



World Library and Information Congress: 69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 085-G
Meeting: 151. Social Science Libraries
Simultaneous Interpretation: Yes

Standardisierung, Heterogenität und die Qualität der Inhaltserschließung. Lösungsansätze für ein zentrales Problem digitaler Bibliotheken

Jürgen Krause

University of Koblenz-Landau and Social Science Information Center (IZ Bonn),
Lennéstr. 30, D-53113 Bonn, Germany

Krause@bonn.iz-soz.de;

<http://www.uni-koblenz.de/~krause/>

Abstract

Lösungen für die Probleme beim Aufbau fachwissenschaftlicher Informationsangebote führen weit über die bisher gewohnten Denkweisen von Informationszentren und Bibliothekaren hinaus. Die diskutierte Leitlinie „Standardisierung von der Heterogenität her zu denken“ und das Paradigma des „Publizierens im Netz“ charakterisieren den Wandel am deutlichsten. Er ist nicht nur technologisch, sondern inhaltlich-konzeptuell. Beide Leitlinien zusammengekommen sind eine Antwort auf die neuen technologischen und Benutzeranforderungen in den veränderten Kontexten des Web und auf die klaren Anzeichen darauf, dass die traditionellen Methoden der Standardisierung nicht mehr ausreichen, die durch die angestrebte Interoperabilität und Datenkonsistenz aber trotzdem erhalten bleiben sollen.

Keywords: virtual library, digital library, content analysis, metadata, heterogeneity, Publizieren im Web; text-fact integration, ViBSoz

1 Einleitung

„Access Point Library“, was soll dieses Motto für die Bibliotheken heute ausdrücken? Was bedeutet es vor allem vor dem Hintergrund ihrer Geschichte der letzten 30 Jahre. Wird die Bibliothek reduziert auf einen access point, deren Inhalte andere gestalten oder soll damit ausgedrückt werden, dass sie die alte zentrale Rolle für die Informationsversorgung der Wis

schaften zurück erhält und die Angebote von Verlagen, Wissenschaftlern im Web und Fachinformationszentren den Bibliotheken wieder untergeordnet werden.

Gerade in Deutschland lässt sich das Zusammenspiel von technologischer Entwicklung, neuen Benutzerwünschen, Verschiebungen in der Organisations- und Angebotsstruktur und die darauf folgenden Veränderungen der Modellierung der informationellen Prozesse deutlich studieren. Im sozialwissenschaftlichen Sinne ist dies ein Teilbereich der Informationsgesellschaft (informational society), wie ihn Castells 2001 als „Netzwerkgesellschaft“ (network society) intensiv aus der Perspektive der Ökonomie und Soziologie beschrieben hat.

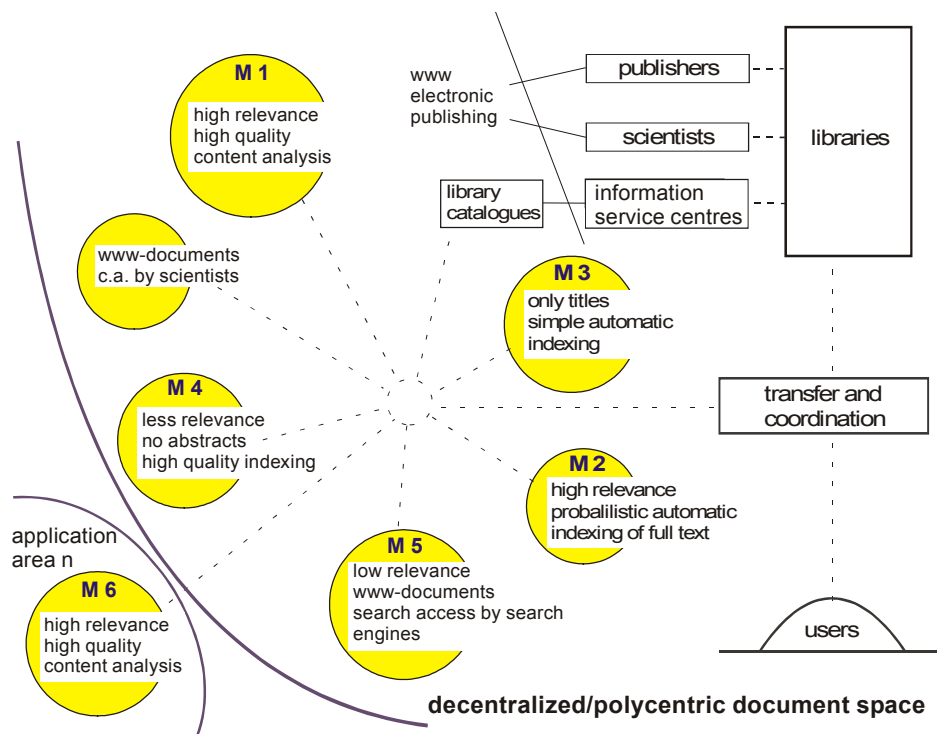
2 Polyzentrische Informationsversorgung

Vor etwa 30 Jahren vollzog sich für die Bibliotheken ein erster aus heutiger Sicht einschneidender Wandel. Sie gaben einen wichtigen Teil der Erschließung von wissenschaftlichen Informationen an Informationsservicestellen und Fachinformationszentren ab, die Literaturdatenbanken zu einzelnen Fachgebieten aufbauten und neben selbständiger Literatur auch Zeitschriftenartikel nachwiesen. In der Regel verblieb die Zeitschrift selbst und die Ausleihfunktion bei der Bibliothek. Warum, kann man sich heute fragen, kam es zu dieser Aufteilung von Funktionen? Warum soll eine Institution vor allem Zeitschriftenartikel nachweisen und die andere nur Bücher? Warum haben die Bibliotheken damals den Schritt in die Informationstechnologie, in den Aufbau von Literaturdatenbanken nicht selbst vollzogen, sondern an neu gegründete Institutionen abgegeben und was besagt das für die heutige Situation? Werden die Bibliotheken den Umbrüchen, die die informationstechnologische Entwicklung eingeleitet hat, gerecht, bleiben sie gewichtiger Mitspieler in der wissenschaftlichen Informationsversorgung oder geben sie weiterhin Funktionen ab, bis sie zu Archiven von physikalischen Dokumenten mutieren, die andere Informationseinrichtungen - solange sie nachgefragt werden - in elektronischer Form anbieten.

