Date : 17/06/2008

Crosskonkordanzen: Terminologie Mapping und deren Effektivität für das Information Retrieval

Philipp Mayr & Vivien Petras

GESIS Social Science Information Centre (GESIS-IZ), Bonn, Germany
philipp.mayr|vivien.petras@gesis.org

Meeting: **129. Classification and Indexing**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

World Library and Information Congress: 74th IFLA General Conference and Council
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

Abstract. *Das Bundesministerium für Bildung und Forschung hat eine große Initiative zur Erstellung von Crosskonkordanzen gefördert, die 2007 zu Ende geführt wurde. Die Aufgabe dieser Initiative war die Organisation, die Erstellung und das Management von Crosskonkordanzen zwischen kontrollierten Vokabularen (Thesauri, Klassifikationen, Deskriptorenlisten) in den Sozialwissenschaften und anderen Fachgebieten. 64 Crosskonkordanzen mit mehr als 500.000 Relationen wurden umgesetzt. In der Schlussphase des Projekts wurde eine umfangreiche Evaluation durchgeführt, die die Effektivität der Crosskonkordanzen in unterschiedlichen Informationssystemen testen sollte. Der Artikel berichtet über die Crosskonkordanz-Arbeit und die Evaluationsergebnisse.*

1. Einführung

Ein ambitioniertes Projekt für die One-stop Suche im wissenschaftlichen Bereich in Deutschland ist das Portal vascoda[1], das gemeinsam vom Bundesministerium für Bildung und Forschung und der Deutschen Forschungsgemeinschaft gefördert wird. Vascoda bietet eine gemeinsame Suchoberfläche für eine Vielzahl von disziplinären und interdisziplinären Datenbanken (z.B. Fachdatenbanken, Bibliothekskataloge, Inhaltsverzeichnisse, Volltexte, usw.) und Sammlungen von Internetquellen (siehe Überblick in Depping, 2007). Seit 2007 ist vascoda Partner im globalen Wissenschaftsportal WorldWideScience.org[2].

Basiskonzept des vascoda Portals ist die Strukturierung und Integration von hochqualitativen Informationsangeboten von mehr als 40 Datenlieferanten innerhalb eines Suchraums (schätzungsweise 81 Millionen Dokumente).

Der Suchraum setzt sich aus verteilten disziplinären Portalen (“Virtuelle Fachbibliotheken”) zusammen und jede der integrierten Kollektionen wird in die vascoda Fachgruppen (Ingenieur- und Naturwissenschaften, Medizin und Lebenswissenschaften, Recht, Wirtschaft

[1] http://www.vascoda.de
[2] http://worldwidescience.org/

und Sozialwissenschaften, Geisteswissenschaften, Regionen und Kulturräume und Fachübergreifende Kollektionen) eingeordnet.

Das vascoda Portal enthält viele Informationsangebote, die sehr individuell entwickelt und strukturiert sind. Diese Kollektionen besitzen ausgeklügelte Metadatenschemata zur Beschreibung und Organisation der Inhalte. Die übergreifende Suchoberfläche bietet jedoch nur eine Freitextsuche, die die ursprünglich präzisen Möglichkeiten des Fachzugangs dieser Informationskollektionen nicht berücksichtigt. Aufgrund des verteilten Aufbaus der Kollektionen und den unterschiedlichen Schemata der inhaltlichen Erschließung wird es als schwerwiegendes technisches und organisationelles Problem gesehen, alle diese heterogenen Informationsangebote mit ihren individuellen und detaillierten Fachzugängen innerhalb einer gemeinsamen Suchoberfläche zu integrieren.

Aktuell werden die großen Internet-Suchmaschinen um neue Kollektionen und erweiterte Möglichkeiten des Fachzugangs ergänzt (z.B. Clustering- und Videosuche in Flickr). Ein weiteres prominentes Beispiel sind die Semantic Web Anwendungen[3], die Deduktionsfunktionen durch den Einsatz von Ontologien und anderen semantischen Daten ableiten. Wie kann es möglich sein, dass die erweiterten Oberflächen für Digitale Bibliotheken in der Entwicklung zurückfallen, wenn moderne Web-Informationssysteme (wie z.B. aus Semantic Web Initiative resultierend) sich gerade bemühen, mehr Struktur und semantische Klarheit zu schaffen?

2004 hat das Bundesministerium für Bildung und Forschung eine größere Initiative zum Terminologie Mapping (das Projekt KoMoHe) am GESIS Informationszentrum Sozialwissenschaften in Bonn gefördert, das Ende 2007 beendet wurde. Eine Aufgabe dieser Initiative war die Organisation, die Erstellung und das Management von Crosskonkordanzen (semantische Querverbindungen) zwischen kontrollierten Vokabularen innerhalb den Sozialwissenschaften und anderen Fachgebieten. Das Hauptziel des Projekts war die Erstellung, Implementation und Evaluation eines Terminologienetzwerks sowie die Ermöglichung semantischer Integration bei der Suche nach heterogenen Quellen in einer typischen Digitalen Bibliothek.

Das Ziel der semantischen Integration ist es, unterschiedliche Informationssysteme über ihre inhaltlichen Metadaten zu verbinden und damit die verteilte Suche über mehrere Informationsangebote zu ermöglichen, ohne die erweiterten Möglichkeiten des Fachzugangs der individuellen Datenbanken zu beschränken. Durch den Überstieg zwischen verschiedenen Fachterminologien wird eine „semantische Kongruenz“ für gemeinsam recherchierbare Kollektionen erreicht. Terminologie Mapping – die Abbildung von Begriffen und Phrasen eines kontrollierten Vokabulars auf Begriffe und Phrasen eines anderen kontrollierten Vokabulars – ermöglicht den reibungslosen Überstieg von einer Ein-Datenbank-Suche in ein verteiltes Suchszenario im Sinne von Digitalen Bibliotheken.

Dieser Artikel beschreibt das Terminologie Mapping Projekt KoMoHe und die involvierten Vokabulare und Datenbanken, die Implementierung der erarbeiteten Crosskonkordanzen für die Suche und die Ergebnisse und Erkenntnisse einer ausführlichen Information Retrieval Evaluation, die Auswirkungen von Terminologie-Überstiegen für die Suche bezüglich Recall und Precision analysiert hat.

2. Semantische Heterogenität

[3] Siehe Übersicht http://www.w3.org/2001/sw/.

Grundsätzlich gibt es zwei Hauptansätze um semantische Heterogenität in Digitalen Bibliotheken zu behandeln: Intellektuelle und automatische Verfahren. Essentiell für alle Bemühungen des Terminologie Mapping ist, dass verbleibende und unveränderbare Unterschiede zwischen verschiedenen Terminologien akzeptiert werden müssen. Aufgrund von Qualitäts- und Kostengründen ist keines der oben genannten Verfahren allein für die Transferproblematik zwischen heterogenen Kollektionen verantwortlich. Nach Krause (2003) sollten sich die Verfahren gegenseitig komplementieren und integriert sein.

- Crosskonkordanzen zwischen kontrollierten Vokabularen: die verschiedenen Begriffssysteme werden im Kontext der Benutzer analysiert und es wird intellektuell versucht, die Konzeptualisierung der Begriffe miteinander in Verbindung zu bringen. Dieses Konzept darf nicht mit der Konstruktion von Metathesauri verwechselt werden. Beim Erstellen von Crosskonkordanzen wird nämlich kein Versuch unternommen die existierenden Begriffswelten zu standardisieren. Crosskonkordanzen gewährleisten nur eine partielle Vereinigung der bestehenden Terminologiesysteme. Sie überbrücken den verbleibenden statischen Teil (Vokabulare) der Transferproblematik. Diese Konkordanzen liefern meist Abbildungen (vgl. Tabelle 1 und 2) im Sinne von Synonym-, Ähnlichkeits- oder Hierarchie-Relationen, die auch als deduzierbare Regelrelationen eingesetzt werden können.
- Quantitativ-statistische Verfahren: das Transferproblem kann generell als ein Vagheitsproblem zwischen zwei Inhaltsbeschreibungssprachen modelliert werden. Um die Vagheit beim Information Retrieval zwischen Termen z.B. zwischen denen der Benutzeranfrage und denen der Dokumentkollektionen zu behandeln, sind verschiedene automatische Verfahren (probabilistische Prozeduren, Fuzzy-Verfahren und neuronale Netze) vorgeschlagen worden, die beim Termtransfer genutzt werden können (Hellweg et al., 2001). Dasselbe individuelle Dokument kann gleichzeitig von zwei Konzeptsystemen indexiert werden und somit können auf Basis der automatischen Verfahren unterschiedliche Systeme, die jeweils anders indexieren, miteinander in Verbindung gebracht werden. Diese Verfahren benötigen allerdings Trainingsdaten. Beim multilingualen IR kann derselbe Text z.B. in zwei unterschiedlichen Sprachen sein.

