

**Jung-ran Park, Ph.D.**  
College of Information Science and Technology, Drexel University  
3141 Chestnut Street, Philadelphia 19104, USA

## **Semantic Interoperability across Digital Image Collections: a Pilot Study on Metadata Mapping<sup>1</sup>**

**Abstract:** The goal of this project is evaluation of the current status of semantic mapping between cataloger-defined field names and Dublin Core metadata elements across digital image collections and identification of the most frequently occurring incorrect and null mappings. A pilot study has been conducted comparing and analyzing 20 digital image metadata templates and 659 metadata item records.

**Résumé :** L'objectif de ce projet est d'évaluer l'état actuel de la mise en correspondance sémantique entre les noms de champs définis par les catalogueurs et les éléments de métadonnées du Dublin Core à travers des collections d'images numériques et d'identifier les correspondances qui sont le plus fréquemment incorrectes et sans valeur. Une étude pilote a été effectuée en comparant et analysant 20 modèles de métadonnées d'images numériques et 659 enregistrements d'éléments de métadonnées.

### **1. Introduction**

Recognition of the critical importance of the linguistic unit 'vocabulary' in information organization and retrieval has long existed (Lancaster, 1986, Furnas et al., 1987, Buckland, 1999) in the library and information science fields. (For the purposes of this study, the term *vocabulary* encompasses information organization schemes such as cataloging and classification, thesauri, ontologies, metadata standards, electronic lexicons, taxonomies, etc.) Recognition has spiked as Web technologies advance toward global interconnection through data exchange and information-sharing across distributed information systems. Active studies of the semantic web, ontology markup language, metadata and ontology engineering, etc., across a variety of disciplines make clear the critical role played by vocabulary in representing and accessing information and knowledge.

The vocabulary uses of synonymy (e.g., author, writer, creator), homographs (e.g., bank [building] vs. bank [river]) and polysemy (multiple meanings of a word that are enumerated in alphabetical order in a typical dictionary entry) in face-to-face human interactions add immeasurably to the richness and creativity of natural language. Any ambiguities and misunderstandings that are engendered are usually resolved smoothly through communication cues provided during social interactions such as repetition and elaboration, social context and non-verbal cues (e.g., facial expressions and gestures). However, in an information retrieval environment these same semantic ambiguities bring about lowered recall and reduced precision, which in turn pose enormous hindrances and challenges in maximizing the full potential of Web and communication technologies for resource sharing and data exchange.

The process of vocabulary mapping across diverse languages and cultures, essential for building multilingual information systems (Hovy, et al., 2001, Purat, 1998, Oard D., et

al., 1999, Bakers, 1997, Matthews, Brian and Michael Wilson, 2000), produces multifold challenges and hindrances due especially to differences in conceptualization and lexicalization patterns across languages (Park, 2002). However, even within the same language the culture and practices of heterogeneous communities are wide-ranging and varied; this is accordingly reflected in disparate vocabulary systems (Friesen, 2002). Furthermore, proliferating vocabulary schemes for accessing networked and digitized resources greatly complicate the goal of semantic interoperability even within the same language and the same community.

Considering the complex nature of semantic interoperability, the scope of this study was narrowed to an examination of equivalent practices of information providers in settings in libraries. The focus will be on metadata mapping for digitized image resources employing CONTENTdm, the digital collection management software.<sup>2</sup>

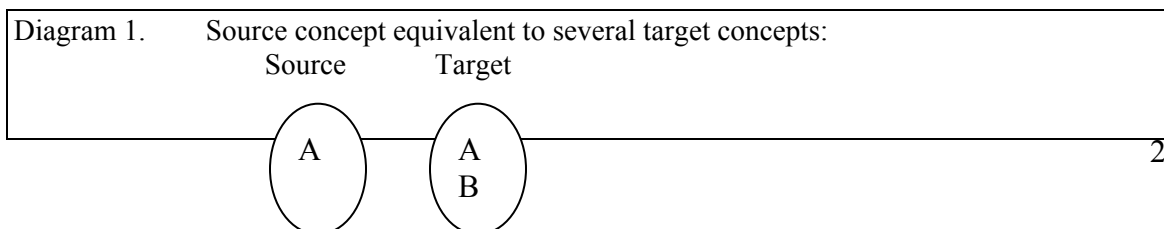
The goal of this ongoing project is to evaluate the current status of semantic mapping between cataloger-defined field names and DC metadata elements across digital image collections and to identify factors producing the most frequent incorrect and null mappings. This goal relates to the issue of semantic interoperability of concept representation across digital collections. For this, as a first step a pilot study has been conducted by comparing and analyzing 20 digital image metadata templates (see Table 3 in Section 3) and 659 metadata item records.

