Heterogenität in wissenschaftlichen Fachdatenportalen*

Stefan Baerisch

Informationszentrum Sozialwissenschaften
Lennéstr. 30, 53113 Bonn
bs@iz-soz.de

Zusammenfassung
Bei der Bereitstellung von Informationsbeständen verschiedener Anbieter in einem gemeinsamen Portal müssen sich Anbieter verschiedenen Anforderungen bei der Behandlung von Heterogenität stellen. Neben der Zusammenführung und Vereinheitlichung auf struktureller Ebene sind auch Fragestellungen der semantischen Heterogenität zu beachten. Eine weitere Heterogenitätsdimension liegt in den unterschiedlichen Nutzergruppen, die auf ein solches Portal zugreifen. In diesem Papier gehen wir auf Konzepte zur Behandlung verschiedener Heterogenitätsdimensionen ein und diskutieren, wie diese Konzepte in einem Portal angewandt werden können. Schwerpunkte sind die Behandlung von struktureller Heterogenität durch semistrukturierte Datenformate und inkrementelle Integrationsprozesse sowie die Behandlung der semantischen Heterogenität durch Crosskonkordanzen. Abschließend wird die Umsetzung dieser Konzepte am Beispiel des sozialwissenschaftlichen Fachportals SOWIPORT erläutert.

1 Einleitung

Eines der zentralen Anliegen der aktuellen Informationslandschaft ist die Integration vorhandener Informationen zu einem gemeinsamen Zugang. Für einen Nutzer von Informationsdiensten ist es nicht zu leisten, für die Recherche zu einem Thema in mehreren Schritten zuerst potentielle Informationsquellen zu identifizieren, diese zu evaluieren und dann erst anzufragen. Diese Problemstellung wird in der Praxis noch verschärft, da die angebotenen Informationsbestände Aktualisierungen unterliegen, somit also eine Wiederholung von Anfragen in regelmäßigen Abständen notwendig wäre. Bei Betrachtung der Anzahl von im Deep Web (siehe [Ragha-

* Veröffentlicht in: Osswald, Achim; Stempfhuber, Maximilian; Wolff, Christian (Hrsg.) (2007). Open Innovation. Proc. 13. Jahrestagung der IuK-Initiative Wissenschaft. Konstanz: UVK, 509-518.

van01]) verfügbaren Quellen wird offenbar, dass ein manuelles Anfragen von Quellen nicht wünschenswert sein kann.

Eine erste Herausforderung bei der Behandlung mehrerer Quellen ist somit die gemeinsame Anfrage dieser Quellen. Zur integrierten Anfrage verteilter Datenquellen existieren in Forschung und Praxis eine Reihe von Ansätzen, etwa die Verwendung von Metasuchen und Föderierter Suche (siehe [Baeza-Yates99]) und verteilten Datenbanken. Eine alternative Möglichkeit besteht in der zentralen Indizierung der zugänglich zu machenden Informationen. Ist eine gemeinsame Anfrage auf allen Informationsbeständen etabliert, stellt sich als nächste Herausforderung die Integration von strukturellen Unterschieden im Datenbestand: Unterschiedliche Datenanbieter verwenden abweichende Modellierungen der beschriebenen Entitäten, insbesondere von Bedeutung ist die Verfügbarkeit von Daten in verschiedenen Detailgraden. Semantische Heterogenität tritt auf, wenn sich die Konzepte und Vokabulare unterscheiden, die zur Beschreibung von Themen verwandt werden. Schlagwortlisten und Thesauri mit verschiedenem Umfang und unterschiedlichen Schwerpunktsetzungen müssen einem Anwender zur Verfügung gestellt werden, ohne dass dieser Kenntnis über die Details der jeweiligen Vokabulare hat.

Ein letzter Punkt der Heterogenität betrifft nicht die bereitgestellten Informationen als solche, sondern die Art und Weise, wie verschiedene Anwendergruppen ihr jeweiliges Informationsbedürfnis an den integrierten Datenbestand ausdrücken. Eine Übersicht der verschiedenen Heterogenitätsarten bietet die Abbildung 1.

