Using Ancillary Text to Index Web-based Multimedia Objects

Lyne Da Sylva
lyne.da.sylva@umontreal.ca

James M Turner
james.turner@umontreal.ca

École de bibliothéconomie et des sciences de l'information, Université de Montréal

1. Introduction

PériCulture is the name of a research project at the Université de Montréal which is part of a larger parent project based at the Université de Sherbrooke. The parent project aimed to form a research network for managing Canadian digital cultural content. The project was financed by Canadian Heritage and was conducted during the fiscal year 2003-2004. PériCulture takes its name from péritexte and culture, péritexte being one of a number of terms used (in French, our working language) to mean ancillary text associated with images and sound. It is a sister project to DigiCulture, another part of the same larger research project which studied user behaviours in interactions with Canadian digital cultural content. The general research objective of PériCulture was to study indexing methods for web-based non-textual cultural content, specifically still images, video, and sound. Specific objectives included:

1. identifying properties of ancillary text useful for indexing;
2. comparing various combinations of these properties in terms of performance in retrieval;
3. contributing to the development of bilingual and multilingual searching environments;
4. developing retrieval strategies using ancillary text and synonyms of useful terms found therein.

Our work in the context of this project focusses on text associated with web-based still images, and builds on previous work in this area of information science (e.g. Goodrum and Spink 2001; Jörgensen 1995, 1998; Jörgensen *et al.* 2001; Turner and Hudon 2002). We identified a number of web sites that met our criteria, i.e., that contained multimedia objects, that had text associated with these objects that was broader than file names and captions, that were bilingual (English and French), and that housed Canadian digital cultural

content. We identified keywords that were useful in indexing and studied their proximity to the object described. We looked at indexing information contained in the <meta> tag and the "alt" attribute of the <img> tag, and whether other tags contained useful indexing terms. We studied whether standards such as the Dublin Core were used. We identified web-based resources for gathering synonyms for the keywords.

2. Background and Context

In computer science, research into indexing images and sound focusses on the low-level approach, performing statistical manipulations on primitives in order to identify semantic content (e.g. Alvarez *et al.* 2004). This approach is also referred to as the content-based approach (e.g. Gupta and Jain 1997, Lew 2001). In information science, research into indexing images and sound focusses on associating textual information with the non-textual elements, and this often involves manipulating ancillary text. This approach is referred to as the high-level or concept-based approach (e.g. Rasmussen 1997, O'Connor, O'Connor, and Abbas 1999).

Although human indexing remains necessary in a variety of situations, a number of factors militate in favour of automating the high-level approach as much as possible. These include the very large volume of web-based materials available and the high cost of human indexing, which in any case is relatively inconsistent. The disparity among cataloguing and indexing methods from one collection of images to another was not a problem until recently, because collections were self-contained, and users learned the indexing languages necessary to reach the information available in the collections they use. Now, however, there is a strong desire to connect repositories worldwide and to provide interfaces which allow users to search multiple collections using a single search strategy. The usefulness of permitting multilingual searches of these same online collections is clear; however, much work remains to be done in setting up the organisational infrastructure of collections to permit this.

Lyne Da Sylva's work in this connection has dealt mainly with automatic indexing of text documents. Her prototype system for automatic indexing (Da Sylva 2004, 2005) constructs a back-of-the-book index for digital documents. It relies extensively on insights from human methodology for indexing books as well as on properties of index entries and semantic relationships between headings in the index. Several techniques for spotting and handling linguistic cues in text documents are directly applicable to the indexing of non-textual elements, given appropriate ancillary text.

James Turner has worked for a number of years on the general problem of using ancillary text to index still and moving images, with emphasis on shot-by-shot indexing of moving images. This work has shown that approaches to indexing images are rather different from those used for indexing text documents. People asked to describe non-art images name the objects seen in the images, as well as persons and events. Of all the names an object might have, only one or a few names are given very often. Furthermore, no significant differences in how the objects are described by various groups such as students and workers or visually-oriented and non-visually-oriented persons were found (Turner 1994). The transfer from visually perceiving the image and naming the objects found in the image seems rather direct (although this is not necessarily true in the case of art images), whereas text indexing requires more strenuous cognitive and intellectual work in interpreting the meaning of the text into indexing terms useful in searching. In studies looking at cross-language differences between English and French, no cultural differences were found in how Canadian English-speakers and French-speakers describe images (Turner and Roulier 1999). In a study of automated translation between English and French of indexing terms for moving images, eight web-based translators performed very well, with success rates between 71% and 89%, in taking the indexing terms in either language and putting them correctly into the other (Turner and Hudon 2002).