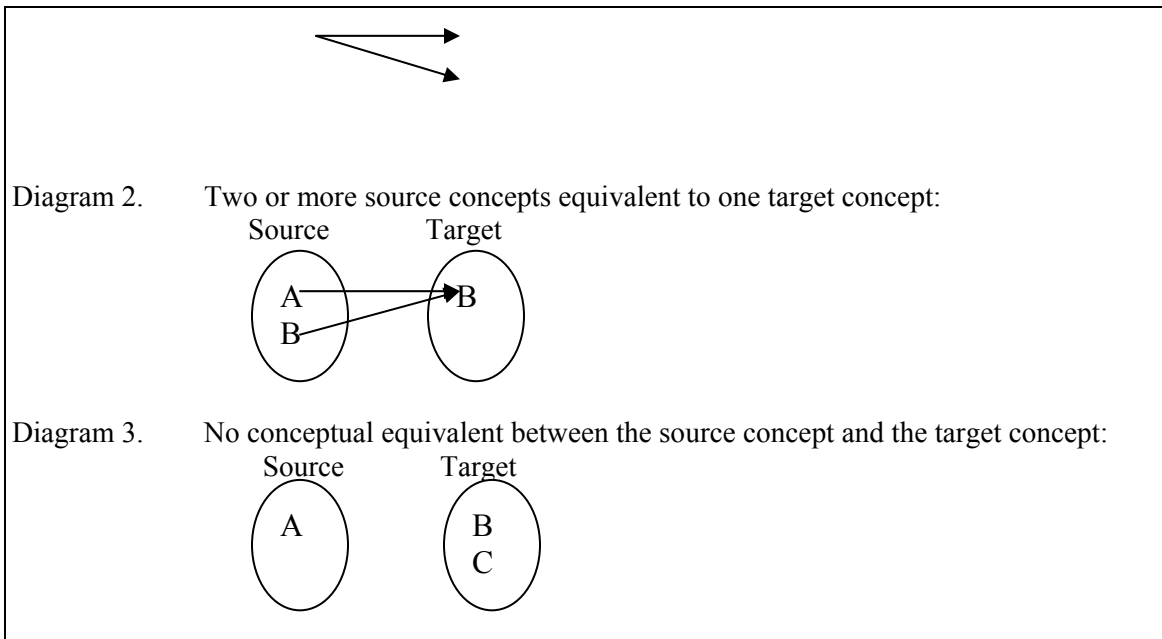
## 2. Related Research to Semantic Mapping

Efforts to increase semantic interoperability across heterogeneous vocabulary systems have dramatically increased through recent and ongoing large-scale projects and initiatives (Lassila, 1998, Miller, 2000, Chan, 2000, Heflin, J. and J. Hendler, 2000, Hunter, 2001, Duval, E. et al., 2002, Godby, C.J., Smith, D., and E. Chidress, 2004, Friesen, 2004, Soergel, D. et al. 2004, OCLC research projects on “interoperability” and “knowledge organization”). Schemes are burgeoning aimed at effecting harmonization and integration of heterogeneous vocabulary systems such as LCSH, LCC, DDC and MeSH and numerous metadata schemes, including Dublin Core, through vocabulary mapping and the creation of crosswalks (Vizine-Goetz, D., et al., 2004, Neuroth, H. and Traugott Koch, 2001, Calhoun, Karen et al., 2001, Burstein, 2003, Getty Research Institute, 2000).

Hindrances in enhancing semantic interoperability have been reported by various studies (Heflin, J. and J. Hendler, 2000, Doerr, 2001, Park, 2002, Vizine-Goetz, D., et al., 2004). Park (2002) presents an overview from a linguistic perspective of the characteristics of natural language, focusing on issues of polysemy and differences in the conceptualization of lexical elements across languages that pose particular challenges in mapping among heterogeneous knowledge organization schemes.

