Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections*

Jung-ran Park

College of Information Science and Technology, Drexel University,
3141 Chestnut Street, Philadelphia PA 19104 USA, <jung-ran.park@cis.drexel.edu>

Jung-ran Park is currently an assistant professor at the College of Information Science and Technology at Drexel University. Prior to joining Drexel University, she worked as a cataloger at Indiana State University. She received a Ph.D. in linguistics, specializing in discourse, pragmatics and semantics, and an MLIS from the University of Hawaii. She has applied her linguistics background to research areas in knowledge organization and representation and computer-mediated communication. Her teaching areas include cataloging and classification, metadata, content representation and information resources in humanities. She received a 2006 IMLS award for conducting metadata quality evaluation for a three-year period.

* The author would like to express her appreciation for the anonymous reviewers' comments, which greatly contributed to enhancing the quality of this paper.

Park, Jung-ran. **Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections.** *Knowledge Organization*, *33(1)* 20-34. 51 refs.

ABSTRACT: This paper is a current assessment of the status of metadata creation and mapping between cataloger-defined field names and Dublin Core (DC) metadata elements across three digital image collections. The metadata elements that evince the most frequently inaccurate, inconsistent and incomplete DC metadata application are identified. As well, the most frequently occurring locally added metadata elements and associated pattern development are examined. For this, a randomly collected sample of 659 metadata item records from three digital image collections is analyzed. Implications and issues drawn from the evaluation of the current status of metadata creation and mapping are also discussed in relation to the issue of semantic interoperability of concept representation across digital image collections. The findings of the study suggest that conceptual ambiguities and semantic overlaps inherent among some DC metadata elements hinder semantic interoperability. The DC metadata scheme needs to be refined in order to disambiguate semantic relations of certain DC metadata elements that present semantic overlaps and conceptual ambiguities between element names and their corresponding definitions. The findings of the study also suggest that the development of mediation mechanisms such as concept networks that facilitate the metadata creation and mapping process are critically needed for enhancing metadata quality.

1. Introduction

Recognition of the vital importance of the linguistic unit 'vocabulary' in knowledge organization and information 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 and taxonomies.) 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 and metadata and ontology engineering across a variety of disciplines make clear the critical role played by vocabulary in representing

and accessing information and knowledge (Hovy et al., 2001).

The vocabulary uses of synonymy (e.g., author, writer, creator), homographs (e.g., bank [building] vs. bank [river]), and polysemy (multiple related meanings of a word that are enumerated in alphabetical order in a typical dictionary entry), in face-to-face human interaction 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 (Svenonius, 2000; Blair, 1999), 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 et al., 1999; Baker, 1997; Matthews and Wilson, 2000), produces multifold challenges and hindrances owing 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 achieving 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 library settings. Digitized image resources from three academic libraries employing *CONTENTdm*, the digital collection management software, are examined. (A background history and an overview of the functionality of this software are available at the site: http://contentdm.com/index.html).

The goal of this project is to assess the current status of metadata creation and mapping between cataloger-defined field names and Dublin Core (DC) metadata elements across three digital image collections and to identify metadata elements that evince the most frequent inaccurate, inconsistent, and incomplete DC metadata application. This project is aimed also at identifying the most frequently occurring locally added metadata elements and associated pattern developments. Implications drawn from the evaluation of the current status of metadata creation and mapping in relation to the issue of semantic interoperability of concept representation across digital image collections is also examined. For this, a randomly collected sample of 659 metadata item records from three digital image collections is analyzed.

2. Semantic interoperability and metadata quality

In natural language, mappings between word forms and meanings can be many-to-many. In other words, the same meaning can be expressed by several different forms (e.g., synonyms), and the same forms may designate different concepts (e.g., homonyms). In addition, the same concept can be expressed by different morpho-syntactic forms (e.g., noun, adjective, compound noun, phrase and clause). In natural language use, this complex mapping between forms and meanings adds intricacy and richness to language use. However, these linguistic phenomena engender confusion and drawbacks in the process of communication between communities in the sense that different communities may use dissimilar word forms to deliver identical or similar concepts or may use the same forms to designate different concepts. These phenomena may be at play even within the same community of practice, as shown in Tables 1 and 2 in the following section: a different form, such as 'neighborhood' in a cataloger defined field name, is used to designate 'spatial coverage' in the DC metadata scheme.

In this paper, the principal aim of semantic interoperability is seen as its ability to disentangle the complex nature of mapping between word forms and meanings in natural language in order to enhance resource exchange and discovery within a community or between communities (see also Miller, 2000; Heflin and Hendler, 2000). In this regard, accurate and consistent metadata mapping between two or more different vocabulary schemes is a vital component in achieving semantic interoperability.

