

Text-Fakten-Integration in Informationssystemen

1. Einleitung

Die Informationsversorgung mit wissenschaftlicher Information - sei es direkt im Kontext von Forschung und Lehre oder im eher anwendungsorientierten bzw. kommerziellen Umfeld - ist auch in Zeiten von Virtuellen bzw. Digitalen Bibliotheken, Semantic Web und Cloud Computing stark geprägt von (fach-)spezifischen Angeboten. Häufig bestimmen Ursprung oder Urheberschaft, Format und Medium, Disziplin, geografischer Raum, Zielgruppe, Versorgungsauftrag und Geschäftsmodell sowohl den Charakter des einzelnen Angebots als auch dessen aktiv betriebene Vernetzung mit weiteren Informationsquellen. Sowohl die institutionell als auch (forschungs-)politisch motivierte Vernetzung wissenschaftlicher Information nach fachlichen (z. B. Virtuelle Fachbibliotheken oder Forschungsdatenzentren) oder geografischen Aspekten (z. B. das Wissenschaftsportal vascoda.de in Deutschland und Europeana.eu auf europäischer Ebene) führen dazu, dass bislang disparate Informationssammlungen verbunden werden und dadurch ein Potential für neue Dienstleistungen entsteht.

Die Zusammenfassung bislang getrennt vorliegender Informationsangebote ist durchaus im Sinne der Nutzerinnen und Nutzer, die darüber hinaus jedoch noch eine Reihe weiterer Anforderungen formulieren. Gerade bei den wissenschaftlichen Nutzern, die Gegenstand mehrerer umfangreicher Studien waren (vgl. IMAC 2002, RSLG 2002, Poll 2004) zeigt sich die Spannweite der – teils auf den ersten Blick widersprüchlichen – Anforderungen, die folgendermaßen zusammengefasst werden können:

- Nutzer wünschen Fachportale, die disziplinspezifisch alle für das jeweilige Fach relevanten Informationen an einer Stelle zusammenfassen. Dabei soll eine fachspezifische, möglichst tiefe Erschließung erfolgen, die dann auch eine sehr detaillierte Suche gestattet.
- Gleichzeitig soll die Information des eigenen Faches mit denen anderer Fächer eng vernetzt sein, so dass die zunehmende Zahl an interdisziplinären Fragestellungen besser beantwortet werden können. Dabei kommt durchaus eine Clusterbildung in Betracht von Fächern, die häufigere und engere Berührungspunkte und Überschneidungen haben als andere.
- Die Integration aller fachlich relevanten Informationen soll auf intelligente Art erfolgen und möglichst alle Informationstypen einschließen. Dies geht meist über reine Literaturinformationen und Volltexte hinaus und schließt auch Daten (z. B. Messreihen oder empirische Studien) und multimediale Informationen ein.
- Bei aller Informationsfülle soll die Qualität der Information und Informationsvermittlung gewährleistet bleiben, es soll also kein „Müll“ geliefert

werden – was sowohl hinsichtlich der wissenschaftlichen und dokumentarischen Qualitätskontrolle der Informationen selbst als auch hinsichtlich des Produzierten Ergebnisses interpretiert werden kann.

- Gleichzeitig soll aber auch die Quantität und Vollständigkeit der Information sichergestellt werden, die Nutzerinnen und Nutzer möchten sich also darauf verlassen, dass keine für sie relevante Information außen vor bleibt.
- Trotz Forderung nach Quantität soll keine Überlastung des Nutzers durch die schiere Menge erfolgen („information overflow“), Information soll also nach Relevanz bewertet und geordnet werden.

Durch die von den Nutzerinnen und Nutzern geforderte Bündelung bislang verteilter Informationen können nicht nur Vorteile wie der einfachere und effizientere Zugriff realisiert werden, auch eine Reihe von hauptsächlich ergonomischen Nachteilen bei der Nutzung von Einzelangeboten ließen sich beseitigen. Hierzu zählt vor allem, dass Suchanfragen auf mehrere Informationssammlungen nur mehr einmal formuliert werden müssen und dadurch für den Nutzer die Notwendigkeit entfällt, sein oft vages Informationsbedürfnis mehrfach in den einzelnen Portalen oder für die einzelnen Datenbanken zu formulieren und die resultierenden Ergebnisse zusammenzuführen. Insbesondere durch unterschiedliche Benutzungsoberflächen der Portale und durch die verschiedenen, zur Inhaltserschließung verwendeten Vokabulare (Schlagwortlisten, Thesauri, Nomenklaturen und Klassifikationen) entsteht eine hohe kognitive Last – der Nutzer muss sich erst in die unterschiedlichen Vokabulare einarbeiten, bevor er befriedigende Suchergebnisse erzielen kann.

Gerade auch in Anwendungsfällen, bei denen zwar die Anfrageformulierung über traditionelle Formulare erfolgt, die durch ihre allgemein bekannte Syntax und inhärente Reduktion der Anfragekomplexität gerade auch wenig geübten Nutzern entgegenkommen, die Ergebnisanzeige aber in Form einer Datenvisualisierung erfolgt (z. B. bei statistischen Daten), tritt sogar innerhalb einer einzelnen Anwendung ein Bruch der Interaktionsmodalität auf. Insbesondere dann, wenn zu einer weiteren Exploration der Daten oder einer Reformulierung der Anfrage der grafische Modus verlassen und zu einer formularbasierten Interaktion zurückgekehrt werden muss.

Betrachtet man schließlich den gesamten Anwendungskontext, in dem Informationsbedürfnisse entstehen, Informationen in unterschiedlicher Form beschafft, verarbeitet, miteinander verknüpft, verdichtet und zu einem Gesamtergebnis integriert werden, so ist die gemeinsame Nutzung von Information in unterschiedlichen Modalitäten (Daten, Texte, Grafiken usw.) und die Produktion multimedialer Dokumente eher die Regel als die Ausnahme. Als Beispiel mag hierfür ein idealtypischer, am Forschungsprozess orientierter Informationskreislauf dienen (Abbildung 1). Er gliedert sich in drei Segmente - Aktivitäten und Akteure, Ergebnisse, Diskurs und Kommunikation - welche prototypische Informations- und Kommunikationsschwerpunkte repräsentieren.

