Auf dem Weg zum wissenschaftlichen Fachportal – Modellbildung und Integration heterogener Informationssammlungen[1]

Philipp Mayr, Maximilian Stempfhuber, Anne-Kathrin Walter

Informationszentrum Sozialwissenschaften, Bonn

Abstract. Aktuelle Programme und Aktivitäten im Bereich der Fachinformation in Deutschland haben zum Ziel, dezentral vorgehaltene und bislang singulär zugängliche Informationsangebote zu aggregieren und dem Nutzer an einer Stelle und integriert anzubieten. Dies wirft die Frage auf, wie heterogene Informationssammlungen zu organisieren und integrieren sind, so dass ein Produkt entsteht, das den Informationsbedürfnissen der Nutzer entspricht und auf Anbieterseite handhabbar bleibt. Als Lösungsansatz wird das Konzept wissenschaftlicher Fachportale vorgestellt. Der Beitrag beschreibt das Konzept, erläutert die zentrale Bedeutung einer vorhergehenden Modellbildung und geht auf Crosskonkordanzen als Maßnahme zur Integration heterogen erschlossener Informationssammlungen im Rahmen von Fachportalen ein. Als Beispiel wird der Einsatz von Crosskonkordanzen für die fachübergreifende Suche im Informationsverbund infoconnex vorgestellt und gezeigt, wie sie sich im praktischen Einsatz auf das Suchergebnis auswirken.

1 Einleitung

Die fortschreitende Vernetzung sowohl im öffentlichen als auch im wissenschaftlichen Bereich hat deutliche Auswirkungen auf die IuD-Landschaft. Heute werden aufgrund der dezentralen, diversifizierten und interdisziplinären Angebotssituation nicht nur ganz andere Recherchestrategien notwendig, vielmehr haben die Informationssuchenden neue, komplexere Informationsbedürfnisse und – aufgrund der zentralen Rolle des Internets als primäres Zugangsmedium – ein stark verändertes Informationsverhalten. Aktuelle Studien (s. Boekhorst et al. 2003) zeigen anschaulich, welchen Service wissenschaftliche Nutzer erwarten und belegen nun auch empirisch, was im Bereich der Informationswissenschaft auf analytischer Ebene bereits seit langem Diskussionsbasis ist. Nutzer erwarten u. a. die Bündelung jeglicher fachlich relevanter Information an einer Stelle, die Integration von Informationssuche und direktem Zugriff auf Primärinformation sowie die Verbindung von Angeboten unterschiedlicher Fächer zur Vereinfachung des Zugangs bei interdisziplinären Fragestellungen oder Problemen der fachlichen Einordnung bei der Suche nach interdisziplinär verorteten Informationen.

Um der aktuellen Zersplitterung der Informationslandschaft zu begegnen, haben sich das Bundesministerium für Bildung und Forschung (BMBF) und die Deutsche Forschungsgemeinschaft (DFG) zur Schaffung eines generellen Wissenschaftsportals und von Fachangeboten 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. Eines dieser Projekte ist das Wissenschaftsportal vascoda[3], das einen generellen Rechercheeinstieg auf oberster Ebene bietet und zu Fachclustern, z. B. dem Informationsverbund Pädagogik – Sozialwissenschaften – Psychologie (infoconnex[4]), und einzelnen Fachdatenbanken und Bibliothekskatalogen weiterleitet.

Die Konsequenz aus diesem umfassenden Ansatz sind hochkomplexe Strukturen und Anforderungen bei der Integration der für vascoda relevanten Informationsangebote. Insbesondere stellt sich die Frage, wie der mit der parallel zur Zahl der Angebote in vascoda wachsenden Heterogenität begegnet werden soll. Sie betrifft die unterschiedliche Struktur und inhaltliche Erschließung der Daten, die von den einzelnen Informationsanbietern geliefert werden, setzt sich aber auf der technisch-funktionalen Ebene fort. Rein technische Lösungen greifen hier zu kurz, da sie qualitativen Anforderungen – besonders aus

[1] Proceedings der 27. DGI-Online-Tagung 2005, Frankfurt am Main
[2] Siehe http://www.virtuellefachbibliothek.de
[3] Siehe http://www.vascoda.de
[4] Siehe http://www.infoconnex.de

Nutzersicht – nicht gerecht werden. Rein organisatorische Lösungen, also z. B. die Standardisierung der formalen und inhaltlichen Erschließung, sind aus vielfältigen Gründen nicht flächendeckend und kurzfristig umsetzbar. Hilfe ist hier nur von einem Rahmenkonzept zu erwarten, das organisatorische, inhaltliche und technische Maßnahmen in Beziehung setzt, Möglichkeiten zur Standardisierung aufgreift und Lösungsvorschläge für den verbleibenden Rest an Heterogenität aufzeigt.

2 Wissenschaftliche Fachportale als zentrales Strukturelement

Auf breiter Front sind in der Fachinformationslandschaft seit geraumer Zeit Aktivitäten zu beobachten, bislang separate Informationsangebote zusammenzufassen oder miteinander zu vernetzen. Nicht zuletzt haben entsprechende Anforderungen der Nutzer, neue technische Möglichkeiten und das Vorbild von allgemeinen Internetsuchmaschinen als „One-Stop-Shop“ für verteilt lagernde Information diese Entwicklung ausgelöst. Erste Beispiele neuer integrierter Informationsangebote sind bereits sichtbar; die in vielen Virtuellen Fachbibliotheken oder Informationsverbünden angebotene Metasuche über Datenbanken, Bibliothekskataloge und Fachinformationsführer, die fachübergreifende Suche im Informationsverbund infoconnex und – auf höchster Ebene – im Wissenschaftsportal vascoda sind nur einige Vertreter. Die Reaktionen der Nutzer sind durchaus unterschiedlich, zum Teil jedoch sehr kritisch, da die Informationsdienste in einigen Bereichen noch hinter den Erwartungen der Nutzer zurück bleiben.

2.1 Anforderungen an integrierte Informationsdienstleistungen

Bei genauer Analyse möglicher Ursachen wird schnell deutlich, dass angesichts der mit der Integration bislang meist unkoordinierter Informationsangebote einhergehenden Herausforderungen jedes andere Resultat überraschend wäre, übersteigt doch die Komplexität der Aufgabe alles bisherige:

- Unterschiedliche Informationstypen wie Literaturnachweise, Volltexte und Internetquellen – zukünftig u. a. auch Forschungsprojekte und Primärdaten – sollen integriert werden.
- Die Formalerschließung ist uneinheitlich; es existiert ein Spannungsfeld zwischen Standardisierung (z. B. Dublin Core[5] und VLIB Application Profile, siehe Becker et al. 2002), fachspezifischen Anforderungen und schwer beeinflussbaren Gegebenheiten in der Praxis.
- Die Inhaltserschließung ist anbieterabhängig und deckt die gesamte Spannbreite von der freien Schlagwortvergabe durch Autoren über fachübergreifende Erschließungsvokabulare (z. B. Schlagwortnormdatei der Deutschen Bibliothek) bis zu Fachthesauri ab.
- Unterschiedliche Informationstypen mit uneinheitlicher Formal- und Inhaltserschließung müssen sinnvoll kombiniert werden.
- Die Informationssysteme bieten bei der Recherche einen unterschiedlichen Funktionsumfang, der es erschwert, ein methodisch adäquates Verfahren für ihre Kopplung zu definieren.
- Nicht immer ist es sinnvoll und möglich, diese Komplexität und Heterogenität vor dem Nutzer zu verbergen. Daher sind neue Konzepte für Benutzungsoberflächen zu entwickeln, die abhängig von Nutzer und Kontext die interne Arbeitsweise von Informationssystemen zugänglich machen und helfen, den Retrievalprozess zu steuern und die Ergebnisse zu interpretieren.