Aber zuerst: wie sieht die Bibliothekswelt und genereller die Literaturversorgung im wissenschaftlichen Bereich heute aus und was sind ihre zentralen Probleme?

Die Welt der Informationsanbieter ist – im Gegensatz zur politischen - nicht mehr zentralistisch oder bipolar, sondern polyzentrisch. Technologisch kann weltweit auf die verschiedensten Informationsquellen relativ problemlos und unmittelbar zugegriffen werden (zu jeder Tageszeit und über jede räumliche Distanz hinweg). Dies vervielfacht - im Vergleich zu den herkömmlichen Medien - die Anzahl von aktiven Verbreitern von Inhalten. Parallel zu anderen Bereichen des E-Commerce „verringert das Internet die Markteintrittsbarrieren“ und wirkt somit gegen bestehende Monopole (Cigan 2002: 15). Anbieter von Informationen können ihr Zielpublikum weltweit direkt erreichen. Gleichzeitig „verschiebt das Internet die Marktmacht vom Produzenten zum Konsumenten“ (Cigan 2002: 19).

Bei der wissenschaftlichen Informationsversorgung erwiesen sich die technischen Veränderungen als Türöffner für eine wachsende Dezentralisierung. Die von Cigan 2002: 9: 15 alternativ bei Wissensmärkten als denkbar angesehene Entwicklung zu einer neuen Monopolbildung tritt hier nicht auf. Bibliotheken mit ihren OPACs und die Datenbanken der IuD-Stellen sind nur noch Teilelemente eines vielfältigen heterogenen Angebots. Benutzer informationeller Dienste stehen damit heute einem hochgradig dezentralisierten und heterogenen Dokumentenraum gegenüber.



Neben die traditionellen Anbieter von Informationen, den Verlagen mit ihren Printmedien, den Bibliotheken, die ihre Buchbestände nach intellektuell vergebenen Klassifikationen erschließen und den Fachinformationszentren, die ihre Datenbanken über Hosts anbieten, sind verstärkt die Wissenschaftler selbst getreten, die in all diesen Bereichen über das WWW eigenständige Dienste unterschiedlichster Abdeckung, Relevanz und Erschließungsverfahren entwickeln. Generell können überall auf der Welt Gruppen auftreten, die zu Spezialgebieten Informationen sammeln. Eine Folge hiervon sind die unterschiedlichsten Konsistenzbrüche:

- Relevante, qualitätskontrollierte Daten stehen neben irrelevanten und eventuell nachweislich falschen. Kein Gutachtersystem sorgt für eine Trennung von Ballast und potentiell erwünschter Information. Wer z.B. als Sozialwissenschaftler Partnerschafts- und Sexualverhalten als Forschungsgebiet hat, weiß, was dies bei der Suche im Web bedeutet.
- 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, mit viel Aufwand intellektuell und qualitativ hochwertig ermittelt wurde, nicht mit dem Term X gleichgesetzt werden, den eine automatische Indexierung aus einem Randgebiet liefert.

Der Benutzer wird trotz solcher Probleme auf die verschiedenen Datenbestände zugreifen wollen, gleich nach welchen Verfahren sie erschlossen oder in welchem System sie angeboten werden. Er hält auch in der Welt dezentralisierter, inhomogener Datenbestände die Forderung an die Informationswissenschaft aufrecht, dafür zu sorgen, dass er möglichst nur die relevanten Dokumente und möglichst alle relevanten bekommt, die seinem Informationsbedürfnis entsprechen.

Wie lässt sich dies bewerkstelligen und welche Änderung der in den Bibliotheken und IuD-Stellen traditionellen und liebgewordenen Vorgehens- und Denkweisen ziehen die neuen Gegebenheiten nach sich?

2.1 Virtuelle Bibliotheken als hybride Bibliotheken

Bedingt durch die technologische Entwicklung waren vor 30 Jahren Bibliotheken und Fachinformationszentren zwangsweise zentralistisch organisiert und im Einklang damit auch konzeptuell auf zentralistische Ansätze der Inhaltserschließung ausgerichtet. Ein zentral aufgestellter Großrechner verwaltete die Daten. Die Klientel wurde über Terminals oder offline über Anfragen an die eine Zentralstelle bedient.

Dem entsprach die theoretische Grundlage der Inhaltserschließung. Nach einem normierten, intellektuell kontrollierten Verfahren, das die Zentralstelle entwickelte und durchsetzte, erfolgte eine einheitliche Erfassung der Dokumente. In diesem Denken kommt der Datenkonsistenz die höchste Priorität zu. Sie und die lückenlose Erfassung der relevanten Dokumente ohne Zeitverzug sicherzustellen, wird im gegenwärtigen Umfeld jedoch immer zeitaufwendiger und schwieriger.

Zentralistische Ansätze im Sinne der Sammlung aller Daten in einer Datenbank durch eine Organisation werden heute kaum mehr vertreten. Auch in der Bibliothekswelt sind diese Vorstellungen durch das Denken in Netzwerken ersetzt worden. Diese Modellvorstellung macht am besten das Konzept der virtuellen Fachbibliotheken klar. Virtuelle (Fach)bibliotheken sollen Wissenschaftlern von ihrem Computer aus einen optimalen Zugang zu den weltweit vorhandenen elektronischen und multimedialen Volltext-, Literaturhinweis-, Fakten- und WWW-Informationen einschließlich der dort vorhandenen Lehrmaterialien, Spezialverzeichnisse, z. B. zu Experten usw. ermöglichen. Virtuelle Fachbibliotheken sind somit hybride Bibliotheken mit einem gemischten Bestand aus elektronischen und gedruckten (und ggf. noch anderen) Daten. Letztere sind über elektronische Dokumentbestell- und -lieferdienste erreichbar. Auf technischer Seite setzt dies u. a. im Netz zugängliche verteilte Datenbanken voraus, auf konzeptueller Seite die Integration verschiedener Informationsgehalte und -strukturen.

