

Rainer Kuhlen (Hrsg.)

Information: Droge, Ware oder Commons?

**Wertschöpfungs- und Transformations-
prozesse auf den Informationsmärkten**

**Proceedings des 11. Internationalen Symposiums
für Informationswissenschaft (ISI 2009)**

Konstanz, 1.–3. April 2009

vwh

Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Informationstechnologische Aspekte der Heterogenitätsbehandlung in Fachportalen

Stefan Baerisch, Peter Mutschke, Maximilian Stempfhuber

GESIS

Lennéstr. 30, 53113 Bonn

E-Mail: {stefan.baerisch | peter.mutschke | max.stempfhuber}@gesis.org

Zusammenfassung

Ein zentrales Problem bei der Integration von Informationsbeständen in Digitalen Bibliotheken ist die Behandlung der semantischen Heterogenität, also von Unterschieden auf der Ebene der Inhaltserschließung. Bei Konzeption und Implementierung eines Moduls zur Behandlung dieser Heterogenität mittels Termtransformationsverfahren gibt es eine Reihe von gegeneinander abzuwägenden Alternativen. Anhand des Wegs von der Konzeptentwicklung für die Heterogenitätsbehandlung in Digitalen Bibliotheken bis zur dienstorientierten Implementierung stellt dieser Beitrag entscheidungsrelevante Kriterien für die informationstechnologische Implementierung von Termtransformationsverfahren vor. Mehrere Fachportale dienen als Beispiele für die Spannbreite möglicher Implementierungen, von einem aufgabenspezifischen, eng gekoppelten System bis hin zu einem dienstorientierten, generisch nachnutzbaren Ansatz.

1 Integration und semantische Heterogenität

Neben der strukturellen Heterogenität, also den Unterschieden im formalen Aufbau von Dokumentsurrogaten (Metadaten) im weitesten Sinne, ist die semantische Heterogenität ein grundsätzliches Problem bei der Recherche in integrierten Informationsbeständen (McCulloch et al. 2005). Diese seman-

tische Heterogenität entsteht durch die Verwendung unterschiedlicher Schlagwortsysteme, die in den Surrogaten zur Repräsentation des Inhalts des Originaldokuments verwendet werden. Sowohl bezüglich des Konzepts eines „Dokuments“ als auch bezüglich der zur Verschlagwortung der Dokumente verwendeten Vokabulare existieren in der Praxis vielfältige parallele und komplementäre Lösungen, die sich an den Informationsbedürfnissen und Arbeitsprozessen der Nutzerinnen und Nutzer genauso orientieren wie am Sprachgebrauch eines Faches oder einer Nutzergruppe. Die im wissenschaftlichen Kontext relevanten Dokumentarten umfassen neben Publikationen auch Primär- und Sekundärdaten (z. B. sozialwissenschaftliche Studien, Messdaten, prozessproduzierte Daten aus statistischen Ämtern und Zeitreihen) sowie im zunehmenden Umfang auch multimediale Daten (Bilder, Video- und Tonaufzeichnungen etc.). Zu ihrer inhaltlichen Beschreibung finden kontrollierte Vokabulare wie allgemeinere Erschließungsvokabulare (z. B. der Schlagwortnormdatei der Deutschen Nationalbibliothek), spezielle Fachthesauri (z. B. Medical Subject Headings (MeSH), der Thesaurus Sozialwissenschaften oder der Standardthesaurus Wirtschaft (STW)) und Nomenklaturen (z. B. für die von internationalen statistischen Ämtern erhobenen Daten) genauso Anwendung wie frei vergebene Stichwörter von Autoren oder Nutzern (social tagging).

Das daraus – für Nutzer wie für Informationsanbieter – entstehende Problem fällt allgemein in den Bereich der Vagheit, speziell in den Bereich der Vagheit bei der Abbildung zwischen Begriffssystemen. Zusätzlich zur traditionell im Information Retrieval betrachteten Vagheit beim Abgleich zwischen dem Informationsbedürfnis eines Nutzers mit der Dokumentbasis (genauer: der Formalisierung und Verbalisierung eines oft unscharfen Informationswunsches mittels dem für die Inhaltsrepräsentation verwendeten endlichen, oft durch Fachtermini bestimmten Vokabulars), im Folgenden Term-to-Concept-Vagheit (V_{tc}) genannt, bestehen beim Retrieval in heterogenen Informationssystemen noch weitere Vagheitsbeziehungen. Sie treten immer dann auf, wenn das Informationsbedürfnis nicht nur mittels eines einzigen kontrollierten Vokabulars ausgedrückt werden muss (z. B. einem Fachthesaurus), sondern dieser Transfer wiederholt auf zusätzliche, unterschiedliche Zielvokabulare geleistet werden muss (z. B. auf einen allgemeinen Bibliotheksthesaurus, auf einen fremdsprachigen Fachthesaurus und auf eine Nomenklatur aus der amtlichen Statistik). Diese im Folgenden als Concept-to-Concept-Vagheit (V_{cc}) bezeichnete Unschärfe stellt (genauso wie die Nutzung unterschiedlicher Rechercheoberflächen und formaler Anfragespra-

