

Erstellung, Wartung und Visualisierung multilingualer Thesauri mittels „THESmain“ und „THESShow“

Rudolf Legat¹, und Hermann Stallbaumer²

Abstract

At present, a General European Multilingual Environment Thesaurus (GEMET) in all the languages of the EU-member states is being 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 emphasis is being put on an environmental Thesaurus based on the UBA-Berlin Thesaurus (Umweltbundesamt Berlin) established 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 THESSaurus MAINTenance (**THESmain**) system as well as tool for visualising thesauri (**THESShow**) had been constructed by the ETC/CDS and the UDK-team Germany/Austria.

THESmain is fully operational since May 1997 for both levels (CDS and UDK). Both GEMET and UDK-Thesaurus are maintained in Vienna.

General information about THESmain:

- THESmain is used for visualisation and maintenance of thesaurus data according to DIN 1462 /ISO 2788 and DIN /ISO 5964
- The software has been developed with Visual Basic 4.0 and may be used under WfW3.11, Windows 95, Windows NT 3.51, Windows NT 4.0 and OS/2. The program incorporates the Microsoft Access 2.5/3.0 Database core which is contained within the programming language.
- The maximum number of entries for descriptors, nondescriptors and relations is limited by the available mass storage space only.

¹ Umweltbundesamt Wien, Spittelauer Lände 5, A-1090 Wien
email: legat@ubavie.gv.at, Internet: <http://udk.ubavie.gv.at>

² Fa. TBHS, Favoritenstraße 182, A-1100 Wien,
email: hermann@tbhs.co.at

- Several (theoretically unlimited) thesauri can be handled simultaneously with the possibility to exchange data between thesauri.
- Each thesaurus may contain terms in up to 30 languages.
- At several user levels the application is password protected. A separate utility to create and maintain users is included for system administration.
- It is possible to associate each term of the thesaurus with external databases. Thus the incorporation of microthesauri is easy. The connection to external databases is done by DDE-links. Examples are provided for Microsoft Access, Microsoft Excel and Microsoft Word. Versatile configuration to meet the demands of external DDE-connections is provided.
- THESmain by itself may be a target to external DDE-requests. It is possible to show every term in every way of visualisation by external commands. By this feature two authorities using THESmain with different thesauri can be synchronised.
- With the DDE-features described above, the attachment of microthesauri is easy.

General information about THESshow:

- THESshow is a tool for visualising thesauri and data indexing and retrieval. It offers hierarchical and alphabetical views, search functions and a collection "basket" for selected terms.
- The software has been developed with Visual Basic 4.0 and may be used under WfW3.11, Windows 95 and Windows NT.

1. Metainformationssysteme im Umweltbereich

Die Suche nach Daten zu umweltbezogenen Themen gestaltet sich für Fachleute, insbesondere aber für die interessierte Öffentlichkeit nicht immer einfach, vor allem, wenn nicht bekannt ist, ob die gesuchten Daten überhaupt erhoben wurden, von wem sie erhoben wurden und wo man sie bekommen kann. Um diesem Informationsbedürfnis gerecht zu werden, wurden in den letzten zehn Jahren in vielen Staaten Umweltdatenkataloge aufgebaut. Diese Metainformationssysteme über umweltrelevante Datenbestände enthalten wichtige Hinweise über die Verwendbarkeit und den Zugriff auf die Daten, wie fachliche Beschreibung, fachlicher Kontext, Raum- und Zeitbezug sowie Angaben zur Zuständigkeit, Verfügbarkeit und Aktualität.

Auch die Gesetzgeber in den Staaten Europas unterstützen aus demokratiepolitischen Gründen die Entwicklung, den Zugang zu Umweltdaten so einfach wie möglich zu gestalten und darüber hinaus eine aktive Umweltinformationstätigkeit der Behörden zu entwickeln.

