SINCERITY:
The making of a search engine for images indexed with a bilingual taxonomy

Authors: Tomasz Neugebauer, Elaine Ménard
Keywords: Digital images, Search interface, Image retrieval, Cross-language information retrieval, Multilingual information, Image collections

Journal: *OCLC Systems & Services*

Published Version: *OCLC Systems & Services: International digital library perspectives* (2015) 31 (3): 112–124.
https://doi.org/10.1108/OCLC-03-2014-0020

Post Print Version

1. Introduction

Image collections have seized upon the latest advances in technology, using computers and the Internet to offer unprecedented access to these visual resources. Image search engines attempt to give access to a wide range of online images available on the Web. The goal of image search is to return relevant and useful images based on the user's query. However, searching for images is somehow different from searching for text-based documents, mainly explained by the exceptionally subjective nature of images. Over the years, several studies (Fidel, 1997; Jörgensen and Jörgensen, 2005; Jansen, 2008; Ménard, 2009) stated that the image retrieval process has ignored several fundamental user issues, including the disparity between text and image retrieval. Actually, the same searching tools and their functionalities are used for both types of resources. For example, image searching is offered by general search engines, such as Google or Yahoo!, and by many specialized search services devoted to non-print or multimedia material. These engines allow users to browse multiple pages of results and provide various filtering options to increase search success.

However, the quality of image search engines' results notably depends on the significance of the textual information surrounding or associated with the images (e.g., filename, nearby text, page title or tags within the HTML code). It is also related to the users' ability to formulate their queries. Inevitably, image searchers are often overwhelmed by massive amount of results they receive searching with a simple query such as "monkey," "business" or "Tiger Woods." Even if search engines currently offer functionalities that are well-designed, these are not particularly well adapted to image searching and browsing. Moreover, most image retrieval is still limited to monolingual search results, that is, image queries are performed in only one language or, at least, a language familiar to the image searcher. This limitation immensely influences the relevance of the results.

This paper presents the third phase of a research project that aims to develop of a new bilingual interface called SINCERITY (Search INterfaCE for the Retrieval of Images with a TaxonomY). In the two previous phases, a best practices review was performed in order to acquire knowledge of the existing image search functionalities and to assess how they could be integrated in the development of a bilingual search interface dedicated specifically to images (Ménard and Smithglass, 2014). In the second phase, we investigated the roles and usefulness of these search characteristics and functionalities for the image search process in a bilingual context from the user's point of view, since image searchers are viewed as informants who can help Web designers fill gaps with their knowledge

(Ménard *et al.*, 2013; Ménard and Khashman, 2014). The combination of the exploration of best practices for image retrieval, the analysis of the behaviours and needs of real users when searching for images provided the foundation for the initial design of the search interface that will be developed in the ultimate stage of the research project.

The paper is structured as follows: Section 2 surveys previous studies in image access; Section 3 describes the methodology used for the initial design and development of the search interface; and Section 4 concludes the paper with suggestions on how the search interface will be tested, as well as future directions for the research project.

2. Related works

Studies on Web search engines reach in scope from technical developments to investigations on the social impact of these tools. Some studies (Goodrum and Spink, 2001; Spink and Jansen, 2004) provided valuable insight about query characteristics performed by image searchers. Other studies related to Web search engines (Machill *et al.*, 2004; Jansen and Spink, 2006) took into account typical user behaviours as possible sources of information to design better interfaces. Some search engine evaluation studies (Bar-Ilan, 2004; Bar-Ilan *et al.,* 2006) tested search engines through comparing their ranked results lists. Approaches to display results from many diverse angles have been examined (Thelwall, 2004; Ginsberg *et al.*, 2009). The literature offers abundant evidence on online image searching processes. In addition, the preceding studies provided valuable insight into approaches and algorithms interrelated to search interface design, including the future users' point of view. Additional studies, however, are needed to investigate how image searchers diverge from text-based searchers.

During the previous phase of this research project, we examined 159 resources provided by four types of organizations (libraries, museums, image search engines and stock photography databases). Search functionalities offered by text-based systems and content-based image retrieval systems have been explored (Ménard and Smithglass, 2014). This examination of existing image retrieval tools aimed to determine what type of functionalities are made available to their users in order to facilitate image retrieval in a multilingual context. This review of best practices was a crucial stage in the interface design because it provided the basic guidelines and characteristics necessary to improve the image retrieval process. All the major image search engines reviewed offer a similar practice for image retrieval: keyword-based query resulting in a grid of image thumbnails. Various query refinements are available, including image size, aspect ratio, colour and different kinds of content (e.g., illustrations, or images containing a face). Image retrieval can also be performed in a multilingual context. Several search engines give limited access to multilingual documents and integrated translation mechanisms

and functionalities in order to support searching in more than one language. Likewise some search engines offer monolingual searches in a number of languages, coupled with a machine translation software that translates the resulting Web pages into English or other languages (Bar-Ilan and Gutman, 2005).

