

Organizing moving image collections for the digital era: research results

Research report from the project of the same name
funded by the
Special Libraries Association
under the
Steven I. Goldspiel Memorial Grant for 1999

submitted December 19, 2001 to
Information Outlook
Special Libraries Association
Information Outlook Submissions
1700 Eighteenth Street, NW
Washington, DC 20009-2514
USA

Authors:

James M. Turner, professeur agrégé
École de bibliothéconomie et des sciences de l'information
Université de Montréal
james.turner@umontreal.ca

Michèle Hudon, professeure adjointe
École de bibliothéconomie et des sciences de l'information
Université de Montréal
michele.hudon@umontreal.ca

Yves Devin, agent de recherche
École de bibliothéconomie et des sciences de l'information
Université de Montréal

Correspondence should be addressed to:
James M. Turner, professeur agrégé
École de bibliothéconomie et des sciences de l'information
Université de Montréal
CP6128, succursale Centre-ville
Montréal, QC H3C 3J7
tel. 514 343 2454
fax 514 343 5753
james.turner@umontreal.ca

Introduction

Pictures have always been used to represent concepts and ideas and to communicate messages, and now that we collect them so extensively, we need to represent the pictures themselves in order to store and retrieve them. Photography, movies, television and now digital images stored on computers have all contributed to the rapid buildup of ever larger collections. Whatever the format or the presentation medium, pictures have become a most important mode of communication in our times. They play a crucial role in such areas as medicine, journalism, advertising, education and entertainment. The notion of picture collections now has to do with a vast world, and an attempt to describe this world is represented by the study poster entitled *The world of visual collections* (GRIV 1998), which takes into account the areas of art, engraving, photography, computer graphics, the types of institutions which collect and the very many professions that use pictures as part of their work.

Moving images are of course a goldmine for many organizations and individuals, and it is important to describe them adequately in databases in order to show their richness and complexity if we are to exploit these collections fully. However, the world of moving image collection organization is one of locally established techniques, with little or no standardization and without communicability between systems. This has not been a problem until now because systems were managed independently of each other, but in the networked world in which we now live, the question of discovery and exchange of information has come to the forefront and needs to be addressed.

Our recently-completed research project was concerned with indexing moving image collections for storage and retrieval. The general goal of the project was to understand the techniques and tools used for representing the content of moving image collections that are indexed shot by shot. We especially wanted to study the question of indexing languages and their structure, as well as techniques for keeping them current. Several more specific objectives were identified:

- to determine how many terms, excluding proper names, are used to describe North American moving image collections indexed at the shot level
- to estimate the rate of growth of term creation in these tools
- to discover to what degree the lexical concepts are similar among the various tools
- to evaluate the possibility of creating a universal indexing vocabulary for general collections of moving images, those that represent everyday objects and events

In this article we look at the background information to the study, after which we describe the method used to collect the data. The results are then given along with discussion of them, followed by conclusions we might draw from this research.

Background

Both still and moving images can be divided into three broad categories: art images, documentary images, and "ordinary" images, each requiring its own type of organization. The proliferation of supports and the changing context brought about first by the arrival of computer technology and then the networking of resources are the driving forces behind a great deal of work in retooling and rethinking work methods, but they do not change this fundamental (if arbitrary) classification. Nor does the shift from analogue to digital images. Thus there is a great deal of work to do, but the guiding principles remain the same.

The many new systems and ever more efficient technologies for capturing and processing moving

images require the establishment of effective management systems. It is necessary to be able to find any specific shot in a particular collection rapidly and efficiently. Without the establishment of new methods for storage and retrieval of moving images, these valuable resources will get lost in a hopeless jumble of useless data.

Research in the area of storage and retrieval of moving images takes place using two distinct approaches with little in common (Cawkell 1992, 180). The low-level or content-based approach is the focus of work by computer science researchers. This approach involves the statistical manipulation of pixels to get information about colour preponderance and arrangement, recognition of textures, patterns, boundaries, objects, scene detection, and so on. The high-level or concept-based approach is the focus of work by information science researchers. This approach involves human generation of metadata substantially assisted by computer technology (semi-automatic), as well as automatic generation of high-level metadata. The general focus of this approach is on finding ways to generate shot-level indexing automatically from text created during the pre-production, production, and post-production stages, such as closed captioning, audio description, and production scripts. The two research streams are complementary, and the best information systems for storage and retrieval of moving images will need to incorporate both approaches.