Entsprechend dem Katalog einer Bibliothek verzeichnet ein Umweltdatenkatalog die bei den Behörden vorhandenen Daten- und Informationsbestände, ohne diese selbst zu enthalten, er ist also ein Informationssystem über Informationsbestände. Im UDK werden die Datenbestände anhand definierter Eigenschaften beschrieben und mit der Adresse eines zuständigen Ansprechpartners verknüpft.

Die wesentlichen Aufgaben und Vorteile eines UDK sind demnach folgende:

- Gewährleistung eines möglichst kompletten Überblicks über die große Mengen umweltrelevanter Datenbestände, die von den Behörden und Institutionen erhoben und gespeichert werden.
- Beinhaltet eine präzise Beschreibung der Datenqualität dieser Datenbestände.
- Ermöglicht die überregionale (internationale) Standardisierung der Beschreibung von Datenquellen.
- Ist ein Informationsinstrument für die Öffentlichkeit nach Umsetzung der EU-Richtlinie „Freier Zugang zu Umweltinformationen“ in nationales Recht.

Wesentlichen Einfluß auf ein zufriedenstellendes Recherchenergebnis im UDK hat eine einheitliche Beschreibung und Beschlagnahme der Datenbestände, die bereits bei der Dateneingabe die unterschiedlichsten Sichtweisen potentieller UDK-Nutzer auf Datenbestände vorwegnehmen muß. Sie muß daher einerseits so detailliert sein, daß sie auch für Experten von Nutzen ist, aber gleichzeitig so allgemein, daß sie auch von Laien des betreffenden Fachgebietes verwendet werden kann. Diese Aufgabe kann nur ein „Thesaurus“ erfüllen, „eine systematisch geordnete Sammlung aller sprachlichen und sonstigen Beziehungen eines bestimmten Anwendungsbereiches“. Dessen Anbindung an ein Metainformationssystem soll einen flächendeckend homogenen Metadatenbestand gewährleisten, um einer Vielzahl von unterschiedlichen Anwendern als adäquates Instrument zur Datenverwaltung, -pflege und -recherche zu dienen. Die Entwicklung und die konsequente Anwendung eines Thesaurus zur Indexierung der Daten stellt daher bei der Realisierung eines UDK ein wichtiges Element dar.

2. Der Umweltdatenkatalog in Deutschland und Österreich

Im Rahmen eines vom Umweltministerium in Bonn geförderten F&E-Vorhabens wurde seit 1991 im Niedersächsischen Umweltministerium mit dem Aufbau des UDK begonnen. An dieser Entwicklung beteiligten sich weitere zehn Bundesländer im Rahmen einer Kooperation, um den UDK flächendeckend in Deutschland einzusetzen. 1995 konnte in Deutschland die Verwaltungsvereinbarung zum UDK zwischen Bund und Ländern beschlossen werden, um den Einsatz, die Erfassung und die Anwendungsentwicklungen zum UDK koordiniert voranzutreiben.

In Österreich wurde, im Rahmen der Umsetzung der Richtlinie 90/313/EWG des Rates der Europäischen Gemeinschaften vom 7. Juni 1990 über den freien Zugang zu Informationen über die Umwelt, das Umweltinformationsgesetz (UIG 1993) beschlossen. Dieses sieht im Sinne einer modernen und offenen Umweltverwaltung sowie einer erleichterten Bürgerpartizipation in § 10 die Einrichtung eines Umweltdatenkataloges als Zugangssystem zu Umweltdaten vor. Das UIG verleiht dem einzelnen durch die Verpflichtung der Behörden und Ämter, ihre Umweltdaten

transparent zu halten, einen neuen Informationsanspruch im Sinne demokratischer Mitgestaltung.

Im August 1993 schlossen Deutschland und Österreich eine „Vereinbarung über die Zusammenarbeit beim Aufbau, bei der Entwicklung und bei der Pflege eines gemeinsamen Umweltdatenkataloges ab. Im Rahmen dieser Kooperation übernahm Österreich die Aufgabe der Koordinierung, Entwicklung und Pflege eines Thesaurus (UDK-Thesaurus) sowie aller dazu erforderlichen Softwarewerkzeuge. Dieses normierte Wortgut hilft dabei, die Daten im UDK einer einheitlichen, semantisch vergleichbaren inhaltlichen Erschließung zu unterziehen und sie gezielt wiederzufinden.