The research results reported here build on this work by studying other aspects of text associated with images in a networked environment to try to gain some understanding of how the ancillary text associated with images on web pages can be exploited to index the corresponding images. We studied this question in the context of web sites that were Canadian, bilingual (English and French), and with patrimonial multimedia content.

3. Methodology

3.1 Constituting the Corpus

In order to conduct our study, we first constituted a corpus of web sites that responded to these criteria. To build this corpus, we did an initial web search on keywords such as “museum”, “Canadian”, and “cultural”, in order to make a list of Canadian museum and government cultural sites. Interesting links found on the pages returned as results were followed, and in the end we reviewed forty (40) sites in order to constitute our corpus.

We then made a preliminary evaluation of the sites in the corpus, assigning to each site a score calculated using a number of criteria. Some of the criteria increased the score, and others decreased it. Criteria such as the presence of both official languages, cultural content, images, video or sound increased the score, as did the use of Dublin Core metadata, HTML, and use of the <meta> tag in HTML. Use of frames, Flash elements, tables, or ASP code decreased the score. Other elements considered were the approximate number of words of text on the page, and the ratio between the number of multimedia elements present and the number of words of accompanying text. In this way, we were able to calculate a score for each site. As a result of this exercise, twenty-five (25) sites were retained for further study, each containing between one and thirty-five (35) pages. The corpus is thus made up of 118 selected web pages from the chosen sites. The pages do not necessarily represent all the multimedia objects available at these sites. For example, in the case of the Musée virtuel canadien/Virtual Museum Canada site, the web pages we examined do not include all those containing paintings; since a number of criteria were used to score each page, not all pages with useful elements to study made it into our study sample, only those with the highest scores were selected.

3.2 Types of Ancillary Text Considered

Web sites are often configured to generate pages dynamically from information in a database in response to a query. In cultural institutions, these collections are often indexed manually (i.e., by humans), to varying degrees of exhaustivity but to the degree that is deemed appropriate by the collection managers or that is possible as a function of the resources available. In the context of the present study, this indexing is not of primary interest to us since it is purposeful, specific to the multimedia objects indexed, and takes into consideration the needs of users of the collection. Here we are interested rather in ancillary text, that which is found in association with multimedia objects but that was created for purposes other than that of indexing the associated objects.

Following are the types of ancillary text we studied:

- the filename of the multimedia object, i.e. the value of the "name" attribute of the <img> tag. File extensions include gif (for image files), mp3, wmf, wav or aif (for audio files) and mpg, mpeg, wmv, qtv (for video files);
- the value contained in the "alt" attribute of the <img> tag, which offers a textual description of the image when the visual display is hindered in some way;

- text from a hyperlink which points to the multimedia object, and more specifically, the text between the anchor tags (this text);
- a legend or other label for the multimedia object;
- the value of metadata elements associated with the HTML file, for example:

  <meta name="Author" content="Gouvernement du canada, Archives nationales du Canada">
  <meta name="Originator" content="Gouvernement du canada, Archives nationales">
  <meta name="Keywords" content="Pavillons, Pavillions, Expo 1967">
- the title of the page as indicated within the <title> tags in HTML;
- headers in the page, as expressed with the tags <h1>, <h2>, and so on;
- text contained in the paragraphs immediately preceding or following a multimedia object.

For each HTML page of the sites we selected, we made a detailed analysis which consisted of identifying all types of ancillary text present and determining whether each text element (e.g. the full title, or the text contained in the "alt" attribute) was useful for describing the multimedia objects on the page. For example, given a photograph of a tepee, it was determined whether expressions such as "typical native lodgings" were useful for indexing. For paragraphs of full text surrounding the object, we counted the number of words in the paragraph, as well as the number of useful indexing terms the paragraphed contained. From this, we calculated a ratio of useful keywords to the number of words in a paragraph. For each case, the research assistants made a judgement call as to whether the term could be considered useful for indexing or not, based on whether the term seemed to be descriptive of the object or not.

