Lynne C. Howarth and Thea Miller
Faculty of Information Studies, University of Toronto
Toronto, Ontario, Canada

Designing a Language-Independent Search Prototype for Accessing Multilingual Resources from Metadata-Enabled Repositories

Abstract: As research described herein suggests, designing a cross-language information retrieval (CLIR) prototype that supports natural language queries in any language, and presents search results in visual category clusters, represents another step towards providing equitable access to the world community by anyone with an Internet connection and an information need.

Résumé : Comme la présente recherche le suggère, la conception d'un prototype de recherche d'information multilingue (RIML) qui permet d'exploiter les requêtes en langage naturel dans n'importe langage, et présente les résultats de recherche sur les grappes de catégories visuelles. Ceci constitue une autre étape pour offrir l'accès équitable à la communauté internationale pour tous ceux qui possèdent une connexion Internet et un besoin informationnel.

1. Introduction and Background to the Research

Since 1986, when the Standard Generalized Mark-up Language (SGML) became an international standard (ISO 8879:1986), there has been steady activity to develop and refine SGML/XML/HTML-based metadata standards for specialised information domains. At the same time, so-called "legacy" metadata schemes, such as Machine-Readable Cataloguing (MARC) (used for the electronic exchange of bibliographic records among libraries), have been mapped to XML to ensure a common syntactic mark-up (data interoperability) standard for digital resources regardless of the metadata scheme employed for original content description. As content metadata schemas, such as Dublin Core, TEI, EAD, GILS, DGMS, etc., have been used more widely, and particularly for projects in non-English language countries, and/or for multilingual resources, the need to develop multilingual versions of the metadata standards, *per se* has been recognized. The Dublin Core (DC) community, for example, has responded vigorously, establishing the Dublin Core Metadata Initiative (DCMI) Localization and Internationalization Special Interest Group for the adoption of Dublin Core to local resources in a local language. Language interoperability is ensured by linking language equivalents for each of the eighteen DC elements to a "universal token"[1]. This (usually) English language (or English-like) token is then accessed by Web crawlers, and the retrieved item returned in response to the original metatag language equivalent. (Baker 1997).

On a broader, international scale, multi-language studies sponsored by the Defense Advanced Research Projects Agency's Translingual Information Detection, Extraction and Summarization program (i.e., DARPA TIDES) continue to develop methods for accelerating cross language automatic translation. Narrowing the contextual focus further to the bibliographic community, Larson, Gey and Chen (2002), with the assistance of Michael Buckland, have been investigating the potential for, "exploiting online library catalogues for multilingual tasks" (p. 186). With their characteristic structured records and controlled vocabularies, OPACs and WebPACs (web-enabled public access catalogues) allow for the harvesting of multiple resources that share similar content in multiple languages. A search for the topic, "Global warming", for example, would yield records for items in numerous languages in library catalogues around the world, assuming that each had applied the same controlled subject term from a standardized thesaurus or subject headings list.

Apart from the preceding use of structured metadata in library catalogues, research into the application and impact of metadata standards as they relate to multilingual electronic resources has been minimal. Recent work has focussed primarily on aspects related to natural language processing and automatic translation[2]. For example, Loukachevitch and Dobrov (2002) explore the creation of monolongual, bilingual, and multilingual thesauri specially designed to be used in automatic processing of large collections of text for information retrieval. A related area of research has been in the area of Cross-Language Information Retrieval (CLIR). As Youssef (2002, 1) notes: "CLIR has many useful applications. For example, multilingual searchers might want to issue a single query to a multilingual collection, or searchers with a limited active vocabulary, but good reading comprehension, in a second language might prefer to issue queries in their most fluent language." Yet, even recognizing that cross-language interoperability is key to universal usability of web-enabled repositories, the semantic aspects, and particularly those that take into account end-user cognitive approaches to language usage and interpretation, remain largely in the experimental phase (Peters and Braschler 2001).