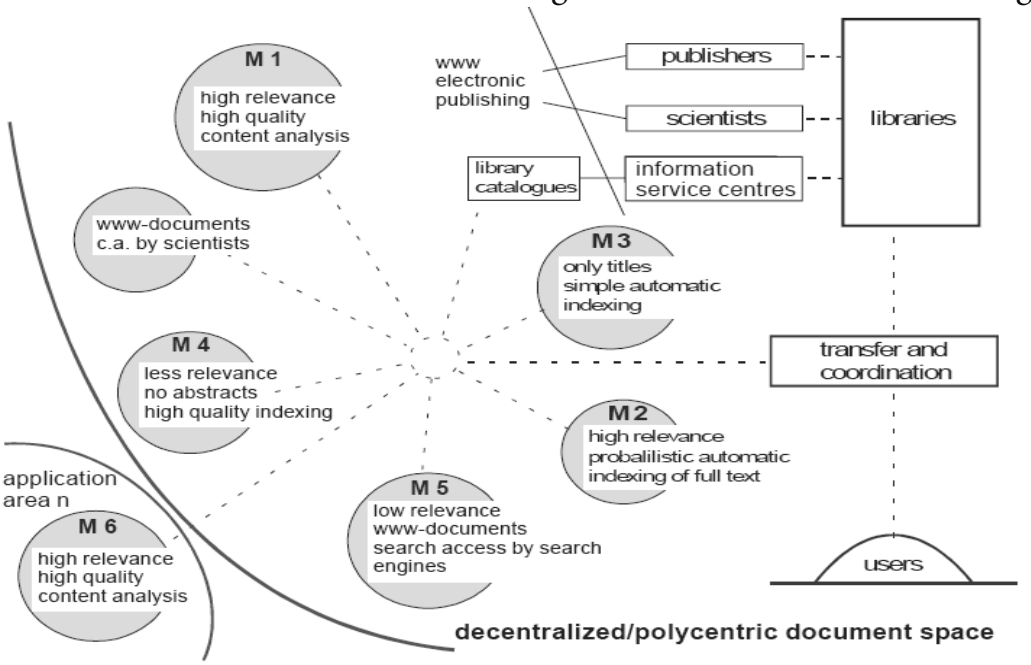

Abbildung 1: Strukturell, semantisch und qualitativ heterogene Datenbestände nach [Krause2004]

In diesem Papier gehen wir auf diese verschiedenen Aspekte oder Dimensionen der Heterogenität ein, die in einem Fachportal behandelt werden müssen. Als Grundlage diskutieren wir die strukturelle Heterogenität in Abschnitt 2 und stellen Ansätze und Integrationsverfahren vor. In diesem Kontext gehen wir auch auf grundsätzliche Aspekte bei der Zusammenführung von Daten ein. Abschnitt 3 geht auf die Behandlung der semantischen Heterogenität ein, im Kern der Betrachtung stehen intellektuelle und statistische Verfahren zur Behandlung zur Erstellung von Crosskonkordanzen. Abschnitt 4 diskutiert den Umgang mit heterogenen Nutzergruppen durch die Mittel der Softwareergonomie und der Oberflächengestaltung. Abschnitt 5 stellt abschließend die gemeinsame Anwendung der diskutierten Konzepte vor, den Hintergrund der Darstellung bildet das wissenschaftliche Fachportal SOWIPORT; vor diesem Hintergrund stellen wir auch unsere Betrachtung von der Informationsintegration als Prozess vor.