3.3 Locating and Making an Inventory of Multimedia Objects and Their Associated Indexing Terms

Locating multimedia objects and identifying their relative position to potentially useful ancillary text were complex tasks, since tables or frames (or other features of HTML mark-up) often blurred what seemed obvious in the resulting page. These tasks were necessary, however, especially as regards paragraphs immediately preceding or following multimedia objects, since these were a primary object of study. To try to maintain some uniformity in the data, the positions of multimedia objects are given by line number (in the original HTML file) and not in terms of the number of intervening words. In cases where the multimedia objects and the ancillary text were separated by JavaScript code rather than HTML code, the number of intervening lines turns out to be smaller than the number of words, which reduces unwarranted discrepancies, although some

do remain. To facilitate processing, line numbers were calculated approximately using the line-numbering function in Word.

We identified images by searching for <img> elements in the HTML code. Where they were found, the corresponding URL was visited to determine whether the page also contained words that were useful for indexing. Other types of multimedia objects were found by manual inspection of the HTML code, since they are generally referred to by a hyperlink. Distinction was made between the following types of multimedia objects: image, sound, video, and link. Links are not multimedia objects, of course, but they represent them in the HTML page. The user may be presented with a QuickTime icon, for example, to represent the multimedia object. In our study, both true multimedia objects and links to them were considered as targets to be indexed. We limited the multimedia objects in our study to two still images. The rationale for this is discussed in the results section.

For each HTML page of the selected sites, we located all types of ancillary text associated with the images (e.g. the text between the <title> tags, the text contained as a value of the "alt" attribute, and so on). Next, we determined whether the text was useful for describing the multimedia objects. Often, the "alt" attribute consisted of a descriptive legend such as "The Honourable Liza Frulla". Sometimes the title of the page was a very general term which could also be used to describe the multimedia objects, such as "Constructing the Canadian Pacific Railroad". Other times, the text was not useful for our purposes, either because it had no meaning (e.g. file names consisting of alpha-numeric sequences such as "TS00254.jpg") or because the ancillary text was ambiguous or non-descriptive. Some multimedia objects could not be considered cultural content but rather were graphic elements used for navigation or identification, such as arrows or labels. Company logos were considered a special case. Although they may be considered multimedia objects representing Canadian entities, they were discarded from the analysis because they often appear on every page of the web site, regardless of the theme of the page. We felt that including them would introduce too much noise in the analysis of multimedia objects.

As we mentioned, not all multimedia objects we considered are visible on the HTML page as it is displayed. Some are only accessible via a click, so we identified each multimedia object as being visible or not. This was especially relevant in the case of images that were not visible but that had an intermediary visible counterpart in the form of a thumbnail of the image. Although the thumbnail is not the image itself as such, we con-

sidered it to be the multimedia object, since it is does appear on the page and since ancillary text can describe both the target image that requires a click to be displayed and the corresponding thumbnail. These methodological acrobatics allowed us to obtain a useful measure of the distance between the multimedia object and the ancillary text which describes it. Links pointing to video clips were treated the same way.

For each multimedia object, we extracted from the corresponding HTML page a list of words that describe it, by reading the entire page. The "Find" function in Word was then used to group occurrences of the words from this list. Finally, we produced a list of pairs, each pair consisting of the candidate indexing term and the type of ancillary text in which the term occurred.

4. Results

Each page of each site was examined to locate the multimedia objects and to determine whether ancillary text was present and if so, whether it could be considered useful for indexing. Ancillary text is not always present. A number of conditions were identified: the title of the page is sometimes left unspecified, a given <img> tag may contain no value for the "alt" attribute, and so on.

Table 1 summarises the results of the pages we examined. The data reveal a number of interesting phenomena which we present in this section.

Table 1. Types of ancillary text for web pages in the corpus with indications of whether types present were considered useful

Types	Present/total	Percentage	Useful/total present	Percentage
File name	109/117	93.2	14/109	12.8
"alt" attribute	77/117	65.8	36/77	46.8
Hypertext	39/117	33,3	39/39	94.9
Legend/label/caption	24/113	21.2	24/24	100.0
<meta> elements	31/116	26.7	31/31	96.8
Title of page	116/117	99.1	94/116	81.0
Paragraphs	116/119	97.5	113/116	97.4

File names for multimedia objects are almost always present (93.2% of the time). However, since they are considered useful (i.e. descriptive of the corresponding object) only 12.8% of the time, their role as indexing terms is rather limited.

In approximately two-thirds of the cases (65.8%) there was a value assigned to the "alt" attribute of the <img> tag. However, it was considered useful less than half the time (46,8%).

