Ali Shiri
School of Library and Information Studies, University of Alberta, Edmonton

Stan Ruecker
Humanities Computing Program, Department of English and Film Studies, University of Alberta, Edmonton

Ximena Rossello
Department of Art and Design, University of Alberta, Edmonton

Matthew Bouchard
Humanities Computing Program, University of Alberta, Edmonton

Paras Mehta
Faculty of Medicine, University of Alberta, Edmonton

Development of A Thesaurus-enhanced Visual Interface for Multilingual Digital Libraries

Abstract: The aim of this paper is to report on the development of a visual user interface enhanced with a thesaurus to support information retrieval in a multilingual digital library context. The interface provides a design to support seamless access to three spaces, namely thesaurus space, query space and result display space.

Résumé : Le but de cette communication est de présenter le développement d'une interface visuelle pour l'utilisateur, améliorée par un thésaurus pouvant assister le repérage d'information dans le contexte d'une bibliothèque numérique multilingue. L'interface offre une conception donnant un accès uniforme à trois espaces, à savoir l'espace thésaurus, l'espace requête et l'espace d'affichage des résultats.

1. Introduction and Related Work

Digital libraries are multifaceted and complex information structures that offer a wide range and variety of information bearing objects. They vary in their content, subject matter, cultural characteristics, language etc. Arms (2000) notes that "a digital library is only as good as the interface it provides to its users." The variety of digital objects and materials in a digital library poses challenges to the design of usable and easy to understand user interfaces. Visual interfaces to digital libraries have recently found widespread attention. This development is mainly due to the fact that Information visualization techniques allow for rich representation of information bearing objects within digital libraries. Borner (2001) suggests that visual interfaces to digital libraries shift users' mental load from slow reading to faster perceptual processes such as visual pattern recognition. Zaphiris, et al. (2004) explore the application of information visualization in digital libraries and identify three key tasks in digital libraries, namely searching, browsing and navigation to which information visualization can make contribution. Visual interfaces to digital libraries exploit powerful human vision and spatial cognition to help humans mentally organize and electronically access and manage large, complex information spaces. The aim is to shift the user's mental load from slow reading to faster perceptual processes such as visual pattern recognition (Borner and

Chen, 2002). Shen et al. (2006) argue that text mining and visualization techniques provide digital libraries additional powerful exploring services, with possible beneficial effects on browsing and searching.

Over the last decade a number of digital libraries and online initiatives have incorporated knowledge organization systems such as thesauri and classification systems into their user interfaces to provide support for query formulation, collection browsing and other search tasks (Hodge, 2000; Hudon, and Hjartarson, 2002; Shiri and Molberg, 2005). A few prototype interfaces have utilized graphical as well as two- or three-dimensional category hierarchies using the MeSH Thesaurus. TraverseNet (McMath et al., 1989), MeSHBrowse (Korn & Shneiderman, 1995), Cat-a-cone (Hearst & Karadi, 1997), Visual MeSH (Xin, 1999), and the Integrated Thesaurus-Results Browser (Sutcliffe et al., 2000) are among the prototype thesaurus-enhanced interfaces. There are also some studies that have found that thesaurus-enhanced search interfaces can support users' query formulation and expansion (Beaulieu, 1997; Shiri, 2006). Jorna and Davis (2001) note that in order to facilitate cross-cultural communication in an increasingly global information society multilingual thesauri can play a significant role.

Thesauri have played an important role in modern information storage and retrieval systems. While initial proposals to utilize thesauri focused on their ability to ensure consistent analysis of documents during input to information retrieval systems, they have increasingly become vital as aids to effective retrieval (Shiri, 2000). Milstead (1998) noted that in the near future, it appears likely that thesauri will be used more during retrieval than at input. The move to increasing use of thesauri as an aid to retrieval has expanded their functional span within information retrieval systems. As Aitchison et al. (1997) have stated, the role of the thesaurus is changing, but it is likely to remain an important retrieval tool. This refocusing of the use of thesauri within information retrieval systems means that it is imperative that professionals take cognizance of the potential of thesauri as essential components of the largest information retrieval environment, namely the World Wide Web (Shiri, 2000).

