

# Indexing and Retrieving Images in a Multilingual World

Elaine Ménard

École de bibliothéconomie et des sciences de l'information,  
Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal,  
Québec, Canada H3C 3J7 <elaine.menard@umontreal.ca>

Elaine Ménard is a Ph. D. Candidate in Information Science (École de bibliothéconomie et des sciences de l'information, Université de Montréal). She also holds a Masters Degree in Information Science, a BA and a Masters Degree in Translation (Université de Montréal). Her main research interests are in multilingual information retrieval, image indexing and metadata.



Menard, Elaine. Study on the Influence of Vocabularies used for Image Indexing in a Multilingual Retrieval Environment. *Knowledge Organization*, 34(2), 91-100. 23 references.

**ABSTRACT:** For many years, the Web became an important media for the diffusion of multilingual resources. Linguistic differences still form a major obstacle to scientific, cultural, and educational exchange. Besides this linguistic diversity, a multitude of databases and collections now contain documents in various formats, which may also adversely affect the retrieval process. This paper describes a research project aiming to verify the existing relations between two indexing approaches: traditional image indexing recommending the use of controlled vocabularies or free image indexing using uncontrolled vocabulary, and their respective performance for image retrieval, in a multilingual context. This research also compares image retrieval within two contexts: a monolingual context where the language of the query is the same as the indexing language; and a multilingual context where the language of the query is different from the indexing language. This research will indicate whether one of these indexing approaches surpasses the other, in terms of effectiveness, efficiency, and satisfaction of the image searchers. This paper presents the context and the problem statement of the research project. The experiment carried out is also described, as well as the data collection methods.

## 1. Introduction

For many years, the Web has become an important media for the diffusion of multilingual resources. Nowadays, we are aware that there are approximately 7,000 living languages throughout the world. Even if it is difficult to consider the exact number of written languages among those, mainly due to the lack of reliable sources of information available, we can suppose that the majority of these languages will possibly generate documents in textual or different other forms. On the Web, linguistic differences still form a major obstacle to scientific, cultural, commercial, and educational exchange. Moreover, the information retrieval process is also confronted to the problem of the superabundance of results. Far from being eclipsed, this problem becomes even more extensive with the emergence of a variety of languages

on the Web. Presently, the Web constitutes a vast universe of knowledge and various human cultures allowing the dissemination of ideas and information and that, in a multitude of languages. Consequently, with the ever increasing size of the Web and the availability of more and more documents in various languages, the access to multilingual information became a gigantic challenge.

Besides this linguistic diversity, more and more databases and collections now contain documents in various formats (text, image, video, etc.), which may also adversely affect the retrieval process. For example, the retrieval of digital images offers different characteristics from the retrieval of textual documents: "Images are notoriously difficult to retrieve with accuracy, as is obvious to anyone who has searched for images on the World Wide Web" (Harpring 2002, 20). Image collections are relatively new

compared to textual ones, but they have developed in an exponential way: “as more and more people and organizations load images onto the Web, the searching and retrieval of images has become a major challenge for researchers, commercial practitioners, and recreational users alike” (Spink and Jensen 2004, 161). It is therefore important to organize all this visual information in order to maximize its accessibility and eventually, its usability. Among the many types of images available on the Web, the ordinary image (i.e., non artistic) occupies an important place in the many searches of Web users. By ordinary images, we mean images representing everyday-life objects, for example: a television set, a lawn mower, a feeding-bottle, etc. The “ordinary” qualifier distinguishes this type of image with the images known as “artistic images” whose first value is to represent an artistic expression (Turner 1998). The growth of the Web highlighted the pressing need to offer suitable tools for image description aiming to facilitate their location since we now find those in the majority of the Web resources: personal pages and sites, virtual libraries, museum collections, commercial services and product catalogues, government information, etc.

In general, two categories of queries are used for retrieving images on the Web: graphic or textual queries. In the former category, the individual submits an “image” query (using an image or a drawing) and the system tries to retrieve a similar image by using certain physical characteristics of the image such as the colour, the shape or the texture. However, the majority of image searches on the Web still use textual queries (Clough et al. 2006). It is this latter category that is studied in this research. When using textual queries the success of the retrieval largely depends on the correspondence between the query of the searcher and the text associated with images. Given the great linguistic diversity existing on the Web, we must expect that the text associated with images exists in many different languages. Consequently, when using a textual query, the user faces a double challenge when he tries to retrieve images. On the one hand, the terms of the query must correspond to the text associated with the images and on the other hand, the language of the query must concur with the language of the text associated with the images. This research project aims to investigate this double problem faced by the user during image retrieval in a multilingual context.