The semantic mapping process is analogous to translating two or more different languages. The following diagram from Park (2002) illustrates some possible conceptual mismatches between two languages:





**Figure 1. Concept Equivalence**

As indicated in Figure 1, precise and equivalent mapping between two languages in translation does not exist; however, an experienced translator can mitigate the semantic ambiguity between source and target languages by utilizing information stored in the mental lexicon (in the case of spoken language) and/or available resource tools such as desktop and online dictionaries, syntax rules, etc., thus enhancing semantic interoperability between the two languages. This can be seen as analogous to the semantic mapping process employed by catalogers when mapping cataloger-defined natural vocabularies (source language) onto DC metadata elements (target language). Mental lexicon and/or desktop dictionaries utilized by translators for successful translation can also be seen as analogous to the mediation mechanism proposed herein that catalogers could refer to during the mapping process to facilitate semantic interoperability across distributed digital collections.

Heflin, J. and J. Hendler (2000) report hindrances in integrating DTDs:

One of the hardest problems in any integration effort is mapping between different representations of the same concepts—the problem of integrating DTDs is no different. One difficulty is identifying and mapping differences in naming conventions. As with natural language, XML DTDs have the problems of polysemy and synonymy. For example, an element such as <spider> might be polysemous: in one document it could mean a piece of software that crawls a web of the silky kind. In general, it is difficult for machines to make determinations of this nature, even if they have access to a complete automated dictionary and thesaurus.

Likewise, Godby et al. (2003) point out challenges in semantic mapping between DC and the cataloger's created natural language fields:

... we have examined approximately 400 Dublin Core records from three data streams that were submitted to one of our test clients from a digital library project. Analysis of the records reveals that only seven of the fifteen Dublin Core elements appear in all three data sets: Identifier, Title, Creator, Subject, Date, Type, and Format. Of these, Subject and Description both contain subject headings and free-text descriptions;

Format and Type both contain names of media types such as photograph; and the data in the Language of the metadata record and the language of the content. Without extensive human-mediated correction, or training that promotes more consistent application of the Dublin Core element semantics when the records are created, even the goal of limited interoperability is compromised. (Underlined emphasis by the author)

Hegg and Knab (2003) echo this argument in their research on cross-collection searches for visual resources by pointing out that the solution for optimal cross-collection searching depends on “the curator’s ability to accurately map the MDID file onto DC elements and refinements.”

As can be seen, studies in the area of semantic interoperability are being actively undertaken, especially through large-scale initiatives and projects aimed at achieving automatic semantic mapping. However, as Godby, et al. (2003) and Hegg and Knab (2003) emphasize, and as pointed out by Heflin, J. and J. Hendler (2000), it is difficult for machines to achieve precise semantic mapping owing to the problems of disambiguating polysemous and synonymous words and senses. Without extensive human-mediated efforts that target the identification of incorrect semantic mapping and the training of catalogers and curators, the goal of enhancing and refining semantic interoperability, even in relatively less complex information environments, will be thwarted.

### 3. Data and Research Methods

A growing number of organizations are building digital collections using both commercial digital collection management software such as CONTENTdm and Encompass and open source software such as Greenstone. The rapidly growing number of distributed digital collections has brought to the fore the critical issues of resource discovery and sharing across these collections.

