EnviroInfo 2004 (Geneva)
Sh@ring – EnviroInfo 2004
Copyright © Editions du Tricorne 2004. ISBN: 28 29 30 275-3

Construction, Maintenance and Visualisation of Multilingual Thesauri Experiences using „SuperThes“

Wolf-Dieter Batschi,[1] Rudolf Legat,[2] Paolo Plini,[3] Hermann Stallbaumer,[4]

Abstract

In the last years, a General Multilingual Environmental Thesaurus (GEMET) in all the languages of the EU-member states has been developed within the working program of the "European Topic Centre for Catalogue of Data Sources & Thesaurus" (ETC/CDS), European Environment Agency, (EEA). GEMET is meant to support indexing of metadata within the CDS system.

At the same time, an environmental Thesaurus based on the Umwelt Thesaurus (UBA - Umweltbundesamt, Berlin) was produced in co-operation between Germany and Austria for their common metainformation system "Environmental Data Catalogue" (Umweltdatenkatalog, UDK). Austrian UDK at http://193.170.161.213:8080/wwwudk/html/de/start.html , German UDK at http://www.udk-gein.de/

To manage and maintain both the CDS-Thesaurus and the UDK-Thesaurus, a THESaurus MAINtenance (**THESmain**) system as well as a tool for visualising thesauri (**THESshow**) had been constructed by TBHS on behalf of the ETC/CDS and the Germany/Austria UDK-team.

THESmain is fully operational since May 1997 for both applications (CDS and UDK). UDK-Thesaurus Editorial Board at http://www.cedar.at/wgr_home.

CNR at present is working on a new Thesaurus for the Environment. To fulfil the requirements of future Thesaurus work, a Memorandum of Understanding (MOU) has been concluded in late 2000 between CNR - Consiglio Nazionale delle Ricerche,

[1] Umweltbundesamt Berlin, Bismarckplatz 1, D-14193 Berlin,
email: wolf-dieter.batschi@uba.de, Internet: www.umweltbundesamt.de

[2] Umweltbundesamt Wien, Spittelauer Lände 5, A-1090 Wien
email: rudolf.legat@umweltbundesamt.at, Internet: http://udk.ubavie.gv.at

[3] CNR Consiglio Nazionale delle Ricerche, Unità Terminologia Ambientale, IIA - Istituto sull'Inquinamento Atmosferico, Area della Ricerca di Roma-Montelibretti, Via Salaria Km 29,300, I-00016 Monterotondo Stazione, Rome - Italy, email: plini@iia.cnr.it,
Internet: http://uta.iia.cnr.it/

[4] Fa. TBHS, Favoritenstraße 182, A-1100 Wien, email: hermann@tbhs.co.at

Rome, UBA Berlin, UBA Vienna and TBHS Vienna. It's aim was to build a co-operation in order to develop a new software tool for building and maintaining multi-lingual, poly-hierarchical Thesauri, based on the experiences of the existing Thesaurus software tools used within the UDK co-operation.
This tool named SuperThes is now ready for work assignment.

1. General information about SuperThes

The Thesaurus of Environmental Data Catalogue Austria and Germany UDKT has been managed since 1995 by the software package THESmain. As, within almost a decade, the software came into age, a successor was designed, using a 4th generation programming language, a more powerful database engine and, hopefully, a more modern and more easily understandable user interface. Of course a multitude of new functions were implemented, most of them coming from requirements of previously working with THESmain.

The new software SuperThes® is currently in use by the environmental agency in Berlin and Vienna and by CNR in Rome. Recently some institutions have adopted SuperThes as platform for their thesaurus work e.g. "Bundesanstalt für Wasserbau", Hamburg /NOKIS or WHO/CEHA in Amman, Jordan.

1.1 SuperThes key features

1. Adaptability
2. Flexibility
3. Interoperability

1.1.1 Adaptability

Typically thesaurus software packages are constructed for a particular application where a single institution pays money for the development. If the packages performance is sufficient, the software is sold to other firms and institutions if it is possible to squeeze their requirements into the existing software. In many cases this is just partially successful –THESmain is such an example.

