

Zum Stand der Heterogenitätsbehandlung in vascoda: Bestandsaufnahme und Ausblick

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

Abstract. Der Beitrag stellt das Verfahren zur Erstellung von Crosskonkordanzen (CK) im Projekt „Kompetenznetzwerk Modellbildung und Heterogenitätsbehandlung“ (KoMoHe)¹ sowie das Netz der bis dato entstandenen Terminologie-Überstiege vor. Neben CK zwischen Indexierungssprachen innerhalb eines Anwendungsgebiets (z.B. Sozial- und Politikwissenschaften), werden Termbeispiele vorgestellt, die Fächer unterschiedlicher Fachgebiete verknüpfen. Es werden weiterhin typische Einsatzszenarien der CK innerhalb von Informationssystemen präsentiert. Die am IZ entstandenen CK sollen künftig über einen Terminologie-Service als Web Service genutzt werden. Der sog. Heterogenitätsservice, der als Term-Umschlüsselungs-Dienst fungieren soll, wird exemplarisch anhand konkreter Fragestellungen vorgeführt.

1 Einleitung

BMBF und DFG haben sich für die Schaffung eines generellen Wissenschaftsportals und von Fachportalen in einem vernetzten Ansatz entschieden, wobei die Projektförderlinien der DFG zu den Virtuellen Fachbibliotheken² und die des BMBF zu den Informationsverbänden zusammengeführt werden sollen. Für den Gesamtkontext wurde der Name vascoda gewählt. Der Ansatz besteht aus einem generellen Einstieg, dem Wissenschaftsportal vascoda³, das zu Fachportalen und Fachclustern weiterleitet.

Die Konsequenz sind hochkomplexe Strukturen und Anforderungen bei der Integration der für vascoda relevanten Informationsangebote, sowohl auf inhaltlicher als auch auf organisatorisch-technischer Ebene. Die Strukturen gehen weit über die hinaus, die in den virtuellen Fachbibliotheken und Informationsverbänden selbst behandelt wurden. Sie führen zu Fragestellungen, für die die vorgefertigten Lösungsmodelle, die bisher die Bibliothekare und die „Macher“ der Informationszentren verwendet haben, nicht mehr greifen. Gleichzeitig stellen sich neue konzeptuelle Fragen der Integration bisher unverbunden entwickelter Module.

Die Klärung dieser Fragen wird im Teilprojekt „Modellbildung und Heterogenitätsbehandlung“ im Kompetenznetzwerk „Neue Dienste, Standardisierung, Metadaten“ bearbeitet und deckt folgende Problemstellungen ab:

- Modellbildung zum Wissenschaftsportal vascoda als Vorbereitung der notwendigen Abstimmungsprozesse, die von der Koordinationsstelle der TIB Hannover moderiert werden. Die Modellbildung soll dabei so prinzipiell angelegt sein, dass ihre Aussagen auf ähnliche Fragestellungen in anderen Verwendungskontexten übertragbar sind (Krause/Mayr, 2006; Mayr, 2006a; Mayr, 2006b).
- Einbringen des spezialisierten Know-hows für die Problembehandlung der Fragen zur Heterogenitätsbehandlung als Ergänzung zur Standardisierung durch einheitliche Metadaten.

Abbildung 1 zeigt ein abstraktes Modell einer Portalinfrastruktur, das alle drei vascoda-Ebenen enthält und auf die Integration von Fachportalen fokussiert. Die erste Form der Integration beginnt unterhalb der Fachportale, indem die einzelnen Angebote (Module) untereinander durch Komponenten der Heterogenitätsbehandlung integriert werden. Die Integration der Angebote erfolgt bilateral (siehe

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 <http://www.virtuellefachbibliothek.de>

³ <http://www.vascoda.de>

Doppelpfeile unter a) in Abb. 1). Auf der Ebene der Fachportale kann die Verknüpfung (Mapping) der kontrollierten Vokabulare direkt zwischen den Referenzthesauri der Fachportale stattfinden (siehe Doppelpfeile b) in Abb. 1).

