

IFLA Satellite Pre-Conference of the Classification and Indexing Section  
“Looking at the Past and Preparing for the Future”  
Florence, Italy, 20-21 August 2009



## **Semiautomatic merging of two universal thesauri: the case of Estonia**

**Sirje Nilbe**

National Library of Estonia  
Tallinn, Estonia  
Sirje.Nilbe@nlib.ee

### **Abstract:**

*The paper deals with a project carried out in Estonia during 2007 to 2009 with the aim of merging two subject indexing tools into one thesaurus in order to facilitate subject search in union catalogues and bibliographic databases and to organise subject indexing and authority control work more economically.*

*The term records of the two thesauri were merged automatically, loaded into a database and the resulting compilation underwent quick human editing. The project was financed by the ELNET Consortium (Consortium of Estonian Libraries Network) and involved two participating libraries – the National Library of Estonia and the University of Tartu Library. The management of the new thesaurus will be the responsibility of those three institutions.*

### **1. Two Estonian-language universal thesauri**

In the 1990s, two universal thesauri were developed in Estonia – the thesaurus of the University of Tartu Library and the Estonian Universal Thesaurus which was managed by the National Library of Estonia. The reason for developing two universal, largely overlapping thesauri was the fact that at the beginning of the 1990s when the first e-catalogues and bibliographic databases were created in Estonia, libraries had no

clear vision of the future trends in information technology development and its impact on library automation. The experience in subject cataloguing and managing controlled vocabularies was rather limited as large libraries have traditionally been using classified card catalogues built up according to the classification system (mostly UDC).

The thesaurus of the University of Tartu Library, also called the INGRID Thesaurus, was initiated together with the library's home-made e-catalogue in 1994 and terms were added to it in the course of current indexing. Only those terms were included in the thesaurus which were needed for subject indexing the documents acquired by the university library. In the library system, the thesaurus module was directly connected with the catalogue module. In 1996 the INGRID was provided with a separate web interface enabling to browse the thesaurus and search terms, but the University of Tartu Library never published the thesaurus as a separate publication and it has not been available to other libraries for indexing. The maintenance of the INGRID was the joint responsibility of the classification and subject indexing team.

During the same period the National Library of Estonia developed the Estonian Universal Thesaurus (*Eesti üldine märksõnastik*, EÜM) according to the example of the Finnish General Thesaurus and the UNESCO SPINES Thesaurus. The EÜM was meant mostly for the National Library (which also acts as a parliamentary library in Estonia) but also for the public libraries network and other libraries. The EÜM was developed with a thesaurus management software designed on the basis of FoxPro but nevertheless the thesaurus was planned to be given out as a printed publication. The EÜM was published at the beginning of 1999, the web version was launched in 2006.

## **2. Necessity and preconditions of merging the thesauri**

After the establishment of the ELNET Consortium and the implementation of INNOPAC (the shared integrated library system of major Estonian libraries), both thesauri were taken into use within this system. INNOPAC was launched at full capacity at the beginning of 1999. Most member libraries started to index using the EÜM which was brand new but not yet tested in practice. The University of Tartu Library continued to use their own thesaurus as it was familiar to the users, trusted as

reliable by the indexers, and the library had already indexed over 30 000 bibliographic records on the basis of the INGRID.

The union catalogue of the ELNET Consortium member libraries consists of two databases – ESTER Tallinn and ESTER Tartu. The differences between the two thesauri have most disrupted the information search, subject indexing and authority control in the Tartu database, and also in the Tallinn database due to the copying of bibliographic records. Also, a lot of duplication has been done by the above two libraries in compiling the thesauri – which we definitely cannot afford, taken the constant lack of qualified staff.

