In: Knorz, Gerhard; Kuhlen, Rainer (Hg.): Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft (ISI 2000), Darmstadt, 8. – 10. November 2000. Konstanz: UVK Verlagsgesellschaft mbH, 2000. S. 179 – 194

Die virtuelle Fachbibliothek Sozialwissenschaften

Matthias N.O. Müller
IZ Sozialwissenschaften
Lennstraße 30
D-53113 Bonn
mr@bonn.iz-soz.de

Wolfgang Meier
Stefan Winkler
TU Darmstadt
Institut ihr Soziologie
Residenzschloss
D-64283 Darmstadt
meierlwinkler@ifs.tu-darmstadt.de

Zusammenfassung

Das DFG-Projekt "Virtuelle Fachbibliothek Sozialwissenschaften" (ViBSoz) zielt auf die integrierte Bereitstellung sozialwissenschaftlicher Literaturinformationen aus verteilten, verschieden strukturierten und inhaltlich unterschiedlich erschlossenen Datenbeständen, die sich in miteinander nicht verbundenen, heterogenen Organisationsstrukturen und Zugänglichkeitskontexten befinden. Schwerpunkte bei den zu lösenden Problemen sind der Aufbau einer geeigneten Systemarchitektur als Vorraussetzung für höhere Dienste, eine Lösung der inhaltlichen Heterogenitätsproblematik mit Hilfe von statistischen und intellektuellen Transferkomponenten sowie die Schaffung einer benutzerorientierten Oberfläche, die der verteilten Struktur Rechnung trägt.

Abstract

The Social Science Virtual Library project aims at presenting an integrated view to distributed, heterogeneous data of social science literature. The main emphasis has been put on solving problems of access to such diverse document sets. As a prerequisite for higher services like this, an adequate system
architecture has to be implemented. The heterogeneity in content will be solved by transfer components, which will realize a switching of vocabulary. Last but not least a user oriented interface will be implemented, which takes the diversity of the given data into consideration.

1. Einleitung

Das DFG-Projekt "Virtuelle Fachbibliothek Sozialwissenschaften" (ViBSoz) zielt auf die integrierte Bereitstellung sozialwissenschaftlicher Literaturinformationen aus verteilten, verschieden strukturierten und inhaltlich unterschiedlich erschlossenen Datenbeständen, die sich in miteinander nicht verbundenen, heterogenen Organisationsstrukturen und Zugänglichkeitskontexten befinden (Institutsbibliotheken, Sondersammelgebiete der Universitätsbibliotheken (SSG), wissenschaftliche Spezialbibliotheken, Referenzdatenbanken, digitale Volltexte).

Mit Hilfe der Virtuellen Fachbibliothek soll dem Benutzer somit die Möglichkeit gegeben werden in nur *einem* Suchvorgang in unterschiedlichen, verteilten Datenbeständen zu recherchieren. Dazu soll er die ihm vertraute Sacherschließung nutzen können, d.h. das System wird in der Lage sein, die vom Benutzer gestellte Anfrage adäquat in andere Sacherschließungen umzusetzen.
Eine geeignete flexible Architektur vorausgesetzt, sind vorrangig zwei Problembereiche zu bearbeiten:

- Lösung der Heterogenitätsproblematik mit Hilfe von statistischen, intellektuellen und neuronalen Transferkomponenten

- eine benutzerorientierte Oberfläche, die der verteilten Struktur Rechnung trägt und die Rechercheintelligenz für den Benutzer transparent macht.

ViBSoz ist ein Gemeinschaftsprojekt des Informationszentrums Sozialwissenschaften, Bonn (Prof. Dr. J. Krause) und des Instituts für Soziologie der TU Darmstadt (Prof. Dr. R. Schmiede). Weitere Kooperationspartner sind das Sondersammelgebiet Sozialwissenschaften der Universitäts- und Stadtbibliothek Köln (USB Köln), die Bibliothek der Friedrich-Ebert-Stiftung (FES), Bonn, das Zentrum für Interdisziplinäre Technikforschung (ZIT) der TU Darmstadt, sowie der Westdeutsche Verlag, Wiesbaden und der Leske + Budrich Verlag, Leverkusen-Opladen. Im Oktober 1999 hat das Wissenschaftszentrum Berlin (WZB) Interesse geäußert, sich ebenfalls an dem Projekt zu beteiligen und ist mit seinen Datenbeständen einbezogen worden.

2. Heterogene Datenbestände

Zwar bemühen sich die Projektpartner, eine weitmögliche Vereinheitlichung bei den verwendeten Datenbanksystemen bzw. Datenmodellen zu erreichen, realistischerweise wird jedoch prinzipiell von einem heterogenen Umfeld in diesem Bereich ausgegangen, um größtmögliche Offenheit und Erweiterbarkeit des Systems zu gewährleisten.

Die Datenbestände der Projektpartner verbleiben vor Ort, d.h. es gibt keinen zentralen Datenserver, Zudem soll die Möglichkeit bestehen, die lokal bereits existierenden Retrievalsysteme zu nutzen. Gegenwärtig sind im Projekt folgende Datenquellen verfügbar:

- Die Datenbestände des Darmstädter Virtuellen Gesamtkatalogs (Monografien und Graue Literatur) waren zunächst aus dem Allegro-Bibliothekssystem in eine relationale Datenbank (Oracle) zu überführen. Es handelt sich dabei um einen OPAC mit den Beständen von über 60 Bibliotheken (490.000 Datensätze), darunter qualitativ hoch erschlossene Fachbibliotheken, wie die des ZIT.

