3. Tagung der deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation (ISKO), Weilburg, 27.-29.10.93

Überlegungen zu einem multilingualen Thesaurus-Konzept

Harald H. Zimmermann

Zusammenfassung

Die Thesaurus-Thematik wird zunächst in den Zusammenhang der gesamten Erschließungs- und Retrievalmöglichkeiten eines Information-Retrieval-Systems gestellt.

Auf dieser Grundlage wird ein multilinguales Thesaurus-Konzept entwickelt und am Beispiel des IRSystems MITI SELECTOR in wichtigen Details und Verfahrensweisen vorgestellt.

Wichtige Elemente sind: die Ermöglichung des Zugangs anhand des Benutzervokabulars, eine systematische, transparente Bedeutungsdifferenzierung und eine Basis-Relationierung anhand einer einzigen ("ausgezeichneten") natürlichen Sprache.

Die wesentlichen Erfahrungen bei der Entwicklung des MITI-Thesaurus und dessen Einbindung und Anwendung im Prototypen MITI SELECTOR werden vorgestellt.

1. Allgemeine Einführung

1.1 Integration von Recherchealternativen

Üblicherweise steht bei Thesauri - monolingual wie multilingual - die Anwendung bei der intellektuellen Erschließung (Indexierung) und einer intellektuellen Recherche (Retrieval) im Vordergrund. Weitere Verfahren, etwa die Volltexterfassung, die automatische Indexierung und die Dokumenterstellung, bleiben weitgehend ausgeklammert

Dies ist im Hinblick auf die neueren Entwicklungen im Information Retrieval insgesamt unzureichend, da in der heutigen Praxis (evtl. abgesehen von der Sammlung externer Quellen) die meisten Materialien auch zur Volltextrecherche verfügbar sind oder zumindest bereitgestellt werden sollten.

Insgesamt erscheinen für die Erschließung und die spätere (Text-)Dokumentsuche folgende Alternativen wünschenswert, die im übrigen bei der Recherche auch kombiniert genutzt werden können:

(1) **Formulargestützte Suche**, wobei nach Dokumenttyp (z.B. Literatur, Anfrage, Fragestunde...), nach Feldinhalten (z.B. Autor, Anfragender, Zusammenfassung, Sitzungsdatum ...) usf. selektiert werden kann.

(2) **Freitextsuche** im weitesten Sinne, d.h. basierend auf Textwortformen. Hierzu stehen bekanntermaßen Volltextretrievalsysteme zur Verfügung (Trunkierung, Abstandsmaße ...)

(3) **Klassifikationsorientierte Suche**. Dies setzt das Vorliegen einer (evtl. facettierten) Klassifikation voraus (Musterbeispiel: die Patentklassifikation).

Die Vergabe (Erschließung; Klassierung) kann im Zusammenhang der (intellektuellen) Deskriptorenvergabe (Indexierung) erfolgen und dabei vom gleichen Personal durchgeführt werden. Sie bedeutet praktisch keinen Mehraufwand, sofern ein Dokument intellektuell indexiert wird.

Im Gegensatz zur **Deskriptorenvergabe** (die als Postkoordinationsinstrument eingesetzt wird) bedeutet die **Klassierung** eine Art Clusterbildung (Präkoordination) der Dokumente. Ein typisches Beispiel mit weltweiter Anwendung ist das Prinzip der Patentklassifikation (Hauptklasse / Nebenklasse).

(4) **Thesaurusbasierte Suche**. Dieses Verfahren ist **obligatorisch** für Dokumente, die nicht auf andere Weise recherchierbar sind, etwa nur graphisch und nicht zeichenbasiert speicherbare Presseartikel; obwohl sich hier inzwischen ja auch einiges "am Markt" entwickelt und Zeitungsverlage ihre eigenen Daten in maschinenlesbarer Form bereitstellen (s.u.).

Da die **intellektuelle Indexierung** als ein *wissensbasiertes Verdichtungsinstrument* gesehen werden kann, ist sie auch dann weiter sinnvoll, wenn Daten maschinenlesbar vorliegen. Automatische Verdichtungssysteme wird es - Ausnahmen in Bereichen, in denen die Quellen (im "Volltext") selbst stark standardisiert sind bzw. ein extrem eingeschränktes Fachgebiet vorliegt, kann es dabei durchaus geben.

Ein moderner Thesaurus muss daher stärker in den Gesamtzusammenhang einer Anwendung gestellt werden: welche Felder welcher Dokumenttypen werden wie recherchierbar, welche Rolle spielt die Freitextrecherche, lässt sich diese mit dem Thesaurus verbinden ...?.

1.2 Maschinen- und plattformunabhängiges Datenformat

Häufig werden elektronische Thesauri in Abhängigkeit von vorhandenen technischen Bedingungen (Software, Betriebssystem ...) entwickelt. Dies ist natürlich für die praktische Umsetzung relevant, darf aber im Grunde nicht Ausgangspunkt einer Konzeption sein oder diese in einem konzeptionellen Stadium zu sehr beeinflussen.

Angesichts der hohen Entwicklungskosten und der Langfristigkeit der späteren Lösungen muss es Ziel sein, sich bezüglich des Einsatzes von Techniken möglichst unabhängig zu machen.

Es ist ein systemunabhängiges Verfahren anzustreben, das es erlaubt, einerseits die Daten bestehender Datenbanken strukturiert zu halten bzw. verfügbar zu machen, andererseits Deskribierungen und Klassifikationen einzubringen und schließlich auch zur Textverarbeitung und zum Papierausdruck (Publishing) bereitzustellen.

Es bietet sich heute an, eine "thesauruspezifische" SGML-Struktur (mit Feldtypisierungen) zu entwickeln, die ggf. angereichert wird durch prozessuale Elemente, soweit dies nicht aus der Struktur des Thesaurus ableitbar ist.

Der Transport der Daten (evtl. auch unter Selektionsgesichtspunkten) von der neutralen Plattform zur konkreten Anwendung (und zurück) muss über Export- und Importfunktionen gesteuert werden.

Ein zu realisierender Thesaurus und ggf. die entsprechende Klassifikation müssen auf ein solches Format abgebildet werden (können), ohne dass die Strukturinformation verloren geht. Beispiele dafür (im Bereich der elektronischen Lexika) gibt es heute genügend.

Insgesamt sollte also bei elektronischen Thesauri eine Unabhängigkeit von einer (wo und von wem auch immer bereitgestellten) Soft- oder Hardware sichergestellt werden, insbesondere in Verbindung mit Export- und Importfunktionen, aber auch durch Loslösung der Pflege von einer spezifischen Anwendung.

1.3 Zur Frage der "Tiefe" eines Thesaurus bzw. einer Deskribierung

Folgende Kriterien werden als zeit- und aufgabengemäß empfunden:

- Wenn ein Thesaurus auch durch nicht auf das Retrievalsystem spezialisierte Nutzer (etwa Wissenschaftler, Parlamentarier ...) genutzt werden soll, reichen die heutigen Standard-Inhalte (und Strukturen) nicht aus. Sie können andererseits einen Ausgangspunkt bilden.