2. Rationale and Objectives for the Research

While the design and implementation of multilingual metadata schemas has proceeded apace, and while work on supporting cross-language information retrieval continues, there remains a paucity of research that deliberately links the two. Metadata-enabled repositories of multilingual digital objects exist largely as distinct silos where the inherent power of metatagging remains under-utilized, and minimally exploited. Recognizing this gap, work was undertaken to extend categorization and modelling techniques used in a *monolingual* English language metadata environment – the foundation on which earlier research focused (Howarth 2004; Howarth and Hannaford 2003; Howarth, Cronin, and Hannaford, 2002) – to the *multilingual* metadata environment where little empirical study has been done. Specifically, the study had as its two objectives:

- to assess how the seventeen-element categorization model (see Table 1) developed previously might be adapted to support language interoperability at the semantic level where end-users are searching metadata-enabled repositories; and
- to design and test a search software tool for retrieving electronic resources in metadata-enabled repositories based on queries posed in languages other than English (the original testbed).

While initial activity centred around the first objective, it became clear very quickly that further refinements to the seventeen-element categorization model, per se, would be required prior to attempting to reconceptualize and reposition it within a multilingual environment. In other words, if there were any ambiguity still remaining with any of the core categories, these would be compounded presumably as we attempted to translate them into other languages. We determined that, only through having end-users evaluate the appropriateness of search results relative to each of the common categories, would we be able to assess the degree of cognitive consonance associated with each element. We were interested in how closely user's own understanding of the term corresponded with the meaning assigned by the research team. This required, then, that the search tool be developed first.

While recognizing that clustering search results relative to their respective English language category (as derived from synthesizing corresponding tags within a set of nine metadata schemas [Howarth, Cronin, and Hannaford, 2002)]) would not be appropriate within a multilingual context, we did want to simulate at least some aspect of language-independent searching. This would ensure that at least a part of the research on language interoperability would be advanced in the shorter term. Consequently, we decided to develop a proof-of-concept search tool that would allow for entry of a query in any language, although results would cluster according to the English language category label. We determined further that, once the prototype was ready, it could be used to test the semantic transparency of the seventeen core categories. Once ambiguities with those had been addressed, we could move to re-conceptualizing (and translating) the categories for testing in language settings other than English. With that sequence of processes having been determined, work on prototype development began. It continued across several stages as the next section will describe in some detail.

Element Label	Definition
Contact Information	Information on how to communicate with someone about a work, i.e., names, phone numbers, etc.
Rights/Restrictions on Use	Legal limitations/rules that affect how you can *use* a work *after* you have been given access to it
Edition	Information on a work's version
Roles	The function of an individual or organization associated with a work
Summary & Description	Details about a work that illustrate its main points
Identifiers	Unique names or numbers assigned to a work so that it can be distinguished from others, for example, its ISBN
Sources, References & Related Works	Other works that are related to the work you are seeking or were used to develop the work you are looking for
Language	The language or dialect of a work
Physical Format	The physical appearance of a work
Subject	The topic of a work; its intellectual content
Date & Time Period	Dates associated with a work, as well as time period information regarding a work's content can be obtained through this category
Terms of Access & Availability	The legal limitations/rules that affect your ability to access a work. This relates to privacy or intellectual property concerns
Methodology	The procedures/techniques used to make or change a work
Genre Type	The nature or style of a work's intellectual content
Names	Names of individuals or organizations associated with a work, such as creators, publishers, sponsors, etc.
Title	The name or phrase assigned to a work for identification purposes
Place	Locations associated with a work, for example, where a work was created, published, is housed, etc.

Table 1: Element Labels and Definitions: The 17 Common Categories Model

3. Design of the Research Prototype (Proof-of-Concept)

a. The prototype

In designing the prototype, we were guided by four factors:

- It should be built using open-source software, that is, software the source of which is publicly accessible.
- It should be lightweight in terms of resource consumption and physical portability.
- It should be flexible both in terms of user functions and application areas.
- The results should be able to be displayed in a graphical format.