chen) aus Nutzersicht eine zusätzliche Belastung dar – sofern er sie selbst durchführen muss. Für das in Abbildung 1 dargestellte Szenario eines integrierten Informationssystems sind also Verfahren zur Heterogenitätsbehandlung zu entwickeln, bei denen trotz der durch die Heterogenität erhöhten Komplexität der Gesamtnutzen für den Anwender höher ist als im Vergleich zur sequenziellen Suche in mehreren Informationssystemen (und der damit jeweils durchzuführenden Neuformulierung der Anfrage).

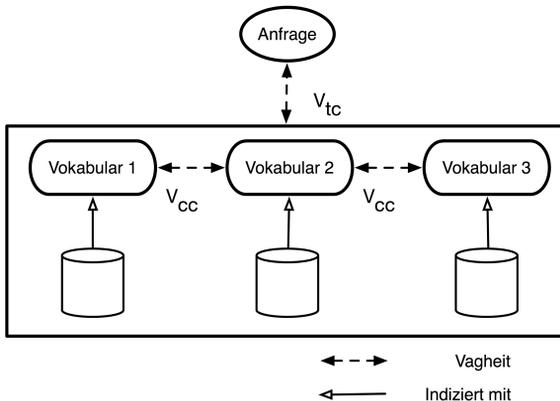


Abbildung 1. Vagheit V_{tc} und Vagheit V_{cc} (nach Hellweg et al. 2001)

Die Behandlung von Heterogenität in Informationssystemen war Gegenstand einer Reihe seit 1995 bei der GESIS durchgeführter Forschungsprojekte.¹ Die Abbildung 2 zeigt die thematische Entwicklung im Überblick.

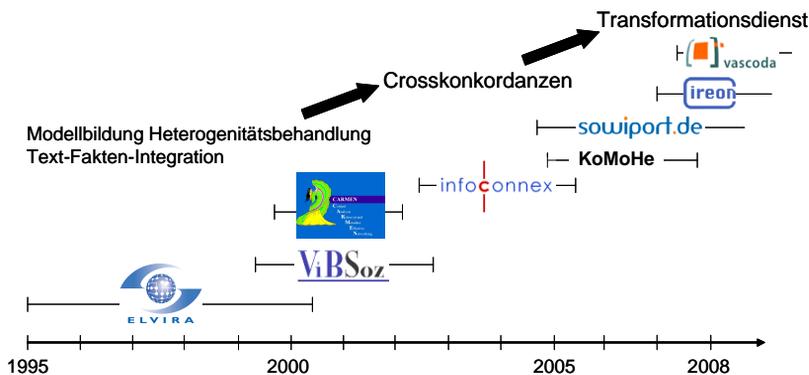


Abbildung 2. Von der Modellbildung zum Termtransformationssdienst

¹ <http://www.gesis.org/forschung-lehre/programme-projekte/informationssysteme/>

Am Anfang stand die Modellbildung zur Heterogenitätsbehandlung und insbesondere zur Text-Fakten-Integration, die am Anwendungsfall „Marktforschung“ im Projekt ELVIRA stattfand (Stempfhuber et al. 2002). Das dort entwickelte Modell orientiert sich einerseits an den in diesem Kontext relevanten Informationsarten (z. B. mit unterschiedlichen Nomenklaturen erschlossene Statistikdaten aus der ganzen Welt und heterogen erschlossene Literaturdatenbanken und Volltexte), andererseits aber auch an konkreten Informationsbedürfnissen und Suchstrategien von Nutzern sowie an Antwortstrategien von Informationsvermittlern (z. B. Strategien zum Schließen von Informationslücken oder zur Erklärung von Effekten in statistischen Daten mithilfe von Textdokumenten). Als Ergebnis liegt ein Informationssystem vor, bei dem Anfragen nach statistischen Informationen mithilfe eines zentralen Vokabulars (switching vocabulary) zwischen etwa 30 Nomenklaturen umgeschlüsselt werden, und in dem mittels statistischer Abbildungen Nomenklatureinträge in Deskriptoren des Standardthesaurus Wirtschaft (und vice versa) umgewandelt werden, sodass eine integrierte Suche nach Texten und statistischen Daten möglich ist. Zusätzlich zur semantischen Ebene wurde Heterogenität auch auf der Ebene der Benutzungsoberfläche behandelt, was zur Entwicklung aufgabenadaptiver, dynamischer Benutzungsoberflächen führte (Stempfhuber 2003).

