Designing a metadata-enabled namespace for accessing resources across domains

Abstract: As Stephen DeRose noted in 1997, "Navigation requires naming, as does access whether by database, catalog, finding aid, or hypertext link. Choosing the right names for information units is perhaps the most crucial issue facing the electronic document community" (p. 301). While recent work emanating from the World Wide Web Consortium (W3C) has focused on the Resource Description Framework (RDF), and metadata maps or "crosswalks" have been created to support the interoperability of metadata standards – thus converting metatags from diverse domains from simply "machine-readable" to "machine-understandable" – the next iteration, to "human-understandable", has not been addressed to the same degree. Recognizing this semantic "naming" challenge, three-phase research was undertaken (Howarth 2001; Howarth 2000; Howarth 1999). This paper reports on the design of a metadata-enabled "namespace" based on the analyses and mapping of eight metadata schemes (i.e., EAD; Dublin Core; GILS; TEI; VRA; CIMI; CSDGM; ONIX). The methods used for, and the particular problems encountered with, developing a multi-domain source-to-target (baseline) crosswalk (St. Pierre and LaPlant, 1998), provide context for the subsequent determination of a natural language "namespace" comprised of eighteen categories. Challenges encountered in determining semantic commonalities across diverse metadata schemes, in resolving semantic ambiguities and term-to-term mapping anomalies, and in creating definitions for subsequent focus-group testing, are described.

1. INTRODUCTION

As information accessible via the Internet has continued to proliferate, and as organizations have developed Intranets for both storing and sharing valuable knowledge assets, the need for creating and fine-tuning naming conventions for content, and for building enterprise-specific taxonomies to identify and access digital information, has assumed greater priority. Paralleling this growth in accessible Web media has been an increased emphasis on designing and implementing metadata-enabled search engines capable of retrieving electronic resources in which a metadata record is embedded, or to which a separately housed metadata record points. Most recently, the World Wide Web Consortium (W3C) has directed its attention to converting "machine-readable" syntax into a "machine-understandable" ontology through the design of the Resource Description Framework (RDF). This framework is intended to support metadata "interoperability" – though largely at the machine level. Interoperability among metadata schemes can also be realized through the creation of crosswalks that map metatags or elements within one scheme to those within other systems.
Whether viewed from the perspective of the RDF, or cross-schema element mapping, the focus on "human-understandable" metadata interoperability has been less evident. This apparent gap has provided the impetus for research (Howarth 2001; Howarth 2000; Howarth, 1999) with the following three 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 in networked knowledge bases, (2) to develop a prototype search tool incorporating that metalevel scheme to provide iterative navigational assistance to end-users when required, and (3) to test whether the prototype 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.

In addressing the first objective, an extensive examination of the literature was undertaken to identify and analyse the structure and content of eight metadata schemes based on HTML/SGML/XML syntax. While numerous metadata schemes could have been included, those chosen cover broad, but somewhat related domains. 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, the Content Standard for Digital Geospatial Metadata (CSDGM), and the ONIX publishing standard were defined and analysed. With the metatags and their descriptions secured, work continued on creating cross-schema mappings as a basis for identifying similarities and differences among the selected standards.

2. THE PROCESS OF CREATING CROSSWALKS

Despite the fact that the eight chosen standards may have somewhat different purposes related to the discipline or domain in which they were initially developed, it was expected that concepts would be found that had similar meanings across schemes. For example, in the bibliographic domain, a commonly used term is "author". Within the schemes that were examined, analogous terms were found, though in some instances, named differently. Thus, "author" (expressed as field 1XX "main entry" in the target scheme MARC) was represented as "originator" (in CSDGM), "creator" (in VRA), or "creatorName" (in CIMI).

