloguing standards distinguish 11 different categories of resources and specify several hundred descriptive elements.[17] When the task of resource description is standardized, the work is distributed among many describers whose results are shared. This principle 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 metadata element set with only 15 elements was proposed as a vastly simpler description vocabulary that people not trained as professional cataloguers could use.[18] 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. In this respect, we can also view the Dublin Core as part of the intellectual foundations for the “crowdsourcing” or “community curation” of resource descriptions by non-professionals (Section 2.5.3.3).

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” results in a substantial loss of precision when the description is created and can cause misunderstanding when the descriptions are reused.[19]

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.

4.3.2 Determining the Purposes of Resource Description

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 2).

4.3.2.1 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 resources are available that 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.[20] 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.

- **Authentication**: Is the resource what it claims to be? (Section 3.5.3). 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 (Section 3.5.4), 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 can also take the perspective of the person creating the resource description and consider his or her primary purpose, which is often to encourage the selection of the resource by someone else. This is what product marketing is about – devising names and descriptions to make a resource distinctive and attractive compared to alternatives. A fish once known as the Patagonian Toothfish became popular in American restaurants when a fish wholesaler began marketing it as the Chilean Sea Bass. Apple has consistently described its products to emphasize experiential or cultural properties, as compared with Intel or other PC manufacturers, whose descriptions emphasize technical specifications.[21]

4.3.2.2 Resource Description to Support Organizing

Often, the activities of organizing resources and designing interactions with them are intertwined, but they are logically separate. We define organizing as specifying the principles or rules for describing and arranging resources in order to create the capabilities on which interactions are based. This lets us treat the design and implementation of resource interactions as if those were separate and subsequent activities. For example, assigning keywords to documents that describe their contents is an organizing activity, while designing and implementing an information retrieval application that uses the keywords is the design of resource interactions.

Physical resources are often organized according to their tangible or perceivable properties like size, color, material of composition, or shape (Section 2.3.1.1).[22] For other types of physical resources, however, such as hazardous materials, it is the descriptions and not the directly perceivable properties that determine or constrain how the resources are organized (Section 2.3.1.2). Similarly, building codes or other regulations associated with physical resources can prescribe or prohibit particular resource arrangements. Rules governing the collection, integration, and analysis of personal information are also resource descriptions that influence the organization of information resources.

Any types of resources that have sortable identifiers can be organized using that descriptive element.

4.3.2.3 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. For bibliographic resources these interactions and the models of resource descriptions needed to support them have been formalized as the Functional Requirements for Bibliographic Records (FRBR).[23] FRBR presents four purposes that apply generically to organizing systems, not just bibliographic ones: Finding, Identifying, Selecting, and Obtaining resources.

- **Finding**: What resources are available that "correspond to the user's stated search criteria" and thus can satisfy an information need? Before there were online catalogs and digital libraries, we found resources by referencing catalogs of printed resource descriptions incorporating the title, author, and subject terms as access points into the collection; the subject descriptions were the most important finding aids when the user had no particular resource in mind. 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 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. In bibliographic contexts this might mean finding the resource that is identified by its citation. 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.

- **Selecting**: In this context we mean 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.[24] A business might select a supplier or distributor that uses the same standard or industry reference model to describe its products or business processes.[25]

- **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.[26]

Elaine Svenonius proposed that a fifth task be added to the FRBR list:

- **Navigation**: 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.[27]

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 XSLT and XPath to address and select data elements, sub-trees, or other structural parts of the document.[28]

The FRBR framework is the most recent formalization of the purposes of resource description that started in 19th century libraries; it is not surprising that it still reflects some historical bias toward interactions with physical bibliographic resources and the descriptions needed to obtain them. With physical resources, any interactions that take place once the primary resources are obtained are outside the scope of the organizing system because they are not directly supported by it.

With digital resources, on the other hand, many of the purposes of resource description are realized with the primary resources. These purposes usually involve processing of the resource content and structure: analysis, summarization, visualization, transformation, reuse, mixing, remixing… far too many purposes to list here. The core principle that underlies all of these purposes is that the variety and functions of the interactions with digital resources depends on the richness of their structural, semantic, and format description (See Section 2.4.1.2).

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

Another important difference between physical resources and digital ones is that interactions with the latter are easily recorded. Usage records from session logs, browsing, or downloading activities are resource descriptions that can be tied to payments for using the resources or analyzed to influence the selection and organizing of resources in future interactions.

