Mapping Knowledge Organization Systems

Philipp Mayr, Anne-Kathrin Walter
GESIS / Informationszentrum Sozialwissenschaften, Bonn

Abstract

Die Vernetzung der Informationssysteme und Datenbanken aus dem wissenschaftlichen Fachinformationsbereich lässt bislang den Aspekt der Kompatibilität und Konkordanz zwischen kontrollierten Vokabularen (semantische Heterogenität) weitgehend unberücksichtigt. Gerade aber für den inhaltlichen Zugang sachlich heterogen erschlossener Bestände spielen für den Nutzer die semantischen Querverbindungen (Mappings /Crosskonkordanzen) zwischen den zugrunde liegenden Knowledge Organization Systems (KOS) der Datenbanken eine entscheidende Rolle. Der Beitrag stellt Einsatzmöglichkeiten und Beispiele von Crosskonkordanzen (CK) im Projekt „Kompetenznetzwerk Modellbildung und Heterogenitätsbehandlung" (KoMoHe)[1] sowie das Netz der bis dato entstandenen Terminologie-Überstiege vor. Die am IZ entstandenen CK sollen künftig über einen Terminologie-Service als Web Service genutzt werden, dieser wird im Beitrag exemplarisch vorgestellt.

1. Einleitung

Eines der klassischen Probleme im Information Retrieval ist die Vagheit, die zwischen einer natürlichsprachigen Anfrage eines Benutzers und den „künstlichen" Termen, die zur Dokumentbeschreibung verwendet werden, besteht. Konkret am Beispiel der Dokumentation kann das bedeuten, dass ein Benutzer für eine beliebige fachliche Fragestellung andere natürlichsprachige Terme verwendet, als derjenige, der die Inhaltserschließung für entsprechende Dokumente vornimmt (vgl. language problem bei Petras 2006). Die Vagheit zwischen Anfrage- und Dokumentebene wird bei Krause V1 genannt (siehe Abbildung 1) und kann beispielsweise durch Verfahren zur Termerweiterung behandelt werden. Dies kann ‚manuell' durch den Nutzer (z.B. durch Verwendung eines Thesaurus mit Synonym- oder Ober-/Unterbegriffsbeziehungen) oder in ähnlicher Weise auch durch das Informationssystem (teil-)automatisch durchgeführt werden (z.B. durch Vorschläge für zusätzliche und alternative Suchbegriffe).

„Jedem Bibliothekar und jedem, der sich mit Information Retrieval befasst, war schon immer klar, dass zwischen den semantischen Termen, die in der Datenbank ein Dokument charakterisieren, und dem Term, den der Benutzer anwendet, nicht immer eine 1:1-Relation besteht." (Krause 2003: S. 10).

Ein weiterer Lösungsvorschlag für die Vagheitsbehandlung der ersten Ebene V1 in der Form von Search Term Recommender Systemen[2] findet sich bei Petras (2006). Handelt es

[1] Der Beitrag ist im Projekt "Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung" entstanden. Dieses Projekt wird vom BMBF unter der Kennziffer 523-40001-01C5953 gefördert. Siehe <http://www.gesis.org/Forschung/Informationstechnologie/KoMoHe.htm>.

[2] "This dissertation describes a mechanism that will provide a translation aid between specialized languages and the documentary language by suggesting appropriate search terms for a searcher's query in relation to the searcher's domain of discourse." (Petras 2006).

sich bei den zu durchsuchenden Dokumentbeständen um homogen erschlossene Datenbanken, sind die Verfahren zur Behandlung von V1 vermutlich ausreichend. Bei inhaltlich heterogen erschlossenen Dokumentbeständen (vgl. Situation in vascoda in Mayr, 2006a) sind andere Verfahren anzuwenden, die in diesem Beitrag beschrieben werden. Durch den Einsatz von unterschiedlichen Thesauri bzw. anderen kontrollierten Vokabularen in Digitalen Bibliotheken entsteht Vagheit/Heterogenität bereits auf der inhaltlichen Beschreibungsebene der Dokumente. Beispielsweise wird ein Dokument bei der Sacherschließung mit Thesaurus A mit dem Deskriptor „Biologieunterricht" beschrieben, während ein anderes kontrolliertes Vokabular B vorschreibt in diesem Fall den Deskriptor „naturwissenschaftlicher Unterricht" zu verwenden (siehe dazu Tabelle 2).

Die Behandlung dieser Vagheit (V2, V3) kann bilateral zwischen den einzelnen Dokumentbeständen geschehen, wie durch Abbildung 1 verdeutlicht wird.

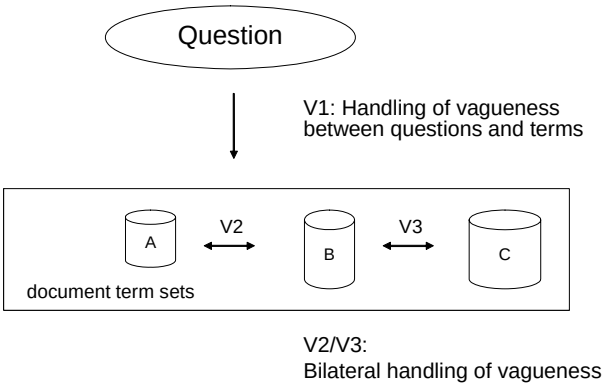

Abb. 1: Mehrstufige[3] Vagheitsbehandlung im Information Retrieval (Zwei-Schritt-Modell aus Hellweg et al. 2001).

Auf semantischer Ebene wird Heterogenitätsbehandlung (Vagheit V2, V3) modelliert, indem Transferkomponenten entwickelt werden, die das kontrollierte Wortmaterial (Deskriptoren) einzelner Vokabulare durch Termtransformationen aufeinander beziehen. Im oberen Beispiel bedeutet das, dass zur Behandlung von V2 alle Terme aus Vokabular A auf Terme in Vokabular B abgebildet werden und umgekehrt alle Terme aus Vokabular B auf A (bilaterale Vagheitsbehandlung).

Zur Erstellung von Termtransformationen sind verschiedene Ansätze erprobt worden (vgl. Hellweg et al. 2001):

- Intellektuell: Für Terme eines Ausgangsvokabulars werden intellektuell Relationen zu passenden Termen eines Zielvokabulars gebildet. Es sind unterschiedliche Relationstypen möglich. Diese Art der Termtransformation wird Crosskonkordanz genannt.
- Statistisch: Bei diesen Verfahren werden semantische Relationen mit Hilfe von statistischen Methoden (Kookkurrenz-Analysen) automatisch erzeugt (vgl. Hellweg et al. 2001; Strötgen 2004; Marx 2005; Zhang 2006).

