Modelling a "Human Understandable" Metalevel Ontology for Enhancing Information Seeking on the World Wide Web

Lynne C. Howarth
Faculty of Information Studies
University of Toronto
e-mail: howarth@fis.utoronto.ca

Abstract

With the explosion of digitized resources accessible via the World Wide Web, and the resultant proliferation of domain-specific schemes, metadata have assumed a prominence not previously experienced. Research into the application and impact of metadata standards as they relate to electronic resources has been minimal. Recent work emanating from the World Wide Web Consortium (W3C) has focussed on aspects that can best be described as being concerned with converting "machine-readable" into "machine-understandable". The next iteration from "machine-understandable" to "***human***-understandable" is noticeably absent. This apparent gap provides the framework for research with the following objectives, namely, (1) to determine and refine a metalevel scheme or terminological ontology which can serve as both a "metadata dictionary" (or "metadata *lingua franca*"), and a switching device for assisting end-users searching for metadata-encoded documents or document-like objects on the World Wide Web, (2) to develop a front-end, pop-up window prototype of that metalevel scheme to provide navigational assistance to searchers when required, and (3) to test whether the prototype ontological software tool enhances the information-seeking process, providing end-users with a greater depth and breadth of search options and/or improving satisfaction with search results and information discovery. This paper describes the components of a three-phase study whiich has recently been undertaken with an ultimate aim of developing an online "human-understandable" tool for more effective Internet searching.

1. Introduction

While the proliferation of electronic resources on the World Wide Web (hereafter referred to as "the Web") has increased the potential range and quantity of readily-accessible information, it has also resulted in what Levy (1990) refers to as a "second flood", threatening to drown the engaged searcher in massive amounts of

material, both useful and irrelevant. In past, perhaps less dramatic iterations of the "information explosion", attention has focused on creating and applying codes and standards to facilitate the identification of, and access to, different types and formats of materials, whether they be housed in library collections, listed in bibliographies or indexes, or stored in electronic databases. Uniformly structured, and consistently devised catalogue records, bibliographic citations, and indexes have served as the surrogates for actual objects or documents, describing their intellectual and physical characteristics, as appropriate. As we near the end of this millennium, however, the utility and relative value of creating bibliographic surrogates based on long-established codes and standards (such as the *International Standard Bibliographic Description*, the *Anglo-American Cataloguing Rules*, *Machine-Readable Cataloguing* [MARC], or *The Chicago Manual of Style*, to name only a select few) has come under scrutiny. Increasingly, reliance on standardized surrogates as "gateways" to multimedia objects or text-based documents is being viewed as cumbersome, and in terms of their creation, time-consuming and expensive.

2. Background to, and Rationale for the Research

The development and refinement of bibliographic codes and standards has occurred across a time-frame exceeding 125 years. In contrast, responses to the meteoric growth in the availability of, and pressing need for access to, electronic resources accessible via the Web have been, by necessity, focussed and relatively rapid. At the same time as the proliferation of electronic resources has strained the capacities of traditional frameworks for describing the intellectual and physical properties of objects or documents, projects for developing metadata schemes for identifying, accessing, and retrieving digital objects or documents have burgeoned. Defined as "data about data" (Miller 1996), metadata schemes provide a conceptual and ontological framework, identifying the "entities" or object types characteristic of a subject domain, assigning physical, intellectual, or logical properties or "attributes" to those entities, and making explicit relationships that may exist between or among entities (Olson and Kent 1998). Meta tags or naming devices are attached to the entity and can be used by search engines (in the World Wide Web context these include Netscape, Lycos, Yahoo, Excite, Infoseek, HotBot, LookSmart, Snap, etc.) to "harvest" electronic resources in response to an online query. Metadata may sit separately from the resources being described or included (embedded) as part of the electronic document or document-like object, *per se*. The encoding and transportation provisions of a metadata scheme may be based on a particular syntax such as the Standard Generalized Markup Language (SGML), developed for the description of mark-up

languages. Mark-up languages represent a formal system by which information or encoding is added to the electronic form of a document in order to represent its meaning and control its processing. SGML allows for mark-up languages to be defined in a way that is independent of any particular device or application, and thus facilitates the interchange and long term conservation of richly structured electronic resources (UKOLN Metadata Group, 1998). The Hypertext Mark-up Language (HTML), and the more recent eXtended Mark-up Language (XML) are derivations of SGML.