The critical issues affecting metadata quality evaluation have been relatively unexplored (Moen et al. 2003; Barton et al., 2003). However, there is a growing awareness of the essential role of metadata quality assurance for successful resource access and

sharing across distributed digital collections. Through examining learning objects and e-prints of communities of practice, Barton et al. (2003) discuss the importance of quality assurance for metadata creation while pointing out the lack of formal investigation into the metadata creation processes. The problems inherent in the metadata creation process, such as inaccurate data entry (e.g., spelling, abbreviations, format of date – date of creation or date of publication, consistency of subject vocabularies) that result in adverse effects on resource discovery are examined. Moen, et al. (2003) also discuss problems of metadata quality through examination of 80 metadata records from the Government Information Locator Service (GILS) using a set of criteria such as completeness, accuracy and currency.

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 and Hendler, 2000; Hunter, 2001; Duval et al., 2002; Godby, et al., 2004; Friesen, 2004; Soergel et al. 2004; OCLC research projects on "interoperability" and "knowledge organization"). Burgeoning schemes are aimed at bringing about harmonization and integration of heterogeneous vocabulary systems such as *Library of Congress Subject Headings (LCSH)*, *Library of Congress Classification (LCC)*, *Dewey Decimal Classification (DDC)*, and *Medical Subject Headings (MeSH)*, as well as numerous metadata schemes, including Dublin Core, through vocabulary mapping and the creation of crosswalks (Vizine-Goetz et al., 2004; Neuroth and Koch, 2001; Calhoun, Karen et al., 2001; Burstein, 2003; Getty Research Institute, 2000).

With the objective of enhancing semantic interoperability and requiring metadata quality assurance, Heery (2004) points out the increasingly rising number of local additions and variants to metadata standards. She emphasizes the necessity of building a mediation mechanism that can be sharable across libraries. Barton et al. (2003) also point out the necessity for guidelines for metadata creation and quality control. Bruce and Hillmann (2004) address challenges in approaching questions of quality by stating "quality standards and measures are sorely missed." In reaction to improving metadata quality, the study suggests examination of documentation procedures and standards documents accompanying best practice guidelines and examples.

Challenges in enhancing access to digital collections have been reported by various scholars (Heflin and Hendler 2000; Doerr 2001; Park 2002; Vizine-Goetz et al., 2004; Hegg and Knab 2003). Park (2002) presents an overview from a linguistic perspective of the characteristics of natural language, focusing on issues of synonymy and polysemy that pose particular challenges in semantic interoperability across heterogeneous knowledge organization schemes. The semantic mapping process is analogous to translating between two or more languages. The following diagram from Park (2002) illustrates some possible conceptual mismatches between two languages:

Diagram 1. Source concept equivalent to several target concepts:

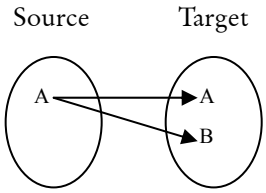

Diagram 2. Two or more source concepts equivalent to one target concept:

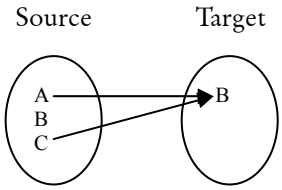

Diagram 3. No conceptual equivalent between the source concept and the target concept:

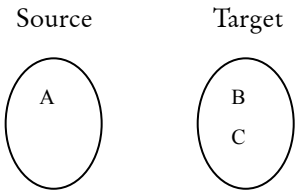

Figure 1. Concept equivalence

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) or available resource tools such as online dictionaries and syntax rules, thus enhancing semantic interoperability between the two languages (see also Dahl-

berg 1996c). 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).

Heflin and Hendler (2000) report hindrances in integrating Document Type Definitions (DTDs) by way of addressing the problems of polysemy and synonymy. They stress the importance of metadata creation by cataloging professionals and human indexers (p. 2): "it is difficult for machines to make determinations of this nature, even if they have access to a complete automated dictionary and thesaurus." McClelland et al. (2002) discuss issues and challenges stemming from *iLumina* project experiences of mismatches of imported metadata from data providers, such as missing and incorrect data values: "metadata will be incomplete and contain errors, don't count on accuracy in data."

Likewise, according to an analysis by Godby et al. of 400 Dublin Core records, incorrect and inconsistent metadata uses occur in the following way (2003, 8; emphasis added):

> *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 *Date* 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*.

Hegg and Knab (2003, 2) 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."

Bui and Park (2006) examine metadata quality of the open source Metadata Repository at the National Science Digital Library (NSDL). The NSDL comprises 111 collection sets submitted from various data providers. The lack of consistency in metadata uses in NSDL is partially due to the fact that metadata in the repository are derived from many different data providers. As well, these data providers utilize a variety of schemes other than the DC metadata scheme. However for data harvesting purposes, all metadata schemes in NSDL are mapped onto the DC scheme. In this mapping process, inaccurate and inconsistent mappings occur (see also Zeng 2006). Such drawbacks in mapping no doubt hinder semantic interoperability even across NSDL collections.

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 essential issues of resource discovery and sharing across these collections. According to a survey based on licensed user groups as of November 2004 (Park 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 (i.e., cataloger-created natural vocabularies or labels) onto DC metadata elements. Table 1 below illustrated this feature. It is derived from a sample digital collection for this study, the *San Fernando Valley History Digital Library* (http://digital-library.csun.edu/metafields.html).

