Designing a Metadata-Enabled Namespace for Enhancing Resource Discovery in Knowledge Bases

by

Lynne C. Howarth, PhD
Faculty of Information Studies
University of Toronto

Abstract: The proliferation of digitized resources accessible via Internet and Intranet knowledge bases, and a pressing need to develop more sophisticated tools for the identification and retrieval of electronic resources, both general purpose and domain-specific metadata schemes have assumed a particular prominence. While recent work emanating from the World Wide Web Consortium (W3C) has focused on the Resource Description Framework (RDF), and metadata maps or Acrosswalks" have been created to support the interoperability of metadata standards -- thus converting metatags from diverse domains from simply Amachine-readable@ to Amachine-understandable@ -- the next iteration, to A*human*-understandable,@ remains a challenge. This apparent gap provides a framework for three-phase research (Howarth, 2000, 1999) to develop a tool which will provide a Ahuman-understandable@front-end search assist to any XML-compliant metadata scheme. Findings from phase one, the analyses and mapping of eight metadata schemes, identify the particular challenges of designing a common Anamespace@, populated with element tags which are appropriately descriptive, yet readily understood by a lay searcher, when there is little congruence within, and a high degree of variability across, the metadata schemes under study. Implications for the subsequent design and testing of both the proposed Ametalevel ontology@ (phase two), and the prototype search assist tool (phase three) are examined.

1. Introduction and Background to the Research

The proliferation of electronic texts, images, sound, and objects accessible via the Internet, has dramatically increased the range and quantity of readily-available multimedia information. To identify and retrieve those resources, search engines, such as AltaVista, Excite, HotBot, InfoSeek, or Northern Lights, have employed web crawlers or robots to gather, and generate concept-based indexes of, World Wide Web text files. More semantically-sophisticated directories, such as Google, WWW Virtual Library or Yahoo!, have been created with the assistance of human intermediaries. Even more precise identification of, and access to, electronic resources has been provided through directory inventories of subject-specific domains, such as those of ADAM[1], or EEVL[2].

Likewise, as Intranet applications have gained ascendancy, and with the concomitant evolution of the Enterprise Information Portal (EIP), creating and fine-tuning naming conventions for content, building taxonomies to appropriately identify resident digital information, and embedding metadata tags to enhance resource discovery, have emerged as key issues to be addressed within the context of information storage and retrieval. The value of well-structured taxonomies, and of appropriately descriptive metatags, is being recognized as their application spreads, and as portal development software, such as IBM/Lotus *Raven*J, PlumTree *Corporate Portal*J, Verity *Corporate J,* or *SageMaker*J are being developed to exploit such functionality.

While metadata -- generally defined as Adata about data@ -- have gained a particular prominence because of recent Aknowledge to the desktop@ initiatives, their potential to address concerns with more precisely identifying electronic resources for enhancing access, has been acknowledged and demonstrated over a period of time. Metadata are manifested through named tags or entities, which have certain characteristics or attributes, which may have a relationship to another entity or entities, and which are assigned specific values.

Metadata can be embedded within the object itself (within the HTML header of the electronic resource), or can reside separately from the information object, linked through bidirectional pointers or hyperlinks. Relationships between metadata schemes, elements and/or records and the objects they identify and describe, as well as relationships between and among the different metadata elements, *per se*, can be documented by registering them with a metadata registry, such as the Metadata Schema Registry (http://metadata.net), or the UK Office for Library and Information Networking site (http://ukoln.bath.ac.uk/metadata/interoperablity).

The past half decade of the AMetadata Movement@, as Baker (1999) describes it, has included the development of general application metadata schemes, such as Dublin Core, GILS (Government Information Locator Service) , or DOI (Digital Object Identifier)[3], as well as domain-specific metadata schemes, such as TEI (Text Encoding Initiative), EAD (Encoded Archival Description), CIMI (Consortium for the Interchange of Museum Information), VRA (Visual Resources Association), etc. Such schemes are based on a common Amachine-readable@ syntax, such as HTML (Hypertext Markup Language), SGML (Standard Generalized Markup Language), or XML (eXtensible Markup Language). Metadata-enabled search engines can thus retrieve by precise metatags and values, those electronic resources in which a metadata record is embedded, or to which a separately housed metadata record points. As Baker (1999, 1) notes, AResearchers today agree that no single type of metadata can suit every application, every type of resource, and every community of users. Rather, the broad diversity of potential metadata needs can best be met by a multiplicity of separate, but functionally focused, metadata packages or schemas.@

