

THESshow und THESmain - Moderne Softwarewerkzeuge zur Erstellung, Wartung und Visualisierung multilingualer Thesauri

von

Wolf - Dieter Batschi (Umweltbundesamt, Berlin), Rudolf Legat (Umweltbundesamt, Wien),
Hermann Stallbaumer (Technisches Büro für Elektrotechnik)

Thesaurusarbeit ist seit Jahrzehnten geprägt von enormem Aufwand für die intellektuelle, fachinhaltliche Durchdringung und Erstellung der Thesaurusbeziehungen, die kontinuierliche Pflege und Aktualisierung des Wortgutes und die Anstrengungen zur Visualisierung der Beziehungen der Begriffe des Thesaurus. Bisher standen zur Unterstützung dieser Arbeiten eine Reihe von Softwarewerkzeugen zur Verfügung, die diesen Prozeß mehr oder weniger stark unterstützten. Sie waren in der Regel auf die Erstellung und Pflege einsprachiger Thesauri ausgerichtet oder in Dokumentationssysteme eingebunden, ohne die Möglichkeit, sie als „stand alone“ - Lösung nutzen zu können.

Im Rahmen eines Projektes der Europäischen Umweltagentur (EUA) zu einem europäischen Umweltdatenkatalog wurde ein Europäisches Themenzentrum Datenquellenkatalog (European Topic Centre on Catalogue of Data Sources ETC / CDS) beim Niedersächsischen Umweltministerium eingerichtet. Es soll die Suche nach umweltrelevanten Daten in einem Europäischen Rahmen ermöglichen und entsprechende Metainformationen speichern. 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.

Wesentlichen Einfluß auf ein zufriedenstellendes Rechercheergebnis in Metainformationssystemen hat eine einheitliche inhaltliche Erschließung der Datenbestände, die bereits bei der Dateneingabe die unterschiedlichsten Sichtweisen potentieller 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. 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 Metainformationssystems ein wichtiges Element dar.

Da es sich bei der Entwicklung des CDS und des dazugehörigen Thesaurus um ein multinationales Projekt mit entsprechenden verschiedensprachigen Beteiligten handelte, war es selbstverständlich, den entstehenden Thesaurus als multilingualen Thesaurus zu konzipieren. Außerdem soll es der Thesaurus allen Nutzern in den Ländern Europas ermöglichen, in ihrer eigenen Sprache den Datenkatalog abzufragen und über die Thesaurusbegriffe bei fremdsprachigen Datenobjekten im Katalog

zumindest einen ersten Hinweis zur Relevanz der gefundenen Informationen zu erhalten. Die inhaltliche Erschließung der Daten erfolgt nicht nur mit Hilfe des Thesaurus sondern darüber hinaus durch 30 (Sach)Gruppen und 40 Themengebiete. Im Rahmen des Projektes zur Erstellung des **General Multilingual Environmental Thesaurus (GEMET)** waren eine Reihe umweltrelevanter Thesauri aus den beteiligten Nationen zu berücksichtigen und zu einem einheitlichen Thesaurus auf europäischer Ebene zu verschmelzen. Hierbei wurden Auszüge aus den umweltrelevanten Thesauri Deutschlands, Frankreichs, Italiens, der Niederlande und Spaniens einbezogen. Der Infoterra - Thesaurus EnVoc der Vereinten Nationen wurde komplett in GEMET übernommen.

Bei der Suche nach einem adäquaten Softwarewerkzeug mußte im Rahmen des Projektes festgestellt werden, daß auf dem Markt kein Produkt existierte, das einer derart anspruchsvollen Aufgabe voll gewachsen war. Außerdem mußte erkannt werden, daß die vorhandenen Tools nicht in dem Umfang erweiterbar waren, wie es für die genannten Zwecke notwendig war.

Auch im Rahmen der bestehenden Kooperationsvereinbarung vom 22. August 1993 über die Zusammenarbeit beim Aufbau, bei der Entwicklung und bei der Pflege eines gemeinsamen Umweltdatenkataloges zwischen der Bundesrepublik Deutschland und der Republik Österreich erwies es sich als notwendig, für den gemeinsamen Thesaurus (UDK - Thesaurus) ein geeignetes Softwarewerkzeug einzusetzen. Österreich hat dabei die Aufgabe übernommen, im Rahmen einer Koordinierungsstelle Thesaurusentwicklung, die Bereitstellung und Pflege des UDK - Thesaurus (der identisch ist mit dem Umweltthesaurus des Umweltbundesamtes in Berlin) sicherzustellen.