Beispiele für die erreichte technische Integration heterogener Datenbestände sind der Bibliotheksverbund KOBV (<http://www.kobv.de/se/cont.html>) oder der virtuelle Bibliotheksverbund NRW.

Was die Bibliotheksverbände und vergleichbare Projekte bisher in der Regel nicht tun, ist, die verschiedenen Inhaltserschließungsverfahren der Teilbestände adäquat zu berücksichtigen.¹ Damit ergibt sich zwar im Vergleich zu den Schwächen der generellen WWW-Suchmaschinen (siehe Krause 2003b: Abschnitt 2) eine Verbesserung bei der Auswahl der relevanten Datenmaterialien, die konzeptuellen Unterschiede zwischen verschiedenen Inhaltserschließungsverfahren werden jedoch auch hier nicht ausgeglichen.

Traditionell wird im Kontext der virtuellen Bibliotheken versucht, konzeptuelle Integration durch Standardisierung sicherzustellen². Wissenschaftler, Bibliotheken, Verleger und die Anbieter von Fachdatenbanken müssten sich z. B. auf Dublin Core-Metadaten und eine einheitliche Klassifikation wie die DDC einigen. Damit würden homogene Datenräume geschaffen, deren Konsistenz qualitativ hochwertige Recherchen erlaubt. Leider gibt es jedoch klare Anzeichen dafür, dass die traditionellen Verfahren der Standardisierung an ihre Grenzen stoßen.

¹ Eine Ausnahme von dieser Regel ist ViBSoz, die virtuelle Fachbibliothek für die Sozialwissenschaften, die erste Heterogenitätskomponenten im Sinne von Abschnitt 3 einsetzt, siehe http://vibsoz.bonn.iz-soz.de/ViBSoz_Start.html.

Einen Überblick über die hier einschlägigen Forschungsprogramme des BMBF und der DFG einschließlich des bisher erreichten Standes der Serviceleistungen gibt Schöning-Walter 2003

² Als Überblick siehe Krause/Niggemann/Schwänzl 2003

Bereits im traditionellen Bibliotheksbereich waren sie häufig mehr Anspruch als Realität. Einerseits erscheinen sie unverzichtbar und haben in Teilbereichen deutlich die Qualität der wissenschaftlichen Informationssuche gesteigert. Andererseits sind sie im Rahmen globaler Anbieterstrukturen von Informationen nur noch partiell durchsetzbar, bei steigenden Kosten. Besonders im Teilbereich der Inhaltserschließung wird deutlich, dass für virtuelle Fachbibliotheken - bei allen notwendigen Bemühungen - nicht von der Durchsetzbarkeit einheitlicher Standards der Inhaltsbeschreibung ausgegangen werden kann. Deshalb muss eine neue Sichtweise auf die bestehen bleibende Forderung nach Konsistenzerhaltung und Interoperabilität gefunden werden. Sie wird uns in Abschnitt 3 unter dem Motto „Standardisierung ist von der verbleibenden Heterogenität her zu denken“ weiter beschäftigen.

2.2 Publizieren im Netz

Das Web geht in seinen Überlegungen zur Modellierung dezentraler Informationsräume deutlich über die Verbindung von Teilbibliotheksbeständen z.B. über eine Z39.50-Schnittstelle hinaus. Systementwicklung und der Datenaufbau von Informationssammlungen sind hier mit dem Paradigma des „Publizierens im Web“ zu verbinden, das am klarsten in den semantic web-Ansätzen, aber auch in Initiativen wie DDI oder OAI (open archive initiative)³ zum Ausdruck kommt.

Die Vision hinter diesen Bemühungen wird z.B. deutlich in den Projekten NESSTAR und FASTER⁴ aus dem Bereich sozialwissenschaftlicher Datenarchive sichtbar, deren Zielvorstellung die folgende Abbildung verdeutlicht. Sie enthält auch die Verbindung von textuellen Elementen (z. B. Publikationen) mit Faktendaten⁵.

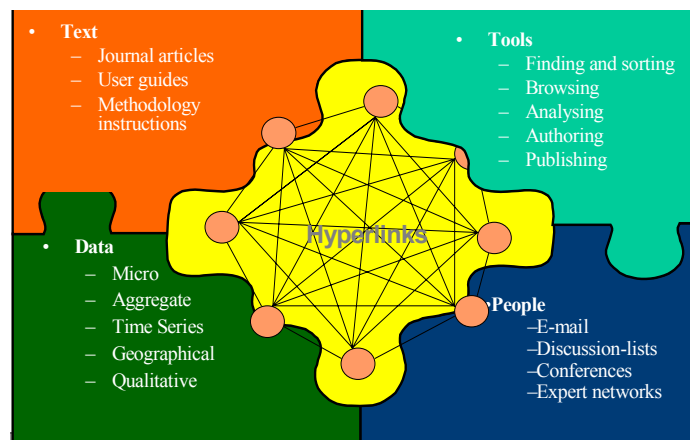


Abbildung 2: NESSTAR aus Ryssevik 2002:14

³ Als Einstieg zum semantic web siehe Matthews 2002; Data Documentation Initiative (DDI) (<http://www.icpsr.umich.edu/DDI/>); Open Archives Initiative (OAI) (<http://www.openarchives.org/>)

⁴ Networked Social Science Tools and Resources (NESSTAR) (<http://www.nesstar.com/>); Flexible Access to Statistics, Tables and Electronic Resources (faster) (<http://www.faster-data.org/>)

⁵ In Deutschland entspricht dieser Ansatz den Planungen der KVI-Projekte (siehe Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik 2001). Bei ihnen geht es um die Verbindung von Faktendaten der statistischen Ämter in Deutschland, die vor allem prozessgeneriert sind, mit den in der Wissenschaft entstehenden (z. B. Umfragen wie der ALLBUS, siehe (<http://www.gesis.org/Datenservice/ALLBUS/index.htm>)). Dabei sollen die bestehenden Datenbanken durch fehlende Metadaten angereichert, weitgehend standardisiert und möglichst integriert angeboten werden. Es sollen neue Metadatensysteme für Mikrodaten an unterschiedlichen Standorten von Sozialwissenschaftlern aus Wissenschaft und statistischen Ämtern realisiert werden.