[3] Mehrstufig bedeutet in diesem Zusammenhang, dass auch eine Kombination von unterschiedlichen Verfahren der Vagheitsbehandlung möglich ist.

- Deduktiv: Bei deduktiven Verfahren wird Textmaterial analysiert und aus den sich ergebenden Zusammenhängen werden mit Hilfe von logischen Schlussfolgerungen Relationen zwischen Termen abgeleitet.

Schwerpunkt der im Projekt KoMoHe erstellten Termtransformationen liegt auf den intellektuell erstellten Crosskonkordanzen, daher wird darauf im Folgenden näher eingegangen.

2. Crosskonkordanzen

Der Trend in der aktuellen Fachinformationslandschaft geht hin zu einer Bündelung der Informationsangebote, sowohl innerhalb eines Fachs (Beispiel Sozialwissenschaften – Fachportal sowiport) als auch interdisziplinär (Beispiel Pädagogik, Psychologie, Sozialwissenschaften – Informationsdienst infoconnex[4]). Ziel ist es, die Recherche für einen Nutzer mit einem bestimmten Informationsbedürfnis zu erleichtern und ihn beim Auffinden der für ihn relevanten Dokumente zu unterstützen (vgl. Mayr et al. 2005; Mayr 2006b). Neben der Integration auf technischer und struktureller Ebene (siehe auch Strötgen 2004), muss zusätzlich eine Integration auf semantischer Ebene vorgenommen werden (siehe dazu Krause 2003). Wie oben bereits skizziert tritt semantische Heterogenität (bei Krause auch „unvermeidlich verbleibende Heterogenität") auf, wenn Informationsangebote unterschiedliche Inhaltserschließungssysteme bzw. Knowledge Organization Systems (KOS[5]) verwenden um die Dokumente fachlich zugänglich zu machen. Im Gegensatz zur strukturellen Heterogenität (z.B. unterschiedliche Metadatenschemata), die vgl. einfach homogenisiert werden kann, verbleiben die unterschiedlichen kontrollierten Vokabulare (KOS) heterogen. Ein Nutzer der seine Anfrage in dem ihm bekannten Vokabular formuliert, findet unter Umständen in den anderen Datenbanken keine Dokumente, da die gesuchten Konzepte dort anders benannt sind oder nicht existieren (vgl. dazu Abbildung 2). Das Konzept der Crosskonkordanzen[6] setzt an dieser Stelle an und versucht die Benutzer zu entlasten, die den Suchraum auf weitere fachlich relevante Datenbanken ausweiten wollen, aber mit ihrem gewohnten Vokabular recherchieren wollen.

2.1 Erstellung von Crosskonkordanzen

Crosskonkordanzen sind gerichtete, relevanzbewertete Relationen zwischen Termen zweier Thesauri, Klassifikationen oder auch anderer kontrollierter Vokabulare. Crosskonkordanzen ermöglichen eine Übersetzung bzw. Transformation von Termen eines Erschließungssystems in Terme eines anderen kontrollierten Vokabulars. Die Erstellung der Relationen erfolgt intellektuell, d.h. ein Terminologie-Experte vergleicht die verschiedenen Begriffssysteme und setzt die Terme zueinander in Beziehung. Die Erstellung der Term-Term Relationen erfolgt in Tabellen. In der linken Spalte sind die

---

[4] <http://www.infoconnex.de>.

[5] "Knowledge Organization Systems (KOS) is a general term referring to the tools that present the organized interpretation of knowledge structures. Three general categories: Term lists, classifications, relational vocabularies." (Zeng/Chan 2004) Siehe dazu auch die Definition in Hodge (2000).

[6] Konkordanzen sind bislang in mehreren Projekten am IZ entwickelt worden, u.a. im Projekt CARMEN AP12 (siehe CARMEN 2002) oder auch für den interdisziplinären Informationsdienst infoconnex (Walter et al. 2006). Eine aktuelle Übersicht der Verfahren und Projekte zur Behandlung von Heterogenität findet sich in Zeng/Chan (2004).

Ausgangsterme eingetragen, in der zweiten Spalte folgt der Typ der Relation, eine Relevanzbewertung und in der rechten Spalte die Entsprechungen im Zielvokabular. Erstellt werden 1:1 und 1:n Relationen, d.h. ein Ausgangsterm kann mit einem oder mehreren Zielkonzepten verbunden werden. Zur Spezifikation der Beziehung zwischen den Termen können vier verschiedene Relationstypen verwendet werden (siehe dazu auch Beispiele in Tabelle 1 und 2, sowie im Anhang):

- Äquivalenzrelation („="): für Terme, die das gleiche Konzept bezeichnen. Die Äquivalenzrelation wird zwischen identischen, synonymen und/oder quasi-synonymen Termen gesetzt.
- Oberbegriffsrelation („<"): für Terme, die in einer Hierarchiebeziehung stehen (Teil-Ganzes, Abstraktion). Eine Oberbegriffsrelation wird ausgehend von einem engeren zu einem weiteren Term gesetzt.
- Unterbegriffsrelation („>"): wie Oberbegriffsrelation. Eine Unterbegriffsrelation wird ausgehend von einem weiteren zu einem engeren Begriff gesetzt.
- Ähnlichkeitsrelation („^"): für Terme die ähnliche oder verwandte Konzepte bezeichnen. Die Ähnlichkeitsrelation wird für verwandte Begriffe gesetzt.

Jede der Relationen wird zusätzlich nach Relevanz bewertet und dadurch eine Aussage über die zu erwartende Relevanz der Treffermenge gemacht (Abstufung: hoch, mittel, gering). Lässt sich keine Entsprechung im Zielvokabular identifizieren, wird die Nullrelation („0") gesetzt. Tabelle 1 zeigt beispielhaft einen Ausschnitt aus einer Konkordanz zwischen Thesaurus Sozialwissenschaften und Standard Thesaurus Wirtschaft. Weitere Crosskonkordanz-Beispiele finden sich in Tabelle 2, in Walter et al. (2006) sowie im Anhang dieses Beitrags.