The creation of distinct silos of metadata schemes would normally require that those who assign or search metatags unique to each domain would need to learn the different conventions. Recent work emanating from the World Wide Web Consortium (W3C), however, has focused on converting Amachine-readable@ syntax into a Amachine-understandable@ ontology through the design of the Resource Description Framework (RDF). The RDF derives from an earlier conceptual model, the Warwick Framework (Lagoze, 1996), which envisioned different metadata schemes, created and maintained by their respective stakeholder communities, with an overarching, or unifying metadata architecture to support interoperability among the schemes regardless of their semantic diversity. As Weibel explains (1999), the Resource Description Framework (RDF) was developed under the auspices of the World Wide Web Consortium (W3C), and adopted as a W3C recommendation in February, 1999, with a similar objective,
that is:

A... to support a broad diversity of metadata semantics within a common syntactic and structural framework. This means that utilities designed to support creation and management of metadata will be integrated into common application software: text editors, image manipulation software, and browsers, for example. Applications will be able to use metadata, and by downloading the schemas for various varieties of metadata the possibility of modular, plug-and-play metadata will come within reach@ (p.8).

Interoperability among metadata schemes can also be realized through the creation of crosswalks which map metatags or elements within one scheme to those within other systems. As Cromwell-Kessler (1998?) observes, metadata systems differ in terms of content and structure, with the latter posing the most difficulty to mapping. Each metadata system is comprised of its own elements, functioning at different levels, and designated in varying and diverse ways. A single metadata element in one scheme, for example, may be represented by two or more concepts in another system. Some metadata schemes utilize more generic elements making mapping from a domain with highly specific metatags problematic, and potentially less useful for search precision. Elements represented in one system may lack any equivalents in another metadata scheme. To further confound these basic concerns, AVariant systems are often found even within a single subject community where competing metadata systems have developed in isolation - and where, before networked access, uniformity was deemed unnecessary@ (Cromwell-Kessler 1998, 1). Though somewhat problematic as a process, creating crosswalks to facilitate the identification and access of resources across a diversity of domains is an important - albeit

inexact - first step to facilitating system interoperability.

Whether interoperability is viewed from the perspective of the Warwick Framework and its successor, the RDF, or facilitated by the creation of metadata scheme crosswalks through element mapping, the goals of seamless translation and interconnection are focused on rendering Amachine-readable@ metadata inherently Amachine-understandable@ (Lassila, 1997; Lassila and Swick, 1999). Is there, however, some opportunity for combining the syntactic interoperability supported by a common, flexible framework, such as XML, with human-generated and/or human-enhanced semantic maps (metadata scheme crosswalks) for purposes of developing a A*human*-understandable@ metadata application?

2. Research Objectives

This question provided the impetus for research (Howarth, 2001, 2000, 1999) with the following three objectives, namely, (1) to determine and refine a metalevel scheme or terminological ontology which can serve as both a Ametadata dictionary@ (or Ametadata *lingua franca*@), and a switching device for assisting end-users searching for metadata-encoded documents or document-like objects in networked knowledge bases, (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 resource discovery.

3. Methodology

Phase one research, which addresses the first objective, has drawn extensively from the literature to identify and analyse the structure and content of seven metadata schemes which are based on HTML/SGML/XML syntax. While numerous metadata schemes could have been included, those chosen cover broad, but somewhat related domains; their selection reflects an approach similar to that of Baca et al. (2000). The entities and attributes of the elements which form the core of the Encoded Archival Description (EAD), the Dublin Core (DC), the Government Information Locator Services (GILS) metadata scheme, the Text Encoding Initiative (TEI) Header, the Visual Resources Association (VRA) Visual Document Description Categories, the Consortium for the Interchange of Museum Information (CIMI) metadata set, and Digital Geospatial Metadata (DGM), were defined and analysed.

The examination of elements within each metadata scheme provided the essential framework for subsequent comparison or mapping of metatags across the standards. The creation of entity maps, or Acrosswalks@ for the study employed the methodology outlined by St. Pierre and LaPlant (1998). As they explain, AForemost in ... crosswalk development is the intellectual task of determining the semantic mapping of elements between the source and target metadata standards. The task involves specifying a mapping of each element in the source metadata standard with a semantic equivalent element in the target metadata standard@ (p. 4). Rather than attempting to map each scheme to every other, the decision was made, in accordance with St. Pierre and LaPlant (1998), to map the seven standards, in turn, to a Acontrol@ metadata vocabulary, namely, the stable, robust, and broadly comprehensive Machine-Readable Cataloging (MARC21) format.