Since 1986, when SGML became an international standard (ISO 8879:1986), there has been steady activity to develop SGML/XML/HTML-based metadata standards for specialised information domains. The Government Locator Services (GILS) (for identifying information resources emanating from the United States and Canadian federal governments), Encoded Archival Description (EAD) (an SGML-based encoding scheme for archive and library finding aids), the Dublin Core (DC) (a simple HTML-based data element set and instructions that authors or publishers can imbed when mounting documents on a network server), the Text-Encoding Initiative (TEI) Header (a SGML-based encoding scheme for complex textural structures), the Visual Resources Association (VRA) Visual Document Description Categories (for describing any entity or event that may be captured in physical form as a visual document of the original work, and including works of art, architecture, and artifacts or structures from material, popular, and folk culture), the Consortium for the Interchange of Museum Information (CIMI) metadata set (for describing digitized museum collections of physical objects, artefacts, and documents), and Digital Geospacial Metadata (DGM) are only a few examples of the many domain-specific metadata schemes which have been developed.

More recently, metadata standards work has been focussed on developing an umbrella framework which would support a variety of target document types and provide a syntactic mechanism for "translating" among the different metadata formats. This conceptual framework was first discussed at the second Dublin Core Workshop in Warwick, England, and came to be known as the "Warwick Framework" (Hakala, Husby, Koch, 1996; Dempsey and Weibel, 1996; Lagoze 1996). Refinements have resulted in the development of the Resource Description Framework (RDF), a foundation for processing metadata geared towards providing interoperability between applications that exchange machine-understandable information on the World Wide Web (Lassila, 1997; Lassila and Swick, 1998). Described as a "work in progress", the RDF has contributed to rendering "machine-readable" metadata inherently "machine-understandable".

While interoperability and flexibility of moving across metadata platforms is a laudable goal, and one which contributes to the evolution of metadata schemes and standards, the "human-understandable" piece is noticeably absent.

Those in the thesaurus construction and classification theory communities have responded to the challenges to information access posed by the vast quantities of electronic resources on the Web by reframing and refining their domain-specific metadata tools (thesauri; subject headings lists; classification systems; etc., with their emphasis on structure and content) to serve as Internet search engines which are more sophisticated than Internet robots or spiders which perform automated indexing on the Web (McIlwaine, 1998; Hudon, 1997; Van der Walt 1998). Traditional subject access systems are examples of "human-understandable" metadata schemes which serve as a target towards which developers of RDF and other interoperable metadata frameworks (for example, the XML-based Conceptual Knowledge Markup Language or CKML) may be aiming (Kent, 1998; Olson and Kent, 1997). But is there an opportunity for combining the two approaches to develop a "machine-understandable" scheme that can also be rendered "human-understandable"?

3. Research Objectives

The challenge posed by the preceding question serves as a launching point for a research program which has recently been undertaken, and frames the following three objectives. These include: (1) to determine and refine a metalevel scheme or terminological ontology which can serve as both a "metadata dictionary" (or "metadata *lingua franca*"), and a switching device for assisting end-users searching for metadata-encoded documents or document-like objects on the World Wide Web; (2) to develop a front-end, pop-up window prototype of that metalevel scheme to provide navigational assistance to searchers when required; and (3) to test whether the prototype ontological software tool enhances the information-seeking process, providing end-users with a greater depth and breadth of search options and/or improving satisfaction with search results and information discovery.

4. Methodology

With the intention of prototyping a software end-user "assist" that will translate "machine understandable" metadata into a "human understandable" terminological framework, and assist searchers in more effectively navigating through the vast array of electronic documents and document-like objects from a broad range of

informational domains that characterize the Web, the study is being conducted in three phases.

In phase one, currently underway, the research is exploring the first objective, drawing extensively from the literature to identify and analyse the structure and content of seven metadata schemes which are based on SGML/XML/HTML syntax. There are numerous metadata schemes which could be selected, but the focus is on those that cover broad but somewhat related domains. Using the extended entity-relationship (EER) data modelling framework (Olson and Kent, 1998), the entities, attributes, and relationships which form the core of the EAD, DC, VRA, CIMI, TEI Header, DGM, and GILS metadata schemes are being analysed and will subsequently be mapped. The research will attempt to identify those elements which match across all schemes, those that correspond between two systems or among three or more, and those that are clearly unique to a domain. High terminological congruence would ensure that a searcher has an open gateway to a broad range of informational domains, and may require a "switching device" to help narrow the search field. The more unique the terminology to one domain, the more targeted the search can be. The modelling process will allow for the construction of a metalevel "metadata dictionary" in which entities are defined by their attributes and further refined by the nature of their terminological relationships. A series of EER models will also result from this first phase of the research.

