Chapter 2

Metadata Models of the World Wide Web

Abstract

The Semantic Web, in standards being developed by the World Wide Web Consortium, is a new way of defining metadata for use and reuse in a networked environment. In this chapter of "RDA Vocabularies for a Twenty-First-Century Data Environment," we'll discuss the definition of metadata and how it involves the creation of domain models (the things and relationships that the metadata will describe) and ontologies (the vocabularies that the metadata will use). The use of standard identifiers, called Uniform Resource Identifiers, creates unambiguous identities for data and statements about data.

The World Wide Web was developed as a web of documents. On this Web, digital documents would link to each other directly, allowing the user to follow the pointers provided by the author from a place in one document to another digital document. In hindsight, it seems obvious that while this ability to navigate the hyperlinks provided is extraordinarily powerful (and achieved something that is not possible in the analog world), the model lacked a key component for discovery, and that is meaningful metadata for the documents themselves. This problem has been partially overcome by the development of search engines that can index the actual text of the documents. Keyword indexing on uncontrolled text, however, lacks precision for searching.

The *Semantic Web* is a result of the realization that there is information in the documents on the Web that could be extremely valuable if it could be made actionable—that is, if there were a way to interact with the information inside documents, not just the documents themselves.[1] The emphasis of the Semantic Web is on topical information within the Web resources: information about persons, places, things, events, and covering the full range of scientific and humanistic thought. To turn the web of documents into a web of data, the Web needs metadata to represent that information. This metadata will not look like standard bibliographic metadata. Bibliographic metadata represents a document or resource. The purpose of the Semantic Web is not to create metadata that represents documents or resources; it is to create metadata for the informational content of those resources.

While Web documents resemble the granularity of articles more than that of books, there is significant overlap in the topics covered by the Web and by libraries. Yet these remain two separate and distinct information spheres. In part this is because libraries hold primarily physical resources. Yet where libraries and the Web could collaborate through an intermingling of digital resources, they are unable to because they use different technologies. The Web relies entirely on search engines and keyword searches, while libraries create metadata in a library-specific record format (MARC) that is stored in closed databases. The development of metadata solutions that are compatible with Web-based technology and can be used both by libraries and on the open Web creates the possibility of making a connection between the two worlds.

In relation to libraries, the Web community is quite late in realizing the importance of metadata. There may have been an advantage to starting to think about metadata for the first time in a fully automated environment. The Semantic Web community began with a kind of metadata *tabula rasa* and a natural tendency to think about machine processing of data at a deep level. Its work began with a study of the basic nature of metadata, or at least the very nature of machine-actionable, networked, interoperable metadata.

Similar to the development of the underlying standards that make the Internet possible, such as TCP/IP, the Semantic Web developers sought to develop the basic structure on which all other metadata would be developed. This basic structure is called the Resource Description Framework, or RDF. RDF itself relies on the Uniform Resource Identifier, the standard identifier format for the Web, and eXtensible Markup Language (XML), a set of rules for encoding documents and data electronically. These form the bottom layer of the "layer cake" of Semantic Web standards (see figure 1).

Resource Description Framework (RDF): W3C Semantic Web Activity
www.w3.org/RDF

Ontologies

To participate in the Semantic Web, a community needs to define an ontology. An ontology, in the sense used by the developers of the Semantic Web, defines the metadata for a particular slice of the knowledge universe. That slice is called a *domain*. Ontologies include a conceptualization of the elements of the domain and the relationships between those elements. The elements, called *entities* in Semantic Web parlance, can be things or concepts. The expression of the ontology creates a controlled vocabulary for describing entities in the domain. The goal is a rigorous knowledge base that can be subjected to computation. Ontologies differ from traditional thesauri and taxonomies primarily in being designed specifically for machine processing, as well as their use of a large variety of relationships between the entities in the defined domain.

The use of relationships in Semantic Web technology adds another dimension to the way the knowledge domain is defined. Where taxonomies are generally two-dimensional and organized in a hierarchy, ontologies can make use of relationships beyond the parent-child relationships that hierarchy implies. Ontologies can express temporal relationships (A happens before B), positional relationships (A is near B), causal relationships (A creates B), and any other relationship you can imagine.