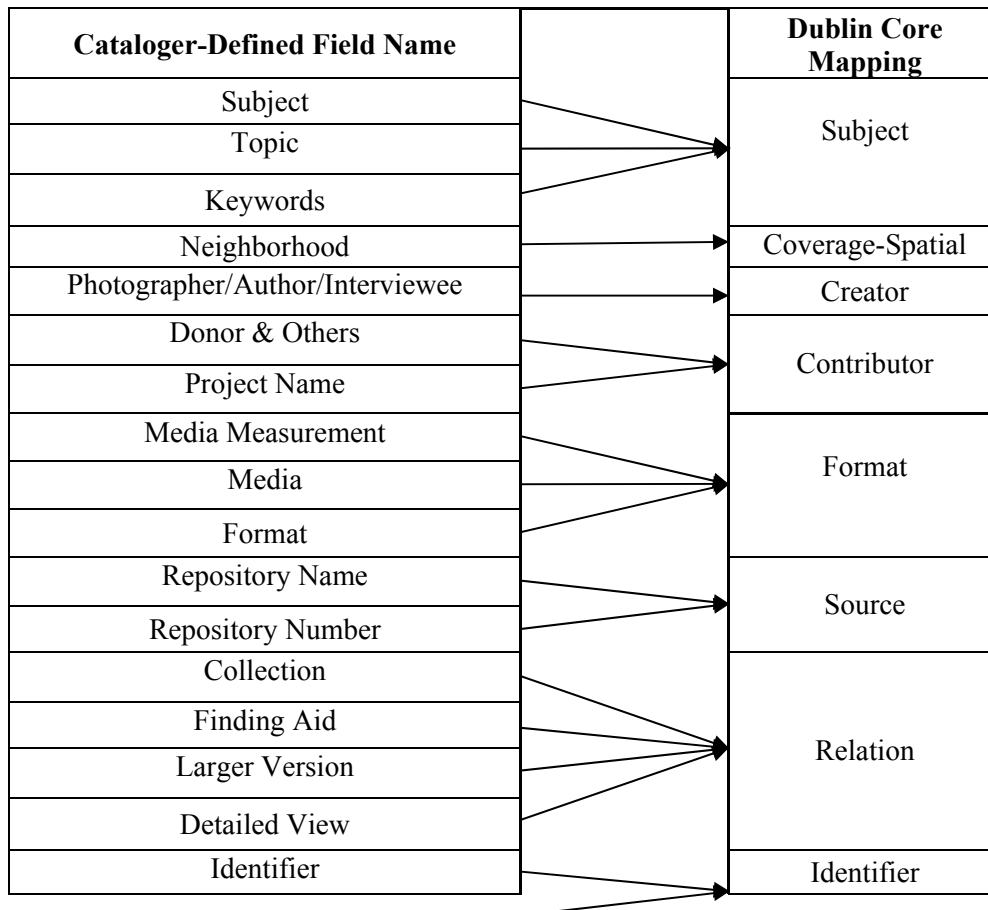
According to a recent survey based on licensed user groups as of November 2004, over 200 organizations, including many academic libraries, are currently building and maintaining digital collections using CONTENTdm software, which utilizes the Dublin Core (DC) metadata scheme. The fact that a significant and growing number of digital collections are using this software demonstrates the need for research on metadata mapping to enhance semantic interoperability for more efficient and successful resource sharing across digital collections using CONTENTdm. The software provides a feature that allows for a cataloger to map cataloger-defined field names onto DC metadata elements as shown in Table 1 below.

Cataloger-Defined Field Names	DC Metadata Elements
Title	DCTitle
Description	DCDescription
Subject	DCSubject
Topic	DCSubject
Keywords	DCSubject
Neighborhood	DCCoverage-Spatial
Date	DCDate
Alternative Dates	DCCoverage-Temporal
Photographer/Author/Interviewee	DCCreator
Donor & Others	DCCContributors

Media	Format-Medium
Media Measurement	Format Extent
Type	DCType
Format	DCFormat
Identifier	DCIdentifier
Language	DCLanguage
Repository Name	Source
Collection	DCRelation
Repository Number	Source
Call Number	Identifier
Finding Aid	DCRelation
Rights	DCRights
Project Name	Contributors
Date Digitized	DCDate-Issued
Publisher	DCPublisher
Detailed View	Relation
Larger Version	Relation

**Table 1. A Metadata Template**

However, the complex nature of natural language, which allows for the representation of a concept in various ways, challenges consistent semantic mapping across digital collections. For instance, twenty metadata templates for this project evince inconsistent and null semantic mapping. The metadata template shown above presents the following mapping between cataloger-defined natural vocabulary and DC metadata elements.



**Table 2. Metadata Mapping Practice**

As shown through information sharing for non-network traditional bibliographic collections through authority control and resource description rules, successful resource discovery and exchange across ever-growing digital collections demands semantic interoperability of concept representation based on unambiguous and consistent resource description. Absent correct mapping of cataloger-defined natural vocabularies onto DC elements, semantic interoperability, even among digital collections employing the identical metadata scheme and the identical digital collection management software configuration will become increasingly problematic, leading to a decrement in information sharing.