Diesen komplexen Strukturen auf Anbieterseite stehen ebenso komplexe wie widersprüchliche Anforderungen der Nutzer gegenüber. Sei es die unterschiedliche Tradition der Fächer in ihrem Streben nach schneller, sofort anwendbarer Information vs. dem Bedürfnis nach einem umfassenden und lückenlosen Überblick über ein Thema, oder die Bereitschaft, in Archiven auf nur lokal verfügbare Information zuzugreifen vs. dem Anspruch, vom Arbeitsplatz aus direkt und sofort digitale Information auf den eigenen Rechner zu übertragen – die Nichtbefriedigung eines Bedürfnisses wird immer zu entsprechender Kritik führen. Doch auch der einzelne Nutzer wird sowohl innerhalb eines Rechercheprozesses als auch über einen längeren Zeitraum der Nutzung seine Anforderungen ändern. Sei es ein Wechsel zwischen einem „einfachen“, fachspezifischen Informationsbedürfnis, das sich im Laufe einer Recherche zu einer komplexen, interdisziplinären Fragestellung entwickelt, oder der längerfristige Gewinn an Expertise, der Nutzer sich sowohl hinsichtlich ihres Faches als auch des Informationssystems

[5] Siehe http://www.dublincore.org

von Anfängern zu Profis weiterentwickeln lässt – Nutzer fordern berechtigterweise in jeder Situation adäquate Hilfestellung und Unterstützung durch das System.

2.2 Das Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung

Doch wie sollen nun diese vielschichtigen, oft widersprüchlichen Anforderungen in konkrete Lösungen umgesetzt werden – falls dies überhaupt möglich ist? Die Klärung dieser Fragen soll durch das Teilprojekt „Modellbildung und Heterogenitätsbehandlung“[6] im Kompetenznetzwerk „Neue Dienste, Standardisierung, Metadaten“ unterstützt werden, das vom BMBF gefördert und seit September 2004 am Informationszentrum Sozialwissenschaften durchgeführt wird. Das Projekt beschäftigt sich schwerpunktmäßig mit:

1. der übergreifenden Modellbildung für komplexe Informationsinfrastrukturen, u. a. am Beispiel des Wissenschaftsportals vascoda mit allen nachgeschalteten Ebenen.
2. Fragen zur Heterogenitätsbehandlung (Integration) als notwendige Ergänzung zur Standardisierung durch einheitliche Metadaten.

Aus dieser Perspektive heraus erarbeitet das Projekt Vorschläge – Modelle – zur Strukturierung wissenschaftlicher Informationsangebote und schafft Rahmenbedingungen und Infrastrukturen zur Förderung der semantischen Integration heterogener Informationssammlungen.

Im Bereich der Heterogenitätsbehandlung (siehe Krause 2003 und Krause 2004) fokussieren die Aktivitäten auf die Erstellung von Crosskonkordanzen zwischen fachlichen Erschließungsvokabularen zur besseren Integration innerhalb und zwischen den Fächern sowie dem Aufbau eines zentralen, internetbasierten Dienstes zur Nutzung dieser Wissensstrukturen in Informationsangeboten. Neben Evaluation und Weiterentwicklung von intellektuellen Verfahren zur Erstellung von Crosskonkordanzen und von statistischen Modellen zur Termtransformation hat das Projekt auch das Ziel, die bislang verfügbaren Verfahren in die Praxis zu überführen und ihre Anwendung zu unterstützen. Zu diesem Zweck stehen Auftragsmittel für die Crosskonkordanzerstellung zur Verfügung, die der gezielten Verbindung unterschiedlich erschlossener Informationssammlungen innerhalb eines Faches und über Fachgrenzen hinweg dienen. Die Kapitel 3 und 4 erläutern das Verfahren zur Crosskonkordanzerstellung sowie ihre Einsatzmöglichkeiten und stellen erste Erfahrungen aus dem Praxiseinsatz vor.

2.3 Modellbildung als Voraussetzung für Integration

Wie bereits dargestellt, stehen Informationsanbieter bei der Integration bislang unverbundener Angebote vor einer großen Herausforderung und komplexen Aufgabe. Standardisierung auf formaler und inhaltlicher Ebene kann – zumindest bei neu entstehenden Angeboten – Vielfalt und damit Komplexität reduzieren, auf bestehende Angebote ist sie aber nur begrenzt anwendbar. Verfahren zur Heterogenitätsbehandlung (z. B. Crosskonkordanzen) helfen, den nicht-standardisierbaren Rest zu integrieren, allerdings schwankt der Aufwand dieser Lösungen sehr stark je nach Einsatzszenario. Vorgefertigte technische Lösungen für die Metasuche helfen, Datenbestände zu integrieren, beschränken sich dabei aber oft auf den kleinsten gemeinsamen Nenner und nivellieren datenbestandsspezifische Funktionen auf i. d. R. niedrigem Niveau. Die bloße Möglichkeit zur Integration wiederum stellt nicht sicher, dass das entstehende Produkt seine Zielgruppe erreicht und akzeptiert und genutzt wird. Was daher bislang fehlt, ist ein informationswissenschaftlich basiertes Rahmenkonzept – ein Modell, wie die Einzellösungen zu einer stimmigen Informationsarchitektur zusammengefügt werden können, die sich an den Bedürfnissen der Nutzer orientiert und die Gegebenheiten auf der Seite der Informationsanbieter berücksichtigt.

