Vocabularies

Term Lists and Thesauri

Abstract

Chapter 4 covers controlled vocabularies, which in linked data are made public on the Web. This allows for machine-checking on validity and also allows communities to describe the meaning of their terms and harmonize them with those of other communities. This chapter describes controlled term lists, subject lists, and thesauri from library and nonlibrary communities.

In the previous chapter we saw that there are data elements available for use covering many things you might want to describe in metadata. Reusing existing data elements is one way to assure that your data will find links on the linked data Web. Another way to increase the meaning of the links is to control the content by using controlled vocabularies as the content of your metadata statements.

Controlled vocabularies in linked data are published—that is, they are made public on the Web. This allows for machine-checking on validity and also gives communities an opportunity to describe the meaning of their terms and harmonize their terms with those of other communities. Let's use the concept of color as an example. If two communities are using the same data element for color but do not share a controlled list of terms, they have no way to know if plum and aubergine are more or less the same color. If the plum people and the aubergine people want to share data, they can each create a controlled vocabulary for their terms, and within these vocabularies they can, using Semantic Web standards, say, "Plum is a close match for aubergine" and "Aubergine is a close match for plum."

If they have followed the recommended practice of linking to broader vocabularies as well, they could have each linked their color vocabulary to a vocabulary of common colors (red, purple, blue, yellow, green, orange)—"Plum is a kind of purple" and "Aubergine is a kind of purple."

In any situation where data is being shared, there is now sufficient information for applications to be written that can use *plum*, *aubergine*, and *purple* interchangeably, if desired.

Apart from the advantages of sharing, controlled vocabularies allow verification at the time of input and receipt of metadata. If the data value must be a member of a list, the input program can verify that whatever is provided meets that rule. Drop-down lists for selection can also aid input so that the person doing input does not have to remember or look up valid values.

Term lists are usually experienced by human users as words from natural language. Having lists that are actually made up of natural language terms, like *red* or *blue,* is not entirely useful on the linked data Web, where one may be sharing data globally. Wherever possible, linked data terms lists use identifiers for the terms. This is not terribly different from the use of codes in the fixed fields in MARC records, although the linked data codes are full URIs. With the members of the list identified in a language-neutral way, user displays can be developed for any desired language. As discussed in chapter 3 in the section on SKOS, for each language, one can define a primary display term and any number of alternate and hidden terms:

```
ex:color rdf:type skos:Concept;
 skos:prefLabel "red"@en;
 skos:altLabel "ruby"@en;
 skos:hiddenLabel "reddish"@en;
 skos:prefLabel "rouge"@fr;
```

```
    skos:prefLabel “rosso”@it;
    skos:altLabel “ciliegia”@it.
```

Controlled lists can be as simple as a single-level list of terms (red, green, blue), or they can be structured with broader and narrower concepts, like a taxonomy.

```
   ex:dogs rdf:type skos:Concept;
    skos:prefLabel “dogs”@en;
    skos:narrower ex:working_dogs;
    skos:broader ex:animals.
```

The color example is, by the way, not at all far-fetched. While the artistic and social perception of colors is hard to define, as this example shows, colors in the computer technology world are easily defined using the RGB coordinate method. In fact, there is a color vocabulary on the LOD cloud called Linked Open Colors. Most likely the plum and aubergine communities of the example express their colors and relationships using such a system, and Linked Open Colors should help them in doing so in a Semantic Web standard way.

Linked Open Colors
http://thedatahub.org/dataset/loc

There is overlap between the concept of a term list and what in libraries we call an authority file. In one sense, a file that controls the identities of personal and corporate creators functions as a term list for input and display. It differs from many term lists, however, because it can provide additional information about the creator that goes beyond the simple selection of a preferred name form. Another difference is that authority files are never considered to be comprehensive, and new entities are added readily whenever a new creator is encountered. Term lists are often considered comprehensive to a point, and adding a new term to, say, a list of colors is done only after great consideration.

Not all metadata elements lend themselves to vocabulary control. The titles of books and articles cannot be predetermined and entered into a list, for example. Elements of this nature must be accepted as uncontrolled text, and little or no automatic verification can be made on these entries. The creative possibilities of language also prevent us from applying much engineering to it in a machine environment. Because some of our information will of necessity be in natural language, it is important to take advantage of every bit of information that can be subjected to algorithmic rigor, since it is these elements that will be the most fruitful for linking and computation.