Das Segment *Aktivitäten und Akteure* repräsentiert dabei die Phase, in der Forschungsaktivitäten geplant und vorbereitet werden. Der Informationsschwerpunkt liegt dabei auf

Strukturwissen über eine Disziplin oder ein Forschungsthema und beantwortet Fragen nach Akteuren in einem Feld, der institutionellen Verortung von Forschungsaktivitäten, Kooperationsbeziehungen, dem Publikationsverhalten, und nachnutzbaren Forschungsergebnissen. Die Modalität der einzelnen Informationen reicht von Fakten in Datenbanken über Studien, statistische Daten und Forschungsprimärdaten bis hin zu Print- und Online-Publikationen.



Abbildung 1: Informationskreislauf im Forschungsprozess

Im Segment *Ergebnisse* liegt der Fokus auf der Produktion eigener Forschungsergebnisse, die überwiegend in Form von Publikationen vorliegen. Der Nachweis der Publikationen erfolgt im Regelfall in Literaturdatenbanken und der Volltext in elektronischer Form ist über die Online-Angebote von Verlagen, zunehmend aber auch über Open Access Repositories - frei zugängliche Dokumentserver, auf denen eine Zweitveröffentlichung von Zeitschriftenartikeln stattfindet - möglich (vgl. Stempfhuber 2009). Die gleichzeitige Veröffentlichung der zugrundeliegenden Datensätze setzt sich trotz entsprechender Aufrufe von Wissenschaftsorganisationen erst langsam durch (vgl. DFG 1998 und OECD 2007).

Den Diskurs über Forschungsergebnisse repräsentiert das Segment *Diskurs und Kommunikation*, das Informationen über wissenschaftliche Veranstaltungen, Diskussionsforen und von Wissenschaftlern selbst aufgebaute Informationskanäle (z. B. thematische Portale wie die von Sonderforschungsbereichen) enthält. Mit den technischen Möglichkeiten neuerer Web-Technologien rückt hier verstärkt die diskursive Auseinandersetzung mit wissenschaftlichen Ergebnissen in den Blickpunkt, die formelle und

informelle Kommunikation zu einem Medium verbindet und den Modusbruch zwischen Publikationen und Datenbanken überbrückt (vgl. Stempfhuber et al. 2008).

Sowohl aus informationswissenschaftlicher als auch nutzerorientierter Sicht stellt sich vor dem Hintergrund des präsentierten Informationskreislaufs die Frage, wie sowohl innerhalb als auch zwischen den Segmenten eine Informationsintegration erreicht werden kann, die auf der Ebene der Information selbst zu einer Überbrückung der Modalitäten führt, gleichzeitig aber auch auf der Ebene informationeller Prozesse eine für den Nutzer rezipierbare Integration bewirkt.

2. Text-Fakten-Integration am Beispiel Wirtschaftsinformation

Im Kontext "Informationsvisualisierung - Grafische Aufbereitung und Analyse von statistischen Daten" stellt sich dennoch die Frage nach der Relevanz textueller aber auch singulärer Fakteninformation (also solcher, die nicht in Form von Zeitreihen oder Datensätzen vorliegt). An einem Beispiel aus der Marktforschung soll daher zunächst dargestellt werden, welche Rolle unterschiedliche Informationstypen und deren Verknüpfung schon bei relativ einfachen Informationsbedürfnissen spielen und welche Art von semantischer Heterogenität sich daraus ergibt. Am Ende der Befriedigung des Informationsbedürfnisses wird gerade in diesem Anwendungsbereich eine Visualisierung stehen - oder der Wunsch, aus einer gefundenen Visualisierung heraus auf weitere, in unterschiedlichen Formaten und Quellen vorliegende Informationen zuzugreifen.

Im Projekt ELVIRA¹ zur Entwicklung eines elektronischen Informationssystems für Industrieverbände, das mittlerweile von ca. 700 Marktforschern in Industrieunternehmen eingesetzt wird, wurden zunächst etwa 100 Anfragen an die Statistikabteilungen der beteiligten Industrieverbände ausgewertet mit dem Ziel, typische Fragestellungen zu ermitteln und festzustellen, welche Arten von Informationen aus welchen Quellen zu ihrer Beantwortung herangezogen werden. Sehr schnell zeigte sich, dass sowohl hinsichtlich der Informationsart, die benötigt wird (z. B. statistische Daten, Literatur, Fakteninformation) als auch bezüglich der Quellen, aus denen diese Information stammen, eine große Spannweite besteht. So kann sich im Vorfeld der Entscheidung, ein Produkt auf einem internationalen Markt einzuführen, folgender mehrstufige informationelle Prozess abspielen:

- Um eine generelle Entscheidung für die Produkteinführung in einen Markt zu treffen, wird zunächst das Marktvolumen, d. h. die Größe des Marktes ermittelt. Hierzu werden der Umfang der Produktion des Produkts im Zielland sowie die entsprechenden Exporte und Importe ermittelt (statistische Zeitreihen).
- Scheint der Markt groß genug für ein Engagement, ist zu prüfen, ob rechtliche Bestimmungen (z. B. Aus-/Einfuhrbeschränkungen, Zollbestimmungen usw.) den

¹ Das Elektronische VerbandsInformations-, Recherche- und Analysesystem (ELVIRA) wurde zwischen 1995 und 2000 gefördert vom Bundesministerium für Wirtschaft (BMWi)

Markteintritt behindern könnten. Hierzu wird in speziellen Literaturdatenbanken recherchiert (Texte, in Texten eingebettete Fakten).

- Zur Kontaktaufnahme vor Ort werden Ansprechpartner in Behörden und Handelskammern oder auch Anwälte benötigt (Fakten).
- Einen Überblick über den derzeitigen Markt und Kundenwünsche geben z. B. Messedatenbanken oder Zeitschriften (Fakten und Texte).
- Zur Ermittlung der größten Mitbewerber auf dem Markt werden nationale und internationale Statistikdaten benötigt (statistische Zeitreihen).

Durch die Aufteilung der benötigten Informationen auf unterschiedliche Informationssysteme und -sammlungen einer Vielzahl von Anbietern im In- und Ausland sieht sich der Nutzer mit der Situation konfrontiert, dass er je nach Informationsart und -quelle sein Informationsbedürfnis mit unterschiedlichen Thesauri oder Nomenklaturen statistischer Ämter wiederholt ausdrücken und zum Zugriff auf die Information unterschiedliche Benutzungsoberflächen nutzen muss. Diese nutzerseitige Last zur Behandlung von semantischer Heterogenität (hier: wiederholte Auswahl meist sehr spezifische benannter Produkte und Waren aus umfangreichen, datenbestandsspezifischen Nomenklaturen) sollte im Projekt ELVIRA durch softwareergonomische und informationswissenschaftliche Maßnahmen reduziert werden, wozu zunächst ein Integrationsmodell entwickelt wurde, das die auftretende semantische Heterogenität beschreibt und damit einen Lösungsraum aufspannt, in dem sowohl softwareergonomische Methoden (z. B. aufgabenadaptive Steuerung des Bildschirmlayouts und der Informationspräsentation) als auch informationswissenschaftliche Verfahren (z. B. Abbildung von Wissensorganisationsystemen) aufeinander abgestimmt zum Einsatz kamen (vgl. Stempfhuber et al. 2002).