Mit dem Ziel, ein derartiges Modell u. a. am Beispiel des Wissenschaftsportals vascoda zu entwickeln, werden im Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung die im Bereich der Fachinformation vorliegenden Strukturen analysiert und auf dieser Basis ein abstraktes Modell für übergreifende, kooperative Angebotsstrukturen entwickelt. Neben eigenen Analysen und Erhebungen gehen hier auch Vorarbeiten aus dem Bereich der Virtuellen Fachbibliotheken und von vascoda ein. Neben einer Bestandsaufnahme relevanter Informationsarten und Angebotsformen sollen vor allem

[6] Siehe http://www.gesis.org/forschung/informationstechnologie/komohe.htm

Konzepte entwickelt werden, wie zum einen nutzerspezifisch komplementäre Informationsarten verbunden und zum anderen iterativ ergänzende Informationsarten in den Rechercheprozess integriert werden können. Dieses Vorgehen geht von einem Schalenmodell (siehe Krause 2003) aus, in dem Informationssammlungen unterschiedlicher inhaltlicher Relevanz und Erschließungsgüte so miteinander in Beziehung gesetzt werden, dass sie dem Nutzer auf unterschiedlichen Stufen des Rechercheprozesses adäquat zur Verfügung stehen.

Als ein essentieller Baustein innerhalb komplexer, fachübergreifender und dezentraler Portale wurden wissenschaftliche Fachportale identifiziert – nicht zuletzt aufgrund der eingangs angeführten Untersuchungsergebnisse. Das Konzept wissenschaftlicher Fachportale findet sich dabei auf einer mittleren Abstraktionsebene zwischen den singulären Informationsangeboten (z. B. Fachdatenbanken, Bibliothekskatalogen und Fachinformationsführer) und fachübergreifenden, generellen Portalen (z. B. dem Wissenschaftsportal vascoda) wieder (siehe Abb. 1). Es geht insoweit über eine Vielzahl bereits existierender Fachportale hinaus, als es den Anspruch erhebt, *alle* aus Nutzersicht relevanten Informationen eines Faches zu integrieren, wobei das Augenmerk auf der inhaltlichen und konzeptuellen Integration liegt und die rein technische Integration als notwendige Voraussetzung in den Hintergrund rückt. Bislang existierende, parallele Portale innerhalb eines Faches wären somit zu einem Fachportal zu integrieren, das dann wiederum in höher liegende Ebenen (z. B. vascoda) integriert würde.

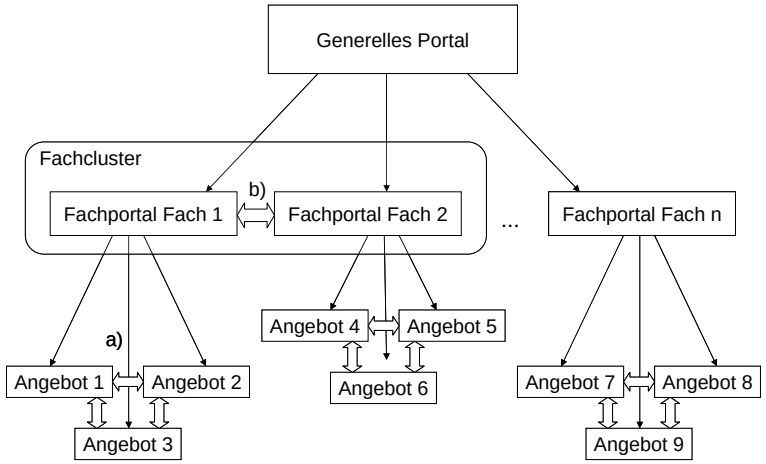

Abbildung 1: Kaskadierendes Modell einer Portalinfrastruktur

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.

Die dargestellten generellen Überlegungen zur Struktur fachübergreifender, integrierter Informationsangebote und dem Teilaspekt wissenschaftlicher Fachportale bieten einen ersten Einblick in Zielstellung und Herangehensweise des Kompetenzzentrums Modellbildung und Heterogenitätsbehandlung. Sie werden im Laufe des Projekts kontinuierlich vertieft und erweitert. Die nachfolgenden Kapitel beschäftigen sich nun mit einem konkreten Aspekt der Modellbildung, der Integration heterogen erschlossener Datenbestände.

3 Heterogenitätsbehandlung mittels Crosskonkordanzen

Die Heterogenitätsbehandlung beschäftigt sich mit konkreten Problemstellungen auf einer Abstraktionsebene unterhalb der übergeordneten Modellbildung und vor allem immer dann, wenn Standardisierungsbemühungen nicht greifen. In diesem Kapitel wird auf Crosskonkordanzen als ein

Verfahren zur Heterogenitätsbehandlung auf semantischer Ebene eingegangen und beschrieben, wie sie aufgebaut und in eine Recherche integriert werden können. Veranschaulicht wird ihr Einsatz am Beispiel der fachübergreifenden Recherche im Informationsverbund infoconnex.

3.1 Einsatz von Crosskonkordanzen zur Termtransformation

Crosskonkordanzen sind intellektuell erstellte Verbindungen zwischen zwei Thesauri oder Klassifikationen. Sie setzen die Terme eines Ausgangsvokabulars mit denen des Zielvokabulars durch Relationen in Verbindung und machen damit Aussagen über die Ähnlichkeit semantischer Konzepte beider Vokabulare. Eingesetzt werden sie als Transfermodule bei der integrierten Suche über Datenbanken, die mit unterschiedlichen Vokabularen inhaltlich erschlossen wurden. Ohne Transformation der Terme zwischen den Vokabularen würden in diesem Szenario lediglich Ergebnisse geliefert, bei denen zur Indexierung eines semantischen Konzepts der gleiche Begriff in exakt der gleichen Ansetzung verwendet wurde. Variationen in der Ansetzung (z. B. Singular vs. Plural) oder die Verwendung unterschiedlicher Begriffe würden zum Nicht-Auffinden von potentiell relevanten Ergebnissen in einzelnen Datenbanken führen.

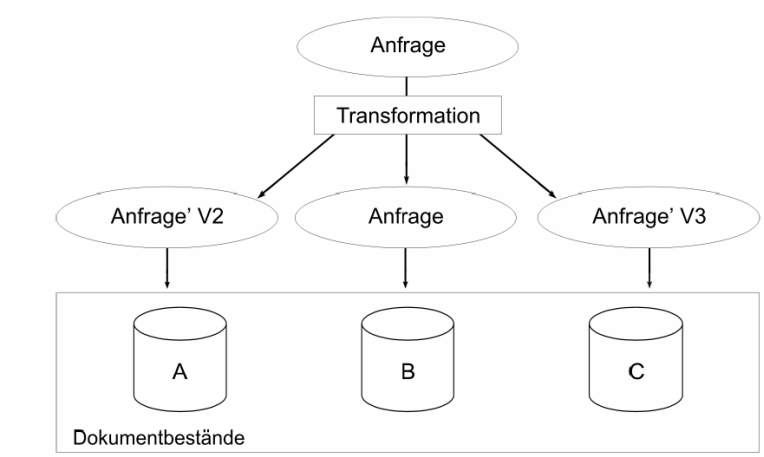

Abbildung 2: Anfrage-Transformation (vgl. Hellweg et al. 2001)

Abbildung 2 verdeutlicht schematisch die Verwendung von Crosskonkordanzen bei der Recherche über drei Dokumentbestände, die mit unterschiedlichen Vokabularen inhaltlich erschlossen sind. Die Anfrage wird zunächst durch ein Transformationsmodul bearbeitet, das Crosskonkordanzen, die jeweils bilateral für die Vokabulare der Datenbanken A, B und C erstellt wurden, enthält. Die Originalanfrage wird bei Bedarf zunächst in die jeweilige Erschließungssprache „übersetzt" und anschließend die transformierten Anfragen (in Abbildung 2 gekennzeichnet durch V2 und V3) für die Suche genutzt.