Das Paradigma des „Publizierens im Web“ macht eines klar: Es war noch nie so schwierig wie heute, neue Informationssysteme zu modellieren und in die Praxis umzusetzen, was vor allem an der Grundforderung jeder Web-Aktivität liegt, die unter dieser Prämisse entsteht. Jedes neue Angebot „... is designed to fit into a wider data input and output environment“ (Musgrave 2003: 5). Frühere Systementwicklungen mussten sich „nur“ darum kümmern, dass ihr System für sich genommen effizient und schnell Anfragen zuließ und die Benutzerbedürfnisse umsetzte. Heute genügt dies nicht. Niemand arbeitet mehr isoliert für sich und seine Benutzergruppe. Jeder ist Teil eines globalen Angebots und erfüllt in diesem fachwissenschaftlichen Informationskontext nur eine kleine, spezielle Aufgabe. Dies gilt auch für die Bibliotheken. Die Benutzer einer speziellen Datenbank werden sich nicht auf dieses eine Angebot beschränken, sondern auf viele vergleichbare Bestände integriert zugreifen wollen. Einige dieser Cluster sind bei Beginn der Entwicklung eines neuen Angebots schon bekannt. Wichtiger ist jedoch, dass mit Sicherheit in den nächsten Jahren nach Fertigstellung des eigenen Angebots neue Informationssammlungen im Web hinzukommen werden, auf die der Benutzer integriert zugreifen will.

Und weil man dies weiß, liegt die eigentliche Schwierigkeit nicht in der konkreten Systemprogrammierung, sondern in der Modellierung des Systems als passfähige Teileinheit. Idealerweise sieht sich die Web-Community als Gemeinschaft von Systemanbietern, deren Einzelbeiträge so modelliert sind, dass jedes Teilelement zum anderen passt, ohne dass es zuvor konkrete Absprachen zwischen den Beteiligten gab – einfach weil es „richtig“ im Sinne des Web-Paradigmas modelliert und programmiert wurde. Jeder Datenbestand soll von jedem Anbieter von Systemleistungen problemlos gelesen und weiterverarbeitet werden können, und jedes programmierte Systemmodul sollte idealerweise von jemand anderem in seine eigenen Systemvorstellungen integriert und weiterentwickelt werden können, ohne die schon geleistete Entwicklungsarbeit noch einmal machen zu müssen, weil das bereits existierende Modul zur Eigenentwicklung nicht passfähig ist.

Kaum noch Schwierigkeiten macht heute unter diesem Paradigma die Protokollebene (z. B. http, JDBC) und weitgehend auch die Syntaxebene (HTML und zumindest bei der Modellierung fast schon durchgesetzt XML). Auf dieser standardisierten und „passfähigen“ Basis arbeiten heutige professionell entwickelte Systeme. Erst damit lassen sich überhaupt Suchmaschinen konstruieren, die ohne vorangegangene Abstimmungsprozesse beliebige Server weltweit besuchen und deren Daten indexieren. Dass diese Standardisierung zu kurz greift und keine für fachwissenschaftliche Zwecke ausreichende Ergebnisqualität erreichen kann, gilt heute als gesichert. Es muss weitergehende Standardisierungen bei der Strukturierung und bei den Inhalten geben. Parallel zu den Überlegungen zur Inhaltserschließung und ihrer verbleibenden Heterogenität beim Leitbild der virtuellen Fachbibliotheken stellt Musgrave für den Beispielbereich der sozialwissenschaftlichen Datenarchive fest:

“On top of the syntax provided by XML and the structure provided by the DDI there is a need to develop more standard semantics via thesauri and controlled vocabularies in order to make for better interoperability of data.” (Musgrave 2003: 1).

Bei den Strukturen ist die internationale Kooperation der Datenarchive dank DDI schon recht weit gediehen (siehe Ryssevik 2002 als Einstieg und die DDI Homepage: <http://www.icpsr.umich.edu/DDI/ORG/index.html>)⁶. Kontrollierte Vokabularien und Thesauri werden sich dagegen in vielen Teilbereichen nicht zu sog. Metathesauri zusammenfassen lassen. Dies ist auch gar nicht nötig, da hier Heterogenitätskomponenten gut greifen (siehe Abschnitt 3).

⁶ Einen Überblick über die Bibliothekslandschaft, die fachwissenschaftlichen Angebote der Informationszentren und die virtuellen Fachbibliotheken gibt Krause/Niggemann/Schwänzl 2003.

Die Grenzen heutiger Entwicklung liegen beim Austausch und der „Passfähigkeit“ der Funktionalität. Ansätze wie die der Agentensysteme und die semantic web-Initiative zeigen den Weg für zukünftige Systeme in Umrissen auf (siehe Matthews 2002).

Fazit: Die diskutierte Leitlinie des „Publizierens im Netz“ geht deutlich über die Dezentralisierungsdiskussion bei den digital libraries hinaus. Sie charakterisiert den IT-Wandel des letzten Jahrzehnts durch die Verbreitung des WWW am deutlichsten. Ihm sind auch alle Bibliotheksaktivitäten unterworfen. Er ist nicht nur technologisch, sondern inhaltlich-konzeptuell. Er lässt sich nur kooperativ, im Zusammenwirken aller bisher an der Informationsversorgung Beteiligten bewältigen, die jeweils ihre Fachkompetenz einbringen und sich neuen Lösungsansätzen öffnen. Die Zeiten einfacher, nur technisch orientierter Lösungen sind auch für jede Art von access point für und an Bibliotheken vorbei, genauso wie die Hoffnung, dass sich informationstechnologisches Know-how für fachwissenschaftliche Angebote allein auf zusätzlich erworbene Programmierkenntnisse von Fachwissenschaftlern reduzieren lässt.