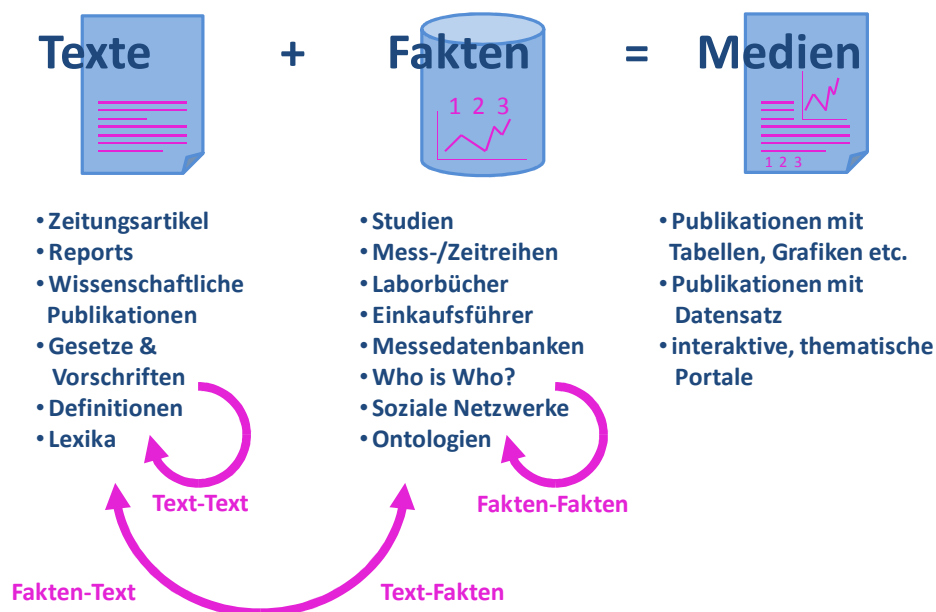


Abbildung 2: Modifiziertes Modell der Text-Fakten-Integration aus ELVIRA

Die Abbildung 2 zeigt eine modifizierte Version des Text-Fakten-Integrationsmodells aus ELVIRA (vgl. Krause et al. 1997). Es geht zunächst davon aus, dass für eine Vielzahl von

Informationsbedürfnissen aus der Erfahrung des Nutzers heraus "Prototypen" von Information existieren, die sein Informationsbedürfnis vermutlich am besten befriedigen. In vielen Standardfällen der Marktforschung, wenn zum Beispiel die Position der eigenen Unternehmung im Markt oder auch die Konjunktorentwicklung beurteilt werden soll, stehen die benötigten Informationsobjekte a priori bereits fest (z. B. Zeitreihen aus der amtlichen Statistik), genauso wie die Präsentation des Ergebnisses (z. B. eine parallel Visualisierung des zeitlichen Verlaufs mehrerer Merkmale). Ebenso eindeutig wird für Hintergrund und Interpretationswissen direkt auf Publikationen (z. B. Zeitschriftenartikel oder wissenschaftliche Literatur) zugegriffen und die gewonnenen Erkenntnisse (z. B. in den Texten enthaltene Fakten) in Textform aufbereitet (z. B. thematischer Pressespiegel). In diesen Standardsituationen werden Nutzer also direkt auf Fakten oder Texte zugreifen - und vermutlich nur selten nach Informationen eines anderen Typs suchen.

Aus informationswissenschaftlicher Sicht interessanter sind allerdings die Fälle, in denen sich entweder die zunächst gewählte Informationsquelle die gesuchte Information nicht enthält, oder eine andere Informationsart herangezogen werden muss. Beispiele hierfür sind der Wechsel von den Daten des Statistischen Bundesamts zu einem anderen, evtl. ausländischen Datenlieferanten, der Bedarf nach Originaldaten zur Überprüfung der Validität einer Aussage in einer Publikation, die Ursachenforschung bei Auffälligkeiten in statistischen Daten und das Schließen von Datenlücken oder zeitlichen Verzögerungen der Datenerhebung durch Rückgriff auf Publikationen, die die gewünschten Fakten näherungsweise enthalten.

Hier unterstützt das Informationssystem idealtypisch den Nutzer indem es ausgehend von der bereits formulierten Anfrage oder vom aktuell betrachteten Informationsobjekt - egal ob Zeitreihe, Text oder ein anderer Informationstyp - möglichst automatisiert eine semantisch äquivalente Anfrage an das selbe oder ein besser geeignetes Informationssystem stellt. Im Projekt ELVIRA erfolgte die erstmalige prototypische Implementierung auf der Basis von Abbildungen zwischen Nomenklaturen der amtlichen Statistik und dem Standardthesaurus Wirtschaft (STW), die sowohl intellektuell als auch mit statistischen Verfahren erstellt wurden. Anfragen nach statistischen Zeitreihen, die mit den amtlichen Nomenklaturen formuliert wurden, wurden dabei automatisch zu Anfragen nach Texten transformiert, wobei gewichtete Schlagwortkombinationen aus dem STW verwendet wurden. Ebenso ermöglichte das System die Recherche von Texten, wobei auf Wunsch die einzelnen Suchbegriffe in Nomenklaturpositionen transformiert wurden - wodurch gleichzeitig mit den Texten auch Zeitreihen nachgewiesen wurden. Eine dahingehend optimierte Benutzungsoberfläche unterstützt die Nutzer bei der Verifikation des Transformationsergebnisses (vgl. Kim et al 2001).

In den Projekten infoconnex und KoMoHe (vgl. Walter et al 2006, Mayr&Petras 2008; zu den technischen Aspekten Baerisch et al. 2010) wurden auf der Basis der Erfahrungen aus ELVIRA die Grundlagen für die breitere Anwendung automatischer Verfahren zur Abbildung von Anfragen zwischen heterogen erschlossenen Informationsquellen gelegt, indem ein umfangreiches Netz von Abbildungen zwischen Thesauri und Klassifikationen –

schwerpunktmäßig für die Sozialwissenschaften, aber auch über eine Vielzahl anderer hinweg – erstellt wurde. Dies ist seit 2003 im Portal infoconnex.de, seit 2007 in sowiport.de und seit 2009 in vascoda.de erfolgreich im Einsatz.