Using open-source software has significant advantages beyond cost-savings. By being able to access the actual application programming code, the knowledgeable software designer is able to implement changes (i.e., to customise) for optimal functionality. This ensures that the application can be very closely tailored to suit the specific needs of the research project. It was desirable to develop a solution that was relatively independent of the environment, both in terms of technical equipment and expertise. This was a consideration especially given that, by its nature, the prototype is a web application, and hence obviously dependent on a server. The greater the relative freedom from such environmental considerations, the greater the ability to design according to the actual needs of the project. In addition, since it was our intention to test the prototype in external environments, it was highly desirable that it be easily physically transported.

Producing a basic model that would enable the various components to be altered relatively easily was regarded as essential in adapting the prototype to user needs, as the perception of these needs changes. Such alteration can be anticipated, both in regard to the design of data collection methods, as well as in response to the results of data collection. In addition, it was desirable to arrive at a solution that was useful to both the research and professional communities.

Finally, because the seventeen core categories (see Table 1) lend themselves especially well to graphical presentation (for example, as nodes), it was important to provide the ability to explore visualisation factors. In particular, two standards were identified as highly relevant, namely, the Resource Description Framework (RDF), and the XTM topic map standard. Using RDF, resources are identified by their link to a resource provider or repository, whereas in using XTM, items are identified within the context of the pool of resources represented by the index (as created by the search engine). The graphical representation of both of these types of XML documents can easily be effected through the Scalable Vector Graphics standard (SVG).

Two aspects we faced, but did not address in this design approach, were the issues of scalability, and the overall development potential of the prototype. In its anticipated area of use, the number of resources an application of this type could access is both very large, and increasing; thorough testing would thus require the ability to process the results for more than 10,000 individual documents at a time. Examples here include applications which run only on one operating system, or which force the user or developer to rely on one software application or programming API. These aspects were not felt to be relevant at the current, proof-of-concept phase of development.

b. The test repository

For testing the prototype, a repository was created using the 107 documents of the British Women Romantic Poets project of the Library of the University of California at Davis. All of these documents were marked up according to the TEI (Text Encoding Initiative) standard, using SGML, and use the metadata associated with the TEIHEADER element. Some additional processing of these documents was required. First, in order to speed up indexing and searching, the body of the documents was removed, since only the metadata were required for the proof-of-concept. Second, because the prototype requires use of XML, the SGML mark-up had to be converted to XML.

c. Indexing and searching

For indexing the repository, Swish-e was used, since it permits considerable flexibility in designing the index. In this particular case, the TEI metadata were simply mapped over aliases to the appropriate come core categories (see Table 1, above). The simplicity of this mapping effectively ensures that the configuration of the indexing can provide for a significant number of differing metadata schemes.

Searching the index is facilitated by a conventional web search interface (an HTML form), in which the user is presented with a simple box for the query expression, and a button for launching the query. At this stage of development, only very simple search expressions are allowed. Clicking the button does not directly launch the query, but instead starts a Perl-CGI script, which, in turn, calls the Swish-e search program as one of its processes. Once the query has been executed, the script then continues by parsing the results, separating the individual metadata content, and packaging it in appropriate XML tags. These tags are, in turn, framed by an appropriate XML header and root element tags, forming a complete XML metadata document. This is then written to an external file. Once the XML document is created, the script continues by transforming it to a SVG document. This is a rather simple process, carried out by a native Perl XSLT parser. Again, the resulting document is written to an external file, available for later use.

d. The display

Finally, the script creates an HTML document frame, which in turn references the newly created SVG file (see Figure 1). The user is then presented with a screen with a notice that the search has been successful, and instructions to click on a link to see the results. Clicking this link will then produce the script-created HTML page, with the embedded SVG document, displaying the results by core element category (see Table 1) in a graphical format. These results, represented by nodes (see Figure 2), can, in turn, be clicked, leading the user (over hyperlinks) to the actual source documents. These documents can then be viewed as intended by the resource provider. The end-user also has the option to view the results in the standard (textual) list form, familiar from common Internet search services.