It was intended that, by identifying commonalities, a common "namespace" framework or template could be developed based on the best-understood term. After retrieving all possible ways of naming a concept, end-users consulted in a subsequent stage of the research could then be asked which term made the most sense to them when searching. Using the above example, we wanted to explore which term a user found most intuitively obvious, namely, "creator", "originator" or "author"?
To find these commonalities, and enable comparison, previously-constructed crosswalks were required. Source crosswalks that compared metadata schemes to the selected baseline or “target” – the Machine Readable Cataloging (MARC) bibliographic format – were sought. This, in fact, proved the first challenge – to find crosswalks that were both authoritative and current. It was necessary to keep in mind that metadata schemes undergo periodic revisions; therefore, mappings were needed that used current element sets for each scheme. Through our research, a relevant and useful site was identified. Maintained by Michael Day, and entitled, “Metadata: Mapping between Metadata Formats”, (http://www.ukoln.ac.uk/metadata/interoperability), the site describes and tracks crosswalking efforts on an international basis. Day’s listing was a rich source for current crosswalks.

Crosswalks were assessed for their reliability. Research needed to be done on the organization or agency that had prepared the crosswalk to ascertain the authority of the mapping. For some, this was a straightforward exercise. For the comparison of MARC to Dublin Core, for example, several crosswalks were located that were prepared by the Library of Congress (LC). The authority for these mappings was obviously well established. As well, the crosswalks were current. Originally prepared in 1999, they had been updated February 2001, only a few months prior to the start of the research.

Other schemes were not as straightforward. When seeking crosswalks comparing the Federal Geographic Data Committee’s (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) to MARC, the only crosswalk that could be found had been prepared in 1996. CSDGM, however, underwent revisions and the FGDC released a second version of the standard in 1998. Therefore, careful analysis had to be made of the differences between the two versions and how this impacted on the crosswalk. Once source material in the form of reliable crosswalks was located, the tags and other descriptive information were initially input into a Microsoft Excel document, then imported into a Microsoft Access database – a transfer that allowed for more ready manipulation and searching.

Developing crosswalks presents a number of challenges. The intent of cross-schema mappings is to facilitate interoperability between two standards. However, as Cromwell-Kessler (1998) points out, some significant issues can arise. First, in the comparison it may be found that two or more concepts in one scheme may be mapped to only one equivalent in another scheme, making it difficult to go back and forth between the two. For example, in the TEI metadata standard, tags “<publicationStmt><publisher>”, and “<publicationStmt><distributor>” are both mapped to the MARC field 260###$b, which includes in its definition both publishers and distributors. Therefore, when working from 260###$b back to the TEI terms, two separate options are available, instead of a neat one-to-one fit. The entity-relationship database modelling requirement for a one-to-one mapping is thus violated, and may have implications for the design of the prototype search tool envisioned from the current research.

Secondly, some elements in one scheme that have specific semantic meanings, may only map to fields with more general coverage in another scheme, as the second metadata
scheme may lack sufficient comparability. For example, in referring to the row corresponding with MARC field 500 (general notes) in the project crosswalk (http://www.fis.utoronto.ca/special/metadata/shortwalk.asp), one notes that terms, such as "geospatial data presentation form" (CSDGM), are mapped to it. This is a domain-specific, technical term whose meaning is completely lost when mapped to the general notes field in the target MARC format (i.e., MARC field 500). Yet the existing MARC structure seems to lack anything that could possibly match the specificity intended by CSDGM.

Lastly, for certain elements, there is no equivalence at all from one schema to another or others. In the project crosswalk, there were a number of tags, which could not be matched to a MARC tag. These "orphans" were important to identify, and required careful scrutiny when categories of common terminology were being devised. Their presence may also impact the design of the prototype search tool as the research progresses.

The crosswalk exercise served to emphasize the differences in detail between schemes. Dublin Core has fifteen elements and accompanying qualifiers. TEI, in contrast, has a 39-chapter manual to explain its multitude of tags. CSDGM, with well over three hundred individual metadata tags, represents yet another level of precision and complexity in both the domain and its vocabulary. These gaps in comparable specificity of tags are problematic from both a data modelling and a semantic perspective, illustrating, yet again, the challenges inherent to identifying commonality where "naming" is applied in different contexts and for diverse purposes. This suggests that syntactic interoperability is more readily achieved/achievable than semantic linking.