## 2. Problem Statement

Image retrieval exhibits some similarities with textual document retrieval, but distinguishes itself in many ways such as the query formulation, the retrieval method, the technique in which the relevance of the results is evaluated, the participation of the individual in the retrieval process, and the fundamental cognitive differences in the interpretation of the visual rather than textual material (Clough and Sanderson 2003). By their very nature, images are language-independent resources. The main difference and difficulty in indexing and classifying images, as opposed to text, is that images do not satisfy the requirements of a specific language. In other words, we cannot automatically label a specific language to this particular document type. For example, we can not say that an image is Chinese or German, as we would say about a textual document written in one of these languages. Nevertheless, the indexing process, with controlled or uncontrolled vocabulary, gives the image some linguistic status similar to any other textual document, which can significantly affect its retrieval in a multilingual context. For example, if a user formulates a query in French and the images are also indexed in French, the cross-lingual problem does not arise. However, if a French query is used and the images are indexed in English (or any other language different from the query language), the retrieval will only be possible if the retrieval system includes a Cross Language Information Retrieval (CLIR) mechanism which allows cross-language mapping between the query terms and the indexing terms.

In order to make this mapping possible, Cross Language Information Retrieval generally proposes two approaches. The first approach consists of translating all the documents in all languages, while the second approach supposes the query translation in the language of the documents to be retrieved. Usually, these two approaches make use of three types of linguistic resources for the document or the query translation: bilingual or multilingual dictionaries, machine translation (MT) systems, or parallel and comparable corpora. For this research project, the approach proposing the query translation is preferred, as well as a machine translation system for the translation process. On the one hand, query translation rather than document translation seems better suited for the purposes of our research, since it can be easily obtained and at a lower cost, by simply adding a machine translation mechanism in the retrieval system. On the other hand, the retrieval

performance using a machine translation mechanism, in a multilingual context, generally ranges between 80 and 100 percent of the retrieval performance in a monolingual context (Nie 2006). However, no matter the approach used (query or document translation) and the linguistic resource employed for the translation (dictionary, machine translation or corpora), the risks of ambiguity remain preoccupying at the time of the translation.

Since the beginning of the 19<sup>th</sup> century, cataloguing, classification and indexing processes were primarily devoted to textual documents. However, the access to audiovisual material raises as much interest for scholars, if not more. According to Guinchat and Menou (1990), the indexing process of visual material poses particular problems because of its own nature and mode of consultation. As we can presume, the choice of suitable index terms for image description is primordial, whether in a monolingual or multilingual retrieval context. This research project proposes to examine what occurs at the retrieval stage when the image is indexed according to one or the other of the two following approaches: 1) traditional image indexing recommending the use of controlled vocabularies, i.e. the indexing terms are chosen from an artificial language whose main function is to generate formal document representations (Hudon 2007), or 2) free image indexing using uncontrolled vocabulary i.e. the indexing terms being used to describe the image do not obey any pre-established rule. We assume that these two indexing approaches show common characteristics, but also differences which may influence retrieval. This research will thus make it possible to indicate whether one of these indexing approaches surpasses the other, in terms of effectiveness, efficiency, and satisfaction of the image searchers.

The first indexing approach studied in this research project recommends the use of the controlled vocabulary. Generally this type of indexing supposes that a metadata schema lays down the structure, the elements and the vocabulary to be used. A metadata schema can be defined as a set of elements designed to meet the needs of specific communities (Smiraglia 2005). Some metadata schemas are of a general nature, while others were designed for specific needs like the description of government information, geo-spatial data, multimedia resources, etc. In addition to the structure definition and the elements to be used, some metadata schemas also propose which vocabularies must be employed for indexing purposes. Typically, the metadata schemas suggest the use of two types of vocabularies: controlled vocabularies or

uncontrolled vocabularies. However, the majority of metadata schemas recommend the use of controlled vocabularies, in order to increase the indexing consistency (Jørgensen 2003).

Controlled vocabularies aim to facilitate the indexing process. Moreover, the use of controlled vocabularies offers many advantages for retrieval, browsing and interoperability. However, controlled vocabularies also present some weaknesses: the main one is to represent the concepts in an artificial way. Indeed, the indexing terms offered by controlled vocabularies often have very few connections with the terms used by individuals in the formulation of their queries (Furnas et al. 1987). Furthermore, controlled vocabularies have a tendency to become quickly outdated. Thus it constitutes another disadvantage since neologisms take a long time before they are integrated in the various controlled vocabularies. Moreover, the use of these vocabularies remains a complex task for the majority of indexers. Finally, most controlled vocabularies suggested by metadata schemas and commonly used for image indexing only exist in English. Consequently, an indexer having little knowledge of the English language and wishing to use these vocabularies will face a major linguistic problem unless a translation mechanism is offered to facilitate their use.