The aim of this paper is to report on the development of a thesaurus-enhanced visual interface for multilingual information retrieval within the context of digital libraries. One of the major issues addressed by this research was to explore the ways in which subject access can be provided based on thesauri as an integral part of the search process. Most of current information access and retrieval systems provide a separate procedure or step to make use of thesauri for searching, browsing or navigation. Some of them have thesauri in advanced search facilities only. This will result in the lack of awareness on the part of users to know about and to be able to make use of semantic support that exists within thesauri. Our objective was to develop an interface which provides seamless access to both thesaurus and document spaces while having control over the search process in a multilingual environment.

2. Methodology

2.1 Theoretical Framework

The design of the interface was based on the concept of 'rich prospect' interfaces in which multiple representations can be used to allow access to a digital collection (Ruecker and Liepert, 2004). In the proposed interface, the aim was to provide the user with the following spaces within the interface:

- Query space: for formulating search statements
- Thesaurus space: for browsing and navigating the thesaurus
- Document space: for viewing document representations

A number of features have been implemented to cater for query formulation and reformulation (Figure 1). These are as follows:

- Language option
- High level facets
- Term browsing
- Scope notes
- Boolean operators
- Retrieved document sort

The form of the interface is tabular, based in part on suggestions by Bertin (2000/2001) that, if implemented with due consideration for the data, a matrix arrangement can provide the user with instantaneous understanding of the underlying structure. In the case of a person attempting to navigate a multilingual thesaurus for the purposes of query enhancement, it is helpful to provide indications of where in the hierarchy a particular thesaurus term resides.

For someone interested in query enhancement, it might be the case that any of these pieces of information is immediately relevant. Having chosen a starting term as a likely candidate to begin the search, the user may wish to expand to a broader search, or narrow to one more specific. In any case, it will be helpful to know what the system considers a preferred term and what is a term to be avoided. In addition, each of these considerations might be addressed in more than one language.

The goal of the interface is therefore to make this wealth of information readily accessible to the user during the process of query formulation or reformulation. A tabular view allows quick navigation through the five kinds of data, and a set of scope notes helps inform the user where a given term falls within the language of the thesaurus. In addition, a side panel presents a list of the highest-level categories in the thesaurus, which gives a user unfamiliar with the system a possible list of starting points.

A language switch provides a means of checking for corresponding terms in another language—these terms are in any case always visible as microtext satellites of the query terms. Their persistent presence in the thesaurus table both reminds the user that more than one language is available and also provides a quick means of switching back and forth between languages. This function is also served by an explicit language selection choice, made with a radio button in the panel to the right of the main table. The language choice for the prototype system is limited to English and French (the languages of the Government of Canada Core Subject Thesaurus). However, the data structure behind the thesaurus and the interface design are intended to be flexible enough to eventually accommodate a considerably wider range of languages.