The rapid buildup of collections and the need to communicate between systems via networks add urgency to the need for development of common methods for storage and retrieval. One important contributor to communication would be a common thesaurus for shot-level indexing.

Analysis and representation of moving images

Images can be analyzed and interpreted in diverse ways. One model widely used in image indexing, based on the work of the art historian Erwin Panofsky (1955), identifies three levels of interpretation. The first, which Panofsky calls pre-iconographic, deals with the primary or “natural” subject of an image. The second, iconographic, has to do with secondary or conventional subject matter. The third, iconologic, deals with symbolic levels of interpretation. Shatford (1986) emphasizes the first and second levels, and translates them into the “ofness” and the “aboutness” of a picture. What is it a picture of? What is the picture about? These levels also correspond rather closely to the ideas of denotation and connotation in the area of semiotics.

Whether they are still or moving, pictures contain a great variety of information and can have different meanings for different viewers (Shatford 1986, 42). This fact can of course be a source of problems in working out uniform methods for describing pictures for purposes of storage and retrieval. For stockshots, it has been suggested that only the primary level is really useful (Turner 1990, 12). Most moving image collections described at the shot level are indexed this way, and the descriptors necessary for representing the visual content simply name the objects, persons and events found in the shots (e.g. a cat or a chair) rather than abstract notions (e.g. comfort or serenity). From this perspective, using a thesaurus as an indexing tool is of great interest. The lexical and structural control such a tool offers can contribute greatly to improving access to the content of collections, to reducing noise (i.e. too many hits) and silence (i.e. too few hits) in retrieval, to improving precision and ultimately to satisfying users by giving them what they need without requiring a great investment of their time.

The development of thesauri is based on rules and principles spelled out in international norms (Hudon 1994, 75-76). Thesauri are dynamic tools that are adaptable to new realities and new needs of the collections that use them. The content can constantly be updated and improved to meet the needs of users, especially when database software is used for managing the thesaurus. However, this tool is relatively exclusive, and is usually developed to manage the vocabulary of a particular area of endeavor and that of a particular group of users (Van Slype 1987, 117). Unlike classification systems and lists of subject headings, thesauri are not encyclopedic. Yet such a thesaurus, general yet

encompassing, is what is needed for the shot-level description of moving image collections.

A few thesauri have been created specifically for indexing visual documents such as art images, photos, slides and plans. Perhaps the best known of these is the *Art and architecture thesaurus* (2001) managed by the Getty Foundation. The *AAT* offers a standardized terminology of about 40,000 expressions covering art and architecture from antiquity to the present day. The *Thesaurus for graphic materials* (2001), published by the Library of Congress, offers descriptors for indexing printed pictures, photos, drawings, cartoons, posters and architectural drawings. In Canada, the National Film Board's thesaurus for indexing stockshots is also of use. In addition, there are a number of visual thesauri available. These take the approach of using pictures instead of text to represent other pictures (Rasmussen 1997, 182), offering a clear advantage in a multilingual environment. The *NASA Visual Thesaurus* is an example of this type of tool.

The work of developing a thesaurus is complex, requiring a sequence of stages and intellectual operations which lead to making a number of decisions. The stage of building the lexicon is of primary importance. Whether the descriptors come from reference sources, from the images that need indexing or from user queries, they need to name objects clearly and with sufficient precision to permit accurate description of the images which contain them. Anecdotal evidence suggests that a limited number of terms is sufficient for describing general collections. This reflects a phenomenon in the use of natural language, in which the number of words available is much greater than the number required for communication and discourse. According to Guiraud (1960, 93), 4000 words account for 97.5% texts, across all languages. Dahl (1979) gives the figure of 8000 words to account for 90% of texts written in everyday English. Thus we can imagine that a smaller number of words, covering only common names, would be sufficient for describing everyday persons, objects and events such as they are represented in general visual collections. If this proves to be true, then a common indexing vocabulary for these collections can probably be created.

