9-2011

Using SKOS to Express Faceted Classification on the Semantic Web

Theresa Putkey
San José State University, tputkey@gmail.com

Using SKOS to Express Faceted Classification on the Semantic Web

Theresa Putkey
San José State University
San José, CA USA

This paper looks at Simple Knowledge Organization System (SKOS) to investigate how a faceted classification can be expressed in RDF and shared on the Semantic Web.

Statement of the Problem

Faceted classification outlines facets as well as subfacets and facet values. Hierarchical relationships and associative relationships are established in a faceted classification. RDF is used to describe how a specific URI has a relationship to a facet value. Not only does RDF decompose "information into pieces," but by incorporating facet values RDF also given the URI the hierarchical and associative relationships expressed in the faceted classification. Combining faceted classification and RDF creates more knowledge than if the two stood alone. An application understands the subject-predicate-object relationship in RDF and can display hierarchical and associative relationships based on the object (facet) value. This paper continues to investigate if the above idea is indeed useful, used, and applicable. If so, how can a faceted classification be expressed in RDF? What would this expression look like?

Literature Review

This paper used the same articles as the paper *A Survey of Faceted Classification: History, Uses, Drawbacks and the Semantic Web* (Putkey, 2010). In that paper, appropriate resources were discovered by searching in various databases for "faceted classification" and "faceted search," either in the descriptor or title fields. Citations were also followed to find more articles as well as searching the Internet for the same terms. To retrieve the documents about RDF, searches combined "faceted classification" and "RDF, " looking for these words in either the descriptor or title.

Research was expanded for this article to include Simple Knowledge Organization System (SKOS) from the W3C.

Research Question(s)

To continue research from the survey paper, following questions will be answered:

- How do faceted classification and RDF work together?
- What is an example of a faceted classification expressed in RDF?
- What are some use cases for it?
- What challenges are there for sharing faceted classification via the semantic web?

Methodology

Based on information from research papers, more research was done on SKOS and examples of SKOS and shared

faceted classifications in the Semantic Web and about SKOS and how to express SKOS in RDF/XML. Once confident with these ideas, the author used a faceted taxonomy created in a Vocabulary Design class and encoded it using SKOS. Instead of writing RDF in a program such as Notepad, a thesaurus tool was used to create the taxonomy according to SKOS standards and then export the thesaurus in RDF/XML format. These processes and tools are then analyzed.

Results

The initial statement of the problem was simply an extension of the survey paper done earlier in this class. To continue on with the research, more research was done into SKOS – a standard for expressing thesauri, taxonomies and faceted classifications so they can be shared on the semantic web.

Defining Faceted Classification

As noted by Louie et al. (2003), information doesn't always fit into well-defined hierarchies that users instinctively know how to browse. In information retrieval, designers don't naturally know how to structure information so it can be found by others. As with any other classification, faceted classification aims to organize objects so they can be retrieved. In faceted classification, there are facets, subfacets (also called arrays) and facet values. Fagan states,

Hearst defines facets as a [*sic*] "a set of meaningful labels organized in such a way as to reflect the concepts relevant to a domain." LaBarre defines facets as representing "the categories, properties, attributes, characteristics, relations, functions or concepts that are central to the set of documents or entities being organized and which are of particular interest to the user group. (2010, p. 58)

Facets are the overall containers for the rest of the values. A facet can contain subfacets that further divide the values a facet might have. Using subfacets makes the faceted classification into a hierarchical faceted classification. Finally, each facet or subfacet has values that are applied to documents.

Broughton defines faceted classification as "…adequate object description (labeling [*sic*] the items to support subject retrieval), providing search tools that support browsing, navigation and retrieval, and, to a more limited extent, the presentation of results" (2006, p. 50). Broughton states that faceted classification helps to: synthesize the complexity of a subject; provide a consistent, logical and regular syntax and structure which can be used by computers; be used in a user interface on a computer or on the Internet; be easily converted into a thesaurus or subject headings; and provide a tool for browsing (2006).

Faceted classification has some standards that need to be followed in order for the classification to be truly faceted. The classification should have the following:

> "the display of useful generic relationships; full and accurate cross-referencing; accurate application of principles of division; a clear citation order; established rules for compounding; and an appropriate notation" (Broughton, 2006, p. 50).