Field Name	**Mapping**
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	DCContributors
Media	Format-Medium
Media Measurement	Format Extent
Type	DCType

→

Field Name	Mapping
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, poses multifold challenges to consistent semantic mapping across digital collections. The metadata template shown in Table 1 is a representation of a mapping between cataloger-defined natural vocabulary and DC metadata elements that is shown below in table 2. (The ordering of data elements is rearranged from Table 1. The use of arrows to indicate mapping is added for the purposes of the paper).

As mentioned, cataloger-defined field names are natural vocabularies (e.g., labels, field names) created by catalogers in local libraries. In user interface for metadata item records, the cataloger-defined field names are displayed. Figure 2 below illustrates a metadata item record in user interface based on the metadata template in Table 1 and on the mapping practice in Table 2.

Field Name	Mapping
Subject	Subject
Topic	
Keywords	
Neighborhood	Coverage-Spatial
Photographer/Author/Interviewee	Creator
Donor & Others	Contributor
Project Name	
Media Measurement	Format
Media	
Format	
Repository Name	Source
Repository Number	
Collection	Relation
Finding Aid	
Larger Version	
Detailed View	
Identifier	Identifier
Call Number	

Table 2. Metadata semantic mapping between cataloger-defined field names and DC metadata elements

Title:	"Fabulous San Fernando Valley"
Description:	Van Nuys, CA: postcard photograph - aerial view of the San Fernando Valley. "Van Nuys, born as a town site on Washington's Birthday, 1911, is located in almost the exact geographic center of the great San Fernando Valley, its position being similar to the hub of a wheel with sixteen other friendly neighboring communities surrounding it on the rim. Recent survey indicates the present population center of the State of California to be at Van Owen and Van Nuys Boulevards, Van Nuys. The city of Van Nuys, with a population of well over 100,000 comprises over one-fifth of the entire Valley's growing population." Published and Distributed by Columbia, Hollywood, CA. Color Photography by Max Mahan (H-1712) Donor: Dr. Tom Reilly. Photographic postcard. 3.5 x 5.5 in.
Subject:	San Fernando Valley (Calif.) Van Nuys (Los Angeles, Calif.) Aerial views
Neighborhood:	Van Nuys (Los Angeles, Calif.)
Date:	1970-1979?
Photographer/Author/Interviewee:	Mahan, Max
Donors & Others:	Reilly, Tom
Media:	Photographic postcard Aerial photograph
Media Measurement:	9 x 14 cm.
Identifier:	SFVC004.jpg
Repository Name:	California State University, Northridge. Oviatt Library. Urban Archives Center
Collection:	San Fernando Valley Collection
Repository Number:	SFV 004
Rights:	http://digital-library.csun.edu/copyright.html
Project Name:	San Fernando Valley History Digital Library
Publisher:	California State University, Northridge. University Library.

Figure 2. A metadata item record example

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

The current study bears this out: The metadata item records as shown in Figure 2 were randomly collected from three digital image collections from July through September 2004: from digital collection A (n=203 records), B (n=215 records) and C (n=241 records). These three sample collections were selected based on the fact that the collections cover the same type of resource (i.e., image) through employment of the same digital collection management software (i.e., *CONTENTdm*). The software utilizes a built-in Dublin Core (DC) metadata scheme. The built-in controlled vocabulary of *CONTENTdm* is the *Library of Congress Thesaurus for Graphic Materials*. As mentioned, the software provides a feature that allows for a cataloger to map cataloger-defined field names onto DC metadata elements (see Tables 1 and 2). The order of fields is flexible, rendering reordering of fields eminently practicable.

The metadata elements comprising the total of 659 metadata item records were exported to an Excel™ database for analysis (Excel™ spreadsheets offer ready-to-view visual inspection). This also allowed the researcher to read a record with all its elements across a page. Scrolling up and down was also helpful in the identification of any anomaly. An Excel™ file can have as many worksheets as the system's capacity allows. This break-up method did not cause any negative effect on data analysis.

The following are the research questions employed for this study:

– What is the current practice of metadata creation and semantic mapping across digitized image collections utilizing *CONTENTdm*?
– Which field names produce the most frequent inaccurate, inconsistent and null mappings from cataloger defined field names onto DC metadata?
– Which types of locally created metadata elements are added to the DC metadata scheme by the three identified user groups of *CONTENTdm*?
– What conceptual ambiguities and semantic overlaps can be found in the DC metadata elements?

The research method for this project is formulated on both a qualitative and quantitative analysis of the usage of the DC metadata scheme and metadata mapping between natural vocabulary field names as defined by catalogers and DC metadata elements. In order to examine the usage and completeness of DC metadata elements, frequency of usage of metadata elements from the 659 metadata item records has been calculated.