2 Datenzusammenführung und Strukturelle Integration

Der Wunsch, verteilte Datenbestände anzufragen, ist seit langem Triebfeder und Thema der Forschung. Unterschieden werden muss hier zwischen der Informationsintegration und der Datenintegration. Die Datenintegration betrachtet vorrangig die Zusammenführung von verteilten, unterschiedlichen Datenbanken, wobei hier im Gegensatz zur Datenbankföderation der lesende Zugriff im Mittelpunkt steht. Ein häufiges Szenario für die Anwendung von Datenintegrationsverfahren im wirtschaftlichen Umfeld ist die Zusammenführung verschiedener Datenbestände in einem Datawarehouse. Im Rahmen der Informationsintegration stehen im Vergleich zur Datenintegration eher semistrukturierte Daten im Mittelpunkt der Betrachtung, da die Daten sind weniger für die automatische Verarbeitung als für die menschliche Information gedacht sind. Eine Folge ist, die Verwendung von automatischen Verfahren zum Schema-Matching erschwert wird. Auch die Anwendung von auf Dateninstanzen basierenden Verfahren wird durch die breite Spannbreite an Ansetzungsformen erschwert.

Die genannten Eigenschaften der bei der Informationsintegration betrachteten Daten erzwingen in der Praxis einen intellektuellen Ansatz zur Datenintegration, was die Frage nach geeigneten Verfahren und Ansätzen aufwirft. Ein aus der Datenintegration entnommener, auch für die strukturelle Informationsintegration anwendbarer Ansatz ist das Konzept der 'Global as View' (siehe[Halevy2006]). Hierbei werden die zu integrierenden Datenbestände in einem gemeinsamen Integrationsschema zusammengefasst. Wrapper nehmen für jeden Datenbestand die Überfüh-

rung in ein gemeinsames Schema vor, führen also entweder Daten zusammen, wenn das Integrationsschema weniger feingranular ist als das Quellschema oder nehmen die Aufspaltung oder Aufbereitung von weniger fein strukturierten Daten vor.

Die Zusammenführung von Daten und der Einsatz von Wrappern kann sowohl im Rahmen einer einmaligen Datenaufbereitung geschehen als auch zum Anfragezeitpunkt. Werden Aspekte wie das Caching vernachlässigt, bedeutet die letztere Lösung eine Erhöhung der Anfragelast auf das jeweilige Einzelsystem. Insbesondere wenn Verfahren zur Relevanzermittlung im Sinne eines statistischen Information Retrieval zum Einsatz kommen, kann der Ressourcenverbrauch einer Anfrage der Integrationsplattform an die Einzelsysteme jenen aus regulären Nutzeranfragen bei Weitem übersteigen. Bei einer solchen Anfrage ist in der Regel neben den eigentlich nutzerrelevanten Informationen die Übertragung von statistischen Hilfsinformationen notwendig. Die Zusammenführung aller Datenbestände an einer Stelle hat unter den Aspekten der Anfrageperformanz Vorteile, ein potentielles Problem ist jedoch neben organisatorischen Aspekten die Aktualität des Datenbestands, da die Konvertierung von Daten zum Indizierungszeitpunkt stattfindet.

Generell liegt die Herausforderung des Prozesses der strukturellen Informationsintegration in der Definition eines geeigneten Integrationsschemas. Die Eignung ist hier sowohl von den Anforderungen abhängig, die sich an das integrierte Datenformat stellen als auch von den zu integrierenden Daten.

3 Semantische Heterogenität

Der Begriff der semantischen Heterogenität bezeichnet unter anderem die Verwendung von verschiedenen Verschlagwortungsvokabularen in den Informationsbeständen eines integrierten Informationssystems. Zur Behandlung der semantischen Heterogenität existieren verschiedene Ansätze. Zu unterscheiden sind in erster Linie die intellektuelle Erstellung von Crosskonkordanzen und die Verwendung von statistischen Verfahren zur Zuordnung von Termen zwischen Vokabularen. Mayr und Walter stellen verschiedene Verfahren vor und gehen auf Methoden zu Evaluation ein (siehe [Mayr07]). Durch die Verwendung von Crosskonkordanzen wird ein Graph definiert, der entlang seiner Kanten die Umsetzung von Anfragen in fremde Verschlagwortungsvokabulare erlaubt, Abbildung 2 stellt einen solchen Graphen dar. Abbildung 3 zeigt die Umwandlung einer Anfrage nach “Bildungseinrichtung”, eines Terms auf dem Thesaurus Sozialwissenschaften. auf verschiedene Zielvokabulare, hier die Schlagwortnormdatei und den Psyndex-Thesaurus.

