The Semantics of Semantic Interoperability: A Two-Dimensional Approach for Investigating Issues of Semantic Interoperability in Digital Libraries

EunKyung Chung, eunkyung.chung@usm.edu
School of Library and Information Science, University of Southern Mississippi

William E. Moen, wemoen@unt.edu
School of Library and Information Science, University of North Texas

Introduction

The networked information environment comprising digital libraries, digital collections, and digital repositories increase people's expectations for information access. Specifically, users anticipate better search capabilities across these networked information resources and the metadata records associated with the resources. The opportunities also exist for improved information exchange or integration in and between communities and applications (Arms et al., 2002; Chen, 1999; Moen, 2001; Tennant, 2001; Zeng & Chan, 2004). To improve capability for information exchange or integration, and to improve search across these resources, it is essential to enhance interoperability of systems and data. This involves addressing not only interoperability at the syntactic and functional levels but at the semantic level (Moen, 2001). Various information services such as federated searching, harvesting metadata, and gathering are often dependent on the capability and quality of semantic interoperability across different collections, systems, domains, and communities (Arms et al., 2002). Improved semantic interoperability provides a basis for more effective information services for users. Effective semantic interoperability is required to meet users' expectations of discovering relevant networked information resources via searches across heterogeneous as well as within homogeneous systems and collections (Chen, 1999; Tennant, 2001).

Although there have been efforts in improving interoperability, many of those focused on protocol and other aspects of interoperability. It is important at this point to examine the current status of semantic interoperability in terms of digital libraries, collections, and repositories across domains, communities, and applications. The purpose of this preliminary study described here is to suggest one approach for addressing some aspects of semantic interoperability and to use this approach to examine four implementations in different communities and applications using Greenstone, Fedora, DSpace, and EPrints. Specifically, four selected implementations were examined with respect to two fundamental dimensions of semantic interoperability: data attribute (i.e., names of fields or elements in the metadata records) and data value (i.e., the actual content of the fields or elements). This study intends to show how this methodological approach can identify potential challenges to semantic interoperability and subsequent studies using this method can yield results and findings that may lead to common guidelines, suggestions, and solutions for the digital library community in terms of semantic interoperability.

Related Research and Research Methodology

Different types or levels of search and data interoperability, such as syntactic, functional and semantic, have been identified (Moen, 2001). While syntactic and functional levels

may focus on protocols used for information retrieval system communication, the semantic level concerns the understandings and meanings of interchanged data. (Ouksel & Sheth, 1999). In the hierarchy of challenges for effective interoperability, semantic interoperability is regarded as a grand-challenge research area for achieving high-quality interoperability (Lynch & Garcia-Molina, 1995). Although there are complex challenges to achieve semantic interoperability, many efforts have been made to improve semantic interoperability (Chen, 1999) to support high quality information exchange and integration across communities.

Semantic interoperability research falls generally into two primary areas: data attributes and data values. The data attribute area addresses questions as to the names, labels, semantics, and granularity of metadata scheme elements and database fields. The data value area addresses the data or information provided in an element or database field. As networked information resources expand in type and numbers in collections, libraries, and repositories, various general or domain-specific, standard and nonstandard metadata schema are being used. Moen (2001) and Zeng and Chan (2004) pointed out, that the diversity of metadata schemes, schemas, and element sets in use means that the data-attribute area is one to be considered to improve semantic interoperability. Accordingly, a range of approaches have been attempted to improve semantic data attribute interoperability, and these approaches can be categorized as either pre-coordinated or post-coordinated. The a priori agreement approach consists of specifying through the use of application profiles or registries or other methods the data attributes (i.e., metadata elements) to be used in systems. Diverse repositories or libraries all agree to use the same data attributes. A post-coordinated approach, on the other hand, endeavors to map, match, or crosswalk different data attributes after specific digital libraries, collections, and repositories are implemented. Buckland (1999) proposed yet another way of dealing with the diversity of data attributes by mapping users' query vocabularies to metadata vocabularies within various systems.

Semantic interoperability related to the data-value area focuses on the data content or information associated with the data attributes, and one of the most critical aspects of this is related to subject-related search and retrieval. Ontologies and controlled vocabularies are constructed with different scopes and levels of abstraction and granularity as needed by various disciplines, and this increases the difficulty to achieve semantic interoperability with data values than with the data-attribute dimension. Proposals for dealing with data value interoperability have been suggested such as Kuhr's (2003) mega-thesaurus to map multiple controlled vocabularies. Doerr (2001) approached the issue with a concept-based mapping between numerous vocabularies for subject retrieval.

Figure 1 illustrates the relationship between the two dimensions—data attribute and data value in the context of semantic interoperability. As semantic interoperability of the two dimensions increases, overall semantic interoperability increases accordingly.

