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

Conceptual and lexical compatibility in thesauri used to describe and access moving image collections

Abstract: The term-to-term comparison method was used to identify various types and levels of conceptual equivalence among five controlled vocabularies used for content representation in collections of non-art moving images. It was found that conceptual overlap is high enough to justify the pursuit of research and development work on a common basic indexing and access language that could be used to name categories of persons, objects, events, and relations most frequently depicted in non art moving image collections.

Résumé: Nous avons utilisé la méthode de comparaison terme-à-terme pour identifier divers types de relations et niveaux d'équivalence conceptuelle entre cinq langages documentaires utilisés pour la représentation du contenu dans des collections d'images en mouvement non artistiques. Les résultats de l'exercice démontrent que la compatibilité conceptuelle est suffisamment élevée pour justifier la poursuite des travaux visant le développement d'un langage documentaire commun utilisable pour nommer les catégories de personnes, objets, événements et relations les plus souvent décrits dans les collections d'images en mouvement non artistiques.

1. Background

It has been suggested by Turner (1990) that only the pre-iconographic or primary level of picture description, the "ofness" level in Shatford's words (1986), was truly useful to provide access to stockshots in moving image collections. Most collections described at the shot level are indeed indexed this way, and the descriptors chosen to represent visual content name the beings, objects and happenings seen in the shots (e.g., cat sleeping on a chair) rather than abstract concepts (e.g. serenity and comfort). Here as in more "traditional" collections, the use of a controlled vocabulary as indexing and retrieval tool is of great interest. The lexical and structural control offered by all types of controlled vocabularies contribute greatly to improving access to the content of collections, to reducing noise and silence in retrieval, to improving precision and ultimately to satisfying users by giving them what they need more quickly and at lower costs. A few thesauri have been created specifically for indexing visual documents such as art images, photos, slides and plans. The best known of these are the *Art and Architecture Thesaurus,* developed and maintained by the J.P. Getty Foundation, and the *Thesaurus for Graphic Materials*, published by the Library of Congress. In Canada, the National Film Board uses its home-grown thesaurus to index its collection of stockshots.

In the global information society, producers and users of moving images agree that it is becoming critical that common methods of shot-level and scene-level description be developed to foster retrieval, and to ultimately facilitate resource sharing. It appears reasonable then to envision a common indexing and access language that could be used to name categories of persons, objects, events, and relations most frequently depicted in non art, ordinary images. Furthermore, anecdotal evidence suggests that a limited number of terms should be sufficient for describing general collections of images, a phenomenon similar to what can be observed in the use of natural language, in which the number of words available is much greater than the number of words required on a daily basis for non specialized communication and discourse (Deweze (1981, 363).

Our current project is a follow-up to a research conducted by Turner and Hudon between 1999 and 2001 under the title “Organizing moving image collections for the digital era: developing metadata standards for moving images”[1]. The general goal of that project was to survey the techniques and tools used for representing the content of moving image collections that are indexed shot by shot. As we were particularly interested in indexing languages and their structure, specific objectives were

- to determine the approximate number of terms, excluding proper names, that are used to describe North American moving image collections indexed at the shot level;
- to estimate the rate of growth of term creation in controlled vocabularies used to index the content of moving image collections;
- to identify terminological and structural patterns in existing controlled vocabularies developed to index the content of moving image collections;
- to assess whether these patterns could contribute to the design of a shared vocabulary usable to describe and access encyclopaedic collections of moving images;
- to evaluate the possibility of creating a universal indexing vocabulary for general collections of moving images, those that represent everyday objects and events.

Thirty-three organizations were initially contacted, and data were eventually collected from 11 organizations through questionnaires and follow-up interviews. These organizations were television networks and movie production studios on both sides of the Canadian-American border, and they managed among themselves a total of 14 collections. Not surprisingly, our data confirmed that the organization and exploitation of moving image collections remains heavily dependent on *ad hoc* information systems structured around locally established methods and tools (Hudon, Turner, and Devin, 2000; Turner and Hudon, 2002).

