„SuperThes“: A New Software for Construction, Maintenance and Visualisation of Multilingual Thesauri

Wolf-Dieter Batschi,[1] Bruno Felluga,[2] Rudolf Legat,[3] Paolo Plini,[4] Hermann Stallbaumer,[5] Konrad L. Zirm[6]

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://udk.ubavie.gv.at.

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

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

[2] 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: bruno.felluga@itbm.rm.cnr.it, Internet: http://www.itbm.rm.cnr.it/HTML/environment/thesaurus/rrdafrme.htm

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

[4] 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@itbm.rm.cnr.it, Internet: http://www.itbm.rm.cnr.it/HTML/environment/thesaurus/rrdafrme.htm

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

[6] Umweltbundesamt Wien, Spittelauer Lände 5, A-1090 Wien
email: zirm@ubavie.gv.at, Internet: http://udk.ubavie.gv.at

(**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). Both GEMET and UDK-Thesaurus are maintained in Vienna. UDK-Thesaurus Editorial Board at http://www.cedar.at/wgr_home.

Since 1999, CNR is working on a new Thesaurus for the Environment, encompassing and extending GEMET. 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, Rome, UBA Berlin and UBA Wien. It has the aim 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.

General information about SuperThes:

SuperThes, the planned Thesaurus Software System, will take into account the long-term experiences with the existing Thesaurus maintenance software THESmain/THESshow, that is in use since almost five years. The new development will be based on the experience gained with the existing software package as well as from demands defined by the partners of the MOU, in order to fulfil the requirements of their Thesaurus works for the next couple of years.

The major design goals for the new software are:

- Compatibility with data maintained in the existing software
- Improved user interface, in order to ease system operability
- Integration of “state of the art” operating system support
- Support to create user-defined Thesaurus structures
- Support to create user-defined export formats
- Support for interfacing with Microsoft standard software
- Provision of maximum flexibility for creating reports.

The program will be generated using a modern 4th generation, object-oriented programming language. The Thesaurus data will be kept in a standard database management system.

1. Technical description of the Thesaurus Maintenance Software SuperThes

Core Programming

The new application will be programmed in Delphi, a Pascal dialect by Borland/Inprise. The programming language offers true object-oriented programming support as well as the stability common to Pascal programs. Unlike many other languages, it does not rely on countless external libraries and controls, thus ensuring long term installation security. The speed performance is as high as expected from a compiler producing processor native code directly.

Data Storage

All persistent data will be kept within a standard relational data base system. This will keep development costs low and gives the application the security of an off-the-shelf database system that is in use by many thousands of users. Various user requests convinced us to step forward to use a more scalable database. Three candidates have been evaluated:

- Oracle
- SQL-Server
- Interbase

One major goal was to achieve a flexible solution, depending on user needs. Either small desktop database directly attached to the application installed on the customers computers or a distributed client server installation.
After evaluating various database on the market we decided to use Interbase.

Platforms

For Delphi products, native processor code platform switching is difficult. Thus, the program will run by default on all modern 32-Bit Microsoft operating systems, such as:

- Windows 95
- Windows 98

- Windows ME
- Windows NT (Service Pack 4+)
- Windows 2000.

Languages

SuperThes as an application will have an English user interface. Thesauri created with this application may contain up to 30 languages. All languages, character sets and glyphs installed on the system may be used. Thus languages like Greek, Russian but also Chinese or Hebrew are no problem. Input method editors, which are common to Asian languages, are possible. Different fonts are also available. Language codes follow ISO-standard 639-2.

Thesaurus Specification

The application provides user interface and functionality to create and maintain monolingual and multilingual thesauri. The structural functionality, as well as the terminology used, conforms to ISO standards 2788 and 5964. The user definable set of hierarchical rules is a superset of ISO standards 2788 and 5964. Other rules, like the minimum requirements for field entry, are also user definable. The adherence to these rules is automatically enforced by the application itself. Main rules are the enforcement of reciprocity, the check for arithmetic loops, the check for duplicate entries, the enforcement for uniqueness of concept, to name just the more important ones. Besides the main Thesaurus structure, a SuperThes Thesaurus may contain other user definable tables, which may be used to describe the main Thesaurus in a more detailed way or to keep attached data in them. An example might be the themes-table in GEMET, or a table containing geographic names or other tables.
Thesaurus data are kept in standard database files, that allow the exchange of thesauri, the creation and deletion of thesauri on a simple file base. SuperThes allows to keep on the same computer system an unlimited number of thesauri and of course of different Thesaurus systems with different data definitions. More than one instance of the program may be used simultaneously on the same computer, showing different data contents e.g. GEMET together with UDK-Thesaurus. All Thesaurus data may contain “translations” (linguistic equivalents) for up to 30 languages using all character sets and glyphs available in the installation system. Input method editors are supported.