3. Ein Schichtenmodell für die Text-Fakten-Integration

Die bisherigen Ausführungen zur Text-Fakten-Integration verdeutlichen deren Relevanz für informationelle Prozesse, in denen heterogene Informationen ggf. aus unterschiedlichen Quellen nutzerseitig zusammengeführt werden sollen. Eine Informationsvisualisierung kann dabei sowohl der Ausgangspunkt sein - wenn sie für den Nutzer relevante Phänomene in Daten enthüllt, zu deren Erklärung weitere Informationen beschafft werden müssen (Fakten zu Text) - oder sie kann das Ziel sein, wenn zu Aussagen in Texten relevante Fakten gefunden und in eine Visualisierung gebracht werden (Text zu Fakten).

In beiden Fällen spielen die Informationsarchitektur und die verfügbaren Wissensorganisationssysteme eine wesentliche Rolle, da sie die Präzision, mit der das Informationssystem die Anfrage des Nutzers zwischen unterschiedlichen Wissensorganisationssystemen transformiert, beeinflussen. Während das oben beschriebene Verfahren der Termtransformation zwischen Thesauri lediglich auf die Ebene der (intellektuellen) Inhalts- bzw. Sacherschließung abzielt, bleiben insbesondere der Entstehungs- und Verwendungskontext von Informationen außer Acht. Ziel muss es daher sein, alle für eine spezifische Datenart verfügbaren Ebenen der Beschreibung mit derer anderer Datenarten in einem Gesamtmodell zu integrieren, so dass bislang unverbunden nebeneinander stehende Verfahren der Wissensorganisation koordiniert innerhalb informationeller Prozesse eingesetzt werden können. Die Abbildung 3 zeigt ein solches Modell, das von der Datenerzeugung bis zur Präsentation im Anwendungskontext die jeweils verfügbaren Wissensorganisationssysteme in Beziehung setzt.

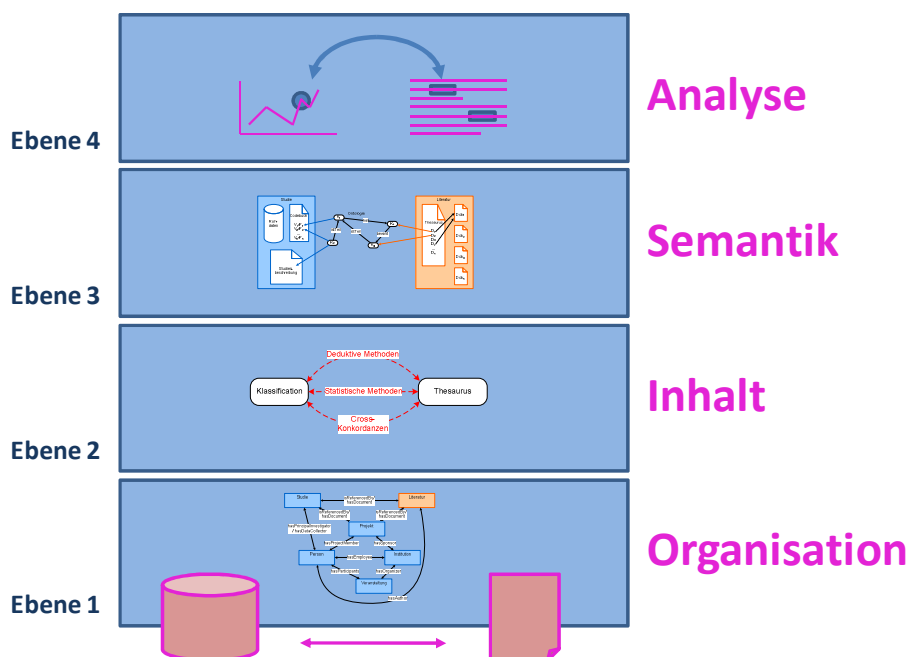


Abbildung 3: Schichtenmodell für die Text-Fakten-Integration

3.1. Ebene 1: Organisation

Die unterste Ebene (Ebene 1) dokumentiert die Datenentstehung in ihrem Kontext und organisiert die einzelnen Informationsobjekte anhand ihres Typs und ihrer semantischen Relationen zu anderen Objekten. Bei wissenschaftlichen Informationen deckt dies in der Regel den gesamten Forschungsprozess ab, von der Antragstellung über die Projektdurchführung bis hin zu den Forschungsergebnissen. Detaillierte Informationen über beteiligte Personen und Institutionen (und deren Kooperationsbeziehungen), Förderorganisationen und -programme, die Projektbeschreibung, angeschaffte oder aufgebaute Forschungsinfrastruktur und Großgeräte, produzierte Forschungsdaten, Publikationen usw. werden semantisch möglichst reichhaltig, vor allem aber vollständig und konsistent beschrieben. Ein europaweit mehr und mehr verwendeter Standard zur Beschreibung ist das CERIF-Datenmodell (Common European Research Information Format), das im Auftrag der Europäischen Kommission entwickelt wurde und von euroCRIS² weiterentwickelt wird. Auch die PolicyGrid Ontologie (vgl. Chorley et al. 2006) wurde zu dem Zweck entwickelt, die Datenerhebung bei evidenz-basierter Forschung in den Sozialwissenschaften möglichst präzise zu dokumentieren.