4.3.2.4. Resource Description to Support Maintenance

Many types of resource descriptions that support selection (Section 4.3.2.1) 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.[29]

Other types of resource descriptions more exclusively associated with maintenance activities include version information and effectivity or useful life information. Usage records are also valuable because they enable the identification of resources that are not being accessed, suggesting that they are no longer needed and can thus be safely archived or discarded.

4.3.3 Identifying Resource Properties for Description

Once the purposes of description have been established, the specific properties of the resources needed to enable them need to be identified. 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. Think about how we might describe something as simple as a chair. In your house, you might describe your chair based on the room it is in, e.g., “the kitchen chair.” When you take the same chair to a potluck dinner at a friend’s house, where all the chairs end up in the kitchen, it becomes “my chair,” or “the wooden chair,” or “the folding chair,” or “the black chair with white trim.” Maybe you inherited it, so when you talk to your family, you call it “grandma’s chair.” If you decide to sell it, you will describe it in a way that is intended to encourage someone to buy it, as an “all-oak antique kitchen chair in mint condition.”

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 Section 4.3.1.3, 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 Section 4.3.6, "Creating Resource Descriptions" and again in Section 4.4.1, "Describing Museum and Artistic Resources."

Fourth, what might seem to be the same property at a conceptual level might be very different at an implementation level. Many types of resources have a resource description that is a surrogate or summary in some respects 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 or thumbnail photo of the original. This surrogate is simple to create and it is easy for users to understand its relationship 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? How is a song summarized? Or a poem? Or a tree?

Two important dimensions for understanding and contrasting resource properties used in descriptions and organizing principles are whether the properties are intrinsically or extrinsically associated with the resource, and whether the properties are static or dynamic. Taken together these two dimensions yield four categories of properties.

4.3.3.1 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 size, color, shape, weight, material of composition, and texture of natural or manufactured objects are intrinsic and static properties that are often used to describe and organize physical resources. If a particular Lego is blue, it is set apart from all the not-blue Legos; a square Lego is physically different from a round one. When bibliographic resources were exclusively physical, their sizes and their number of pages were common physical properties in their descriptions.

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. These intrinsic static properties cannot be directly perceived. The author or creator of a resource, the context of its creation, and the duration of a song are other examples of intrinsic and static resource properties.

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. Those statistics about content properties would still be intrinsic descriptions even though an external process creates them. Similarly, "image signatures' or "audio fingerprints" are intrinsic descriptions (Section 4.4).

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 parts, both physical ones like manufactured objects and digital ones like hierarchical documents or databases, 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 syntactically as relations. Chapter 5, "Describing Relationships and Structures," discusses them in great detail.

4.3.3.2 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 different 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.

4.3.3.3 Intrinsic Dynamic Properties

Intrinsic dynamic properties change over time, such as developmental personal characteristics like a person's height and weight, skill proficiency, or intellectual capacity. Because these properties aren't 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, or an academic honor society where membership is based on current grade point average, are using extrinsic dynamic properties to describe and organize the resources.

4.3.3.4 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.

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

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.[30] For the Lego boys familiar with the Star Wars movies, "light saber" was just the obvious word for a long, neon tube with a handle. However, someone who has never seen or heard of Star Wars would not understand this description, and he would describe the piece some other way. Sometimes a cultural description lasts longer than its salience, so it loses its power to evoke anything other than puzzlement about what it might mean.[31]

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."[32] 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 (Section 3.5); 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 5, "Describing Relationships and Structures."

4.3.4 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. Section 3.4, "Naming Resources," 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.

4.3.4.1 Principles of Good Description

In *The Intellectual Foundations of Information Organization*, Svenonius proposes a set of principles or "directives for design" of a description language.[33] Her principles, framed in the narrow context of bibliographic descriptions, still 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 the 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**: Use the same properties and terms for all types of resources whenever possible

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?

4.3.4.2 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 6).

Some resource descriptions are designed to be used by computers or other machines, which seemingly reduces the importance of design principles that consider user preferences or common uses. However, the Standardization and Integration principles become more important for inter-machine communication because they enable efficient processing, reuse of data and software, and increased interoperability between organizing systems.[34]

4.3.4.3 Controlled Vocabularies and Content Rules

As we defined in Section 3.4.4.2, 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 controlled vocabulary (MeSH).[35]

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 5, “Describing Relations and Structures.”