Thesaurus Definition

In SuperThes, a Thesaurus may be defined due to the required data model. There are only two fields required for any model: one is the record neutral identifier; the other is a field for the decimal notation controlling the hierarchies, the content of which is generated dynamically by the application. All further fields are definable. The following data types are available:

- String — Text field with up to 255 characters
- Large String — Text field with up to 32kByte
- Integer — Contains numbers in the range from –32768 to + 32767,
- Long — Contains numbers in the range from -2^{31} to $2^{31}-1$
- Boolean — True or False
- Sound — Wave files
- Picture — Picture files
- List — Multiple entries derived from pick lists (data type table)
- Table — Multiple reference to Table (data type table).

The data type table is definable by itself. It contains at least a field containing a record identifier. Up to 255 additional fields may be defined from the following data types:

- String — Text field with up to 255 characters
- Large String — Text field with up to 32kByte
- Integer — Contains numbers in the range from –32768 to + 32767,
- Long — Contains numbers in the range from -2^{31} to $2^{31}-1$
- Boolean — True or False

String fields, Large String fields, Lists and Tables may be defined as language dependent or language independent fields. In the second case, the field or table will be generated for all languages that are currently defined.

Indexes, either unique or allowing multiple entries, are allowed for String-, Integer-, Long- and Boolean fields.

Additional to the standard relations BT/NT, RT and USE/UF/UC, modifiers for relations may be used. Thus, an unlimited number of relation types may be defined.

The software package will contain predefined templates for GEMET, UDK-Thesaurus and CNR Thesaurus for the Environment.

User Interface

The program is designed to create and edit thesauri using a graphical user interface according to rules defined for Microsoft Windows. The program functions are hierarchically ordered. Each window provides context sensitive help. State of the art features, like drag and drop and right click context menus, will be used where useful. All functions are available *via* mouse control, as well as via keyboard control. Operation of similar functions will be similar in all windows.
The following key functions will be implemented:

- Management for thesauri, including creation, deletion, and copying thesauri
- Definition of thesauri due to the required data model, including auxiliary tables
- Definition of languages
- Visualisation of all data contents of a Thesaurus, either in tabular or in hierarchical view
- Editing all Thesaurus data
- Generation of reports
- Data exchange with external applications.

Visualisation in tabular form

All Thesaurus data may be shown in tabular form. The visualisation table may be defined by the user, who can select which fields in which languages are currently shown. The table may be sorted in any column that contains an indexed field. Sorting is accomplished by simply double clicking the desired column header. Special sorting orders are also available. An example might be: Descriptors/Non-descriptors/Unused Terms. Column arrangement and column width may be changed on the fly. When the program is closed, all settings are stored and used for the next launch. Different languages using different character sets may be used simultaneously. To retrieve specific data, a high-speed find algorithm is implemented. To reduce the number of displayed items, a “Select” function is included. This function exhibits a graphical editor to specify parameters attached to any field used in the Thesaurus. From this editor, an SQL-expression is defined, which is also editable directly. The database then returns all records that fit into this expression. The selections may also be stored for later use.

Editing tabular entries

Fields may be edited directly in the table. In the case of a field displayed directly in the table, like numbers or strings, double clicking the entry allows to directly edit the field content. In all other cases, special field editors for sounds, pictures … are launched.

Visualisation of relations

To visualise relations, a modified tree view will be used, displaying hierarchies in a convenient way. Special user interface functions, to expand and collapse whole branches or to expand all data up to a specified hierarchy, will be included. In a second window, all detail data of an activated leaf in the tree view will be displayed. Entries representing relations in the details window may be used as hypertext entries into the tree view. The data content shown in the detail view is again configurable by the user. The ordering within the tree view is also user selectable.

Editing relations

Relations are edited by selecting one part of the relation, either by mouse or by keyboard. A context menu is presented to select the kind of relation. The selected term is bound to the second part of the new relation, either per drag and drop or by clicking and inserting. If a term is moved within the hierarchy, a user may select what happens to the “hole”. Either it is simply closed by moving up all subsequent entries by one hierarchy level, or the whole branch is moved, or the rest is moved to the top [upper] level hierarchy. When changing hierarchical properties, status information on terms will be calculated automatically by SuperThes.
The application is able to open two data contents at the same time. These may be two different thesauri, but also two different views of the same Thesaurus. Dragging and dropping of hierarchical elements may be done across the borders of the two displayed data sets.

