THE DISCIPLINE OF ORGANIZING
PROFESSIONAL EDITION

Edited by ROBERT J. GLUSHKO

This version of the 4th Edition (2016) is available under Creative Commons license CC BY-NC - https://creativecommons.org/licenses/by-nc/4.0/
Chapter 9
The Forms of Resource Descriptions

Ryan Shaw
Murray Maloney

9.1 Introduction
Throughout this book, we have emphasized the importance of separately considering fundamental organizing principles, application-specific concepts, and details of implementation. The three-tier architecture we introduced in §1.6 is one way to conceptualize this separation. In §6.7, we contrasted the implementation-focused perspective for analyzing relationships with other perspectives that focus on the meaning and abstract structure of relationships. In this chapter, we present this contrast between conceptualization and implementation in terms of separating the content and form of resource descriptions.

In the previous chapters, we have considered principles and concepts of organizing in many different contexts, ranging from personal organizing systems to cultural and institutional ones. We have noted that some organizing systems have limited scope and expected lifetime, such as a task-oriented personal organizing system like a shopping list. Other organizing systems support broad uses that rely on standard categories developed through rigorous processes, like a product catalog.

By this point you should have a good sense of the various conceptual issues you need to consider when deciding how to describe a resource in order to meet the
goals of your organizing system. Considering those issues will give you some
sense of what the content of your descriptions should be. In order to focus on
the conceptual issues, we have deferred discussion of specific implementation
issues. Implementation involves choosing the specific form of your descriptions,
and that is the topic of this chapter.

We can approach the problem of how to form resource descriptions from two
perspectives: structuring and writing. From one perspective, resource descrip-
tions are things that are used by both people and computational agents. From
this perspective, choosing the form of resource descriptions is a kind of design.
This is easy to see for certain kinds of resource descriptions, notably signs and
maps found in physical environments like airport terminals, public libraries, and
malls. In these spaces, resource descriptions are quite literally designed to help
people orient themselves and find their way. But any kind of resource descrip-
tion, not just those embedded in the built environment, can be viewed as a de-
signed object. Designing an object involves making decisions about how it
should be structured so that it can best be used for its intended purpose. From
a design perspective, choosing the form of a resource description means making
decisions about its structure.

In §6.5, we took a structural perspective on resources and the relationships
among them. In this chapter, we will take a structural perspective on resource
descriptions. The difference is subtle but important. A structural perspective on
resource relationships focuses on how people or computational processes asso-
ciate, arrange, and connect those resources. A structural perspective on re-
source descriptions focuses on how those associations, arrangements, and con-
nections are explicitly represented or implemented in the descriptions we cre-
ate. Mismatches between the structure imposed on the resources being organ-
ized and the structure of the descriptions used to implement that organization
could result in an organizing system that is complex, inefficient, and difficult to
maintain, as you will see in our first example (Example 9.1, Description struc-
tured as a dictionary).

The structures of resource descriptions enable or inhibit particular ways of in-
teracting with those descriptions, just as the descriptions themselves enable or
inhibit particular ways of interacting with the described resources. (See
§3.4 Designing Resource-based Interactions (page 127), and Chapter 10, Interac-
tions with Resources) Keep in mind that resource descriptions are themselves
information resources, so much of what we will say in this chapter is applicable
to the structures and forms of information resources in general. Put another
way, the structure and form of information resources informs the design of re-
source descriptions.

From another perspective, creating resource descriptions is a kind of writing. I
may describe something to you orally, but such a description might not be very
useful to an organizing system unless it were transcribed. Organizing systems need persistent descriptions, and that means they need to be written. In that sense, choosing the form of a resource description means making decisions about notation and syntax.

Modern Western culture tends to make a sharp distinction between designing and writing, but there are areas where this distinction breaks down, and the creation of resource descriptions in organizing systems is one of them. In the following sections, we will use designing and writing as two lenses for looking at the problem of how to choose the form of resource descriptions. Specifically, we will examine the spectrum of options we have for structuring descriptions, and the kinds of syntaxes we have for writing those descriptions.

9.2 Structuring Descriptions

Choosing how to structure resource descriptions is a matter of making principled and purposeful design decisions in order to solve specific problems, serve specific purposes, or bring about some desirable property in the descriptions. Most of these decisions are specific to a domain: the particular context of application for the organizing system being designed and the kinds of interactions with resources it will enable. Making these kinds of context-specific decisions results in a model of that domain. (See §5.3.1.2 Abstraction in Resource Description (page 251).)

Over time, many people have built similar kinds of descriptions. They have had similar purposes, desired similar properties, and faced similar problems. Unsurprisingly, they have converged on some of the same decisions. When common sets of design decisions can be identified that are not specific to any one domain, they often become systematized in textbooks and in design practices, and may eventually be designed into standard formats and architectures for creating organizing systems. These formally recognized sets of design decisions are known as abstract models or metamodels. Metamodels describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain. While any designer of an organizing system will usually create a model of her specific domain, she usually will not create an entirely new metamodel but will instead make choices from among the metamodels that have been formally recognized and incorporated into existing standards. The resulting model is sometimes called a “domain-specific language.” Reusing standard metamodels can bring great economical advantages, as developers can reuse tools designed for and knowledge about these metamodels, rather than having to start from scratch.

In the following sections, we examine some common kinds of structures used as the basis for metamodels. But first, we consider a concrete example of how the structure of resource descriptions supports or inhibits particular uses. As we
Figure 9.1. A Batten Card.

[Table showing a Batten Card]

An example of a punch card used by Batten to describe a particular patent in a patent collection. Each card represented an individual description term, and each punch position on a card represented a particular patent.

explained in Chapter 1, the concept of a resource de-emphasizes the differences between physical and digital things in favor of focusing on how things, in general, are used to support goal-oriented activity. Different kinds of books can be treated as information resources regardless of the particular mix of tangible and intangible properties they may have. Since resource descriptions are also information resources, we can similarly consider how their structures support particular uses, independent of whether they are physical, digital, or a mix of both.

During World War II, a British chemist named W. E. Batten developed a system for organizing patents. The system consisted of a language for describing the product, process, use, and apparatus of a patent, and a way of using punched cards to record these descriptions. Batten used cards printed with matrices of 800 positions (see Figure 9.1). Each card represented a specific value from the vocabulary of the description language, and each position corresponded to a particular patent. To describe patent #256 as covering extrusion of polythene to produce cable coverings, one would first select the cards for the val-
uses polythene, extrusion, and cable coverings, and then punch each card at the 256th position. The description of patent #256 would thus extend over these three cards.

The advantage of this structure is that to find patents covering extrusion of polythene (for any purpose), one needs only to select the two cards corresponding to those values, lay one on top of the other, and hold them up to a light. Light will shine through wherever there is a position corresponding to a patent described using those values. Patents meeting a certain description are easily found due to the structure of the cards designed to describe the patents.

Of course, this system has clear disadvantages as well. Finding the concepts associated with a particular patent is tedious, because every card must be inspected. Adding a new patent is relatively easy as long as there is an index that allows the cards for specific concepts to be located quickly. However, once the cards run out of space for punching holes, the whole set of cards must be duplicated to accommodate more patents: a very expensive operation. Adding new concepts is potentially easy: simply add a new card. But if we want to be able to find existing patents using the new concept, all the existing patents would have to be reexamined to determine whether their positions on the new card should be punched: also an expensive operation.

The structure of Batten’s cards supported rapid selection of resources given a partial description. The kinds of structures we will examine in the following sections are not quite so elaborate as Batten’s cards. But like the cards, each kind of structure supports more efficient mechanical execution of certain operations, at the cost of less efficient execution of others.
9.2.1 Kinds of Structures

Sets, lists, dictionaries, trees, and graphs are kinds of structures that can be used to form resource descriptions. As we shall see, each of these kinds is actually a family of related structures. These structures are abstractions: they describe formal structural properties in a general way, rather than specifying an exact physical or textual form. Abstractions are useful because they help us to see common properties shared by different specific ways of organizing information. By focusing on these common properties, we can more easily reason about the operations that different forms support and the affordances that they provide, without being distracted by less relevant details.

9.2.1.1 Blobs

The simplest kind of structure is no structure at all. Consider the following description of a book: Sebald’s novel uses a walking tour in East Anglia to meditate on links between past and present, East and West. This description is an unstructured text expression with no clearly defined internal parts, and we can consider it to be a blob. Or, more precisely, it has structure, but that structure is the underlying grammatical structure of the English language, and none of that grammatical structure is explicitly represented in a surface structure when the sentence is expressed. As readers of English we can interpret the sentence as a description of the subject of the book, but to do this mechanically is difficult. On the other hand, such a written description is relatively easy to create, as the describer can simply use natural language.

A blob need not be a blob of text. It could be a photograph of a resource, or a recording of a spoken description of a resource. Like blobs of text, blobs of pixels or sound have underlying structure that any person with normal vision or hearing can understand easily. But we can treat these blobs as unstructured, because none of the underlying structure in the visual or auditory input is explicit, and we are concerned with the ways that the structures of resource descriptions support or inhibit mechanical or computational operations.

9.2.1.2 Sets

The simplest way to structure a description is to give it parts and treat them as a set. For example, the description of Sebald’s novel might be reformulated as a set of terms: Sebald, novel, East Anglia, walking, history. Doing this has lost much of the meaning, but something has been gained: we now can easily distinguish Sebald and walking as separate items in the description. This makes it easier to find, for example, all the descriptions that include the term walking. (Note that this is different from simply searching through blob-of-text descriptions for the word walking. When treated as a set, the description Fiji, fire walk-
ing, memoir does not include the term walking, though it does include the term fire walking.)

Sets make it easy to find intersections among descriptions. Sets are also easy to create. In §8.1.2 we looked at “folksonomies,” organizing systems in which non-professional users create resource descriptions. In these systems, descriptions are structured as sets of “tags.” To find resources, users can specify a set of tags to obtain resources having descriptions that intersect at those tags. This is more valuable if the tags come from a controlled vocabulary, making intersections more likely. But enforcing vocabulary control adds complexity to the description process, so a balance must be struck between maximizing potential intersections and making description as simple as practical.

A set is a type or class of structure. We can refine the definition of different kinds of sets by introducing constraints. For example, we might introduce the constraint that a given set has a maximum number of items. Or we might constrain a set to always have the same number of items, giving us a fixed-size set. We can also remove constraints. Sets do not contain duplicate items (think of a tagging system in which it does not make sense to assign the same tag more than once to the same resource). If we remove this uniqueness constraint, we have a different structure known as a “bag” or “multiset.”

9.2.1.3 Lists

Constraints are what distinguish lists from sets. A list, like a set, is a collection of items with an additional constraint: their items are ordered. If we were designing a tagging system in which it was important that the order of the tags be maintained, we would want to use lists, not sets. Unlike sets, lists may contain duplicate items. In a list, two items that are otherwise the same can be distinguished by their position in the ordering, but in a set this is not possible. For example, we might want to organize the tags assigned to a resource, listing the most used tag first, the least frequently used last, and the rest according to their frequency of use.