Using MARC as the Abaseline@ structure, or Atarget@ metadata standard, all metadata elements across the seven schemes, or Asource@ metadata standards, were analysed and compared using Microsoft Access database programming. The resulting crosswalks 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. In formulating the research, it was hypothesized that high terminological congruence would imply that a searcher has an open gateway to a broad range of informational domains, and

may require a Aswitching device@ to help narrow the search field. The corollary was that, the more unique the terminology to one domain, the more targeted the search can be.

4. Findings

Analyses of the crosswalks focused on the two components of (1) degree of overlap in among metatag elements. The former considered where there was a metadata element tag corresponding to a particular MARC field. As Table 1 illustrates, the number of metatags unique
to one metadata scheme only, far exceeded instances of overlap. Nearly two-thirds of the total number of metadata elements (i.e., 201 out of 293, or 68.60%) had no corresponding metatag in any other scheme. Metatags which populated two or three schemes represented 21.84% of the
total (or 64 out of 293), while 23 of 293 (7.85%) elements overlapped four, five, or six systems. Full overlap across the seven standards occurred in five instances for 1.17% of the total.

Table 1
Degree of Overlap in Metatags Across the Schemes

Degree of Overlap in Metatags Across the Schemes	Number	Percentage
No overlap (i.e., metatags unique to one metadata scheme only)	201	68.60 %
Minimal overlap (i.e., equivalent metatags populate 2-3 metadata schemes)	64	21.84 %
Moderate overlap (i.e., equivalent metatags populate 4-6 metadata schemes)	23	7.85 %
Full overlap (i.e., equivalent metatags populate all (7) metadata schemes)	5	1.17 %
TOTAL:	293	100 %

Table 2 presents elements that were represented in five, six, and seven of the schemes, respectively. Determining degree of overlap highlighted a number of element-to-element mapping issues also identified by Cromwell-Kessler (1998?), and St. Pierre and LaPlant (1998). While a one-to-one mapping, is the ideal from the perspective of harmonizing standards, it is rare across the metadata schemes under examination within the present research. Such congruence can be observed where one scheme is based on, or largely derived from another, such as with CIMI which uses, in addition to content-specific museum metadata, elements defined by the Dublin Core. Successful one-to-one mapping also occurred with what might be described as ubiquitously occurring elements, such as Atitle@ (MARC tag 245 a).

More usual, however, were instances of one-to-many linkages. The Asource@ DC metadata scheme, for example, includes the element Asubject@. This same entity is represented in the MARC standard by several manifestations of Asubject@ as expressed through the 6XX field tag series. While each of the Asource@ metadata schemes were formally mapped to the Atarget@ MARC standard, one can observe the potential for the Amany-to-one@ problem between, for example, DC and DGM or GILS. As Cromwell-Kessler notes (1998?), mapping from a more inclusive system, such as MARC (or DGM or GILS) to a less inclusive system, such as DC (or CIMI) is less problematic than the reverse.

A third element-to-element mapping issue which emerged from the research was that of source elements that do not map to any appropriate element in the target standard. This

occurred in mapping the TEI and EAD schemes to the MARC standard. This problem of Aextra elements

Table 2
Overlap of Elements Across Five, Six, or Seven Metadata Schemes
as Mapped to the MARC Format as Baseline

MARC *	CIMI	DGM	DC	EAD	GILS	TEI	VRA
041 a - 5	language		language	langusage	languageOf Resource	language	
260 a – 5		pubplace	publisher	publicatio nstmt	placeOf Publication	publication stmt	
655 a – 5	type		type		Medium	textclass	work type
720 a – 5	contributor	origin	contributor		originator	editor	
856 u – 5	identifier	onlink	rights		linkage	availability	
260 b – 6	publisher	publish	publisher	publisher	distributor	distributor	
520 a – 6	description	abstract	description	scopecont ent	abstract		visual documen
651 a – 6		placekey	subject	geogname	placeKey- word	keywords	current site
700 a – 6	contributor	origin	publisher	sponsor		editor	creator
710 a – 6	contributor	origin	publisher	sponsor		resp	creator
245 a – 7	title	title	title	titlestmt	title	title	title
260 c – 7	date	pubdate	date	unitdate	dateOf Publication	date	date
500 a – 7	coverage	supplinfo	relation	notestmt	supplemen- talInformati	editorial decl	notes
650 a – 7	subject	tempkey	subject	controlacc ess	controlled Term	keywords	subject
653 a – 7	subject	themekey	subject	index	uncontrol- ledTerm	term	subject

* Indicates number of metadata schemes relative to MARC format baseline field tag and subfield code across which element occurs

in source@ (St. Pierre and LaPlant 1998, 6), reflects, in large part, not only the different environments which the respective source metadata schemes support (TEI and publishing; EAD and archives) in contrast with the target metadata standard (MARC and libraries), but also the diverse applications and purposes for which each metadata scheme is intended (TEI for formatting electronic texts; EAD for creating archival finding aids; MARC for communicating/exchanging bibliographic information). In order to achieve interoperability, this lack of equivalence for elements expressed in one system but not in another or others requires resolution, and may warrant addressing how, specifically, values will be added to the target metadata scheme. This is, likewise, a key issue to be addressed as the present research continues toward developing a metalevel Aswitching device@ which will, itself, require mapping of queries to appropriate elements within, between, and among different

metadata schemes.