- Die Oracle-Datenbanken des IZ Sozialwissenschaften, Bonn (SOLIS und FORTS) enthalten ca. 300.000 Monografien, mit Grauer Literatur, Zeitschriftenaufsätzen und Sammelwerksbeiträgen.

- Der SISIS OPAC der USB Köln (Gesamtbestand ca. 1,5 Mio) verweist mit ca. 700.000 Einträgen auf Dokumente des Sondersammelgebiets Sozialwissenschaften.

- Der OPAC der FES-Bibliothek verzeichnet 500.000 Dokumente (Monografien, Graue Literatur und Zeitschriftenaufsätze in mehreren Sprachen), die über ein Allegro-System zugänglich sind. Hinzu kommen 600 eigene Publikationen als elektronische Volltexte.

- Im OPAC des WZB werden insgesamt 278.700 Dokumente nachgewiesen. Als Information-Retrieval-System wird BRS/Search der Firma IHS verwendet, für das auch eine Z39.50 Schnittstelle zur Verfügung steht.

Hinsichtlich der im Projekt vertretenen unterschiedlichen Sacherschließungen siehe Kapitel 5.1

3. Systemarchitektur

Die technische Implementierung eines verteilten Retrieval-Systems muss die für das Projekt zentralen Überlegungen zur Heterogenitätsbehandlung berücksichtigen. Die Programmlogik muss in der Lage sein, zwischen der Benutzeranfrage und den jeweiligen Informationsquellen zu vermitteln, wobei deren formale und inhaltliche Unterschiede zu berücksichtigen sind.

Zwei grundlegende Ziele sind daher:

- dass der Benutzer seine Anfrage in einer allgemeinen und adäquat zu bedienenden graphischen Benutzungsoberfläche formulieren kann, ohne im Einzelnen über die formalen und inhaltlichen Besonderheiten der angesprochenen Quellen informiert sein zu müssen.

- dass weitere Informationsquellen jederzeit in die Architektur aufgenommen werden können. Die Schnittstellen zwischen ViBSoz und den Systemen der teilnehmenden Projektpartner müssen deshalb allgemein anerkannten Standards genügen.

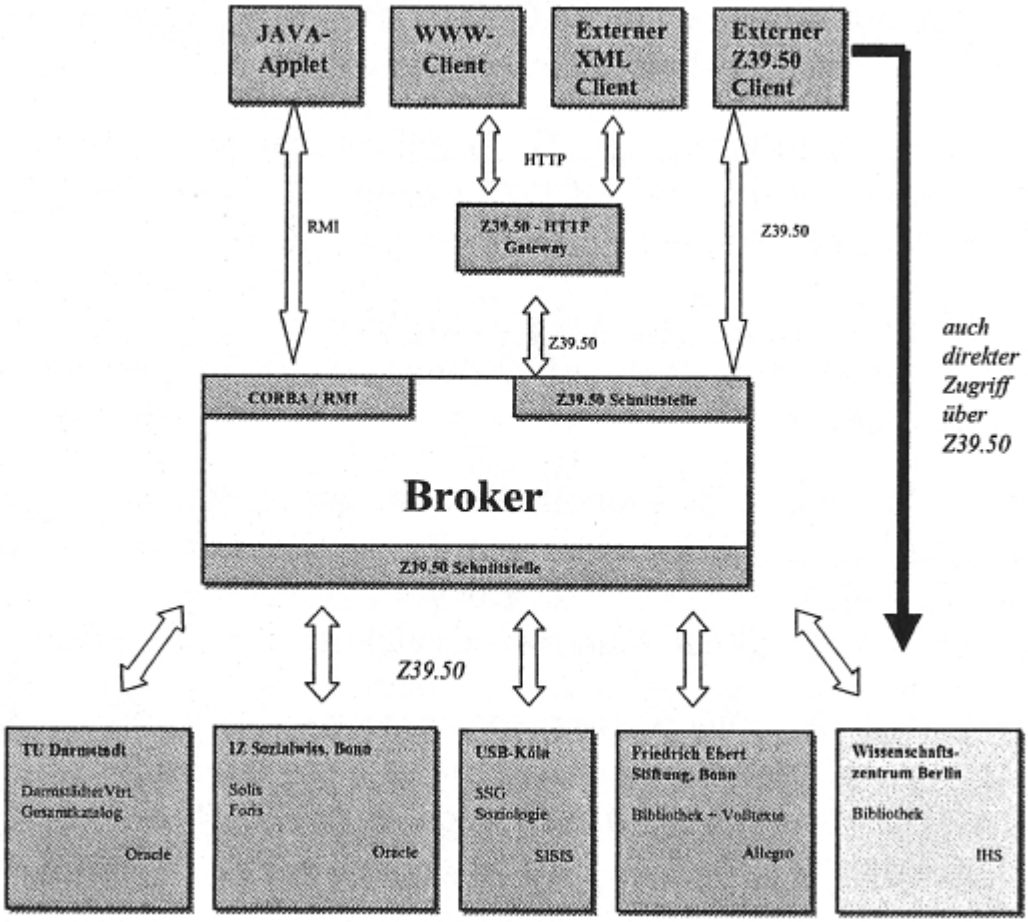

Abbildung 1: Systemarchitektur

Die Projektpartner halten ihre Datenbestände jeweils auf unterschiedlichen Informationssystemen. Vertreten sind relationale Datenbanken, verschiedene Bibliothekssysteme und Volltextdatenbanken (siehe Kapitel 2). Um erweiterbar zu sein, muss die Architektur deshalb von den speziellen Eigenschaften und Möglichkeiten des betreffenden Systems abstrahieren. Damit auf diese Systeme über eine einheitliche Schnittstelle zugegriffen werden kann, wird das im Bibliotheksbereich verbreitete Z39.50 Protokoll als Kommunikationsprotokoll verwendet.

