Jürgen Krause

Konkretes zur These, die Standardisierung von der Heterogenität her zu denken

Foto privat

Jürgen Krause

Prospective developments in the area of scientific information are specialized innovative and integrative portals which are collectively accessible through one scientific gateway and bring together the basic interdisciplinary access paths in libraries with subject-based gateways. Within this structure the customer has access to scientific information (bibliographic citations, references to experts and research programs, online texts, materials, data, facts, lists of links, etc.) via high quality search and select tools. Both theory-based analyses and state-of-the-art reports of the information scene as well as the findings of newer user surveys regarding information-seeking behavior and information needs among scientists indicate that such a development is desirable. Today there is a broad consensus about what goals are to be achieved. The challenge for further developments lies not in finding acceptance for these goals, but in deciding how they are to be achieved. This paper discusses the development of bi-lateral transfer components for the treatment of heterogeneous semantics among document collections with different descriptive cataloguing styles and presents a viable solution for an essential portion of the question as to how this goal can be achieved. The paper reports on theoretical and practical aspects, describes the current state of developments and presents specific instances for application.

Die Entwicklungsperspektive für den Bereich wissenschaftlicher Information sind innovative, integrierende Fachportale, die in einem Wissenschaftsportal zusammengefasst werden und die allgemeinen, fachübergreifenden Zugänge der Bibliotheken mit spezifischen Fachzugängen verbinden. In dieser Struktur kann der Kunde mit qualitativ hochwertigen Such- und Selektionsinstrumenten auf wissenschaftsrelevante Informationen (Literaturnachweise, Experten und Forschungsreferenzen, Volltexte, Materialien, Daten, Fakten, Linklisten etc.) zugreifen.

Sowohl theoriegeleitete Analysen und Bestandsaufnahmen der wissenschaftlichen Informationslandschaft als auch die Ergebnisse der neueren Benutzerumfragen zum Informationsverhalten und zum -bedarf von Wissenschaftlern weisen auf die Wünschbarkeit solch einer Entwicklung hin. Heute ist ein weitgehender Konsens über das anzustrebende Ziel erreicht. Die Herausforderung für die Weiterentwicklung ist somit nicht die Akzeptanz der angestrebten Zielvorstellung, sondern die Frage, wie sie zu erreichen ist.

Die im Folgenden diskutierte Entwicklung von bilateralen Transferkomponenten zur Behandlung semantischer Heterogenität zwischen Dokumentensammlungen mit unterschiedlicher Inhaltserschließung zeigt für einen wesentlichen Teil der Frage nach dem »Wie« der Zielerreichung eine tragfähige Lösungsstrategie auf. Sie wird theoretisch und praktisch konkretisiert, der Entwicklungsstand beschrieben und die konkreten Einsatzmöglichkeiten werden aufgezeigt.

»Nebeneinander« von Informationsangeboten

Ausgangslage

Am Beginn der Einrichtung fachbezogener Informationsservicestellen wie dem Informationszentrum Sozialwissenschaften in Bonn (IZ) standen vor über 30 Jahren neue Möglichkeiten der informationstechnischen Entwicklung. Suchten Wissenschaftler bis dahin ihre Literatur in den Katalogen der Bibliotheken, eröffneten jetzt Datenbanken ökonomischere Möglichkeiten der Erfassung, Erschließung und Suche nach wissenschaftlichen Dokumenten. Die Bibliotheken stellten sich diesen Innovationsanforderungen damals nicht. Dies führte in der Konsequenz zu einer neuen Ausdifferenzierung der wissenschaftlichen Informationsversorgung. Fachbezogene Informationsservicestellen und Fachinformationszentren bauten Literaturdatenbanken zu einzelnen Fachgebieten auf und wiesen neben selbständiger Literatur auch Zeitschriftenartikel nach. Die Zeitschriften selbst und die Ausleihfunktion verblieben bei den Bibliotheken.

In den letzten Jahren stellt erneut ein einschneidender informationstechnischer Entwicklungsschub – vor allem bedingt durch die Ausbreitung des WWW – die althergebrachten Formen der wissenschaftlichen Informationsversorgung in Frage. In diesem Kontext erweist sich besonders die Trennung von Funktionen der wissenschaftlichen Informationsversorgung, die zu separaten Teilangeboten führt, als ärgerliche Hürde. Warum soll eine Institution Zeitschriftenartikel und Bücher nachweisen (und die Benutzer zum Lesen wieder an die Bibliotheken verweisen) und die andere nur Bücher? Reformuliert man diese Frage aus dem Blickwinkel der Benutzer fachwissenschaftlicher Angebote und weitet sie um Anbieter wie Verlage und die Wissenschaftler selbst und auf die Überlappungsbereiche verschiedener Fachdatenbanken aus, lautet sie: Was bewirkt dieses Nebeneinander von Informationsangeboten, die mittlerweile alle elektronisch zugänglich gemacht wurden, für den Wissenschaftler?

Ein Beispiel: Ein Benutzer des IZ sucht in der sozialwissenschaftlichen Literaturdatenbank SOLIS mit dem Begriff *Staatsfunktion* und findet hierzu auch einige Nachweise. Da er wegen der zwischen den Fächern bestehenden Überschneidungsbereiche und vermuteter interdisziplinärer Arbeiten auch im Bereich Pädagogik nach Dokumenten suchen möchte, geht er in die Datenbank FIS Bildung des DIPF (Deutsches Institut für Internationale Pädagogische Forschung), die eigene, von der des IZ abweichende Schlagwörter benutzt. *Staatsfunktion* führt zu keinem Treffer, wohl aber die Verbindung von *Staat* und *Funktion* oder *Staat* und *Aufgabe*. Bei Verlagsangeboten ist die Frage, ob er mit einer dieser Schlagwortvarianten etwas findet, völlig offen, da sie in der Regel weder der Bibliotheks- noch der IZ-Norm folgen. Die Inhaltserschließungssysteme

der verschiedenen Anbieter können sich also beträchtlich unterscheiden. Unterschiede bestehen bereits zwischen den Bibliotheken und dem IZ hinsichtlich ihrer Erschließungsart und -intensität (die bibliothekarischen Regeln für den Schlagwortkatalog RSWK, Fachthesauri z. B. für die Sozialwissenschaften, verschiedene Klassifikationen, freie Schlagwörter, strukturierte Begriffsmengen usw.) wie auch in Bezug auf Inhalt und Struktur (reine Sachtitel, Kurzreferate, Volltexte). Sie verstärken sich deutlich, wenn weitere Informationsanbieter hinzukommen. Will der Benutzer somit Literaturinformationen aus unterschiedlichen nicht verbundenen, heterogenen Organisationsstrukturen und Zugänglichkeitskontexten, den Universitätsbibliotheken, den Fachdatenbanken, den Verlags- und Wissenschaftlerservern im WWW usw. recherchieren, muss er diese Unterschiede kennen, in Suchstrategien umsetzen und bei der mehrfachen Anfrageformulierung berücksichtigen. Dies ist in der Praxis die Ausnahme und bei mehr als zwei Anbietern auch eine Zumutung.

Durch die Entwicklung der letzten Jahre verschärfte sich das Problem erheblich:

Auf dem WWW-Institutsserver der Universitäten stellen Wissenschaftler ihre Forschungsergebnisse und Inhalte aller Art in verschiedenster Form zur Verfügung.

Auch beim WWW war internationale Standardisierung zur Konsistenzerhaltung die Leitidee. Durch neue Metadaten-Ansätze wie die Dublin Core Initiative (DC) soll die im WWW verloren gegangene Konsistenz aus der Welt der Bibliotheken und Fachdatenbanken teilweise neu etabliert werden (s. Krause / Niggemann / Schwänzl 2003). Im Gegensatz zu den Bibliotheken und Fachdatenbanken sind die Ziele der Etablierung von WWW-Metadaten bescheidener: Hier sind sie (oft nicht eingehaltene) Übereinkünfte, bestimmte Merkmale eines Dokumentbestandes in einer verabredeten Form bei den eigenen Daten auszuweisen, wie verschieden sie in Bezug auf andere Merkmale auch immer sein mögen.

Virtuelle Fachbibliotheken, in denen Bibliotheken mit den anderen Informationsanbietern kooperieren, wollen nicht nur mit intellektuell erstellten Verzeichnissen von WWW-Quellen, den Fachinformationsführern, Orientierungshilfe leisten. Sie wollen nicht nur textuelle Informationen verschiedenster Art erschließen, sondern idealiter gleichzeitig Fakten (z. B. die Zeitreihen von Umfragen) oder auch Lehrmaterialien integriert in verlässlicher bibliothekarischer Qualität nachweisen. Erreicht haben sie dieses Ziel allerdings bei weitem noch nicht (s. www.virtuellefachbibliothek.de/).

Informationsverbünde entwickeln eigene Volltextserver zu kommerziellen Zeitschriften, in der Mehrzahl der Fälle zusätzlich zum bestehenden Angebot einer fachwissenschaftlichen Literaturdatenbank. Sie werden seit August 2003 zusammen mit den Fachinformationsführern der virtuellen Fachbibliotheken unter dem einheitlichen Dach des Wissenschaftsportals vascoda technisch integriert angeboten (s. www.vascoda.de).

Die Verlagsserver der großen Anbieter wie Elsevier erschließen ihr Angebot nach eigenen Standards und mit eigener Zugriffssoftware.

Literaturrecherche: eine Zumutung

Internationale Datenbanken wie Sociological Abstracts, die Datenbank für anglo-amerikanische soziologische Veröffentlichungen, bieten wiederum eigene Zugriffswege auf der Basis unterschiedlicher Inhaltserschließungsverfahren an.

