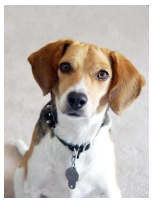
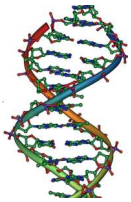
Chapter 2.
Symbol-based Representations and Descriptions

Representations are the foundation of information systems. There are many possible representations. One important distinction for representations is whether they are discrete (qualitative) or continuous (quantitative). Qualitative representations are easier to work with and they seem to work for the way people use categories and language. Because of their connection to logic, they are also called symbolic representations. These also include explicit relationships between the concepts. Symbol processing has been very useful, but is not the only approach. Non-symbolic processing focuses on the similarity as an alternative to categories. Entity classes vs. instances.





English:	Dog
Spanish:	Perro
Scientific:	Canis Lupus
ASCII:	104 157 147

Figure 2.1: Some representations and descriptions for "dog". Some are for a specific dog; others are for the class of dog. Some are symbol-based and some are not. (check permission)

Good representations capture important information in an effective way. Representations can provide information to users within an appropriate context; they can be copied and, in some cases decomposed and reassembled. In this chapter we focus on symbolic representations but there are also non-symbolic representations such as equations and distributed representations[8]. In addition, some representations now include behavioral elements and that allows many variations (3.9.3). In short, there are many alternatives to the symbol-based model and many reasons to criticize it, but it is so widely used that we need to start with it.

2.1. Categories and Classes

2.1.1. Categories and Classes are Representational Frameworks

When we interact with the world we encounter individual objects. But, those objects fall into groups. Some of the groups are ad hoc clusters say, all the objects which are on a desk. If the clustering seems important or if there is a similarity among the objects. we put them together in a category.

Figure 2.2: Grocery stores often use ad hoc categories for organizing their shelves. (check permission)

Natural lines of fracture versus artificial constructs. Classes and classification systems as a social artifact. classification of organisms ((sec:biologicalclassification)), of diseases (9.9.2), and of business (8.12.0). Categories are often based on ad hoc similarity but we are often interested sets of entities which fit a pre-defined system. Categories and classes usually involve similarity based on several attributes. Classifiers. Feature extraction is the process of determining which features to focus on when doing categorization or classification.

Categories are probably the simplest type of representation. Categories and classes make life easier, people do not have to judge individual situations separately. They can instead, categorize the situation and follow the rules which apply to it. Suppose you are organizing your kitchen. You would probably try to put similar things together: the spices on one shelf, the canned soups on another, and so on. Eventually, the categories help to simplify the complexity of the natural world. Rather than remembering or communicating every detail about a complex situation, the categories provide sufficient detail to allow a person to develop reasonable expectations about that situation. Categorization is the first step in knowledge representation. To create a database, for instance, we must categorize to what entity class each entity belongs (3.9.1). Later, we will consider related topics such as categories in human information processing. Classification is the process of assigning objects to classes. Classes are formalized than categories and are often based on consensus from members of a group.

2.1.2. Categories and Classes as Defined by Attributes: Aristotelian Categories

The simplest type of categories, "Aristotelian categories," are determined solely by attributes or characteristics inherent to the items to be included in the group. These "defining attributes", those attributes that define whether or not an item can be included in an Aristotelian category, must be universal for the entire category. That is, all the members of an Aristotelian category must share all of the defining attributes that make up that category. This leads us to distinguish between attributes that are required for category membership, i.e., defining attributes, and attributes which, though often associated with a category, are not required for membership in that category. These are called "characteristic attributes".

Where do the attributes come from? Scientific knowledge is often thought of as identifying attributes and processes and Aristotle is regarded as one of the founders of scientific reasoning (9.2.0). In some cases, classes are based on underlying processes such as evolution being the basis of biological classifications (9.8.1). Formally, Aristotelian categories are defined as a conjunction of attributes. Such attributes should be able to be combined and they could be used for logical inference.

While a category system may be very useful for one community or for one application, it may leave out aspects which are crucial for other applications. Not every object fits neatly into a category; sometimes there has to be a forced fit; such categorization is biased by the available choices for representations.

Classes extend categories by applying a conceptual framework. They are "top-down". A classification could be based on counties of the world. Classification should be differentiated from categorization or clustering which are purely data dependent. Typically, classes are based on a formal classification system while categories are based just on ad hoc similarity[16].

2.1.3. Other Approaches to Categories

Figure 2.3: Plato (left) and Aristotle (right) shown in a detail from *The School of Athens* by Raphael. Plato is pointing upward to signify his belief in prototypes (Platonic Ideals) whereas Aristotle gestures to the ground to indicate his emphasis on empirical attributes. (check permission)

While models based on Aristotelian categories dominate many information-system applications such as databases, many other models have been proposed for categories although these are not often employed in information systems. These also move away from simple models of symbol processing. Categories as used by people don't always seem to follow the Aristotelian approach. We will discuss the implications of

this more when we consider human cognition (4.3.0). Is a whale a fish? Although whales are mammals based on attributes such as feeding milk to their young many people think of them as fish. People don't seem to use purely attribute-defined categories; rather, they seem to interact with entities as "prototypes". A prototype is an idealized form. This is Plato's approach and unlike Aristotle's approach in which an object is either entirely in or out of a category, there is a degree of similarity or typicality in category membership. That is that some attributes are more typical than others. The distinction has implications across many areas of information systems. Similarity rather than attribute-based. Generally, Aristotelian categories have been very successful in natural science and are the basis of much of out thinking about laws. However, in addition, to the alternative Plato presents, there are several concerns about the nature of Aristotelian categories (Fig. 2.4). Statistical analyses and categories. Not linearly separable. Several of these other approaches can be modeled with non-symbolic methods such as neural networks. The role of protoypes in categorization and language processing remains widely debated [Lakoff-WFDT].

Label	Description	Example
Continuous	Some attributes do not have distinct boundaries.	An example is colors. Even seemingly distinct attributes may be continuous (Fig. 2.5).
Abstract	Some categories we cannot define with specific attributes.	Beauty. Many social categories.
Functional	Defined by function rather than by attributes.	Is a tree branch a chair (Fig. 2.6)? Are all tree branches chairs?
Radial	Radial categories are extended from a central example or prototype (Fig. 2.7). These are the result of analogy and metaphor.	
Family Resemblance	Some categories do not seem well defined by a single set of attributes[38]. These are thought to show similarity like the resemblance among members of a family so these are termed "family resemblance" categories. No one attribute is always associated with the categories. That is, these are a disjunction of conjunctions.	The definition of games (Fig. 2.8).

Figure 2.4: A variety of other issues for categories.

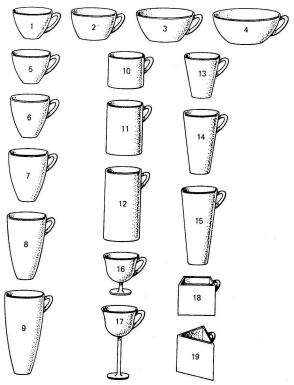


Figure 2.5: At what point does a cup become a glass, a goblet, or a jug? (check permission)(redraw)

2.1.4. Semantic Relationships among Classes

Classes can be part of a larger set of inter-related concepts. there are other concepts and relationships among them. Some common types of relationships can be identified. Indeed, relationships are so important that many of their attributes can be described. From very general to very specific. Related concepts versus named relationships. Binary, n-ary. Relationship among composite objects. [**?**]. From

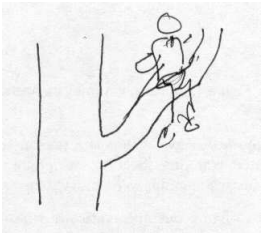


Figure 2.6: Is a tree branch a “chair”? A category may be defined by its function.

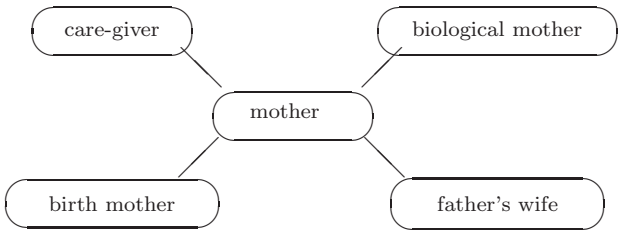


Figure 2.7: Sets of radial categories have a central theme and related concepts, but the related concepts are not differentiated by simple attributes.

game	**chess**	**soccer**	**card solitaire**	**Farmville**
teams		x		
physical space and activity		x		
competitive	x	x		

Figure 2.8: No single set of attributes seems to define a ”game”. Rather, there are subsets of attributes which games possess. (not finished)

semantic relationships to semantic networks. Recent activity in identifying semantic relationships with FrameNet (6.2.3).

Grouping allows complex objects to be understood and organized more easily by reducing their complexity. Another way to simplify the complexity of the natural world is through grouping. We have already seen, hierarchy and aggregation are illustrated in Fig. 2.9. Hierarchies show “is-a” relationships while aggregations show “has-a” or “part-of” relationships. Aggregation groups together objects that are part of a broader conceptual unit. Another type of relationship among objects is an ordering.

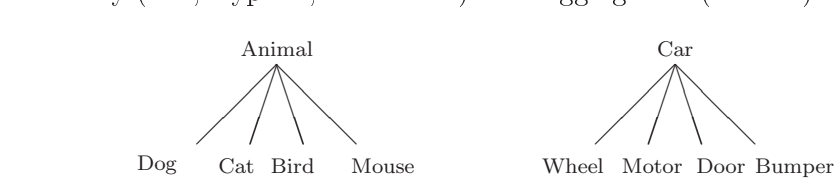


Figure 2.9: Two types of grouping relationships: hierarchy and aggregation.

Inheritance

In hierarchical relationships attributes may be carried, or inherited, from more general classes to more specific ones. An animal is the “parent” of a bird and a bird is the “parent” of a canary. Inheritance

is an efficient way to store information because characteristics (such as laying eggs) do not need to be stored with every instance, but only with the parents. By continuing with this logic, we might get even more specific and refer to a particular canary. By doing so, we would move from types (of birds) to tokens (specific examples). This is also similar to networks of concepts ((sec:conceptualnetwork)).

Partonomies

Several ways in being part-of. Parts within levels. System analysis.

Semantic Network

Semantic relationships explicitly describe the inter-relatedness of concepts (Fig. 6.17).

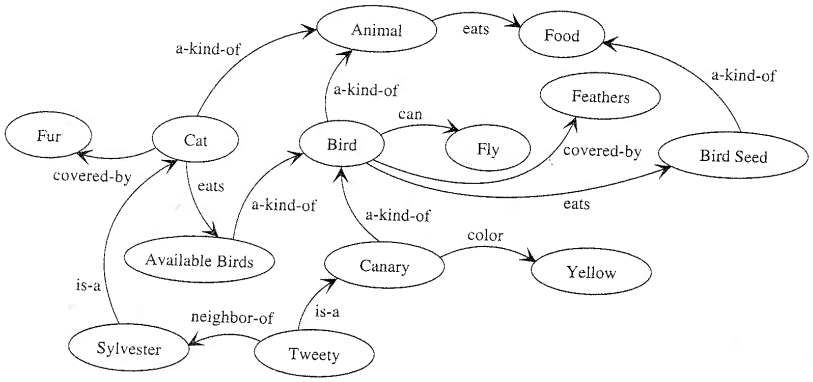


Figure 2.10: One type of semantic network identifies words and the relationships between them. This is also similar to conceptual models which we will discuss later. (redraw)

2.2. Knowledge Organization Systems and Knowledge Representation

Formal systems have been developed many of these approaches to description. Sets of categories can form descriptions of complex areas. Systems of semantic relationships. Specifically, this refers to sets of categories and classes are useful for describing things. Descriptions often reflect representations but they should also facilitate access [37]. They need to be tailored to the needs of the people who will be using them. There are many types of descriptions and we consider them at many places in this book. Some descriptions, those we consider in this section, are simply a few words. Descriptions would also include metadata (2.4.0) and abstracts (2.5.5). Descriptions of entire resources versus the contents with semantic annotation. Different ways of describing things. Epistemology. Frameworks for describing knowledge. Systematic classes. Most often this would be part of specifically selected set of terms. There are several ways these models can be structured. Here we consider three approaches: Taxonomies, Thesauri, and Ontologies; Many nuances are not able to be expressed and there can be drift of meaning across time [29]. Classification systems as boundary objects. Knowledge organizing systems can be applied information resources. In the previous section, we looked at the basic units of information: entities and attributes but useful descriptions require interrelated sets of attributes. These are Knowledge Organization Systems (KOSs).