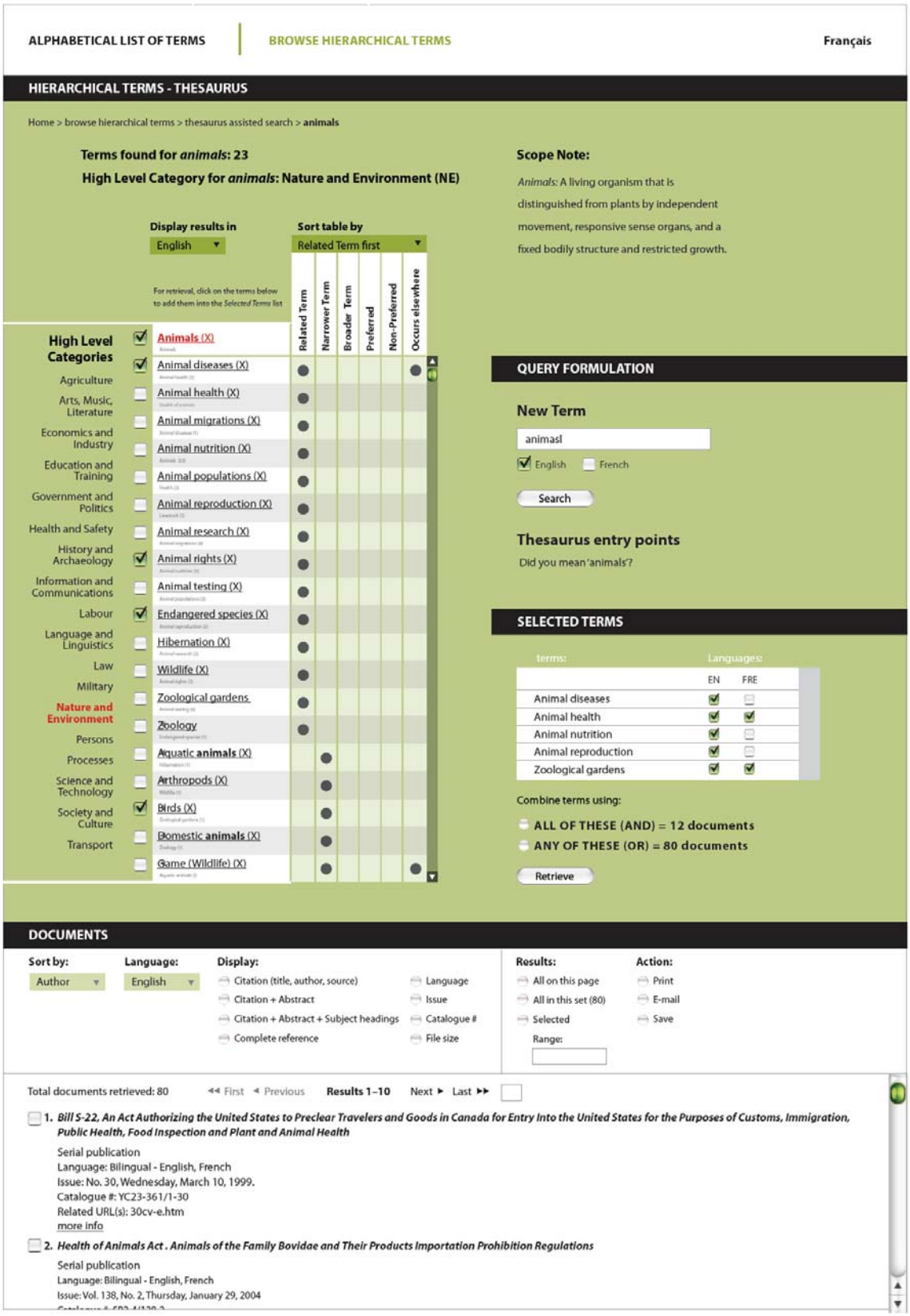

Figure 1. A Thesaurus-enhanced Visual Interface for Digital Libraries

The formulation and reformulation of queries is the primary goal of the interface. But in addition, once the user has constructed and executed a query, the system also provides some tools for managing the result set. These are found in the Document space, which

forms the third section of the screen, running across the bottom below the Thesaurus space on the left and the Query space to the right. The Document space shows the returns for the current query, in whichever languages have been selected. The documents are represented by standard bibliographic information about the author, title, and date; they can also be sorted in various ways, and each item serves as a link to the actual document.

By offering the combination of the three information spaces, namely query space, thesaurus space and documents space and in particular placing a query formulation mechanism within the context of a tabular display of the thesaurus terms and their interrelations, this study suggests possibilities for the efficient integration of multilingual thesauri into the query formulation process.

The Government of Canada Core Subject Thesaurus was chosen for this project for the following reasons:

- It is a bilingual thesaurus
- It is associated with a number of digital collections such as the Government of Canada Publications and the Government of Canada site
- It is a general purpose thesaurus in the sense that it is not confined to a specific subject area
- It represents a standard thesaurus with standard term relationships
- It is a faceted thesaurus which makes it suitable for browsing and searching

2.2 System design

The implementation of the system can be conceived in three layers: the database layer (or backend), the database-interface layer (which facilitates communication between the backend and the front-end), and the user-interface layer (or front-end). At each layer, the goal was to maximize flexibility and scalability to ensure that changes and additions could be elegantly handled.

Backend work on the project was developed starting a version of the Government of Canada Core Subject Thesaurus database provided in a plain-text format. This information was transferred into a MySQL database for faster access and easier interaction with web technologies.

