Indexing and Retrieving Images in a Multilingual World
Elaine Ménard

Introduction
This communication presents the problem statement, the methodology and the preliminary results of a research project aiming to compare two different approaches for indexing images, namely: traditional image indexing with the use of controlled vocabularies, or free image indexing using uncontrolled vocabulary. The experiment intends to measure their respective performance for image retrieval, in a multilingual context.

Problem Statement
Images have always been a powerful communication tool. They present a multifaceted and actual value. They serve not only as a source of information but, with the development of more accurate visualization techniques, they also enhance the understanding of that information. For a few years, the diffusion of images has increased, mainly because of the development of digital technologies and the unprecedented growth of the Web. Confronted to this profusion of images, individuals now speculate on how to retrieve images with efficacy and efficiency. In general, two categories of queries are used to retrieve images on the Web: graphic or textual queries. In the first category, the individual submits a graphic 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 color, shape or texture. However, since Content Based-Image Retrieval Systems (CBIR systems, i.e. systems based on images’ physical characteristics) suffer from many caveats, the majority of image searches on the Web still use textual queries and the retrieval’s success depends on the match between the query terms and the text associated (ancillary text or indexing terms) with the images. 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 the images. Since images do not always include a caption or any kind of ancillary text, the indexing process remains crucial. Image indexing has, so far, been divided between two camps: the proponents of “controlled vocabulary”, and the proponents of “uncontrolled vocabulary”. The former method focuses on indexing terms extracted from thesauri, classification schemes or subject heading lists, while the latter focuses on terms drawn from the natural language.

Consequently, to retrieve an image the individual first has to transpose in words what he is looking for. This is the first challenge. The second obstacle comes from the “language” of the image. By their very nature, images are considered to be language-independent resources. Nevertheless, the text associated with the images gives the image a linguistic status similar to any other textual document, which can significantly affect its retrieval. And given the great linguistic diversity existing on the Web, we must expect that the text associated with the images exists in many different languages. For example, if a user formulates a query in English and the images to be retrieved are associated with English text, the cross-lingual problem does not arise. However, if an English query is used and the associated text is in Italian (or any other language different from the query language), the retrieval will not be possible, unless the retrieval system includes a Cross Language Information Retrieval (CLIR) mechanism which allows cross-language mapping between the query terms and the associated text.
**Research Objectives**

For this research, three specific objectives were defined: identifying the characteristics of each of the two approaches used for image indexing, in a multilingual context; exploring 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 finally, determining 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 context for retrieving images. This study supposes that the two approaches of indexing show common characteristics, but also differences that can influence the image retrieval.

**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. For this research, three data collection methods are used.

Firstly, an analysis of the vocabularies used for image indexing is employed 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. The objective of this analysis is to identify the specific characteristics of each indexing approach being studied in this project. This analysis is conducted on the indexing terms assigned to the 3,950 ordinary images drawn from a commercial online catalogue and included in the database images prepared for this research. Each image was indexed in four different ways: with controlled vocabulary (French and English), and with uncontrolled vocabulary (French and English). Following the indexing process, the assigned indexing terms have been examined. A grid analysis was developed and applied on all the indexing terms assigned to the 3,950 images of the database. Three levels of analysis were determined: terminological, perceptual and interpretative.

Secondly, a simulation of the retrieval process involving a subset of images, indexed according to each indexing approach studied, is performed with 60 participants. For the retrieval simulation, 30 images have been randomly selected from the image database and are searched by each participant. 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 the combination of controlled and 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 the combination of controlled and uncontrolled combined). The quantification
of the retrieval performance of each indexing approach is based on the usability measures recommended by the standard ISO 9241-11, i.e. effectiveness, efficiency, and satisfaction of the user.

Finally, 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 order to gather information on searcher satisfaction during and after the retrieval process and to complement the data collected during the retrieval simulation. This questionnaire is submitted to the participants in two stages. Initially, since we wish to measure the satisfaction for each indexing approach being studied, the participants are asked to evaluate their degree of satisfaction regarding the retrieval results they obtained after each image searched. The question “satisfaction” introduced at this stage of the simulation makes it possible to collect spontaneously the impressions of the participants on the retrieval task which was just performed. Secondly, a comprehensive questionnaire including open and closed questions is presented to the participants, at the end of the retrieval tasks. This questionnaire consists of questions about their general perception of the retrieval tasks, their regular search practices and familiarity with image retrieval and their demographic profile. The preliminary results of the analysis of the indexing terms, the data recorded during the retrieval simulation and with the questionnaire are presented.

**Contribution**

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. This research underlines the 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. 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. In short, this research emphasizes the pressing need to optimize the methods used for image processing, in order to facilitate their retrieval and their dissemination in a multilingual context.