Grundlage des UDK-Thesaurus ist der Thesaurus des Umweltbundesamtes Berlin, der seit vielen Jahren für die Datenbanken des Umweltplanungs- und Informationssystems (UMPLIS) eingesetzt wird.

Mit Hilfe des UDK und des UDK-Thesaurus lassen sich mehr Informationen von höherer Qualität auf effektivere Art und Weise beschaffen und verwalten als vorher. Dies führt zu einer spürbaren Verbesserung der Informationsversorgung sowohl der interessierten Bevölkerung als auch der Umweltverwaltungen und damit zu einem effizienteren Umweltschutz.

Dem Stand der Technik entsprechend ist der UDK Österreich seit März 1996 unter der Adresse <http://udk.ubavie.gv.at> im Internet zugänglich. Bislang wurden etwa 550.000 Zugriffe aus 60 Staaten von 12.000 verschiedenen Hosts registriert. Der UDK beinhaltet etwa 12.000 Datensätze (UDK-Objekte) und 1100 Adressinformationen (UDK-Adressen).

3. Das Europäische Metainformationssystem CDS und der Europäische Umweltthesaurus GEMET

Das große Interesse europäischer Staaten an der Entwicklung des technisch und organisatorisch weit fortgeschrittenen UDK trug zur Entscheidung der Europäischen Umweltagentur (EEA) in Kopenhagen bei, das Metainformationssystem „Catalogue of Data Sources“ (CDS) auf der Basis des UDK zu entwickeln. Dazu wurde ein entsprechendes „European Topic Centre“ (ETC/CDS) im Niedersächsischen Umweltministerium eingerichtet.

Der CDS wird ausgewählte Umweltinformationen der Mitgliedstaaten enthalten, die auf der Ebene der EU von Relevanz sind. Das vorrangige Ziel des CDS besteht darin, eine nahtlose Suche in den verteilten europäischen Katalogen zu ermöglichen und Teil eines „Global Information Locator System“ zu sein.

Die Voraussetzung dafür, mittels des CDS die Sprachgrenzen im EU-Raum zu überwinden, ist der Aufbau und Einsatz eines multilingualen Umweltthesaurus. Diese Aufgabe soll GEMET (General European Multilingual Environment Thesaurus) leisten, der im Rahmen von CDS entwickelt wird. GEMET wird etwa

5000 Begriffe in allen Sprachen der EU-Mitgliedstaaten enthalten und von einer umfassenden Terminologie-Datenbank unterstützt werden. Die Version 1 steht seit Ende 1997 zur Verfügung, die Version 2 in vorerst elf Sprachen ist für Mai 1999 zu erwarten.

4. Der Thesaurus des Umweltdatenkataloges (UDK-Thesaurus)

Allgemein formuliert ist ein Thesaurus ein „hierarchisch strukturierter, begrenzter Wortschatz, welcher der natürlichen Sprache entnommen ist und der ein Hilfsmittel für das vereinheitlichte Beschreiben („Indizieren“) und Auffinden von Informationen eines bestimmten Fachgebietes mittels normierter Begriffe („Deskriptoren“) darstellt.“

In großem Umfang entstand z.B. der Bedarf an Thesauri aus Effizienzgründen bei multinationalen Konzernen aus dem Pharma-, Automobil-, sowie Luft- und Raumfahrtbereich.

Der UDK-Thesaurus ist ein wesentlicher Bestandteil des UDK. Wie bereits erwähnt, dient als dessen Grundlage der Umweltthesaurus des Umweltbundesamtes Berlin, welcher auch maßgeblich in die Entwicklung des europäischen Umweltthesaurus „GEMET“ eingeflossen ist.

