Library of Congress controlled vocabularies and their application to the Semantic Web

By Corey A. Harper and Barbara B. Tillett

SUMMARY:

This article discusses how various controlled vocabularies, classification schemes and thesauri can serve as some of the building blocks of the Semantic Web. These vocabularies have been developed over the course of decades, and can be put to great use in the development of robust web services and Semantic Web technologies. The article covers how initial collaboration between the Semantic Web, Library and Metadata communities are creating partnerships to complete work in this area. It then discusses some cores principles of authority control before talking more specifically about subject and genre vocabularies and name authority. It is hoped that future systems for internationally shared authority data will link the world's authority data from trusted sources to benefit users worldwide. Finally, the article looks at how encoding and markup of vocabularies can help ensure compatibility with the current and future state of Semantic Web development and provides examples of how this work can help improve the findability and navigation of information on the World Wide Web.

KEYWORDS: Controlled vocabularies, Semantic Web building blocks, authority control

1 Introduction: Library of Congress Tools and Launching the Semantic Web

> An essential process is the joining together of subcultures when a wider common language is needed. Often two groups independently develop very similar concepts, and describing the relation between them brings great benefits… The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web.

Thus concludes Tim Berners-Lee’s seminal 2001 *Scientific American* article on the Semantic Web (Berners-Lee, Hendler & Lasilla, 2001). The concepts established here are strong ones, and the vision of a thoroughly interconnected Web of data that these concepts suggest, could prove a catalyst for the way we research, develop, interact with, and build upon ideas, culture and knowledge. The idea presented here, of independent groups working with similar concepts, is applicable to the very technologies that can serve as the Semantic Web’s underpinnings. The Semantic Web communities and library communities have both been working toward the same set of goals: naming concepts, naming entities, and bringing different forms of those names together. Semantic Web efforts toward this end are relatively new, whereas libraries have been doing work in this area for hundreds of years. The tools and vocabularies developed in libraries, particularly those developed by the Library of Congress, are sophisticated and advanced. When translated into Semantic Web technologies they will help to realize Berners-Lee's vision.

Semantic Web technologies are now, in their own right, starting to reach a state of maturity. Berners-Lee, the director of the World Wide Web Consortium (W3C) and Eric Miller, W3C Semantic Web Activity Lead, frequently describe these technologies in terms of the Semantic Web Stack (see Figure 1). Many of the components depicted as layers in the Semantic Web Stack are already in place, although not nearly as widely implemented as most Semantic Web proselytizers would like. Development on the various levels depicted in this graphic has been a long time in the making. In a recent interview with Andrew Updegrove (2005) in the *Consortium Standards Bulletin*, Berners-Lee identifies one cause of the slow rate of development, implying that each layer is dependant on the layers below it. "We were asked to hold up the query and rules work because people didn't want to start on it until the ontology work had finished, so for some we were in danger of going too fast".

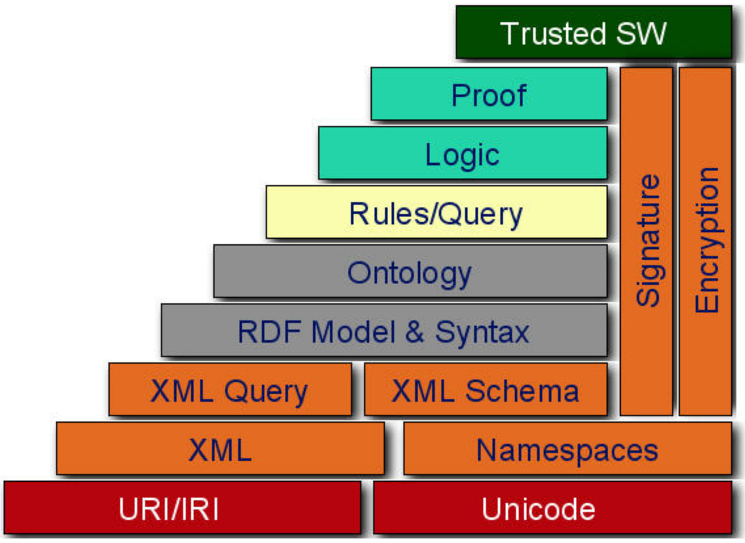


Figure 1: W3C Semantic Web Stack
Taken from W3C website, licensed under CC Attribution 2.5 License

As the technologies represented by the Semantic Web Stack continue to mature, there is a tremendous potential for the library community to play a significant role in realizing Berners-Lee's vision. In a presentation at Dublin Core 2004 (Miller, 2004, Slide 26), as well as in a number of other presentations over the years, Eric Miller implored the library community to become active in Semantic Web development. Miller outlines the role of libraries in the Semantic Web as follows:

- "Exposing collections – use Semantic Web technologies to make content available
- Web'ifying Thesaurus / Mappings / Services
- Sharing lessons learned
- Persistence"