Subject Lists and Thesauri

Nonlibrary and General

Topical thesauri are the sine qua non of controlled vocabularies. Topical thesauri have much in common with the science of taxonomy and with various attempts, since the times of ancient Greece, to create a comprehensive map of the world and of our knowledge about the world.

Defining the world as a strict hierarchy is rarely the goal of subject lists today. Topical access is instead a component of language-based retrieval. Keyword access, plus the ability of query engines to include facets and other advanced search capabilities, has made it unimportant that the list fit a strict top-down structure.

Most subject lists are ontologically incomplete; that is, they do not describe the whole of reality (for any definition of reality). Even the library classifications and subject lists are bounded by the topical coverage of the items in the library, which, although vast, is still a view limited to the knowledge that has been written or recorded in some way. Subject lists tend to grow organically as material is acquired or gathered together. This is definitely true for the current crop of linked data resources: each develops subjects as needed for its own applications that are appropriate to the materials its metadata describes. That said, there are interesting overlaps between these since we are all living in approximately the same real world environment. This section provides a small number of examples of topical lists that might be of interest for linking with library data. Note, however, that every dataset in the Linked Data cloud in a sense represents topical information and that the separation between subject heading and subject is not without ambiguity.

DBpedia Ontology

- **Name**: DBpedia Ontology
- **Creators**: various
- **URL**: http://wiki.dbpedia.org/Ontology
- **Updated**: September 11, 2011 (version 3.7)

Being the very center of the Linked Data cloud, you might guess that DBpedia covers a broad topic area. That is indeed the case. The basis of DBpedia is a compilation of all of the facts from Wikipedia. Most of these facts are derived primarily from the structured data that you see in the info boxes on each Wikipedia entry. The term list has 1,830,000 entries overall, 526,000 for places, 416,000 for persons, 262,000 for works, 183,000 for species, and 169,000 for organizations.

DBpedia is described as a shallow ontology because it has over three hundred top-level classes, each of which has a small number of subclasses. The shallowness of the ontology means that some terms are associated with a large number of resources.

The DBpedia term list reflects the nature of the user-produced Wikipedia; thus, it has a number of high-level entry points for sports and none for philosophy, although there is an entry for philosopher as a kind of person, but none for poet. This may seem odd or even wrong to some, but we actually experience this kind of topical skewing whenever we search on the Web, where a search for python returns first sites for the programming language and only later those for the snake of the same name. DBpedia, like the Web and like Wikipedia, is a growing organism whose contents are determined by its users.

Freebase Types

- **Name**: Freebase types
- **Creators**: various
- **URL**: http://www.freebase.com/schema
- **Updated**: continuously

Freebase is a linked data database with a graphical user interface. It takes data from a number of different sources (including DBpedia) and has information on over two million persons and six million books, among other topics. Freebase articles are "typed," or given topical headings, both algorithmically and by Freebase users. The upper level of the Freebase topical schema is called a domain. As in DBpedia, the domains reflect the nature of the current state of linked data culture: the first domain listed in alphabetical order is American football. While some domains cover what most people would consider "real" subjects, like Geology, Education, and Medicine, others are definitely more unusual: Digicams, Fictional Universes, Metaweb System Types. That the upper-level domains are not equal in importance can sometimes be intuited by the number of subcategories in the domain. These range from as few as one (for Wikipedia) to over 160 (for Location). While the structure of Freebase may be somewhat quirky, individual domains may prove to be interesting for others who wish to categorize resources. As an example, the Theater domain has categories for types of theatrical offerings (Musical comedy) and entries for different roles, such as Choreographer and Designer.

The Freebase schema entries are available as a SKOS vocabulary, and work is being done to link the Freebase topics to other linked data topical vocabularies so that a maximum of linking can be achieved.

New York Times Subjects

- **Name**: New York Times Linked Open Data
- **Creator**: New York Times
- **URL**: http://data.nytimes.com
- **Updated**: January 13, 2010

Newspapers are, of course, information institutions. One back-room function of a newspaper is its morgue, an archive of the published content of the paper, generally organized by date and by topic. The morgue is where editors go for photographs and background information to accompany stories.

In today's newspaper, the morgue logically becomes a digital archive. The *New York Times* has made available approximately ten thousand topics organized by the paper's own subject entries. Most of these topics are related to people, organizations, and locations; a minority are topical in nature.