Derzeit enthält der UDK-Thesaurus etwa 24.500 Begriffe (ca. 8.500 Deskriptoren und 16.000 Non-Deskriptoren).

Der UDK-Thesaurus 3.0 ist ein deutschsprachiger Thesaurus, der ins Englische übersetzt wurde. Dabei wurden die dt. Deskriptoren übersetzt, den Übersetzungen aber eigene Non-Deskriptoren (N.) beigefügt. Daher ist die Struktur des Thesaurus in beiden Sprachen gleich, jeder Term hat somit in beiden Sprachen die gleichen Ober- und Unterbegriffe. Die zu einem Deskriptor gehörenden Synonyme (Non-Deskriptoren) sind jedoch für beide Sprachen verschieden, d.h. die Non-Deskriptoren wurden nicht übersetzt, sondern für jede Sprache stehen unterschiedliche N. zur Verfügung. Die Menge der N. ist somit in jeder Sprache unterschiedlich.

Bei der Weiterentwicklung des UDK-Thesaurus muß auf die Entwicklung des UDK Bedacht genommen werden. Erweiterungen werden notwendig, wenn

- Begriffe, die zur Beschlagwortung des UDK notwendig sind, nicht im Thesaurus vorhanden sind und
- Gewisse Fachgebiete, die von allgemeinem Interesse sind, im Thesaurus unvollständig oder nicht in genügendem Ausmaß vorhanden sind.

Um die Inhalte des UDK-Thesaurus ständig an die Bedürfnisse der UDK-Nutzer (sowohl der indizierenden Behörden als auch der Informationssuchenden) anzupassen, wurde das internationale Gremium „Wortgutredaktion“ (WGR) gegründet, welches sich aus Behördenvertreter und Umweltfachleuten der Kooperationspartner zusammensetzt, gegründet.

Die Arbeit der Wortgutredaktion hat die folgenden Schwerpunkte:

- Themenpolitik: Sicherstellung der Darstellung umweltrelevanter Fachbereiche mit geeigneten Deskriptoren.
- Begriffs- und Wortpolitik: Sicherstellung einer hohen sprachlichen Qualität des Thesaurus sowie der Übersetzung der Thesaurusbegriffe in andere Sprachen.
- Release-Politik: Versionsmanagement

Grundlage der Arbeit sind die internationalen Normen ISO 2788 (documentation – Guidelines for the Establishment and Development of Monolingual Thesauri) und ISO 5964 (Documentation – Guidelines for the Establishment and Development of Multilingual Thesauri) bzw. die entsprechende DIN-Norm 1463 Teil 1 und 2.

5. Die Entwicklung der Softwarewerkzeuge THESmain und THESshow

5.1. Anforderungen an die Software

Wie bereits erwähnt hat Österreich innerhalb der Kooperation zum UDK die Aufgabe übernommen, eine „Koordinierungsstelle Thesaurus“ einzurichten mit dem Ziel, einen Thesaurus für den Umweltdatenkatalog zu entwickeln, bereitzustellen und zu pflegen. Dies beinhaltet auch die Bereitstellung aller erforderlichen Softwarewerkzeuge.

Thesauri sind langlebige Produkte, die über viele Jahre (Jahrzehnte) hinweg gepflegt werden müssen. Die Pflege erfolgt meist zentral von einer kleinen Anzahl von Personen (Experten). Im Gegensatz dazu steht eine meist große Anzahl von Anwendern, die den Thesaurus benutzen aber keinesfalls Änderungen an den Daten durchführen dürfen. Früher wurden Thesauri ausschließlich in gedruckter Form verwendet, heute stehen elektronische Medien mit all ihren Vorzügen im Vordergrund. Eine Software zur Thesauruspflege besteht daher sinnvollerweise aus zwei Teilen:

- Einem Pflegemodul, das die Erstellung und Wartung eines Thesaurus ermöglicht und nur an der zentralen Pflegestelle (Wortgutredaktion) installiert ist und
- einem Visualisierungsmodul, welches den Nutzern zur Verfügung steht.