The second indexing approach studied constitutes an alternative to traditional indexing using controlled vocabularies. Indeed, free indexing which, in addition, does not impose any structure or mandatory element, generally uses uncontrolled vocabulary. This long-neglected form of indexing is currently becoming very trendy. As an example of free indexing, let us mention the indexing we encounter in photo sharing sites, like *Flickr* ([www.flickr.com](http://www.flickr.com)). These sites allow massive image storage and diffusion on the Web. In these systems, the users upload their own images and index them using their own terms. It is also possible to make these images public, i.e. the images can be seen by all users or by a group of people chosen by the system user, thus forming a vast and communal image database. In a *flickr* system, the user who uploads images can thus determine who will have access to these images by stating certain rules of access control. In parallel, other users of the system have the possibility to update the image indexing by adding other keywords (tags) or comments to any image they have access to. These annotations assigned by the uploader or any other user of the *flickr* system constitute a form of free indexing called collaborative tagging. Moreover, in this type of photo sharing system, the index-

ing can be done in one language or combine several languages. However, despite its growing popularity and much like controlled vocabulary indexing, free indexing with uncontrolled vocabulary also presents several gaps. For example, some ambiguities emerge because the same keyword is often employed by several individuals, but in various contexts. In the same vein, the lack of synonymic control involves the use of many different keywords to describe the same concept. Consequently, the free indexing is often considered to be of poor quality.

The use of controlled or uncontrolled vocabularies raises a certain number of difficulties for the indexing process. These difficulties will necessarily entail consequences at the time of image retrieval. Indexing with controlled or uncontrolled vocabularies is a question extensively discussed in the literature. Several scholars explored the question of the best vocabulary to be used for the indexing process. The results of many studies indicate that uncontrolled vocabularies offer an output higher than the controlled vocabulary (Savoy 2005). This point of view was also shared by Rao Muddale (1998), while other scholars (Markey et al. 1980, Kamps 2004) consider that the controlled vocabularies improve the results precision at the time of retrieval. However, it is clear that many searchers recognize the advantages of either form of vocabulary according to circumstances (Arsenault 2006). Though it seems that the many difficulties associated with free indexing using uncontrolled vocabularies can only be understood via a comparative analysis with the controlled vocabulary indexing (Macgregor and McCulloch 2006).

Two valuable elements come out from this analysis of the situation of image indexing and retrieval. Firstly, we noticed that we know very few things about the influence of the vocabulary used for image indexing, at the time of retrieval in a monolingual context, and even less in a multilingual context. Secondly, we pointed to the great variability in the process of image retrieval. Indeed, for a few years, we distinguished an evolution in the manner of formulating the queries used to retrieve images. This transformation brings us to reconsider the way in which the image must be indexed and to wonder whether the controlled vocabulary traditionally employed for image indexing is still well adapted to this particular type of document. Therefore, the study of the Web image searcher's behaviours seems necessary, in order to conceive better retrieval systems. In the same way, it is as essential to sustain the study of image searchers retrieval strategies in order to lay down a

better indexing and more specifically, a vocabulary better adapted to the real and current needs and behaviours of the image searchers.

Taking into account what has just preceded, several questions are raised about image retrieval in a multilingual context. Thus, we can wonder whether the choice of the vocabulary used for image indexing will influence the retrieval results in a multilingual context and if such is the case, what this influence is. Moreover, we can speculate about which image indexing approach it is preferable to implement to facilitate the retrieval of images in a multilingual context. To our knowledge, few studies address these questions. Nevertheless, the image indexing process, allowing effective, efficient and satisfactory retrieval in a multilingual context, remains a delicate question which deserves more attention.

### 3. Research Objectives

This research project aims to verify the existing relations between each of these two indexing approaches: 1) traditional image indexing recommending the use of controlled vocabulary and 2) free image indexing using uncontrolled vocabulary, and their respective performance for image retrieval, in a multilingual context. In order to achieve this goal, three specific objectives were defined:

1. Identify the characteristics of each of the two approaches used for image indexing, in a multilingual context;
2. Explore the existence of relations between the characteristics of each of the two approaches used for image indexing and the retrieval performance, in a multilingual context; and,
3. Determine the direction of the relations between the characteristics of each of the two approaches used for image indexing and the retrieval performance, in a multilingual context.

The achievement of these three objectives will make it possible to empirically measure the influence of each of the indexing approaches in a multilingual image retrieval context. This study supposes that the two approaches of indexing show common characteristics, but also differences that can influence the image retrieval. Eventually, it will be possible to extend this research to other linguistic environments as well as other types of images. This research thus enables us to underline the essential elements of a well-adapted process to collections of digital images.

These collections offer unequalled informational richness as they represent essential elements of the collective memory and world inheritance.

#### 4. Methodology

This research compares image retrieval within two contexts: a monolingual context, i.e., where the language of the query (French) is the same as the indexing language (French); and a multilingual context, i.e., where the language of the query (French) is different from the indexing language (English). However, the monolingual retrieval context is only considered here in order to establish a comparison with the multilingual retrieval context. Besides, this research uses a quantitative approach. This category of approach is regularly used in Information Science to measure and compare the retrieval performance of textual and visual documents. The concepts are studied and tested using research hypotheses based on observable and clearly defined variables. Three data collection methods are used: 1) the analysis of the vocabularies used for image indexing in order to examine the multiplicity of term types applied to images (generic description, identification, and interpretation) and the degree of indexing difficulty due to the subject and the nature of the image; 2) the simulation of the retrieval process with a subset of images indexed according to each indexing approach studied, and finally, 3) the questionnaire to gather information on searcher satisfaction during and after the retrieval process.

