Text as a Tool for Organizing Moving Image Collections

James M. Turner, Michèle Hudon, and Yves Devin

Université de Montréal

Abstract

In the rapidly-growing networked environment, it is critical to develop common methods for shot-level and scene-level description of moving image documents, in order to foster discovery and retrieval of these resources worldwide. In this paper, we describe the methodology used and report preliminary results obtained in a project designed to study existing tools for indexing moving images at the shot level. A number of institutions holding moving image collections were recruited as partners, and visits to them were then made to complete a questionnaire and gather other information during a structured interview. Preliminary results show that a variety of tools are used for indexing moving images, and that those institutions which have been able to invest in building good indexing tools over the years have seen the quality of those tools suffer in the wake of cutbacks in resources needed to maintain them.

1. Introduction

As we enter the new millenium, we observe that the organization of moving image collections is still characterized by ad hoc information systems. In the rapidly-growing networked environment, it is critical to develop common methods of shot-level and scene-level description, in order to foster retrieval and ultimately, to share resources worldwide. Researchers are aware of the problem, and two streams of research which complement each other address the issues involved. The first stream focuses on low-level access to images using methods from computer science and concentrates on statistical techniques for deriving characteristics of images that help promote retrieval. The second stream focuses on high-level access to images using methods from library and information

science, and concentrates on the use of text to create information useful for retrieval, information which is especially valuable since it is not available from the images themselves.

The materials contained in film libraries, television libraries, and both film and television stockshot libraries is usually of a general nature. Thus managing the terminology used to index these collections involves descriptor lists and thesauri which are often constructed from scratch to reflect the particular local reality of the collection. For the researcher who looks in many collections to find material for use in film and video production, this means relying heavily on the resource persons who work with each collection and learning a number of different retrieval systems and indexing vocabularies.

It has been thought that a thesaurus for indexing everyday film and video materials would become unmanageable after a certain point because there are so many kinds of persons, objects and events to describe that it would eventually become impossible to manage the semantic relationships between them. However, anecdotal evidence suggests that term creation levels off once sufficient terminology for indexing most shots has been created. This paper describes a research project which belongs to the second or high-level research stream and which aims to study the characteristics of thesauri used for indexing moving image collections at the shot level.

What is the point at which term creation levels off? How many terms for describing moving images does a thesaurus need to contain in order to be considered complete enough to describe a general collection adequately? Are the terms similar from one thesaurus to the next, or are collections so particular that an individual tool is required for each collection? Would it be reasonable to try to construct a general thesaurus of everyday persons, objects and events that could be shared among moving image collection managers? These are the research questions our project addresses.

The general goal of the study is to reach an understanding of the organization of existing vocabulary-management tools for moving image collections. The specific objectives are:

- to discover how many terms, excluding proper names, are contained in a controlled vocabulary for managing general moving images collections before term creation levels off;
- to identify patterns among terms in the existing thesauri created for moving image collections;
- to assess how patterns found can contribute to building a shared vocabulary useful for special collections containing general material.

2. Background

Moving image collections are largely found in movie and television production facilities, but they are also to be found in many other contexts and environments, such as corporate libraries, government agencies, documentation centres of research groups, holocaust museums, religious archives, and so on. Librarians have developed a great deal of expertise with managing print collections, but there are no generally- and widely-accepted tools available for organizing moving image collections. Visual resources librarians and researchers are working on problems related to art collections and to slide collections of works of art, but moving image and other non-art picture collections need urgent attention. This is due to the profusion of production and the consequent mushrooming of such collections in recent decades, including television news libraries and stockshot libraries. To be of help to their user base, these collections need to be catalogued and indexed at the shot level.

In the context of working groups around the planet trying to work out metadata standards for the management of all kinds of digital materials, high-level metadata standards for moving image databases have not been studied. We hope that the results of our study will make a contribution in helping establish a theoretical basis on which standards can be built.

3. Methodology

The work involved in this project is divided into four phases: planning, data collection, data analysis, and dissemination of results. At the time of writing, the second phase is nearing completion. In the planning stage, a literature search for statistical and other information on thesauri was conducted, to gain a broad understanding of the use of existing thesauri, of their users, of their contents and of their structure. In addition, linguistic data on the number of terms ordinary people need to name common objects in order to function in society was sought, so that we could later determine whether there is some correlation between these numbers and the number of terms found in thesauri used for managing general moving image collections.

Using personal contacts and the membership directory of the Association of Moving Image Archivists (AMIA), moving image collections in North America were then identified and holding institutions were asked to participate in this study. The data collection instrument was built during this second phase of the project. It consists of an information package about the project, and a twelve-page questionnaire. Since we wanted to collect data in both English and French, a separate version of the questionnaire was produced in each of these languages. The questionnaire was designed with a view to obtaining precise information concerning a profile of the holding institution, the characteristics of its collections, its collection management policies, and the characteristics of the thesaurus used for vocabulary management. The information package was sent ahead to institutions that had agreed to participate in the study, as well as to those who were at least potentially willing to collaborate.