Neben dem Aspekt der Abbildung einer Anfrage auf unterschiedliche Inhaltserschließungswerkzeuge ergibt sich für den Informationssuchenden zusätzlich der Vorteil, dass er Altwissen – die Kenntnis eines Fachvokabulars – nutzen kann, um in noch unbekannten Informationssammlungen zu suchen und dadurch der Informationszugriff wesentlich erleichtert wird. Das Verfahren ist generell einsetzbar, z. B. auch für die Integration von Literatur- und Faktendaten (siehe Stempfhuber et al. 2002).

3.2 Direkte und indirekte Transfers

Sind bei einer Suche unterschiedliche Erschließungssysteme beteiligt, gibt es im Idealfall zu jeder Erschließungssprache eine Crosskonkordanz zu allen anderen Erschließungssprachen. Eine Anfrage kann somit direkt übersetzt werden (*direkter* Transfer). Dieser Idealfall liegt allerdings nicht immer vor, da es nicht immer möglich und sinnvoll ist, sämtliche Erschließungssprachen untereinander bilateral zu verknüpfen. Im Bereich der Sozialwissenschaften z. B. besteht der Wunsch, sowohl die mit Fachthesauri erschlossenen Fachdatenbanken als auch den mit der Schlagwortnormdatei (SWD) indexierten Bibliothekskatalog des Sondersammelgebiets Sozialwissenschaften gemeinsam zu durchsuchen. Der Aufwand, jeden Fachthesaurus auch mit der SWD zu verbinden, wäre nicht leistbar. Am Beispiel des

Thesaurus und der Datenbank des Deutschen Zentralinstituts für soziale Fragen (DZI)[7] lässt sich zeigen, wie eine Anfragetransformation durch *indirekten* Termtransfer, also durch die Anwendung mehrerer Crosskonkordanzen nacheinander, realisiert werden kann (siehe Abbildung 3). Im ersten Schritt erfolgt eine Transformation zum Thesaurus Sozialwissenschaften, wobei zur Sicherstellung einer möglichst hohen Präzision zunächst nur Äquivalenzrelationen zur Anwendung kommen. In einem zweiten Schritt wird danach die Crosskonkordanz zwischen Thesaurus Sozialwissenschaften und SWD verwendet, um die Abbildung auf das gewünschte Zielvokabular zu erreichen.

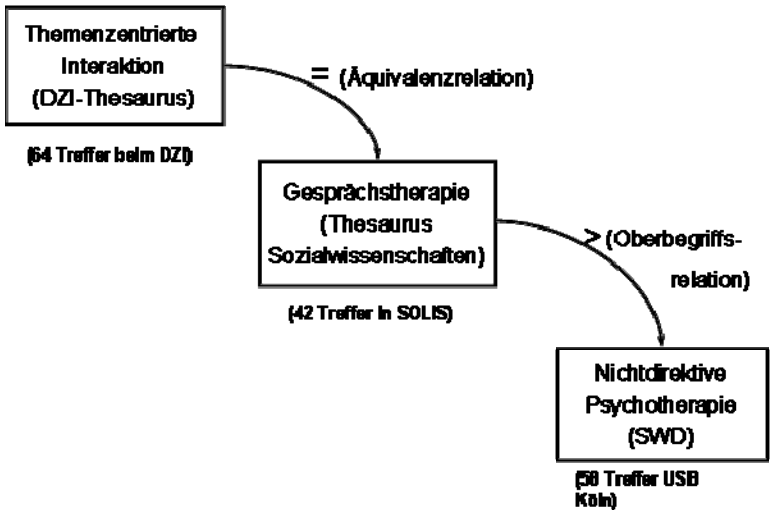

Abbildung 3: Beispiel eines indirekten Termtransfers

Da nicht für jeden Transformationsschritt Äquivalenzrelationen verwendet werden können, muss auch auf Ober- oder Unterbegriffsrelationen zurückgegriffen werden. Die Auswirkungen auf die Präzision der Gesamttransformation werden im Rahmen des Kompetenzzentrums am IZ noch eingehender geprüft.

3.3 Erstellung von Crosskonkordanzen

Eine Crosskonkordanz wird bilateral zwischen zwei Erschließungssystemen A und B erstellt (siehe Schott&Schröder 2004). Bilateral bedeutet, dass zunächst die Terme von Erschließungssystem A als Ausgangsmenge gewählt und mit Deskriptoren von Erschließungssystem B verknüpft werden. Im Anschluss daran wird die Rückrichtung erstellt. Beide Richtungen sind notwendig, da durch die erste Richtung nur ein Ausschnitt des Zielthesaurus abgedeckt wird. Fachlich relevante Terme, die durch diese Richtung nicht erreicht werden, deckt die Rückrichtung ab. Zudem liegen zwischen den Deskriptoren der einzelnen Thesauri mitunter deutliche semantische Unterschiede vor, die sich in unterschiedlichen Relationen und Relationstypen ausprägen.

Im Projekt CARMEN (siehe Strötgen 2002) wurde ein Verfahren für die Erstellung von Crosskonkordanzen vorgeschlagen, das so auch in infoconnex und im Kompetenzzentrum zum Einsatz kommt. Es definiert mehrere Möglichkeiten, Terme zweier Erschließungssysteme mit Relationen zu verknüpfen:

- **Äquivalenzrelation**: Diese Relation wird zwischen identischen, synonymen und/oder quasi-synonymen Bezeichnungen gesetzt.
  - Identische Bezeichnungen für identische Sachverhalte

sozialer Konflikt	=	Sozialer Konflikt

  - Verschiedene Benennungen für identische Sachverhalte

Legislaturperiode	=	Wahlperiode

[7] Siehe http://www.dzi.de/

- Verschiedene Benennungen für ähnliche Sachverhalte, die in einem der Thesauri gleichgesetzt sind.

konformes Verhalten	=	Konformität

- **Oberbegriffsrelation**: Eine Oberbegriffsrelation wird von einem engeren zu einem weiteren Begriff gesetzt.

Hochadel	<	Adel

- **Unterbegriffsrelation**: Eine Unterbegriffsrelation wird von einem weiteren zu einem engeren Begriff gesetzt.

Kampagne	>	Werbekampagne

- **Ähnlichkeitsrelation**: Die Ähnlichkeitsrelation wird für verwandte Begriffe gesetzt.