A database of ordinary images including images drawn from a commercial online catalogue was prepared. A total of 3,950 images were downloaded in the database between December 26, 2006 and January 10, 2007. These images represent objects drawn from the eight sections of the catalogue: Sports and recreation, House and home, Auto, Tools, Kitchen and bath, Electronics, Outdoor living and Apparel. The image selection was mainly based on three specific criteria: 1) visual quality of the image illustrating a specific object, 2) ease of identification of the image by a non-specialist, and 3) maximum diversity of the images included in the database. Each image was then indexed in four different ways: with controlled vocabulary (French and English), and with uncontrolled vocabulary (French and English). For this study, French and English were chosen for two reasons. On the one hand, the French language is the language used by the sample of participants who will perform the retrieval tasks and on the other hand,

the majority of the controlled vocabularies designed for the indexing process only exist in English. Consequently, the inclusion of English as the second indexing language was considered mandatory.

The indexing process using French uncontrolled vocabulary was carried out by a French-speaking indexer, while the indexing using English uncontrolled vocabulary was carried out by an English-speaking indexer. The indexing with French and English controlled vocabulary was carried out by a bilingual indexer. The *Nouveau dictionnaire visuel multilingue* was selected as the controlled vocabulary. Three reasons justify this choice. Firstly, this dictionary contains the appropriate terms for the image type contained in the database, i.e. ordinary images of everyday-life objects. Secondly, the *Visuel* offers a form of standardization of the terms which allows a clear and precise identification of the objects, while exerting a maximum control on regionalisms, archaisms, Anglicisms, etc. Finally, the *Visuel* includes, in parallel, both the French and English terminologies which facilitated the indexing process, while ensuring a certain form of interlinguistic consistency. The indexing process with controlled vocabulary was completed using the paper dictionary and the CD-ROM version (*Le Visuel 3*). An example of the controlled vocabulary (CD-ROM version) is presented in Figure 1.

Once all the images were indexed with each one of the four indexing forms, all the indexing terms of the French controlled and uncontrolled vocabularies were combined, as well as the indexing terms of the English controlled and uncontrolled vocabularies. Therefore, we obtained two more indexing forms. The image database thus associates six indexing forms to each one of the 3,950 images included. Following the indexing process, a grid analysis for the indexing terms was developed. The objective of this analysis is to identify the specific characteristics of each indexing approach being studied in this project.

An image retrieval system was also built to interact with the database. This system tries to reproduce the image retrieval conditions we usually find in general Web search engines. This retrieval system includes: 1) a search interface (in French) allowing the image searcher to formulate a textual query (Figure 2); 2) a translation mechanism of the queries from French to English; 3) a mapping module for the query terms (or the query terms translation) and the indexing terms associated with the images; and 4) a display interface of the retrieval results where the user can select the images judged satisfying (Figure 3).

