How Many Terms Are Enough? Stability and Dynamism in Vocabulary Management for Moving Image Collections

Abstract: Most moving image collections have existed for less than a century, and as we enter the new millennium we observe that the organisation of these collections is still characterized by ad hoc practices. An important stream of research in this area focuses on high-level access to images using methods from library and information science, and using text to create information useful for retrieval. It has been established that common names for objects seen in the image are the key to retrieval in such collections. On a day-to-day basis, those responsible for collection management build indexing vocabularies, creating terms as necessary, and often structuring them loosely into a thesaurus. Discussions with moving image collection librarians have led us to believe that there may be an optimal number of common names a thesaurus for managing general collections of moving images should contain, and that the terms may even be the same from one thesaurus to the next. In this paper, we describe the methodology adopted for studying this question, and report preliminary results.

1. Introduction and Background

Most moving image collections have existed for less than a century. They are mostly found in movie and television production facilities, but also in many other environments, such as corporate libraries, government agencies, documentation centres of research groups, museums, religious archives, and so on. As we enter the new millennium, we observe that the organization and exploitation of these collections is based on 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 this issue, and at present there are two complementary streams of research tackling the problem. The first, called content-based, focuses on low-level access to images using methods borrowed from computer science, and concentrates on statistical techniques for deriving characteristics of images that help promote retrieval. The second stream, identified as concept-based, focuses on high-level access to images, borrowing methods from library and information science, and concentrating on the use of text to create metadata useful for retrieval, information which is especially important since it is not available from the images themselves.

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 productions and the consequent mushrooming of such collections in recent decades, including television news libraries and stockshot libraries. To be useful, these collections need to be catalogued and indexed at the shot level. A number of traditional library science methods are in use, such as the establishment of a controlled list of descriptors, the application of a document classification scheme, or the construction of a thesaurus. Lists of descriptors prove insufficient once the collection becomes large. Traditional classification schemes adapt poorly to describing both still and moving image materials. Thesauri work best with specialized collections, while the materials contained in film libraries, television libraries, and both film and television stockshot libraries is of a general nature. It has been thought that a thesaurus for
indexing everyday film and video materials would also become unmanageable after a certain
point because there are so many kinds of persons, objects and events to describe that the
semantic networks would become uncontrollable. However, film and television librarians
often create a thesaurus from scratch when building an information system for their
collection, and term creation levels off at a certain point because terminology for indexing the
great majority of shots has been created by then.

What is the point at which this happens? 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 a unique tool is required for each collection? Would it be
reasonable to try to construct a general thesaurus naming everyday objects, events, and
categories of persons that could be shared among moving image collection managers? These
research questions are the focus of this study.

2. Goals and Objectives of the Study

The general goal of this study is to reach an understanding of the organization of
existing vocabulary management tools in special libraries with moving image materials, by
identifying and studying patterns, and with a view to building a uniform vocabulary available
for broad use by those who manage collections of moving images.

The specific objectives are:
• to discover how many terms, excluding proper names, are contained in a controlled
vocabulary for managing general moving image collections when term creation
starts to level off;
• to identify terminological and structural patterns in existing thesauri created to
describe moving image collections;
• to assess how these patterns can contribute to the design of a shared vocabulary
useful for special collections containing general material.

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 consisted 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 on the
following: the personality of the holding institution itself, the characteristics of its collections,
the collection management policies, 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.
The research assistant for the study followed up with visits to the participating institutions,
where structured interviews were conducted with resource persons. Where possible, the
vocabulary management tools were also examined. Specific quantifiable data were sought, including the number of common names found in the thesaurus, the number of proper names, and the number of semantic networks. In addition, the entire content of entries for randomly-selected letters of the alphabet was collected, to permit a comparison of the various lexicons at the data analysis stage. The research assistant made further notes on other aspects of importance or of use to this study, such as the way geographical information was organized, whether there were authority lists for proper names, any noteworthy methods used in the construction or management of the vocabularies, and so on. Where participating institutions permitted it, a copy of the entire thesaurus was 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 much expenses. Data collection is not yet finished, but we note that some of the most useful data received to date has come from the local institutions. 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. A participant was also found in San Francisco, and that city was added to the original list.

The third phase of this project will involve an analysis of all 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 will also want 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 will be undertaken. A merged list of the terms accompanied by a count of frequency across the thesaurus 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

At the time of this writing, a total of seven participating institutions have been visited. The results reported here are obviously incomplete. They have not been analysed, and correlations between various data elements cannot yet be established.

Of the seven institutions that have provided data, four identify themselves as stockshot libraries, two as Archives, and one as News Video Archives. Five of the seven institutions have been founded after 1980.

Collections are characterized as general in contents in all but one institution. Three settings also declare that some areas of their collection are more developed than others. Collections contain films (five out of seven), videos (six out of seven), and new formats such as laser discs are listed by one institution. Wide differences (from 750 to 100,000) are observed in numbers of titles/hours/volume available in those seven institutions; we note that most institutions can only provide an estimate of what is available in their own collection. Fiction and non-fiction cohabit in three institutions while a separate category Current News/Events is present in four settings. Two centres cannot, or will not, provide information on the nature of their collection.

Moving image collections can be processed at one or more of three levels: at the title level, at the sequence level and at the shot level. In our sample so far, four libraries declare that they index at all three levels, one at the shot level only, and one at the macro-level of "library rolls". One institution does not provide any human-based indexing in the processing of its collection. Of those six institutions where indexing is done, four use a list of keywords, three use an in-house thesaurus, two use a list of subject headings, and one an in-house
classification. Table 1 shows that most institutions do in fact use more than one indexing tool concurrently.

