# Einsatzmöglichkeiten von Crosskonkordanzen

*Philipp Mayr, Anne-Kathrin Walter*
gesis - Informationszentrum Sozialwissenschaften, Bonn

## Zusammenfassung

Der Beitrag stellt Einsatzmöglichkeiten und spezifische Problembereiche 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. Des Weiteren wird ein Testszenario samt Evaluationsdesign beschrieben über das der Mehrwert von Crosskonkordanzen empirisch untersucht werden kann.

## 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 Fachbibliotheken[2] und die des BMBF zu den Informationsverbünden zusammengeführt werden sollen. Für den Gesamtkontext wurde der Name vascoda[3] gewählt. Der Ansatz besteht aus einem generellen Sucheinstieg, dem Wissenschaftsportal vascoda, das zu Fachportalen und Fachclustern weiterleitet.

Die Konsequenz dieser Zusammenführung 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 Informationsverbünden selbst behandelt wurden. Gleichzeitig stellen sich neue konzeptuelle Fragen der Integration bisher unverbunden entwickelter Informationsmodule (vgl. Mayr et al., 2005).

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

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.

Der Trend in der aktuellen Fachinformationslandschaft geht hin zu einer Bündelung der Informationsangebote, sowohl innerhalb eines Fachs als auch interdisziplinär. Ziel ist, die Recherche für einen Nutzer mit einem bestimmten Informationsbedürfnis zu erleichtern und ihn beim Auffinden der für ihn relevanten Dokumente zu unterstützen. Neben der Integration auf technischer und struktureller Ebene (siehe auch Strötgen, 2004), muss ebenfalls eine Integration auf semantischer Ebene vorgenommen werden (siehe dazu Krause, 2003). Semantische Heterogenität (bei Krause auch „unvermeidlich verbleibende Heterogenität“) tritt auf, wenn Informationsangebote unterschiedliche Inhaltserschließungssysteme verwenden. Im Gegensatz zur technischen Heterogenität (z.B. unterschiedliche Metadatenschemata), die vgl. einfach homogenisiert werden kann, verbleiben die unterschiedlichen kontrollierten Vokabulare zunächst 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 (vgl. dazu Abbildung 1).

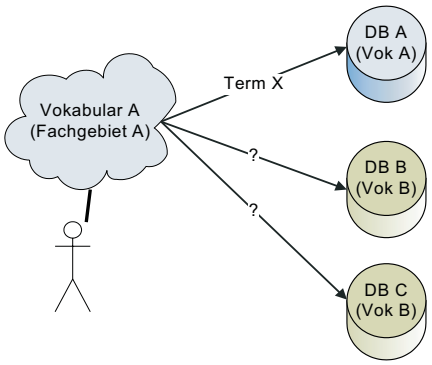


**Abb. 1: Heterogenität kontrollierter Vokabulare**

Es gibt mehrere Möglichkeiten, semantische Heterogenität zu behandeln. Ziel dabei ist, geeignete Termtransformationen in die entsprechenden kontrollierten Vokabulare der zu durchsuchenden Datenbanken bereit zu stellen.

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 Crosskonkordanzen

### 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 infoconnex[4] (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 Aus-

4 http://www.infoconnex.de

gangsterm 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):

- Äquivalenzrelation („="): für Terme, die das gleiche Konzept bezeichnen
- 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 zwischen Thesaurus Sozialwissenschaften und Standard Thesaurus Wirtschaft. Weitere Crosskonkordanz-Beispiele finden sich in Tabelle 2 sowie in Walter et al. (2006).

**Tabelle 1: Beispiel für Crosskonkordanz-Relationen**

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

| | | | |
|---|---|---|---|
| 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-Liste |
| Biologieunterricht | <+o | Biologie + Unterrichtsstunde | BISp-Liste |
| Biologieunterricht | <+ | Biologie + Schule | DZA |
| Biologieunterricht | ^+ | Biologie + Unterricht | FES |

## 2.2 Übersicht: verbundene Vokabulare und semantisches Netz der Crosskonkordanzen

Mittlerweile sind insgesamt 18 kontrollierte Vokabulare aus acht[5] Fachgebieten (siehe auch Abbildung 2) durch Crosskonkordanzen verbunden worden.

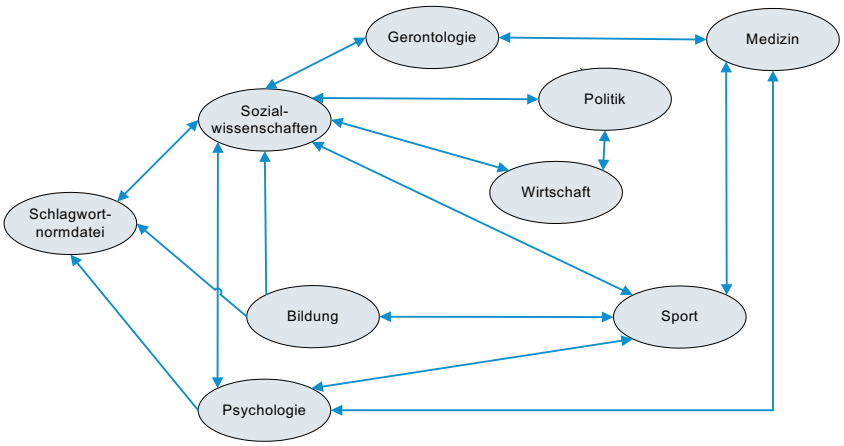


**Abb. 2: Vernetzung der Fachgebiete durch CK**

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

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). Zur Speicherung der Crosskonkordanzen siehe auch Abschnitt 3.2. Tabelle 3 zeigt eine Aufstellung der Vokabulare die durch Crosskonkordanzen verbunden sind.