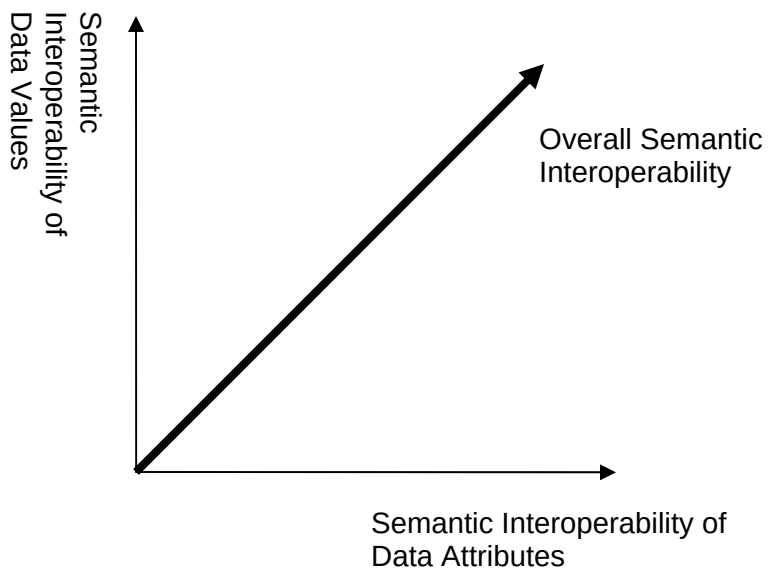

Figure 1. Dimensions of Semantic Interoperability

Table 1 specifies components of the two dimensions. The data attribute dimension's components are: semantics and contents (Chan & Zeng, 2006). The semantics aspect is designed to constrain the attributes according agreed upon meaning and the data to be associated with that attribute. Using a searching example, if a user wants to search across repositories for "author" information, the target repositories should execute the search against the author attribute (or element or field). Thus, the system maintains the user's semantic intention with the search (and not expand the search to other attributes). The other variation is to match users' searching intentions with semantically related attributes. There might be a case in which a user wants to search information with an "author" attribute, but the target systems label that attribute as "originator," "creator," or "writer." To achieve a higher degree of semantic interoperability between these systems, the seemingly dissimilar but semantically related attributes should be connected (or "mapped") with each other in a proper way. The contents aspect is defined as "declarations or instructions of what and how values should be assigned to the elements" (Chan & Zeng, 2006). For instance, the "creator" attribute could be prescribed differently in terms of the data values in it depending on specific collections, communities, and domains.

The data-value dimension of semantic interoperability is more complex and difficult to achieve, since the main concerns of the dimension are dependent on specific controlled vocabularies and ontologies. In addition to addressing various controlled vocabularies and ontologies, the data-value dimension is interrelated with aspects of linguistics and meanings of words such as polysemy, synonymity, and granularity, especially in subject-related attributes. Based on the elements of the data-value dimension, with respect to interoperability among different collections, domains, and communities, it becomes much more complicated to achieve semantic interoperability than in homogeneous settings.

Table 1. The Elements in two dimensions

Dimension	Element	Example
Data Attribute	Semantics	The meaning of "Creator" is defined as the party or parties responsible for creating the specific resource
	Content	The "Creator" attribute should be assigned to individuals or organizations responsible for the intellectual contents of a specific resource, rather than the physical contents
Data Value	Vocabularies and Ontologies	Different controlled vocabularies and ontologies
	Granularity	Different levels of terms assigned in the same concept (e.g., vehicle vs. sedan)
	Word	Polysemy and synonymity

Applying the Two Dimensional Approach

Digital libraries and repositories have been implemented in many communities and domains for a variety of collections, and using several software platforms and applications such as Greenstone, DSpace, Eprints, and Fedora. Four implementations, The Writing University Archive, Duke Law Faculty Scholarship Repository, Drexel University E-repository and Archives, and Tufts Digital Library were selected for analysis using the two-dimensional perspective. These collections primarily contain scholarly works from specific institutions and were designed to assist teaching and research activities. Table2 summaries key details of these implementations.