Das in ELVIRA entwickelte Modell zur Heterogenitätsbehandlung durch Termtransformation wurde in den Projekten CARMEN (Hellweg et al. 2001) und Virtuelle Fachbibliothek Sozialwissenschaften (ViBSoz, Müller 2003) aufgegriffen und methodisch weiterentwickelt, hier insbesondere die intellektuelle Erstellung von bilateralen Abbildungen zwischen Erschließungsvokabularen (Crosskonkordanzen, Zeng/Chan 2004) und deren prototypische Anwendung in einem verteilten Informationssystem. Diese Vorarbeiten kamen im Informationsverbund Bildung-Sozialwissenschaften-Psychologie (infoconnex) 2003 erstmalig in einem größeren, fachübergreifenden Kontext zum Einsatz. Die im Rahmen der Projekte CARMEN und infoconnex entwickelten Crosskonkordanzen wurden im Projekt KoMoHe (Mayr et al. 2005 und Mayr/Petras 2008a) fortgeführt und systematisch durch Konkordanzen zu einer Vielzahl weiterer Fachvokabulare ergänzt. Diese Crosskonkordanzen bilden die Basis der Heterogenitätsbehandlung im sozialwissenschaftlichen Portal sowiport.

Der Nachnutzungsaspekt semantischer Verfahren der Heterogenitätsbehandlung in unterschiedlichen Kontexten führte schließlich zu deren Modularisierung als eigenständiger Termtransformationdienst, der im Fach-

portal ireon und im Wissenschaftsportal vascoda eingesetzt wird. Die nachfolgenden Abschnitte sollen erläutern, welche Implementierungsalternativen in der Praxis zur Verfügung stehen.

2 Einsatz von Crosskonkordanzen in Fachportalen

Crosskonkordanzen sind intellektuell von Experten erstellte Zuordnungen zwischen semantisch verwandten Konzepten unterschiedlicher Thesauri oder Klassifikationen. Zwei Anforderungsbereiche bestimmen die Art der Behandlung der semantischen Heterogenität durch Crosskonkordanzen in einem Fachportal: grundsätzliche konzeptionelle Fragestellungen unter dem Gesichtspunkt des Information Retrieval und softwaretechnische Fragestellungen. Unter dem Blickwinkel des Information Retrieval ist ein System so zu gestalten, dass Nutzer bezüglich ihres Informationsbedürfnisses mit möglichst allen relevanten Informationen versorgt werden, wobei bereits die Wahl eines geeigneten Relevanzmaßes für eine nutzerorientierte oder systemorientierte Evaluation (Manning et al. 2008) nicht trivial ist (Cooper 1973, Voorhees 2002). Unter dem Blickwinkel der Softwaretechnik sollte eine Implementierung nicht nur performant und zuverlässig sein, sondern muss auch nachnutzbar in anderen Anwendungskontexten sein.

2.1 Relationstypen und Vokabulare

Bei der Verwendung von Crosskonkordanzen ist die Entscheidung von Bedeutung, welche Vokabulare und Relationen zwischen ihnen für die Bearbeitung einer Anfrage berücksichtigt werden sollen. Die folgenden Kriterien sind in diesem Zusammenhang relevant:

- Welche Relationstypen werden berücksichtigt? Neben einer eindeutigen Zuordnung von semantisch gleichen Konzepten zweier Vokabulare durch Äquivalenzrelationen bestehen mit Ober- und Unterbegriffsrelationen die Möglichkeit, die Präzision der Nutzeranfrage systemseitig oder auf Nutzerfeedback hin zu variieren.