As an example, Ian Davis and Eric Vitiello have created a vocabulary for describing relationships between people that they call simply "RELATIONSHIP." The vocabulary contains thirty-five possible relationships, from family relationships ("grandparent of") to less stable relationships ("has met," "would like to know").

RELATIONSHIP: A Vocabulary for Describing Relationships between People
http://vocab.org/relationship/.html

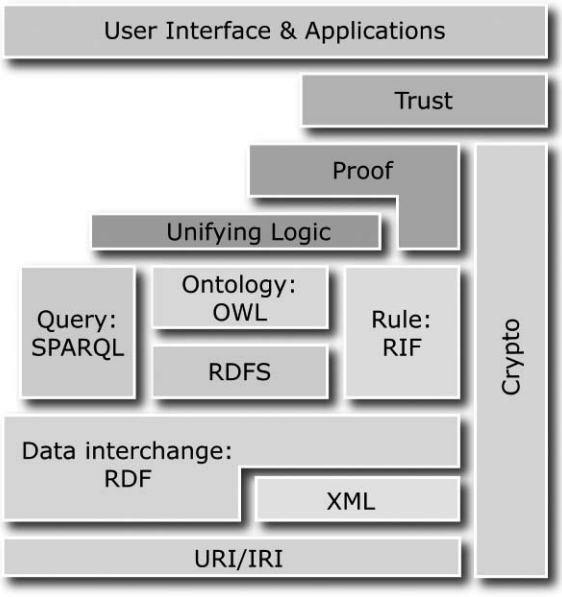

Figure 1
The Semantic Web "layer cake" model. Source: www.w3.org/2007/03/layerCake.png (accessed Dec. 15, 2009). Copyright © 2007 World Wide Web Consortium (Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University). All Rights Reserved. http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231.

The Resource Description Framework (RDF) is the Semantic Web standard for defining an ontology in machine-readable form.

RDF Knowledge Representation

In some ways, RDF reflects classic thinking about the nature of knowledge and how we represent it. It models knowledge as classes of things and relationships between things. Members of a class all have the characteristics that define the class.

The Role of Identifiers

Although RDF is described in terms that can also be expressed in human language (*subject, object*), what distinguishes it from natural language is that it is intended to be processed by machines. For that reason, RDF does not make use of natural language for the concepts and things it describes. Instead, each element of the RDF statement must be expressed with a unique identifier. This unique identifier has two primary advantages: (1) it overcomes the inherent ambiguity of human language (Pluto the celestial body vs. Pluto the Disney character)

URI	What It Identifies
http://purl.org/dc/terms/title	the Dublin Core metadata term "title"
http://id.loc.gov/authorities/sh85103579	the LC subject authority entry for "Pluto (Dwarf planet)"
http://id.loc.gov/authorities/sh96010495	the LC subject authority entry for "Pluto (Fictitious character)"
http://xmlns.com/foaf/spec/#term_name	the Friend of a Friend vocabulary term for a name

Table 1
Examples of identifiers used on the Web and the Semantic Web.

Subject	Predicate	Object
Vladimir Nabokov	is author of	*Lolita*

Table 2
"Vladimir Nabokov is the author of *Lolita*"

Subject	Predicate	Object
http://en.wikipedia.org/wiki/Vladimir_Nabokov	http://rdvocab.info/roles/authorWork	http://lccn.loc.gov/56024827

Table 3
Nabokov statement with URIs.

and (2) it allows for internationalization of the metadata, because the same identifier can be used even though the language of the display form is different (*computer* versus *ordinateur*).

On the Web, the standard identifier is called a Uniform Resource Identifier (URI). A URI follows a prescribed syntax: it begins with a URI scheme name, followed by a colon, followed by a string in a format that is particular to that scheme. It so happens that the URL, with its "http:" at the beginning, is a valid URI. URLs are commonly used as identifiers in Web-compatible applications (see table 1).