As well as these attributes, an important point is that each facet and its values must stand alone and cannot overlap with another facet. The facets must be mutually exclusive. The values within a facet must be exhaustive. "If the analysis is accurate there should be no difficulty about this. Enumerative systems on the other hand often produce groupings of classes that are not mutually exclusive, and that is a sure sign of a "non-faceted" structure" (Broughton, 2006, p. 54). This value would not be repeated elsewhere in another facet. If a thesaurus or classification system states that it is faceted, but re-uses the same value in two different facets, it should not be considered truly faceted.

Faceted Classification Example

Below is a faceted taxonomy created in the author's Vocabulary Design class. In this class, the work product was referred to as a faceted taxonomy instead of a faceted classification. The same faceted taxonomy term is used here, recognizing that the faceted taxonomy does not have the notation standard that a true faceted classification would have. In this way, the term "faceted taxonomy" is used in a looser way than one would use "faceted classification." The faceted taxonomy is given here for reference. This taxonomy is meant for an online boutique retailer selling home furnishings and decorations. After this taxonomy is displayed, Semantic Web and RDF are defined.

| Product Facet | Brands Facet | Occasion Facet |

Subfacet	Subfacet	Subfacet	Subfacet	Subfacet	Brands	Collection	Anniversary
Ambiance					Ana	Collection	Anniversary
	Candles				Aquiesse		Baby Shower
		Accompaniments			Archipelago Botanicals		Birthday
			Lighter				Everyday
			Snuffer			AB Home	Graduation
			Wick Trimmer			Baroque	Holiday
		Ceramic				Boticario De Havana	Housewarming
		Jar				Excursion	Mother's Day
		Multi-wick				Private Reserve	Romance
		Pedastal				Royal	Sympathy
		Taper				Signature	Thank You
		Tea			Doudou et Compagnie		Wedding
		Tin/Metal			Ellington		
		Travel					
			Mini			Hailey	
			Regular			Hepburn	
		Votive				Ramoli	
	Diffusers					Sophie	
		Oil Diffuser				Sorrento	
			Burner			Stella	
			Oil			Territory	
				Ceramic	Hobo International	Vero	
				Soapstone		Mixed Media	
				Terracotta		Soft Pebble	
						Vintage	

Tables

- Coffee Tables
- Side Tables

Home Decor

- Candle Holders
- Fabric by the Yard
  - Silk
  - Velvet
- Lighting
- Magnets
- Rugs
- Wall Decor
  - Art
  - Colorwash

Linens

- Blankets
  - Bed Blankets
  - Stadium Blankets
  - Throw Blankets
  - Travel Blankets
- Kitchen Linens
  - Runners
  - Tablecloths

Table Top

- Napkin Rings

Place Card Holders

Runners

Serving Dishes

Tablecloths

Trivets

Utensils

Defining the Semantic Web and RDF

The Semantic Web concentrates on relationships and connections between items. Applications can know something about data.

The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources... It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing. (W3C, http://www.w3.org/2001/sw/)

For the Semantic Web to work, relationships and bridges need to be built between applications. Since the same language is not used in each application, a common language and standard needs to be created so applications can exchange information with each other.

The Semantic Web defines these relationships using the Resource Description Framework (RDF). "RDF is intended for situations in which this information needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning" (W3C, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/). RDF helps applications exchange information even if their underlying structures are different.

Indeed, one of the main driving forces for the Semantic [W]eb, has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed [*sic*] by machines. RDF's serialization format – its syntax in XML – is a very suitable format for expressing relational database information. (Berners-Lee, http://www.w3.org/DesignIssues/RDFnot.html)

RDF can be used to assign attributes and values to resources on the Internet and to express relationships between items. RDF allows computers to know something about a subject. "RDF is nothing more than a general method to decompose information into pieces. The emphasis is on general here because the same method can be used for any type of information" (Tauberer, http://www.rdfabout.com/intro/?section=3). RDF requires a subject (or resource), predicate (or relationship) and object (or value). "RDF tools are ignorant of what these names mean, but they can still usefully process the information" (Tauberer, http://www.rdfabout.com/intro/?section=3). RDF may not know what the names mean, but holds information to allow applications to communicate with other applications to display the information to the end-user, who is then able to make sense of the information.