The New York Times has made its subject list available in linked data format using some SKOS features and some locally defined elements. The announcement of the service mentions that these terms are the result of 150 years of careful curation of the newspaper's archives. Many of the terms are for people, organizations, and locations, but nearly 500 are topical descriptors. These topical descriptors have no structure (that is, no broader or narrower structure) and vary in their breadth (*reality television* and *poetry* and *poets*). Some terms, however, have a link to the first and latest uses by the newspaper, which could be of interest for historical study.

Where possible, the subject terms have been linked to related terms in the Freebase and DBpedia schemes. These in turn link to other sources of information. For example, the New York Times term for the Wimbledon tennis tournament links to terms in both DBpedia and Freebase. The DBpedia term links to the GeoNames entry for the location. Wimbledon is topically narrower than Tennis Tournaments in the Freebase subject schema, and that leads to topics for other tennis tournaments, which gives us the graph in figure 4.1.

In addition, both Freebase and DBpedia link to Wikipedia entries for the same topic in languages other than English. It's easy to see how a simple subject entry in one file can be expanded by association.

Name entries for people account for a majority of the Times subject headings. The entry for J. Edgar Hoover (entered as Hoover, J Edgar) provides a brief biographical note and a link to a Times page about him. It also includes links to key Times articles about Hoover. All of this is retrieved with the URI that identifies the person. In addition, the Times includes links to the related URIs from DBpedia and Freebase. The link to DBpedia links to the Wikipedia entry, and the Wikipedia entry includes a link to the Virtual International Authority File (VIAF), a file of library name authority records, including the Library of Congress Name Authority data.