Names found in hypertext links are present only one-third of the time (33.3%). However, when they are present they are considered useful almost all the time (94.9%).

Text in legends, labels, or captions is also relatively rare in the data we examined, occurring only 21.2% of the time. However, when such text is present, it is a very reliable indicator of the content (100% of the cases were deemed useful). This is not surprising, since the purpose of such text is to identify information in the picture. In our study his category of data was identified by human observation, but a way of identifying it automatically needs to be found, otherwise the data would have to be distributed among the other categories. The explicit nature of this kind of text (e.g. a legend created specifically to give descriptive information about the corresponding image) militates in favour of finding some way to help an algorithm identify it as such, even if this means developing a special tag for it, for example.

Similarly, data is included in the <meta> tag only about one-quarter of the time (26.7%), but when it is present it is almost always useful (96.8% of the time in the data we analysed). Again, the usefulness of such text is not surprising, since the tag exists to facilitate adding searchable terms for describing the content. The low rate of use of the <meta> tag is not surprising either, since widespread abuse of the tag has caused almost all popular search engines to disregard it (Sullivan 2002).

Titles of HTML pages as expressed in the <title> tag are virtually always present (99.1% of the time). In addition, the text contained in these titles is very reliable for indexing purposes. In our data, the text of titles was deemed useful 81% of the time. However, in almost all cases the words were classed as "general" so that their usefulness in indexing the specific content of images is limited. For example, the title "Canadian Pacific Railway" gives some helpful general information but does not describe an image of the inside of a waiting room in a rural train station that might be found on the page. Another example, in which text is only vaguely useful or even outright misleading is a page entitled "Our roots" and containing a photograph entitled "Along

the Fifth" (i.e. Fifth Avenue). While the connection as general indexing can be made without much trouble, it is clear that users searching for images of roots will be disappointed if they land on this page.

Paragraphs immediately preceding or following multimedia objects appear to be the most promising sources of useful indexing text in the data we studied. Text adjacent to multimedia objects was both very frequent (it occurred 97.5% of the time) and very useful (97.4% of the time there were indexing terms considered useful). However, this figure for usefulness requires some explanation. The length of paragraphs (identified by the <p> tag for our purposes) varied greatly. In addition, we allowed untagged paragraphs to be counted in our analysis. Thus in some cases a very large paragraph containing only one or a few useful indexing terms is considered useful.

In Table 2 we try to account more precisely for these variations, since we believe they warrant closer analysis. Here we seek to determine the distance from the multimedia object at which useful text can still be found, and whether the text preceding the multimedia object is more useful than the text following it. To determine this, we examined closely the size of paragraphs preceding and following images in terms of the number of words they contained and how many of these can be considered keywords useful for indexing.

Table 2. Mean and standard deviation for number of words and number of useful keywords

	Paragraph before image 1	Paragraph after image 1	Paragraph before image 2	Paragraph after image 2
No. web pages in sample	116	109	22	22
Mean no. words in paragraph	117.31	284.43	133.68	124.33
Standard deviation	185.47	553.50	144.23	89.74
Mean no. keywords in parag.	8.89	18.66	6.65	11.46
Standard Deviation	15.43	39.45	5.20	11.05
Percentage useful keywords	7,58	6.56	4.98	9.22
Ratio useful keywords	1:13	1:15	1:20	1:11

In order to make the task manageable, we limited the analysis to the first two images in each web page. This strategy also allows us to compare the behaviour of initial images with that of images in the main body of a web page. The number of images in each web page varied, of course, and since not all pages contained more than two, this approach allowed us to work with a more uniform pool of data.

Our sample of 116 pages represents those that included at least one image (i.e., two pages out of the 118 had no images), and twenty-two (22) of these pages included at least one more image. We did not remove stop words in this analysis because of limited resources available for processing. However, it would be helpful to compare the results against another analysis in which stop words are removed, as the figures we obtained would probably change considerably.

