Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren

Binder, Gisbert; Marx, Jutta; Mutschke, Peter; Riege, Udo; Strötgen, Robert; Kokkelink, Stefan; Plümer, Judith

GESIS

IZ InformationsZentrum Sozialwissenschaften

IZ-Arbeitsbericht Nr. 24

Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren

G. Binder, J. Marx, P. Mutschke, U. Riege, R. Strötgen
(Informationszentrum Sozialwissenschaften);
S. Kokkelink, J. Plümer (Universität Osnabrück)

April 2002

Abschlussbericht zum Projekt CARMEN (Content Analysis, Retrieval and MetaData: Effective Networking), Arbeitspaket 11

InformationsZentrum
Sozialwissenschaften

Lennéstraße 30
D-53113 Bonn
Tel.: 0228/2281-0
Fax.: 0228/2281-120
email: soe@bonn.iz-soz.de
Internet: http://www.gesis.org

ISSN: 1431-6943

Herausgeber: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)

Druck u. Vertrieb: Informationszentrum Sozialwissenschaften, Bonn
Printed in Germany

Inhalt

1 Zielstellung des Projekts

Es gibt klare Anzeichen dafür, dass die traditionellen Verfahren der Standardisierung und Normierung an ihre Grenzen stoßen. 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 Informationsverbünde, virtuelle Fachbibliotheken und Wissenschaftlernetze des WWW wie Mathnet - bei allen notwendigen Bemühungen – nicht von der Durchsetzbarkeit einheitlicher Standards der Inhaltsbeschreibung ausgegangen werden kann. Es wird immer Anbieter geben, die sich den vorgegebenen Standards der Mehrheit nicht unterordnen, auf deren Daten der Benutzer im Rahmen einer integrierten Recherche jedoch nicht verzichten möchte. Deshalb muss eine neue Sichtweise auf die bestehend bleibende Forderung nach Konsistenzerhaltung und Interoperabilität gefunden werden. Sie lässt sich durch die folgende Prämisse umschreiben, die in CARMEN[1] arbeitsleitend war: Standardisierung ist von der verbleibenden Heterogenität her zu denken. Erst im gemeinsamen Zusammenwirken von intellektuellen und automatischen Verfahren zur Heterogenitätsbehandlung und Standardisierung ergibt sich eine Lösungsstrategie, die den heutigen technischen, politischen und gesellschaftlichen Rahmenbedingungen gerecht wird.

Ziel des CARMEN Arbeitspaketes 11 (kurz: CAP11) ist vor diesem Hintergrund eine (semantische) Verbesserung der Ausweitung von Recherchen in verschiedenen, dezentralen Informationsbeständen (insbesondere in Fachdatenbanken und Internet-Beständen) durch Komponenten der Heterogenitätsbehandlung – überall da, wo die Homogenisierung durch Metadatenabsprachen wie die DC Initiative nur teilweise oder gar nicht erfolgreich war. Hierzu wird in drei Schritten vorgegangen:

- Das Erzeugen fehlender Metadaten aus den Dokumenten: Über deduktiv-heuristische Verfahren werden Metadaten (Titel, Autor, Institution, Keywords und Abstract) automatisch aus Dokumenten generiert. Nach einer genauen Analyse der in exemplarischen Dokumenten vorgefun-

[1] Das Projekt CARMEN wurde im Rahmen der Sonderfördermaßnahme Global-Info vom BMBF gefördert (Laufzeit des Arbeitspaketes 11: 1.12.1999-31.12.2001).

denen Heterogenität werden Heuristiken zum Auffinden fehlender Metadaten erstellt (s. Kap. 5.2.).

- Heterogenitätsbehandlung mithilfe von statistisch-quantitativen Methoden: Sie bilden unterschiedliche Verwendungen von Thesaurus- und Klassifikationstermen in den verschiedenen Beständen aufeinander ab. Für mathematische Dokumente liegen teilweise Doppelkorpora vor, eine Voraussetzung zur Anwendung dieser Verfahren. Für sozialwissenschaftliche Quellen werden sie mittels einer kommerziellen probabilistischen Volltextdatenbank simuliert. Über Wort-Kookurrenz lassen sich Transferbeziehung zwischen den einzelnen Freitext-Termen und den Deskriptoren eines Sacherschließungssystems wie dem IZ-Thesaurus oder der Schlagwortnormdatei ableiten (s. Kap. 5.3).
- Einbau der in CARMEN Arbeitspaket 12 (kurz: CAP12[2]) intellektuell entwickelten Crosskonkordanzen als dritte Methode.

Voraussetzung für die Bearbeitung der genannten Aufgabenstellung ist das Vorliegen eines Testkorpus, an dem die Heterogenitätsphänomene analysiert, die Module zu ihrer Behandlung erstellt und erprobt werden können. Zur Überprüfung der Übertragbarkeit der Ergebnisse wird dem Korpus aus der Domäne Mathematik ein sozialwissenschaftlicher Korpus gegenübergestellt.

2 Erstellung von Test-Datenbanken

2.1 Sozialwissenschaften

Aus dem für CAP11 erstellten Gesamtkorpus Sozialwissenschaften (12.631 Dateien) wurde ein Beispielkorpus mit 3661 Internetquellen aufgebaut. Ein quantitativer Überblick über den Gesamtkorpus (Anzahl der Dateien/Dateiformate pro Themenbereich) findet sich in Tabelle 1. Inhaltlich wurde der Korpus auf die Themenbereiche des GIRT-Scopes, d.h. auf die Themen „Frauenforschung“, „Migration“ und „Industrie- und Betriebssoziologie“ eingeschränkt. Startpunkt für den Aufbau des Testkorpus war das Clearinghouse *SocioGuideWest* der Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS), welches eine Sammlung von ausgewählten Internet-Adressen auf dem Gebiet der Sozialwissenschaften ist (http://www.gesis.org/SocioGuide/).

---

[2] Weiterführende Infos zu CAP12 s. http://www.gesis.org/Forschung/Informationstechnologie/CARMEN-AP12.htm

Der CARMEN-Testkorpus beschränkt sich auf Internetquellen, die eine inhaltliche Beschreibung einer wissenschaftlichen Aktivität darstellen. Um die institutionelle Kontextinformation zu erhalten, wurde die jeweilige Homepage des Instituts, von der aus alle ausgewählten URLs zu einer Quelle erreicht werden können, mit erfasst. Bezüglich Dokumenttyp und Dateiformat wurden keine Einschränkungen festgelegt, d.h. der Korpus enthält sowohl Projektbeschreibungen und Literatur(nachweise), aber auch Beschreibungen der Arbeitsschwerpunkte sozialwissenschaftlicher Institutionen, kommentierte Vorlesungsverzeichnisse, Seminararbeiten, Lizentiate, Themenschwerpunkte von Konferenzen, Publikationslisten, verschiedene Berichtstypen (Erfahrungs-, Forschungs-, Konferenz-, Länder-, Tätigkeitsberichte etc.), Diskussionslisten usw., sofern sie in einem projektbezogenen oder institutionellen Kontext stehen. Zeitschriften/Newsletter sind mit unterschiedlichem Informationsumfang vertreten, d.h. mit Inhaltsverzeichnissen, Beitragsabstract und/oder Volltexten etc. Vereinzelt finden sich auch komplette Bücher und virtuelle Ausstellungsprojekte. Allein die Vielfalt der Ressourcentypen weist bereits eine sehr hohe internetspezifische Heterogenität auf. Es sind überwiegend Dokumente im HTML-Format im Korpus vertreten, aber auch Dokumente im TXT-, RTF-, PDF-, PS- und Word-Format. Hier fällt besonders auf, dass HTML-Dokumente häufiger Teildokumente einer umfangreicheren Informationseinheit sind und aus inhaltlicher Perspektive hinsichtlich darin enthaltenen Textes häufig nur in Verbindung mit weiteren HTML Dokumente gesehen werden können. Dieser Aspekt ist für die Qualität entsprechender Metadaten wie z.B. Titel höchst relevant, da viele Seitentitel als einzelne keine vollständige Aussagekraft haben, wohl aber im gesamten „site-kontext“ eines Informationsanbieters. Dagegen stehen TXT-, RTF-, PDF-, PS- und Word-Dokumente zwar im thematischen Kontext, stellen in sich in der Regel aber geschlossene Texteinheiten dar. Die Internetdokumente wurden im Dateisystem und in einer Oracle-Datenbank abgespeichert.

Gesamtkorpus	.htm /.html	.doc	.txt	.rtf	.ps	.pdf	.zip	.ps. gz	So	**Zwischen-summe**	img	∑
Frauenforschung	*2111*	*36*	*20*	*0*	*0*	*10*	*6*	*0*	*23*	**2206**	*731*	**2937**
Industrie- und Betriebssoziologie	*1911*	*89*	*0*	*3*	*8*	*248*	*22*	*7*	*31*	**2319**	*987*	**3306**
Migration	*3887*	*114*	*240*	*13*	*0*	*314*	*1*	*0*	*76*	**4645**	*1743*	**6388**
Gesamt	**7909**	**239**	**260**	**16**	**8**	**572**	**29**	**7**	**130**	**9170**	**3461**	**12631**

Tabelle 1: Quantitativer Überblick (Gesamtkorpus)

2.2 Mathematik

Aus dem Bereich **Mathematik** standen verschiedenartige Testdaten zur Verfügung. Der Preprintdienst MPRESS[3] enthält etwa 40.000 Dokumente. Inhaltlich ist dieser Datenpool homogen, da es sich bei allen Dokumenten um mathematische Preprints bzw. um deren Abstracts handelt. Formal sind die Daten aber sehr heterogen, die Originaltexte liegen in der Regel im PostScript-Format vor. Die Abstracts sind in HTML-Files gespeichert und enthalten Dublin Core Metadaten. Bei einigen Dokumenten liegen sowohl Originaltext, als auch ein Abstract mit DC Metadaten vor.

Als weitere Datenpools standen Daten zur Verfügung, die aus mathematischen Online Journalen von einem Harvester gesammelt wurden. Als dritter und sicherlich heterogenster Datenpool boten sich die Internetseiten von mathematischen WWW-Servern in Niedersachsen und Bremen an.

Diese drei Bestände werden durch Sammelagenten der Harvest Software generiert. Die einzelnen Datensätze liegen daher im flachen SOIF (Summary Object Interchange Format) vor. Die SOIF enthalten auch die URLs der Originaldokumente; diese können zur Analyse dann jeweils lokal gespeichert werden.

Komplementär zur Domäne Sozialwissenschaften, die sich auf die Metadatenextraktion aus HTML-Files konzentrierte (s.u.), wurde für die Mathematik ein Testkorpus erstellt, der aus Dateien im Postscript Format besteht. Dazu wurden aus dem Preprintdienst MPRESS zufällig 400 Datensätze ausgewählt, die der Index aus Postscript Dateien generiert hat.

Zur technischen Erstellung des Testkorpus wurde mit Hilfe eines Perl Scripts aus den SOIF dieser Dokumente die URLs extrahiert. Das Script holt dann über das http-Protokoll die Originaldokumente und speichert sie im Filesystem ab.

Vergleichend hierzu wurde eine Datenanalyse im Bereich Physik durchgeführt (50 Dokumente aus PhysNet). Diese ergab, dass sich die Struktur der Daten im Bereich Physik nicht wesentlich von der in der Mathematik unterscheidet. Aus diesem Grund wurde zunächst auf Testdaten aus der Physik verzichtet.

---

[3] Weitere Informationen unter http://MathNet.preprints.org

3 Planung und Ablauf des Vorhabens

Die Behandlung der Aufgabenstellung erfolgte in 4 Schritten:

- Analyse der Heterogenität
- Metadatenextraktion
- Behandlung der verbleibenden Heterogenität mittels Transfermodule
- Test der Module

Der Antrag sah ursprünglich eine parallele Bearbeitung von Metadatenextraktion und Transferkomponenten vor, wobei im ersten Jahr einfache und im zweiten Jahr komplexe Module erstellt werden sollten. Demgegenüber erwies sich bei der Metadatengenerierung ein stärker iteratives Vorgehen als vorteilhaft, weshalb durch das Vorziehen eines Unterarbeitspaketes (UAP 11) in einem fließenden Übergang von einfachen zu komplexen Extraktionsregeln das Paket Metadatengenerierung in einem Schritt abgearbeitet wurde.

4 Entwicklung der relevanten wissenschaftlichen und technischen Bereiche während der Projektlaufzeit

Für die Zielsetzung von CAP11 waren insgesamt drei Forschungsbereiche zu beobachten und nach verwertbaren Lösungsansätzen zu analysieren:

a) Behandlung semantischer Heterogenität mittels quantitativ-statistischer Verfahren (probabilistische Verfahren und neuronale Netze)

b) automatische Generierung von Metadaten

c) Thesaurus-Management-Systeme zur Verwaltung von intellektuell erstellten Crosskonkordanzen

Auf dem Gebiet der neuronalen Verfahren wurde mit der Universität Hildesheim kooperiert, Als Ergebnis entstand ein umfassender state of the art Bericht als Ausgangspunkt für weitere Integrationsüberlegungen (cf. Mandl

2000[4]). Ein weiterer Arbeitsbericht befasst sich konkret mit der praktische Umsetzung neuronaler Ansätze für CARMEN (cf. Krause 2000[5]).