The development of the database layer was focused carefully on scalability. Neither the number of language representations of a thesaurus entry nor the languages available in the database themselves were hard-coded into the database structure. This flexibility will make it relatively simple to add languages to the database and to add corresponding language representations of existing terms. The database-interface layer was written in PHP, a programming language that is a standard for generating dynamic web content and a common method for interacting with MySQL. In keeping with scalability and to ease the maintenance and upgrading of the system, the PHP layer was implemented using Object Oriented programming. Objects were created that represent a thesaurus term, a language representation of that term, and a document in the collection to be searched. The power of using objects is the abstraction of lower-level system concepts from higher-level system functionality. This means that you can easily modify low-level implementation details (such as the nature of a term representation) without changing high-level functions. Figure 2 shows the architecture of the system.

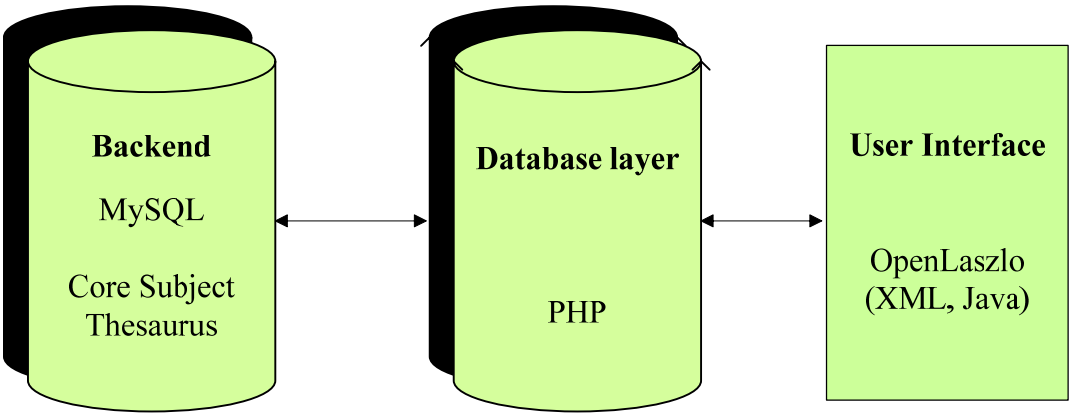

Figure 2. Architecture of the Visual Thesaurus-enhanced Search System

One important design choice, made to increase the efficiency of collection searches and to provide more information about search terms in the interface, was the pre-indexing of the document collection against the search terms. Having this information in the database allows us to show the user how many documents match a particular search term or combination of terms before they complete the search. This dynamic information provides the user with much richer browsing and searching environment, hopefully providing them with more useful results.

The front-end of the project used OpenLaszlo (or OL) as a protoyping tool. OL can be used to create rich Flash applications, but the programmer need not know Flash or ActionScript. The whole program is written in XML and Javascript which OL compiles into Flash. This allows the creation of web applications that have the functionality of desktop applications. By leveraging the power of Flash, OL makes programs that can be run on any browser on any operating system. OL also includes several useful interface structures (like buttons and drop-down menus) pre-built and easy to deploy. In addition to its portability, OL is written from the ground up as an interface prototyping language. The important advantage of this is that the base units of the language are all visual. There is no need to envision how the code will be translated into on-screen elements. The code is simply a description of what the screen will look like. Of course, this leads to the main disadvantage that things which are more easily implemented in a traditional, procedural language are more difficult in OL. As the language matures, these things will become easier as the creators of OL and its community fully explore its potential. Similar to the PHP layer, the user-interface layer was designed using Object Oriented methods to provide a front-end that easily accommodates more languages and more searchable collections. The objects are generated dynamically through XML that is produced by the database-interaction layer. OL fully implements Xpath and can quickly and easily access different levels of data.

3. Conclusion

The increasing number of prototype visual interfaces for digital libraries points to the growing popularity and acceptance of these types of interfaces for multimedia, multimodal and multilingual information repositories and digital libraries. In this paper we presented the methodological, technical and informational aspects of the development of a thesaurus-enhanced visual interface. The design of the interface takes advantage of a knowledge organization system and visualization techniques. The design provides a theoretical framework for creating semantically rich representation of digital collections.

The next stage of this project will focus on conducting a user-centred evaluation to assess the usability, usefulness and utility of the interface for carrying out various information search and retrieval tasks.

