International Journal of Internet, Broadcasting and Communication Vol.10 No.2 1-6 (2018)
http://dx.doi.org/10.7236/IJIBC.2018.10.2.1

IJIBC 18-2-1

Handling Semantic Ambiguity for Metadata Generation

Gi-Chul Yang[1]*, Jeong-Ran Park[2]

[1]**Department of Convergence Software*, *Mokpo National University, Korea*
gcyang@mokpo.ac.kr
[2]*Colledge of Computing and Informatics, Drexel University, USA*
Jp365@drexel.edu

Abstract

The following research questions are examined in this paper. What hinders quality metadata generation and metadata interoperability? What kind of semantic representation technique can be utilized in order to enhance metadata quality and semantic interoperability? This paper suggests a way of handling semantic ambiguity for metadata generation. The conceptual graph is utilized to disambiguate semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. The mechanism introduced in this paper has the potential to alleviate issues dealing with inconsistent metadata application and interoperability across digital collections.

Keywords: *Metadata, Semantic Ambiguity, Digital Library, Conceptual Graph.*

1. Introduction

Please refer to this document and follow the specifics outlined below when submitting your final draft. These guidelines include complete descriptions of the fonts, spacing, and related information for producing Metadata is used to manage, organize and discover information resources. Metadata allows the organization of resources using precise description. This in turn facilitates effective resource discovery. Generation of quality metadata is critical in a networked world. Metadata interoperability, which is a critical factor for data sharing and exchange, is one of the top challenges in a networked environment. Quality metadata generation is an even more critical issue in the aggregated environment. In order to ensure meaningful metadata outside of its local context, it is essential that metadata interoperability be based on accurate and consistent resource description. Current metadata practices, however, still do not attain conditions for interoperability. The studies reveal that metadata interoperability remains a major challenge in the current networked environment despite growing awareness of its importance (Park, 2006).

The impact of metadata quality on resource discovery is significant. Problems inherent in the manual metadata creation process, such as inaccurate data entry (e.g., spelling, abbreviations), format of date (e.g., date of creation or date of publication) and consistency of metadata application result in adverse effects on resource discovery (Park, 2006). One of the advantages of automatic metadata generation is that it is apt to produce consistent metadata application which conditions metadata interoperability.

This study aims at presenting a mechanism of automatic metadata extraction which has a good potential to alleviate issues dealing with inconsistent metadata application and semantic interoperability across digital repositories. The following research questions are examined in this paper:

Manuscript Received: Feb. 10, 2018 / Revised: Feb. 17, 2018 / Accepted: Mar. 3, 2018
Corresponding Author: gcyang@mokpo.ac.kr
Tel: +82-61-450-2355, Fax: +82-61-450-2358
Department of Convergence Software, Mokpo National University, Korea

1. What hinders quality metadata generation and metadata interoperability?
2. What kind of semantic representation technique can be utilized in order to enhance metadata quality and semantic interoperability?

2. Semantic Ambiguity and Interoperability

The key components of metadata standards are semantics of metadata elements, content standards used for supplying appropriate data values for each metadata element and syntax used for encoding metadata elements and values. Metadata semantics concerns the semantics associated with metadata such as concepts, conceptual relations and definitions of metadata elements. Metadata semantics greatly affects consistent and accurate metadata application (Park and Childress, 2009).

As evinced through information sharing for non-networked traditional bibliographic collections in libraries, successful resource discovery and exchange across distributed digital repositories demands semantic interoperability based on consistent and accurate resource description. The flexibility and complex structure of natural language allows for the representation of a concept in various ways. In natural language, mapping between word forms and meanings can be many-to-many. That is, the same meaning can be expressed by several different forms (e.g., synonyms) and the same forms may designate different concepts (e.g., homonyms). These linguistic phenomena may engender confusion in the sense that different communities may use dissimilar word forms to deliver identical or similar concepts, or may use the same forms to convey different concepts.

For instance, Bui and Park (Park, 2006) examine metadata quality of the open source Metadata Repository at the National Science Digital Library (NSDL). The NSDL comprises over one hundred 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 derives from many different data providers. As well, these data providers utilize a variety of metadata schemes other than the Dublin Core (DC) metadata scheme. However for data harvesting purposes, all metadata schemes in NSDL are mapped onto the DC metadata scheme. In this mapping process, inaccurate and inconsistent mappings occur. Such drawbacks in mapping no doubt hinder semantic interoperability even among digital repositories employing identical vocabulary scheme such as Dublin Core metadata elements and identical digital collection software configuration. Park (Park, 2006) evidence this complexity of mapping between word forms and meanings and its impact on metadata application and semantic interoperability.