Artificial Intelligence: OpenCyc and UMBEL

- **Name**: OpenCyc
- **Creator**: Cyc Foundation

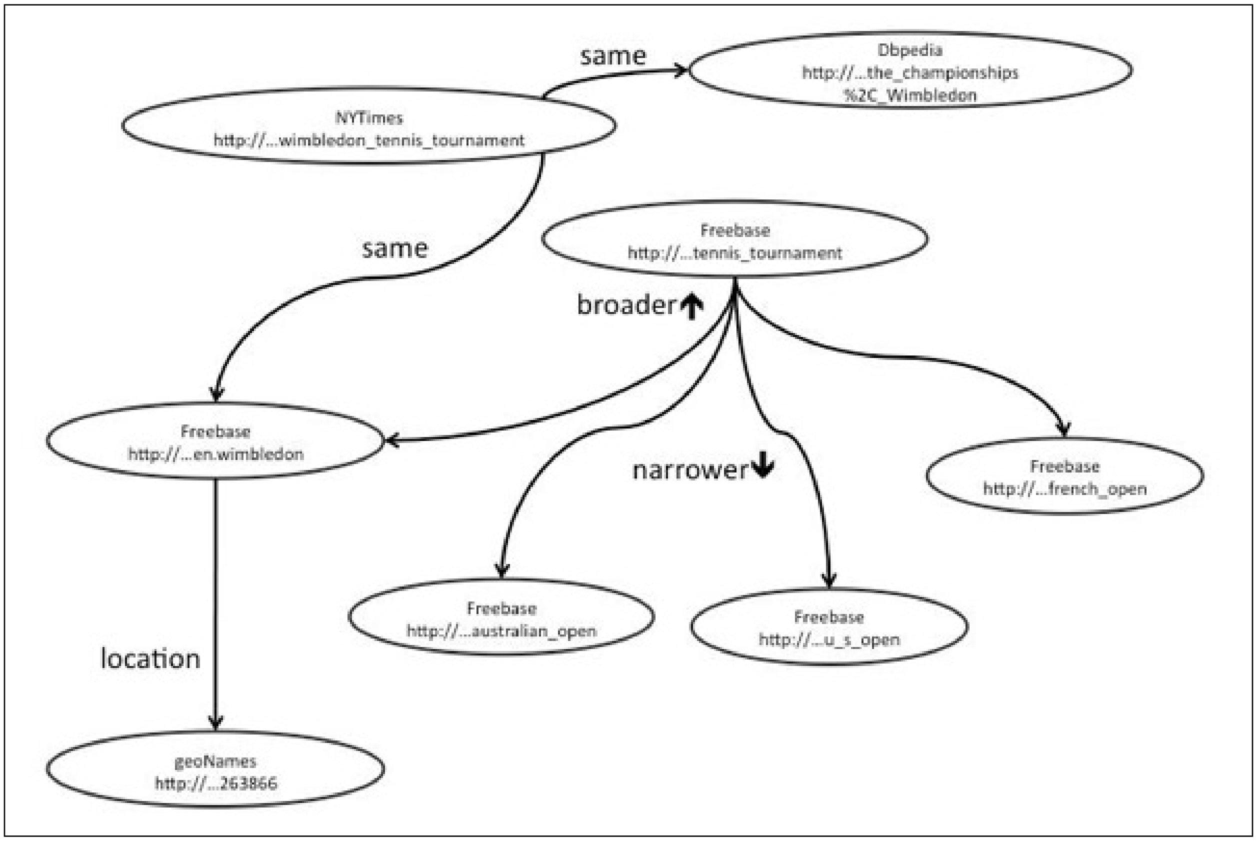


Figure 4.1
A graph of related subjects starting with *Wimbledon* from the New York Times thesaurus.

- **URL**: http://sw.opencyc.org
- **Created**: 2009
- **Updated**: 2009

- **Name**: UMBEL
- **Creators**: Michael Bergman, Frédérick Giasson
- **URL**: http://umbel.org
- **Created**: 2008
- **Updated**: February 14, 2011 (version 1.00)

It is hard to resist the challenge of creating the ultimate thesaurus that will work for every possible resource online or off. While DBpedia and Freebase have built their term lists from the bottom up, there are other groups that set about to create thesauri that are generalized in nature. Two of these are UMBEL and OpenCyc.

OpenCyc is a Semantic Web output from the Cyc project in artificial intelligence. The Cyc project attempts to define the world of everyday knowledge for use in artificial intelligence applications. OpenCyc is a subset of the same data produced for linking in the Linked Data cloud.

UMBEL stands for Upper Mapping and Binding Exchange Layer. It is a subset of OpenCyc, designed with linking in mind. UMBEL has about 28,000 concepts that could provide linking points between more specialized vocabularies. For example, in UMBEL there is a category for dog, with a subcategory herding dog, and bottom-level concepts like German shepherd and Bearded collie. These terms could find links to library subject headings and to library classifications. Because UMBEL comes from the artificial intelligence world, it includes common world concepts that may not normally appear in library catalogs, such as a category door that has subcategories garage door, screen door, and back door key, among others. It isn't hard to imagine some interaction between UMBEL and the Web of commerce that helps a personal bot find exactly the piece of hardware you need.

Library-Specific Subject Headings

Libraries have begun to make their subject lists available, at least in part because the SKOS standard provides a ready data format for these. These subject lists are generally more comprehensive than those that represent individual datasets on the linked data Web, but it is important to recognize that even these subject lists have grown organically over time, although in the case of libraries the time stretches to centuries.

Two features of linked data make the expression of library subject headings in this format especially powerful: the ability to create links between entries in the subject heading lists, and the ease of managing multilingual thesauri. While each subject list is necessarily bound by the collection it describes, together the lists create a broader view of the world and allow the creation of applications that move seamlessly (or so we hope) through these various views and repositories of knowledge.

Just the few subject heading lists detailed in this document can create a potentially rich web of interlinked topics in multiple languages.

Library of Congress Subject Headings (LCSH)

- **Name**: Library of Congress Subject Headings
- **Creator**: Library of Congress
- **URL**: http://id.loc.gov
- **Created**: April 2009
- **Updated**: April 6, 2011

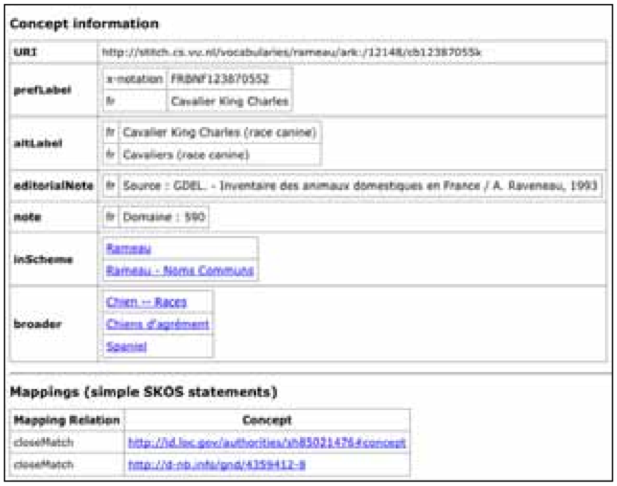
In 2008, an experimental version of LCSH was published in SKOS by Ed Summers. Summers, employed at LC, was a member of the SKOS standard committee, so making use of the standard for the library's topical list was undoubtedly something that he had been thinking about for a while. That version was unofficial, but LC issued an official version of LCSH in SKOS in April 2009.