Data Exchange

Data exchange may be performed in several ways:

- SGML Export and Import

For the use in the European Environmental Agency, a data exchange format utilising SGML has been designed. This format is already used in a number of related applications. SuperThes will also be able to write and read this data exchange format. Benefits of this method is the general available document type definition and the possibility to read and write every detail of a Thesaurus. The drawback is the amount of work that must be put into a custom reader or writer.

- Microsoft compatible Export and Import

Import and Export of “flat” (non-hierarchical) data directly to and from Microsoft applications like MS-Excel or MS-Access. This method was already used with great success in porting GEMET into other languages. It may also be used to import any flat file (list of terms or similar) into the system.

- Interface to attached applications

For applications designed to be used in conjunction with SuperThes, a specialized interface will be available. This feature may be used for THESshow, the standalone visualizer, as well as for the Web visualizer, “THESweb”, which is in the design phase.

Report Generation

Beside the need to display data in electronic form, generating formatted printed output is always necessary. A fully user definable approach is used, where every field may be defined for size, font, colour, frame and so on. All fields belonging to a record are combined into a band. Again, for this band, geometric issues like spacing, frames and so on may be defined. In a third step, the page outline has to be defined. This includes size, headers / footers page numbering and similar. Definitions may be stored for later use. Target is either the attached system printer or a file in portable document format. As templates, the standard print formats for GEMET, UDK-Thesaurus and CNR Thesaurus for the Environment are included.

Data integrity

From a systems point of view, the application will provide Thesaurus backup mechanisms as well as database assisted functions, to repair corrupt files. The use of a standard database module will assure consistent data management within the tables. From the users point of view, all changes in a Thesaurus is logged, describing who has done what and when. If possible, a “back” function will be included.

Data access

The application will present a Login window at program start. The user supplied username and password will be used to define access rights for a given user in the Thesaurus. Also the database itself is secured, so that access without correct username and password even from external third party programs will not be possible. The log-file will contain every attempt to open a Thesaurus. In all changes to data, the username is recorded in the log-file.
Several levels of user access rights will be established:

- Administrator: everything granted
- Active User: read and write access to thesauri, no management functions
- Passive User: read access to thesauri, report definition and generation.

The user management (immission and deletion of users) is performed in an external utility. User rights may be granted at system level as well as at Thesaurus level.

Literature

Batschi, W-D. (1994): „Environmental Thesaurus and Classification of the Umweltbundesamt (Federal Environmental Agency), Berlin“, Berlin 1994

Batschi, W-D. (1995): „Development and State-of-the-Art of the German Environmental Thesaurus (UBA-Thesaurus) and User Experience in Germany“, Berlin 1995

Bundesgesetz über den Zugang zu Informationen über die Umwelt (Umweltinformationsgesetz – UIG 1993), BGBL 495/93, Wien 1993

Bundesministerium für Umwelt, Jugend und Familie (1993): „Das Recht auf Umweltinformation“, Informationsbroschüre, Wien 1993

Bundesministerium für Umwelt (1995): „Grundlagen und Methodik des Umweltdatenkataloges“, Schriftenreihe zum UDK, Band 1, Wien 1995

Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (1998): „UDK Version 4.0, Benutzerhandbuch“, Bonn, Hannover, 1998

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 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.

EEA, European Environment Agency (1996): Newsletter issue 8, Copenhagen, June 1996

Günther, O. (1995): "Gutachten zur Entwicklung des Umwelt-Datenkataloges (UDK)", Humboldt-Universität zu Berlin, Jänner 1995

Hashemi-Kepp, H., Legat, R. (1996): „Der Umweltdatenkatalog, ein Anwendungsbeispiel für Metainformationssysteme", Informatikforum Band 10, Wien September 1996

Legat, R., Hashemi-Kepp, H. (1994): „Der Umweltdatenkatalog – Ein bundesweites Metainformationssystem über umweltrelevante Datenbestände", VGI – Österreichische Zeitschrift für Vermessung & Geoinformation, Heft 1+2/94, Wien 1994

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.

Schober, W., Lopatta, H. (1994): Umweltinformationsgesetz. Verlag Österreich, Wien 1994

Umweltbundesamt Berlin, Umweltbundesamt Wien (1997): „Thesaurus des Umweltdatenkatalogs (UDK-Thesaurus 3.0)", Band I bis III, Berlin, Wien 1997

Umweltbundesamt Berlin, Umweltbundesamt Wien (1998): „THESshow", Thesaurus des Umweltdatenkatalogs (UDK-Thesaurus 3.0)", CD-ROM, Berlin, Wien 1998