Die Konsequenz dieser Vielfalt ist, dass Benutzer informationeller Dienste heute einem hochgradig dezentralisierten und heterogenen Dokumentenraum gegenüberstehen, mit der Folge unterschiedlichster Konsistenzbrüche:

heterogener Dokumentenraum mit Konsistenzbrüchen

Relevante, qualitätskontrollierte Daten stehen neben irrelevanten und eventuell nachweislich falschen. Kein Gutachtersystem sorgt für eine Trennung von Ballast und potentiell erwünschter Information.

Ein Deskriptor X kann in einem solchen System die unterschiedlichsten Bedeutungen annehmen. Auch im engen Bereich der Fachinformation kann ein Deskriptor X, der aus einem hochrelevanten Dokumentenbestand mit viel Aufwand intellektuell und qualitativ hochwertig ermittelt wurde, z. B. nicht mit dem Term X gleichgesetzt werden, den eine automatische Indexierung aus einem Randgebiet liefert.

Dass der Benutzer trotzdem auf keines dieser Teilangebote verzichten will, weil jedes für sich die Chance erhöht, die für ihn relevanten Informationen zu erhalten, zeigt eine Reihe von Benutzerbefragungen der letzten Zeit. Von besonderem Interesse ist hierbei die Studie zum Informationsverhalten und Informationsbedarf der Wissenschaft, über die im gleichen Heft berichtet wird (Poll 2004, Boekhorst / Kayß / Poll 2003; s. auch Machill / Welp 2003 für die allgemeine, nichtwissenschaftliche Nutzung des WWW und Stahl et al. 1998 und 2002, IMAC 2002 für den sozialwissenschaftlichen Bereich).

Wünsche nach tieferer Erschließung

Fasst man die in unserem Kontext wesentlichen Ergebnisse der genannten Benutzerbefragungen zusammen, ergibt sich – exemplarisch auf die Sozialwissenschaften bezogen – folgendes Bild:

Zusätzlich zum Konzept der Bibliotheken, allgemeine, fachübergreifende Zugänge zu erarbeiten, wird ein fachwissenschaftlicher Zugang gewünscht, der über tiefere Erschließungsinstrumente verfügt.

Benachbarte Fächer mit Überschneidungsbereichen wie Mathematik – Physik oder Sozialwissenschaften – Pädagogik – Psychologie – Wirtschaft sollen für die Abfrage ein integriertes Cluster bilden.

kritische Sicht auf Suchmaschinen

Die Qualität der Ergebnisse muss deutlich über der heutiger allgemeiner Internet-Suchmaschinen liegen (kein »Müll«). Interessant ist, dass die frühere unkritische Sichtweise auf die allgemeinen Internet-Suchmaschinen bei den neuesten Umfragen einer realistischeren, kritischen Einstellung gerade bei jüngeren Wissenschaftlern, die das WWW deutlich häufiger nutzen, weicht. Sie werden nicht mehr als Ersatz von OPACs oder Fachdatenbanken angesehen, sondern als Ergänzung.

Nicht nur Metadaten und Abstracts werden bei der Literatursuche gefordert, sondern auch der direkte Zugang zu Volltexten. Der Beschaffungsaufwand durch den Gang in die Bibliothek oder durch die Fernbestellung und der damit verbundene Zeitverzug sollen entfallen.

Nicht nur Literatur-Fachdatenbanken und Bibliotheks-OPACs, sondern auch Forschungsprojektdaten, Institutionenverzeichnisse, WWW-Quellen und Faktendaten sollen unter einem fachwissenschaftlichen Portal integriert werden.

hochintegriertes Angebot von Teilkomponenten

Alle Teilkomponenten sind möglichst hochintegriert am Arbeitsplatz anzubieten. Der Benutzer will – wie bei der Beratung durch einen Informationsvermittler oder Experten – nicht zwischen verschiedenen Datentypen unterscheiden und mehrfach seine Frage unterschiedlich formulieren, sondern sein Informationsbedürfnis direkt und nur einmal ausdrücken: »Ich möchte Informationen zum Thema X«.

Diese Forderungen gehen deutlich über die heutigen Angebote der Bibliotheken und der fachspezifischen Informationszentren hinaus. Besonders interessant ist, dass sich die erfahrensgeleiteten Einschätzungen der Mehrzahl der wissenschaftlichen Nutzer, die sich in den Ergebnissen der Benutzerbefragungen widerspiegeln, derzeit weitgehend mit den analytischen Ableitungen aus Bestandsaufnahmen wie Krause/Niggemann/Schwänzl 2003 und den theoriegeleiteten Erkenntnissen von Informationswissenschaft und Information Retrieval decken. Gleichzeitig gibt uns die technische Entwicklung die Mittel in die Hand, solche Vorstellungen und Benutzerwünsche zumindest längerfristig auf hohem Qualitätsniveau umzusetzen.

Somit besteht heute ein weitgehender Konsens über das zu erreichende Ziel. Die Entwicklungsperspektive sind neben den generellen, fachübergreifenden Einstiegspunkten der OPACs Fachportale, die die obigen Benutzerwünsche erfüllen. Mit ihnen kann der Wissenschaftler mit innovativen Such- und Selektionsinstrumenten auf wissenschaftsrelevante Informationen (Literaturnachweise, Experten und Forschungsreferenzen, Volltexte, Materialien, Daten, Fakten, Linklisten etc.) zugreifen. Gleichzeitig müssen fachbezogene Informations- und Dokumentationsdienstleistungen zunehmend um interdisziplinäre (Clusterbildungen wie infoconnex, Wissenschaftsportal vascoda) und internationale Komponenten erweitert und mit den generellen Zugängen der Bibliotheken verknüpft werden, eine Strategie wie sie z. B. das deutsche Wissenschaftsportal vascoda verfolgt (s. www.vascoda.de).

Perspektive: Fachportale

Zentral bei dem hier vertretenen Leitbild von Fachportalen als Basis jeder weiter gehenden Integration ist die Formel »qualitativ und intelligent integrierend« und ihre Auslegung. Nicht gemeint sind rein technische Integrationsleistungen, wie sie bei allgemeinen Internet-Suchmaschinen oder den heutigen Bibliotheksverbünden zu beobachten sind. Schon hier wird von »Portalen« als integrierendem Zugriff auf Teilbereiche von Wissen oder als Einstiegspunkt im Sinne eines *»one stop shops«* gesprochen. Dieser Begriffsinflation – und mit ihr der Gefahr von Missverständnissen – ist nur schwer zu begegnen. Der Ausweg, die Terminologie zu wechseln, ist aber versperrt. Der Portalbegriff ist weit verbreitet und an sich adäquat, die Zielvorstellung damit semantisch-pragmatisch gut erfassbar. Zu begegnen ist seiner Trivialisierung, die das »Wie« der Zielerreichung inadäquat simplifiziert.

Die Herausforderung für die Weiterentwicklung der heutigen Informationslandschaft ist somit nicht die Akzeptanz der angestrebten Zielvorstellung, sondern die Frage, wie sie zu erreichen ist. Die im Folgenden diskutierte Entwicklung von Transferkomponenten zur Behandlung semantischer Heterogenität zwischen Dokumentensammlungen mit unterschiedlicher Inhaltserschließung zeigt für einen wesentlichen Teil der Frage nach dem »Wie« der Zielerreichung eine neue tragfähige Lösungsstrategie auf. Sie soll in den folgenden Abschnitten konkretisiert, der Entwicklungsstand beschrieben und die praktischen Einsatzmöglichkeiten sollen aufgezeigt werden.

Technische Integration und Grenzen der Standardisierung bei dezentraler Organisation

Bemühungen, die Homogenität und Konsistenz in der heutigen dezentralen Informationswelt wiederherzustellen, setzen auf die Schaffung geeigneter Informationssysteme, die mit verteilten Datenbeständen effizient umgehen können und auf die Einhaltung von Normierungen und Standards.

Ersteres Vorgehen steht für die technikorientierte Sichtweise von Problemlösungen. Man sorgt – wie bei

den generellen WWW-Suchmaschinen, aber begrenzt auf Server der Wissenschaften – dafür, dass physikalisch auf die verschiedenen Dokumentenräume gleichzeitig zugegriffen werden kann und dass dies effizient geschieht. Beispiele für die heute erreichte technische Integration heterogener Datenbestände sind der Bibliotheksverbund KOBV (www.kobv.de/se/cont.html) oder der virtuelle Bibliotheksverbund NRW. Was die Bibliotheksverbünde und vergleichbare Projekte bisher nicht tun, ist, die verschiedenen Inhaltserschließungsverfahren der Teilbestände adäquat zu berücksichtigen. Damit ergibt sich zwar im Vergleich zu den Schwächen der generellen WWW-Suchmaschinen eine Verbesserung bei der Auswahl der relevanten Datenmaterialien, die konzeptuellen Unterschiede zwischen verschiedenen Inhaltserschließungsverfahren werden jedoch auch hier nicht ausgeglichen.

Einen Schritt weiter gehen die Ansätze zur Einführung einheitlicher Metadaten, die über die bisherigen Bibliotheksverbünde und das traditionelle Web-Angebot hinausreichend standardisieren möchten. Metadaten sind Übereinkünfte, bestimmte Merkmale eines Dokumentbestandes in einer verabredeten Form bei den eigenen Daten auszuweisen, wie verschieden sie in Bezug auf andere Merkmale auch immer sein mögen. Ein Beispiel aus der Bibliotheksszene ist der derzeit versuchte Anschluss der deutschen Bibliotheken an die amerikanische Klassifikation DDC (Dewey Decimal Classification) oder für WWW-Quellen das Metadatenschema Dublin Core (DC, s. Weibel 1997). Aber nicht einmal den Bibliotheken ist es in der Vergangenheit für den deutschsprachigen Raum gelungen, den Standardisierungsanspruch wirklich durchzusetzen (s. Krause/Niggemann/Schwänzl 2003).