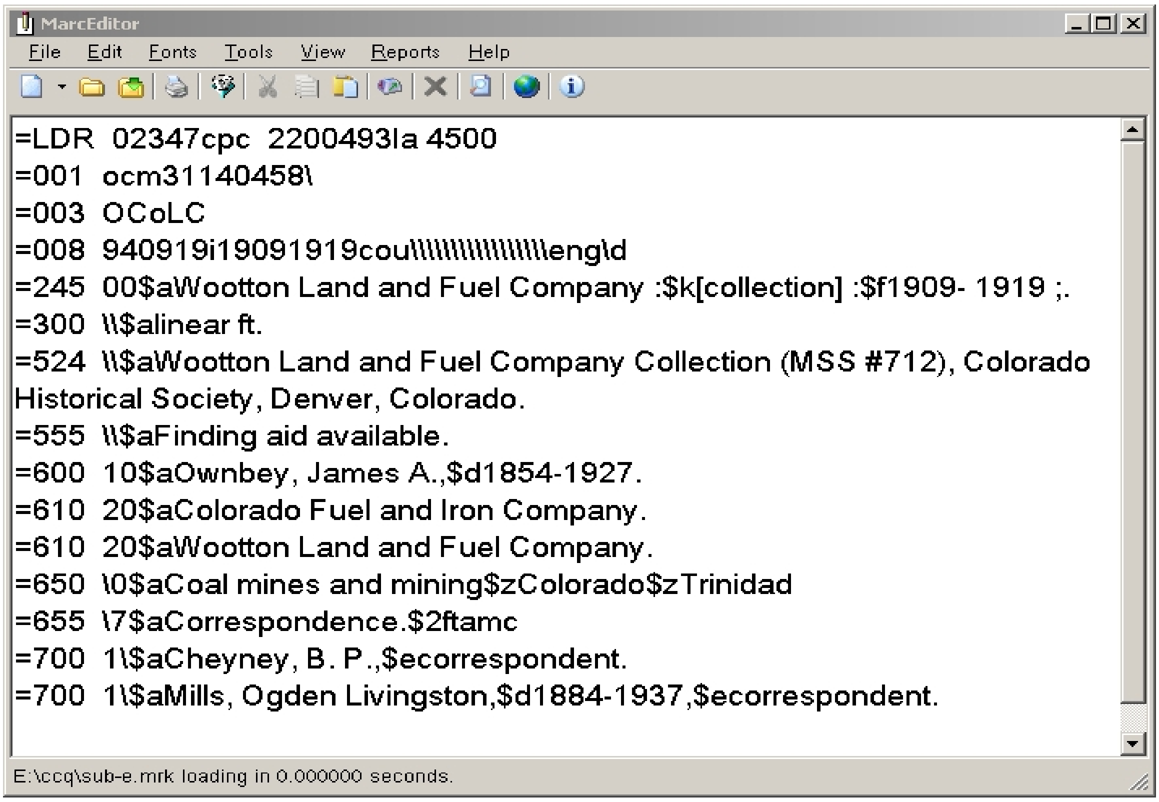
While all of these roles are significant, the idea of moving thesauri, controlled vocabularies, and related services into formats that are better able to work with other Web services and software applications is particularly significant. Converting these tools and vocabularies to Semantic Web standards, such as the Web Ontology Language (OWL), will provide limitless potential for putting them to use in myriad new ways. This will enable the integration of research functionality – such as searching and browsing diverse resources, verifying the identity of a particular resource's author, or browsing sets of topics related to a particular concept – into all sorts of tools, from online reference sources and library catalogs to authoring tools like those found in Microsoft's Office Suite.

Miller (2004, Slide 27) also emphasizes the role that libraries can play in helping to realize the trust layer in the Semantic Web Stack, stating, "Libraries have long standing trusted position that is applicable on the Web". The Semantic Web has a lot to gain by recruiting libraries and librarians and involving them in the development process. The W3C's stated mission is "to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth for the Web". This focus on protocols and guidelines helps explain why the Semantic Web Stack includes little to no mention of content. For example, it includes an ontology layer, which is primarily represented by OWL – a Web Ontology *Language* – a specification for adding Semantic Web enabling functionality to existing ontologies. More recently, another Semantic Web technology– Simple Knowledge Organization System (SKOS) Core – has been designed for the encoding of the contents of thesauri. The W3C's emphasis has been on how to encode ontologies, which fits with their stated mission. The source of ontologies and vocabularies is outside the scope of the W3C's concerns, although the usefulness of such ontologies is certainly dependant on their validity and trustworthiness. This is where Miller's thoughts on the role of libraries seem most relevant. Libraries have a long-standing history of developing, implementing, and providing tools and services that make use of numerous controlled vocabularies. Presumably, part of the process of "web'ifying thesaurus, mappings and services" involves converting existing tools into Semantic Web standards, such as OWL and SKOS. Miller, and other vocabulary experts recognize that this would be of tremendous value to Semantic Web initiatives. Taking these steps would reduce the need for Semantic Web development to revisit decisions made over the centuries that libraries have been organizing and describing content, which ties into the related idea of "sharing lessons learned".

2 Incremental Progress – Initial Developments

Progress on bridging the gap between the Semantic Web community and the library community has been underway for some time. A variety of projects are in progress, or completed, that will help to bring more of the tools that libraries develop into the Semantic Web and more general Web Services spheres. Many of these projects are being developed within the Dublin Core Metadata Initiative (DCMI), which draws heavily on both the Semantic Web and library communities, as well as a variety of other information architecture and metadata communities.