Im Bereich probabilistischer Verfahren zur Heterogenitätsbehandlung konnte auf die Erfahrungen der Projekte ELVIRA (cf. Krause/Stempfhuber 2000[6]) und ViBSoz (cf. Müller/Meier/Winkler 2000[7]) zurückgegriffen werden. Aus ELVIRA, das den Transfer zwischen Zeitreihendaten und textueller Information für ein Verbandsinformationssystem untersucht, wurde ein Tool zur Aufbereitung eines Datenbestandes für die statistische Analyse als Ausgangspunkt verwendet und an die CARMEN-Spezifika angepasst (cf. Kap. 5.3.1).

Von dritter Seite konnten keine weiteren Forschungsaktivitäten auf diesem Gebiet beobachtet werden.

Bei der Diskussion um die Verwendung von Metadaten im WWW, ist zu beobachten, dass deren Notwendigkeit auf allgemeinen Konsens stößt. Die Anstrengungen konzentrieren sich jedoch die intellektuelle Vergabe bzw. auf Unterstützungsverfahren und Kontrollmechanismen. Daneben sind Umfang und Scope der Metatags als Erweiterung von Dublin Core ein oft diskutiertes Thema (cf. z.B. Borbinha/Baker 2000[8]). Dass diese Normierungsbemühungen nicht ausreichen, ein befriedigendes Rechercherergebnis zu gewährleisten,

---

[4] Mandl, Thomas (2000): Einsatz neuronaler Netze als Transferkomponenten beim Retrieval in heterogenen Dokumentenbeständen. (IZ-Arbeitsbericht Nr. 20, November 2000). Bonn.

[5] Krause, Jürgen (2000): Integration von Ansätzen neuronaler Netzwerke in die Systemarchitekturen von ViBSoz und CARMEN. (IZ-Arbeitsbericht Nr. 21. Dezember 2000). Bonn.

[6] Krause, Jürgen; Stempfhuber, Maximilian (erscheint): Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA (IZ-Forschungsberichte 4). Bonn.

[7] Müller, Matthias; Meier, Wolfgang; Winkler, Stefan (2000): Die virtuelle Fachbibliothek Sozialwissenschaften. In: Knorz, Gerhard; Kuhlen, Rainer (eds.): Informationskompetenz - Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationalen Symposiums für Informationswissenschaft. Schriften zur Informationswissenschaft Bd. 38. Konstanz.

[8] Borbinha, José; Baker, Thomas (eds.) (2000): Research and Advanced Technology for Digital Libraries. 4th European Conference, ECDL 2000, Lisbon, Portugal, September 2000. Proceedings. Berlin, Heidelberg, New York.

macht allein schon der aktuelle Zustand der Internetquellen für den Bereich Sozialwissenschaften deutlich (cf. Kap. 5.1.1).

Zur Verwaltung der im Rahmen von CAP 12 erstellten Crosskondordanzen diverser Fach- und Allgemeinthesauri sowie Klassifikationen wurde der Markt auf Softwaresysteme hin gesichtet. Die Suche gestaltete sich recht schwierig, da fast alle Systeme ausschließlich für die Verwaltung eines Thesaurus ausgerichtet sind und keine Interthesaurusbeziehungen vorsehen. Die Wahl fiel schließlich auf das System SIS-TMS von ICS-FORTH[9], da es eine umfangreiche Funktionalität zur Verwaltung, Speicherung und Anzeige von Termrelationen bereithält und zudem eine Konfiguration des zugrundeliegenden Modells zulässt, so dass es auf die Bedürfnisse von CARMEN angepasst werden konnte (siehe Kap. 5.4.1.3).

5 Erzielte Ergebnisse

5.1 Analyse der vorhandenen Heterogenität

5.1.1 Sozialwissenschaften

Die Analyse der Heterogenität sozialwissenschaftlicher Internetdokumente stützte sich auf den oben beschriebenen Testkorpus (Dokumente im HTLM-Format als häufigster Formattyp). Nach einer ersten Durchsicht bezüglich des Vorkommens von Metadaten konnten die Parameter „Dokumententitel“, „Autor“, „Institution“, „Schlüsselwörter (keywords)“ und „Abstract“ als grundsätzlich extrahierbar identifiziert werden

Hier zeigte sich, dass die Unterschiede der Codierung von Metadaten innerhalb einer WWW-Site einer Institution oder eines Autors relativ einheitlich verwendet werden, während vor allem zwischen den verschiedenen Sites große Unterschiede bestehen. Für einen Pretest wurde daher eine Zufallsstichprobe von 100 Dokumenten, bestehend aus einer oder mehreren HTML-Dateien pro Site, gebildet. Für diese wurden die Extraktionsmöglichkeiten von Metadaten aus HTML-Quelltext näher analysiert, wobei folgendes festzustellen war:

---

[9] Institute of Computer Science, Foundation for Research and Technology - Hellas (ICS-FORTH), http://www.ics.forth.gr/

- Nur ein geringer Prozentsatz der untersuchten sozialwissenschaftlichen Internetquellen weisen Meta-Tags auf. Es werden ausschließlich die Tags „Author“, „Description“ und „Keywords“ gefunden, Meta-Tags nach Dublin-Core wurden zum Testzeitpunkt in fast keinen Dokumenten verwendet. Für die bei der Generierung behandelten Tags („Title“, „Keywords“, „Description“) ergibt sich ein Mittelwert von ca. 15%.
- Vielfach werden nicht inhaltsbeschreibende HTML-Tags (wie <ADDRESS>, <Hx> o.Ä.), sondern formatierende (wie <FONT SIZE=X>, <B> o.ä.) verwendet, die eine inhaltliche Analyse wesentlich erschweren und von Site zu Site ausgesprochen unterschiedlich verwendet werden.
- In den Internet-Dokumenten kann nicht durchgehend korrektes, fehlerfreies HTML erwartet werden. So wurden beispielsweise <TITLE>-Tags im <BODY> statt im <HEAD> vorgefunden, wo sie formal falsch und daher für HTML-Parser möglicherweise nicht identifizierbar sind.
- HTML-<META>-Tags wurden – wo überhaupt – nicht immer korrekt verwendet. Institutionen z.B. wurden häufig im Meta-Tag „Description“ vorgefunden.
- Kontextinformationen wie Titel, Autor oder Datum fehlen in einer Vielzahl von Dokumenten völlig.

Für den Dokumententitel beispielsweise wurden in 88 Dokumenten folgende Auszeichnungen vorgefunden (Mehrfachauszeichnungen möglich):

Tag	TITLE	H1	H2	H3	H4	H5	EM	STRONG	eingebettete Abbildung	sonstiges
Häufigkeit	72	5	6	8	1	1	1	1	2	13

Tabelle 2: Pretest Dokumenttitel

13 Dokumente wiesen keinerlei Dokumenttitel auf. Die sonstigen Auszeichnungen waren vor allem formatierender Art z.B. Zeichengröße, Schriftart oder Zentrierung.

Der Autor als weiteres Beispiel wurde durch folgende Auszeichnungen markiert:

Tag	META „Author“	TITLE	EM	STRONG	sonstiges
Häufigkeit	4	4	1	1	28

Tabelle 3: Prestest Autor

56 Dokumente sind ohne erkennbaren Autor. Die sonstigen Auszeichnungen sind hier neben formatierenden Tags auch E-Mail-Links.

Zusammenfassend kann festgehalten werden, dass nach der ersten Analyse der Heterogenität die Möglichkeiten und Schwierigkeiten der Extraktion von Metadaten identifiziert, die zu extrahierenden Metatags festgelegt wurden und Material für die Entwicklung von Heuristiken für diese Extraktion vorbereitet wurde. Daran schloß sich die Implementierung der konkreten Heuristiken sowie ein detaillierter Test der Extraktionsergebnisse an.

5.1.2 Mathematik

Im Bereich Mathematik wurden aus dem Preprintindex MPRESS 400 PostScript Dokumente zufällig ausgewählt. Da mathematische Originalarbeiten im Netz zum überwiegenden Teil im Postscript Format vorliegen, stellt dies eine vertretbare Einschränkung auf eine Dokumentart dar. Die Dokumente lassen sich bezüglich der Qualität der Erschließung in folgende zwei Klassen einteilen:

- Postscript Dokumente mit vorhandenen DC Metadaten (im zusätzlichen Abstract File im HTML Format)
- Postscript Dokumente ohne Metadaten

MPRESS ermöglicht bisher nur eine strukturierte Suche (z.B. nach Autor, Titel oder Klassifikation) für Dokumente, die mit Metadaten versehen sind. Um eine hochwertige Erschließung des gesamten Datenbestandes zu ermöglichen, wurden in CAP11 Verfahren zur Extraktion von Metadaten aus Postscript-Dokumenten realisiert.

5.2 Metadaten-Extraktion

5.2.1 Heuristiken

5.2.1.1 Sozialwissenschaften

Erste Analysen des Testkorpus ergaben, dass eine sehr große Anzahl von Dokumenten über einen spezifizierten Dokumenttitel (title-tag) verfügten. Aus diesem Grunde konnte der Versuch unternommen werden, durch eine detaillierte Heuristik von vorneherein auch qualitative Unterschiede in der Relevanz des spezifizierten Titels für den Dokumentinhalt zu ermitteln. Für die verwendeten Schlüsselwörter (keywords-tag) und Inhaltsangaben (description-tag) der Dokumente sind die Heuristiken vergleichsweise einfach. Die Extraktion von Begriffen aus dem Text und ihre Identifikation als Schlüsselbegriffe scheint zumindest in den Dokumenten der Domäne Sozialwissenschaften nahezu unmöglich. Bei inhaltsbeschreibenden Informationen (Abstract, Zusammenfassung) wurde zumindest der Versuch unternommen, solche Textpassagen im Gesamtdokument zu finden.

Nach der stichprobenartigen Analyse konnten erste Entwürfe für Heuristiken formuliert und implementiert werden. In einem iterativen Prozess wurden die Heuristiken über den Testkorpus getestet, die Ergebnisse überprüft und daraufhin durch eine Reformulierung der Heuristiken Fehler korrigiert und neue Extraktionsmöglichkeiten hinzugefügt.

Bei der Überprüfung der Testläufe wurden die extrahierten Metadaten mit den im Dokument vorhandenen Daten verglichen. Zu unterscheiden waren hier falsche oder ungültige Metadaten, die extrahiert wurden und eine Korrektur der Heuristik erforderten, von richtigen, gewünschten Metadaten, die im Dokument vorhanden, aber nicht extrahiert wurden und daher eine Erweiterung der Heuristik nötig machten.

Folgende Tafeln zeigen die Endfassung der Heuristiken, die auch der qualitativen Bewertung der Extraktionsergebnisse zugrunde lag. Die vermuteten Wahrscheinlichkeiten für die extrahierten Daten wurden nach einer Evaluation korrigiert (siehe Kapitel 5.2.4.1.2).

5.2.1.1.1 Titel-Heuristik

```
If (<title> vorhanden && <title> enthält nicht „untitled“ && HMAX
    vorhanden){
    /* 'enthält nicht „untitled“' wird case insensitive im <titel>
       als Substring gesucht */
    If (<title>==HMAX) {
      <1> Titel[1,0]=<title>
    } elsif (<title> enthält HMAX) {
      /* 'enthält' meint hier immer case insensitive als Substring */
      <2> Titel[0,8]=<title>
    } elsif (HMAX enthält <title>) {
      <3> Titel[0,8]=HMAX
    } else {
      <4> Titel[0,8]=<title> + HMAX
    }
  } elsif (<title> vorhanden && <title> enthält nicht „untitled“ &&
     S vorhanden) {
     /* d.h. <title> vorhanden UND es existiert ein Eintrag S mit
      //p/b, //i/p usw. */
      <5> Titel[0,5]=<title> + S
  } elsif (<title> vorhanden) {
      <6> Titel[0,5]=<title>
  } elsif (<Hx> vorhanden) {
      <7> Titel[0,3]=HMAX
  } elsif (S vorhanden)
  {
      <8> Titel[0,1]= S
    }
  }
```

Abbildung 1: Titel-Heuristik

Dieser Algorithmus leistet folgendes:

Die Methoden <1> bis <6> untersuchen Dokumente, die über einen qualifizierten Seitentitel verfügen, d.h. der Seitentitel des Dokumentes existiert und ist nicht mit „untitled“, „neue Seite x“ oder ähnlichem belegt.

Im ersten Schritt wird nun die gekennzeichnete Überschrift höchster Ordnung ermittelt (header-tag) und mit dem Seitentitel verglichen. Je nachdem, ob Überschrift und Seitentitel identisch sind, einer ein Teil des anderen ist oder beide unterschiedlich sind, werden die Metadaten für den Titel bestimmt (Methoden <1> bis <4>) und die vermutete Relevanz zugewiesen. In einem zweiten Schritt werden die Dokumente bearbeitet, die zwar über einen Seitentitel verfügen, aber keine Überschriften haben. Stattdessen werden Textteile gesucht, die von ihrer Formatierung als Überschriften interpretiert werden können, z.B. fett oder kursiv geschriebene Texte oder Tabellenüberschriften. Falls solch ein Textteil gefunden werden kann, wird dieser dem Seitentitel hinzugefügt (Methode <5>), falls nicht, wird ausschließlich der Seitentitel als extrahiertes Metadatum verwendet (Methode <6>).