Three observations are in order here. First, the mean number of words in the paragraphs preceding image 1 (117.31 words) is smaller than that of paragraphs following it (284.43 words). This is true despite the fact that if the text following image 1 was the same text as that preceding image 2, it was only counted as the latter, and thus the text following image 1 was set to zero (the same was done for image 2, in cases where there existed a third image). Figures for image 2 show the inverse pattern, although the difference in the figures is not as great: on average, 133.68 words precede the image and 124.33 words follow it. We speculate that this may reflect properties of the first image, often situated near the beginning of a page, with most of the page's text following it, while the properties of other images (image 2, in our analysis) are less predictable. An alternate explanation is that the pages we analysed were relatively short, so that if a second image were present, there remained little text after it. Interestingly, in our sample the length of text before each image was comparable (between 117 and 133 words). This corresponds to roughly a paragraph of 12 lines on letter-sized paper. However, this comparable length does not correlate with comparable usefulness, as we shall see presently.

The second observation has to do with the percentage of words useful for indexing. This varies greatly within adjacent paragraphs. For the first image, 7.58% of words preceding it are useful for indexing, compared to 6.56% in the paragraphs following it. It is reasonable to expect initial paragraphs to be more informative; however, it is counterintuitive when one considers the large size (284.43 words on average) of the following paragraph, which then seems wordy without being descriptive of the image preceding. For the second image, however, the percentages are reversed. Only 4.98% of words preceding image 2 were deemed relevant

for indexing (despite the comparable size of preceding paragraphs for each image), while 9.22% of words in the following paragraph were considered useful. This may have to do with some type of conclusive matter in the paragraphs following image 2, which in most cases is the last of the page.

The third observation we wish to make is that the standard deviation for the number of words in the paragraphs is quite high, ranging from three quarters of the mean (89.74 for a mean of 124.33) to almost double (553.50 for a mean of 284.43). This indicates great variation in the number of words the paragraphs contain. For the number of useful keywords, there is strong variation in the paragraph following the first image (39.45 standard deviation for a mean of 18.66), but it is smaller in the paragraph preceding it (15.43 standard deviation for a mean of 8.89), and in either paragraph adjacent to image 2 (5.20 standard deviation for a mean of 6.65 and 11.05 standard deviation for a mean of 11.46 respectively). This needs to be taken into account when building algorithms for seeking out keywords.

5. Discussion

As we have seen, some of the elements of ancillary text studied are very useful as sources of indexing terms, and others not very useful at all. All types make some contribution to the pool of indexing terms that can be derived from ancillary text.

Not surprisingly, the most useful elements are those that are designed to hold descriptive content, such as the <meta> tag and legends or their equivalent. The least useful is the name of the file, found to be useful only 12.8% of the time. The profile of some of elements is that they are not always present, but when they are, they are very useful.

The "alt" attribute of the image tag allows creators of web pages to include some description of an image for the benefit of users who cannot see the image for any of a number of reasons. As we noted, in our data there was a value assigned to this tag 65.8% of the time, and when there was a value it was found to be useful only 46.8% of the time. This data suggests that the "alt" attribute is underexploited, an observation that has been made in other contexts, and which further suggests that those responsible for creating web page content should be made aware of the potential benefits of using of this attribute.

Perhaps the richest source of potential keywords for indexing is the text of the paragraphs surrounding the images on web pages, because these contain the greatest number of words, although only a percentage of

these words are useful for indexing. As we noted, we did not remove the stop-words in this study, but once this is done the percentages and ratios of useful keywords would improve considerably. We made a number of casual observations about the properties of the text of paragraphs, and these observations might now be formulated more rigourously as research questions to help get a better understanding of the nature of this text.

We notice that the first image of the web pages we studied occurred rather near the beginning of the page, so that the text surrounding the first image may well have different properties from those of the text surrounding additional images on the page. We also observe that the paragraphs following the second image seem to be the most informative, followed by those preceding the first image, although variations were considerable in our data. We also observe that the paragraphs preceding the second image are less useful for indexing than those preceding the first image, for a comparable number of words. Finally, paragraphs following the first image are typically quite large but not very useful for indexing.

6. Conclusions

Our study found that a large number of useful indexing terms are available in the ancillary text of web sites with cultural content. We evaluated various types of ancillary text as to their usefulness in retrieval. Our results suggest that these terms can be manipulated in a number of ways in automated retrieval systems to improve search results.

Cross-language comparison of the results reinforces our previous research results, which suggest that indexing in other languages can be generated automatically from a single language using web-based tools.

Rich information that can be used for retrieval is available in many places on web sites with cultural content, from the file name to explicit information in captions to descriptive information in surrounding text to the contents of various HTML tags. Algorithms need to be developed to exploit this information in order to improve retrieval.