Our research assistant is following up with visits to the participating institutions, where structured interviews are conducted with resource persons. Where possible, vocabulary management tools are also examined. Specific quantifiable data is sought, including the number of common names found in the thesaurus, the number of proper names, and the number of semantic networks, as well as the nature of these last and the degree to which they are developed. In addition,

the entire content of entries for randomly-selected letters of the alphabet is being collected, to permit a comparison of the various lexicons at the data analysis stage. Further notes are made concerning other aspects of importance or of use to this study, such as the way geographical information is organized, whether there are authority lists for proper names, any noteworthy methods used in the construction or management of the vocabularies, and so on. Where participating institutions permit it, a copy of the entire thesaurus is obtained.

Fortunately, several institutions in the Montréal area manage moving image collections, and we were lucky in that the most important ones agreed to participate in the study. We were thus able to use these as a testbed for the data collection instrument without incurring a great deal of expense. At the time of writing, data collection is not yet finished. Within the limits of the budget available for the study, and based on the concentration of important moving image collections found there, the following North American cities were targeted for visits: Montréal, Toronto, New York, Atlanta and Los Angeles. Participants were also found in Boston and San Francisco, and these cities were added.

The third phase of this project involves an analysis of quantifiable data to determine whether there is some reasonable equivalence among the number of common names found in thesauri used for moving image description. We also hope to determine whether the terms are the same from one thesaurus to another. Thus, a comparison of all the entries for randomly-selected letters of the alphabet is being undertaken. A merged list of the terms accompanied by a count of frequency across the thesauri will yield this information. The management of proper names will be analysed in terms of patterns, comparison of authority lists, and procedures for updating. Finally, a qualitative analysis will be undertaken to identify patterns in work practices and organizational methods.

4. Preliminary results

Results to date are compiled from notes made after meetings with the participating institutions, as well as from information the participants gave on the questionnaires they received. As we mentioned in the methodology section, the

questionnaire is designed to collect data in four categories: a profile of the holding institution, characteristics of its collections, its collection management policies, and characteristics of the thesaurus used for vocabulary management. In general, we note a great variety in vocabulary management tools, a reflection of the ad hoc systems found in these collections. Because of this reality, comparison of the tools is difficult and our analysis tends to be more qualitative than quantitative. In this section we summarize results from the eight institutions from whom data has been collected so far.

4.1 Profile of the holding institutions

Four of the eight institutions supplying data classified themselves as stockshot libraries, two as archives, one as a special library, and one as a documentation centre. Concerning the dates the collections were started, one institution could not supply the information and of the remaining seven, five were founded after 1960.

4.2 Characteristics of the collections

As regards the types of collections, two are characterised as holding material of a general nature, two as holding material of a specialised nature, three collections are described as mixed, and one is classified as "other". The kinds of materials held in the collections are shown in table 1. The figures indicate how many of the eight institutions hold each type of material.

Table 1: Types of material held in the collections visited

Type and format	Number of institutions using
Film	
8 mm	2

16 mm	5
35 mm	6
72 mm	0
Other film formats	4
Video	
3/4 inch U-matic	6
1 inch	3
2 inch	1
Betacam	7
Other video formats	5
Other supports	4

From this table we can see that the most widely-held film formats are 16mm and 35mm, and the most widely-held video formats are 3/4 inch U-matic and Betacam. These figures correspond to what is generally believed to be the composition of such collections. As can be expected, other formats are also found.

Concerning the size of the collections, information is sketchy, partly because of the variety of ways this can be measured, and partly because collections typically do not maintain rigorous statistics on this information. We tried to collect data on the number of titles or records, the number of hours of viewing time, and the shelf space occupied by the material in linear metres. Concerning this last metric,

none of the institutions visited so far was able to supply a figure. Six institutions gave the number of titles or records, the smallest holding slightly under 5000 and the largest holding 100 000. The average number for the six institutions reporting such a figure is 36 260. Only three institutions were able to estimate a number of hours of viewing time for their collections. These are as follows: 3800 hours (for 14 000 records), 750 hours (for 11 755 records), and 17 500 hours (for 50 000 records).

4.3 Collection management policies

The material in moving image collections can be described at one or more of three levels, the title, the sequence, and the shot. Of the eight institutions for which data is reported here, six index at the title level, four index at the sequence level, five index at the shot level, and four index at all three levels. Two institutions reported indexing at "other" levels, such as a physical storage unit for the material (e.g. a reel or cassette).

Concerning the type of indexing used, one institution reported using no human indexing of materials, and of the remaining seven, four use key words, one has an in-house classification, three use a thesaurus, and five reported using some "other" indexing instrument. Table 2 summarizes how many of the eight institutions use each type of indexing tool.