The natural vocabularies that catalogers create do not necessarily correspond to DC metadata elements and are variable across distributed digital collections. Inaccurate, inconsistent and null mappings between cataloger-defined natural vocabulary field names and DC metadata elements have been identified and analyzed to discern any pattern development. Locally added metadata elements to the DC metadata scheme have also been identified. A pattern development as defined here concerns particular field names or metadata elements that evince frequent errors in application of the DC metadata scheme.

Conceptual ambiguities and semantic overlaps in the DC metadata elements have been examined utilizing qualitative analysis: the DC metadata element name and its corresponding definition have been examined by utilizing linguistic semantic analysis. As well, the results from the analysis of 659 metadata item records and from other related studies (Bui and Park, 2006; Zeng, 2006; Howarth, 2003, Caplan 2003) have been factored into this analysis.

4. Discussion

The following sections discuss the analysis of 659 metadata item records and associated findings. The conceptual ambiguities and semantic overlaps inherent in some DC metadata elements is also discussed in relation to semantic interoperability drawn from the analysis of metadata item records of the study.

4.1 Analysis and Findings

The analysis of 659 metadata item records brings to the fore the vital issues at play in terms of resource sharing and discovery across digital collections owing to inaccurate and inconsistent metadata usage

and mapping. To reiterate, in this study semantic interoperability across just three digital image collections using the same DC metadata scheme and the same digital collection management software, *CONTENTdm*, was seen to be hindered. The most frequently occurring inaccurate and inconsistent field names and metadata elements are listed below:

- The 'physical description' field is mapped onto either 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 locally added metadata elements concern provenance information such as: '*contact information*,' '*ordering information*,' and '*acquisition*.' In addition, the following are frequently identified as null mapping field names: '*full text*,' '*note*,' '*scan date*,' '*full resolution*,' and, '*image modification*'. These null mapping fields, other than provenance-related field names, raise the essential issue of educating cataloging professionals: these field names can indeed be mapped onto pertinent DC metadata elements.

Table 3 below represents the usage of DC metadata elements by three digital image collections (A, B and C).

Percentage of the Total Number of DC Metadata Elements Used by Three Collections (A, B, C)

DC Elements	A N=203	% of total number of DC elements used N=3476	B N=215	% of total number of DC elements used N=2721	C N=241	% of 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
Locally added elements	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 digital image collections:

Collection Name	Top Five Metadata Elements	Percentage
Collection A	*Subject, Format, Source, Contributor, Identifier*	54% of all metadata elements within the collection
Collection B	*Subject*, *null mapping fields*, *Coverage, Publisher, Title*	53% of all metadata elements within the collection
Collection C	*Subject, Format, Description, Title, Coverage*	71.2% of all metadata elements within the collection

Table 4. Top five metadata elements and the percentage of usage

Among the three collections, the following metadata elements are the most frequently employed, in descending order: *Subject, Description, Title, Format* and *Coverage* across the 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 use of *Description* and *Format* does not precisely reflect actual usage owing to inconsistent and inaccurate metadata mapping among the 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 serve to 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 indicates creation of the title by the cataloger.

The results of this study highlight the critical need for a mediation mechanism that catalogers can refer to during the process of metadata creation and 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,' and 'topic,' suggests the need for future studies in this area, especially since the high usage of this particular data element is derived from the combination of different types of controlled vocabulary schemes. Such practices will necessarily involve the vital issue of interoperability across heterogeneous controlled vocabulary schemes.

4.2 Semantic overlaps in DC metadata elements

As shown in the previous section, the inherent conceptual ambiguities and semantic overlaps in some of the DC metadata elements affect semantic interoperability across the three digital image collections. In addition to the lack of surrounding context in which a DC metadata element and its usage (i.e., definition) occur, semantic overlap among certain DC metadata element names and their corresponding definitions create conceptual ambiguity and consequently hinder accurate, consistent and complete application of the DC metadata scheme. Caplan (2003, 78) also points out the issue of semantic overlap: "Despite the simplicity of the Dublin Core Scheme, certain problems have arisen repeatedly in applications. One issue concerns the overlap in meaning in the definition of some elements."

As illustrated in an earlier section, there is great confusion in the usage of DC metadata elements such as *Format* and *Type*. The following are the definitions of these element names from both qualified and unqualified DC metadata schemes (DCMI, 2005):

- *Format* is "physical or digital manifestation of the resource" – unqualified DC metadata: *Format*
- *Type*: "image may include both electronic and physical representations" – qualified DC metadata: *Type, DCMI type vocabulary* on image

According to the crosswalk (LOC, 2001) from MARC to qualified DC, *physical description* (i.e., 300$a) in the MARC field can be mapped onto the DC *Format* element.

The definitions above, as well as the crosswalk evince semantic overlaps among *Type* and *Format,* and *Physical Description* (MARC 300$a) in the sense that the semantic boundaries among these elements are fuzzy and not clear cut; consequently, they may be used interchangeably with resulting confusion. The examples below of the *Format* element from the NSDL metadata repository illustrates the mixed us-

age among these elements (i.e., *Format, Type, Physical Description* (MARC 300$a)), indicating confusion due to semantic overlaps among these elements: (Bui and Park, 2006):

text
text/html
digital TIFF
image/tiff
1 v. (various pagings)
10 p., [6] p. of plates
MPEG-4

Table 5. Mixed use of DC metadata elements: Format with Type and Physical Description (MARC 300$a)

As mentioned before, the analysis of 659 Dublin Core metadata item records also shows evidence of frequent inaccurate usage between *Type* and *Format*. The confusion of application of these metadata elements is also reported by Zeng's (2006) analysis of the NSDL metadata repository: "Particular areas where confusion occurs are between *Type* and *Format*."