Statements

The basic building block of RDF is the *statement*. RDF statements are semantic units in a simple form: subject + predicate + object. Like simple molecules, the statements are interconnecting building blocks that can create complex networks. A statement says something simple, like

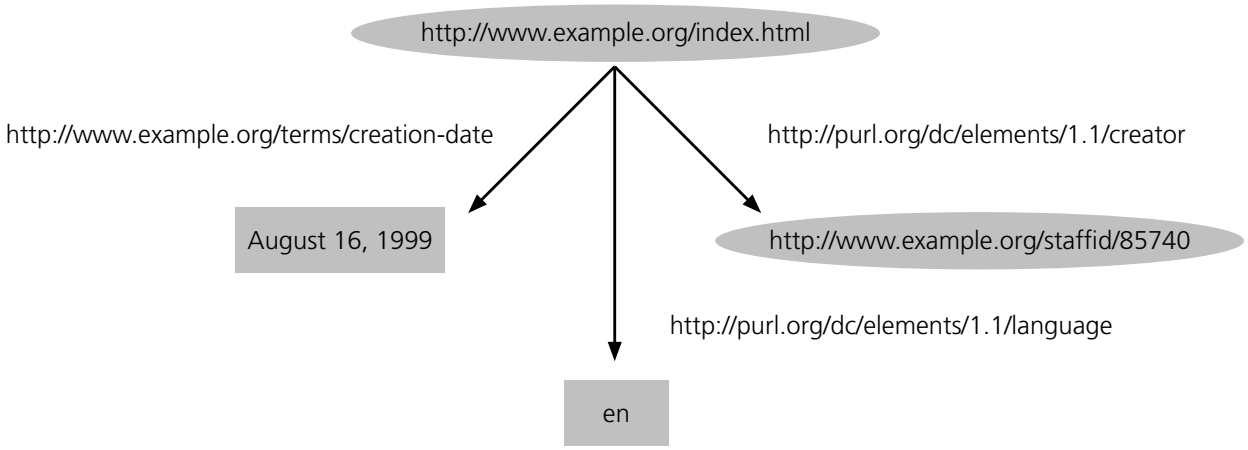

Figure 2
A simple RDF diagram. Source: "RDF Primer: W3C Recommendation 10 February 2004," figure 3, www.w3.org/TR/rdf-primer (accessed Dec. 15, 2009). Copyright © 2004 World Wide Web Consortium (Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University). All Rights Reserved. http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231.

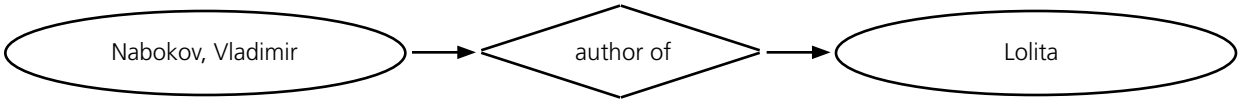

Figure 3
Lolita information represented in RDF graph.

"Vladimir Nabokov is the author of *Lolita"* (see table 2).

Note that in actual machine-readable RDF, each element of the statement would be represented by a URI (see table 3).

Because each statement is made up of three parts, they are often referred to as *triples*. Statements are commonly represented with diagrams (see figure 2).

Because using accurate URIs would make the examples in this document difficult to read, most examples that follow will use abbreviated values in the place of actual URIs.

With current library data in MARC21 records, the same data is present but expressed differently, in part because of the record structure that binds separate statements to each other. In a MARC21 record, the two statements below are semantically the same as the RDF graph shown in figure 3.

100 $a Nabokov, Vladimir

245 $a Lolita

What differs, and significantly so, is that the RDF statement contains the authorship relationship explicitly, while the two separate fields in the MARC21 record are held together only because they are contained within the same record. Outside of the record structure, they lose their connection to each other. The explicit inclusion of the relationship between the two things in our statement, the name of the person and the title of the book, creates a meaningful information unit that is not dependent on a record format.

Metadata in RDF

The Resource Description Framework is neither a data format nor an application. RDF provides a basic level of structure for metadata on which actual metadata can be built. It is so simple that it defines only three types of data that can be used in a statement: literal values (free text), structured values (text, but with structure like date and time), and identifiers in URI format. While the first two are essentially kinds of strings, the last can represent anything that has a Web-compatible identifier.

The Dublin Core Metadata Initiative built its abstract model (DCAM) on top of RDF and added a few more details that could be of use in library metadata. In particular, DCAM adds values that are controlled vocabularies.

Value Types

Literal

When an element is defined as taking a literal value, it means that the value will be free text, such as titles of documents, descriptive notes, or reviews. Knowing that this element will be free text tells developers that there is limited "computing" that can be applied to the data. This is a field for human readers, not for specific machine processing. The developer then needs to understand the meaning and intent of the field in order to determine how and when to present it to users, whether it might be useful as a searchable field, and so on.