As a process, mapping is very detail oriented. Because it can be prone to error, great care must be taken not to miss a match, or to erroneously map one schema value to another – in this case a source metadata scheme element to the wrong field tag in the target MARC format. The process is also exacerbated by the multitude of formats of crosswalks. Some versions were prepared in chart form, sometimes lacking sufficient detail, and, consequently, requiring careful interpretation. For example, TEI has a hierarchy with repeating elements. Yet, sometimes the higher level of the hierarchy was not placed into the crosswalk, requiring that TEI and MARC definitions be consulted in order to determine which parts of the hierarchy had not been expressed in the source (TEI) crosswalk. Other crosswalks were presented in paragraph format, and contained a great amount of detail. This detail aided in the clarification, but meant careful transcription of these notes as the formatting of the MARC record could be completely changed if the additional information was lost.

Errors were also possible by the those who had originally created the source crosswalk. For example, one EAD source crosswalk mapped the tag, "<organization>" to MARC field 351 – "organization and arrangement of materials". Another mapped "<organization>" to MARC 110$a – "main entry – corporate name". The term "organization", then, has been understood in two very different ways, depending on which EAD to MARC crosswalk one references. The meaning of EAD tag
"<organization>" is: "how the described materials have been subdivided into smaller units ... to identify the logical or physical groupings of the described materials within a hierarchical structure" (http://lcweb.loc.gov/ead/tglib/tlin101.html). It therefore becomes clear that MARC field 351 is the more appropriate tag in this case. However, this EAD to MARC mapping is a good example of how easy it is to make an inaccurate or misleading match related to a misinterpretation of meaning.

Another issue that required resolution was the discovery that, while some of the selected crosswalks mapped the schemes to USMARC, others mapped to MARC21. Research, then, had to be done into the differences between these two MARC formats to ensure that there was no 'muddying of the waters' with multiple versions of MARC. In-depth investigation unearthed few differences that would negatively impact the ongoing compilation of a master crosswalk.

Compiling an eight source to one target crosswalk – the “master crosswalk” – clearly underscored the importance of finding user-friendly naming conventions. Seemingly straightforward tags with element names, such as “publisher” (DC; CSDGM), or “<PublisherName>” (ONIX), or “editionStmt” (TEI), were located. In contrast, other less semantically transparent tags, such as “<fsdDecl>” were also identified. The TEI tag, <fsdDecl>, actually refers to the ‘feature system declaration’ and relates to methods and editorial principles that govern the transcription or encoding of the text. The name of the term, however, provides no sense of this at all. Often, then, it was necessary to go back to the documentation for the scheme to assess the semantic meaning of the element.

Definitions that included the tag name, per se, in the definition of the element sometimes confounded understanding. For example, the CSDGM tag, “Modified Stereographic for Alaska”, has the following definition: “contains parameters for the Modified Stereographic for Alaska projection”. The definition repeats the tag title in the content and, therefore, does not provide enough clarification to be useful.

The approach to mapping varied across sources. It seemed that those who developed the initial schema-to-MARC crosswalk aimed to capture all fits with MARC, while others sought the best fit. Therefore, some schemes had tags that were cross mapped to several MARC tags, while other scheme tags were mapped to only one MARC tag.

Despite the many issues that arose during the process, it was possible to create a master crosswalk which had been validated through the consultation of more than one source schema-to-MARC mapping for each metadata scheme selected for the project. With this foundation, we were then in a position to evaluate similarities across the master crosswalk. All MARC tags were extracted that had at least two elements from other schemes mapped to it. This created a “like terms” crosswalk. Where slight ambiguity arose, a “common sense” aggregating of MARC tags was done. For example, some schemes mapped to 500, while others mapped to a greater level of granularity (e.g., MARC field 500$a). These were categorized together. It was also found that some schemes mapped to MARC in a very general way. For example, the CIMI tag, “creatorGeneral”, is mapped to MARC field 1XX – in other words, to all of the 100 level
MARC tags (i.e., personal name [MARC 100]; corporate name [MARC 110]; conference name [MARC 111]; uniform title main entry [MARC 130]. This required additional interpretation as to whether one or more of the target MARC field tags was applicable or appropriate. A separate “scheme-specific” crosswalk was created, to accommodate those instances where an element of one scheme only mapped to a single MARC tag. Thus, the outcomes derived from the master compilation of eight source metadata schemes to the target MARC “baseline”, were (1) a “like terms” crosswalk, and (2) a separate “scheme-specific” crosswalk (http://www.fis.utoronto.ca/special/metadata/shortwalk.asp). These three cross-domain listings provided the foundation for isolating, evaluating, and subsequently defining “common terminology” categories as candidates for the project’s intended natural language “namespace” – as the following section will describe.