- Werden indirekte Relationen berücksichtigt? Wie aus Abbildung 3 zu erkennen, besteht beim Aufbau von Crosskonkordanzen zwischen unterschiedlichen Fachgebieten keine Zuordnung zwischen allen Fachgebieten und damit auch nicht zwischen allen verwandten Vokabularen. Da die Zahl von notwendigen Zuordnungen quadratisch mit der Zahl der Dokumentenbestände wächst, können Crosskonkordanzen zwischen allen Vokabularen in komplexen Szenarien nicht mit vertretbarem Aufwand erstellt und gepflegt werden. Auch die Verwendung eines einheitlichen switching vocabularies ist nicht immer möglich. Somit ist zu klären, ob auch über mehrere Vokabulare hinweg verkettete Transformationen berücksichtigt werden können.

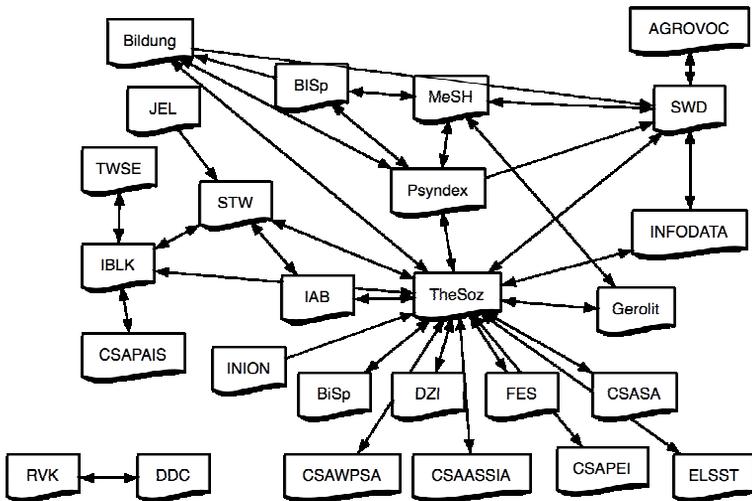


Abbildung 3. Netzwerk von Crosskonkordanzen zwischen Vokabularen

- Welche Vokabulare werden für die Recherche verwendet? Neben einer ausschließlichen Verwendung von Schlagwörtern besteht insbesondere für die Volltextsuche die Möglichkeit, unkontrollierte Vokabulare zur Anreicherung der Anfrage (Termexpansion) zu verwenden.

Da vor der eigentlichen Behandlung der Vagheit V_{cc} noch die Behandlung der Vagheit V_{ic} erfolgen muss, etwa durch die Anwendung linguistischer Methoden oder den Einsatz unterschiedlicher Sprachlexika, ergeben sich zahlreiche Alternativen zur Einbindung von Crosskonkordanzen und für Strategien zur Transformation einer Nutzeranfrage auf eine Vielzahl unterschiedlich erschlossener Datenbestände (Abbildung 4).

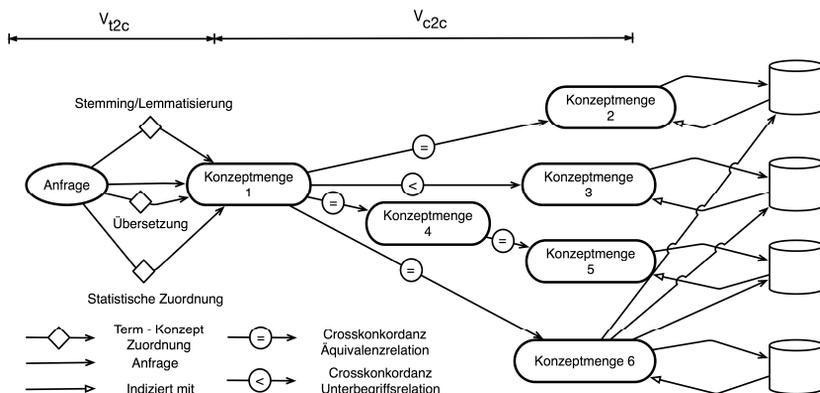


Abbildung 4. Strategien zur Termtransformation mit Crosskonkordanzen

Die anhand der obigen Fragestellungen gewählte Vorgehensweise hat signifikanten Einfluss auf die interne Umsetzung von Nutzeranfragen auf einen aus heterogenen Einzelbeständen aufgebauten, integrierten Datenbestand, bei dem also optimierte Teilanfragen auf spezifische Bereiche des Suchraums angewendet werden. Der Einsatz jedes Verfahrens erzwingt einen Kompromiss hinsichtlich der Optimierung zwischen Recall und Precision einer Anfrage. So führt die Verwendung von Relationen über Äquivalenzrelationen hinaus zur Aufnahme von mehr Suchtermen in eine Anfrage und damit auch zu einer erhöhten Komplexität der Anfragelogik. Eine für alle Portale, Nutzergruppen und Informationsbedürfnisse gültige Lösung kann es somit nicht geben. Dies gilt im Besonderen bei der Verkettung von Transformationen über mehrere Thesauri hinweg.