Thesaurus Sozialwissenschaften	**Relation**	**Relevanz**	**Standard Thesaurus Wirtschaft**
Abgaben	=	h	Gebühr
Deutsche Bundesbank	=+	h	Zentralbank + Deutschland
Abitur	<	m	Bildungsabschluss
Entschuldung	^	h	Schuldenerlass
Katastrophe	>	g	Naturkatastrophe
Pädagogische Faktoren	0		

Tabelle 1: Beispiel für Crosskonkordanz-Relationen.

Für die Crosskonkordanz-Erstellung können grundsätzlich unterschiedliche Sichtweisen eingenommen werden, d.h. es gibt unterschiedliche Faktoren, die Einfluss auf die Art und Weise der Relationierung zwischen Termen bzw. Konzepten nehmen können.

- Ziel ist, eine möglichst exakte Abbildung von Konzept A, das durch D(A) benannt wird, in Thesaurus B zu erhalten.
- Ziel ist es nicht, Indexierungsfehler über die Konkordanz auszugleichen. Es werden aber – wo dies möglich ist – Indexierungsregeln der Vokabulare berücksichtigt. Im Idealfall kann für jede Relation in der Datenbank recherchiert werden und so lange optimiert werden, bis voraussichtlicher Recall und Precision befriedigend sind. Da dies aufgrund des Aufwands nicht für alle Terme möglich ist, wird meist auf der Wortebene verbunden und nur in Problemfällen in der Datenbank recherchiert.
- Ziel ist es ebenfalls nicht, den Scope bzw. das Fachgebiet einer Datenbank über das kontrollierte Vokabular anzugleichen. Um die durch den Zieldeskriptor gefundenen Dokumente auf das Fachgebiet des Ausgangsthesaurus einzugrenzen, wurde in

Erwägung gezogen, den Zieldeskriptor mit einem weiteren Deskriptor zu kombinieren, der das Fachgebiet des Ausgangsthesaurus beschreibt. Beispielsweise würde man in der Konkordanz vom Sport zu den Sozialwissenschaften eine Kombination mit dem Deskriptor „Sport" aus dem Thesaurus Sozialwissenschaften bilden, damit die gefundenen Dokumente sportrelevant sind: D(A) → D(B) + Sport. Diese Vorgehensweise wurde nicht weiter verfolgt. Gleiches gilt für andere Fachgebiete.

Weitere Regeln bei der Crosskonkordanz-Erstellung finden sich in der Präsentation[7] für die Wiener ISKO-Tagung 2006.

Thesaurus Sozialwissenschaften	Relation	Zielterm(e)	Zielvokabular
Biologieunterricht	<	Unterricht	DZI
Biologieunterricht	<	Unterricht	Standard Thesaurus Wirtschaft
Biologieunterricht	=	Biologieunterricht	Schlagwortnormdatei
Biologieunterricht	<+	Biology + Teaching	CSA
Biologieunterricht	=+	Naturwissenschaftlicher Unterricht + Biologie	Psyndex Terms
Biologieunterricht	=+	Fachunterricht/Unterrichtsfach + Biologie	IBLK
Biologieunterricht	=+o	Biologie + Schulfach	BISp
Biologieunterricht	<+o	Biologie + Unterrichtsstunde	BISp
Biologieunterricht	<+	Biologie + Schule	DZA/Gerolit
Biologieunterricht	^+	Biologie + Unterricht	FES

Tabelle 2: Beispiel für Crosskonkordanz-Relationen ausgehend von dem Deskriptor „Biologieunterricht" des Thesaurus Sozialwissenschaften. Die Kürzel der rechten Spalte sind über die Tabelle 3 aufzulösen.

2.2 Übersicht der Crosskonkordanzen

Seit Ende 2004 werden innerhalb des Projekts „Kompetenznetzwerk Modellbildung und Heterogenitätsbehandlung" eine Vielzahl an Crosskonkordanzen zwischen unterschiedlichen Fächern bearbeitet. Mittlerweile sind insgesamt 18 kontrollierte Vokabulare aus acht[8] Fachgebieten durch Crosskonkordanzen verbunden worden. Abbildung 3 bietet eine Übersicht der verbundenen Vokabulare und das somit entstehende semantische Netz der Crosskonkordanzen.

Es existieren 21 Crosskonkordanzen (bilaterale Konkordanzen) sowie drei unilaterale Konkordanzen. Vier der Crosskonkordanzen und zwei der unilateralen Konkordanzen wurden bereits in den Projekten CARMEN/infoconnex erstellt (Sozialwissenschaften, Psychologie, Bildung, SWD), alle übrigen sind im Projekt KoMoHe entstanden. Insgesamt existieren momentan ca. 200.000 Relationen zwischen 80.000 Konzepten (Stand Dezember 2006). Tabelle 3 zeigt eine Aufstellung der Vokabulare die durch Crosskonkordanzen verbunden sind.

[7] <http://www.ib.hu-berlin.de/~mayr/arbeiten/mayr-walter_isko06.pdf> (Folien 10ff.).

[8] Die Schlagwortnormdatei als einziges universelles Vokabular im Projekt spielt aufgrund ihrer Größe und fachlichen Zuordenbarkeit eine Sonderrolle und wird daher gesondert aufgeführt.

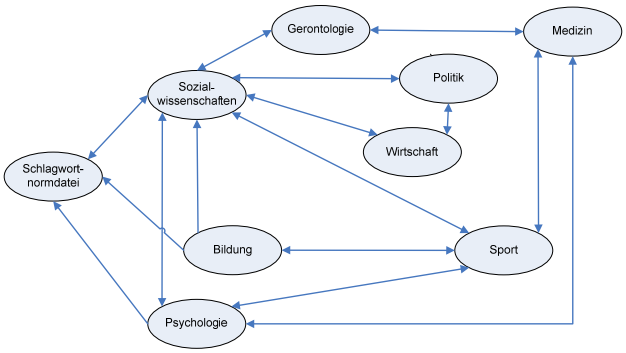

Abb. 3: Vernetzung der Fachgebiete in vascoda durch Crosskonkordanzen.