These techniques are examples of knowledge representation. Linked data is also a type of knowledge representation. Domain models, conceptual models, and user models all applications of knowledge representation. The term knowledge representation is also associated with inference systems (-A.7.0) based on those representations but there can be considerable value to systems of description without considering inference.

2.2.1. Objects, Things, Entities, Instances, and Names

Description is one of the great challenges of information. As we shall see, there are many approaches. We start with the basic units to be described. Data models (3.9.2). FRAD to match and extend FRBR.

Instances have specific values for each attribute. The attribute and its associated value are known as attribute-value pairs. Identity. Attributes (2.1.2). Names identify specific entities. Naming implies a degree of acknowledgment and recognition. The properties of a name depends on how it will be used. Some names, such as "Bob," are informal. This name is useful in some contexts, but it would not be helpful in other contexts (at a convention of people named Bob, for instance). In more formal situations, we want to manage a system of names. To be most useful, a name should be distinctive and persistent (i.e., has it persisted through time). Some physical objects and categories, such as people and places, have proper names. These, however, are often neither distinctive nor permanent. Consider the number of towns in the United States that are named "Springfield". Enough information should be included to make a name a distinct identifier. A related, problem is that many variations of common names may be used. The name of the painter we usually know as Rembrandt appears on paintings with many variations. Concepts (1.1.4, 4.4.1).

The terms applied for common objects given by ordinary users vary widely [9]. These can often be names. Names should be unique, at least in a given context. Social implications of naming.

2.2.2. Knowledge Structures and Knowledgebases

Concepts do not exist in isolation. Rather, than describing separate descriptions, we need sets of related descriptions. Classification policies. Classification model. Description logic. Conceptual frameworks. These are basic models for networks of concepts. It's also worth noting that these descriptive system reflect social efforts and help to define the world for members of the social groups. Sets of classes must be drawn to adequately cover a field. Beyond classes to processes (3.9.3, 8.11.2).

Knowledge structures. Ordered and unordered lists. Schematics. Useful in schematics. Two of the most important knowledge structures are taxonomies and frames. We consider taxonomies below and frames in (4.4.1)(-A.7.1). Knowledge organization systems (2.2.0). Decisions about classification systems for information organizations Indeed, there are more subtle issues in knowledge structures such as inheritance.

Hierarchical Classification and Taxonomies

Grouping relationships can be stacked one on another to form a hierarchical classification. Such hierarchical classification is particularly easy to understand and navigate. An obvious example is library classification system which we discuss in the next chapter (2.5.1). Most classification systems are hierarchical. Indeed, the system of biological classification is so strictly hierarchical that we say it is a taxonomy (Fig. 2.11).

Taxonomies are composite knowledge organizing structures which demonstrate inheritance of attributes. For instance, we know one of the defining characteristics of animals is that they breath so every instance of an animal should have that property. However, inheritance relationships are not always so simple. While it is true that almost all birds can fly, there are exceptions. Penguins and ostriches are birds that cannot fly. A special attribute would be needed to mark such exceptions. such as a subclass of birds that cannot fly, such as penguins and ostriches, be developed? Sometimes, entities may inherit properties from more than one parent (2.5.2). A taxonomy should have a purpose. For instance, it should predict functionalities.

However, it is also worth noting the scientific taxonomies have come to be organized more by evolutionary heritage than by inheritance of visible attributes (9.8.1).

Kingdom: Animal
Phylum: Chordate
Class: Mammal
Order: Carnivsore
Family: Canide
Genus: Canus
Species: familiarus

Figure 2.11: The zoological taxonomy for dogs.

Thesauri

A thesaurus is a descriptive vocabulary about a specific domain. Terms thesaurus terms are developed which describe aspects of the domain and in some cases, there may be loosely specified relationships among the terms. For instance, there many be NTs (Narrower Terms [children]) and BTs (Broader Terms [parents]); define a hierarchical relationship. Typically, a thesaurus also includes RTs (Related Terms). The familiar *Roget's Thesaurus* lists words which are similar to the given word (6.2.1). Also SN, UF. Later, we will implement thesauri with the XML-based SKOS package (2.3.3).

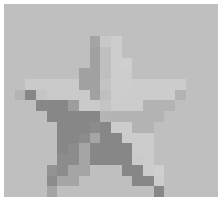


Figure 2.12: Example thesaurus.

Thesauri may also provide a conceptual structure for a domain. Thesauri may facilitate text searches by providing a standard controlled vocabulary (2.5.3) for the concepts in that domain (Fig. 2.13). Not all concepts can be identified. The appropriate concepts can be selected by examining the questions people use. This is another example of identifying orthogonal, hierarchical concepts and then composing them into more complex objects. Thesauri are used in text retrieval for query expansion (10.7.2, -A.6.4). There can be multiple controlled vocabularies.

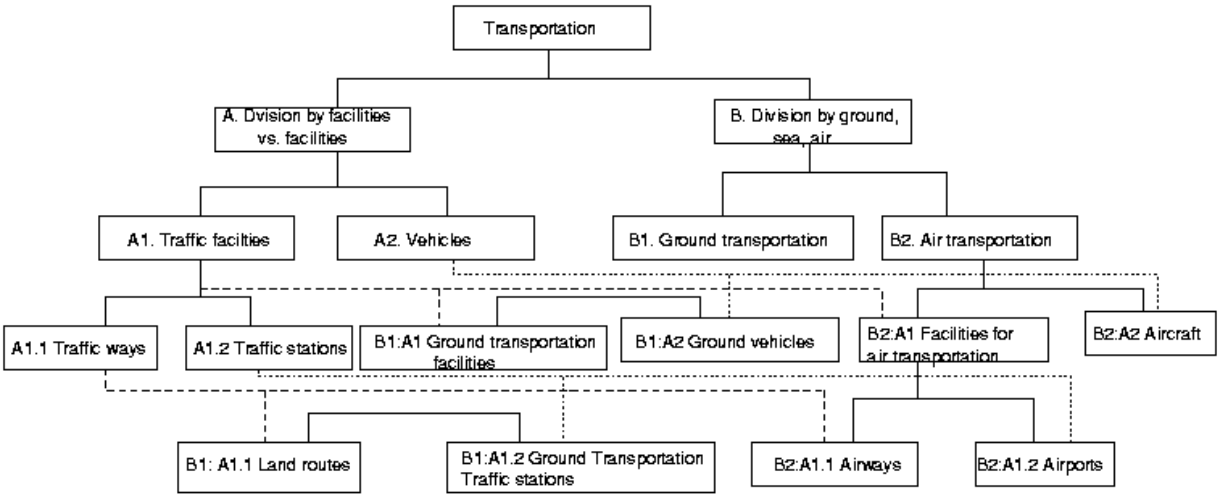


Figure 2.13: A concept hierarchy for aspects of transportation can generate thesaurus terms[26]. Some of the resulting concepts can be composed to form complex concepts. "A1.2 B2 Airports" combines "A1.2 Traffic Stations" and "B2 Air Stations". (check permission)

Formal Ontologies

There are several senses of the term "ontology". While the term ontology is often used loosely to include all types of knowledge organizing systems, the formal definition ontologies extend the semantic network shown in Fig. **??**. Specifically, ontologies provide the content for predicate logic (-A.7.1), which is the deconstruction of natural language to its actionable elements, thus formalizing and codifying its meaning. Linked to other sets of concepts. Ontologies are discussed further when we introduce XML and RDF-related tools(2.3.3). Merging and mapping ontologies.

Predicates and Knowledge Representation Languages

Taxonomies and thesauri have relatively simple relationships among entities but we also need to consider a broader range of relationships among entities. Semantic relationships (2.1.4). Natural Language (6.1.0). Predicates. Statements (Fig 2.14).

Figure 2.14: more complex structures require a predicate. The triangle is next to the circle. (not finished)

Make statements and inferences about the objects being described. KOS elements can be combined. Languages (6.5.2). Description logic. LOOM classifier logic.

Automated inferences from the knowledge on the web. Such named relationships can be useful for logic; indeed, the KR often results in a "knowledge base" which represents the world by the combination of the facts in it and the inference mechanisms which operate on those facts. Several KR languages have been developed. Some of them may be used with natural language processing systems (6.2.3), the Semantic Web, expert systems and logical inference (2.2.2, -A.7.0).

Inference with Knowledge Representation

Inference has always proven difficult. Symbolic representation and logic. Brittleness.

Inference based on knowledge representation. Ontologies with predicate calculus.

The Semantic Web and Semantic Technologies

The Semantic Web has pushed semantic technologies into new domains. Most importantly, the Semantic Web expect that such descriptions are machine processable. For instance, in supporting interactive systems for interacting with corpora. The annotations provide an indication of similarity. The broader goal of using the Semantic Web for inference is largely unrealized. This is certainly not a formal ontology or even thesaurus in the usual sense because it includes complex concepts. Particularly, used for technical fields with large data sets (9.6.0).

The Semantic Web also addresses many of these issues and it is often used in applications beyond those normally considered by traditional information specialists. Moreover, the Semantic Web emphasizes making the tags machine readable. On the other hand, sometimes the lessons of the traditional approaches are lost in the study of the semantic web. However, the very strengths of controlled vocabularies also suggest limitations. The Semantic Web has brought many advantages of automatic processing and management of terminology. However, that automated processing has allowed great inconsistencies to come in.

Linked data. Beyond linked data to linked processes and events.

Importantly, the semantic web focuses on automatically processed statements. This allows automated evaluations of the vocabulary system. For instance, it can check integrity constraints. It also allows manipulation of basic values such as conversion of units.

Like thesauri, ontologies, are task or domain specific. This is because for a given domain or task, the terms are usually relatively unambiguous. Event ontology. However, coordination across domains can be difficult as are attempts to develop ontologies for general applications, because the terminology can be ambiguous. Furthermore, unlike people for which language is highly fluid, ontologies do not adapt to context or new situations; thus, we say they are brittle. Coordinating disjoint ontologies. This is less of a problem for thesauri since they do not try to be as exact. Indeed, this many also represent the social uses of language and concepts[24]. Furthermore, there may be a combinatoric explosion[2]!

There might be multiple vocabulary systems. Integrated vocabulary modeling[10]. Vocabulary ecosystem. Ontology server. Concept bank. Vocabulary registry and repository. Vocabulary provenance.

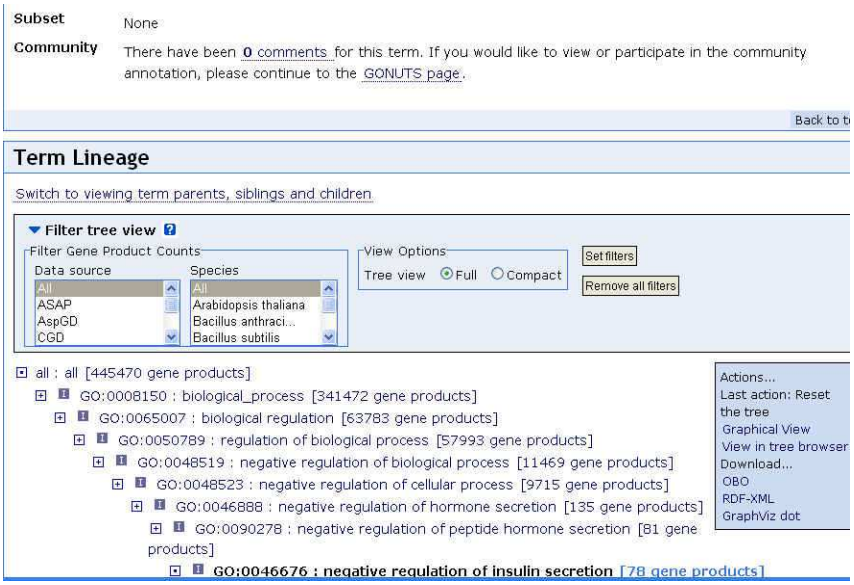


Figure 2.15: GeneWiki[5] uses an "ontology" for describing genes. (crop)

Lexical resources coded with RDF. DBPedia.