Methodology

Thirty-three organizations were identified as potential participants in our study. Criteria for participation in the study included having a collection of non-art moving images that had been in operation for at least five years, and that the collection be indexed at the shot level. Each organization received an information kit about the project, along with an invitation to participate. Follow-up e-mails and telephone calls completed our attempts to recruit participants. Twenty-two responses were obtained (67%). Eleven organizations (50%) agreed to participate, and nine others (41%) declined. These latter were mostly from the private sector, for which the advancement of research is not necessarily a priority and for which information about collections is often considered proprietary. Two organizations (9%) which showed initial interest were unable to follow up. The eleven organizations which agreed to participate managed a total of fourteen collections.

Participating organizations responded to a questionnaire containing four distinct sections: identification of the milieu, characterization of the collections, methods of collection management, and characterization of the lexical tools used for indexing and retrieval. Structured follow-up interviews with personnel closely associated with the indexing tools in these organizations helped complete the information from the questionnaire. These persons also had intimate knowledge of the milieu, the collections, and the tools used for collection management. The on-site interviews were also helpful in obtaining additional useful information related to the study. Sometimes these visits also made it possible to obtain useful documentation, such as a copy of part of the thesaurus used or of the indexing policy in force. The participating organizations were very helpful and cooperative in sharing information useful to our study, when such information was available.

Results and discussion

Each of the organizations that supplied data has its own structure and work methods. Time and production constraints are very important. Sometimes these are such that work is duplicated because material cannot be found in time to meet a deadline, for example within an hour of a request from production staff. In addition, budget restraints always mean pressure to try and do more with less. These severe constraints are difficult to reconcile with the investment of time and effort necessary for developing good tools for managing collections.

In our study, several names were given for the sites where collections are housed, sometimes even within an organization. The term “stockshot library” was used by seven of the organizations. “Archives center” was the preferred term for five of the sites. Other denominations included “video art distributor” and “news video archive”.

The collections. Seven of the fourteen collections that supplied data are mixed, covering both general and specific topics. The material includes film, clips of news footage and other public-interest material. The collections in our study were closely tied to television networks and film production studios. A broad range of general and specific subjects are necessary to respond to the needs of the varied clienteles of these collections. Three collections (21%) were of a general nature, and only two (14%) described themselves as specialized. Interestingly, two of the collections were unable to fit themselves into these categories. Twelve of the fourteen collections (86%) were less than fifty years old.

Table 1 provides an overview of the material available in the various collections we studied.

Table 1. Formats used in the collections.

Type	Number of collections
8 mm film	4
16 mm film	8
35 mm film	8
72 mm film	1
Other film formats	6
3/4" U-Matic video	11
1-inch video	8
2-inch video	5
Betacam	13
Other video formats	10
Optical disk and other formats	5

The material is varied, ranging from 8 mm film to optical disks, and this is characteristic of the broad range of supports on which moving images can be stored. Not surprisingly, the most representative supports are 16 mm and 35 mm film, and the video formats 3/4" U-Matic and Betacam.

The size of the collections we studied was impressive. The data given in table 2 provides only a partial view of the importance of the collections managed by the participating organizations.

Table 2. Size of the collections.

Collection number	Number of titles	Viewing time (hours)
1	4 962	n/a
2	14 000	3 800
3	n/a	n/a
4	36 848	n/a
5	11 755	750
6	100 000	n/a
7	n/a	n/a
8	n/a	n/a
9	50 000	17 500
10	n/a	n/a
11	18 500	17 848
12	94 732	n/a
13	5 600	n/a
14	n/a	n/a

For the most part, the only way to measure this was by the number of titles. Information about the number of viewing hours was either summary or not available (n/a), and surprisingly, no institution was able to supply data about the volume of the collections (measured in linear feet or meters).