In den Methoden <7> und <8> wird versucht, aus den (wenigen) Dokumenten, die nicht über eine Seitentitel verfügen, Metadaten zu extrahieren. Falls diese Dokumente Überschriften haben (header-tag) wird die höchster Ordnung als Titelinformation verwendet (Methode <7>). Falls nicht, wird nach Textteilen gesucht, die überschriftähnlich formatiert sind (Methode <8>).

5.2.1.1.2 Keyword-Heuristik

```
If (<meta name="keywords"> vorhanden) {
  <1> Keywords[1,0]= <meta name="keywords">
}
```

Abbildung 2: Keyword-Heuristik

Gesucht wird mit dieser Heuristik nach dem verwendeten keywords-tag. Die dort vergebenen Begriffe werden als die vom Autor bestimmten Schlüsselwörter interpretiert. Der Versuch, Schlüsselbegriffe aus dem Seitentext zu extrahieren, erscheint zumindest für die der Domäne Sozialwissenschaft eigene Art zu publizieren, relativ aussichtslos.

5.2.1.1.3 Abstract - Heuristik

```
if Knoten = {//meta/@name="description"} nicht leer
 String S = join ("\n", getContent(Knoten))
 (Method 1) return (0.8 , S)
elsif Knoten = {//p oder //li} nicht leer
Suche in Knoten nach dem ersten Element k
  mit getText(k) contains "Zusammenfassung" oder "Summary"
  oder "Abstrakt" oder "Abstract". Uberspringe Knoten,
  in dem der Treffer zwischen <A HREF="xxx"> </A> steht.
if treffer:
 Fasse alle Folgeelemente k_1, ,..., k_n zu
 einem String S zusammen fuer die gilt: getText(k_i)>99 Zeichen.
 (Method 2) return (0.5 , S)
else:
 if Knoten = {//h1 oder //h2 oder ... //h6 } nicht leer
 String S = join ("\n", getContent(Knoten))
 (Method 3) return (0.1 , S)
```

Abbildung 3: Abstract-Heuristik

Methode <1> übernimmt die Einträge des description-tag als Metadaten, falls dieser nicht leer ist. In Methode <2> wird versucht, aus dem Dokumenttext Passagen zu identifizieren, die als Inhaltsangabe verwendet werden können. Extrahiert werden dazu die Absätze, die auf die Schlüsselbegriffe „Zusammenfassung“, „summary“ und „abstract“ folgen. Tests ergaben, dass die Begriffe „Inhalt...“ oder „content“ ungeeignet waren. Die Methode <3> „addiert“ sämtliche Überschriften (header-tag) des Dokumentes in der Vermutung, so inhaltsbeschreibende Aussagen zu konstruieren.

5.2.1.1.4 Voruntersuchungen zur Autoren- und Institutionen- Heuristik

```
1. AUTOREN = {//meta/@name="author"}
2. if AUTOREN nicht leer:
 2.1 foreach autor k in AUTOREN:
  2.1.1 if k in Personenliste return (1, getContent(k))
  2.1.2 ABSAETZE = {//p}. Suche nach einem Absatz P
        aus ABSAETZE, in dem getContent(k) und eines
        der Wörter {"Webmaster", "Webmistress",
        "Webeditor", "Web Editor"} vorkommt.
  2.1.3 if P gefunden: return (0.2 , getContent(k))
  2.1.4 else
   2.1.4.1 if getContent(k) entspricht Namenspattern*:
           return (0.5, getContent(k))
3. ADDRESSES = {//address} nicht leer:
 3.1 foreach addr in ADDRESSES:
  3.1.1 if getContent(addr) not contains {"Webmaster", "Webmistress",
                "Webeditor", "Web Editor"}:
   3.1.1.1 if getContent(addr) in Personenliste:
                   return (1, getContent(addr))
   3.1.1.2 elsif getContent(addr) entspricht Namenspattern*:
                   return (0.5, getContent(addr))
   3.1.1.3 else return (0.2 , getContent(addr))
4. TITEL = {//title} nicht leer:
 4.1 foreach title in TITEL:
  4.1.1 if title in Personenliste: return (0.2, getContent(title))
5. X = {//strong/p ...} (vergleiche Titel-Heuristik) nicht leer:
 5.1 foreach x in X:
  5.1.1 if x in Personenliste: return (0.1, getContent(x))

Namenspattern: Bis zu vier Wörter, alle in Großschreibung; eines darf
klein geschrieben sein, wenn in {"von", "zu"};
ein Komma oder Semikolon darf zwischen den Wörtern stehen;
Wortzahl-1 dürfen abgekürzt sein.
```

Abbildung 4: Autoren-Heuristik

Dieser Entwurf einer Autoren-Heuristik nutzt ein einfaches Namenspattern und die Autorenliste des IZ. Sie wurde aus Zeitgründen nicht mehr vollständig implementiert (siehe Kapitel 5.2.4.1).

Entsprechendes gilt für die Institutionen-Heuristik. Sie ist ähnlich der Autoren-Heuristik aufgebaut und nutzt die Institutionen-Liste des IZ.

5.2.1.2 Mathematik

Für die PostScript-Dokumente aus dem Bereich Mathematik wurden folgende Heuristiken entwickelt:

- *Abstract*. Suche in den ersten drei Seiten nach Paragraphen, die mit den Wörtern 'abstract', 'summary', oder 'zusammenfassung' beginnen (case insensitive). Überprüfe, ob die Anzahl der Wörter des gefundenen Para-

graphen größer 10 ist. Falls die Anzahl der Wörter kleiner 10 ist, iteriere diesen Prozess so lange mit den folgenden Paragraphen, bis diese Bedingung erfüllt ist.
Erkenne den Paragraphen als Anfang eines Abstracts. Überprüfe, ob der Paragraph mit einem Punkt endet oder die Anzahl der Zeilen des Paragraphen größer 1 ist. Falls nicht, addiere solange die folgenden Paragraphen zu dem gefundenen Abstract-Anfang, bis diese Bedingung erfüllt ist.
Lösche evtl. eines der Wörter 'abstract', 'summary', oder 'zusammenfassung'.

- *Keywords*. Suche in den ersten drei Seiten nach Zeilen, die eines der Wörter 'key' und 'words' oder 'schlüsselwörter' enthält (case insensitive). Lösche den Text der gefundenen Zeile bis zu dem ersten Vorkommen von "." oder ":" und erkenne diese Zeile als Anfang von Keywords.
  Überprüfe, ob die Zeile mit einem Punkt endet oder letzte Zeile des Paragraphen war. Falls nicht, addiere solange die folgenden Zeilen, bis diese Bedingung erfüllt ist.
  Falls diese Strategie keinen Treffer liefert, gehe wie folgt vor: Addiere die ersten drei Seiten in ein Textfragment und suche in diesem Textfragment nach einer Zeile (im Sinne von Text, nicht von Prescript!), die mit dem Wort 'Key' beginnt. Addiere zu dieser Zeile die nächsten fünf Zeilen in ein neues Textfragment. Lösche in diesem Textfragment alles, was nach dem ersten Auftreten eines "
", eines <P> oder eines "." kommt. Erkenne den Rest als Keywords.
- *MSC Klassifikation*. Suche in den ersten drei Seiten nach Paragraphen, die mindestens zwei der Wörter "ams","msc","classification","subject" oder "mathematics" enthalten. Suche in diesem Paragraphen nach Wörtern der Form (D)DCE(E) und erkenne diese als MSC Klassifikationen. (D=Ziffer, C=Buchstabe oder "-", E=Ziffer oder "x", ()=ein- oder keinmal).
- *References*. Suche in den letzten vier Seiten des Dokumentes nach einem Paragraphen, der eines der Wörter (bibliography, references, Literatur) enthält und dessen Länge kleiner 20 ist. Erkenne den nächsten Paragraphen oder die nächste Zeile als potentiellen Anfang einer Referenz-Sequenz. Iteriere solange über die folgenden Paragraphen und Zeilen, bis eine Zeile mit [oder \d+\s oder \d+\. beginnt. Erkenne diese Zeile als Anfang einer Referenz. Das Ende der aktuellen Referenz ist durch den Anfang der nächsten definiert. Die letzte erkannte Referenz kann evtl. zu viel Daten enthalten. Falls die Anzahl der Zeilen in der letzten erkannten Referenz >=2 ist, lösche in dieser Referenz alles bis auf die erste Zeile.

5.2.2 Implementierung

5.2.2.1 Sozialwissenschaften/HTML

Grundvoraussetzung für die Extraktion von Informationen aus HTML Dokumenten ist die Fähigkeit, die Struktur von HTML Dokumenten zu analysieren. Die Analyse der Testkorpora zeigte, dass dieser Voraussetzung zwei Tatbestände entgegenstehen:

- Die meisten HTML Dokumente im WWW sind nicht spezifikationskonform.
- Es existiert momentan keine standardisierte Anfragesprache für HTML.

Im Umfeld von XML sieht die Situation wesentlich besser aus. Die Korrektheitsbegriffe *well-formed* und *valid* sind ein essentieller Bestandteil von XML. Es ist unwahrscheinlich, dass diese Begriffe von XML Applikationen so ignoriert werden, wie dies zum Beispiel bei Browsern bezüglich HTML der Fall ist. Weiterhin existieren eine Fülle standardisierter Technologie im Umfeld von XML, unter anderen:

- *DOM*: Das Document Object Model ist ein Application Programming Interface für XML und ermöglicht die Navigation in und die Modifikation von XML Dokumenten.
- *XPath* : Anfragesprache für XML Dokumente.

Aufgrund der Verwandtschaft von HTML und XML lag es nahe, vorhandene XML Technologien für die Analyse von HTML Dokumenten zu nutzen. Dazu waren im wesentlichne nur die (häufig nicht spezifikationskonformen) HTML Dokumente in well-formed XML Dokumente zu konvertieren. Als Ausgangsformat sollten bei diesem Prozess diejenigen HTML Dokumente zulässig sein, die von den gängigen HTML Browsern noch dargestellt werden.

Da die hohe Anzahl der notwendigen Iterationen (Tests -> Redesign -> Implementierung -> Tests ->...) zur Verbesserung der Heuristiken eine kurze Implementierungszeit voraussetzt, wurde der Entschluss gefasst, die bis dahin entwickelten Heuristiken unter Ausnutzung standardisierter Technologien neu zu implementieren. Um die Einsetzbarkeit der Software in die Retrieval-Komponente aus CAP7[10] zu gewährleisten, wurde in enger Zusammenarbeit

[10] Infos zu CARMEN AP7 siehe http://ls6-www.cs.uni-dortmund.de/ir/projects/carmen/wp7.html.de

mit CAP7 ein Softwarepaket gesucht, welches die oben beschriebenen XML Standards DOM und XPath implementiert. Die Wahl fiel schließlich auf das Perl Modul LibXML, das neben den geforderten Implementierungen von DOM und XPath zusätzlich einen HTML-Parser enthält, der HTML Dokumente direkt in das DOM Modell abbilden kann. Die folgende Abbildung gibt einige Beispiele, wie dieses Modul HTML Fragmente auf eine baumartige XML Struktur (DOM) abbildet.

HTML Fragment	XML Dokument
<html> <head><title>Der Titel</title></head> <body><P><B>Text</B></body> </html>	<html> <head><title>Der Titel</title></head> <body> Text </body> </html>
LEERE DATEI	<html><head/> <body/> </html>
123456	<html><head/> <body>123456 </body></html>
<h1 my="2" my="3"> 123456	<html><head/><body> <h1 my="3"> 123456</h1> </body></html>
Titel	<html><head/><body> Titel </body></html>
<P ohne Ende <P> hier wieder alles ok	<html><head/><body> hier wieder alles ok </body></html>
Ä Ä ß	<html><head/><body>ä ä ß </body></html>
<body>Text</bo>	<html><head/> <body>Text </body> </html>

Tabelle 4: HTML-XML-Konvertierung

Diese (syntaktisch) homogenisierten Daten lassen sich nun mit Hilfe von XPath-Anfragen auswerten. Z.B returniert die XPath-Anfrage *//meta[@name="author"]* alle <meta> Elemente, die das Attribut name="author" besitzen. Insgesamt wurde folgende Vorgehensweise zur Extraktion der Daten angewandt:

1. Spezifikation der für eine spezielle Heuristik (z.B. für Titel) relevanten Teilbäume der HTML Dokumente mit Hilfe von XPath Anfragen (Hierbei ist zu berücksichtigen, dass diese Anfragen auf den DOM Bäumen operieren, in die die HTML Do-

kumente konvertiert wurden.) Das Ergebnis einer einzelnen Anfrage ist immer eine geordnete Liste (in der Reihenfolge des Vorkommens im Dokument).

2. Formulierung von Heuristiken auf Basis der im Schritt 1 generierten Daten. Das Ergebnis einer solchen Heuristik ist immer eine mit Wahrscheinlichkeiten gewichtete Liste von Ergebnissen.

Im Projektzeitraum konnten Heuristiken für die Extraktion des Titels, der Schlüsselwörter und des Abstracts implementiert werden. Für die Heuristiken zur Extraktion von Autoren und Institutionen ergaben sich erste Lösungsansätze. Es zeigte sich, dass einfache Heuristiken zur Extraktion dieser Daten nicht funktionieren. Deshalb wurde der Grundgedanke weiterverfolgt, die mögliche Autoren- und Institutionsnamen mit einer bekannten Liste (Datenbank) von Autoren- und Institutionsnamen abzugleichen. Erste einfache Implementierungen dieser Idee erwiesen sich auf Grund ihrer Komplexität und der daraus resultierenden Häufigkeit an notwendigen Datenbankabfragen als nicht effizient genug. Auf eine mögliche Weiterentwicklung wurde zu Gunsten einer Verbesserung der Heuristiken für Titel, Schlüsselwörter und Abstract verzichtet.