An earlier report of findings from phase one of the research noted that, A... even where some, most, or all of schemes assign tags descriptive of the same types of elements (e.g., title; edition; date of publication; geographic subject headings; language; etc.), there is high variability in naming conventions@ (Howarth 2000, 6). While this inconsistency could prove as problematic to an end-user as having to learn the metadata elements specific to a particular domain, it does support the proposition for a Aswitching device@ to seamlessly translate and direct the query to the appropriate domain(s). Additional analyses of the mappings provide insight into the degree or extent of terminological match between and among metatags as Table 3 summarizes. Number totals and percentages represent those instances where a term or Alabel@ used for an element in one metadata scheme exactly matched than in another or others. The highest degree of terminological match occurred with metatags populating all seven schemes within the study.
It was anticipated, in contrast, that greater congruence would be evident with terms covering two metadata standards. Full equivalency between CIMI and DC elements was expected, but did not always occur. This suggests that, while CIMI is based on DC, it also contains elements unique to the museums domain; in some instances, special constraints posed by museums resources require a different label expression than that provided for in the more generic DC.

Table 3
Degree of Terminological Match Between/Among Metatags

Degree of Terminological Match Between/Among Metatags	Number/ of Total	Percentage
Match with equivalent metatags across two metadata schemes	16/48	33.33 %
Match with equivalent metatags across three metadata schemes	10/48	20.83 %
Match with equivalent metatags across four metadata schemes	20/52	38.46 %
Match with equivalent metatags across five metadata schemes	9/25	36.00 %
Match with equivalent metatags across six metadata schemes	10/42	23.81 %
Match with equivalent metatags across seven metadata schemes	16/35	45.71 %
Overall terminological match:		33.02 %

Overall, 33.02% of terms matched exactly, element label to element label. This somewhat precise criteria for matching meant that some synonymous terms were excluded from the total. For example, three metadata schemes (CIMI; DC; TEI - see Table 2, MARC tag 041 a), use the term Alanguage@, while the EAD label is Alangusage@, and GILS employs the descriptor, AlanguageOfResource@. The terminological match was calculated as 3 out of 5. However, some leeway was tolerated where the same term was represented in truncated form. In the instance where three schemes (CIMI; DC: EAD - see Table 2, MARC tag 260 b), use the term, Apublisher@, CSDGM truncates to Apublish@, while two schemes (GILS; TEI), refer to Adistributor@, the terminological match was considered as 4 out of 6. These distinctions are somewhat arbitrary, but the number of cases in which they needed to be applied were so minimal as to have little effect on the calculation of terminological match, based on an exact element to element correspondence. Arguably, different domains require the flexibility, and must retain the right, to express their inherent characteristics,

histories, and environmental and/or operational constraints through unique terminologies. From the perspective of developing a Ametadata *lingua franca@*, however, having different vocabularies or labels to describe the same element value in more than one metadata scheme poses a particular semantic challenge for resolution.

5. Conclusion

Phase one of the present research began with the hypotheses that (1) high terminological congruence across the seven metadata schemes would necessitate the design of a Aswitching device@ to assist in narrowing the search field, and (2) the more unique the terminology to one domain, the more targeted a search could be. Findings suggest that the number of metadata elements unique to one metadata scheme only, far exceeded instances of overlap. While this would perhaps facilitate targeted and precise searching, it also highlighted the problems, from the perspective of metadata interoperability, of having no equivalencies in a target scheme for values represented in the source standards. Moreover, even where there was moderate to full overlap across the schemes under study, there was high variability in naming conventions. This correspondingly low degree of terminological match could prove as problematic to an end-user as having to learn the metadata elements specific to a particular domain. The impetus for creating a metalevel ontology or Aswitching device@ to provide a common vocabulary gateway for the 67% of metadata elements which populate two or more metadata schemes, but employ diverse naming conventions was underscored.