SKOS turned out to be only barely adequate for the complexity of LCSH, in particular the use of precoordinated facets. Each SKOS entry is an entire LCSH heading entry, with double dashes between segments of the precoordinated heading and no way to differentiate the types of subheadings (topical, geographical, or time period). Nevertheless, LCSH in SKOS is a milestone in presenting library data in linked data format and has found a place on the Linked Data cloud diagram.

Although LCSH in SKOS is imperfect, it does allow LCSH to be linked to other subject heading schemes, such as RAMEAU, from the Bibliotheque Nationale de France.

Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU)

- **Name**: RAMEAU
- **Creator**: Bibliotheque Nationale de France
- **URL**: http://www.cs.vu.nl/STITCH/rameau
- **Created**: 2008



Concept information

URI	http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb12387055k
prefLabel	x-notation FRBNF123870552 fr Cavalier King Charles
altLabel	fr Cavalier King Charles (race canine) fr Cavaliers (race canine)
editorialNote	fr Source : GDEL. - Inventaire des animaux domestiques en France / A. Raveneau, 1993
note	fr Domaine : 590
inScheme	Rameau Rameau - Noms Communs
broader	Chien -- Races Chiens d'agrément Spaniel

Mappings (simple SKOS statements)

Mapping Relation	Concept
closeMatch	http://id.loc.gov/authorities/sh85021476#concept
closeMatch	http://d-nb.info/gnd/4359412-8

Figure 4.2
A RAMEAU entry with links to LCSH and German subject headings

The Bibliotheque Nationale de France makes its subject headings available in SKOS through an experimental service called STITCH, Semantic Interoperability to Access Cultural Heritage. The service includes other vocabularies and endeavors to create links between them as a way to increase interoperability between Europe's libraries and museums.

RAMEAU has been linked to LCSH and to the terms from the National Library of Germany (see figure 4.2).

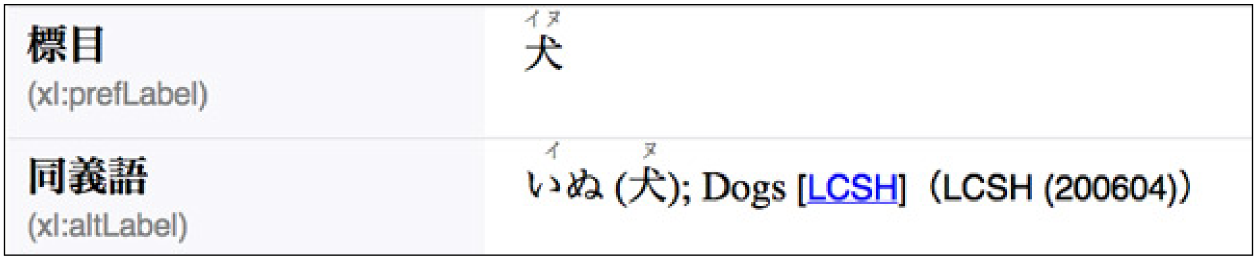
Deutsche Nationalbibliothek Subject Headings

- **Name**: Deutsche Nationalbibliothek subject headings
- **Creator**: Deutsche Nationalbibliothek
- **URL**: https://wiki.d-nb.de/display/LDS
- **Created**: 2010
- **Updated**: 2011

The release of the subject headings from the National Library of Germany (Deutsche Nationalbibliothek) is part of a larger plan to produce all of the library's data in linked data form. Both name authority and subject authority files have been published and are available for download. The subject headings (Schlagwortnormdatei, or SWD) are defined using SKOS elements for broader, narrower, and related headings. The subject headings are linked to both LCSH and RAMEAU. Names are linked to DBpedia and Wikipedia and are also included in VIAF.

National Diet Library List of Subject Headings (NDLSH)

The National Diet Library List of Subject Headings (NDLSH) is the subject list from the national library of Japan. The list contains all topical headings and some name headings. NDLSH is available in linked data format. It has some links to LCSH, which then provide English language terms for searching (see figure 4.3).