Durch die Langlebigkeit von Thesaurusdaten empfiehlt sich die Verwendung eines weit verbreiteten Datenbankformats, um bei neuen Betriebssystemen eine problemlose Portierung zu gewährleisten. Da besonders das Visualisierungsmodul von einer großen Anzahl von Nutzern verwendet wird, welche nicht immer EDV-Experten sind, empfiehlt sich eine standardisierte Benutzerschnittstelle um eine intuitive Bedienung der Software zu erleichtern.

Marktrecherchen im Jahr 1995 haben ergeben, daß zu diesem Zeitpunkt keine Thesaurussoftware am Markt verfügbar war, die geeignet gewesen wäre, die

speziellen Anforderungen der deutsch-österreichischen Zusammenarbeit sowie die des ETC/CDS zu erfüllen. Diese sind im Besonderen:

- Pflege- und Visualisierungsmodul
- Mehrsprachigkeit
- Standarddatenbankformat
- Software unter MS-Windows

5.2 Pflegesoftware THESmain

Zur Erstellung und Wartung des UDK-Thesaurus im EDV-technischen Sinn wurde das Pflegeprogramm „THESmain“ in gemeinsamer Beauftragung durch das Bundesministerium für Umwelt, Jugend und Familie/Umweltbundesamt und das „ETC/CDS“ der Europäischen Umweltagentur (EEA) entwickelt. THESmain dient zur Wartung sowohl des UDK-Thesaurus als auch des GEMET. Geplant ist weiters, künftig auch die Verwaltung des ENVOC-Thesaurus der UNEP mittels THESmain vorzunehmen.

Die Software kann generell zur Einrichtung, Pflege und Visualisierung von mono- und multilingualen Thesauri eingesetzt werden, es kann mehrere Thesauri gleichzeitig verwalten und Daten zwischen mehreren Thesauri austauschen. Jeder Thesaurus kann Begriffe in bis zu 30 Sprachen enthalten. Externe Datenbanken (Microthesauri) können einem Begriff zugeordnet werden.

Einige weitere Merkmale mögen der Aufstellung im Abstract dieser Publikation entnommen werden. Eine Demoversion sowie die Dokumentation des Tools können über einen Server des Umweltbundesamtes Wien bezogen werden.

5.3 Visualisierungssoftware THESshow

Die Erstellung dieses Tools wurde vom UBA Wien beauftragt mit dem Ziel, den UDK-Thesaurus Version 3.0 in Form einer CD-ROM den Partnern sowie den interessierten Nutzern anzubieten.

THESshow eignet sich zur Präsentation der mittels THESmain erzeugten Datenbestände und zum Arbeiten mit dem Thesaurus beim Indizieren oder Suchen in einer Datenbank.

Ebenso wie THESmain ist es multilingual bezüglich der Inhalte. Wird die Darstellung „Deutsch“ gewählt, werden die Begriffe in deutscher Sortierung und mit deutschsprachigen Relationen gezeigt. Bei englischer Darstellung werden die Begriffe mit englischen Relationen, englischer Sortierung usw. dargestellt. Im Detailfenster werden die deutschen Übersetzungen und Synonyme gezeigt. Somit ist der UDK-Thesaurus auch für englische Fachliteratur verwendbar.

Die Benutzeroberfläche kann wahlweise in mehreren Sprachen beim Programmstart gewählt werden.

THESshow bietet folgenden Funktionsumfang:

- Hierarchische und alphabetische Darstellung der Thesaurusdaten
- Suchfunktion
- Sammelkorb für ausgewählte Terme

THESshow wird in Kürze auch mit GEMET erhältlich sein.

Literaturverzeichnis

- 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
- European Environment Agency (1996): Newsletter issue 8, Copenhagen, June 1996
- Günther, O. (1995): „Gutachten zur Entwicklung des Umweltdatenkataloges (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
- 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