The semantic interoperability can be seen as its ability to disentangle the complex nature of mappings between word forms and meanings in natural language in order to enhance resource exchange and discovery within a community or between communities. Accurate and consistent metadata mapping between two or more different vocabulary schemes is a critical factor in achieving semantic interoperability. Efforts to increase semantic interoperability across heterogeneous vocabulary schemes have dramatically increased through harmonization and integration of heterogeneous vocabulary and metadata schemes by utilizing mapping mechanisms (e.g., Vizine-Goetz et al., 2004).

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

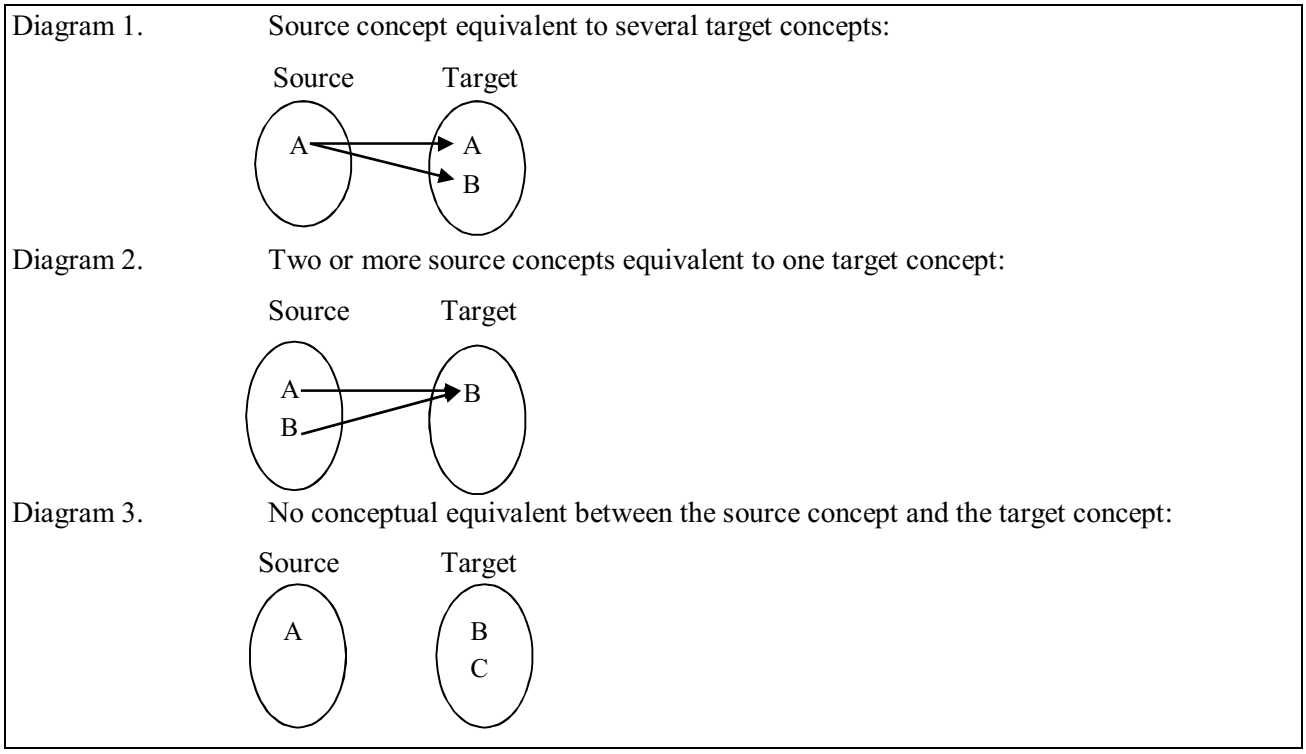

Figure 1. Concept Equivalence from Park (2002)

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 various tools that show conceptual relations and contextual attributes among terms thus enhancing semantic interoperability between the two languages. One of major factors hindering metadata mapping concerns the lack of a surrounding context in which a metadata element name and its usage occur (Howarth, 2003). This lack of contextual attributes may bring forth semantic ambiguity and further hinder quality metadata generation through accurate, consistent and complete metadata application.

The salient characteristics of the Dublin Core (DC) metadata scheme, which was approved as an ANSI/NISO standard (Z39.85) in 2001, emerge from its simplicity, flexibility, and interoperability. The functionalities of the DC metadata elements are easy to implement owing to this simplicity, which creates high compatibility across multiple repositories. However, despite such characteristics, there are inherent conceptual ambiguities and semantic overlaps in some of the DC metadata elements. In other words, 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. This consequently hinders accurate, consistent and complete application of the DC metadata scheme (Park and Childress, 2009).

In order to attain quality metadata and semantic interoperability, a mediation mechanism that provides contextual relations among metadata elements and their corresponding usage is essential. A conceptual network/graph is one of the formal languages that represent the meanings of natural language. Knowing and locating where a vocabulary element is visually placed in a concept network/graph is an essential part of acquiring the meaning of the term (Miller et al., 1990). Accordingly, concept networks/graphs has potential for being utilized as a mediation mechanism that facilitates the metadata creation and mapping process and automatic metadata extraction by disambiguating semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. In the following section, we will present conceptual graphs in detail.