Nowadays, the traditional way to search for images implies two steps: the user enters a query into a box and results are displayed (Morville and Callender, 2010). Content-based image retrieval (CBIR) systems were also explored in the previous phase of the research. CBIR is defined as a process that searches and retrieves images on the basis of automatically derived low-level characteristic features such as colour, texture and shape (Enser *et al.*, 2007; Jain and Singh, 2011), while text-based retrieval systems are based on the use of the verbal language to perform the queries (Jörgensen, 2003). Computer researchers have been working on technologies that would go beyond metadata to determine what objects are in a picture, and most efforts to date have only been moderately successful. However, our exploration revealed that CBIR systems are promising but still present some weaknesses (Mehyar and Atoum, 2012). Most CBIR systems still have a number of imperfections and remain at the experimental stage (Jörgensen, 2003). Consequently, text-based queries remain the most common way to retrieve images.

Our literature review drew attention to the fact that while search engines have evolved considerably over the years, so have the image searchers. During the second phase of our research project, we conducted a survey and interviews in order to investigate image searchers' needs and behaviours when searching for online images. The data analysis suggested several interesting implications for the design of the interface. For example, the analysis of participants' responses revealed the incontestable importance of keeping the search interface as simple as possible. The majority of the interviewed participants mentioned a box where the image searcher can enter a few keywords. However, one of the main outcomes of this second phase of the project revealed that image searchers still have difficulty formulating an effortless query that would lead them to the image they are looking for. Moreover, some respondents candidly admitted they would be interested in using predetermined subject categories to initiate their queries. Consequently, we hypothesize that including a taxonomic structure to begin the search process would be an interesting feature to include in our search interface (Ménard and Khashman, 2014). The idea of incorporating a taxonomy to help image retrieval has been suggested by many image searchers who participated in the exploration of the roles and usefulness of functionalities for image searching in a bilingual context (Ménard *et al.*, 2013). Very few search engines offer their users the opportunity to browse a taxonomic structure to initiate their queries.

Furthermore, the data from the survey and interviews illustrated the strong reliance by participants on keyword-based searching and underlined the lack of interest of the respondents in most content-based searching (CBIR) methods (e.g., using low-level characteristics such as shapes, colours, etc.). The image searchers prefer to search with various metadata-based keywords related to the content of the image they are looking for, the events taking place or people that appear in the picture. Even if image retrieval systems would greatly benefit from offering CBIR techniques, as reported in the literature review carried out for this study, image searchers seemed very sceptical about the effectiveness of initiating their search with a non-textual query, for example, with an image or a drawing. This result is important and explains why CBIR methods are not included in the search interface we developed.

There are abundant search engines that are currently in existence; however, few support true cross-language information retrieval (CLIR) that allows users to make queries in one language and retrieve documents in one or more other languages. For example, many search engines are monolingual but have the added functionality of carrying out translation of the retrieved pages from one language to another, for example, Google, Yahoo!, etc. (Zhang and Lin, 2009). Many research projects on query translation for different language pairs have been conducted in the last decade, as well as on the many linguistic resources integrated in CLIR systems (bilingual or multilingual dictionaries, machine translation [MT], parallel or comparable corpora). These linguistic devices all demonstrated advantages and limitations (Pirkola *et al.*, 2001; Hedlund *et al.*, 2004; Chen and Gey, 2004; Zhang and Vines, 2004; Xu and Weischedel, 2005). Most CLIR systems use an automatic translation device. Studies revealed that it usually performs well and can be used for translating queries, summaries and even complete documents (Bracewell *et al.*, 2008). Consequently, the integration of a query translation device using a machine translation device seems to be an obvious choice.

The survey and interviews we conducted prior to the interface development revealed that most image searchers formulate queries only in their native language when searching for images. Nevertheless, a majority of respondents expressed interest in the possibility of retrieving images associated with words from different languages. Consequently, our search interface model facilitates searching for images indexed in two different languages (English and French), giving the image searchers the opportunity to easily access a variety of visual resources, regardless of the image indexing language.