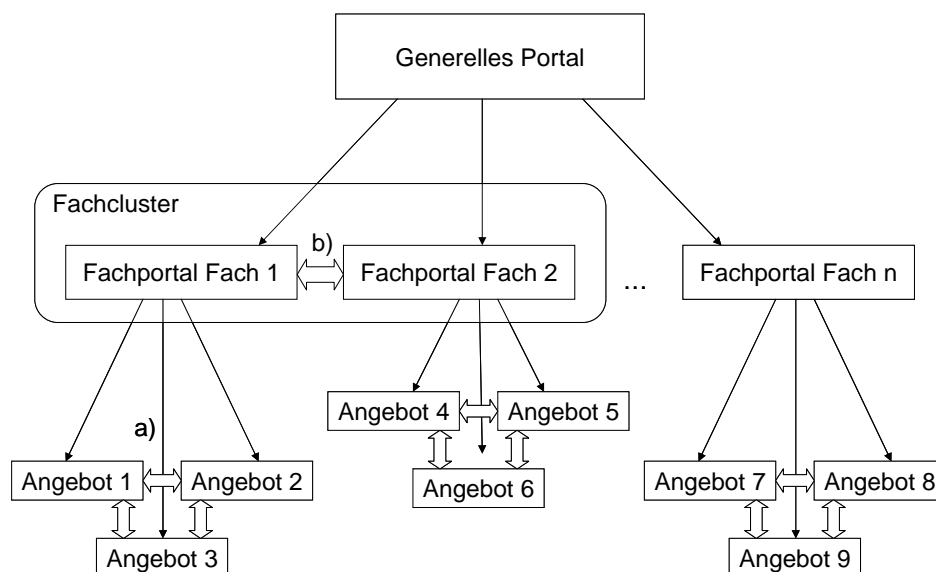


Abb. 1: Kaskadierendes Modell einer Portalinfrastruktur (aus Mayr et al., 2005)

Die Überlegungen zu Fachportalen als zentrale Verknüpfungspunkte gehen u. a. von der Beobachtung aus, dass Komplexität und Aufwand für eine hochwertige Integration heterogener Informationsangebote zunimmt, je abstrakter die betrachtete Integrationsebene ist. Gleichzeitig sinkt die Möglichkeit, Spezifika innerhalb eines Faches, z. B. domänenspezifische Restriktionen, die zu einer lokalen Komplexitätsreduktion oder Qualitätssteigerung führen können, auf höherer Ebene auszunutzen. In Abbildung 1 können z. B. verwandte Thesauri (a) innerhalb eines Faches mit sehr viel weniger Aufwand und höherer Präzision aufeinander abgebildet werden, als dies in der Menge aller Thesauri aller Fächer möglich ist. Zusätzlich reduziert sich der Aufwand zur fachübergreifenden Verbindung von Erschließungswerkzeugen erheblich, sobald dies mit wenigen „Referenzthesauri“ (b) der Fächer geschehen kann, anstatt mit allen Thesauri aller Fächer.

2 Heterogenitätsbehandlung

Eines der klassischen Probleme im Information Retrieval ist die Vagheit, die zwischen einer Anfrage eines Benutzers und den Termen, die ein Dokument beschreiben, besteht. Konkret bedeutet dies, dass ein Benutzer für eine beliebige Fragestellung andere Terme verwendet, als derjenige, der die Inhaltserschließung für entsprechende Dokumente vornimmt (vgl. language problem bei Petras, 2006). Die Vagheit zwischen Anfrage- und Dokumentenebene wird bei Krause V1 genannt (siehe Abbildung 2) und kann durch Verfahren zur Termerweiterung behandelt werden (2003). Lösungsvorschläge für die Vagheitsbehandlung auf der Ebene V1 in der Form von Search Term Recommender Systemen finden sich z. B. bei Petras (2006).