2.2.3. Process Models for Description

2.3. Information Resources

2.3.1. Documents

We make complex statements about the world and collect those into documents. Documents are a important construct for in the study of information. We will consider them in two perspectives – as structured information resources and as resources with a social purpose.

There are many ways in which information is captured such as pictures, blocks of text, Web pages, databases, mashups, video, simulations, and software. Some of these such as most pictures or blocks of text are characterized simply as information objects. Such information objects are often combined into more complex objects. In many cases, the composite information objects are documents.

Figure 2.16: A variety of document types: a) A passport has the information necessary for crossing international borders. b) A journal article is structured typically focuses on presenting new information. c)

There are rules, data models, for the ways in which these objects can be combined. Because documents

are structured and have distinct components, they can often be tagged with XML and presented electronically. While traditional documents remain static, when presented electronically they can be interactive. Indeed, several pages can be linked together to form hypertexts, which allow richer models of interactivity. Document communities (5.8.2). In a broad sense, documents help to structure society.

At one level, documents are simply structured presentations of information which have permanence. We are issued documents at birth, another at death, and countless ones in the time between. They are very common and highly varied. Examples include passports, books, drivers licenses, newspapers, course listings and technical reports. Documents as a conceptual unit (9.0.0). Genres. Wikis and blogs can be considered as genres for the Web.

When multiple copies of an information resource are made especially when they are made for distribution, it may be helpful to distinguish the original from its copies. By comparison to documents, works are intellectual or artistic creations. There is also a a close connection between works and collections; works are the basic units of a collection is a work.

2.3.2. "Social Life of Documents"

Documents are more typically created to accomplish a certain task or to suit a given function. Increasingly, documents go through many versions and there are intermediate types of content such as coordinated fragments of documents forming mashups. Where a document typically emphasizes the utility of a document for transmitting structured information across organizational boundaries. Thus, we call them boundary objects. When taken out of context, some documents may be difficult to understand; consider emails. Some materials may be designed specifically to be unambiguous and to easily cross boundaries. These "boundary objects" can be understood outside of a narrow context. With its photograph, official-looking seals and concise information, a passport is understood and accepted in many countries. Indeed, such boundary objects allow the transfer of processes across separate systems so that they can become sub-systems of a combined system. Limitations of the effectiveness of boundary objects. However, it is also important to note that documents may end up being used in ways far beyond the intentions of the author[30].

2.3.3. XML: eXtensible Markup Language

Typically, documents are highly structured. That structure can be encoded with XML which is the eXtensible Markup Language has become Here, we approach XML as it is applied to documents. Later, we will see that XML is useful to describing domains can be encoded with XML and it is also useful as a database interchange tool.

Structuring Documents with XML

It helps to think of the content of a document as separate from its layout. A business letter, for instance, has distinct components such as a return address, date, greeting, body, and signature. However, the content of those specific components, whose name, which date, what address, will vary from letter to letter. Useful services can be developed by tagging specific components of the document's structure without considering how or in what order they will be displayed. The presentation can be controlled separately. Tagged content can be useful in developing indexes or a table of contents, or text marked as section headers can be displayed in a style different from the rest of document.

XML separates the components of a document from their layout. Fig. 2.17 shows an example of an XML-tagged document. XML is basically hierarchical: that is, it defines the broadest aspect of an object first, followed by the second-broadest and so on down to the most specific aspect and XML creates document structures like Fig. 2.18. It is a common language (i.e., an interchange standard) for Web-based artifacts. We will discuss several of the applications of XML in later sections.

XML Pattern Documents XML tags provide a type of semantic annotation An XMLSchema provides a simple framework for defining the structure of the document tree. One of the uses for XML schemas is to define the structure of documents (Fig. 2.19) after tagging the components they contain. The notation

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="poem.xsl"?>
<POEM
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi=noNamespaceSchemaLocation="poem.xsd" >
  <TITLE>Sonnet #49</TITLE>
  <AUTHOR> W. Shakespeare </AUTHOR>
  <STANZA>
     <LINE>Against that time, if ever that time come,</LINE>
     <LINE>When I shall see thee frown on my defects,</LINE>
     <LINE>Whenas thy love hath cast his utmost sum,</LINE>
     <LINE>Called to that audit by advised respects;</LINE>
     <LINE>Against that time when thou shall strangely pass</LINE>
     <LINE>And scarcely greet me with that sun thine eye,</LINE>
     <LINE>When love, converted from the thing it was,</LINE>
     <LINE>Shall reasons find of settled gravity:</LINE>
     <LINE>Against that time do I ensconce me here</LINE>
     <LINE>Within the knowledge of mine own desert,</LINE>
     <LINE>And this my hand against myself uprear</LINE>
     <LINE>To guard the lawful reasons on thy part.</LINE>
     <LINE>To leave poor me thou hast the strength of laws,</LINE>
     <LINE>Since why to love I can allege no cause.</LINE>
  </STANZA>
</POEM>
```

Figure 2.17: Poem tagged with XML tags as defined by the XML Schema.

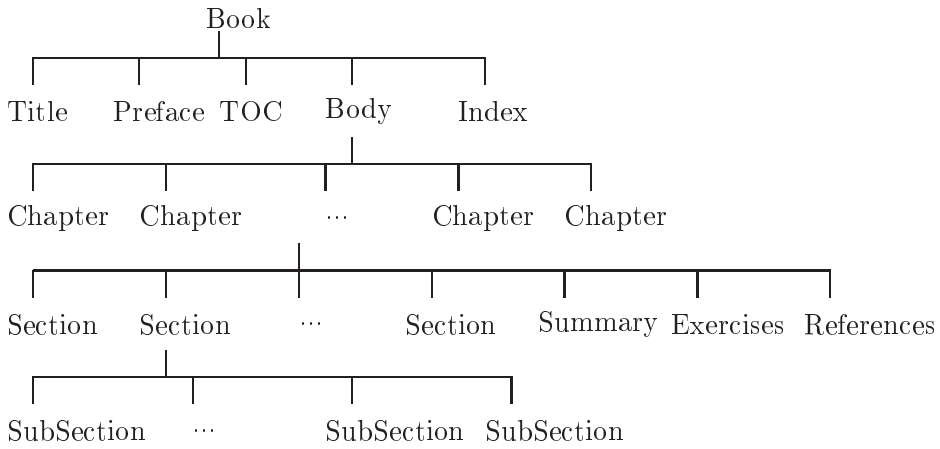


Figure 2.18: A traditional hierarchical document tree applied to the structure of this textbook. While this is an easy structure to understand and browse, it ignores the cross-links between sections such as references to other material

says that a poem has a TITLE, one or more AUTHORS, and one or more STANZAs. STANZAs are made up of one or more LINEs. XML documents need to make sure they conform to the DTD. Developing a standard of elements facilitates the interoperability of documents. DTDs implement hierarchical structures like that in Fig. 2.18. In most cases, individual users do not create their own DTDs but apply pre-established ones. Indeed, many publishers provide standard DTDs for their content to ensure consistency.

```
<?xml version="1.0"?>
<xs:schema
 xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="POEM">
  <xs:complexType>
   <xs:sequence>
     <xs:element ref="TITLE" minOccrs="1" maxOccurs="1"/>
     <xs:element ref="AUTHOR" minOccurs="1" />
     <xs:element ref="STANZA" minOccurs="0" />
   </xs:sequence>
 </xs:complexType>
 </xs:element>

 <xs:element name="STANZA">
  <xs:complexType>
   <xs:sequence>
     <xs:element ref="LINE" type="xs:string" minOccurs="0" />
   </xs:sequence>
 </xs:complexType>
 </xs:element>

 <xs:element name="AUTHOR" type="xs:string"/>
 <xs:element name="LINE" type="xs:string"/>
 <xs:element name="TITLE" type="xs:string"/>
</xs:schema>
```

Figure 2.19: XML Schema which defines the tags used in the POEM document.

Specification of Document Layouts: XSL and XSLT Information should be presented in a way that is most convenient and logical for the recipient. Now that we have defined the components of a document, we can turn to the presentation of a document's content. While the physical layout of a document generally reflects its logical structure, several different physical structures are possible. Fig. 2.20 shows two layouts of a business letter. The two panels reflect two different but equally accepted styles for positioning the return address and signature line. A business letter may have its return address in either the top left or the top right corners.

Because the physical layout should be separate from the logical structure, a special language is needed to describe layouts. The XML Style Language (XSL) was created for this purpose. This language is used to determine the presentation style of an XML document. Sets of XSL specifications are often collected into style sheets. XSLT is the XSL Transformation Language; it allows XML to generate other, multiple formats. An XSLT element can display the title of a document in HTML, and documents can be converted to an electronic-book format or even submitted to a database. Fig. 2.21 illustrates that an XSLT script can generate a variety of formats from an XML file. This is a type of dissemination service. Later we will consider synthesis and publishing of entire publications (8.13.4).

The Resource Description Framework (RDF)
Many packages are built on top of XML. One of the more important of these is RDF, the Resource Description Framework. Fig 2.23. RDF provides a way to associate metadata with digital resources. RDF allows a standard approach to the creation of defining relationships among resources. But this extra capability is not always needed and some services can be implemented in either XML or RDF. XML has many applications beyond documents and information objects. It is particularly helpful for describing semantic relationships. Its applications are shown in the so called "layer cake" diagram in Fig. 2.24.

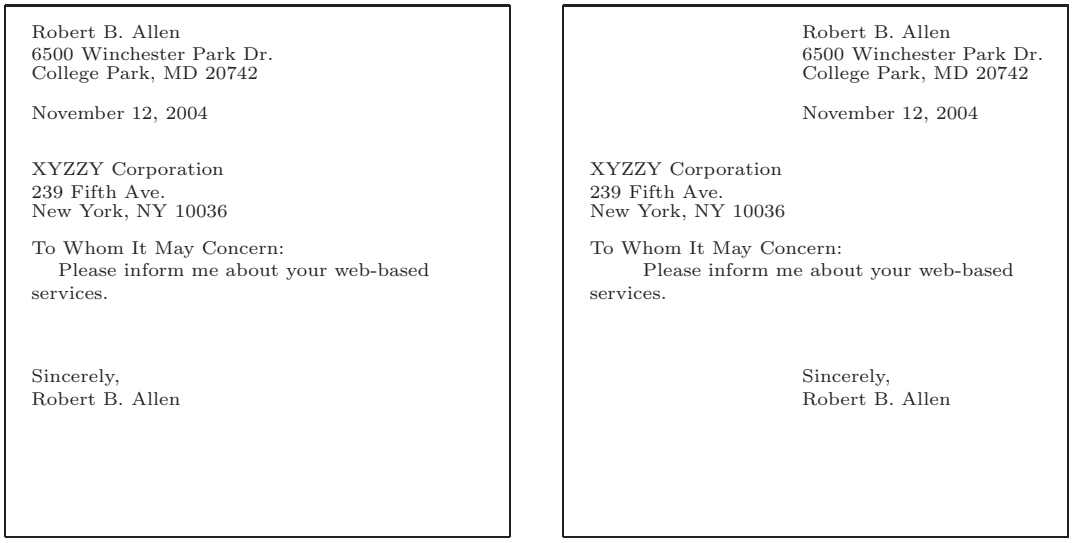


Figure 2.20: Two common layouts for a business letter; the content is identical, but the formatting differs. If the letters are coded with XML, the layouts can be generated with different XSLT scripts.

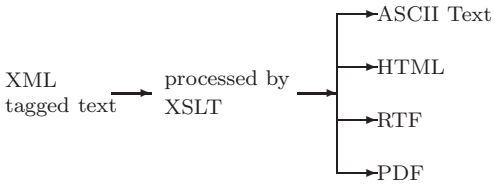


Figure 2.21: XSLT also allows a single tagged XML file to be converted to several different display formats such as ASCII, HTML, RTF, and PDF.