In this aspect SuperThes is very different. After installation there are no predefined tables or fields. There is also no predefined hierarchy.

A thesaurus developer is able to define tables and fields according to his requirements. There is no limitation in the number of tables. Fields are limited to 64 kBytes per line. Thus using Memo fields 16384 fields may be defined per table. The internal data storage is Unicode compliant enabling the use of every written language as long operating system support is existing. Currently about 200 languages are defined in SuperThes.

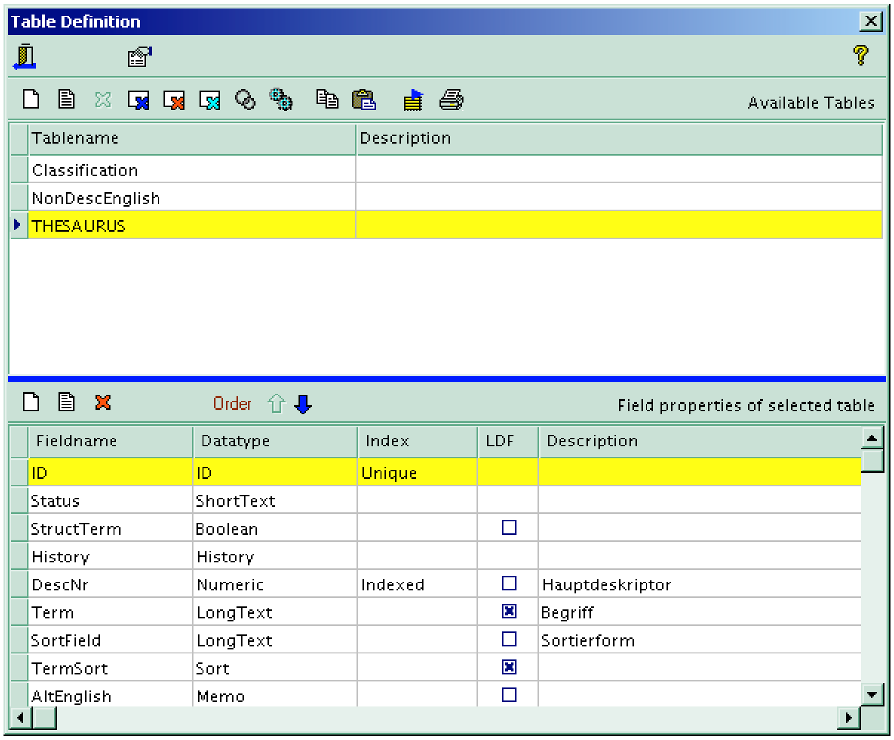


Table Definition

Available Tables

Tablename	Description
Classification	
NonDescEnglish	
THESAURUS	

Order

Field properties of selected table

Fieldname	Datatype	Index	LDF	Description
ID	ID	Unique		
Status	ShortText			
StructTerm	Boolean		□	
History	History			
DescNr	Numeric	Indexed	□	Hauptdeskriptor
Term	LongText		☒	Begriff
SortField	LongText		□	Sortierform
TermSort	Sort		☒	
AltEnglish	Memo		□	