Konfliktverhalten	^	Konfliktfähigkeit

- **Nullrelation**: Eine Nullrelation wird gesetzt, wenn es keine sinnvolle Entsprechung des Begriffs im Zielthesaurus gibt.

Jeder Relation wird zusätzlich eine Relevanz zugeordnet, durch die eine Aussage über die Treffermenge gemacht wird, die durch die Verknüpfung zu erwarten ist. Es sind drei Stufen vorgesehen:

- **Hohe Relevanz**: der gesuchte Sachverhalt ist der oder ein wesentlicher Hauptaspekt im gefundenen Dokument
- **Mittlere Relevanz**: der gesuchte Sachverhalt ist ein nachrangiger, aber nicht marginaler Aspekt im gefundenen Dokument
- **Geringe Relevanz**: der gesuchte Sachverhalt ist kein oder nur ein marginaler Aspekt im gefundenen Dokument

4 Untersuchung zur Wirksamkeit von Crosskonkordanzen

Die Wirksamkeit von Crosskonkordanzen nach dem geschilderten Verfahren wurde bislang noch nicht systematisch evaluiert. Als erster Schritt wurden daher auf der Basis der in infoconnex erzeugten Crosskonkordanzen zwischen drei Thesauri unterschiedlicher Fächer Untersuchungen zur Steigerung der Zahl der nachgewiesenen Dokumente in der Ergebnismenge durchgeführt. Auf Basis dieser Tests soll sowohl das Verfahren zur Crosskonkordanzerstellung überprüft als auch die Methodik nachfolgender Retrieval Tests definiert werden.

4.1. Überblick über die Datenbanken und Thesauri in infoconnex

Im Informationsverbund Pädagogik – Sozialwissenschaften – Psychologie (infoconnex, siehe Stempfhuber 2003c) sind drei Fachdatenbanken integriert recherchierbar:

1. SOLIS, die sozialwissenschaftliche Fachdatenbank des Informationszentrums Sozialwissenschaften (IZ), ist mit dem Thesaurus Sozialwissenschaften erschlossen und umfasst über 300.000 Nachweise aus den Fachgebieten Soziologie, Methoden der Sozialwissenschaften, Politikwissenschaft, Sozialpolitik, Sozialpsychologie, Bildungsforschung, Kommunikationswissenschaften, Demographie, Ethnologie, Historische Sozialforschung, Arbeitsmarkt- und Berufsforschung sowie aus weiteren interdisziplinären Gebieten der Sozialwissenschaften.
2. PSYNDEX, die psychologische Fachdatenbank des Zentrums für Psychologische Information und Dokumentation (ZPID), beinhaltet ca. 185.000 Nachweise und ist mit den Psyndex Terms, der deutschen Übersetzung des Thesaurus of Psychological Index Terms, erschlossen.
3. FIS Bildung, die Datenbank des Deutschen Instituts für internationale pädagogische Forschung (DIPF) enthält mehr als 500.000 Nachweise zu empirischer Bildungsforschung, Bildungspolitik und

Bildungsverwaltung, Berufsausbildung und Berufsfortbildung, Erwachsenenbildung, Medienpädagogik, Fernstudium, Hochschulwesen, Sozial- und Sonderpädagogik, Schule und Unterricht.

4.2. Vergleich der Thesauri in infoconnex

In infoconnex kann zusätzlich zur Recherche in einer Datenbank auch eine übergreifende Suche in allen drei Datenbanken durchgeführt werden. Dabei stehen die Deskriptoren aller drei Thesauri gemeinsam zur Verfügung, wobei die überwiegende Zahl von Deskriptoren nur in der jeweiligen Fachdatenbank Treffer liefert. Dennoch gibt es zwischen den drei Fachgebieten (und damit auch zwischen den drei Thesauri) inhaltliche Überschneidungen, die dazu führen, dass mit einem Deskriptor aus einem der drei Thesauri auch in den anderen beiden Datenbanken relevante Dokumente gefunden werden können. Dies ist der Fall, wenn es sich um einen Deskriptor handelt, der in allen drei Vokabularen zeichengleich vorhanden ist. Je größer die Überlappung der Vokabulare ist, desto höher ist die Wahrscheinlichkeit, dass bei einer übergreifenden Suche auch ohne zusätzliche Maßnahmen aus mehreren Datenbanken Treffer geliefert werden.

Überlappung	TheSoz	PsyT	TheBild
TheSoz	**7.568** (1.0)	763 (0.15)	5.520 (0.10)
PsyT	763 (0.10)	**4.970** (1.0)	1.390 (0.03)
TheBild	5.520 (0.73)	1.390 (0.28)	**52.643** (1.0)

Tabelle 1: Überlappung der Vokabulare Thesaurus Sozialwissenschaften (TheSoz), Psyndex Terms (PsyT) und Thesaurus Bildung (TheBild) auf der Basis zeichengleicher Terme.

Tabelle 1 stellt Größe und Überlappung der Thesauri durch zeichengleiche Terme dar. Die Zellen auf der Diagonale zeigen die Zahl (fett gedruckt) der Deskriptoren des jeweiligen Thesaurus (Nichtdeskriptoren werden im Folgenden nicht betrachtet, da sie für die Crosskonkordanzen keine Rolle spielen). Die Tabelle ist spaltenweise zu lesen und zeigt paarweise die Überlappungen der Vokabulare als absolute Termzahl und Anteil des Ausgangsvokabulars. Der Thesaurus Sozialwissenschaften (TheSoz) umfasst momentan 7.568 Deskriptoren. Die beiden Thesauri TheSoz und die Psyndex Terms (PsyT) überschneiden sich mit 763 Deskriptoren, so dass rund 10% zeichengleicher Terme (siehe Klammerwert) der PsyT im TheSoz enthalten sind. Die PsyT enthalten 4.970 Terme, wovon 1.390 zeichengleich im Thesaurus Bildung (TheBild) vorkommen. TheBild ist somit zu 28% in den PsyT enthalten.

Auffällig sind die generell großen Überschneidungsbereiche mit dem Thesaurus Bildung. Diese betragen 73% des Vokabulars des TheSoz und 28% des Vokabulars der PsyT. Ein Grund hierfür ist die große Anzahl der Deskriptoren im Thesaurus Bildung, dadurch finden sich dort viele zeichengleiche Terme. Auch in den folgenden Aufstellungen erzeugt die große Menge an Deskriptoren auffällige Zahlen.