Handelt es sich bei den zu durchsuchenden Datenbeständen um homogen erschlossene Datenbanken, sind die Verfahren zur Behandlung von V1 ausreichend. Bei heterogenen Dokumentenbeständen sind andere Verfahren anzuwenden. Durch den Einsatz von unterschiedlichen Thesauri und kontrollierten Vokabularen 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. Die Behandlung dieser Vagheit geschieht bilateral zwischen den einzelnen Dokumentenbeständen, wie durch Abbildung 2 (V2, V3) verdeutlicht wird. Mehrstufig bedeutet in diesem Zusammenhang, dass auch eine Kombination von Verfahren möglich ist.

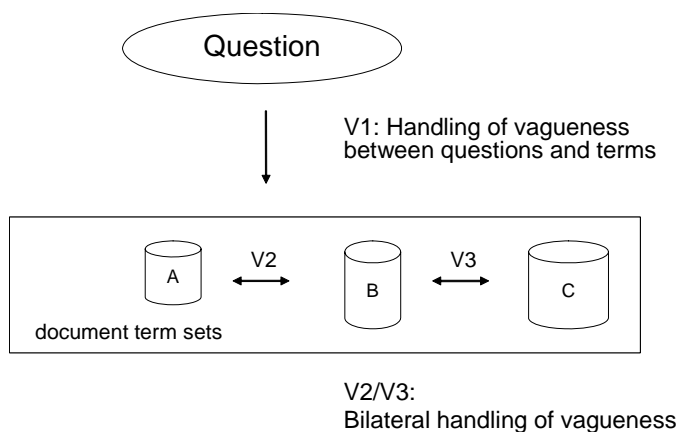


Abb. 2: Vagheitsbehandlung im Information Retrieval (Zwei-Schritt-Modell aus Hellweg et al., 2001)

Auf semantischer Ebene wird Heterogenitätsbehandlung (Vagheit V2/3) durchgeführt, indem Transferkomponenten entwickelt werden, die das kontrollierte Wortmaterial (Deskriptoren) einzelner Vokabulare durch Termtransformationen aufeinander beziehen. Im oberen Beispiel (Abb. 2) 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. Zur Erstellung der Termtransformationen gibt es verschiedene Ansätze (vgl. Hellweg et al., 2001):

- Intellektuell: Für Terme eines Ausgangsvokabulars werden intellektuell Relationen zu passenden Termen eines Zielvokabulars gebildet. Es sind unterschiedliche Relationstypen möglich (siehe auch Abschnitt 2.1). Diese Art der Termtransformation wird Crosskonkordanz genannt.
- Statistisch: Bei diesen Verfahren werden semantische Relationen mit Hilfe von statistischen Methoden (Co-occurrence Analysen) automatisch erzeugt (vgl. Strötgen, 2004; Marx, 2005, Zhang, 2006).
- 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.

2.1 Crosskonkordanzen als intellektuelles Verfahren zur Heterogenitätsbehandlung

Crosskonkordanzen sind gerichtete, relevanzbewertete Relationen zwischen Termen zweier Thesauri, Klassifikationen oder auch anderer kontrollierter Vokabulare. Die Erstellung der Relationen erfolgt intellektuell. Konkordanzen sind bislang in mehreren Projekten am IZ entwickelt worden, u.a. in CARMEN AP12 (siehe CARMEN 2002) oder auch für den interdisziplinären Informationsdienst infoconnex⁴ (eine Übersicht der Verfahren und Projekte findet sich in Zeng/Chan, 2004). Seit Ende 2004 werden innerhalb des Projekts „Kompetenznetzwerk Modellbildung und Heterogenitätsbehandlung“ eine Vielzahl an Crosskonkordanzen zwischen unterschiedlichen Fächern bearbeitet (siehe auch Abschnitt 2.2).