Trotzdem ist im gesamten Bereich der Fachinformation die klassische Normierungs- und Standardisierungsphilosophie, der auch die theoretischen Grundlagen der Inhaltserschließung in den Informations- und Bibliothekswissenschaften verhaftet sind, ungebrochen. Nach einem normierten, intellektuell kontrollierten Verfahren, das eine Zentralstelle entwickelt und durchsetzt, erfolgt eine einheitliche Erfassung der Dokumente. In diesem Denken kommt der Datenkonsistenz die höchste Priorität zu, wodurch der Benutzer (idealiter) immer einem homogenisierten Datenbestand gegenübersteht. Dieses Modell der Inhaltserschließung, das z. B. auch allen Bibliothekserschließungen zugrunde liegt, erwies sich in Teilbereichen wie bei den OPACs und Fachdatenbanken durchaus als ein gangbarer Weg, der sich in den letzten zwanzig Jahren bewährte. Was sich jedoch geändert hat, sind die Rahmenbedingungen. Normierungsbemühungen und Initiativen zur Akzeptanz und Verbreitung einheitlicher Metadaten sind fraglos wichtig und eine Voraussetzung für übergreifende Suchprozesse in einer täglich dezentraler und polyzentrischer werdenden Informationswelt. Sie versuchen, die verloren gegangene Datenhomogenität und Konsistenz der Inhaltserschließung durch freiwillige Absprachen aller am Informationsprozess Beteiligten wiederherzustellen. Solange man sich darüber im Klaren ist, dass dies nur teilweise gelingen kann, spricht alles für Initiativen dieser Art. Ganz gleich jedoch, wie erfolgreich die Einführung von Metadaten in einem Fachgebiet sein wird, die verbleibende Heterogenität z. B. in Bezug auf verschiedene Arten der Inhaltserschließung (automatisch, verschiedene Thesauri, verschiedene Klassifikationen, Unterschiede der erfassten Kategorien) wird zu groß sein, um sie zu vernachlässigen.

mangelnde Berücksichtigung der verschiedenen Inhaltserschließungsverfahren

Heterogenitätsbehandlung und bilaterale Transfermodule

Im Kontext fachwissenschaftlicher Informationen ist die Problematik der heterogenen und mehrfachen Inhaltserschließung generell besonders kritisch, weil die Heterogenität der Datentypen hoch ist. Z. B. sind Faktendaten, Literatur- und Forschungsprojektdaten gleichzeitig anzusprechen und neben die traditionellen Benutzergruppen der Fachinformation treten Zielgruppen, die nicht fachsprachlich, sondern umgangssprachlich anfragen werden.

Heterogenität als Zwangsläufigkeit

Zentral für die Behandlung dieser nicht vermeidbaren, nach Standardisierungsbemühungen zwangsläufig verbleibenden semantischen Heterogenität sind intelligente Transferkomponenten zwischen den verschiedenen Formen der Inhaltserschließung, die den semantisch-pragmatischen Differenzen Rechnung tragen. Sie interpretieren die technische Integration zwischen den einzelnen Datenbeständen mit unterschiedlichen Inhaltserschließungssystemen zusätzlich konzeptuell. Die Begriffswelt der fachspezifischen und generellen Thesauri, Klassifikationen, eventuell auch thematische Begriffsfelder und Abfragestrukturen begrifflicher Datensysteme usw. sind aufeinander zu beziehen. Das System muss z. B. wissen, was es heißt, wenn Term X aus einer fachspezifischen Klassifikation oder einem Thesaurus zur intellektuellen Indexierung eines Zeitschriftenaufsatzes benutzt wurde, die WWW-Quelle aber nur automatisch indexiert werden konnte. Term X dürfte sich nur zufällig in den Termen des Fließtextes finden lassen und dennoch gibt es konzeptuelle Bezüge zwischen den beiden Termgruppen, die auszuwerten sind.

Generell gibt es hierfür drei Verfahrensweisen. Keines der Verfahren trägt die Last des Transfers allein. Sie sind ineinander verschränkt und wirken zusammen.

Crosskonkordanzen zu Klassifikationen und Thesauri

Die verschiedenen Begriffssysteme werden im Anwendungskontext analysiert und der Versuch gemacht, ihre Begrifflichkeit intellektuell aufeinander zu beziehen. Crosskonkordanzen decken damit den statisch bleibenden Teil der Transferproblematik ab. Bei der Recherche bieten solche Verzeichnisse die Möglichkeit, Terme des einen Begriffssystems auf die des anderen auszuweiten, im einfachsten Fall im Sinne einer Synonymie- oder Ähnlichkeitsrelation, aber auch als deduktive Regelbeziehung.

Quantitativ-statistische Ansätze

Das Transferproblem lässt sich allgemein als Vagheitsproblem zwischen zwei Inhaltsbeschreibungssprachen beschreiben. Für die im Information Retrieval (IR) behandelte Vagheit zwischen den Termen der Benutzeranfrage und denen des Datenbestandes sind verschiedene Verfahren vorgeschlagen worden (probabilistische Verfahren, Fuzzy-Ansätze und neuronale Netze (s. Mandl 2001), die sich auf die Transferproblematik anwenden lassen.

Verfahren dieser Art benötigen Trainingsdaten, bei denen einzelne Dokumente nach zwei Begriffsschemata erschlossen sind. Für das multilinguale IR kann dies z. B. der gleiche Text in zwei Sprachen sein. Die Ausgangssituation für solche Verfahren ist bei der Zusammenführung von Bibliotheksbeständen mit Fachdatenbanken der Informationszentren in der Regel besonders günstig, da Informationszentren neben den Aufsätzen immer auch die selbstständige Literatur aufgenommen haben, die damit zumindest doppelt verschlagwortet vorliegt. Bei der Virtuellen Fachbibliothek Sozialwissenschaften sind z. B. alle Dokumente des Sondersammelgebiets der Universität Köln gleichzeitig in der sozialwissenschaftlichen Literaturdatenbank SOLIS erfasst.

Forderung nach Konsistenz und Interoperabilität

Im Unterschied zu den Crosskonkordanzen basiert die statistische Transformation nicht auf allgemeinen intellektuell ermittelten semantischen Beziehungen, sondern die Wörter werden in einen gewichteten Termvektor transformiert, der die Verwendung des Terms im Datenbestand widerspiegelt.

Qualitativ-deduktive Verfahren

Empirische Untersuchungen am Textmaterial einer virtuellen Fachbibliothek dürften deduktive Zusammenhänge offen legen, die mit Techniken aus dem Bereich der Expertensysteme zu behandeln sind.

Deduktive Komponenten finden sich beim Intelligenten Information Retrieval (Ingwersen 1996), bei intelligenten Recherchesystemen wie OSIRIS (Ronthaler/Zillmann 1998) und im Bereich der Expertensysteme.

Wesentlich ist, dass die postulierten Transfermodule bilateral auf der Ebene der Datenbestände arbeiten. Sie verbinden Terme der verschiedenen Inhaltsbeschreibungen. Dies ist konzeptuell – und in der praktischen Auswirkung – etwas anderes als die Behandlung der Vagheitsproblematik klassischer IR-Ansätze zwischen Benutzeranfrage und dem Dokumentenbestand der Datenbank. So können die Transfermodule z. B. zwischen einem Dokumentbestand, der mit einer generellen Schlagwortliste wie der Schlagwortnormdatei (SWD) indexiert wurde und einem zweiten, dessen Indexierung auf einem speziellen fachspezifischen Thesaurus beruht, durch qualitative Verfahren wie die Crosskonkordanzen und Deduktionsregeln aufeinander bezogen werden und der Rechercheralgorithmus kann die Verbindung zur Benutzerterminologie mit einem probabilistischen Verfahren herstellen. Diese Möglichkeit, dem Problem unterschiedlicher Begriffssysteme auf der Ebene der Dokumentenbestände begegnen zu können, ist ein wesentlicher Unterschied zu den bisherigen IR-Lösungen, die undifferenzierter bei der benutzerseitigen Behandlung des Vagheitsproblems ansetzen (s. auch Krause 2003).

Standardisierung von der verbleibenden Heterogenität her denken

Mit diesen Überlegungen zu den Heterogenitätsmodulen wurde eine neue Sichtweise auf die bestehen bleibende Forderung nach Konsistenzerhaltung und Interoperabilität gefunden, die sich durch die folgende Prämisse umschreiben lässt: Standardisierung ist von der verbleibenden Heterogenität her zu denken. Sie wird in Krause/Niggemann/Schwänzl 2003 ausführlicher erläutert und aus einer Bestandsaufnahme der heutigen dezentralen Informationslandschaft abgeleitet. Sie geht vor allem von der klaren Abkehr der fachwissenschaftlichen Benutzer vom geringen Qualitätsstandard rein technisch orientierter heutiger Web-Lösungen aus und adressiert die Art der Inhaltserschließung, ihre faktisch existierende und partiell nicht aufhebbare Unterschiedlichkeit zwischen den verschiedenen bestehenden Dokumentenbeständen und dem einzuschlagenden Lösungsweg.

Einzelverfahren und erreichter Stand

Als prinzipielle Verfahren zur bilateralen Bearbeitung konzeptueller Heterogenität wurden bisher Crosskonkordanzen und statistische Verfahren in einer Reihe von DFG- und BMBF-Projekten bis zur Einsatzreife entwickelt (s. den Überblick in Krause 2003). Erste Versuche gab es zum Einsatz neuronaler Netze (s. Mandl 2001). Sie bleiben hier genauso ausgeklammert wie die deduktiven Verfahren, die am Problem der au-

tomatischen Metadatenextraktion im BMBF-Projekt CARMEN erprobt wurden (s. Binder et al. 2002: Abschnitt 5.2). Bei der Metadatenextraktion haben die Transfermodule einen anderen Stellenwert in der Gesamtarchitektur. Man zieht sie heran, um fehlende Informationen (z. B. Metatag »Titel«) direkt bei der Dokumentenbeschreibung zu ergänzen.