```
<xsl:template match="POEM">
    <HTML>
      <xsl:apply-templates>
    <HTML>
</xsl:template>

<xsl:template match="TITLE">
  <H1> <FONT COLOR="Green">
    <xsl:value-of/>
  </FONT> </H1>
</xsl:template>
```

Figure 2.22: Part of an XSLT description for a poem and the title (as used in Fig. 2.17) which generates HTML. The stanzas are composed of additional templates as indicated by "xsl:apply-templates" and the title is a literal string as indicated by "xsl:value-of".

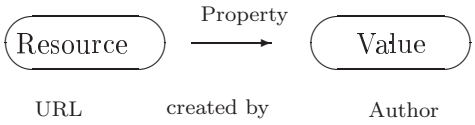


Figure 2.23: RDF associates metadata with a resource. Specifically, it has triples composed of: Resource, Property, Value.

High-Level XML Packages: SKOS and OWL

Several frameworks have been developed for knowledge representation but it is natural to use an approach with is consistent with XML. For instance, for ontologies (2.2.2), Taxonomies and thesauri can be described in RDF with the Simple Knowledge Organization System (SKOS) (Fig. 2.25). OWL, the Web Ontology Language, does that. Specifically, OWL implements a description logic, that is a formal method for creating descriptions. OWL is built on RDF Schema (RDFS) [32] which extends RDF. OWL allows the creation of Classes such as "Mother" or "Father". Furthermore, OWL allows the specification of types of properties such as functional properties (Fig. 2.26). Using OWL for conceptual descriptions.

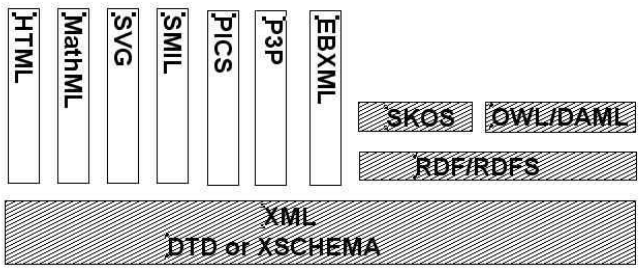


Figure 2.24: This "layer-cake" diagram shows that XML is a unified framework that provides structure and descriptions for many Web-based objects (adapted from[34]). The specific components shown and described in the text.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#">

  <skos:Concept rdf:about="http://www.my.com/#dog">
    <skos:prefLabel>dog</skos:prefLabel>
    <skos:altLabel>canine</skos:altLabel>
  </skos:Concept>
</rdf:RDF>
```

Figure 2.25: This example uses two XML packages: RDF and SKOS to define a concept (dog) and two labels ("dog" and "canine") associated with it.

```
<owl:Class rdf:ID="Operetta">
  <rdfs:subClassOf rdf:resource="#MusicalWork"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasLibrettist" />
      <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">
      </owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Class>
      <owl:complementOf rdf:resource="#Opera"/>
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
```

Figure 2.26: An example of an OWL statement. This defines an Operetta as a Musical Work which much have Librettist and which is a complement of an Opera[33]. (check permission).

2.4. Data Schemas and Metadata

Data schemas are structured descriptions of objects and metadata are structured descriptions of information resources. The author of a book is an attribute of that book, and can therefore be a piece of metadata. We do not need to be too strict about the distinction between data and metadata; the important point is that metadata describe and supports the primary information contained in a system or collection. In some cases, the distinction between data and metadata is blurred. For instance, descriptions about people or about locations.

2.4.1. Data Schemas

A schema is a template for an entity with a selected set of attributes. Schema.org

Micro-data. Frames (4.4.1).

Broad range of items to classify. Metadata for non-traditional objects. Comic books. T-shirts.

Criteria for a good classification system. Metadata for data repositories.

Inheritance hierarchy (2.1.4). For example:

Thing > Person,
FOAF,

Thing > Creative Work > Book

Descriptions of scientific results. Description of geography. Descriptions of museum objects (7.6.1).

2.4.2. Metadata

When information resources are being described, we describe the attributes as metadata. These systems have been particularly weel worked out. Reasons for metadata: find, identify, select, obtain, explore. There are several types, levels, and applications of metadata. Describing content and then repurposing it for different platforms such as mobile, smart TV. Semantic publishing (**??**).

Library metadata, archival metadata (7.5.4)design and process metadata ((sec:designmetadata)).

Figure 2.27: The meaning of a picture is different from the elements that appear in the picture. This picture illustrates the metaphor that the "broom" of woman's suffrage will "sweep clean" prostitution, gambling, and drunkenness[18]. This illustrates the difference in describing the "ofness" and the "aboutness".

More generally, different types of metadata have value at different stages of the lifecycle of the information resource. Some description systems are based on the content of the information the system contains, while others describe attributes of the resource itself, such as the creator or the date. Metadata description is a representation. It is a description of information resources. Thus it is a secondary representation. Semantic annotation. Descriptions of scientific data sets. Tagging versus annotation.

Information resources and metadata associated with that. Constrained sets of attributes have been

developed to guide the content of any given description. We will first focus on descriptive systems for information resources and then turn to more general description frameworks. We have emphasized the importance of representations. Let us consider document representations; they should be discriminative, descriptive, complete, and correct. Metadata are attribute values used to describe information resources[14]. Metadata can be described as data about data. The set of metadata used to describe an entity is an information model.

Metadata supports services and user needs. Physical objects can also be described by metadata; museum artifacts, for example, need descriptors (7.5.4). Metadata is clustered into groups (Fig. 2.28). When we want to describe a collection of documents so we need a flexible set of terms. Knowledge organizing systems (2.2.0). Developing metadata descriptions in the context of a complex collection of objects is more difficult than describing individual objects. Simple metadata that is consistent across users and collections facilitates access for a variety of users. Any system of metadata should cover the scope of a field and should be coordinated across domains. Furthermore, descriptive systems need to serve a community.

2.4.3. Library-Oriented Bibliographic Metadata

Information resources have attributes in common which typically fit well together. Here, we focus on library resources which are often books with attributes such the publisher and the date of publication.

Standards for provide consistency across environments. However, the standards have become increasingly varied and complex. The MARC (Machine Readable Cataloging) record is a library standard for organizing bibliographic records. Metadata composites for complex objects (7.8.0).

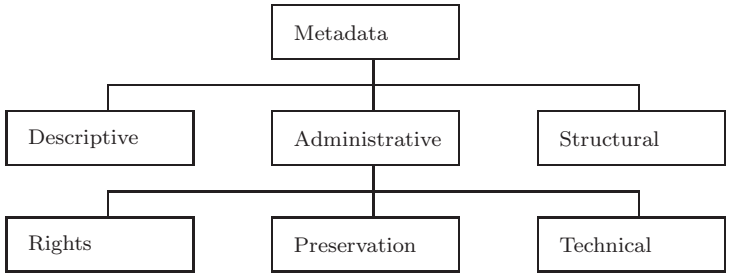


Figure 2.28: One way of characterizing types of bibliographic metadata[17].

Bibliographic records are standard descriptions while presenting pertinent information about collections of information resources. Bibliographic theory.

Bibliographic Works and Records

One of the distinctive features of published materials is that there are many closely related copies of closely related material. When we describe such material, at some points we want to describe the original work which is being reproduced and at other times we may want to describe individual copies.

Traditional publications produce multiple nearly identical copies. Metadata may be organized by a data model (Fig. 2.29) (3.9.1). Functional requirements (7.9.1). As indicated on the right side of the figure, different types of metadata are associated with each level of the hierarchy. FRBR: item level, collection level[6]. The original version of a creative work (2.4.3) is distinct from all subsequent instances of that work For traditional texts, such as books and documents, the concept of a "work" is generally clear. On the Web, however, it is not always so clear. Sometimes, the individual page might be considered a work, and at other times, the entire Web site might be considered a "work". As we will see, defining the original work is an important part of organizing the metadata that pertains to it. A "derivative work" is not entirely original, but involves adding intellectual effort to an original work. A translation is a derivative of the work being translated. A superwork includes many related versions of work. Ability to include a broader range of materials in a catalog. Describe relationships among entities. Works also generally have social significance [**?**].

Beyond book, other collections of information resources have related lateyed structures.

Entities as the basis for the functional requirements (7.9.1). Bibliographic relationships help to create an entity-relationship model (3.9.1). Relationships among information resource include [?]: Equivalence, Derivative, Descriptive, Whole-part, Accompanying, Sequential. Derivative relationships can be subdivided into ...

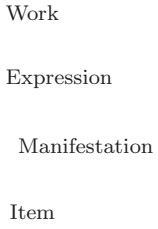


Figure 2.29: When many copies of an information object are made and especially when there are many versions of that information object, metadata can keep that straight. Some attributes belong to individual copies and others apply to the entire work. That is, some of the metadata values are inherited from the higher levels. Typical metadata attributes for formally published materials are shown in parentheses at each of the level.

Bibliographic Control and Authority Files

Consistency across the records in a catalog. An example of semantic tools. Bibliographic control ensure quality and consistency. Cataloging rules provide standard definitions and encourage consistency in catalog records[7]. One example is the "Rule of 3," which specifies that any author list that contains three or more names should be simplified by stating the first author's name followed by "et al.". If a database has a fields for first, middle, and last names, consider the difficulty of entering the following names: Madonna, George Herbert Walker Bush, Sitting Bull. For formal indexing, explicit policies should be created. Principles not just rules. Work languages, Document languages, Subject languages. Cross-cultural conceptions of authorship and classification [?].

Authority files provide standardized forms of entities. Specifically, name authority files provide standard spelling for a name (Fig. 2.30).

Catalogs

Nowadays these may be in digital repository (7.8.0). Typically, access points for collections are grouped along dimensions such as title, author, or subject. This are attributes which reflect common information access behavior of users. user needs or use cases. Applying successive levels of restrictions can be a way to specify a search (Fig. 2.31). Cooperative cataloging. Use cases (3.10.2) for content development.

Catalogs for collections present standardized metadata for the objects in that collection. ICP: Convenience of the user, Common usage, Representation, Accuracy, Sufficiency and necessity, Significance, Economy, Consistency and standardization, Integration. The metadata used in a catalog should be constructed to help users to find items in that collection. More discussion about metadata when we consider complex digital objects (??). Union catalog.

2.4.4. Dublin Core Metadata System and Schema.org/Book

Dublin Core was designed as a light-weight metadata system for describing Web pages and not necessarily for full works. However, it is so common that we will include it here. For the Web, the known in the Dublin Core. There are 15 elements of Dublin Core (Fig. 2.32), the metadata system that is often used for Web objects. As its name suggests, these 15 elements are intended as a core and that core can be extended to cover a wide range content types including visual resources and educational materials (5.11.6). Dublin Core attributes can also be "qualified" by sub-attributes. "dc.creator" can be qualified

Paul Rembran	Rembrandt Harmenszoon van Rijn	Rembrant Van Rin
Paul Rembrandt	Rembrandt Harmensz Van Rijn	Rembrardt
Rambrandt	Rembrandt Harmensz van Rijn	Rembrat
Rebranch	Rembrandt Harmensz. van Rijn or Rhijn	Rembrdandt
Reimbrant	Rembrandt Hermanszoon van Rijn	Remdrandt
Rem.	Rembrandt Hermansz van Rijn	Reymbram olandes
Rembrach'	Rembrandt Olandese	Rijmbrand
Rembradt	Rembrandt Van Rhyn	Rijn, Rembrandt Harmensz. van
Rembrand	Rembrandt van Rijn	Rijn, Rembrandt van
Rembrande	Rembrandt van Ryn	School of Rembrandt
Rembrands	Rembrant	Van Rhyn Rhembrandt
Rembrandt	Rembrants	Van Ryn, Paul Rembrandt
	Rembrant van Rhijn Rembrant van Rijn	

Figure 2.30: The painter Rembrandt and variations in the spelling of his name[15]. (check permission)

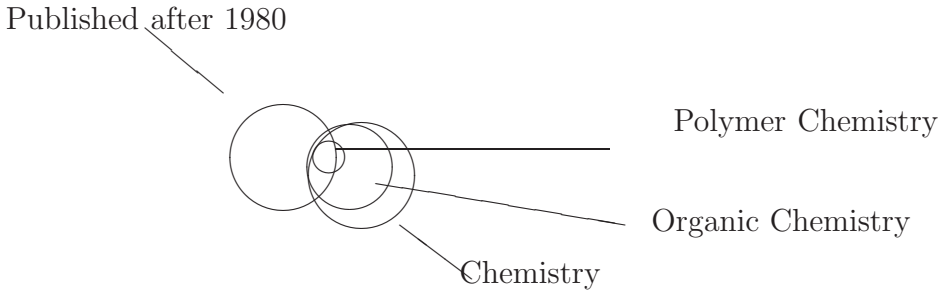


Figure 2.31: Levels of hierarchical metadata can be useful for controlling scope during retrieval. We could first search on topics relating to organic chemistry published after 1980 before moving on to the narrower search for research on polymer chemistry.

as "dc.creator.illustrator".