Die Definition der Heuristiken (und deren Implementierung) stehen zum Download auf der Homepage von CAP11[11] bereit.

5.2.2.2 Mathematik/Postscript

Ausgangspunkt für die Behandlung von Postscript Dokumenten ist das von der New Zealand Digital Library[12] entwickelte Programm `prescript`[13], das Postscript Dokumente in Text- oder HTML-Dokumente konvertiert.

Das Konvertierungsprogramm `prescript` versucht bei der Konvertierung von Postscript nach HTML die Informationen über Seiten, Paragraphen und Zeilen des ursprünglichen Postscript Dokumentes zu erhalten. Erste Versuchen zeigten jedoch, dass vor allem Umlaute, Sonderzeichen und mathematische Formeln dem Programm besondere Probleme bereiten.

Aus diesem Grund wurde zunächst die Software `prescript`, die im Quellcode in der Programmiersprache `python` vorliegt, modifiziert. Dabei müssen

---

[11] http://www.bonn.iz-soz.de/research/information/carmen/ap11/

[12] Nähere Informationen siehe unter http://www.nzdl.org/

[13] Siehe auch http://www.nzdl.org/html/prescript.html

Umlaute oder Buchstaben mit Akzenten anders behandelt werden als Sonderzeichen wie „ß“. In Postscript werden Buchstaben mit Akzenten durch ein Übereinanderschreiben von zwei Zeichen - dem Grundvokal und dem Akzent - dargestellt. Buchstaben und Sonderzeichen werden durch spezielle Postscript Nummernkürzel kodiert oder durch das Einbinden spezieller Zeichensätze erzeugt. Die Sonderzeichen, die durch das Einbinden spezieller Zeichensätze erzeugt werden, können durch die Software `prescript` grundsätzlich nicht erkannt werden. Zu diesen Sonderzeichen zählen mathematische Symbole und der überwiegende Teil griechischer Buchstaben.

Die Software wurde nun so erweitert, dass die Sonderzeichen und Buchstaben mit Akzenten, die grundsätzlich erkannt werden können, in ihrer UTF-8 Kodierung ausgegeben werden. Dieses Vorgehen ist für die HTML-Darstellung optimal und entspricht der Vorgehensweise von CAP 7. In der von CAP 7 entwickelten Retrievalmaschine können UTF-8 Kodierungen von Zeichensätzen verarbeitet werden.

Durch diese Modifikation der Software ist es gelungen, die Probleme mit den Umlauten und einigen weiteren Sonderzeichen (z.B. Å, Æ, µ) zu lösen. Eine Extraktion von mathematischen Formeln aus Postscript-Dokumenten erscheint jedoch mit diesem Hilfsmittel unrealistisch, da schon die Zeichenerkennung nicht möglich ist.

Für die Anwendung von Heuristiken ist es unumgänglich, einen wohldefinierten Zugang zu der Dokumentenstruktur der von `prescript` erzeugten HTML Dokumente zu haben. Deshalb wurde die Perl Bibliothek `PrescriptStructure` (siehe unter http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP11/) erstellt, die HTML Dokumente auf die folgende Datenstruktur abbildet:

Ein Objekt aus der Klasse `PrescriptStructure` besteht aus einem Array [1..n] von n Seiten. Einer Seite ist evtl. eine Seitenzahl zugeordnet, falls diese im ursprünglichen Postscript Dokument angegeben war.

1. Jede Seite besteht aus einem Array [1..m] von m Paragraphen.

2. Jeder Paragraph besteht aus einem Array [1..k] von k Zeilen. Die Textdaten des ursprünglichen Postscript Dokumentes sind in diesen Zeilen enthalten.

Die Klasse `PrescriptStructure` stellt eine Reihe von Methoden zur Verfügung, mit deren Hilfe auf diese Datenstruktur zugegriffen werden kann.

Aufbauend auf die Klasse `PrescriptStructure` wurde die Klasse `MathHeuristics` erstellt, die Methoden zur Extraktion von Abstract, Schlüsselwörter und MSC Klassifikationen aus den erzeugten HTML Dokumenten anbietet.

Die Extraktion der Autoren und des Titels eines Dokumentes bereitete große Probleme, da das Konvertierungsprogramm `prescript` keine Informationen über Zeichensätze und Zeichengröße erhält. Dieses Problem könnte allenfalls durch Kombination mit einer Analyse der Literaturangaben gelöst werden, was im Projektzeitraum allerdings nicht machbar war.

Im Bereich Schlüsselwörter wird eine Extraktion wie bisher beschrieben zur Erschließung der Volltexte in MPRESS nicht ausreichen, denn in vielen Dokumenten sind keine Schlüsselwörter vom Autor kenntlich gemacht worden. Dennoch ist eine Zuordnung von Schlüsselwörtern bzw. -phrasen wünschenswert und einer Suche nach Volltexttermen vorzuziehen.

Eine Möglichkeit zur Lösung dieses Problems wird durch den Einsatz der in Neuseeland entwickelten Software `Kea`[14] gesehen. `Kea` extrahiert mit Hilfe probabilistischer Verfahren Schlüsselphrasen aus Texten. Anhand einer Trainingsmenge von Texten mit ausgewählten Schlüsselwörtern werden Wahrscheinlichkeiten berechnet, die Aussagen darüber machen, an welcher Stelle eines Textes - sowohl in Bezug auf seine Gesamtlänge, als auch in Bezug auf die Position in einem Absatz - Schlüsselphrasen vorkommen. Diese Wahrscheinlichkeiten hängen offensichtlich von der Dokumentart ab. Wahrscheinlich sind sie aber auch abhängig von Schreibgewohnheiten und Ausdrucksformen in unterschiedlichen Themenbereichen. Aus diesem Grund kann `Kea` auf spezielle Themengebiete und Textarten trainiert werden. Die Trainingsmenge muss nur einen Umfang von ca. 50 Dokumenten haben, wie detaillierte Analysen der Entwickler zeigen. Dies macht ein Training der Software für unterschiedliche Bereiche realistisch. Im Test mit mathematischen Preprints liefert `Kea` sehr gute Ergebnisse.

Ein Problem ist derzeit allerdings die Bewertung der Ergebnisse: Wie sollen in einer Retrievalmaschine durch Heuristiken extrahierte Schlüsselwörter im Vergleich zu durch probabilistische Verfahren generierte Schlüsselwörter behandelt werden? Die Vergabe von hohen Relevanzgewichtungen analog zur Extraktion von Metadaten aus HTML Seiten im Bereich Sozialwissenschaften erscheint hier sinnvoll.

---

[14] Nähere Informationen siehe unter http://www.nzdl.org/Kea/

Während die Heuristiken für die Extraktion der Referenzen gute Ergebnisse lieferten, erwies sich die Extraktion der Autorennamen wie bei den HTML Heuristiken als sehr problematisch und wurde schließlich auf Grund der schlechten Ergebnisse aufgegeben.

In enger Zusammenarbeit mit CAP9[15] wurde an Heuristiken zur Extraktion von PACS Klassifikationen (analog zu der Extraktion von MSC Klassifikationen) gearbeitet. Hier konnten gute Ergebnisse erzielt werden, die für CAP 9 in Bezug auf die Integration von Diensten aus dem Bereich der Mathematik und Physik von zentraler Bedeutung sind.

Die Definition und die Implementierung der PostScript-Heuristiken stehen zum Download auf der Homepage von CAP11 bereit.

5.2.3 Einsatz der Extraktionswerkzeuge in der Gatherer-Komponente aus CAP7

Laut Vorhabensbeschreibung sollen die Extraktionswerkzeuge von CAP11 in der Gatherer-Komponente des Retrieval-Systems aus CAP7 zum Einsatz kommen. Aus diesem Grund wurden sehr früh mit CAP7 die notwendigen technischen Voraussetzungen für einen solchen Einsatz abgestimmt. Nach Bereitstellung der technischen Spezifikationen aus CAP7 wurde mit der Implementierung von Summarizern begonnen, die auf die Extraktionswerkzeuge aufbauen und sich direkt in die Gatherer-Komponente einbinden lassen. Dazu wurden zwei Perl Bibliotheken geschrieben, die die von CAP7 spezifizierte Summarizer-Schnittstelle implementieren. Für die HTML Dokumente heißt der Summarizer *CAP11HTML* und für die PostScript Dokumente *CAP11Postscript*. Beide Summarizer sind inzwischen in der Distribution der Gatherer-Komponente enthalten. Der folgende Konfigurationsausschnitt zeigt, wie sich diese Summarzier in der Gatherer-Komponente benutzen lassen:

```
<Summarizer>
      text/html CAP11HTML
      TYPECAP11Prescript CAP11Prescript
      application/postscript RunProg, prescript_wrapper \\
            %infile %outfile, TYPECAP11Prescript
</Summarizer>
```

Für die Indexierung der Daten müssen diese in eine XML Repräsentation abgebildet werden. Deshalb wurde ein Transformationsmodul *CAP11* nach

[15] Infos zu CARMEN AP9 s. http://www.physik.uni-oldenburg.de/carmen/ap9/

Vorgaben von CAP7 erstellt, das die im Gatherer enthaltenden Daten (des HTML Dokumentenbestandes) auf XML Dokumente abbildet, die der folgenden DTD genügen:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT description ( (title |
                         subject |
                         abstract )* ,
                         full-text,
                         url
)>
<!ELEMENT url (#PCDATA)>
<!ELEMENT title (#PCDATA | weight |  method | value)*>
<!ELEMENT subject (#PCDATA | weight |  method | value)*>
<!ELEMENT abstract (#PCDATA | weight |  method | value)*>
<!ELEMENT full-text (#PCDATA)>
<!ELEMENT weight (#PCDATA)>
<!ELEMENT method (#PCDATA)>
<!ELEMENT value (#PCDATA)>
```

Ein Beispieldokument sieht z.B. folgendermaßen aus:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<description>
      <title>
            <weight>0.8</weight>
            <method>2</method>
            <value> the extracted title </value>
      </title>
      <subject>
            <weight>1</weight>
            <method>4</method>
            <value> some keywords </value>
       </subject>
      <url>http://www.myorg.org/doc.html</url>
</description>
```

Die so generierten Daten sind von HyREX erfolgreich indexiert worden. Die dazu notwendigen Konfigurationsdatei für die Document Definition Language (DDL) ist im Anhang und auf der Homepage von CAP11 zu finden.

5.2.4 Ergebnisse/Tests

5.2.4.1 Sozialwissenschaften

In dem oben beschriebenen iterativen Verfahren zur Verbesserung der Heuristiken werden diese an die Osnabrücker Projektpartner gesendet, dort in das Extraktions-Tool integriert und die Ergebnisse über eine XML-Datei zurück an das InformationsZentrum Sozialwissenschaften übermittelt.

Im InformationsZentrum Sozialwissenschaften wurde als Java-Programm ein Test-Tool entwickelt (siehe Abbildung 5), welche die Auswertung wesentlich vereinfachte. Die extrahierten Metadaten werden in eine Baumstruktur geladen und können komfortabel durchblättert werden. Die jeweils ausgewählten Dokumente können in einem Browser- und in einem HTML-Quelltext-Fenster zu dem jeweiligen Eintrag der XML-Datei betrachtet werden, so dass sich die Extraktionsergebnisse schnell und einfach mit den zugrunde liegenden Dokumenten vergleichen lassen. Geplante Erweiterung erlauben das automatische Vergleichen verschiedener Transfer-Durchläufe und die Nutzung für andere Datei-Formate als HTML.