Again, we can introduce constraints to refine the definition of different kinds of lists, such as fixed-length lists. If we constrain a list to contain only items that are themselves lists, and further specify that these contained lists do not themselves contain lists, then we have a table (a list of lists of items). A spreadsheet is a list of lists.
9.2.1.4 Dictionaries

One major limitation of lists and sets is that, although items can be individually addressed, there is no way to distinguish the items except by comparing their values (or, in a list, their positions in the ordering). In a set of terms like Sebald, novel, East Anglia, walking, history, for example, one cannot easily tell that Sebald refers to the author of the book while East Anglia and walking refer to what it is about. One way of addressing this problem is to break each item in a set into two parts: a property and a value. So, for example, our simple set of tags might become author: Sebald, type: novel, subject: East Anglia, subject: walking, subject: history. Now we can say that author, type, and subject are the properties, and the original items in the set are the values.

author
  Sebald

type
  novel

subject1
  East Anglia

subject2
  walking

subject3
  history

This kind of structure is called a dictionary, a map or an associative array. A dictionary is a set of property-value pairs or entries. It is a set of entries, not a list of entries, because the pairs are not ordered and because each entry must have a unique key. Note that this specialized meaning of dictionary is different from the more common meaning of “dictionary” as an alphabetized list of terms accompanied by sentences that define them. The two meanings are related, however. Like a “real” dictionary, a dictionary structure allows us to easily find the value (such as a definition) associated with a particular property or key (such as a word). But unlike a real dictionary, which orders its keys alphabetically, a dictionary structure does not specify an order for its keys.

Dictionaries are ubiquitous in resource descriptions. Structured descriptions entered using a form are easily represented as dictionaries, where the form items' labels are the properties and the data entered are the values. Tabular data with a “header row” can be thought of as a set of dictionaries, where column headers are the properties for each dictionary, and each row is a set of corresponding values. Dictionaries are also a basic type of data structure found in nearly all programming languages (referred to as associative arrays).
Again, we can introduce or remove constraints to define specialized types of dictionaries. A sorted dictionary adds an ordering over entries; in other words, it is a list of entries rather than a set. A *multimap* is a dictionary in which multiple entries may have the same key.

### 9.2.1.5 Trees

In dictionaries as they are commonly understood, properties are terms and values are their corresponding definitions. The terms and values are usually words, phrases, or other expressions that can be ordered alphabetically. But if generalize the notion of a dictionary as abstract sets of property-value pairs, the values can be anything at all. In particular, the values can themselves be dictionaries. When a dictionary structure has values that are themselves dictionaries, we say that the dictionaries are *nested*. Nesting is very useful for resource descriptions that need more structure than what a (non-nested) dictionary can provide.

*Figure 9.2. Four Nested Dictionaries.*

When a dictionary contains other dictionaries, they are said to be nested.

*Figure 9.2, Four Nested Dictionaries.* presents an example of nested dictionaries. At the top level there is one dictionary with a single entry having the property *a*. The value associated with *a* is a dictionary consisting of two entries, the first having property *b* and the second having property *c*. The values associated with *b* and with *c* are also dictionaries.

If we nest dictionaries like this, and our “top” dictionary (the one that contains all the others) has only one entry, then we have a kind of *tree* structure. *Figure 9.3, A Tree of Properties and Values.* shows the same properties and values...
as Figure 9.2, this time arranged to make the tree structure more visible. Trees consist of nodes (the letters and numbers in Figure 9.3) joined by edges (the arrows). Each node in the tree with a circle around it is a property, and the value of each property consists of the nodes below (to the right of) it in the tree. A node is referred to as the parent of the nodes below it, which in turn are referred to as the children of that node. The edges show these “parent of” relationships between the nodes. The node with no parent is called the root of the tree. Nodes with no children are called leaf nodes.

**Figure 9.3. A Tree of Properties and Values.**

An alternative representation of nested dictionaries is as a tree. The lowest level or leaf nodes of the tree contain property values.

As with the other types of structures we have considered, we can define different kinds of trees by introducing different types of constraints. For example, the predominant metamodel for XML is documents is a kind of tree called the **XML Information Set** or Infoset. 525[Com]

The XML Information Set defines a specific kind of tree structure by adding very specific constraints, including ordering of child nodes, to the basic definition of a tree. The addition of an ordering constraint distinguishes XML trees from nested dictionaries, in which child nodes do not have any order (because
dictionary entries do not have an ordering). Ordering is an important constraint for resource descriptions, since without ordering it is impossible to, for example, list multiple authors while guaranteeing that the order of authors will be maintained. Figure 9.3 depicts a kind of tree with a different set of constraints: all non-leaf nodes are properties, and all leafs are values. We could also define a tree in which every node has both a property and a value. Trees exist in a large variety of flavors, but they all share a common topology: the edges between nodes are directed (one node is the parent and the other is the child), and every node except the root has exactly one parent.

Trees provide a way to group statements describing different but related resources. For example, consider the description structured as a dictionary here:

**Example 9.1. Description structured as a dictionary**

- author given names → Winfried Georg
- author surname → Sebald
- title → Die Ringe des Saturn
- pages → 371

The dictionary groups together four property-value pairs describing a particular book. (The arrows are simply a schematic way to indicate property-value relations. Later in the chapter we look at ways to “write” these relations using some specific syntax.)

But really the first two entries are not describing the book; they are describing the book’s author. So, it would be better to group those two statements somehow. We can do this by nesting the entries describing the author within the book description, creating a tree structure:

**Example 9.2. Nesting an author description within a book description**

- author →
  - given names → Winfried Georg
  - surname → Sebald
- title → Die Ringe des Saturn
- pages → 371

Using a tree works well in this case because we can treat the book as the primary resource being described, making it the root of our tree, and adding on the author description as a “branch.”
We also could have chosen to make the author the primary resource, giving us a tree like the one in Example 9.3.

**Example 9.3. Nesting book descriptions within an author description**

- given names → Winfried Georg
- surname → Sebald
- books authored →
  1. title → Die Ringe des Saturn
     pages → 371
  2. title → Austerlitz
     pages → 416

Note that in this dictionary, the value of the books authored property is a list of dictionaries. Making the author the primary or root resource allows us to include multiple book descriptions in the tree (but makes it more difficult to describe books having multiple authors). A tree is a good choice for structuring descriptions as long as we can clearly identify a primary resource. In some cases, however, we want to connect descriptions of related resources without having to designate one as primary. In these cases, we need a more flexible data structure.

9.2.1.6 Graphs

Suppose we were describing two books, where the author of one book is the subject of the other, as in Example 9.4, Two related descriptions:

**Example 9.4. Two related descriptions**

1. author → Mark Richard McCulloch
   title → Understanding W. G. Sebald
   subject → Winfried Georg Sebald
2. author → Winfried Georg Sebald
   title → Die Ringe des Saturn

By looking at these descriptions, we can guess the relationship between the two books, but that relationship is not explicitly represented in the structure: we just have two separate dictionaries and have inferred the relationship by matching property values. It is possible that this inference could be wrong: there might be two people named Winfried Georg Sebald. How can we structure these descriptions to explicitly represent the fact that the Winfried Georg Sebald that is the subject of the first book is the same Winfried Georg Sebald who authored the second?

One possibility would be to make Winfried Georg Sebald the root of a tree, similar to the approach taken in Example 9.3, Nesting book descriptions within an author description, adding a book about property alongside the books authored.
Stop and Think: Social Network Properties

Compare the concept of “friend” in Facebook with that of “follower” in Twitter, in terms of the semantic properties discussed in §6.3.2 Properties of Semantic Relationships (page 308) and the graph properties discussed in this section.

one. This solution would work fine if people were our primary resources, and it thus made sense to structure our descriptions around them. But suppose that we had decided that our descriptions should be structured around books, and that we were using a vocabulary that took this perspective (with properties such as author and subject rather than books authored and books about). We should not let a particular structure limit the organizational perspective we can take, as Batten’s cards did. Instead, we should consciously choose structures to suit our organizational perspective. How can we do this?

If we treat our two book descriptions as trees, we can join the two branches (subject and author) that share a value. When we do this, we no longer have a tree, because we now have a node with more than one parent (Figure 9.4, Descriptions Linked into a Graph.). The structure in Figure 9.4, Descriptions Linked into a Graph. is a graph. Like a tree, a graph consists of a set of nodes connected by edges. These edges may or may not have a direction (§6.6.3 Directionality (page 331)). If they do, the graph is referred to as a “directed graph.” If a graph is directed, it may be possible to start at a node and follow edges in a path that leads back to the starting node. Such a path is called a “cycle.” If a directed graph has no cycles, it is referred to as an “acyclic graph.”

A tree is just a more constrained kind of graph. Trees are directed graphs because the “parent of” relationship between nodes is asymmetric: the edges are arrows that point in a certain direction. (See §6.3.2.1 Symmetry (page 308).) Furthermore, trees are acyclic graphs, because if you follow the directed edges from one node to another, you can never encounter the same node twice. Finally, trees have the constraint that every node (except the root) must have exactly one parent.

In Figure 9.4, Descriptions Linked into a Graph. we have violated this constraint by joining our two book trees. The graph that results is still directed and acyclic, but because the Winfried George Sebald node now has two parents, it is no longer a tree.

Graphs are very general and flexible structures. Many kinds of systems can be conceived of as nodes connected by edges: stations connected by subway lines, people connected by friendships, decisions connected by dependencies, and so on. Relationships can be modeled in different ways using different kinds of graphs. For example, if we assume that friendship is symmetric (see §6.3.2.1 Symmetry (page 308)), we
Figure 9.4. Descriptions Linked into a Graph.

Descriptions can be linked to form a graph when the value assigned to two different properties is the same.

would use an undirected graph to model the relationship. However, in web-based social networks friendship is often asymmetric (you might “friend” someone who does not reciprocate), so a directed graph is more appropriate.

Often it is useful to treat a graph as a set of pairs of nodes, where each pair may or may not be directly connected by an edge. Many approaches to characterizing structural relationships among resources (see §6.5.3 Structural Relationships between Resources (page 324)) are based on modeling the related resources as a set of pairs of nodes, and then analyzing patterns of connectedness among them. As we will see, being able to break down a graph into pairs is also useful when we structure resource descriptions as graphs.

In §9.4.1 we will use XML to model the graph shown in Figure 9.4, Descriptions Linked into a Graph, by using “references” to connect a book to its title, authors and subject. This will allow us to develop sophisticated graphs of knowledge within a single XML document instance. (See also the sidebar, Inclusions and References (page 486))
9.2.2 Comparing Metamodels: JSON, XML and RDF

Now that we are familiar with the various kinds of metamodels used to structure resource descriptions, we can take a closer look at some specific metamodels. A detailed comparison of the affordances of different metamodels is beyond the scope of this chapter. Here we will simply take a brief look at three popular metamodels—JSON, XML, and RDF—in order to see how they further specify and constrain the more general kinds of metamodels introduced above.

