Harald H. Zimmermann, Saarbrücken

Conception and application possibilities of classification concordances in an OPAC environment

In: Dahlberg, Ingetraud; ISKO (Hrsg, 1996): **Compatibility and integration of order Systems: research seminar proceedings of the TIP/ISKO Meeting. Warschau 13-15 September 1995. Warschau** Wydaw. SBP
Version: 1.0; Date: 1995-9-13; DOC: E37ZWS1/warschau.txt

1. **Introduction**

An online public access catalog (OPAC) is defined as an electronic data bank based on machine readable catalog sheets of a library.

Libraries normally used to provide such sheets (or cards) for internal administrational purposes as well as for users. The main function - similar to the "paper" or sheet version of the catalog - still is to identify the lending number of a book to support the lending process in a library.

Compared to data of special information services, the content of an OPAC is very poor (e.g. no abstracts). In addition, the data itself is based on monographs (normally, an OPAC doesn't refer to scientific articles), i.e. it doesn't represent the "world of knowledge" in an area. Instead, it represents the "books" available in a specific library collected by some librarians and institutes which are using this library as service institute.

Besides, the full text retrieval with an OPAC is limited to the words of the titles of a book. In addition, there might exist a kind of catch word or key word category so that the content might be represented by such words.

To some degree, also notations of classification systems (homemade or external ones) are used to describe the content of OPAC entries. But there is, at least in the area of "scientific librarianship" in Germany, no agreement about a standard classification system to be used.

In addition, within (smaller) libraries, domain specialists or experts are not available to be able to provide high-sophisticated classification. As a result, OPACs exist using different classification systems and with a different depth of classification.

2. OPACs and the future role of libraries in classification

The more a user is equipped with techniques like online access or CD-ROM facilities, the more he is looking for appropriate information systems.

Most users (especially students) have in mind that "their" OPAC (for example, the OPAC of a university library), would provide an interesting tool for information. To some degree, i.e. for some "first" access to information, this might be true. But in general, such an "idiosyncratic" OPAC does not solve sufficiently the completeness problem of a data base, especially compared to other services, e.g. patent information or domain specific information services.

On the other hand, the "cumulation" of OPACs (i.e. the building of so-called shared catalogs) could reduce the problem of incompleteness, at least on behalf of monographs.

There are many reasons for cooperation between libraries, especially to reduce administration costs. The different classification systems used for several years are an obstruction for such a cooperation.

There are the following ways to solve the problem:

1. One is the "ideal" one, but very unrealistic: All libraries decide on an international standard of classification to be applied worldwide in their OPAC developments. They agree on a distributed coding system where high-sophisticated specialists are coding "books" in there specific domains on the deepest level possible within the classification.

2. Libraries abandon the creation and coding of own classifications and rely on existing classifications in use at important information services and accepted by the experts of such a domain (like in medicine).

3. This would lead to different classifications in such a case where an object belongs to more than one of these "basic" domains, but it would allow the libraries to profit from the expert knowledge when building their (selected) OPACs.
   In this case, the creation and use of concordances between the different classification systems is needed if a user accesses different OPACs or OPACs are cumulated.

4. Libraries continue in creating classifications of their own, but on a distributed way in the sense that they agree on the most-possible depth (within a specific classification) and on doing it by experts of the relevant domain. This leads to relatively "balanced" coding and high compatibility between the different systems and reduces the cost of coding. Also in this case, if OPACs of different types are accessed or cumulated, relatively "simple" concordances will allow a content based query by the user (accessing the system by one of these classifications).

5. Libraries continue in creating classifications of their own on a very shallow level (in general) or in an unbalanced way, i.e. dependent on the experts or even specialisations existing (or not) within the relevant library (as today). If OPACs of such types are accessed or even cumulated, highly sophisticated concordances have to be developed to reduce the content based query problems of the user.

3. The use of a "classification" thesaurus as a concordance tool

In the following, a concordance (or correspondence) between different classification systems is considered as a specialised thesaurus in the sense that the classification elements (= notations) are handled as "artificial words" or entries of this thesaurus and relations are used to indicate the (type of correspondence between the notations of different classification systems.

To explain this concept in general, the structure of a classification system is taken:

Let us assume that A is a class element (notation) of the classification C; A.A and A.B are also elements of this classification, also A.A.A, A.A.B, A.B.A and A.B.B are elements of this classification. By using the relations BT (broader term) resp. NT (narrower term), one can relate these elements in a hierarchical form by building the pairs

(i) A (BT) A.A
A (BT) A.B
A.A (BT) A.A.A
A.A (BT) A.A.B
A.B (BT) A.B.A
A.B (BT) A.B.B

Let us assume that N is a class element of the classification K; N.A and N.B are also elements of this classification, also N.A.A, N.A.B, N.B.A and N.B.B are elements of this classification. By using the relations BT (broader term), one can relate these elements in a hierarchical form by building the pairs

(ii) N (BT) N.A
N (BT) N.B
N.A (BT) N.A.A
N.A (BT) N.A.B
N.B (BT) N.B.A
N.B (BT) N.B.B

When creating a concordance between both classification systems C and K, the simplest case would be if only the names of the notations differ, but the content referred is the same. In this case, the synonym relation SY can be used to generate the concordance, where the reference to the relevant classification is indicated by an attribute:

(iii) A (C) (SY) N (K)
A.A (C) (SY) N.A (K)
etc.