Semantic overlap also occurs with the following DC metadata elements: *Source* and *Relation*. The definitions of these elements below are from both unqualified and qualified DC metadata schemes (DCMI, 2005):

- *Source* is "a reference to a resource from which the present resource is derived." – unqualified DC scheme: *Source*
- *Relation* is "the described resource is a physical or logical part of the referenced resource." – qualified DC scheme: *Relation*, *is Part of Relation* is "the described resource is a version, edition, or adaptation of the referenced resource." – qualified DC scheme: *Relation*, *is Version of*

These definitions present semantic overlaps between *Source* and *Relation* stemming from the way *Source* is seen as a particular type of *Relation*.

According to the results of this study, the DC metadata elements *Source* and *Relation* are the infrequently employed metadata elements in digital image collections: *Source* (54.9%) and *Relation* (33.8%) out of the total usage of DC metadata element in the three digital image collections (see table 3). *Source* and *Relation* are also interchangeably used. This hinders semantic interoperability across digital collections and has a negative effect on metadata quality. Infrequent and inaccurate usage of *Source* and *Relation* is also the case in the NSDL metadata repository. According to Bui and Park (2006), the usage of *Source* in 111 collections is less than 15% and that of *Relation* is less than 7%. Zeng's (2006) analysis of the NSDL metadata repository also presents this: "Particular areas where confusion occurs are between … *Relation* and *Source* …."

The DC metadata elements *Creator, Contributor*, and *Publisher* also present semantic overlaps, as shown below. All definitions are from the unqualified DC metadata scheme (DCMI, 2005):

- *Creator* is "An entity primarily responsible for making the content of the resource."
- *Contributor* is "An entity responsible for making contributions to the content of the resource."
- *Publisher* is "An entity responsible for making the resource available."

According to these definitions, *Creator* can be seen as both a particular type of *Contributor* and *Publisher*. Caplan (2003) points out that, at one point, a proposal to combine the elements *Creator, Contributor* and *Publisher* into a single element called "agent" was considered and rejected due to complications. Thus, at this point there are no refining qualifiers to specify the meaning of these elements. This semantic overlap engenders confusion and inaccuracy in the usage of the DC metadata elements *Creator, Contributor* and *Publisher*.

As presented in the previous section, the inherent semantic overlaps in DC metadata elements affect semantic interoperability across digital image collections by contributing to inaccurate and inconsistent metadata description. The consequences of such inherent semantic overlaps and the conceptual ambiguities of some DC metadata elements are reflected in this empirical study of 659 metadata item records from three digital image collections and in the study of metadata quality analysis of the NSDL digital repositories (Bui and Park, 2006; Zeng, 2006).

5. Issues and implications

The following sections cover issues and implications drawn from the result of the study. A mediation mechanism that facilitates proper interpretation of metadata concepts and accurate and consistent usage of data elements during the metadata creation and mapping process is discussed. As well, the evolving nature of DC metadata semantics is presented by

utilizing analogy to the characteristics of the evolution of a natural language.

5.1 Concept networks as a mediation mechanism for knowledge organization

The results of this study suggest the critical need for mediation mechanisms that provide contextual relations among metadata elements and their corresponding definitions and usage in order to facilitate metadata creation and mapping process through the mitigation of semantic ambiguity. Current utilization of unqualified Dublin Core metadata, especially by non-cataloging professionals, brings to the fore the necessity for such a mediation mechanism. As an example, a significant number of data providers for Open Archive Communities, such as OLAC (Open Language Archives Community), are non-cataloging professionals who lack education and practicum related to information organization and access (Park, 2004a).

As illustrated in Howarth (2003), hindrances and problems in metadata mapping result from the absence of the context in which a metadata element name and its usage (i.e., definition) occur. This lack of contextual attributes creates semantic ambiguity and consequently hinders accurate, consistent and complete metadata application. Knowing and locating where a vocabulary element is visually placed in a concept network is an essential part of acquiring the meaning of the term (Miller et al., 1990). Knowledge of the meaning of a term in a concept network has great potential to improve usage of metadata elements and consequently improve the metadata creation and mapping process between DC metadata elements and author or cataloger generated vocabularies. In this sense, concept networks can be utilized as a mediation mechanism that enhances the metadata creation and mapping process by disambiguating semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context (Park, Forthcoming).