IZ Bonn und TU Darmstadt halten ihre Daten auf Oracle Datenbanken. Um auf diese Daten über Z39.50 zugreifen zu können, wurde von uns ein eigenes Z39.50-Server-Modul implementiert. Dieses dient zunächst dazu, ankommende Z39.50-Anfragen auf die entsprechenden SQL-Kommandos umzusetzen, bzw. Datensätze in verschiedenen Formaten (UNIMarc, MAB, XML) aus der Datenbank auszugeben. Der Z39.50-Server wurde so implementiert, dass er auch außerhalb des Projekts von Nutzen sein kann, sowohl, um andere Datenbanksysteme mit einer Z39.50-Schnittstelle auszustatten, wie auch, um eine Übersetzung zwischen Z39.50 und anderen Protokollen zu gewährleisten.

An der TU Darmstadt wurde ein erster Prototyp der Broker-Komponente fertiggestellt (siehe Abb. 1). Realisiert sind bereits die parallele Abfrage der Targets und die Zusammenführung der Ergebnisse.

Da die teilnehmenden Server unterschiedliche Datenformate (Marc, MAB, XML) unterstützen, werden alle Titelsätze zunächst in ein internes Format transformiert. Dieses basiert auf XML, da somit die Verarbeitung möglichst vieler unterschiedlicher Datensatzstrukturen gewährleistet ist. Der Broker muss lediglich wissen, wie er die für Sortierung und Dublettenkontrolle notwendigen Informationen (Titel, Autor, ISBN etc.) aus dem jeweiligen Datensatz gewinnen kann. Die Konvertierung von XML nach Marc, MAB etc. und umgekehrt wird durch editierbare Stylesheets abgearbeitet, wodurch Anpassungen ohne Änderungen am Programmcode möglich sind.

Aus Sicht der Clients ist der Broker ein „normaler" Z39.50-Server und kann von jeder Z39.50-fähigen Software angesprochen werden. Durch die offene Architektur können die Benutzer zwischen diversen Möglichkeiten des Zugriffs wählen, z.B. über eine Java-Oberfläche, ein Web-Interface, spezielle Z39.50-Clients oder über Bibliothekssoftware mit entsprechender Schnittstelle.

4. Benutzungsoberflächen

4.1 Java Oberfläche

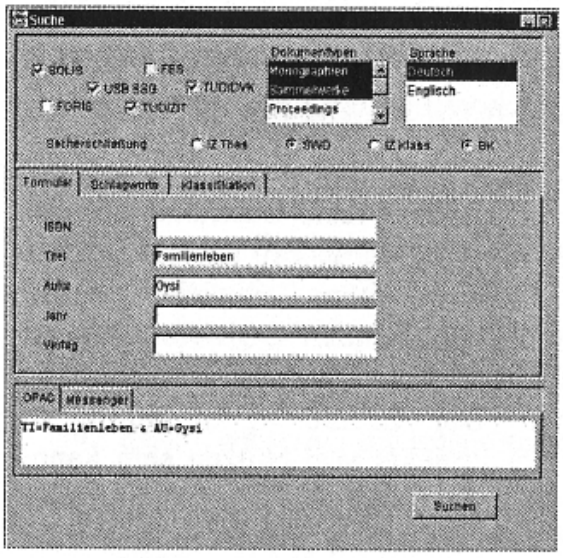

Abbildung 2: Formulareingabe

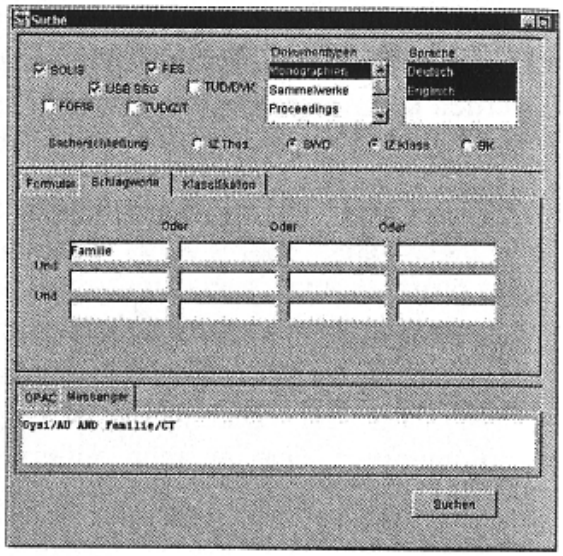

Abbildung 3: Eingabe-Grid

Der primäre Zugang zu ViBSoz wird über eine Java-Applikation erfolgen. Die Verwendung von Java eröffnet gegenüber einem einfachen HTML Zugang alle Möglichkeiten einer eigenständigen Applikation, wie z.B. eine detailreiche, dynamische Oberfläche oder das Speichern und Laden einer Anfrage und deren Ergebnis. Die Umsetzung basiert auf dem WOB-Modell (Krause 1997), welches grundlegende softwareergonomische Prinzipien umsetzt. Dadurch wird eine dem

Problem angemessene und intuitive Benutzung gewährleistet. Der vorliegende Entwurf dient als Arbeitsgrundlage um zu prüfen, in wie weit die Anforderungen des WOB-Modells und die Bedürfnisse der Domäne 'Literatursuche in verteilten Datenbanken' mit Java umgesetzt werden können. Er spiegelt also nur das Grundgerüst der entgültigen Oberfläche wider.