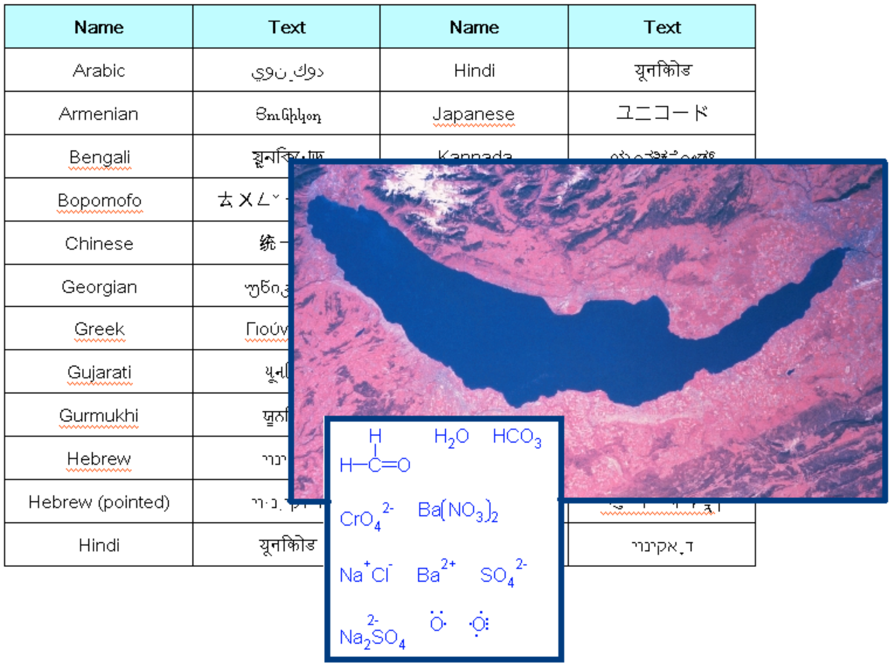
Figure 1
Table and Field Editor

1.1.2 Flexibility

Traditional thesaurus software packages contain few data types only. Typically there are data types for numbers, for text and memo fields for larger chunks of text. It is not even possible to write simple chemical formulas like H_2O. If data is found in electronic sources like Microsoft Word documents or in web pages, the text is likely to be not in ASCII any more. So thesaurus workers have to convert the RTF or UTF8 encoded data into their legacy ASCII systems.

SuperThes contains many powerful data types beside the standard field types like “Text” or “Memo”. So fields may contain images, sounds or formatted documents in Rich Text format (RTF) allowing to incorporate documents created in Microsoft

Word or in StarOffice. For the ease of operation drop down selectors are also implemented.



Name	Text	Name	Text
Arabic	يونِكود	Hindi	यूनिकोड
Armenian	Յունիկոդ	Japanese	ユニコード
Bengali		Kannada	
Bopomofo	ㄊㄨㄥˇ		
Chinese	统		
Georgian			
Greek	Γιούν		
Gujarati			
Gurmukhi			
Hebrew			
Hebrew (pointed)			
Hindi	यूनिकोड		יוניקאָד

Figure 2: Document and Image Editors

1.1.3 Interoperability

Almost all document and image data are kept within data processing equipment. Imagine finding a definition from an online encyclopaedia including an image fitting a term in your data collection. Traditional software requires you to write the definition and to forget about the image. SuperThes enables you to include document and image either per cut and paste or by simple dragging the data into the appropriate fields. Of course this features are bi-directional enabling drag and drop operation from SuperThes into Microsoft Word, Excel etc.

Another important item is bulk data exchange. SuperThes supports two different methods for data exchange of bulk data:

1.1.3.1 XML data exchange

SuperThes produces XML output from its tabular and structural data and of course it is possible to import from XML sources. The XML format is contained in a Document type definition which of course is published and available to everyone.

The document type definition is in fact a metadefinition describing a method to describe a thesaurus' actual configuration. An SuperThes XML file consists of three parts:

The document type definition

The description of tables and fields of a particular thesaurus

The data (tabular and structural) itself

Over all of the advantages of XML data exchange there is one big disadvantage: the "other" program has to have a XML parser. Well you might use some public available tool kit like SAX. But the data path from XML into your database engine must be programmed. So there is most likely a high initial cost to implement XML data exchange.

If it comes to actions like to get data from an old system into SuperThes most people prefer to use a simple text format.

XML data may contain images, documents and sounds. This type of data are encoded as Base64 which is adapted from RFC1421. This type of encoding is compliant to RFC2045 (Internet Message Bodies)

1.1.3.2 Textual data exchange

Data which is contained in tables is very frequently found as a candidate to import into a thesaurus. Such data collections coming typically from text processing software like MS-Word or from a spreadsheet program like MS-Excel may be very easily imported into SuperThes. Various types of character encoding, date and time formats, field and record separator may be selected. Text may be encoded in every code page available on Microsoft operating systems and of course in Unicode. Unicode is supported either in UCS-16 or in UTF-8.