Structured Value

A structured string is one with a defined set of elements, like "yyyy-mm-dd" for a date. There may be value rules, such as limiting the characters allowed to numbers and hyphens. The structure often allows for certain operations to be performed, like presenting the data in an ordered list either alphabetically or numerically. Structure is also valuable for the creation of displays. For example, "2009-02-14" could be displayed as "February 14, 2009" or "14 febbraio, 2009" or "14/02/2009."

Application programs usually exert control over input of the data in structured strings, making sure that the data matches the defined structure perfectly so that subsequent processing will produce accurate results.

What makes up the structure can be nearly anything, including other data elements. The bibliographic element "publication statement" is a structured element consisting of the elements "place," "publisher," and "publication date," each of which could be a defined element represented by a URI.

URI

Oftentimes the actual value of an RDF property will be an identifier, as in the example in table 3 where we identified our book author with the URI "http://en.wikipedia.org/wiki/Vladimir_Nabokov." Where possible, this is the preferred method for representing data on the Semantic Web.

Although anyone can create URIs, their value for sharing and linking data arises from the authority of the agency assigning the identifier. The domain portion of a URL ("id.loc.gov") generally belongs to the assigning agency, which is also often the agency that has created the

data being identified. Many commonly needed data types have not yet been assigned an identifier, such as standard lists for languages, and this is something of a stumbling block in the development of the Semantic Web.

Controlled List

In the controlled list data type, the value itself is taken from a previously determined finite list. The simplest of these are lists like those for languages or language codes, musical instruments, or audience types.

Some lists have structural relationships between their entries. For example, a thesaurus is a controlled vocabulary with structural relationships between entries (broader terms, narrower terms), and it often contains alternate forms of display for the entries and definitions. The name authority files used in libraries are controlled lists in which a given name has a great deal of information associated with it in the name authority record.

Controlled lists are used in metadata creation applications to assure that the value entered is indeed one of the values on the list. Applications making use of metadata that has controlled values take advantage of additional information that is provided related to the value. To do so, however, the controlled value must be unambiguously identified, preferably with a URI, and the information must be available in a machine-readable form on the Web. The Library of Congress is the maintenance agency for many controlled lists used by library cataloging, including the Subject Authorities, which are now available defined in accordance with Semantic Web technology. Each entry in the list has a unique identifier, and the authority record data is available for human readers and for machine processing. Figure 4 shows an LC subject authority entry in human-friendly format, and

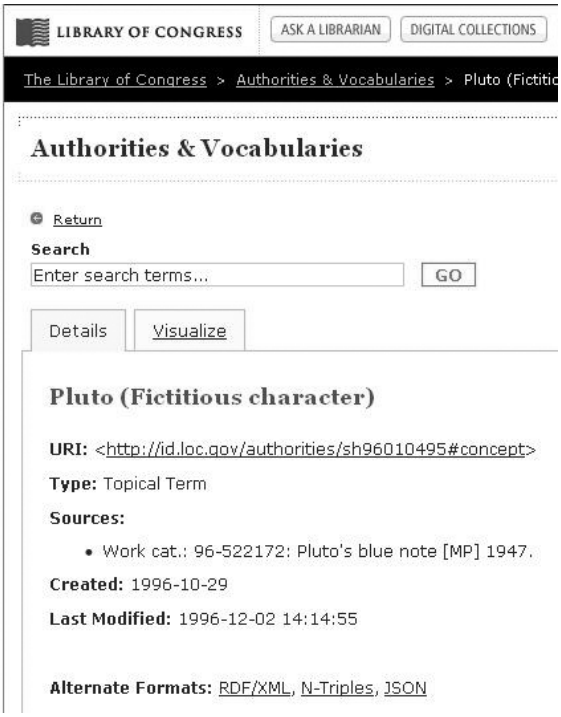

Figure 4
An LC subject authority entry on the Web in human-friendly format. Source: http://id.loc.gov/authorities/sh96010495.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
>
   <rdf:Description
rdf:about="http://id.loc.gov/authorities/sh96010495#concept">
     <skos:prefLabel xml:lang="en">Pluto (Fictitious character)</skos:prefLabel>
     <owl:sameAs rdf:resource="info:lc/authorities/sh96010495"/>
     <dcterms:modified
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1996-12-02T14:14:55-04:00</dcterms:modified>
     <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
     <dcterms:created
rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1996-10-29T00:00:00-04:00</dcterms:created>