Table 2. DL examples across different applications

DL Title	URL	Application	Data Attributes
The Writing University Archive	http://iwp.info-science.uiowa.edu/cgi-bin/library	Greenstone	Titles, Authors, Country, Year
Duke Law Faculty Scholarship Repository	http://eprints.law.duke.edu/	Eprints	Title, Authors, Abstract, Keywords, Subjects, Type, Conference, Department, Editors, Status, Refereed, Publication, Year
Drexel University E-repository and archives	http://dspace.library.drexel.edu/	DSpace	Collections, Titles, Authors, Subjects
Tufts Digital Library	http://dl.tufts.edu/	Fedora	Creator, Collection, Date, Description, Organizations, People, Places, Topics, Title, Type

Table 2 also lists the data attributes identified through an examination of each of these implementations. This shows that data attributes occur that are semantically related but given different attribute names such as authors, creators, and contributors. Secondly, certain data attributes may occur in one or more implementations but not in all. For example, there is no subject attribute in The Writing University Archive, while it occurs in the others. In addition, there are more granular attributes in an implementation that could be subsumed by more broadly defined attributes in other implementations. For example, the Tufts Digital Library has data attributes of topics, people, places, and organization that could be considered subject-related attributes. The contents component in data attributes were examined in terms of how and what to assign the values in. In the case of the author or creator data attribute in The Writing University Archive and Drexel University E-repository and archives, the primary description method is in the order of last name and first name, but there is a differently abbreviated description issue, especially in the first name.

In addition, the data-value dimension was examined with respect to subject or topic data attributes. There is variation in which controlled vocabularies and ontologies are used in the different implementations for assigning subject terms as data values to the appropriate data attributes. The Tufts Digital Library defines the data value of the subject attribute as "terms or values selected for identification of broader categorization of an object"[1] without specifying any controlled vocabularies. The Duke Law Faculty Scholarship Repository uses Library of Congress Subject Headings for terms to use in the subject attribute. In addition, Drexel University E-repository and archives uses in-house controlled vocabulary for the subject attribute. This preliminary study primarily examined the utilization of various controlled vocabularies and ontologies, rather than granularity and word elements in the data-value dimension, but this method will support examining those other components in this dimension.

Conclusion

Semantic interoperability plays a key role in providing quality information services in the networked information environment. This preliminary study related to semantic interoperability examined four implementations within similar domains of knowledge from a two-dimensional perspective. The data-attribute dimension, semantics and contents of the data attribute were examined in terms of different naming practices, occurring and non-occurring attributes, granular and more general attributes that may be semantically related. and abbreviated-designation issues. The data-value dimension was examined with respect to different controlled vocabularies and ontologies in the subject-related attributes. Future research using this two-dimensional perspective can investigate issues and potential challenges for semantic interoperability in digital libraries, repositories, and archives across domains of knowledge, communities, and applications. Extending and expanding this research can contribute to a more comprehensive survey these networked information resources, and the findings from the research can result in suggestions and solutions to improve semantic interoperability for these collections and systems.

[1] http://dca.tufts.edu/tdl/search_help/index.html#advanced_search

References

Arms, W. Y., Hillmann, D., Lagoze, C., Kraft, D., Marisa, R., Saylor, J., et al. (2002). A spectrum of interoperability. *D-Lib Magazine, 8*(1). Retrieved Feb 10 2007, from http://www.dlib.org/dlib/january02/arms/01arms.html.

Buckland, M. (1999). Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine, 5*(1). Retrieved Retrieved Feb 10 2007, from http://www.dlib.org/dlib/january99/buckland/01buckland.html.

Chan, L. M. & Zeng, M. L. (2006). Metadata interoperability and standardization – a study of methodology I: Achieving interoperability at the schema level. *D-Lib Magazine, 12*(6). Retrieved Feb 11 2007 from http://www.dlib.org/dlib/june06/zeng/06zeng.html.

Chen, H. (1999). Semantic research for digital libraries. *D-Lib Magazine*, 5(10). Retrieved Feb 9 2007 from http://www.dlib.org/dlib/october99/chen/10chen.html.

Doerr, M. (2001). Semantic problems of thesaurus mapping. *Journal of Digital Information, 1*(8). Retrieved Feb 1 2007 from http://jodi.tamu.edu/Articles/v01/i08/Doerr/.

Kuhr, P. S. (2003). Putting the world back together: Mapping multiple vocabularies into a single thesaurus. In *Subject Retrieval in a Networked Environment: Proceedings of the IFLA Satellite Meeting*, *Dublin, OH* (pp. 37-42). München: K.G. Saur.

Lynch, C., & Garcia-Molina, H. (1995). Interoperability, scaling, and the digital libraries research agenda. Retrieved February, 1, 2007, from http://wwwdiglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html.

Moen, W. (2001). Mapping the interoperability landscape for networked information retrieval. In *Proceedings of 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 50-51). New York: ACM Press.

Ouksel, A. M., & Sheth, A. (1999). Semantic interoperability in global information systems: A brief introduction to the research area and the special section. *SIGMOD Record, 28*(1), 5-12.

Tennant, R. (2001). Different paths to interoperability. *Library Journal, 126*(3), 118-119.

Zeng, M. L., & Chan, L. M. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology, 55*(5), 377-395.