Wenn semantische Heterogenitätsbehandlung (z.B. Crosskonkordanzen) in einem verteilten Suchszenario implementiert ist, kann das System mit den verschiedenen Informationskollektionen mit dem Fachvokabular durchsucht werden, mit dem der Suchende vertraut ist. Terminologie Mappings können damit die verteilte Suche auf unterschiedliche Arten unterstützen. Zuerst und vor allem soll die nahtlose Suche in Datenbanken mit unterschiedlichen kontrollierten Vokabularen ermöglicht werden. Zusätzlich können die Terminologie-Überstiege als Tool für die Vokabularerweiterung dienen, da sie ein Terminologienetz aus äquivalenten, weiteren und engeren sowie verwandten Termverbindungen repräsentieren (siehe dazu die Termbeispiele in Tabelle 1 und 2). Drittens kann das Vokabularnetz mit seinen semantischen Abbildungen auch zur Expansion und Reformulierung von Anfragen genutzt werden.

Für den Fall, dass andere Informationskollektionen in die Digitale Bibliothek eingebunden sind, wird die Anfrage nicht nur in präzisen Suchstatements formuliert, sondern durch den Terminologie-Service automatisch in alle implementierten Terminologien übersetzt. Ein Recherchierender kann damit nahtlos zwischen verschiedenen Informationsressourcen wechseln, weil die semantische Übersetzung zwischen den unterschiedlichen Terminologien automatisch ausgeführt wird.

Für interdisziplinäre Informationssysteme erhöht die semantische Integration nicht nur die Erfolgschancen verteilter Suchen über Kollektionen mit unterschiedlichen Erschließungsschemata, sondern eröffnet dem Suchenden einen Einblick in die anderen disziplinären Kontexte und die domänenspezifische Sprache, vorausgesetzt die abgebildeten Vokabulare werden zur Verfügung gestellt (siehe z.B. Abbildung 1).

Semantische Abbildungen spielen zusätzlich eine große Rolle beim Anbieten von Transfermethodologien zwischen fremdsprachigen Datenbanken. So wie ein Mapping zwischen kontrollierten Vokabularen unterschiedlicher Datenbanken oder Disziplinen erstellt werden kann, können Mappings auch eine Übersetzung im klassischen Sinne bieten: zum Beispiel in Tabelle 1 von einer deutschsprachigen Terminologie zu einer englischsprachigen.

Tabelle 1 präsentiert zwei Ausgangsterme (linke Spalte) in deutschsprachigen Thesaurus Sozialwissenschaften (TheSoz) und intellektuell verbundene Endterme in den abgebildeten Vokabularen (Endvokabularen). Die Relationstypen zwischen den Termen werden genauer erklärt in Tabelle 2.

Start term TheSoz	**Relation**	**End term**	**End vocabulary**
Weiterbildung engl: "further education"	=	Weiterbildung	Psyndex, STW, Infodata, SWD, BISp, DZI
	^	Berufsfortbildung	FES
	=	Further education	CSA-ASSIA
	=	Continuing education	CSA-PEI
	=	Adult Education	CSA-SA
	<	Education	CSA-WPSA
	=	Erwachsenenbildung	IBLK
Meinungsforschung engl: "opinion research"	0		Psyndex
	^	Einstellungsforschung	IAB
	=	Opinion Polls	CSA-ASSIA
	=	Opinions + Research	CSA-SA
	<	Research	CSA-PEI
	=	Public Opinion Research	CSA-WPSA
	=	Public Opinion Polls	ELSST
	=	Meinungsumfrage/Meinungs-forschung	IBLK

Tabelle 1. Start- oder Ausgangsterm im TheSoz Vokabular und eine Auswahl von Endtermen (semantische Mappings).

In den letzten Jahren haben unterschiedliche Institutionen Bemühungen im Bereich der semantischen Integration von Informationssystemen unternommen. In den Vereinigten Staaten hat OCLC das Terminology Services[4] Projekt gestartet (Vizine-Goetz, 2004, 2006) um Web Services für Terminologie Mappings zwischen unterschiedlichen kontrollierten Vokabularen wie z.B. DDC, LCC, LCSH oder MeSH anzubieten. In Europa hat das Delos2 Network of Excellence in Digital Libraries Programm ein Arbeitspaket dem Problem von

[4] http://www.oclc.org/research/projects/termservices/

Wissensextraktion und semantischer Operabilität gewidmet (Patel et al., 2005). Ein anderer Bericht beauftragt durch JISC bietet einen Überblick über Terminologie Services mit Fokus auf Projekte in Großbritannien (Tudhope, Koch et al., 2006). Andere Projekte sind das CRISSCROSS[5] Projekt an der Deutschen Nationalbibliothek und der Fachhochschule Köln, dass ein multilinguales thesaurus-basiertes Forschungsvokabular zwischen der Schlagwortnormdatei (SWD) und den Notationen der Dewey Dezimal Klassifikation (DDC) erstellt (siehe Panzer, 2008). Die Agricultural Information Management Standards Abteilung der FAO[6] ist in unterschiedliche Terminologie Mapping Initiativen involviert (z.B. Liang & Sini, 2006). Das High-Level Thesaurus Projekt (HILT[7]) an der University of Strathclyde ist ein weiteres Beispiel eines Projekts mit langjähriger Erfahrung bei der Entwicklung von Terminologie Mapping Technologien (Macgregor et al., 2007).

3. Terminologie Mapping am GESIS-IZ

Semantische Interoperabilität kann auf unterschiedlichen Wegen erreicht werden. Einen guten Überblick über unterschiedliche Terminologie Mapping Methodologien und Projekte finden sich in Zeng & Chan (2004, 2006a, 2006b), Doerr (2001, 2004), und Hellweg et al. (2001).

Das Projekt KoMoHe fokussiert auf Crosskonkordanzen. Wir definieren Crosskonkordanzen als intellektuell und manuell erstellte Verbindungen, die Äquivalenz, Hierarchie und Verwandtschaft zwischen Termen zweier kontrollierter Vokabulare über Relationen bestimmen. Typischerweise werden die Vokabulare bilateral verbunden, d.h. eine Crosskonkordanz verbindet Terme eines Vokabulars A zu einem Vokabular B und eine weitere Crosskonkordanz verbindet Terme von Vokabular B zurück zu A. Bilaterale Relationen sind nicht notwendigerweise symmetrisch. Z.B. wird der Term ‚Computer' aus System A auf den Term ‚Information System' in System B abgebildet, aber der gleiche Term ‚Information System' in System B wird auf einen anderen Term z. B. ‚Data base' in System A zurück verbunden.

Unser Verfahren erlaubt die folgenden 1:1 oder 1:n Relationen:

- Äquivalenz (=) bedeutet Identität, Synonymie oder Quasi-Synonymie
- Hierarchie (Weitere Terme <; engere Terme >)
- Verwandtschaft (^) für ähnliche Terme
- Eine Ausnahme ist die Null (0) Relation, die bedeutet, dass ein Term nicht auf einen anderen Term abgebildet werden kann (z.B. Relation Nummer 4 in Tabelle 2).

Zusätzlich muss jede Relation mit einem Relevanz-Rating (hoch, mittel und gering) versehen werden. Die Relevanz-Ratings sind ein sekundäres und relativ schwaches Instrument um die Qualität der Relationen auszudrücken. Sie werden in unseren aktuellen Implementationen nicht ausgewertet.

Tabelle 2 präsentiert typische unidirektionale Crosskonkordanzen zwischen zwei Vokabularen A und B.

[5] http://www.d-nb.de/eng/wir/projekte/crisscross.htm
[6] http://www.fao.org/aims/
[7] http://hilt.cdlr.strath.ac.uk/

No	Vocabulary A	Relation	Vocabulary B	Description
1	hacker	=	Hacking	Equivalence relationship
2	hacker	^+	computers + crime	2 association relations (^) to term combinations (+)
3	hacker	^+	Internet + security	
4	isdn device	0		Null-relation. Concept can't be mapped, term is too specific.
5	isdn	<	telecommunications	Narrower term relationship
6	documentation system	>	abstracting services	Broader term relationship

Tabelle 2. Crosskonkordanz-Beispiele (unidirektional).