All of our participating organizations managed structurally complex databases which fostered more or less effective retrieval of pictures representing specific situations or objects. Almost all of the collections (11/14 or 79 percent) were catalogued and indexed by title or whole document; this was to be expected, given the ease in obtaining title information and of its importance for retrieval. Some of the collections were described and indexed more deeply, at the sequence level (5/14 or 36 percent) or at the shot level (8/14 or 57 percent). Controls of indexing practices included using the *Library of Congress Subject Headings*, a commercial or home-grown thesaurus, a list of keywords developed for the purpose, or some simple classificatory structure, as well as combinations of various techniques. Although severe time, budget and human resource constraints obviously made it difficult to invest as much as needed in the development of standardized tools for collection management and exploitation, we found that six out of 11 organizations (55 percent) were using one or more controlled language tools for content representation.

We noted easily the very broad range of domains described in these language tools, a reflection of the range of general and specific subjects necessary to respond to the needs of the varied clienteles of television networks and movie production studios. We observed as well the important proportion of proper names in the lists (close to a third in each vocabulary). We finally noted the extremely small proportion of non-descriptors or lead-in terms; from this we concluded that the tightening of the vocabulary by synonym control had most likely not been effected.

In the last phase of the project, a simple analysis was conducted to determine whether there was a significant degree of lexical redundancy in the six vocabularies used for naming everyday objects, events and relations depicted in our collections. As sample lexicon, we used the set of terms beginning with the letters F, I, and R in all six vocabularies. These letters were chosen at random within the 15 letters used as initial letter in a minimum of 900 and a maximum of 5 000 words in the English language. Numbers, names of persons and institutions, titles, geographic names, etc. were first identified and removed from each list. The remaining terms were then combined into a single list of 2 292 distinct terms. Of this number, 1 858 (81 percent) represented concrete objects or entities, and 434 (19 percent) expressed abstract concepts. Frequency counts yielded surprising information: out of a total of 1 858 nouns, only seven were present in all six lexicons, while 1 680 appeared only once in the merged list of all terms beginning with an F, and I, or an R. This result obviously weakened our assumption that the number of terms necessary to describe the content of non art images was limited, and that these terms were likely to be the same in all collections of images describing daily life objects and events. Because of the simplicity and non scientific character of the process, and because we had observed that synonym control did not appear to have been effected, we suggested that a deeper analysis of the lexicons, taking into account concepts rather than just lexical forms, was needed before we could declare the incompatibility of those six indexing and access tools.

In our current project[2], we applied principles and methods described in previous research on the compatibility of indexing languages. Most popular in the seventies, this area of research has become active again over the past decade, in the wake of global expansion of information networks and of the increasing ease with which distinct collections can be accessed simultaneously, or even merged, whether in actual practice or virtually. In such a context, Lancaster and Smith have been proven right over and over again, they who observed that “while controlled vocabularies tend to promote internal consistency within information systems, they also tend to reduce intersystem compatibility” (1983).

In this project, our main goal was not to try and reconcile, harmonize, unify several indexing languages, as was the case in most previous applied research endeavours concerned with compatibility. Our primary objective was rather to estimate levels of conceptual redundancy in several controlled vocabularies currently used to index moving image collections. Common concepts and shared verbal representations could indeed be used as a basis in the development of a common controlled vocabulary usable for representing categories of persons, objects and events depicted in general collections of non art moving images. A secondary objective was to test the efficiency of a simple methodology for estimating levels of conceptual redundancy in controlled indexing and retrieval languages more generally.

2. Methodology

Compatibility of controlled indexing and retrieval languages can be measured at one or more of four levels. Lexical compatibility exists between terms; conceptual compatibility goes beyond terms to uncover similarities and differences in sets of concepts represented; structural compatibility is to be found in the network of interpreting and conceptual relations, and subject compatibility refers to the possibility for two or more indexing languages to represent the same subject, whether by means of a single descriptor or by a combination of terms (latter type defined by Riesthuis, 1996). In this project, we were interested in lexical and in conceptual compatibility, while indirectly looking also at subject compatibility.