標目 (xl:prefLabel)	イヌ 犬
同義語 (xl:altLabel)	イ ヌ いぬ (犬); Dogs [LCSH]（LCSH (200604)）

Figure 4.3
A portion of an entry for NDLSH showing link to LCSH

Special Library Thesauri

Many special libraries or specialist communities create their own thesauri owing in part to the fact that general thesauri often do not provide the detail that a specialist needs. While these thesauri serve their populations well, they also create a separation between the special library and its general counterpart, as well as a separation between special libraries that could benefit from sharing. The linked data solution has the potential to create navigable links from general to special and vice versa, as well as between specialist libraries. This could also increase the visibility of special libraries to users who may enter through a more general library portal without knowing that special libraries or collections exist in the area they are exploring.

National Agriculture Library (NAL)

- **Name:** National Agricultural Library Thesaurus
- **Creator:** National Agricultural Library
- **URL:** http://agclass.nal.usda.gov/download.shtml
- **Updated:** 2012

The US National Agriculture Library (NAL) provides its thesaurus and glossary in linked data format in both English and Spanish. The linked data standard for thesauri, SKOS, allows for both authoritative and alternate forms to be in any number of languages:

```
<skos:prefLabel xml:lang=
"es">perros</skos:prefLabel>
<skos:prefLabel xml:lang=
"en">dogs</skos:prefLabel>
```

The preferred label in each language is the one that would be displayed to users in a system serving users in that language.

NAL's thesaurus entries are connected to LCSH and have connections from AGROVOC.

AGROVOC

- **Name:** AGROVOC
- **Creator:** Food and Agriculture Organization (FAO)
- **URL:** http://aims.fao.org/standards/agrovoc/linked-open-data
- **Updated:** 2011

The Food and Agriculture Organization's thesaurus for agricultural information, AGROVOC, is a highly sophisticated concept scheme with about 40,000 concepts in over twenty languages. With such a large thesaurus to manage, FAO has long been a leader in the adoption of new technologies for managing and using its data. The unique position of a United Nations agency in reaching out to the entire globe with information to promote development makes its participation in the Linked Data cloud even more valuable.

AGROVOC in linked data format is a participant in the Linked Data cloud and is linked to a number of other vocabularies there such as DBpedia, the Geopolitical Ontology, and the STW Thesaurus for Economics.

Thesaurus for Graphic Materials (TGM)

- **Name**: Thesaurus for Graphic Materials
- **Creator**: Library of Congress
- **URL**: http://id.loc.gov
- **Updated**: May 16, 2011

The Library of Congress Prints and Photographs Division has contributed its Thesaurus for Graphic Materials to the list of linked data term lists managed by LC.

Library Classification Schemes

Library classification schemes are rich, structured views of the knowledge universe that sits on library shelves. They have unfortunately been relegated to shelf location schemes because of the difficulty of navigating them as subject guides. In addition, in most cases there is not an actual index to the classification that is part of the user experience in libraries: most US libraries use LCSH for subject access, but classify books by the Dewey Decimal Classification or the Library of Congress Classification system. While there are cross-references between the subject headings and the classification, these aren't readily visible to library users. Because linked data facilitates linking between disparate vocabularies, it may in the end aid in creating a better subject navigation experience using subject headings and classification schemes.

Dewey Summaries

- **Name**: Dewey Summaries
- **Creator**: OCLC
- **URL**: http://dewey.info
- **Created**: 2010

The first three digits of the Dewey Decimal Classification (DDC) are available as a SKOS vocabulary. This is the portion of DDC called the Dewey Summaries that is made publicly available by OCLC. DDC has been translated into many languages, and we've seen that Semantic Web standards are designed to be multilingual. When you browse the online entries, you see the same term in all of the available languages. This linked data version of DDC gives access to the summary level of the twenty-third edition of Dewey in eleven languages and to the current abridged edition in three languages. It is possible to search the classification online, and there is also a SPARQL endpoint that allows programs to search and download entries. There is a sample SPARQL search linked from the project homepage.

Dewey Summaries project
http://www.oclc.org/dewey/webservices

Universal Decimal Classification (UDC)

- **Name**: Universal Decimal Classification (summary)
- **Creator**: UDC Consortium
- **URL**: http://www.udcc.org/udcsummary/exports.htm
- **Created**: November 18, 2011

The summaries of the Universal Decimal Classification will be released gradually beginning in 2012. The first entries will have the basic thesaurus structure with the classification notation, display terms, broader and narrower term links, and application notes that describe the allowed usage of the class. In a future release the published terms will have additional data, including administrative metadata (dates of introduction and update and other change information).