If the classification systems or parts of it differ in the sense that one notation has more than one representation within the other classification, but the "cumulative content" is the same, the relation "partly synonym" (PSY) can be used to describe the relation:

(iv) A.B (C) (PSY) N.A (K)
A.B (C) (PSY) N.B (K)

If the classification systems or parts of it differ in the sense that one class name has one or more than one representation within the other classification and the "content" is overlapping with other classification elements of the related classification, the relation "quasi synonym" (QSY) could be used:

(v)A.B.A (C) (QSY) N.B.A (K)
A.B.A (C) (QSY) N.B.B (K)

What happens, if one classification (within the same domain) is much more differentiated on behalf of the depth (within a subclass)?

To demonstrate the possible solution of such a case, let us assume that classification C consists only of the category A, whereas classification K consists of at least the elements N, N.A and N.B.

Because there exists the relation BT:

(vi) N (K) (BT) N.A (K) and
N (K) (BT) N.B (K),

it will be sufficient to relate A (C) by the relation (SY) to N (K):

(vii) A (C) (SY) N (K),

if there exists a general rule that all narrower terms (notations) of a system where no explicit concordance relation exists are handled as being included in the relation of the notation which is in broader term relation to them.

If such a software doesn't exist, the rule can be used to build explicit PSY relations between all referenced elements of such a classifications:

(viii) A (C) (PSY) N (K)
A (C) (PSY) N.A (K)
A (C) (PSY) N.B (K)

The last general rule to be handled is the case where there doesn't exist a relevant content classification within one of these classifications. This normally indicates that there will be no reference (document) with such a content available in the OPAC. For quality assurance reasons, a "NULL" relation could be used.

5. Application possibilities

5.l Application within a user query

To be able to use such a "classificational thesaurus" for data base access (by the user), the relevant OPAC must indicate the type (= attribute) of "its" classification system to the retrieval (and thesaurus) system.
If the user asks a question by using "his" classification system, the retrieval system applies the concordance rule to transfer this part of the query into the query relevant for the system. This transfer could be indicated to the user.

5.2 Application in creating cumulative OPACs

If cumulative OPACs are developed based on data from different libraries with different classification systems (e.g. in cooperation between libraries), there could be a decision made o applying one (the more explicit) classification out of the different "input" classifications, so that during the retrieval, the notation of this system would be the only resulting notation.

The alternative could be to cumulate all the different notations with the source (or type) indicator. This leads to more complicated queries (because all possible alternatives have to be created with "OR"), but the data are available in their "original" form.

5.3 Application within distributed OPACs

If queries are used to access different (distributed) OPACs by one query, the replacement rule mentioned above (5.1) could be used - dependent on the indicator of the OPAC actually accessed - to provide the "right" query.

5.4 Enrichment procedures

The concordance technique described above can also be used to "enrich" an OPAC data base with "deeper" notations. Let us assume that "classification system C" is "shallow" in general (i.e. with very "broad" content notations, because there were no people or money available to describe the data on a very specialised level), and that there exists a classification system K which is more or less compatible to C in the sense that the notations of "C" used correspond to a high degree with the "broader notations" which are part of the more complex system K. By using the concordance (and by replacing the classification system), the "weaker" notation of the OPAC could be replaced by the subtle notation, if the same document is detected.

6. Integration within "classical" thesaurus applications

The thesaurus based classification concept has an additional advantage: It can be used together with "classical" terms to provide the user with access tools from his "wording" (in natural language) - "via" a (natural language based) descriptor to the relevant (or possible) notation of the classification(s). Highly sophisticated systems would even be able to provide this access in the user's natural language.
Therefore, guided by the "word-based" thesaurus, he or she will be able to select the relevant notation (by browsing techniques) and prepare the "right" search by using this notation.

7. **Final remarks**

The concept looks very simple, but the problems are in the detail. Fore instance, it will not make any sense to build concordances between a very week system and a highly sophisticated system, if the "enrichment rule" is not applied (or not possible). The relations described have to be seen as examples; for practical purposes, this list might be extended.
In my opinion, it makes no sense to struggle for "the only general and universal system" of classification, because the "concordance classification thesaurus" is able to overcome, to a high degree, the problem on the user's side. On the other hand, the rule of "garbage in, garbage out" also is valid in this subject.