4.3. Vokabularerweiterung durch Crosskonkordanzen

Für alle drei Thesauri wurden im Projekt infoconnex paarweise Crosskonkordanzen erstellt. Auf ihrer Basis kann nun untersucht werden, ob und welchen Einfluss sie auf eine Ausweitung des Suchvokabulars bei der datenbankübergreifenden Suche haben. In einer ersten Phase werden in infoconnex bislang nur Äquivalenzrelationen für den Termtransfer eingesetzt, da durch sie das qualitativ hochwertigste Ergebnis geliefert wird. Die Verwendung der weiteren Relationstypen (Oberbegriffe, Unterbegriffe, Ähnlichkeiten) soll dem Nutzer später stufenweise ermöglicht werden. Vor ihrem Einsatz gilt es aber zu prüfen, in wieweit sie das Potential haben, dem Nutzer zusätzliche, potentiell relevante Treffer zu liefern. Dies soll erste Anhaltspunkte geben, welche Möglichkeiten zur Feinparametrisierung des Crosskonkordanzeinsatzes dem Nutzer an die Hand gegeben werden müssen, um ein überschaubares Ergebnis zu erhalten. Qualitative Aspekte in Form von Recall- und Precision-Messungen sind für eine darauf aufbauende, künftige Evaluation vorgesehen.

Richtung	DES	ÄQU	BT	NT	RT	%
TheSoz-PsyT	7.568	**584 (**7,7%**)**	1.084 (14,3%)	1.048 (13,9%)	1.628 (21,5%)	**57,4%**
TheSoz-TheBild	7.568	**854 (**11,3%**)**	210 (2,8%)	46 (0,6%)	142 (1,9%)	**16,5%**
PsyT-TheSoz	4.970	**301 (**6,0%**)**	1.451 (29,2%)	37 (0,7%)	1.843 (37,1%)	**73,1%**
PsyT-TheBild	4.970	**1.769 (**35,6%**)**	1.075 (21,6%)	18 (0,4%)	1.680 (33,8%)	**91,4%**
TheBild-TheSoz	52.643	**551 (**1,0%**)**	3 (0,0%)	45 (0,1%)	88 (0,2%)	**1,3%**
TheBild-PsyT	52.643	**554 (**1,0%**)**	994 (1,9%)	379 (0,7%)	832 (1,6%)	**5,3%**

Tabelle 2: Effekte der Crosskonkordanzen auf die Erweiterung von Anfragen

Die Tabelle 2 zeigt die Auswirkung der Relationsarten auf die Erweiterung einer Anfrage. Hierzu wurden sämtliche Kombinationen von Thesauri betrachtet und jeweils die Menge der Deskriptoren des Ausgangsthesaurus (DES) sowie die durch jede der vier Relationsarten (Äquivalenz = ÄQU, Oberbegriff = BT, Unterbegriff = NT, Ähnlichkeit = RT) erreichte Ausweitung der Ausgangsmenge angetragen. Die Spalte rechts enthält die durch die Relationen der Crosskonkordanzen insgesamt erreichbare Vergrößerung des Vokabulars. Beispielsweise erweitern alle Relationen der Crosskonkordanz TheSoz → PsyT das Suchvokabular des Nutzers bezüglich des TheSoz um ca. 57,4%. Hierbei ist zu beachten, dass die Anzahl der Relationen nicht unbedingt der Anzahl der Ausgangsterme entspricht, da einem Ausgangsterm auch mehrere Zielterme zugewiesen sein können (1:n-Relation).

Zum jetzigen Zeitpunkt lassen sich noch schwer allgemeingültige Aussagen über die durch Crosskonkordanzen erreichbare Ausweitung des Suchvokabulars treffen. Sie hängt stark von der Anzahl der Relationen ab, die wiederum von der Überschneidung und Größe der Thesauri abhängt. Aufgrund der stark unterschiedlichen Größe zwischen TheBild und den anderen beiden Thesauri profitieren diese sehr stark von der Crosskonkordanz, wohingegen die Erweiterung des TheBild durch die Crosskonkordanzen weitaus geringer ausfällt. Neben den Thesauri selbst, so zeigte sich, beeinflussen auch Faktoren wie die Sichtweise des Bearbeiters der Konkordanzen (Nutzerebene, Termebene usw.), unterschiedliches semantisches Begriffsverständnis oder Kenntnis der Thesauri die Anzahl und Art der Relationen. Im Schnitt befinden sich aber die Äquivalenz-, Ähnlichkeits- und Oberbegriffsrelationen ungefähr in einer Größenordnung, wogegen nur sehr wenige Unterbegriffsrelationen gebildet werden. Die einzige Ausnahme bei den hier betrachteten Crosskonkordanzen ist die Konkordanz TheBild → TheSoz, da hier nur von einem sehr begrenzten Bereich des Thesaurus Bildung aus relationiert wurde. Auch wurden bislang ausschließlich Crosskonkordanzen zwischen Thesauri unterschiedlicher Fächer untersucht. Bei Vorliegen der ersten Crosskonkordanzen innerhalb eines Faches wären die bislang gemachten Beobachtungen zu überprüfen.

4.3. Auswirkungen der Crosskonkordanzen auf die Treffermenge

Vor dem Hintergrund des Aufwands für die Crosskonkordanzerstellung stellt sich die Frage, ob durch formale Kriterien a priori Aussagen über besonders lohnende Bereiche eines Thesaurus für die Termerweiterung getroffen werden können. Abbildung 4 zeigt daher die 606 Deskriptoren aus dem Thesaurus Sozialwissenschaften, die Äquivalenzrelationen in die beiden anderen Vokabulare aufweisen, sowie deren Trefferzahlen in den drei Datenbanken SOLIS, PSYNDEX und FIS Bildung, die aus der Anwendung der Äquivalenzrelationen resultieren. Die Abbildung visualisiert somit die zusätzlichen Treffer durch ausgewählte Termtransfers, die der Recherchierende in infoconnex erhält, wenn er bei der übergreifenden Suche Deskriptoren aus dem Thesaurus Sozialwissenschaften zur Anfrageformulierung verwendet.