Imported data may be appended to an existing table. Columns may be referenced to already existing data and data may be merged with existing data.

Data to export are defined by selecting a table and fields.

Textual data exchange may also contain image or document data which are encoded as Base64 (See XML data exchange for details)

2. Development of the thesaurus UDKT-8 by using SuperThes

The Thesaurus of Environmental Data Catalogue Austria and Germany (UDK-Thesaurus Version 8) is originally based on environmental thesaurus UMTHES® of UBA Berlin and developed within a co-operation between Germany and Austria since 1993. Version 8 was released on 25. February 2004.

The first goal was to export the data from THESmain into SuperThes. Although it would have been easy to export tables from THESMain and import these table into SuperThes a conversion utility was programmed. The basic idea of creating a utility was to assist all THESmain users in upgrading to SuperThes with an automatic conversion utility.

Thesaurus Editorial Board on the web: http://www.cedar.at/wgr_home/

2.1 UDK structure

2.1.1 UDK-T 8 basic structure

UDK-T descriptors (preferred terms, polyhierarchically, with filing sections for extended structuring, English translations), references of synonymous non-descriptors to descriptors, references between related descriptors /related terms, definitions, grammatical information for text analysis.

2.1.2 UDK-T 8 content coverage

Waste. Soil. Environmental chemicals/pollutants. Environmental aspects of energy and raw materials. Environmental aspects of genetically changed organisms, viruses and genes. Noise/vibration. Environmental aspects of the agriculture and forestry, fishery, food. Air. Nature and landscape/regional development. Radiation. General and overlapping environmental questions. Environment law. Environmental economics. Water. ...

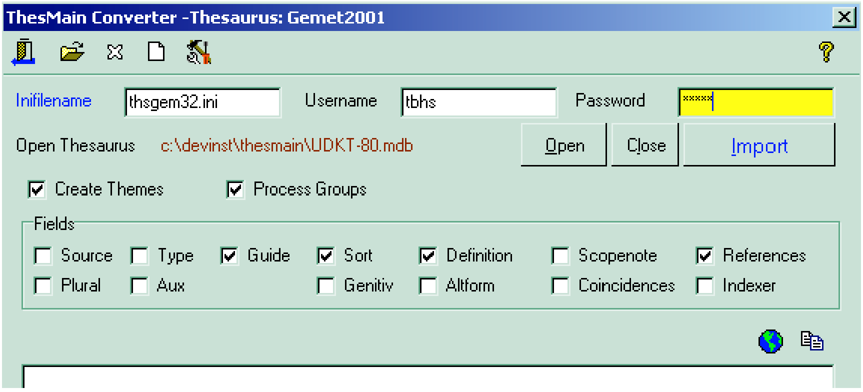
2.1.3 size (no. of terms) (February. 2004)

- Descriptors (preferred terms): **8,903**
- Filing Sections: 110
- Top Terms 33
- Non-descriptors (synonyms of descriptors): 19,366
- Part Terms for text analysis: 13,300
- sum total **41,712**

The next step was the import of the data delivered from the environmental agency Berlin. They provide an MS-Excel like data content from their ADIS system.

The data now is held in three tables namely THESAURUS, Classification and NonDescEnglish.

For the thesaurus table 26 fields were created.
The total time for creating table and fields was half an hour.



ThesMain Converter -Thesaurus: Gemet2001

Inifilename: thsgem32.ini
Username: tbhs
Password: *****
Open Thesaurus: c:\devinst\thesmain\UDKT-80.mdb

☑ Create Themes ☑ Process Groups

Fields
□ Source □ Type ☑ Guide ☑ Sort ☑ Definition □ Scopenote ☑ References
□ Plural □ Aux □ Genitiv □ Altform □ Coincidences □ Indexer

Figure 3
THESmain-SuperThes Converter

To import the data content again a utility was programmed. Also here it would have been possible to get the data in with SuperThes native capabilities only.