Ausgeklammert bleiben auch weiter gehende Verfahren wie die in infoconnex geplanten Autorennetzwerke (s. Mutschke 2000) oder Cognitive-Mapping Ansätze (s. Mutschke/Quan-Haase 2001). Sie liefern nicht nur Relationen zwischen einzelnen Termen, sondern verknüpfen Begriffsgruppen mit zentraler Semantik zu semantischen Strukturen.

Crosskonkordanzen und qualitative Heterogenität

Crosskonkordanzen sind intellektuell erstellte Verbindungen zwischen Termen zweier Thesauri oder Klassifikationen. Sie enthalten neben reinen Synonymrelationen auch Oberbegriffs-/Unterbegriffs- und Ähnlichkeitsrelationen.

Die verschiedenen Begriffssysteme werden im Anwendungskontext analysiert und der Versuch gemacht, ihre generelle semantische Begrifflichkeit intellektuell aufeinander zu beziehen. Beispiele hierfür aus der Crosskonkordanz zwischen SWD und dem Thesaurus Sozialwissenschaften des IZ sind (Abb. 1):

Geschlechterrolle	Geschlechtsrolle
Sowjetunion	UdSSR
Berufliche Fortbildung	Berufliche Weiterbildung
Haushalt	Privathaushalt
Führung	Personalführung
Jugendarbeitslosigkeit	Jugendlicher + Arbeitslosigkeit

Abb. 1: Beispiele intellektuell erstellter Crosskonkordanzen

Das Konzept der Crosskonkordanzen darf nicht mit dem der Metathesauri verwechselt werden. Crosskonkordanzen streben keine neue Standardisierung bestehender Begriffswelten an. Sie erfassen die partielle Verbindung zwischen bereits bestehenden Terminologiesystemen, deren Vorarbeit genutzt wird und decken damit den statisch bleibenden Teil der Transferproblematik ab. Bei der Recherche bieten solche Verzeichnisse die Möglichkeit, Terme des einen Begriffssystems in die des anderen umzusetzen, im einfachsten Fall im Sinne einer Synonymie- oder Ähnlichkeitsrelation.

Bei Crosskonkordanzen spielt es im Gegensatz zur Planung von Metathesauri keine Rolle, ob sich 90 % der Begriffe oder nur 20 % aufeinander abbilden lassen, weil z. B. zwischen zwei fachverwandten Thesauri die semantischen Überschneidungen eher gering sind oder weil es sich bei einem Thesaurus um ein sehr generelles, alle Fächer übergreifendes Begriffssystem handelt (wie die SWD), das mit einem tief erschließenden Fachthesaurus verbunden wird. Dies spielt schon deshalb keine Rolle, weil immer daran gedacht wird, verschiedene Typen von Transfermodulen gleichzeitig einzusetzen, so dass nicht ein Modul die gesamte Transferlast tragen muss (s. Abschnitt Transfer-Architektur und Einbettung in dezentrale und zentralistische Ansätze).

Ein interessanter Nebeneffekt der Erarbeitung von Crosskonkordanzen ist das Aufdecken von Ähnlichkeiten zwischen miteinander konkurrierenden Begriffssystemen. Die für Crosskonkordanzen notwendige Analyse zeigt sehr genau auf, wie hoch der Anteil gleicher Begriffssegmente ist, aller hochemotionalen Schulenbildung für die eine oder andere Lösung zum Trotz.

Crosskonkordanzen für infoconnex und ViBSoz

Intellektuell erzeugte Crosskonkordanzen für Thesauri und für Klassifikationen wurden bisher unter anderem zwischen SWD und den Thesauri der Pädagogik, der Psychologie, der Sozial- und der Wirtschaftswissenschaften entwickelt. Im Rahmen des Wissenschaftsportals vascoda werden sie derzeit im Cluster infoconnex als Transfer zwischen Pädagogik, Psychologie und Sozialwissenschaften eingesetzt wie auch in der Virtuellen Fachbibliothek Sozialwissenschaften ViBSoz, die darüber hinaus Verbindungen zwischen der Basisklassifikation und der Klassifikation Sozialwissenschaften des IZ integriert hat (s. Müller/Marx 2001).

Text-Fakten-Integration bei ELVIRA

Die Text-Fakten-Integration behandelt bisher nur ELVIRA, das **»El**ektronische **V**erbands**i**nformations-, **R**echerche- und -**A**nalysesystem« (gefördert durch das Bundesministerium für Wirtschaft, s. Stempfhuber/Hellweg/Schaefer 2002, Stempfhuber 2003, URL: www.gesis.org/Forschung/Informationstechnologie/ELVIRA.htm). Das System, das mittlerweile von über 700 Firmen eingesetzt wird, verbindet Nomenklaturen und die Länderliste des Statistischen Bundesamtes und der Verbände durch Crosskonkordanzen (z. B. im Elektrobereich: Außenhandel EU-Länder – alte Kennziffern bis 1996 <–> Außenhandel EU-Länder – neue Kennziffern ab 1996).

Statistische Transferverfahren

erheblicher Aufwand führt zu Verzicht

Wie in Abschnitt Crosskonkordanzen und qualitative Heterogenität diskutiert, sind Crosskonkordanzen intellektuell erstellte Verbindungen zwischen Termen zweier Thesauri oder Klassifikationen. Der häufigste Grund, warum auf intellektuell erstellte Crosskonkordanzen verzichtet wird, ist der nicht unerhebliche Aufwand. Schon deshalb muss eine andere Möglichkeit,

den Transfer durchzuführen, zur Verfügung stehen: der Einsatz automatischer statistischer Verfahren. Dabei werden die jeweiligen Begriffspaare nicht intellektuell festgelegt, sondern mithilfe mathematischer Verfahren berechnet. Die eingesetzten Algorithmen beruhen im Wesentlichen auf der Häufigkeit des »Zusammen-Vorkommens« von Begriffen in sog. Parallelkorpora (Term-Kookkurrenzen), die als Trainingsdaten dienen.

Parallelkorpora: zweifach indexierte Dokumente

Parallelkorpora sind Dokumentsammlungen, deren einzelne Dokumente nach zwei Begriffssystemen indexiert sind (z. B. SWD und Thesaurus Sozialwissenschaften des IZ). Sie werden nicht durch intellektuelle Doppelindexierung neu geschaffen – was wiederum kostenintensiv wäre –, sondern man findet sie bereits vor und nutzt diesen Umstand. Die Ausgangssituation ist bei der Zusammenführung von Bibliotheksbeständen mit Fachdatenbanken in der Regel besonders günstig, da Informationszentren neben den Aufsätzen immer auch die selbstständige Literatur aufnehmen, die damit zumindest doppelt verschlagwortet vorliegt. So konnte für die Verbindung von SWD und dem Thesaurus Sozialwissenschaften des IZ in ViBSoz ein Parallelkorpus von 33.328 Dokumenten zugrunde gelegt werden. Aber auch die einzelnen Bibliotheksverbünde in Deutschland produzieren doppelte Indexierungen, die sich als Grundlage nutzen lassen.

Realisierung statistischer Transfermodule mit Jester

Eine spezielle Art von Parallelkorpora lässt sich durch eine zusätzliche Freitextindexierung von Dokumenten erreichen, die bereits intellektuell mit einem Thesaurus/einer Klassifikation erschlossen wurden. Die errechnete Beziehung betrifft dann die Relation zwischen den Schlagwörtern des Thesaurus / der Klassifikation und dem Freitextterm.

Auch der gleiche Text in mehreren Sprachen kann die Basis für einen Parallelcorpus bilden, in dem verschiedensprachliche Freitextterme aufeinander bezogen werden. Für das multilinguale IR ergänzt somit der gleiche bzw. hilfsweise ein ähnlicher Text in einer Fremdsprache den Parallelcorpus. Am generellen Prinzip ändert sich dadurch nichts. So basieren alle bisher in der Virtuellen Fachbibliothek Sozialwissenschaften ViBSoz und in ELVIRA getesteten Module auf einem Ansatz, der von Sheridan/Ballerini 1996 für das multilinguale Retrieval vorgeschlagen wurde.

Vereinfacht ausgedrückt gilt: Je häufiger ein Begriffspaar aus zwei Erschließungssprachen in einem Parallelcorpus vorkommt oder je häufiger sich Begriffe aus zwei unterschiedlichen Dokumenten aufeinander beziehen lassen (Term-Kookkurrenzen), desto wahrscheinlicher ist es, dass es sich um eine sinnvolle Verbindung handelt (= der Wahrscheinlichkeitswert in der letzten Spalte von Abb. 2). Hinzu kommen Parameter wie die Größe des Korpus oder die Verteilung der Begriffe innerhalb des Textes.

Für die Realisierung statistischer Transfermodule steht das am IZ entwickelte Werkzeug »Jester« (**J**ava **E**nvironment for **S**tatistical **T**ransf**ER**s) zur Verfügung. Jester erlaubt Schwellenwerte der Gewichtung anzugeben und bestimmte Terme (z. B. Allgemeinbegriffe) aus der Matrix auszuschließen. Eine ausführliche Beschreibung findet sich in Hellweg et al. 2001.

Im Unterschied zu den Crosskonkordanzen in Abschnitt Crosskonkordanzen und qualitative Heterogenität ergeben sich die statistischen Transformationen nicht durch allgemeine intellektuell ermittelte semantische Beziehungen, sondern die Term-Term-Matrix spiegelt die errechneten Term-Kookkurenzen wider. Teilweise entsprechen sie den semantischen Beziehungen, die auch intellektuell ermittelt wurden. Es ergeben sich aber auch davon abweichende Relationstypen, die zu anderen relevanten Treffern führen (Abb. 2).