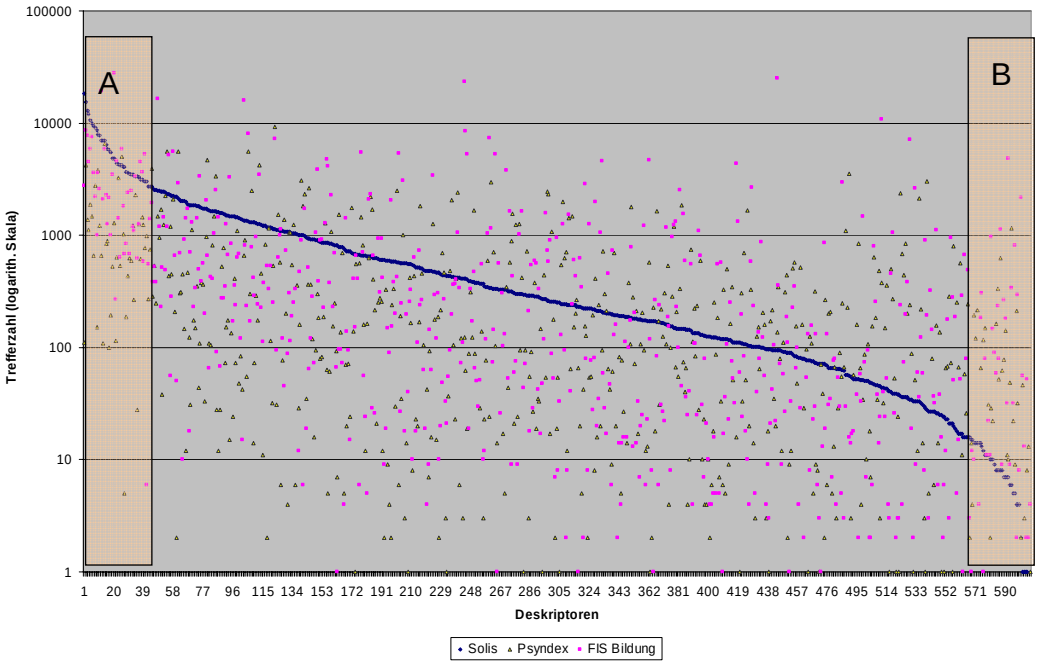

Abbildung 4: Trefferzahlen von 606 TheSoz Deskriptoren in SOLIS (absteigend sortiert nach Trefferzahl) und der über Äquivalenzrelationen in FIS Bildung und PSYNDEX erreichten Trefferzahlen.

Der Vergleich der ausgehend von einem SOLIS-Deskriptor in den anderen Datenbanken erreichbaren Trefferzahl zeigt ein uneinheitliches Bild mit großer Streuung. Trotz großer Überlappung zwischen TheSoz und TheBild, einer größeren Zahl von Äquivalenzrelationen zwischen beiden Thesauri und einer größeren Dokumentmenge in FIS Bildung gegenüber PSYNDEX kann über das ganze Spektrum der Deskriptoren hinweg kein quantitativer Vorteil der einen Crosskonkordanz gegenüber der anderen gezogen werden oder ein Bereich identifiziert werden, in dem die Crosskonkordanzerstellung zu einem spezifischen Thesaurus besonders lohnend wäre.

Für eine nähere Betrachtung wurden zwei Bereiche aus der Menge der Deskriptoren ausgewählt. Der Bereich A in Abbildung 4 umfasst die 30 Terme, die beim Indexieren am häufigsten in SOLIS vergeben wurden, also sehr verbreitet sind. Der Bereich B beinhaltet die Deskriptoren, die am seltensten in SOLIS vergeben sind. Die beiden folgenden Abbildung 5 und 6 verdeutlichen exemplarisch die Mengen an Treffern, die dem Recherchierenden zusätzlich durch die Äquivalenzrelationen der Crosskonkordanzen geliefert werden.

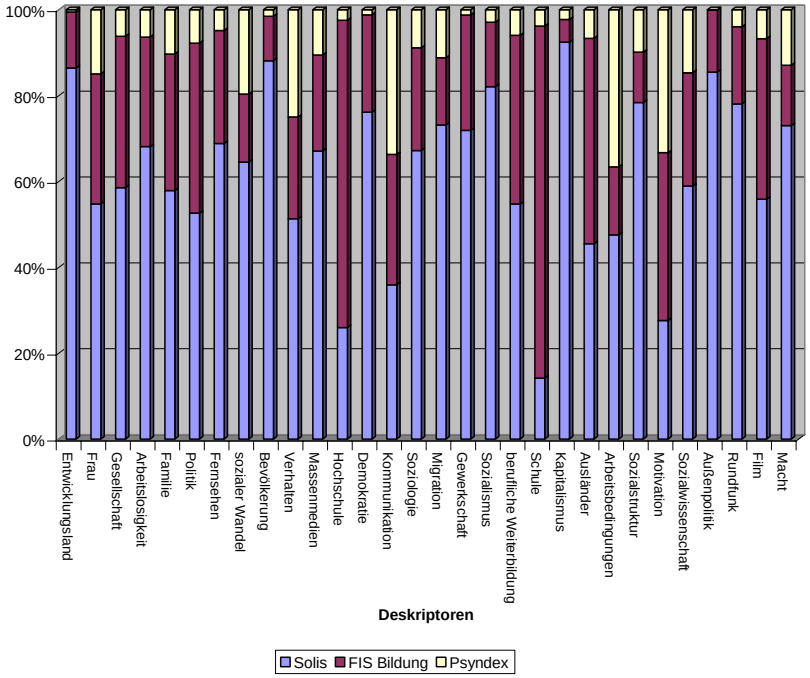

Abbildung 5: Bereich A aus Abbildung 4. Trefferzahlen für Äquivalenzrelationen der 30 häufigsten Deskriptoren aus SOLIS (absteigend sortiert nach absoluter Trefferzahl).

Die Abbildung 5 zeigt, dass für häufig zur Indexierung in SOLIS verwendete Deskriptoren der Zugewinn durch Äquivalenzrelationen bei der übergreifenden Suche relativ gering ist. Im Schnitt liefert der Ausgangsterm aus SOLIS 62,1% der Treffer, FIS Bildung trägt 27,7% und Psyndex nur noch 10,2% zusätzlich bei. Aufgrund der hohen Verwendungsfrequenz bei der Indexierung sind in diesem Bereich vorwiegend mehr allgemeine semantische Konzepte zu finden, von denen eine Entsprechung in den anderen Thesauri wahrscheinlicher wäre.

Ein genau spiegelverkehrtes Bild zeigt sich für die 30 am seltensten zur Indexierung in SOLIS verwendeten Deskriptoren. Die Abbildung 6 zeigt den Anteil dieser Deskriptoren und der Termtransformationen am Ergebnis einer datenbankübergreifenden Suche mit diesen Termen. Trotz der hohen zu erwartenden Selektivität der Begriffe existieren lohnenswerte Äquivalenzrelationen zu den anderen Thesauri, so dass SOLIS lediglich 8,9% zum Ergebnis beiträgt, wohingegen FIS Bildung 58,3% und Psyndex 32,8% liefern, so dass der Nutzer einen Großteil der Treffer aus den andere Datenbanken erhält. Der Mehrwert für den Nutzer ist klar ersichtlich.