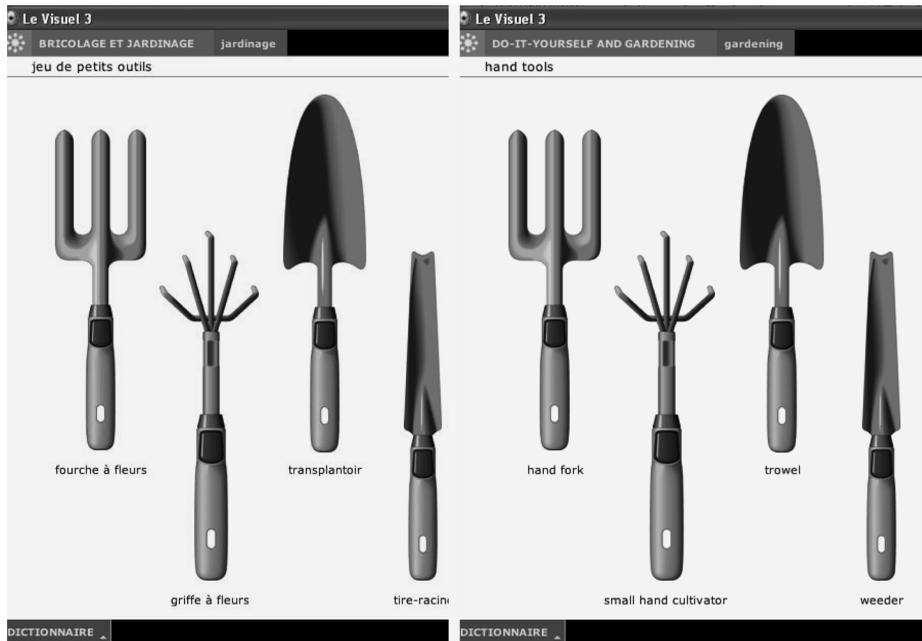


Figure 3. Example of the controlled vocabulary

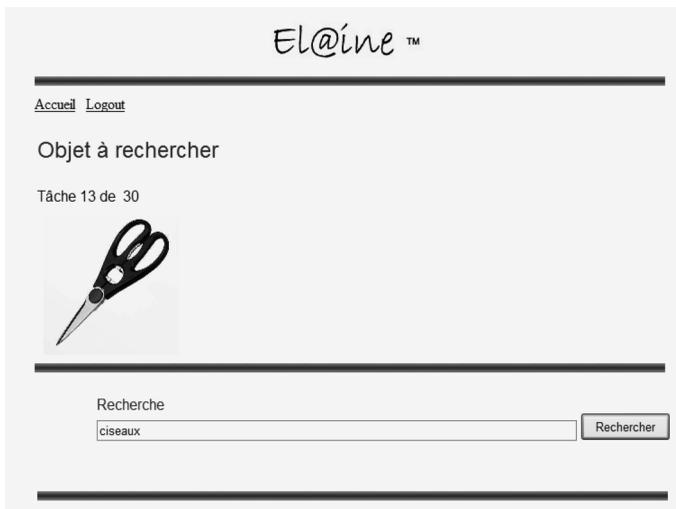


Figure 3. Search Interface

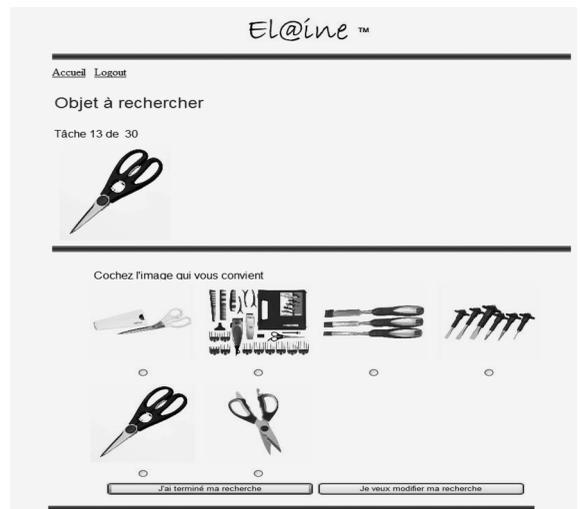


Figure 3. Display Interface

For the retrieval simulation, a subset of 30 images has been randomly selected from the image database and will be searched by each participant. The diversity of the images to be retrieved was also considered in order to neutralize the learning effect, and to privilege the selection of images presenting the same level of retrieval difficulty. During the retrieval task each of the 30 images is evenly associated at random with one of the 6 indexing forms (i.e., 5 images for each indexing form). These 30 permutations (image/indexing form) constitute the 30 retrieval tasks that each participant will perform during the retrieval simulation. Each image is successively shown

to the participant who tries to retrieve the image using a textual query (in French). All participants are native French speakers. Half of the images to be retrieved are mapped to one of the three French indexing forms (controlled, uncontrolled, and controlled + uncontrolled combined), while the queries for the other half of the images to be retrieved are first translated using the integrated machine translation mechanism, and then mapped to one of the three English indexing forms (controlled, uncontrolled, and controlled+uncontrolled combined), as shown in Figure 4.