In the second phase of the study, the metalevel "metadata dictionary" (called metalevel because it is placed above the level of the metadata scheme) will be refined with the assistance of focus groups. Previous studies conducted at this Faculty and involving the identification and ranking of bibliographic elements - themselves examples of bibliographic metadata - suggest that five groups of not more than eight participant (n = 40) provide sufficient data to inform the research process (Luk, 1996; Stoyanova, 1998). Twenty graduate and twenty undergraduate student volunteers who have had no direct exposure to the concept of metadata (i.e., no students from the Master of Information Studies Program will participate) will be asked to review the terminology of the metalevel "metadata dictionary" and to provide feedback as to its clarity and directional potential. Based on their comments, refinements will be made and work will begin on the development of a prototype pop-up software tool which will identify each entity, describe it according to its attributes and relationship to other entities, and provide a directional link to an appropriate corresponding informational domain or domains (the "switching" function). A metalevel "metadata dictionary" and prototype of end user-assist software which will be the "front-end" to a World

Wide Web search engine will result from this phase of the study.

In the third and final phase of the study, the prototype will be tested, and evaluated, using as a framework the four generic tasks associated with any search, namely: find; identify; select; and obtain (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1997). Thirty student volunteers (15 graduate; 15 undergraduate), different from those participating in phase 2 of the research, will be randomly recruited as participants. They will be asked to perform a series of searches involving a mix of target domains, and will do so both with the prototype end-user assist and without. Student assistants will observe the searching behaviours of the participants and transaction logs will also be examined. Data will be analysed to determine what effect, if any, the use of the prototype has had on information discovery. A positive impact will be observed if the participant judges that he or she has retrieved a greater number of relevant electronic documents or document-like objects, or has been directed to an informational domain which he or she might not otherwise have accessed. Findings will be used to iteratively enhance the prototype.

5. Conclusion

Metadata, *per se*, have been within the portfolios of the bibliographic control and knowledge representation communities for over a century. But with the explosion of digitized resources accessible via the Web, and the resultant proliferation of domain-specific schemes, metadata have assumed a prominence not previously experienced. Research into the application and impact of metadata standards as they relate to electronic resources has been minimal. Recent work has focussed on aspects that can best be described as related to "machine understanding" rather than to "human understanding". The proposed research is innovative in its intention to examine possibilities for the latter, and has the potential to make a significant contribution to the scholarly literatures of knowledge representation, information seeking strategies, and information discovery. While the study will itself focus on a sample of undergraduate and graduate students, the research can subsequently be extended to other types and levels of end-users to ensure greater generalizability and prototype applicability. With the design, testing, and subsequent enhancement of a prototype to assist end-users in information seeking and knowledge discovery on the Web, the research may also be of interest to Internet product vendors. The derivation of a tool which can assist with more effectively navigating massive amounts of Web-based electronic resources can potentially benefit any beleaguered Internet searcher with an information need.

Acknowledgement

The author gratefully acknowledges the financial assistance provided by the Social Sciences and Humanities Research Council of Canada in funding this research project (SSHRC SRG # 410-99-1287).

Bibliography

Albrechtsen, H., and Jacob, E.K. 1998. "The Role of Classificatory Structures as Boundary Objects in Information Ecologies." In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998.* Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wűrzburg: Ergon Verlag, pp. 1-3.

American Library Association. ALCTS CC:DA Metadata Taskforce. 1998. *Dublin Core, Metadata, and the Cataloging Rules: Draft Report.* Available at URL: http://www.lib.virgiia.edu/ccda/about/draft1.html

Dempsey. L. and Weibel, S. 1996. "The Warwick Metadata Workshop: A Framework for the Deployment of Resource Description." *D-Lib Magazine*, July/August, 1996. Available at URL: http://www.dlib.org/dlib/july96/07/weibel.html

Fischer, D.H 1998. "From Thesauri towards Ontologies?" In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998.* Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wűrzburg: Ergon Verlag, pp. 18-30.

Gaynor, E. 1994. "Cataloging Electronic Texts: the University of Virginia Experience." *Library Resources and Technical Services* 38 (4): pp. 403-413.

Hakala, J., Husby, O., and Koch, T. 1996. *Warwick Framework and Dublin Core Set Provide a Comprehensive Infrastructure for Network Resource Description: Report from the Metadata Workshop II, Warwick, UK, April 1-3, 1996.* Available at URL: http://www.nlc-bnc.ca/ifla/document/cataloging metadata/warwick.htm

Howarth, L.C. 1998. "Metadata Structures and User Preferences: Designing User-Focussed Knowledge Access Systems." In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998.* Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wűrzburg: Ergon Verlag, pp. 360-366.

Hudon, M. 1997. "Multilingual Thesaurus Construction: Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts." *Knowledge Organization* 24 (2): 84-91.

IFLA Study Group on the Functional Requirement for Bibliographic Records. 1997. *Functional Requirements for Bibliographic Records.* Munich: K.G. Saur.