3. Methodology

3.1 Image database and indexing process

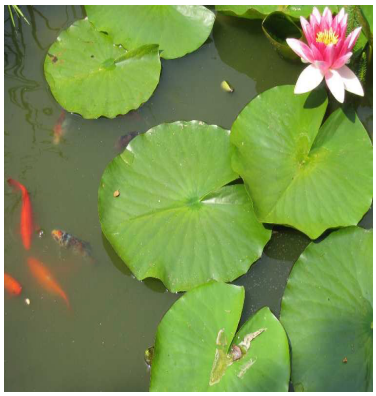
In the summer of 2011, 14 people agreed to give copies of their photos to the IDOL image database (Images DOnated Liberally). A total of 8,932 images was first collected and reduced to 6,015 by removing duplicates. Subsequently, all images have been indexed using TIIARA (Taxonomy for Image Indexing And RetrievAl), a taxonomy created for the indexing of ordinary images (Ménard, 2012). The development of TIIARA was based on an extensive analysis of existing specialized terminology used by professional indexers to describe images, as well as the tags employed by regular Internet users. This exploration was undertaken to determine how these terminologies could be integrated in the development of the taxonomy. An evaluation of 150 vocabulary resources that organize and describe images (libraries, museums, search engines and commercial websites) was carried out. This examination of best practices for the organization of digital images used by indexing specialists and non-specialists alike was a crucial step, since it provided the basic guidelines and standards for the categories, formats of terms and relationships to be included in the new bilingual taxonomy (Ménard and Smithglass, 2012).

Two graduate students (an English and a French native speaker) from the McGill University School of Information Studies, with the necessary indexing skills, were chosen to index the IDOL image database using TIIARA. The indexing process had two main steps:

1) *Image visualization and selection of the main 'subject'*. Since the primary goal of the indexing policy was to encourage *precision*, only the main subject and/or concept needed to be identified when visualizing the images.

2) *Translation of the main 'subject' into an indexing term* chosen from the TIIARA main categories (English or French version) at the level of the most *specific* available term .

To check the reliability of the indexing, an inter-indexer consistency test was conducted on the basis of 300 images (5%) randomly extracted from the IDOL database. The two project indexers achieved a consistency rate of 65%. Of the 300 images, 195 images were indexed with the same English and French terms, 105 images with different English and French terms. This result is consistent with the rates generally observed (Lancaster, 2003; Ménard, 2008). The test did highlight some gaps in TIIARA that would need to be addressed, but the results were sufficiently encouraging to justify asking the indexers to proceed to the next stage, indexing the entire IDOL image database. At the end of the indexing process, all images from the IDOL image database were indexed with two (one English and one French) TIIARA subcategories as shown in Figure 1.



IDOL-1752 English indexing term: Flowers (L3) French indexing term: Nature (L1)	IDOL-5817 English indexing term: Weddings and commitment ceremonies (L4) French indexing term: Mariage et cérémonie d'engagement (L4)

Figure 1. Examples of images and indexing terms associated with them

3.2 Implementation

The implementation of SINCERITY shown in Figure 2 is the result of the following requirements that were defined during the earlier phases of the research:

- Multilingual indexing
- Taxonomy-based browsing
- Customizable user interface
- Query translation between French and English
- Scalability to large image collections