4.1.1 Suchfenster

Das Suchfenster gliedert sich in die drei Bereiche:

- Filter
- graphisch orientierten Eingaben
- formalsprachliche Eingaben

Die im Suchfenster oben liegenden Filter dienen der vorherigen Festlegung, in welchen Datenbeständen gesucht werden soll. Hierzu zählen die zu durchsuchenden Datenbanken, die für den Benutzer relevanten Dokumenttypen (Monographien, Sammelwerke, usw.) und die Sprache der einzelnen Werke. Hinzu kommt die Angabe, welche Sacherschließung für die Anfrageformulierung benutzt werden wird[1]. Der Benutzer kann somit einerseits vorab den Suchraum auf die Dokumente einschränken, die für ihn von Interesse sind, andererseits kann er auch während des Suchprozesses den Suchraum erweitern oder weiter einschränken, ohne die Suchanfrage selbst verändern zu müssen. Die Formulierung der eigentlichen Suchanfrage geschieht wahlweise auf graphisch, formularorientierte Weise oder formalsprachlich.

Zu den graphisch orientierten Eingaben zählen die Formulareingabe (Abbildung 2, Mitte) und das Eingabe-Grid (Abbildung 3, Mitte). Die aus der Formalerschließung bekannten Felder wie Autor/Herausgeber, Titelstichwort, Erscheinungsjahr oder Verlag können über das Formular gesucht werden.

Für die Formulierung einer Schlagwort- und/oder einer Klassifikationssuche können zwei Grids genutzt werden. Durch vorgegebene syntaktische Strukturen erlauben sie die einfache Formulierung einer Boole'schen Anfrage, wodurch der Großteil der von Benutzern formulierten Anfragen abgedeckt wird.

Syntaktisch komplexere Anfragen können in der formalsprachlichen Anfrage durch verschiedene formale Sprachen formuliert werden. Vorgesehen sind die formale Sprache des SISIS Opacs, welche die Benutzer der USB Köln und zum Teil die Benutzer aus Darmstadt gewöhnt sind, und die Sprache MESSENGER des STN[2], die Kunden des IZ -Sozialwissenschaften vertraut ist.

[1] Mehr zur Sacherschließung siehe Abschnitt 5.1.

[2] The Scientific & Technical Information Network. URL: http://www.fizkarlsruhe.de/stn.html

Alle drei Eingabeebenen sind durch das Prinzip der kontextsensitiven Durchlässigkeit[3] miteinander verbunden, d.h. Eingaben in eines der Elemente werden systemgesteuert in die anderen Elemente übertragen und gegebenenfalls transferiert. Eingaben in das Formular oder eines der Grids z.B. werden automatisch in eine formalsprachliche (Teil-)Anfrage übersetzt und im formalsprachlichen Eingabefenster angezeigt und vice versa. Der Benutzer kann also z.B. den Grundstock seiner Anfrage bequem über die graphisch orientierten Eingaben aufbauen, und diesen dann formalsprachlich präzisieren.

4.1.2 Ergebnisdarstellung

Die Ergebnisdarstellung gliedert sich in zwei Bereiche, den formalsprachlichen Dialog und die eigentliche Ergebnispräsentation.

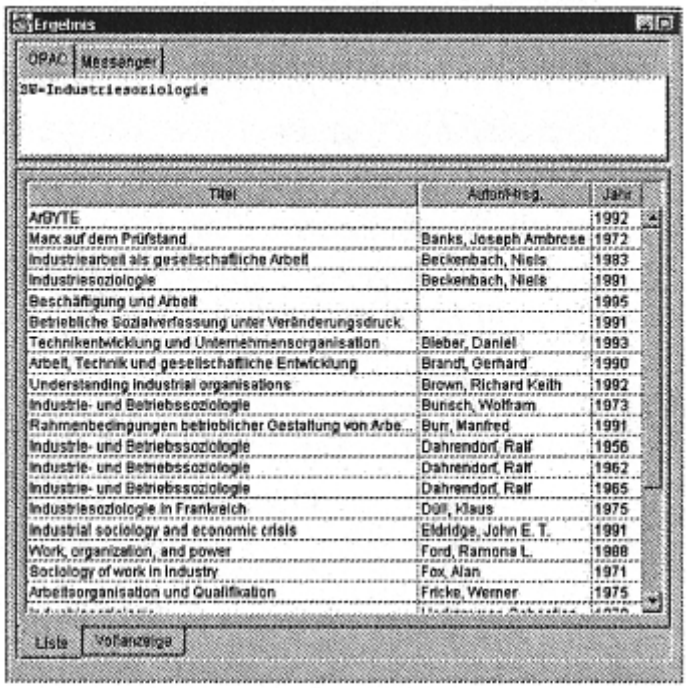

Abbildung 4: Ergebnisfenster

Der formalsprachliche Dialog gleicht exakt dem aus der Suchanfrage, d.h. er dient hier einerseits als Zustandsanzeige über die bislang erstellte Anfrage[4], andererseits kann der Experte hiermit die Anfrage umformulieren und das Ergebnis der Umformulierung gleich bewerten. Eine solche editierbare Zustandsanzeige erspart also im iterativen Retrieval den zusätzlichen Interaktionsschritt der Rückkehr zur Anfrageformulierung und den erneuten Wechsel zur Ergebnisanzeige.