Simple Knowledge Organization System (SKOS)

SKOS is a W3C standard for entering thesauri, taxonomies and classification schemes in an RDF format. This format makes these thesauri machine-readable and shareable between applications.

The Simple Knowledge Organization System (SKOS) is an RDF vocabulary for representing semi-formal *knowledge organization systems* (KOSs), such as thesauri, taxonomies, classification schemes and subject heading lists. Because SKOS is based on the Resource Description Framework (RDF)…these representations are machine-readable and can be exchanged between software applications and published on the World Wide Web. (http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/)

SKOS gives a standard for entering concepts, concept schemes, broader, narrower and related terms, as well as

scope notes and definitions. A typical scenario for using SKOS is: a library, such as the Library of Congress, wants to share their classification scheme with other organizations. Instead of sending it off in a Word document and expecting someone on the other end to put it into another computer system, the classification scheme can be sent in RDF format and imported into an application. Another scenario might be creating crosswalks between taxonomies. For example, a content publisher may want to aggregate content on a website. Although this content publisher has its own content and classifies it with its own taxonomy, other websites from which the content publisher pulls information use different taxonomies. All articles about cats and felines need to be displayed together, so the content publisher can map its term "cat" to the other taxonomy's term "felines." The content publisher displays all articles about cats, whether they are classified with the terms "cats" or "felines."

SKOS uses certain terms to identify everything in a taxonomy in RDF format. Using a portion of the faceted taxonomy from above, here's how SKOS sees the values in the taxonomy:

| Faceted Taxonomy				SKOS Mapping	
Product Facet				Property:Concept Scheme	
Subfacet	Subfacet	Subfacet	Subfacet		
Ambiance				Property:Top Concept	
	Candles			Property: Concept	
				Is also: NT of Ambiance	
		Accompaniments		Property: Concept	Is also: BT of Lighter, Snuffer and Wick Trimmer
				Is also: NT of Candles	
			Lighter	Property: Concept	Is also: RT of Snuffer and Wick Trimmer
				Is also: NT of Accompaniments	
			Snuffer	Property: Concept	Is also: RT of Lighter and Wick Trimmer
				Is also: NT of Accompaniments	
			Wick Trimmer	Property: Concept	Is also: RT of Lighter and Snuffer
				Is also: NT of Accompaniments	

SKOS uses the following properties to identify values:

- Concept scheme is skos:ConceptScheme
- Top Concept is skos:hasTopConcept
- Concept is skos:Concept
- A broader term (BT) is skos:broader
- A narrower term (NT) is skos:narrower
- A related term (RT) is skos:related

Preferred, Alternative and Hidden labels can also be used. This faceted taxonomy example does not give alternative or hidden labels and all terms are the preferred labels with the property skos:prefLabel.

In SKOS, transitive hierarchies need to be explicitly defined. As an example, "cats" can have a broader term of "mammals" which can have a broader term of "animals." In SKOS, without explicit statement, you cannot infer that "animals" is a broader term for "cats."