- Die Themen des Gegenstandsbereichs sollten möglichst "problemnah" und in den verwendeten Benennungen beschrieben werden. Der Thesaurus dient damit zwar *zur Beschreibung der relevanten Themen eines Dokuments* (i.S. einer - notwendigen - *Verdichtung* auf das als wesentlich Angenommene), bei den Benennungen bleibt er jedoch - stellvertretend für die Begriffe / Themen - möglichst präzise, problem- und und anwendungsnah.

- Der Thesaurus sollte nicht *präskribierend*, sondern *deskribierend* sein. Grundlage ist der zentrale Gegenstandsbereich der Dokumente (d.h. die konkreten Themen), Qualitätsmaßstab ist die "Precision" in Verbindung mit dem "Recall" beim Retrieval. (Beim Aufbau kann man ggf. auf bestehende Materialien auch unter statistischen Aspekten zurückgreifen.)

- Die Relationierungen sollten in erster Linie im Hinblick auf das Retrieval entwickelt werden; natürlich sind dies und die ständige Pflege eine Aufgabe für Spezialisten (Linguisten / Dokumentare).

1.4 Bedeutung der Multilingualität

Der Sicherstellung des multilingualen Zugangs kommt im Zusammenhang mit der internationalen Kommunikation (d.h. natürlich auch der Vermarktung von Informationssystemen) sowie in Ländern mit Mehrsprachigkeit (Schweiz, Kanada, die EG ...) eine besondere Bedeutung zu.

Für das kontrollierte Vokabular des Thesaurus bedeutet dies - insbesondere unter dem Aspekt, dass Nicht-Dokumentare einen Zugang dazu haben sollen - u.a.:

- Aufbau einer **Sprachschnittstelle "Allgemeinwortschatz - Thesauruswortschatz**" (Disambiguierung, Bedeutungserklärung ...)

- Bereitstellung von Übersetzungsäquivalenten (bzw. Quasi-Äquivalenten) zu den Deskriptoren (und verzeichneten Nicht-Deskriptoren) des Thesaurus in den betreffenden Sprachen.

Da die *Freitextrecherche* (bezogen auf maschinenlesbare Quellen) in zukünftigen Anwendungen einen wichtigen Teilaspekt ausmachen wird, ist neben dem Thesaurus-Vokabular ein am Datenmaterial orientiertes (mehrsprachiges) *elektronisches Wörterbuch* zu entwickeln bzw. zu integrieren, das als *Zugangshilfe* in solchen Fällen dient, in denen Volltexte nicht in allen relevanten Sprachen verfügbar sind.

Die Verwendung einer *Klassifikation* ist "an sich" sprachneutral. Dennoch bedarf es (auch hier kann als Exemplum auf die Patentklassifikation verwiesen werden) eines (sprachbezogenen) *Stichwortverzeichnisses,* um einen Nicht-Spezialisten in der Klassifikation (letztlich aber auch den Dokumentar) zur Klassifikation hinzuführen, wozu im Thesaurus ggf. eine Beziehung "Deskriptor < - > Klassifikation" aufgebaut werden kann. Da diese Stichwörter auch in den relevanten Sprachen verfügbar sein müssen, bietet sich entweder eine Verbindung mit einem elektronischen Übersetzungswörterbuch oder aber mit dem Thesaurus (oder aber beides) an.

Wenn man den multilingualen Thesaurus als ein offenes semantisches Netz versteht, könnte man auch versuchen, beide Komponenten (Thesaurus und Wörterbuch) in einem Gesamtsystem - differenziert - darzustellen. Die dazu notwendige Differenzierung ist in einem "elektronischen" System kein besonderes Problem.

Die ISO-Norm - dies zeigt v.a. die Diskussion in Deutschland - ist für die "klassischen" Thesauri entwickelt worden und muss für elektronische Anwendungen in jedem Falle erweitert werden.

2. Exemplum: Der multilinguale Thesaurus des MITI SELECTOR

2.1 Problemstellung für MITI

Wer heute als nicht-professioneller Anwender in elektronischen Datenbanken (Fachinformationsbanken) recherchieren möchte, ist kaum in der Lage, das Angebot zu überblicken, geschweige denn - wenn einmal die Auswahl getroffen ist - eine sachgerechte Abfrage durchzuführen.

Die Gründe hierfür sind vielfältig:

Probleme bei der Auswahl der relevanten Datenbank

- Angesichts der inzwischen bestehenden Vielfalt von Online-Datenbanken - der Cuadra verzeichnet allein über 2.000 - gibt es außer kurzen Angaben in generellen Übersichten kaum systematisch-inhaltliche Beschreibungen. Etwas ausführlicher informieren meist die sog. Bluesheets, doch bilden diese selten die Grundlage einer Erstauswahl.

- Für einen Gelegenheitsnutzer ist die Auswahl besonders dann schwer, wenn er spezielle Fragen hat.

- Ein Zugang ist nur über die Beschreibungssprache der Quelle (heute meist Englisch) möglich. Dies setzt voraus, dass man in der entsprechenden (englischsprachigen) Terminologie des Sachgebiets "aktiv' bewandert ist.

Probleme bei der Suche in einer Datenbank:

- Die *Anfragesprache* (Kommandosprache) des Information-Retrieval-Systems wird vom Nutzer nicht ausreichend beherrscht.

- Möglichkeiten der *Suche mit datenbankspezifischen Feldern* werden aufgrund fehlender bzw. unzureichender Kenntnisse nur eingeschränkt genutzt.

- Die *Recherche mit Freitext* verlangt umständliche technische Handlungen (ein Beispiel ist die Trunkierung).

- *Fachbegriffe* in der "Datenbanksprache" sind dem Nutzer nicht gegenwärtig.

- *Thesauri und Klassifikationen* lassen sich aufgrund unzureichender Erfahrung nur unzureichend einbeziehen.

Dies führt gegenwärtig zu folgenden Konsequenzen:

(1) Die Nutzerschulung ist aufwendig und strapaziös.

(2) Gelegenheitsnutzer werden - trotz grosser Fachkenntnisse - von der unmittelbaren, d.h. eigenständigen Nutzung abgehalten.

(3) Sofern aus diesen Gründen ein "Informationsvermittler" zwischengeschaltet wird, verteuert sich die Recherche unnötig (es gibt natürlich häufig gute Gründe, einen Recherche-Spezialisten hinzuzunehmen, doch sollten diese nicht dadurch bestimmt sein, dass man ein Retrievalverfahren technisch unzureichend beherrscht).

(4) Angesichts der durch diese Handling-Probleme mit verursachten relativ geringen Nutzung sind heute die Datenbankpreise sehr hoch.

Die Problematik ist vielfach bekannt und beschrieben. Es gibt zudem inzwischen einige Ansätze und Verfahren, die für Abhilfe sorgen sollen:

- Fast schon "klassisch" ist die Definition und Einführung einer sog. "Common Command Language": statt verschiedener hostspezifischer An- und Abfragesprachen kann eine - gegenüber der speziellen Sprache meist leicht eingeschränkte - neutrale Sprache verwendet werden.