The discussion over the necessity of a shared controlled vocabulary and the possible methods for creating it dates back to the year 2000, initially arising in the Classification and Indexing Working Group of the ELNET Consortium. Negotiations were also held between the owners of the two thesauri – the National Library of Estonia and the University of Tartu Library. It was finally decided that the most appropriate organisation for carrying out the project was the ELNET Consortium, and the fastest method was the merging of the two thesauri by a computer programme and the human editing of the resulting compilation.

Several preconditions existed for carrying out this complex project:

- the typological and structural similarity of both thesauri;
- good cooperation between the editorial teams of both thesauri;
- the possibility to involve a software designer with extensive experience in creating thesaurus management software;
- the interest of all member libraries of the Consortium in the project and their consent to cover the project costs from the Consortium's budget.

### **3. Feasibility study**

In the autumn of 2007 the Consortium carried out a feasibility study on the merging of the two thesauri.

The aims of the feasibility study were the following:

- to identify the compatibility of the data structures and to find the best option for merging the data;
- to identify the overlap of terms and relationships between them;
- to identify the approximate amount of logical mistakes evolving in the merging process in order to evaluate the manpower needed for human editing.

During the testing of automatic merging, approximately one third of the most important data of records of the two thesauri were merged (the productive first letters in the Estonian language A, K and T): preferred terms, nonpreferred terms and relationships between terms. It appeared that 32% of the terms overlapped, two thirds of the terms occurred either in one or the other thesaurus. There were 30% of overlapping relationships, 68% of relationships occurred only in one of the thesauri, and 2% of the relationships were conflicting. A conflicting relationship means that one and the same term is in relationship with another term in two different ways, e.g. in both hierarchical and associative relationship.

On one hand, the one-third overlap was a surprise, because everyday work had given the impression that the similarity of the thesauri was more extensive. On the other hand, the reasons for the differences are clear – the INGRID has been designed for the specific needs of a multidisciplinary library of a classical university, while the development of the EÜM has mostly been influenced by the Estonian publishing output and the acquisition policy of the National Library of Estonia as a research library for the humanities and social sciences. The indexing vocabulary requirements of public libraries are more similar to those of the National Library than those of the University of Tartu Library.

In addition, all terms and relationships of the subject field Computer Science were merged during the test. The overlap here was even smaller, including 17% of terms and 15% of relationships. A more exhaustive analysis revealed that one important reason for this difference was the large amount of names in computer science, e.g. the names of computer programmes and programming languages. These names had often been formulated into thesaurus terms according to different rules, resulting in a lot of

names re-occurring in different forms. Another reason for the small overlap was the fact that the content covered by the tested subject field in the EÜM was more extensive than that in the INGRID, including also automatic control.

	<b>A K T</b>		<b>Computer Science</b>	
<b>Terms</b>	13 767		1476	
Overlapping	4446	32%	253	17%
EÜM	6450	47%	943	64%
INGRID	2871	21%	280	19%
<b>Relationships</b>	50 827		4015	
Overlapping	15 087	30%	609	15%
Only in one	34 648	68%	3321	83%
Conflicting	1092	2%	85	2%

Figure 1. Results of the test merging of data

The feasibility study showed that

- the semiautomatic merging of the data of both thesauri is possible;
- the merging must be preceded by the harmonisation of the list of subject fields and extensive manual correction of the subject field specifiers of terms;
- the merged data compilation contains about 3500 logical mistakes;
- the editing cannot confine to the correction of logical mistakes, it should also involve the merging of synonymous terms;
- the capacity of the required editing is approximately 2400 work hours, plus the time needed for programming and computing.

#### **4. Main project**

The time initially planned for software designing, data merging and human editing was one year. According to this the use of the new controlled vocabulary was designed to start at the beginning of 2009. However, the development of the project was not as smooth as desired. The preparation of the documentation and the thesauri, data merging and the realisation of the web design of the new thesaurus were more

time-consuming than expected. On the other hand, the creation of editing software and the editing process itself followed the initial timescale. The new thesaurus, called the Estonian Subject Thesaurus (*Eesti märksõnastik*, EMS) was opened for public use on 14 May 2009.