	Kürzel	Name des Vokabulars	Größe d. Vok. (ca.)	Datenbank	Anbieter
1	Bildung	Thesaurus Bildung	55,000	FIS Bildung	DIPF Frankfurt/M
2	BISp	Deskriptoren des Bundesinstituts für Sportwissenschaft	7,500	SPOLIT	BISp Bonn
3	CSA-ASSIA	CSA Thesaurus Applied Social Sciences Index and Abstracts	17,000	ASSIA	CSA, IZ
4	CSA-PAIS	CSA Thesaurus PAIS International Subject Headings	7,000	PAIS	CSA, IZ
5	CSA-PEI	CSA Thesaurus Physical Education Index	1,800	PEI	CSA, IZ
6	CSA-SA	Thesaurus of Sociological Indexing Terms	4,000	SA	CSA, IZ
7	CSA-WPSA	CSA Thesaurus of Political Science Indexing Terms	3,150	WPSA	CSA, IZ
8	DZI	Thesaurus des Deutschen Instituts für soziale Fragen	2,000	SoLit	DZI, IZ
9	ELSST	European Language Social Science Thesaurus	3,200	Madiera	
10	FES	Deskriptoren der Friedrich-Ebert Stiftung	4,000	Digitale Bibliothek FES	Friedrich-Ebert-Stiftung Bonn, IZ
11	GEROLIT	Thesaurus des Deutschen Zentrums für Altersfragen	2,000	GEROLIT	DZA Berlin
12	IBLK	Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)	9,000	World Affairs Online (WAO)	SWP Berlin
13	MeSH	Medical Subject Headings	22,000	ZB Med Katalog, Medline	ZB Med Köln, NLM
14	Psy	Psyndex Terms	5,300	Psyndex	ZPID Trier
15	STW	Standard Thesaurus Wirtschaft	5,600	Econis	ZBW Kiel
16	SWD	Schlagwortnormdatei	400,000[9]	div. OPACs	Deutsche National Bibliothek
17	TheSoz	Thesaurus Sozialwissenschaften	7,500	SOLIS	IZ
18	TWSE	Thesaurus für wirtschaftliche und soziale Entwicklung	2,800	InWEnt	InWEnt – Internationale Weiterbildung und Entwicklung Bonn

Tabelle 3: Überblick über die verbunden Vokabulare.

[9] Bislang wurde nur der sozialwissenschaftliche Ausschnitt der SWD-Terme (ca. 8.000) in die Datenbank importiert.

2.3 Zusammenfassung

Zusammenfassend lässt sich sagen, dass die weite Definition von Knowledge Organization Systems (KOS) bei Hodge (2000) positive und negative Effekte hat.

> "The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies. Because knowledge organization systems are mechanisms for organizing information, they are at the heart of every library, museum, and archive." (Hodge 2000).

Zum einen ist es im Sinne einer größeren Verbreitung und Bekanntheit sicher nützlich, alle Typen und Formen von kontrollierten Vokabularen unter einem eingängigen Konzept (nämlich KOS) zu subsumieren, zum anderen entstehen neue Probleme, die insbesondere beim Mapping, also bei der Erstellung von Crosskonkordanzen auffallen. Das Hauptproblem beim Mapping von KOS besteht darin, das die unterschiedliche Tiefe der Strukturiertheit/Organisiertheit der KOS zu Vagheitsproblemen führen, für die es bestenfalls Heuristiken zur Abminderung semantischer Verluste gibt. Um in der Terminologie von Hodge zu bleiben; es ist pragmatisch sehr vage, einfache Termlisten (term lists) mit Klassifikationen (classifications) oder ausdefinierten Thesauri (relationship lists) zu verbinden. Ein gutes semantisches Verständnis der Konzepte unterschiedlicher KOS kann sich im Prinzip nur auf der Basis ausdefinierter Deskriptoren ergeben, dies kann per definitionem von KOS nur näherungsweise erreicht werden.

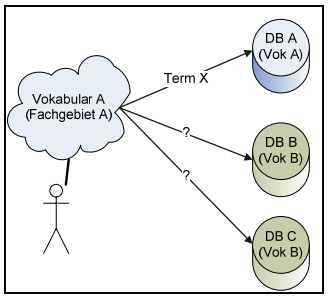

Abb. 2: Heterogenität kontrollierter Vokabulare.

3. Heterogenitätsservice

Die erstellten Crosskonkordanzen werden über einen Terminologie-Dienst, den sogenannten Heterogenitätsservice (HTService) verfügbar gemacht. In diesem Abschnitt wird anhand eines Einsatzszenarios dessen Funktionalität vorgestellt und die Datenbasis beschrieben, auf die er zugreift und die gleichzeitig das Speicherformat der Crosskonkordanzen darstellt.

3.1 Funktionalität

Es gibt mehrere Einsatzmöglichkeiten für den Heterogenitätsservice. Basisfunktionalität ist der Dienst des Terminologie-Mappings (Term-Umschlüsselung) für Fachportale. Weiterhin ist der Einsatz des Service als Rechercheunterstützung für den Nutzer denkbar. Das durch die Crosskonkordanzen entstandene semantische Netz zwischen Suchtermen kann bei der Formulierung von Suchanfragen hilfreich sein. Ferner könnte der Service in Zukunft Funktionen zum Update der Konkordanzen umfassen. Der Schwerpunkt der ersten Version des Service liegt bei der Funktionalität des Terminologie-Mappings. Anhand des im Folgenden beschriebenen Szenarios (siehe auch Abbildung 4) werden Entscheidungen zur technischen Realisierung, zur Schnittstelle und zur Architektur des Service erläutert.

> Ein Nutzer hat ein Informationsbedürfnis und formuliert seine Anfrage in dem ihm vertrauten Vokabular A (Ausgangsvokabular), das Dokumente der Datenbank A inhaltlich erschließt. Die Datenbanken B und C sind mit anderen Vokabularen erschlossen. Ziel des Fachportals, das die drei Datenbanken zur integrierten Recherche anbietet, ist es, dem Nutzer alle relevanten Dokumente bezogen auf sein Informationsbedürfnis zu liefern. Bevor es die Anfrage an die Datenbanken weitergibt, wird der Heterogenitätsservice nach Termtransformationen in die Vokabulare (Zielvokabulare) der Datenbanken B und C gefragt. Falls andere Terme für die Datenbanken vorhanden sind, wird die Anfrage pro Datenbank modifiziert und anschließend die Abfrage gestartet.

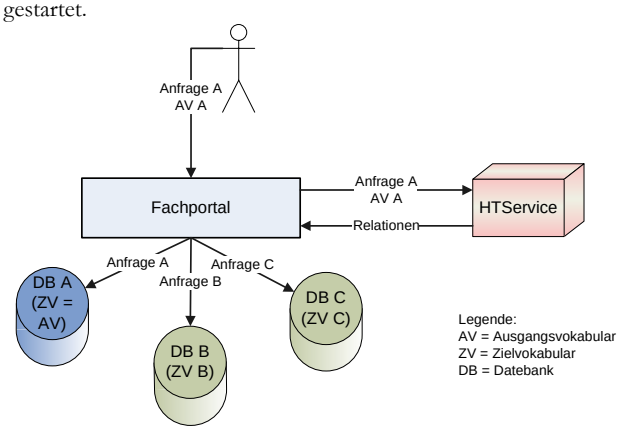

Abb. 4: Einsatzszenario des Heterogenitätsservice.