Some of our previous work on how people assign indexing terms to images suggests that noun phrases are probably the most useful indexing constructs of all. Including in the search algorithm some kind of parser that could identify noun phrases would undoubtedly be helpful.

Building further on previous and present results, indexing terms that have been identified as such could then be filtered through a bilingual dictionary in order to provide indexing in the other language. This princi-

ple can probably be extended to create additional indexing in other languages; however, the universal feasibility of this principle needs to be demonstrated empirically.

A further step that can be undertaken to improve performance is to filter the indexing terms through an online thesaurus, in order to pick up synonyms and hierarchically-related words. For example, it would be helpful to be able to manipulate specific indexing terms such as "sparrow" so that users searching for images of any birds could also find these.

For ancillary text which is sometimes but not always useful, such as the "alt" attribute of the <img> tag, one possible direction for research would be to analyse words to estimate their usefulness in a given page: for example, does the word appear in a thesaurus of the domain? Does it confirm or is it compatible with information already present in other ancillary text? Such analysis would require external terminological resources.

It is clear that much more can be done to improve the performance of search algorithms for finding multimedia objects in a networked environment. We hope the results from this study make some contribution, even a modest one, to solving this problem. Enough knowledge has been gained to assure us that investing more effort in the area of exploiting ancillary text for indexing web-based multimedia objects is an investment that will surely pay off.

Acknowledgements

We thank Canadian Heritage for funding this project with a grant received via CoRIMedia, a research consortium based at the Université de Sherbrooke which focusses on access to multimedia information. We also thank our research assistants Nawel Nassr and Stéphane Boivin for their contribution to this work.

References and Bibliography

Alvarez, C., Oumohmed, A. I., Mignotte, M., and Nie, J.-Y. (2004). Toward cross-language and cross-media image retrieval. In *Working Notes for the CLEF 2004 Workshop,* 15-17 September, Bath, UK.

Da Sylva, L., and Doll, F. (2005). A document browsing tool: using lexical classes to convey information. In Lapalme, G. and Kégl, B. (eds), *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005 (Proceedings)*. New York: Springer-Verlag, *pp.* 307-318.

Da Sylva, L. (2004). Relations sémantiques pour l'indexation automatique: définition d'objectifs pour la détection automatique, *Document numérique, Numéro spécial « Fouille de textes et organisation de documents »*, 8(3): 135-155.

Goodrum, A. and Spink A. (2001). Image searching on the Excite web search engine, *Information Processing and Management*, 27(2): 295-312.

Gupta, A. and Jain, R. C. (1997). Visual information retrieval, *Communications of the ACM*, 40(5): 71-79.

Jörgensen, C. (1995). *Image attributes: an investigation*. PhD thesis, Syracuse University.

Jörgensen, C. (1998). Image attributes in describing tasks: an investigation, *Information Processing and Management*, 34(2/3): 161-174.

Jörgensen, C., Jaimes, A., Benitez, A. B., and Chang, S.-F. (2001). A conceptual framework and empirical research for classifying visual descriptors, *Journal of the American Society for Information Science and Technology (JASIST)*, 52(11): 938-947.

Lew, M. S. (2001). *Principles of visual information retrieval.* New York: Springer.

Marsh, E. E., and White, M. D. (2003). A taxonomy of relationships between images and text, *Journal of Documentation*, 59(6): 647-672.

O'Connor, B. C., O'Connor, M. K., and Abbas, J. M. (1999). User reactions as access mechanism: an exploration based upon captions for images, *Journal of the American Society for Information Science*, 50(8): 681-697.

Rasmussen, E. M. (1997). Indexing images. In Williams, M. E. (ed.), *Annual Review of Information Science and Technology*, 32. Medford, NJ: Learned Information, pp. 169-196.

Sullivan, D. (2002). Death of a meta tag, *SearchEngineWatch*, http://searchenginewatch.com/sereport/article.php/2165061 (first published October 1, 2002; last accessed 13 January 2006).

Turner, J. M. and Hudon, M. (2002). Multilingual metadata for moving image databases: preliminary results. In Howarth, L.C., Cronin, C., Slawek, A. T. (eds), *L'avancement du savoir : élargir les horizons des sciences*

de l'information, Travaux du 30e congrès annuel de l'Association canadienne des sciences de l'information, Toronto: Faculty of Information Studies, pp. 34-45.

Turner, J. M. and Roulier, J.-F. (1999). La description d'images fixes et en mouvement par deux groupes linguistiques, anglophone et francophone, au Québec, *Documentation et bibliothèques*, 45(1): 17–22.