As Baca et al. (2000) noted in the development of their multi-scheme metadata crosswalk, even when elements correspond, the conceptual and semantic models determined by the communities from which they derive, and the different purposes for which the schemes were designed, may undermine the supposed equivalencies of the terms. Mapping is an iterative process involving ongoing refinement, revision, and, even rethinking of element matches. As the present research revealed, the potential for having even as broad and comprehensive a metadata scheme as the MARC format serve as a foundation for the content, and perhaps even some of the vocabulary for, a Ametalevel ontology@ as envisioned in phase two of the research, will not be appropriate. Consequently, the project has been refocused to designing a neutral, common Anamespace@ which will be defined and enabled within the XML document standard. The latter syntactic mapping will pose a lesser challenge than that of crafting a semantic bridge (the Anamespace@) to link not only a diversity of metadata standards, but also a mélange of historical and operational environments which characterize each of the domains and purposes for which the schemes were uniquely developed. This metalevel Ametadata dictionary@ will be essential to the phase three design of a prototype Afront-end@ search software tool to assist end-users in more effectively navigating the vast amounts of information available through Internet and Intranet knowledge bases.

Acknowledgements
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), and invaluable research assistance from Julie Hannaford and Christopher Cronin, graduates students at the Faculty of Information Studies, University of Toronto, Canada.

Endnotes
1. Art, Design, Architecture & Media Information Gateway (ADAM) [http://adam.ac.uk/].
2. Edinburgh Engineering Virtual Library (EEVL) [http://www.eevl.ac.uk/].
3. Digital Object Identifier Metadata Workgroup - or DOI metadata scheme: see also [http://www.doi.org] for additional information. Accessed 11/25/99.

References
Baca, Murtha, et al. (2000). A Crosswalk of Metadata Standards. In *Introduction to Metadata: Pathways to Digital Information*. [http://www.getty.edu/gri/standard/intrometadata/crosswsd.htm]. Accessed 03/02/00.

Baker, Thomas. (1999). *Organizing the Web: an Update on the Metadata Movement.* [http://www.cs.ait.ac.th/~tbaker/kyunghee.html.gz]. Accessed 11/29/99.

Cromwell-Kessler, Willy. (1998?) Crosswalks, Metadata Mapping, and Interoperability: What Does it all Mean? In *Introduction to Metadata: Pathways to Digital Information*. [http://www.getty.edu/gri/standard/intrometadata/define.htm]. Accessed 03/02/00.

Howarth, Lynne C. (2001). Project Website for "Modelling a Metalevel Ontology". [http://www.fis.utoronto.ca/special/metadata/index.htm]. Accessed 11/16/01.

Howarth, Lynne C. (2000). Designing a >Human Understandable= Metalevel Ontology for Enhancing Resource Discovery in Knowledge Bases. In *Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International Conference of the International Society for Knowledge Organization (ISKO), University of Toronto, Toronto, Canada, 10-13 July, 2000.* Edited by N.J. Williamson, C. Beghtol, and L.C. Howarth. W_rzburg: Ergon Verlag. pp. 391-397.

Howarth, Lynne C. (1999). Modelling a >Human Understandable= Metalevel Ontology for Enhancing Information Seeking on the World Wide Web. In *Information Science: Where Has it Been, Where is it Going? Proceedings of the 27th Annual Conference of the Canadian Association for Information Science, Congress of the Social Sciences and Humanities, Université de Sherbrooke, Sherbrooke, Québec, 9-11 June, 1999.* Montréal: CAIS. pp. 115-124.

Lagoze, Carl. (1996). The Warwick Framework: A Container Architecture for Metadata. *D-Lib Magazine*, July/August, 1996. [http://www.dlib.org/dlib/july96/07/lagoze/07lagoze.html]. Accessed 09/14/99.

Lassila, Ora. (1997). *Introduction to RDF Metadata*. Technical Report. AW3C Note 1997-11-13.@ [http://www.w3.org/TR/NOTE-rdf-simple-intro-971113.html]. Accessed 09/14/99.

Lassila, Ora, and Swick, Ralph R. (1999). *Resource Description Framework (RDF) Model and Syntax Specification.* Technical Report. AW3C Recommendation 22 February, 1999.@ [http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.html]. Accessed 09/14/99.

St. Pierre, Margaret, and LaPlant, William P. (1998). Content Metadata. In *NISO Standards: Issues in Crosswalking*. [http://www.niso.org/crswalk.html]. Accessed 10/11/99.

Weibel, Stuart. (1999). The State of the Dublin Core Metadata Initiative. *D-Lib Magazine*, April 1999. [http://www.dlib.org/dlib/april99/04weibel.html]. Accessed 11/25/99.