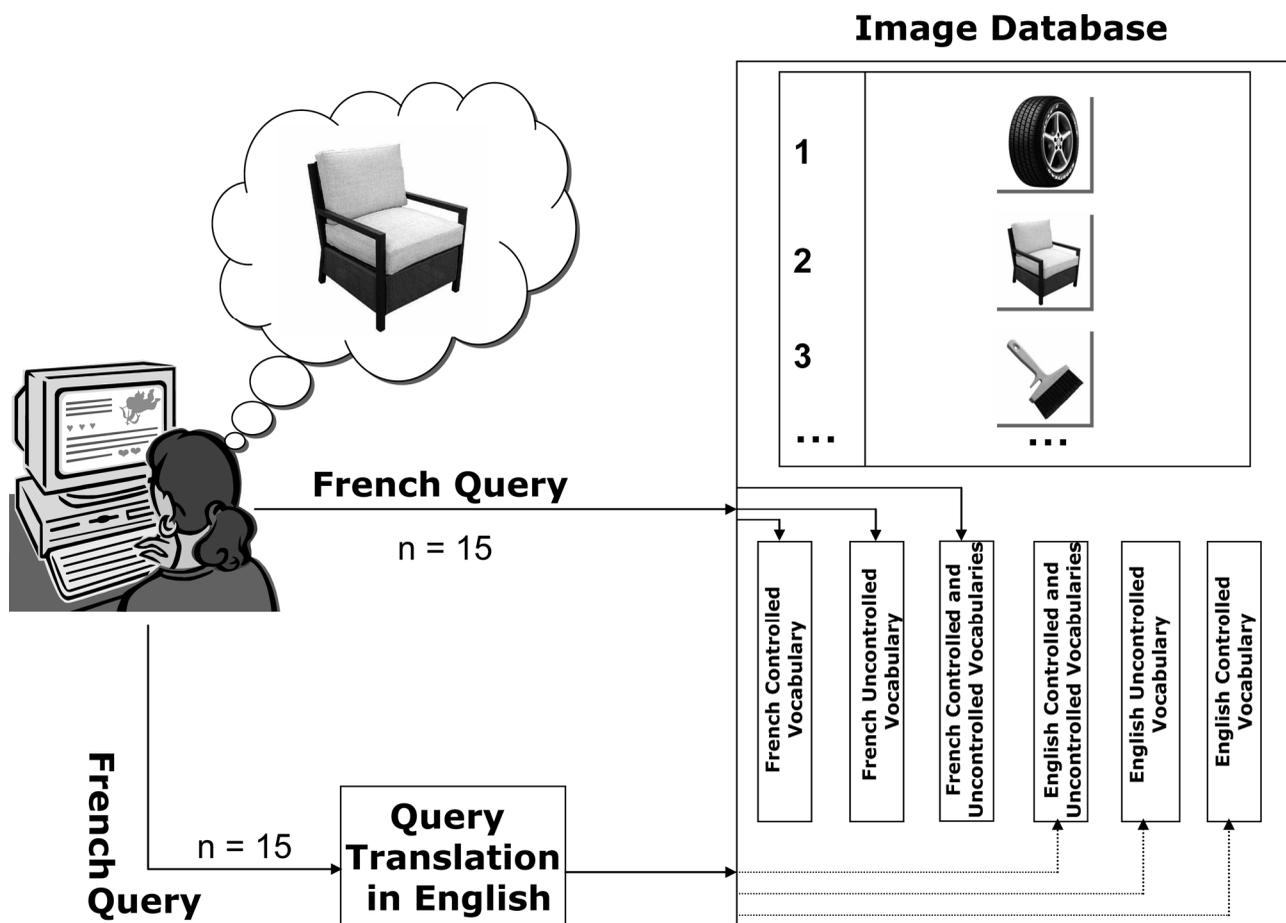


Figure 4. Schema of the Retrieval Simulation

The retrieved images are then displayed on the interface. The participant must then select the images judged satisfying. Each image deemed satisfactory by the participant will contribute to an estimate of the retrieval success rate. The participant can also submit another query or record his results. In order to reduce the learning effect, the order of the image retrieval tasks is defined randomly and modified for each participant. Each of the six combinations image/indexing forms is searched by a group of seven participants thus requiring 42 participants. The retrieval simulation will take place in the fall of 2007. The participants selection for the retrieval simulation will be done according to these 4 criteria: 1) 18 years and older; 2) native French speakers; 3) undergraduate students of the Université de Montréal; and 4) given the nature of the task to be carried during the retrieval experiment, the participants should not have any professional experience in a field using image retrieval.

Finally, in order to complete the data collected during the retrieval simulation, a questionnaire was

developed to collect the impressions of the participants concerning the retrieval process and the obtained results. This questionnaire is submitted to the participants in two stages. Initially, since we wish to measure the satisfaction for each indexing approach being studied, we ask the participants to evaluate their degree of satisfaction regarding the retrieval results they obtained after each image search task. Secondly, a comprehensive questionnaire including open and closed questions will be presented to the participants, at the end of the retrieval tasks. This questionnaire consists of questions about: 1) their general perception of the retrieval tasks; 2) their regular search practices and familiarity with image retrieval; and, 3) their demographic profile.

### 5. Observations

Before starting the indexing process with controlled vocabulary, the indexer received as a directive to use the indexing policy specially conceived for this research project and which recommends the use of the

*Nouveau dictionnaire visuel multilingue* for indexing the 3,950 images of the database. The indexing process with the French and English controlled vocabularies was carried out by the same indexer. However, in order to measure the level of interindexer consistency, a proportion of images (7 percent of the total number of images included in the database) were also indexed by a second indexer. The interindexer consistency can be defined as the capacity for several indexers to analyze the same document in the same way and to represent the same concepts by the same terms (Bertrand-Gastaldy 1986). The calculation of the interindexer consistency level is the relation, expressed with a percentage, between the total number of terms commonly used (C) by two indexers (A and B), multiplied by two, divided by the total number of terms chosen by these two indexers, according to the following formula:  $2C \div (A + B)$  (David 2003). For the French and English controlled vocabularies, we obtained an interindexer consistency rate of 73.1 percent. This rate is calculated with the identical indexing terms only, i.e. the terms commonly chosen by the two indexers. For example, “tennis racket” and “racket” are not considered as common indexing terms. We also calculated the terminological consistency level. In that case, “white plate” and “plate” are considered as common indexing terms since the main substantive is the same. We obtained a rate of 78.1 percent. Finally, we calculated the conceptual consistency level (same concept, different word for the main substantive) and we obtained 89.2 percent. These rates are the same for both English and French since the controlled vocabulary is bilingual. These rates are presented in Table 1.