Die Crosskonkordanzen im Projekt KoMoHe involvieren die gesamten oder große Teile der Vokabulare. Bevor das Mapping der Terme beginnt, werden die Vokabulare bezüglich thematischen und syntaktischen Überlappungen untersucht. Die Mappings werden von Wissenschaftlern oder Terminologie-Experten erstellt. Es ist essentiell für ein erfolgreiches Mapping, dass die Bedeutung und Semantik der Terme und der internen Relationen der beteiligten Vokabulare vollständig verstanden werden. Das setzt syntaktische Prüfung von Wortstämmen und semantisches Verständnis bei der Suche nach Synonymen und anderen verwandten Begriffen voraus. Der Mapping-Prozess basiert auf einem Set von praktischen Regeln und Richtlinien (siehe z.B. Patel et al., 2005). Während des Mapping der Terme werden alle Intra-Thesaurusrelationen (Scope Notes eingeschlossen) verwendet. Recall und Precision der erstellten Relationen sollen in den entsprechenden Datenbanken geprüft werden. Dies ist insbesondere für Kombinationen von Termen (1:n Relationen) wichtig. 1:1 Termrelationen sollen bevorzugt werden. Wortgruppen und Relevanz-Ratings müssen konsistent umgesetzt werden. Zum Schluss wird die semantische Korrektheit der Crosskonkordanzen von Experten kontrolliert, zudem werden Stichproben empirisch auf Dokument Recall und Precision geprüft. Wenn alle Regeln berücksichtigt werden, ist es ein hochqualitativer aber kosten- und zeitintensiver Aufwand ein Terminologie-Netz alleine auf Basis von Crosskonkordanzen zu erstellen.

3.1 Ergebnisse der Mapping-Initiative

Bis heute wurden 25 kontrollierte Vokabulare aus 11 Disziplinen und 3 Sprachen (Deutsch, Englisch und Russisch) miteinander verbunden, wobei die Vokabulargrößen zwischen 1.000 - 17.000 abgebildeter Terme pro Vokabular betrugen (Abbildung 2 zeigt einen detaillierten Überblick über die Mappings). Mehr als 513.000 Relationen wurden in 64 Crosswalks erstellt (30 bilaterale[8] und 4 unidirektionale Crosskonkordanzen). Abbildung 1 stellt das erstellte Netz an Crosskonkordanzen pro Disziplin dar.

[8] Eine bilaterale Crosskonkordanz wird als zwei Crosswalks gezählt.

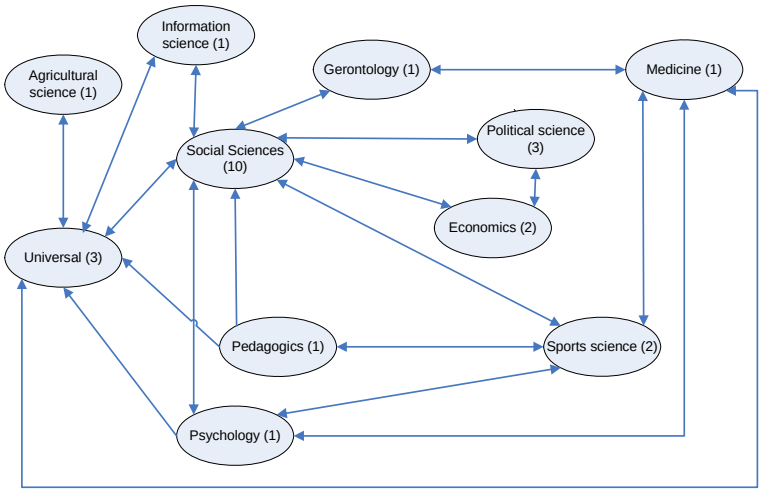

Abbildung 1.Netz der Terminologie-Mappings im KoMoHe Projekt. Die Zahlen und den Klammern enthalten die Anzahl der abgebildeten kontrollierten Vokabulare einer Disziplin.

Das Projekt hat Crosskonkordanzen zwischen kontrollierten Vokabularen (Thesauri, Deskriptorenlisten, Klassifikationen und Subject Headings) erstellt, die alle eine Rolle in den fachspezifischen Kollektionen von vascoda spielen. Mehrere Crosskonkordanzen der Vorgängerprojekte CARMEN[9] und infoconnex[10] wurden in das Netz integriert.

Die im Projekt KoMoHe verwendeten Vokabulare sind hauptsächlich in Deutsch, Englisch (N=8), Russisch (N=1), oder multilingual (z.B. AGROVOC, IBLK, DDC). Einige Vokabulare beinhalten englische oder deutsche Übersetzungen der Terme (z.B. THESOZ, PSYNDEX, MESH, INION, STW).

Involvierte Thesauri (N=16):

- AGROVOC Thesaurus (AGROVOC): A vocabulary in the *agricultural* domain which contains round 39,000 terms. Mapping to: SWD.
- CSA Thesaurus Applied Social Sciences Index and Abstracts (CSA-ASSIA): A vocabulary in the *social science* domain which contains round 17,000 terms. Mapping to: THESOZ.
- CSA Thesaurus PAIS International Subject Headings (CSA-PAIS): A vocabulary in the *political science* domain which contains round 7,000 terms. Mapping to: IBLK.
- CSA Thesaurus Physical Education Index (CSA-PEI): A vocabulary in the *sports science* domain which contains round 1,800 terms. Mapping to: THESOZ.
- CSA Thesaurus of Political Science Indexing Terms (CSA-WPSA): A vocabulary in the *social and political science* domain which contains round 3,100 terms. Mapping to: THESOZ.
- European Language Social Science Thesaurus (ELSST): A vocabulary in the *social science* domain which contains round 3,200 terms. Mapping to: THESOZ.
- INFODATA Thesaurus (INFODATA): A vocabulary in the *information science* domain which contains round 1,000 terms. Mapping to: THESOZ and SWD.
- Psyndex Terms (PSYNDEX): A vocabulary in the *psychological* domain which contains round 5,400 terms. Mapping to: THESOZ, SWD, BISP, MESH and BILDUNG.
- Standard Thesaurus Wirtschaft (STW): A vocabulary in the *economics* domain which contains round 5,700 terms. Mapping to: THESOZ, SWD, IAB and IBLK.
- Thesaurus Bildung (BILDUNG): A vocabulary in the *pedagogic* domain which contains round 50,000 terms. Mapping to: THESOZ, SWD, PSYNDEX and BISP.
- Thesaurus Internationale Beziehungen und Länderkunde (IBLK): A vocabulary in the *political science* domain which contains round 8,400 terms. Mapping to: THESOZ, STW, TWSE and CSA-PAIS.

[9] http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en
[10] http://www.infoconnex.de/

- Thesaurus Sozialwissenschaften (THESOZ): A vocabulary in the *social science* domain which contains round 7,700 terms. Mapping to: GEROLIT, DZI, FES, CSA-WPSA, CSA-ASSIA, CSA-SA, CSA-PEI, ELSST, IAB, IBLK, STW, SWD, BILDUNG, PSYNDEX, INFODATA and BISP.
- Thesaurus für wirtschaftliche und soziale Entwicklung (TWSE): A vocabulary in the *political science* domain which contains round 2,800 terms. Mapping to: IBLK.
- Thesaurus of Sociological Indexing Terms (CSA-SA): A vocabulary in the *social science* domain which contains round 4,300 terms. Mapping to THESOZ.
- Thesaurus of the Deutschen Instituts für soziale Fragen (DZI): A vocabulary in the *social science* domain which contains round 1,900 terms. Mapping to THESOZ.
- Thesaurus of the Deutschen Zentrums für Altersfragen (GEROLIT): A vocabulary in the *gerontology* domain which contains round 1,900 terms. Mapping to THESOZ and MESH.

Involvierte Deskriptorenlisten (N=4):

- Descriptors of the Bundesinstitut für Sportwissenschaft (BISP): A vocabulary in the *sports science* domain which contains round 7,400 terms. Mapping to THESOZ, MESH and BILDUNG.
- Descriptors of the Friedrich-Ebert Stiftung (FES): A vocabulary in the *social science* domain which contains round 4,000 terms. Mapping to THESOZ.
- Descriptors of the Institut für Arbeitsmarkt- und Berufsforschung (IAB): A vocabulary in the *social science* domain which contains round 6,800 terms. Mapping to THESOZ and STW.
- Descriptors of the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences (INION): A vocabulary in the *social science* domain which contains round 7,000 terms. Mapping to THESOZ.

Involvierte Klassifikationen (N=3):

- Dewey Decimal Classification (DDC): An *universal* vocabulary which contains thousands of notations. Mapping to RVK.
- Journal of Economic Literature Classification System (JEL): A vocabulary in the *economics* domain which contains round 1,000 notations. Mapping to STW.
- Regensburger Verbundklassifikation (RVK): An *univesral* vocabulary which contains thousands of notations. Mapping to DDC.

Involvierte Schlagwortnormdateien (N=2):

- Medical Subject Headings (MESH): A vocabulary in the *medicine* domain which contains round 23,000 terms. Mapping to PSYNDEX, GEROLIT, BISP and SWD.
- Schlagwortnormdatei (SWD): An *universal* vocabulary which contains round 650,000 terms. Mapping to THESOZ, MESH, STW, AGROVOC and INFODATA.