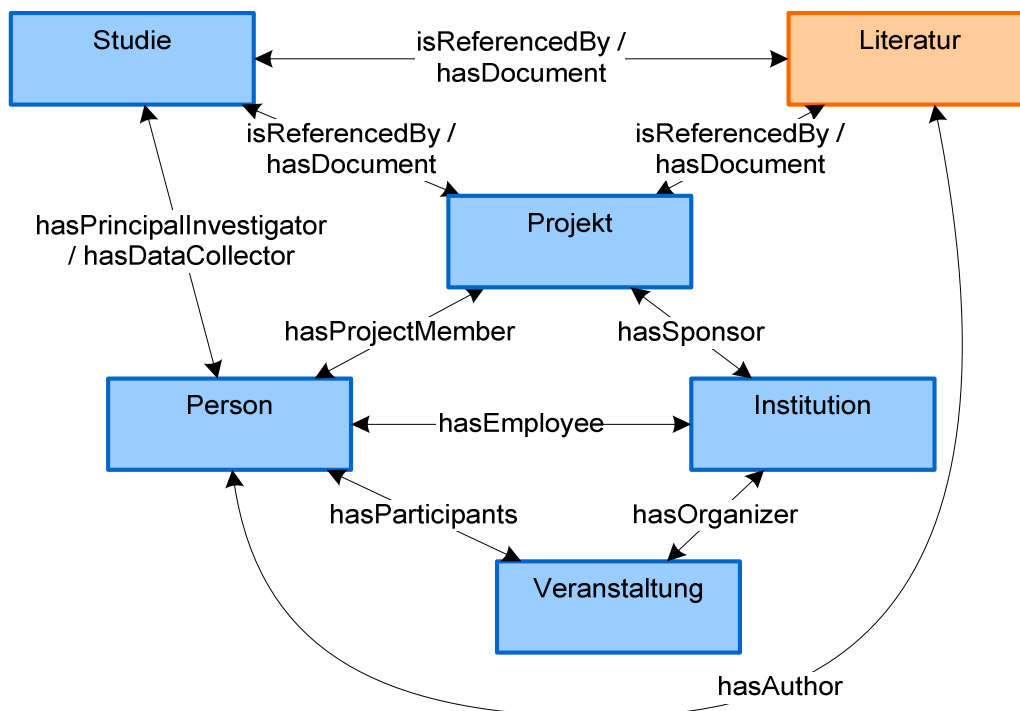


Abbildung 4: Entitäten im Forschungsprozess

Die Abbildung 4 zeigt ein Beispiel für im Forschungsprozess identifizierte Entitäten und deren semantische Beziehungen untereinander. In der Praxis sind viele der Relation oder Rollen zwischen den Entitäten mit Zeitmarken versehen, da Rollen wie die Mitarbeit in einem Projekt oder die Zugehörigkeit zu einer Institution in der Regel einen definierten Beginn und ein definiertes Ende haben - und deren Fehlen zu einer Falschinterpretation der

² www.eurocris.org

Informationen führen könnte (z. B. wenn eine Person innerhalb eines Projekt die Rolle wechselt oder die Institution verlässt).

Auf der Basis einer detaillierten Beschreibung der organisatorischen Ebene können nicht nur strukturelle und strategische Informationen gewonnen werden (Erfolg einer Einrichtung bei der Projekteinwerbung, Nachwuchsförderung, Forschungsleistung), sie kann vielmehr auch im Kontext der Informationssuche benutzt werden um Vagheit aufzulösen und Verknüpfungen zwischen Informationsobjekten herzustellen, die sich über ihre Metadaten nicht direkt referenzieren.

3.2. Ebene 2: Inhalt

Die Ebene 2 repräsentiert die traditionellen Verfahren der Inhalts- und Sacherschließung wie sie bei Literaturdatenbanken, Bibliothekskatalogen und Datenarchiven in großem Umfang zum Einsatz kommen. Vor dem Hintergrund einer integrierten Zugänglichkeit von Texten und Fakten kommt der Homogenität der Inhalterschließung hohe Bedeutung zu. In der Praxis ist dieses Desiderat aber oft nur hinsichtlich eines einzelnen Datenbestandes zu erreichen, und auch hier treten nicht selten Brüche in der Erschließung auf wenn aus inhaltlichen oder organisatorischen Gründen das zugrundeliegende Wissensorganisationssystem gewechselt werden musste (z. B. beim Umstieg von einer eigenen Schlagwortliste auf einen standardisierten (Fach-)Thesaurus).

Die Heterogenität verstärkt sich sobald mehrere Datenquellen bzgl. ihrer Erschließung vernetzt werden sollen. Dies trifft sowohl für die Integration gleicher Datentypen (z. B. mehrere Literaturdatenbanken) als auch unterschiedlicher Datentypen (z. B. Studien und Literatur) zu. Die zur Erschließung verwendeten Wissensorganisationssysteme unterscheiden sich in ihrer Fachspezifität, im Detaillierungsgrad der Modellierung des Gegenstandsbereichs, in den internen Relationen (z. B. Ober- und Unterbegriffshierarchien) und in der Sprache. Da ein Wechsel des Erschließungssystems und die damit verbundene Rückerschließung wegen des großen Ressourcen- und Zeitbedarfs kaum durchgeführt wird, müssen andere Verfahren eingesetzt werden um der Heterogenität zu begegnen.

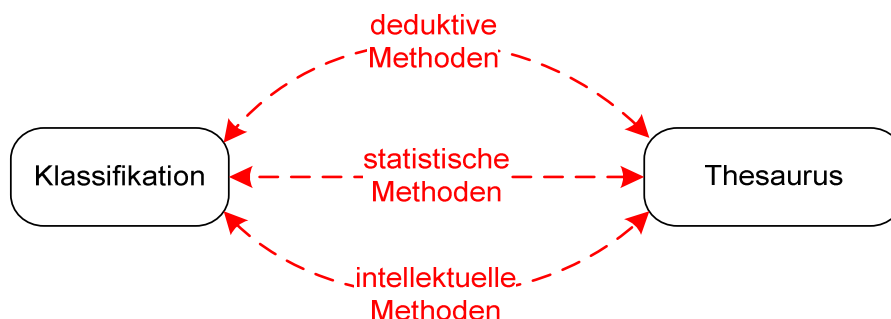


Abbildung 5: Abbildung zwischen Wissensorganisationssystemen

Abbildung 5 zeigt drei prinzipielle Möglichkeiten, unterschiedliche Erschließungssysteme aufeinander abzubilden, wobei in der Praxis Abbildungen sowohl zwischen mehreren Thesauri, mehreren Klassifikationen und zwischen Thesauri und Klassifikationen möglich

sind. Am meisten verbreitet sind intellektuelle und statistische Verfahren. Bei intellektuellen Verfahren werden entweder Paare von Erschließungssystemen in einer oder beide Richtungen aufeinander abgebildet, oder mehrere Systeme werden zu einem neuen, universalen zusammengefasst. Bei der Integration heterogen erschlossener Informationssammlungen bietet sich die paarweise Abbildung von Erschließungssystemen aufeinander an, wobei zur Vermeidung eines exponentiell wachsenden Aufwands auch nur auf einzelne, zentrale Systeme abgebildet werden kann und später deduktiv geprägte Verfahren verwendet werden können, um zu einer Abbildung zwischen nicht direkt verbundenen Systemen zu gelangen.