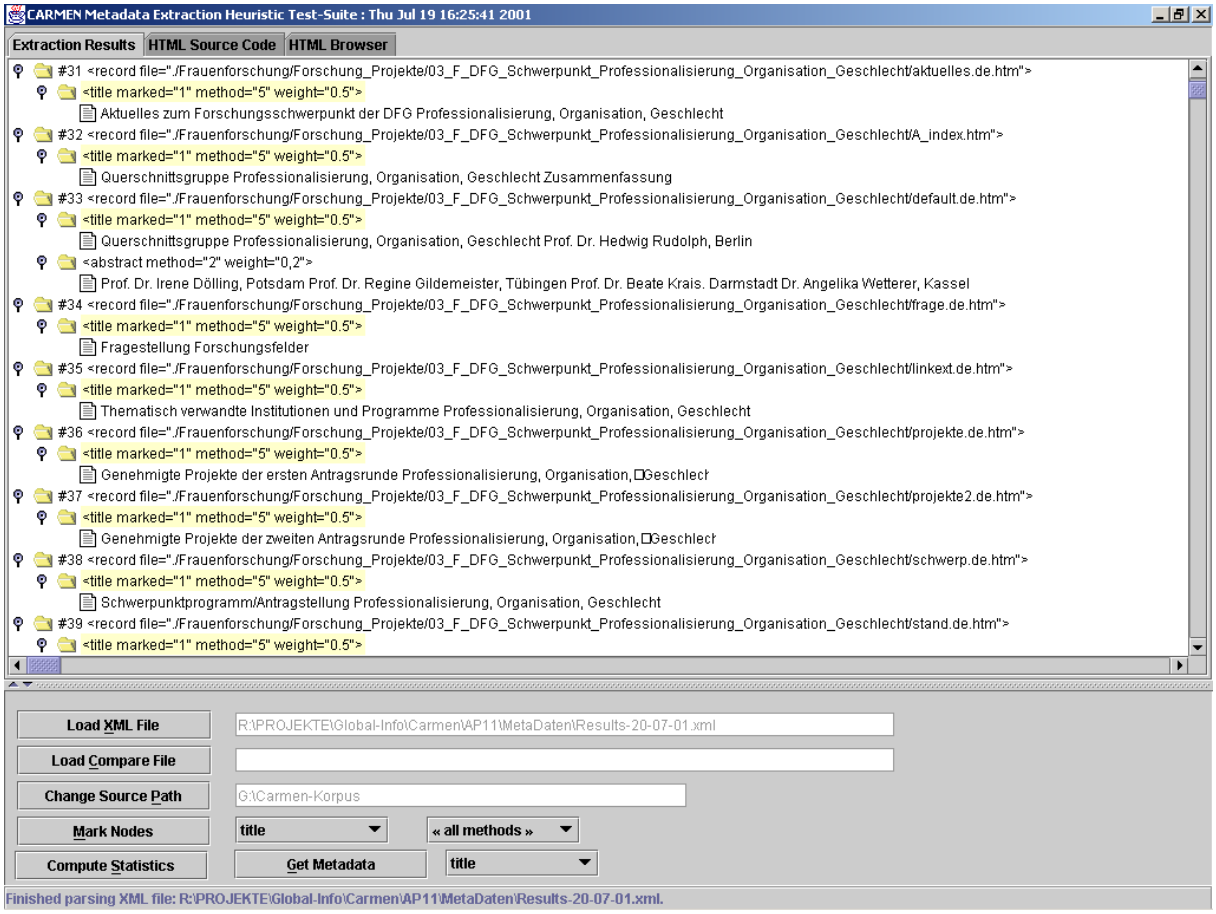

Abbildung 5: Test-Tool für die Überprüfung der Heuristiken

Überprüft werden dabei die Vorkommen der Metadaten Titel, Keywords und Abstract.

5.2.4.1.1 Evaluationsverfahren

Bei der Bewertung der Metadaten wurden unterschiedliche Verfahren je nach Metadatentyp angewandt.

5.2.4.1.1.1 Titel

Aufgrund der sehr hohen Fallzahlen des Vorkommens eines Titels in den Metadaten (in dem zu testenden CARMEN-Korpus von 3661 html Dokumenten wurden 3537 Titel extrahiert) wurde eine Stichprobe in einem Zufallsverfahren in 10er Schritten angewandt:

- zunächst wurde eine Zahl n zwischen 1 und 10 durch „Würfeln“ ermittelt
- diese Zahl entspricht der Dokumentnummer des ersten zu bewertenden Dokumentes
- folgend werden die n+(x mal 10). Dokumente bewertet (für x = 1 bis Ende)
- falls gerade dieses Dokument keinen Titel hat, wird stattdessen das nächste darauf folgende Dokument mit Titel ausgewählt und als neues Startdokument festgelegt
- die weiteren Bewertungen beginnen beim neuen Startdokument nach dem gleichen Algorithmus

Bewertet wurde die syntaktische **und** die semantische Relevanz des extrahierten Titels. Syntaktische Relevanz bedeutet hier, ob eine korrekte der jeweiligen heuristischen Methode entsprechende Extraktion des Titels gegeben ist. Dies war in allen überprüften Dokumenten der Stichprobe der Fall. Bewertet wurde in 3 Stufen: relevant, nicht relevant, nicht bewertbar. Anders verhält es sich bei der Bewertung der semantischen Relevanz. Es kann zunächst nur von einer systemorientierten Bewertung ausgegangen werden. Hier wurde nach 4 Stufen bewertet:

- Relevanzstufe 1: hohe Präzision und Vollständigkeit des Titels;
  hier ist eine hohe Rückgewinnungs- (recall)[16] und Trefferquote (precision)[17] zu erwarten
- Relevanzstufe 2: hohe Präzision für einen Teil der Titelextraktion, aber unvollständig und/oder auch Extraktion nicht relevanter Anteile
- Nicht relevant
- Nicht bewertbar

---

[16] Recall = Zahl der relevanten ausgegebenen Dokumenten dividiert durch Zahl der relevanten Dokumente im Speicher

[17] Precision = Zahl der relevanten ausgegebenen Dokumenten dividiert durch Zahl der insgesamt ausgegebenen Dokumente

Bei der Entscheidung für eine semantische Relevanzstufe fielen Allgemeinbegriffe wie z.B. Zeitschriften, Einträge, Colloquium, Geographika, Projekte, Einrichtungen, Recherche, New, Zu dieser Ausgabe, Bildung, Netzwerk, Forschungsprojekte, ... unter die Bewertung „nicht relevant“. Diese Begriffe sind zwar nicht völlig irrelevant, erhalten ihre Bedeutung jedoch erst im Kontext mit weiteren Begriffen, welche im Titel aber nicht extrahiert sind. Akronyme z.B. von Institutionen erfuhren in der semantischen Relevanzbewertung keine Abstriche, setzen allerdings im Falle einer benutzerorientierten Bewertung (subjective view) entsprechende Kenntnis voraus.

5.2.4.1.1.2 Keywords

Von den im Testkorpus vorkommenden keywords wurde die Häufigkeit durch manuelle Auszählung ermittelt und die ausgewählten Worte stichprobenartig auf ihren Sinngehalt überprüft. Da keywords durch den Erzeuger der html-Seite explizit vergeben werden müssen und nicht durch html-Generatoren automatisch erzeugt werden (wie z.B. der Seitentitel „neue Seite 1“ durch MS-Frontpage), kann man bei der Analyse davon ausgehen, dass diese inhaltlich korrekt sind.

5.2.4.1.1.3 Abstract

Bei abstracts wurden je nach Größe der Ergebnismenge pro Methode verschiedene Verfahren angewandt. Für die Methoden <1> und <2> wurden die Häufigkeiten ausgezählt und analog zum Verfahren bei anderen Heuristiken auf Plausibilität hin überprüft. Aufgrund der hohen Fallzahlen des Vorkommens von Überschriften (Methode <3>) in den Metadaten (1139 Dokumente des CARMEN-Testkorpus) wurde analog zum Verfahren unter Kapitel 5.2.4.1.1.1 eine Stichprobe gebildet.

5.2.4.1.2 Ergebnisse

5.2.4.1.2.1 Titel

Von den 3661 Dokumenten konnten 3537 Titel extrahiert werden. In der Stichprobe wurden mit Zufallsauswahl 360 überprüft, die syntaktisch zu 100% korrekt extrahiert waren. Unter Anwendung des in Kapitel 5.2.4.1.1.1 beschriebenen Verfahrens kann hinsichtlich der semantischen Relevanz der entwickelten Titelheuristiken festgehalten werden:

Relevanz	Anteil in %	Anzahl Dokumente
Stufe 1	38	138
Stufe 2	42	151
Nicht relevant	8	29
Nicht bewertbar	12	42

Tabelle 5: Prozentualer Anteil der Bewertungsstufen

Methode	Relevanzbewertung
M 1	die Relevanz liegt überwiegend im mittleren Bereich (Stufe 2)
M 2	die Relevanz liegt verteilt im hohen (Stufe 1) und mittleren Bereich (Stufe 2)
M 3	die Relevanz liegt überwiegend im mittleren Bereich (Stufe 2), insbesondere weil der Titelextraktion in den meisten Fällen noch eine weitere Information für eine hohe Relevanz (Stufe 1) fehlt.
M 4	die Relevanz liegt überwiegend im hohen Bereich (Stufe 1).
M 5	die Relevanz liegt überwiegend im mittleren Bereich (Stufe 2). Besonders auffällig ist hier, dass in einem Teil der Titelextraktion zu viele nicht relevante Anteile enthalten sind (in 41 von 121 Fällen)
M 6	durchschnittlich in allen Relevanzstufen. Allerdings sind hier erheblich viele nicht bewertbare Dateien infolge leerer Dokumente vorhanden.

Tabelle 6: Überwiegende Bewertungsstufen bei den Methoden 1-6

Aufgrund der semantischen Relevanzbewertungen wurden die vermuteten Wahrscheinlichkeiten für das Vorkommen der Titelextraktion nach den jeweiligen Heuristiken korrigiert. Dabei wurde für Relevanzstufe 1 der Faktor 1,

für Relevanzstufe 2 Faktor 0,5, für nicht relevante Titel der Faktor 0 (alternierend: - 0,5) und für nicht bewertbare Extraktionen Faktor –1 angenommen. Es wurden folgende Wahrscheinlichkeitskorrekturen errechnet:

Methode	vermutet	korrigiert	Alternierender Wert bei Faktor -0.5 für nicht relevante Titel
M 1	1.0	0.4	0.4
M 2	0.8	0.7	0.7
M 3	0.8	0.6	0.5
M 4	0.8	0.9	0.8
M 5	0.5	0.5	0.4
M 6	0.5	- 0.05	- 0.15

Tabelle 7: Wahrscheinlichkeiten (M 1 bis M 6)

Für die Methoden 7 und 8 wurde aufgrund der geringen Fallzahlen im xml-Testfile (Methode 7: 20 Dokumente; Methode 8: 2 Dokumente) eine Bewertung für **alle** nach diesen Methoden extrahierten Titel durchgeführt.

Methode	Relevanzbewertung
M 7	die Relevanz liegt überwiegend im hohen (Stufe 1) und mittleren Bereich (Stufe 2)
M 8	die Relevanz liegt für ein Dokument bei Bewertungsstufe 2, das andere ist nicht relevant.

Tabelle 8: Überwiegende Bewertungsstufen bei den Methoden 7 und 8

Für die korrigierten Wahrscheinlichkeiten lässt sich folgende Aussage treffen:

Methode	vermutet	korrigiert	Alternierender Wert bei Faktor - 0.5 für nicht relevante Titel
M 7	0.3	0.7	0.6
M 8	0.1	0.3	0

Tabelle 9: Wahrscheinlichkeiten (M 7, M 8)

5.2.4.1.2.2 Keywords

Keywords konnten mit der in Kapitel 5.2.4.1.1.2 beschriebenen Methode im Testkorpus in 871 (23,8%) von 3661 Dokumenten extrahiert werden.

In 64 weiteren Fällen wurde zwar ein keyword-tag extrahiert, dieser ist aber ohne Inhalt [„...“]. Da diese Keyword-Vergabe ausnahmslos vom gleichen Informationsanbieter erstellt wurde, ist zu vermuten, dass die Metadaten zum Zeitpunkt des Downloads noch nicht vergeben waren. Ein Gegencheck der relevanten Seiten im aktuellen Internetbestand (ca. 15 Monate nach dem Download), ergab, dass diese Seiten, sofern sie im Internet noch identisch vorhanden sind, inzwischen vergebene keywords haben, die noch zusätzlich nach dem Dublin-Core-Standard vergeben sind.

Die Dokumente der Titelstichprobe wurden zusätzlich auch auf Vorkommen und Sinngehalt der vergebenen keywords geprüft. Von den 360 überprüften Dokumenten sind in 86 Fällen keywords vorhanden. Mit einer Ausnahme entsprechen alle vergebenen keywords dem Informationsangebot. Es lassen sich aber zwei Tendenzen feststellen:

- Tendenz 1: die angegebenen keywords sind bei allen Seiten eines Informationsanbieters identisch
- Tendenz 2: die angegebenen keywords von einem Informationsanbieters sind Seiteninhaltsentsprechend spezifiziert.

5.2.4.1.2.3 Abstract

Mit der Heuristik-Methode 1 müssen alle Dokumente gefunden werden, die einen description-tag im header der Seite enthalten. Leider hat die Implentation nicht korrekt funktioniert, sodass keine quantitativen Aussagen gemacht werden können. Bei der jetzigen Version konnten 391 Dokumente identifiziert werden. Mindestens weitere 380 Dokumente enthalten jedoch einen syn-

taktisch korrekt verwendeten description-tag, ohne dass diese identifiziert werden konnten.