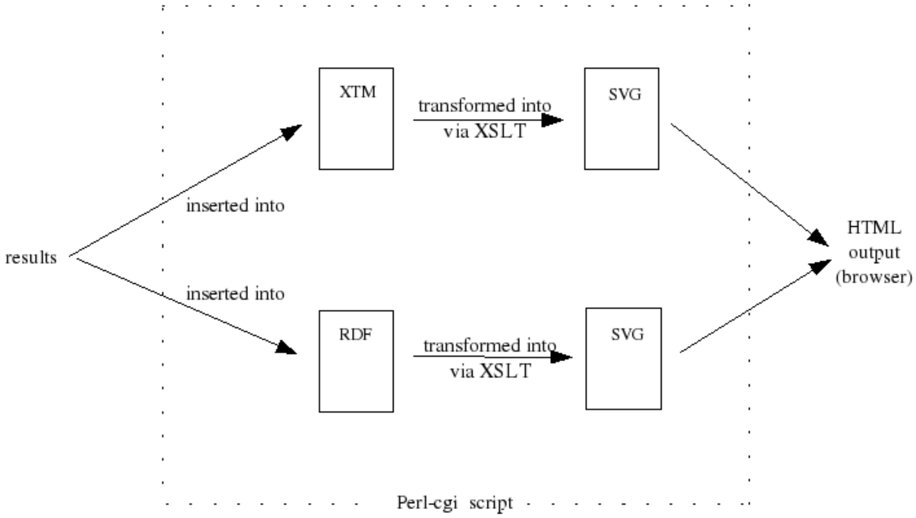

Figure 1: Processing the results (overview)

Within the basic processing model, the methods used to create output in RDF and XTM differ in detail. Given the constraints of producing an initial proof-of-concept, no effort was made in optimising either the graphical presentation itself, or the implementation of the RDF and XTM standards. In particular, with regard to the creation of RDF documents from the search results, significant issues remain, including the handling of namespaces, and the consequent nesting of various resource providers within a single parent resource description element. The topic map presentation, on the other hand, appears to be less problematic, as the seventeen core element categories (see Table 1) are conceptually strongly related to topics.

The visual representation of query results, grouped together according to commonly recognized concepts[3], such as "Names", "Title", "Place", "Sources", "Roles", etc.[4] (see Figure 2), provides the searcher with a literal picture of the number of objects that populate each category, as derived from the total of all repositories selected by the user. While the language of the query and of the objects retrieved is of relevance to the searcher, it is not of any consequence to the CLIR prototype or the structure (tagging) of the metadata records in specified target repositories. Nonetheless, should the common categories be confirmed – as a result of further user testing – as useful conceptual "buckets" or collocating devices, and particularly as they are visually represented through topic maps, then the language of each category label and definition will become important.

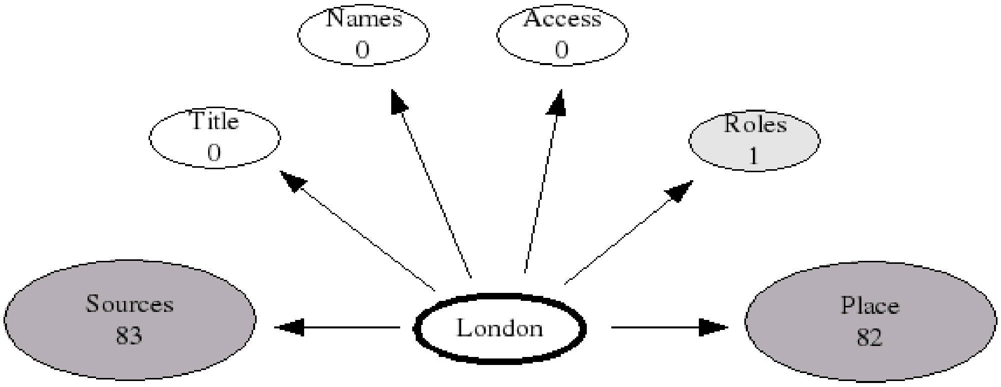

Figure 2: Results displayed as a topic map (XTM)