9.2.2.1 JSON

*JavaScript Object Notation (JSON)*

*JavaScript Object Notation (JSON)* is a textual format for exchanging data that borrows its metamodel from the JavaScript programming language. Specifically, the JSON metamodel consists of two kinds of structures found in JavaScript: lists (called “arrays” in JavaScript) and dictionaries (called “objects” in JavaScript). Lists and dictionaries contain values, which may be strings of text, numbers, Booleans (true or false), or the null (empty) value. Again, these types of values are taken directly from JavaScript. Lists and dictionaries can be values too, meaning lists and dictionaries can be nested within one another to produce more complex structures such as tables and trees.

Lists, dictionaries, and a basic set of value types constitute the JSON metamodel. Because this metamodel is a subset of JavaScript, the JSON metamodel is very easy to work with in JavaScript. Since JavaScript is the only programming language that is available in all web browsers, JSON has become a popular choice for developers who need to work with data and resource descriptions on the web. (See §9.3.2 Writing Systems (page 494) later in this chapter.) Furthermore, many modern programming languages provide data structures and value types equivalent to those provided by JavaScript. So, data represented as JSON is easy to work with in many programming languages, not just JavaScript.

9.2.2.2 XML Information Set

The *XML Information Set* metamodel is derived from data structures used for document markup. (See §5.2.2.2.) These markup structures—*elements* and *attributes*—are well suited for programmatically manipulating the structure of documents and data together.\[528\]

*XML Infoset*

The *XML Infoset* is a tree structure, where each node of the tree is defined to be an “information item” of a particular type. Each information item has a set of type-specific properties associated with it. At the root of the tree is a “document item,” which has exactly one “element item” as its child. An *element* item has a set of *attribute* items, and a list of child nodes. These child
nodes may include other element items, or they may be character items. (See §9.2.1 Kinds of Structures (page 472) below for more on characters.) Attribute items may contain character items, or they may contain typed data, such as name tokens, identifiers and references. Element identifiers and references (ID/IDREF) may be used to connect nodes, transforming a tree into a graph. (See the sidebar, Inclusions and References (page 486)).

Figure 9.5, A Description Structure. is a graphical representation of how an XML document might be used to structure part of a description of an author and his works. This example demonstrates how we might use element items to model the domain of the description, by giving them names such as author and title. The character items that are the children of these elements hold the content of the description: author names, book titles, and so on. Attribute items are used to hold auxiliary information about this content, such as its language.

![Figure 9.5. A Description Structure.](image)

An XML document can be described as a tree in which elements are nodes that can contain character content directly or attributes that contain character content.

This example also demonstrates how the XML Infoset supports mixed content by allowing element items and character items to be “siblings” of the same parent element. In this case, the Infoset structure allows us to specify that the
book description can be displayed as a line of text consisting of the original title and the translated title in parentheses. The elements and attributes are used to indicate that this line of text consists of two titles written in different languages, not a single title containing parentheses.

If not for mixed content, we could not write narrative text with hypertext links embedded in the middle of a sentence. It gives us the ability to identify the sub-components of a sentence, so that we could distinguish the terms “Sebald,” “walking” and “East Anglia” as an author and two subjects.

Using schemas to define data representation formats is a good practice that facilitates shared understanding and contributes to long-term maintainability in institutional or business contexts. An XML schema represents a contract among the parties subscribing to its definitions, whereas JSON depends on out-of-band communication among programmers. The notion that “the code is the documentation” may be fashionable among programmers, but modelers prefer to design at a higher level of abstraction and then implement.

The XML Infoset presents a strong contrast to JSON and does not always map in a straightforward way to the data structures used in popular web scripting languages. Whereas JSON’s structures make it easier for object-oriented programmers to readily exchange data, they lack any formal schema language and cannot easily handle mixed content.

9.2.2.3 RDF

In Figure 9.4, Descriptions Linked into a Graph., we structured our resource description as a graph by treating resources, properties, and values as nodes, with edges reflecting their combination into descriptive statements. However, a more common approach is to treat resources and values as nodes, and properties as the edges that connect them. Figure 9.6, Treating Properties as Edges Rather Than Nodes. shows the same description as Figure 9.4, Descriptions Linked into a Graph., this time with properties treated as edges. This roughly corresponds to the particular kind of graph metamodel defined by RDF. (§5.2.2.4 Resource Description Framework (RDF) (page 241))

We have noted that we can treat a graph as a set of pairs of nodes, where each pair may be connected by an edge. Similarly, we can treat each component of the description in Figure 9.6, Treating Properties as Edges Rather Than Nodes. as a pair of nodes (a resource and a value) with an edge (the property) linking them. In the RDF metamodel, a pair of nodes and its edge is called a triple, because it consists of three parts (two nodes and one edge). The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the subject of the triple, and the other node (the one to which the edge is pointing) as its object. The edge is referred to as the predicate or (as we have been saying) property of the triple.
We can treat each component of a description as a pair of nodes (a resource and a value) with an edge (the property) linking them. Here, we have two book resources that are related to four values through five properties. The single value node, “Winfried George Sebald” is the subject of one book while being the author of the second book. The books are depicted as boxes, the edges as labeled arrows and the values as text strings.

Figure 9.7, Listing Triples Individually. lists separately all the triples in Figure 9.6. However, there is something missing in Figure 9.7. Figure 9.6 clearly indicates that the Winfried George Sebald who is the subject of book 1 is the same Winfried George Sebald who is the author of book 2. In Figure 9.7, Listing Triples Individually, this relationship is not clear. How can we tell if the Winfried George Sebald of the third triple is the same as the Winfried George Sebald of the triple statement? For that matter, how can we tell if the first three triples all involve the same book 1? This is easy to show in a diagram of the entire description.
Figure 9.7. Listing Triples Individually.

Lists each of the triples individually. Here, each statement relates one resource to one value through an edge. Thus, we have two distinct “Winfried George Sebald” value nodes. The books are depicted as boxes, the edges as labeled arrows and the values as text strings.

graph, where we can have multiple edges attached to a node. But when we dis-aggregate that graph into triples, we need some way of uniquely referring to nodes. We need identifiers (§4.4.3 Choosing Good Names and Identifiers (page 209)). When two triples have nodes with the same identifier, we can know that it is the same node. RDF achieves this by associating URIs with nodes. (See §5.2.2.4 Resource Description Framework (RDF) (page 241))

The need to identify nodes when we break down an RDF graph into triples becomes important when we want to “write” RDF graphs—create textual representations of them instead of depicting them—so that they can be exchanged as data. Tree structures do not necessarily have this problem, because it is pos-
Inclusions and References

An XML Infoset is typically the result of processing a well-formed XML document instance. Schemas associated with XML document instances “inform” the corresponding XML Infoset. Thus, the “truth value” of any XML Infoset is dependent upon its related schemas. Traditionally, any documentation that is related to the schema is considered to be part of the schema definition and, at least notionally, informs human understanding and interpretation of corresponding documents.

The XML family offers several mechanisms to create inclusion relationships: by employing element references; by way of entity definition and reference; by using XML Inclusions (XInclude) or XLink. These inclusions and references can also inform the XML Infoset, if they are processed.

Any XML node may refer to another node simply by referencing it by its assigned ID. Assuming attributes are declared, the Infoset exposes this information as a references property as an ordered list of element information items. That is to say that an element may contain other element nodes by subordination, or by reference.

XInclude “specifies a processing model and syntax for general purpose inclusion. Inclusion is accomplished by merging a number of XML information sets into a single composite infoset.” XInclude offers the most versatile mechanism for addressing whole documents, specific information items, ranges of information items, and even parts of information items, which has led to its widespread adoption in document processing.

XLink “allows elements to be inserted into XML documents in order to create and describe links between resources. It uses XML syntax to create structures that can describe links similar to the simple unidirectional hyperlinks of today’s HTML, as well as more sophisticated links.” Entities are similar to macros found in many programming languages; a value is assigned to a token, the token is referenced wherever the value is needed, and macro expansion happens when the XML document instance is read into the Infoset. Entities are a handy feature, but since they are expanded on their way in, entities do not survive as information items in the XML Infoset. The ID/IDREF feature is more popular than the use of entities because it carries more information into the XML Infoset.

It is possible to textually represent a tree structure without having to mention any node more than once. Thus, one price paid for the generality and flexibility of graph structures is the added complexity of recording, representing or writing those structures.
9.2.2.4 Choosing Your Constraints

This tradeoff between flexibility and complexity illustrates a more general point about constraints. In the context of managing and interacting with resource descriptions, constraints are a good thing. As discussed above, a tree is a graph with very specific constraints. These constraints allow you to do things with trees that are not possible with graphs in general, such as representing them textually without repeating yourself, or uniquely identifying nodes by the path from the root of the tree to that node. This can make managing descriptions and the resources they describe easier and more efficient—if a tree structure is a good fit to the requirements of the organizing system. For example, an ordered tree structure is a good fit for the hierarchical structure of the content of a book or book-like document, such as an aircraft service manual or an SEC filing. On the other hand, the network of relationships among the people and organizations that collaborated to produce a book might be better represented using a graph structure. XML is most often used to represent hierarchies, but is also capable of representing network structures.

9.2.3 Modeling within Constraints

A metamodel imposes certain constraints on the structure of our resource descriptions. But in organizing systems, we usually need to further specify the content and composition of descriptions of the specific types of resources being organized. For example, when designing a system for organizing books, it is not sufficient to say that a book’s description is structured using XML, because the XML metamodel constrains structure and not the content of descriptions. We need also to specify that a book description includes a list of contributors, each entry of which provides a name and indicates the role of that contributor. This kind of specification is a model to which our descriptions of books are expected to conform. (See §5.3.1.2 Abstraction in Resource Description (page 251).)

When designing an organizing system we may choose to reuse a standard model. For example, ONIX for Books is a standard model (conforming to the XML metamodel) developed by the publishing industry for describing books.

If no such standard exists, or existing standards do not suit our needs, we may create a new model for our specific domain. But we will not usually create a new metamodel: instead we will make choices from among the metamodels, such as JSON, XML, or RDF, that have been formally recognized and incorporated into existing standards. Once we have selected a metamodel, we know the constraints we have to work with when modeling the resources and collections in our specific domain.
9.2.3.1 Specifying Vocabularies and Schemas

Creating a model for descriptions of resources in a particular domain involves specifying the common elements of those descriptions, and giving those elements standard names. (See §5.3 The Process of Describing Resources (page 246)) The model may also specify how these elements are arranged into larger structures, for example, how they are ordered into lists nested into trees. Meta-models vary in the tools they provide for specifying the structure and composition of domain-specific models, and in the maturity and robustness of the methods for designing them. RDF and XML each provide different, metamodel-specific tools to define a model for a specific domain. But not every metamodel provides such tools.