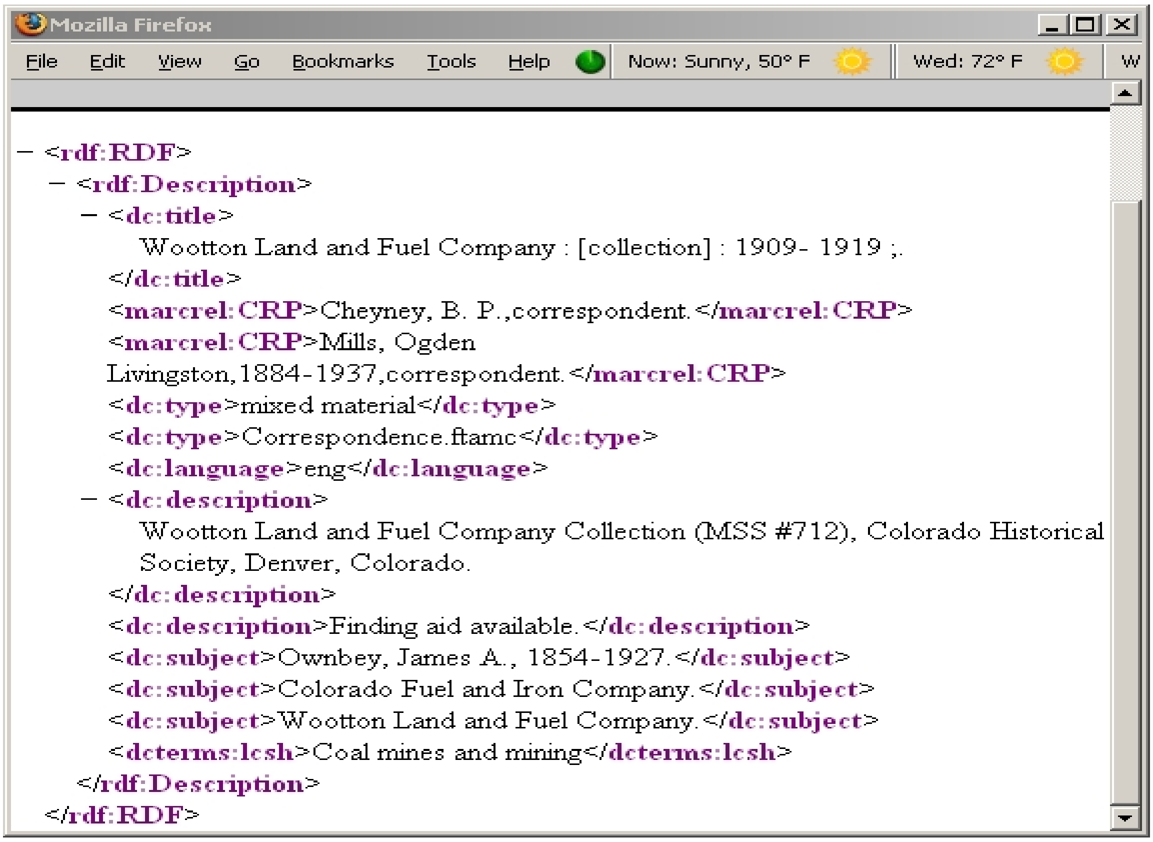
One example of such collaboration is the expression of a sub-set of MARC Relator Terms (Network Development and MARC Standards Office – Library of Congress, 2006) in RDF for use as refinements of the DC contributor element. Relator terms allow a cataloger to specify the role that an individual played in the creation of a resource, such as illustrator, calligrapher, or editor. Allowing some of these terms to be used as refinements of contributor allows the expression of much more specific relationships between individuals and the resources they create. An example of the use of MARC Relator Terms, both in a MARC record and in Dublin Core, can be seen in Figure 2. Figure 2a depicts a mnemonic MARC record with personal name added entries for correspondents. The 700 tags at the end of this record each include a subfield e, which contains a MARC Relator term identifying the role of these individuals. The corresponding Dublin Core Record in 2b represents this same information using the ‘marcrel’ namespace with the Relator code ‘CRP’, which corresponds to the term ‘correspondent.’ The concept of author added entry, represented by MARC tag 700 in 2a, is implicit in the Dublin Core example because the Relator terms are all element refinements of dc:contributor.



```
=LDR  02347cpc  2200493Ia 4500
=001  ocm31140458\
=003  OCoLC
=008  940919i19091919cou\\\\\\\\\\\\\\\\\eng\d
=245  00$aWootton Land and Fuel Company :$k[collection] :$f1909- 1919 ;.
=300  \\$alinear ft.
=524  \\$aWootton Land and Fuel Company Collection (MSS #712), Colorado
Historical Society, Denver, Colorado.
=555  \\$aFinding aid available.
=600  10$aOwnbey, James A.,$d1854-1927.
=610  20$aColorado Fuel and Iron Company.
=610  20$aWootton Land and Fuel Company.
=650  \0$aCoal mines and mining$zColorado$zTrinidad
=655  \7$aCorrespondence.$2ftamc
=700  1\$aCheyney, B. P.,$ecorrespondent.
=700  1\$aMills, Ogden Livingston,$d1884-1937,$ecorrespondent.
```

E:\ccq\sub-e.mrk loading in 0.000000 seconds.

Figure 2a: Relator Terms in and MARC (shown in MARC tag 700, subfield e) [1]



```
– <rdf:RDF>
  – <rdf:Description>
    – <dc:title>
        Wootton Land and Fuel Company : [collection] : 1909- 1919 ;.
      </dc:title>
      <marcrel:CRP>Cheyney, B. P.,correspondent.</marcrel:CRP>
      <marcrel:CRP>Mills, Ogden
      Livingston,1884-1937,correspondent.</marcrel:CRP>
      <dc:type>mixed material</dc:type>
      <dc:type>Correspondence.ftamc</dc:type>
      <dc:language>eng</dc:language>
    – <dc:description>
        Wootton Land and Fuel Company Collection (MSS #712), Colorado Historical
        Society, Denver, Colorado.
      </dc:description>
      <dc:description>Finding aid available.</dc:description>
      <dc:subject>Ownbey, James A., 1854-1927.</dc:subject>
      <dc:subject>Colorado Fuel and Iron Company.</dc:subject>
      <dc:subject>Wootton Land and Fuel Company.</dc:subject>
      <dcterms:lcsh>Coal mines and mining</dcterms:lcsh>
    </rdf:Description>
  </rdf:RDF>
```

Figure 2b: MARC Relator Terms in and DC (shown in <marcrel:CRP> tag)

Another example of collaboration can be seen in the ongoing work to bind the Metadata Object Description Schema (MODS) metadata element set (Library of Congress, 2006, June 13) to RDF and to the DC Abstract Model so that MODS terms, or alternately new DC properties derived from or related to these MODS terms, could be available to Dublin Core Application Profiles (Heery & Patel, 2000). Similar collaborations between the Institute of Electrical and Electronics Engineers Learning Object Metadata group (IEEE-LOM) (IEEE Learning Technology Standards Committee , 2002) and DCMI resulted in an RDF binding of IEEE-LOM, which essentially serves to make IEEE-LOM metadata statements and records useable within the context of the Semantic Web. A summary of this process was reported at the Ariadne Conference in 2003 (Nilsson, Palmer & Brace, 2003, November). This presentation resulted in a joint-task force between the IEEE and DCMI communities to formally map the IEEE-LOM Schema to the DCMI Abstract Model as well. Progress on this process can be tracked on the DC Education Working Group’s Wiki.(Joint DCMI/IEEE LTSC Taskforce, 2006, March). However, both of these examples apply to metadata elements and resource descriptions themselves. The progress that has been made on bringing tools from the library world into the Semantic Web has, thus far, been entirely focused on the idea of exposing library collections. Incorporating elements used in library resource descriptions into the sets of resource properties that are used to enable the Semantic Web is a large step, and enables library metadata to interoperate with Dublin Core and other RDF encoded metadata. Authority records, library thesauri, and library controlled vocabularies, if converted into formats that support Semantic Web technologies, have an even greater potential for revolutionizing the way users – and machines – interact with information on the Internet.

3 Authority Control – Core Principles

The benefits and virtues of authority control have been debated and restated for decades. When we apply authority control, we are reminded how it brings precision to searches, how the syndetic structure of references enables navigation and provides explanations for variations and inconsistencies, how the controlled forms of names, titles, and subjects help collocate works in displays, how we can actually link to the authorized forms of names, titles, and subjects that are used in various tools, like directories, biographies, abstracting and indexing services, and so on. We can use the linking capability to include library catalogs in the mix of various tools that are available on the Web. In order to enable these capabilities in a Web-based environment, it will be necessary to move the records that facilitate linking and aggregating in the library into more universal and generalized encoding. The Library of Congress is taking steps in this direction, particularly through work on MARCXML for Authority Records (Library of Congress, 2005, April 22) and the Metadata Authority Description Schema (MADS) (Library of Congress, 2005, December 14), along with crosswalks to go between them. A next logical step is to begin work on translating this data into the Resource Description Framework (RDF), which is not a trivial task.