Unter der Annahme, dass Nutzer grundsätzlich mit kontrolliertem Vokabular suchen (Schlagwörter), wäre ein Ansatz ausreichend, der Anfrageterme jeweils nur auf die Schlagwörter der angesprochenen Einzelbestände expandiert. Da aber ca. 46% der Anfragen in sowiport Freitextsuchen sind, stellt die Expansion einer Anfrage um Zielterme aus allen zur Verfügung stehenden Datenbanken eine zu prüfende Alternative für die Termexpansion und damit Steigerung des Recall dar.

2.2 Modifikation von Datenbestand oder Anfragen

Unabhängig von der zuvor diskutierten Frage, welche Strategie der Anfrage-
transformation mit Crosskonkordanzen Verwendung findet, ist zu klären, wie

diese Crosskonkordanzen in das Informationssystem eingebunden werden. Hierzu gibt es grundsätzlich zwei Möglichkeiten:

- Die Einbindung erfolgt durch die Erweiterung der Dokumentsurrogate um Terme aus anderen Vokabularen – d.h. zum Indexierungszeitpunkt.
- Die Einbindung erfolgt on-the-fly durch Erweiterung der Nutzeranfrage vor der Ausführung durch das Informationssystem – d.h. zum Anfragezeitpunkt.

Die Einbettung von Crosskonkordanzen in die Dokumentsurrogate während der Indexierung hat den Vorteil, dass zur Anfragezeit keine Transformation der Nutzeranfrage vorgenommen werden muss, die die Antwortzeit des Retrievalsystems verlängert. Allerdings erfordert die Verfügbarmachung neuer Crosskonkordanzen oder die Veränderung der Transformationsstrategie eine Modifikation der Dokumentsurrogate mit anschließender Neuindexierung. Für die Einbindung von Crosskonkordanzen in die laufende Anfragebearbeitung (also in „Echtzeit“) spricht die höhere Flexibilität. So können unterschiedliche Expansionsstrategien parallel realisiert werden, ohne eine entsprechende Anpassung der Dokumentsurrogate durchführen zu müssen – allerdings bei erhöhter Anfragekomplexität und -zeit.

2.3 Komponente oder Dienst

Für die Einbindung von Crosskonkordanzen in die Anfrageverarbeitung stehen grundsätzlich zwei Alternativen zur Verfügung:

- Das System zur Heterogenitätsbehandlung wird als Komponente direkt in die Portalsoftware eingebunden. Bei dieser Form der Einbindung wird die Heterogenitätsbehandlung auf die konkreten Anforderungen des jeweiligen Fachportals hin ausgerichtet und an dessen spezifische Architektur angepasst sein.
- Die Heterogenitätsbehandlung wird als zentraler Dienst für externe Softwaresysteme zugänglich gemacht: “Services are self-describing, platform-agnostic computational elements that support rapid, low-cost composition of distributed applications” (Papazoglou 2003). Ein Termtransformationsdienst nach dieser Definition ist auch durch andere Fachportale oder andere Systeme verwendbar. Allerdings ist ein solcher Dienst als Produkt und Teil eines Programmiersystems mit höherem Aufwand zu implementieren als eine Bibliothek oder Komponente (Brooks/Phillips 1995).

Die Entscheidung zwischen einer Implementierung der Heterogenitätsbehandlung als Komponente oder als Dienst hängt neben den verfügbaren Ressourcen vor allem davon ab, ob eine Nutzung außerhalb des ursprünglichen Anwendungskontexts vorgesehen ist. Ist dies der Fall, so ist die Implementierung als Dienst sinnvoll, da die Pflege paralleler Installationen der Komponente (ggf. in unterschiedlichen Versionen) bei verschiedenen Nachnutzern entfällt. Auch sinkt das Risiko einer Fragmentierung der Codebasis in verschiedene Mitglieder einer Programmfamilie (Parnas 1976), einer Sammlung ähnlicher Systeme ohne systematische Wiederverwendung von Implementierungswissen und Artefakten. Ein Dienst hat darüber hinaus den Vorteil, dass er trotz höherem initialen Implementierungsaufwand über einen längeren Zeitraum betrachtet mit geringerem Ressourcenaufwand betrieben werden kann, da die Wartungskosten aus den oben beschriebenen Gründen geringer sind.