When tags from different metadata systems are included in a given document, it is necessary to be clear about what system they come from. This is defined by the "namespace" (xmlns) and the namespace package identifier is included with the tag. dq:creator is the creator tag as defined by the Dublin Core metadata system.

Linking Works with Metadata Attributes

RDF. Semantic graph.

Resource Description and Access (RDA) proposes rules for developing systematic metadata. Low-level attributes at the item level.

FRBR describes Entities. Creating catalogs.

Machine processable. Dublin Core abstract model. As the name suggests, RDF used to apply resource descriptions such as Dublin Core to documents. This is accomplished using an "about" clause that governs the relationship between the resources and attributes.

Element	Description	Example
Title	A name given to the resource.	Information: A Fundamental Construct
Creator	An entity primarily responsible for making the content of the resource.	Robert B. Allen
Subject	The topic of the content of the resource.	Information science and systems
Description	An account of the content of the resource.	A textbook.
Publisher	An entity responsible for making the resource available.	Robert B. Allen
Contributor	An entity responsible for making contributions to the content of the resource.	Robert B. Allen
Date	A date associated with an event in the life cycle of the resource.	1/1/07
Type	The nature or genre of the content of the resource.	textbook
Identifier	An unambiguous reference to the resource within a given context.	ISBN
Format	The physical or digital manifestation of the resource.	LaTeX
Source	A reference to a resource from which the present resource is derived.	Authored
Language	The language of the intellectual content of the resource.	English
Relation	A reference to a related resource.	PPTs
Coverage	The extent or scope of the content of the resource.	"Information Science, Information Systems, Web Science"
Rights	Information about rights held in and over the resource.	Robert B. Allen

Figure 2.32: The base set of Dublin Core metadata attributes[4]. Here, an example is filled in. Not every element is included in many semi-formal collections. (check permission)

```
<META NAME="DC.creator">
<META NAME="DC.creator.illustrator">
<META NAME="DC.subject" CONTENT="lcsh-heading" SCHEME= "LCSH">
<META NAME="DC.subject" CONTENT="mesh-heading" SCHEME= "MESH">
```

Figure 2.33: The base set of DC attributes can be qualified with subdivisions as creator.illustrator. Further attributes can be extended. For the subject tag CONTENT and SCHEME which describe the system used for the content description (LCSH and MESH are systems of subject descriptors).

Extended DC

Figure 2.34: Extended DC.

Going forward, such efforts will facilitate making resources more available from Web based search and, thus, will be able to satisfy more information needs and this has been a significant concern for academic librarians.

Metadata Application Profile

A metadata application profile specifies the range or applications to which a set of metadata is typically applied. It is related to the community interests which the collection is expected to serve. Dublin Core application profiles.
Singapore application profile framework. The MPEG standards body has defined MPEG-A as a framework for new MPEG applications. Functional requirements
Domain model
Description Set Profile

```
<?xml version='1.0'?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dc = "http://purl.org/dc/elements/1.0/"
     xmlns:dcq = "http://purl.org/dc/qualifiers/1.0/">
  <rdf:Description rdf:about = "http://doc">
    <dc:creator>
      <rdf:Description>
        <rdf:value> Pat Jones </rdf:value>
        <dcq:creatorType> Photographer </dcq:creatorType>
      </rdf:Description>
    </dc:creator>
  </rdf:Description>
</rdf:RDF>
```

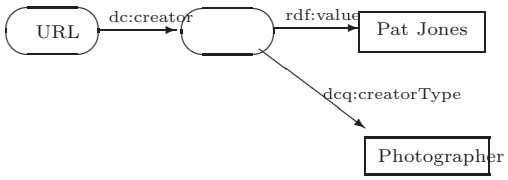


Figure 2.35: RDF can be applied with qualified and extended Dublin Core. The dc:creator attribute is qualified dcq:CreatorType with the value of "Photographer".

Usage guidelines
Encoding syntax guidelines

2.4.5. Documentary Languages

2.5. Subject Languages: Descriptions Based on Document Content

The metadata examined thus far has not focused on the content of the information resources but about attributes such as the year of publication and author's name. Tools which use such description include indexes, abstracts, and classification.

Applying knowledge organizing systems (2.2.0). We have already considered thesauri (2.2.2). SKOS (2.3.3). In additional to information resources, cultural objects such found in museums (7.6.1) and architecture.

Language help to define communities.

The "Semantic Web" is often associated with ontologies, but it frequently goes beyond these to cover all types of descriptions [21]. Beyond indexing to semantic annotations (7.8.4). This identify semantic units within the text. Alphabetic languages versus topic-oriented languages.

Description of other resources Data sets (9.6.0).

Sensory, perceptual, emotional dimensions. MPEG-7.

There are a variety of semantic technologies ranging from classification systems to controlled vocabularies to ontologies. Each of these has strengths and may usefully be applied in different situations.

2.5.1. Hierarchical Subject (Topic) Classification

Classification is used for many kinds of objects and information, such as videos in a video store, food in a grocery store, topics in a newsgroup, or items in online auctions. Classification systems are frequently used to organize books and other materials in libraries; you are probably familiar with the subject classification system used for books in your local library. Formal classification systems, such as those used in libraries, are often hierarchical (2.2.2). Classification systems: broad, close, design.

Library Classification Systems

Libraries (7.2.1) have been particularly active in developing large-scale classification systems. The largest and most widely used classification systems are simple hierarchies. It is likely that your library uses one of the two most common systems: the Dewey Decimal System or the Library of Congress Classification Systems (LCC). The Dewey Decimal Classification (DDC) system is used in most public libraries in the U.S. As the word "decimal" suggests, the DDC has no more than 10 items per level. The top-level categories for DDC are shown on the left side of Fig. 2.36. Books and other documents with numbers between 000 and 099 fall into the category called "Generalities". Although library classification systems are primarily hierarchical, faceting (2.5.3) is sometimes added to them. This crosses the main classification dimension with other dimensions. Mining might be subdivided by a category such as geographic region (e.g., mining in Asia, mining in North America, etc.). Classification systems may describe the same concept in rather different ways; we need a guide for how terms from the two systems are related. Such guides are called crosswalks.

Number	Description
000	Generalities
100	Philosophy and Related Subjects
200	Religion
300	Social Sciences
400	Language
500	Mathematics
600	Technology
700	The Arts
800	Literature and Rhetoric
900	General Geography and History

Figure 2.36: Top-level of Dewey Decimal Classification.

Primary Labels	Secondary Labels
Arts & Humanities	Literature, Photography...
Business & Economy	Companies, Finance, Jobs...
Computers & Internet	Internet, WWW, Software, Games...
Education	Universities, K-12, College Entrance...
Entertainment	Cool Links, Movies, Humor, Music...
Government	Military, Politics, Law, Taxes...
Health & Medicine	Diseases, Drugs, Fitness...
News¡ & Media	Full Coverage, Newspapers, TV...
Recreation & Sports	Sports, Travel, Autos, Outdoors...
Reference	Libraries, Dictionaries, Quotations...
Regional	Countries, Regions, US States...
Science	Biology, Astronomy, Engineering...
Social Science	Archaeology, Economics, Languages...
Society & Culture	People, Environment, Religion...

Figure 2.37: Top-level of Yahoo.com classification (as of January, 1999).

In addition to the DDC and LC, there are several other comprehensive library classification systems such as the UDC and Colon Classification.

Structure and Evolution of Subject Classification Systems

Decisions about library classification structures are often based on the notion of warrant. Semantic warrant, literary warrant.

A classification schedule from the 1950s would not have much about space travel; one from 1980 wouldn't mention HIV. While being dynamic enough to change as needed, a subject classification system should be static enough to be predictable for users. Although the top-level subject classification systems are

static, the Dewey Decimal Classification is revised frequently as new areas of knowledge emerge. A recent expansion included Eastern Religions, which had not been covered fully in the earlier editions. Fig. 2.38 shows the changes in a section of the classification system used in the rapidly changing the field of computer science from 1964 to 1998. Evolution of terminology is even more rapid in descriptions of popular music.

3.7 Information Retrieval
3.70 General
3.71 Content Analysis
3.72 Evaluation of Systems
3.73 File Maintenance
3.74 Searching
3.75 Vocabulary
3.79 Miscellaneous

H.3 INFORMATION STORAGE AND RETRIEVAL
H.3.0 General
H.3.1 Content Analysis and Indexing
H.3.2 Information Storage
H.3.3 Information Search and Retrieval
H.3.4 Systems and Software
H.3.5 Online Information Services
H.3.6 Library Automation
H.3.7 Digital Libraries
H.3.m Miscellaneous

Figure 2.38: Here is a classification developed for the rapidly developing field of Computer Science. Fragment of the ACM Classification in 1964 (left) and the corresponding section in 1998 (right). Note how much the classifications changed in the space of 34 years. Topics such as "online information services" did not appear at all in the earlier classification[1].

2.5.2. Poly-hierarchies, Multiple Inheritance, and Facets

One of the strengths of simple single hierarchies such as those used in traditional library classification systems is that the items are located in one and only one position. However, it may be difficult to find a single specific location in a hierarchy because an item seems to belong to several categories. Pneumonia is both an infectious disease and a lung disease. Sharing properties from several parent categories is known as "multiple inheritance," and the structures formed from multiple inheritance are called "polyhierarchies" (Fig. 2.39). Some classification systems attempt to avoid multiple inheritance because of the complications in overlapping attributes.

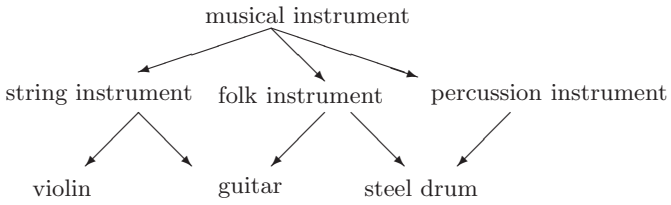


Figure 2.39: A guitar can be part of a polyhierarchy under string instrument and folk instrument.

Facets and Facet Classification

About facets. EBay.

Facets can be systematically developed with semantic factoring can create a faceted, controlled vocabulary by identifying orthogonal underlying terms. Many works and collections are better characterized by independent facets. These faceted systems have orthogonal dimensions. That is, they categorize their concepts with a series of seemingly unrelated concepts. With such a system, minerals for instance, could be considered according to the regions in which they are found. Ideally, each dimension would be independent of the others as shown in the example of a faceted thesaurus (Fig. 2.40).

Wikipedia topic structure as a DAG.

2.5.3. Index Terms and Indexing Languages

The term "index" is used in several ways. An index can be a data structure used by a document retrieval system, a pointer to topics in one document, or a catalog for access to information resources such and those in a document or collection. an index provides an organization of the literature of an

Facet Name Associated Hierarchies
Associated Concepts
Associated Concepts
Physical Attributes
Attributes and Properties
Conditions and Effects
Design Elements
Color
Styles and Periods
Styles and Periods
Agents
People
Organizations
Activities
Disciplines
Functions
Events
Physical Activities
Processes and Techniques

Facet Name Associated Hierarchies
Materials
Materials
Objects
Object groupings and systems
Object genres
Settlements and landscapes
Built complexes and districts
Single built works
Open spaces and site elements
Furnishings
Costume
Tools and equipment
Weapons and ammunition
Measuring devices
Containers
Sound devices
Recreational artifacts
Transportation vehicles
Visual works
Exchange media
Information forms

Figure 2.40: Top-level facets from the *Art and Architecture Thesaurus*[22]. Note that the facets are designed to be independent from each other.

entire field. An index may be measured by "exhaustivity," or the extent to which it covers all of the concepts included in a work and by its "specificity," that is, the level of detail, the depth, or ichness of the indexing. Indexing functionality.