Die Erstellung der Term-Term Relationen erfolgt in Tabellen. In der linken Spalte sind die Ausgangsterme eingetragen, in der zweiten Spalte folgt der Typ der Relation, eine Relevanzbewertung und in der rechten Spalte die Entsprechungen im Zielthesaurus. 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:

- Äquivalenzrelation („=“): für Terme, die das gleiche Konzept bezeichnen

⁴ <http://www.infoconnex.de>

- Oberbegriffsrelation („<“): für Terme, die in einer Hierarchiebeziehung stehen (Teil-Ganzes, Abstraktion)
- Unterbegriffsrelation („>“): wie Oberbegriffsrelation
- Ähnlichkeitsrelation („^“): für Terme die ähnliche oder verwandte Konzepte bezeichnen
- Nullrelation („0“): wird gesetzt, wenn sich keine Entsprechung im Zielthesaurus identifizieren lässt.

Jede der Relationen wird zusätzlich nach Relevanz bewertet und dadurch eine Aussage über die zu erwartende Relevanz der Treffermenge gemacht (Abstufung: hoch, mittel, gering). Tabelle 1 zeigt beispielhaft einen Ausschnitt aus einer Konkordanz. Weitere Crosskonkordanz-Beispiele finden sich in Walter et al. (2006).

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

2.2 Übersicht: verbundene Vokabulare und semantisches Netz der Crosskonkordanzen

Mittlerweile sind insgesamt 18 kontrollierte Vokabulare aus acht⁵ Fachgebieten (siehe auch Abbildung 3) durch Crosskonkordanzen verbunden worden.

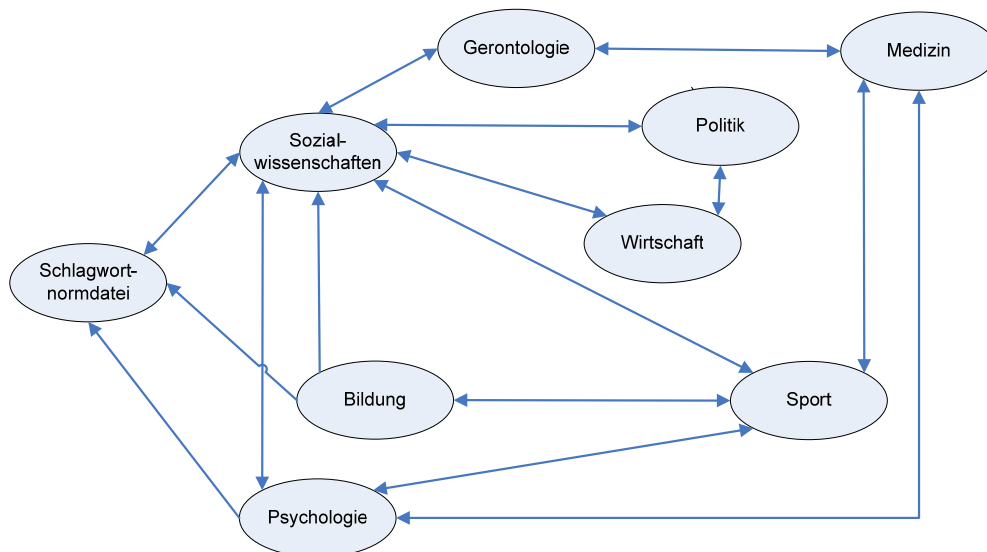


Abb. 3: Vernetzung der Fachgebiete durch CK

Es existieren 21 Crosskonkordanzen (bilaterale Konkordanzen) sowie drei unilaterale Konkordanzen.. Vier der Crosskonkordanzen und zwei der unilaterale Konkordanzen wurden bereits in 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 Ende 2006). Zur Speicherung der Crosskonkordanzen siehe auch Abschnitt 3.2. Tabelle 2 zeigt eine Aufstellung der Vokabulare die durch Crosskonkordanzen verbunden sind.