Im Rahmen der semantischen Integration muss auch die Behandlung von Datenbeständen unterschiedlicher Qualität betrachtet werden, insbesondere bezüglich der Qualität der inhaltlichen Erschließung, Krause definiert hierfür das Schalenmodell (siehe [Krause2004])

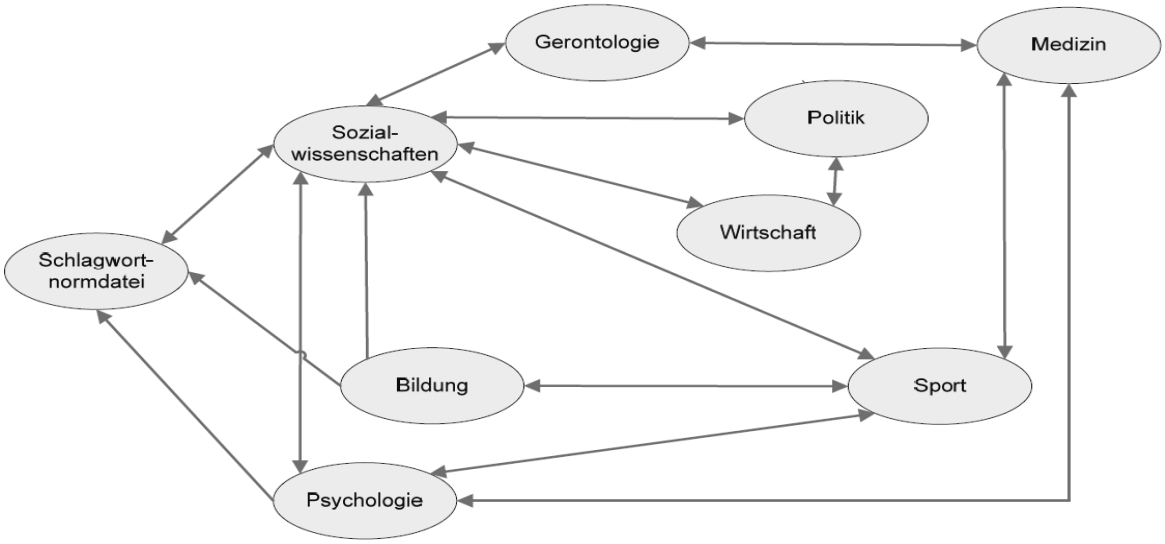

Abbildung 2: Von Crosskondordanzen aufgespannter Graph zur Behandlung semantischer Heterogenität aus [Mayr2007]