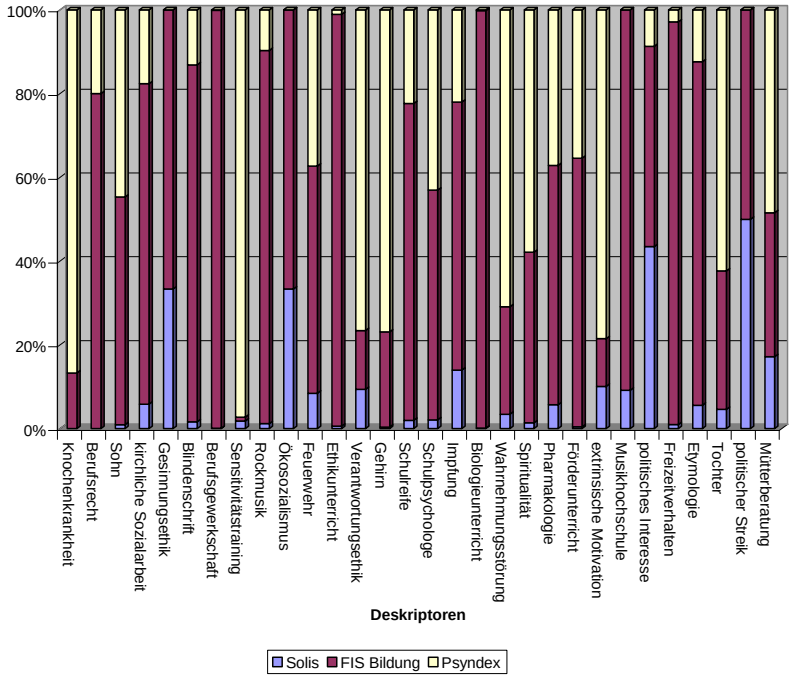

Abbildung 6: Bereich B aus Abbildung 4. Treffer für Äquivalenzrelationen der 30 seltensten Deskriptoren aus SOLIS (absteigend sortiert nach absoluter Trefferzahl).

4.4 Vorläufiges Resümee

Die Untersuchung konnte belegen, dass im interdisziplinären Bereich Crosskonkordanzen zu einer beträchtlichen Erhöhung der Treffermenge führen und dem Nutzer daher den einfachen Zugang zu Information ermöglichen, die er sonst nur – falls überhaupt – durch zusätzlichen intellektuellen Aufwand erreicht hätte.

Die bislang gemachten Erfahrungen reichen noch nicht aus, um a priori für die Crosskonkordanzerstellung besonders lohnende Bereiche in den Ausgangsthesauri zu ermitteln. Evtl. könnten sich bei Thesauri innerhalb eines Faches oder bei einer stärker inhaltlich orientierten Analyse entsprechende Hinweise ergeben.

5 Ausblick

Die ersten empirischen Ergebnisse des Einsatzes von Crosskonkordanzen bei der Suche in heterogenen Fachdatenbanken am Beispiel des Informationsverbunds infoconnex geben Erfolg versprechende Hinweise auf den Mehrwert dieses Verfahrens zur Heterogenitätsbehandlung. Als nächster Schritt werden sowohl innerhalb eines Faches als auch fachübergreifend Untersuchungen zur Auswirkungen von Crosskonkordanzen auf Recall und Precision durchgeführt, wobei auch die weiteren, in infoconnex bislang nicht verwendeten Relationen (Oberbegriffe, Unterbegriffe, Ähnlichkeiten) einbezogen werden sollen. Damit verbunden sind auch Überlegungen zur Gestaltung von Benutzungsoberflächen, die den Nutzern den fein abgestimmten Einsatz der Relationen in unterschiedlichen Phasen der Recherche ermöglichen. Erste Vorüberlegungen hierzu wurden bereits durchgeführt (siehe Stempfhuber 2003a und 2003b). Die Ergebnisse werden in die weitere Modellbildung einfließen und sollen Informationsanbietern sowohl Unterstützung beim Aufbau von Informationsdienstleistungen bieten als auch die Integration von Angeboten auf fachlicher Ebene und interdisziplinär fördern. Sicherlich kann der vorliegende Beitrag aber nur als ein erstes Argument pro Crosskonkordanzen bzw. Heterogenitätsbehandlung angesehen werden. Eine schlichte Treffererhöhung, also Steigerung des Recalls durch die Anfragetransformation

wird vom Nutzer voraussichtlich positiv aufgenommen werden, kann aber unserer Ansicht nach nur ein Ziel auf dem Weg zu wissenschaftlichen Fachportalen sein.

Literatur

Becker, Hans Jürgen; Hengel, Christel; Neuroth, Heike; Weiß, Berthold; Wessel, Carola (2002): Die Virtuelle Fachbibliothek als Schnittstelle für eine fachübergreifende Suche in den einzelnen Virtuellen Fachbibliotheken: Definition eines Metadaten-Kernsets (VLib Application Profile). Bibliotheksdienst 36. Jg. (2002). URL: http://bibliotheksdienst.zlb.de/2002/02_01_03.pdf

Boekhorst, Peter te; Kayß, Matthias; Poll, Roswitha (2003): Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. URL: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/aktuelles/download/ssg_bericht_teil_1.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 (2004): Konkretes zur These, die Standardisierung von der Heterogenität her zu denken. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie 51, Nr. 2, S. 76 - 89

Schott, Hannelore; Schröder, Albert (2004): Crosskonkordanzen von Klassifikationen und Thesauri. S. 41 - 49. In: Budin, Gerhard; Ohly, H. Peter (Hrsg.): Wissensorganisation in kooperativen Lern- und Arbeitsumgebungen: Proceedings der 8. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation, Regensburg, 9.-11. Oktober 2002. Würzburg: Ergon Verl. (Fortschritte in der Wissensorganisation; Bd. 8 (FW - 8)) ISBN 3-89913-411-7

Stempfhuber, Maximilian (2003a): Adaptive and Context-Aware Information Environments based on ODIN - Using Semantic and Task Knowledge for User Interface Adaption in Information Systems. In: Harris, Don; Duffy, Vincent; Smith, Michael; Stephanidis, Constantine (eds.): Human-Computer Interaction: Cognitive, Social and ergonomic Aspects; Volume 3 of the Proceedings of HCI International 2003, 10th International Conference on Human Computer Interaction, 22-27 June 2003, Crete, Greece, pp. 864 - 868.

Stempfhuber, Maximilian (2003b): Objektorientierte Dynamische Benutzungsoberflächen ODIN: Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie. (Dissertation) Bonn: IZ Sozialwissenschaften. 337 S. (Forschungsberichte; 6)

Stempfhuber, Maximilian (2003c): Der Informationsverbund Bildung- Sozialwissenschaften- Psychologie als Beispiel für eine dezentrale interdisziplinäre Bibliothek, (11. März). - 9. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaften in Deutschland 2003, Osnabrück, 10. - 13. März 2003 URL: http://www.iwi-iuk.org/iuk2003/program/stemp/ppt/

Stempfhuber, Maximilian; Hellweg, Heiko; Schaefer, André (2002): ELVIRA: User Friendly Retrieval of Heterogenous Data in Market Research. In: Callaos, Nagib; Harnandez-Encinas, Luis; Yetim, Fahri (eds.): SCI 2002: The 6th World Multiconference on Systemics, Cybernetics and Informatics; July 14-18, 2002, Orlando, USA; Proceedings, Vol. I: Information Systems Development I, pp. 299 - 304.

Strötgen, Robert (2002): Metadatenextraktion und Anfragetransfers: Ergebnisse der Heterogenitätsbehandlung im Projekt CARMEN AP11, (18. April). Heterogenitätsbehandlung - Schritte zum Schalenmodell, Informationszentrum Sozialwissenschaften, Bonn, 18. April 2002