Gerade der Bereich der eigentlichen Ergebnisdarstellung ist noch nicht im Detail entworfen, da das Vorgehen und die Einflussmöglichkeiten auf das Ranking noch nicht endgültig geklärt sind, sie jedoch Auswirkungen auf die Darstellung des Ergebnisses haben.

[3] Mehr zum Thema kontextsensitive Durchlässigkeit siehe zum Beispiel (Müller 1999).

[4] Zur Funktion der Zustandsanzeige im WOB-Modell siehe zum Beispiel (Marx 1996).

Sicher festgelegt werden kann aber schon die Aufteilung in eine Listendarstellung und eine Detailansicht. In der Listendarstellung werden alle Treffer tabellarisch zusammengestellt. Der Benutzer kann sich somit einen Überblick über die Treffermenge und deren Inhalt machen. Für einzelne Dokumente wird in die detailliertere Vollansicht gewechselt. Dort wird alle verfügbare Information über ein einzelnes Dokument angezeigt. Dazu zählen neben der Formalerschließung z.B. die unterschiedlichen Sacherschließungen und - falls vorhanden - ein Verweis auf den Volltext.

4.2 Z39.50-Web-Gateway

Um gelegentlichen Benutzern, die die Installation und Einarbeitung in die funktionsreichere Java Oberfläche scheuen, einen einfachen Zugang zu ermöglichen, wurde ein rein XML basierter Zugang geschaffen.
Da der in ViBSoz verwendete Broker die Datensätze wie erwähnt intern im XML-Format verarbeitet (siehe Kap. 3), kann dieses Format auch für die externen Schnittstellen genutzt. Diese Variante ermöglicht einerseits die Kommunikation mit in naher Zukunft zu erwartenden XML-Browsern, und andererseits die Schaffung eines einfachen HTML Zugangs mit Hilfe von Standard Technologien. Für den HTML-Zugang werden die XML-Daten serverseitig per XSLT nach HTML konvertiert. Zukünftige Webbrowser werden diesen Schritt nicht mehr benötigen.
Generell haben Systeme, die verschiedene Varianten von Benutzungsoberflächen anbieten, das Problem, dass Benutzer potentiell zwischen den beiden Zugängen wechseln können. Deshalb müssen die Grundprinzipien der Bedienung miteinander kompatibel sein. Für ViBSoz bedeutet dies, dass das Design des Z39.50-Webclients und der Java Applikation aufeinander abgestimmt sein müssen. Diese Abstimmung erfolgt auf der Basis der ersten Benutzertests im zweiten Projektabschnitt.
Das Z39.50-Web-Gateway implementiert einen standardkonformen Z39.50-Client, der als XML-/Web-Schnittstelle zu jedem beliebigen Z39.50-Server dienen kann. Die Verwendung von XML ermöglicht eine größtmögliche Trennung von Form, Inhalt und Funktion. Das Layout der Suchseite, sowie der verschiedenen Datenansichten ist vollständig veränderbar, da es gänzlich über XSL-Stylesheets[5] definiert wird.

[5] Siehe: W3C Working Draft (2000). Extensible Stylesheet Language (XSL) Version 1.0, 27 March 2000, http://www.w3.org/TR/xsl/