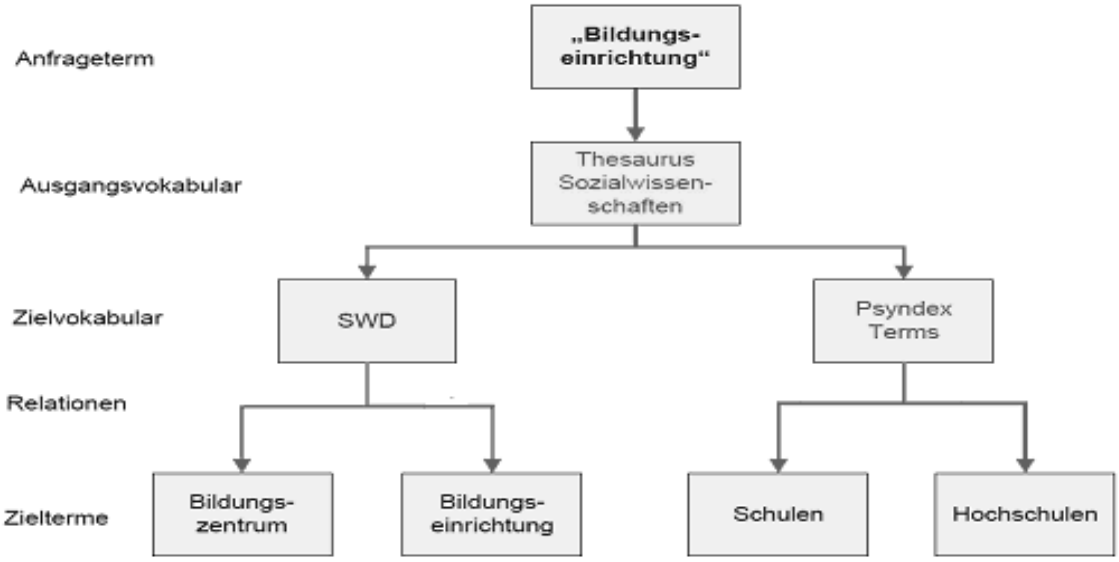

Abbildung 3: Beispielanfrage bei Verwendung von Crosskonkordanzen nach [Mayr2007, modizifiert]

4 Lösungsansätze für verschiedene Anwendergruppen

Die Anwendergruppen, die ein System nutzen, sind in hohem Maß unterschiedlich. Geht man davon aus, dass das betrachtete Integrationsprojekt vertikal orientiert ist,

also inhaltlich oder thematisch spezialisiert ist, so bestehen noch immer Unterschiede im konkreten Informationsbedürfnis und in den bekannten oder bevorzugten Methoden, dieses Informationsbedürfnis auszudrücken. Internetsuchmaschinen wie Google haben bezüglich der Anfragedarstellung einen Standard gesetzt, an dem sich Informationsanbieter wohl oder übel messen lassen müssen, zugleich ist die Verwendung einer einfachen Suchzeile, wie sie von Google propagiert wird, nicht in allen Fällen geeignet für die Formulierung komplexer Anfragen.

Ein Mittel, die verschiedenen Nutzererwartungen und -anforderungen zu behandeln, ist die Bereitstellung von spezifischen Anfrageschnittstellen für unterschiedliche Nutzergruppen. Typische Beispiele sind hier die *Erweiterte Suche* oder die *Expertensuche*, die von vielen Informationsanbietern in Ergänzung zu einer *Einfachen Suche*, meist einer einzelnen Suchzeile, angeboten werden.

Eine Herausforderung an die verschiedenen Recherchezugänge ist die Beantwortung der Frage, wie ein Anwender bei der Formulierung seines Informationsbedürfnisses unterstützt werden kann. Ein Beispiel für eine solche Unterstützung ist die automatische Anfrageerweiterung, etwa die Umwandlung eingehender Nutzeranfragen mittels Crosskonkordanzen so dass der Recall einer Anfrage erhöht wird. Bei der Anwendung von Anfrageerweiterungen muss beachtet werden, dass diese für den Anwender nachvollziehbar sind und ihn gegebenenfalls in die Lage zu versetzen, in Zukunft selbst optimierte Anfragen zu erstellen.

Neben der Anfrageerweiterung besteht ein weiterer Ansatz, mit unterschiedlichen Voraussetzungen seitens der Anwender umzugehen in der Bereitstellung von Informationen über den Inhalt der Datenbank. Ein Mittel, dies zu leisten sind Facetten oder Navigatoren. Navigatoren bieten eine intuitive Möglichkeit, eine weit gefasste Anfrage einzuschränken; zugleich bieten sie Information zum Inhalt der Datenbank, die Nutzern bei der Formulierung von neuen Anfragen hilfreich sein können. Abbildung 4 zeigt drei Navigatoren aus dem SOWIPORT Portal die des erlauben eine Treffermenge anhand der Quelldatenbank oder anhand von Personen oder Schlagworten aus den bei einer Anfrage gefundenen Dokumenten einzuschränken.