⁵ Die Schlagwortnormdatei als einziges universelles Vokabular im Projekt spielt aufgrund ihrer Größe und fachlichen Zuordenbarkeit eine Sonderrolle und wird daher gesondert aufgeführt.

	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	ZB Med Köln
14	Psy	Psyndex Terms	5,300	Psyndex	ZPID Trier
15	STW	Standard Thesaurus Wirtschaft	5,600	Econis	ZBW Kiel
16	SWD	Schlagwortnormdatei	400,000 ⁶	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 2: Überblick über die verbundenen Vokabulare

3 Heterogenitätsservice

Die erstellten Crosskonkordanzen werden über einen Dienst, den sogenannten Heterogenitätsservice 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 ist.

3.1 Funktionalität

Es gibt mehrere Einsatzmöglichkeiten für den Heterogenitätsservice. Basisfunktionalität ist der Dienst des Terminologie-Mappings für Fachportale. Weiterhin ist ein Einsatz des Service als Rechercheunterstützung für den Nutzer denkbar. Das durch die Crosskonkordanzen entstandene semantische Netz kann bei der Formulierung seines Suchbedürfnisses 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 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 Datenbank A erschließt. Die Datenbanken B und C sind verschieden 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 Transformationen in die Vokabulare (Zielvokabulare) der

⁶ Bislang wurde nur der sozialwissenschaftliche Ausschnitt der SWD-Terme (ca. 8,000) in die Datenbank importiert.

Datenbanken B und C gefragt. Falls vorhanden, wird die Anfrage pro Datenbank modifiziert und anschließend die Abfrage gestartet.
Abbildung 4 verdeutlicht das Szenario.

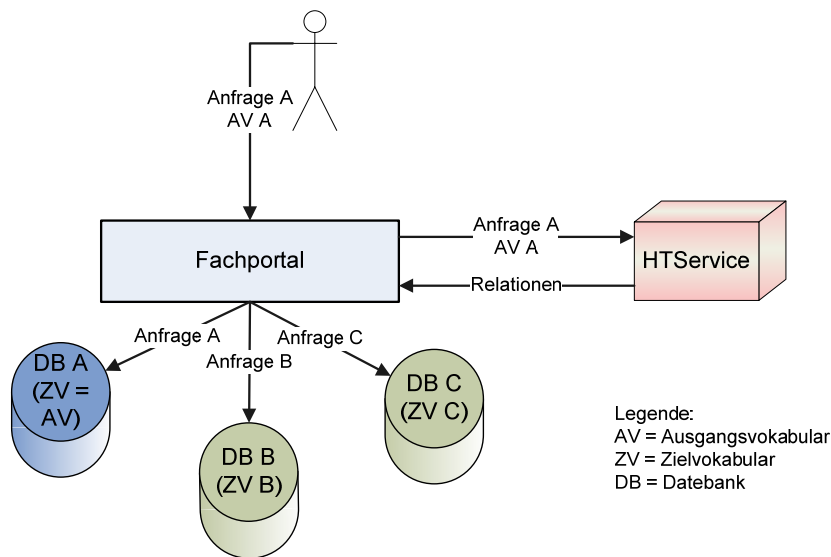


Abb. 4: Einsatzszenario des Heterogenitätsservice

Technische Realisierung

Grundlage für den Heterogenitätsservice ist die Web Service-Technologie. Das Kommunikationsprotokoll SOAP⁷ 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 es ein offener Standard, der ohne Einschränkungen zugänglich ist. Die Realisierung als Web Service bietet einen weiteren Vorteil für die automatisierte Kommunikation zwischen Anwendungen: es gibt ein standardisiertes Format zur Beschreibung der Schnittstelle, d.h. es ist spezifiziert, welche Funktionen der Service anbietet, wie sie aufgerufen werden und wie die Antwort aufgebaut ist. Auf diese Weise kann sehr einfach eine Anfrage an den Dienst erfolgen.