The research questions employed for this pilot study were:

- How are cataloger-defined field names mapped onto DC metadata elements?
- Which field names produce the most frequent incorrect mappings and null mappings?
- Which factors produce the most frequent incorrect mappings?
- To what extent do ambiguities of concept in relation to the specific object and the general collection described by the field name engender incorrect semantic mapping?

The research method for this project is formulated on both qualitative and quantitative analysis of metadata mapping between natural vocabulary field names as defined by catalogers and DC metadata elements. As shown in Table 1 above, 20 metadata templates of digital collections using CONTENTdm software were compared. In order to examine usage of DC metadata elements and to determine the accuracy of the mapping between cataloger-defined natural vocabulary field names and DC metadata elements, a total of 659 metadata item records from three digital image collections were also examined.

The natural vocabularies that catalogers create do not necessarily correspond to DC metadata elements and are variable across distributed digital collections. Both incorrect and null mapping fields were identified and analyzed to discern any pattern development. A pattern development as here defined concerns particular field names or metadata elements that evince frequent incorrect and null mappings.

#### **4. Discussion**

The analysis of 20 metadata templates and 659 metadata item records shows evidence of frequent incorrect and null metadata mappings. Some examples: the 'physical description' field is either mapped onto DC 'description' or 'format.'; there is great confusion in employing the DC elements 'type' and 'format' and they are interchangeably used; the DC elements 'source' and 'relation' are inconsistently mapped onto various cataloger-defined fields; the DC element 'relation' is interchangeably used with cataloger-defined field names such as 'digital collection', and 'example issues.'; the DC 'subject' element is mapped by a variety of cataloger-defined natural vocabularies such as 'topic', 'category,' and 'keyword.'

The most frequently identified null mapping field names are: ‘contact information’, ‘ordering information’, ‘full text’, ‘note’, ‘scan date’, ‘full resolution’, ‘acquisition’ and ‘image modification’.

Table 3 below represents the usage of DC metadata elements by three digital image collections (A, B and C). The total of 659 metadata item records were collected thus: from digital collection A (n/203 records), B (n/215 records) and C (n/241 records).

Percentage of the Total Number of DC Metadata Elements Used by Three Collections (A, B and C)								
DC Element	A n/203	% of the total number of DC elements used n/3476	B n/215	% of the total number of DC elements used n/2721	C n/241	% of the total number of DC elements used n/2606	Total n/659	% of total usage of DC
Title	203	5.8	217	8.0	241	9.2	661	100.3
Creator	196	5.6	148	5.4	30	1.2	374	56.8
Subject	580	16.7	416	15.3	448	17.2	1444	219.1
Description	203	5.8	210	7.7	263	10.1	676	102.6
Publisher	203	5.8	231	8.5	0	0.0	434	65.9
Contributor	289	8.3	100	3.7	19	0.7	408	61.9
Date	201	5.8	113	4.2	236	9.1	550	83.5
Type	0	0.0	150	5.5	235	9.0	385	58.4
Format	384	11.0	139	5.1	417	16.0	940	142.6
Identifier	265	7.6	107	3.9	7	0.3	379	57.5
Source	362	10.4	0	0.0	0	0.0	362	54.9
Language	63	1.8	0	0.0	5	0.2	68	10.3
Relation	121	3.5	98	3.6	4	0.2	223	33.8
Coverage	203	5.8	281	10.3	241	9.2	725	110.0
Rights	203	5.8	215	7.9	241	9.2	659	100.0
Non-Mapping	0	0.0	296	10.9	219	8.4	515	78.1
Total	3476	100.00	2721	100.0	2606	100.0	8803	1335.8

**Table 3: Dublin Core Metadata Usage in Three Digital Image Collections**

The following is the usage percentage of the top five metadata elements in the above three collections:

Collection A: subject, format, source, contributor, identifier (54% out of all metadata elements within the collection)

Collection B: subject, null mapping fields, coverage, publisher, title (53% out of all metadata elements within the collection)

Collection C: subject, format, description, title, coverage, (71.2% out of all metadata elements within the collection)

Among the three collections, the following metadata elements are the most frequently employed, in descending order: subject, description, title, format and coverage across

three digital image collections. Usage of these five metadata elements constitutes over 50% of all the DC metadata elements. However, as stated earlier, the percentage of ‘description’ and ‘format’ does not precisely reflect actual usage owing to inconsistent and incorrect metadata mapping among the total 659 metadata item records.