UDC is issued in whole in twenty-six languages using nine scripts. The summaries are in a total of forty-six languages. UDC uses faceting techniques that result in similar problems to those faced in the linked data version of LCSH.

Other Controlled Vocabularies

In libraries we rely heavily on controlled vocabularies. We use them for things like document types (dictionary, bibliography, biography) or to describe the particular relationship between a person and a resource (editor, illustrator, composer). In linked data these relationships tend to be treated as actual things rather than as vocabulary terms. As an example, RDA includes a list of roles for persons and corporations. In RDF these are not stored as a list because the roles actually represent relationships between entities and will be used as predicates in the RDF triple:

J.S. Bach → [is] composer [of] → Mass in B minor
Mass in B minor → [has] composer → J. S. Bach
Alfred Hitchcock → [is] director [of] → Spellbound
Spellbound → [has] director → Alfred Hitchcock

There isn't always a clear distinction between a controlled list of terms and a set of data. For example, DBpedia gathers all of the facts from the Wikipedia boxes and creates a very large dataset of interlinked data elements. Those elements can also be used as identified terms in metadata.

Some of the data element sets that were presented in chapter 3 contain concepts that are treated as lists of terms in current library data. As library data takes on more of the Semantic Web structure, we can expect some lists to be replaced by elements. Other lists are appropriate as vocabularies, at least in some contexts. Lists of colors for images, or of tape and groove widths for recordings, will continue to be needed. Systems like GeoNames have been mentioned elsewhere in this report for their element sets, but they are listed in this section for the controlled vocabularies that they contain.

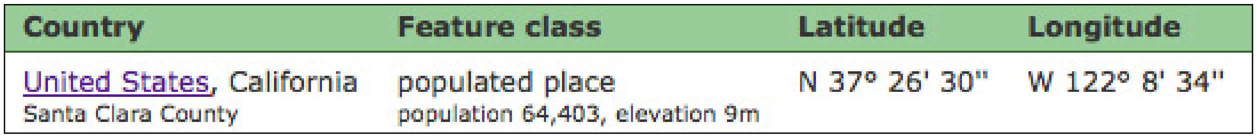
Non–Library-Based Lists

GeoNames

Geographic names are needed in data and metadata from government and scientific data to leisure and entertainment. Fortunately there is a strong geographic vocabulary in the Linked Data cloud, GeoNames (figure 4.4). It contains names, coordinates, and a categorization (populated place, lake, administrative division). GeoNames has been compiled from a number of different sources and contains some duplicate data, and uses expert volunteers, called ambassadors, throughout the world to provide quality control.

BBC Wildlife Ontology

- **Name:** BBC Wildlife Ontology
- **Creator:** BBC
- **URL:** http://www.bbc.co.uk/ontologies/wildlife/2010



Country	Feature class	Latitude	Longitude
United States, California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"

Figure 4.4
A GeoNames entry for the city of Palo Alto

- **Created:** 2010
- **Updated:** 2010

The BBC Wildlife Ontology is a relatively simple vocabulary for describing biological species on the BBC Wildlife website. It aims to be scientifically accurate but not overwhelming for the website user. It includes such information about the species as habitat, conservation status, and adaptation to its environment. The vocabulary is applied to information on the BBC Wildlife site, which links to and supports BBC programs in this area. The information beyond the species taxonomy makes this vocabulary suitable for many types of learning environments.

Dublin Core Type Vocabulary

- **Name**: Dublin Core Type Vocabulary
- **Creator**: DCMI
- **URL**: http://dublincore.org/documents/dcmi-type-vocabulary
- **Created**: October 11, 2010

A common function in resource description is to characterize the type of thing you are describing. The Dublin Core metadata identifies one dozen resource types, such as *text*, *image*, and *event*. These are broad categories, in keeping with Dublin Core's philosophy of providing a high-level metadata description that can be extended by more specific metadata schemes.