4 Subject & Genre Vocabularies for the Semantic Web

There are a vast number of controlled vocabularies for various forms of subject access to library materials. Some of these vocabularies are classification schemes, such as *Dewey Decimal Classification* (DDC) and *Library of Congress Classification* (LCC). Others are controlled lists of subject headings or terms, which adhere to national and international guidelines for thesaurus construction, like the *Library of Congress Subject Headings* (LCSH), the two *Thesauri for Graphic Materials: Subject Terms* (TGM I) and *Genre & Physical Characteristic Terms* (TGM II), *Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc.* (GSAFD), and the *Ethnographic Thesaurus*. Some of the subject thesauri published by The Getty, such as *The Getty Thesaurus of Geographic Names* (TGN) and *The Art and Architecture Thesaurus* (AAT) are truly hierarchical thesauri.

Many of these controlled vocabularies are registered as DCMI Encoding Schemes in the DCMI Terms namespace. Registered Encoding Schemes qualifying the DC Subject element include (DCMI Usage Board, 2005, January 10):

- Dewey Decimal Classification (DDC),
- Library of Congress Classification (LCC),
- Library of Congress Subject Headings (LCSH),
- Medical Subject Headings (MeSH),
- National Library of Medicine Classification (NLM),
- The Getty Thesaurus of Geographic Names (TGN),
- Universal Decimal Classification (UDC).

However, when these vocabularies are used as values of DC Subject in metadata records outside of the library context, the syndetic structure of the source vocabulary is all but lost. There is no way for search tools or other applications to make use of information about related terms and variant forms as part of the entry vocabulary. In some cases, item metadata for Dublin Core records is included within a larger system that relies heavily on MARC data, such as OCLC WorldCat. When this is true, many of the Library of Congress controlled vocabularies are indirectly linked to these Dublin Core records by virtue of the availability of MARC authority records for the controlled vocabularies in those systems. Systems outside of these application environments need to be able to retrieve and translate or otherwise make use of MARC authority record data to make effective use of the hierarchies, equivalency relationships, and structures in the content of controlled vocabularies.

These vocabularies present tremendous potential for improving access to web resources and Semantic Web data, as well as enhancing networked applications. Search engine results could be dramatically improved, both in terms of precision and recall. Different subject vocabularies covering the same concept space could be merged together or associated, providing an environment where differences in terminology between different communities would provide less of a barrier to effective browsing of resources. Front-end interfaces could be built for a variety of online reference tools that take advantage of the rich structure of relationships between topics that is provided by controlled vocabularies. In some regards, libraries are only now realizing the full potential of catalog records to provide innovative and new browsing interfaces. An example can be seen in North Carolina State Library's new, Endeca powered, catalog interface (North Carolina State University [NCSU] Libraries, 2006), which is presently built entirely on the structure of bibliographic records. This new catalog interface allows users to refine a search by navigating through record clusters that share a particular property. Drawing on the idea of faceted classification, clusters can be grouped by subtopic, genre, language, time-period, geographic region, and in a variety of other ways. These facets are derived from information in various access points, subject headings and subdivisions in a particular bibliographic result set. It is easy to envision a similar technique being used to broaden searches and even to present initial browse interfaces to specific collections of information resources. When interfaces start to leverage the power of authority control in new and interesting ways, the benefit to users will be immense.

Another example of initiatives to leverage authority control is OCLC Research's Terminology Services project to "offer accessible, modular, Web-based terminology services" (OCLC Research, 2006). The Terminologies Pilot Project in October 2005 explored techniques for encoding sample vocabularies, means of mapping between them to help users identify relationships, and methods for incorporating the resulting services into other applications and tools. These services are the beginning of a rich tapestry of semantics that can be delivered within a user's current context, whatever that context is. In some cases, the services may be entirely carried out by server-side applications. Almost any application that serves content dynamically could include a navigation system that draws on the hierarchies and relationships between terms in LCSH, and could provide search interfaces that draw on term equivalencies to retrieve a broader set of resources when doing keyword searches.

Additionally, there could be services that exist within client side applications, pulling vocabulary structures from the network and integrating them into authoring tools. One such service component prototyped for the OCLC pilot uses the Microsoft Office 2003 Research Services Pane to access genre terms. "… if a college student wishes to categorize a reading list of fiction titles based on genre, he could copy the titles into a Microsoft Excel 2003 workbook, open the Research services pane, send a search to the OCLC Research GSAFD vocabulary service, and then place the results into his document" (Vizine-Goetz, 2004). This provides the user with the ability to browse and search genre categorizations without having to leave the application in which the search results will be used. Additionally, the more sources of data are made available in this way, the more automated the process can be. If the Research Services Pane had access to bibliographic records, genre or subject categorization could be fully automated and available at the touch of a button. A cataloger could use a similar tool to suggest appropriate terms from a controlled vocabulary, which will lead to lots of cost saving opportunities.

Similar applications could be developed as browser plug-ins or extensions. A sidebar application could be built for FireFox that could harness the power of controlled vocabularies when browsing Web resource that provide some degree of keyword tagging or folksonomy support. Imagine browsing a resource like Flickr (Flickr, 2006), and being able to query LCSH for relationships that may be defined between subject terms through the tags labeling the subject of a particular photographic image and tags for various related concepts. The sidebar could include hyper-linked broader, narrower, equivalent, and associative terms that would pull together additional photographs tagged with those related terms. In a different context, such an application could attempt to scan html source code for word frequency and try to guess the primary topics, returning related terms from a controlled vocabulary, and perhaps linking to search engine results for the concepts represented. Additionally, a similar service could be provided using a word highlighted by the user. The possibilities are endless.