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 datatype (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.

4.3.4.4 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." [36]

These techniques might sound imposing and they are computationally complex, but they all have the same simple concept at their core. What they do is analyze the correlations between resource descriptions to transform a large set into a much smaller set of uncorrelated ones. In a way this implements the principle of Sufficiency and Necessity we mentioned in Section 4.3.4.1 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.

4.3.5 Designing the Description Form and Implementation

At this step in the process of resource description we already 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 8, “The Form of Resource Description.”

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 created, stored, and managed. We will take this more pragmatic perspective in Chapter 10, “The Organizing System Roadmap,” but until then, we will continue to focus on design issues and defer discussion of the implementation choices.

4.3.6 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. From the traditional perspective of library and information science with its emphasis on bibliographic description, these modes of creation imply different levels of description complexity and sophistication; Taylor and Joudrey suggest that professionals create **rich** descriptions, untrained users at best create **structured** ones, and automated processes create **simple** ones.

This classification reflects a disciplinary and historical bias more than reality. "Simple" resource descriptions are "no more than data extracted from the resource itself… the search engine approach to organizing the web through automated indexing techniques."[37] It might be fair to describe an inverted index implementation of a Boolean information retrieval model as simple, but it is clearly wrong to consider what Google and other search engines do to describe and retrieve web resources as simple.[38] A better notion of levels of resource description is one based on the amount of interpretation imposed by the description, an approach that focuses on the descriptions themselves rather than on their methods of creation. We will discuss this sort of approach in Section 4.4.1 in the context of describing museum and artistic resources.

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 social description 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.

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 9. In the discussion that follows we focus on the creation of new resource descriptions where none yet exist.

4.3.6.1 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 cataloguers 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.[39]

4.3.6.2 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 cataloguers would use.

4.3.6.3 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 physical "in the world" 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.[40] 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 what musical genre that file belongs to. 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.[41]

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.[42] 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.

4.3.6.4 Computational and Automated Resource Description

When a digital camera takes a picture, it creates a description in the EXIF file format using properties associated with the camera and its settings, as well as some properties of the context in which the photo is taken. Creating the description by hand would be laborious, especially if constructed retroactively. The downside, however, is that the automated description does not capture the meaning or purpose of the photo. The automated description might contain information about time and place, but not that that people in the picture were on a honeymoon vacation. The difference between automated and human description is called the semantic gap (Section 3.4.2.5).

Some computational approaches create resource descriptions that are similar in purpose to those created by human describers. For example, Hu and Lui created a text mining and summarization system for customer comments about products for sale on the web. Thousands of comments about a particular digital camera are reduced to a list of the most important features.[43] 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.[44]

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 Section 4.4 when we discuss "Describing Non-Text Resources" and in Chapter 9.

4.3.7 Evaluating Resource Descriptions

Evaluation is implicit in many of the activities of organizing systems we described in Chapter 2 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, but the most commonly used ones are accuracy, completeness, and consistency.[45] 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 that would enhance usability.

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.

4.3.7.1 Evaluating the Creation of Resource Descriptions

When resource descriptions are created by professionals in a centralized manner, as they have long been for 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, other quality considerations emerge.

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. [46] 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.[47]

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.[48] 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.[49]

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.

4.3.7.2 Evaluating the Use of Resource Descriptions

The user's perspective is embodied in the FRBR statement of the purposes of bibliographic description (Section 4.3.2.3), but the problems of resource description on the Web have highlighted this point of view. The most important quality criteria are now functional ones – do the resource descriptions satisfy their intended purposes in a usable way?

In many ways, the answer is a disappointing no. In one of the earliest revisions to the original HTML specification, a <META> tag was added to allow website creators to define a set of key terms to describe a web site or web page, thus helping the site's position in search rankings when a user searched for one or more of those terms. 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 altogether, but many other techniques that use false resource descriptions continue to plague web users. (See "Web Resources with Bad Intent" in Section 2.3.3.5).

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.

4.3.7.3 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. 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.

4.4 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 are 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 but cannot (yet) describe "aboutness" in a reliably complete manner.

4.4.1 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.