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 in Abschnitt 2.1 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

⁷ <http://www.w3.org/TR/soap12-part1/>

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 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.

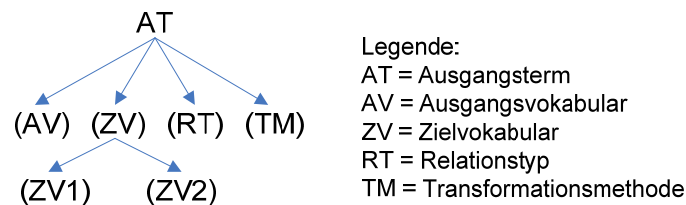


Abb. 5: Format der Anfrage

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- zu 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.

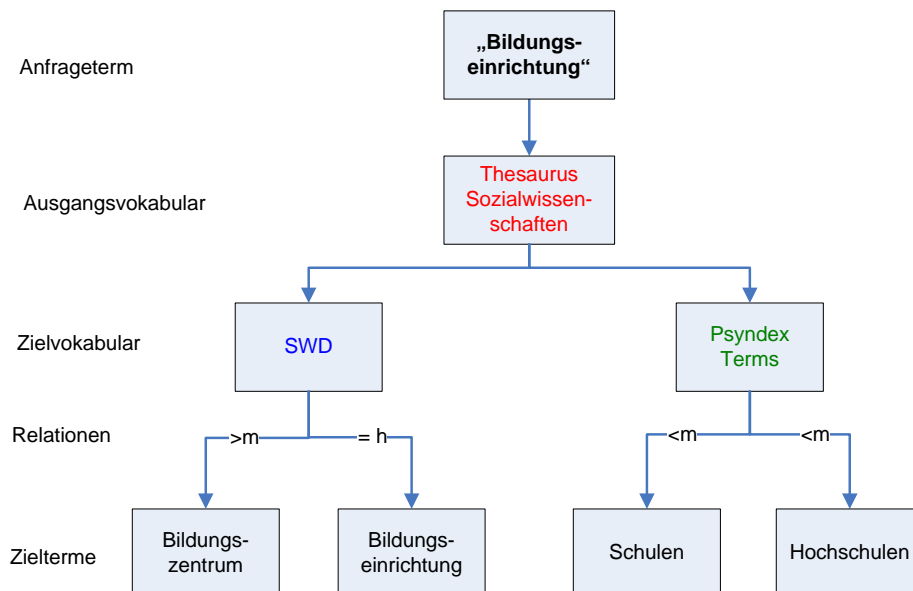


Abb. 6: Beispiel für die Rückgabe

3.2 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.

- (1) 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.
- (2) 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 3.1—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 CK TheSoz-SWD).

Die oben beschriebenen Tabellen werden allerdings nicht nur als Datenbasis für den Heterogenitätsservice genutzt, auch für die Evaluation der Crosskonkordanzen wird darauf zurückgegriffen. Neben quantitativen Auswertungen wird eine qualitative Analyse durchgeführt, die im folgenden Abschnitt beschrieben wird.

4 Evaluation der Crosskonkordanzen

Anfang 2005 konnte bereits mit einer ersten Voranalyse der in infoconnex entstandenen Crosskonkordanzen begonnen werden (siehe Mayr et al., 2005 und Walter et al., 2006). Die bisherigen Arbeiten konnten empirisch zeigen:

- Die Crosskonkordanzen bringen trotz gewisser Überlappungen zwischen den Vokabularen eine signifikante Vokabularerweiterung (Erweiterung des Suchraums), die dem Recherchierenden ausgehend von jedem verbundenen Vokabular zur Verfügung steht (siehe dazu Abbildung 7 und 8).
- Die Crosskonkordanzen erweitern die Treffermenge für Schlagwort-Anfragen und erhöhen damit den Recall bei der datenbankübergreifenden Suche.
- Besonders im Bereich interdisziplinärer Fragestellungen konnte exemplarisch gezeigt werden, dass Crosskonkordanzen einen informationellen Mehrwert bieten, da sie Nischen eines Fachgebiets mit potentiell zentraleren Bereichen eines anderen Fachgebiets verbinden können.