2.3 Standardisierung bei dezentraler Organisation

Auch unter dem Paradigma des „Publizierens im Web“ setzen Bemühungen, die Homogenität und Konsistenz in der heutigen dezentralen Informationswelt wiederherzustellen, auf die Schaffung geeigneter Informationssysteme, die mit verteilten Datenbeständen effizient umgehen können und auf die Einhaltung von Standards.

- Ersteres Vorgehen steht für die technikorientierte Sichtweise von Problemlösungen. Man sorgt dafür, dass physikalisch auf die verschiedenen Dokumentenräume gleichzeitig zugegriffen werden kann und dass dies effizient geschieht. Die meisten Publikationen zu einer Literaturrecherche über „verteilte Datenbanken, integrierte Informationssysteme“ befassen sich hiermit. Diese technikorientierten Lösungen des Problems dezentraler Dokumentenräume sind eine unabdingbare Voraussetzung, lösen jedoch das Hauptproblem der inhaltlichen und konzeptuellen Differenzen zwischen den einzelnen Dokumentenbeständen noch nicht.
- Einen Schritt weiter gehen die Ansätze zur Einführung von Metadaten. 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 hierfür ist der Dublin Core (<http://dublincore.org/>). Metadaten unterstützen zumindest ein Minimum an technischer und konzeptueller Austauschbarkeit. Z. B. sieht Jeffery 1998 in der Akzeptanz von Metadaten die wesentliche Lösungsstrategie für einen Verbund europäischer Projektnachweisdatenbanken.

Standardisierungsbemühungen und Initiativen zur Akzeptanz und Verbreitung von Metadaten sind fraglos wichtig und eine Voraussetzung für übergreifende Suchprozesse in einer täglich dezentraler und polyzentrischer werdenden Informationswelt. Im Prinzip versuchen sie - auf einem niedrigeren Niveau - dasselbe wie die zentralistischen Ansätze der 70er Jahre, die auf Kooperationsabsprachen abstellen, ohne jedoch über hierarchische Durchsetzungsmittel zu verfügen. Besonders deutlich im Bereich der Inhaltserschließung versuchen sie, die verloren gegangene Datenhomogenität und Konsistenz durch freiwillige Absprachen aller am Informationsprozess Beteiligten wiederherzustellen. Obige Überlegungen zur Heterogenitätsbehandlung mit der klassischen Forderung nach übergreifenden Standardisierungsbemühungen zu begegnen, liegt auch nahe und ist per se nicht verkehrt: Wenn alle den gleichen Thesaurus benutzen oder die gleiche Klassifikation, brauchen wir keine Heterogenitätskomponenten. Wenn die einzelnen Anbieter von dieser Grundprämisse jedes Standardisierungsansatzes ab

gewichen sind, müsse man sie „irgendwie“ dazu bringen (sie zwingen), sich an die klassischen Spielregeln zu halten.

Solange man sich darüber im Klaren ist, dass dies nur teilweise gelingen kann, spricht alles für Initiativen dieser Art. Ganz gleich, 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. Der breite Zugang zu den Netzwerken wirkt einer zentralistischen Doktrin der Informationserschließung - die deutlich auch in den Metadatenaktivitäten spürbar bleibt - per se entgegen. Hier sind auch die Bemühungen der deutschen Bibliotheksszene positiv einzuordnen, zu einer Aufgabe der RAK-WB und des MAB-Formats zu gelangen und den Übergang zu AACR2 und MARC21 verbindlich für alle deutschen Bibliotheken festzulegen⁷.

Überall auf der Welt können Gruppen auftreten, die zu Spezialgebieten Informationen sammeln. Der Benutzer wird auf sie zugreifen wollen, gleich nach welchen Verfahren sie inhaltlich erschlossen oder in welchem System sie angeboten werden. Das obige Kooperationsmodell würde verlangen, dass die zuständige Informationsservicestelle mit diesen Anbietern Kontakt aufnimmt und sie überzeugt, bestimmte Normen der Dokument- und Inhaltserschließung einzuhalten (z.B. den Dublin Core). Das mag im Einzelfall funktionieren, jedoch nie als generelle Strategie. Es wird immer eine Fülle von Angeboten geben, die sich vorgegebenen Leitvorstellungen nicht unterordnen lassen. Früher lehnten die zentralen Informationsservicestellen Dokumente ab, die nicht bestimmte Regeln der Erschließung einhielten, wodurch der Benutzer (idealiter) immer einem homogenisierten Datenbestand gegenüberstand. Darauf war die gesamte IuD-Methodik, einschließlich der Verwaltungsstruktur der Bibliotheken und Fachinformationszentren ausgerichtet. Ob man dies für richtig oder falsch hält, diese Ausgangssituation ist in einem System weltweiter Vernetzung auch in der abgeschwächten Form der Absprache von Metadaten nicht mehr gegeben. Das Postulat der Datenkonsistenz als wesentlicher Eckpfeiler heutigen IuD-Handelns erweist sich schon aus diesem Grund als Illusion.

Auf diese Veränderung muss die heutige IuD-Landschaft reagieren. Es sind Konzepte zu entwickeln, die mit statt gegen diese Deregulation arbeiten. Denkbar erscheinen höchstens Metadatenvereinbarungen, die in abgeschwächter Form einen Teil der verlorengegangenen Datenkonsistenz zurückgewinnen. Somit stellt sich die Frage, welche Modellvorstellungen für den verbleibenden Teil der Heterogenität auf den verschiedensten Ebenen entwickelt werden können.