The 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:

- **Primary subject matter**: At this level, we describe the most basic elements of a work in a generic way that would be recognizable by anyone regardless of expertise or training. The painting *The Last Supper*, for example, might be described as "13 people having dinner."

- **Secondary subject matter** or **identification**: Here, we introduce a level of basic cultural understanding into a description. Someone familiar with a common interpretation of the Bible, for example, could now see *The Last Supper* as representing Jesus surrounded by his disciples.

- **Intrinsic meaning** or **interpretation**: At this level, context and deeper understanding come into play—including what the creator of the description knows about the situation in which the work was created. Why, for example, did this particular artist create this particular depiction of *The Last Supper* in this way? Panofsky posited that professional art historians are needed here, because they are the ones with the education and background necessary to draw meaning from a work.[50]

In other words, Panofsky saw the need for many different types of descriptors—including physical, cultural, and contextual—to work together when making a full description of an artifact.

Professionals who create descriptions of museum and artistic resources, architecture and other cultural works typically use the VRA Core from the Library of Congress, or the Getty "Categories for the Description of Works of Art" (CDWA), a massive controlled vocabulary with 532 categories and subcategories. A CDWA-Lite has been developed to create a very small subset for use by non-specialists.[51]

4.4.2 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 Lightroom. All of these sites provide some facilities for users to assign tags to images or arrange them in named groups.

Computer image analysis techniques are increasingly used to create content-based descriptions. The "visual signature" of an image is 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 objects or settings.[52]

For computers to identify the objects or people in images – creating the kinds of resource descriptions that people want – requires training with tags or labels. Louis van Ann devised a clever way to collect large amounts of labeled images with a web-based game that randomly pairs people to suggest labels or tags for an image. Typically, the obvious choices are removed from contention, so a photo of a bird against a blue sky might already strike "bird" and "sky" from the set of acceptable words, leaving users to suggest words such as "flying" and "cloudless."[53] This technique was later adopted by Google to improve its image search. An analogous effort to automatically identify people in Facebook photos uses photos where people identified themselves or others in photo.

4.4.3 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. A DJ, for example, might care greatly about the beats per minute in each song. If you're making a playlist for a road trip, you might be seeking songs that you'd describe as "good for driving" —though you'd have to figure out what "good for driving" means first, which is a highly subjective description. If you're looking for recommendations for new bands, you might want to know how to find music that's somehow like music you already know you love.

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.[54]

On the other hand, the online radio service Pandora uses music experts, not computers, to create text-based descriptions. The company employs an army of coders, including trained musicologists, who listen to individual pieces of music and determine which words from Pandora's highly controlled vocabulary for musical description apply to a given song. The result is Pandora's "Music Genome," an algorithm that ultimately recommends songs for its users by stripping down the songs they say they like to their component parts and suggesting, for example, more songs with "driving bass" or "jangly guitars." [55]

4.4.4 Describing Video

Video is yet another resource domain where work to create resource descriptions to make search more effective is ongoing. Identifying the content of 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.[56]

4.4.5 Describing Resource Context

As we discussed in section 3.4, sensors are now making all sorts of objects "smarter," capable of reporting their status, their location, or other important descriptive data. Many applications of smart resources are still in their infancy, and that makes them interesting to study for how descriptions can be created automatically and processed (automatically, by people, or both) down the line.

Some sensors create relatively simple descriptions: A pass that registers a car's location at a toll booth, for example, doesn't need to do much besides communicate that the particular sensor has arrived at some location. In essence, it acts as an identifier.

The information created or tracked by sensors can be more complex. Some sensors can calculate location using GPS coordinates and satellite tracking, while others can take readings of temperature, pollution, or other environmental measurements. These readings can be used separately or combined into richer and more detailed descriptions of a resource or event. The tradeoffs in creating these descriptions likely sound familiar by now: More descriptors can create a fuller and more accurate picture of the world, but they require more processing power not only to collect the necessary information but also to render it into a meaningful form.[57]

4.5 Key Points in Chapter Four

- 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.
- 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
- 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.
- 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
- A bibliographic description of an information resource is most commonly realized as a structured record in a standard format that describes a specific resource
- The Functional Requirements for Bibliographic Records (FRBR) presents four purposes that apply generically: Finding, Identifying, Selecting, and Obtaining resources
- Metadata is structured description for information resources of any kind, which makes it a superset of bibliographic description
- The Standard Generalized Markup Language introduced the Document Type Definition (DTD) for describing the structure and content elements in hierarchical document models. SGML was largely superseded by the eXtensible Markup Language (XML)
- 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
- RDF can be used for bibliographic description, and some libraries are exploring whether RDF transformations of their legacy bibliographic records can be exposed and integrated with resource descriptions on the open web
- 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
- A relational database schema is designed to restrict resource descriptions to be simple and completely regular sets of attribute-value pairs