Subject Categories and Controlled Vocabularies

Topic descriptions versus other attributes. Which attributes to select and include in a set of metadata. Systems of metadata (2.4.3). It is useful to have a standard set of descriptive terms as a controlled vocabulary. Although there are differences among concepts, in a controlled vocabulary, these distinctions may be helpful. This process of selecting optimal terms is similar to the process of defining entities. We need to extract terms for a set of documents that are pre-defined as referring to that set. Fig. 2.42 shows the stages for such a systematic development of a thesaurus. Another basis for a developing a controlled vocabulary is by examining the words people use to ask questions. Coordinating with lingustic tools such as FrameNet (6.2.3).

abode, address, apartment, asylum, bungalow, cabin, castle, cave, commorancy, condo, condominium, cottage, crash pad, diggings, digs, domicile, dormitory, dump, dwelling, farm, fireside, flat, habitation, hangout, haunt, hearth, hideout, home plate, homestead, hospital, house, hut, igloo, illahie, joint, living quarters, manor, mansion, nest, orphanage, pad, palace, parking place, place, residence, resort, roof, rooming house, roost, shanty, shelter, trailer, turf, villa.

Figure 2.41: Terms that may be used to describe a "home" (adapted from Roget). While the variants have slightly different senses, for indexing it is usually clearer to use just one standard term.

Many concepts are combinations of other concepts. The concept of "doctor" or "nurse" combines the concepts of "person" and "medical treatment". Each concept is independent, i.e., orthogonal, from the others. This process of identifying the underlying dimensions is known as "semantic factoring". Recall

	Examples	
	Original Terms	**Final Term**
1. Combine related terms	Aesthetics and Esthetics	Aesthetics
2. Combine related concepts	Aesthetics and Production Values	Production Values

Figure 2.42: Steps in vocabulary reduction for creating controlled vocabulary word lists.

that semantics is the study of meaning in language. The concept of "hospital" could be decomposed into "building" and "medical treatment".

Tools for managing large-scale collections of vocabularies.

Subject descriptors are standard terms that cover the major topics in a collection. They are usually not hierarchical and are properly an example of "enumeration" rather than classification. The Library of Congress Subject Headings (LCSH) are the most widely used set of subject descriptors Several subject descriptors may be combined for a specific document, several subject headings may be used (Fig. 2.43). An index may include concepts which do not actually appear in the document.

France–History–Revolution, 1789-1799–Songs and music
Motion pictures–Law and legislation–Japan

Figure 2.43: Library of Congress Subject Headings may be combined into composite descriptions. The second example above would be for a document about laws concerning motion pictures in Japan. The order of the terms identifies which concepts are most important with respect to the object which is being indexed. (new example)

Subject Analysis and Facet Analysis

In order to classify it, we need to determine what a book or document is about. Indeed, classification systems such as the Dewey Decimal System identify single positions in the hierarchies. A subject classification system requires identifying what a work is about. "Subject analysis" determines the subject of a work and assigns it to a subject classification system. It would be nice to assume that a work has only a single subject, but resources are often complex and contain many attributes, making it difficult to assign only one subject category. There may simply not be a single topic, and viewpoint classification may be ambiguous from the user's viewpoint. Finding the book on a given topic via text processing. What would people want to use this book for? Epistemological potential [**?**]. [13]. In some approaches, facets may be combined to create complex statements about the topic of a book.

2.5.4. Creating Metadata and Metadata Systems

Developing a consistent large-scale metadata system is very difficult. Authority implies care and attention to details.

Communities of practice define metadata systems appropriate to their needs.

Good metadata supports interoperability. Metadata comes from many sources. in other cases, it is the result of systematic effort by professionals. Indeed, there are formal organizations for considering metadata standards. In other cases, metadata is loosely defined. The amount of effort invested in creating metadata depends on the importance of the collection and the needs of the users. Some metadata are harder to define than others.

It is surprisingly difficult to generate accurate metadata. There are three problems in doing this: the feature may not be known, there may be true ambiguity about the feature, or the metadata may be assigned carelessly. "Content guidelines" facilitate consistency of the metadata but care may be needed to assign even with such guidelines (Fig. 2.44). Using controlled vocabularies Validation lists for checking the actual terms entered.

Costs of systematic metadata development. There is a chance of systematic attacks of organization of information. Automatic capture of metadata at creation.

> If in doubt about what constitutes the title, repeat the Title element and include the variants in second and subsequent Title iterations. If the item is in HTML, view the source document and make sure that the title identified in the title header is also included as a meta title (unless the DC metadata element is to be embedded in the document itself).

Figure 2.44: Content guidelines for the Title Element in the Dublin Core[35].

Cooperative cataloging for sharing metadata records which are used in library catalogs. Cost-benefit for developing metadata.

Open metadata.

Socially Constructed Metadata

Traditionally, the metadata for formal collections hase been carefully constructed by professionals. Another approach, is to let the users create the metadata. Social indexing. The sets of metadata generated in this way is known as folksonomies. This is certainly much cheaper and more flexible, but it has other implications. These may be reflect cultural biases or of intentionality, persuasion and bias.

Need for consistency in metadata. Groundswell of popular trends and emergent metadata. Limitations of folksonomies[20].

The Web is a highly dynamic environment. Separate taxonomies could be developed quickly for separate interest groups. Ad hoc taxonomies. This can be helpful when systematic descriptions are not possible. The Open Directory Project (DMOZ)[3]. Social tagging and finding objects: del.icio.us. Comparison of social tagging to policies for traditional classification[23]. The danger is that social tags may reflect a popularity context rather than systematic classification. Another approach for generating metadata is "Games with a purpose"[31] (Fig. 2.45). Games (11.7.0). Semantic relationships (6.2.3). Game-oriented crowdsourcing.

Figure 2.45: "Games with a Purpose" generate descriptors in which web-mediated participants try to match descriptive terms. (check permission)

Workflow models.

Coordinating Across Systems of Metadata

Linked data.

2.5.5. Making Resources and Collections Usable

Content Coordination

Techniques for supporting interaction with content. Interface tools for interacting with information resource content. This internal structure can be captured with Coordination Widgets. Across re-usable content objects [?] Information architecture (1.1.3) and semantic publishing. Books (8.13.6). Annotations of several sorts. Reader annotations.

Tables of contents support access to it the components of a work such as its chapters. Structure often cannot be separated from meaningful presentations. Table of figures. Table of (legal) cases TOC for video.

Knowledge organization widgets in encyclopedias.

Back-of-the-Book Indexes

As suggested earlier, the term index is used in several ways. In general use, an index is most often a back-of-the-book index. Subject indexes do not simply select keywords from the text. Problem of indexing mentions. The phrase: "But John Major was no Winston Churchill..." should not be indexed under 'Churchill'.

Indexes across collections of books. Metadex.

User-centered indexing. Adaptive hypertexts for personalized indexes. Task-oriented abstracts.

Meta-dex.

Catalogs

Snippets and Surrogates A document surrogate stands in place of a document. It might be a thumbnail image of a document but it is most often a bundle of metadata which follows the information model for that type of documents. When documents are arranged in collections, surrogates may be organized into a catalog.

Web page summaries often include snippets.

Abstracts

Descriptions beyond metadata. Abstracts can help users maintain "current awareness" of work in a field as new documents are written. Like other information resources, abstracts should serve information needs. Abstracts may be characterized by the type of description they provide (Fig. **??**) [**?**]. This is especially important for scholarly literature (9.1.1). An "informative abstract" attempts to convey as much of the information of the larger document as possible. An "indicative abstract" simply indicates the topics which are covered. It is most often used for material which is difficult to summarize such as the contents of a database. An evaluative abstract critiques the ideas and gives an indication of what is contained in the article without necessarily describing the contents. Abstracts should cover the major points in the work they refer. Some abstracts are structured; that is, they may discuss specific issues based on the structure of the original document. An abstract of a scientific publication might require descriptions of the hypotheses, procedures, results, and conclusion sections.

Some abstracts are structured so readers can focus on the essential aspects of the research[12]. This has become especially common for medical applications. One example of a structured abstract style sheet requires that the following categories be included: Background, Purpose, Research Design, Setting, Study Sample, Intervention, Control or Comparison Condition, Data Collection and Analysis, Findings, Conclusions, Citation.

2.6. Hypertext and The Web

Linking supports browsing. Stand-alone documents are effective for many applications, but a wider range of user needs can be supported with linking those documents to others. Hypertexts are sets of information objects that are linked together. Many types of services can be developed to support interaction in these hypertexts. Links in hypertext serve multiple functions. They provide a navigational path but they also provide signals of association between concepts. In a real sense, knowledge is stored in the network of links. Links are similar to semantic relationships (6.2.3). Hypertext structures provide a types of information organization which support browsing. Hypertext as a literary genre (6.3.7).

2.6.1. Links and Anchors

The simplest links connect two documents. We have briefly seen Xlinks. A more complex type of link, embedded or contextual links, connect regions within documents. Fig. 2.46 shows familiar HTML HREF links and anchors which as embedded links. The end points of a link are known as "anchors". Anchors can be single points within a document, sections of a document, or temporal locations for

scenes in a video or other multimedia objects. For HTML documents, the location of anchors may mean linking to a whole document or only to a section within a document.

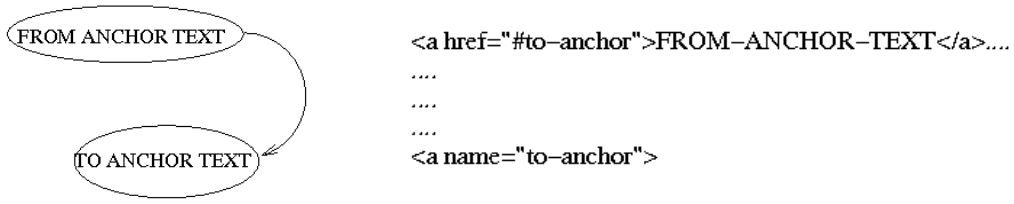


Figure 2.46: Anchors are end points of embedded links. This is illustrated with HREFs in HTML. (redraw)(check permission)

The collection of all the links in a hypertext forms the structure of that hypertext. "Referential integrity" checks whether the links are complete; that is, whether or not each link (reference) contained in a hypertext is composed of an object that it is linking from, as well as an object that it is linking to.

Links in a hypertext can have attributes. An electronic book might have a special type of link for providing definitions of words. When hypertext systems have typed-links, the link types are often drawn from a predefined set. XLink, the link framework for XML (Fig. 2.47) shows the specification for an XLink. The links can be defined to have attributes; that is, they can take on "roles" or functions, such as the simple "dictionary-definition" function that links a word to its definition in a dictionary program. Beyond simple HREF's there are many variations of linking. The links may be multidimensional (a single link may connect to several other sites) (10.4.3) or links may be adaptive (they may be displayed for only some users or situations). Link roles may be compared to semantic relationships (2.1.4). Multi-headed links and OHS.

```
<!ELEMENT student ANY>
<!ATTLIST student
  xmlns:xlink CDATA #FIXED "http://www.w3.org/1999/xlink/namespace/"
  xlink:type CDATA #FIXED "simple"
  xlink:href CDATA #REQUIRED
  xlink:role CDATA #IMPLIED
  xlink:title CDATA #IMPLIED
  xlink:show () "replace"
  xlink:actuate () "onRequest" >

<students xlink:href="studentList.xml">
  The list of students.
</students>
```

Figure 2.47: An XLink definition and an example of its use. The "student" tag has an argument which is the HREF of a file called "studentList.xml".

2.6.2. Composite Hypertext Structures

HTML implements a simple model for linking notes in a hypertext. Other types of hypertexts can introduce additional structure. Several of these are summarized in Fig. 2.48. Formally, hypertexts may even be specified with data models (3.9.0). Basic hypertext is easily modeled as a graph (-A.3.0). Composites [11] are higher-level objects, such as indexes and tables of content. Composites can also introduce their own navigational structures. Instead of a link simply navigating the user to a new document a link in a composite might bring up a schematic on a split-screen to allow comparison with the content of the composite. These hypertext composites help users to contextualize knowledge. Visual information, especially as seen in visualization has similarities to hypertext (11.2.5).