A specific policy for the free indexing using uncontrolled vocabularies was also elaborated for this research. This policy provides that the indexer uses his own vocabulary to represent the concepts of each image contained in the database, without any form of control of the sense or the form. The English and French indexing using uncontrolled vocabulary was done independently by one indexer that had the required linguistic knowledge to achieve this task in the specified indexing language. Similar to the indexing process with controlled vocabulary and in order to measure the interindexer consistency level, a number of images (10.0 percent of the database) was indexed with uncontrolled vocabulary by a second indexer. The interindexer consistency rates (identical, terminological, and conceptual) were also calculated for both French and English uncontrolled vocabularies. The results are also presented in Table 1.

<i>Vocabulary</i>	<i>Identical Consistency %</i>	<i>Terminological Consistency %</i>	<i>Conceptual Consistency %</i>
<i>French Controlled Vocabulary</i>	73.1	78.1	89.2
<i>English Controlled Vocabulary</i>	73.1	78.1	89.2
<i>French Uncontrolled Vocabulary</i>	09.8	65.3	85.3
<i>English Uncontrolled Vocabulary</i>	11.0	61.4	82.9

Table 1. *Interindexer Consistency Rates*

The divergence between the interindexer consistency rates for controlled vocabulary and uncontrolled vocabulary duplicates the conclusions of the studies describing the use of controlled vocabulary which maintain that the interindexer consistency rate increases with the use of such a vocabulary, comparatively to the use of uncontrolled vocabulary (David 2003). The indexing term analysis process is currently in progress. Once completed, this analysis will enable us to characterize each indexing approach being studied. It will also highlight the similarities and the differences of the two indexing approaches that could influence the image retrieval in a multilingual context. The results of this analysis will be available in the near future.

The quantification of the retrieval performance of each indexing approach being studied is based on the measures usually employed to evaluate the performance of the retrieval systems (precision and recall), as well as the usability measures recommended by the standard ISO 9241-11, i.e. effectiveness and efficiency (AFNOR 1998). The third measure, satisfaction, is quantified using the third collection method used by this research project, the questionnaire. Generally, effectiveness refers to the capacity to achieve a given goal, whereas efficiency refers to the capacity to carry out a specific task with minimal effort. Traditionally, several indicators are taken into account for these two usability measures. For this research, the success rate, namely the capacity to achieve partially or completely the objective of the retrieval task, is considered to be the principal indicator of effectiveness. Concerning efficiency, a distinction was made between two specific forms: (1) human efficiency measured by effectiveness divided by human effort, i.e. the total number of queries

formulated for a given retrieval task; and (2) temporal efficiency measured by effectiveness divided by the time in seconds spent to carry out the selection of a satisfying result for a given retrieval task. For each retrieval task, the following elements are recorded directly by the retrieval system.:

- Query formulation: terms employed in each query formulated to retrieve a specific image;
- Retrieved images: identification number of the images selected by the participant for each retrieval task;
- Number of queries: total number of queries formulated to retrieve a specific image;
- Retrieval duration: time spent, in seconds, to perform the retrieval task, from the first to the last query, for each image to be retrieved;
- Time devoted to each retrieved image: ratio calculated using the retrieval duration of the retrieval divided by of the number of retrieved images.

The retrieval simulation will take place in the fall of 2007. The results of the data recorded analysis during the retrieval simulation will be available in the summer of 2008.

The main goal of the questionnaire is to collect information for the third usability measure recommended by the standard ISO 9241-11 (AFNOR 1998), i.e. the satisfaction of the image searcher. The questionnaire primarily aims to quantify the satisfaction measure. For that, three indicators are used:

- The degree of satisfaction regarding the results obtained: measured by a scale of the participant satisfaction about the obtained results for each retrieval task;
- The degree of learning facility: measured by a scale of the participant perception about the facility with which he succeeds in achieving each retrieval task;
- The degree of tolerance to errors: measured by a scale of the participant tolerance towards errors during each retrieval task.

The results of the questionnaire analysis will also be available in the summer of 2008.