     <dcterms:source xml:lang="en">Work cat.: 96-522172: Pluto's blue note [MP] 1947.</dcterms:source>
     <skos:inScheme
rdf:resource="http://id.loc.gov/authorities#conceptScheme"/>
     <skos:inScheme rdf:resource="http://id.loc.gov/authorities#topicalTerms"/>
   </rdf:Description>
</rdf:RDF>
```

Figure 5
The same data as that shown in figure 4 in RDF/XML for use in computer applications.

figure 5 shows the same data in RDF/XML for use in computer applications.

Some Metadata Implementations Using RDF

RDF provides a foundation but is not itself a metadata implementation. There are numerous metadata standards and applications that are being developed using the RDF concepts and rules. Some of the ones of greatest interest to library data developers are listed here.

SKOS

The World Wide Web consortium (W3C), the standards body that develops Semantic Web standards and is also responsible for RDF, is creating some key data formats that use RDF. Of these, one of great importance to libraries is Simple Knowledge Organization System (SKOS). SKOS is a standard way to present organized data such as thesauri, classification schemes, and subject heading schemes. With SKOS you can represent hierarchical relationships and provide indexing terms, entry vocabulary, and definitions. Because the basis of SKOS is RDF, SKOS makes use of the RDF concepts of classes, properties, and values.

SKOS is being used for the implementation of Library of Congress Subject Headings on the Web. It is also being used for the encoding of the vocabularies that are part of the new library cataloging standard, Resource Description and Access (RDA). Both of these will be discussed in greater detail later on.

OWL

Another W3C standard is the Web Ontology Language, OWL. OWL contains additional features for the expression of vocabularies and relationships between terms in a way that facilitates the development of machine applications that use the vocabularies. The implementations of OWL to date tends to focus on scientific vocabularies, where the precision of OWL is needed.

Linked Data

By far the most commonly used implementation of RDF is that of linked data. Linked data is a fairly simple expression of data using the basic concepts of RDF: that data is expressed in the RDF triple format (subject–predicate–object) and that the parts of the triple should be represented by standard identifiers where available. The starting point for linked data as a concept is in a 2006 essay by Tim Berners-Lee on the W3C website.[2] In that short essay, Berners-Lee laid out the essential rules for linked data, which include the use of URLs to identify elements. One great advantage of using URLs as identifiers is that the identifier can also serve as a link to further information about the thing being identified. While using the same string as both an identifier and a location can also create some confusion, this method has been used already for hundreds of data sets.

Whereas the Semantic Web, at least as initially described by Berners-Lee, was intended to create a web

OWL overview
www.w3.org/TR/owl2-overview

List of OWL ontologies
http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library

Linked Data website
http://linkeddata.org

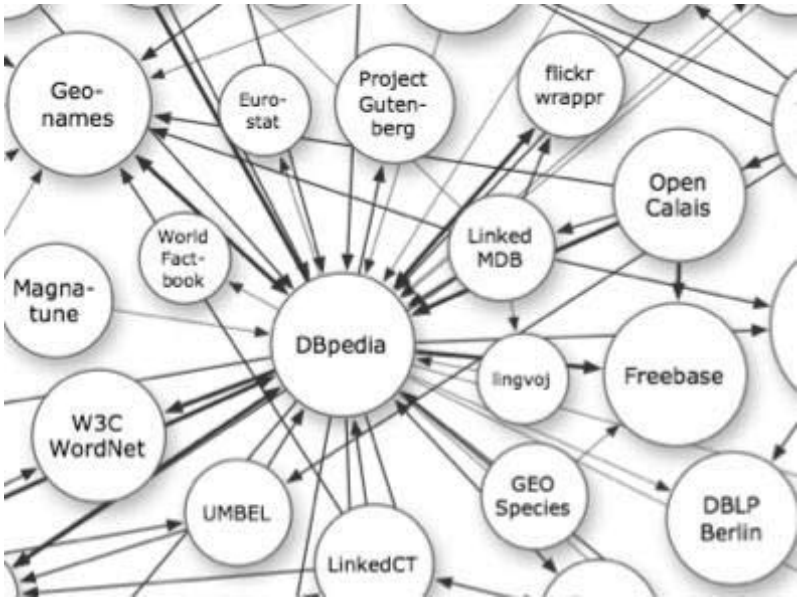