3. Conceptual Graph

Font of the paper title must be in 14-point Times New Roman, boldfaced, centered, and multiple-spacing at 1.25. Leave 2 spaces after the paper title in 11-point, then type author/s name/s. Author/s name/s must be A conceptual graph (CG) is a finite, directed, connected, bipartite graph proposed by John Sowa(Sowa, 1984). There are two kinds of nodes; concept nodes (displayed as a box in graph notation) which represent entities, attributes, states, and events, and relation nodes (displayed as a circle in graph notation) which represent the relationship among concept nodes.

A CG can be constructed by assembling percepts. In the process of assembly, concept relations specify the role that each percept plays, and concepts represent percepts themselves. A concept can be generic or individual. The function referent maps concepts into a generic marker * or a se I = {#1, #2, #3, …}, the elements of which are individual markers. The function *type* maps concepts into a set of type labels. A concept c with *type* (c) = t and *referent* (c) = r is displayed as [t : r] in the linear form. The function type also can be applied to relation. A relation r with *type* (r) = t is displayed as (t) in the linear form. Types of concepts (type labels) are organized into a hierarchy called type hierarchy. Type hierarchy forming operations include conjunction and disjunction operations, so the type hierarchy will be a formal lattice. The type hierarchy constitutes a partial ordering and becomes a type lattice when all the intermediate types are introduced.

A CG can be represented in three different forms. There is a graphic notation called the *display form* (DF), a more compact notation called the *linear form* (LF) as well as a concrete syntax called the *conceptual graph interchange form* (CGIF), which has a simplified syntax and a restricted character set designed for compact storage and efficient parsing. Both DF and LF are designed for communication with humans or between humans and machines. For communication between machines, the CGIF has a simpler syntax

Figure 2 shows the display form of a conceptual graph that represents the prepositional content of the English sentence *Tom is going to New York by car*.

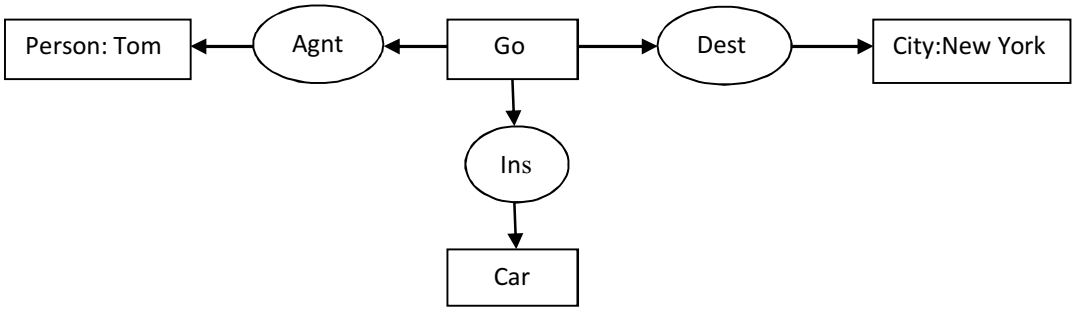

Figure 2. CG Display form for "Tom is going to New York by car

In DF, concepts are represented by rectangles: [Go], [Person: Tom], [City: New York], and [Car]. Circles or ovals represent conceptual relations: (Agnt) relates [Go] to the agent Tom, (Dest) relates [Go] to the destination New York, and (Inst) relates [Go] to the instrument car. The linear form for CGs is intended as a more compact notation than DF, but with good human readability. Following is the LF for Figure 2:

[Go]-

(Agnt)->[Person: Tom]

(Dest)->[City: New York]

(Inst)->[Car].

In this form, the concepts are represented by square brackets instead of boxes, and the conceptual relations are represented by parentheses instead of circles. A hyphen at the end of a line indicates that the relations attached to the concept are continued on subsequent lines. Following is the CGIF for Figure 1:

[Go *x] (Agnt ?x [Person: Tom]) (Dest ?x [City: New York]) (Inst ?x [Car])

or (Agnt [Go] [Person: Tom]) (Dest [Go] [City: : New York]) (Inst [Go] [Car])

It is beneficial to use CGIF to represent internal data when developing a system based on CGs since CGIF has machine friendly notation as we seen above. There are many knowledge handling CG applications such as in (Uta Priss et. Al., 2007) and (Yang & Choi, 1996).

4. Handling Semantic Ambiguity by using Conceptual Graphs

Unlike manually generated standard metadata, we are interested in to generate input data for digital collection management software (such as CONTENTdm) directly from information resources in order to produce outputs tagged with standard metadata such as Dublin Core (DC).