3 Verbleibende Heterogenität im Bereich der Inhaltserschließung

Will man Literaturinformationen - und später Fakteninformationen und multimediale Daten - aus verteilten und inhaltlich unterschiedlich erschlossenen Datenbeständen, die sich in miteinander nicht verbundenen, heterogenen Organisationsstrukturen und Zugänglichkeitskontexten befinden (Institutsbibliotheken, Sondersammelgebiete der Universitätsbibliotheken, wissenschaftliche Spezialbibliotheken, Referenzdatenbanken, digitale Volltexte) mit einer Anfrage integriert recherchieren, müssen die Probleme des inhaltlichen Zugriffs auf verteilte Dokumentenbestände gelöst werden. Ein vom Benutzer gewähltes Stichwort (Deskriptor) X kann in den verschiedenen Dokumentenbeständen die unterschiedlichsten Bedeutungen annehmen. Auch im engen Bereich der Fachinformation kann ein Term X, der aus einem hochrelevanten,

⁷ Ein erster Beschluss in diese Richtung wurde Ende 2001 vom „Standardisierungsausschuss“, dem zuständigen Gremium der deutschen Bibliotheken gefasst. Es sind allerdings noch viele Fragen offen, die derzeit in einem von der Deutschen Bibliothek koordinierten Projekt bis Ende 2003 beantwortet werden sollen.

mit viel Aufwand qualitativ hochwertig ermittelten Dokumentenbestand stammt, nicht mit dem Term X gleichgesetzt werden, den z. B. eine automatische Indexierung auf der Basis von Titeln aus einem Randgebiet liefert. Deshalb genügt eine rein technologische Verknüpfung verschiedener Dokumentenbestände und die formale Integration unter einer Benutzungsoberfläche allein nicht. Sie führt zum fehlenden Nachweis relevanter Dokumente und zu einer Fülle von irrelevanten Treffern.

Im Kontext fachwissenschaftlicher Informationen ist die Problematik der heterogenen und mehrfachen Inhaltserschließung generell besonders kritisch, weil die Heterogenität der Datentypen besonders 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.

Trotz dieser heterogenen Ausgangslage soll der Benutzer z. B. nicht gezwungen werden, sich zuerst in das Erschließungssystem einer Bibliothek einarbeiten zu müssen, um dann bei einer Erweiterung seiner Suchintention auf unselbständige Publikationen ein zweites System der Inhaltserschließung lernen und in geeignete Suchstrategien umsetzen zu müssen, und wiederum ein anderes, wenn er weitere fachwissenschaftliche Datenbanken ergänzend durchsuchen möchte.

Deshalb muss die bestehende Heterogenität der verschiedenen Inhaltserschließungssysteme durch geeignete Maßnahmen aufeinander bezogen werden. Dabei stellt die Integration von Fachdatenbanken und Bibliotheksbeständen nur eine erste Stufe dar. Sie ist durch Internetquellen und Faktendaten (z. B. Zeitreihen zu Umfragen wie in NESSTAR) zu ergänzen, generell um alle Datentypen, die wir heute bei virtuellen Fachbibliotheken, auf den unterschiedlichsten Fachportalen oder elektronischen Marktplätzen finden.

3.1 Bilaterale Transfermodule

Das im Folgenden kurz skizzierte Modell stellt einen allgemeinen Rahmen dar, in dem sich bestimmte Klassen von Dokumenten mit unterschiedlicher Inhaltserschließung analysieren und algorithmisch aufeinander beziehen lassen. Zentral sind intelligente Transferkomponenten zwischen den verschiedenen Formen der Inhaltserschließung, die den semantisch-pragmatischen Differenzen Rechnung tragen und sich als eigenständige Agenten modellieren lassen. 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.

Deshalb sind Transfermodule zwischen jeweils zwei Datenbeständen unterschiedlichen Typs zu entwickeln, die den Übergang nicht nur technisch, sondern konzeptuell gestalten (Genauer zum „bilateralen“ Transfer in Krause 2003a).

Generell gibt es drei Verfahrensweisen, die in Bezug auf ihre Wirksamkeit im Einzelfall zu überprüfen und zu implementieren wären. 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. Das Konzept

darf nicht mit dem der Metathesauri verwechselt werden. Es wird keine neue Standardisierung bestehender Begriffswelten angestrebt. Crosskonkordanzen enthalten nur die partielle Verbindung zwischen bestehenden Terminologiesystemen, deren Vorarbeit genutzt wird. Sie 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 modellieren. Für die im Information Retrieval behandelte Vagheit zwischen den Termen der Benutzeranfrage und denen des Datenbestandes sind verschiedene Verfahren vorgeschlagen worden (probabilistische Verfahren, Fuzzy-Ansätze und neuronale Netze (Mandl 2000), die sich auf die Transferproblematik anwenden lassen. Verfahren dieser Art benötigen Trainingsdaten, bei denen einzelne Dokumente nach zwei Begriffsschemata erschlossen oder bei denen zwei unterschiedliche und unterschiedlich erschlossene Dokumente in Beziehung gesetzt 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 Sonder-sammelgebiets der Universität Köln gleichzeitig in SOLIS erfasst.

In ELVIRA⁸, einem IZ-Projekt, bei dem es um die Zusammenführung von Texten und Fakten (Zeitreihen) für ein Verbandsinformationssystem geht, wurde eine Kookkurrenzanalyse zwischen Zeitreihen und automatisch indextierten Texten eingesetzt. Alle bisher getesteten Module basieren auf einem Ansatz, der von Sheridan/Ballerini 1996 für das multilinguale Retrieval vorgeschlagen wurde. Im Unterschied zu den Crosskonkordanzen basiert die 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.

Welches Verfahren sich als erfolgversprechend erweist, lässt sich nur empirisch auf der Basis realer Daten klären.