- 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
- 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
- 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
- The variety and functions of the interactions with digital resources depends on the richness of their structural, semantic, and format description.
- 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
- 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.
- 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
- Professionally created resource descriptions, author or user created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs
- 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
- The most commonly used criteria for evaluating resource descriptions are accuracy, completeness, and consistency. Other typical criteria are timeliness, interoperability, and usability.
- Sensors that assign more resource descriptors can create a fuller and more accurate picture of the world, but they require more processing power to collect the necessary information and render it into a meaningful form

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

[2] [CogSci] This is much more than just a "kids say the darndest 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.

[3] [CogSci] Reaney and Wilson (1991) 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.

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

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

[6] [LIS] An excellent source for both the history and theory of bibliographic description is *The Intellectual Foundation of Information Organization* by Elaine Svenonius (2000),

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

[8] [Citation] See (Sen 2004), (Laskey 2005).

[9] [Citation] The official source for all things RDF is the W3C RDF page at http://www.w3.org/RDF/

[10] [Computing] 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 decision is a controversial question among RDF architects and users.

[11] [Citation] Heath and Bizer (2011) and linkeddata.org are excellent sources.

[12] [LIS] Byrne and Goddard (2010) present a balanced analysis of the cultural and technical obstacles to the adoption of RDF and linked data in libraries. Yee (2009) is a highly specific technical demonstration of converting bibliographic descriptions to RDF. A detailed analysis / rebuttal of Yee's article is at http://futurelib.pbworks.com/w/page/13686677/YeeRDF

[13] [LIS] Taylor's book on *The Organization of Information*, now in its 3rd edition (with co-author Daniel Joudrey), has been widely used in library science programs for over a decade.

[14] [LIS] Gilliland 2008.

[15] [Computing] 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 2005), especially Chapter 3, "Introduction to Web Services Technologies."

[16] [LIS] Creating descriptions that can keep pace with the growth of a collection has been an issue for librarians for years, as libraries moved away from describing simply "whatever came across a cataloger's desk" to cataloging resources for a national and even international audience (Svenonius 2000, p. 31).

[17] [LIS] The Anglo-American Cataloguing Rules (AACR2) have rules for books, pamphlets, and printed sheets; cartographic materials; manuscripts and manuscript collections; music; sound recordings; motion pictures and videorecordings; graphic materials; electronic resources; 3-D artifacts; microforms; and continuing resources. *The Concise AACR2 (4th Edition)* is the most accessible treatment of these very complex rules (Gorman 2004). The Resource Description and Access (RDA) vocabularies have been proposed as the successor to AACR2 and make even finer distinctions among resource types. See http://rdvocab.info/vocabulary/list.html

[18] [Citation] See the Dublin Core Metadata Initiative at http://dublincore.org/.

[19] [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 (Hillman 2005) that shows "Shakespeare,Wiliam" and "Hubble Telescope" as creators.

[20] [Computing] 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).

[21] [Business] Real estate advertisements are notorious for their creative descriptions; a house "convenient to transportation" is most likely next to a busy highway, and a house in a "secluded location" is in a remote and desolate part of town.

[22] [CogSci] Typically, this takes place in an unanalyzed or unreflective manner. We can stack boxes on top of each other only if they are of certain relative and absolute sizes. Even if we consciously focus on resource properties when we follow organizing principles, our experience is that of directly arranging the resources, not organizing them on the basic of implicit resource descriptions. Even in the case when the physical resources have descriptions of their sizes and other properties (like labels on boxes or in clothes), when we arrange boxes or clothes, it is still the primary resources that we are organizing, not these descriptions.

[23] [LIS] We encountered FRBR several times in previous chapters (especially in Section 3.1.1.2) where we asked "what is this thing we call 'Macbeth'" and described FRBR's four-level abstraction hierarchy of the "work."

[24] [Law] The Creative Commons nonprofit organization defines six kinds of copyright 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 site 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/).