Das folgende Beispiel (Abb. 3) stammt aus dem BMBF-Projekt CARMEN, in dem sowohl intellektuelle Crosskonkordanzen zwischen MACS (Mathematische Klassifikation) und PACS (Physik) von Bibliothekaren erarbeitet als auch statistische Term-Term-Matrizen auf der Basis eines Parallelkorpus errechnet wurden (s. Binder et al. 2002: Abschnitt 5.3.3 mit weiteren Beispielen und genauen Angaben zum Testdesign). Einzelne mathematische Methoden traten gehäuft in Verbindung mit inhaltlichen Forschungsgebieten auf, eine für das Retrieval wertvolle Relation, die bei der intellektuell erstellten Crosskonkordanz nicht berücksichtigt wurde.

Aber auch statistische Kookkurrenzen, die sich nicht mehr systematisch erklären lassen, können zu relevanten Dokumenten führen, wie das Beispiel Abb. 4 aus einem Retrievaltest von sozialwissenschaftlichen Texten (Freitext) zeigt. Der Ausgangsbegriff ohne die Zuschaltung der Transferbegriffe war **Dominanz** und führte zu 16 relevanten Treffern.

Term aus SWD	Term aus Thesaurus Sozialwissenschaften des IZ	Wahrscheinlichkeitswert
	Abfallwirtschaft,	0.6
	Umweltpolitik	0.6
Jugendarbeitslosigkeit	Jugendlicher,	0.88
	Arbeitslosigkeit	0.83
Wissensbasiertes System	Informationssystem,	1.0
	Künstliche Intelligenz	1.0

Abb. 2: Beispiele für statistische Transferbeziehungen, die den intellektuell ermittelten entsprechen

Statistisch ermittelte Crosskonkordanzen als Spezialfall des statistischen Transfers

Die in Abschnitt Statistische Transferverfahren als Alternative zum Einsatz intellektuell erstellter Crosskonkordanzen eingeführte statistische Transferbeziehung zeichnet sich nicht zuletzt durch die gefundenen Relationen aus, die nicht zwingend auf semantisch-pragmatischen Beziehungen basieren. Der Bearbeiter einer intellektuell erstellten Crosskonkordanz würde sie deshalb nicht angeben. **Wüste** und **Mongolei** in Abb. 4 machen für den Benutzer, der die Termerweiterungen vor ihrer Verwendung prüfen möchte, in der Regel keinen Sinn, können aber dennoch zu relevanten Dokumenten führen. Deshalb liegt es nahe, Prozesse dieser Art als Hintergrundprozesse ablaufen zu lassen, vor allem dann, wenn zusätzlich zum statistischen Transfer eine intellektuell erstellte Crosskonkordanz zur Verfügung steht. Ein statistischer Transfer als Hintergrundprozess schließt die Möglichkeit nicht aus, dass der Benutzer über die Termerweiterungen informiert wird (wie bei ViBSoz) und die vom System vorgeschlagene Liste verändern kann.

PACS 62.30.+d	Mechanical and elastic waves; vibrations (Mechanische und elastische Wellen, Schwingungslehre)
MSC 74S15	Boundary element methods (Randelementmethode)

Abb. 3: Statistische Methodenbeziehungen MSC <-> PACS

Transferbegriffe	Dominanz, Messen, **Mongolei**, Nichtregierungsorganisation, Flugzeug, Datenaustausch, Kommunikationsraum, Kommunikationstechnologie, Medienpädagogik, **Wüste**
Zahl zusätzliche relevante Treffer	7
Anteil der zusätzlichen relevanten Treffer an den zusätzlichen Treffern	50 %
Erklärungstext	Mitglieder des Vereins wom@n reisten zur UNO-Frauenkonferenz nach Beijing. Auf der Fahrt durch die **Mongolei** und die **Wüste** ...

Abb. 4: Beispiel *Dominanz*

Eine andere Ausgangssituation entsteht, wenn Kostengründe die Entwicklung einer intellektuell erstellten Crosskonkordanz verhindern. Die Term-Term-Matrix des statistischen Verfahrens kann dann – in der Regel nach einer zwischengeschalteten intellektuellen Überarbeitung – diejenigen semantisch-pragmatischen Relationen liefern, die Benutzer in jedem Thesaurus bzw. jeder Klassifikation erwarten. In diesem Fall sprechen wir von statistisch erstellten Crosskonkordanzen. Sie wurden im Projekt CARMEN zusätzlich zur intellektuell erstellten Crosskonkordanz zwischen der mathematischen Klassifikation MSC und der physikalischen PACS erarbeitet. Das Motiv war hier die geringe Ausgefächertheit der Begriffe, die Klassifikationen inhärent ist.

aus Kostengründen: statistisch statt intellektuell erstellte Crosskonkordanzen

In ELVIRA entstand aus Kostengründen eine statistische Crosskonkordanz für die Terminologie des Verbands Deutscher Maschinen- und Anlagenbau (VDMA). Sie ersetzt eine ursprünglich geplante, intellektuell erstellte Crosskonkordanz.

Zusammenspiel der Einzelverfahren

Die bisher geschilderten Heterogenitätsverfahren werden heute noch alternativ eingesetzt, d.h. eine Transformation besteht aus einem von möglicherweise mehreren für einen Übergang zur Verfügung stehenden Verfahren. Welche Transfers z.B. in ViBSoz zugeschaltet werden, kann der Benutzer durch Auswahlfelder beeinflussen, wenn er das möchte. Nur bei Nullantworten sollten systemseitig alle verfügbaren Transfers – in der Reihenfolge ihrer Relevanz – angestoßen werden, bis sich Dokumente nachweisen lassen.

Darüber hinaus wurden in allen Einsatzbereichen einige Ad hoc-Festlegungen zur Parametrisierung beider Einzelverfahren getroffen, ohne dass hierzu bereits systematische Untersuchungen vorliegen. So ist beim Einsatz intellektueller Crosskonkordanzen zu entscheiden, ob bzw. welche der Relationen Synonymie, Ober-/Unterbegriff, Ähnlichkeit standardmäßig zugeschaltet werden, um das bestmögliche Transferergebnis zu erreichen. Bei den statistischen Verfahren spielen die

Schwellenwerte der Wahrscheinlichkeiten, die Verbindung mit linguistischen Verfahren (als vorgeschaltete Termbereinigung) und die Gewichtung der Art der Inhaltserschließung (z. B. Schlagwörter höher gewichten als Stichwörter) eine Rolle.

ETB nur als Prototyp

Auf der Basis der empirisch motivierten Optimierung der Einzelverfahren, die sich je nach Anwendungskontext und Fachgebiet deutlich unterscheiden wird, kann dann der eigentliche »Mehrwert« der hier vorgeschlagenen Heterogenitätsbehandlung auf der Basis bilateraler Transfers zwischen einzelnen Dokumentenmengen zum Tragen kommen. Ihr Vorteil liegt gerade darin, dass die unterschiedlichsten Mischformen möglich sind. So könnten die Transfermodule zwischen einem Dokumentbestand, der mit einer generellen bibliothekarischen Schlagwortliste wie SWD indexiert wurde und einem zweiten, dessen Indexierung auf einem speziellen fachspezifischen Thesaurus beruht, durch qualitative Verfahren wie Crosskonkordanzen aufeinander bezogen werden. Die Einbindung der dritten Datenquelle (z. B. WWW-Angebote wie Clearinghouses oder graue Literatur) könnte wiederum über statistische Transfers oder neuronale Netzwerke erfolgen. Die Verbindung zur Benutzerterminologie ließe sich abschließend mit einem probabilistischen Verfahren herstellen. Diese Möglichkeit, die einzelnen Stufen des Transfers jeweils differenziert an die speziellen Gegebenheiten anzupassen, ist ein wesentlicher Unterschied zu den bisherigen IR-Lösungen. Im gesamten Forschungskontext ist jedoch bisher ungeklärt, wie die verschiedenen Verfahren kombiniert werden sollen, um größtmögliche Effizienz zu erreichen. Dass eine solche Kombination deutliche Vorteile verspricht, darauf verweisen die Ergebnisse der jährlich stattfindenden TREC- und DELOS-Evaluationen (s. Kluck 2004). In einem objektivierten Kontext können hier Entwickler von IR-Software die Leistungsfähigkeit ihrer Systeme im Vergleich zu denen der anderen Entwicklungsgruppen testen. Es stellt sich immer wieder heraus, dass sich verschiedene IR-Verfahren nicht ausschließen, sondern unterschiedliche Mengen relevanter Dokumente liefern. Auch die ersten empirischen Ergebnisse aus CARMEN mit dem Vergleich alternativer Zugänge weisen in diese Richtung (s. Binder et al. 2002).

Verbindung zur Benutzerterminologie

nachweisbare Qualitätsverbesserung

Bereits heute lässt sich feststellen, dass im ViBSoz- und CARMEN-Kontext empirische Belege für die Qualitätsverbesserung der Recherche durch den Einsatz der Transfermodule in ihrer jetzigen – gegenüber dem geplanten Ausbau noch sehr beschränkten – Architektur nachweisbar sind. Der eigentliche Mehrwert des Verfahrens wird sich jedoch erst nach der empirischen Fundierung verschiedener Parametrisierungen der Einzelmodule und der möglichen Kombinatorik der unterschiedlichen Transfertypen entfalten können. Hierzu ist am IZ eine Dissertation in Arbeit, deren Ergebnisse im Sommer 2004 vorliegen dürften.