In XML, models are defined in separate documents known as schemas. An XML schema defining a domain model provides a vocabulary of terms that can be used as element and attribute names in XML documents that adhere to that model. For example, Onix for Books schema specifies that an author of a book should be called a Contributor, and that the page count should be called an Extent. An XML schema also defines rules for how those elements, attributes, and their content can be arranged into higher-level structures. For example, the Onix for Books specifies that the description of a book must include a list of Contributor elements, that this list must have at least one element in it, and that each Contributor element must have a ContributorRole child element.

If an XML schema is given an identifier, XML documents can use that identifier to indicate that they use terms and rules from that schema. An XML document may use vocabularies from more than one XML schema. Associating a schema with an XML instance enables validation: automatically checking that vocabulary terms are being used correctly.

If two descriptions share the same XML schema and use only that schema, then combining them is straightforward. If not, it can be problematic, unless someone has figured out exactly how the two schemas should “map” to one another. Finding such a mapping is not a trivial problem, as XML schemas may differ semantically, lexically, structurally, or architecturally despite sharing a common implementation form. (See Chapter 6, Describing Relationships and Structures.)

Tree structures can vary considerably while still conforming to the XML Infoset metamodel. Users of XML often specify rules for checking whether certain patterns appear in an XML document (document-level validation). This is less often done with RDF, because graphs that conform to the RDF metamodel all have the same structure: they are all sets of triples. This shared structure makes it simple to combine different RDF descriptions without worrying about checking structure at the document level. However, sometimes it is desirable to check descriptions at the document level, as when part of a description is required. As with XML, if consumers of those descriptions want to assert that they expect
those descriptions to have a certain structure (such as a required property), they must check them at the document level.

Because the RDF metamodel already defines structure, defining a domain-specific model in RDF mainly involves specifying URIs and names for predicates. A set of RDF predicate names and URIs is known as an RDF vocabulary. Publication of vocabularies on the web and the use of URIs to identify and refer to predicate definitions are key principles of Linked Data and the Semantic Web. (Also see §6.8.1, as well as later in this chapter.)

For example, the Resource Description and Access (RDA) standard for cataloging library resources includes a set of RDF vocabularies defining predicates usable in cataloging descriptions. One such predicate is:

\[
\text{<http://rdvocab.info/Elements/extentOfText>}
\]

which is defined as “the number and type of units and/or subunits making up a resource consisting of text, with or without accompanying illustrations.” The vocabulary further specifies that this predicate is a refinement of a more general predicate:

\[
\text{<http://rdvocab.info/Elements/extent>}
\]

which can be used to indicate, “the number and type of units and/or subunits making up a resource” regardless of whether it is textual or not.

JSON lacks any standardized way to define which terms can be used. That does not mean one cannot use a standard vocabulary when creating descriptions using JSON, only that there is no agreed-upon way to use JSON to communicate which vocabulary is being used, and no way to automatically check that it is being used correctly.

9.2.3.2 Controlling Values

So far, we have focused on how models specify vocabularies of terms and how those terms can be used in descriptions. But models may also constrain the values or content of descriptions. Sometimes, a single model will define both the terms that can be used for property names and the terms that can be used for property values. For example, an XML schema may enumerate a list of valid terms for an attribute value.

Often, however, there are separate, specialized vocabularies of terms intended for use as property values in resource descriptions. Typically these vocabularies provide values for use within statements that describe what a resource is about. Examples of such subject vocabularies include the Library of Congress Subject Headings (LOC-SH) and the Medical Subject Headings (MeSH). Other vocabularies may provide authoritative names for people, corporations, or places.
Classification schemes are yet another kind of vocabulary, providing the category names for use as the values in descriptive statements that classify resources.

Because different metamodels take different approaches to specifying vocabularies, there will usually be different versions of these vocabularies for use with different metamodels. For example the LCSH are available both as XML conforming to the Metadata Authority Description Standard (MADS) schema, and as RDF using the Simple Knowledge Organization System (SKOS) vocabulary.

Specifying a vocabulary is just one way models can control what values can be assigned to properties. Another strategy is to specify what types of values can be assigned. For example, a model for book descriptions may specify that the value of a pages property must be a positive integer. Or it could be more specific; a course catalog might give each course an identifier that contains a two-letter department code followed by a 1-3 digit course number. Specifying a data type like this with a regular expression narrows down the set of possible values for the property without having to enumerate every possible value. (See the sidebar.)

In addition to or in lieu of specifying a type, a model may specify an encoding scheme for values. An encoding scheme is a specialized writing system or syntax for particular types of values. For example, a model like Atom for describing syndicated web content requires a publication date. But there are many different ways to write dates: 9/2/76, 2 Sept. 1976, September 2nd 1976, etc. Atom also specifies an encoding scheme for date values. The encoding scheme is RFC3339, a standard for writing dates. When using RFC3339, one always writes a date using the same form: 1976-09-02.

Encoding schemes are often defined in conjunction with standardized identifiers. (See §4.4.3.1 Make Names Informative (page 210).) For example, International Standard Book Numbers (ISBN) are not just sequences of Arabic numerals: they are values written using the ISBN encoding scheme. This scheme specifies how to separate the sequence of numerals into parts, and how each of these parts should be interpreted. The ISBN 978-3-8218-4448-0 has five parts, the first three of which indicate that the resource with this identifier is 1) a product of the book publishing industry, 2) published in a German-speaking country, and 3) published by the publishing house Eichborn.

Encoding schemes can be viewed as very specialized models of particular kinds of information, such as dates or book identifiers. But because they specify not only the structure of this information, but also how it should be written, we can also view them as specialized writing systems. That is, encoding schemes specify how to textually represent information.

In the second half of this chapter, we will focus on the issues involved in textually representing resource descriptions—writing them. Graphs, trees, dictio-
Regular Expressions

Regular expressions have been used to describe patterns in text documents since the early days of computing and came into widespread use when Ken Thompson incorporated them into early UNIX text processing tools, such as ed and grep. There are too many variations of regular expression syntax for us to detail them here, but it is worthwhile to consider them briefly while we are on the subject of controlling values.

Regular expressions are employed by modern text processing tools for selection and retrieval purposes. In search and replace applications, one might search for the string “Chapter [1-5]” to express your intent to select chapters 1 through 5, or “it[’]s” to locate every use of “it’s” and “its” in a manuscript; this capability is highly valued by anyone who has had to edit a book. Programmers and data modelers use regular expressions to describe expected encoding schemes when they design documents, data elements, databases, and encoding schemes. You experience regular expression processing when you enter a phone number or postal code into a Web-based form. Many data modeling, programming and XML schema languages employ regular expressions to control data entry and validation of values. In the context of controlling values, we can use regular expressions to describe data values as varied as identifiers, names, dates, telephone numbers, and postal codes. We can, likewise, define rules for white space handling and punctuation within a data value.

naries, lists, and sets are general types of structures found in different metamodels. Thinking about these broad types and how they fit or do not fit the ways we want to model our resource descriptions can help us select a specific metamodel. Specific metamodels such as the XML Infoset or RDF are formalized and standardized definitions of the more general types of structures discussed above. Once we have selected a metamodel, we know the constraints we have to work with when modeling the resources and collections in our specific domain. But because metamodels are abstract and exist only on a conceptual level, they can only take us so far. If we want to create, store, and exchange individual resource descriptions, we need to make the structures defined by our abstract metamodels concrete. We need to write them.
9.3 Writing Descriptions

Suppose that I am organizing books, and I have decided that it is important for the purposes of this organizing to know the title of each book and how many pages it has. Before me I have a book, which I examine to determine that its title is Die Ringe des Saturn and it has 371 pages. Example 9.5, Basic ways of writing part of a book description, lists a few of the ways to write this description. Let us examine these various forms of writing to see what they have in common and where they differ.

Example 9.5. Basic ways of writing part of a book description.

<table>
<thead>
<tr>
<th>The title is Die Ringe des Saturn and it has 371 pages.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ book: {&quot;title&quot;:&quot;Die Ringe des Saturn&quot;,&quot;pages&quot;:371} }</td>
</tr>
<tr>
<td>&lt;div class=&quot;book&quot;&gt;The title is &lt;span class=&quot;title&quot;&gt;Die Ringe des Saturn&lt;/span&gt; and it has &lt;span class=&quot;pages&quot;&gt;371 pages&lt;/span&gt;. &lt;/div&gt;</td>
</tr>
</tbody>
</table>

We examine the notations, writing systems and syntax of each of these description forms, and others, in the following sections.

9.3.1 Notations

First, let us look at the actual marks on the page. To write you must make marks or—more likely—select from a menu of marks using a keyboard. In either case, you are using a notation: a set of characters with distinct forms. The Latin alphabet is a notation, as are Arabic numerals. Some more exotic notations include the symbols used for editorial markup and alchemical symbols. The characters in a notation usually have an ordering. Arabic numerals are ordered 1 2 3 and so on. English-speaking children usually learn the ordering of the Latin alphabet in the form of an alphabet song.

A character may belong to more than one notation. The examples in Example 9.5, Basic ways of writing part of a book description, use characters from a few different notations: the letters of the Latin alphabet, Arabic numerals, and a handful of auxiliary marks: . { } " :< > / $. Collectively, all of these characters—alphabet, numerals, and auxiliary marks—also belong to a notation called the American Standard Code for Information Interchange (ASCII).
ASCII is an example of a notation that has been codified and standardized for use in a digital environment. A traditional notation like the Latin alphabet can withstand a certain degree of variation in the form of a particular mark. Two people might write the letter A rather differently, but as long as they can mutually recognize each other’s marks as an “A,” they can successfully share a notation. Computers, however, cannot easily accommodate such variation. Each character must be strictly defined. In the case of ASCII, each character is given a number from 0 to 127, so that there are 128 ASCII characters. When using a computer to type ASCII characters, each key you press selects a character from this “menu” of 128 characters. A notation that has had numbers assigned to its characters is called a character encoding.