5 Name Authority for the Semantic Web

The benefits of authority control described above – search precision, more powerful navigation, collocation, and linking between various tools and resources – apply to metadata about the creators of resources as well as to subject access. The library community is well positioned to play a significant role in these developments. Libraries have been dealing with identification, disambiguation, and collocation of names of content creators since the beginning of cataloging. The different forms of name used by the same creator in various print publications and other types of resources have always led to some degree of difficulty in grouping works together. The syndetic structure of name authority files has proven itself a very useful tool to help collocate works by an author regardless of the form of name on a particular item. The proliferation of resources on the Web extends the scope of the collocation problem.

Initiatives are appearing in the Web community to help provide better mechanisms for identifying persons, families, and corporate entities that have a role with respect to information resources. InterParty is a European Commission funded project exploring the interoperation of "party identifiers," which would provide standard identification numbers to serve the same purpose that authorized forms of names serve in library applications: to help collocate and disambiguate individual content creators (Information Society Technologies, 2003). More recently, some of the InterParty members and others submitted a similar proposal for International Standard Party Identifiers (ISPI) as an ISO standard (Lloret & Piat, 2006). More directly related to Semantic Web Development is the Friend of a Friend (FOAF) project. FOAF is about "creating a Web of machine-readable homepages describing people, the links between them and the things they create and do" (The Friend of a Friend [foaf] Project, n.d.). Finding ways to integrate these initiatives with existing mechanisms for name authority control in libraries can help to bring library catalogues into the mix of tools available on the Web. Additionally, the availability of library authority data in a more Web-friendly format has the potential to positively influence the organization of the broad spectrum of Web content already available. The development of a virtual international authority file (VIAF) has been a key idea moving forward this initiative.

A Virtual International Authority File (VIAF)

Presently, authority files are maintained and developed by a large range of national bibliographic agencies. To make the most of this potential, it is useful to first integrate the somewhat disparate sources of authority data that exist even within the library community. These agencies develop files that are generally focused on the creators of content relevant to a particular national and cultural identity. As geographic boundaries become more and more porous, and culture becomes much more international, there is increasingly overlap between the authority files maintained by these agencies. The concept of a Virtual International Authority File (VIAF) has been discussed since the 1970's within the International Federation of Library Associations and Institutions (IFLA). Initially, IFLA envisioned a single shared file; more recently the concept has evolved into one of linking existing national and regional authority files. The primary objective of this vision is to facilitate sharing the workload and reducing cataloging costs within the library community. The community is expanding, especially in Europe, where libraries are viewed as one of many "memory institutions", along with archives, museums, and rights management agencies. Ideally, authority files can be freely shared among all of these communities. A shared file would reduce the cost of doing authority work by avoiding repetition of effort while combining various forms of names that are particular to resource published within the context of a particular region, culture, or nation. Combining or clustering the forms of name will result in a much richer set of authority information, enabling users to access information in the language, scripts, and form they prefer. Additionally, a single international authority file system will be far more useful when integrated into various Web retrieval tools and Web content descriptions. Such a tool could be used by a wide range of Web systems to improve the precision of user's searches and to provide the user's preferred display of the language and script of names.

Authority records are used to collocate resources that utilize varying forms of name, but collocation does not need to dictate the script or language used by the end user display. Figure 3 shows how a single entity – whether it is a concept, person, place, or thing – has a variety of labels that identify it in different languages and scripts. Traditionally, library systems have relied on a preferred label for all entities, although the preference would vary depending on the geographic context of the system. Merging or linking records that utilize different languages and scripts allows the end-user to select the language and script used to display information about entities irrespective of system's default preference. This is appropriate in a truly global Web, where geographic and national boundaries are considerably less significant.

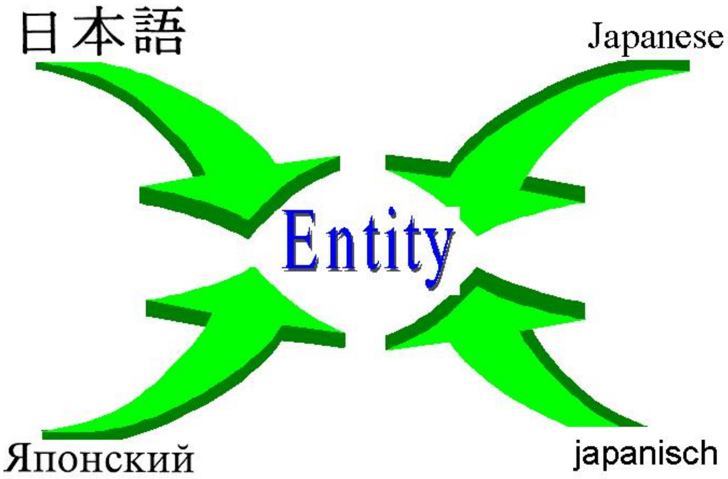


Figure 3: One Entity - Many Labels[2]

A variety of projects have sought to address this language challenge in recent years, exploring mechanisms to combine individual authority files. One such project, the European Commission funded AUTHOR project converted a sample of authority records from five European bibliographic agencies in France, England, Belgium, Spain, and Portugal into the UNIMARC format and made them available as a searchable file. “The challenge was that each library has its own language, cataloging rules, bibliographic record format and local system for its online authority file” (Tillett, 2001). Combining these records into a single UNIMARC file required a large amount of record normalization. No attempt was made to link the records for the same entity.

More recently, OCLC Online Computer Library Center Inc., the Library of Congress, and Die Deutsche Biblioteck (DDB) began a joint project to test the idea of a VIAF. OCLC used matching algorithms to link name authority records of these two national bibliographic agencies and built a server to store the combined records. Additional phases of the project will involve ongoing maintenance to update the central file when either source is updated and possibly the development of a multilingual end user interface (Morris, 2003). Following the evaluation of the project, the addition of new partners will be explored, particularly those potential partners with non-roman authority records.

6 Leveraging a VIAF on the Semantic Web

When combined with developments in the broader metadata, Web-design, and Semantic Web communities, the power and utility of VIAF outside of libraries becomes clear. Authority record data can be associated more easily with a variety of Web resources, allowing users and potentially machines to immediately start to evaluate the information they are looking at. A quick search of bibliographic data related to a given resource author allows the retrieval of her dissertation, which could be mined for data about the degree granting institution. Other bibliographic records could be retrieved to help evaluate the original Web resource and related works could offer pathways to additional relevant resources.