Da die Projekte GEMET und Weiterentwicklung des UDK - Thesaurus sehr ähnlich sind, hat es sich angeboten, eine gemeinsame Entwicklung eines modernen Softwarewerkzeugs für die Erstellung, Pflege und Visualisierung von Thesauri zu betreiben.

Auf der Basis der Vorgaben der Thesaurusfachleute aus den Umweltbundesämtern Berlin und Wien sowie der italienischen Experten vom Consiglio Nazionale delle Ricerche (CNR), Rom erfolgte die Entwicklung der Programme durch die Firma Technisches Büro für Elektrotechnik (TBHS), Wien. Das Softwarepaket besteht, neben einer Reihe von Utilities, aus den Programmen *THESmain* für die Erstellung und Wartung und *THESshow* für die Visualisierung. Die Erstellung eines Thesaurus wird üblicherweise von einem kleinen Team zentral durchgeführt. Der fertige Thesaurus wird von einer großen Anzahl von Benutzern verwendet. Dies führt zu unterschiedlichen Anforderungen an die Software für die Erstellung und für die Benutzung eines Thesaurus, wobei noch festzuhalten ist, daß die meisten kommerziell erhältlichen Thesaurusverwaltungsprogramme den Nutzern den Thesaurus ohnedies nur auf Papier zur Verfügung stellen. Deshalb kam es zu der Entwicklung eines Softwarepaketes für beide Nutzungsarten, wobei die wesentlichsten Elemente und Funktionalitäten in beiden Programmen identisch sind.

Diese Elemente und Funktionen sind nachfolgend dargestellt:

Erstellungs- und Wartungsprogramm *THESmain*

- Definition von Thesauri
- Vielseitige Editiermöglichkeiten mit Sicherstellung der Datenintegrität
- Flexible Gestaltung von Zugriffsrechten
- Export- und Importfunktionen
- Kompatibilität zu gängigen Programmen (MS-Access, MS-Excel)

Visualisierungsprogramm THESshow

- Leicht verständliche Darstellung des Thesaurusinhalts
- Einfache Bedienung
- Gute Online Hilfe
- Einfache und robuste Installation
- Unempfindlichkeit gegen Eigenheiten von Rechnern
- Keine Möglichkeiten zum Editieren des Datenbestands

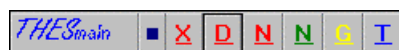
THESmain: eine Anwendung zur Thesaurusverwaltung

Wesentliche Funktionen

- THESmain dient der Erstellung und Wartung von Thesauri gemäß DIN 1462/1, DIN 1462/2 sowie ISO5964.
- Das Programm wurde unter Visual Basic 4.0 entwickelt und ist unter WfW 3.11, Windows 95/98 und Windows NT lauffähig. Die Daten werden in einem zu MS-Access kompatiblen Datenbankkern abgelegt.
- Das Programm besitzt eine grafische Benutzeroberfläche mit hierarchisch gegliederter Funktionalität. Zu jedem Fenster steht kontext - sensitive Hilfe zur Verfügung. Die aktuellen Fenstereinstellungen können auf Wunsch zur Wiederverwendung gespeichert werden. Alle Programmfunktionen können sowohl mit der Maus als auch über Tastaturkürzel aufgerufen werden.
- Die maximale Anzahl von Begriffen und Relationen ist nur durch den verfügbaren Massenspeicher begrenzt.
- Mehrere (theoretisch unbegrenzt viele) Thesauri können gleichzeitig bearbeitet werden.
- Jeder Thesaurus kann bis zu 30 Sprachen beinhalten.
- Das Programm ist Passwortgeschützt und verfügt über mehrere Klassen von Zugriffsrechten. Ein separates Hilfsprogramm zum Erzeugen und Verwalten von Benutzerberechtigungen steht für den Systemadministrator zur Verfügung.
- Es ist möglich, Verbindungen zu externen Datenbeständen herzustellen. Mit diesem Verfahren können Mikrothesauri problemlos eingebunden werden.

Kontrolleiste

Die einzelnen Unterprogramme von THESshow können mittels einer Kontrolleiste, ähnlich der von Microsoft Office, aufgerufen werden. Durch Anwahl von Schaltflächen dieser Kontrolleiste werden die entsprechenden Funktionen entweder gestartet oder das dazugehörige Fenster wird in den Vordergrund gestellt. Dadurch können mehrere offene Fenster auch auf kleinen Bildschirmen leicht verwaltet werden. Aufgerufen werden die Tabellen für Deskriptoren und Nondeskriptoren, das Navigationsfenster, welches die Relationen zeigt, das Fenster für die grafische Darstellung und das Fenster für die Werkzeuge.