- Den kommandoorientierten Retrievalsprachen werden Anfragemenüs zu- oder vorgeschaltet. Der Nutzer muss hierbei die Datenbankstruktur nicht mehr kennen, doch ist er in seinen Anfragemöglichkeiten etwas eingeschränkt. Typisch für diese Art der Benutzer-

schnittstelle sind beispielsweise die sog. OPACs (Online Public Access Catalogues) für Literaturanfragen in Bibliothekskatalogen.

- In neueren Entwicklungen der hostbezogenen Retrievalsysteme werden die Nutzer zunehmend besser unterstützt, v.a. zur Suche in Spezialdatenbanken, etwa mit chemischen Formeln.

Dennoch bleibt ein Teil der Möglichkeiten, die spezifische Systeme und Datenbanken *grundsätzlich* bieten, für den Standardanwender unzugänglich.

2.2 Das MITI-Konzept

MITI steht für ein Produktkonzept, bei dem ein Nutzer in "seiner" natürlichen Sprache ohne technische Retrievalkenntnisse einerseits die Auswahl der relevanten Datenbank treffen und dann in ihr so qualifiziert suchen kann wie der professionelle Informationsvermittler. Auch ein professioneller Nutzer soll jedoch von diesen Möglichkeiten profitieren können.

Für den ersten Teil (die *Datenbankselektion*) steht die Entwicklung des **MITI SELECTOR**. Die forschungsbezogenen Entwicklungen zum MITI SELECTOR sind mit dem Vorliegen eines Prototypen abgeschlossen; sie werden in diesem Beitrag ausführlicher vorgestellt. Für den zweiten Aufgabenbereich, die *Suche in verschiedenen Datenbanken auf unterschiedlichen Hosts,* steht die Entwicklung des Prototypen **MITI RETRIEVER** Dieser Teil des Projekts soll hier nur kurz vorgestellt werden.

Die Entwicklungen von MITI erfolgten im Rahmen eines europäischen Forschungsprojekts. Das Projekt wurde im Rahmen des EG-Förderprogramms IMPACT finanziell unterstützt. Folgende Firmen und Forschungsteams waren an den Entwicklungen beteiligt:

- SOFTEX GmbH, Saarbrücken als Hauptkontraktor und Koordinator: zuständig für den *Aufbau der MITI-Wissensbank "MITI-KDB*" (auf der Basis einer relationalen Datenbank), für die *linguistischen* Teile (inkl. der elektronischen Wörterbücher) sowie den **MITI SELECTOR** als Stand-alone-Teillösung;

- INFOPARTNERS S.A., Luxemburg: zuständig für die Konzeption und Entwicklung des **MITI RETRIEVER**;

- GMD (IPSI), Darmstadt: zuständig für die spezifische Oberflächengestaltung (Bereitstellung einer entsprechenden generalisierten Rahmensoftware) und für die Entwicklung einer Subversion (MITI CORDIS);

- UPS (Universität Paul Sabatier), Toulouse: zuständig für die Entwicklung des Add-On-Systems EURISKO zum Relevance-Feedback im Rahmen des MITI RETRIEVER;

- EUROBROKERS S.a.r.l., Luxemburg: zuständig für Datenbereitstellung, Evaluierung und Marketing.

Mit MITI werden folgende Konzepte realisiert:

Nutzung standardisierter Rahmensoftware und -module

- Die Produkte werden (zunächst) unter UNIX in den Varianten SCO-UNIX (mit OSF-MOTIF und X-WINDOWS, überlagert von dem TOOLKIT der GMD) und unter SUN (MITI CORDIS) bereitgestellt. MS-WINDOWS-Lösungen sowie Lösungen unter OS-2 sind für die Zukunft vorgesehen.

- Der Nutzer bewegt sich dadurch in einer "gewohnten" und weitgehend im Handling standardisierten Umgebung; neue Bedienertechniken müssen nicht erlernt werden.

- Der GMD-TOOLKIT erlaubt die Einbringung und Anwendung verschiedener *Interaktionssprachen*. Realisiert sind eine deutsche, eine französische, eine englische und eine spanische Oberfläche.

Gemeinsame Wissensbank "MITI-KDB"

- Den Entwicklungen MITI SELECTOR und MITI RETRIEVER liegt in weiten Teilen (d.h. dort, wo sich aus der Selektion einer Datenbank für das weitere Retrieval in dieser Datenbank inhaltliche Beziehungen bzw. Überschneidungen ergeben) eine gemeinsame Wissensbank (MITI-KDB) zugrunde. Für den Prototypen wird INGRES verwendet (die Subversion MITI CORDIS benutzt SYBASE); eine Unterlegung mit anderen SQL-basierten Datenbanken ist möglich.

- Die relationale Datenbank umfasst Beschreibungen der jeweiligen Datenbank bzgl. ihrer Struktur (Feldeinteilung) und bzgl. der relevanten Inhalte (z.B. Fachgebiete).

- Die Datenbankbeschreibungen sind den jeweiligen Hosts und Hostbeschreibungen zugeordnet, die ebenfalls in der Wissensbank entsprechend erfasst sind.

- Wenn ein Nutzer eine Datenbank und einen Host über den MITI SELECTOR selektiert hat, wird dieses Ergebnis (u.a. die spezifizierten Feldnamen und deren Verwendung bei der Datenbankabfrage auf dem Host) dem MITI RETRIEVER bei Bedarf zugänglich.

Komfortable Datenbankselektion: MITI SELECTOR

- Durch Verwendung des GMD-Shells, wie er analog in der TORI-Anwendung von CORDIS realisiert wurde, erfolgt der Zugang über angepasste bzw. leicht weiter anpassbare Formulare.

- Die Abfragen im MITI SELECTOR können unter alternativer Verwendung deutscher, französischer, englischer oder spanischer Fachbegriffe erfolgen. Die Zuordnung erfolgt über inhaltlich spezifische *MITI-Lexika*, die weitgehend (einschließlich der Zuordnung der Übersetzungen) den **Möglichkeiten von Thesauri** entsprechen: Das interne Suchvokabular ist also "kontrolliert", über Thesaurusrelationen erfolgen Zuordnungen von Nutzersuchwörtern zu den zugelassenen (systemseitig eindeutigen) Termini.

- Der Benutzer wird bei der Auswahl nach Themenbereichen über sein (freies) Suchwort zu einem Deskriptor hingeführt, der einen Treffer "verspricht".

- Das (triviale) Darstellen graphematisch korrekter Schreibweisen (Umlaut, Akzentuierung ...) kann (soweit eindeutig) entfallen (interne Verwendung eines Spellcheckers).

- Ein Browsen in verschiedenen Richtungen (Oberbegriff, Unterbegriff, "Siehe-auch") ist möglich.

- Für die interne Beschreibungssprache wurde Deutsch zugrundegelegt. Über den Klärungsdialog wird sichergestellt, dass ein Nutzer mit "seiner" Anfragesprache (Französisch, Englisch oder Spanisch) auskommt, ohne in diesen Fälle die interne Sprache kennen zu müssen.