3.2 Technische Realisierung

Grundlage für den Heterogenitätsservice ist die Web Service-Technologie. Das Kommunikationsprotokoll SOAP[10], als deren Basis, ermöglicht es, dass Fachportal und HTService unabhängig vom unterliegenden Übertragungsprotokoll und lokal verwendeten Technologien kommunizieren. Da SOAP ein XML-basiertes Protokoll ist, bleibt die Kommunikation menschenlesbar, ist aber auch für Maschinen prozessierbar. Zudem ist SOAP ein offener Standard, der ohne Einschränkungen zugänglich ist. Die Realisierung als Web Service bietet einen weiteren Vorteil für die automatisierte Kommunikation zwischen Anwendungen: Es existiert ein standardisiertes Format zur Beschreibung der Schnittstelle, d.h. es ist spezifiziert, welche Funktionen der Service anbietet, wie die

[10] <http://www.w3.org/TR/soap12-part1/>.

Funktionen aufgerufen werden und wie die Antwort aufgebaut ist. Auf diese Weise kann sehr einfach eine Anfrage an den Dienst erfolgen.

3.3 Inhaltliche Realisierung

Die Anfrage eines Fachportals an den Heterogenitätsservice kann je nach Suchanfrage des Nutzers unterschiedlich strukturiert sein. Immer enthalten ist natürlich der Ausgangsterm, der transformiert werden soll. Abhängig von der Suche, die der Nutzer durchführt, können weitere Einschränkungen angegeben sein.

- Relationstyp: Wie im oberen Abschnitt beschrieben, gibt es unterschiedliche Relationstypen, die die Deskriptoren verbinden. Ober- und Unterbegriffsrelationen liefern weitere oder engere transformierte Terme, daher ist davon auszugehen, dass die Treffermenge bezüglich des Ausgangsterms und damit bezüglich des Informationsbedürfnisses des Nutzers, zu groß, bzw. zu speziell ist. Das gleiche gilt für die Ähnlichkeitsrelation: Sie liefert ein verwandtes Konzept zur ursprünglichen Anfrage. Die beste Abbildung wird durch die Äquivalenzrelation erbracht. Es ist daher empfehlenswert, nur letztere automatisiert einzusetzen und dem Nutzer die weiteren Relationen zur Verfeinerung bzw. Ausweitung seiner Suche anzubieten. Es muss daher möglich sein, die Anfrage an den Heterogenitätsservice auf einen bestimmten Relationstyp einzuschränken.
- Bei der erweiterten Suche kann ein Nutzer die Datenbanken auswählen, in denen er suchen möchte. Durch die Auswahl sind die Zielvokabulare bekannt, in die transformiert werden soll, d.h. die Relationen können bei der Anfrage an den Heterogenitätsservice auf diese eingeschränkt werden.
- Eventuell hat ein Nutzer seine Suchterme aus einem Online-Thesaurus oder Search Term Recommender (vgl. Petras 2006) ausgewählt und auf diese Weise das Ausgangsvokabular, von dem aus transformiert werden soll, vorgegeben. Da Terme in mehreren Vokabularen vorkommen können, sollte auch das Ausgangsvokabular in der Anfrage festgelegt werden können.
- Längerfristig soll der Heterogenitätsservice auch andere Transformationen als die intellektuell erstellten zurückgeben (z.B. durch statistische Verfahren ermittelte Relationen), daher wird in der Anfrage noch ein Feld vorgesehen, in dem die Transformationsmethode spezifiziert werden kann.

Für das Format von Anfrage und Rückgabe wird ebenfalls XML gewählt. Es gelten die gleichen Vorteile: die Kommunikation ist sowohl durch Anwendungen prozessierbar, aber auch menschenlesbar und XML ist ebenfalls ein offener, frei zugänglicher Standard.
Abbildung 5 zeigt das Format der Anfrage, der Übersichtlichkeit nicht in XML, sondern als Baumstruktur dargestellt. Die Klammern bedeuten, dass dieser Parameter optional ist.

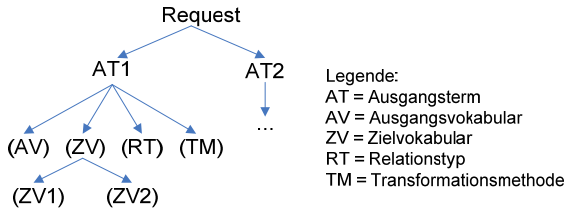

Abb. 5: Format der Anfrage.

Um das Auswerten des Ergebnisses zu erleichtern, sollte das Format der Rückgabe einheitlich sein, unabhängig davon, wie viele Einschränkungen (z.B. Zielvokabular, Relationstyp) in der Anfrage spezifiziert wurden. Es ist allerdings nicht ausreichend, nur die transformierten Terme zurück zu geben, da sonst unklar ist, für welches Zielvokabular sie sind. Weiterhin sollte eine Zuordnung von Ausgangs- zur Zielvokabular erfolgen, damit ersichtlich ist, welche Konkordanz angewendet wurde. Für die Rückgabe ergibt sich damit eine Baumstruktur, die anhand des Anfrageterms „Bildungseinrichtung“ in Abbildung 6 beispielhaft dargestellt ist.

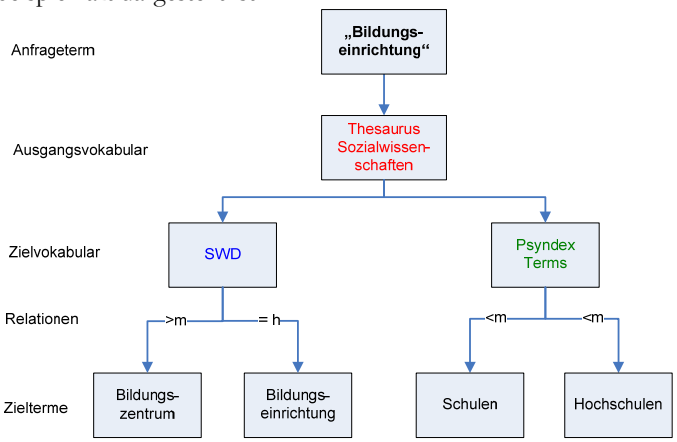

Abb. 6: Beispiel für die Rückgabe (Baumdarstellung).