[25] [Business] 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 Chapter 5,"How Models and Patterns Evolve."

[26] [Computing] 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.

[27][Citation] Svenonius (2000, pages 18-19).

[28] [Citation] Ken Holman's *Definitive XSLT and XPath* (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.

[29] [LIS] The PREMIS standard for preservation metadata is maintained by the US Library of Congress at http://www.loc.gov/standards/premis/. A good place is to start is the 2011 PREMIS Data Dictionary (http://www.loc.gov/standards/premis/v2/premis-2-1.pdf).

[30] [CogSci] 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 isn't immediately meaningful to anyone who doesn't 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.

[31] [CogSci] A particular type of geometrically patterned Turkish rug came to be known as a "Holbein carpet" after the German Renaissance painter Hans Holbein, who often depicted the rugs in his work. Holbein was very famous in his time, and his commissioned paintings of the English King Henry VIII have Henry standing on such rugs. But the rugs themselves existed (and were even painted by others) long before Holbein painted them, and today Holbein is much less famous than he once was. (http://en.wikipedia.org/wiki/Holbein_carpet)

[32] [Computing] (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.

[33] [Citation] Svenonius 2000, Chapter 5.

[34] [Citation] Laskey 2005.

[35] [Citation] http://www.ncbi.nlm.nih.gov/mesh/

[36] [Computing] 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.

[37] [LIS] Taylor and Joudrey 2009, p. 91.

[38] [Computing] See Chapter 4 of (Buttcher, Clarke, and Cormack 2010) for a description of a simple Boolean information retrieval model and Chapter 14 and 15 for descriptions of Google-scale ones. For a popular discussion of the Google algorithm see (Levy 2010),

[39] [Business] 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:

- Libraries that use national bibliographic standards to satisfy requirements set by industry associations or other accreditation bodies (ref)
- Companies that follow industry standards for information or process models, product classification or identification to be eligible for government business (ref)
- Legislative documents that conform to National or European Community standards for structure, naming, and description (ref)
- ICANN, the Internet Corporation for Assigned Names and Numbers, and its policies for operating the domain name system that make it possible for every web site to be located using its logical name (like "berkeley.edu" rather than using an IP address like 169.229.131.81). (ref)

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 or the W3C industry consortia to establish specifications or technical recommendations for their information systems or web services (REF)

[40] [LIS] Many organizing systems describe and arrange their physical or information resources in ad hoc ways because the person or institution determining the arrangement is completely autonomous. This is the domain of organizing systems embraced by David Weinberger in *Everything is Miscellaneous* (Weinberger, 2007).

[41] [Computing] (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.

[42] [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 aren't neutral, metrics influence results, (and) there's more than one way to describe something."

[43] [Citation] (Hu and Lui 2004).

[44] [Citation] http://www.amazon.com/gp/search-inside/sipshelp.html/

[45] [Citation] Park (2009)

[46] [Business] 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

[47] [Business} 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. 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).

[48] [Citation] Schmitz and Black 2008.

[49] [Citation] (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."

[50] [Citation] Panofsky 1939.

[51] [Citation] For CDWA, see Baca, M., & Harpring, P. (Eds.). (2005) and http://www.getty.edu/research/publications/electronic_publications/cdwa/. For CDWA-Lite, see http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.pdf

[52] [Computing] 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. The company's Multicolr search, for example, returns a set of stock photos with similar color combinations to the ones selected dynamically by the user. See http://www.ideeinc.com/

[53] [Citation] (von Ahn and Dabbish, 2008)

[54] [Citation] (Cano et al 2005).

[55][Citation] (Walker 2009)

[56] [Business] 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).

[57] [Business] One interesting service that uses sensors to create descriptions of location is the NextBus transportation tracking service in San Francisco, which tells transit riders exactly when vehicles will be arriving at particular stops. NextBus uses sensors to track the GPS coordinates of buses and trains, then compares that to route and information from transportation providers and estimates the time it will take for a vehicle to arrive at some selected location. To offer information that will be useful to riders, NextBus must figure out how to describe the location of a vehicle and the distance between that location and some intended target, as well as creating descriptions of transit routes (name, number, and/or direction of travel) and particular stops along the route. In some areas, NextBus incorporates multiple descriptors for a given stop by allowing users to search by route, by location, or by an ID number assigned to the stop.