Beim gegenwärtig realisierten Prototypen erfolgt die Pflege (Erweiterung der Datenbankbeschreibungen, Vergabe der internen Deskriptoren, Kontrolle des Beschreibungsvokabulars, Zuordnung natürlichsprachiger Begriffe zum Systemvokabular ...) durch die Entwickler.

Es ist jedoch vorgesehen, dem Nutzer selbst über eine Pflegekomponente eigenständige Erweiterungen und Modifikationen zu ermöglichen.

Entlastung von Trivialitäten bei der Datenbanksuche: MITI RETRIEVER

- Durch Verwendung des GMD-Toolkits wird auch hier eine Interaktion in verschiedenen Benutzersprachen möglich (zunächst realisiert mit deutscher, französischer, englischer und spanischer Oberfläche).

- Bei der Suche kann zwischen einer einfachen formalen Anfragesprache und einer Menüoberfläche gewählt werden. Ist die relevante Datenbank bekannt, werden dem Nutzer die hier möglichen suchbaren Felder (zur Auswahl) angegeben.

- Soweit Felder verzeichnet sind, die *Deskriptoren oder Klassifikationen* enthalten, werden - soweit vorhanden - entsprechende Auswahlverfahren (mit Browsing) angeboten.

- Im *Freitextbereich* wird bei Bedarf eine automatische Trunkierung (alternativ: eine *Grundformenermittlung*) zur Verfügung gestellt.

- Dem System ist (über die MITI-KDB) bekannt, in welcher Sprache (und in welcher Schreibvariante, z.B. bei Deutsch: mit / ohne Umlaut; bei Französisch: mit / ohne Akzente) eine Datenbank recherchiert werden muss.

- Soweit dies terminologisch möglich ist, werden Anfragen (d.h. Wörter), die nicht in der Suchsprache der Datenbank erfolgen, in diese Sprache übersetzt. Dem Nutzer ist es möglich, unter den zur Verfügung gestellten Äquivalenten eine Auswahl zu treffen (sofern er die Suchsprache des Systems beherrscht).

- Über eine *Wörterbuchpflegekomponente* wird der Nutzer in die Lage versetzt, weitere Äquivalente für spätere Übersetzungen in das relevante Wörterbuch einzutragen.

Im Gegensatz zum MITI SELECTOR ist beim MITI RETRIEVER eine Kontrolle des Suchvokabulars (i.S. der Verwendung eindeutiger Termini) nicht möglich.

2.3 Technische Rahmenbedingungen

Für die Nutzung des MITI SELECTOR und des MITI RETRIEVER (als Prototyp) bildet derzeit ein PC 486 mit SCO UNIX und INGRES als relationalem Datenbanksystem die technische Grundlage.

Während der MITI SELECTOR stand-alone genutzt werden kann, ist für die Nutzung des MITI RETRIEVER anwenderseitig ein Anschluss an DFÜ-Netze (etwa DATEX-P) sowie die Berechtigung zur Datenbanknutzung auf einem Host erforderlich. An die Stelle der Hostkennung kann auch ein Mailboxanschluss treten.

2.4 Status, Testphase, weiterer Ausbau

Die Prototypen können nach Fertigstellung (dies ist für den MITI SELECTOR seit März 1993 der Fall) Interessenten zur probeweisen Nutzung bereitgestellt werden.

Gegenwärtig sind beim MITI RETRIEVER Übersetzungen nur in den Sprachpaaren Deutsch / Englisch, Deutsch/Französisch und Deutsch/Spanisch möglich.

Für den ersten Test wurden 4 Hosts und Beschreibungen von 20 Datenbanken ausgewählt. Perspektivisch ist ein Ausbau auf alle relevanten Datenbanken und wichtige Hosts vorgesehen. Dies soll über eine enge Zusammenarbeit mit den Datenbankanbietern und Hosts erreicht werden, sofern entsprechendes Interesse besteht.

2.5 Zur linguistischen Verfahrensweise des MITI SELECTOR

Die im folgenden behandelten linguistischen Fragen lassen sich in die drei Problemkreise *monolinguale Lösungen, multilinguale Lösungen* und *Thesaurusanwendung* untergliedern. Es ist zunächst darauf hinzuweisen, dass sich der MITI SELECTOR hierbei deutlich von den Anwendungsmöglichkeiten des MITI RETRIEVER unterscheidet:

Beim MITI SELECTOR handelt es sich um ein weitgehend *autarkes* System, insofern alle inhaltlichen Daten unter der Kontrolle der Entwickler stehen. Dies bot (u.E. erstmals) die Möglichkeit, alle Varianten der Thesaurus-Anwendung zu erproben, ausgehend von der Deskriptorenvergabe (Auswahl und Schreibform der Benennungen über die Differenzierung von Bedeutungen bis hin zu den Übersetzungen und dem Retrieval.

Der MITI RETRIEVER ist demgegenüber abhängig von den Daten (auch Klassifikationen und Deskriptoren) der Datenbanken und Hosts, der jeweiligen Sprache(n) der Abstracts usf.

3. Das Thesaurus-Konzept des MITI-SELECTOR

Die Strategie von MITI sieht beim MITI SELECTOR folgendes vor:

(1) Ein Nutzer soll *möglichst früh* wissen, ob ein von ihm benutztes Wort auch ein möglicher *Deskriptor* ist bzw. in der *Hierarchie des Thesaurus* (beim Browsing) einen Sinn macht.

(2) Es muss vermieden werden, dass *Bedeutungsvarianten nicht erkannt werden* und der Nutzer in die Irre geht (etwa, wenn er unter "Gericht" etwas anderes versteht als das System).

(3) Dem Nutzer kann andererseits zugemutet werden, dass sein originäres Wort zum Zwecke der weiteren Anwendung in die systemrelevante (ein-eindeutige) Form umgewandelt wird, soweit dies automatisch geschieht und er entsprechend darauf hingewiesen wird. Dieser Systemeintrag kann eine andere Schreibform sein (orthographische Variante), ein Pluraleintrag statt des Singular (auch umgekehrt) oder auch eine Synonym-Vorzugsbenennung.

(4) Bei all diesen Vorgängen bewegt sich der Nutzer in *seiner* Anfragesprache, ohne jemals die internen Äquivalente "sehen" zu müssen.

Der Nutzer kann andererseits von folgendem ausgehen:

(a) Wenn ein *sprachlich korrektes Wort* bereits rechtschreiblich nicht identifiziert wird, ist über den MITI SELECTOR kein Treffer möglich (dies macht es erforderlich, dass alle Wörter (Deskriptoren wie Übersetzungen) in jedem Falle beim Aufbau auch über den jeweiligen Spellchecker gehen).

(b) Wenn er eine Vereindeutigung (einer Bedeutungsvariante) vornimmt, bei der *in seiner Anfragesprache* der Hinweis "nicht suchbar" erfolgt, gibt es zu dieser Variante später keinen Treffer.