Für die inhaltliche Bewertung ergibt sich ein analoges Problem, wie bei den Schlüsselbegriffen. Bei der Vergabe der description-tags lassen sich nach stichprobenartiger Durchsicht fünf Tendenzen feststellen:

- Tendenz 1: Vergabe eines description-tags ausschließlich auf der Homepage / Indexseite eines Informationsanbieters.
- Tendenz 2: Vergabe eines description-tags inhaltsidentisch mit dem keyword-tag
- Tendenz 3: "Zufalls"vergabe auf irgendeiner Seite innerhalb eines Webangebotes
- Tendenz 4: Vergabe seitenspezifischer description-tags für alle Seiten eines Webangebotes
- Tendenz 5: Vergabe eines gleichen description-tag für alle Seiten eines webstrukturellen Teilbereichs

Die Implementation der Heuristiken der Methode 2 gelang innerhalb des Projektzeitraumes nicht zufriedenstellend. Die Logik und deren Implementation hätten weiter ausgefeilt werden müssen, um akzeptable Resultate zu erzielen. Im derzeitigen Zustand wird der gesamte Text nach den Signalbegriffen (Zusammenfassung, Summary, Abstrakt oder Abstract) durchsucht und der dann folgende Absatz als Inhaltsbeschreibung genommen. Das bringt natürlich in den allermeisten Fällen im Sinne einer Inhaltsbeschreibung keine verwertbaren Ergebnisse. Besser wäre es, wenn man derart einschränken könnte, dass die Signalbegriffe nicht im Fließtext, sondern hervorgehoben stehen (fett, als Überschrift, kursiv, also in enger Nachbarschaft zu '>' und '<', o.ä.). Das bleibt einem Nachfolgeprojekt überlassen. Prinzipiell kann man aber davon ausgehen, dass das angedachte Verfahren auch in sozialwissenschaftlichen Texten akzeptable Ergebnisse zeigen würde. Pretests über den CARMEN-Korpus ergaben eine Menge von ca. 130 Dokumenten, in denen zusammenfassende Abschnitte mit den o.g. Signalbegriffen eingeleitet werden. Diese Abschnitte sind dann qualitativ sehr gute Zusammenfassungen der betreffenden Seite. Die hinreichend sichere Identifikation dieser Abschnitte bereitete jedoch größere Probleme, als erwartet. Ursachen dafür liegen vor allem in der nicht spezifikationskonformen Verwendung von HTML.

Die Abstract-Extraktion nach Methode 3 (summierende Liste aller Überschriften der Seite) ergab eine Treffermenge von 1139 von 3661 Dokumenten (entspricht 31%). Die Bewertung der Ergebnisse vollzog sich für

(entspricht 31%). Die Bewertung der Ergebnisse vollzog sich für eine 10er Stichprobe (114 Dokumente).

Methode	Anzahl gesamt	Mit Zufallsauswahl gefunden	Relevanzbewertung Syntax			Relevanzbewertung Semantik			
			relevant	Nicht relevant	Nicht bewertbar	Relevanz Stufe 1	Relevanz Stufe 2	Nicht relevant	Nicht bewertbar
M 3	1139	114	108	6	0	15	80	17	2
%		10%	94,7%	5,3%	0	13,2%	70,2%	14,9%	1,7%

Tabelle 9: Relevanzbewertung der Abstract Heuristik, Methode 3

Bewertet wurden analog der extrahierten Titel die syntaktische und semantische Relevanz des extrahierten Abstract. Bei der syntaktischen Überprüfung wurden 6 Extraktionen (5,3%) von 114 überprüften Dokumenten als nicht relevant eingestuft. In 3 Fällen wurde mindestens eine der vorhandenen Überschriften nicht extrahiert, davon trat in einem Fall zusätzlich eine Extraktionsdopplung auf. In einem weiteren Fall wurde ein nicht vorhandenes Wortfragment extrahiert, dessen Ursprung nicht ermittelbar ist. In den zwei weiteren Fällen wurde der angewandten Heuristik folgend eine falsche Methode verwandt. Das Abstract wurde nach Methode 3 extrahiert, obwohl in den Metadaten ein description-tag vorhanden ist.

Die Bewertung der semantischen Relevanz wurde mit den gleichen Abstufungen wie bei der Titelbewertung vorgenommen. Die Qualität eines extrahierten Abstracts nach dieser Methode ist nicht mit dem Abstraktcharakter z.B. eines wissenschaftlichen Zeitschriftenaufsatzes zu vergleichen, es sollte aber eine das Dokument in seinen wesentlichen inhaltlichen Zügen entsprechende Beschreibung erkennen lassen.

Weitere Probleme:

- **Webstruktur/ Textausschnitt/ - menge/-struktur**
  Die wenigsten html-Dokumente sind in sich geschlossene Texte. Viele Seiten sind Teil einer umfassenderen Text-/ Informationsstruktur, deren Bezug bei der Anwendung der Heuristiken außen vor bleiben muss.
- **Html Formatierung**
  Html wird häufig nicht standardgemäß angewandt bzw. durch Verwendung z. B. grafisch orientierter Editoren nicht beachtet. Es entstehen Formatierungsungenauigkeiten, die sich auf die Extraktionen auswirken;

z.B. ein Dokument hat eine Gliederungsstruktur mit entsprechend vorhandenen Überschriften (meint hier nicht <hx>, sondern Textbezogene Überschriften), denen aber nicht <hx> zugewiesen ist

- **Titelidentität**
  Da in einigen angewandten Methoden der Titelheuristiken auch Überschriften verwandt wurden wie hier in der Methode 3, tritt insbesondere bei Dokumenten, die nur eine Überschrift enthalten, Titelidentität auf. Ein Dokument erhält dadurch keinen höheren Beschreibungswert.

5.2.4.2 Mathematik

Im Bereich der Mathematik wurden die PostScript Extraktionstools an einem Testkorpus bestehend aus mathematischen Preprints (im PostScrip-Format).getestet. Dazu wurden zunächst 240 (zufällig ausgewählte) Dokumente mit dem Programm `prescript` in HTML Dokumente konvertiert. Danach wurden mit den entwickelten Extraktionstools die Metadaten Abstracts, Keywords, MSC Klassifikation und Referenzen extrahiert. Die Ergebnisse können als sehr positiv bewertet werden, insbesondere wurden 37 der 240 Dokumente vollständig ausgewertet:

Die folgende Tabelle zeigt die Anzahl des Vorkommens der einzelnen Metadaten, die Anzahl der davon korrekt erkannten Metadaten, die Anzahl der falsch bzw. unvollständig erkannten Metadaten und die Anzahl der nicht erkannten Metadaten.

	Vorhanden	Richtig erkannt	Unvollst. erkannt	Nicht erkannt
Abstracts	30 (81%)	27 (90%)	2	1
Keywords	21 (58%)	20 (95%)	1	0
MSC	22 (59%)	22 (100%)	0	0
Referenzen	30 (81%)	26 (86%)	2	2

Tabelle 10: Extrahierte Metadaten

Damit wurden 90% der vorhandenen Abstracts, 95% der vorhandenen Keywords, 100% der vorhandenen MSC Klassifikationen und 86% der vorhandenen Referenzen korrekt erkannt. Ein ausführliche Dokumentation steht im Internet unter der Adresse http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP11/ zur Verfügung.

5.3 Statistisch erzeugte Transferbeziehungen

Die Auswertung der Metadatenextraktion zeigt zwar bei den erfolgreich generierten Daten eine gute bis sehr gute Qualität, allerdings sind die Quoten, bei denen die Heuristiken greifen, teilweise sehr gering (Keywords 23,8%, Abstract 31%). Die somit verbleibende Heterogenität des Datenmaterials versucht der Ansatz von CAP11 mit statistischen Transferbeziehungen zwischen Termen verschiedener Textkorpora weiter einzuschränken. Diese lassen sich über Doppelkorpora erstellen, d.i. Dokumentsammlungen, deren einzelne Dokumente nach mehreren Begriffssystemen indexiert sind. Das grundsätzliche Vorgehen bei der Erzeugung statischer Transferbeziehungen ist, die den Dokumenten zugeordneten Terme zu extrahieren, und das gemeinsame Auftreten von Termen aus verschiedenen Begriffssystemen zu analysieren. Dadurch ist es möglich, Beziehungen zwischen verschiedenen Begriffssystemen über eine Term-Term-Matrix herzustellen.

5.3.1 Evaluierung von Term-Term-Beziehungen mit JESTER

Für die Realisierung statistischer Transfermodule wurde auf das Werkzeug „Jester“ (**J**ava **E**nvironment for **S**tatistical **T**ransf**ER**s) zugegriffen, das im Kontext des Projekts Elvira II erstellt wurde (cf. Hellweg 2000[18]). Für den Einsatz im Projekt Carmen wurde das Werkzeug angepasst und erweitert, insbesondere um die Berücksichtigung von Doppelkorpora zwischen Freitexttermen und Sacherschließungssystem (cf. Kap.5.3.2.1).

Jester ist ein Java-basiertes Tool, das bei der Neuerschließung eines Datenbestandes oder bei der Neuaufnahme eines Inhaltserschließungssystems eingesetzt werden kann, um die statistischen Transferbeziehungen zu erstellen, die dann über die Transfermodule in das Carmen-Retrieval eingehen. Es erstellt über eine Term-Term-Matrix Termerweiterungsregeln, die später für das Retrieval benutzt werden können.

Als Maß für den statistischen Zusammenhang zweier Terme wird die „bedingte Wahrscheinlichkeit“, mit der zwei Terme gemeinsam auftreten, berechnet. Für einen Deskriptor **a** bestimmt **P(a)** die Wahrscheinlichkeit, dass **a** einem Dokument zugeordnet ist. Sie lässt sich ermitteln, indem die Zahl der

---

[18] Hellweg, Heiko (erscheint): "Statistische Transferbeziehungen zur Text-Fakten Integration". In: Krause, Jürgen; Stempfhuber, Maximilian (erscheint): Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA (IZ-Forschungsberichte 4). Bonn.

Dokumente, in denen **a** auftritt durch die Gesamtzahl der Dokumente geteilt wird. **P(a ∧ b)** bezeichnet die Wahrscheinlichkeit, dass zwei Terme **a** und **b** gemeinsam auftreten. Ist das Auftreten der beiden Terme unabhängig voneinander, so sollte in etwa **P(a ∧ b) = P(a)* P(b)** gelten. Weichen diese Werte erheblich voneinander ab, so liegt ein systematischer Zusammenhang zwischen den Termen vor. Treten sie häufiger gemeinsam auf, als vorhergesagt, so sind sie wahrscheinlich nah verwandt. Treten sie deutlich seltener gemeinsam auf, so liegt nahe, dass sie eine gegensätzliche Beziehung haben und einander ausschließen.

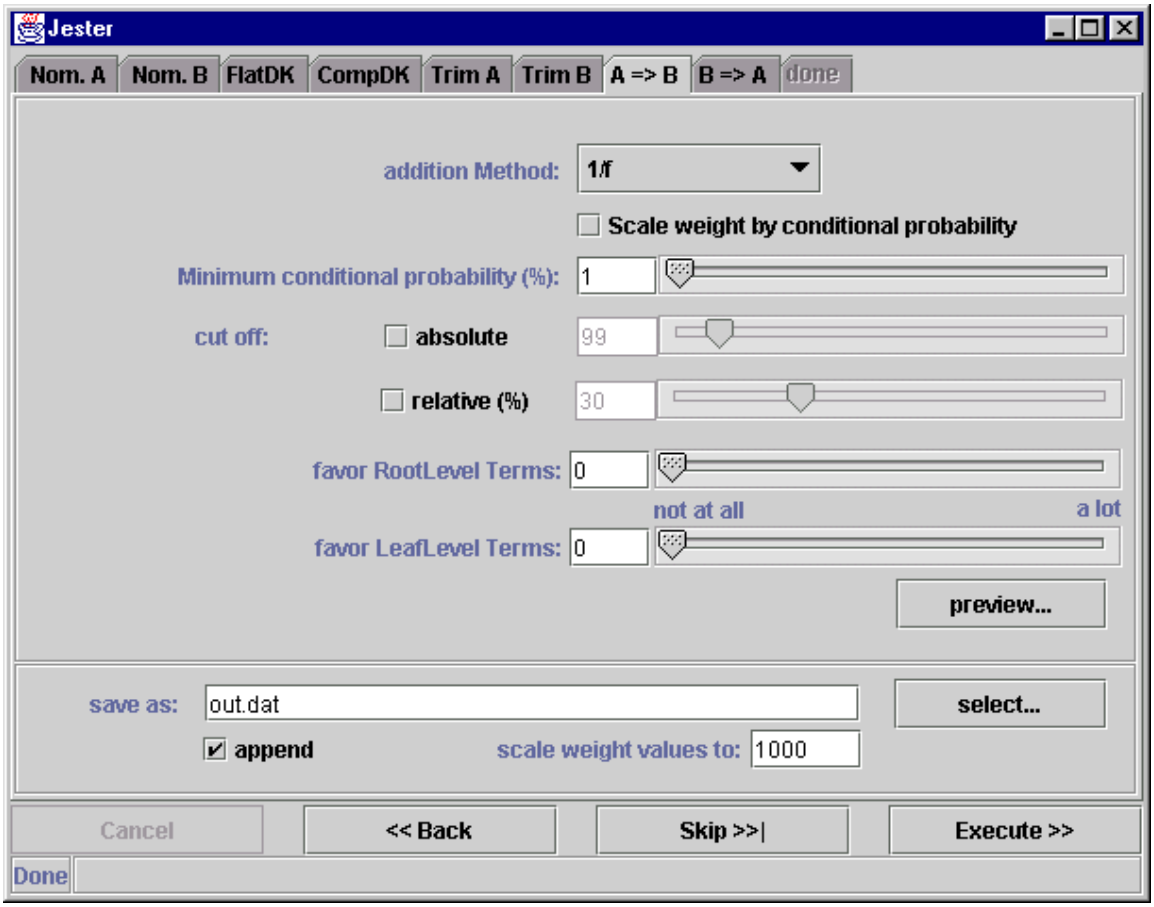

Abbildung 1: Bildschirmkopie Jester Parameterauswahl für Schwellenwerte <neuer Screenshot>

Anstatt nur diese symmetrische Beziehung zu betrachten, die bei geeigneter Wahl eines Schwellwertes die Ermittlung von Termpaaren erlaubt, die quasi Synonym verwendet werden, kann auch die gerichtete bedingte Wahrscheinlichkeit betrachtet werden. Hier wird nur betrachtet, wie wahrscheinlich es ist, dass, wenn Deskriptor **a** vergeben wurde, ebenfalls Deskriptor **b** auftritt. **P(a) / P(a ∧ b)** beschreibt diesen Zusammenhang. Je größer dieser Wert ist, um so

höher ist die gerichtete Abhängigkeit, die z.B. auftritt, falls **a** semantisch ein Unterbegriff von **b** ist.