Turner, J. (1994). *Determining the subject content of still and moving image documents for storage and retrieval: an experimental investigation*. PhD thesis, University of Toronto.

APPENDIX

Following is a list of the sites we studied. These were selected on the basis of the following criteria: the presence of multimedia objects, text in both English and French, sufficient text, HTML code, Dublin Core metadata and absence (or relative non-importance) of Flash objects or other dynamic code.

Sites

Bank of Canada Currency Museum / Musée de la monnaie de la Banque du Canada

http://www.museedelamonnaie.ca/fre/index.php

Bonjour Québec (Québec government official tourist site / Site touristique officiel du gouvernement du Québec)

http://www.bonjourquebec.com/francais/attraits/

http://www.bonjourquebec.com/francais/restauration/index.html

http://www.bonjourquebec.com/francais/regions/index.html

Canadian Conservation Institute / Institut canadien de conservation

http://www.cci-icc.gc.ca/html/

Canadian Museum of Civilization / Société du Musée canadien des civilisations

http://www.civilisations.ca :

Selected pages including : « Histoire des autochtones du Canada »

(http://www.civilisations.ca/archeo/hnpc/npint00f.html); « Salles des trésors »

(http://www.civilisations.ca/tresors/tresorsf.asp); « Lois Etherington Betteridge - Orfèvre » (http://www.civilisations.ca/arts/bronfman/better1f.html)

Canadian Museum of Nature / Musée canadien de la nature

http://nature.ca/

Selected pages : « Nos trésors préférés - Une histoire merveilleuse » (http://nature.ca/discover/treasures/trsite_f/trmineral/tr3/tr3.html); « Nos trésors préférés - De minuscules terreurs » (http://www.nature.ca/discover/treasures/trsite_f/tranimal/tr2/tr2.html); « Nos trésors préférés - Une histoire merveilleuse » (http://nature.ca/discover/treasures/trsite_f/trmineral/tr4/tr4.html)

http://www.mcq.org/roc/fr/plan.html

Canada Science and Technology Museum / Musée des sciences et de la technologie du Canada

http://www.science-tech.nmstc.ca

Exposition virtuelle - Maîtres de l'art populaire

http://pages.infinit.net/sqe1rl2/

Government of Canada - The National Battlefields Commission / Gouvernement du Canada - Commission des champs de bataille nationaux

http://www.ccbn-nbc.gc.ca/

http://www.nlc-bnc.ca/jardin/h11-2006-f.html.

Maison Saint-Gabriel

http://www.maisonsaint-gabriel.qc.ca/maison/000_e.html

Maritime Museum of British Comlumbia

http://mmbc.bc.ca

Musée acadien of the Université de Moncton / Musée acadien de l'Université de Moncton

http://www.umoncton.ca/maum/INDEX.html

Musée de la nature et des sciences

http://www.mnes.qc.ca/index.html

Musée virtuel du C.F.O.F.

http://www.cfof.on.ca/francais/navbar/museetest.htm

Museum of New France - Canadian Museum of Civilization Corporation / Musée de la Nouvelle-France - Société du Musée canadien des civilisations

http://www.civilization.ca/vmnf/vmnff.asp

National Archives of Canada / Archives nationales du Canada

http://www.archives.ca/

Selected pages : « Expo 1967 - Pavillons » (http://www.archives.ca/05/0533/05330202_f.html); « Expo 1967 - Activités » (http://www.archives.ca/05/0533/05330203_f.html); « Expo 1967 - Projet d'une exposition universelle à Montréal et mise en candidature » (http://www.collectionscanada.ca/05/0533/0533020101_f.html)

http://www.archives.ca/05/0509_f.html :

Selected pages including : « Dictionnaire montagnais-français, v. 1678, par le père Antoine Silvy, missionnaire jésuite » (http://www.collectionscanada.ca/05/0509/050951/05095104_f.html)

Old Montreal / Vieux Montréal

http://www2.ville.montreal.qc.ca/vieux/histoire/

Old Port of Montréal / Vieux-Port de Montréal

http://www.vieuxportdemontreal.com/histoire_patrimoine/

Our roots - Canada's local histories online / Nos racines - Les histoires locales du Canada en ligne

http://www.ourroots.ca/f/intro2.asp.

Société des musées québécois

http://www.smq.qc.ca/