Figure 6
Partial view of the "linked data cloud" from http://linkeddata.org.

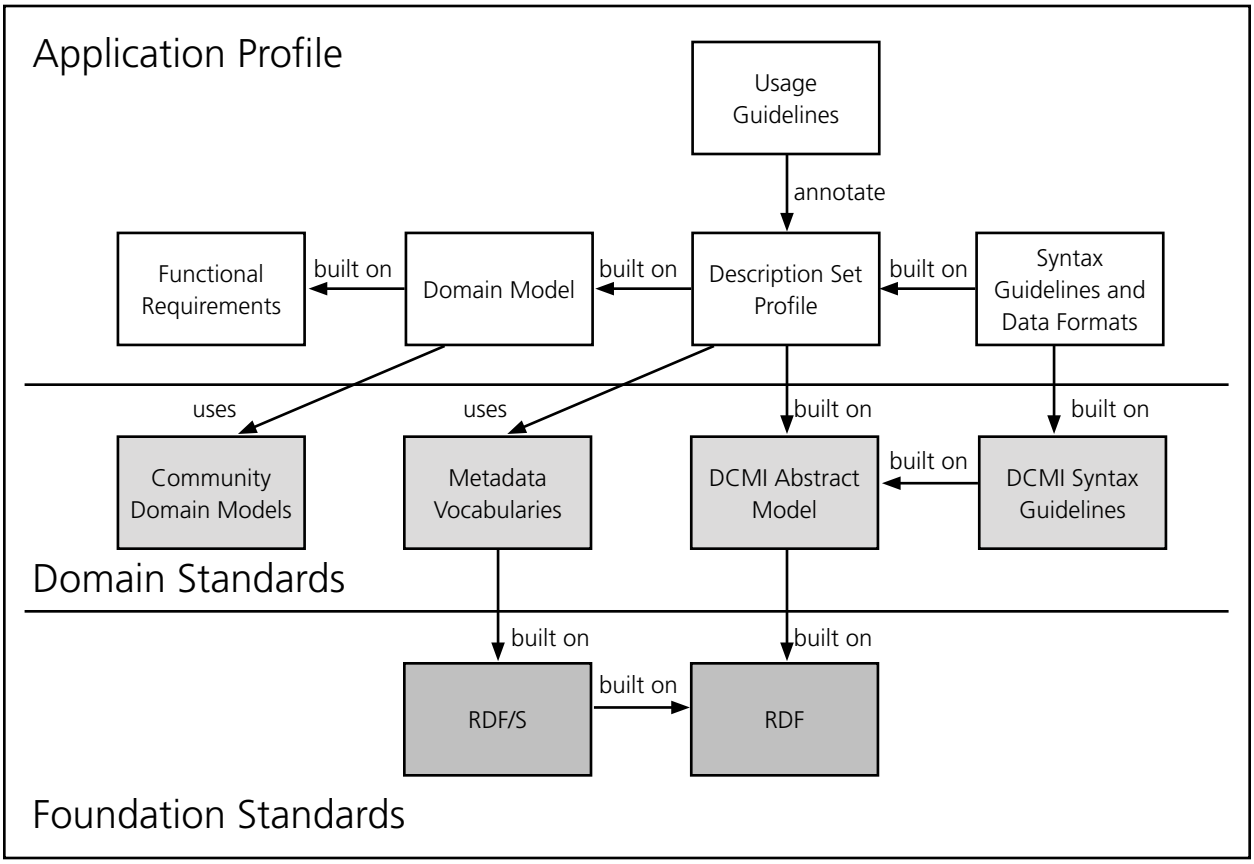

Figure 7
The Singapore Framework. Source: http://dublincore.org/documents/singapore-framework. Copyright © 2007 Dublin Core Metadata Initiative. All Rights Reserved. http://www.dublincore.org/about/copyright/.

from the information currently buried in the many millions (or billions?) of documents on the Web, linked data is taking on a somewhat simpler task by allowing those with data and metadata that is already in a structured format to place that data on the Web. Once on the Web in a standard format, data from different sources can be linked together to create new information views. The data available as linked data varies greatly, from data sets representing popular music and movies to scientific data like that of Bio2RDF, covering human and mouse genome information (see figure 6).

The Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) has embraced RDF principles in its work and has transformed the initial fifteen metadata elements that make up the original core set into an extensible and flexible RDF-compliant set of metadata.[3] Its new work goes far beyond the creation of a core of metadata elements, although the work does include the definition of the Dublin Core metadata terms using Semantic Web standards. Of particular interest is the "big picture" model with Foundation Standards, Domain Standards, and Application Profiles, shown in figure 7. The diagram is known as the Singapore Framework because it was first presented at the Dublin Core conference in Singapore in 2007.[4]

The Singapore Framework diagram helps make sense of the complex of elements that make up a functional metadata description. It also introduces the concept of an application profile as the cohesive element for metadata applications.

Some elements of the diagram are specific to the thinking of the Dublin Core community, in particular the DCMI Abstract Model and the DCMI Syntax Guidelines. The basic structure and components, however, are very helpful for understanding the creation of a metadata set in an environment where the metadata must be defined for machine processing. The foundation of the model is RDF, which provides the basic concepts of metadata components in terms of things and relationships. The

next layer up defines domain standards, such as a general community domain model and the vocabularies that will be used in the application. In the library community, the Functional Requirements for Bibliographic Records (FRBR) and its companion models of Functional Requirements for Authority Data (FRAD) and Functional Requirements for Subject Authority Records (FRASAR) are the models of our domain. They specify the components and delineate the boundaries of library metadata. Above this level is that of the application profile. It is here that the somewhat abstract definition of terms and structures becomes an operational metadata activity, with a selection of terms and the presence of guidelines for the creation of metadata for that community.

CIDOC CRM