Datenvisualisierung

Die Daten eines Thesaurus werden in THES*main* auf drei Arten gezeigt:

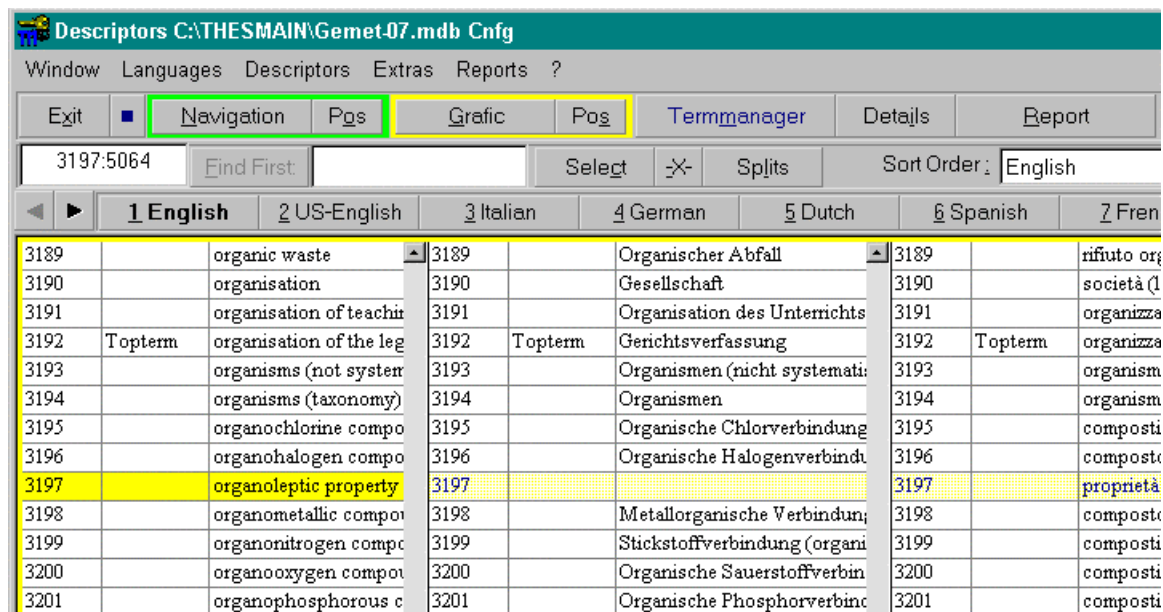
- Tabellarische Darstellung von Deskriptoren und Nondeskriptoren
- Im Navigationsfenster, um die Relationen zu zeigen
- In grafischer Darstellung

Tabellarische Darstellung von Deskriptoren und Nondeskriptoren

Wesentliche Funktionen

- Unabhängige Tabellen für Deskriptoren und Nondeskriptoren
- Hohe Darstellungsgeschwindigkeit (etwa 0.2 Sek. für Suchen mit einem Sprung über 18 000 Einträge)
- Darstellung von mehreren Sprachen mit deren Fonts und Zeichensätzen gleichzeitig (etwa Griechisch)
- Vielfältige Sortiermöglichkeiten
- Suchfunktionen unter Verwendung von Wildcards
- Die Möglichkeit Untermengen zu definieren. Dies kann entweder mit dem eingebauten Query Generator geschehen oder durch die direkte Eingabe von SQL - Abfragen. Einmal definierte Abfragen können zur weiteren Verwendung gespeichert werden
- Möglichkeit des automatischen Positionierens von einem Fenster in ein anderes

Das nachfolgende Bild zeigt Deskriptoren in drei Sprachen mit englischer Sortierung.