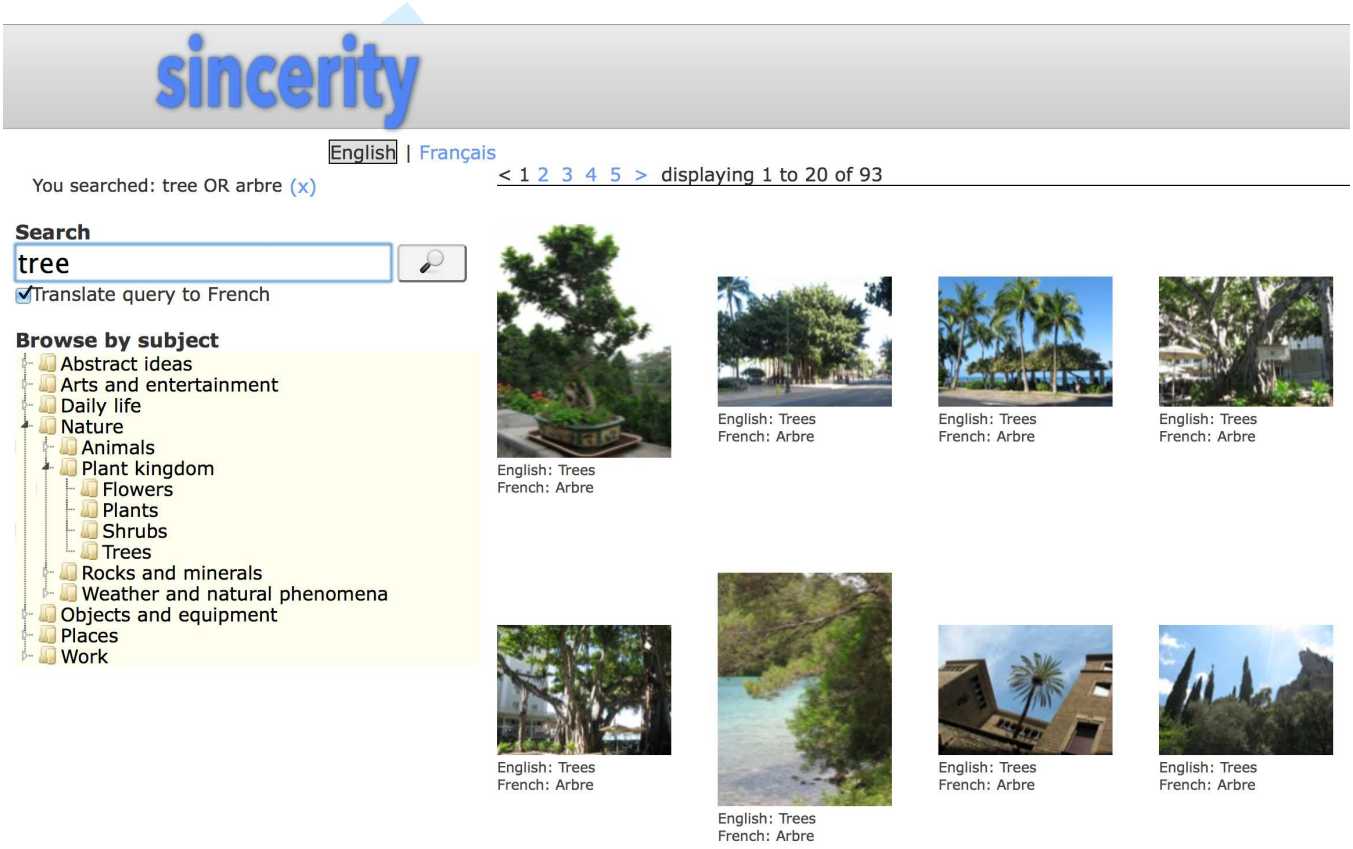


Figure 2. SINCERITY user interface

The strategy of using open-source software components as much as possible for the solution was chosen for the advantages of this approach: low initial cost and accessibility to evaluate and develop enhancements independently and driven by research objectives rather than financial viability.

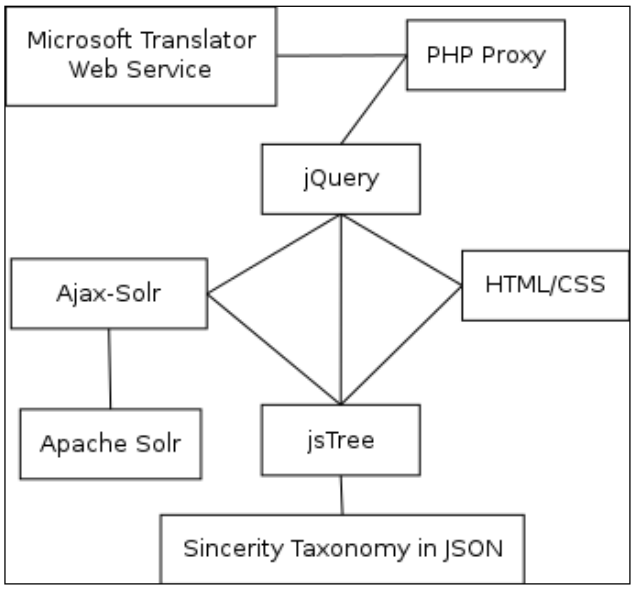


Figure 3. SINCERITY components

Figure 3 shows the main components of the SINCERITY implementation. The Lucene search library has been used to build image search utilities for the Windows operating system (Ng, 2008). However, given the research context of image searching on the Web, the decision to embed the SINCERITY interface in a web browser rather than a Windows interface made sense. Developing SINCERITY using HTML/CSS that runs on a browser ensures that the interface is accessible across operating systems and could be moved onto the Web in the future.

Apache Solr[1] is a well-known open-source search server built with the Lucene search library at its core. Solr was chosen as the indexing technology for SINCERITY. Among the many features of Solr, the following were particularly important: proven scalability, an active open-source community, extendible plug-in architecture, flexibility and adaptability of the indexing configuration. Solr runs as a stand-alone search server that is easy to use through HTTP with other programming languages. The use of Solr made the addition of French and English word stemming relatively straightforward, for example.