4. Next Steps: Testing the Proof of Concept

The ultimate intent of the research is to support the following *matrix of search scenarios*, with results being grouped within commonly understood core categories, and displayed in scalable topic maps:

- English language query retrieves monolingual English results
- English language query retrieves multilingual results
- Other language query retrieves monolingual results by language of query
- Other language query retrieves multilingual results
- Multiple languages query retrieves results in languages defined by query
- Multiple languages query retrieves multilingual results

With the proof-of-concept having been developed, the next stage of research is to expand the test repository to include those collections with resources encoded with a metadata schema or schemas additional to TEI. The inclusion of Dublin Core-enabled digital objects is a priority because of the growing number of DC implementations. In fact, any of the original nine schemas used in creating the master crosswalk and subsequent seventeen-category model (Howarth, Cronin, and Hannaford, 2002) would be an appropriate candidate for testing the search prototype, particularly to the extent that schema has been rendered XML compliant.

With sample repositories representing a broader base of metadata-enabled resources – including those in languages other than English – the search prototype will undergo end-user testing. Searchers will be asked (1) to assess results relative to their inclusion in a particular category cluster, and (2) to evaluate the relative utility and interpretability of corresponding topic maps. Depending on outcomes from user testing activities, research activities will focus on how best to transform the English language core categories to other languages. This will entail more than word-to-word translation, given that the categories are themselves derived with reference to, and best understood and interpreted within, some kind of context. Indeed as focus groups from an earlier stage of the research had emphasized, "… when you're actually using it [an element label and/or its definition], the context always does help" (Howarth and Hannaford, 2003 [conference presentation slide]). Carrying linguistic and cultural assumptions from one language milieu to another, is not only misleading, but also profoundly misguided in the absence of context. It is difficult fine tuning meaning around a monolingual concept; crossing languages will necessarily pose a set of challenges that may not be resolved readily or even at all as the present research continues.

We are mindful that digital objects are being represented by and captured in structured metadata (e.g., descriptive, administrative, preservation, technical, etc.), that are, themselves, a representation of consensual interpretation from experts in a domain. Meaning, like communication, becomes less clear the further one moves from the source to the interpretation or representation of the original. Thus, the digital object is described in the language of the domain within which it is associated. What essential meaning is changed or lost when that description is reinterpreted or "translated" into another language or languages? Preserving meaning within *context* offers a temporary stopgap, though not a solution, particularly if context itself can be considered fluid. Prototype testing should assist us in determining the extent to which the presentation of visual, scalable nodes situated relative to each other in a physical sense may add some kind of "other level" meaning – though limited, we acknowledge, by the text of the labels currently associated with each node. Can we exploit some kind of visual context to minimize the distortion or loss of meaning that translation may invoke? Further activities within the present research will focus on those key issues.

5. Implications of the Research and Conclusions

The research as described in this paper, is innovative in its intention to examine possibilities for creating a cross-language metadata framework encompassing several information domains. Unlike studies which focus on natural language processing or machine translation applications for accessing indexed Internet resources or full text, this research targets the metadata that are used to mark-up digital content in knowledge repositories, and focuses on exposing them through the lens of readily understood core categories. This work also aims at providing a transparent, language neutral query interface (or multilingual "gateway"), and the presentation of search results in visual category clusters (i.e., topic maps) for an end-user who may know little about the metadata *per se*. Building on previous work, the research has the potential to make a significant contribution to the scholarly literatures of knowledge representation, information seeking strategies, and information discovery within multi- or trans-lingual contexts. With the testing, and subsequent enhancement of a cross-language search tool to assist end-users in knowledge discovery on the Web, the research may also be of interest to both Internet content providers, such as those responsible for government online initiatives, and product vendors. The derivation of a tool that can assist with more effectively navigating massive amounts of multilingual, Web-based resources can potentially benefit any overwhelmed searcher with an information need.

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 # 410-03-1413). The co-authors of this paper extend particular thanks to previous members of the research team: Christopher Cronin, Christine Dumovich, Julie Hannaford, and Annie Ng.