- 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 (Belkin 1996, Ingwersen 1996), bei intelligenten Recherchesystemen wie OSIRIS (Zillmann 1997, 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 Vagheitspro

⁸ siehe <http://www.gesis.org/Forschung/Informationstechnologie/ELVIRA.htm>

blematik zwischen Benutzeranfrage und dem Dokumentenbestand der Datenbank, der in den Recherchealgorithmus integriert ist. So können die Transfermodule z. B. zwischen einem Dokumentbestand, der mit einer generellen Schlagwortliste wie der SWD⁹ indexiert wurde und einem zweiten, dessen Indexierung auf einem speziellen fachspezifischen Thesaurus beruht, durch qualitative Verfahren wie den Crosskonkordanzen und Deduktionsregeln aufeinanderbezogen werden und der Recherchealgorithmus die Verbindung zur Benutzerterminologie mit einem probabilistischen Verfahren herstellen. Diese Möglichkeit, den Transfer jeweils differenziert an die Gegebenheiten der Verbindung unterschiedlicher Datentypen anpassen und dem Problem unterschiedlicher Begriffssysteme nicht nur undifferenziert im Recherchealgorithmus begegnen zu können, ist ein wesentlicher Unterschied zu den bisherigen IR-Lösungen.

3.2 *Standardisierung von der verbleibenden Heterogenität her denken*

Heterogenitätskomponenten eröffnen eine neue Sichtweise auf die bestehen bleibende Forderung nach Konsistenzerhaltung und Interoperabilität. Sie lässt sich durch die folgende Prämisse umschreiben: Standardisierung ist von der verbleibenden Heterogenität her zu denken. Da fachwissenschaftliche Angebote heute in unterschiedlichen Kontexten mit verschiedenen Inhaltserschließungstraditionen entstehen (Bibliotheken, Fachinformationszentren, Web-Community), treffen in ihnen die Regeln und Standards aufeinander, die für die verschiedenen Welten jeweils Gültigkeit haben.

Die Quintessenz „Standardisierung ist von der verbleibenden Heterogenität her zu denken“ wird in Krause/Niggemann/Schwänzl 2003 weiter erläutert und in Krause 2003a mit dem heute erreichten Stand der Heterogenitätsbehandlung bei fachwissenschaftlichen Informationsangeboten verbunden. Sie betrifft vor allem die klare Abkehr der Benutzer vom geringen Qualitätsstandard herkömmlicher genereller Suchmaschinen im Web und hierbei die Art der Inhaltserschließung, ihre faktisch existierende und partiell nicht aufhebbare Unterschiedlichkeit zwischen den verschiedenen bestehenden Dokumentenbeständen und dem einzuschlagenden Lösungsweg:

„Trotz freiwilliger Absprachen aller am Informationsprozess Beteiligten ist jedoch eine durchgehende Homogenität der Daten nicht mehr herzustellen. Der verbleibenden und unvermeidlichen Heterogenität muss daher durch verschiedene Strategien begegnet werden. Neue Problemlösungen und Weiterentwicklungen sind deshalb in beiden Bereichen nötig:

- den Metadaten
- und den Methoden des Umgangs mit der verbleibenden Heterogenität.“

(Krause/Niggemann/Schwänzl 2003: 27)

Krause/Niggemann/Schwänzl 2003 stellen auch fest, dass beide Anforderungen eng zusammenhängen. Durch die Fortentwicklung im Bereich der Metadaten soll einerseits die verlorengegangene Konsistenz partiell hergestellt werden. Andererseits lassen sich mit Verfahren zur Behandlung von Heterogenität Dokumente unterschiedlichen Niveaus der Datenrelevanz und Inhaltserschließung aufeinander beziehen.

⁹ SWD ist die von den deutschen wissenschaftlichen Universalbibliotheken kooperativ aufgebaute Schlagwortnormdatei auf der Basis des Regelwerks RSWK „Regeln für den Schlagwortkatalog“, siehe Krause/Niggemann/Schwänzl 2003.

4 Fazit und Ausblick

Die Probleme beim Aufbau fachwissenschaftlicher Informationsangebote – sei es als access point bei den Bibliotheken oder als „Markplatz“ oder Wissenschaftsportal anderer Informationsanbieter - führen weit über die bisher gewohnten Denkweisen von Informationszentren und Bibliothekaren hinaus. Die diskutierte Leitlinie „Standardisierung von der Heterogenität her zu denken“ und das Paradigma des „Publizierens im Netz“ charakterisieren den Wandel am deutlichsten. Er ist nicht nur technologisch, sondern inhaltlich-konzeptuell. Er lässt sich nur kooperativ, im Zusammenwirken aller bisher an der Informationsversorgung Beteiligten bewältigen, die jeweils ihre Fachkompetenz einbringen und sich neuen Lösungsansätzen öffnen. Die Zeiten einfacher, nur technisch orientierter Lösungen ist vorbei, genauso wie die Hoffnung, dass sich informationstechnologisches Know-how für fachwissenschaftliche Angebote auf Programmierkenntnisse reduzieren lassen.

Benutzerbefragungen der letzten Zeit zeigen deutlich, dass Kunden von Informationsdienstleistungen die folgenden Ziele für fachwissenschaftliche Informationen verfolgen¹⁰:

- Der primäre Einstieg wird über ein fachwissenschaftliches Portal gewünscht
- Benachbarte Fächer mit Überschneidungsbereichen wie Mathematik – Physik oder Sozialwissenschaften – Bildung – Psychologie - Wirtschaft sollen für die Abfrage ein integriertes Cluster bilden
- Die Qualität der inhaltlichen Erschließung muss deutlich über der heutiger genereller Suchmaschinen liegen (kein „Müll“)
- Nicht nur Metadaten und Abstracts sind bei der Literatur gefordert, auch der direkte Zugang zu Volltexten
- Nicht nur Bibliotheks-OPACs und Literaturfachdatenbanken, sondern auch Forschungsprojektdaten, Institutionenverzeichnisse, WWW-Quellen, Faktendaten sollen unter einem fachwissenschaftlichen Portal integriert werden.
- Alle Teilkomponenten sind möglichst hochintegriert anzubieten. Der Benutzer will wie bei der menschlichen Beratung 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“