Much research in image indexing is focused on CBIR. Even though CBIR is currently not a requirement for SINCERITY, the potential to easily include these indexing techniques in the future is a technological consideration. Choosing Apache Solr as a search server provides the additional advantage of the potential to leverage CBIR technologies, such as the LIRE open-source library (Lux and Chatzichristofis, 2008), recently made available as a Solr plug-in.[2]

The SINCERITY end-user interface is implemented with the AJAX-Solr JavaScript framework for creating user interfaces to Solr.[3] This framework uses jQuery[4] Ajax to pass and retrieve search results from Solr, based on end-user search and browse queries. Moreover, jQuery is the foundation

library for the jsTree plug-in[5] used for implementing the interactive browse tree for the taxonomy. The taxonomy was encoded using JSON format so that it would be compatible with jsTree.

Lastly, the requirements included the need to optionally translate user queries between English and French before passing them to the search server. The Google Translate web service requires a monthly payment for any usage, while the Microsoft Translate[6] web service is free for up to 2 million characters per month. Sending the translation web service requests to the translator requires a PHP proxy due to the authentication routine involved in Microsoft Translate as well as the browser security restrictions preventing cross-domain JavaScript.

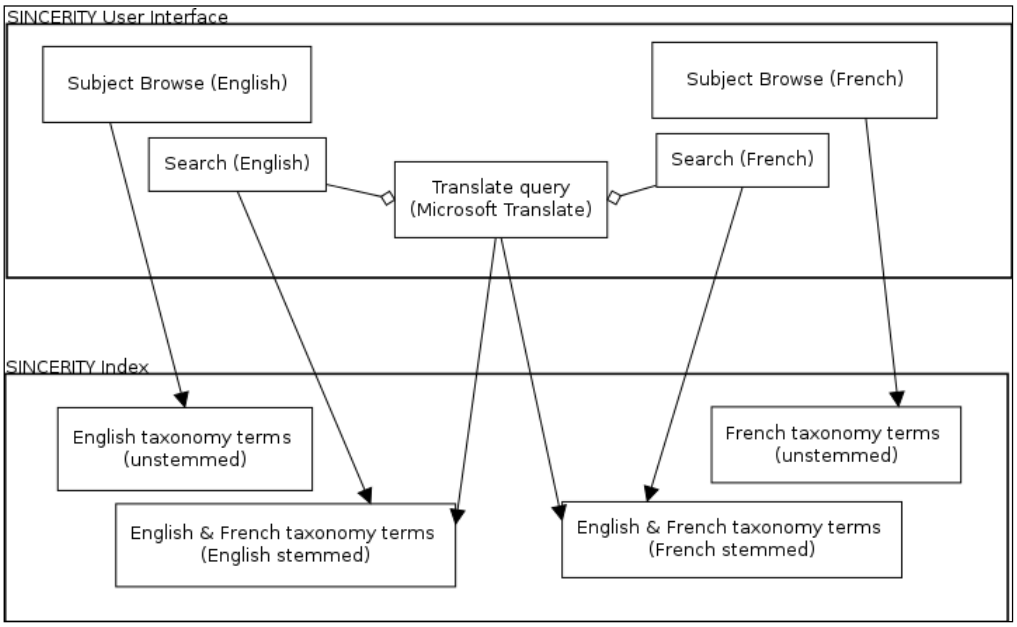


Figure 4. SINCERITY user interface and its Solr indexes

There is a scarcity of research papers describing the indexing implementation details of hierarchical faceted search applications (Ben-Yitzhak *et al.*, 2008). Solr's facets are flat, so the indexing of hierarchical taxonomies with Solr involves indexing the taxonomy *paths* for each image.

In order to keep the implementation of the taxonomy navigation as simple as possible, the taxonomy structure was encoded in JSON and rendered with jsTree. Figure 4 shows how the main functions on the end-user interface relate to the Solr index. The Subject Browse in English and French is implemented by linking the jsTree interactive tree with the underlying English and French terms indexed in Solr without stemming. All of the taxonomy levels are included in this unstemmed index, allowing the user to drill down to more specific terms. For example, the index for an image with the hierarchy "Daily life > Food and drink > Alcohol" would contain all three of these terms. The category names in the taxonomy are almost entirely unique, and thus the jsTree implementation needs only to send the final term, "Alcohol" in this example, to obtain the appropriate results. In case of terms that

might repeat in more than one place in the taxonomy, jsTree sends the additional search parameters consisting of the more general terms, in the example, “Daily life” and “Food and drink.”