Figure 1. Detailing broader and broaderTransitive relationships from http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/broaderNonTransitive.jpg
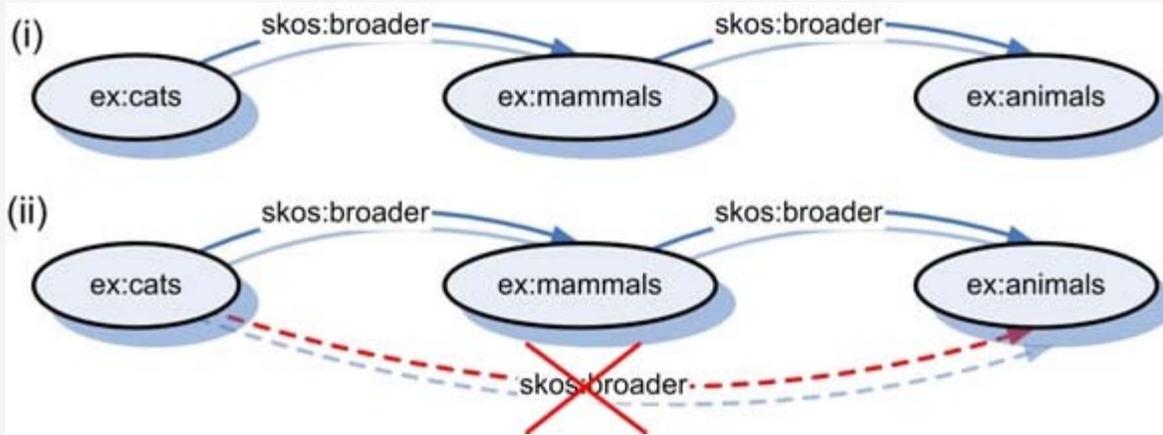
To indicate that broader and narrower terms are transitive, the properties skos:broaderTransitive and skos:narrowerTransitive, respectively (http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/#sectransitivebroader). This idea comes into effect with PoolParty's application and RDF/XML.

PoolParty

Instead of writing out the RDF/XML for SKOS, an online application called PoolParty was used to help input the above faceted taxonomy and then export it to RDF/XML. This would prevent many typing mistakes and help create the correct RDF structure without having to know RDF or XML in detail. When entering this information, the program quite helpful, but also had some drawbacks.

PoolParty allows companies to create and maintain thesauri, taxonomies and classification schemes. The company offers a demo account to try the software. "PoolParty is a thesaurus management system and a SKOS editor for the Semantic Web including text mining and linked data capabilities. The system helps to build and maintain multilingual thesauri providing an easy-to-use interface. PoolParty server provides semantic services to integrate semantic search or recommender systems into systems like CMS, DMS, CRM or Wikis" (http://poolparty.punkt.at/). PoolParty is based in Vienna, Austria and has been in business since 1998 (http://poolparty.punkt.at/company).

When entering a faceted taxonomy into PoolParty, PoolParty ensures it is added according to SKOS and other standards. A screenshot from PoolParty Online Retail Boutique faceted taxonomy home page shows the overall structure and the Dublin Core metadata tracked with the taxonomy.
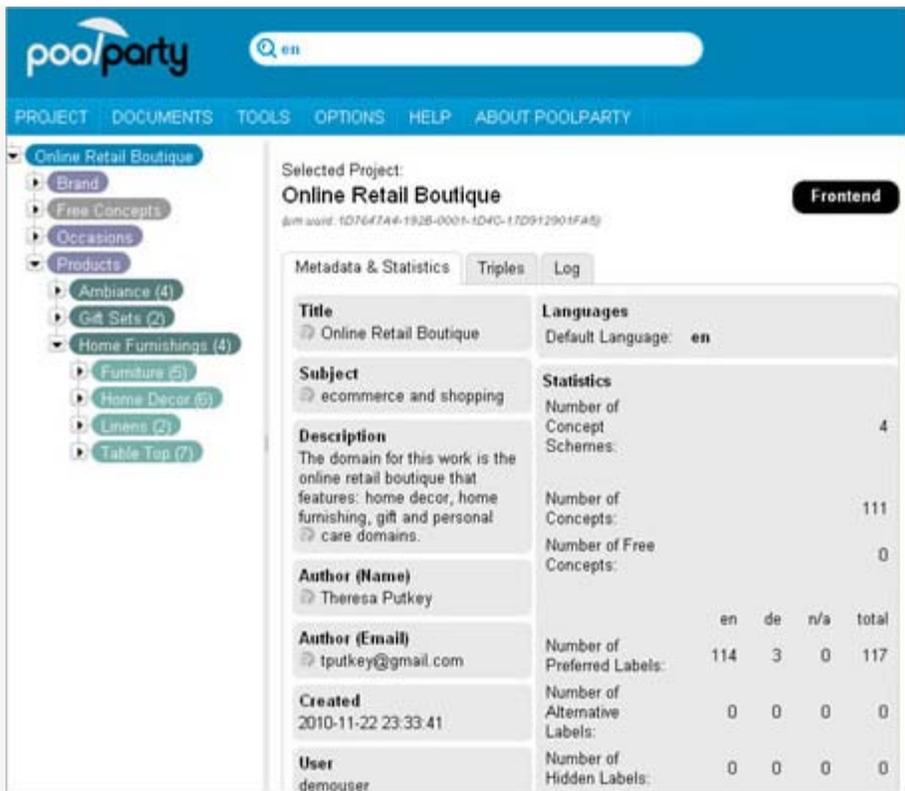
Figure 2. Home page for the Online Retail Boutique faceted taxonomy in PoolParty.