Bei statistischen Verfahren wird im günstigsten Fall ein Parallelkorpus verwendet, also zwei (möglichst) identische Ausgangsmengen von Dokumenten, die mit unterschiedlichen Systemen erschlossen sind. Durch Analyse der Kookkurrenzen von Schlagwörtern in beiden Systemen und auf der Basis gleicher Dokumente können statistische Abhängigkeiten zwischen Schlagwörtern und Schlagwortgruppen beider Systeme ermittelt werden, von denen dann eine hohe semantische Ähnlichkeit angenommen wird ("Bei vielen Dokumenten, bei denen Schlagwort A aus Thesaurus 1 vergeben wurde, wurden auch die Schlagwörter X und Y aus Thesaurus 2 vergeben"). Liegt kein Parallelkorpus vor, so kann dieser unter Umständen dadurch simuliert werden, dass mit einem Begriff des ersten Thesaurus in einer Datenbank gesucht wird und dann eine Kookkurrenzanalyse zwischen dem Suchbegriff und den Schlagwörtern der gefundenen Dokumente durchgeführt wird.

Durch die Abbildung der zur Inhaltserschließung verwendeten Systeme aufeinander wird die Suche über Datenbank- und auch Sprachgrenzen hinweg möglich. Dies erlaubt Nutzern ihre Anfrage mit dem ihnen am besten vertrauten Vokabular zu formulieren - das System transformiert die Anfrage dann zur Laufzeit in die Erschließungssysteme der zu durchsuchenden Informationssammlungen. Gleichzeitig wird die Suche nach verwandten Informationen über inhaltliche Aspekte ermöglicht, wenn das Informationssystem zum Beispiel ausgehend von einem als relevant eingestuften Dokument automatisch Suchanfragen nach ähnlich erschlossenen Dokumenten generiert. Dies erfolgt komplementär zu den durch Ebene 1 ermöglichten Suchstrategien, die nicht auf inhaltliche sondern organisatorische Aspekte abheben (z. B. Such nach Publikationen oder Datensätzen aus dem gleichen Projekt, unabhängig von deren Inhalt).

3.3. Ebene 3: Semantik

Die Ebene 3 (Abbildung 6) behandelt spezifische Unterschiede in der semantischen Ausdrucksstärke von Thesauri, Klassifikationen, Codebüchern etc., die immer dann relevant für die Informationssuche werden, wenn sie bei einer integrierten Recherche gemeinsam verwendet werden sollen. Sie treten zutage, wenn implizite semantische Informationen zu Forschungsdaten (z.B. die wissenschaftliche Intention hinter der Frage in einer Studie) auf die weniger ausdrucksstarken Terme eines Thesaurus abgebildet werden müssen, mit dem Dokumente indiziert sind.

Das Problem dieser unterschiedlichen Ausdrucksstärke wird erkennbar, wenn Suchbegriffe zu einer großen Trefferzahl bei Daten führen, da sie bei einer Vielzahl von Studien in Studienbeschreibung, Frageformulierung oder Variablenlabel verwendet wurden, bei der Textrecherche dieselben Suchbegriffe aber relativ selektiv sind. Oft zeigt auch erst eine genaue Analyse der Studie, ob diese relevant für das ursprüngliche Informationsbedürfnis ist, da z. B. erst von einer Frageformulierung, die den Suchbegriff enthält, auf die sozialwissenschaftliche Intention der Frage geschlossen werden muss. Das bloße Auftreten eines Suchbegriffs im Fragebogen ist keine Gewähr dafür. Und schließlich ist bei der Suche nach Texten die kognitive Belastung des Nutzers beim Bewerten der Relevanz der Ergebnisse sehr viel geringer als bei Studien, da oftmals ein schneller Blick auf Titel und Abstract genügt, wogegen bei Studien evtl. große Teile der Dokumentation zurate gezogen werden müssen.

Mittels Ontologien können solche konkreten und inhaltlich komplexeren Aspekte einer Studie modelliert werden und als Verknüpfung zwischen den Termen eines Thesaurus auf der einen Seite und den Dokumentationsstrukturen auf Studiensseite dienen. Eine wesentliche Rolle kommt dabei der Visualisierung der Ontologie für den Nutzer zu, da sie zur Interpretation des Suchergebnisses zwingend notwendig ist.