Description and indexing. Given the size of the collections and their fast rate of growth, the role of computers in managing them has been considered essential for some time now. The participating institutions all had databases with complex structures which fostered more or less effective retrieval of pictures representing specific situations or objects. This is reflected in the description and indexing practices. Almost all of the collections (11/14 or 79%) are catalogued and indexed by title or whole document. This is not surprising in view of the ease in obtaining this information and of its importance for retrieval. Some of the collections we studied were described and indexed more deeply, at the sequence level (5/14 or 36%) or at the shot level (8/14 or 57%). Five organizations catalogue and index at all three levels (title, sequence, shot). Five organizations also said they index at other levels than those we had suggested, for example using a reel of film as the indexing unit.

Most of the collections were indexed at the first and second levels of meaning borrowed from Panofsky. Surprisingly, five of the collections were also apparently indexed at the third or symbolic level, usually thought to be limited to the world of art. However, none indexed only at this level without also using the other levels.

Four collections indexed using five descriptors or fewer (on average) per shot. The highest average number of descriptors permitted per shot was 15, and this figure was the case for three collections. Two other collections had no maximum number. In all cases, the maximum number of descriptors actually assigned depended on the indexing policy, when there was one, or on the capacity of the automated information system in use.

Regarding the degree to which indexing practice was structured, a continuum of situations was found in the collections studied. At one extreme, there was no control at all over indexing practice (everything is indexed, or nothing is indexed). At the other end of the continuum was indexing tailored to the collection using a thesaurus specifically developed for this purpose. Between these two extremes, controls on indexing practice included using the Library of Congress Subject Headings or some adaptation of these, using a list of keywords developed for the purpose, or some simple

classificatory structure, as well as combinations of various techniques. This continuum of practice reflects the double tendency highlighted by Cawkell (1992) of maximum use of computer technology combined with techniques well established in the area of information science. Only rarely is a formal indexing policy in use, and unfortunately we were unable to consult any at all.

Table 3 shows the types of indexing languages used for representing the content of the collections in our study.

Table 3. Types of indexing languages used in the collections.

Type of indexing	Number of collections
Keywords	7
Classification	3
Commercial thesaurus	2
In-house thesaurus	5
Mixed thesaurus	1
Subject headings	6
Other	6

Keywords, taken from natural language and with no controls over the form they take nor the precise meaning assigned, are used in 7/14 (50%) of the collections. Subject headings are used in 6/14 (43%) of the collections. A classification scheme is used for indexing 3/14 (21%) of the collections. The classification schemes we found in use are in-house schemes developed in response to local needs. During the visits our research assistant made to follow up on the questionnaire participants filled out, we were able to confirm that the systems in use for managing at least six of the collections also permitted full-text searching; however, only two had mentioned this possibility in responding to the questionnaire. Finally, vocabularies in the “other” category include a list of technical terms specific to the area of film, and a list of geographic descriptors.

In our study, an organization generally used more than one indexing tool for each or several of its collections. The majority of participating institutions said they use between two and six different vocabulary-management tools to represent the content of their collections.

Thesauri. Six of eleven organizations used one or several tools identified by them as being a thesaurus. Two collections used a commercial thesaurus, one used a mixed thesaurus, and five had an in-house thesaurus. Two other collections were indexed using thesaurus-like tools. The six thesauri we were able to consult were presented in the usual form of alphabetical lists of descriptors. Specialized coverage was found in five of these tools, while only one could be described as truly general. The terminology can be described as “everyday” in three of the thesauri, and was more formal in the other three.

Several of the questions in our questionnaire had to do with the lexical content of the thesauri (e.g. total number of terms, of terms that were not descriptors, of proper names, and so on). Unfortunately, it was next to impossible to collect precise data about this type of information from the participating institutions. As it turned out, most of the thesauri are managed by proprietary software that was unable to generate statistics useful to us. Because of this, the figures given in table 4 are mostly estimates derived from lexical samples taken from the thesauri to which we were given access. The figures given are thus presented as only an attempt at an indication of the great variation in the number of terms found in the tools in use.

Table 4. Types of terms found in the thesauri.