As other resources start including metadata that uses identifiers or headings to link to a VIAF, the opportunity to connect more interesting bits of information can add significant value to any Web-based information resource. Wikipedia entries, journal articles, Who’s Who biographical info, an individual's blog, their homepage, or the homepage of their place of work can all be interconnected, as well as linked to journal articles, bibliographic records in catalogs and in e-commerce sites, and a variety of other scholarly resources. These interconnections have extensive implications for research. Once there is a corpus of biographical information combined into a data store that is connected to authority data (as well as associated bibliographic data), the information can be used to make inferences about any document, article, Web page, or blog entry that turns up when searching for information. For example, imagine a blog post that includes information about its author's identity. This information could be referenced against available biographical data and used to make inferences about the veracity and objectivity of the post's content. If the author were affiliated with the Recording Industry Association of America (RIAA) a trade group representing the U.S recording industry, or the Electronic Frontier Foundation (EFF), a non-profit legal organization focused on defending "digital rights", heavily involved in fighting bad uses of Digital Rights Management technology, and opposing limitations on fair use, the statement is likely to be much less objective than content posted by a Harvard law professor. While an agent or search tool couldn't necessarily flag such resources as potentially biased without additional information about the affiliate organization, it would still be very useful to present these additional facts to the user when returning search results.

The metadata community has a number of initiatives underway for describing people, both as agents of resource creation and for the utility of describing relationships between people, describing connections between people and organizations, and for capturing other contact information and other descriptive information about individuals and groups.

Examples of such initiatives include the Dublin Core Agents Working Group's work on defining a metadata standard for agent description and identification. Ultimately, this work should result in the development of an Application Profile for agent description. Also, work has been completed on "Reference Models for Digital Libraries: Actors and Roles" within the

DELOS/NSF Working Group. This work culminated in Final Report, issued in July 2003. Interestingly, this work goes well beyond the scope of authors and content creators; instead the DELOS/NSF model categorizes Actors as Users, Professionals, and Agents. In this context, agents are the traditional content creators that help populate a digital library and professionals are the developers of the digital library itself and the providers of digital library services. The scope of their work may be much deeper than is necessarily relevant to discussions of authority control on the Semantic Web, although the models used and conclusions drawn may prove to play an important role in future discussions about authority control for names.

Another initiative, the Friend Of A Friend (FOAF) project described earlier, is of particular use to the Semantic Web community, because it was conceived with the Semantic Web in mind and is built upon the RDF data model. FOAF expresses identity through any property value pair, allowing you to aggregate data about individuals using any unique property, such as an email address or the URL for a home page. FOAF is primarily designed for community building, but when the possible privacy issues of sharing personal data are resolved there is much potential for FOAF to help aggregate public information about individuals.

FOAF could be used to aggregate all sorts of resources, both by and about individuals—Again, resources such as, Wikipedia entries, journal articles, Who’s Who biographical info, the individual’s homepage, their blog, or their place of work. This information could be glommed into a program like Piggy Bank (Simile, 2006) – a FireFox extension for viewing, collecting and merging RDF encoded data, developed by the Simile (Semantic Interoperability of Metadata and Information in unLike Environments) project – or any other Semantic Web enabled tool, and processed along with other local or remote data stores. For example, a local data store of vcards and/or FOAF data might include private data, such as phone number, calendaring system, and email address. The very presence of that particular identifier in a local store of RDF data about people might change the context of any information interaction with a Wikipedia or Who’s Who entry.

Using any "unique" property to identify entities will certainly help to aggregate most of them, but it would miss entities that are not described using a particular piece of identifying information. Inferences about which descriptions represent the same entity will only go so far in establishing a positive match. The ability to completely aggregate such data becomes even more powerful when the identifying properties can be referenced against the VIAF to determine alternate forms of name and representations in alternate scripts. This allows a query of available RDF data to be much more comprehensive when deciding what pieces of data should be aggregated.

7 Markup and Encoding of Authority Data

One of the most valuable activities that libraries and librarians can engage in, both to help realize the Semantic Web and to generally increase the findability of electronic resources in general, is the process of creating versions of vocabularies in machine-readable format. Thesauri, authority files, classifications schemes, and subject heading lists – collectively referred to as Knowledge Organization Systems – have enormous potential for enhancing the discoverability and organization of resources in a networked environment. The potential only increases when such systems are provided in formats designed for emerging Web technology, such as OWL and SKOS – two of the ontology schema of the Semantic Web.

> Knowledge organization systems can enhance the digital library in a number of ways. They can be used to connect a digital library resource to a related resource. The related information may reside within the KOS itself or the KOS may be used as an intermediary file to retrieve the key needed to access it in another resource. A KOS can make digital library materials accessible to disparate communities. This may be done by providing alternate subject access, by adding access by different modes, by providing multilingual access, and by using the KOS to support free text searching. (Hodge, 2000)

The perceived benefits of knowledge organization systems listed above are of particular importance, and apply to networked resources beyond the scope of digital library materials. The availability of machine-readable representations of various thesauri and other controlled vocabulary enables more effective search and retrieval, better browse functionality and general organization of materials online, and the automatic creation of context sensitive linkages between available resources and data sets. Additionally, and most importantly for Semantic Web development, providing controlled vocabularies outside of traditional library systems enables a variety of applications to more effectively merge and manipulate data and information from disparate sources.

There are many steps that need to be taken to realize this set of goals. Firstly, the library and Semantic Web communities must agree on how best to encode these vocabularies. Many of the subject schemes listed above already exist in one or more machine-readable representations. Much of LCSH exists in MARC records in library catalogs and the databases of cataloging services such as OCLC’s WorldCat and RLIN 21. All of the LC controlled vocabularies (LC classification, LCSH subject authority, and the name authority records) are available as complete files of MARC 21 or MARCXML formatted authority records through the Cataloging Distribution Service of the Library of Congress (2005) (free test files are also available). The Getty (2003) provides licensed access to its vocabularies in three formats: XML, relational tables and MARC, and provides sample data from each vocabulary for free. At first glance, these formats do not appear to be of much use in the context of RDF, but the standardization and global use of MARC makes it possible to convert these into RDF-friendly data. One approach is the creation of URIs to identify each terminal node on the source XML structure as a unique concept that can be used as a property in the RDF and DC Abstract Model sense. This prospect has the potential to retain as much of the detail available in the source format as possible, but may prove unnecessary and undesirable due to the complexity of the resultant sets of properties. On the other hand, the information from MARC records may in fact be useful for automatic processes and machine activities. The richness of MARC authority data, and the time and effort invested in developing, encoding, and sharing this data, provides a unique and powerful set of vocabularies. However, it remains to be seen whether the complexity of the resulting XML records would be an impediment to interoperability.