The structure of a concept network can be designed to comprise conceptual categories that share a core semantic property. To illustrate: the concept of 'contributor,' 'creator' or 'performing body' may share the core semantic property of 'intellectual responsibility of a work;' this can be categorized into the same conceptual category under 'name.' Conceptual categories can be organized into a hierarchical structure, i.e., conceptual taxonomy. A concept network may also consist of a description that represents conceptual relations among terms and categories. Such concept description should be designed for disambiguating any semantic overlaps among metadata elements (e.g., creator, contributor and performing body) that share a core semantic property (e.g., intellectual responsibility).

Instances, a brief definition, and a scope note if necessary for further disambiguation of a concept can also be part of the structure of concept networks. Conceptual relations are expressed by a variety of semantic features such as thing (i.e., object), people (i.e., agent, actor), event (i.e., process), time/aspect, place (i.e., location), and instrument. Concept networks can be visually expressed by employing a small number of notations and symbols. For instance, a rectangular box may represent a conceptual category. Conceptual relations between concepts and conceptual categories can be represented by nodes (points) and conceptual relations between concepts can be expressed by links. Instances may also be represented by a vertical line.

The concept networks can be modified and enhanced through an iterative process of analyzing conceptual structure. In other words, addition or deletion of the instantiation of a concept can affect the structure of concept networks in aspects such as conceptual taxonomy, conceptual relations, definition and scope note. Concept networks have good potential to facilitate proper interpretation of metadata concepts and accurate and consistent usage of the data elements during the metadata creation and mapping process, for both non-cataloging professionals as well as cataloging professionals.

5.2 DC metadata scheme as an evolving language

One of the salient characteristics of natural language is that it is akin to a living organism in the sense that word meanings are constantly evolving through the extension of core meaning, through the coining of new words and through the obsolescence of older words. For instance, the physical sight sense of the perceptual verb *see* has extended to comprise the mental sight of understanding and knowledge (e.g., *I see = I understand*). As the language evolves, the concrete core meaning of physical sight has extended to the abstract meaning of mental vision through metaphorical semantic extension. As Sweetser (1990, p. 21) points out, metaphorical semantic extension from physical activities to mental activities does not occur arbitrarily but rather with motivational ground; that is, through "shared structural properties" between the two physical and abstract domains

(e.g., physical sight vs. mental vision), specifically "our ability to focus our mental and visual attentions." Newly coined words or obsolescent and dead words and expressions are similarly representative of the constant evolution of language.

The above process constitutes the phenomenon of semanticization (Hopper and Traugott, 1993). The semanticization phenomenon underscores inherent properties of natural languages: flexibility and creativity of language use, as well as complexities of lexical meaning. In other words, by extending a given lexeme or word, language communities fulfill needs for expressing new and abstract concepts such as moral values, perspectives, attitudes and beliefs. As a consequence of the semanticization process, multiple and extended meanings of a concept may develop over time.

DC metadata can also be seen as akin to a language in the sense that the DC metadata scheme is continuously evolving. Baker (1998) employs a creative analogy, utilizing a phenomenon in natural language, to describe the basic DC 15 metadata elements and their evolution into more sophisticated metadata scheme: The DC metadata scheme is analogous to a very simple Pidgin language, which lacks the principal grammatical and lexical features of a standard language. For example, Hawaiian Pidgin originated from a multicultural environment owing to waves of immigration from different countries and ethnically heterogeneous plantation life. In this environment, communication was facilitated by employing only the core semantic meaning of English lacking any structured grammatical elements such as defined word order or morphemic rules that specify and refine a meaning of a word.

Users in the digital universe represent diverse linguistic and cultural backgrounds; they are like tourists in the sense that they utilize very simple vocabularies to communicate across domains, languages and cultures in the process of seeking and sharing information. In this sense, the simple vocabulary scheme is analogous to a Pidgin lacking fully fledged grammatical and lexical features (NISO, 2004).

Over time, by frequent use, a Pidgin may evolve into a Creole, which employs fully-fledged grammatical and lexical features and is structured virtually to the same extent as an established language. In analogy, the Pidgin-like simple DC metadata scheme may evolve into a more sophisticated Creole in order to facilitate the needs of diverse communities and provide means for effective communication to information users across languages, cultures and domains, as Hawaiian Creole developed from plantation Pidgin. Qualified DC metadata through refinement and usage of an encoding scheme that specifies and refines metadata element values can be seen as equivalent to some elements of the grammatical and lexical features of a fully fledged Creole. As well, Baker's analogy of creolization vis-à-vis the DC metadata scheme denotes one of the major characteristics of natural language: constant evolution through language use in the real context of the universe of discourse.

As discussed in earlier sections, semantic interoperability across digital collections utilizing the DC metadata scheme is hindered partially due to the drawbacks inherent in the semantics of the scheme. To become a fully fledged Creole, referring to Baker's analogy, the DC metadata scheme needs to further evolve in order to disambiguate the semantic relations of the DC metadata elements that present semantic overlaps. Conceptual ambiguities between some DC element names and their corresponding definitions also need to be disambiguated based on an empirical analysis of usage of the DC metadata scheme in the digital information sphere.