The international museum community is also working on new models for its data under the International Council of Museums. The CIDOC Conceptual Reference Model (CRM) defines an extensible semantic framework for the scientific documentation of cultural heritage collections. The CIDOC CRM has been developed in cooperation with the DCMI and FRBR communities, among others. The CIDOC ontology for cultural heritage information has been established as ISO 21127. The CIDOC CRM ontology is available as a file in RDF.

International Council of Museums
http://icom.museum

CIDOC Conceptual Reference Model, v. 5.0.1 (Nov. 2009)
http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Nov09.pdf

CIDOC CRM's domain model covers description, object management, and preservation. CIDOC has also created an extension of FRBR called FRBRoo, for "object-oriented." FRBRoo has many additional entities that are required by the museum community, including entities for individual and complex works and for events. These entities reflect needs of the museum community that were not part of the library community's analysis.

The museum community is of particular interest to library metadata development because there is an overlap between the metadata needs of museums (which own objects and documents) and libraries (which own mainly documents but also some objects). The CIDOC CRM has the potential to provide an excellent testbed for the concept of linking between the library and the museum communities using a FRBR- and RDF-based metadata model.

ISO 21127
www.iso.org/iso/catalogue_detail.htm?csnumber=34424

CIDOC CRM v3.3.2 Encoded in RDFS
www.cidoc-crm.org/docs/xml_to_rdfs/CIDOC_v3.3.2.rdfs

RDF and Library Data

In the past, library cataloging has focused almost exclusively on the creation of usage guidelines in the form of cataloging rules. Usage guidelines are the instructions on how to make decisions about the content of the metadata. This is an area where libraries excel, and the rules cover cases that most other communities handling bibliographic data have never considered.

Until recently, a well-developed domain model did not exist, but this has been described by FRBR and its companion functional models. The addition of FRBR to the library metadata toolkit provides both an opportunity and a challenge: the opportunity to rethink the structure and content of library metadata, and the challenge to actually restructure that metadata based on that rethinking.

RDA, as an implementation of the FRBR model, provides a chance to move into a more modern style of metadata development and usage. As with previous library cataloging rules, RDA is primarily in the form of usage guidelines: a document for the catalogers who will make decisions about the content of library metadata. From the document, however, one can extract the information necessary for the creation of the metadata vocabularies.

FRBRoo Model, v. 1.0 (draft), May 2009
www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0_draft__2009_may_.pdf

These vocabularies can initially be defined apart from any particular data or record format. It is the combination of the vocabularies, the model, and an eventual application profile that will form the basis for the future of bibliographic data.

The remainder of this report will focus on the possible transformation of library data through Semantic Web and linked data principles.

Continued on page 36