Library-Based Lists

MARC Term Lists

- **Name:** MARC Countries
- **Creator:** Library of Congress
- **URL:** http://id.loc.gov/vocabulary/countries.html
- **Updated:** April 26, 2011

- **Name:** MARC Geographic Areas
- **Creator:** Library of Congress
- **URL:** http://id.loc.gov/vocabulary/geographicAreas.html
- **Updated:** April 26, 2011

- **Name:** MARC Languages
- **Creator:** Library of Congress
- **URL:** http://id.loc.gov/vocabulary/languages.hhtml
- **Updated:** April 26, 2011

- **Name:** MARC Relators
- **Creator:** Library of Congress
- **URL:** http://id.loc.gov/vocabulary/relators.html
- **Updated:** April 26, 2011

There are about 200 different controlled lists used in the MARC 21 fixed fields. These vary from extensive lists, like the list of language codes, to the binary values like the Festschrift or the index values. Some key term lists have been published at the Library of Congress's Authorities and Vocabularies site, and it is expected that more will appear over time. Each term in the lists is individually identified with a URI, positioning these lists for use in linked data.

The lists available to date are

- MARC Countries
- MARC Relators
- MARC Geographic Areas
- MARC Languages

RDA Vocabularies

- **Name:** RDA vocabularies
- **Creator:** DCMI/JSC Working Group
- **URL:** http://rdvocab.info
- **Created:** 2009
- **Updated:** 2011

The RDA cataloging rules define some seventy different term lists for use in RDA data. Some of these overlap with MARC term lists, but there are enough differences that these can be considered distinct. All of these lists have been published in linked data format using the SKOS language, and all terms within the lists have URIs in the domain `rdvocab.info`. These terms can be used today in linked data if anyone wishes; however, the lists are still under revision by the JSC, and individual lists are being given the status of published as they are reviewed.

Name Authorities

A person's name is the public identity that is most associated with that individual. Yet personal names are not good identifiers because they are not unique. In all data dealing with people, which is quite a lot of data types, it is necessary to create an individual identity for persons and for corporate bodies, especially in bibliographic data, where these are the actors and creators of knowledge resources.

Virtual International Authority File (VIAF)

- **Name:** Virtual International Authority File
- **Creator:** OCLC
- **URL:** http://viaf.org
- **Created:** 2009
- **Updated:** 2011

Libraries have made an early entry in the Linked Data cloud in the area of name authorities (including persons, organizations, places, and works) through the Virtual International Authority File (VIAF). Nearly two dozen national libraries and specialist libraries have contributed their name authority data to VIAF, where it is merged and made available as linked data. VIAF includes name data from many European libraries as well as from libraries in Egypt and Israel. With nearly twenty million name clusters, VIAF is surely the largest dataset of controlled names in the Linked Data cloud. Names are an obvious linking point for data from nearly all communities, and VIAF links to sources of information about people and corporations such as *Wikipedia*. *Wikipedia* articles also can link to VIAF entries using the *Wikipedia* name authority template.

New York Times People and Organizations

- **Name:** New York Times Linked Open Data
- **Creator:** *New York Times*
- **URL:** http://data.nytimes.com
- **Updated:** January 13, 2010

Accurate identification of people and organizations is of paramount importance to a news organization like the New York Times. While the controlled list of names at the New York Times is much smaller than what can be found in VIAF, around six thousand, it is of no less importance to the activities of the organization that created it. The New York Times makes its authority file available as linked data, along with links to key articles about the person or organization. The Times data includes links to DBpedia and Freebase.

Preservation

Libraries have developed controlled vocabularies in areas that few others have yet ventured into. As these term lists become available in linked data formats, they could contribute to data quality in other communities that do not yet have terms for these concepts. For that to happen, of course, the library community needs to make others aware of these lists. One example where libraries lead the way is in creating standards for preservation, in particular preservation of digital resources.

LC Preservation Level Role and Preservation Events

- **Name**: Library of Congress Preservation Events
- **Creator**: Library of Congress
- **URL**: http://id.loc.gov/vocabulary/preservationEvents.html
- **Created**: June 1, 2011

- **Name**: Library of Congress Preservation Level Role
- **Creator**: Library of Congress
- **URL**: http://id.loc.gov/vocabulary/preservationLevelRole.html
- **Created**: June 1, 2011

These preservation vocabularies provide terminology for preservation activities such as decryption, replication, and migration. Digital preservation consists of emerging techniques and takes place in many different organizations and venues. Given the leadership role of libraries and archives in this area, these small vocabularies may be both helpful and educational for users who are developing preservation metadata for the first time.