Descriptors C:\THESMAIN\Gemet-07.mdb Cnfg																				
Window Languages Descriptors Extras Reports ?																				
Exit			Navigation		Pgcs		Grafic		Pos		Termmanager		Details		Report					
3197:5064			Find First		Select		-X-		Splits		Sort Order: English									
1 English			2 US-English			3 Italian			4 German			5 Dutch			6 Spanish			7 Fren		
3189		organic waste	3189		Organischer Abfall	3189		rifiuto org												
3190		organisation	3190		Gesellschaft	3190		società (1												
3191		organisation of teachin	3191		Organisation des Unterrichts	3191		organizza												
3192	Topterm	organisation of the leg	3192	Topterm	Gerichtsverfassung	3192	Topterm	organizza												
3193		organisms (not system	3193		Organismen (nicht systemati	3193		organism												
3194		organisms (taxonomy)	3194		Organismen	3194		organism												
3195		organochlorine compo	3195		Organische Chlorverbindung	3195		composti												
3196		organohalogen compo	3196		Organische Halogenverbind	3196		composti												
3197		organoleptic property	3197			3197		proprietà												
3198		organometallic compo	3198		Metallorganische Verbindun	3198		composti												
3199		organonitrogen compo	3199		Stickstoffverbindung (organi	3199		composti												
3200		organooxygen compo	3200		Organische Sauerstoffverbin	3200		composti												
3201		organophosphorous c	3201		Organische Phosphorverbin	3201		composti												

In diesem Fenster können mittels des Termmanagers Deskriptoren und Nondeskriptoren angelegt, geändert und gelöscht werden. Änderungen können nicht nur auf den gerade gewählten Term angewendet werden. Es ist auch möglich Änderungen auf die gerade gewählte Selektion von Termen anzuwenden.

Beim Erstellen oder Ändern eines Begriffs werden Prüfungen bezüglich der Konsistenz durchgeführt:

- Prüfung, ob ein solcher Begriff schon vorhanden ist
- Prüfung, ob der Begriff Teil eines schon vorhandenen ist
- Prüfung, ob der Begriff sich nur in einem Zeichen von einem anderen unterscheidet

Folgende Felder stehen pro Term zur Verfügung:

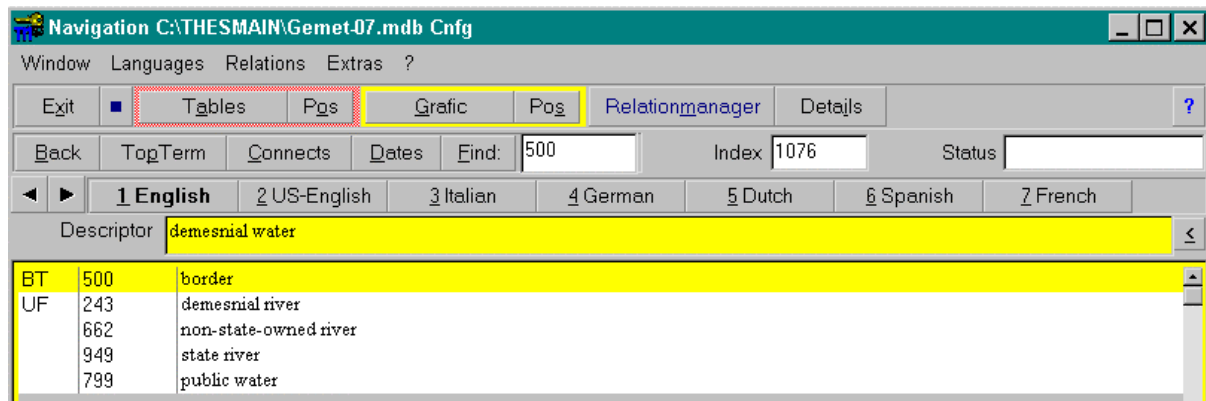
- | | |
|-------------------|------------------------|
| - Der Term selbst | - Type |
| - Sort | - Coincidence |
| - Scopenote | - Themes |
| - Definition | - Groups |
| - Indexer | - Coincidences |
| - Genitiv | - Themes |
| - Plural | - Groups |
| - Alternate Form | - Thesaurus references |
| - Source | |

Navigations Fenster

Wesentliche Funktionen

- Darstellung eines Terms mit seinen Relationen
- Datenaustausch mit anderen Programmen über die Windows - Ablage
- Positionieren im Kontext durch Doppelklicken der gewünschten Relation
- Positionieren außerhalb des Kontexts durch Eingabe der Nummer des gewünschten Terms
- Wahl von Sprachen in sehr kurzer Zeit. Font und Zeichensatz werden automatisch angepaßt
- Positionieren in das grafische oder tabellarische Fenster
- Anzeige von angebondenen externen Datenbeständen
- Anzeige der Topterms der gewählten Hierarchie
- 16 - stufige Undo - Funktion

Das nachfolgende Bild zeigt eine typische Darstellung im Navigationsfenster.