Dies wird dadurch erreicht, dass in der Phase des Disambiguierungsdialogs anhand des anfragesprachrelevanten Relationenwörterbuchs die *potentielle* ein-eindeutige Wortform dahingehend (ggf. über die Übersetzung) daraufhin überprüft wird, ob ihr ein Deskriptor unmittelbar oder über eine zugelassene Hierarchie zugeordnet werden kann. Ist ein Wort nicht im MITIR_D-Wörterbuch verzeichnet, ist dies der Nicht-Relevanz gleichzusetzen.

Damit wird es erforderlich, auch in der Datenbank nicht verfügbare (nicht-relevante) Bedeutungsvarianten mit ein-eindeutigen Zuordnungen zu versehen und diese auch zu übersetzen, ohne dass dazu allerdings eine Beziehung aufgebaut wird. Für den Fall, dass ein Deskriptor ohne Thesaurusrelationen eingebracht wird, ist die *Notlösung* einer Pseudo-Relation vorgesehen.

3.1 Funktion und Einbindung der Wörterbücher in MITI SELECTOR

3.1.1 Grundlagen

Die Wörterbücher des MITI SELECTOR stellen einen auf die textuellen Datenbankbeschreibungen und auf die Feldnamen des MITI SELECTOR abgestimmten (und ggf. leicht modifizierten) **Subset** (z.Z. - wegen der fehlenden analogen Systematisierung der allgemeinen Wörterbücher - einen variierten und stärker systematisierten Subset) der allgemeinen SOFTEX-Wörterbücher dar. Eine Ausnahme macht die Benutzung des jeweiligen monolingualen Identifikationswörter-

buchs zur Rechtschreibkontrolle und Lemmatisierung (hier wird das allgemeine Verfahren angewendet).

Bei den "Wörtern" des MITI SELECTOR handelt es sich um ein kontrolliertes Vokabular in dem Sinne, dass alle Deskriptoren jeweils nur eine Bedeutung haben und zeichenmässig ebenfalls eindeutig sind.

Die Deskriptoren in der MITI-KDB (d.h. der mit SQL recherchierbaren INGRES-Datenbank) sind nur *in deutscher Sprache* festgehalten (Deutsch ist also die o.a. "ausgezeichnete" Sprache für die hierarchischen Relationierungen). Im Prinzip hätte es auch eine der anderen Sprachen sein können, doch wurde Deutsch aus *rein pragmatischen Gründen* ausgewählt: Bei SOFTEX ist das maschinenlesbare Grundinventar, aus dem bereits vorhandene Übersetzungen selektiert werden, überwiegend mit Deutsch als einem Quell- oder Zielsprachenteil vorhanden.

Es werden - neben den allgemeinen, MITI-unabhängigen Rechtschreibwörterbüchern - folgende *spezifische* MITI-spezifischen Wörterbücher unterschieden:

Übersetzungswörterbücher:

MITI_DE und MITI_ED für Deutsch / Englisch
MITI_DF und MITI_FD für Deutsch / Französisch
MITI_DS und MITI_SD für Deutsch / Spanisch

Relationenwörterbücher.

MITIR_D für Deutsch (zur nutzerseitigen Disambiguierung und Thesaurusrelationierung; umfangreich, s.u.)
MITIR_E für Englisch (nur zur nutzerseitigen Disambiguierung)
MITIR_F für Französisch (nur zur nutzerseitigen Disambiguierung)
MITIR_S für Spanisch (nur zur nutzerseitigen Disambiguierung)

Anm.: Für den MITI *RETRIEVER* sind zunächst nur die MITI-Übersetzungswörterbücher relevant, da hierüber die Übersetzungen der Feldnamen ermittelt werden. Bei einem weiteren - vorgesehenen - Ausbau des MITI RETRIEVERS kann man sich vorstellen, dass auch das Relationenwörterbuch mit Bezug zu den Klassifikationen und Thesauri (Nutzung zum inhaltlichen Ausfüllen entsprechender Felder) relevant wird.

3.1.2 Relationstypen

Die Relationenwörterbücher verfügen derzeit über folgende im vorliegenden Zusammenhang relevante Relationen:

001 Synonym (alle Richtungen)
002 Teilwort - > Kompositum *3)
003 Kompositum - > Teilwort *3)
004 Derivation (alle Richtungen) *3)
005 Akronym - > Langform
006 Langform - > Akronym
007 Antonym (alle Richtungen)
008 Sprachvarianten (Deutsch/Schweiz)
009 Siehe-Auch-Relation (nur "klass." Thesauri)

010 Orthographische Varianten (alle Richtungen)
011 Stichwort - > Nichtstichwort
012 Nichtstichwort - > Stichwort
013 (Syntakt.) Homographen (alle Richtungen)
014 Quasi-Synonym (alle Richtungen)
015 Unterbegriff - > Oberbegriff (hierarchische Relation)
016 Oberbegriff - > Unterbegriff (hierarchische Relation)
017 Grundform - > Ablaut/Umlaut (f. unregelm. Konj. u. Deklination)
018 Ablaut/Umlaut - > Grundform (f. unregelm. Konj. u. Deklination)
019 Wort (Begriff) - > Klassifizierungscode
020 Klassifizierungscode - > Wort (Begriff)
021 Wort - > Worterläuterung (nur eine Richtung) *2)
022 Einzelwort - > Mehrwort (Teilwortverweis) *3)
023 Mehrwort - > Einzelwort (Elementverweis) *3)
024 Synonym als Stichwort - > Synonym als Nicht-Stichwort
025 Synonym als Nicht-Stichwort - > Synonym als Stichwort *4)
026 Homonym (alle Richtungen; mit Semantik-Differenzierung) *5)
031 Unregelmässiger Nominativ Plural - > Nominativ Singular * 1)
032 Nominativ Singular - > Unregelmässiger Nominativ Plural* 1)
035 Korrekte Schreibweise - > Rechtschreibfehler
036 Rechtschreibfehler - > Korrekte Schreibweise
037 Zusammenbildung - > Element *3)
038 Element - > Zusammenbildung
039 chemische Formel - > Langform
040 Langform - > chemische Formel
041 Mehrwort kanonisch - > Mehrwort Text *7)
042 Mehrwort Text - > Mehrwort kanonisch *7) 044 Pluralbildung
045 Übersetzung *8)
046 Übersetzung (falsch) *8)
047 Nichtstichwort (Singular) - > Stichwort (Plural)
048 Stichwort (Plural) - > Nichtstichwort (Singular)
049 Nichtstichwort (Plural) - > Stichwort (Singular)
050 Stichwort (Singular) - > Nichtstichwort (Plural) *4)
051 Deskriptor (ohne Relationierung) *4)
052 Nichtdeskriptor (ohne Relationierung) *4)

Erläuterungen zu speziellen Anwendungsbereichen:

*1) u.a. verwendet bei der automatischen Trunkierung (MM RETRIEVER)
*2) u.a. bei Worterläuterung / Stilhilfe
*3) u.a. bei Indexierung
*4) u.a. bei MITI SELECTOR
*5) im zweiten Wortlaut stehen eindeutige "Synonyme" (bedeutungsdifferenziert)
*7) für OPAC-Anwendungen (auch für Thesauri)
*8) vorgesehen für den Fall, dass Einträge im "Übersetzungswörterbuch" über ein Relationenwörterbuch dargestellt werden.