3.4 Datenbasis des Heterogenitätsservice

Die Erstellung von Crosskonkordanzen erfolgt in Tabellen, die allerdings als Datenbasis für den Heterogenitätsservice nicht geeignet sind, da sie leicht verändert oder einfach verschoben, bzw. gelöscht werden können. Für eine persistente Speicherung, die gleichzeitig einen zuverlässigen Zugriff ermöglicht, bietet es sich an, die Crosskonkordanzen in einer Datenbank abzulegen. Ein weiterer Vorteil davon ist, dass eine Selektierbarkeit und Auswahl der Relationen nach unterschiedlichen Kriterien möglich ist.

Die Speicherung in einer Datenbank erfordert ein Tabellen-Schema, das folgenden Anforderungen genügen muss.

- Kein Informationsverlust gegenüber den Tabellen, in denen die Konkordanzen erstellt werden: Sämtliche Angaben über Relationen, Relevanzen und Zielterme müssen in der Datenbank wieder zu finden sein.
- Selektierbarkeit: Die Crosskonkordanzen sollten nach verschiedenen Kriterien selektierbar sein.
    - Ausgangsterm: Die Transformation einer Anfrage muss bearbeitet werden können, ohne jede Crosskonkordanz einzeln durchsuchen zu müssen, daher werden alle Relationen in einer einzigen Tabelle abgespeichert. Terme, die aus unterschiedlichen Thesauri kommen, sich aber nur in der Groß-/Kleinschreibung oder hinsichtlich der Schreibweise von Umlauten unterscheiden, müssen ebenfalls durch eine einzelne Abfrage zu ermitteln sein. Neben der Originalschreibweise werden sie daher auch in einer normierten Schreibweise (Großschreibung und ohne Umlaute) vorgehalten.
    - Ausgangs- und Zielvokabular: Die Speicherung aller Relationen in einer Tabelle erfordert, dass eine Zuordnung von Relation zu Konkordanz

möglich ist. Daher wird für jede Transformation Ausgangs- und Zielvokabular in extra Spalten gespeichert.

- Relationstyp: Da die verschiedenen Relationstypen unterschiedliche Auswirkungen auf die Treffermenge haben, sollte es möglich sein, die Relationen auf einen Typ (siehe Abschnitt Inhaltliche Realisierung), z.B. die Äquivalenzrelation, zu begrenzen.

Vor dem Laden in die Datenbank werden sowohl Terme als auch Relationen und Relevanzen auf syntaktische Korrektheit überprüft, d.h. die richtige Schreibweise für die Terme, sowie nur die erlaubten Relationen und Relevanzen. Erwähnenswert ist, dass nicht alle Terme eines Thesaurus auch in den Termtransformations-Tabellen zu finden sind, da zum Teil nur Ausschnitte von Thesauri verknüpft wurden (z.B. sozialwissenschaftlicher Ausschnitt der SWD in der Crosskonkordanz TheSoz-SWD, Ausschnitte der Medical Subject Headings).

3.5 Indirekte Termtransformationen

Da der Aufwand für eine vollständige Verknüpfung aller Vokabulare in der Regel zu groß ist, besteht als konzeptuelle Erweiterung der Crosskonkordanzen die Möglichkeit indirekte Termtransformationen anzuwenden. Beispielsweise wird ein Ausgangsterm in Thesaurus C gefunden (siehe auch Abbildung 7), es gibt aber keine direkte Transformation in Thesaurus A, allerdings besteht eine Konkordanz zwischen B und A. Thesaurus B könnte in dem Fall als sogenannte „Switching Language“ benutzt werden, um ebenfalls Termtransformationen in Thesaurus A zu erhalten.

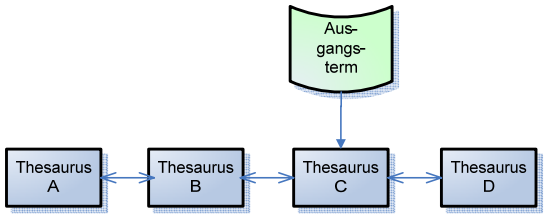

Abb. 7: Mapping zwischen Ausgangsterm und kontrolliertem Vokabular.

3.6 Kontext des Ausgangsterms: Problem der duplizierenden Abbildung

Erfolgt bei der Abfrage einer Term-Transformation keine Einschränkung auf ein Ausgangsvokabular (fehlendes Mapping von Ausgangsterm zu kontrolliertem Vokabular, vgl. Abbildung 8) erhält man das Problem der „duplizierenden Abbildung“. Im Fall der „duplizierenden Abbildung“ ist es möglich, dass der Ausgangsterm syntaktisch zeichengleich in mehreren Thesauri vorkommt (Beispiel: Deskriptor „Internet“), z.B. in Abbildung 8 in Thesaurus B und Thesaurus C. Die Folge ist, dass entschieden werden muss, welche Transformationen angewendet werden: die von Thesaurus C nach Thesaurus A und D (durchgezogene Linie in Abbildung 8) oder die von Thesaurus B nach Thesaurus A und D (gestrichelte Linie in Abbildung 8).

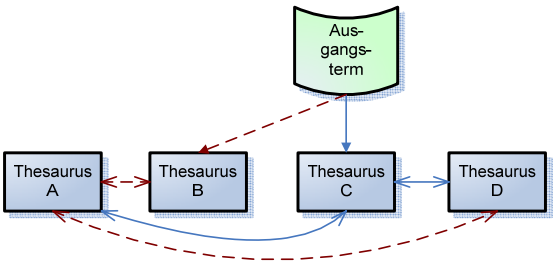

Abb. 8: Kontext des Anfrageterms.

Sind die jeweiligen Relationen von B und C nicht nur vom gleichen Typ, sondern zeigen auch auf dieselben Zielterme, hätte sich das Problem erübrigt. Durch die Vagheit bei der Konkordanzerstellung und durch unterschiedliche Fachkontexte des Thesaurus ist dies leider nur selten der Fall.