Thesaurus number	Total number of terms	Terms other than descriptors	Proper names (personal, geographic, etc.)
1	6 969	1 451	2 244
2	344 500	8 850	220 000
3	3 222	n/a	660
4	1 163	42	744
5	704	89	n/a
*6	3 680	1 346	1 204

*For Thesaurus 6, only words beginning with the letters F, I, and R are counted

We observe the important proportion of terms (almost a third in each tool) that are proper nouns (names of persons, of institutions, or of geographical places). We also note the small proportion of terms other than descriptors. From this we can deduce that the tightening of vocabulary by synonym control has not taken place and that the efficiency of the tools for retrieval is thus weakened. Terms that are included in the various thesauri come from a number of sources, such as general and specialized reference sources, user queries, and existing semantic networks such as those found in other thesauri.

Most of the thesauri in use in the participating organizations had an explicit relational structure connecting descriptors by relations of equivalence, of hierarchy, and of various kinds of associations. Only one had cross-language associations (a bilingual English-French thesaurus; all others used only English), but all six had hierarchical and associative relationships, and four of the six had some control of synonyms. The fact that only four of the six thesauri used some kind of control over conceptual and terminological equivalence again suggests that semantic control is only partial and so it is probably somewhat ineffective. However, the similar structures suggest that norms for thesaurus construction were nevertheless considered.

The effectiveness of tools for vocabulary management can only be maintained if the lexical and relational content is kept up to date. The responses to our question concerning the frequency and the regularity of updates showed that for three of the six thesauri, this was done as needed, changes being immediate and integrated dynamically into the database. In the other three cases, one was updated daily, one weekly, and one irregularly.

For three of the thesauri, a single person was responsible for updating the thesaurus and for making decisions about controlling and expanding the semantic networks. In the case of another thesaurus, two persons were responsible for its upkeep, and for the two remaining thesauri all the users contributed to the updating operations. Formal procedures or guidelines for updating these tools were not always available.

Updating a thesaurus has largely to do with creating new descriptors. Responses to our question about the number of new descriptors added annually were rather surprising. Half of the thesauri were increased by a maximum of 50 new descriptors annually, while the other half were increased by more than 300. We might wonder about the causes for the disparity in these tools which, conceptually at least, should be rather similar one to the other. However, managers of the thesauri we studied were unable to say with any certainty which proportion of the terminology included at the time the data was collected had been included by the end of the first, the third, and the fifth years of the existence of the thesaurus, nor at what moment the rate of term creation had leveled off and attained its present level. While it may be fairly clear that the number of terms necessary for indexing a general collection of moving images reaches a peak after which only few new terms need to be added, the data we

obtained do not permit us to identify where that peak is situated.

Lexical analysis. An analysis of the entire lexical content of the various tools that were made available to us would be ideal. However, in view of the resources available for doing this and the large number of terms contained in the tools we studied, such an undertaking was impossible. The rather summary observations that follow are based on a sample of all the terms beginning with the letters F, I and R in the seven tools (thesauri, keyword and subject headings lists) to which we had access. The three letters of the alphabet we used were randomly selected from the fifteen letters of the alphabet which fell in the mid-range of a minimum of 900 English words to a maximum of 5000 words beginning with those letters. The letters outside this range were eliminated on the basis of having too few words beginning with them to be useful, or too many words beginning with them to be manageable. The calculations were based on the content of three English-French dictionaries, *Larousse* (1990), *Harrap's* (1978) and *Robert-Collins* (1987).

Words that were represented by numbers, as well as proper names of persons, organizations or titles of books, songs, films, etc. were removed from the data. The remaining terms were combined into a single list of 2292 distinct terms. Of this number, 1858 (81%) represent concrete objects or entities and 434 (19%) represent abstract notions of the kind that are more useful for indexing at the second or third levels borrowed from Panofsky.

Table 5 gives the frequency of terms in seven vocabulary management tools.

Table 5. Frequency of the terms.