Table 9.1. ASCII

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NUL</td>
<td>DLE</td>
<td>space</td>
<td>0</td>
<td>@</td>
<td>P</td>
<td>`</td>
</tr>
<tr>
<td>1</td>
<td>SOH</td>
<td>DC1</td>
<td>!</td>
<td>1</td>
<td>A</td>
<td>Q</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>STX</td>
<td>DC2</td>
<td>&quot;</td>
<td>2</td>
<td>B</td>
<td>R</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>ETX</td>
<td>DC3</td>
<td>#</td>
<td>3</td>
<td>C</td>
<td>S</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>EOT</td>
<td>DC4</td>
<td>$</td>
<td>4</td>
<td>D</td>
<td>T</td>
<td>t</td>
</tr>
<tr>
<td>5</td>
<td>ENQ</td>
<td>NAK</td>
<td>%</td>
<td>5</td>
<td>E</td>
<td>U</td>
<td>e</td>
</tr>
<tr>
<td>6</td>
<td>ACK</td>
<td>SYN</td>
<td>&amp;</td>
<td>6</td>
<td>F</td>
<td>V</td>
<td>f</td>
</tr>
<tr>
<td>7</td>
<td>BEL</td>
<td>ETB</td>
<td>’</td>
<td>7</td>
<td>G</td>
<td>W</td>
<td>g</td>
</tr>
<tr>
<td>8</td>
<td>BS</td>
<td>CAN</td>
<td>(</td>
<td>8</td>
<td>H</td>
<td>X</td>
<td>h</td>
</tr>
<tr>
<td>9</td>
<td>HT</td>
<td>EM</td>
<td>)</td>
<td>9</td>
<td>I</td>
<td>Y</td>
<td>i</td>
</tr>
<tr>
<td>A</td>
<td>LF</td>
<td>SUB</td>
<td>*</td>
<td>:</td>
<td>J</td>
<td>Z</td>
<td>j</td>
</tr>
<tr>
<td>B</td>
<td>VT</td>
<td>ESC</td>
<td>+</td>
<td>;</td>
<td>K</td>
<td>[</td>
<td>k</td>
</tr>
<tr>
<td>C</td>
<td>FF</td>
<td>FS</td>
<td>,</td>
<td>&lt;</td>
<td>L</td>
<td>\</td>
<td>l</td>
</tr>
<tr>
<td>D</td>
<td>CR</td>
<td>GS</td>
<td>-</td>
<td>=</td>
<td>M</td>
<td>]</td>
<td>m</td>
</tr>
<tr>
<td>E</td>
<td>SO</td>
<td>RS</td>
<td>.</td>
<td>&gt;</td>
<td>N</td>
<td>^</td>
<td>n</td>
</tr>
<tr>
<td>F</td>
<td>SI</td>
<td>US</td>
<td>/</td>
<td>?</td>
<td>O</td>
<td>_</td>
<td>o</td>
</tr>
</tbody>
</table>

The most ambitious character coding in existence is Unicode, which as of version 6.0 assigns numbers to 109,449 characters. Unicode makes the important distinction between characters and glyphs. A character is the smallest meaningful unit of a written language. In alphabet-based languages like English, characters are letters; in languages like Chinese, characters are ideographs. Unicode treats all of these characters as abstract ideas (Latin capital A) rather than specific marks (A A A A A). A specific mark that can be used to depict a character is a glyph. A font is a collection of glyphs used to depict some set of characters. A Unicode font explicitly associates each glyph with a particular
number in the Unicode character encoding. The inability of computers to use contextual understanding to bridge the gap between various glyphs and the abstract character depicted by those glyphs turns out to have important consequences for organizing systems.

Different notations may include very similar marks. For example, modern music notation includes marks for indicating the pitch of note, known as accidentals. One of these music notation marks is ♯ (“sharp”). The sharp sign looks very much like the symbol used in English as an abbreviation for the word number, as in We’re #1! If you were to write a sharp sign and a number sign by hand, they would probably look identical. In a non-digital environment, we would rely on context to understand whether the written mark was being used as part of music notation, or mathematical notation, or as an English abbreviation.

Computers, however, have no such intuitive understanding of context. Unicode encodes the number sign and the sharp sign as two different characters. As far as a computer using Unicode is concerned, ♯ and # are completely different, and the fact that they have similar-looking glyphs is irrelevant. That is a problem if, for example, a cataloger has carefully described a piece of music by correctly using the sharp sign, but a person looking for that piece of music searches for descriptions using the number sign (since that is what you get when you press the keyboard button with the symbol that most closely resembles a sharp sign).

9.3.2 Writing Systems

A writing system employs one or more notations, and adds a set of rules for using them. Most writing systems assume knowledge of a particular human language. These writing systems are known as glottic writing systems. But there are many writing systems, such as mathematical and musical ones, that are not tied to human languages in this way. Many of the writing systems used for describing resources belong to this latter group, meaning that (at least in principle) they can be used with equal facility by speakers of any language.

Glottic writing systems, being grounded in natural human languages, are difficult to describe precisely and comprehensively. Non-glottic writing systems, on the other hand, can be described precisely and comprehensively using an abstract model. That is the connection between the structural perspective taken in the previous section, and the textual perspective taken in this section. A non-glottic writing system is described by a particular metamodel, and structures that fit within the constraints of a given metamodel can be textually represented using one or more writing systems that are described by that metamodel.

Some writing systems are closely identified with specific metamodels. For example, XML and JSON are both 1) metamodels for structuring information and 2)
writing systems for textually representing information. In other words, they specify both the abstract structure of a description and how to write it down. It is possible to conceive of other ways to textually represent the structure of these metamodels, but for each of these metamodels just one writing system has been standardized.\textsuperscript{555}\textsuperscript{[Com]}

RDF, on the other hand, is \textit{only} a metamodel, not a writing system. RDF only defines an abstract structure, not how to write that structure. So how do we write information that is structured as RDF? It turns out that we have many choices. Unlike XML and JSON, several different writing systems for the RDF metamodel have been standardized, including N-Triples, Turtle, RDFa, and RDF/XML.\textsuperscript{556}\textsuperscript{[Com]} Each of these is a writing system that is abstractly described by the RDF metamodel.

Writing systems provide rules for arranging characters from a notation into meaningful structures. A character in a notation has no inherent meaning. Characters in a notation only take on meaning in the context of a writing system that uses that notation. For example: what does the letter \textit{I} from the Latin alphabet mean? That question can only be answered by looking at how it is being used in a particular writing system. If the writing system is American English, then whether \textit{I} has a meaning depends on whether it is grouped with other letters or whether it stands alone. Only in the latter case does it have an assignable meaning. However in the arithmetic writing system of ancient Rome, which also uses as a notation the letters of the Latin alphabet, \textit{I} has a different meaning: \textit{one}.

This example also serves to illustrate how the ordering of a notation can differ from the ordering of a writing system that uses that notation. According to the ordering of the Latin alphabet, the twelfth letter \textit{L} comes before the twenty-second letter \textit{V}. But in the Roman numeric writing system, \textit{V} (the number 5) comes before \textit{L} (the number 50). Unless we know which ordering we are using, we cannot arrange \textit{L} and \textit{V} “in order.”\textsuperscript{557}\textsuperscript{[Ling]}

\textbf{Table 9.2. Roman Numerals}

<table>
<thead>
<tr>
<th>Roman Number</th>
<th>Arabic Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>5</td>
</tr>
<tr>
<td>X</td>
<td>10</td>
</tr>
<tr>
<td>L</td>
<td>50</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
</tr>
<tr>
<td>D</td>
<td>500</td>
</tr>
<tr>
<td>M</td>
<td>1000</td>
</tr>
</tbody>
</table>
This kind of difference in ordering can arise in more subtle ways as well. When we alphabetically order names, we first compare the first character of each name, and arrange them according to the ordering of the writing system. The first known use of alphabetical ordering was in the Library of Alexandria about two thousand years ago, when Zenodotus arranged the collection according to the first letter of resource names.\footnote{558\text{Ling}} If the first characters of two names are the same, we compare the second character, and so on. We can also apply this same kind of ordering procedure to sequences of numerals. If we do, then 334 will come before 67, because 3 (the first character of the first sequence) comes before 6 (the first character of the second sequence) according to the ordering of our notation (Arabic numerals). However, it is more common when ordering sequences of numerals to treat them as decimal numbers, and thus to use the ordering imposed by the decimal system. In the decimal writing system, 67 precedes 334, since the latter is a greater number.

This difference is important for organizing systems. Computers will sort values differently depending on whether they are treating sequences of numerals as numbers or just as sequences. Some organizing systems mix multiple ways of ordering the same characters. For example, Library of Congress call numbers have four parts, and sequences of Arabic numerals can appear in three of them. In the second part, indicating a narrow subject area, and fourth part, indicating year of publication, sequences of numerals are treated as numbers and ordered according to the decimal system. In the third part, however, sequences of numerals are treated as sequences and ordered “notationally” as in the example above (334 before 67).

Differences in ordering demonstrate just one way that multiple writing systems may use the same notation differently. For example, the American English and British English writing systems both use the same Latin alphabet, but impose slightly different spelling rules.\footnote{559\text{Ling}} The Japanese writing system employs a number of notations, including traditional Chinese characters (\textit{kanji}) as well as the Latin alphabet (\textit{rōmaji}). Often, writing systems do not share the same exact notation but have mostly overlapping notations. Many European languages, for example, extend the Latin alphabet with characters such as Å and Ü that add additional marks, known as diacritics, to the basic characters.\footnote{560\text{Com}}

In organizing systems it is often necessary to represent values from one writing system in another writing system that uses a different notation, a process known as\textit{ transliteration}. For example, early computer systems only supported the ASCII notation, so text from writing systems that extend the Latin alphabet had to be converted to ASCII, usually by removing (or sometimes transliterating) diacritics. This made the non-ASCII text usable in an ASCII-based computerized organizing system, at the expense of information loss.
Even in modern computer systems that support Unicode, however, transliteration is often needed to support organizing activities by users who cannot read text written using its original system. The Library of Congress and the American Library Association provide standard procedures for transliterating text from over sixty different writing systems into the (extended) Latin alphabet.

9.3.3 Syntax

The examples in Example 9.5, Basic ways of writing part of a book description, express the same information using different writing systems. The examples use the same notation (ASCII) but differ in their syntax: the rules that define how characters can be combined into words and how words can be combined into higher-level structures.\footnote{Com}

- Consider the first entry: The title is Die Ringe des Saturn and it has 371 pages. The leading capital letter and the period ending this sequence of characters indicate to us that this is a sentence. This sentence is one way we might use the English writing system to express two statements about the book we are describing. A statement is one distinct fact or piece of information. In glottic writing systems like English, there is usually more than one sentence we could write to express the same statement. For example, instead of it has 371 pages we might have written the number of pages is 371. English writing also enables us to construct complex sentences that express more than one statement.\footnote{Ling}

In contrast, when we create descriptions of resources in an organizing system, we generally use non-glottic writing systems in which each sentence only expresses a single statement, and there is just one way to write a sentence that expresses a given statement.\footnote{Com} These restrictions make these writing systems less expressive, but simplify their use. In particular, since there is a one-to-one correspondence between sentences and statements, we can drop the distinction and just talk about the statements of a description.

Now we return to our example and look at the structure of the statement, The title is Die Ringe des Saturn and it has 371 pages. Spaces are used to separate the text into words, and English syntax defines the functions of those words. The verb is in this statement functions to link the word title to the phrase Die Ringe des Saturn. This is typical of the kind of statements found in a resource description. Each statement identifies and describes some aspect of the resource. In this case, the statement attributes the value Die Ringe des Saturn to the property title.

As we saw when we looked at description structures, we can analyze descriptions as involving properties of resources and their corresponding values or content. In a writing system like English, it is not always so straightforward to determine which words refer to properties and which refer to val-
ues. (This is why blobs are not ideal description structures.) Writing systems designed for expressing resource descriptions, on the other hand, usually define syntax that makes this determination easier. In our dictionary examples above, we used an arrow character \( \rightarrow \) to indicate the relationship between properties and values.