The following screenshot shows a value in the taxonomy.

Figure 3. The Cabinets value with broader terms, narrower terms and related terms.

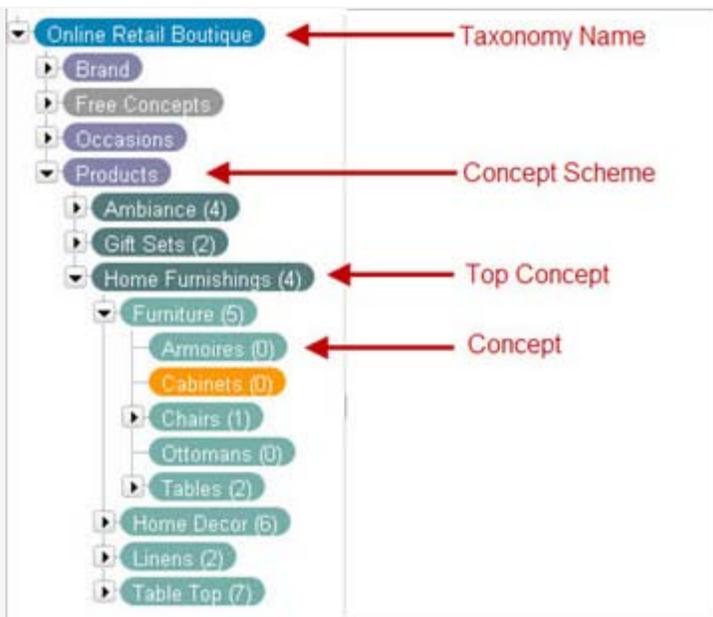The following image shows how SKOS properties are displayed in PoolParty.

Figure 4. How SKOS values are represented in PoolParty.

Once added according to the program and SKOS standards, the taxonomy can be exported and shared with other applications in either XML or Triples. This paper only looks at the RDF/XML format. After exporting to RDF/XML, the code appears as:

Figure 5. RDF/XML exported from PoolParty for the Online Retail Boutique

Looking closely at part of the XML, we can see how the subject, object and predicate are laid out in the XML.
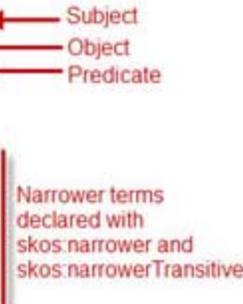
Figure 6. Illustration of the subject, object and predicate and narrower terms declared.

Discussion

Attempting to understand faceted classification, XML, RDF and SKOS was a daunting task. In order to understand the principles of SKOS, one must have a good understanding of how to create thesauri, taxonomies and faceted classifications. In order to set up the thesaurus properly in SKOS, one needs to be able to properly map the thesaurus to the properties in SKOS. One must also know RDF and XML. With some exposure to HTML and XML, reading XML can be relatively straightforward. However, understanding RDF and how it is reading it needs to be learned.

SKOS is unique standard for expressing thesauri, taxonomies and classification schemes in XML. It seemed to be an excellent way to share thesauri instead of recreating them in each application. However, sharing thesauri between systems can be challenging when trying to create bridges between them. How does one know that two concepts refer to the same thing? How does one determine this? It must be challenging for an information scientist to create mappings between the two concepts, either in theory or in software.