User search queries are sent to their respectively treated indexes in Solr. Thus, the French-language interface queries against the French-treated index in Solr. English terms are also included in this index, making it possible for the end-user to make an English query even though they are on the French interface. However, given that the terms are stemmed as English on the English interface only, the search results could be better in some cases when the interface language matches the query language. For example, if the user types in “designing” on the French interface, they will not get any results, because retrieval depends on English stemming: there are images actually indexed with the English term “Designed products.” The same query on the English interface will perform the appropriate stemming and so retrieve the relevant images.

When the user activates the runtime query translation with Microsoft Translate, the results of the translation are queried against the Solr index with the appropriate language. Thus, if the user enters “house” on the English language interface with the “translate query” set, the interface will query “house” on the English stemmed index and the result of the translation (“maison”) on the French stemmed index.

Besides the prepared taxonomies in English and French, the interface needed the functionality to translate custom user queries before sending them to Solr for retrieval. There are a few technologies available to choose from for this purpose, including the Google Translate API.[7] The Microsoft Translator service was chosen. It is a statistical machine translation system that integrates human translation to improve the quality of the translation, made available to the public by Microsoft since 2007 (Wendt, 2010). The reasons for choosing the Microsoft Translate service for SINCERITY include the following:

- Well-documented AJAX API
- Cost-free for under 2 million characters per month
- Robust system that can translate and detect 45 languages

In addition to the technical availability and no initial cost, Microsoft Translate performed comparably to other translation systems when tested with English-to-French translation of user-generated content, and even surpassed other systems significantly with the English-to-German language pair (Roturier and Bensadoun, 2011).

4. Conclusion and further work

The context of use of the proposed model is that of image search and browse tasks through a collection of images, indexed with multiple languages, on a personal computer, commercial website or the site of a museum or a special collection. SINCERITY is a research prototype for testing the usability of a search and retrieval tool for collections of images indexed in more than one language, at the level of everyday familiarity with objects and events.

The current implementation does not make use of the language detection feature of Microsoft Translate. Instead, the assumption is made that users enter queries in the language matching that of the selected language of the interface. If testing reveals this assumption to be problematic, the language detection features of Microsoft Translate could be used to detect if the user entered a French-language or English-language term, and only then sending the query to the appropriate index in Solr.

The images in SINCERITY are indexed with only one taxonomy path per image. Therefore, no attempt was made to provide additional relevancy ranking within each path. User testing could reveal the need for additional relevance ranking and sorting of search results. The addition of this functionality would require reindexing the images with multiple taxonomy paths per image or the addition of content-based search parameters.

In the next phase of the research project, we plan to continue improving our prototype, in terms of both performance and aesthetics. A two-phased user testing strategy will be conducted to ensure that the final product is clear, comprehensive and consistent. The first stage of user testing will consist of a group of image searchers formulating queries to retrieve a sample of images using the interface. A questionnaire will be developed in order to gather the comments of the users. Unstructured interviews will also be conducted to deepen our knowledge of the ins and outs of the search interface. Upon completion of this first evaluation, the search interface will be refined according to the comments and suggestions received.

The second stage of testing will entail an evaluation of the performance of SINCERITY, involving a usability test under experimental conditions. An exhaustive simulation of the retrieval process will be performed. From the image database, a sample of realistic image retrieval tasks will also be determined to serve as topics that can be submitted as queries to retrieve images in order to eventually evaluate the retrieval performance supplied by the new search interface. The objective of this testing is to ask a representative sample of image searchers to complete typical image retrieval tasks using the new interface in order to measure the degree of effectiveness, efficiency and user satisfaction. The interface testing is expected to identify the usability inconveniences of SINCERITY

that may not be revealed by less formal testing (Sproull, 2002). The experiment will also aim to evaluate the structure of the interface and the quality of the translation.

The search interface model is intended to be an innovative and powerful tool for image searchers who are looking for ordinary images or other types of images. Eventually, documentary images related to a specific area (e.g., sports news, medical imaging) or artistic images (e.g., museum objects, famous works) could be included in the image database if we were able to obtain a better understanding of the best approach to adopt to optimize image retrieval.

Other languages could eventually be added to SINCERITY in order to offer members of other linguistic communities equivalent image retrieval opportunities and subsequently bridge the information divide that still exists. The methodology proposed by this project could also be applied to other types of documents, such as audio files or videos, for example, in order to study how to index and retrieve these multimedia documents and make the necessary suggestions to improve their organization and, eventually, their retrieval.