Implicit structure versus full visualization of the structure. The "language of selection" [19]. Formal

Type	Description (Section)
Table of contents	Structure of links (2.5.5)
Guided tours	A predetermined chain of related pages. (2.6.2)
Templates	Links mapped to regions in a graphical structure. (2.6.2)
Spatial hypertexts	Implicit links based on proximity. (9.10.0)
Hypertext maps	Overview of link structure. (2.6.2)
Argumentation systems	Typed links that describe the components of an "argument". (6.3.5)

Figure 2.48: Composite hypertext and related structures.

Hypertext Models Open hypertext models. Mappings between different hypertext models.

Menus allow the selection of options from a set of brief descriptions. Menus can be used to explore documents that are organized hierarchically (2.1.2). One common example of a menu is a table of contents. A menu with more breadth contains more choices per page, but fewer pages. (Fig. 2.49). A menu with more depth contains fewer choices per page, but more pages. Users are generally able to find items in menus with high breadth faster than in menus with high depth, as it requires fewer clicks to reach a given point. In addition, user satisfaction often decreases as the number of required clicks increases. However, a menu with a greater depth often allows for a more logical, sequential progression of choices, decreasing the possibility of user confusion. There is a tradeoff between depth and breadth in the efficacy of menu organization, and it may be found that certain menu styles are more suited to particular tasks than others.

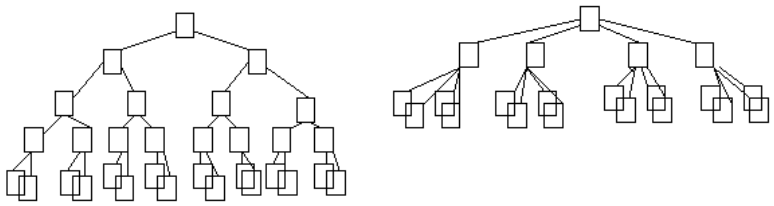


Figure 2.49: Two structures of menus that allow the user to reach 16 nodes. The one with high depth (left) has more layers but fewer choices at each layer. The other, with high breadth (right), has fewer layers but more options at each layer.

Several of these structures are the basis for coordination widgets (2.5.5).

A guided tour follows a predetermined path through a collection of information resources; it can be considered a type of composite hypertext. The simplest guided tour has a single path, which is presented straight through from beginning to end. Other guided tours allow you to "choose your own adventure," and are more branched and complicated. Examples of guided tour composites include lectures, novels, broadcast television news programs, and movies.

Hypertext Maps, Templates, and Spatial Hypertexts

Interfaces for interacting with arguments. Graphical views of arguments.

Argumentation vs inference. The structure of arguments is captured in argumentation systems. As their name implies, argumentation systems are often used for describing group discussions. Fig. 2.50 shows a tagged fragment of the discussion about rebuilding the Reichstag in Berlin. Fig. 2.51 shows an argumentation system that helps students to develop scientific explanations collaboratively by illustrating the connections between seemingly disparate facts. Group argumentation systems are used for education (2.6.2).

Hypertext maps provide an overview of several nodes. Some hypertexts are composed of templates that reflect specific knowledge structures related to the tasks. These may be schematics. Fig. 2.52 shows a workspace filled with templates representing information about individual countries. Spatial

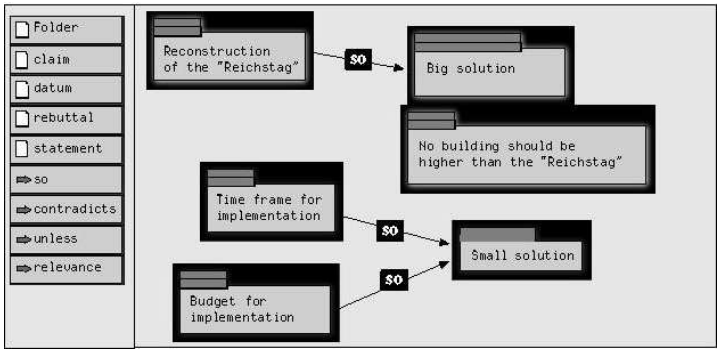


Figure 2.50: An argumentation system is a hypertext map (adapted from[27]) which lays out aspects of an argument. Note the objects types (folder, claim, datum, rebuttal, statement) and the link types (so, contradicts, unless, reference). (redraw) (check permission)

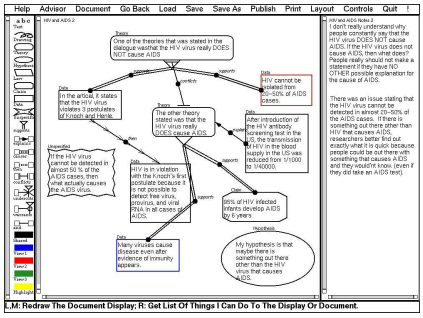


Figure 2.51: An argumentation system can support student learning about scientific reasoning[28]. (check permission)

layout organizes the templates; thus, these sets of templates form a spatial hypertext in which the user is guided by the structure rather than by explicit links.

Structure and interactivity are introduced to hypertext maps these become interactive schematics and visualization systems (11.2.5).

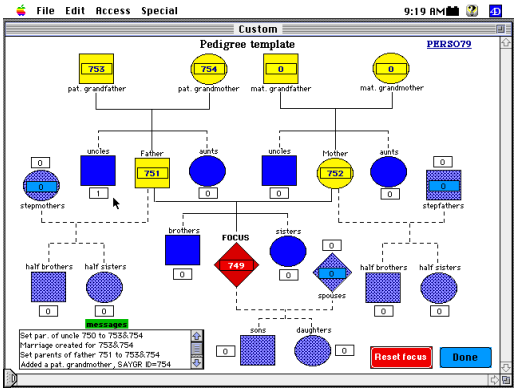


Figure 2.52: A schematic can provide a visual structure for facilitating page-based browsing[25].

Adaptive Hypertexts

Adaptive hypertexts support reconfiguration of the nodes and links based on user characteristics and history. Prioritizing links on a page based on user preferences. Effectively, this becomes a model of the

user's knowledge. Personalization (4.10.2). These can be useful in teaching and are related to adaptive tutoring systems (5.11.3).

2.6.3. The Web as a Common-Use Hypertext

The Web is more than a simple collection of documents in a hypertext or library. It provides many kinds of information ranging from recipes to reservations to digital libraries. Thus, the Web is known as a common-use hypertext. The Web does not have a simple unified architecture, but XML is being expanded to provide a unified framework.

Web-Page and Web-Site Design

Information design and information architecture. Visual languages (11.2.4). Information architecture (1.1.3). The goal of layout is to allow the user to identify and easily access the content of a Web site. Web sites have many applications, some focused on specific users and some broadly based for the public. To build an effective Web site, we need to decide how, and by whom, it will be used. We then need to provide access points for meeting the information needs of the user group. The interface in Fig. 2.53 allows users to search for movies by title, by actors, and by locations. The content of the Web site should be highlighted in the interface and clues or instructions given to users about navigation.

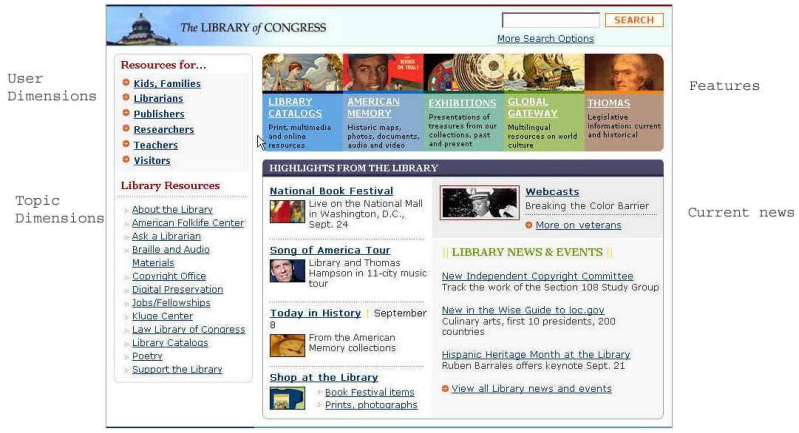


Figure 2.53: Access dimensions for a browser display should reflect the underlying content.

Just as library catalogs have different dimensions for access, web pages should be designed with consideration of the types of material users will want to access. This is similar to the specification of use cases in software applications (3.10.2). Whatever the design chosen, it should remain consistent across the entire site, and there should also be no dead-end links. A well designed site will highlight its core information, while at the same time providing diversions and subordinate information in easily accessible links. Interaction design (4.8.1). General principles for design of applications beyond the Web will be considered later (4.8.0).

Layout for disjoint information objects. The layout of a newspaper — how the stories or sections are organized on a single page or throughout the issue — contributes to a reader's ability to both find articles of interest and to understand the relationship of various news items. Layout, in the news or other media, is often used (or manipulated) to aid reading or to make associations for viewers; an effective layout is one that highlights a recurring theme. The theme of newspapers is generally one of importance: information that is deemed to be important is given a special place — the front page — while news that is considered less important is moved toward the back. Visuals are used in a way that contributes to the advancement of the overall theme and creates a synergy between text and images. The photographs of a newspaper typically support the information that the news articles contain; in other media, such as comic strips or satires, the text may contradict the image to create irony. A layout need not be simply visual, but may include audio or even tactile presentations, the latter existing

mostly in the world of art. The dynamic elements of interactivity make layout and design decisions more complex: interactive electronic documents are now designed for a specific user's preferences and actions, rather than to an entire group. Interactivity leads us from documents to hypertexts, which we shall consider in the next section. Interaction design (4.8.1).

Design and patterns.

Discourse relationships can help structure layout (6.3.2) to support comprehension (10.2.3) . Document analysis (10.1.5).

Link Semantics Creating a link adds meaning. It suggests that there is a significant relationship between two documents. Links can be an indication of similarity (10.10.2). In some hypertext models different types of links perform different actions. Some links, such as a "Submit" button, commit the user to action. Other links, such as a back button or a chapter heading, simply navigate to a new location. However, all links, as the operable elements of a hypertext, share a common purpose: to support information access and task completion by users, and not just provide a formal model. Following a link has two effects on a user: it shifts the attention to a new topic while at the same time retaining the context of the previous page.

A link should be easily distinguishable from the text in which it occurs. This is often accomplished with different-colored font or underlining. In addition, because interactive documents and hypertexts allow users to jump to information that is of particular interest to them, a link should provide clues to the user about where it leads. Fig. 2.54 shows an example of "link visualization". This is one of many general user interface principles (4.8.0).

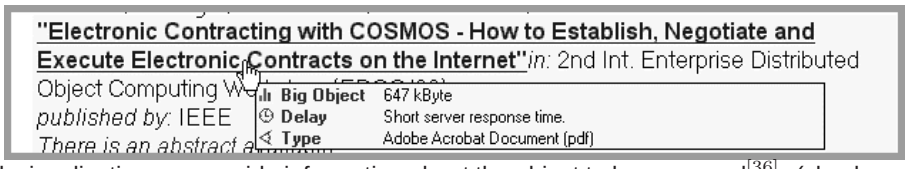


Figure 2.54: Link visualization can provide information about the object to be accessed[36]. (check permission).

Hypertext provides an alternative to traditional linear documents. It allows a great deal of flexibility in allowing users to browse through a set of inter-related concepts. Thus, there is a usability tradeoff in the flexibility provided by hypertext rather than the simple linear order of traditional documents.

Emergent Structure of Information Networks

The Web is the result of many people and organizations independently designing sites and posting material of interest to those sites. The Web is an information network. Nonetheless, it is not entirely chaotic; patterns emerge. We can count Web objects such as pages, servers, and links; we can count how frequently these objects change; and we can record user interaction with the Web. The resulting patterns allow us to identify different elements of the World Wide Web. It is helpful to characterize the Web as a graph (-A.3.0). Specifically, the web is a small-world graph. Social networks (5.1.0). Characterizing aggregate structure of the Web (Fig. 2.55). Because the Web is so large, we can look at the number of in-links and out links across a large number of nodes.