Abbildung 2 gibt einen Überblick über alle 64 Crosskonkordanzen. Der Thesaurus Sozialwissenschaften (THESOZ) ist das Vokabular mit den meisten eingehenden und ausgehenden Verbindungen und aufgrund seiner Zentralität wird der THESOZ in der Mitte des Netzes dargestellt. Andere Vokabulare wie die SWD oder der PSYNDEX spielen eine zentrale Rolle, um in andere Fachdomänen zu wechseln. Die Crosskonkordanz DDC-RVK ist die einzige Crosskonkordanz, die nicht mit dem Netz verbunden ist. Möglicherweise kann die Terminologiearbeit, die im Projekt CRISSCROSS zwischen SWD und DDC entstanden ist, genutzt werden, um dieses unverbundene Paar in das Netz einzugliedern. Die Konkordanz JEL-STW ist ein Beispiel für eine unidirektionale (one-way) Crosskonkordanz ausgehend von JEL zu STW.

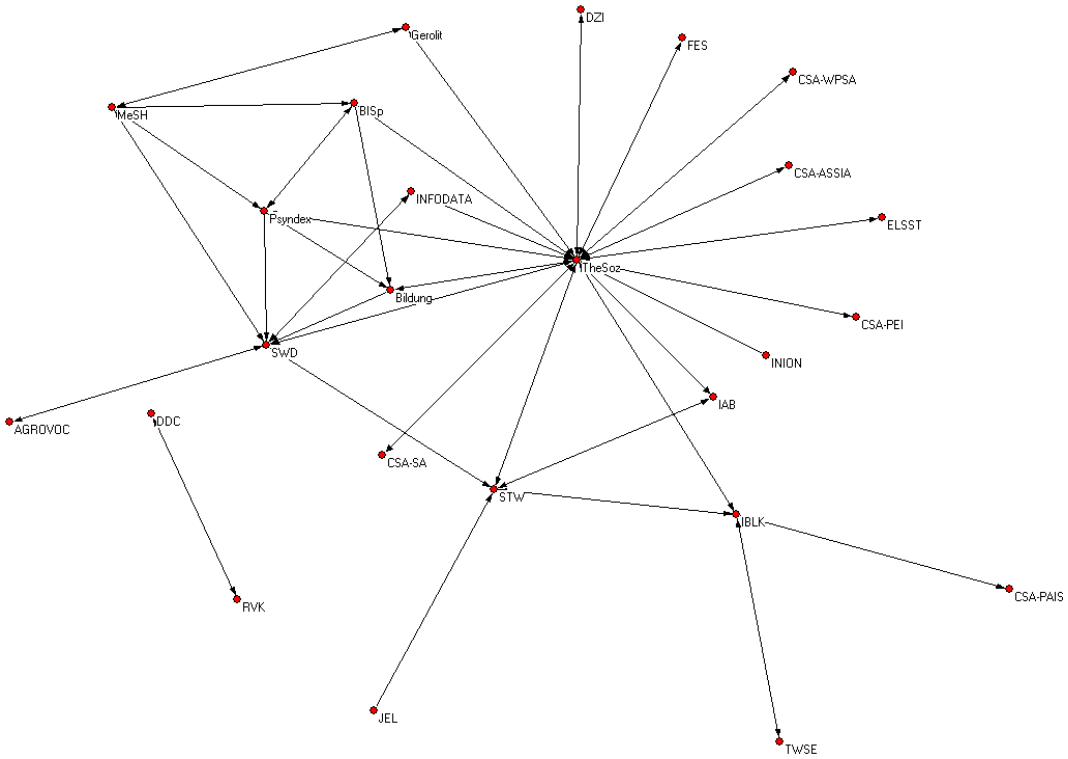

Abbildung 2. Netz der verbundenen Vokabulare im KoMoHe Projekt.

Die 513.000 Relationen, die in unserer Crosskonkordanz-Datenbank verfügbar sind, beinhalten mehr als 181.000 einzelne suchbare Konzepte (einzelne kontrollierte Deskriptoren oder Deskriptor-Kombinationen oder Notationen). Im Durchschnitt (pro Crosskonkordanz) werden 6.500 Startterme zu 3.600 Termen im Endvokabular verbunden (1,2 Relationen pro Term).

Abbildung 3 zeigt die Verteilung der Relationstypen im Projekt (vgl. Tabelle 2). Äquivalenzrelationen (ca. 45%) sind der häufigste Relationstyp zwischen Termen. Lediglich 12% von allen Relationen sind ‚Null Relationen' (ein Term kann nicht verbunden werden).

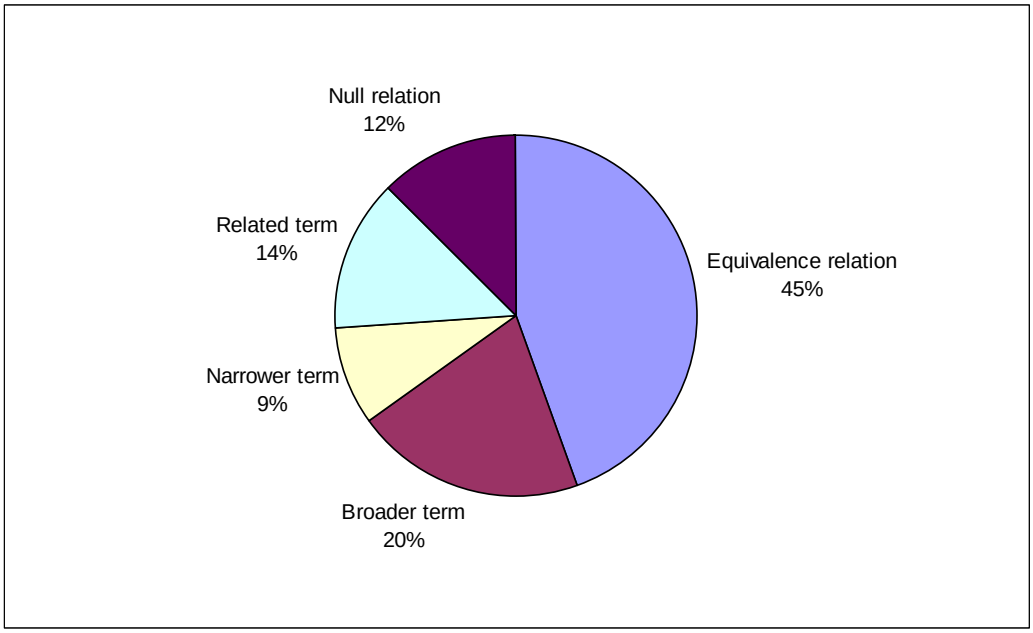

Abbildung 3. Verteilung der Relationstypen über alle Crosskonkordanzen.

3.2 Implementation der Crosskonkordanzen

Um die Crosskonkordanzen für den späteren Gebrauch zu speichern, wurde eine relationale Datenbank erstellt. Es stellte sich heraus, dass die relationale Struktur geeignet war, die Menge der unterschiedlichen kontrollierten Vokabulare, Terme, Kombinationen von Termen und Beziehungen angemessen abzubilden. Die Vokabulare und Terme werden in Listenform repräsentiert, wobei diese unabhängig voneinander und ohne Berücksichtigung der syndetischen Struktur der involvierten Vokabulare gespeichert werden. Orthographie und Groß-/Kleinschreibung der Terme der kontrollierten Vokabulare werden normalisiert. Kombinationen aus Termen (z.B. „computers + crime“ als verwandte Kombination für den Term „hacker“) werden auch als separate Konzepte gespeichert.

Um die Terminologie-Daten in der Datenbank zugänglich zu machen, wurde ein Web Service (in Abbildung 4 Heterogenitäts-Service oder HTS genannt, siehe dazu Mayr & Walter, 2008) entwickelt, der die Crosskonkordanz-Suche nach individuellen Starttermen, abgebildeten Termen, Start- und Zielvokabularen, sowie den unterschiedlichen Relationstypen unterstützt. Eine Implementation des HTS, die ausschließlich Äquivalenzrelationen nutzt, schlägt Suchterme in den Termlisten der kontrollierten Vokabularen nach und fügt alle äquivalenten Suchterme aus allen zugänglichen Vokabulare zu der ursprünglichen Suche. Wenn die kontrollierten Vokabulare in unterschiedlichen Sprachen vorliegen, bietet der HTS eine Übersetzung des Originalterms zu dem bevorzugten kontrollierten Term in der anderen Sprache. Auch wenn die ursprüngliche Anfrage einen booleschen Operator enthält, bleibt dieser nach der Anfrageerweiterung intakt (jeder Suchbegriff wird separat expandiert). Aufgrund von Performance-Gründen wird die Suchterm-Erweiterung durch Crosskonkordanzen nicht für jede einzelne Datenbank und ihr entsprechendes kontrolliertes Vokabular unterschieden, es werden aber alle äquivalenten Terme zu der Anfrage hinzugefügt. Im Prinzip erweitert die Benutzung des Terminologie-Netzes eine beliebige Term-Anfrage mit Synonymen oder Quasi-Synonymen.