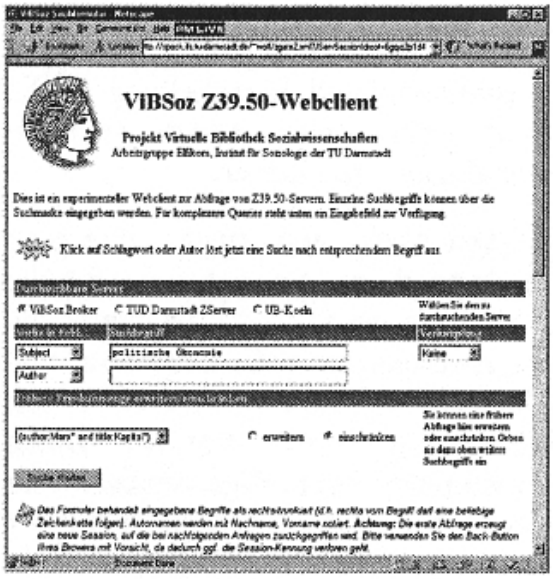

Abbildung 5: Suchmaske des WebGateways

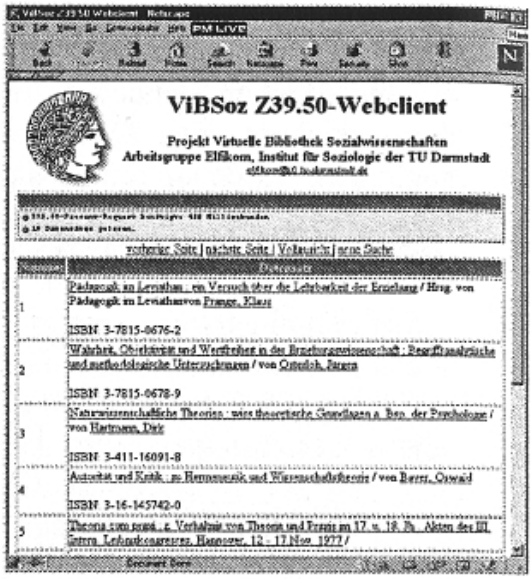

Abbildung 6: Kurzanzeige der Trefferliste

Queries können derzeit wahlweise über einfache HTML-Forms oder eine den populären Internet-Suchmaschinen (Altavista „Advanced Search") nachempfundene Query-Language eingegeben werden. Sie ermöglicht Boole'sche Algebra, Feldsuche, Trunkierung und beliebig viele Klammerebenen. (vgl. Abbildung 5). Die Trefferlisten können mit Namen versehen werden, die wiederum mit neuen Suchbegriffen kombiniert werden können.

5. Behandlung semantischer Heterogenität

Neben der strukturellen Vielfalt ist im Bibliotheksbereich mit einer bereiten Palette verwendeter Sacherschließungssysteme zu rechnen. Ohne geeignete intelligente Verfahren zum Transfer zwischen diesen muss das Rechercheergebnis zwangsläufig unzureichend bleiben.

5.1 Sacherschließungen

Im Projekt Virtuelle Fachbibliothek Sozialwissenschaften ist eine Vielzahl von Sacherschließungen vertreten. Das Spektrum reicht von allgemeinen Regelwerken, wie der Schlagwortnormdatei (SWD), über fachspezifische, wie dem Thesaurus Sozialwissenschaften (IZ-Thesaurus), bis zu nicht reglementierten freien Schlagwörtern. Gleiches gilt für die Klassifikationen, die mit verschiedenen Ausprägungen der Basisklassifikation (BK) als Allgemeinklassifikation, der Klassifikation Sozialwissenschaften (IZ-Klassifikation) als Fachklassifikation und den weniger spezifischen Aufstellsystematiken im Projekt vertreten sind. Die Sacherschließungen der Beteiligten bieten also einen guten Querschnitt der in der

Bibliothekslandschaft vorkommenden Sacherschließungen. Eine große Übertragbarkeit der Ergebnisse des Projekts ist somit gewährleistet.

Im Bereich der Verbalerschließung bilden die Schlagwortnormdatei, der Thesaurus Sozialwissenschaften und das von der FES verwendete Regelwerk einen Schwerpunkt. Im Hinblick auf eine Integration sind hier vor allem das Zusammenspiel bzw. die Gegensätze zwischen der SWD als allgemeinem Sammelwerk und dem IZ-Thesaurus als fachspezifischem Sammelwerk von Interesse. Auch das System der FES, als Sonderfall eines eigenen, auf die Bedürfnisse einer einzelnen Institution abgestimmten Systems, benötigt besonderes Augenmerk. Im Bereich der Klassifikation bilden die Basisklassifikation, mit seinen beiden Ausprägungen PICA/GVK und PICA/HEBIS, sowie die Klassifikation Sozialwissenschaften den zweiten Schwerpunkt. Auch hier findet sich wieder das Problem der sehr groben Allgemeinklassifikation gegenüber der detailreichen, stark fachbezogenen Klassifikation.

5.2 Transfermodule

5.2.1 Problembereich

Im vorherigen Abschnitt wurde die Vielfältigkeit der Sacherschließungen innerhalb des Projekts vorgestellt. Diese Vielfalt an Systemen für die inhaltliche Beschreibung von Dokumenten erzeugt für den Benutzer Probleme bei der Suche nach Information. Möchte er in allen beteiligten Beständen eine tiefergehende Suche durchführen, so war er bisher gezwungen, alle verwendeten Sacherschließungssysteme zu erlernen. Er kann nicht einfach das ihm vertraute System auf die anderen Bestände übertragen. Würde zum Beispiel ein Benutzer der USB Köln mit dem SWD Schlagwort 'Berufliche Fortbildung' in der Datenbank SOLIS des IZ-Sozialwissenschaften suchen, so würde er keine Treffer erhalten, denn dort ist der Begriff als 'Berufliche Weiterbildung' vergeben worden. Ebenso verhält es sich mit vielen Eigennamen und Bezeichnungen, wie z.B. der 'Sowjetunion' die in SOLIS unter 'UdSSR' zu finden ist. Besonders problematisch sind mehrdeutige Begriffe wie 'Haushalt'. Er kann einerseits das Budget andererseits den Haushalt im Privaten bezeichnen. In der USB wird er als Haushalt im Privatbereich verstanden. Um diese Lesart abzudecken, muss beim IZ-Sozialwissenschaften mit dem Schlagwort 'Privathaushalt' gesucht werden, andernfalls erhält der Benutzer nicht die gewünschte Information. Bei steigendem Recall sinkt die Precision.

Aufgabe des zu entwickelnden Systems ist die Unterstützung des Benutzers bei der "Übersetzung" von einem Sacherschließungssystem zum anderen. Idealiter geschieht diese Übersetzung innerhalb des Systems, so dass der Benutzer zwar auf Wunsch darüber informiert wird, nicht aber selbst eingreifen muss. Er kann dann seine Anfrage an das Gesamtsystem in seiner gewohnten Weise formulieren, und die Umsetzung bzw. Anpassung an die verschiedenen beteiligten Systeme erfolgt automatisch.