Using the two utilities it takes about a working day to import and syntactically check the huge data set consisting of 22629 terms and 36779 relations.

The following working days doing quality checking, comparison with previous years data, removing bugs from the data content showed that this work can be done quicker than using THESmain. It also turned out that SuperThes was able to handle the quite large amount of data without problems severe bugs or hang-ups.

After creating the Version 8 the terms according to new German orthography were imported. In this course several new fields were defined to hold the old data. Also this task could be finished within short time and without problems.

The production data were exported in textual format. The data are already imported into the UDK and in public use.

3. Application of SuperThes to the Environmental Application Reference Thesaurus EARTh

EARTh Thesaurus is based on a faceted structure. Its classification scheme was built utilizing a deductive-inductive approach and is based on the identification and adoption of a system of categories. The Thesaurus vertical structure is founded on categories; it is organized in a framework of different levels and classification knots, and it comprises hierarchical relationships.

The proposed model envisages the possibility of complementing the faceted structure with a system of themes; the latter, by crossing with the vertical structure, would form a matrix system.

Although an operational proposal was created for a specific utilization, the system of themes, as it was conceived, should be developed according to the specific needs of the applicative context, which precisely defines the theme, i.e. the object of interest.

In a thematic approach, the terms linked to a specific sector, are reassembled, while the facet structure tends to scatter them under the more general referral concept.

For EARTh, the implementation of an extended set of semantic relationships was planned and is at present in progress. The associative, equivalence and hierarchical relationships will be arranged in a series of sub-relationships, whose logical content is specified. The enrichment and the logical clarification of the relationships, reinforce their potential utilization for navigating on a conceptual basis. Besides, it will increase the possibility of using and reusing them also within the context of artificial intelligence which is based on semantic networks for knowledge modelling, in the perspective of developing a machine-readable version of the thesaurus.

From a different point of view, the increase and the specification of the associative relationships will strengthen the capacity of the system to represent the interconnected feature of the conceptual area. It will also face in a better way the requirements posed by the networked information on the Internet.

The EARTh terminological content is derived from various multilingual and monolingual sources of controlled environmental terminology: GEMET - GEneral Multilingual Environmental Thesaurus; the Italian Earth Sciences Thesaurus; EnvDev, the Terminological Bulletin for the Rio de Janeiro Conference on Environment and Development, plus other documents concerning specific sectors.

The selection, as well as the semantic and morphological processing of this terminology, still in progress, requires the management of problems connected with the utilization of multiple sources, developed in different operational fields and characterized by different approaches, concerning conceptual structuring and terminological representation.

EARTh has been uploaded into SuperThes and is currently successfully handled with it since the software ensures the possibility to implement all the features developed into EARTh and to work in a multilingual environment which is required by the EARTh user community.

Bibliography

Angrik, M., Bös, R., Rüther, M., Bandholtz, T. (2002): "Semantic Network Services (SNS)". In: EnviroInfo 2002, TU Vienna 2002.

Bandholtz, T., Bös, R., Rüther, M. (2000): "The German Environmental Information Network (GEIN)." In: Cremers, A.; Greve, K. (Eds.): Computer Science for Environmental Protection '00. 2000.

Batschi, W-D., Legat, R., Stallbaumer, H., (1999): „THESshow und THESmain", Moderne Softwarewerkzeuge zur Erstellung, Wartung und Visualisierung multilingualer Thesauri. In: 51. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI), Hamburg 21. Bis 23. September 1999, Seite 285ff, Frankfurt am Main: DGI, 1999. Available at: http://www.cedar.at/wgr_home/pub/hamburg.htm.

Batschi, W.D., Felluga, B., Legat, R., Plini, P., Stallbaumer, S., Zirm, K.L., (2002) - "SuperThes": A New Software for Construction, Maintenance and Visualisation of Multilingual Thesauri". 16th International Conference of the GI TC 4.6 "Informatics for Environmental Protection", Vienna 2002. "Environmental Communication in the Information Society".