2.4 Abwägung der angebotenen Mehrwertdienste

Bei der Implementierung eines Dienstes zum Termtransfer sind die Schnittstellen, die ein solcher Dienst anbietet, von großer Bedeutung. Dabei ist zu klären, welche Funktionalität der Termtransferdienst anbieten soll. Bezüglich der öffentlich zugänglichen Funktionalität stehen die folgenden Alternativen zur Verfügung:

- Nur die für den Hauptanwendungsfall benötigte Funktionalität wird durch die Schnittstellen des Dienstes veröffentlicht. Die Herausforderung bei diesem Ansatz besteht darin, einen Hauptanwendungsfall zu ermitteln, der für alle Nutzer des Dienstes zutrifft. Wie zuvor dargestellt, unterliegt die Einbindung von Crosskonkordanzen einer Reihe von Entscheidungen, die von den individuellen Anforderungen und Kenntnissen der Nutzer eines Informationssystems und den zu durchsuchenden Datenbeständen abhängt.
- Die Schnittstellen decken die gesamte konzeptionell wünschenswerte und technisch realisierbare Funktionalität ab. Dies setzt eine umfassendere – als die reine Freigabe der Kernfunktionalität – Implementierung des Heterogenitätsdienstes voraus und bedingt damit einen größeren Ressourcenaufwand. Unter dem Aspekt der Nachhaltigkeit wiegt jedoch noch schwerer, dass eine derart umfassende Implementierung erfordert, dass die Schnittstellen des Dienstes inklusive aller Parametrisierungen vollständig spezifiziert und dokumentiert werden.

Sowohl die Implementierung eines Dienstes auf einen Hauptanwendungsfall hin, als auch die Implementierung aller möglichen Verwendungsszenarien, haben in der Praxis erhebliche Nachteile. Vor diesem Hintergrund bietet sich die inkrementelle Entwicklung (Larman/Basili 2003) der Schnittstellen wie auch der Funktionalität an. Hierbei werden die Erfahrungen aus Entwicklung und Verwendung in einer ersten Version des Heterogenitätssdienstes als Grundlage für dessen Weiterentwicklung verwendet.

3 Entwicklung von Termtransferkomponenten

Anhand der Portale infoconnex², sowiport³, IREON⁴ und schließlich der für vascoda⁵ entwickelten Variante soll nachfolgend die Entwicklung der Heterogenitätsbehandlung mit Crosskonkordanzen in Bezug auf Integration, Dienstorientierung und Nachnutzbarkeit illustriert werden.

3.1 infoconnex

Ziel des interdisziplinären Informationsverbundes „Pädagogik-Sozialwissenschaften-Psychologie“ (infoconnex) ist es, dem Benutzer innerhalb einer dezentral organisierten Digitalen Bibliothek einen integrierten Zugang zu allen relevanten Informationen – hier Informationen aus den Datenbanken SOLIS/SOLFIS (Sozialwissenschaften), FIS Bildung (Pädagogik) und Psyndex (Psychologie) – zu ermöglichen. Ein Problem dieses interdisziplinären Informationsdienstes ist die semantische Heterogenität, die durch die unterschiedlichen Erschließungssysteme der Datenbanken entsteht: Jede der drei Fachdatenbanken ist mit einem eigenen Thesaurus erschlossen. Das Vokabular der Thesauri unterscheidet sich nicht nur aufgrund der unterschiedlichen fachlichen Ausrichtung, sondern auch aufgrund unterschiedlicher Verwendung von Komposita bzw. Präkombination. So entspricht der

2 <http://www.infoconnex.de>

3 <http://www.sowiport.de>

4 <http://www.ireon-portal.de/>

5 <http://www.vascoda.de>

Deskriptor „Kommunikationsgeschichte“ aus dem FIS-Bildung-Thesaurus im Thesaurus Sozialwissenschaften dem Term „Mediengeschichte“ und im Psyndex-Thesaurus der Deskriptorkombination „Kommunikation UND Geschichte“. Für den Benutzer bedeutet dies, dass er mit einer Schlagwortsuche ohne semantische Heterogenitätsbehandlung unter Umständen nur in einer der drei Datenbanken Treffer erhält (für „Kommunikationsgeschichte“ also nur in FIS-Bildung).