„Weiterhin fällt auf, dass die Überführung der TheSoz-Deskriptoren in das Vokabular des PsyT (CK) terminologisch schwieriger ist und folglich viel häufiger Deskriptorkombinationen verwendet werden müssen, um die Semantik der TheSoz-Deskriptoren auszudrücken. Beispiele hierfür sind „Arbeitslosigkeit + Arbeiter“, „Gewerkschaft + Politik“ oder „Modell + Entwicklung“ (vgl. Walter et al. 2005).

Zusätzlich zur quantitativen Analyse ist für die nächsten Monate im Projekt KoMoHe eine qualitative Analyse der erstellten Crosskonkordanzen geplant. 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 und CLEF-Studien über externe Dokumentbewertungen evaluiert werden. Unsere Arbeitshypothese lautet:

Die eingesetzten Crosskonkordanzen verbessern das Sucherlebnis, indem sie mehr und präzisere Suchergebnisse (Dokumente) besonders in den durch Termtransformationen verbundenen Datenbanken liefern. Die Crosskonkordanzen verbessern das Sucherlebnis umso mehr, je deutlicher sich die so verbundenen Datenbanken im Fachgebiet, Scope und Größe unterscheiden. Als Konsequenz einer verstärkt integrierten Suche wird die Resultatsmenge interdisziplinärer, d.h. es werden mehr relevante Dokumente aus benachbarten Fachgebieten gefunden.

Folgende Tests sind vorgesehen:

1. Test innerhalb der Sozialwissenschaften: Es soll getestet werden an Anfragen und Datenbanken aus dem disziplinären Bereich der Sozialwissenschaften. Die natürlichsprachigen Nutzeranfragen und Topics werden von IuD-Experten in Deskriptoren des Thesaurus Sozialwissenschaften übersetzt und in die Vokabulare anderer Datenbanken transformiert (siehe dazu Abbildung 7). Jeweils drei Anfragen werden operationalisiert und an die entsprechenden Datenbanken geschickt: 1) Die natürlichsprachige Anfrage, 2) die übersetzte Anfrage (bestehend aus Deskriptoren) und 3) die transformierte Anfrage (bestehend aus Deskriptoren) werden im Freitextfeld (natürlichsprachige Anfrage) und im Schlagwortfeld der Zieldatenbanken gesucht. Die nachfolgende Relevanzbewertung der Ergebnisdokumente erfolgt durch die Nutzer (alternativ Sachexperten) des Informationsangebots. Die Datenbanken werden anhand ihrer disziplinären Abdeckung und ihrer unterschiedlichen Dokumententypen gewählt. Zusätzlich können die entgegen gesetzten Konkordanzen zum Thesaurus Sozialwissenschaften evaluiert werden, indem die Anfragen in Deskriptoren der anderen Datenbanken übersetzt, in Deskriptoren des Thesaurus Sozialwissenschaften transformiert und dann in der Datenbank SOLIS (der Literaturdatenbank des IZ) gesucht werden.