Kent, R.E. 1998. "Organizing Conceptual Knowledge Online: Metadata Interoperability and Faceted Classification." In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998.* Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wárzburg: Ergon Verlag, pp.388-395.

Lagoze, C. 1996. "The Warwick Framework: A Container Architecture for Metadata." *D-Lib Magazine*, July/August, 1996. Available at URL: http://www.dlib.org/dlib/july96/07/lagoze/07lagoze.html

Lassila, O. *1997. Introduction to RDF Metadata.* Technical report. "W3C Note 1997-11-13." Available at URL: http://www.w3.org/TR/NOTE-rdf-simple-intro-971113.html.

Lassila, O. And Swick, R.R. 1998. *Resource Description Framework (RDF) Model and Syntax Specification.* Technical report. "WC3 Working Draft 08 October, 1998." Available at URL: http://www.w3.org/TR/WD-rdf-syntax/

Levy, Pierre. 1990. *Les Technologies de l'Intelligence.* Paris: La Découverte.

Library of Congress. Network Development and MARC Standards Office. 1997. *Dublin Core/MARC/GILS Crosswalk.* Available at URL: http://lcweb.loc.giv/marc/dccros.html.

McIlwaine, I.C. 1998. "Knowledge Classifications, Bibliographic Classifications and the Internet." In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998.* Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wűrzburg: Ergon Verlag, pp.97-105.

Miller, P. 1998. "An Introduction to the Resource Description Framework (RDF). *D-Lib Magazine*, May 1998. Available at URL: http://www.dlib.org/dlib/may98/miller/05miller.html

Luk, Annie T. 1996. *Evaluating bibliographic displays from the users' point of view: a focus group study.* Master of Information Studies Research Project Report. Toronto: Faculty of Information Studies, University of Toronto.

Miller, P. 1996. "Metadata for the Masses." *Ariadne,* Issue 5 (September 1995). Available at URL: http://www/ukoln.ac.uk/ariadne/issue5/metadata-masses/

Neuss, C., and Kent, R.E. 1995. "Conceptual analysis of resource meta-information." *Computer Networks and ISDN Systems* 27: 973-984.

Olson, J.D., and Kent, R.E. 1998?. "Conceptual Knowledge Markup Language." Available as a "White Paper" at URL: http://wave.eecs.wsu.edu/WAVE/Ontologies/CKML/White Papers/CKML.html

Olson, J.D., and Kent, R.E. 1997. "Conceptual Knowledge Markup Language: an XML application." Unpublished presentation given at the XML Developers' Day, August 21, 1997, Montreal, Canada. Available at: URL: Http//wave.eecs.wsu.edu/WAVE/Ontologies/CKML/WhitePapers/CKML/CKML.html

Stoyanova, Penka. 1998. *Content of bibliographic records for serials: the users' point of view.* Master of Information Studies Research Project Report. Toronto: Faculty of Information Studies, University of Toronto.

Taylor, A.G. 1994. "The Information Universe: Will we Have Chaos or Control?" *American Libraries*, 25 (7): 629-632.

Trant, J., and Bearman, D. 1997. "The Art Museum Image Consortium: Licensing

Museum Digital Documentation for Educational Use. *Spectra*, Fall 1997. Available at URL: http://www.archimuse.com/papers/amico.spectra.9708.html

UCLA. *SWISH-E: Simple Web Indexing Software for Humans - Enhanced.* Available at URL: http//sunsite.berkeley.edu/SWISH-E

UKOLN Metadata Group. 1998. *A Review of Metadata: a Survey of Current Resource Description Formats*. Technical report. "Work package 3 of Telematics for Research Project DESIRE." Available at URL: http://www.ukoln.ac.uk/metadata/desire/overview/

Van der Walt, M. 1998. "The Structure of Classification Schemes used in Internet Search Engines." In *Structures and Relations in Knowledge Organization: Proceedings of the Fifth International Conference of the International Society of Knowledge Organization, Université Charles-de-Gaulle, Lille, France, 25-29 August, 1998*. Edited by W. Mustafa el-Hadi, J. Maiez and S.A. Pollitt. Wűrzburg: Ergon Verlag, pp. 379-387.

Veltman, K. H. 1997. "Frontiers in Conceptual Navigation. *Knowledge Organization* 24 (4): 225-245.

Weibel, S. 1995. "Metadata: the Foundations of Resource Description." *D-Lib Magazine*, July 1995. Available at: URL: http://www.cnri.reston.a.us/home/dlib/July 95/07weibel.html

Weibel, S., and Hakala, J. 1998. "DC-5: The Helsinki Metadata Workshop: A Report on the Workshop and Subsequent Developments." *D-Lib Magazine*, February 1998. Available at URL: http://www.dlib.org/dlib/february98/02/weibel.html