This ease of distinguishing properties and values comes at a price, however. The syntax of English is forgiving: we can read a sentence with somewhat garbled syntax such as \( 371 \text{ pages it has} \) and often still make out its meaning. This is usually not the case with writing systems intended for expressing resource descriptions. These systems strictly define their rules for how characters can be combined into higher-level structures. Structures that follow the rules are **well formed** according to that system.

- Take for example the second entry in Example 9.5, **Basic ways of writing part of a book description**.

  
  
  
  
  { book: {"title":"Die Ringe des Saturn","pages":371} }

  
  
  
  This fragment is written in JSON. As explained earlier in this chapter, JSON is a metamodel for structuring information using lists and dictionaries. But JSON is also a writing system, which borrows its syntax from JavaScript. The JSON syntax uses brackets to textually represent lists \([1,2,3]\) and braces to textually represent dictionaries \({\text{title}:"Die Ringe des Saturn","pages":371}\). Within braces, the colon character : is used to link properties with their values, much as was used in the previous example. So "pages":371 is a statement assigning the value 371 to the property pages.

- The third fragment is written in **XML**.

  
  
  
  

  
  
  
  Like JSON, XML is a metamodel and also a writing system. Here we have XML elements and attributes. XML elements are textually represented as *tags* that are marked using the special characters <, > and /. So, this fragment of XML consists of a book element with a child element, title, and a pages attribute, each of which has some text content. In this case, pages="371" is a statement assigning the value 371 to the property pages. The difference is syntax is subtle; quotation marks surround the value and equal sign = is used to assign the property to its value.

- The fourth is a fragment of **HTML**.

  
  
  
  
  <div class="book">The title is <span class="title">Die Ringe des Saturn</span> and it has <span class="pages">371 pages.</span></div>

  
  
  
  
  498 Chapter 9 — The Forms of Resource Descriptions
The writing system that HTML employs is close enough to XML to ignore any differences in syntax. In this example, the CLASS attribute contains the property name and the property value is the element content.

- The fifth entry is a fragment of Turtle, one of the writing systems for RDF.

```
<http://lccn.loc.gov/96103072>
<http://rdvocab.info/Elements/title> "Die Ringe des Saturn"@de ;
```

Turtle provides a syntax for writing down RDF *triples*. Each triple consists of a subject, predicate, and object separated by spaces. Recall that RDF uses URIs to identify subjects, predicates, and some objects; these URIs are written in Turtle by enclosing them in angle brackets `< >`. Triples are separated by period `.` characters, but triples that share the same subject can be written more compactly by writing the subject only once, and then writing the predicate and object of each triple, separated by a semicolon `;` character. This is what we see in Example 9.2, Nesting an author description within a book description: two triples that share a subject.

The two fragments in Example 9.6, Writing part of a book description in Semantic XML, demonstrate namespaces, terms from the Dublin Core and DocBook namespaces, and the facility with which XML embraces semantic encoding of description resources.

**Example 9.6. Writing part of a book description in Semantic XML.**

```
  <dc:title>Die Ringe des Saturn</title>
  ...
</book>

  <bookinfo>
    <title>Die Ringe des Saturn</title>
    <pagenums>371 p.</pagenums>...
  </bookinfo>
  ...
</book>
```

- The first example extends the third fragment from Example 9.5, Basic ways of writing part of a book description.; the xmlns:dc="..." segment is a namespace declaration, which is associating dc with the quoted URI, which happens to be the Dublin Core Metadata Initiative (DCMI); the child `<dc:title>` element and the attached dc:extent="371" tell us that the corresponding values are attributable to the title and extent properties, respectively, from the Dublin Core namespace.

- The next fragment employs DocBook DTD namespace; we now have a `<pagenums>` element for which the meaning is contextually obvious; the title is still
Microformats, RDFa and Microdata

When Tim Berners-Lee deployed HTML, its syntax contained the basic elements and attributes needed to make formal statements about the document as a whole by using `<LINK/>`, or about specific parts of the document by using the `<A>` element. Each of these elements have four attributes in common: the famous HREF attribute contains a URI that names an object resource; the NAME attribute allows the element to be the target end of a link; the REL and REV attributes contain descriptions of the link relations.

Microformats, RDFa and Microdata are the latest generation of metadata extensions to HTML. Each approach is widely used on the web and by search engines. As such, they are potential targets when transforming into HTML from richer semantic formats.

Microformats are the simplest of the three. It uses controlled vocabularies of terms in REL/REV, and in the CLASS attribute, to declare high-level information types.

RDFa is RDF in Attributes. That is, RDFa is a formal specification for writing RDF expressions by using attributes in XML and HTML documents. It uses an ABOUT attribute to name the subject of the relation; the REL and REV attributes; HREF is joined by SRC and RESOURCE to name the object of the link; a TYPEOF attribute declares a type; PROPERTY and CONTENT attributes are used to attribute a value to an object’s property.

Microdata is similar, inasmuch as it uses attributes extensively. The presence of an ITEMSCPE attribute identifies an item while the ITEMTYPE attribute value identifies its type; ITEMID declares an item’s name or unique identifier; ITEMPROP is a name value pair, and; ITEMREF relates this item to other elements that are outside of the scope of the container element.

The two fragments in Example 9.7, Writing part of a book description in RDFa or microdata, demonstrate RDFa and microdata formats, which each rely upon specific attributes to establish the type of the property values contained by the HTML elements. In each example, the book title is contained by a `<span>` element. Whereas RDFa relies upon the property attribute, the microdata example employs the itemprop attribute to specify that the contents of the element is, effectively, a “title” in exactly the same sense as we know that the contents of `<dc:title>` is a “title.”
9.4 Worlds of Description

In the previous two sections we have considered descriptions as designed objects with particular structures and as written documents with particular syntaxes. As we have seen, there are many possible choices of structure and syntax. But these choices are never made in isolation. Just as an architect or designer must work within the constraints of the existing built environment, and just as any author must work with existing writing systems, descriptions are always created as part of a pre-existing “world” over which any one of us has little control.

In the final part of this chapter, we will consider how choices of structure and syntax have converged historically into broad patterns of usage. For lack of a better term, we call these broad patterns “worlds.” “World” is not a technical term and should not be taken too literally: the broad areas of application sketched here have considerable overlap, and there are many other ways one might identify patterns of description structure and syntax. That said, the three worlds described here do reflect real patterns of description form that influence tool and technology choices. In your own work creating and managing resource descriptions, it is likely that you will need to think about how your descriptions fit into one or more of these worlds.

9.4.1 The Document Processing World

The first world we will consider is concerned primarily with the creation, processing and management of hybrid narrative-transactional documents such as instruction manuals, textbooks, or annotated medieval manuscripts. (See The Document Type Spectrum (page 182)). These are quite different kinds of documents, but they all contain a mixture of narrative text and structured data, and they all can be usefully modeled as tree structures. Because of these shared qualities, tools as different as publishing software, supply-chain management software, and scholarly editing software have all converged on common XML-based solutions. (“The XML world” would be another appropriate name for the document-processing world.)
This convergence was no accident, because XML was designed specifically to address the problem of how to add structure and data to documents by “marking them up.” XML is the descendant of Standard Generalized Markup Language (SGML), which in turn descended from International Business Machines (IBM)’s Generalized Markup Language, which was invented to enable the production and management of large-scale technical documentation. The explicitness of markup makes it well-suited for representing structure and content type distinctions in institutional contexts, where the scope, scale, and expected lifetime of organizing systems for information implies reuse by unknown people for unanticipated purposes.

The abstract data model underlying XML is called the XML Information Set or Infoset. The Infoset defines a document as a partially ordered tree of “information items.” Every XML document can thus be understood as a specific kind of tree, although not every tree structure is expressible as an XML document. As we discussed in Inclusions and References (page 486), XML has the ability to describe graphs by incorporating the use of ID and IDREF attribute types to create references among element information items within the same document. This modest form of hypertext linking allows us to present the following document fragment that approximates the graph we saw modeled in Figure 9.4, Descriptions Linked into a Graph.

**Example 9.8. XML implementation of a biblio-graph**

```
<person id="WG.Sebald">Winfried George Sebald</person>
<person id="MR.McCulloch">Mark Richard McCulloch</person>

<book>
  <title>Understanding W.G. Sebald</title>
  <subject idref="WG.Sebald"/>
  <author idref="WG.Sebald"/>
  <author idref="MR.McCulloch"/>
</book>

<book pages="371">
  <title lang="de">Die Ringe des Saturne</title>
  <title lang="en">The Rings of Saturn</title>
  <author idref="WG.Sebald"/>
</book>

<book pages="416">
  <title lang="de">Austerlitz</title>
  <author idref="WG.Sebald"/>
</book>
```
As one might expect, tools and technologies in the document-processing world are optimized for manipulating and combining tree structures. A “toolchain” is set of tools intended to be used together to achieve some goal.

**The XML Toolchain**

The XML toolchain is quite comprehensive. It consists of tools for creating XML documents (XML editors), tools for expressing logical document and data models (DTD, XML Schema, *REgular Language for XML Next Generation (RELAX NG)*, Schematron), tools for transforming XML documents (XSLT), tools for describing document processing “pipelines” (*XProc: An XML Pipeline Language*), and tools for storing and querying collections of XML documents (XML databases, queried using XML Query Language (XQuery)). Used together, these tools provide very powerful means of working with tree-structured documents. XML editors incorporate knowledge of DTDs, schemas, transformations, style sheets, queries, databases and pipelines. Pipelines choreograph the plumbing and inter-dependencies involved in processing a complex dataset and publishing a useful result in one or more output formats.

For programmers who do not to use the XML toolchain, other programming languages also provide libraries for working with XML. This fact has led some to propose, and others to believe, that XML is a kind of universal format for exchanging data among systems. However, programmers have observed that a random XML Infoset does not map easily to the data structures commonly found in many programming languages. “Working with XML” frequently means translating from XML tree structures to data structures native to another language, usually meaning lists and dictionaries. This translation can be problematic and often means giving up many of the strengths of XML. By the same token, there are decades more practical experience working with markup languages and institutional publishing than there is with JSON and RDF.

XML is not a universal solution for every possible problem. That does not mean that it is not the best solution for a wide variety of problems, including yours. To gauge whether your resource descriptions are, or ought to be, part of the document-processing world, ask yourself the following questions:

- Do my resource descriptions contain mixtures of narrative text, hypertext, structured data and a variety of media formats?
- Can my descriptions easily be modeled using tree structures, hypertext links, and transclusion?
- Are the vocabularies I need or want to use made available using XML technologies?
• Do I need to work with a body of existing descriptions already encoded as XML?
• Do I need to interoperate with processes or partners that utilize the XML toolchain?
• Do I need to publish my resource descriptions in multiple formats from a single source?

If the answer to one or more of these questions is “yes,” then chances are good that you are working within the document processing world, and you will need to become familiar with conceptualizing your descriptions as trees and working with them using XML tools.