Our sample was composed of five of the six indexing vocabularies used in the project described in the previous section of this paper. Although these indexing vocabularies are all referred to as thesauri within the organization that created, maintains, and uses them, three out of five are in fact lacking the relational structure that would make them into a true thesaurus. All five indexing tools have been designed in-house for the purpose of representing the content of encyclopaedic collections of non art moving images in the television and movie industries, and they do so at various levels of specificity. At the request of the commercial organizations which participated in the 1999 study, the thesauri are here again identified only as T1 to T5. As already mentioned, all tools include a significant number of names.

The sixth thesaurus was eventually dropped from this study because it has been created most recently and did not include the minimal number of terms which would have made any comparison interesting.

As source lexicon, we used the set of terms beginning with the letters F, I, and R, in T1, which offers the most extensive list of potential descriptors. T1, a very large tool containing more than 344 000 terms, is one of the least "controlled" of all indexing tools under examination. It is updated on a daily basis and at will by a number of people who are using it at the time of cataloguing, in a profit-oriented organization. The very large list of approximately 11 500 terms thus obtained was reduced by eliminating most names (of persons, programs, brands, ships, etc.) and all titles (songs, books, movies, etc.) provided in this indexing language. Country names were kept however, since they tended to be listed in all five thesauri. Acronyms and terms including obvious spelling or typographical errors (such as e.g. *Indondesian people* or *Fource ouvrière*) were also deleted since the developed and correct forms for most of them were found elsewhere in the list. Long series of quasi-identical terms (e.g. *Flags (Afghanistan)*, *Flags (Alabama)*, *Flags (Alaska)*, *Flags (Albania)*, etc.) were cut short as we judged that having them all (close to 150 occurrences for Flags only) added nothing to the comparison.

The next step consisted of further reducing the source list by identifying lexical and conceptual equivalents within T1. Lexical equivalents are terms that appear in various forms such as *Farm houses* and *Farmhouses*, or *Factory* and *Factories;* once lexical equivalents had been identified, plural and singular forms, pre-coordinated and post-coordinated forms, and direct and inverted forms of the same term had been brought together so that they would count as one source concept / term only. Conceptual equivalents are synonyms and quasi-synonyms, such as *Farm workers* and *Farm labourers*, *Receptionists* and *Reception clerks*, *Foreign relations* and *International relations*. Identification of conceptual equivalents within our source list further reduced our bank of concepts / terms to a total of 1 424 units which were then used as a basis for comparing the contents of all five thesauri.

The term-to-term comparison method described by Dégez (1998) was chosen to compare the lexical and conceptual contents of our five language tools. Each term from T1 was first used as access point to search the Wordnet[3] term base; this provided a more extensive set of the various forms of expression of a particular concept. Each component of this synset was then used to search the entire lexical content of each one of T2, T3, T4, and T5.

To identify concept / term equivalents, we relied on the simple process of determining which term(s) would be selected to index a resource providing information on the source concept (i.e. T1 concept / term) in, for example, T2, if T2 did not include in its own lexicon the term suggested in T1. The comparison involved identification and coding of the following types of relationships:

1. Exact lexical equivalence (e.g. T1Federal buildings, T2Federal buildings=);

2. Exact conceptual equivalence (e.g. T1Felicity, T5Happiness=; T1Infirmaries, T4Hospitals=);

3. Exact equivalence through term combinations (e.g. T1Federal buildings, T4Federal+Buildings=; T1Reefs [Artificial], T4Artificial+Reefs);

4. Partial equivalence through hierarchy (narrow to broad) (e.g. T1Fast food restaurants, T2Restaurants>; T1Flower shows, T2Exhibitions>);

5. Partial equivalence through hierarchy (broad to narrow) (e.g. T1Ranches, T3Horse ranches<; T1Flight crews, T5Pilots<);

6. Partial equivalence through association (e.g. T1Recipes, T3Food preparation–; T1Referendums,T5Voting–; T1Rampart, T2Defence –);

7. Non equivalence (e.g. T1Face recognition; T1Recounts; T1Intramuros).

Hierarchical and associative relations were established in conformity with international guidelines for thesaurus development (International Organization for Standardization, 1986), with the exception that the whole-part relation was systematically considered as an acceptable hierarchical relation.