An alternative approach involves simply cross-walking the XML data into an already defined RDF-friendly form. Along the way, detail about relationships between terms will likely be lost, but the end product will probably be much simpler to work with. One possible target RDF vocabulary, the Simple Knowledge Organization System (SKOS) Core (W3C Semantic Web Activity, 2004, February), has much potential in this context. SKOS Core provides a model for expressing the structure of what they refer to as a ‘Concept Scheme’. “Thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary are all examples of concept schemes” (Miles, Mathews, Wilson & Brickley, 2005, September). SKOS Core provides a means of expressing most of the semantic relationships included in most library subject vocabularies. For example, “prefLabel” and “altLabel” represent “use” and “use for” references, while “broader” and “narrower” are used to identify hierarchical relationships. SKOS allows for the creation of new labels and the encoding of more specific types of relationships as well. Additionally, SKOS provides mechanisms for various types of notes – scopeNote, definition, example and note. The SKOS Core community has drafted documentation on “Publishing a Thesaurus on the Semantic Web,” (Miles, 2005, May) which provides a guide for using SKOS to both describe and encode vocabularies in RDF.

Converting large controlled vocabularies into RDF data is certainly a good way to get sample data sets to use to build prototype services. However, the long-term maintenance of such data stores may be problematic. As changes are made to the source vocabulary, those changes need to be propagated through to all the various formats that the vocabulary is made available in. In the context of SKOS, this is a manageable task. SKOS extensions have been proposed to allow for versioning and the tracking of changes made to controlled vocabularies over time (Tennis, 2005). However, in cases where vocabularies are likely to be managed in a variety of different formats, the SKOS extensions are less helpful. Another approach is to harvest the source data in its native format and translate it into a variety of output formats either as nightly batch processes or on-the-fly as data is requested. Progress is being made in this area, such as OCLC’s work on exposing vocabularies in a variety of delivery formats, including MARCXML and SKOS (Dempsey et al., 2005).

It doesn’t matter whether the data store is MADS, MARC, RDF, or some arbitrary flavor of XML, it can be transformed into another format either on-the-fly or as a batch process. A centralized name authority database could be created from the national authority files available in various library communities and stored as normalized MARC 21. If this data store is deemed useful to the Friend of a Friend (FOAF) community, the data can be turned into RDF. Similarly, Web services like Flickr could convert and makes use of Library of Congress Subject headings to augment both the searching and development of their folksonomies.

8 Conclusion

Berners-Lee suggested that, “The vast bulk of data to be on the Semantic Web is already sitting in databases … all that is needed [is] to write an adapter to convert a particular format into RDF and all the content in that format is available” (Updegrove, 2005). The data, metadata, and thesauri available in various library databases and systems present a unique opportunity to take a large step forward in the development of the Semantic Web.

The realization of the Semantic Web vision, which isn’t too far from Berner-Lee’s original vision for the World Wide Web itself, involves a remarkably broad set of goals. Part of the Semantic Web vision is about aiding resource discovery by creating tools to help searcher’s refine and develop their searches, and to aid in the navigation of search results. These improvements will be augmented by the improved metadata that will result from making these same tools and vocabularies available to resource authors. Information professionals are too few in numbers to describe and catalog all of the Web's resources. Resource authors will have to play an active role in describing the materials they publish, perhaps having their descriptions refined and further developed by automated processes and by information professionals. Research in this area has been taking place in a variety of author communities, including scientific, government, and educational institutions (Greenberg & Robertson, 2002). Such collaborative efforts need not be limited to authors and metadata professionals. Other domain experts can add further descriptive information through the process of tagging and reviewing materials. This marriage of the folksonomy and controlled vocabularies would serve as a step towards what Peter Morville has elegantly referred to as "the Sociosemantic Web" (Morville, 2005).

Another large part of the Semantic Web vision is about enabling "agents" or systems to insert a searcher's/user's individual context or perspective into a search for information. This necessarily involves interacting with the elements that make up that context, such as schedules, contacts, group membership, profession, role, interests, hobbies, location, etc. Systems can then be developed that "understand" the searcher's needs, based on who the searcher is and the searcher's "context" or demographics. Developing this kind of machine understanding involves encoding the vast wealth of information available electronically in such a way that it can be negotiated according to a searcher's individual "context". Even if privacy issues hinder our ability to automate the incorporation of personal information into the "context" of a specific information interaction, there is still a tremendous amount of value in making external information sources more readily accessible for machine processing, and making the information more interoperable, easier to interpret and ultimately combined and used in novel and interesting ways. More importantly, even if it is possible to automate the user side of this process, there will undoubtedly be a user base that chooses not to trust these context-dependant decisions to Semantic Web "agents" described in Berners-Lee's writings. In the case of either of these two scenarios, the tools that support Semantic Web technology will still make most searchers' experiences more pleasant and much less frustrating.

Acknowledgements

The authors are indebted to the sound boarding, feedback, and editorial input of many people, and would particularly like to thank Lori Robare and Jane Greenberg for their input.

NOTES → REFERENCES

[1] MARC Record Shown in MarcEdit, freely distributed MARC Record editing software. Available online at: http://oregonstate.edu/~reeset/marcedit/html/

[2] Figure from Tillett, Barbara B. "Authority Control: State of the Art and New Perspectives," co-published simultaneously in *Cataloging & Classification Quarterly*, v. 38, no. 3/4, 2004, p. 23-41; and *Authority Control in Organizing and Accessing Information: Definition and International Experience* (ed.: Arlene G. Taylor and Barbara B. Tillett). Haworth Press, 2004, p. 23-41 (figure on p. 34).