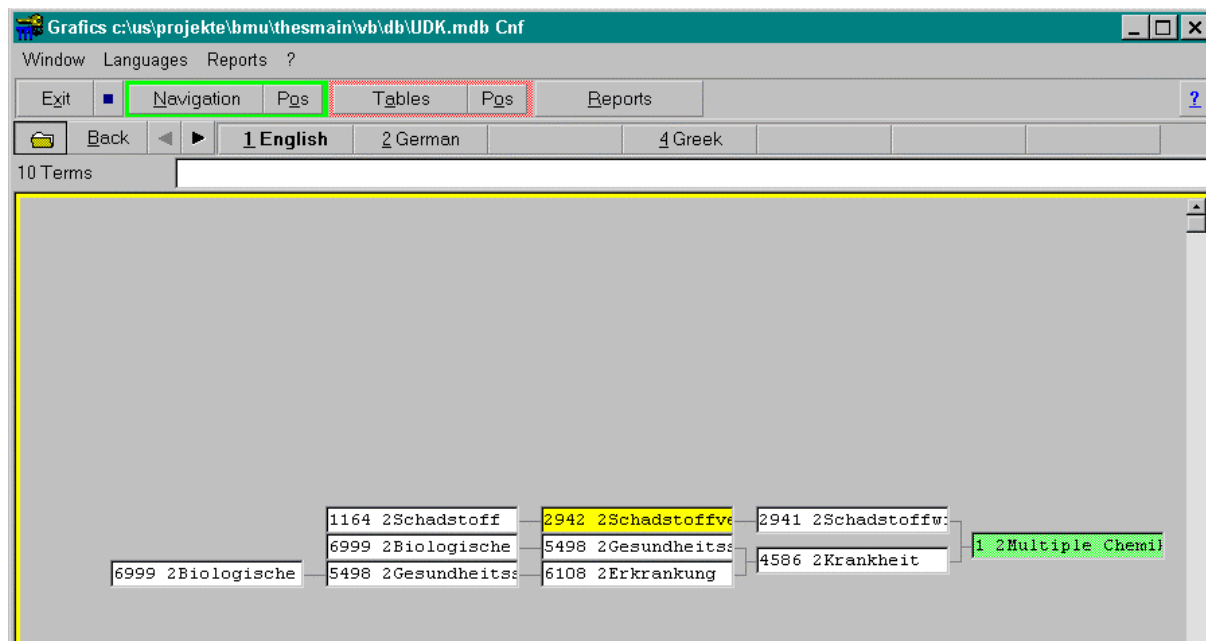
Im Navigationsfenster werden auch Relationen geändert, angelegt oder gelöscht. Dies wird mit der Funktion „Relationmanager“ erreicht. Änderungen werden pro Termpaar aber auch für ganze Selektionen von Termen durchgeführt. Reziproke Einträge werden automatisch erzeugt.

Grafische Darstellung

Wesentliche Funktionen

- Display aller Ober- und Unterbegriffe in grafischer Darstellung. Wahlweise können entweder alle Hierarchieebenen des Terms oder nur die übergeordnete und untergeordnete Hierarchieebene angezeigt werden
- Die Darstellung erfolgt in Textfeldern, die Relationen werden durch graue Linien dargestellt. Die Größe und der Abstand der einzelnen Elemente kann eingestellt werden
- Wahlfreie Anwahl von Begriffen und Selektion mit Maus oder über die Tastatur
- Lange Begriffe werden in der Statusleiste angezeigt
- 16 - stufige Undo - Funktion

Das nachfolgende Bild zeigt eine typische Darstellung im Grafikfenster:



Einige Zusatzfunktionen

Der Reportgenerator

Wesentliche Funktionen

- Alle im Thesaurus enthaltenen Daten können in eine Datei oder auf dem Drucker ausgegeben werden
- Das Layout, Kopf - und Fußzeile, Seitennumerierung usw. sind von Benutzer einstellbar
- Standardformate sowie die Druckformate für den Thesaurus des Umweltdatenkataloges bzw. von GEMET sind vordefiniert
- Bei Verwendung des Standardformats ist jedes Feld selektierbar
- Konfigurationen können für späteren Gebrauch gespeichert werden
- Jeder Ausdruck wird über eine Druckvorschau erzeugt.
- In der Druckvorschau können verschiedene Parameter eingestellt werden
- Die Ausgabe erfolgt auf den Windows - Systemdrucker
- Dateninterfaces zu MS-Word und MS-Excel sind verfügbar
- Einfache Erzeugung einer Konkordanzliste mit der Möglichkeit, Stopwortlisten anzugeben

Sprachen

- Bis zu 30 Sprachen können definiert werden
- Jede Sprache kann ihren eigenen Font und Zeichensatz verwenden

- Die Reihenfolge der Sprachen in den Tabellen und im Navigationsfenster kann vom Benutzer eingestellt werden.