## 6. Conclusion

The need to retrieve a particular image from a collection is shared by several user communities including teachers, artists, journalists, scientists, historians,

filmmakers and librarians, all over the world. Image collections also have many areas of application: commercial, scientific, educational, and cultural. Until recently, image collections were difficult to access due to limitations in dissemination and duplication procedures. The advent of the Web highlighted the pressing need to develop suitable tools for the description of digital images, since these can be found in the majority of available resources. The expected contributions of this research project are the following: 1) at the theoretical level, this research identifies the essential characteristics of the digital image indexing in order to facilitate its retrieval in a multilingual context; 2) this research also tries to bring a methodological contribution with an evaluation method of the preferable vocabulary to be used for image description; and finally, 3) this research reveals the essential elements of a tailored process for the collections of digital images. In short, our research underlines the pressing necessity to optimize the methods used for image processing, in order to facilitate the images' retrieval and their dissemination in multilingual environments. The results of this study will offer preliminary information to deepen our understanding of the influence of the vocabulary used in image indexing. In turn, these results can be used to enhance access to digital collections of visual material in multilingual environments.

## Acknowledgements

The author would like to thank the members of the Research Committee, Mr. Clément Arsenault, Ms. Christine Dufour and Mr. James Turner. The author would also like to acknowledge the hard work and dedication of the retrieval system programmer, Ms. Geneviève Bastien, and the members of the indexing team, Mr. Clément Arsenault, Ms. Manon Gagnon, Ms. Rita Maalouf and Ms. Chantal Ménard. A special thanks to Mr. Philippe St-Germain who took the time to read this article and correct my many mistakes. This research is graciously supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

## References

- Arsenault, Clément. 2006. L'utilisation des langages documentaires pour la recherche d'information. *Documentation et bibliothèques* 52: 139-48.
- AFNOR. 1998. *Exigences ergonomiques pour travail de bureau avec terminaux à écrans de visualisation*

- (TEV) – partie 11: lignes directrices relatives à l'utilisabilité. Genève: Organisation internationale de normalisation.
- Bertrand-Gastaldy, Suzanne. 1986. De quelques éléments à considérer avant de choisir un niveau d'analyse ou un langage documentaire. *Documentation et bibliothèques* janvier-juin: 3-23.
- Clough, Paul et al. 2006. *Multilingual interactive experiments with Flickr*. [http://www.sics.se/jussi/newtext/working\\_notes/11\\_cloughEtAl.pdf](http://www.sics.se/jussi/newtext/working_notes/11_cloughEtAl.pdf).
- Clough, Paul and Sanderson, Mark. 2003. *The CLEF 2003 cross language image retrieval task*. [http://clef.isti.cnr.it/2003/WN\\_web/45.pdf](http://clef.isti.cnr.it/2003/WN_web/45.pdf).
- David, Claire. 2003. *Analyse cognitive de la cohérence interindexeurs lors de l'indexation de documents*. Montréal: Université de Montréal.
- Furnas, George W. et al. 1987. The vocabulary problem in human-system communication. *Communications of the ACM* 30: 964-71.
- Guinchat, Claire and Menou, Michel. 1990. *Sciences et techniques de l'information et de la documentation*. Paris: Unesco.
- Harpring, Patricia. 2002. The language of images: enhancing access to images by applying metadata schemas and structured vocabularies. In Baca, Murtha ed., *Introduction to art image access issues, tools, standards, strategies*. Los Angeles: Getty Research Institute, pp.20-39.
- Hudon, Michèle. 2007. *Notes de cours BLT - 6055 Analyse et représentation documentaires 2*. Montréal: Université de Montréal.
- Jørgensen, Corinne. 2003. *Image Retrieval Theory and Research*. Lanham, Md: Scarecrow.
- Kamps, Jaap. 2004. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In McDonald, Sharon and Tait, John eds., *Advances in Information retrieval: 26<sup>th</sup> European Conference on IR research, ECIR 2004*. Berlin: Springer, pp. 283-95.
- Macgregor, George and McCulloch, Emma. 2006. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55: 291-300.
- Markey, Karen et al. 1980. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review* 4: 225-36.
- Nie, Jian-Yun. 2006. *Notes de cours IFT 6255 – Présentation sur SMART*. <http://www.iro.umontreal.ca/~nie/IFT6255/CLIR.html>.
- QA International. 2003. *Le Nouveau dictionnaire visuel multilingue*. Montréal: Québec Amérique.
- QA International. 2006. *Le Visuel 3 multimédia (V3.5)*. Montréal: Québec Amérique.
- Rao Muddamalle, Manikya. 1998. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. *Journal of the American Society for Information Science and Technology* 49: 881-87.
- Savoy, Jacques. 2005. Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing & Management* 41: 873-90.
- Smiraglia, Richard P. 2005. Introducing metadata. *Cataloging & Classification Quarterly* 40: 1-15.
- Spink, Amanda and Jansen, Bernard J. 2004. *Web search: public searching of the Web*. Boston: Kluwer Academic.
- Turner, James. 1998. *Images en mouvement : stockage, repérage, indexation*. Sainte-Foy: Presses de l'Université du Québec.
- Yahoo! Inc. 2007. *Flickr!*. <http://www.flickr.com/>.