5.2.2 Cross-Konkordanzen

Eine Möglichkeit eine solche Umsetzung zu realisieren ist die Verwendung von Cross-Konkordanzen. Dabei vergleichen Fachreferenten intellektuell jeweils zwei Sacherschließungssysteme und definieren Übereinstimmungen zwischen diesen. Somit können neben den Synonymierelationen auch Oberbegriffs- / Unterbegriffs- und Ähnlichkeitsrelationen berücksichtigt werden.

Im Projekt wird eine (partielle) intellektuelle Cross-Konkordanz zwischen der SWD und dem Thesaurus Sozialwissenschaften sowie der Basisklassifikation und der Klassifikation Sozialwissenschaften erarbeitet.

5.2.3 Quantitativ-Statistische Verfahren

Eine andere Möglichkeit einer solchen Umsetzung ist der Einsatz quantitativ-statistischer Verfahren, die den Transfer zwischen der Sacherschließung der Benutzereingabe zu den Sacherschließungen der anderen Datenquellen berechnen. Dabei werden die jeweiligen Begriffspaare nicht nach qualitativen Maßstäben erzeugt, wie bei der intellektuellen Erstellung, sondern nach ihrer Quantität bezogen auf einen Korpus, also nach der Häufigkeit ihres Vorkommens. Vereinfacht ausgedrückt: Je häufiger ein Begriffspaar in einem Korpus vor kommt, desto wahrscheinlicher ist es, dass es sich um eine sinnvolle Verbindung handelt. Hinzu kommen Parameter wie die Größe des Korpus oder die Verteilung der Begriffe innerhalb dessen.

Voraussetzung für diese Verfahren ist ein Korpus, in dem die Dokumente nach beiden Sacherschließungen indexiert sind (Parallelkorpus). Er entsteht durch den Abgleich zweier Korpora und die Extraktion von Paaren gleicher Dokumente. Jedes Dokument (-paar) ist dann nach zwei Erschließungssystemen indexiert.

Die Entwicklung der Verfahren zur Erstellung der Parallelkorpora für ViBSoz ist weitgehend abgeschlossen. Zur Zeit liegt ein Parallelkorpus USB Köln – IZ-Sozialwissenschaften mit ca. 15 Tausend Einträgen vor.

Aus einem solchen Korpus werden dann Relationen zwischen einzelnen oder Gruppen von Schlagwörtern / Klassifikationen abgeleitet. Als Verfahren hierfür werden in ViBSoz vorwiegend statistische Verfahren eingesetzt, experimentell aber auch Neuronale Netze.

5.2.3.1 Statistische Verfahren

Statistische Verfahren analysieren die Häufigkeit des gemeinsamen Auftretens zweier Schlagwörter[6] innerhalb des Parallelkorpus und übertragen die

[6] Im Weiteren wird das Vorgehen bei der Analyse von Paaren einzelner Schlagwörter betrachtet. Gleiches gilt aber äquivalent für Klassifikationen und Gruppen von Schlagwörtern / Klassifikationen.

gewonnenen Relationen später auf neue Dokumente. Vergleicht man dabei zwei Schlagwörter aus unterschiedlichen Sacherschließungen, so lassen sich statistische Abhängigkeiten zwischen den einzelnen Begriffen und somit zwischen den Sacherschließungen ermitteln.

Beispielsweise wurden für das Dokument *"Gysi, Jutta: Familienleben in der DDR, zum Alltag von Familien mit Kindern, Akademie Verlag Berlin, 1989, ISBN 3-05-000771-0 ."* an der USB Köln die Schlagwörter 'Deutschland <DDR>' und 'Familie' vergeben. Im IZ-Sozialwissenschaften erhielt das gleiche Dokument die Schlagwörter 'Arbeitsteilung', 'Ehe', 'Familie', 'DDR' und 'Partnerschaft'. Bei diesem Dokument treten also unter Anderem das SWD Schlagwort 'Deutschland <DDR>' und das IZ Schlagwort 'DDR' gemeinsam auf. Ist dieses gemeinsame Auftreten auch bei anderen Dokumenten zu beobachten, so handelt es sich hierbei höchst wahrscheinlich um eine sinnvolle Transferbeziehung. Tatsächlich tritt dieses Paar im derzeitigen Parallelkorpus von USB Köln und IZ Sozialwissenschaften noch bei 272 weiteren Dokumenten auf.

Wie aus der Statistik bekannt, reicht diese absolute Häufigkeit jedoch nicht aus, um die Güte einer solchen Transferbeziehung zu bestimmen. Sie muss z.B. mit der Häufigkeit der Vorkommen der Terme oder der Gesamtzahl der Relationen in Bezug gesetzt werden. Dazu werden in ViBSoz die bedingte Wahrscheinlichkeit und der Äquivalenzindex getestet.

Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit mit der ein Begriff B aus der Sachschließung X auftaucht, wenn auch der Begriff A aus der Sacherschließung Y aufgetaucht ist. Bezogen auf unser Beispiel ist es dann die Frage: Wie wahrscheinlich ist es, dass der IZ Term 'DDR' vergeben wurde, wenn die USB Köln den Term 'Deutschland <DDR>' vergeben hat. Verwendung findet dieses Vorgehen z.B. im System AIR/PHYS (Biebricher/Fuhr/Lustig et al. 1988), welches sich mit automatischer Indexierung befasst.