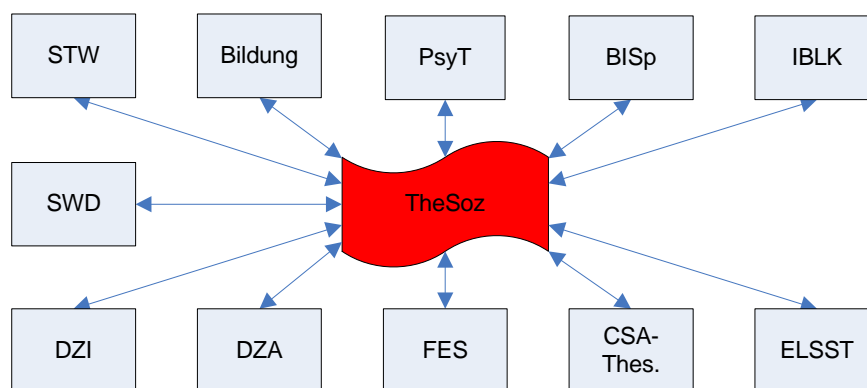


Abb. 7: Netz der Crosskonkordanzen, ausgehend von Thesaurus Sozialwissenschaften

2. interdisziplinärer Test ausgehend von den Sozialwissenschaften: Weiterhin sollen Ausgangsanfragen (reale natürlichsprachige Nutzeranfragen übersetzt in Terme kontrollierter Vokabulare) aus anderen Fachgebieten in mit dem Thesaurus Sozialwissenschaften kompatible Suchanfragen transformiert werden und diese in SOLIS recherchiert werden. Zusätzlich wird auch hier eine Freitext-Suche der natürlichsprachigen Nutzeranfrage getestet. Eine Relevanzbewertung der Ergebnismenge erfolgt wiederum durch den Nutzer (alternativ Sachexperten) des Informationsangebots.

3. Interdisziplinärer Test ohne Beteiligung der Sozialwissenschaften (siehe dazu Abbildung 8): Crosskonkordanzen, die nicht den Thesaurus Sozialwissenschaften involvieren, sollen nach dem gleichen Verfahren stichprobenweise evaluiert werden.

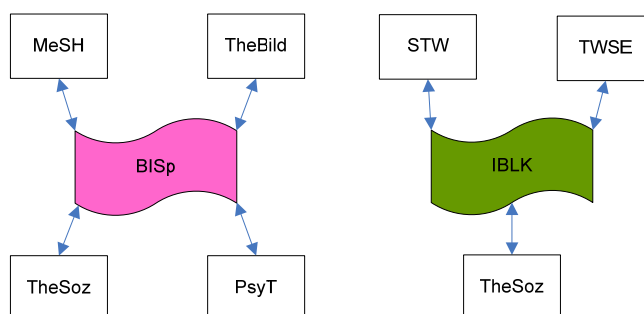


Abb. 8: Netz der Crosskonkordanzen, ausgehend vom BISp- und IBLK-Vokabular

Der besondere Fokus bei der Evaluation auf den Sozialwissenschaften liegt vor allem in der einfachen Verfügbarkeit der Datenbanken begründet.

Die Evaluation der Crosskonkordanzen gliedert sich grob in folgende Schritte:

1. Lieferung realer Nutzeranfragen von den IZ- und Crosskonkordanz-Partnern. Die Partner wurden gebeten, die Nutzeranfragen möglichst operationalisiert in Deskriptoren zu liefern.
2. Formulierung und Pretest der Suchanfragen zu den Evaluations-Szenarien.
3. Suche mit den ausgewählten Suchanfragen (drei Anfragen je evaluierter Nutzeranfragen) in den entsprechenden Datenbanken und Download der Dokumente.
4. Import der Dokumente in das Assessment-Tool und externe Relevanzbewertungen der Dokumente.
5. Auswertung der Relevanzbewertungen.

Wir erwarten im August 2007 erste Ergebnisse der Evaluation der Crosskonkordanzen vorlegen zu können.

Literatur

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; Mayr, Philipp (2006): Allgemeiner Bibliothekszugang und Varianten der Suchtypologie - Konsequenzen für die Modellbildung in vascoda. Bonn: Informationszentrum Sozialwissenschaften. 52 S. (IZ-Arbeitsbericht; Nr. 38) URL: http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_38.pdf

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/Publicationen/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

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/>

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/forschunguebersichten/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