As exemplified in Table 1 below, no equivalent for the source concept / term *Faith healing* could be found in either one of T2, T3 and T4. On the other hand, the source concept / term *Film festival/s*, which appears in both singular and plural form in T1, has an equivalent in each one of the other thesauri: T5 lists an exact lexical equivalent, while a lexical equivalent is made available in T2 and T4 through a combination of terms. T3 provides a generic term which we establish as partial equivalent (narrow to broad). *Infanticides* and *Rocking horses* are other examples of equivalents obtained by a combination of otherwise unrelated terms.

Source concepts / terms *Fuselage*, *Icelandic people*, *Racoons*, and *Receptionists*, show through several examples that we have used a broad definition of the hierarchical relation to determine partial equivalence (narrow to broad); the source concept / term *Racoons* is related to *Land mammals* through a standard generic relation, while *Fuselage* is a <part of> an *Aircraft*. Finally, the *Recycling* example shows a standard partial equivalence through association, demonstrating that broad context relations were also taken into account. All of these relations were seen as indicative of conceptual overlap.

Where source terms could not be disambiguated (e.g. *Indexing*, *Interns*, *Runways*), any term found in the target tools which could be considered equivalent to any one of the potential meanings of the source term was coded as exact conceptual or partial equivalent.

T1	T2	T3	T4	T5
Faith healing	-	-	-	Faith healing =
Film festival/s	Films + Fairs and festivals =	Festivals >	Film + Festivals =	Film festivals =
Fuselage	Aircraft >	Fuselage (Airplanes) =	Fuselage =	Aircraft >
Icelandic people	People >	Ethnic groups >	Scandinavian + People >	-
Infanticide	Murder + Babies =	Crime >	Murdering + Infants =	Infanticide =
Racoons	Land mammals >	Racoons =	Racoons =	Racoons =
Receptionists	Clerks >	Receptionists =	Reception + Clerk =	Office workers >
Recycling centers	Recycling –	Recycling –	-	Recycling –
Rocking horses	-	Toys + Horses =	Rocking + Horses =	Toys + Horses =
Runways	Airports >	Runways (Aeronautics) =	Runways =	Airports >

Table 1. Examples of coding

Our original plan to code fine distinctions between terms representing objects, persons, happenings or events, and abstract concepts, with a view to establishing correlations between, for example, type of concept represented and non equivalence, had to be temporarily cast aside for lack of time and resources.

3. Results

Once all 1 424 source concepts / terms had been examined and coding of equivalents had been completed, we used the total of exact and partial equivalents to estimate the degree of conceptual compatibility between T1 and each one of T2, T3, T4, and T5. At this time, internal comparison within T2 and T3, T2 and T4, etc., has yet to be made.

Tables 2 and 3 show the results of the comparison process, first with actual numbers of concepts / terms, then using percent figures. The value of *N* represents the number of distinct terms present in T1. Numbers given for T2, T3, T4 and T5 are not indicative, however, of the actual number of distinct terms present in these lexicons; in the target thesauri, the same term may appear in different types of relations (e.g. T1Insurance, T2Insurance=; T1Insurance [Life], T2Insurance >), and the same term may be used as lexical or partial equivalent for several source terms (e.g. T3Occupations is considered a partial equivalent through hierarchy (narrow to broad) for T1Financial planners, T1Image consultants, and T1Repairmen among others). The few cases of multiple equivalence within one lexicon (e.g. T1Ions, T5Atoms and T5Particles) have been counted as one equivalent only.

In the following tables, levels of compatibility can be estimated by reading the Sub-total lines in each table.