The DC metadata standard is a simple yet effective element set for describing a wide range of networked resources (Dublin Core Metadata Initiative, 2005). The Dublin Core standard comprises fifteen elements. Following are some examples:

Contributor: The DC element used to designate Person(s) or organization(s) in addition to those specified in the CREATOR element who have made significant intellectual contributions to the resource.

Creator: The person or organization primarily responsible for creating the intellectual content of the resource.

Title: The DC element used to designate the name given to the resource.

These elements can be matched with relation nodes of conceptual graph. Figure 3 shows the overall workflow of semantic ambiguity handling process.

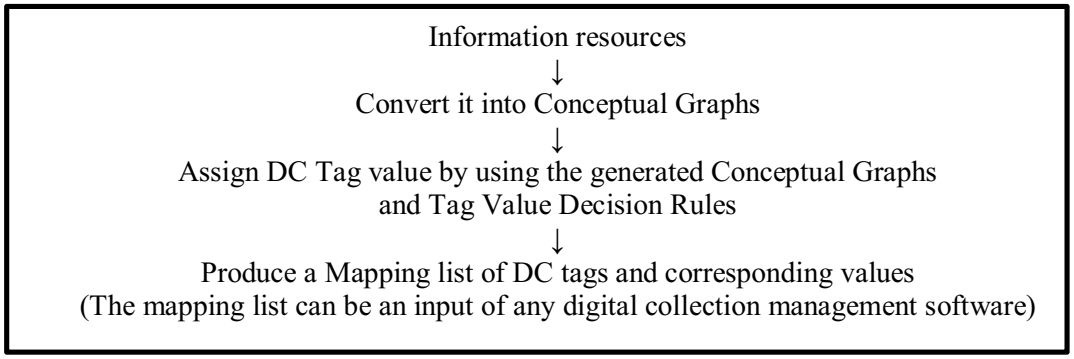

Figure 3. Workflow of Semantic Ambiguity Handling

As shown in Figure 3, the information resources should be converted into conceptual graphs. The remaining process is mapping between DC tag value and corresponding relation node values of the conceptual graph. In this process Tag Value Decision Rules are utilized. Hence, the capability of representing semantics of conceptual graph is very useful to handling the semantic ambiguity and interoperability for metadata

generation.

5. Conclusions

In this paper we discussed semantic ambiguity and interoperability for metadata generation. Metadata generation is one of important problems need to overcome for digital library collections. As evidenced through rapidly growing digital repositories and web resources, automatic metadata generation is becoming more critical, especially considering the costly and complex operation of manual metadata creation. The conceptual graph is utilized to disambiguate semantic ambiguities caused by isolation of a metadata element and its corresponding definition from the relevant context. The way of handling mechanism of semantic ambiguity for metadata generation introduced in this paper will provide a strong base for development of automatic metadata generation system.

Acknowledgement

This Research was supported (in part) by Research Funds of Mokpo National University in 2017.

References

[1] Park, Jung-ran. Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections. *Knowledge Organization,* Vol. 33 (1): 20-34. 2006.

[2] Park, Jung-ran. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Special Issue on Metadata and Open Access Repositories (M.S. Babinec and H. Mercer Eds.). *Cataloging and Classification Quarterly,* Vol. 47(3): 213-228. 2009.

[3] Park, Jung-ran and Childress, Eric. Dublin Core Metadata Semantics: An Analysis of the Perspectives of Information Professionals. *Journal of Information Science,* Vol. 35(6): 727-739. 2009.

[4] Bui, Yen and Park, Jung-ran. An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository. In Haidar Moukdad (Ed.). *Information Science Revisited: Approaches to Innovation*, CAIS/ACSI 2006 Proceedings of the 2006 annual conference of the Canadian Association f or Information Science held with the Congress of the Social Sciences and Humanities of Canada at the York University, Toronto, Ontario. June 1 - 3, 2006. http://www.cais-acsi.ca/proceedings/2006/bui_2006.pdf 2006.

[5] Vizine-Goetz, D., Hickey, C., Houghton, A., and Thompson, R. Vocabulary mapping for terminology servi ces. *Journal of Digital Information* 4(4), Themes: Digital libraries, information discovery. 2004.

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

[7] Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K., *Introduction to WordNet : An on-line lexical database*. Journal of Lexicography, 3, 235-244. 1990.

[8] Sowa, J. Conceptual Structure: Information Processing in Mind and Machine, Addison Wesley, Massachusetts, 1984.

[9] Uta Priss, Simon Polovina, Richard Hill, Conceptual Structures: Knowledge Architectures for Smart Applications, Lecture Notes in Computer Science, Volume 4604, 2007.

[10] Yang, Gi-Chul & K.S. Choi, Construction of a Knowledge Base by using Korean Text, AAAI96- Fall Symposium, MIT, U.S.A. 1996.