Die Erfüllung von Wünschen dieser Art bedeutet auch unter dem Paradigma, Standardisierung von der verbleibenden Heterogenität her zu denken, und bei der Akzeptanz der Leitlinie des „Publizierens im Netz“, dass noch viele zusätzliche Fragen offen bleiben. Z.B. ist das Problem des Zusammenspiel universaler Bibliotheksangebote und das der fachspezifisch aufbereiteten Literatursammlung von Fachinformationszentren zu klären, wenn man übergreifende Wissensportale wie VASCODA in Deutschland (<http://www.vascoda.de/>) realisieren möchte. Beide Leitlinien schaffen jedoch einen tragfähigen Ausgangspunkt. Die Konsequenz der Veränderungen, die sich in den obigen Benutzerforderungen widerspiegeln, sind hochkomplexe Strukturen, die auch im Detail zu Fragestellungen führen, für die es keine vorgefertigten Lösungsmodelle mehr gibt, auf die bisher die Bibliothekare und die „Macher“ der Informationszentren zugegriffen haben.

¹⁰ Zu den vorläufigen Ergebnissen der Studie des Bibliotheksausschusses DFG siehe den Bericht aus der DFG in ZfBB H. 1, 2003, S. 41 - 44; zu den BMBF-Studien siehe Schöning-Walter 2003; zur GESIS-Studie: Binder et. al. 2001; zur Infoconnex-Studie s. IMAC 2002.

Literatur

- Belkin, Nicolas J. (1996): Intelligent Information Retrieval: Whose Intelligence? In: Krause, Jürgen; Herfurth, Matthias; Marx, Jutta (Hrsg.): Herausforderungen an die Informationsgesellschaft. Konstanz 1996, S. 25 - 31.
- Binder, Gisbert; Klein, Markus; Porst, Rolf; Stahl, Matthias (2001): GESIS-Potentialanalyse: IZ, ZA, ZUMA im Urteil von Soziologieprofessorinnen und -professoren, GESIS-Arbeitsbericht, Nr. 2. Bonn, Köln, Mannheim.
- Castells, Manuel (2001): Der Aufstieg der Netzwerkgesellschaft. Teil 1. Opladen, S. 31-82.
- Cigan, Heidi: Der Beitrag des Internet für den Fortschritt und das Wachstum in Deutschland. Hamburg: Hamburg Institute of International Economics, 2002. (HWWA-Report 217)
- IMAC (2002): Projekt Volltextdienst. Zur Entwicklung eines Marketingkonzepts für den Aufbau eines Volltextdienstes im IV-BSP. IMAC Information & Management Consulting. Konstanz. September 2002.
- Ingwersen, Peter (1996): The Cognitive Framework for Information Retrieval: A Paradigmatic Perspective. In: Krause, Jürgen; Herfurth, Matthias; Marx, Jutta (Hrsg.): Herausforderungen an die Informationswirtschaft. Konstanz 1996, S. 25 - 31.
- Jeffery, Keith G. (1998): ERGO: European Research Gateways Online and CERIF: Computerised Exchange of Research Information Format, ERCIM News No.35.
- Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.) (2001): Wege zu einer besseren informationellen Infrastruktur. Nomos, Baden-Baden.
- Krause, Jürgen (2003a): Wie mache ich das konkret, die Standardisierung von der Heterogenität her zu denken? Bilaterale Transferkomponenten als Basis. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 50, Nr. 3.
- Krause, Jürgen (2003b): Suchen und "Publizieren" fachwissenschaftlicher Informationen im WWW. In: Medieneinsatz in der Wissenschaft: Tagung: Audiovisuelle Wissensmedien online; Informationsveranstaltung der IWF Wissen und Medien gGmbH, Göttingen, 03.12. - 04.12.2002. Wien: Lang. (IWF: Menschen, Wissen, Medien) erscheint Spätsommer 2003.
- Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland (2003): Normierung und Standardisierung in sich verändernden Kontexten: Beispiel: Virtuelle Fachbibliotheken. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 50, Nr. 1, S. 19 - 28.
- Mandl, Thomas (2001): Tolerantes Information Retrieval. Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Dissertation. Konstanz: UVK, Univ.-Verl. (Schriften zur Informationswissenschaft; Bd. 39).
- Matthews, Brian M. (2002): Integration via Meaning: Using the Semantic Web to deliver Web Services. In: Wolfgang Adamczak, Nase, Annemarie (eds.): Gaining Insight from Research Information. Proceedings of the 6th International Conference on Current Research Information Systems, Kassel University Press. Kassel, S. 159 - 168.
- Musgrave, Simon (2003): NESSTAR Software Suite. <http://www.nesstar.org> (January 2003)
- Ronthaler, Marc; Zillmann, Hartmut (1998): Literaturrecherche mit OSIRIS; ein Test der OSIRIS-Retrievalkomponente. In: Bibliotheksdienst, 32 (7): 1203 - 1212.

Ryssevik, Jostein (2002): Bridging the Gap Between Data Archive and Official Statistics Metadata Traditions. Power Point Presentation. <http://www.nesstar.org> (January 2003).

Schöning-Walter; Christa (2003): Die DIGITALE BIBLIOTHEK als Leitidee: Entwicklungslinien in der Fachinformationspolitik in Deutschland. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 50, Nr. 1, S. 4 - 12.

Sheridan, Páraic; Ballerini, Jean Paul (1996): Experiments in Multilingual Information Retrieval using the SPIDER System. In: Frei et al. 1996. S. 58 - 65.

Zillmann, Hartmut (1997): OSIRIS. Ein gemeinsames Projekt der Universitätsbibliothek Osnabrück und des Instituts für Semantische Informationsverarbeitung der Universität Osnabrück (ISIV). Osnabrück 1997.
http://www.ub.uni-osnabrueck.de/osiris/osiris_papers.html