3. CREATING COMMON “NAMESPACES”

With the crosswalk of “like terms” thus developed, the project’s next task was to translate that crosswalk from its format in Microsoft Access into a more accessible “metadata dictionary”, using natural language “namespaces”, or categories, to communicate the semantic commonalities between or among the tags of different metadata schemes. In other words, rather than relating metadata tags to a baseline of MARC, as was done with both the original source and compiled master crosswalks, the objective became to create categories that searchers would be able to understand without a specialized knowledge of either MARC or any of the eight metadata schemes, per se. The result of this analysis was seventeen categories (see Table 1) into which it was thought the majority of the tags would appropriately fit. An eighteenth category, Additional Information, was established to accommodate those tags that did not fit semantically into any of the other categories. Categories were also given definitions to describe the tags contained therein. This process was an iterative one, as category definitions—and even the names themselves—were repeatedly refined during the course of the categorization process.

One example of how initial concepts of these categories were modified was in the review of the Sources, References & Related Works category. The first definition of this category addressed only those tags that provide either bibliographic citations or citations to people or places from which the work was obtained. However, categorization showed that some standards also have elements providing references to other objects. CIMI, for instance, has a tag, “relatedObject”, which is defined as: “other works connected to the object as part of a collection.” An example of a “relatedObject”, then, might be a panel that is part of an altarpiece. Dublin Core has the tag, “Relation–IsPartOf”, which describes a resource that, “is a physical or logical part of the referenced resource.” Based on such instances, the initial definition of Sources, References & Related Works was revised to reflect that some references are not merely bibliographic citations, or references solely between works and their owners, but also between the object being described, and other objects to which it is directly related.
<table>
<thead>
<tr>
<th>&quot;Namespace&quot; or Common Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Information</td>
<td>how a user may get in touch with someone connected to the work</td>
</tr>
<tr>
<td>Date &amp; Time Period</td>
<td>a particular day, month, or year associated with a work; may also describe the time period or historical coverage of the work's content</td>
</tr>
<tr>
<td>Edition</td>
<td>information on the version of the particular work</td>
</tr>
<tr>
<td>Genre Type</td>
<td>the nature or style of the intellectual content of the work</td>
</tr>
<tr>
<td>Identifiers</td>
<td>unique names or numbers associated with the work for the purposes of records management</td>
</tr>
<tr>
<td>Language</td>
<td>language, sublanguages, dialects, etc., of the intellectual content of the work</td>
</tr>
<tr>
<td>Methodology</td>
<td>procedures, methods and/or techniques employed in the construction or alteration of the work</td>
</tr>
<tr>
<td>Names</td>
<td>the names assigned to an individual or organization associated with the work</td>
</tr>
<tr>
<td>Physical Format</td>
<td>the physical appearance or construction of the work</td>
</tr>
<tr>
<td>Place</td>
<td>geographical locations associated with the work</td>
</tr>
<tr>
<td>Rights &amp; Restrictions On Use</td>
<td>limitations and legal rules that affect how the work (or in CSDGM, the metadata describing the work) is to be used after access has been granted</td>
</tr>
<tr>
<td>Roles</td>
<td>the function of an individual or organization associated with the work</td>
</tr>
<tr>
<td>Sources, References &amp; Related Works</td>
<td>citations to persons or entities from which the work (or information contained within the work) was obtained; may also refer to other works or objects to which the work at hand is in some way related</td>
</tr>
<tr>
<td>Subject</td>
<td>terms, keywords, or phrases that describe, identify, or interpret the intellectual content of the work and what it depicts or expresses</td>
</tr>
<tr>
<td>Summary &amp; Description</td>
<td>a short account or commentary giving the main points about the work</td>
</tr>
<tr>
<td>Terms Of Access &amp; Availability</td>
<td>restrictions and legal prerequisites that affect one's ability to access the work (or in CSDGM, the metadata describing the work)</td>
</tr>
<tr>
<td>Title</td>
<td>the identifying name or phrase given to the work (or codeset in some cases in CSDGM); it may or may not contain a subtitle</td>
</tr>
<tr>
<td>Additional Information</td>
<td>miscellaneous information about the work that does not easily fit into one of the other categories</td>
</tr>
</tbody>
</table>