The project group involved two computer specialists, one of them responsible for data merging, coding and analysis, and the other for preparing the editing software and realising the final user interface and web design. This enabled them to work in parallel which was an advantage as their involvement in this project was an addition to their everyday job.

The harmonising of subject fields was completed at the beginning of September 2008. On 8 September the last amendments and corrections were made in the EÜM and the INGRID, after that both were “frozen”. The merging of data and the loading of the merged data into the database of editing software were carried out during September to December 2008.

	<b>EÜM</b>	<b>INGRID</b>	<b>EMS</b>
<b>Preferred term</b>			
Term	+	+	+
Subject field specifier	repeatable field, numbers and words	repeatable field, words	repeatable field, numbers and words
English equivalent	repeatable field	all in one field, separated by ;	repeatable field
Scope note	+	+	+
Editor's note	+	-	+
UF	+	+	+
BT	+	+	+
NT	+	+	+
RT	+	+	+
<b>Nonpreferred term</b>			
Term			
Subject field specifier	repeatable field, numbers and words	repeatable field, words	repeatable field, numbers and words
English equivalent	repeatable field	all in one field, separated by ;	repeatable field
Editor's note	+	-	+
USE	+	+	+

Figure 2. Comparison of the data elements of term records

The statistical analysis carried out after the data merging of the two thesauri indicated that the overlap of the terms of both thesauri was actually even less than one-third.

Merged	12 223	25,84 %
EÜM	24 399	51,59 %
INGRID	10 674	22,57 %
Total	47 296	100 %

Figure 3. The proportional division of the origin of the terms in the merged thesaurus

### **5. Editing of the new merged thesaurus**

The editing team consisted of 8 persons, all of them previously involved in the management of the initial two thesauri. The whole editing team thus had the necessary competence and experience. The editing period lasted for three months - from the beginning of January until the end of March in 2009. It was additional work for all editors and all of them were paid extra for that. Access to the database was possible via a regular browser which fortunately enabled to do the editing also at home.

At the first stage the workload was divided by subject fields, trying to take into account the knowledge and practical work experience of each editor. At the second stage the thesaurus was divided alphabetically between the editors and reviewed. A specially compiled editing guide could not solve all problems and the editors had to make a lot of independent substantive decisions.

The software created for editing was based on the software of the EÜM, containing several improvements which facilitated editing.

The origin and nature of the terms was indicated by colour coding – overlapping terms were black, those occurring only in the INGRID were blue, and the terms occurring only in the EÜM were green. Red denoted the terms with conflicting

relationships and the terms acting as preferred terms in one thesaurus and as nonpreferred terms in the other. Lilac designated the terms with a merged scope note.