Eine wesentliche Eigenschaft der Spracheinstellung ist die Möglichkeit, Sprachen mit unterschiedlichen Zeichensätzen gleichzeitig darzustellen. Dazu muß aber auch das Betriebssystem des Rechners vorbereitet werden.

- Zur Verfügung stellen des Windows „Multilanguage Support feature“
- Zur Verfügung stellen der nötigen Zeichensätze

Export / Import

Alle Daten eines Thesaurus können exportiert und importiert werden. Als Format steht ein SGML Austauschformat zur Verfügung, das auch in anderen Anwendungen, wie etwa dem Umweltdatenkatalog, sowie den Programmen der europäischen Umweltagentur zur Anwendung kommt.

Benutzerverwaltung

Die Thesaurusdatenbank ist immer mit Nutzernamen und Passwörtern geschützt. Mittels eines mitgelieferten Tools können neue Nutzer mit entsprechenden Zugriffsrechten erzeugt werden.

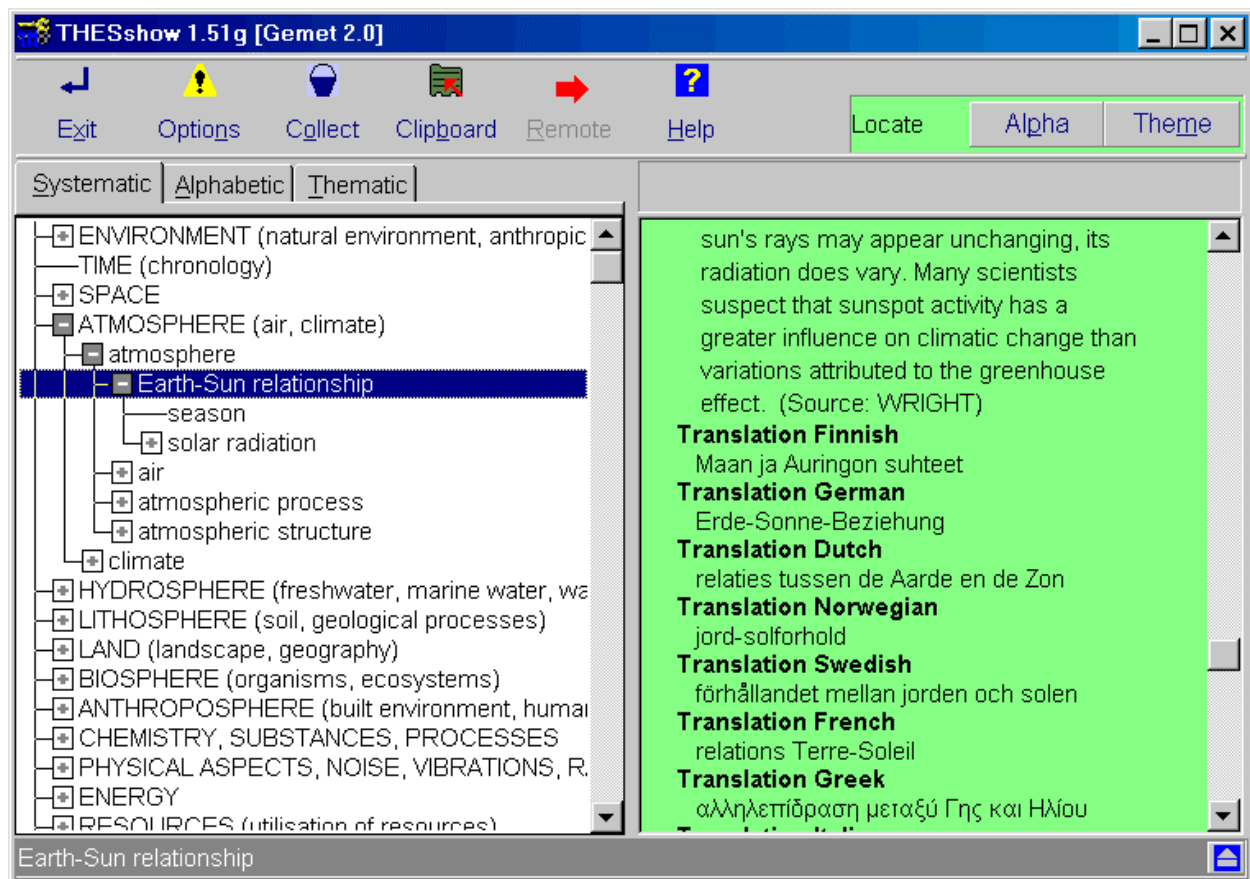
Derzeit sind folgende Nutzergruppen verfügbar:

- Administrator
- Configuration (darf keine Nutzerrechte vergeben)
- Management (darf keine Thesaurusstrukturen ändern)
- Processing (darf keine Thesauri anlegen oder löschen)
- Report (darf keine Änderungen an den Daten vornehmen)
- Read (darf nur lesen)

THESshow: eine Anwendung zur Thesaurusvisualisierung

THESshow ist das Visualisierungswerkzeug für THESmain basierende Thesauri. Es wird derzeit für den Thesaurus des Umweltdatenkataloges (UDK T, Version 4.0) sowie für den Thesaurus der europäischen Umweltagentur (GEMET, Version 2.0) verwendet. Es gestattet dem Nutzer in einfacher Weise den Datenbestand zu durchsuchen. Zum Einstieg in die Daten eignet sich entweder die systematische Darstellung, wo von wenigen Topterms aus die Menge der Deskriptoren durch Durchwandern der Hierarchien erschlossen werden kann, oder die alphabetische Darstellung, wo durch Eingabe eines Wortes ein Einstiegspunkt gefunden werden kann. Es ist dabei möglich, per Knopfdruck von einer Darstellung in die andere zu wechseln, wobei auf den gleichen Term positioniert wird.

Das folgende Bild zeigt eine typische Ansicht in systematischer Darstellung. Beachten Sie bitte auch die gleichzeitige Darstellung verschiedener Zeichensätze im grau unterlegten Detailfenster.