References

Bar-Ilan, J. (2004), "The use of web search engines in information science research", in Cronin, B. (Ed.), *Annual review of information science and technology*, Information Today, Medford, NJ, Vol. 38, pp. 231-288.

Bar-Ilan, J. and Gutman, T. (2005), "How do search engines respond to some non-English queries? ", *Journal of Information Science*, Vol. 31 No. 1, pp. 13-28.

Bar-Ilan, J., Mat-Hassan, M. and Levene, M. (2006), "Methods for comparing rankings of search engine results", *Computer Networks*, Vol. 50 No. 10, pp. 1448-1463,

Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., Sheinwald, D., Shekita, E., Sznajder, B. and Yogev, S. (2008), "Beyond Basic Faceted Search", paper presented at the *2008 International Conference on Web Search and Data Mining*, Palo Alto, USA, available at: http://dl.acm.org/citation.cfm?id=1341539 (accessed 28 May 2014).

Bracewell, D. B., Ren, F. and Kuroiwa, S. (2008), "A low cost machine translation method for cross-lingual information retrieval", *Engineering Letters*, Vol. 16 No. 1, pp. 160-165.

Chen, A. and Gey, F. C. (2004), "Multilingual information retrieval using machine translation, relevance feedback and decompounding", *Information Retrieval*, Vol. 7 No. 1/2, pp. 149-182.

Enser, P.G.B., Sandom, C.J., Hare, J.S. and Lewis, P. H. (2007), "Facing the reality of semantic image retrieval", *Journal of Documentation,* Vol. 63 No. 4, pp. 465-481.

Fidel, R. (1997), "The image retrieval task: implications for the design and evaluation of image databases", *The New Review of Hypermedia and Multimedia*, Vol. 3 No. 1, pp. 181-99.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7634, pp. 1012-1014.

Goodrum, A. and Spink, A. (2001), "Image searching on the EXCITE Web search engine", *Information Processing & Management*, Vol. 37 No. 2, pp. 295-311.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. and Järvelin, K. (2004), "Dictionary-based cross-language information retrieval: learning experiences from CLEF 2000-2002", *Information Retrieval*, Vol. 7 No. 1/2, pp. 99-119.

Jain, M. and Singh, S. K. (2011), "A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques For Large Data sets", *International Journal of Managing Information Technology,* Vol. 3 No. 4, pp. 23-39.

Jansen, B. (2008), "Searching for digital images on the web", *Journal of Documentation*, Vol. 64 No. 1, pp. 81-101.

Jansen, B. and Spink, A. (2006), "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", *Information Processing & Management*, Vol. 42 No. 1, pp. 248-263.

Jörgensen, C. (2003), *Image retrieval - theory and research*, Lanham, Scarecrow Press.

Jörgensen, C. and Jörgensen, P. (2005), "Image querying by image professionals", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 12, pp. 1346-59.

Lancaster, F. W. (2003), *Indexing and abstracting in theory and practice*, Champaign, Ill., Graduate School of Library and Information Science, University of Illinois.

Lux, M. and Chatzichristofis, S.A. (2008), "Lire: lucene image retrieval: an extensible java CBIR library", in *Proceedings of the 16th ACM international conference on Multimedia (MM '08)*. ACM, New York, NY, pp.1085-1088.

Machill, M., Beiler, M. and Zenker, M. (2008), "Search-engine research: A European-American overview and systematization of an interdisciplinary and international research field", *Media, Culture & Society*, Vol. 30 No. 5, pp. 591-608.

Ménard, E. (2013). TIIARA for an IDOL: Indexing adventure of a small digital image collection, *The Indexer*, 31 (1), 2-11.

Ménard, E. (2012), "TIIARA: The "making of" a bilingual taxonomy for retrieval of ordinary images", *Library Hi Tech*, Vol. 30 No. 4, pp. 643-54.

Ménard, E. (2009), "Image Retrieval: A comparative study on the influence of indexing vocabularies", *Knowledge Organization,* Vol. 36 No. 4, pp. 200-213.

Ménard, E. (2008), *Étude sur l'influence du vocabulaire utilisé pour l'indexation des images en contexte de repérage multilingue,* Doctoral Dissertation, Montréal, Université de Montréal, available at: https://papyrus.bib.umontreal.ca/jspui/bitstream/1866/2611/1/ menard-e-these-indexation-reperage-images.pdf (accessed 28 May 2014).

Ménard, E. and Khashman, N. (2014), "Image retrieval behaviours: The real users are leading the way", *Library Hi Tech*, Vol. 32 No. 1, pp. 50-68.