REFERENCES

Berners-Lee, T., Hendler, J., & Lasilla, O. (2001). The Semantic Web [Electronic version]. Scientific American, 284(5), 34-43. Retrieved April 15, 2002, from: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Cataloging Distribution Service – Library of Congress. (2005). MARC Distribution Services: Your Source for Machine Readable Cataloging Records via FTP. Retrieved March 22, 2006, from: http://www.loc.gov/cds/mds.html

DCMI Usage Board. (2005, January 10). DCMI Metadata Terms. Retrieved March 22, 2006, from: http://dublincore.org/documents/dcmi-terms/

Dempsey, L., Childress, E., Godby, C.J., Hickey, T.B., Houghton, A., Vizine-Goetz, D., & Young, J. (2005). Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape. Forthcoming in LITA guide to e-scholarship (working title), ed. Debra Shapiro. Retrieved April 15, 2006, from: http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf

Flickr. (2006). Popular Tags on Flickr Photo Sharing. Retrieved April 1, 2006, from: http://www.flickr.com/photos/tags/

The Friend of a Friend (foaf) Project. (n.d.). Retrieved April 12, 2006, from: http://www.foaf-project.org/

The Getty. (n.d.). Obtain the Getty Vocabularies. Retrieved March 22, 2006, from: http://www.getty.edu/research/conducting_research/vocabularies/license.html

Greenberg, J. & Robertson, D.W. (2002) Semantic Web Construction: An Inquiry of Authors' Views on Collaborative Metadata Generation. In: Metadata for e-Communities: Supporting Diversity and Convergence. Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002, Florence, Italy. October 13-17. Retrieved April 15, 2006, from: http://www.bncf.net/dc2002/program/ft/paper5.pdf

Heery, R. & Patel, M. (2000). Application profiles: mixing and matching metadata schemas." Ariadne, 25 Retrieved April 24, 2006, from: http://www.ariadne.ac.uk/issue25/app-profiles/

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries. The Digital Library Federation. Retrieved April 12, 2006, from: http://www.clir.org/pubs/reports/pub91/contents.html

IEEE Learning Technology Standards Committee. (2002, July). Draft Standard for Learning Object Metadata. Retrieved March 22, 2006, from: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Information Society Technologies. (2003). InterParty. Retrieved April 12, 2006, from: http://www.interparty.org/

Joint DCMI/IEEE LTSC Taskforce. (2006, March). DCMI Education Working Group Wiki. Retrieved March 22, 2006, from: http://dublincore.org/educationwiki/DCMIIEEELTSCTaskforce

Library of Congress. (2005, April 22). MARCXML – MARC 21 XML Schema: Official Web Site. Retrieved April 15, 2006, from: http://www.loc.gov/standards/marcxml/ (includes Test authority data for Classification, Names, and Subjects).

Library of Congress. (2005, December 14). MADS – Metadata Authority Description Schema: Official Web Site. Retrieved April 15, 2006, from: http://www.loc.gov/standards/mads/ (includes MARCXML Authorities to MADS crosswalk).

Library of Congress. (2006, June 13). MODS: Metadata Object Description Schema. Retrieved June 30, 2006, from: http://www.loc.gov/standards/mods/

Lloret, R., & Piat, S. (2006). Outline for ISO Standard ISPSI (International Standard Party Identifier). Retrieved April 12, 2006, from: http://www.collectionscanada.ca/iso/tc46sc9/docs/sc9n429.pdf

Miles, A. (2005, May). Quick Guide to Publishing a Thesaurus on the Semantic Web: W3C Working Draft 17 May 2005. Retrieved April 15, 2006, from: http://www.w3.org/TR/2005/WD-swbp-thesaurus-pubguide-20050517/

Miles, A., Mathews, B., Wilson, M., & Brickley D. (2005, September) SKOS Core: Simple Knowledge Organisation for the Web In: Proceedings of the International Conference on Dublin Core and Metadata Applications, Madrid, Spain, 12-15 September 2005. p. 5-13. Retrieved April 15, 2006, from: http://www.slais.ubc.ca/PEOPLE/faculty/tennis-p/dcpapers/paper01.pdf

Miller, E. (2004, October). The Semantic Web and Digital Libraries. Keynote presentation from The International Conference on Dublin Core and Metadata Applications, 2004. Shanghai, China, 11-14 October 2006. PowerPoint presentation retrieved April 1, 2006 from: http://dc2004.library.sh.cn/english/prog/ppt/talk.ppt

Morris, S. (2003, September). Virtual International Authority [press release]. Retrieved April

15, 2006, from: http://www.loc.gov/loc/lcib/0309/authority.html

Morville, P. (2005). Ambient Findability. Sebastopol, CA: O'Reilly.

Network Development and MARC Standards Office – Library of Congress. (2006, June 23). MARC Code Lists for Relators, Sources, Description Conventions. Retrieved June 30, 2006, from: http://www.loc.gov/marc/relators/

Nilsson, M., Palmer, M. & Brase, J. (2003, November). The LOM RDF binding – principles and implementation. Paper presented at The 3rd Annual Ariadne Conference, Leuven, Belgium. Retrieved March 22, 2006, from : http://rubens.cs.kuleuven.ac.be/ariadne/CONF2003/papers/MIK2003.pdf

North Carolina State University (NCSU) Libraries. (2006). Endeca at the NCSU Libraries. Retrieved March 6, 2006, from: http://www.lib.ncsu.edu/endeca/

OCLC Research. (2006). Terminology Services. Retrieved March 22, 2006, from: http://www.oclc.org/research/projects/termservices/

Tennis, J. (2005). SKOS and the Ontogenesis of Vocabularies. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, Madrid, Spain, 12-15 September 2005. Retrieved April 15, 2006, from: http://purl.org/dcpapers/2005/Paper33

Tillett, B. (2001). Authority control on the Web. In Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium : Confronting the Challenges of Networked Resources and the Web, Washington, D.C., November 15-17, 2000. Sponsored by the Library of Congress Cataloging Directorate. Edited by Ann M. Sandberg-Fox. Washington, D.C. : Library of Congress, Cataloging Distribution Service, p. 207-220. Retrieved April 15, 2006, from: http://www.loc.gov/catdir/bibcontrol/tillet_paper.html

Simile. (2006). Piggy Bank. Retrieved March 22, 2006, from: http://simile.mit.edu/piggy-bank/

Updegrove, A. (2005, June). The Semantic Web: An Interview with Tim Berners-Lee. Consortium Standards Bulletin, 5(6), Retrieved February 9, 2006, from: http://www.consortiuminfo.org/bulletins/semanticweb.php

Vizine-Goetz, D. (2004). Terminology services: Making knowledge organization schemes more accessible to people and computers. OCLC Newsletter, 266. Retrieved March 22, 2006, from http://www.oclc.org/news/publications/newsletters/oclc/2004/266/

W3C Semantic Web Activity. (2004, February) Simple Knowledge Organisation System (SKOS). Retrieved March 22, 2006, from: http://www.w3.org/2004/02/skos/