Ein Ansatz zur Behebung des Problems könnte das Auswerten der Relationstypen sein. Bestehen die Relationen von Thesaurus B aus Äquivalenzrelationen, die von Thesaurus C aber aus einer Äquivalenz- und einer Oberbegriffsrelation, sollte Thesaurus B als Ausgangsvokabular gewählt werden. Eine weitere Möglichkeit wäre, sämtliche Zielterme pro Thesaurus mit „ODER"-Relationen zu verknüpfen, d.h. sowohl den der Relation von Thesaurus B als auch den der Relation von Thesaurus C. Dies könnte allerdings negative Auswirkungen auf die Qualität des Suchergebnisses haben. Welche Strategien und Heuristiken am besten eingesetzt werden können, muss im Projekt noch ermittelt werden.

4. Ausblick

Der nächste Meilenstein im Projekt KoMoHe fokussiert auf die qualitative Evaluation einzelner Crosskonkordanzen. Im Mittelpunkt der qualitativen Evaluation steht die Untersuchung der durch Termtransformationen für den Nutzer erreichbaren Dokumente. Diese zusätzlichen Dokumente sollen durch Relevanzmessungen gemäß dem Verfahren der TREC[11] und CLEF[12]-Studien über externe Dokumentbewertungen evaluiert werden (zum Evaluationsdesign siehe Mayr/Walter, erscheint). Erste Ergebnisse der Evaluation der Crosskonkordanzen werden für August 2007 erwartet.

Die Evaluation und Untersuchung der zuletzt angesprochenen indirekten Termtransformationen sowie weiterer Spezifika der Crosskonkordanzen soll im Anschluss an das Projekt erfolgen. Im Rahmen dieser Untersuchungen soll auch über die Integration des Gesamtkonzepts „semantische Heterogenitätsbehandlung" in Standardisierungsbemühung des Semantik Webs (vgl. Miles 2006: SKOS[13]) nachgedacht werden (Krause 2006).

Die Auszählung der 10-Jahres Bibliographie der ISKO aus dem Jahr 1999 von Riesthuis/Schmitz-Esser zeigt für den Zeitraum 1989-1999 rechtlich deutlich, dass sich die Literatur rund um die Gesellschaft für Wissensorganisation hauptsächlich mit den Grundlagen von Kompatibilitätsfragen beschäftigt hat und die Erstellung und Evaluation von Crosskonkordanzen und anderen Modulen zur Heterogenitätsbehandlung in der Literatur eine bislang untergeordnete Rolle gespielt haben (vgl. Tabelle 4). Es bleibt zu

[11] <http://trec.nist.gov/>.
[12] <http://www.clef-campaign.org/>.
[13] <http://www.w3.org/2004/02/skos/>.

hoffen, dass die nächste Auszählung dieser Bibliographie vermutlich im Jahr 2009/2010 Fortschritte in der Wissensorganisation und damit in diesen bisher unterentwickelten Systemstellen zeigen kann.

class	description	frequency
281	Objectives and Nature of Systems Compatibility	10
282	Intermediate Languages	1
283	Compatibility in Classing and Indexing	3
284	Establishment of Concordances	5
285	Correlative Indexes. Mapping	
286	Systems Reconciliation, e.g. beteewn Classification Systems and Thesauri, Linking Terms	1
287	Organised Compilation of Compatible Classification Systems and Thesauri, Integration	
288	Compatibility in Subject Areas	1
289	Evaluation of Compatibility	

Tabelle 4: Auszählung der ISKO-Bibliographie von 1999. Systemstelle 28 „Compatibility and Concordances between Indexing Languages“ samt Untergruppen (Riesthuis/Schmitz-Esser, 1999).

5. Literatur

CARMEN-Projekt: CARMEN - Abschlussbericht des Arbeitspakets 12 (AP 12) Crosskonkordanzen von Klassifikationen und Thesauri, 2002. 44 S. URL: <http://www.opus-bayern.de/uni-regensburg/volltexte/2003/242/pdf/CARMENAP12_Abschlussbericht_Netz.pdf>.

Hellweg, Heiko; Krause, Jürgen; Mandl, Thomas; Marx, Jutta; Müller, Matthias N.O.; Mutschke, Peter; Strötgen, Robert (2001): Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften. 47 S. (IZ-Arbeitsbericht; Nr. 23) URL: <http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf>.

Krause, Jürgen (2003): Standardisierung von der Heterogenität her denken: Zum Entwicklungsstand Bilateraler Transferkomponenten für digitale Fachbibliotheken. Bonn: IZ Sozialwissenschaften. 32 S. (IZ-Arbeitsbericht; Nr. 28) URL: <http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_28.pdf>.

Krause, Jürgen (2006): Shell Model, Semantic Web and Web Information Retrieval. S. 95-106. In: Harms, Ilse; Luckhardt, Heinz-Dirk; Giessen, Hans W. (Hrsg.): Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann. München: K. G. Saur.

Marx, Matthias N.O. (2005): Empirische Ergebnisse zu Evaluation semantischer Transformationen. Bonn: IZ Sozialwissenschaften. (unveröffentlichter IZ-Arbeitsbericht).

Mayr, Philipp (2006a): Informationsangebote für das Wissenschaftsportal vascoda - eine Bestandsaufnahme. Bonn: Informationszentrum Sozialwissenschaften. 67 S. (IZ-Arbeitsbericht Nr. 37) URL: <http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_37.pdf>.

Mayr, Philipp (2006b): Thesauri, Klassifikationen & Co – die Renaissance der kontrollierten Vokabulare? S. 151-170. In: Hauke, Petra; Umlauf, Konrad (Hrsg.): Vom Wandel der Wissensorganisation im Informationszeitalter. Festschrift für Walther Umstätter zum 65. Geburtstag. Bad Honnef: Bock + Herchen Verlag. (Beiträge zur Bibliotheks- und Informationswissenschaft: Band 1) URL: <http://edoc.hu-berlin.de/miscellanies/vom-27533/151/PDF/151.pdf>.

Mayr, Philipp; Stempfhuber, Maximilian; Walter, Anne-Kathrin (2005): Auf dem Weg zum wissenschaftlichen Fachportal – Modellbildung und Integration heterogener Informationssammlungen.

S. 29-43. In: Ockenfeld, Marlies (Hrsg.): 27. DGI-Online-Tagung. Frankfurt am Main: DGI. URL: <http://www.ib.hu-berlin.de/~mayr/arbeiten/mayr_etal_dgi05.pdf>.