Diese semantische Heterogenität wird in infoconnex durch Transformation der Suchbegriffe auf der Basis von intellektuell erstellten Crosskonkordanzen überwunden. Die Heterogenitätsbehandlung erfolgt während der Anfragebearbeitung (und nicht zum Indexierungszeitpunkt), d. h. die Anfragelogik wird für jede Zieldatenbank auf der Basis der Crosskonkordanzen entsprechend transformiert. Verwendet werden dabei nur Äquivalenzrelationen, da diese die höchste Präzision der Transformation darstellen. Wird kein konkordanter Begriff gefunden, wird der Term des Nutzers an die Zieldatenbank weitergeleitet. Termtransformationen werden in infoconnex nicht nur bei der Schlagwortsuche eingesetzt, sondern auch bei der Volltextsuche. Durch die semantische Heterogenitätsbehandlung konnte der Recall erheblich erhöht werden (Mayr et al. 2005, Mayr/Walter 2007).

3.2 sowiport und ireon

Im sozialwissenschaftlichen Portal sowiport der GESIS und dem Portal ireon der Stiftung Wissenschaft und Politik kamen verschiedene Verfahren zur Termtransformation parallel zum Einsatz. Konzeptionell wurde nach einer Versuchsphase eine Implementierung gewählt, die alle durch direkte Äquivalenzrelationen erreichbaren Konzepte für einen Anfrageterm auflöst und in die Dokumentsurrogate einbindet. Die Implementierung erfolgte in drei Phasen. Zunächst wurden die Informationsbestände in sowiport mithilfe der Crosskonkordanzen um alle Begriffe aus dem Thesaurus Sozialwissenschaften angereichert, die äquivalent zu den nativ verwendeten Schlagwörtern sind. Dieser Teil der Heterogenitätsbehandlung fand also während des Aufbaus des Datenbestands, d.h. zum Indexierungszeitpunkt statt.

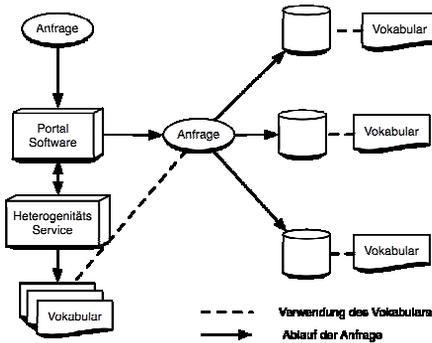


Abb. 5: Generische Anfragen

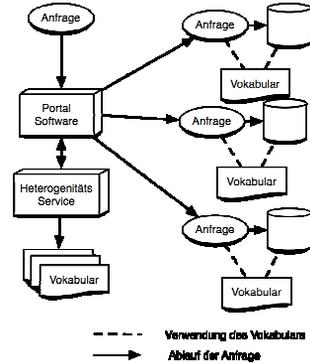


Abb. 6: Spezifische Anfragen

Im Anschluss daran wurden verschiedene Strategien zur Anfrageexpansion in Echtzeit erprobt und als fester Bestandteil des jeweiligen Fachportals unter Berücksichtigung der spezifischen Anforderungen implementiert. Die in diesen Bibliotheken enthaltene Funktionalität wurde später extrahiert und in einem Dienst gekapselt, der über eine HTTP-Schnittstelle angesprochen werden kann. Die Expansion von Anfragen findet wie in Abbildung 5 dargestellt für alle angesprochenen Datenbestände identisch statt. Die Weiterentwicklung des Dienstes hin zu spezifischen Anfragen an die Teildatenbanken (Abbildung 6) wird zzt. insbesondere unter dem Gesichtspunkt des Laufzeitverhaltens und des Einflusses verschiedener Anfragestrategien auf Recall und Precision geprüft.

4 Fazit

Die Behandlung der semantischen Heterogenität in Fachportalen hat sich seit dem Onlinegang von infoconnex im Jahr 2003 beständig weiterentwickelt. Die Einbindung in sowiport und die Nachnutzung in ireon, vifanord und vascoda demonstrieren den praktischen Wert der Heterogenitätsbehandlung für wissenschaftliche Fachportale. Unter technischen Gesichtspunkten entwickelte sich die Heterogenitätsbehandlung hin zu einem nachnutzbaren Dienst mit plattformunabhängigen Schnittstellen, der die Nutzung des in den Crosskondordanzen codierten semantischen Wissens in vielfältigen Kon-

texten erlaubt. Der Dienst liefert zzt. zu einem Ausgangsterm alle damit verknüpften Konzepte in den Zielthesauri. Die Erweiterung dieser Basisfunktionalität wird, ausgehend von den Erfordernissen von Nachutzern und weiteren Evaluationsergebnissen (insbesondere bei Anwendung linguistischer Verfahren), Gegenstand zukünftiger Arbeiten; so ist eine Verwendung der Heterogenitätskomponente in vascoda.de noch in 2009 vorgesehen. Erste Ergebnisse (s. Mayr/Petras 2008b) dokumentieren den positiven Einfluss von Crosskonkordanzen zur Heterogenitätsbehandlung auf Recall und Precision.