T1		T2	T3	T4	T5
N =1424	Exact lexical equivalence	74	223	376	265
	Exact conceptual equivalence	27	84	95	64
	Exact equivalence (combination)	16	22	446	43
	Partial equivalence (hierarchy broad)	399	429	271	444
	Partial equivalence (hierarchy narrow)	4	9	2	5
	Partial equivalence (association)	191	141	71	152
	SUB-TOTAL	711	908	1261	973
	Non equivalence	713	516	163	451
	Total	1424	1424	1424	1424

Table 2. Results of the comparison (actual number of concepts / terms)

T1		T2	T3	T4	T5
1424=100%	Exact lexical equivalence	5.2	15.7	26.4	18.6
	Exact conceptual equivalence	1.9	5.9	6.7	4.5
	Exact equivalence (combination)	1.1	1.5	31.3	3.0
	Partial equivalence (narrow to broad)	28.0	30.1	19	31.2
	Partial equivalence (broad to narrow)	0.3	0.6	0.1	0.4
	Partial equivalence (association)	13.4	9.9	5.0	10.7
	SUB-TOTAL	50	63.7	88.5	68.4
	Non equivalence	50	36.3	11.5	31.6

Table 3. Results of the comparison (percentages)

The following observations can be made:

1. As could be expected, conceptual compatibility (i.e. the total of Exact conceptual equivalence, Exact equivalence (combination), and all three cases of Partial equivalence) is higher than exact lexical equivalence.

2. T1 and T4 show the greatest conceptual overlap (88.5 percent). The very high number of conceptual equivalents in T4 is explained by the nature of this tool, in fact an extensive list of keywords (uniterms) which can be combined in sets of two, three or four terms to form a very large number of compound and complex terms and expressions.

3. T1 and T2 are the least conceptually redundant tools. T2 is a fully structured bilingual thesaurus in which we have observed significant gaps, particularly in the hierarchies.

4. Partial equivalence through hierarchy (narrow to broad) (e.g. T1Ragweed, T4Weeds) is the most frequent relation in three thesauri out of four. This suggests that the most important difference between T1 and the other tools is a difference in specificity rather than a difference in coverage.

5. Overall, and when taking all types of equivalence into account, levels of conceptual compatibility (or redundancy) reach 50 percent or more in all four target thesauri. Although there is no benchmark figure available for this type of study, we believe that this would be considered a particularly good level of compatibility in indexing languages of an encyclopaedic nature.

6. The value of standard related terms as conceptual equivalents could be contested. However, even if the numbers for partial equivalence through association are excluded from the final

count (see Table 4), levels of conceptual compatibility (or redundancy) still reach 50 percent or more in three target thesauri out of four, and in all cases, there remains a noticeable difference between percentages of conceptual equivalence and percentages of non equivalence.

T1		T2	T3	T4	T5
1424=100%	Exact lexical equivalence	5.2	15.7	26.4	18.6
	Exact conceptual equivalence	1.9	5.9	6.7	4.5
	Exact equivalence (combination)	1.1	1.5	31.3	3.0
	Partial equivalence (narrow to broad)	28.0	30.1	19.0	31.2
	Partial equivalence (broad to narrow)	0.3	0.6	0.1	0.4
	SUB-TOTAL	36.5	53.8	83.5	57.7
	Non equivalence	50	36.3	11.5	31.6

Table 4. Results of the comparison (percentages) – Excluding Partial equivalence (association)

A total of 517 T1 concepts / terms (36.3 percent) have one or more equivalents in each one of the four other tools. On the other hand, cases of absolute non equivalence (concept / term in T1 having no equivalent in any one of the other tools) are rare (97 or 6.8 percent). Non equivalence is observed with highly specific concepts / terms (e.g. *Fetal protection plan*, *Introverts, Ransom notes*), as well as with ill-defined or very general ones (e.g. *Flywheels*, *Imagery, Ratification*).