Eine Herausforderung, die nicht durch Navigatoren behandelt werden kann, ist die Darstellung der verbliebenen Heterogenität von Datenbeständen in Fällen, in denen die strukturelle oder semantische Heterogenität nicht in vollem Umfang behandelt werden konnte. So soll einem Anwender etwa konkret vermittelt werden, dass in einer bereitgestellten Literaturdatenbank grundsätzlich keine Zeitschriftenartikel enthalten sind oder welche Selektivität die konkreten Terme einer booleschen Suchanfrage aufweisen. Die Integration entsprechender Informationen in

Benutzungsoberflächen in einer für den Anwender verständlichen Form ist noch immer Gegenstand der Forschung, ein Beispiel für einen entsprechenden Ansatz ist ODIN (siehe [Stempfhuber2002]), dargestellt in Abbildung 5.

Abbildung 4: Navigatoren zur Einschränkung einer Treffermenge anhand von Personennamen, Schlagworten und der Quelldatenbank eines Ergebnisses

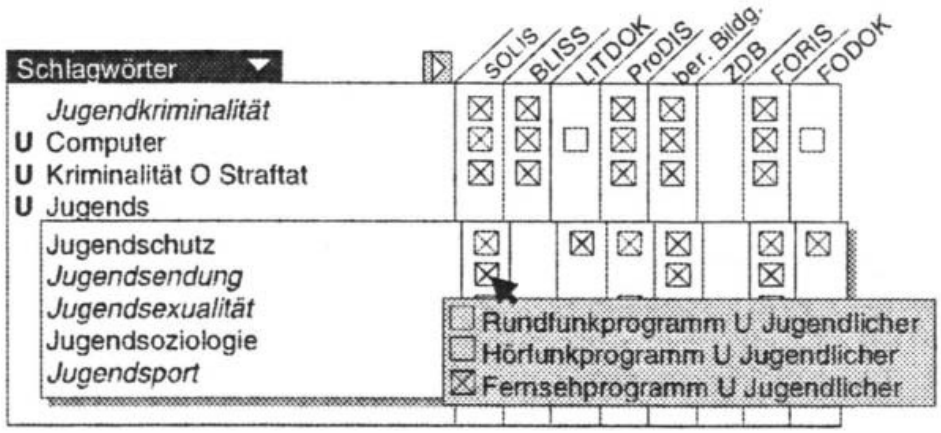

Abbildung 5: Behandlung semantischer Heterogenität durch die ODIN-Benutzungsoberfläche nach [Stempfhuber2002]

Eine Kernfrage moderner Informationssysteme ist die Frage, wie die Relevanz von Informationen in Bezug auf das Informationsbedürfnis eines Anwenders ermittelt werden kann. Gegenüber dem Booleschen Modell finden aktuell Modelle wie das statistische Information Retrieval oder das erweiterte boolesche Modell zunehmend Verbreitung. Diese Modelle bieten eine gestufte Bewertung der Relevanz von Ergebnisdokumenten, womit sich die Frage stellt, anhand welcher Eigenschaften eines

Dokuments sich die Relevanz ermitteln lässt. Das Information Retrieval sieht die Ermittlung eines Relevanzwertes anhand des Inhalts eines Dokuments vor, wobei in der Regel unterschiedliche Dokumentenbestandteile als unterschiedlich wichtig angesehen werden. So wird ein Term im Titel eines Dokuments als wichtiger erachtet als ein Treffer im Textkörper. Ergänzend zum Inhalt können die Positionen von Wörtern im Verhältnis zueinander betrachtet werden, so dass im Dokument nahe beieinander stehende Suchbegriffe eine höhere Relevanz erhalten. Topologische Verfahren wie Googles Pagerank betrachten die Verlinkung von Webseiten untereinander, im Bereich der Literaturrecherche können durch Zitationsanalysen ähnliche Verfahren angewandt werden. Externe Kriterien zur Dokumentenbewertung beziehen das Alter eines Dokuments in die Bewertung ein, wobei die Annahme zugrunde liegt, dass ältere Dokumente weniger relevant sind als jüngere, eine weitere Möglichkeit ist die Einbeziehung eines intellektuell oder automatisch vergebenen Qualitätswertes in die Relevanzermittlung.