Endnotes

[1] *See* language equivalency crosswalks for 30 languages currently available through the DCMI site at: http://dublincore.org/groups/languages/ Accessed 13 April, 2005.

[2] *See*, in particular, the Cross Language Evaluation Forum (CLEF), an activity of the DELOS Network of Excellence for Digital Libraries. http://clef.isti.cnr.it/ One of the objectives of CLEF is, "to provide an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts." Accessed 13 April, 2005.

[3] The seventeen common categories were also focus-group tested for semantic transparency, and modified as required (Howarth, Cronin, Hannaford, 2002; Howarth and Hannaford, 2003).

[4] The common category labels were shortened to fit within the nodes, while also preserving their essential meaning.

References

Baker, T. 1997. "Metadata Semantics Shared across Languages: Dublin Core in Languages other than English." *Dublin Core Metadata Initiative (DCMI).* Available at: http://dublincore.org/groups/languages/mr-19970303.shtml Accessed 13 April, 2005.

Howarth, L.C. 2004. "Modelling a Natural Language Gateway to Metadata-enabled Resources." In *Knowledge Organisation and the Global Information Society: Proceedings of the Eighth International Conference of the International Society of Knowledge Organization, University College London, London, England, UK, 13-16 July, 2004.* Edited by I. McIlwaine and K. Lottman. Würzburg: Ergon Verlag. Pp. 61-66

Howarth, L.C., and Hannaford, J. 2003. "Deriving a Multilingual Gateway to Cultural Repositories." In *Bridging the Digital Divide: Equalizing Access to Information and Communication Technologies: Proceedings of the 31st Conference of the Canadian Association for Information Science/Association canadienne des sciences de l'information – held with the Congress for the Social Sciences and Humanities of Canada, Dalhousie University, Halifax, Nova Scotia, 30 May-1 June, 2003.* Edited by L.Spiteri and W. Peekhaus. Toronto: CAIS. Pp. 196-209.

Howarth, L.C., Cronin, C., and Hannaford, J. 2002. “Designing a Metadata-Enabled Namespace for Accessing Resources Across Domains.” In *Advancing Knowledge: Expanding Horizons for Information Science: Proceedings of the 30th Annual Conference of the Canadian Association for Information Science/Association canadienne des sciences de l’information – held with the Congress for the Social Sciences and Humanities of Canada, University of Toronto, Toronto, Canada, 30 May-1 June, 2002.* Edited by L.C. Howarth, A. Slawek, and C. Arsenault. Pp. 223-232.

Larson, R.R., Gey, F., and Chen, A. 2002. “Harvesting Translingual Vocabulary Mappings for Multilingual Digital Libraries.” In JCDL ’02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, Oregon, 13-17 July, 2002. Pp. 185-190. Available at: http://delivery.acm.org/10.1145/550000/544259/p185-larson.pdf?key1=544259&key2=7992163111&coll=GUIDE&dl=GUIDE&CFID=42110664&CFTOKEN=91403834 Accessed 13 April, 2005.

Loukachevitch, N, and Dobrov, B. 2002. “Cross-Language Information Retrieval Based on Multilingual Thesauri. In *Cross-Language Information Retrieval: A Research Roadmap Workshop at SIGIR-2002, Tampere Finland August 15, 2002* Available at: http://ucdata.berkeley.edu:7101/sigir-2002/sigir2002CLIR-06-loukachevitch.pdf Accessed 13 April, 2005.

Peters, C., and Braschler, M. 2001. "Cross-language system evaluation: The CLEF campaign", *Journal of the American Society for Information Science and Technology*, 52/12 (2001), 1067-1072.

UCLA. *Swish-e: Simple Web Indexing Software for Humans - Enhanced.* Available at URL: http://swish-e.org/ Accessed 13 April, 2005.

Youssef, M. A. 2002. "Cross Language Information Retrieval", in: *Universal Usability in Practice*, April, 2002. Available at: http://www.otal.umd.edu/UUPractice/clir/ Accessed 13 April, 2005.