3.2 Rahmenbedingungen

Es gelten folgende Rahmenbedingungen:

(1) Die in der Wissensbank abgelegten Deskriptoren werden in der Form abgelegt, wie sie ein Anwender üblicherweise abfragen würde. Nicht immer also ist dies der Nominativ Singular, sondern häufig auch der Nominativ Plural. (Dies kompliziert den Prozess, wobei dies aber seitens des Anwenders nicht bemerkt wird. Vgl. aber unten.)

(2) Ein Nutzer in einer "Fremdsprache" (aus Sicht des "deutschen" Basissystems, z.B. in Englisch oder Französisch) kann das System benutzen, *ohne jemals die deutschsprachigen Deskriptoren zu "sehen*".

Dies setzt voraus, dass auch das Übersetzungsvokabular entsprechend kontrolliert wird: Den Deskriptoren ist jeweils nur eine Übersetzung je Sprache zugeordnet. Das Vokabular ist - analog den Regeln für Thesauri - natürlichsprachig "**basiert**".

(2) Soweit bei der Eingabe - gleich in welcher der zugelassenen Sprachen, also auch in Deutsch - vom Nutzer Wörter verwendet werden, die im Verlauf der Recherche den Bedingungen der Eindeutigkeit nicht genügen würden, müssen sie in einem Klärungsdialog (d.h. bei systemseitig erkannter vorliegender Mehrdeutigkeit) oder aber automatisch (wenn dies "voreingestellt" ist, um den fachspezifischen Nutzer von trivialen Rückfragen zu entlasten) den zugelassenen Deskriptoren zugeordnet werden.

Es kann dabei geschehen, dass die Originaleingabe leicht verändert werden muss (etwa durch Ersetzung des Singulars durch den Plural u.a.m.). Hierauf ist der Nutzer ggf. entsprechend hinzuweisen. Da dies ein automatischer Vorgang ist, wird es ihn (hoffentlich) kaum stören.

(3) Es kann geschehen, dass ein Wort *nicht* als Deskriptor aufgeführt ist, sondern als Suchhilfe wegen einer inhaltlichen Beziehung zu einem Deskriptor (z.B. Hierarchierelation) eingeführt ist. Um zu vermeiden, dass ein Nutzer vergeblich damit arbeitet (d.h. damit sucht), werden Wörter, die "echte" Deskriptoren sind, entsprechend markiert. Hierzu wird intern ein "neues" Stilmerkmal verwendet ("Deskriptor").

Auch **Feldnamen von Datenbanken** werden derzeit in das Relationenwörterbuch übernommen. Hier liegt ein Grund darin, dass ein solcher Name in Datenbanken gelegentlich im Plural verwendet wird. Aus systematischen Gründen (Lemmatisierung, Rechtschreibkorrektur) wird jedoch stets auch die Singularform notiert. Dementsprechend wird hierfür die Relation Nichtdeskriptor (Singular) - > Deskriptor (Plural) genutzt (und umgekehrt), Ziel- und Quelleintrag erhalten das zusätzliche (neue) "Stilmerkmal" Feldname".

Ist ein Wort sowohl als Feldname als auch Element des MITI-Thesaurus zu sehen, werden beide Stilmerkmale vergeben.

4 Verfahrensweise bei Retrieval

4.1 Identifizierung zugelassener Wörter

Zunächst wird versucht, in einem Identifizierungsvorgang das jeweilige (Benutzer-)Wort mit einem Systemwort zu identifizieren.

(a) Prüfung, ob eine Vereindeutigung vorgenommen werden muss. Dazu ist eines der o.a. Relationenwörterbücher (MITIR_D, MITIR_F, MITIR_E, MITIR_S) anzusprechen.

Alle dort verzeichneten "Vorzugsverweise", (techn. Anm.: Dies geschieht so, dass Wörter mit einer Relation, die in der "Punktmarkierung, '.vorz" steht - z.B. Relation 25 oder 12, als solche erkannt werden), werden alternativ wie folgt behandelt:

- Wenn es mehr als einen Vorzugsverweis gibt, wird diese Liste dem Nutzer zur Auswahl (Entscheidung) angeboten. Er muss sich also für eine Variante entscheiden. Diese wird dann an die Stelle seines Ursprungseintrags gesetzt. Die Varianten sind ggf. kommentiert (Zusatzangabe), um die Auswahl zu erleichtern.

  Da man an dieser Stelle bei der Nutzereingabe mit Homonymen bzw. Polysemen rechnen muss, gilt folgende Regelung: Auch dann, wenn nur eine der Varianten für die weitere Suche relevant ist, müssen alle "denkbaren" (sinnvollen) Bedeutungsvarianten hier aufgeführt werden. Gibt es zu einer Variante keinen direkten oder indirekten "Treffer", so wird der Nutzer hier gleich informiert. *Dies ist die einzige Stelle, wo diese Angaben in allen Suchsprachen vorliegen müssen, um dem Nutzer unnötige Suchschritte zu ersparen.* Als Zusatzmerkmal wird die Kennung (Stilvariante) "Nondeskriptor" verwendet (nicht zu verwechseln mit der entsprechenden Relation).

- Gibt es nur *einen* Vorzugsverweis (d.h. die entsprechende "Liste" besteht nur aus einem Eintrag), wird dieser automatisch übernommen mit einem Hinweis an den Nutzer, dass sein Wort durch eine Vorzugbenennung ersetzt wurde.

(b) Im nachfolgenden Schritt wird zunächst unterschieden, ob die Suchanfrage in Deutsch oder einer "Fremdsprache" erfolgt:

- Nur bei Fremdsprachenanalyse: Es erfolgt eine Suche im Übersetzungswörterbuch, unabhängig davon, ob Schritt (a) zu einer Ersetzung geführt hat oder nicht. Wird das Wort im Übersetzungswörterbuch identifiziert, so wird es für die weiteren Schritte zunächst "intern" durch sein deutsches Äquivalent ersetzt.

- Bei allen Sprachen: Es erfolgt ein (erneuter) Abgleich mit dem deutschen Relationenwörterbuch. (Techn. Anmerkung: Hierbei werden nur solche Relationen akzeptiert, die in der Punktmarkierung ".thes" verzeichnet sind.)

  Wird das Wort im Relationenwörterbuch entsprechend identifiziert, so ist es in Ordnung. In diesen Fällen wird die Lemmatisierung nicht aktiviert.

Ist das Wort im Relationenwörterbuch verzeichnet, ohne dass die Relation (über ".thes") zugelassen ist, wird zu "Wort bei der Suche nicht verwendbar" verzweigt (Neueingabe).

(c) Führt (b) zu einem total negativen Ergebnis, wird die Lemmatisierung (Grundformenermittlung, automatisch) aufgerufen. Führt diese zu einem Ergebnis, werden die Prozeduren (a) und (b) nacheinander wiederholt.