It provides links between information resources. The Web is the most obvious example but there are many others. For instance, in traditional scientific research articles the citations form links. Two notable types of sites are "authorities" and "hubs" (Fig. 2.56). "Authorities" are linked to by many other pages; that is, they have a lot of inward links. Moreover, the greater the number of different pages linking to an "authority" is an indication of that page's quality. "Hubs" are the opposite of authorities. They link to many other pages. The quality of a hub may be measured by the quality of the authorities to which it points. This insight is the basis for the PageRank algorithm, which is used to rank documents following a Web search (10.10.2, -A.3.5).

Exercises

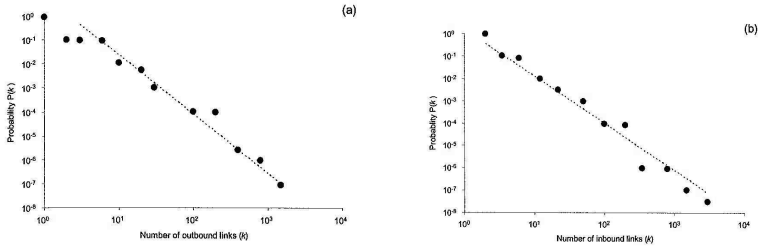


Figure 2.55: Graph of frequency versus number of in-links and out-links for Web pages. These are log-log plots so the data shows a power law. (check permission)

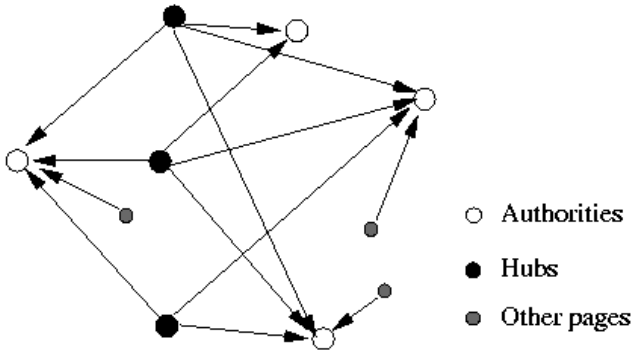


Figure 2.56: We can treat the web as a large complex system. Schematic of the structure of the Web. Authorities (open circles) have many other pages pointing to them. Hubs (black circles) point to many other pages. (redraw)

Short Definitions:

Abstract (document)
Abstraction
Access point (collection)
Aggregation (document)
Attributes
Attribute-value pair
Authority file
Data model
Cataloging
Classification
Collection
Common-use hypertext
Content guideline
Controlled vocabulary
Data dictionary
Database
Derivative work
Document
Document Type Definition(DTD)
Dublin Core
Entity (databases)
Epistemology
Folksonomy
Facet (classification)
Guided tour
Inheritance (KR)
Information Model
Knowledgebase
Menu
Metadata
Multiple inheritance
Namespace
Ontology
Procedural knowledge
Prototypes
Query language
Representational bias
Resource Description Framework (RDF)
Schema (data)
Semantic factoring
Surrogate
Symbolic representation
Taxonomy
Thesaurus
Typed-link
XLINK
XML
XSLT
XMLSchema
Work (metadata)

Review Questions:

1. List some defining and characteristic attributes for an automobile. (2.1.2)
2. Describe the relative advantages of "classification" and "key word" systems. (2.1.2)
3. Give additional examples of the grouping relationships we described. (2.1.4)

4. What are some of the difficulties in a single, simple hierarchical topic classification system. (2.1.2)
5. Identify the elements of this chapter that should be included in a DTD. (2.3.3)
6. Compare DTDs and XMLSchemas for describing the structure of documents. (2.3.3)
7. Explain the difference between logical structure and presentation structure for documents. (2.3.3)
8. What are some different ways a person could be a "creator" of an information object. (2.4.4)
9. Compare the process of identifying entities for a database and selecting a controlled vocabulary. (2.5.3)
10. Compare the structure of the a folksonomy subject classification system with the structure of formal library classification systems such as the LC or Dewey Decimal Systems. (2.5.1)
11. What are some of the advantages and disadvantages of a controlled vocabulary for a given topic? (2.5.3)
12. What are the relative advantages of informative and indicative abstracts? (2.5.5)
13. Explain what is meant by a "composite hypertext". Give an example. (2.6.2)
14. List several elements of effective Web site design. (2.6.3)
15. Give some examples of Web sites that are "hubs" and other sites that are "authorities". (2.6.3)

Short-Essays and Hand-Worked Problems:

1. What are some of the advantages and difficulties in the standard ("Aristotelian") approach to categorization. (2.1.1)
2. Explain how you would identify the category of "airport". Is an aircraft carrier an airport? (2.1.1)
3. Can you identify any truly unambiguous categories? (2.1.1)
4. What are some examples of prototypes as a model of categorization? (2.1.3)
5. Describe the pros and cons of classification into a single hierarchy versus facets. (2.1.2, 2.5.3)
6. Consider the objects around you as you read this. Briefly describe those objects and propose a classification system for them. (2.1.2)
7. Consider the books you own. Make a subject classification system for organizing them. What are the difficulties? (2.1.2)
8. Critique the effectiveness of the library subject classification system used in your university library or in your town's public library. Pick a work from the shelf and explain how it might have classified in a different location. (2.1.2)
9. Give an example of a classification system you have used that is confusing or ambiguous. How could that be improved? (2.1.2)
10. What are the advantages and disadvantages of using subject classification systems as a primary information access technique? (2.1.2)
11. Ask two friends to develop subject classification systems for the same topic independently from each other. For instance, they might make a classification system for games. Compare the results. (2.1.2)
12. Hierarchies are widely used as a navigation structure for hypertext. Describe why it is useful and what are some of the difficulties in using it. (2.1.4)
13. Pick a section of the Dewey Decimal System and attempt to explain why classification may have been selected. (2.1.2)
14. What makes an effective classification system? (2.1.2)
15. Will search engines replace the need for metadata? (2.1.2, 10.7.4)
16. Develop a system for categorizing the food stored in your kitchen (or your parent's kitchen). (2.2.0)
17. Explain the distinction between "types" and "tokens". (2.2.1)
18. Should subjective metadata reflect the creator's view of the material or the user's likely view of that information?(2.2.0)
19. Select a small domain about which you are very familiar and build an ontology of the concepts for it. (2.2.2)
20. Explain how you might create a thesaurus of (a) your personal photographs and (b) Web objects. (2.2.2)
21. Choose a topic and build a thesaurus for it. The terms should show complete coverage of the area without being redundant. Hint: Use a systematic strategy such as that illustrated in Fig. 2.42. (2.2.2)
22. How is a thesaurus different from an ontology? (2.2.2)
23. Some knowledge representation projects have attempted to map all knowledge. What are some of the difficulties of doing this? (2.2.2)
24. What is a "fact"? (2.2.2)
25. Why are people inconsistent about assigning names? (2.2.1, 2.2.2, 6.2.3)
26. Contrast the definition of documents. (2.3.1)
27. Create a DTD for this chapter of the text. Entities should include: chapter, sections, subsections, exercises, notes, readings, and references. (2.3.3)
28. Explain the difference between DTD and XSLT files. (2.3.3)
29. Create Dublin Core metadata for your course home page. (2.4.0)
30. What is the appropriate metadata for an electronic thesis or dissertation? (2.4.0)

31. What is the relationship between "North by Northwest" and "Der unsichtbare Dritte". (2.4.0)
32. What techniques could you use to ensure the consistency of metadata? (2.4.0)
33. Describe a system of metadata for describing a collection of cartoons. (2.4.3)
34. What is the main advantage of RDF over basic XML? (2.3.3, 2.4.4)
35. What are some of the possible ways the "Date" attribute in Dublin Core could be used? (2.4.4)
36. Develop a Dublin Core description for your home page, a book, or a document. Develop one for a DVD (2.4.4)
37. Explain the differences between simple, qualified, and extended Dublin Core. What are the strengths and weaknesses of each approach? (2.4.4)
38. Using the approach in Fig. 2.42, develop your own controlled vocabulary for either a sport of your choice or for an educational resource used at your university. (2.5.3)
39. Pick a site which you believe supports browsing of different sorts of users. Discuss what categories of users it is aimed for and how it supports each of those groups. (2.6.3)
40. Identify the types of users who are likely to go to a computer company Web site and their information needs. (2.6.3)
41. Describe some of the clues that can be provided to users to support navigation in hypertexts. (2.6.3)
42. How is navigation with a map related to navigation of a hypertext? How might navigation of a hypertext be improved using ideas from a map of physical space? If documents are to be created only for audio presentation, how would they be different from text and image documents? (2.6.3, 9.10.5)

Practicum:

1. **Objectives and Skills:**
2. Do classification. Create metadata.
3. XML for documents.
4. Build a thesaurus. (2.2.2)
5. Layout.
6. Simple XML, (2.3.3)

Going Beyond:

1. Do you agree with statement that "A record of any type of human thought is a document?" Explain. (2.3.1)
2. Describe some of the difficulties in transforming a complex object such as a table from one format into another second format. (2.3.3)
3. (a) Describe a program that would validate whether a document has XML tags which are consistent with a DTD. (b) Build it. (2.3.3, 10.4.2)
4. How would you develop metadata for a movie which is based on a book? (2.4.0)
5. The proliferation of XML standards may lead to a "tower of babble" in the use of different metadata schemes. How could that possibility be minimized? (2.3.3, 2.4.3)
6. Metadata is sometimes described as "data about data". Is that a good description? (2.4.3)
7. If you were developing a system of metadata what terms would you include? (2.4.3)
8. The Dublin Core "Type" attribute is often criticized as being vague. Explain whether or not you agree. (2.4.4)
9. Generate an example of Dublin Core using RDF. (2.4.4)
10. Should classification systems and tools that support them such as data description languages, support multiple inheritance? (2.5.2)
11. Describe and contrast how topics in mythology are cataloged by the Dewey and LCC classification systems. (2.5.1)
12. Develop a subject classification system for Web pages and build a tool to classify them. (2.5.1)
13. Some people argue that the non-linearity of hypertext frees readers from the limitations of linear thinking imposed by traditional documents. Do you agree with this criticism? (2.6.0, 10.2.0)
14. Build an application in frames and Javascript to present guided tours of Web pages. (2.6.2)
15. Pick two Web pages at random and find a path of links that goes between them. Is that the shortest path? (2.6.3)
16. Sample about 20 Web random pages and count how many links they have and report then in a bar chart. (2.6.3)

Teaching Notes

Objectives and Skills: The student should develop an understanding of document structure and learn the basics of XML and RDF, Making effective descriptions using metadata. Developing classification systems.

Instructor Strategies: The threads of XML and collaboration could be emphasized. Advanced practice with XML. Many of the themes of hypertext will be revisited later in other contexts and could be previewed here.

Related Books

- Bowker, G., and Starr, S. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge MA, 1999
- Brachman, R., and Levesque, H. *Knowledge Representation*. Morgan-Kaufmann, San Francisco, 2004.
- Jonassen, D.H., Beisser, K., and Yacci, M. *Structural Knowledge: Techniques for Representing. Conveying, and Acquiring Structural Knowledge*. Erlbaum Associates, Hillsdale NJ, 1993.
- Kent, W. *Data and Reality*. 1stBooks, 2000.
- Lakoff, G. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago, 1990.
- Levy, D. *Scrolling Forward: Making Sense of Documents in the Digital Age*. Arcade Publishing, New York, 2001.
- Lynch, P.J., and Horton, S. *Web Style Guide: Basic Design Principles for Creating Web Sites*. Yale University Press, New Haven CT, 1999.
- O'Connor, B.C. *Explanations in Indexing and Abstracting: Pointing, Virtue, and Power*. Libraries Unlimited, Westport CT, 1996.
- Rosenfeld, L. and Morville, P. *Information Architecture*. 2^{nd} ed. O'Reilly, Sebastopol CA, 2002.
- Sowa, J. *Knowledge Representation: Logical, Philosophical. and Computational Foundations*. Brooks/Cole, Pacific Grove CA, 2000.
- Svenonius, E. *The Intellectual Foundation of Information Organization* MIT Press, Cambridge MA, 2000.
- Taylor, A.G. *The Organization of Information*. Libraries Unlimited, Westport CT, 1999.