6. Conclusion and future studies

Assessment of metadata creation and mapping based on an analysis of 659 metadata item records shows that the metadata elements that engender particular difficulty and significant confusion during the metadata creation process are *Format*, *Type*, *Description*, *Source* and *Relation* (see also Bui and Park 2006, Zeng 2006). The most frequently occurring locally added metadata elements concern provenance information such as *Contacts*, *Acquisition Date* and *Ordering Information*.

The high usage of the '*subject*' data element in this study indicates that cataloging professionals are especially cognizant of the value of subject access. However, the high usage of this particular data element is derived from the combination of different types of controlled vocabulary schemes. These practices bring to the fore the inevitable issue of interoperability across heterogeneous controlled vocabulary schemes. Some of the following aspects of controlled vocabulary schemes for subject access need to be studied further: current schemes for subject access (e.g., *Library of Congress Subject Headings*, *Art & Architecture Thesaurus*, *Thesaurus for Graphic Materials*) relating to the nature of resources; the effectiveness and consequences of utilizing several different types of controlled vocabulary schemes in describing the

same type of resources; and the most desired controlled vocabulary schemes for describing digital image collections.

As reflected in this empirical study of 659 metadata item records from three digital image collections and in the study of metadata quality analysis of the NSDL digital repositories (Bui and Park, 2006; Zeng, 2006), conceptual ambiguities between some DC metadata element names and their corresponding definitions and semantic overlaps among some DC metadata elements affect the accurate, consistent and complete application of DC metadata. This in turn has great potential to hinder semantic interoperability vis-à-vis metadata quality.

Concept networks have good potential as a mediation mechanism that can facilitate proper interpretation of metadata concepts and accurate and consistent application of data elements during the metadata creation and mapping process. Concept networks can serve to disambiguate semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. The development of such a mediation mechanism calls for further studies on cataloger metadata creation and mapping practices and user studies on image searches.

Some of the areas for possible future studies relate to the application of metadata quality factors: metadata application guidelines (i.e., content specification) and procedures for cataloging professionals to follow during the creation of descriptive metadata elements and application of controlled vocabularies; employment of metadata guidelines and controlled vocabulary schemes in relation to the nature of the collection (e.g., manuscripts, dictionaries, photos) and resource media type (e.g., sound, text, image); relevance of a selected metadata standard to digital collections; criteria and reasoning behind local additions and variation of metadata element values to and from selected metadata and controlled vocabulary schemes; and measures and methods used by libraries for metadata quality control; adequate training of cataloging professionals and expectations of catalogers regarding a support and mediation mechanism for metadata creation and mapping.

References

Baker, Thomas. 1997. Metadata semantics shared across languages: Dublin Core in languages other than English, at http://dublincore.org/documents/1997/03/multilingual-semantics/.

Baker, Thomas. 1998. Languages for Dublin Core. In *D-lib magazine*. http://www.dlib.org/dlib/december98/12baker.html

Barton, J., Currier, S., and Hey, J.M.N. 2003. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. *2003 Dublin Core Conference*. http://purl.oclc.org/dc2003/03barton.pdf

Blair, D.C. 1990. *Language and representation in information retrieval*. Amsterdam: Elsevier.

Bruce, Thomas R. and Hillmann, Diane I. The continuum of metadata quality: defining, expressing, exploiting. In Diane Hillmann & Elaine L. Westbrooks, eds. *Metadata in practice.* Chicago: American Library Association.

Buckland, Michael. 1999. Vocabulary as a central concept in library and information science. In Arpanac, T. et al. eds. *Digital libraries: interdisciplinary concepts, challenges, and opportunities: proceedings of the Third International Conference on Conceptions of Library and Information Science*, Dubrovnik, Croatia, 23-26 May 1999. Zagreb: Lokve, pp 3-12. Also available at: http://www.sims.berkeley.edu/~buckland/colisvoc.htm.

Bui, Yen and Park, Jung-ran. 2006. An assessment of metadata quality: a case study of the National Science Digital Library Metadata Repository. In Moukdad, Haidar, ed. *Information science revisited: approaches to innovation*, *CAIS/ACSI 2006 Proceedings of the 2006 annual conference of the Canadian Association for Information Science*, York University, Toronto, Ontario. June 1 - 3, 2006. http://www.cais-acsi.ca/proceedings/ 2006/bui_2006.pdf

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). Place: Publisher, pp. 261-8.

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.*

Caplan, Priscilla. 2003. *Metadata fundamentals for all librarians*. Chicago: American Library Association.

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

Dahlberg, I. 1996c. Compatibility and integration of order systems 1960-1995: an annotated bibliography. *Compatibility and integration of order systems* (Research Seminar Proceedings of the TIP/ISKO Meeting), Warsaw, 13-15 September, 1995. Warsaw Wydawnictwo SBP. Warsaw, 13-15 September, 1995. Warsaw: Wydawnictwo.