4. Evaluation der Crosskonkordanzen

4.1 Allgemeine Fragestellungen

Obwohl die Notwendigkeit für Terminologie Mapping von der Community allgemein anerkannt wird und viele Mapping-Projekte durchgeführt wurden, wurde die Effektivität und Nützlichkeit der Projektergebnisse selten stringent evaluiert.
Viele Fragenstellungen ergeben sich, wenn man ein Terminologie-Netz und Mapping erstellt, z.B.:

- Wie viele Ausdrücke finden sich für ein Konzept?
- Welche Konzepte können verbunden werden?
- Sind die Vokabulare im Scope weiter oder enger?
- Welche Terminologien sind einander sehr ähnlich?
- Welche Disziplinen/Fächer grenzen aneinander oder sind weit voneinander entfernt?
- Wie sehr überlappen sich unterschiedliche Datenbanken oder kontrollierte Vokabulare in einem bestimmten Fach?

Die wichtigste Frage und gleichzeitig der Grund warum die meisten Mappings erstellt werden, ist die Frage nach der Effektivität und Nützlichkeit der Mappings in einer beliebigen Suche. In einem Informationsportal mit vielen unterschiedlichen Datenbanken ist insbesondere entscheidend, ob eine Crosskonkordanz eine verteilte Suche ermöglichen kann. Können Crosskonkordanzen z.B. die Sprachbarriere überbrücken um eine nahtlose Suche mit der gleichen Anfrage in unterschiedliche Datenbanken zu ermöglichen?

Bei der Evaluation von Terminologie Mappings ist der analytische Startpunkt entscheidend. Es ist zu klären, was untersucht wird: die Qualität der Mappings an sich oder die Qualität der anschließenden Suche? Die Qualität der Mappings ist eine Grundvoraussetzung für eine Verbesserung der Qualität der Suche. Daher wurden die Crosskonkordanzen im Projekt KoMoHe alle von den Fachexperten der Partnerinstitutionen geprüft. Die manuelle Erstellung und sorgfältige Prüfung gewährleistet, dass die Mappings sinnvoll, angemessen und von konsistenter Qualität sind.

Die intrinsischen Merkmale von Crosskonkordanzen (und ihre Auswirkung auf die Suche) können sich abhängig von den abgebildeten kontrollierten Vokabularen und externen Faktoren im Erstellungsprozess der Crosskonkordanzen unterscheiden. Zum Beispiel kann das Erstellungsdatum der Crosskonkordanz die Anzahl der Relationen pro Startterm beeinflussen. Im Projekt wurden beispielsweise zu Beginn weniger Relationen erstellt als zum Ende. Die Crosskonkordanzen des Vorgängerprojekts (CARMEN) wurden z. T. in der Expertengruppe intensiv diskutiert und deutlich selektiver erstellt als im KoMoHe Projekt. Änderungen in den kontrollierten Vokabularen oder in der Indexierungspraxis kann die Qualität der Crosskonkordanz zudem beeinflussen. Andere Unterschiede können beobachtet werden:

- Größe der Start- und Zielvokabulare
- Unterschiede im Grad der Prä- und Postkoordination der Vokabulare
- Anzahl der Relationen
- Anzahl der abgebildeten Zielterme (Abdeckung / Überlappung)
- Verteilung der Relationen (Äquivalenz, breiterer Term, engerer Term, verwandter Term, Null Relation)
- Verteilung der Relevanzen (hoch, mittel, gering)
- Identische Terme

- Unterschiede in der Spezifität (z.B. Vokabulare mit einem sehr allgemeinen oder spezifischen Scope)
- Kombinationen von abgebildeten Termen (Mappings, die aus mehr als einem Endterm bestehen).

Eine quantitative Analyse kann einige Aufschlüsse über die Grundmerkmale einer Crosskonkordanz ergeben, die Qualitätsverbesserungen durch die spezifischen Mappings für die Suche können damit aber nicht bestimmt werden. Wir haben daher einen Information Retrieval Test entworfen, der das Ziel hat, die Anwendung von Crosskonkordanzen in einem realen Suchszenario zu evaluieren.

4.2 Design des Information Retrieval Tests

Wenn die Qualität von Terminologie Mappings evaluiert werden sollen, kommen bei der Suche mehrere Faktoren zum Tragen: die Crosskonkordanzen an sich, aber auch die Inhalte der involvierten Datenbanken, deren Abdeckung oder Überlappung der Inhalte, die Suchoberfläche oder das eingesetzte Retrieval Ranking. Das Ziel war, die Auswirkungen der Crosskonkordanzen zu evaluieren und gleichzeitig die aktuellen Retrievalbedingungen (Oberfläche, Rankingmethode, usw.) soweit wie möglich stabil zuhalten.

Die Hauptidee der Nutzung der Crosskonkordanzen ist es, die Suchbegriffe in die anderen Terminologien zu übersetzen um damit die Suche über unterschiedliche Datenbanken und Terminologien zu ermöglichen. Der Einsatz der Crosskonkordanzen soll den Suchraum erweitern, Ambiguitäten und Ungenauigkeiten in der Anfrageformulierung korrigieren und folglich mehr relevante Dokumente zu einer bestimmten Anfrage finden.

Die Anwendung von Crosskonkordanzen bei der Suche stellt auch einen Mehrwert dar: sie können die Geschwindigkeit und Einfachheit des Suchprozesse beeinflussen. Eine Prämisse für die technische Implementation der Terminologie Mappings sollte sein, dass die Mappings durch den Suchenden unkompliziert genutzt werden können. Die Mappings sollten das Sucherlebnis verbessern ohne die Komplexität des Informationssystems für den Nutzer zu erhöhen. Die Nutzung der Crosskonkordanzen für die Evaluation folgte einem strikt automatischen Ansatz, d.h. es war kein manuelles Eingreifen in Form von intellektueller Anfrage-Reformulierung notwendig.

Zwei Information Retrieval Tests wurden festgelegt, um die Qualität der Crosskonkordanzen bei der Suche zu evaluieren:

Test 1: *Verbessert der Einsatz der Term-Mappings die Suche gegenüber einer nichttransformierten fachlichen (z.B. über ein kontrolliertes Vokabular) Suche?*

In Test 1 wurde eine Anfrage in Terme eines kontrollierten Vokabulars (A) übersetzt und dann in den Schlagwort-Feldern einer bibliographischen Datenbank gesucht, die mit dem kontrollierten Vokabular (B) erschlossen ist. Die Suche wurde mit Hilfe der Crosskonkordanz A→B wiederholt, wobei die Suchterme des ursprünglichen kontrollierten Vokabulars in die kontrollierten Terme der Zieldatenbank übersetzt wurden. Abbildung 4 zeigt eine graphische Repräsentation dieses Prozesses. Die Retrievalergebnisse wurden daraufhin verglichen.

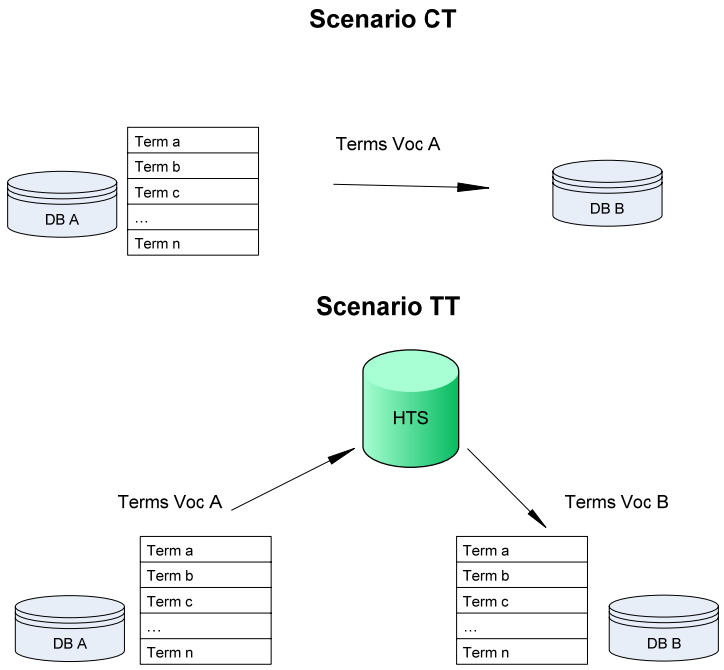

Abbildung 4. Aufbau der Information Retrieval Evaluation mit Crosskonkordanzen. Scenario CT = Suche mit kontrollierten Termen. Scenario TT = Suche mit Termtransformation. HTS = Heterogenitätsservice.