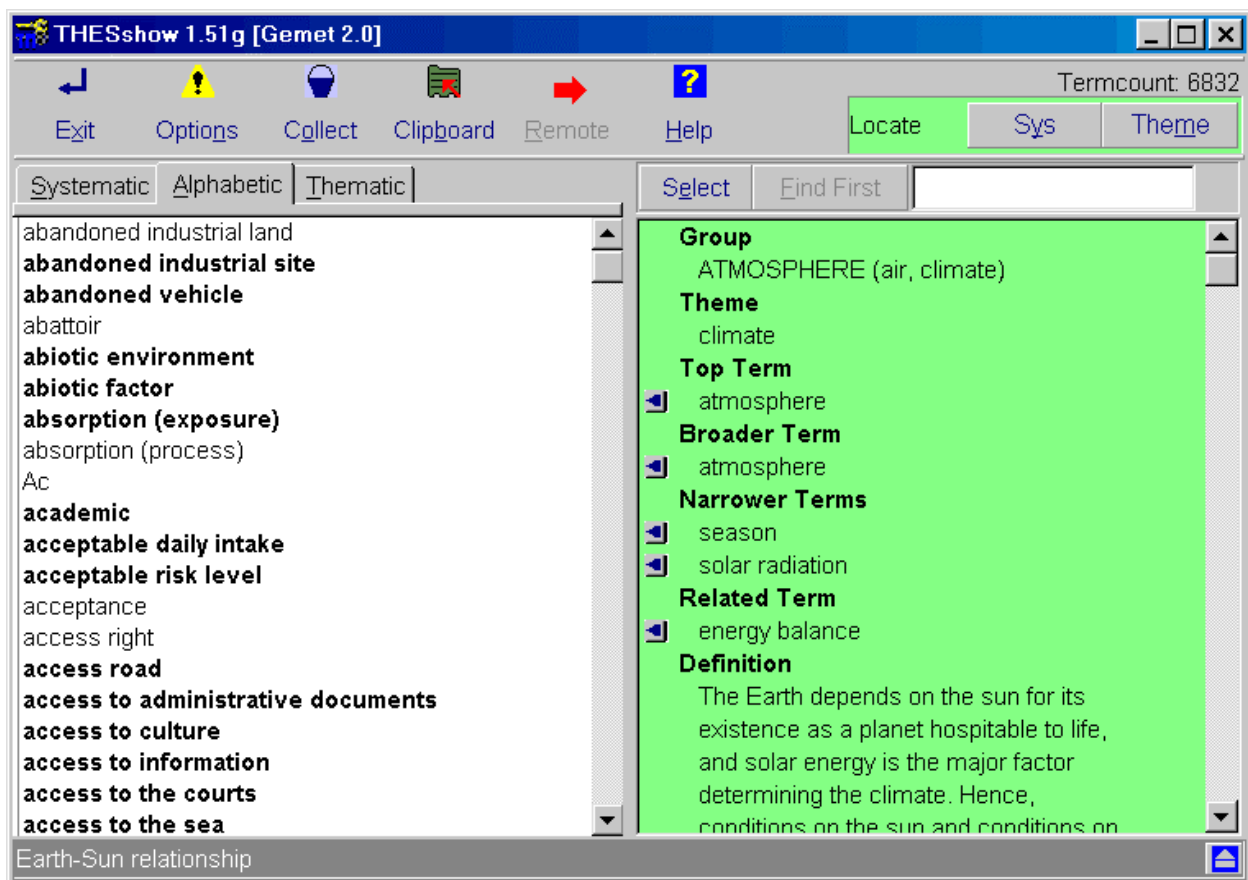


Wesentliche Funktionen

- Verfügbar auf CD-ROM und als installierbare Datei via Internet
- Einfache, automatische Installation
- Der Datenbestand kann wahlweise von CD-ROM gelesen werden. In diesen Fall ist der Platzbedarf nur wenige Megabyte
- Vielsprachig
- Einfache und flexible Konfiguration durch den Benutzer
- Möglichkeit der Erstellung von Teildatenbeständen
- Möglichkeit zur Darstellung polyhierarchischer Daten
- Darstellung des Thesaurus in systematischer, alphabetischer und thematischer Form
- Sehr schnelle Suchfunktion in alphabetischer Darstellung
- Automatisches Auffinden eines Begriffs in anderen Darstellungen
- Möglichkeit des Einbindens von Mikrothesauri
- Möglichkeit der Verwendung mehrerer Thesauri

- Darstellung der Details eines Terms inklusive aller Übersetzungen und deren Synonyme unabhängig von Sprache und Zeichensatz (Griechisch!, Cyrillisch!)
- Datenbestände in allen europäischen Sprachen können visualisiert werden, sofern das Betriebssystem diese Sprachen unterstützt
- Auf Knopfdruck können durch Relationen verknüpfte Begriffe lokalisiert werden

Das folgende Bild zeigt eine typische Ansicht in alphabetischer Darstellung. Deskriptoren sind dabei fett dargestellt. Die Schaltflächen mit den Pfeilen dienen zur Lokalisierung von Ober- und Unterbegriffen.



Dieses Softwarewerkzeug läßt sich nicht nur für die Erstellung, Pflege und Visualisierung von Thesaurusdaten einsetzen. Jegliche hierarchische Struktur wie z.B. Organigramme großer Institutionen sind mit ihm leicht und optisch überzeugend darstellbar.

Nicht nur das ETC / CDS zur Verwaltung und Pflege des zwölfsprachigen Europäischen Umweltthesaurus GEMET und die Deutsch / Österreichische Kooperation zum UDK verwenden THESmain. Geplant ist weiters, künftig die Verwaltung des Infoterra - Thesaurus EnVoc des Umweltprogramms der Vereinten Nationen UNEP mittels THESmain vorzunehmen. Auch die amerikanische Umweltbehörde will THESmain zur Erweiterung des GEMET um Sprachen des pazifischen Raumes einsetzen.

Die Produkte „THESshow“ und „THESmain“ können auch für kundenspezifische Anforderungen adaptiert und zur Verfügung gestellt werden. Lizenzierungsmodelle für den Bereich Forschung und

Lehre sowie für die kommerzielle Nutzung können unter der e-Mail-Adresse: legat@ubavie.gv.at angefordert werden.

**

Namen und Adressen der Autoren:

Wolf - Dieter Batschi, Umweltbundesamt, Bismarckplatz 1, D - 14193 Berlin

Tel: ++ 49 (0)30 8903 - 2423; Fax: ++ 49 (0)30 8903 - 2102; e - Mail: wolf-dieter.batschi@uba.de

Internet: <http://www.umweltbundesamt.de>

Rudolf Legat, Umweltbundesamt, Spittelauer Lände 5, A - 1090 Wien

Tel: ++ 43 (0)1 31304 - 5404; Fax: ++ 43 1 31304 - 5400; e - Mail: legat@ubavie.gv.at

Internet: <http://www.udk.ubavie.gv.at>

Hermann Stallbaumer, Technisches Büro für Elektrotechnik TBHS, Favoritenstraße 182, A - 1100 Wien

Tel: ++ 00 43 (0)2236 76232 - 32; Fax: ++ 43 (0)2236 76232 - 76; e - Mail: hermann@tbhs.co.at