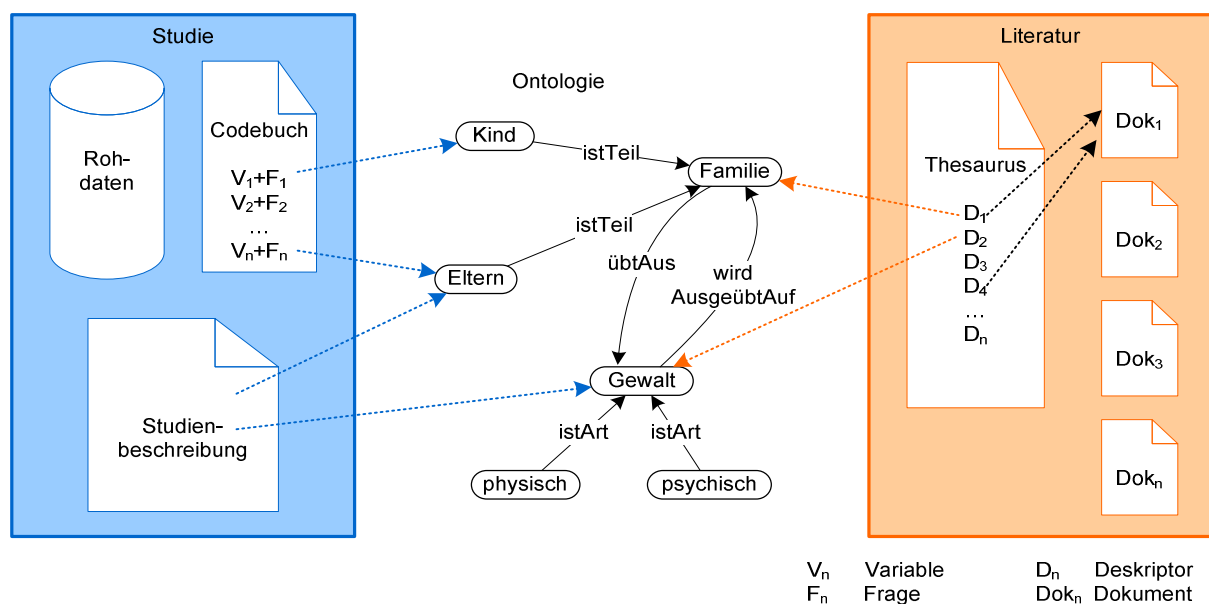


Abbildung 6: Integration auf der Ebene von domänenspezifischen Konzepten

3.4. Ebene 4: Analyse

Während die Ebenen 1 bis 3 alleine auf die Beschreibung und Erschließung von Informationen mit Metadaten und Wissensorganisationssystemen und deren Verbindung im Kontext einer integrierten Informationsarchitektur abzielen, fokussiert die Ebene 4 auf die eigentlichen Inhalte der Dokumente (hier im weitest möglichen Sinn interpretiert). Im Falle von Zeitreihen also auf die einzelnen Datenpunkte und im Falle von Textdokumenten auf den Inhalt der Volltexte. Beides wird für den Nutzer interessant sobald die gefundenen Informationen visualisiert werden – bei numerischen Daten also zum Beispiel Tabellen oder Diagramme oder bei Texten der eigentliche Volltext angezeigt wird. Beides stellt den

Ausgangspunkt für eine weitere Exploration des Informationsraums dar, wobei aber gängige Systeme zur Informationsvisualisierung und Datenexploration ein Vielfaches der Funktionalität von Textrecherchesystemen zur Verfügung stellen. Gleichwohl sind aus konzeptueller Sicht beide Richtungen – also Fakten zu Text und Text zu Fakten – gleichrangig, unterscheiden sich jedoch in Menge und Art der Operatoren.

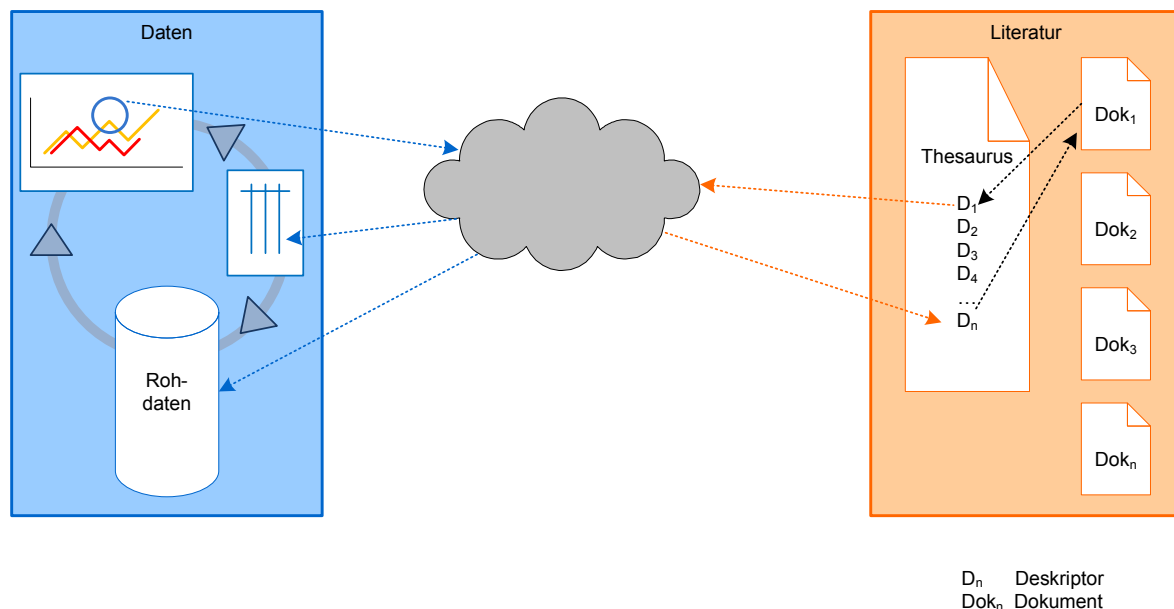


Abbildung 7: Transformation von Dateneigenschaften in Suchanfragen

Die Abbildung 7 zeigt schematisch den prototypischen Ablauf einer Datenexploration mit Text-Fakten-Integration. Im Falle eines primären Informationsbedürfnisses nach Faktendaten wird der Nutzer zunächst relevante Datensätze selektieren und auf geeignete Weise visualisieren, sei es als Tabellendarstellung oder Grafik. Bereits hier setzt ein iterativer Prozess ein, wenn zu einem Datensatz weiter recherchiert werden, die ein gefundenes Phänomen erklären oder stützen sollen (Fakten-Fakten-Integration). Die Erweiterung der Recherche kann dabei mit Hilfe der Ebenen 1 bis 3 erfolgen (Metadaten und Inhaltserschließung), es bieten sich aber auch Verfahren der Mustererkennung (Data Mining) an, um Datensätze mit ähnlichen Eigenschaften auf Werteebene zu finden. Prinzipiell stehen äquivalente Verfahren (Text Mining) auch auf der Seite textueller Daten zur Verfügung (Text-Text-Integration).

Einen informationellen Mehrwert leistet die Ebene 4 darüber hinaus, wenn es gelingt, Datenmuster wie Anstiege, Maxima oder Korrelationen so in eine textuelle Repräsentation umzusetzen, dass diese für die Textrecherche verwendet werden kann. Grundlage dafür sind sowohl die den Datensätzen zugeordneten Beschreibungsmerkmale der Ebenen 1 bis 3, mit deren Hilfe primär ähnliche Informationen gefunden werden können, darüber hinaus aber auch semantische Repräsentationen der in den Daten (oder Texten) gefundenen Muster (oder Ausreißer) selbst. Diese Transformation von Mustern zu Anfragen wird stark von der konkreten Domäne und der Richtung der Transformation bestimmt, sie wird oftmals aber auch von temporalen Aspekten determiniert, denkt man zum Beispiel an den zeitlichen

Versatz der amtlichen Statistik und der Tagespresse. In diesem Fall wäre also nicht nur eine textuelle Repräsentation eines Datenmusters (z. B. dem starken Anstieg eines Wertes von Donnerstag auf Freitag) zu generieren (z. B. „Anstieg“, „hohes Handelsvolumen“, „schloss fester“, „konnte steigern“), sondern der Suchraum für Texte auch entsprechend zu definieren (z. B. Tagespresse vom darauf folgenden Montag). Gleiches gilt prinzipiell auch für die Richtung von Texten zu Fakten (Transformation von Aussagen in Texten zu prototypischen Datenmustern, nach denen dann gesucht werden kann).