Wenn die Ergebnisse der beiden Suchen identisch sind, dann hat die Anwendung der Crosskonkordanzen für die Termtransformation keinen Effekt. Wenn sich die Suchergebnisse verschlechtern, dann haben die Crosskonkordanzen einen negativen Effekt. Wenn sich das Suchergebnis verbessert, was erwartet wird, dann hat der Einsatz der Crosskonkordanzen einen positiven Effekt auf die Suche.

Test 2: *Verbessert der Einsatz der Term-Mappings die Freitext-Suche?*

In Test 2 wurde die Originalanfrage in einem Szenario mit Freitext-Suche genutzt (die meisten Suchenden verwenden nicht die entsprechenden kontrollierten Vokabulare). Bei der Freitextsuche werden die Suchbegriffe nicht ausschließlich in den kontrollierten Vokabularfeldern gesucht, sondern zudem in den Titel- und Abstract-Feldern. In unserem Experiment wurde die Originalanfrage zuerst in den Freitext-Feldern der Datenbank gesucht. Anschließend wurden die Anfrageterme in den Crosskonkordanzen nach vorhandenen Term-Mappings geprüft. Die neuen Terme aus der Crosskonkordanz wurden zu der Originalanfrage hinzugefügt.

Ein Beispiel illustriert die Unterschiede zwischen den beiden Tests 1 und 2: eine natürlichsprachige Anfrage sucht nach Dokumenten zum Thema „family relations“ (dt. „Familienbeziehungen“). „Family relations“ ist bereits ein kontrollierter Term in dem Vokabular A und braucht deshalb nicht in einen kontrollierten Term übersetzt werden, er wird somit als Term für die erste Suche im Schlagwort-Feld der Datenbank B verwendet: *Test 1 CT: Family relations.*

Die Crosskonkordanz A→B bildet die Phrase „family relations“ aus Vokabular A auf die Termkombination „family“ AND „social relations“ in Vokabular B ab. Die zweite Suche in Datenbank B ist folglich: *Test 1 TT: Family AND social relations.*

Für den Test 2 wird die Originalanfrage in den Freitext-Feldern der Datenbank B gesucht (Titel, Abstract und kontrollierte Terme): *Test 2 FT: Family relations.*

Da die Anfrageterme in Vokabular A existieren, kann ein Term-Mapping für die Anfrage in den Crosskonkordanzen gefunden werden. Die Terme werden zu der Originalanfrage hinzugefügt und in den Freitext-Feldern der Datenbank B gesucht: *Test 2 FT+TT: Family relations OR (Family AND social relations).*

Test 1 durchsucht nur die Schlagwort-Felder wohingegen Test 2 zusätzliche andere Felder durchsucht, in denen unter Umständen die Anfrageterme vorkommen (Titel und Abstract). Test 2 ist generell ein schwächerer Test, da die abgebildeten Terme an die Anfrage gehängt werden aber nicht die Originalanfrage ersetzen wie in Test 1. Da Test 2 auch die Felder aus Test 1 durchsucht, kann über Test 2 ausgesagt werden, dass er Test 1 subsummiert. In Test 2 ergeben sich jedoch weniger Term-Erweiterungen, weil nicht alle Anfrageterme im dem Original kontrollierten Vokabular auftauchen und somit weniger Term-Mappings gefunden werden.

Um zu gewährleisten, dass realistische Anfragen verwendet werden, wurde zur Anfrageerstellung die Hilfe der Produzenten oder Anbieter der getesteten Datenbanken erbeten. Die Anbieter wurden gebeten, 3-10 Anfragen (Durchschnitt: 6-7) aus ihrer täglichen Arbeit zur Verfügung zu stellen. Diese Anfragen wurden in die kontrollierten Vokabulare der getesteten Datenbanken übersetzt. Die natürlichsprachigen Freitext-Anfragen enthielten ca. 1-3 Terme pro Anfrage, während die Booleschen Anfragen für die Schlagwort-Suche ca. 2-6 Terme enthielten. Für alle Information Retrieval Experimente wurden ausschließlich Äquivalenzrelationen der Term-Mapping verwendet. Es wurden aus allen Informationssystemen jeweils nur die ersten 1000 gelisteten Dokumente einbezogen. Zum Schluss wurden die Dokument-Ergebnismengen aus jedem Experiment bzgl. ihrer Relevanz zur Anfrage bewertet.

Um den Effekt der Crosskonkordanzen zu evaluieren, wurden die klassischen Information Retrieval Maße Recall und Precision auf Basis der Relevanzurteile der gefundenen Dokumente errechnet. Die folgenden Maßzahlen wurden analysiert:

- Retrieved: durchschnittliche Anzahl der gefundenen Dokumente (über alle Suchvarianten)
- Relevant: durchschnittliche Anzahl der relevanten gefundenen Dokumente (über alle Suchtypen)
- Rel_ret: durchschnittliche Anzahl der relevanten gefundenen Dokumente eines bestimmten Suchtyps
- Recall: Anteil der relevanten gefundenen Dokumente an allen relevanten Dokumenten (Durchschnitt über alle Anfragen eines Suchtyps)
- Precision: Anteil der relevanten gefundenen Dokumente an allen gefundenen Dokumenten (Durchschnitt über alle Anfragen eines Suchtyps)
- P10: Precision at 10 = Precision nach 10 gefundenen Dokumenten
- P20: Precision at 20 = Precision nach 20 gefundenen Dokumenten

P10 und P20 wurden berechnet, um ein realistisches Suchszenario zu repräsentieren, in dem ein Nutzer normalerweise nicht mehr als eine oder zwei Ergebnisseiten betrachtet. Für Retrievalsysteme, die nicht ranken sondern die Ergebnisse nach Jahr oder Autoren sortieren, sind P10 und P20 bedeutungslos.

4.3 Crosskonkordanzen und Datenbanken im Test

Für beide Experimente wurden die Crosskonkordanzen nach Disziplin (intra- oder interdisziplinär) und nach Sprache (ein- oder mehrsprachig) aufgeteilt. Intradisziplinäre Crosskonkordanzen im Test setzen sich hauptsächlich aus Vokabularen in den Sozialwissenschaften zusammen, da sich die meisten Crosskonkordanzen im Projekt in dieser Disziplin verorten. Die interdisziplinären Crosskonkordanzen bilden Vokabulare in den Disziplinen Wirtschaftswissenschaft, Medizin, Politikwissenschaft, Psychologie und den Sozialwissenschaften ab. Die einsprachigen Crosskonkordanzen schließen deutschsprachige Vokabulare ein; die zweisprachigen Crosskonkordanzen beinhalten deutsch- und englischsprachige Vokabulare. Tabelle 3 gibt einen Überblick über die Anzahl der getesteten Crosskonkordanzen pro Experiment:

Test 1: Controlled term search	Schlagwortsuche
Intradisciplinary cross-concordances	5 (1 bilingual)
Interdisciplinary cross-concordances	8
Test 2: Free-text search	Freitextsuche
Intradisciplinary cross-concordances	6 (1 bilingual)
Interdisciplinary cross-concordances	2

Tabelle 3. Anzahl der getesteten Crosskonkordanzen

Der Hintergrund zur Aufteilung der Crosskonkordanzen in dieser Weise liegt in der Hypothese begründet, dass Crosskonkordanzen zwischen Vokabularen in der gleichen Disziplin (intradisziplinär) sich vermutlich mehr überlappen und mehr identische Terme enthalten und folglich einen geringeren Effekt auf die Retrievalergebnisse haben als interdisziplinäre Crosskonkordanzen. Term-Mappings zwischen Vokabularen in unterschiedlichen natürlichen Sprachen (z.B. Englisch → Deutsch) oder zwischen unterschiedlichen Notationssystemen (z.B. DDC → LCC) werden vermutlich einen sehr viel größeren Effekt haben, weil das Vorkommen von identischen Termen oder eine Überlappung unwahrscheinlicher sind.
Die in den Experimenten genutzten bibliographischen Datenbanken werden hauptsächlich in Deutschland produziert und gehostet und enthalten zwischen 70.000 - 16 Millionen Dokumente. Unter den Datenbanken befinden sich Fachdatenbanken aber auch Bibliothekskataloge. Tabelle 4 gibt einen Überblick über die getesteten Datenbanken und ihre assoziierten Vokabulare:

Vokabular	**Disziplin**	**Datenbank**	**Dokumente in DB**
TheSoz – Thesaurus Sozialwissenschaften (GESIS-IZ)	Social Sciences	SOLIS	345.086
DZI – Thesaurus des Deutschen Instituts für soziale Fragen	Social Sciences	SoLit	151.925
SWD – Schlagwortnormdatei	General (Social Sciences Excerpt)	USB Köln Sowi OPAC	72.729
CSA – Thesaurus of Sociological Indexing Terms (Cambridge Scientific Abstracts)	Social Sciences	CSA Sociological Abstracts	294.875
Psyndex - Psyndex Terms	Psychology	Psyndex (ZPID)	Ca. 200.000
STW – Standard Thesaurus Wirtschaft	Economics	Econis (ZBW Kiel)	Ca. 3.000.000
IBLK - Thesaurus Internationale Beziehungen und Länderkunde (Euro-Thesaurus)	Political Science	World Affairs Online WAO (SWP Berlin)	643.420
Mesh – Medical Subject Headings	Medicine	Medline (Dimdi)	Ca. 16.800.000

Tabelle 4. Vokabulare und Datenbanken im KoMoHe IR Test

Viele Crosskonkordanzen und ihre zugehörigen Datenbanken konnten in-house getestet werden, indem die Dokumente mit dem Open-Source Information Retrievalsystem Solr[11] indexiert wurden. Es wurden die gleichen Processing und Ranking-Module für alle Datenbanken verwendet. Für die Datenbanken, die nicht in-house verfügbar waren, haben wir die Hosts gebeten, gerankte Ergebnislisten für die vorbereiteten Anfragen zu liefern.