Führt die Lemmatisierung nicht zu einem positiven Ergebnis (Wort ungleich Ausgangswort), wird zu dem Ergebnis "Wort unbekannt oder Schreibfehler" verzweigt (Neueingabe).

4.2 Aufbau der Suchzeile(n); Benutzung des Thesaurus

Es war Aufgabe von SOFTEX, die für die Suche relevanten Felder im Suchmenü des MITI SELECTOR aufzubereiten.

Es wurden dabei zwei inhaltliche Felder vorgesehen:

(1) Das Feld "Hauptfachgebiete": Dieses Feld dient dazu, das Themenfeld *grob* zu selektieren. Es bleibt natürlich dem Nutzer überlassen, ob er dieses Feld überhaupt belegen möchte. Wenn er es leer lässt, wird es durch "?" für die SQL-Abfrage belegt.

Bei jedem suchbaren Wort wird angegeben (Stil-Zusatzmarkierung "Hauptfachgebiet"), ob ein solches Wort "suchbar" ist, d.h. bei einer Datenbankbeschreibung in diesem Feld abgelegt wurde. Ist dies nicht der Fall, kann sich der Nutzer mit seinem Suchwort so weit in der Hierarchie "hochhangeln", bis er ein solches Wort identifiziert, das er dann in dieses Suchfeld übernehmen kann.

Es wird ihm anschließend angeboten, die Prozedur zu wiederholen, um weitere Wörter einzutragen ...

(2) Das Feld "Themen / Fachgebiete": Dieses Feld dient dazu, den Themenbereich *fein* zu selektieren. Die Begriffe des Hauptfachgebiets können hier ebenfalls verwendet werden, haben aber eine andere Bedeutung. Handelt beispielsweise eine Datenbank über die Geschichte der Medizin, so reicht es, hier, "Geschichte" anzugeben, wenn im Hauptfachgebiet "Medizin" angegeben wurde. Bleibt das Hauptgebiet offen, so werden alle Datenbanken angegeben, die irgendwie auch "Geschichte" beinhalten.

4.2.1 Allgemeine Regel

Da im Thesaurus Relationen ggf. nur über die Singularschreibform angegeben sind, wird zuvor - wenn ein Pluraleintrag vorliegt - "intern" das Suchwort durch die Singularvariante ersetzt (Relation 31). Wird zu dem relationierten Wort ebenfalls ein Plural angeboten (Relation 47), so wird diese optisch angezeigt bzw. bei der Übersetzung zugrundegelegt.

Ist die Nutzersprache *nicht Deutsch*, so ist durch einen Eintrag der Übersetzung der Pluralform im Übersetzungswörterbuch sicherzustellen, dass die richtige Schreibform angeboten wird. (Nach den ersten Erfahrungen bzgl. des Pflegeaufwands wird man in Zukunft zwei Möglichkeiten verfolgen:

- automatische Abbildung eines Pluraleintrags auf die Singularform (anstelle der Angabe der Übersetzungen der Pluralformen selbst)
- Umsetzung auf die Singularform unter alleiniger Verwendung der Singularschreibweise bei der Übersetzungs-Zuordnung. Dies bedeutet dann zwar eine etwas "geringere" Benutzerfreundlichkeit, reduziert andererseits den Pflegeaufwand erheblich.

4.2.2 Relationstypen

Die Verwendung der Synonymie- und Polysemie-Relationen zur Disambiguierung usf. wurde schon geschildert.

Folgende Relationstypen zur Auswahl eines thematisch relevanten Begriffs (repräsentiert durch seine Benennung) werden "aktiv" unterstützt:

(1) Oberbegriff - > Unterbegriff
(2) Unterbegriff-Oberbegriff (invers zu (1))
(3) Siehe-auch-Relation
(4) virtuelle Relation: "Nebenbegriff'

Bei den Relationstypen Oberbegriff / Unterbegriff kann zudem die "Relationstiefe" (Kette) benutzerseitig variiert werden.

Die virtuelle Relation "Nebenbegriff" resultiert aus Fällen, bei denen zwei oder mehr "Begriffe" den gleichen Ober- oder aber Unterbegriff aufweisen.

Im konkreten Fall (MITI-Selektor) können in einem Such- bzw. Präsentationsvorgang zwei scrollbare "Fenster" (und damit max. 2 Relationen) zu einem Suchwort aktiviert werden. Dies ist aber eine eher technische und bildschirmbezogene Einschränkung.

4.2.3 Aktivierung des Thesaurus

Der Thesaurus kann im MITI SELECTOR über verschiedene Wege aktiviert werden:

(1) Auch dann, wenn ein als Suchwort zugelassener Deskriptor getroffen wurde. Man kann über diesen "einsteigen" und durch Aktivierung des Thesaurus zusätzlich "browsen", um weitere Wörter in die Suchfrage einzubeziehen.

(2) Es gibt bei einem "zugelassenen" Deskriptor das Problem, dass man mit ihm als Suchwort ggf. nichts in der MITI SELECTOR Datenbank findet. Daher werden solche Suchwörter, die Deskriptoren darstellen, über das (Stil-)Merkmal "Deskriptor" gekennzeichnet. (Die Vergabe dieses Merkmals - das ja abhängig ist von den realen Deskriptoren der Daten-

bank - erfolgt derzeit noch "halbautomatisch", lässt sich aber weiter automatisieren). Versucht ein Benutzer, ein Wort in die Suchfrage zu übernehmen, das nicht dieses Merkmal enthält, so wird der Versuch abgelehnt mit dem Hinweis, zu "browsen", d.h. entweder nach einem Oberbegriff oder einem Unterbegriff oder aber auch einem ähnlichen Begriff zu forschen, zu dem ein Treffer möglich ist.

Beim Verzweigen über die o.a. Relationen (einer Art "Blättern") kann man natürlich wieder auf "Nichtdeskriptoren" stoßen, d.h. solche ohne die Kennung "Deskriptor". Die fehlende Kennung kann erneut als Indiz verwertet werden (Hinweis an den Nutzer), damit er nicht vergeblich versucht, das Wort in die Suchfrage zu übernehmen (das Ankreuzfeld wird beispielsweise am Bildschirm "verwaschen" angezeigt, ähnlich der SPELL-Markierung). Ein solches Wort steht also - um es erneut zu sagen - allein wegen der Hierarchie-Beziehungen im Thesaurus.

Es ist nicht zu vergessen, dass bei nicht-deutscher Eingabe zwischenzeitlich (ohne dass dies vom Nutzer bemerkt wird) immer eine Übersetzung erfolgen muss. Über die Pflege des Wörterbuchs ist sicherzustellen, dass hierbei keine Mehrdeutigkeit entsteht.

4.3 Übernahme in die Suchanfrage

Es gibt im MITI SELECTOR grundsätzlich drei Möglichkeiten, ein weiteres Wort in die Suchanfrage zu übernehmen:

- mit UND-Verknüpfung (Schnittmenge)
- mit ODER-Verknüpfung (Vereinigungsmenge)
- mit UND-NICHT-Verknüpfung (Ergänzungsmenge).