The least used elements in ascending order are: language, relation, source, creator, and identifier.

The low usage of ‘creator’ is likely owing to inaccessibility of its data value from image documents. Unlike text-oriented materials such as books, image documents tend not to represent themselves by explicating title, creator or other descriptive data elements that identify image documents. On the other hand, the high usage of ‘title’ can be derived from cataloger-assigned titles by enclosing them with square brackets which indicate the creation of title from cataloger.

The results of this pilot study strongly suggest the critical need for a mediation mechanism in the form of metadata mapping guidelines and a mediation model (e.g., concept maps) that catalogers can refer to during the process of mapping cataloger-defined field names onto DC metadata elements with the goal of increasing semantic mapping consistency and enhancing semantic interoperability across digital image collections. As well, the high percentage [see Table 3] of usage of ‘subject’ by cataloger-defined natural vocabulary field names such as ‘keyword,’ ‘category,’ ‘topic,’ etc., suggests a critical gap in terminology and the pressing need to develop terminology for accessing digital image collections. The development of such a mechanism for metadata semantic mapping and of vocabulary for subject access calls for future studies on cataloger metadata mapping practices and user studies on image searches.

## **5. Future Study**

Survey and phone interviews with catalogers are a necessary step in identifying factors producing null and incorrect metadata semantic mapping. The following areas of inquiry relate to identifying such factors: procedures, steps and methods catalogers follow in creating field names and mapping them onto DC metadata elements; the concept held by catalogers of the role played by the semantic mapping process; metadata elements that engender cataloger difficulties during the mapping process; catalogers’ expectation on a support and mediation mechanism geared toward the mapping task from both digital collection management software developers and LIS educators.

Based on research and consultation with catalogers through surveys and interviews in order to elicit factors that engender null and incorrect mapping, development of metadata semantic mapping guidelines and other mediation mechanisms such as concept maps that facilitate the mapping process are seen as critically needed.

The high usage of the ‘subject’ data element suggests a critical need for future study in this area. The following aspects of inquiry need to be studied further: current subject schemes (e.g., Library of Congress Subject Headings, The Arts & Architecture Thesaurus) catalogers employ for image resources, effectiveness of such subject schemes in retrieving image resources and development of subject terminology for effective access to digital image collection.

In addition, the results of this pilot study needs to be further pursued through future projects aiming at the enhancement of semantic interoperability across heterogeneous digital collection management software systems.

## References

- Bakers, Thomas. 1997. Metadata semantics shared across languages: Dublin Core in languages other than English, at <http://dublincore.org/documents/1997/03/multilingual-semantics/>.
- Buckland, Michael. 1999. Vocabulary as a central concept in library and information science. In *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science*.
- Burstein, Mark H. 2003. The many faces of mapping and translation for semantic web services. In *Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE'03)*.
- Calhoun, Karen et al. 2001. Mixing and mapping metadata to provide integrated access to digital library collections: an activity report. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*.  
<http://www.nii.ac.jp/dc2001/proceedings/Contents.html>.
- Chan, Lois Mai. 2000. Exploiting LCSH, LCC, and DDC to retrieve networked resources: Issues and challenges. [http://lcweb.loc.gov/catdir/bibcontrol/chan\\_paper.html](http://lcweb.loc.gov/catdir/bibcontrol/chan_paper.html)
- Doerr, M. 2001. Semantic problems of thesaurus mapping. In *Journal of Digital Information* 1(8) <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/>
- Duval, E., Hodgins, W., Sutton, S., and S.L. Weibel. 2002. Metadata principles and practicalities. In *D-Lib Magazine* 8(4).
- Friesen, Norm. 2002. Semantic interoperability and communities of practice.  
<http://www.cancore.ca/documents/semantic.html>

Friesen, Norm. 2004. CanCore: Semantic interoperability for learning object metadata. In *Metadata in Practice*, Diane Hillman & Elaine L. Westbrook (eds.). Chicago: American Library Association.

