International Conference on Library Automation in Central & Eastern Europe
April 11-13, 1996
Budapest, Hungary

THESAURUS FOR SUBJECT SEARCHING AND INDEXING AS PART OF AN INTEGRATED LIBRARY SYSTEM

Sirje Nilbe
Tartu University Library

Abstract

Thesaurus software developed at Tartu University Library as part of integrated library system INGRID offers significant aid for thesaurus construction, subject indexing and searching. Software facilities are described. The most intellectual problems related to thesaurus content are touched upon.

Introduction

Subject access problems have hold much attention among researchers and designers of library systems. However one must allow that the practical results are far from satisfactory. In general the libraries have continued to use the same controlled vocabularies and indexing methods as in card catalogues, indexed by pre-coordinated subject headings like LCSH. There are two essential reasons for this. First, the design of new vocabularies that are more suitable for post-coordinated indexing and searching is expensive and time-consuming. Secondly, the re-indexing all library materials in process of retrospective cataloguing is assumed to be a too time-consuming work. Libraries have always and everywhere experienced lack of staff and funds. So, most efforts for improving subject access are made post factum - by automated processing of records' text and by developing of user interfaces.

The large academic libraries of Estonia have arranged their subject card catalogues systematically by UDC over 40 years. It is considered reasonable to add UDC numbers to electronic records too because there is need for standardized and language independent subject access point on a national and international scale. UDC tables in Estonian will be published this year. UDC numbers have some good searching facilities for experienced user to broaden or narrow the query with ease. But for an unexperienced user the numbers are meaningless. It is clear that there is not much enthusiasm about UDC number searching even among librarians 1. That is why the main facility for subject access for end-users should be verbal.

Software

There is an in-house developed integrated library system INGRID at Tartu University Library. It is based on INFORMIX database management system and runs on UNIX platform. Up to now following modules are implemented: acquisition, cataloguing, including serials control and cataloguing, and OPAC. For vocabulary control and subject searching there is a thesaurus module as part of INGRID too.

A thesaurus software can be very complicated or less complicated. In our system minimal requirements established on thesaurus software are accomplished. It is possible to create conventional relationships as USE, BT, NT and RT. Reciprocal relationship is forming automatically when input is made in one side. J. L. Milstead 2 declares, that automatic reciprocation of relationships is the fundamental criterion for considering a software package to support thesaurus management.

In addition to fields for referred terms, there are fields for English equivalent, scope note and broad subject domain names in a term record. Term validation capabilities are good - it is not possible to create new record for a term already existing in the database, or refer to a term not included to the thesaurus yet. However, relationship validation is not realized automatically. For example errors like the same term standing both on the BT field and RT field are not prevented.

Main display of the thesaurus is alphabetical. We can retrieve any string in terms using left, right and middle truncation. Truncation at right is implicit while searching thesaurus database (not while searching for references in catalogue). Rotated listings formed by this way provide an entry not only for every word but for every part of word. The last is very important because there is a large number of compound words and derivative suffixes in Estonian.

For systematic grouping of terms there are the subject domains, for instance chemistry, economy, medicine. Further classification is implied by BT/NT relationships only.

The use of the thesaurus for indexing is similar to that of for searching. The user opens a new window and displays a desirable part of alphabetical listing of terms. Then he/she can navigate through thesaurus using relationships among terms.

While indexing a document it is possible to add only descriptors included in thesaurus by selecting terms from the list. It means that uncontrolled terms cannot be input. A non-preferred term if used by indexing or searching is automatically replaced by the preferred term. While searching for references, selecting from the list is not compulsory. One can write the search terms on searhing form by using truncation and logical operators AND, OR, NOT as well.

It is possible to display print preview of any term with its relationships and other appropriated information. There is also a facility for printing thesaurus or a part of it, but it is only for staff.

Terms in thesaurus are both in Estonian and in English, the former as the dominant language. The indexing language is Estonian. The English terms are linked to Estonian ones and are searchable only from OPAC if English is chosen for the dialog language.

Due to the very dynamical nature of our thesaurus the full integration of the thesaurus maintenance and database updating is utmost essential. It works so that a change in a thesaurus term causes a global change to the database - all documents indexed by the old version will be changed to the new version of the term. Deleting a term from the thesaurus is possible only if there are no documents linked to this term. It means that one must edit the corresponding records first. It is time-consuming but useful for database quality.

It seems that facilities for vocabulary control offered by INGRID thesaurus module are not more limited than offered by most corresponding modules of commercially available library systems 3.

Content of the thesaurus

We have to face the difficulties connected with the thesaurus content and structure, that are more serious than software problems. There isn't any completed thesauri in Estonian except one for indexing literature on Finno-Ugristics for international reference database URBIS. The general Estonian thesaurus for libraries is being developed in the National Library but it was at initial stage when we began indexing for our online catalogue in November 1994, and it is still far from being completed. So we must build the thesaurus and index documents at the same time.

We use international standards ISO 2788, ISO 5964, ISO 5963, and well-known handbooks 4;5 as guidelines. But there are some specific factors affecting the indexing and thesaurus construction environment in our library. These are the very broad subject and language coverage of collections and the lack of fixed scientific terminology in Estonian.

Tartu University Library is an old and large library. In 2002 it will come to pass 200 years from its foundation. The collections of the library - 3,7 million items - include research monographs, textbooks, fiction and other nonscientific materials, pictures and manuscripts etc. Subject and language coverage comprises exact sciences, natural sciences, social sciences, humanities in Estonian, English, German, Russian, Finnish, Swedish etc. It is very difficult to build a unified conceptual system for indexing all this universe of knowledge.

The thesaurus standards do not work very well in universal library environment. They are oriented to development of indexing thesauri for specialized bibliographic databases. A higher pre-coordination level is needed in a library catalogue than in a specialized database. There are cases where the strong semantical restrictions for establishment of BT/NT relationships may hinder the effectiveness of thesaurus navigation. It seems that application of the end-user oriented thesauri in libraries is increasing. There is imperative need for guidelines and handbooks related to this area.

The other problem in connexion with content of our thesaurus is less library dependent. Great changes are taking place in the conceptual basis of the most subject fields in present-day Estonia. In language usage it brings about extensive synonymity and vagueness in

meanings. Fortunately terminology as a field of applied linguistics has a quite long tradition in Estonia. Today our terminologists and subject specialists are working intensively at vocabularies of several subject fields. We hope that our terminology is soon set and our thesaurus will be an useful tool for indexing and searching in online library catalogue.

References

1. Buxton, A. B. Computer searching of UDC numbers. Journal of Documentation, vol. 46, no. 3, p. 193 - 217.
2. Milstead, J. L. Thesaurus management software. In: Encyclopedia of library and information science, vol. 51. New York: Marcel Dekker 1993, p. 389 - 407.
3. Library systems in Europe: a directory & guide. London: TFPL Publishing 1994, 401 p.
4. Aitchison, J., Gilchrist, A. Thesaurus construction: a practical manual. 2nd ed. London: Aslib 1987, 173 p.
5. Lancaster, F. W. Vocabulary control for information retrieval. 2nd ed. Arlington: Information Resources Press 1986, 270 p.