Je nach Datenbank-Beschreibung kann es folgende Situation geben:

(1) Sowohl der Oberbegriff als auch der Unterbegriff sind bei einer Datenbankbeschreibung im Feld "Themenbereiche" verzeichnet. Handelt etwa eine Datenbank (im folgenden DB-A) von Medizin (allgemein) und dabei besonders von Neurologie, so sind beide Deskriptoren vergeben.

(2) Nur der Unterbegriff ist beispielsweise in einer (speziellen) Datenbank verzeichnet, etwa einer Datenbank, die nur den Bereich Neurologie umfasst (im folgenden DB-N). Dann würde man mit einer Anfrage "Medizin" UND "Neurologie" im "Themenbereichs-Feld" die Datenbank DB-N nicht finden. Hätte man "Medizin" dagegen im Feld "Hauptfachgebiete" eingegeben und "Neurologie" im Feld "Themenbereiche", so wäre es auch in diesem Falle zu einem Treffer gekommen.

Wenn sich ein Nutzer also für eine Datenbank interessiert, die allgemein über Medizin handelt und speziell auch Neurologie einbezieht, so wird er bei der Anfrage "Medizin" UND "Neurologie" im Feld "Themenbereiche" die Datenbank DB-A finden, bei "Medizin" nur die Datenbank DB-A. Er kann sich natürlich auch damit behelfen, dass er alle Unterbegriffe zu "Medizin" im Thesaurus "anklickt" und dann die Anfrage "Medizin" ODER "Neurologie" ODER ... stellt. Will er alle Datenbanken, die irgendwie über Medizin handeln (ob speziell oder allgemein oder beides), so füllt er nur das Feld "Hauptfachgebiete" mit "Medizin" aus.

Die vom Nutzer gewünschte Kombination muss über einen entsprechenden Dialog abgefragt werden. Die komplette Anfrage-Formel wird zunächst "extern" aufgebaut, ehe sie an das SQL-Feld abgegeben wird.

4.5 Thesaurus-Pflege und Qualitätskontrolle

Zum gegenwärtigen Zeitpunkt werden folgende Funktionen systemseitig unterstützt:

(1) Abgleich der Datenbank-Benennungen gegen das bestehende Inventar (die MITI-Wörterbücher).
(2) Falls (1) keinen Treffer aufweist: Bereitstellung von Wörtern und möglichen Übersetzungen zur Vokabularerweiterung aus den allgemeinen elektronischen Lexika.

Soweit eine (intellektuelle) Erweiterung der MITI-Wörterbücher notwendig wurde:

(3) Konsistenzüberprüfung der Übersetzungswörterbücher
(4) Konsistenzüberprüfung der Relationenwörterbücher. In diesem Falle werden sog. TOP-Terms ausgenutzt (bei Fachgebieten ist dies das Term "ALLGEMEIN". Erreicht eine hierarchisch verknüpfte Benennung nicht eines der TOP-Terms, so wird auf "Fehler" verzweigt.

Nicht sichergestellt werden kann (natürlicherweise) die Vollständigkeit der semantischen Disambiguierung. Diese stellt - neben dem Nachtragen von fehlenden Übersetzungen - überhaupt den Hauptaufwand bei der intellektuellen Bearbeitung dar.

Einige eher technische, wenn auch letztlich nicht unbedeutende Fragen sind bei der Anwendung im MITI SELECTOR "nebenher" aufgetreten:

- Relationale Datenbanken können Schwierigkeiten mit dem Zeichensatz haben. So musste an der Schnittstelle zu INGRES beispielsweise eine komplizierte "Umlautdarstellung" erfolgen.

- Es gibt (fast natürlicherweise) Kollisionen in Fällen, wo bei unterschiedlichen Sprachen Wörter unterschiedlich differenziert werden. Das englische Wort "libraries" bedeutet beispielsweise einerseits "Plural von library = Bibliothek", andererseits auch "Bibliothekswesen". Eine automatische Grundformermittlung reicht also hier nicht aus, vielmehr muss in solchen Fällen (bei Englisch als Ausgangssprache) schon die Pluralform disambiguiert werden, und nur im Falle der Bedeutung "Bibliothek" kann man dann zum Singular verzweigen usf.

- Es ist zu überlegen, ob man, statt - wie bisher im MITI SELECTOR geschehen - systemseitig "eindeutige" natürlichsprachige Synonyme zu wählen, nicht besser einen künstlichen Index (z.B. eine nachgestellte Ziffer) einführt. Im Zweifelsfall kann sich der Anwender über eine Hilfefunktion (Worterklärung) diesen Index erklären lassen. Beispiel (von oben): libraries - > library_1(= Bibliothekswesen) und library_2 (= Bibliothek) usf.

Nach unserer Auffassung "rechnet" sich diese - entsprechend noch leicht angepasste - Verfahrensweise der Thesauruserstellung bereits bei einem voll ausgebauten (d.h. auf die in der Welt "verfügbaren" Datenbanken angewendeten) MITI SELECTOR:

- Es ist flexibel (d.h. an die Themen anpassbar) und bleibt dennoch konsistent in der Beschreibung

- Es eröffnet ohne große Probleme Anwendungen in den unterschiedlichsten natürlichen Sprachen.

- Es ist benutzerfreundlich, da es den Zugang und den Dialog in der Nutzersprache erlaubt und stets Erklärungen parat hat, wenn es um den Gebrauch von Benennungen geht.

5. Zusammenfassung und Ausblick

Die ersten Erfahrungen mit dem Prototypen MITI SELECTOR sind sehr ermutigend. Vor allem die Vorgehensweise, einerseits dem Benutzer über "seine" Benennungen einen Zugang zu eröffnen, ihn andererseits möglichst frühzeitig über mögliche Treffer zu informieren (praktisch eine einfache Kopplung zwischen Thesaurus und Datenbank - die man natürlich technisch auch anders lösen könnte), hat sich bei den Vorführungen und Tests als wichtig erwiesen.

Natürlich muss ein solches Konzept weiter erprobt und ggf. weiter verallgemeinert werden. Vorschläge zu einer allgemeinen Standardisierung der Terminologie (auch bzgl. einer festgelegten Bedeutungsdifferenzierung) sowie zur weiteren Typisierung / Aspektierung von Thesaurusrelationen wurden in meinem Beitrag auf dem Deutschen Dokumentartag in Jena 1993 gemacht, so dass hier darauf verwiesen werden kann.

> **Nach meiner Erkenntnis - und dazu hat die prototypische Entwicklung des MITI-Systems wesentlich beigetragen - gibt es in Zukunft zu derartigen Verfahrensweisen keine grundsätzliche Alternative.**

Angesichts der inzwischen - technisch gesprochen - in "Massen" auftretenden eher trivialen Systemen wird sich die vorgestellte Differenzierung - ganz gleich, welche technischen Werkzeuge auf welchen Plattformen auch immer genutzt werden - als Wettbewerbsvorteil herausstellen.

Harald H. Zimmermann, Universität des Saarlandes

Weilburg, den 26. Oktober 1993 (T31WB1)