Der Äquivalenzindex hat seinen Ursprung in der Kookkurrenzanalyse und findet heute vielfältige Verwendung. Er bezieht, im Gegensatz zur bedingten Wahrscheinlichkeit, auch die Vorkommenshäufigkeit des zweiten Terms in die Berechnung mit ein.

Beiden Verfahren gemeinsam ist das Problem, dass der richtige Schwellwert gefunden werden muss, ab dem eine statistische Konkordanz als sinnvoll angesehen werden kann. Dieser Wert ist abhängig von dem verwendeten Verfahren und dem zugrundeliegenden Korpus. Er muss also für alle Sacherschließungspaare einzeln gefunden werden.

Zur Zeit werden die beiden Verfahren anhand des Parallelkorpus USB Köln — IZ-Sozialwissenschaften geprüft und es werden Vorgehen entwickelt, die entsprechenden Schwellwerte zu bestinunen. Erste Ergebnisse, wie bei der IuK-Tagung im März 2000 präsentiert, sind sehr vielversprechend.

5.2.3.2 Neuronale Netze

Die Ausgangsbasis für statistische Verfahren und neuronale Netze ist die gleiche: ein Parallelkorpus. Im Gegensatz zu statistischen Verfahren arbeiten neuronale Netze aber nicht mit mathematischen Formeln, die statistische Relationen berechnen, sondern sie entstehen durch das Training eines neuronalen Netzes mit den Daten des Parallelkorpus. Dabei werden die Sacherschließungen paarweise an ein Backpropagation-Netz angelegt. Das Netz ‚lernt' bei jedem Anlegen die Verbindungen der einzelnen Begriffe und Begriffspaare. Durch häufiges Anlegen bilden sich so zwischen einzelnen Neuronen des Netzes stärkere Verbindungen heraus als bei anderen. Durch die unterschiedliche Stärke der neuronalen Verbindungen ist es in der Lage sich zu ‚merken', welche Begriffe mit welchen anderen korrespondieren. Legt man dann später Begriffe aus einem Sacherschließungssystem an das Netz an, so kann es die zugehörigen Begriffe eines anderen Sacherschließungssystems bestimmen.

Das in ViBSoz erprobte System basiert auf dem COSIMIR Modell (Mandl 1998) und zeigt in ersten Voruntersuchungen ermutigende Leistungen.

6. Ausblick

Nächster Schritt bei der Umsetzung der Projektziele ist die rasche Integration und Ergänzung der vorliegenden Teile zu einem funktionsfähigen Prototypen. Auf seiner Basis können die erarbeiteten Verfahren getestet, verbessert und wo möglich generalisiert werden.

Im Bereich des Brokers liegen die Schwerpunkte zum Einen bei der Integration der noch fehlenden Datenbestände des IZ Sozialwissenschaften, der FES und des WZB. Zum anderen müssen die Verfahren zur Ergebniszusammenführung umgesetzt werden.

Im Bereich der Transferkomponenten stehen die Implementierung und Einbettung der erarbeiteten Cross-Konkordanzen und statistischen Transferbeziehungen in das Gesamtsystem an nächster Stelle. Die Verfeinerung der statistischen Verfahren zur Erstellung letzterer und ihre Anwendung auf die noch verbleibenden Sacherschließungen zählen zu den nächsten Schritten.

Somit ist in den kommenden Monaten mit der Fertigstellung eines funktionsfähigen Prototypen zu rechnen, der dann zur Evaluation zur Verfügung steht.

Informationen zum Projekt sind über den WWW Server des Projekts unter dem URL http://vibsoz.bonn.iz-soz.de/ zu finden. Neben kurzen Beschreibungen der Zielsetzungen des Projekts sind auch Publikationen online abrufbar.

Literatur

Biebricher, P., Fuhr, N., Lustig, G., *et al.* (1988). *'The Automatic Indexing System AIR/PHYS - From Research to Application'.* 11th International Conference on Research & Development in Information Retrieval. 1988, Grenoble, France.

Grievel, L., Mutschke, P. und Polanco, X. (1995). *'Thematic Mapping on Bibliographic Databases by Cluster Analysis: A Description of the SDOC Environment with SOLIS'.* Knowledge Organisation 22: 8.

Krause, J. (1997). *'Das WOB-Modell'* Vages Information Retrieval und graphische Benutzeroberflächen: Beispiel Werkstoffinformation. Konstanz: Universitätsverlag. pp. 59--88.

Mandl, T. (1998). *'Learning Similarity Functions in Information Retrieval'.* 6th European Congress on Intelligent Techniques and Soft Computing. , Aachen, Germany.

Marx, J. (1996). *'Bidirektionale Sprache. Faktenrecherche und Informationsdarstellung durch dynamische Erzeugung korrigierbarer Zustandsanzeigen in natürlicher Sprache'.* Hildesheim: Olms Verlag.

Meier, W., Müller, M.N.O. und Winkler, S. (2000). *'Virtuelle Fachbibliothek Sozialwissenschaften: Problembereich und Konzeption'.* Bibliotheksdienst 34: 1236-1244.

Müller, M.N.O. (1999). *'Konsistenzerhaltung in multimodalen Benutzungsoberflächen an Beispiel KO-NEXIS'.* Fachbereich Informatik, Universität Koblenz-Landau.