Das Vorhandensein dieser Vielzahl von verschiedenen Relevanzkonzepten macht es schwer, ein allen Nutzeranforderungen entsprechendes, integriertes Konzept zu entwickeln. Vielmehr stellt sich die Frage, wie die einzelnen Aspekte der Relevanzermittlung einem Anwender gegenüber präsentiert werden können, um diesem die Auswahl einer geeigneten Methode der Relevanzermittlung zu gestatten.

5 Heterogenitätsbehandlung in SOWIPORT

Bei der Konzeption und Umsetzung des sozialwissenschaftlichen Fachportals SOWIPORT sah sich das IZ Sozialwissenschaften mit den in diesem Papier genannten Dimensionen konfrontiert. Dieser Abschnitt stellt die Verfahren der Heterogenitätsbehandlung dar, die bei SOWIPORT Verwendung fanden und gibt einen Ausblick des geplanten weiteren Vorgehens.

SOWIPORT verfolgt einen zentralisierten Ansatz der Datenintegration: Alle im Portal durchsuchbaren Daten werden für eine Indizierung gesammelt und in ein gemeinsames Datenformat gebracht. Aufgrund der Heterogenität der enthaltenen Datenbestände und um bezüglich der zu unterstützenden Anfrageverfahren größtmögliche Flexibilität zu wahren, wurde ein Datenformat mit einem begrenzten Kernbestand von Konzepten formuliert das im Bedarfsfall für die Besonderheiten der jeweiligen Datenbestände spezialisiert werden kann. Die Zweistufigkeit von grundlegenden Konzepten zusammen mit der Möglichkeit der Spezialisierung erlaubt es, beim Entwurf des Schemas noch ungekannte Konzepte auszudrücken, hält das Datenformat zugleich aber übersichtlich.

Zurzeit wird für SOWIPORT eine zentrale Indizierungsstrategie verfolgt, alle beteiligten Partnerdatenbanken werden in einem zentralen, gemeinsamen Index zusammengefasst. Die Datenintegration findet dabei sowohl auf Ebene eines für die Darstellung verwandten, detaillierten XML-Datenformats als auch im Rahmen der Indizierung durch die FAST Suchmaschine statt, Abbildung 6 stellt den Datenfluss dar. Für die Zukunft ist hierüber hinaus auch die Einbindung externer Datenquellen vorgesehen die neben dem FAST Index angesprochen werden.