References

Aitchison, J., Gilchrist, A. and Bawden, D. (1997), *Thesaurus Construction and Use: A Practical Manual*, 3rd ed., Aslib, London.

Arms, William Y. (2000) *Digital Libraries*. M.I.T. Press.

Beaulieu, M. (1997). Experiments of interfaces to support query expansion. *Journal of Documentation*, 53 (1), 8-19.

Bertin, J.(2000/2001). Matrix Theory of Graphics. *Information Design Journal*. 10(1), 5-19.

Börner, K. and Chen, C. Visual Interfaces to Digital Libraries, *Lecture Notes in Computer Science* 2539, Berlin Heidelberg: Springer-Verlag, 2002.

Borner, K.; Chen, C. (2001) Visual Interfaces to Digital Libraries– The *First International Workshop at the First ACM/IEEE Joint Conference on Digital Libraries*, 2001. Available at: http://www.sigir.org/forum/S2001/DLVis.pdf

Hearst, M.A. & Karadi, C. (1997). Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In: *SIGIR '97: Proceedings of the 20th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, 27–31 July 1997 (ACM, New York, 1997).

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*, CLIR, Washington, Available at: http://www.clir.org/pubs/abstract/pub91abst.html

Hudon, M., & Hjartarson, F. (2002). Governments meet people: developing metathesauri in the framework of "government online" initiatives. In *Advancing knowledge : Expanding horizons for information science : Proceedings of the 30th Annual Conference of the Canadian Association for Information Science, 30 May - 1 June 2002, Toronto, Canada*, p.46-60. Edited by Lynne C. Howarth, Christopher Cronin, and Anna T. Slawek. Toronto, ON: CAIS, 2002.

Jorna, K. & Davies, S. (2001). Multilingual thesauri for the modern world - no ideal solution? *Journal of Documentation*. 57(2), 284-295.

Korn, F. & Shneiderman, B. (1995). *Navigating terminology hierarchies to access a digital library of medical images*, University of Maryland Technical Report HCIL-TR-94- 03, 1995.

Lin, X. (1999). Visual MeSH. In: M. Hearst, F. Gey and R. Tong (eds), *SIGIR'99: Proceedings of 22nd Annual International ACM/SIGIR Conference on Research and*

Development in Information Retrieval, Berkeley, CA, 15–19 August 1999 (ACM, New York, 1999).

McMath, C.F.; Tamaru, C.S. & Rada, R. (1989). A graphical thesaurus-based information retrieval system, *International Journal of Man–Machine Studies.* 31, 121–147.

Milstead, J. (1998). Thesauri in a full-text world. In Cochrane, P.A. and Johnson, E. (Eds), Visualizing Subject Access for 21st Century Information Resources, Proceedings of the 1997 Annual Clinic on Library Applications of Data Processing, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, IL, pp. 28-38.

Ruecker, S. & Liepert, S. (2004). Strategies for Creating Rich Prospect in Interfaces. *Joint International Conference of the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH)*. Göteborg, Sweden. June 11-16, 2004.

Shen, R., Srinivas, N., Vemuri, S. N., Fan, W., Torres, R. S. , Fox, E. A. (2006). Exploring digital libraries: integrating browsing, searching, and visualization. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries* JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, 1-10.

Shiri, A. A.; Revie, C. (2000) Thesauri on The Web: Current Developments and Trends. *Online Information Review*, 24(4): pp.273-279.

Shiri, A.; Molberg, Keri (2005). Interfaces to Knowledge Organization Systems in Canadian Digital Library Collections. *Online Information Review*. 29(6), 604-620.

Shiri, A. & Revie, C. (2006). Query Expansion behaviour within a thesaurus-enhanced search environment: A user-centred evaluation. *Journal of the American Society for Information Science and Technology*, 57 (4), 462-478.

Sutcliffe, A.G. , Ennis, M. & Hu, J. (2000). Evaluating the effectiveness of visual user interfaces for information retrieval, *International Journal of Human–Computer Studies* 53(5), 741–763.

Zaphiris, P., Gill, K., Ma, T. H. , Wilson, S., Petrie, H. (2004). Exploring the use of information visualization for digital libraries. *New Review of Information Networking*, 10 (1), May 2004, 51 – 69.