Table 1. "Namespace" Categories: Common Terms and their Definitions
Another challenge associated with tag categorization was that several showed characteristics that would place them in more than one of the designated categories. For example, across the eight schemes, there were a total of three tags relating to the concept of provenance, generally used to relate the history of ownership of an object. Because information encoded in these tags often express the chronology of ownership of an object, it was felt that they should be listed under the category of Date & Time Period. Because all three tags provide the names of people and/or organizations associated with that ownership, it was anticipated that users might also search for such tags under the category Names. Finally, because all three tags describe the history of sources of the object or work, it was determined that still other users might require the tags under the category Sources, References & Related Works. In order to accommodate all possible uses of such tags, they were placed in all relevant categories and cross-referenced accordingly. While it is expected that the search interface deriving from the research will be able to support having one tag mapped to multiple categories, it will be important to test whether users, themselves, find such cross-listings a hindrance, or useful to, their searching.

It was noted, previously, that only those tags in the “like terms” crosswalk were used to populate the categories. However, when reviewing the crosswalk of “scheme-specific” tags, commonalities were identified between “scheme-specific” terms, the “like terms”, and indeed the categories themselves. One pertinent example is the qualified Dublin Core element, “Date-Created”, which was mapped by the Library of Congress to MARC field 260$a. Because no other schemes had a tag mapped to this MARC field, it was not included in the “like terms” crosswalk, and therefore not initially included in our categories. However, one cannot deny the importance of creation date information for the purposes of description and searching. As such, we endeavoured to determine how many such instances there were of “scheme-specific” tags that, when removed from the context of their mapping to MARC, did indeed have something semantically in common with our categories and the tags therein. In fact, the number of tags ultimately translated from the “scheme-specific” crosswalk was surprisingly high (i.e., n= 218/348 or 62.64%).

This unexpectedly high number of inclusions of what might be considered semantically unique terms in a process devoted to formulating categories based on “common terminology”, speaks to the particular subjectivity of developing crosswalks. Identifying matches and creating semantic links is open to human interpretation – and, likewise, misinterpretation. Cross-domain mappings reflect the different ways people and organizations “translate” metatag definitions and values. Moreover, for those who develop the initial metadata schemes within a particular field or application, what one may mean to say in the naming convention, may not necessarily emerge intact from the process of interpretation. “What’s in a name?” – while important to address in a proscribed environment – assumes even greater relevance when extending beyond that domain, and across several applications.
4. SUMMARY

Throughout the processes of compiling a master crosswalk, deriving separate mappings of "like" and "scheme-specific" terms, and establishing categories of "common terminologies", we have been especially aware of the values and pitfalls of naming, as described by DeRose (1997): "Navigation requires naming, as does access whether by database, catalog, finding aid, or hypertext link. Choosing the right names for information units is perhaps the most crucial issue facing the electronic document community" (p. 301).

The next steps in the research, therefore, will be to conduct focus groups to review and validate the categories and definitions (see Table 1, above) for their usefulness. The ultimate goal of the project is to develop, through an iterative process of testing and enhancing, a prototype computer-based system that aids in Internet/Intranet searching. Focus groups will first help to determine whether the category names and definitions are useful, and then will evaluate how well those categories work within the search tool itself. Do the tags we have placed in each category belong there, or are they better listed in other categories? Is it useful having some tags cross-listed in multiple categories? Above all, do the categories and the iterative search-assist prompts together aid in navigating diverse metadata schemes and cross-disciplinary domains where the precise language of each may be unknown or unfamiliar to the end-user?

Simplifying access to repositories of metadata-enabled electronic resources through a search interface powered by natural language query boxes can open a world of separate, yet potentially related knowledge bases to novice or expert information seekers. While this goal remains to be tested through the remaining two phases of the research project, it continues to show promise and intrinsic good value as it moves towards realization.

ACKNOWLEDGEMENTS

The principal investigator (Howarth) 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). The co-authors of this paper extend particular thanks to Christine Dujmovich who was instrumental in the design of the original cross-domain mappings.

REFERENCES