Furnas G, T.K. Landauer, L.M. Gomez, S.T. Dumais. 1987. The vocabulary problem in human-system communication. In *Communications of the ACM*. (30). pp. 964-71.

Getty Research Institute. 2000. Metadata Standards Crosswalks.

[http://www.getty.edu/research/institute/standards/intrometadata/3\\_crosswalks/index.html](http://www.getty.edu/research/institute/standards/intrometadata/3_crosswalks/index.html)

Godby, C. J., Smith, D., and E. Childress. 2003. Two paths to interoperable metadata. In *DC-2003: Supporting Communities of Discourse and Practice—Metadata Research & Applications*, September 28-October 2, in Seattle, Washington (USA).

<http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf>.

Heflin, J. and J. Hendler. 2000. Semantic interoperability on the Web. In *Proceedings of Extreme Markup Languages*. Graphic Communications Association.

<http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf>

Hegg, K.J. and A.R. Knab. 2003. Using Dublin Core to facilitate cross-collection searches in an enterprise image repository. In *Dublin Core Conference: Supporting Communities of Discourse and Practice--Metadata Research & Applications*. September 28-October 2, 2003, Seattle, Washington. Syracuse, NY: Information Institute of Syracuse. <http://dc2003.ischool.washington.edu/Archive-03/03hegg.pdf>

Hovy et al. 2001. *Multilingual Information Management: Current Levels and Future Abilities*. Pisa: Italy Insituti Editoriali e Poligrafici Internazionali.

Hunter, Jane. 2001. MetaNet-A metadata term thesaurus to enable semantic interoperability between metadata domains. In *Journal of Digital Information* 1(8).

Lancaster, F. Wilfrid. 1986. *Vocabulary Control for Information Retrieval*, Arlington, Va.: Information Resources Press.

Lyons, John. 1977. *Semantics*. Cambridge University Press.

- Lassila, O. 1998. Web metadata: A matter of semantics. In *IEEE Internet Computing* 2(4): 30-37.
- Matthews, Brian and Michael Wilson. 2000. Multilingual metadata to access social science data. In *Information Systems Engineering, CLRC-Ral*.
- Miller, Paul. 2000. Interoperability: What is it and why should I want it? In *Ariadne* 24 <http://www.ariadne.ac.uk/issue24/interoperability/intro.html>
- Neuroth, Heike and Traugott Koch. 2001. Metadata mapping and application profiles: Approaches to providing the cross-searching of heterogeneous resources in the EU project *renardus*. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*. <http://www.nii.ac.jp/dc2001/proceedings/Contents.html>.
- Oard, Douglas et al. 1999. Multilingual information discovery and access. In *D-Lib Magazine* 5(10). <http://www.dlib.org/dlib/october99/10oard.html>
- Park, Jung-ran. 2002. Hindrances in semantic mapping among metadata schemes: A linguistic perspective. In *Journal of Internet Cataloging*, Vol. 5(3): 59-79.
- Purat, Jacek. 1998. The world of multilingual environmental thesauri. [http://www.sims.berkeley.edu/~purat/world\\_multilingual\\_environmental](http://www.sims.berkeley.edu/~purat/world_multilingual_environmental)
- Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., and S. Katz. 2004. Reengineering thesauri for new applications: the AGROVOC example. In *Journal of Digital Information* Vo. 4(4): Themes: Digital Libraries, Information Discovery.
- Vizine-Goetz, D., Hickey, C., Houghton, A., and R. Thompson. 2004. Vocabulary mapping for terminology services. In *Journal of Digital Information* 4(4), Themes: Digital libraries, information discovery. <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/>

---

<sup>1</sup> I would like to express my appreciation to research assistant Sang-Joon Park for his assistance with this project.

<sup>2</sup> For background history and overview of the functionality of this software, visit the site: <http://contentdm.com/index.htmls>

---

College of Information Science and Technology, Drexel University  
3141 Chestnut Street, Philadelphia 19104, USA  
Submitted on April 15, 2005