**Tabelle 3: Überblick über die verbunden Vokabulare**

| | 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 |

## 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 Suchterme 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 3) 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 Term-Transformationen in die Vokabulare (Zielvokabulare) der Datenbanken B und C gefragt.

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

Falls andere Terme für die Datenbanken vorhanden sind, wird die Anfrage pro Datenbank modifiziert und anschließend die Abfrage gestartet.

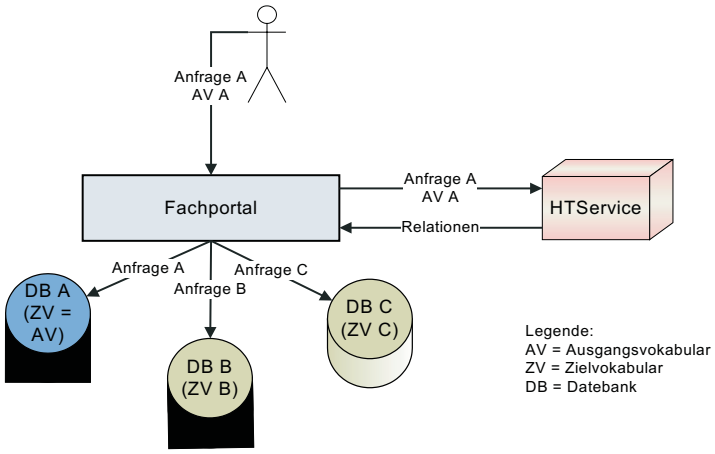


**Abb. 3: Einsatzszenario des Heterogenitätsservice**

### Technische Realisierung

Grundlage für den Heterogenitätsservice ist die Web Service-Technologie. Das Kommunikationsprotokoll SOAP[7] 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 Funktionen 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.

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

- 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 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 4 zeigt das Format der Anfrage, der Übersichtlichkeit nicht in XML, sondern als Baumstruktur dargestellt. Die Klammern bedeuten, dass dieser Parameter optional ist.

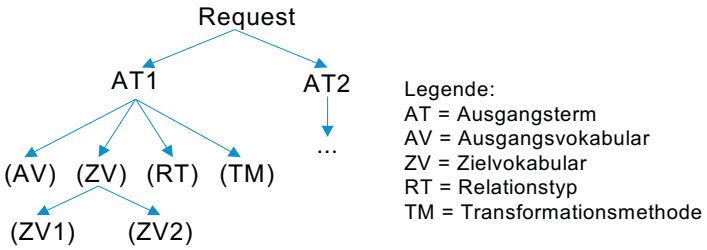


**Abb. 4: Format der Anfrage**

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 5 beispielhaft dargestellt ist.

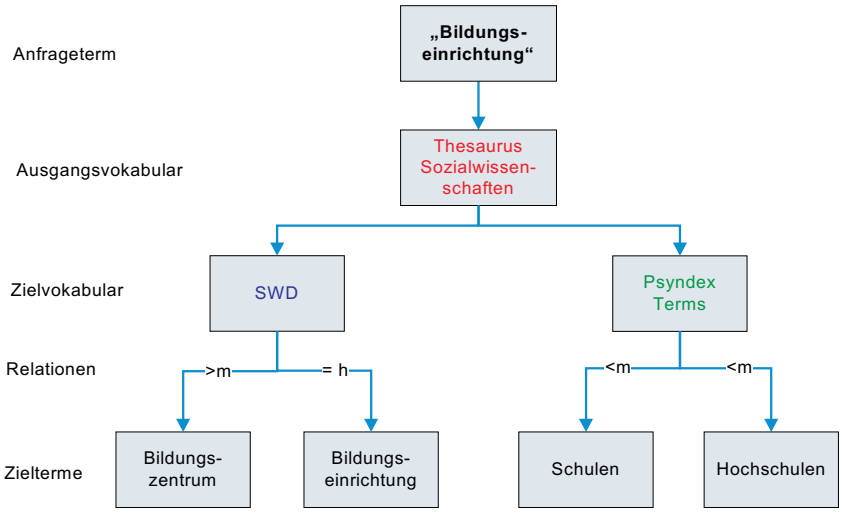


**Abb. 5: Beispiel für die Rückgabe (Baumdarstellung)**

## 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 Crosskonkordanz TheSoz-SWD, Ausschnitte der Medical Subject Headings).