Transfer-Architektur und Einbettung in dezentrale und zentralistische Ansätze

Varianten der beschriebenen Transfermodule wurden bisher in fünf verschiedene Architekturen eingebettet: infoconnex (als Bestandteil von vascoda), ELVIRA, CARMEN, ViBSoz und ETB (»The European Schools Treasury Browser«, EU-Projekt, s. Kluck/Strötgen 2002, Bandholm/Lund 2000). Drei davon sind heute im praktischen Einsatz (CARMEN HYREX und die ETB-Transfers kamen nicht über die Prototypphase hinaus).

Da die Transfermodule unabhängig von der Systemarchitektur des Gesamtsystems sein sollten, um portabel zu sein, war eine möglichst unabhängige Integrationsebene zu finden. Abb. 5 zeigt exemplarisch die Einbindung in die CARMEN-Architektur, die wie ViBSoz dezentrale Datenbanken bzw. Web-Angebote miteinander verbindet: Ein sozialwissenschaftlicher Benutzer ist z. B. mit den Schlagwörtern der SOLIS-Datenbank des IZ vertraut und formuliert deshalb seine Anfrage mit den Begriffen des Thesaurus Sozialwissenschaften, die er – wie in ViBSoz möglich – aus einem Thesaurusbrowser auswählt, will aber gleichzeitig in der Universität Köln mit den dortigen OPACs suchen und in einem dritten Dokumentenbestand mit wiederum unterschiedlicher Inhaltserschließung. Die Anfrage wird unverändert an die Datenbank SOLIS des IZ geschickt (Recherche Bestand B), als Variante 2 nach einem bilateralen Transfer Thesaurus Sozialwissenschaften <–> SWD (z. B. intellektuell erstellte Crosskonkordanz) an den OPAC der Universität Köln (A) gerichtet usw.

Abb. 5 macht deutlich: Die Transferkomponente wird als Termerweiterungs- bzw. Veränderungsverfahren bei den Eingabebegriffen des Nutzers realisiert und verteilt die Ergebnisse auf Untermengen der Datenbestände.

Beim EU-Projekt ETB ging es im Unterschied zu der dezentralen Brokerarchitektur von CARMEN und ViBSoz um eine zentralistische Architektur (s. Abb. 6). Die einzelnen Anbieter von Bildungsdatenbanken (Site A, B usw.) wollten ein zentrales Angebot aufbauen mit einer einheitlichen Zugangssprache (ETB Native Repository). Eigentlich widerspricht dieser Ansatz den hier und in Krause / Niggemann / Schwänzl 2003 vorgetragenen Überlegungen, da er vordergründig nur auf eine neue übergeordnete Standardisierung setzt. Heterogenitätsmodule kamen dennoch für alle Altbestände,

die die neue ETB-Erschließungssprache als Zielpunkt des Transfers auf Dokumentenebene einsetzten, zum Zuge und für all jene, die zweigleisig bleiben wollten, d.h. ihre eigene Form der Inhaltserschließung beibehielten und die Arbeit für eine Doppelindexierung scheuten. Auch in diesem Fall erzeugen die Transferkomponenten die Zielstrukturen.

Das Beispiel zeigt, dass auch bei zentralistischen Architekturen, die an einer zentralen Datenbank und einer einheitlichen Beschreibungssprache festhalten, Transfermodule ihren Platz haben.

Wie konzeptualisiert der Benutzer seine Anfrage an einen dezentralen, heterogenen Informationsraum?

Solange der Benutzer einem Dokumentenbestand gegenübersteht, den er als einen Informationstyp empfindet, wird er die Wünsche an die Benutzungsoberfläche aus der Vorstellung eines exemplarischen Vertreters des Dokumentenbestandes ableiten. Bei der Literaturrecherche erwartet er bestimmte Strukturen und Merkmale (Autor, Titel, Schlagwörter usw.), die über alle Dokumentenbestände verschiedener OPACs oder Fachinformationssysteme hinweg angenommen werden. Die einheitliche Rechercheoberfläche muss dieser Erwartungshaltung Rechnung tragen.

Bei der Clusterrecherche von infoconnex wird z.B. davon ausgegangen, dass im Regelfall Pädagogen mit den ihnen vertrauten Begriffen aus ihrer Datenbank FIS Bildung suchen. Wird die Recherche auf die Sozialwissenschaften oder die Psychologie ausgeweitet, richtet sich die Transformation ausgehend von der Pädagogik auf die dem Benutzer nicht vertrauten Inhaltserschließungssysteme der Nachbardisziplinen. Die eigene Fachterminologie prägt somit das exemplarische Vorbild des Benutzers, das durch den Anfragebildschirm zu unterstützen ist.

Man könnte zusätzlich zu diesen Einstiegen davon ausgehen, dass es Benutzer gibt, die ihr Informationsbedürfnis abstrahiert von den verschiedenen Inhaltserschließungssystemen des dezentralen Informationsraums konzeptualisieren wollen, auch wenn sie zumindest eines der Inhaltserschließungssysteme genauer kennen. Sie würden dann quasi ihr eigenes Begriffsystem als Startpunkt aller Transformationsprozesse, die systemseitig angestoßen werden, setzen. Dies klingt plausibel und scheint bei der Gestaltung des Suchbildschirms problemlos umsetzbar.

Schwieriger wird es jedoch, wenn neben textuelle Dokumente Faktendaten treten, wie dies bei ELVIRA der Fall ist.

ELVIRA wurde in der ersten Stufe als Fakteninformationssystem für die Firmen der Elektronikindustrie entwickelt. Es sollte einen benutzerfreundlichen Zugang zu Produktions-, Außenhandels-, Konjunktur- und Strukturdaten ermöglichen. Der Recherchezugang ist durch die spezifische Datenstruktur und die Art der Inhaltserschließung der bei den Verbänden intellektuell ermittelten Daten gekennzeichnet. Die Zeitreihen-Faktentabellen werden nicht über ihre Zellenwerte und den Tabellennamen angesprochen, sondern indirekt über intellektuell vergebene Deskriptoren zu