EEA, European Environment Agency. Felluga B. & Batschi W.D. (Eds.), (1999) - GEMET, General European Multilingual Environmental Thesaurus. CNR, Rome – UBA, Berlin, Version 2.0. CD-ROM edition. Available on the Internet at http://www.mu.niedersachsen.de/cds/etc-cds_neu/software.html.

EEA, European Environment Agency. Felluga, B. & W.-D. Batschi, Eds. GEMET, GEneral Multilingual Environment Thesaurus. Version 2.0, August 1999. Printed edition. pp. xxvi + Volume 1: Systematic List of Descriptors, pp. 44; Volume 2: Thematic List of Descriptors, pp. 78; Volume 3: Alphabetical List of Terms, pp. 550; Volume 4: Concordance List, pp. 127; Volume 5: Multilingual List of Descriptors, pp. 536. CNR, Rome, April 2000.

Felluga, B., Plini, P., Lucke, S., Palmera, M., Eds. (2002), EARTh, Environmental Applications Reference Thesaurus. Pp. vi + Volume 1: Elenco Sistematico, pp. 37; Volume 2: Elenco tematico, pp. 159; Volume 3: Elenchi Alfabetici brevi, Italiano-Inglese e Inglese-Italiano, pp. 113 + 113; Volume 4: Elenco alfabetico bilingue, pp. 1418; Volume 5: Elenco delle Concordanze, pp. 115; Volume 6: Elenco Termini liberi – assegnati a liste – esclusi/cancellati, pp. 20; totale: vi + 1975. CNR, Roma, Aprile 2002.

Hashemi-Kepp, H., Legat, R. (2000): "Der Österreichische UmweltDatenKatalog, Erhebung und Strukturierung der Daten" UI 2000 - 14. Symposium Informatik für den Umweltschutz 2000 in Bonn, at: http://www.cedar.at/wgr_home/pub/hashemih.htm.

Kruse, F., Karschnick O., Spöringer A., Eichler M., Behrens S., "The UDK, Present Status and Future Development - An Overview." In: EnviroInfo 2002, TU Vienna 2002.

Legat, R., Batschi, W.-D., Hashemi-Kepp, H., Kruse, F., Nikolai, R., Nyhuis, D., Pultz, S., Stallbaumer, H., Swoboda, W., Zirm, K. (1999): "Der Umweltdatenkatalog in Österreich", 5 Jahre Erfahrungen, Workshop Umweltdatenbanken im Web, FZI Karlsruhe. Available at http://www.cedar.at/wgr_home/pub/karlsruhe.htm.

Rüther, M.,(2004): "Sharing Environmental Vocabulary". In: EnviroInfo 2004, Geneva, 2004

Plini P., Lucke S., Baffioni C. & Felluga B. (2001): T-REKS: a Contribution to the Environmental Information Management through a Computer-Supported Modular Knowledge Organisation System for the Environment. 15th International Symposium Informatics for Environmental Protection, Zurich 2001. "Sustainability in the Information Society". Hilty L.M., Gilgen P. W. (Eds.). October 9-12, 2001. Pp. 691-698.

Plini, P., Mazzocchi, F., (2004) - EARTh Environmental Applications Reference Thesaurus. A proposal for a new environmental thesaurus model. UNEP Thesaurus/Terminology Workshop. Geneva, 14-15.IV.2004 at: http://eea.eionet.eu.int/Public/irc/envirowindows/jad/library?l=/ecoinformatics_indicator/thesaurusterminology/overview_scenarios&vm=detailed&sb=Title

Swoboda, W., Kruse, F., Legat, R., Nikolai, R., Behrens, S. (2000): "Harmonisierter Zugang zu Umweltinformationen für Öffentlichkeit, Politik und Planung: Der Umweltdatenkatalog UDK im Einsatz", UI 2000 - 14. Symposium Informatik für den Umweltschutz 2000 in Bonn at: http://www.cedar.at/wgr_home/pub/ui2000.htm.