Für die meisten Datenbanken wurden die Term-Mappings in beide Richtungen getestet, ausgehend von Vokabular A nach B (A → B) und umgekehrt von Vokabular B nach A (B → A). Da die Crosskonkordanz-Richtungen unterschiedliche Suchen (unterschiedliche Anfragen, die vom jeweiligen Startvokabular abhängen) und Datenbanken involvieren, sind sie unabhängig voneinander zu betrachten.

5. Ergebnisse der Evaluation

5.1 Test 1: Suche mit kontrollierten Termen

Test 1 evaluierte, ob die Ersetzung einer Anfrage aus Termen des Vokabulars A (CT) durch kontrollierte Terme aus dem Vokabular B (Transformationen durch Term-Mappings) die Suche in Datenbank B verbessern kann. Wenn das Term-Mapping unpräzise oder doppeldeutig ist oder sich die Vokabulare überlappen, dann kann die Übersetzung der Originalanfrage in die abgebildete Anfrage „Rauschen“ in die Anfrageformulierung bringen, was sich negativ auf die Qualität der Suche auswirken kann.

Tabelle 5 zeigt einen Überblick über die gemittelten Ergebnisse aller 13 getesteten Crosskonkordanzen. Die letzte Zeile zeigt die Unterschiede zwischen den Suchtypen in Prozentpunkten:

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	156.5	144.8	42.0	0.3152	0.2214	0.1987	0.1748
TT	325.4	144.8	88.2	0.6047	0.3391	0.3052	0.2848
				91.8%	**53.2%**	**53.6%**	**62.9%**

Tabelle 5. Ergebnisse der Test 1 Evaluation für alle Crosskonkordanzen (N=13)

Die Suche, die die Termtransformation nutzt, verdoppelt die Anzahl der gefundenen Dokumente, d.h. mehr Dokumente, die die Anfrageterme enthalten werden gefunden. Der Recall verbessert sich um fast 100% während sich die Precision um mehr als 50% verbessert. Der Einsatz einer Crosskonkordanz in dieser bestimmten Suche findet nicht nur mehr relevante Dokumente (Recall) sondern ist zudem noch präziser (Precision) als eine Suche ohne Termtransformation.

Diese enorme Verbesserung geht jedoch teilweise auf die Übersetzung zwischen Englisch und Deutsch in den bilingualen Crosskonkordanzen zurück. Monolinguale Term-Mappings dürften teilweise ineffektiv sein, weil die abgebildeten Terme identisch sind, was bei übersetzten Mappings nicht der Fall ist. Tabelle 6 zeigt die Retrievalergebnisse wenn die bilingualen Crosskonkordanzen aus dem Test entfernt werden:

[11] http://lucene.apache.org/solr/

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	169.6	141.2	45.5	0.3415	0.2399	0.2153	0.1894
TT	320.5	141.2	87.6	0.6113	0.3431	0.3126	0.2877
				79.0%	**43.1%**	**45.2%**	**51.9%**

Tabelle 6. Ergebnisse der Test 1 Evaluation für alle monolingualen Crosskonkordanzen (N=12)

Aufgrund der Termüberlappung sollten sich die Retrievalergebnisse für die Crosskonkordanzen zwischen zwei Disziplinen (interdisziplinär) bzw. Crosskonkordanzen innerhalb der gleichen Disziplin (intradisziplinär) unterscheiden. Wenn die Testergebnisse nach Disziplinen aufgeteilt werden, können wir signifikante Unterschiede in den Retrievalergebnissen erkennen. Der Recall und die Precision bei den intradisziplinären Crosskonkordanzen steigen ebenfalls, allerdings weniger stark. Eine kleinere oder negative Änderung der Precision ist aber erwartungsgemäß, da normalerweise beim Information Retrieval Precision und Recall invers miteinander verbunden sind (wenn der Recall steigt, sinkt die Precision).

Tabelle 7 zeigt die durchschnittlichen Precision- und Recall-Werte für alle und nur einsprachige intradisziplinäre Crosskonkordanzen. Für einsprachige intradisziplinäre Crosskonkordanzen steigen Precision und Recall immer noch an, aber sehr viel weniger im Vergleich zur Gesamtheit der Crosskonkordanzen.

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	126.6	101.3	36.2	0.3726	0.2491	0.2002	0.1637
TT	238.9	101.3	59.9	0.5189	0.3335	0.2784	0.2352
				39.3%	**33.9%**	**39.1%**	**43.7%**
Monolingual							
CT	158.2	79.7	45.3	0.4657	0.3113	0.2503	0.2046
TT	202.9	79.7	51.0	0.5174	0.3441	0.2939	0.2315
				11.1%	**10.5%**	**17.4%**	**13.2%**

Tabelle 7. Ergebnisse der Test 1 Evaluation für intradisziplinäre Crosskonkordanzen (N=5)

Eine außergewöhnliche Verbesserung von Recall und Precision kann bei der Anwendung der interdisziplinären Crosskonkordanzen beobachtet werden. Recall und Precision steigen signifikant mehr an als der Durchschnitt der Crosskonkordanzen (siehe Tabelle 8):

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	175.2	171.9	45.6	0.2794	0.2041	0.1978	0.1817
TT	379.4	171.9	105.9	0.6583	0.3426	0.3220	0.3157
				135.6%	**67.8%**	**62.8%**	**73.7%**

Tabelle 8. Ergebnisse der Test 1 Evaluation für interdisziplinäre Crosskonkordanzen (N=8)

Der Einsatz der Crosskonkordanzen hat mehr als einen positiven Effekt auf die Suche mit kontrollierten Termen. Die Ergebnismenge ist nicht nur größer sondern auch präziser. Der größte Effekt kann für die Crosskonkordanzen beobachtet werden, die mehr als eine Disziplin umfassen.

5.2 Test 2: Freitext-Suche

Test 2 evaluierte, ob das Hinzufügen von kontrollierten Termen, die aus dem Mapping natürlichsprachiger Anfrageterme zu dem kontrollierten Vokabular einer Datenbank (FT-CK) gewonnen wurden, zu einer Freitext-Anfrage (FT) die Retrievalergebnisse verbessern kann. Einige der individuellen Anfragen in den Tests konnten nicht erweitert werden, weil keine passenden kontrollierten Terme gefunden werden konnten. Tabelle 9 zeigt die Retrievalergebnisse für alle 8 getesteten Crosskonkordanzen:

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	155.3	106.4	56.2	0.6026	0.4551	0.4101	0.3682
FT-CK	266.8	106.4	72.8	0.7273	0.3934	0.3203	0.3083
				20.7%	**-13.6%**	**-21.9%**	**-16.3%**

Tabelle 9. Ergebnisse der Test 2 Evaluation für alle Crosskonkordanzen (N=8)

Die Ergebnisse zeigen, dass nicht nur mehr sondern auch mehr relevante Dokumente gefunden wurden. Der durchschnittliche Recall erhöht sich nach wie vor um 20%. Kontrollierte Terme, die einfach zu einer Anfrage hinzugefügt werden, können ganz generell die Retrievalergebnisse verbessern. Es wurde jedoch ein Abfall der Precision beobachtet, der trotzdem nicht so groß ist wie der Anstieg des Recall.

Tabelle 10 zeigt die Retrievalergebnisse für Crosskonkordanzen, die Terme in der gleichen Disziplin abbilden, während Tabelle 11 die Ergebnisse von 2 interdisziplinären Crosskonkordanzen zeigt:

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	163.8	115.8	60.2	0.5934	0.5025	0.4635	0.4090
FT-CK	244.9	115.8	77.1	0.7096	0.4449	0.3826	0.3681
				19.6%	**-11.5%**	**-17.5%**	**-10.0%**

Tabelle 10. Ergebnisse der Test 2 Evaluation für intradisziplinäre Crosskonkordanzen (N=6)

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	129.9	78.2	44.3	0.6303	0.3129	0.2500	0.2459
FT-CK	332.4	78.2	59.8	0.7805	0.2388	0.1333	0.1292
				23.8%	**-23.7%**	**-46.7%**	**-47.5%**

Tabelle 11. Ergebnisse der Test 2 Evaluation für interdisziplinäre Crosskonkordanzen (N=2)

Im Gegensatz zu der Analyse in Test 1 sind die Unterschiede zwischen intradisziplinären und interdisziplinären Crosskonkordanzen nicht so groß.