Doerr, M. 2001. Semantic problems of thesaurus mapping. *Journal of digital information*, 1(8) http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

Dublin Core Metadata Initiative. 2005. *DCMI metadata terms*. http://dublincore.org/documents/dcmi-terms/

Dublin Core Metadata Initiative. 2005. *Using Dublin Core-Dublin Core qualifiers*. http://dublincore.org/documents/usageguide/qualifiers.shtml

Duval, E., Hodgins, W., Sutton, S., and Weibel, S.L. 2002. Metadata principles and practicalities. *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 Hillmann, Diane & Elaine L. Westbrooks eds.. *Metadata in practice.* Chicago: American Library Association.

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

Getty Research Institute. 2000. *Metadata standards crosswalks*. http://www.getty.edu/research/institute/standards/intrometadata/3_crosswalks/index.html

Godby, C. J., Smith, D., and Childress, E. 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 .

Heery, Rachel. 2004. Metadata future: steps towards semantic interoperability. In Hillmann , Diane and Elaine L. Westbrooks, eds. *Metadata in practice*. Chicago: American Library Association.

Heflin, J. and Hendler, J. 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 Knab, A.R. 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

Hopper, P. J. and Traugott, Elizabeth C. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Hovy et al. 2001. *Multilingual information management: current levels and future abilities*. Pisa: Italy Insituti Editoriali e Poligrafici Internazionali.

Howarth, Lynne C. 2001. Designing a metadata-enabled namespace for enhancing resource discovery in knowledge bases. In Guerrini and Sardo eds. *Proceedings International Conference Electronic Resources: Definition, Selection and Cataloguing.* http://w3.uniroma1.it/ssab/er/relazioni/howarth_eng.pdf

Howarth, Lynne C. 2003. Designing a common namespace for searching metadata-enabled knowledge repositories: an international perspective. *Cataloging & classification quarterly*, 37(1/2): 173-185.

Hunter, Jane. 2001. MetaNet-a metadata term thesaurus to enable semantic interoperability between metadata domains. *Journal of digital information*, 1(8).

Lancaster, F. Wilfrid. 1986. *Vocabulary control for information retrieval*, Arlington, Va.: Information Resources Press.

Lassila, O. 1998. Web metadata: a matter of semantics. In *IEEE internet computing*, 2(4). 30-37.

Library of Congress. 2001. *MARC to Dublin Core crosswalk*. http://www.loc.gov/marc/marc2dc.html

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

Matthews, Brian and Wilson, Michael. 2000. Multilingual metadata to access social science data. In *Information systems engineering, CLRC-Ral.* http://www.dl.ac.uk/TCSC/datamanagement/workshop/mathews1.html.

McClelland, M. et al. 2002. Challenges for service providers when importing metadata in digital libraries. In *D-lib magazine* 8(4). http://www.dlib.org/dlib/april02/mcclelland/04mcclelland.html

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

Miller, G.A. et al. 1990. Introduction to WordNet: an on-line lexical database. *International journal of lexicography* 3: 235-44.

Moen, W.E., Steward, E.L., and McClure, C.R. 1997. The role of content analysis in evaluating metadata for the U.S. Government Information Locator Service: results from an exploratory study. http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.

National Information Standards Organization. 2004. *Understanding metadata*. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Neuroth, Heike and Koch, Traugott. 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. *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. *Journal of internet cataloging* 5(3): 59-79.

Park, Jung-ran. 2004a. ALA accredited schools offering a metadata course: a survey based on curriculum information provided by 53 ALA accredited LIS program websites. Unpublished manuscript.

Park, Jung-ran. 2004b. Language-related open archives: impact on scholarly communities and academic librarianship. *E-JASL: The electronic journal of academic and special librarianship* 5(2/3). http://southernlibrarianship.icaap.org/content/v05n02/park_j01.htm

Park, Jung-ran. 2005. Semantic interoperability across digital image collections: a pilot study on metadata mapping. In Vaughan, Liwen ed. *Data, information, and knowledge in a networked world*, *Proceedings of the 2005 annual conference of the Canadian Association for Information Science*, University of Western Ontario, London, Ontario, June 2 - 4, 2005. http://www.cais-acsi.ca/proceedings/2005/park_J_2005.pdf

Park, Jung-ran. (Forthcoming). Evolution of concept networks and implications for knowledge representation. *Journal of documentation*.

Purat, Jacek. 1998. The world of multilingual environmental thesauri. http://www.sims.berkeley.edu/~purat/world_multilingual_environmental_

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., and Katz, S. 2004. Reengineering thesauri for new applications: the AGROVOC example. *Journal of digital information*, 4(4). Themes: Digital Libraries, Information Discovery.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization.* Cambridge: MIT Press.

Sweetser, Eve. 1990. *From etymology to pragmatics: metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Vizine-Goetz, D., Hickey, C., Houghton, A., and Thompson, R. 2004. Vocabulary mapping for terminology services. *Journal of digital information* 4(4), Themes: Digital libraries, information discovery. http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

Zeng, Marcia. 2006. Metadata quality study for the National Science Digital Library (NSDL) Metadata Repository. Presented paper at the *Research and Teaching Talk Series* in Information Science and Technology at Drexel University.