PoolParty had some benefits and drawbacks for creating an SKOS-complaint faceted taxonomy. This tool definitely helps one create a taxonomy without coding errors and it was rather easy to use. It was somewhat tedious to enter all the values manually, but they can also be imported if the taxonomy is in the proper RDF format. One major drawback of PoolParty was the seeming inability to make a hierarchy not transitive. Earlier, this paper discussed that concepts are transitive only if explicitly stated. In SKOS, one can employ the skos:narrowerTransitive or skos:broaderTransitive properties to express transitive values. In PoolParty, looking at the RDF/XML, we see that all narrower terms are set as skos:narrowerTransitive. PoolParty did not have an way to take this property out of the taxonomy.

Although outside the scope of this research paper, it would have been interesting to have time and space to investigate Triples further as well as alternatives to expressing SKOS in RDF/XML. SKOS can be expressed in Triples, which could possibly create a more succinct XML file.

Conclusion

Overall, the investigation and research into SKOS was quite interesting. One must understand the principles of SKOS before it can be put into practice. Looking at the "real world," programmers might be the ones creating the XML but information scientists are the ones with the understanding of thesauri principles. However, both the programmer and information scientist must understand enough of each other's subject matter to work together to create an accurate and appropriate thesaurus in SKOS.

References

Berners-Lee, T. (1998). *What the semantic web can represent*. Retrieved October 9, 2010, from http://www.w3.org/DesignIssues/RDFnot.html

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings, 58*(1/2), 49-72. Retrieved September 20, 2010, from ABI/INFORM Global. (Document ID: 1127261641).

Denton, W. (2009). *How to make a faceted classification and put it on the web*. Retrieved October 9, 2010, from http://www.miskatonic.org/library/facet-web-howto.html

Fagan, J. (2010). Usability Studies of faceted browsing: A literature review. *Information Technology & Libraries*, *29*(2),

58-66. Retrieved September 20, 2010, from Library, Information Science & Technology Abstracts with Full Text database

Getty Trust. (2010). *About the AAT*. Retrieved October 11, 2010, from http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html

Getty Trust. (2010). *Art & architecture thesaurus online.* Retrieved October 11, 2010, from http://www.getty.edu/research/conducting_research/vocabularies/aat/

Louie, A., Washington, W., & Maddox, E. (2003, March). *Using faceted classification to provide structure for information architecture.* Paper presented at the Information Architecture Summit 2003, Portland, OR. Retrieved September 20, 2010, from http://depts.washington.edu/pettt/presentations/conf_2003/IASummit.pdf

Mills, J. (2004). Faceted classification and logical division in information retrieval. *Library Trends*, *52*(3), 541-570. Retrieved September 20, 2010 from Library, Information Science & Technology Abstracts with Full Text.

National Information Standards Organization. (2005). *ANSI/NISO Z39.19: Guidelines for the construction, format, and management of monolingual thesauri.* Retrieved September 12, 2010, from http://www.niso.org/kst/reports/standards/

Perugini, S. (2010). Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links. *Information Processing & Management*, *46*(1), 22-43. doi:10.1016/j.ipm.2009.06.007. Downloaded September 13, 2010 from Science Direct.

Petersen, T. (1990). Developing a new thesaurus for art and architecture. *Library Trends, 38*(4), 644-658. Downloaded September 19, 2010 from Angel

PoolParty. (2010). *PoolParty website*. Retrieved November 19, 2010 from http://poolparty.punkt.at/

Putkey, T. (2010). A Survey of Faceted Classification: History, Uses, Drawbacks and the Semantic Web. Unpublished manuscript.

Spiteri, L. (1998). A simplified model for facet analysis: Ranganathan 101. *Canadian Journal of Information and Library Science 23*(1/2), 1-30.

Tauberer, J. (2005) *RDF:about*. Retrieved October 11, 2010 from http://www.rdfabout.com

Uddin, M., & Janecek, P. (2007). Faceted classification in web information architecture. *Electronic Library*, *25*(2), 219-233. doi:10.1108/02640470710741340. Retrieved September 20, 2010 from Emerald.

W3C. *RDF Primer*. Retrieved October 11, 2010 from http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

W3C. W3C *Sematic Web Activity.* Retrieved October 11, 2010 from http://www.w3.org/2001/sw/

W3C. *SKOS*. Retrieved October 11, 2010 from http://www.w3.org/2004/02/skos/

W3C. *SKOS Simple Knowledge Organization System Primer*. Retrieved October 11, 2010 from http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/