<table>
<thead>
<tr>
<th></th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
<th>I5</th>
<th>I6</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of keywords</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>In-house thesaurus</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>List of subject headings</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-house classification</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncontrolled vocabulary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table 1. Indexing tools**

Indexing policies and guidelines are not available in half of these six institutions. Not surprisingly, Table 2 shows that the depth of indexing (i.e. the average and total number of subject descriptors use to describe images) varies considerably.

<table>
<thead>
<tr>
<th></th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
<th>I5</th>
<th>I6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing policy</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Average number of descriptors</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>?</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Maximum number of descriptors</td>
<td>No mx</td>
<td>No mx</td>
<td>No mx</td>
<td>No mx</td>
<td>10</td>
<td>No mx</td>
</tr>
</tbody>
</table>

**Table 2. Depth of indexing**

As seen in Table 1, three institutions only, to date, declare using a thesaurus for moving image collections management. At first sight, these thesauri appear to vary more widely than we expected in their essential characteristics (see Table 3).

When asked what sources were used to develop the thesaurus, all three institutions reported using a mix of primary documents, of users requests, and of terms found in other thesauri; this is in accordance with current thesaurus development practices. None of our interviewees could tell us off hand how many of the total number of terms available in the thesaurus were common names, but one institution could specify that 220 000 of its 344 000 terms were persons names. The number of lead-in terms identified by two institutions is so small in comparison to the number of descriptors that their real usefulness is questionable. Not surprisingly, the oldest thesaurus has a full relational structure while the newest one shows only associative relationships at this time. A complete analysis will be done, however, to determine whether these structures do indeed conform to international standards in the area of thesaurus development.

Vocabulary managers were asked to estimate the rate of growth of the thesaurus lexicon by indicating what percentage of thesaurus terms was already present in the thesaurus at the end of its first year of existence, at the end of its third year of existence, at the end of its fifth year of existence, and so on. This proved to be a more difficult question than we had anticipated; it appears that proprietary software used to manage the vocabularies does not provide precise enough chronological data on the integration of terms in the lexicon. Our first resource person was confident in giving us a figure of 70%, while the second resource person did not even risk a guess at what this number was. The third thesaurus has not yet been in existence for a full year.
<table>
<thead>
<tr>
<th></th>
<th>Thesaurus 1 (15)</th>
<th>Thesaurus 2 (16)</th>
<th>Thesaurus 3 (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of creation</td>
<td>1987</td>
<td>1980</td>
<td>1999</td>
</tr>
<tr>
<td>Sources of terms</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Coverage</td>
<td>Mixed</td>
<td>General</td>
<td>Mixed</td>
</tr>
<tr>
<td>Total no of terms</td>
<td>13 444</td>
<td>344 500</td>
<td>1 250</td>
</tr>
<tr>
<td>Nouns</td>
<td>Not known</td>
<td>Not known</td>
<td>Not known</td>
</tr>
<tr>
<td>Lead-in terms</td>
<td>Not known</td>
<td>8 850</td>
<td>50</td>
</tr>
<tr>
<td>Geographic names</td>
<td>1 716</td>
<td>Not known</td>
<td>Not known</td>
</tr>
<tr>
<td>Persons names</td>
<td>Not known</td>
<td>220 000</td>
<td>Not known</td>
</tr>
<tr>
<td>Relational structure</td>
<td>Full</td>
<td>Full</td>
<td>Associative only</td>
</tr>
<tr>
<td>Rate of growth</td>
<td>70% first year</td>
<td>Not known</td>
<td>100% first year</td>
</tr>
<tr>
<td>Updating</td>
<td>As needed</td>
<td>Daily</td>
<td>As needed</td>
</tr>
<tr>
<td>Responsibility</td>
<td>one person</td>
<td>More than 2 people</td>
<td>two people</td>
</tr>
<tr>
<td>Descriptors added/year</td>
<td>less than 50</td>
<td>More than 300</td>
<td>more than 300</td>
</tr>
<tr>
<td>Software support</td>
<td>Proprietary</td>
<td>Proprietary</td>
<td>commercial</td>
</tr>
</tbody>
</table>

**Table 3. Thesaurus characteristics**

All three thesauri are updated regularly. In one institution (15), updating is a centralized process for which one person only is responsible; less than 50 new terms are currently added every year. In stark contrast, all indexers are allowed to update the vocabulary in a second institution (16), with the result that more than 300 terms are still added to the lexicon on a yearly basis even though the thesaurus has been developing since 1980.

**5. Discussion / Conclusion**

The data obtained so far support our assertion that the indexing of moving image collections is still, in most environments, an ad hoc, little standardized, process.

Based on what we have seen until now, the most important tendency that seems to be taking shape is that public institutions have the best and most elaborate tools while collections in private institutions suffer from the expediency of the profit motive. Because of drastic budget cuts in the past decade, with the consequent reduction of available resources, the quality of the tools used in all institutions may have deteriorated to the extent that collection managers are no longer able to provide the level of detail in description and indexing that is essential to high-quality user services.

The weakening of the once carefully built tools used for managing moving image collections in public institutions and the evident lack of long-term organizational strategies for managing collections in private institutions point to the urgency of developing ways to improve the situation. The rather sloppy methods now in use will break down as collections become too large to manage without good tools and stricter processing guidelines. One source of hope for improvement may reside in preliminary results obtained in another one of our projects, which indicate that high-quality indexing can be derived automatically from text sources created for other purposes during the production and post-production stages. If these resources are tapped, and if good algorithms are written to derive the indexing and connect it
to the shots to which it refers, much of the work now done manually can be automated. A thesaurus listing the most frequently used names of objects, categories of people, events, etc. could then serve as a behind-the-scene authority to standardize forms and meanings, and to ultimately facilitate retrieval.

Acknowledgments
This work is 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 of Atlanta, and Renée Provitt and Janice Simpson of Los Angeles, for which we thank them.