## 3.3 Spezifika von Crosskonkordanzen

### Indirekte Term-Transformationen

Da der Aufwand für eine vollständige Verknüpfung aller Vokabulare in der Regel zu groß ist, besteht als konzeptuelle Erweiterung der Crosskonkordanzen die Möglichkeit indirekte Term-Transformationen anzuwenden. Beispielsweise wird ein Ausgangsterm in Thesaurus C gefunden (siehe auch Abbildung 6), es gibt aber keine direkte Transformation in Thesaurus A, allerdings besteht eine Konkordanz zwischen B und A. Thesaurus B könnte indem Fall als sogenannte „Switching Language“ benutzt werden, um ebenfalls Term-Transformationen in Thesaurus A zu erhalten.

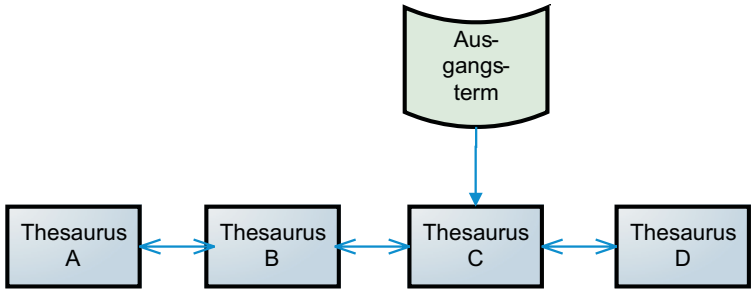


**Abb. 6: Mapping zwischen Ausgangsterm und kontrolliertem Vokabular**

### 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 7) 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 7 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 7) oder die von Thesaurus B nach Thesaurus A und D (gestrichelte Linie in Abbildung 7).

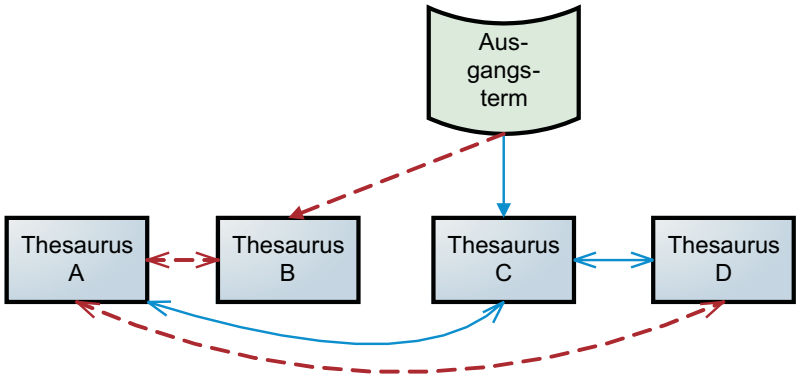


**Abb. 7: Kontext des Anfrageterms**

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.

### Mehrsprachige Deskriptoren

Ein Nutzer sucht beispielsweise in einem Fachportal, das Datenbanken integriert, die mit unterschiedlichsprachigen kontrollierten Vokabularen erschlossen sind. Z.B.

- Datenbank A, erschlossen mit Thesaurus AE (englischsprachig)
- Datenbank B, erschlossen mit Thesaurus BD (deutschsprachig), es existiert eine englischsprachige Übersetzung
- Datenbank C, erschlossen mit Thesaurus CD (deutschsprachig)
- Datenbank D, erschlossen mit Thesaurus DD (deutschsprachig)

Es existieren weiterhin folgende Crosskonkordanzen:

- Zwischen A und B
- Zwischen B und C
- Zwischen B und D

Ein englischsprachiger Nutzer verwendet dabei Deskriptoren des Thesaurus A, dementsprechend wird auch nur die Konkordanz zu Thesaurus B angewendet. In C und D werden aufgrund der englischen Terme in der Regel keine Dokumente gefunden. Hätte Thesaurus B keine englischsprachige Übersetzung, müsste man B als Switching Vocabulary anwenden, um auch etwas in C und D zu finden. Da es aber eine Übersetzung gibt, könnte man einem Nutzer diese ebenfalls als Suchvokabular anbieten, so dass die direkten Transformationen angewendet werden können. Ob dies einen Mehrwert gegenüber der indirekten Transformation bringt, muss ebenfalls noch ermittelt werden.

## 5 Ausblick

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 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., 2006).

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 Term-

transformationen 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 8). 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 Dokumenttypen 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.

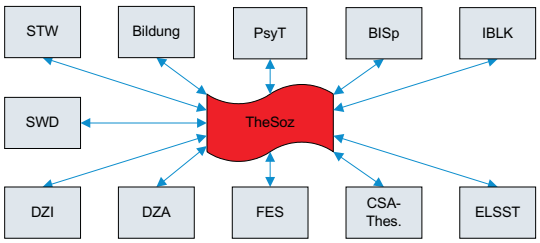


**Abb. 8: Netz der Crosskonkordanzen, ausgehend von Thesaurus Sozialwissenschaften**

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: Crosskonkordanzen, die nicht den Thesaurus Sozialwissenschaften involvieren, sollen nach dem gleichen Verfahren stichprobenweise evaluiert werden.

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

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; 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/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

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/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. (Forschungsberiichte; Bd. 8) ISBN 3-8206-0149- X