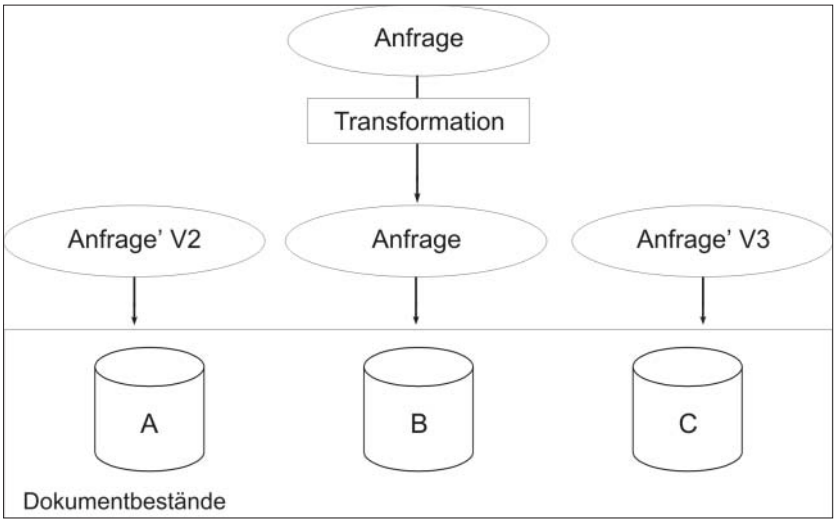

Abb. 5: Anfrage-Transformation aus Hellweg et al. 2001

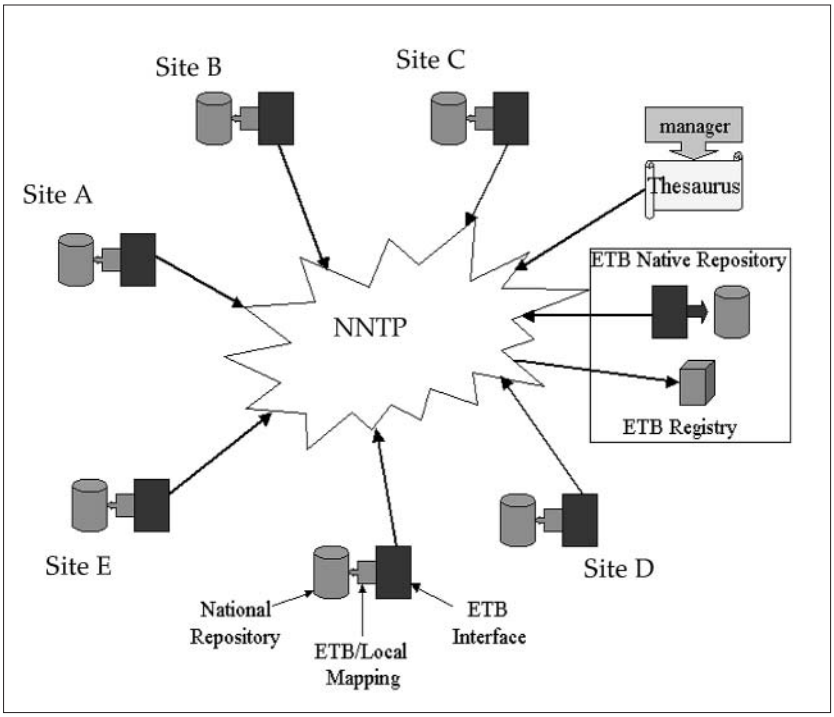

Abb. 6: ETB-Architektur aus Bandholm/Lund 2000, S. 7

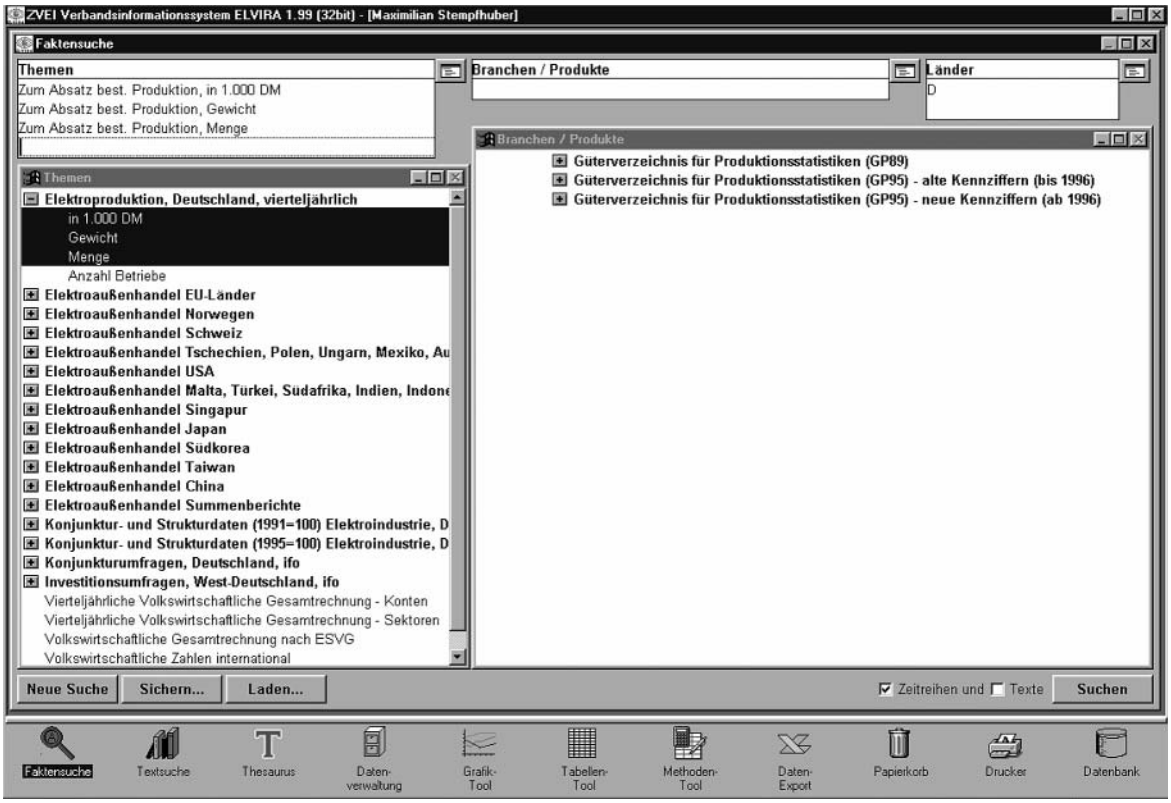

Abb. 7: ELVIRA-Faktenrecherche, Stempfhuber 2003

den drei Kategorien: betroffenes Thema (z. B. Export), Branche / Produkt (z. B. Mikrowellengeräte) und Land (s. Stempfhuber 2003).

Rechercheformulierung durch den Benutzer

Benutzertests zeigten trotz hoher Akzeptanz rasch, dass die Verbandskunden neben den Zeitreihen andere Informationsquellen zur Lösung ihrer Problemstellungen fordern. Beim VDMA waren dies z. B. textuelle Außenwirtschaftsinformationen zur qualitativen Marktbeurteilung. Deshalb wurde ELVIRA um einen Textzugang erweitert und eine integrierte Suche auf der Basis des in den vorangegangenen Abschnitten vorgestellten Modells von Transfermodulen zugelassen.

Einen Sinn macht die Integration von Fakten und Texten allerdings erst, wenn eine zweite Lücke der heute üblichen Zugangssysteme geschlossen wird. Es muss ein adäquates Modell für die Behandlung des Problems der Rechercheformulierung durch den Benutzer gefunden werden – oder einfacher ausgedrückt: Wie sieht der einheitliche Anfragebildschirm in diesem Fall aus?

Zusammen mit Nutzern und Informationsvermittlern bei den Verbänden wurden zunächst die relevanten Textbestände durch die Auswertung von Anfragen an die Verbände und durch Benutzerinterviews ermittelt. Die Untersuchung zeigte, dass es sowohl Fragestellungen gab, bei denen der Benutzer entweder *nur* nach Zeitreihen oder *nur* nach Texten sucht, dass aber auch Mischformen eine Rolle spielen. Die beiden Reinformen, die Fakten- und die Text-Recherche, dienten als Ausgangspunkt für die Integration von Text- und Faktenrecherche.

Für unser Problem sind die Fälle von besonderem Interesse, bei denen der Benutzer ausgehend von einem Ergebnistyp (z. B. Texten) plötzlich auch Ergebnisse eines anderen Typs (z. B. Zeitreihen) in die Suche einschließen möchte. Dabei macht die Gestaltung

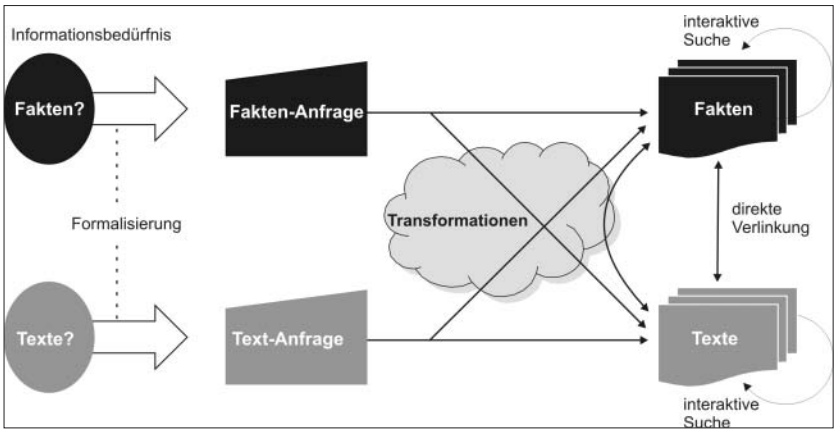

Abb. 8: Reinformen

des Bildschirms der Ausgangsfrage bei den folgenden Übergangstypen keine Probleme:

- Fakten-Anfrage – > (Fakten- und) Text-Ergebnis
- Text-Anfrage – > (Text- und) Fakten-Ergebnis
- Fakten-Ergebnis – > Text-Ergebnis
- Text-Ergebnis – > Fakten-Ergebnis

All dies sind keine »echten« Integrationen auf der Ebene des einheitlichen Anfragebildschirms. Eine echte Integration liegt dann vor, wenn der Benutzer sein Informationsbedürfnis völlig abstrakt formuliert (s. Abb. 9).

Erste empirische Untersuchungen ergaben Hinweise, dass solche Bedürfnisse im Anwendungsfall von ELVIRA vorkommen (s. Krause/Mandl/Stempfhuber 1997). Der Versuch, diese echte Integration zusammen mit den Benutzern als Anfragebildschirm zu erarbeiten, zeigte aber, dass im konkreten Fall einer definierten Recherche der Benutzer immer entweder vom Text oder von der Faktendarstellung als Vorstellung ausgeht. So gibt er zwar an, er wolle »alles« zu einem Land wissen. Bei der Rückfrage, wie er sich dazu den idealen Anfragebildschirm wünscht, wechselt er jedoch wieder zu einer konkreten Ausgangsfrage im Text- oder Faktenmodus. Nach den bisherigen Erfahrungen spricht deshalb vieles dafür, dass bei der Suchformulierung der Text-Fakten-Integration bereits in Beispielen gedacht wird und damit das abstrakte Informationsbedürfnis auf der Ebene der Suchformulierung keine Entsprechung mehr hat.

Damit würden die jeweiligen Reinformen der Recherche als Anfragebildschirme genügen. Ob sich diese Beobachtung in anderen Heterogenitätskontexten bestätigt, muss abgewartet werden.

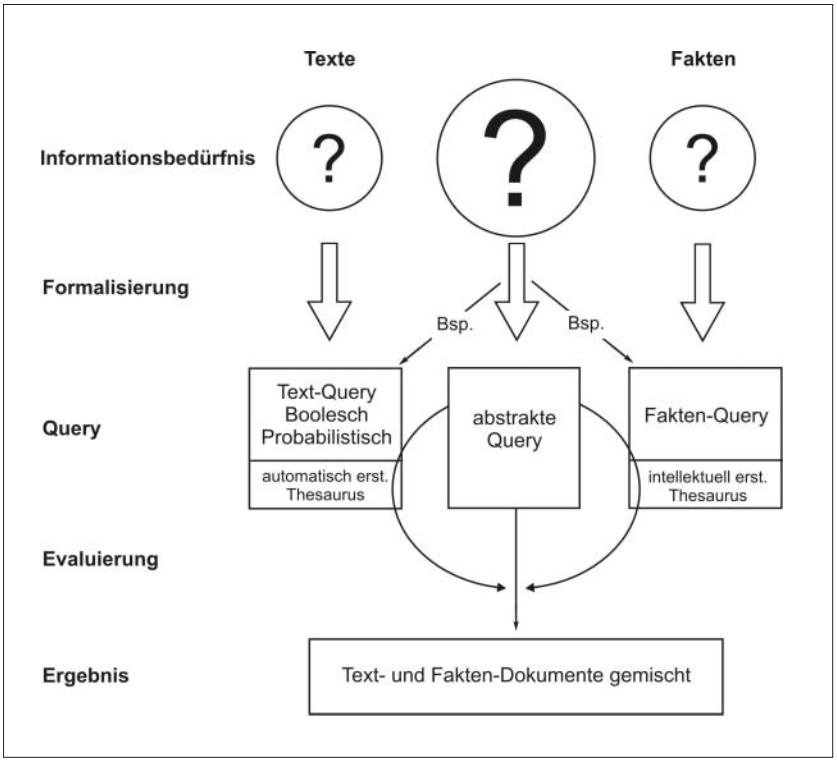

Abb. 9: »Echte« Integration

Fazit Transfermodule und Ausblick Schalenmodell

Das Konzept der bilateralen Transfermodule ist heute – von der Modellbildung und von den praktischen Einsatzmöglichkeiten her – so weit fortgeschritten, dass es sich konkret bei virtuellen Fachbibliotheken, Fachportalen, Informationsverbünden und im Wissenschaftsportal vascoda mit Gewinn einsetzen lässt. Es operationalisiert den Leitsatz »Standardisierung ist von der verbleibenden Heterogenität her zu denken« auf einer ersten praktisch einsetzbaren Stufe. Mit infoconnex wurde das Konzept im August 2003 Bestandteil von vascoda und wird in diesem Kontext weiter ausgebaut werden.