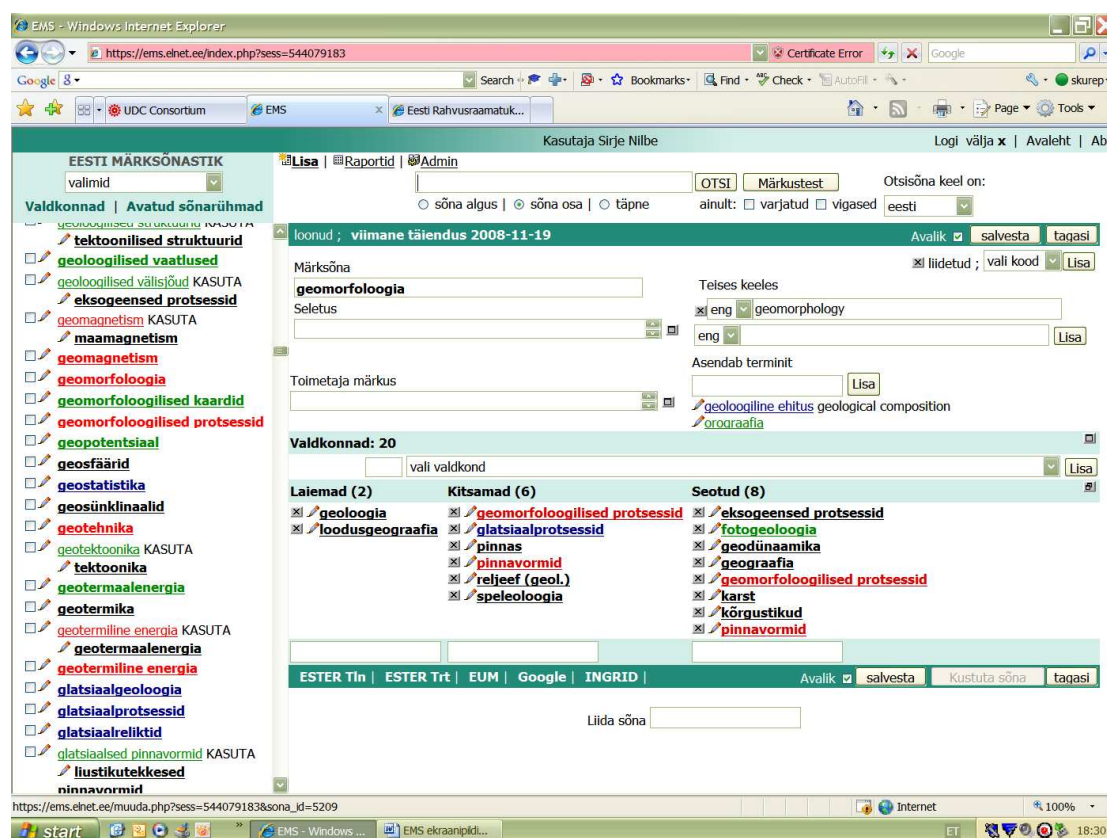


Figure 4. Colour coding of terms in the database

The database enabled to present only the logical mistakes evolved in the course of merging, the arising of new mistakes was blocked. When a logical mistake was corrected, it turned from red to black.

The editors could additionally merge term records if they considered the corresponding terms to be synonymous. In this case the relationships of the two terms were merged automatically and the other term could be left as a nonpreferred term.

## 6. The main characteristics of the Estonian Subject Thesaurus (EMS)

The EMS includes 34 000 preferred terms and 12 700 nonpreferred terms – altogether 46 700 terms, divided into 48 subject fields. It is a universal controlled vocabulary for indexing and searching in Estonian various library material.

The database of the thesaurus enables the users

- to browse subject terms by subject fields;
- to search terms by the beginning or part of word, or by exact match;
- to search terms by English equivalent;
- to view search results as word lists or as full records;
- to search by every term in the online catalogue ESTER, in the database of Estonian articles ISE or in Google;
- to print, e-mail or save into file selected word lists or full records;
- to subscribe current awareness service for new, changed and deleted subject terms.