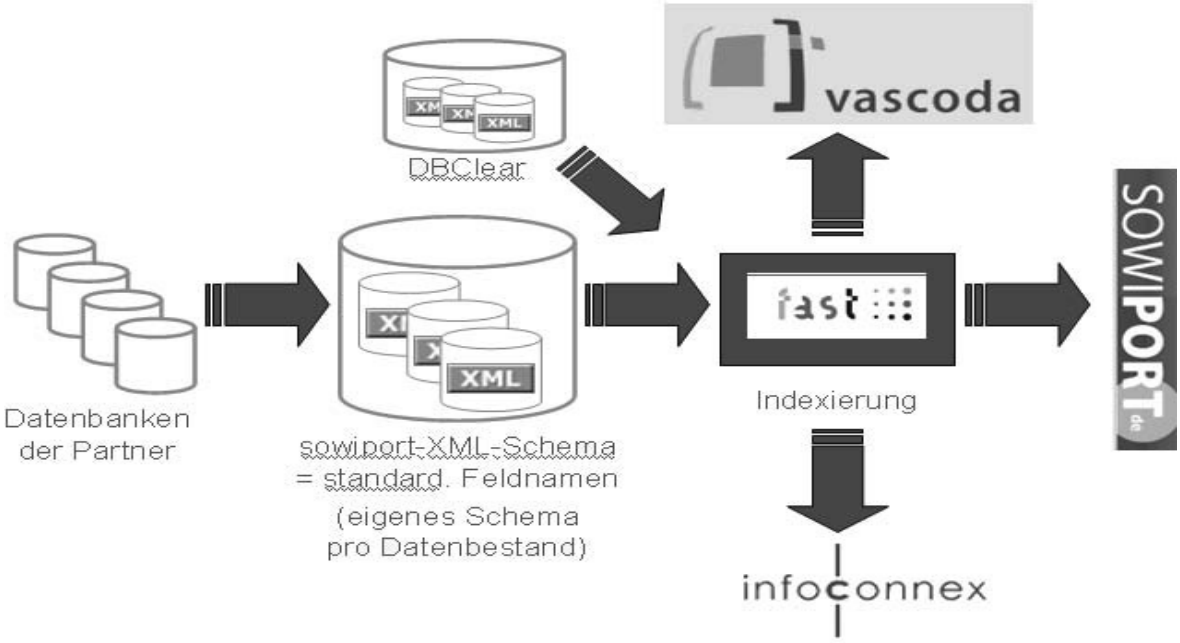

Abbildung 6: Verarbeitungs- und Indizierungspfade zur Bereitstellung heterogener Daten in SOWIPORT.

Bezüglich der Anfrage- und Ergebnisdarstellung stellt SOWIPORT zurzeit eine feldbasierte Suche in einer einfachen und einer erweiterten Version bereit um verschiedene Nutzergruppen anzusprechen. Mittels verschiedener, auswählbarer Kriterien zur Relevanzbestimmung werden diesbezüglich unterschiedliche Nutzeranforderungen unterstützt; die Verwendung von Facetten bietet die Möglichkeit zur Einschränkung von Suchanfragen.

Die Behandlung der semantischen Heterogenität geschieht mittels Crosskonkordanzen, diese werden bei der Anfrageabarbeitung automatisch auf Schlagwortsuchen angewandt. Die Verwendung der Crosskonkordanzen soll in weiteren Ausbaustufen durch die Bereitstellung eines Heterogenitätsservice erweitert werden, neben der Erweiterung von Anfragen wird dieser auch in der Lage sein, Anwender in einem interaktiven Prozess bei Anfragen mittels kontrollierten Vokabularen zu unterstützen.

6 Literatur

R. Baeza-Yates; B. Ribeiro-Neto; et al.: Modern information retrieval, 1999 Verlag Addison-Wesley Harlow, England

Alon Halevy; Anand Rajaraman; Joann Ordille: Data Integration: The Teenage Years. In: Conference on Very Large Data Bases '06, 12-15 September, 2006, Seoul, Korea.

Krause, Jürgen: Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution. In: IFLA Journal: Official Journal of the International Federation of Library Associations and Institutions 30, Nr. 4, 2004, S. 310 – 318

Mayr, Philipp; Walter, Anne-Kathrin (erscheint): Zum Stand der Heterogenitätsbehandlung in vascoda: Bestandsaufnahme und Ausblick. In: Bibliothek & Information Deutschland (Hrsg.): 3. Leipziger Kongress für Information und Bibliothek, 19. – 22. März 2007. Leipzig: Verlag Dinges & Frick.

S. Raghavan; H. Garcia-Molina: Crawling the Hidden Web. In: Proceedings of the 27th International Conference on Very Large Data Bases, 2001, Rom, S. 129—138

Stempfhuber, Maximilian: Objektorientierte Dynamische Benutzeroberflächen ODIN, Dissertation an der Universität Koblenz-Landau, 2002