Ménard, E. and Smithglass. M. (2014), "Digital image access: An exploration of the best practices of online resources", *Library Hi Tech*, Vol. 32 No. 1, pp. 98-119.

Ménard, E. and Smithglass, M. (2012), "Digital image description: A review of best practices in cultural institutions", *Library Hi Tech*, Vol. 30 No. 1, pp. 291-309.

Ménard, E., Khashman, N. and Dorey, J. (2013), "Two solitudes revisited: A cross-cultural exploration of online image searchers behaviours", in Marcus, A. (Ed.), *15th International Conference on Human-Computer Interaction 2013, Las Vegas, NV, July 21-26, 2013*, Lecture Notes in Computer Science, Vol. 8013, Berlin, Springer, pp. 79-88.

Mehyar, T. and Atoum, J. O. (2012), "An enhancement on content-based Image retrieval using color and texture features", *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 3 No. 4, pp. 488-496.

Morville, P. and Callender, J. (2010), *Search Patterns*, O'Reilly, Sebastopol, CA.

Ng, A. (2008), "Indexing and Searching Image files", Dr. Dobb's, available at: http://www.drdobbs.com/windows/indexing-and-searching-image-files/210603684?pgno=1 (accessed 28 May 2014).

Pirkola, A., Hedlund, T., Keskustalo, H. and Järvelin, K. (2001), "Dictionary-based cross-language information retrieval: problems, methods, and research findings", *Information Retrieval,* Vol. 4 No. 3/4, pp. 209-230.

Roturier, J. and Bensadoun, A. (2011), "Evaluation of MT Systems to Translate User Generated Content", in *Proceedings of Machine Translation Summit XIII, Xiamen, China*, pp. 244-251, available at: http://mt-archive.info/MTS-2011-Roturier.pdf (accessed 28 May 2014).

Spink, A. and Jansen, B. J. (2004), *Web search: public searching of the Web*, Kluwer Academic, Boston.

Sproull, N. L. (1995), *Handbook of research methods: a guide for practitioners and students in the social sciences,* Scarecrow Press, Metuchen, NJ.

Thelwall, M. (2004), *Link analysis: An information science approach*. *Library and information science*, Academic Press, Amsterdam.

Wendt, C. (2010), "Better translations with user collaboration – Integrated MT at Microsoft", paper presented at *AMTA 2010*, Denver, available at: http://mt-archive.info/AMTA-2010-Wendt.pdf (accessed 28 May 2014).

Xu, J. and Weischedel, R. (2005), "Empirical studies on the impact of lexical resources on CLIR performance", *Information Processing & Management,* Vol. 41 No. 3, pp. 475-487.

Zhang, J. and Lin, S.Y. (2007), "Multiple language supports in search engines", *Online Information Review*, Vol. 31 No. 4, pp. 516-532.

Zhang, Y. and Vines, P. (2004), "Using the web for automated translation extraction in cross-language information retrieval", paper presented at the *SIGIR'04*, Sheffield, UK, available at: http://tinyurl.com/jv9l7 (accessed 28 May 2014).

Acknowledgments
This research project is funded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

Biographical Details
Tomasz Neugebauer is an Associate Librarian, Digital Projects & Systems Development at Concordia University (Montreal, Canada). His main research interests focus on digital libraries, metadata and open source software development. He has published in various scholarly journals, including *Information Technology and Libraries*, *International Journal on Digital Libraries*, *International Journal of Digital Curation, Art Libraries Journal*, and *The Indexer.*

.

Elaine Ménard is an Associate Professor at the School of Information Studies, McGill University (Montreal, Canada). Her teaching expertise includes cataloguing, indexing, classification and information retrieval. Her main research interests deal with cross-language information retrieval, image indexing and metadata. She has published in a number of scholarly journals, including *The Indexer*, *Journal of Information Ethics*, *Knowledge Organization*, *Library Resources & Technical Services* and *Library Hi Tech.*

---

[1] Apache Solr, available at http://lucene.apache.org/solr/ (accessed 28 May 2014).
[2] LIRE Solr plug-in, available at https://bitbucket.org/dermotte/liresolr (accessed 28 May 2014).
[3] AJAX-solr, available at https://github.com/evolvingweb/ajax-solr (accessed 28 May 2014).
[4] jQuery, available at http://jquery.com (accessed 28 May 2014).
[5] jsTree, available at http://www.jstree.com (accessed 28 May 2014).
[6] Microsoft Translator, available at http://datamarket.azure.com/dataset/bing/microsofttranslator (accessed 28 May 2014).
[7] Google Translate API, available at https://developers.google.com/translate/ (accessed 28 May 2014).