Bei beiden Sets erhöht sich der Recall gegenüber der einfachen Freitext-Suche während die Precision absinkt. Der Recall erhöht sich leicht für die interdisziplinären Crosskonkordanzen während die Precision stärker abfällt. Dies kann aber auch daran liegen, dass nur 2 interdisziplinäre Crosskonkordanzen evaluiert wurden, was die Ergebnisse verzerren und sicher nicht als Trend gewertet werden kann.

Die Ergebnisse der Experimente mit der Freitext-Suche könnten wahrscheinlich sehr verbessert werden, wenn die kontrollierten Terme und die natürlichsprachigen Terme besser bei der Anfrageformulierung integriert werden würden, anstatt sie nur aneinander zufügen. Eine Möglichkeit wäre die natürlichsprachigen Terme (automatisch) a priori in kontrollierte

Terme zu übersetzen, damit eine größere Chance besteht, unterschiedliche kontrollierte Vokabulare und Datenbanken aufeinander abzubilden (siehe Ansatz in Mayr et al., 2008).

Zusammenfassend kann man sagen, dass die Information Retrieval Experimente einen positiven Effekt des Einsatzes von Crosskonkordanzen bei heterogenen Datenbanken zeigen. Die Retrievalergebnisse verbessern sich für alle Crosskonkordanzen, es zeigt sich jedoch, dass interdisziplinäre Crosskonkordanzen einen deutlich stärkeren (positiven) Effekt auf die Suchergebnisse haben. Für alle Crosskonkordanzen in beiden Testszenarios wurden, verglichen mit den Anfragetypen ohne Nutzung der Crosskonkordanzen, mehr relevante Dokumente gefunden; in bestimmten Fällen waren die Suchergebnisse sogar noch präziser (Anstieg der Precision).

6. Schlussfolgerung und Ausblick

Nachdem wir gezeigt haben, dass der Einsatz von Crosskonkordanzen bei der Suche einen positiven Effekt auf die Suchergebnisse haben kann, planen wir alle Crosskonkordanzen im interdisziplinären Portal vascoda zu implementieren. Im deutschen sozialwissenschaftlichen Portal sowiport[12], das bibliographische und andere Informationsressourcen (insgesamt 15 Datenbanken mit 10 unterschiedlichen Vokabularen und ca. 2,5 Million bibliographischen Referenzen) anbietet, nutzen wir bereits viele der Crosskonkordanzen innerhalb der Suche.

Die Implementation in sowiport nutzt – wie in den hier beschriebenen Experimenten – nur die Äquivalenzrelationen, indem die Suchterme in den kontrollierten Vokabularen nachgeschlagen werden und anschließend automatisch alle äquivalenten Terme aller verfügbaren abgebildeten Vokabulare zu der Anfrage hinzugefügt werden. Dies ist ähnlich der Methodologie, die in Experiment 2 (Freitext-Suche) angewendet wurde. Boolesche Operatoren fungieren als Trenner zwischen den Anfragephrasen, das bedeutet auch, dass die Operatoren nach der Anfrageerweiterung intakt bleiben (z.B. jeder Anfrageteil wird separat expandiert). Das Term-Mapping wird automatisch ausgeführt und bleibt weitestgehend unsichtbar für den Suchenden, ein kleines Icon symbolisiert die erfolgte Transformation (per Click auf das Icon werden die angefügten Anfrageterme aufgelistet).

Für die weitere Forschung und Speicherung wurde eine relationale Datenbank konzipiert, die die Crosskonkordanzen zur Verfügung stellt. Um die Terminologie-Daten in der Datenbank zu finden, wurde ein Web Service (Heterogenitäts-Service, siehe Mayr & Walter, 2008) entwickelt, der die Suche in den Crosskonkordanzen nach individuellen Starttermen, abgebildeten Termen, Start- und Zielvokabularen und unterschiedlichen Relationstypen unterstützt. Die Datenbank kann zudem auch direkt abgefragt werden. Die Terminologie Mappingdaten sowie der Web Service können für Forschungszwecke zugänglich gemacht werden. Einige der Mappings werden bereits im Domain-specific Track der CLEF (Cross-Language Evaluation Forum) Retrievalkonferenz (Petras, Baerisch & Stempfhuber, 2007) genutzt. Weitere Features und Anwendungen der Crosskonkordanzen wie z.B. Switching, Interaktion oder Manipulation werden in späteren Forschungsarbeiten exploriert.

Eine weitere Option für die Speicherung und Anfragebearbeitung ist der Semantic Web basierte SKOS Standard (Simple Knowledge Organization System)[13]. Ziel des SKOS Standards, der aktuell als W3C Working draft vorliegt, ist es, einen Standard zu formulieren, der den Gebrauch von kontrollierten Vokabularen zur weiteren Implementierung in Semantic

[12] http://www.sowiport.de/
[13] http://www.w3.org/2004/02/skos/

Web Anwendungen unterstützt. Der Draft enthält zudem einen Abschnitt, der das Mapping von Vokabularen beschreibt. Wenn sich der SKOS Standard durchsetzt, werden wir unsere Mappingdaten in diesem Format zur Verfügung stellen.

Ein interessanter Bereich ist die Erstellung von Mappings auf Basis eines Pivot-Vokabulars für Ressourcen, für die keine direkten Mappings zur Verfügung stehen. Wenn beispielsweise Vokabular A zu Vokabular B gemappt ist und B zu Vokabular C gemappt ist, dann könnte es möglich sein, ein Mapping A→C zu erstellen, indem man die Mappinginformationen des Pivotvokabulars B nutzt. Wir hoffen, dass mit der Entwicklung eines Standards für die Präsentation und den Austausch mehr Mappings und Vokabulare für die weitere Forschung zur Verfügung gestellt werden.

Danksagung

Das Projekt wurde vom BMBF unter der Kennziffer 01C5953 gefördert. Wir bedanken uns bei allen unseren Projektpartnern für die Zusammenarbeit. Wir sind besonders dankbar für die Unterstützung unserer Kollegen Anne-Kathrin Walter und Stefan Baerisch, die das meiste der technischen Infrastruktur implementiert haben und uns bei der Evaluation unterstützt haben.

Literaturverzeichnis

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level. D-Lib Magazine, 12(6).

Depping, R. (2007). vascoda.de and the system of the German virtual subject libraries. In A. R. D. Prasad & D. P. Madalli (Eds.), International Conference on Semantic Web & Digital Libraries (ICSD 2007) (pp. 304-314). Bangalore, India: Documentation Research & Training Centre, Indian Statistical Institute.

Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1(8).

Doerr, M. (2004). Semantic interoperability: Theoretical Considerations.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., et al. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften.

Krause, J. (2003). Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. Paper presented at the IFLA 2003, World Library and Information Congress: 69th IFLA General Conference and Council, Berlin.

Liang, A. C., & Sini, M. (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review in Hypermedia and Multimedia, 12(1), 51-62.

Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. Bangalore, India.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed Digital Libraries: treatment of term vagueness and document re-ranking. Library Review, 57(3), 213-224. URL: http://www.emeraldinsight.com/10.1108/00242530810865484

Mayr, P., & Walter, A.-K. (2008). Mapping Knowledge Organization Systems. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte der Wissenorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 80-95). Würzburg: Ergon.

Panzer, M. (2008). Semantische Integration heterogener und unterschiedlichsprachiger Wissensorganisationssysteme: CrissCross und jenseits. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.),

Fortschritte in der Wissensorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 61-69). Würzburg: Ergon.

Patel, M., Koch, T., Doerr, M., & Tsinaraki, C. (2005). Semantic Interoperability in Digital Library Systems.

Petras, V., Baerisch, S., & Stempfhuber, M. (2007). The Domain-Specific Track at CLEF 2007, Cross Language Evaluation Forum Workshop (CLEF) 2007. Budapest.

Tudhope, D., Koch, T., & Heery, R. (2006). Terminology Services and Technology: JISC state of the art review.

Vizine-Goetz, D., Hickey, C., Houghton, A., & Thompsen, R. (2004). Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4).

Vizine-Goetz, D., Houghton, A., & Childress, E. (2006). Web Services for Controlled Vocabularies. ASIS&T Bulletin, 2006 (June/July).

Zeng, M. L., & Chan, L. M. (2004). Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. Journal of the American Society for Information Science and Technology, 55(3), 377-395.

Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. D-Lib Magazine, 12(6).