Bilaterale Transfermodule sind von der Modellbildung her sehr einfache Grundbausteine, die jedoch durch ihre Variationsbreite und eine additive Anwendung der Einzelelemente schnell zu komplexen Strukturen führen können. Sie sind bei den bisherigen Anwendungen mit ihrer beschränkten Vielfalt der zu integrierenden Inhaltserschließungstypen noch übersichtlich analysierbar und handhabbar. Bei der sehr viel größeren Anzahl von Variationen, die uns erwarten, wenn wir die Integrationsmöglichkeiten des WWW ernst nehmen, dürfte sich das jedoch rasch ändern.

Schalenmodell

Deshalb braucht das vorgeschlagene Modell auf einer zweiten Stufe abstraktere Ordnungsansätze, die auf einem höheren Niveau der Zusammenfassung arbeiten. Dies soll das Schalenmodell leisten, auf das hier nicht weiter eingegangen werden konnte. Es bezieht neben der informationswissenschaftlichen Ebene organisatorische und wissenschaftspolitische Dimensionen mit ein. Das Schalenmodell ergänzt die bilateralen Transfermodule um einige zusätzliche Annahmen: Verschiedene Niveaus der Inhaltserschließung und Dokumentenrelevanz werden zu sog. Schalen zusammengefasst, die untereinander durch höherstufige Transfermodule verbunden werden (genauer Krause 2000). Im Gegensatz zu den bilateralen Transfermodulen mit ihrer praktischen Umsetzbarkeit geben die weiter gehenden Anforderungen des Schalenmodells heute allerdings noch eher eine Denkrichtung an, die den weiteren Ausbau eines Richtungswechsels in der Fachinformation kennzeichnen soll.

Literatur

Bandholm, Anders; Lund, Tommy Byskov: WP9 Architecture of the Metadata Networking Infrastructure. European Treasury Browser. Deliverable No. D9.1 2000. URL: www.eun.org/etb/Output and Documents

Binder, Gisbert; Marx, Jutta; Mutschke, Peter; Riege, Udo; Strötgen, Robert; Kokkelink, Stefan; Plümer, Judith: Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren. In: IZ-Arbeitsbericht 24 (2002). URL: www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_24.pdf

Boekhorst, Peter te; Kayß, Matthias; Poll, Roswitha: Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft. Universitäts- und Landesbibliothek Münster und infas Institut für angewandte Sozialwissenschaft GmbH. Seit Dezember 2003 als Langfassung einsehbar unter: www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/aktuelles/download/ssg_bericht_teil_1.pdf

Hellweg, Heiko; Krause, Jürgen; Mandl, Thomas; Marx, Jutta; Müller, Matthias N.O.; Mutschke, Peter; Strötgen, Robert: Treatment of Semantic Heterogeneity in Information Retrieval. In: IZ-Arbeitsbericht 23 (2001). URL: www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_23.pdf

IMAC: Projekt Volltextdienst : Zur Entwicklung eines Marketingkonzepts für den Aufbau eines Volltextdienstes im IV-BSP. Konstanz : IMAC Information & Management Consulting, September 2002. Management summary (infoconnex-Studie)

Ingwersen, Peter: The Cognitive Framework for Information Retrieval : A Paradigmatic Perspective. In: Krause, Jürgen; Herfurth, Matthias; Marx, Jutta (Hrsg.): Herausforderungen an die Informationswissenschaft. Konstanz : Univ.-Verl., 1996, S. 25–31

Kluck, Michael: Die Evaluation von Cross-Language-Retrieval-Systemen mit Hilfe der GIRT-Daten des IZ. In: IZ-Arbeitsbericht **(2004)** (erscheint)

Kluck, Michael; Strötgen, Robert: Report on the solution for transfer components and description of the underlying software, European Treasury Browser, Deliverable No. D6.2 2002. URL: www.eun.org/etb/Output and Documents

Krause, Jürgen: Standardisierung von der Heterogenität her denken – Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken. In: IZ-Arbeitsbericht 28 (2003). URL: www.ge sis.org/Publikationen/Berichte/IZ_Arbeitsberich te/pdf/ab_28.pdf

Krause, Jürgen: Sacherschließung in virtuellen Bibliotheken – Standardisierung versus Heterogenität. In: Rützel-Banz, Margit (Hrsg.): Grenzenlos in die Zukunft. Frankfurt am Main : Klostermann, 2000, S. 202–212

Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland: Normierung und Standardisierung in sich verändernden Kontexten : Beispiel Virtuelle Fachbibliotheken. In: Zeitschrift für Bibliothekswesen und Bibliographie 1 (2003), S. 19–28

Krause, Jürgen; Mandl, Thomas; Stempfhuber, Maximilian: Text-Fakten-Integration in ELVIRA. In: IZ-Arbeitsbericht 12 (1997). URL: www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab12

Machill, Marcel; Welp, Carsten (Hrsg.): Wegweiser im Netz : Qualität und Nutzung von Suchmaschinen. Gütersloh : Verl. Bertelsmann-Stiftung, 2003

Mandl, Thomas: Tolerantes Information Retrieval : Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Dissertation. Konstanz : UVK-Verl.-Ges., 2001 (Schriften zur Informationswissenschaft; 39)

Müller, Matthias N.O.; Marx, Jutta: The Social Science Virtual Library Project : Dealing with Semantic Heterogeneity at the Query Processing Level. In: Proceedings of the Third DELOS Network of Excellence Workshop »Interoperability and Mediation in Heterogeneous Digital Libraries«, Darmstadt, Germany; 8–9 September 2001, Sophia-Antipolis, USA : ERCIM. (DELOS Network of Excellence on Digital Libraries Workshop Series), S. 19–24

Mutschke, Peter: Cognitive and Social Structures in Social Science Research Fields and their Use in Information Systems. – 5th International Conference on Logic and Methodology »Social Science Methodology in the New Millenium«, Köln, 03. – 06. Oktober 2000

Mutschke, Peter; Quan-Haase, Anabel: Collaboration and Cognitive Structures in Social Science Research Fields : Towards Socio-Cognitive Analysis in Information Systems. In: Scientometrics 52, 3 (2001), S. 487–502

Poll, Roswitha: Informationsverhalten und Informationsbedarf der Wissenschaft: Teil 1 der Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. In: Zeitschrift für Bibliothekswesen und Bibliographie 2 (2004), S. 59–75

Ronthaler, Marc; Zillmann, Hartmut: Literaturrecherche mit OSIRIS : ein Test der OSIRIS-Retrievalkomponente. In: Bibliotheksdienst 7 (1998), S. 1203–1212

Sheridan, Páraic; Ballerini, Jean Paul: Experiments in Multilingual Information Retrieval using the SPIDER System. In: Frei, Hans-Peter; Harman, Donna; Schäuble, Peter; Wilkinson, Ross (Hrsg.): Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '96, August 18–22, 1996, Zurich, Switzerland.

New York: ACM Press (Special Issue of the SIGIR Forum), 1996, S. 58–65
Stahl, Matthias; Binder, Gisbert; Cosler, Detlev; unter Beratung von Dr. Joachim Scharioth: TRI:M-Studie zur Kundenzufriedenheit (Mehrfachkunden) 1997. IZ-Arbeitsbericht 13 (1998)
Stahl, Matthias; Binder, Gisbert; Marx, Jutta: Das Informationszentrum Sozialwissenschaften im Urteil von Soziologieprofessorinnen und -professoren aus Deutschland, Österreich und der Schweiz. IZ-Arbeitsbericht 25 (2002). URL: www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab_25.pdf
Stempfhuber, Maximilian: Objektorientierte Dynamische Benutzungsoberflächen – ODIN : Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie. IZ-Forschungsbericht 6 (2003)
Stempfhuber, Max; Hellweg, Heiko; Schaefer, André: ELVIRA : User Friendly Retrieval of Heterogenous Data in Market Research. In: Callaos, Nagib; Hernandez-Encinas, Luis; Yetim, Fahri (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, 2002, S. 299–304
Strötgen, Robert: Meta-Data Extraction and Query Translation : Treatment of Semantic Heterogeneity. In: Agosti, Maristella; Thanos, Constantino (Hrsg.): Research and Advanced Technology for Digital Libraries : 6th European Conference, ECDL 2002, Rome, Italy, September 16–18, 2002; Proceedings. Berlin : Springer, S. 362–373
Weibel, Stuart: Dublin Core – State of the art after DC5. Proceedings Workshop »Meta Data : Qualifying Web-Objects«, Osnabrück, 1997. URL: www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html.

Der Verfasser

Prof. Dr. Jürgen Krause ist Direktor des IZ Sozialwissenschaften, Lennéstraße 30, 53113 Bonn, und Professor für Computervisualistik in Koblenz.
jk@Bonn.IZ-Soz.de