Table 2: Types of indexing tools used in the collections visited

Type of tool	Number of institutions using
Key words	4
Classification	1
Thesaurus	3
Subject headings	0
Other	5

From these figures, we can get an idea of the difficulty involved in conducting an analysis on the thesauri used. We note the tendency of institutions to use more than one type of indexing tool to describe their collections, in order to respond to different types of user needs.

Concerning the number of indexing terms assigned per record, only one institution said it had a maximum number of ten, and the other institutions had no maximum. Of the five institutions able to say how many terms they actually assign to records typically, the lowest figure reported (one institution) was five, the highest figure reported (one institution) was twenty, and the three remaining institutions reported figures in the range of 10 to 13 terms.

4.4 Characteristics of the thesaurus

From the data collected so far, it seems clear already that we will not be able to achieve our goals concerning the nature and characteristics of thesauri used to manage moving image collections. This is partly because fewer than half of the institutions from which data has been collected (three of eight) actually use a thesaurus to manage the terminology for providing subject indexing to their material, and partly because even those who do use a thesaurus are not able to provide quantitative data on its history, management, characteristics, and use.

Of the three institutions using a thesaurus as a tool for managing indexing vocabulary, one stated that the material covered was of a general nature, and two stated that the material was both general and specialised. Two of the thesauri have equivalence, hierarchical and associative relationships (the standard thesaurus format), and the other expresses only associative relationships. In all three cases the organisation of the vocabulary entries is alphabetical.

Updating the thesaurus is done as needed in all three cases. In answering our question, two institutions responded Dynamically and the third Other (in this case meaning "as needed"). The question we asked concerning who held responsibility for updating the thesaurus had possible answers of One person, Two people, and More than two people. For the three organisations using a

thesaurus, one fell into each of these three categories. In the case of the organisation in which more than two people had responsibility, there was some coordination among those responsible, although new terminology is simply created as needed to respond to new indexing situations. This reflects the needs of the industry to have up-to-date information quickly, and the subsequent elimination of discussions and meetings to accept candidate descriptors as required. The availability of computers to manage the vocabulary permits this streamlining of the process, but the largely ad hoc nature of decision-making means that the quality of the expression of semantic relationships cannot be tightly controlled.

In response to our question on the number of descriptors added annually, one organisation responded Fewer than 50 per year, and the two others responded Over 300 per year. Interestingly, the organisation that adds fewer than 50 new descriptors per year is a public institution that has a thesaurus with rich semantic relationships, while the two other thesauri are found in private institutions with little control over semantic relationships. Finally, two of the organisations use proprietary software to manage their terminology, and the third uses a commercial database product.

5. Discussion

The public-sector partners we recruited have been glad to provide data and to help us in many other ways to complete our study. The difficulty we have had in finding private-sector partners reflects the idea that research is not a priority for them. Those managers who were able to agree have been helpful and willing partners, but the competitive nature of their institutions has meant that they were sometimes unable to share with us data considered proprietary. Both public-and private-sector partners were not always able to invest the resources required to compile data we needed to complete our study.

However, the scanty data available to date does support the richer anecdotal evidence gathered in numerous conversations with managers of moving image collections to the effect that, despite the lack of training in information science

methods and working under the pressures of the industry, the private sector institutions still manage to hobble along using indiscriminate methods, retrieve usable material quickly enough, and turn a profit. Managers of collections in public institutions that have made the investment in high quality tools over the years have seen their careful work deteriorate appreciably as pressure to show short-term gain has required them to dismiss competent employees, "simplify" complex indexing strategies, and come to terms with the notion that any retrieval results are good enough, as long as a sale takes place.

6. Conclusions

It seems clear that short-sighted, haphazard policies imposed on managers of moving image collections will continue to impede their work for some time to come. Just as whatever results a Web search engine returns are deemed "good enough" by most users, the results returned from ad hoc systems with makeshift indexing will remain the norm.

The key to providing high-quality access to the rich collections of moving images that already exist and to those which are being built resides in developing automated techniques for storage and retrieval. Some of the groundwork in demonstrating the theoretical underpinnings for recycling text for use in shot-level indexing has been done, and some existing information management systems have begun to make use of networked vocabulary-management tools to attach keywords to moving images. Ongoing work will hopefully produce useful tools that can be shared in a networked environment, as well as provide some basis for standardisation of the high-level metadata needed for resource discovery and sharing. What we are not able to accomplish manually for lack of resources may be able to be done by computers, if we can train them properly.

Acknowledgments

This work is being carried out under the Steven I. Goldspiel Memorial Research Grant for 1999, awarded by the Special Libraries Association, Washington, DC, to whom grateful acknowledgment is made. In addition, special contributions to

the work described in this paper were made by Kathy Christensen, Renée Provitt and Janice Simpson, for which we thank them.