Number of tools	Number of descriptors	Percentage this represents
1	1680	73
2	338	15
3	134	6
4	72	3
5	47	2
6	14	0.6
7	7	0.3
Total	2292	99.9

As we can see from this table, a large proportion of the terms are found in only one tool, and only 7 of the 2292 terms are present in all seven vocabulary-management tools we analyzed. Considering the similarity of the type of content held in the collections managed by the participating institutions, these results are surprising. In addition, they do not support our hypothesis to the effect that the number of terms necessary for indexing general collections of moving images might be rather limited, and that the terms might well be the same from one tool to another.

However, we are unable to conclude anything definitive from this analysis. More than likely there are a large number of synonyms among the descriptors with a frequency of 1. Grouping the terms by concepts these terms represent might shed some light on this. Furthermore, additional grouping would become possible once terms representing identical or similar notions but at different hierarchical levels were identified.

Conclusion

Although we were unable to complete this study as we had hoped for lack of available data, we were able to confirm the great disparity in tools and methods used for representing the content of moving image collections that practitioners observe widely.

The data we were able to collect confirm that despite the absence of personnel formally trained in information management techniques, and despite a great deal of pressure from the very competitive environment in which they function, the participating organizations managed to retrieve useful information within reasonable delays. Often this seems to be more a function of fast computer technology than of good information management. We do not know whether the material retrieved is the best available from the database, only that it will do for the purposes at hand. Even public-sector organizations have sometimes been reduced to this state because severe budget cuts in the last decade or so have forced them to loosen the tight management methods they had developed and to abandon their careful methods of analysis in favor of more summary practices.

All the participating organizations in our study showed interest in our project and its hypotheses. In addition, all found interesting and doable the idea of a common thesaurus for managing general moving image collections of everyday persons, objects and events. In the context of other research projects we have undertaken, and in light of the results of analysis of other lexical data, we are studying the possibility of creating and testing such a thesaurus

Acknowledgements

We are grateful for the help of the Special Libraries Association, which financed this research under its Steven I. Goldspiel Memorial Grant for 1999. We also thank the personnel in the organizations who supplied data and who generously shared their time in helping us do this project.

References

- Art and architecture thesaurus. 2001. Available at <http://www.getty.edu/research/tools/vocabulary/aat/index.html> (page accessed 2001 12 06).
- Cawkell, A.E. 1992. Selected aspects of image processing and management: review and future prospects. *Journal of Information Science* 18, 179-192.
- Dahl, H. 1979. *Word frequencies for spoken American English*. Detroit: Gale Research Company.
- GRIV (Groupe départemental de recherche en information visuelle). 1998. *The world of visual collections*. Study poster, 91 x 61 cm. Montréal: École de bibliothéconomie et des sciences de l'information, Université de Montréal.
- Guiraud, P. 1960. *Problèmes et méthodes de la statistique linguistique*. Paris: Presses universitaires de France.
- Harrap's. 1978. *Harrap's new shorter French and English dictionary*. High Holborn, London: Harrap & Co.
- Hudon, Michèle. 1994. *Le thésaurus: conception, élaboration, gestion*. (Clé en main). Montréal: ASTED.

- Larousse. 1990. *Dictionnaire français-anglais*. Nouvelle édition enrichie. Paris: Larousse.
- Panofsky, Erwin. 1955. *Meaning in the visual arts: papers in and on art history*. Garden City, NY: Doubleday Anchor Press.
- Rasmussen, Edie M. 1997. Indexing images. *Annual Review of Information Science and Technology* 32, 169–196.
- Robert-Collins. 1987. *Robert-Collins dictionnaire français-anglais, anglais-français*. Nouvelle édition. Paris: Dictionnaire Le Robert.
- Shatford, Sara. 1986. Analysing the subject of a picture: a theoretical approach. *Cataloging & Classification Quarterly* 6, no. 3, 39–62.
- Theasurus for graphic materials. 2001. Available at <http://lcweb.loc.gov/rr/print/tgm1/> (page accessed 2001 12 06).
- Turner, James M. 1990. Representing and accessing information in the stockshot database at the National Film Board of Canada. *Canadian Journal of Information Science* 15, no. 4, 1-22.
- Van Slype, G. 1987. *Les langages d'indexation: conception, construction et utilisation dans les systèmes documentaires*. Paris: Éditions d'Organisation.