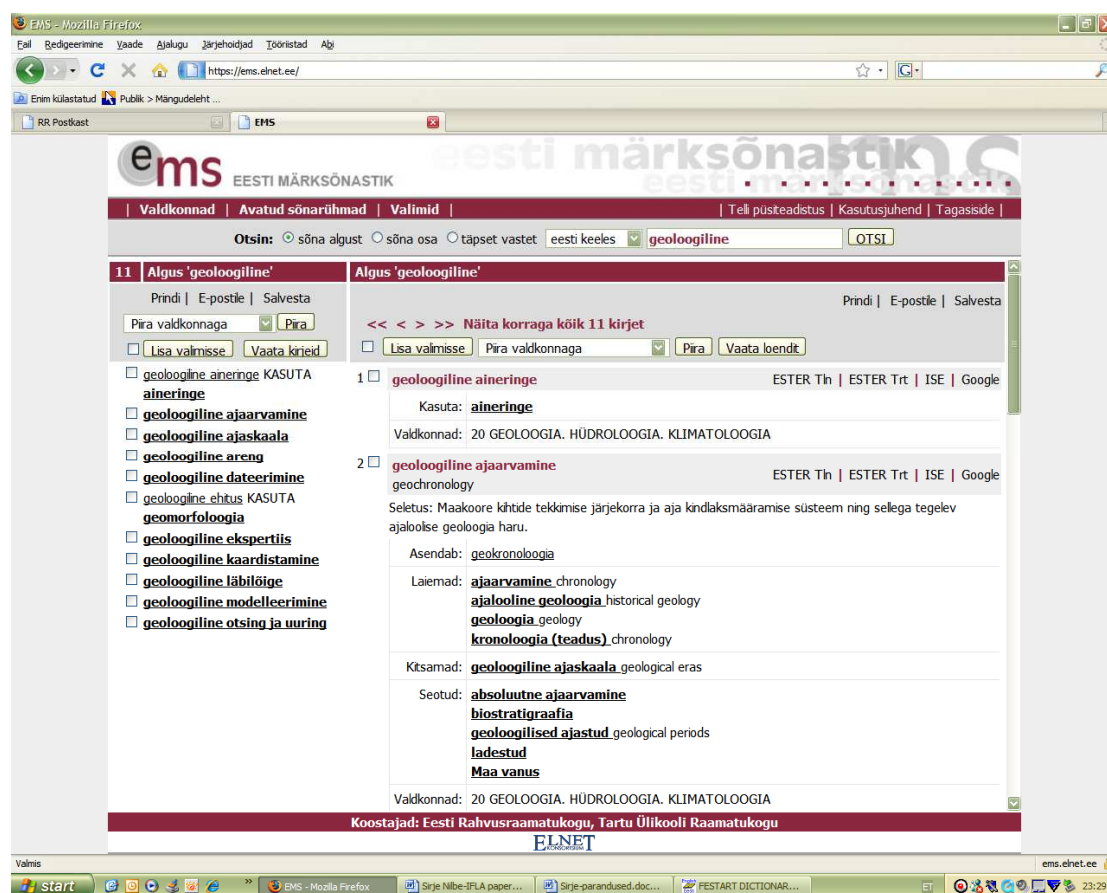


Figure 5. Example of the display of the EMS

The thesaurus is jointly managed by the ELNET Consortium, the National Library of Estonia and the University of Tartu Library, it is used as a standard in the Consortium's union catalogue ESTER (ester.nlib.ee; ester.utlib.ee) and in the database of articles ISE (ise.elnet.ee). As the EMS replaces the EÜM, all libraries previously using the EÜM will continue to index by the new EMS.

The Estonian Subject Thesaurus is freely accessible on the web (ems.elnet.ee).

## **7. Conclusion**

The project of merging two thesauri could be considered successful. In a relatively short period a shared thesaurus was created which is suitable for indexing and information search. The programmatic merging of the thesauri and subsequent inevitable editing is the fastest method for creating a shared controlled vocabulary. This is the best way to preserve the compliance between the authority data and the hitherto done subject indexing, as only a small number of subject terms are changed and deleted.

In current e-environment it is more practical to use shared indexing languages which are universal, multifunctional and flexible, instead of later facing the compatibility problems of different indexing languages and the task of achieving interoperability. This principle is particularly sensible in a small country with a small language community as Estonia.

---

### **Short biography**

I have academic degrees in Estonian language and information science. From 1986 to 1997 I worked at the University of Tartu Library, from 1998 at the National Library of Estonia. My professional fields are authority control, classification and indexing, development of thesauri. Current positions: Head of the Authority Control Department of the National Library of Estonia, Chief of the Classification and Indexing Working Group of the ELNET Consortium, Manager of the EMS Thesaurus.

Sirje Nilbe