Literaturverzeichnis

- Brooks, F.P./Phillips, F. (1995). *The mythical man-month: essays on software engineering*. Addison-Wesley.
- Cooper, W.S. (1973). On Selecting a Measure of Retrieval Effectiveness. *Journal of the American Society for Information Science*, **24** (6), 413–424.
- Hellweg, H./Krause, J./Mandl, T./Marx, J./Müller, M.N.O./Mutschke, P./Strötgen, R. (2001). *Treatment of Semantic Heterogeneity in Information Retrieval*. Technical Report 23. Gesis-IZ Sozialwissenschaften.
- Larman, C./Basili, V.R. (2003). *Iterative and Incremental Development: A Brief History*. *Computer*, **36** (6), 47–56.
- Manning, C.-D./Raghavan, P./Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mayr, P./Petras, V. (2008a). Building a terminology network for search: the KoMoHe project. *International conference on dublin core and metadata applications 2008*. (erscheint)
- Mayr, P./Petras, V. (2008b). Cross-concordances: terminology mapping and its effectiveness for information retrieval. *IFLA World Library and Information Congress*. (erscheint)
- Mayr, P./Stempfhuber, M./Walter, A.K. (2005). Auf dem Weg zum wissenschaftlichen Fachportal – Modellbildung und Integration heterogener Informationssammlungen. S. 29–43. In: Ockenfeld, M. (Hrsg.): *Leitbild Informationskompetenz: Positionen, Praxis, Perspektiven im europäischen Wissensmarkt*; 27. Online-Tagung der DGI; 67. Jahrestagung der DGI, Frankfurt am Main, 23.–25. Mai 2005; Proceedings: Dt. Ges. für Informationspraxis.

- Mayr, P./Walter, A.K. (2007). Zum Stand der Heterogenitätsbehandlung in vascoda: Bestandsaufnahme und Ausblick. In: 3. Leipziger Kongress für Information und Bibliothek, Leipzig, 19.–22. März 2007 [Conference Paper].
- McCulloch, E./Shiri, A./Nicholson, D. (2005). Challenges and issues in terminology mapping: a digital library perspective. *Electronic library*, **23** (6), 671–677.
- Müller, M. N. O. (2003): Integration unterschiedlich erschlossener Datenbestände am Beispiel der Virtuellen Fachbibliothek Sozialwissenschaften. S. 335–345. In: Schmidt, Ralph (Hrsg.): *Competence in Content; proceedings/25*. Online-Tagung der DGI, Frankfurt am Main, 3.–5. Juni 2003. Frankfurt am Main: DGI.
- Papazoglou, M.-P. (2003). *Service-Oriented Computing: Concepts, Characteristics and Directions*. WISE '03: Proceedings of the Fourth International Conference on Web Information Systems Engineering. Washington, DC, USA: IEEE Computer Society.
- Parnas, D.-L. (1976). On the Design and Development of Program Families. *IEEE transactions on software engineering*, **2** (1), 1–9.
- Stempfhuber, M. (2003): *Objektorientierte Dynamische Benutzungsoberflächen ODIN: Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie*. Bonn: IZ Sozialwissenschaften. 337 S. (Forschungsberichte; 6).
- Stempfhuber, M./Hellweg, H./Schaefer, A. (2002): ELVIRA: User Friendly Retrieval of Heterogenous Data in Market Research. S. 299–304. In: Callaos, N./ Hernandez-Encinas, L./Yetim, F. (Hrsg.): *SCI 2002: The 6th World Multiconference on Systemics, Cybernetics and Informatics; July 14–18, 2002, Orlando, USA; Proceedings, Vol. I: Information Systems Development I*. Orlando: TPA Publ.
- Voorhees, E.-M. (2002). *The Philosophy of Information Retrieval Evaluation*. CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems. London, UK: Springer-Verlag.
- 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** (5), 377–395.