Jester führt den Bearbeiter schrittweise durch die notwendigen Abläufe und unterstützt bei der Auswahl von Schwellenwerten. Auf diese Weise kann direkt bei der Manipulation der Parameter die Auswirkung beobachtet werden und interaktiv eine optimale Auswahl getroffen werden.

5.3.2 Erstellung der Doppelkorpora und Tranferbeziehungen

5.3.2.1 Sozialwissenschaften

Für den Bereich Sozialwissenschaften bestand die Sonderproblematik, dass die sozialwissenschaftlichen Internet-Quellen im Carmen-Testkorpus nach keinem Deskriptor- oder Klassifikationssystem indexiert wurden, so dass auch kein Doppelkorpus vorlag (siehe Abbildung 6). Lediglich die Freitextterme des Korpus konnten über einen Volltextindexierer gewonnen werden.

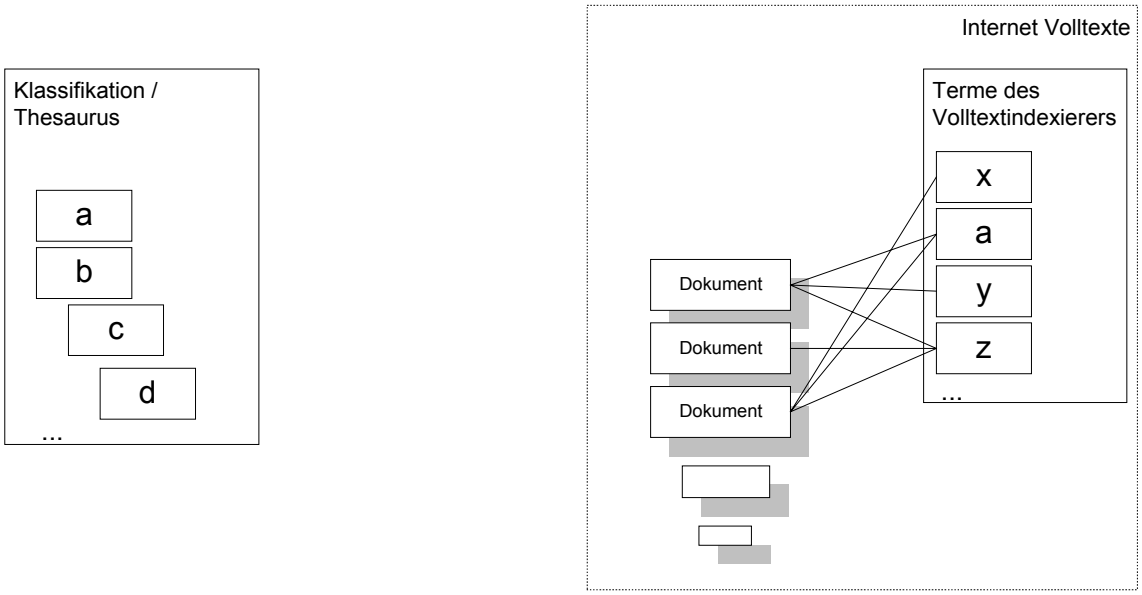

Abbildung 6: CARMEN-Testkorpus: Sozialwissenschaftliche Volltexte im Internet ohne "Verschlagwortung"

Einen solchen Korpus nachträglich intellektuell zu erstellen, wäre vom Aufwand her nicht sehr sinnvoll gewesen. Für die Sozialwissenschaften ist es aber ein erklärtes Ziel, die Internet-Dokumente besser zu „verschlagworten", d.h. mittels Dokumentationssprachen inhaltlich auszuzeichnen. Damit wäre ein Korpus vorhanden, das sich über die oben beschriebenen Verfahren bearbeiten ließe und einen statistischen Transfer von Thesaurus bzw. Klassifikati-

on zu den Freitexttermen ermöglichte. Da aber derzeit keine „verschlagworteten“ Internet-Quellen für die Sozialwissenschaften vorliegen, musste der Doppelkorpus, d.h. eine intellektuelle Indexierung simuliert werden. Die Simulation des Doppelkorpus geht von dem Szenario aus, dass ein Benutzer, der seine Suche mit kontrolliertem Vokabular aus einem Thesaurus beginnt, auch relevante Internet-Dokumente findet, die nicht intellektuell indexiert wurden.

Die Simulation der intellektuellen Indexierung impliziert, dass ein schwächeres Verfahren implementiert werden musste, das *vage* Deskriptor-Dokument-Relationen erzeugt – im Unterschied zur intellektuellen „Verschlagwortung“, wo Deskriptor-Dokument-Zuordnungen i.d.R. ungewichtet, d.h. mit 1 gewichtet sind. Ähnlich wie ein Volltextindexierer, produziert der Carmen-Indexierer dagegen Relationen, die auf einer [0,1]-Skala gewichtet sind (siehe Abbildung 7).

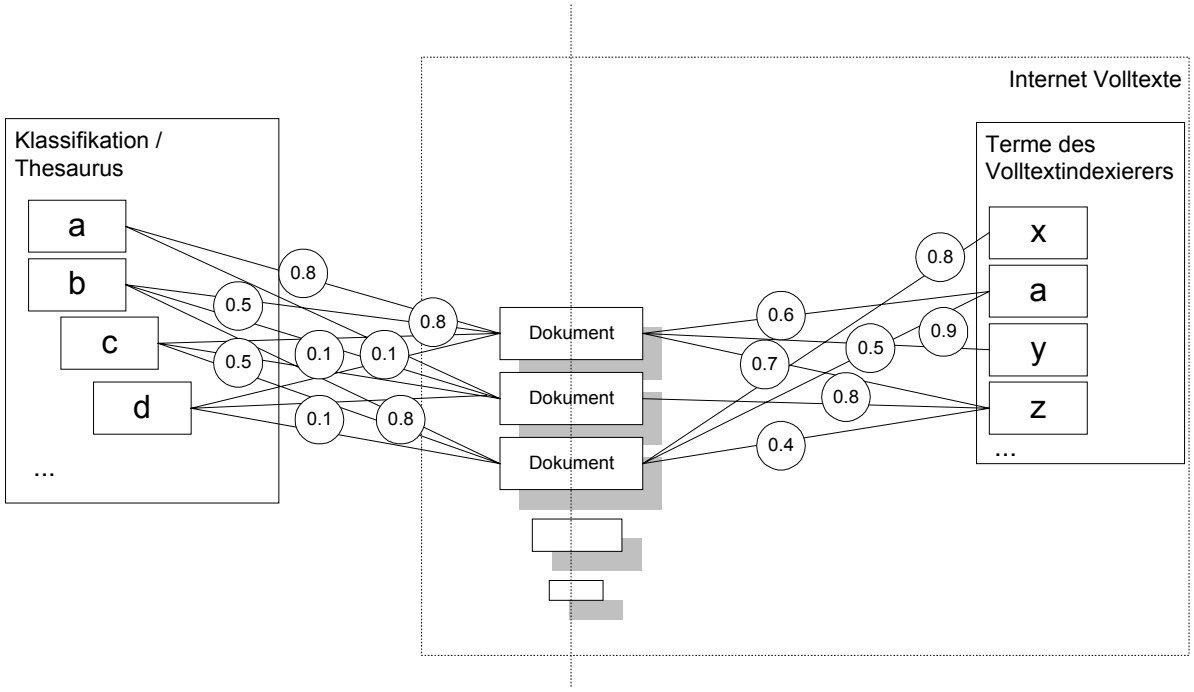

Abbildung 7: Parallel-Korpus-Simulation mit vagen Deskriporen und Volltexttermen

Für die automatische Zuordnung von kontrolliertem Vokabular zu sozialwissenschaftlichen Internet-Dokumenten wurde der Thesaurus des Informationszentrums Sozialwissenschaften herangezogen. Das grundsätzliche Vorgehen bei der Simulation der intellektuellen „Verschlagwortung“ ist, jeden einzelnen

Deskriptor aus dem Thesaurus als Anfrage gegen den Carmen-Testkorpus zu betrachten. Als Retrievalmaschine wurde die Volltextsuchmaschine Fulcrum SearchServer 3.7 verwendet, die die Dokumente im Korpus nach ihrer Ähnlichkeit zum Anfragevektor (Rankingwert) absteigend sortiert. Jedes Dokument in der Ergebnismenge, dessen Rankingwert einen bestimmten Schwellwert überschreitet, wurde mit dem Ausgangsterm indexiert. Die Deskriptor-Dokument-Relation wurde mit dem Rankingwert gewichtet.

Da dieses Vorgehen jedoch nicht die Relevanz eines Deskriptors für das jeweils gefundene Dokument berücksichtigt, sondern nur das Vorkommen des Terms im Dokument, wurden die Anfragevektoren jeweils um thematische Kontextinformationen erweitert. Dieses Verfahren soll sicherstellen, dass Dokumente nur mit den Deskriptoren indexiert werden, mit denen das Dokument inhaltlich wohl-beschrieben ist.

Zu diesem Zweck wurde ein weiterer Parallelkorpus mit SOLIS-Dokumenten aufgebaut, die thematisch zum Carmen-Korpus passen. Die SOLIS-Dokumente sind mit Deskriptoren aus dem IZ-Thesaurus indexiert. Thematische Kontextinformationen zu jedem IZ-Deskriptor wurde durch eine Co-Word-Analyse der den SOLIS-Dokumenten zugeordneten Deskriptoren gewonnen. Dabei wurden die Co-Word-Relationen einer Jester-Analyse unterzogen und der initiale Anfragevektor (mit dem Ausgangsterm) um die Co-Begriffe des Ausgangsterms angereichert, die sich auch in der von Jester produzierten Term-Term-Matrix wiederfanden. Für die Gewichtung der Relationen wurde der Äquivalenzindex herangezogen.

Mit den um solche Zusatzterme erweiterten Anfragevektoren wurde nun im FULCRUM-indexierten Carmen-Korpus gesucht. Als Retrievalmaschine wurde das FULCRUM-Vektorraummodell verwendet. Die gefundenen Dokumente wurden, soweit deren Rankingwert einen bestimmten Schwellwert nicht unterschritt, mit dem Ausgangsterm indexiert. Als Gewichtung wurde der Rankingwert des Dokuments bezüglich der Anfrage genommen (siehe Abbildung 7).

Die Volltextterme des Carmen-Korpus wurden mit Hilfe eines Stringtokenizers gewonnen und mittels eines Stemmingverfahrens normalisiert. Um das (sehr umfangreiche) Wortmaterial aus den Volltexten auf sinnvolle Terme zu reduzieren, wurden nur Terme berücksichtigt, die auch im IZ-Thesaurus (deutsch und englich) sowie in der Liste der „freien“ Deskriptoren des IZ vorkommen.

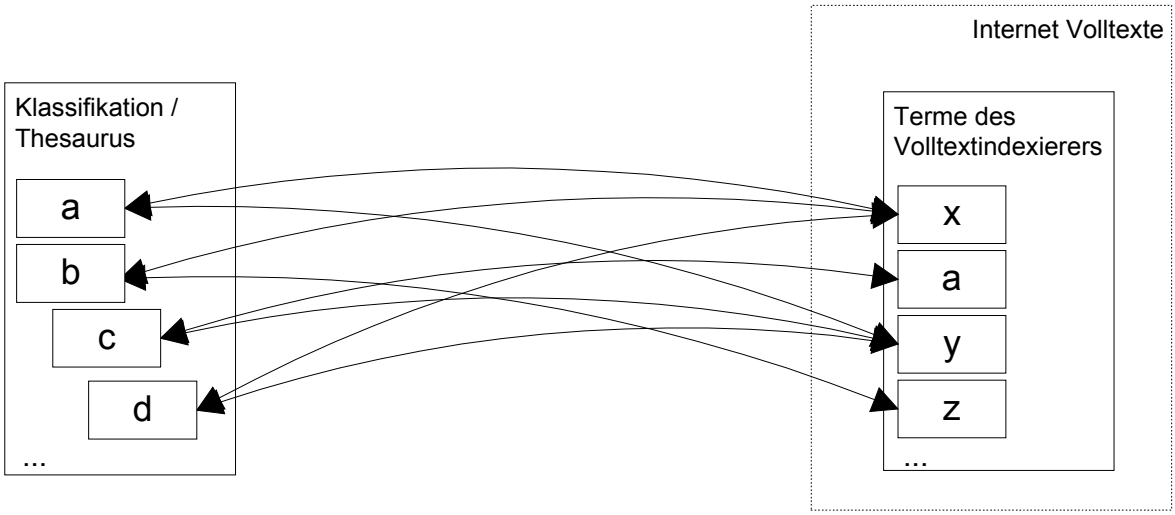

Abbildung 8: Relationen zwischen Deskriptoren und Volltexttermen (Term-Term-Matrix)

Auf der Basis des auf diese Weise simulierten Doppelkorpus' wurden Relationen zwischen Deskriptoren und Freitexttermen extrahiert und mit Jester analysiert. Für die Erzeugung von Term-Term-Beziehungen (siehe Abbildung 8) wurde der Äquivalenzindex auf der Basis der bedingten Wahrscheinlichkeit, mit der zwei Terme kookkurrieren, herangezogen. Die auf diese Weise produzierte Term-Term-Matrix wurde in einer Oracle-Tabelle abgespeichert, die wiederum von den in Kapitel 5.4.1.2 beschriebenen Transfermodulen für den Überstieg von Thesaurustermen auf Freitextterme und umgekehrt abgefragt wird.