Mayr, Philipp; Walter, Anne-Kathrin (erscheint): Einsatzmöglichkeiten von Crosskonkordanzen. In: Stempfhuber, Maximilian (Hrsg.): Lokal - Global Vernetzung wissenschaftlicher Infrastrukturen - 12. IuK-Jahrestagung. Bonn: IZ Sozialwissenschaften

Miles, Alistair (2006): SKOS: Requirements for Standardization. In: URL: <http://isegserv.itd.rl.ac.uk/public/skos/press/dc2006/camera-ready-paper.pdf>.

Petras, Vivien (2006): Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages. University of California, Berkeley Berkeley, USA, URL: <http://www.sims.berkeley.edu/~vivienp/diss/>.

Riesthuis, Gerhard; Schmitz-Esser, Winfried (1999): Bibliography of 10 Years International Society for Knowledge Organization. In: Knowledge Organization 26, Nr. 4, S. 203-260.

Strötgen, Robert (2004): ASEMOS. Weiterentwicklung der Behandlung semantischer Heterogenität. S. 269-281. In: Bekavac, Bernard; Herget, Josef; Rittberger, Mark (Hrsg.): 9. Internationales Symposium für Informationswissenschaft (ISI 2004). Chur (Schriften zur Informationswissenschaft) URL: <http://www.stroetgen.de/Dokumente/isi2004.pdf>.

Walter, Anne-Kathrin; Mayr, Philipp; Stempfhuber, Maximilian; Ballay, Arne (2006): Crosskonkordanzen als Mittel zur Heterogenitätsbehandlung in Informationssystemen. S. 205-225. In: Stempfhuber, Maximilian (Hrsg.): In die Zukunft publizieren - 11. IuK-Jahrestagung. Bonn: IZ Sozialwissenschaften. URL: <http://www.gesis.org/information/forschungsuebersichten/tagungsberichte/publizieren/iuk_tagungsband_11_walter.pdf>.

Zeng, Marcia Lei; Chan, Lois Mai (2004): Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. In: Journal of the American Society for Information Science and Technology 55, Nr. 3, S. 377-395.

Zhang, Xueying (2006): Rough set theory based automatic text categorization and the handling of semantic heterogeneity. Bonn: IZ Sozialwiss. 151 S. S. (Forschungsberichte; Bd. 8) ISBN 3-8206-0149-X.

6. Anhang

Ausschnitte aus der Crosskonkordanz Thesaurus Sozialwissenschaften (TheSoz) zu Standard Thesaurus Wirtschaft (STW)

TheSoz	Relation	Relevanz	STW
Bundeskanzler	=	h	Regierungschef
Bundeskompetenz	<	g	Föderalismus
Bundeskriminalamt	0		
Bundesland	=	h	Teilstaat
Bundesministerium	=	h	Ministerium
Bundesnachrichtendienst	=+	h	Staatsschutz + Deutschland
Bundespolitik	<	g	Politik
Bundespräsident	=	h	Staatsoberhaupt
Bundesrat	^	m	Parlament
Bundesrecht	<	m	Recht
Bundesregierung	=	h	Regierung
Bundesrepublik Deutschland	=	h	Deutschland
Bundesrepublik Jugoslawien	=	h	Serbien-Montenegro
Bundessozialgericht	^	h	Sozialgericht
Bundessozialhilfegesetz	=	h	Sozialhilferecht
Bundesstaat	=	m	Föderalismus
Bundestag	=	h	Parlament
Bundestagswahl	=	h	Wahl
Bundesverfassungsgericht	^	h	Verfassungsgericht
Bundesversammlung	0		
Bundesverwaltung	<	g	Öffentliche Verwaltung
Bundesverwaltungsgericht	^	h	Verwaltungsgericht
Bundeswehr	=	h	Militär
Bundeszentrale für politische Bildung	0		
Bund-Länder-Beziehung	^	h	Föderalismus
Bund-Länder-Kommission	^	g	Föderalismus
Bündnis 90/ Die Grünen	=+	h	Ökologische Partei + Deutschland
Eurokommunismus	=	m	Kommunismus
Europa	=	h	Europa
Europäische Zentralbank	=+	h	Zentralbank + EU-Staaten
Europäer	0		
europäische Institution	^+	m	Internationale Organisation + EU-Staaten
europäische Integration	=	h	Europäische Integration

...

Ausschnitte aus der Crosskonkordanz Psyndex Terms zu Medical Subject Headings (MeSH)

Psyndex deutsch	Relation	Relevanz	MeSH
AIDS	=o	h	Acquired Immunodeficiency Syndrome
AIDS	>o	h	AIDS-Related Complex
AIDS	>o	h	AIDS Dementia Complex
AIDS	>o	m	HIV Wasting Syndrome
AIDS	^o	h	HIV Infections
AIDS-Demenz	=	h	AIDS Dementia Complex
AIDS-Prävention	<	m	Acquired Immunodeficiency Syndrome
Akademisches Fachpersonal	<o	m	Occupational Groups
Akademisches Fachpersonal	<o	m	Faculty
Akademisches Fachpersonal	^o	g	Professional Role
Akademisches Fachpersonal	^o	m	Education, Professional
Akathisie	^	h	Psychomotor Agitation
Akkulturation	=	h	Acculturation
Akrophobie	<	m	Phobic Disorders
Aktivismus	<	m	Politics
Aktivitätsniveau	<	g	Human Activities
Aktivitätstheorie	<+	g	Human Activities + Aging
Akupunktur	=	h	Acupuncture
Akustik	=	h	Acoustics
Akustische Displays	<	g	Acoustic Stimulation
Akustische Halluzinationen	=+	h	Acoustics + Hallucinations
Akustischer Nerv	=	h	Cochlear Nerve
Akustischer Reflex	=	h	Reflex, Acoustic
Akute Alkoholvergiftung	=	h	Ethanol/poisoning
Akute Psychose	<	h	Psychotic Disorders
Akute Schizophrenie	=o	h	Schizophrenia
Akute Schizophrenie	^o	m	Schizophrenia and Disorders with Psychotic Features
Akute Schizophrenie	^o	m	Schizophrenia, Disorganized
Akute Schizophrenie	^o	m	Schizophrenia, Catatonic
Akute Schizophrenie	^o	m	Schizophrenia, Childhood
Akute Schizophrenie	^o	m	Schizophrenia, Paranoid
Akute Stressstörung	^	h	Stress Disorders, Traumatic, Acute
Alanine	=	h	Alanine
Alarmreaktionen	^	m	Escape Reaction
Alaska-Bevölkerung	<	m	Alaska

...