4. Discussion and conclusion

The determination of conceptual equivalents is a complex process which involves a good amount of subjectivity. In our project, rules were established as we were working through the list and retroactive editing was often needed. Although great efforts were made to ensure high consistency levels in identification of relation and equivalent types, for example through a second, independent coding of the data by a different individual, interpretation and coding errors may still have resulted from the lack of definitions and relational structure in T1, T4, and T5, and the ensuing difficulty of determining the extension of a concept and the exact meaning of its verbal representation (as with, for example, *Impact records*, *Radicals*, *Receiving lines*, *Runways*).

In this project, we worked with a restricted sample for reasons of convenience: a few sections only of the source thesaurus T1 were made available to us. While recognizing this limitation, we are confident that our methodology and the care taken in comparing the lexicons confer a reasonable degree of validity to our results. The term-to-term comparison approach is a simple methodology that has proven quite efficient in our study, producing clear and easily interpretable results.

A closer examination of our thesauri also led to observations on their own internal quality and coherence. T2 and T3 offer a standard relational structure but since their lexicons lack many important concepts, hierarchies are remarkably porous. T1 and T4 are alphabetical lists of words and/or expressions that may be easy to use but are in need of serious editing, on the conceptual as well as on the lexical levels. T5, also an alphabetical list of descriptors, has been carefully edited, but remains incomplete as far as essential concept representations are concerned.

Managers of moving image collections are aware of these deficiencies in their indexing tools. As early as 1999, several of them were already expressing their interest in a common controlled vocabulary for representation and access which they cannot, however, develop and maintain themselves because they lack the expertise, time, money, and human resources to do so. The results of this small project have allowed us to estimate more precisely, fairly, and realistically the actual degree of conceptual compatibility in five controlled vocabularies currently used to describe and access the content of non art moving image collections in North America. It was found that the conceptual overlap in these tools appear high enough to justify the pursuit of research and development work on a common basic indexing and access language that could be used to name categories of persons, objects, events, and relations most frequently depicted in these collections.

Notes

[1] Project funded by the Special Libraries Association, through its Stephen I. Goldspiel Memorial Grant. James M. Turner, principal investigator, and Michèle Hudon, co-investigator.

[2] Project funded through an internal grant from the Faculté des Arts et Sciences, Université de Montréal. M. Hudon, investigator. Sincere thanks to Touria Fadaili and Daniel Gagnon, research assistants for this project.

[3] http://www.cogsci.princeton.edu/~wn/

References

Dégez, Danièle. 1998. Compatibilité des langages d'indexation : Mariage, cohabitation ou fusion? : Quelques exemples concrets. *Documentaliste & Sciences de l'information* 35 (1) : 3-14.

Deweze, André. 1981. *Réseaux Sémantiques : Essai de modélisation : Application à l'indexation et à la recherche documentaire*. Lyon : Université Claude Bernard.

Hudon, Michèle., James M. Turner and Yves Devin. 2000. How many terms are enough? Stability and dynamism in vocabulary management for moving image collections. In *Dynamism and stability in knowledge organization : Proceedings of the Sixth International ISKO Conference.* Wurzburg, Germany : Ergon. pp. 333-338.

International Organization for Standardization. 1986. *Guidelines for the establishment and development of monolingual thesauri.* ISO 2788-1986. Geneva : ISO.

Lancaster, F.W. and L.C. Smith. 1983. *Compatibility issues affecting information systems and services*. Paris : UNESCO.

Riesthuis, Gerhard J.A. 1996. Theory of compatibility of information languages. In *Compatibility and integration of order systems : research seminar proceedings of the TIP/ISKO meeting, Warsaw, 13-15 September, 1995; Warsaw, Poland*. Warsaw: WYDAWNICTWO SBP. pp. 23-31.

Shatford, Sara. 1986. Analysing the subject of a picture : a theoretical approach. *Cataloging & Classification Quarterly* 6, 3 : 39-62.

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 (4): 1-22.

Turner, James M., and Michèle Hudon. 2002. Organizing moving image collections for the digital era : research results. *Information Outlook* 6 (8) : 14-25.