9.4.2 The Web World

The second “world” emerged in the early 1990s with the creation of the World Wide Web. The web was developed to address a need for simple and rapid sharing of scientific data. Of course, it has grown far beyond that initial use case, and is now a ubiquitous infrastructure for all varieties of information and communication services. (“The browser world” would be another appropriate name for what we are calling the Web World.)

Documents, data, and services on the web are conceptualized as resources, identified using Uniform Resource Identifiers (URI), and accessible through representations transferred via Hypertext Transfer Protocol (HTTP). Representations are sequences of bytes, and could be HTML pages, JPEG images, tabular data, or practically anything else transferable via HTTP. No matter what they are, representations transferred over the web include descriptions of themselves. These descriptions take the form of property-value pairs, known as “HTTP headers.” The HTTP headers of web representations are structured as dictionaries.

Dictionary structures appear many other places in web infrastructure. URIs may include a query component beginning with a ? character. This component is used for purposes such as providing query parameters to search services. The query component is commonly structured as a dictionary, consisting of a series of property-value pairs separated by the & character. For example, the following URI:

https://www.google.com/search?q=sebald&tbs=qdr:m

includes the query component q=sebald&tbs=qdr:m. This is a dictionary with the properties q and tbs, respectively specifying the search term and temporal constraints on the search.

Data entered into an HTML form is also structured as a dictionary. When an HTML form is submitted, the entered data is used either to compose the query
component of a URI, or to create a new representation to be transferred to a web server. In either case, the data is structured as a set of properties and their corresponding values.

HTML documents are structured as trees, but descriptions embedded within HTML documents can also be structured as dictionaries. HTML documents may include a dictionary of metadata elements, each of which specifies a property and its value. Recently support for microdata was added to HTML, which is another method of adding dictionaries of property-value pairs to documents. Using microdata, authors can annotate web content with additional information, making it easier to automatically extract structured descriptions of that content. Microformats are another method for doing this by mapping existing HTML attributes and values to (nested) dictionary structures.

Dictionary structures are easy to work with in any programming language, and they pervade various popular frameworks for programming the Web. In the programming languages used to implement web services, HTTP headers and query parameters are easily mapped to dictionary data structures native to those languages. On the client side, there is only one programming language that runs within all web browsers: JavaScript. The dictionary is the fundamental data structure within JavaScript as well.

Thus it is unsurprising that JSON, a dictionary-structured, JavaScript-based syntax, has become the de facto standard for application-to-application interchange of data on the web in contexts that do not involve business transactions. Web services providing structured data intended for programmatic use can make that data available as JSON, which is well-suited for use either by JavaScript programs running within browsers, or by programs written in other languages running outside of browsers (e.g., smart phone applications).

It is now commonly accepted that there are useful differences of approach between the document-processing world and the Web World. This does not mean that the two worlds do not have significant overlaps. Some very important web representation types are XML-based, such as the Atom syndication format. Trees will continue to be the structure of choice for web representations that consist primarily of narrative rather than transactional data. But for structured descriptions that are intended to be accessed and manipulated on the Web, dictionary structures currently rule.

To gauge whether your resource descriptions are or ought to be part of the Web world, ask yourself the following questions:

- Is the web the primary platform upon which I will be making my descriptions available?
- Are my resource descriptions primarily structured, transaction-oriented data?
• Can my descriptions easily be modeled as lists of properties and values (dictionaries)?
• Are the vocabularies I need or want to use made available primarily using HTML technologies such as microdata or microformats?
• Do I need to make my descriptions easily usable for use within a wide array of programming languages?

If the answer to one or more of these questions is “yes,” then chances are good that you are working within the Web World, and you will need to become familiar with conceptualizing your descriptions as dictionaries and working with them using programming languages such as JavaScript.

9.4.3 The Semantic Web World

The last world we consider is still somewhat of a possible world, at least in comparison with the previous two. While the document processing world and the web world are well-established, the Semantic Web world is only starting to emerge, despite having been envisioned over a decade ago.

The vision of a Semantic Web world builds upon the web world, but adds some further prescriptions and constraints for how to structure descriptions. The Semantic Web world unifies the concept of a resource as it has been developed in this book, with the web notion of a resource as anything with a URI. On the Semantic Web, anything being described must have a URI. Furthermore, the descriptions must be structured as graphs, adhering to the RDF metamodel and relating resources to one another via their URIs. Advocates of Linked Data further prescribe that those descriptions must be made available as representations transferred over HTTP.569[Web]

This is a departure from the web world. The web world is also structured around URIs, but it does not require that every resource being described have a URI. For example, in the web world a list of bibliographic descriptions of books by W.G. Sebald might be published at a specific URI, but the individual books themselves might not have URIs. In the Semantic Web world, in addition to the list having a URIs, each book would have a URI too, in addition to whatever other identifiers it might have.570[Web]

Making an HTTP request to an individual book URI may return a graph-structured description of that book, if best practices for Linked Data are being followed. This, too, is a departure from the web world, which is agnostic about the form representations or descriptions of resources should take (although as we have seen, dictionary structures are often favored on the web when the clients consuming those descriptions are computer programs). On the Semantic Web, all descriptions are structured as RDF graphs. Each description graph links to other description graphs by referring to these related resources using
their URIs. Thus, at least in theory, all description graphs on the Semantic Web are linked into a single massive graph structure. In practice, however, it is far from clear that this is an achievable, or even a desirable, goal.

Although the Semantic Web is in its infancy, a significant number of resource descriptions have already been made available in accordance with the principles outlined above. Descriptions published according to these principles are often referred to as “Linked Data.” Prominent examples include: DBpedia, a graph of descriptions of subjects of Wikipedia articles; the Virtual International Authority File (VIAF), a graph of descriptions of names collected from various national libraries’ name authority files; GeoNames, a graph of descriptions of places; and Data.gov.uk, a graph of descriptions of public data made available by the UK government.571[Web]

Despite the growing amount of Linked Data, tools for working with graph-structured data are still immature in comparison to the XML toolchain and Web programming languages. Although there is an XML syntax for RDF, using the XML toolchain to work with graph-structured data is generally a bad idea. And just as most programming languages do not support natively working with tree structures, most do not support natively working with graph structures either. Storing and querying graph-structured data efficiently requires a graph database or triple store.

Still, the Semantic Web world has much to recommend it. Having a common way of identifying resources (the URI) and a single shared metamodel (RDF) for all resource descriptions makes it much easier to combine descriptions from different sources. To gauge whether your resource descriptions are or ought to be part of the Semantic Web world, ask yourself the following questions:

- Is the web the primary platform upon which I will be making my descriptions available?
- Is it important that I be able to easily and freely aggregate the elements of my descriptions in different ways and to combine them with descriptions created by others?
- Are my descriptions best modeled as graph structures?
- Have the vocabularies I need or want to use been created using RDF?
- Do I need to work with a body of existing descriptions that have been published as Linked Data?

If the answer to one or more of these questions is “yes,” then chances are good that you should be working within the Semantic Web world, and you ought to become familiar with conceptualizing your descriptions as graphs and working with them using Semantic Web tools.
9.5 Key Points in Chapter Nine

- We can approach the problem of how to form resource descriptions from two perspectives: structuring and writing. (See §9.1 Introduction (page 467))
- Metamodels describe structures commonly found in resource descriptions and other information resources, regardless of the specific domain. (See §9.2 Structuring Descriptions (page 469))
- Blobs, sets, lists, dictionaries, trees, and graphs are all kinds of structures that can be used to form resource descriptions. (See §9.2.1 Kinds of Structures (page 472))
- A list, like a set, is a collection of items with an additional constraint: their items are ordered. (See §9.2.1.3 Lists (page 473))
- A dictionary, also known as a map or an associative array, is a set of property-value pairs or entries. (See §9.2.1.4 Dictionaries (page 474))
- Nested dictionaries form a tree. (See §9.2.1.4 Dictionaries (page 474))
- Trees consist of nodes joined by edges. (See §9.2.1.5 Trees (page 475))
- JSON consists of two kinds of structures: lists (called arrays in JavaScript) and dictionaries (called objects in JavaScript). (See §9.2.2.1 JSON (page 481))
- The XML Infoset is a tree structure, where each node of the tree is defined to be an information item of a particular type. (See §9.2.2.2 XML Information Set (page 481))
- Using schemas to define data representation formats is a good practice that facilitates shared understanding and contributes to long-term maintainability. (See §9.2.2.2 XML Information Set (page 481))
- The RDF metamodel is a directed graph, so it identifies one node (the one from which the edge is pointing) as the subject of the triple, and the other node (the one to which the edge is pointing) as its object. The edge is referred to as the predicate or (as we have been saying) property of the triple. (See §9.2.2.3 RDF (page 483))
• An “encoding scheme” is a specialized writing system or syntax for particular types of values. Encoding schemes specify how to textually represent information.
(See §9.3.1 Notations (page 492))

• A writing system employs notations, and adds a set of rules for using them.
(See §9.3.2 Writing Systems (page 494))

• Differences in ordering demonstrate just one way that multiple writing systems may use the same notation differently.
(See §9.3.2 Writing Systems (page 494))

• Syntax is the rules that define how characters can be combined into words and how words can be combined into higher-level structures.
(See §9.3.3 Syntax (page 497))

• The document processing world is concerned primarily with the creation, processing and management of hybrid narrative-transactional documents.
(See §9.4.1 The Document Processing World (page 501))

• In the web world, documents, data, and services are conceptualized as resources, identified using Uniform Resource Identifiers (URI), and accessible through representations transferred via the Hypertext Transfer Protocol (HTTP).
(See §9.4.2 The Web World (page 504))

• The Semantic Web world unifies the concept of a resource as it has been developed in this book, with the web notion of a resource as anything with a URI. Descriptions must be structured as graphs, adhering to the RDF metamodel and relating resources to one another via their URIs.
(See §9.4.3 The Semantic Web World (page 506))

Endnotes for Chapter 9

[516][Com] This discussion of Batten’s cards is based on (Lancaster 1968, pages 28-32). Batten’s own explanation is in (Batten 1951).


[518][Ling] The technique of diagramming sentences was invented in the mid-19th century by Stephen W. Clark, a New York schoolmaster; (Clark2010) is an exact reprinting of a nearly 100 year old edition of his book A Practical Grammar. A recent tribute to Clark is (Florey 2012).
It is easy to underestimate the incredible power of the human perceptual and cognitive systems to apply neural computation and knowledge to enable vision and hearing to seem automatic. Computers are getting better at extracting features from visual and auditory signals to identify and classify inputs, but our point here is that none of these features are explicitly represented in the input “blob” or “stream.”

As we commented earlier, an oral description of a resource may not be especially useful in an organizing system because computers cannot easily understand it. On the other hand, there are many contexts in which an oral description would be especially useful, such as in a guided tour of a museum where visitors can use audio headsets.

What was lost was the previously invisible structure provided by the grammar, which made us assign roles to each of these terms to create a semantic interpretation.

It is rarely practical to make things as simple as possible. According to Einstein, we should endeavor to “Make everything as simple as possible, but not simpler.”