Die automatische Generierung von explorativen Anfragen bei der Text-Fakten-Integration ist – so zeigt bereits dieses einfache Beispiel – stark von explizit codiertem Domänenwissen abhängig. Die auf den Ebenen 1 bis 3 eingesetzten Wissensorganisationsstrukturen beschreiben die in einer Domäne als relevant erachteten Konzept in ihrer Allgemeinheit, können jedoch mit einfachen Mitteln keine direkte Verbindung zur Repräsentation des Konzepts in einzelnen Datensätzen oder Dokumenten herstellen. Ontologien scheinen aufgrund ihrer semantischen Ausdrucksmächtigkeit bislang am besten für den Einsatz auf Ebene 4 geeignet, jedoch existieren bislang nur wenige Beispiele, in denen umfangreiches Domänenwissen erfolgreich mit Ontologien ausgedrückt und zum Einsatz gebracht werden konnte. Zudem gestaltet sich der Entwicklungsprozess einer Ontologie – falls er auf einen Konsens in der Fach-Community abzielt – durchaus langwierig. Im Rahmen der Text-Fakten-Integration besteht jedoch die Möglichkeit, gezielt nur einzelne Aspekte der Domäne unter direkter Einbeziehung der Nutzer zu modellieren und deren Wirksamkeit bei der Datenexploration durch einen engen Zyklus von Nutzertests zu verifizieren.

4. Zusammenfassung

Ziel des Beitrags war, die enge Verbindung zwischen Informationsvisualisierung, Datenexploration und Information Retrieval aufzuzeigen und die spezifische Rolle der Text-Fakten-Integration herauszuarbeiten. Es wurde gezeigt, dass die Datenexploration – orientiert sie sich am Informationsbedürfnis des Nutzers – die Grenzen einzelner Datentypen hinter sich lassen muss. Das Information Retrieval liefert hierfür sowohl geeignete Informationsarchitekturen als auch methodische Grundlagen.

Das vorgestellte Schichtenmodell stellt eine Informationsarchitektur für die integrierte Recherche – und damit auch für die Datenexploration – in heterogenen Informationssammlungen dar. Hervorzuhebendes Merkmal der Informationsarchitektur ist, dass sie bislang getrennt voneinander betrachtete Verfahren im Bereich Wissensorganisation zusammenführt und ihre Vorteile kombiniert. Der Erfolg dieses Integrationsansatzes wird durch Retrievaltests zu belegen sein.

Literatur

Baerisch, S.; Mutschke, P.; Stempfhuber, M. (2010, erscheint): Informationstechnologische Aspekte der Heterogenitätsbehandlung in Fachportalen, in: Proceedings des 11.

Internationalen Symposiums für Informationswissenschaft (ISI 2009), Konstanz, 1.-3. April 2009.

Chorley, A.; Edwards, P.; Hielkema, F.; Philip, L.; Farrington, J. (2008): 'Developing Ontologies to Support eSocial Science: The PolicyGrid Experience', in Proceedings of the 4th International Conference on e-Social Science, Manchester, 2008. <<http://www.ncess.ac.uk/events/conference/programme/fri/3cedwards.pdf>> (abgerufen am 20.01.2010).

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

DFG (1998): Deutsche Forschungsgemeinschaft (1998): Sicherung guter wissenschaftlicher Praxis. Denkschrift.1998. <http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf> (abgerufen am 07.01.2010).

Kim, D-W.; Krause, J.; Mandl, T.; Schaefer, A.; Stempfhuber, M. (2001): Usability Design for Information Systems for the Retrieval of Texts and Numerical Data. S. 163 - 168. In: HCI International 2001: 9th International Conference on Human-Computer Interaction; Poster Sessions - Abridged Proceedings. New Orleans.

Krause, J.; Mandl, T.; Stempfhuber, M. (1997): Text-Fakten-Integration in ELVIRA. IZ Arbeitsbericht Nr. 12, Dezember 1997, 27 Seiten (Printversion vergriffen) <http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/iz_arbeitsberichte/ab12.pdf> (abgerufen am 07.01.2010).

Mayr, P.; Petras, V. (2008): Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: 74th IFLA World Library and Information Congress. Québec, Canada. <http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf> (abgerufen am 08.01.2010)

OECD (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding. <<http://www.oecd.org/dataoecd/9/61/38500813.pdf>> (abgerufen am 07.01.2010).

Poll, R. (2004): Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung, Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft. In: ZfBB 51 (2004), S. 59 - 75.

RSLG (2002): Researchers' Use of Libraries and other Information Sources: Current Patterns and Future Trends. Final Report / Education for Change Ltd.; SIRU, University of Brighton & The Research Partnership, 2002 <<http://www.rslg.ac.uk/research/libuse/>> (abgerufen am 07.01.2010).

Stempfhuber, M. (2003): Objektorientierte Dynamische Benutzungsoberflächen ODIN: Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den

Mitteln der Softwareergonomie. IZ Forschungsberichte Nr. 6. Bonn: IZ Sozialwissenschaften. 337 Seiten.

Stempfhuber, M. (2009): Die Rolle von "open access" im Rahmen des wissenschaftlichen Publizierens. S. 116 - 131. In: Alexander-von-Humboldt-Stiftung (Hrsg.): Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen. Zweite erweiterte Auflage. Bonn (Diskussionspapiere der Alexander von Humboldt-Stiftung; Nr. 12/2009) <http://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publicationsverhalten2_kompr.pdf> (abgerufen am 07.01.2010).

Stempfhuber, M.; Hellweg, H.; Schaefer, A. (2002): ELVIRA: User Friendly Retrieval of Heterogeneous Data in Market Research. S. 299 - 304. 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: TPA Publ.

Stempfhuber, M.; Schaer, P.; Shen, W. (2008): Enhancing visibility: integrating Grey Literature in the SOWIPORT Information Cycle. S. 23 - 30. In: Farace, Dominic J.; Frantzen, Jerry (2008): Ninth International Conference on Grey Literature: Grey Foundations in Information Landscape, 10 - 11 December 2007, House of the Province, Antwerp, Belgium. Amsterdam: TextRelease, GL conference series No. 9.

Walter, A-K.; Mayr, P.; Stempfhuber, M.; Ballay, A. (2006): Crosskonkordanzen als Mittel der Heterogenitätsbehandlung in Informationssystemen. S. 205 - 226. In: Stempfhuber, Maximilian (Hrsg.): In die Zukunft publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften; 11. Kongress der IuK-Initiative der Wissenschaftlichen Fachgesellschaften in Deutschland. Bonn: Informationszentrum Sozialwissenschaften (Tagungsberichte; Bd. 11)