This structural metamodel only allows one value for each property, which means it would not work for books with multiple authors or that discuss multiple subjects.

Going the other direction is not so easy, however: just as real dictionaries do not support finding a word given a definition, neither do dictionary structures support finding a key given a value.

The XML Information Set (Cowan2004) RDF/XML is one example where meta models meet. In Document Design Matters, (Wilde and Glushko 2008b) point out that “If the designer of an exchange format uses a non-XML conceptual metamodel because it seems to be a better fit for the data model, XML is only used as the physical layer for the exchange model. The logical layer in this case defines the mapping between the non-XML conceptual model, and any reconstruction of the exchange model data requires the consumer to be fully aware of this mapping. In such a case, it is good practice to make users of the API aware of the fact that it is using a non-XML metamodel. Otherwise they might be tempted to base their implementation on a too small set of examples, creating implementations which are brittle and will fail at some point in time.”

Technically, what is described here is referred to as “rooted tree” by mathematicians, who define trees more generally. Since trees used as data structures are always rooted trees, we do not make the distinction here.
This feature relies upon the existence of an XML schema. An XML schema can declare that certain attributes are of type ID, IDREF or IDREFS. Whether an XML DTD or one of the many schema languages that have been developed under the auspices of the W3C or ISO.

http://www.w3.org/TR/xml-infoset/.

The XML Infoset is one of many metamodels for XML, including the DOM and XPath. Typically, an XML Infoset is created as a by-product of parsing a well-formed XML document instance. An XML document may also be informed by its DTD or schema with information about the types of attribute values, and their default values. Attributes of type ID, IDREF and IDREFS provide a mechanism for intra-document hypertext linking and transclusion. An XML document instance may contain entity definitions and references that get expanded when the document is parsed, thereby offering another form of transclusion.

A well-formed XML document instance, when processed, will yield an XML Information Set, as described here. Information sets may also be constructed by other means, such as transforming from another information set. See the section on Synthetic Infosets at http://www.w3.org/TR/xml-infoset/#intro.synthetic for details.

The Infoset contains knowledge of whether all related declarations have been read and processed, the base URI of the document instance, information about attribute types, comments, processing instructions, unparsed entities and notations, and more.

A well-formed XML document instance for which there are associated schemas, such as a DTD, may contribute information to the Infoset. Notably, schemas may associate data types with element and attribute information items, and it may also specify default or fixed values for attributes. A DTD may define entities that are referenced in the document instance and are expanded in-place when processed. These contributions can affect the truth value of the document.

The SGML standard explicitly stated that documentation describing or explaining a DTD is part of the document type definition. The implication being that a schema is not just about defining syntax, but also semantics. Moreover, since DTDs do not make possible to describe all possible constraints, such as co-occurrence constraints, the documentation could serve as human-consumable guidance for implementers as well as content creators and consumers.

Attribute types may be declared in an XML DTD or schema. Attributes whose type is ID must have a valid XML name value that is unique within that XML document; an attribute of type IDREF whose value corresponds to a unique ID has a “references” property whose value is the element node that corresponds to the element with that ID. An attribute of type IDREFS whose value
corresponds to a list of unique ID has a “references” property whose value is a list of element node(s) that corresponds to the element(s) with matching IDs.

XML Inclusions (XInclude) is (Marsh, Orchard, and Veillard 2006).

XML Linking Language (XLink) is (DeRose, Maler, Orchard, and Walsh 2010).

Within the document’s DTD, one simply declares the entity and its corresponding value, which could be anything from an entire document to a phrase and then it may be referenced in place within the XML document instance. The entity reference is replaced by the entity value in the XML Infoset. Entities, as nameable wrappers, effectively disappear on their way into the XML Infoset.


Do not take on the task of creating a new XML model lightly. Literally thousands of XML vocabularies have been created, and some represent hundreds or thousands of hours of effort. See (Bray 2005) for advice on how to reduce the risk of vocabulary design if you cannot find an existing one that satisfies your requirements.

See (Glushko and McGrath 2005) for a synthesis of best practices for creating domain-specific languages in technical publishing and business-to-business document exchange contexts. You need best practices for big problems, while small ones can be attacked with ad hoc methods.

Unless an XML instance is associated with a schema, it is fair to say that it does not have any model at all because there is no way to understand the content and structure of the information it contains. The assignment of a schema to an XML instance requires a “Document Type Declaration.” If some of the same vocabulary terms occur in more than one XML schema, with different meanings in each, using elements from more than one schema in the same instance requires that they be distinguished using namespaces. For example, if an element named “title” means the “title of the book” in one schema and “the honorific associated with a person” in another, instances might have elements with namespace prefixes like <book:title>The Discipline of Organizing</book:title> and <hon:title>Professor</hon:title>. Namespaces are a common source of frustration in XML, because they seem like an overly complicated solution to a simple problem. But in addition to avoiding naming collisions, they are important in schema composition and organization.

What “correctly” means depends on the schema language used to encode the conceptual model of the document type. The XML family of standards includes several schema languages that differ in how completely they can en-
code a document type’s conceptual model. The Document Type Definition (DTD) has its origins in publishing and enforces structural constraints well; it expresses strong data typing through associated documentation resources. XML Schema Definition Language (XSD) is better for representing transactional document types but its added expressive power tends to make it more complex.


[543][Com] Attribute values can be constrained in a schema by specifying a data type, a default value, and a list of potential values. Data types allow us to specify whether a value is supposed to be a name, a number, a date, a token or a string of text. Having established the data type, we can further constrain the value of an attribute by specifying a range of values, for a number or a date, for example. We can also use regular expression patterns to describe a data type such as a postal code, telephone number or ISBN number. Specifying default values and lists of legal values for attributes simplifies content creation and quality assurance processes. In Schematron, a rule-based XML schema language for making test assertions about XML documents, we can express constraints between elements and attributes in ways that other XML schema languages cannot. For example, we can express the constraint that if two <title> elements are provided, then each must contain a unique string value and different language attribute values.


[546][Com] There is no single authority on the subject of regular expressions or their syntax. A good starting point is the Wikipedia article on the subject: http://en.wikipedia.org/wiki/Regular_expression.

[547][Com] The terminology here and in the following sections comes from (Harris 1996).


[549][CogSci] Entitled “The ABC,” the song was copyrighted in 1835 by Boston music publisher Charles Bradlee. It is sung to a tune that was originally developed by Wolfgang Amadeus Mozart, and is commonly recognizable as Twinkle, Twinkle, Little Star.

Only 95 of these characters are actually “marks” in the sense of being visible and printable. The other 33 ASCII characters are “control codes” that indicate things like horizontal and vertical tabs, the ends of printed lines, form feeds, and transmission control. We can think of many of these as special auxiliary marks, similar to the kind of symbols editors and proofreaders use to annotate texts.

The Unicode standard is maintained by a global non-profit organization. Everything you need to know is at http://www.unicode.org/.

The Chinese character 井 (water well) looks like the # character too. The # symbol was historically used to denote pounds, the Imperial unit of weight, as in 10# of potatoes. In the United Kingdom, the # character is called “hash.” We could go on, but we will leave it to you to discover more.

To add to the confusion, while the American standard (ASCII) places the # character at position 23, the British equivalent (BS 4730) places the currency symbol £ at the same position. As a result, improperly configured computers sometimes display # in place of £ and vice versa.

Recently, an alternative writing system for XML-structured data has been standardized: Efficient XML Interchange (EXI). However it is not yet widely used.

RDF/XML is a bit confusing; it is a writing system that uses XML syntax to textually represent RDF structure. This means that while XML tools can read and write RDF/XML, they cannot manipulate the graph structures it represents, because they were designed to work with XML’s tree structures.

Although we use alphabetic characters today to represent Roman numerals, originally they were represented by unique symbols.

It took a few hundred years before alphabetization became recursive and applied to letters other than the first (Casson 2002, p. 37). Alphabetization relies on the ordering of the writing system, not the notation. For example, Swedish and German are two writing systems that assign different orderings to the same notation.

For example, the American spelling of the words “center” and “color” contrasts slightly with the English spelling of “centre” and “colour.” There are too many examples to include here. Wikipedia has a comprehensive analysis of American and British spelling differences at http://en.wikipedia.org/wiki/American_and_British_English_spelling_differences.

ASCII’s 128 characters are insufficient to represent these more complex character sets, so a new family of character encodings was created, ISO-8859, in which each encoding enumerates 256 characters. Each encoding thus has more space to accommodate the additional characters of regionally-specific no-
tations. ISO 8859-5, for example, has extensions to support the Cyrillic alphabet.

[561][Com] In discussions of glottic writing systems, “syntax” usually refers only to the rules for combining words into sentences. In discussions of programming languages, “syntax” has the broader sense we use here.

[562][Ling] Compound sentences contain two independent clauses joined by a conjunction, such as “and,” “or,” “nor,” “but.” For example: I went to the store and I bought a book.” Complex sentences contain an independent clause joined by one or more dependent clauses. For example: “I read the book that I bought at the store.”

[563][Com] In truth, even non-glottic writing systems designed to encode resource descriptions unambiguously can have variant forms of the same statement. For example, XML permits some variation in the way the same Infoset may be textually represented. Often these variations involve the treatment of content that may under some circumstances be treated as optional, such as white space. The difference is that in writing systems designed for resource description, these variations can be precisely enumerated and rules developed to reconcile them, while this is not generally true for glottic writing systems.

[564][Ling] Fortunately for Yoda. There are many web services for converting English to Yoda-speak; an example is http://www.yodaspeak.co.uk/.

[565][Com] DocBook (Walsh 2010) is widely used to publish academic, commercial, industrial book, scientific, and computing book, papers and articles. The book that you are reading is encoded with DocBook markup; complete bibliographic information for the book is contained within the source files, ready to be extracted on the way into one of the latest ebook formats.

[566][Com] It should be noted that the content of the Infoset for a given document may be affected by knowledge of any related DTDs or schemas. That is to say that, upon examination of a given XML document instance, its Infoset may be augmented with some useful information, such as default attribute values and attribute types. (See Inclusions and References (page 486).)

[567][Web] Microdata is an invention of WHATWG and exists and part of what they call a “living standard.” It was supported by Google, so it was widely used and there exist numerous controlled vocabularies, including those for creative works, persons, events and organizations. Support for microdata has since been withdrawn from Apple Safari and Google Chrome browsers.

[568][Web] Microformats is a non-standard that emerged from the community and has been sponsored by CommerceNet and Microformats.org.

[569][Web] (Bizer, Heath, and Berners-Lee 2009).
It is worth noting that URIs are not required to have anything at their endpoints. Resolvability of URIs is evangelized as a best practice for Linked Data but not a requirement within the broader Semantic Web paradigm. Merely asserting that a URI is associated with a book is enough. If the URI can return a description or a resource, so much the better, but if not, at least you can talk about the book by referring to the same URI.

Many more available datasets are listed at linkeddata.org.