5.3.2.2 Mathematik

Im Unterschied zu sozialwissenschaftlichen Dokumenten im Internet lagen für den Bereich Mathematik bereits klassifizierte Internet-Quellen vor. Für diese wurde ein Transfer zwischen Klassifikation und Freitexttermen geschaffen. Hierzu wurden die Freitexte mit Hilfe eines Stringtokenizers in Einzelterme zerlegt, wobei Stoppwörter herausgefiltert wurden. Die verbliebenen Einzelterme wurden mittels eines Porter-Stemming-Verfahrens auf die entsprechende Grundform reduziert. Die auf diese Weise bereinigten Freitextterme wurden mit den Klassifikationstermen in Beziehung gesetzt. Diese Co-Word-Relationen wurden ebenfalls mit Jester statistisch analysiert und auf die semantisch gehaltvollen Relationen reduziert.

Sobald „verschlagwortete“ Internet-Dokumente auch für die Sozialwissenschaften vorliegen, kann das beschriebene Verfahren analog eingesetzt werden.

5.3.2.3 Abstimmung von CAP9 und CAP11

Im Rahmen von CAP9 wurde zur Erstellung einer gemeinsamen Suchoberfläche für mathematische und physikalische Dokumente eine Konkordanz zwischen MSC und PACS benötigt, um bei der Suche nach Klassifikationen den vollen Dokumentraum erschließen zu können. Ziel war die Verbindung der Dienste MPRESS (http://MathNet.preprints.org/) und PhysDoc (http://physnet.uni-oldenburg.de/PhysNet/physdoc.html). Dem Nutzer sollen bei einer Suche mit Hilfe der mathematischen Klassifikation auch entsprechende Klassifikationsterme im Bereich Physik angeboten werden und umgekehrt.

Geplant war die Nutzung von Termerweiterungsregeln, die die von CAP12 bzw. von CAP9 (nach dem Muster von CAP11) selbst erzeugten Crosskonkordanzen und. Transferbeziehungen benutzen.

Die von CAP12 gelieferte Crosskonkordanz ist sehr grob; sie orientiert sich nur an den Top-Level-Kategorien der Klassifikationen. Dies ist für eine Suche nicht ausreichend differenziert. Zur Erstellung einer statistisch generierten Crosskonkordanz wurden vom FIZ Karlsruhe Datensätze zur Verfügung gestellt, die von CAP9 aufbereitet wurden. Der so entstandene kompakte Parallelkorpus wurde mit Hilfe des Werkzeugs Jester aus CAP11 für die Generierung von Transferrelationen genutzt. Hierzu wurde Installationshilfe und inhaltliche Beratung von CAP11 geleistet.

Diese Transfers liefern im Gegensatz zu der Crosskonkordanz aus CAP12 sehr viele Verknüpfungen zwischen den Klassifikationssystemen, die dadurch begründet sind, dass die Mathematik in der Physik oft als Anwendung genutzt wird. Dies hat zur Folge, dass die Klassifikationen in vielen Bereichen quer zueinander liegen. Aus diesen Gründen wurde die Nutzung von Termerweiterungsregeln zur Retrievalzeit von CAP9 abgelehnt. Alternativ werden die Transferrelationen genutzt, um den Endnutzer in seiner Suchstrategie zu beraten. Sucht er nach einer bestimmten MSC Klassifikation, so werden ihm die Treffer aus den beiden Datenpools geliefert, zusätzlich bekommt er aber noch den Hinweis, dass es sinnvoll sein könnte, nach bestimmten PACS Klassifikationen zu suchen.

Abbildung 9: Ergebnis der parallelen Suche in MPRESS und PhysDoc (CAP9)

Darüber hinaus entstand in CAP9 der Wunsch, eine statistisch erstellte Crosskonkordanz zwischen MSC Klassifikation und Freitexttermen zu bekommen. Ein Testkorpus wurde zu diesem Zweck in Zusammenarbeit mit CAP9 aus dem Datenbestand von MPRESS generiert. Zur Reduktion der Volltextphrasen auf ihre Grundformen wurden Teile der Software KEA genutzt (weiteres s. Kap. 5.3.2.2).

5.3.3 Ergebnisse/Tests

5.3.3.1 Sozialwissenschaften

5.3.3.1.1 Fragestellung und Vorgehensweise

Die Effizienz der statistisch erstellten Transferbeziehungen wurde in einem Benutzungstest untersucht. Gegenstand dieses Tests war die Frage, ob die Einbeziehung zusätzlicher Transferbegriffe in die Freitextrecherche einen Zuwachs an relevanten Dokumenten erbringt. Grundlage des Tests waren Suchbegriffe zu den Carmen-Domänen Frauenforschung, Industriesoziologie und Migration. Die Recherche wurde von einem mit Benutzungstests vertrauten Sozialwissenschaftler des IZ Sozialwissenshaften, Bonn mit FULCRUM (Vektorraum-Modell) durchgeführt.

In einem ersten Schritt wurde eine einfache Suche mit einem Ausgangssuchbegriff (z.B. Dominanz) aus der domänspezifischen Termliste (Ausgangsterme der statistisch erstellten Term-Term-Matrix) durchgeführt. Die Ergebnisliste enthält Dateinamen von Dokumenten ausschließlich aus dem Carmen-Corpus. Der Suchbegriff wurde nicht trunciert. Die gefundenen Treffer wurden ex definitione als „relevant“ klassifiiert, da sie zumindest einmal den Suchbegriff (in der Stammform) im Text enthalten.[19]

In einem zweiten Schritt wurde die Suche um die in der Term-Term-Matrix enthaltenen Transferbegriffe (nicht trunciert) erweitert. Diese erweiterte Suche enthielt in jedem Fall auch den Ausgangssuchbegriff (vgl. Kap. 5.4.1). Die Ergebnisliste wurde auf alle (im Vergleich zur einfachen Suche) zusätzlich angezeigten Dokumentnamen durchgesehen. Ein cut-off konnte nicht definiert werden, da sich das Ergebnis-Ranking mit FULCRUM als instabil erwies d.h. bei der Testwiederholung eine veränderte Reihenfolge der Ergebnisse zeigte. Alle bei der erweiterten Suche zusätzlich gefundenen Dokumente wurden in jedem Einzelfall aufgerufen und auf ihre inhaltliche Relevanz überprüft. Ein Dokument wurde als relevant eingestuft, wenn es inhaltlich einen Zusammenhang mit demjenigen Sachverhalt aufweist, der mit dem Ausgangssuchbegriff angesprochen ist. Beispielsweise wurde bei der erweiterten Suche mit dem Ausgangssuchbegriff „Dominanz“ ein Dokument als relevant eingestuft, wenn der Text zumindest einmal das Wort „Herrschaft“ (auch als

[19] In wenigen Ausnahmefällen erbringt diese Suche im Freitext allerdings auch nicht-relevante Treffer: Bei der Suche z.B. nach „Schweiz“ wird ein Dokument angezeigt aufgrund der Textpassage „...aber ohne Berücksichtigung der Schweiz“.

„Männerherrschaft“) enthält. Die jeweils zur Relevanzbewertung in diesem Sinn herangezogenen Begriffe (semantisches Feld) sind bei der Ergebnisdarstellung der einzelnen Anfragen angegeben. Jedes zusätzlich gefundene Dokument wurde auf den oder die im Text enthaltenen Transferbegriff(e) untersucht, um Aussagen über ertragreiche und weniger ertragreiche Transferbegriffe treffen zu können (Welche Transferbegriffe führen zu zusätzlichen relevanten Dokumenten?). Schließlich wurde für jeden Transferbegriff die Häufigkeit seines Vorkommens in der gesamten Domäne ermittelt.

Die Ergebnisdarstellung enthält für 6 Recherchen (jeweils 2 Recherchen aus jeder der 3 Domänen) folgende Informationen:

bei der einfachen Suche

- Zahl der gefundenen relevanten Treffer

bei der erweiterten Suche:

- Zahl der Treffer insgesamt (ohne Dubletten)
- Zahl der zusätzlichen Treffer im Vergleich zur einfachen Suche (ohne Dubletten)
- Zahl der relevanten Treffer (alle Treffer aus der einfachen Suche und zusätzliche als relevant eingestufte Treffer aus der erweiterten Suche)
- Zahl der zusätzlichen relevanten Treffer (relevante Treffer aus der erweiterten Suche, die in der einfachen Suche nicht enthalten waren)
- Anteil der zusätzlichen relevanten Treffer an den zusätzlichen Treffern (Prozentanteil der in der erweiterten Suche zusätzlich gefundenen relevanten Treffer an der Gesamtzahl der in der erweiterten Suche gefundenen Treffer)
- Aufzählung der bei der Relevanzprüfung eingesetzten (intellektuell vergebenen) Suchbegriffe

bei den Transferbegriffen:

- Nennung der Transferbegriffe (damit: Zahl der Transferbegriffe pro Ausgangsbegriff)
- Häufigkeit des Vorkommens in der jeweiligen Domäne (ohne Dubletten)
- Zahl der durch jeden Transferbegriff (einzeln oder in Kombination mit anderen Transferbegriffen) erzielten zusätzlichen Treffer
- Zahl der durch jeden Transferbegriff (einzeln oder in Kombination mit anderen Transferbegriffen) erzielten zusätzlichen relevanten Treffer

5.3.3.1.2 Ergebnisse aus der Domäne Frauenforschung

5.3.3.1.2.1 Erstes Thema: Dominanz

Einfache Suche	
Suchbegriff	Dominanz
Zahl der relevanten Treffer	16

Erweiterte Suche	
Transferbegriffe	Dominanz, Messen, Mongolei, Nichtregierungsorganisation, Flugzeug, Datenaustausch, Kommunikationsraum, Kommunikationstechnologie, Medienpädagogik, Wüste
Zahl der Treffer (bereinigt)	30
Zahl der zusätzlichen Treffer (bereinigt)	14
Zahl der relevanten Treffer insges.	23
Zahl der zusätzlichen relev. Treffer	7
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	50%

Semantisches Feld: dominant, (Vor-)Herrschaft, herrschen, Unterordnung, Überordnung, Unterdrückung, Abhängigkeit, Unabhängigkeit, „nur Männer“, Patriarchat, Matriarchat, patriarchalisch, matriarchalisch

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Dominanz	16	-	-
Messen	6	5 (4E, 1K)	3 (2E, 1K)
Medienpädagogik	6	4 (4E)	0
Wüste	5	4 (3E, 1K)	3 (2E, 1K)
Kommunikationstechnololgie	3	1 (1E)	1 (1E)
Mongolei	2	1 (1K)	1 (1K)
Nichtregierungsorganisation	2	1 (1E)	1 (1E)
Flugzeug	1	0	0
Datenaustausch	1	0	0
Kommunikationsraum	0	0	0

Die 9 zusätzlichen Suchbegriffe führen zu 14 weiteren Treffern, von denen 7 relevant sind. Bezogen auf die Ausgangsmenge von 16 relevanten Treffern, die mit dem Suchbegriff „Dominanz“ gefunden wurden, beträgt der Zugewinn 44%.

Der Anteil der relevanten Treffer an den neugefundenen weiteren Treffern beträgt 50% (7 von 14).

5 der 9 zusätzlichen Transferbegriffe führen zu den 7 neugefundenen relevanten Treffern, einer davon (Mongolei) allerdings nur in Kombination mit anderen Transferbegriffen. 2 Transferbegriffe (Messen und Wüste) erbringen 5 der 7 relevanten neuen Treffer. Beide Begriffe gehören zu denjenigen 3 Transferbegriffen, die in der Domäne mindestens 5mal vorkommen. Als unergiebig erweist sich der ebenfalls häufiger vorkommende Transferbegriff „Medienpädagogik“.

Die Einbeziehung der 9 zusätzlichen Suchbegriffe erbringt ein deutlich besseres Suchergebnis als die Suche mit „Dominanz“ alleine. Allerdings hätte eine Suche mit „Dominanz Messen Wüste“ ausgereicht, um 21 der 23 relevanten Treffer zu finden. Einige der unergiebigen Suchbegriffe kommen in der Domäne gar nicht oder nur ein- oder zweimal vor und hatten damit auch nur eine ganz geringe Chance, zu einem relevanten Treffer zu führen. Aus dieser Gruppe erbringt nur „Nichtregierungsorganisation“ einen relevanten Treffer.

Durch die Einbeziehung der zusätzlichen Suchbegriffe werden 7 nichtrelevante Treffer gefunden. Die precision der Recherche liegt bei 77% (23 relevante von 30).

5.3.3.1.2.2 Zweites Thema: Frauenhaus

Einfache Suche	
Suchbegriff	Frauenhaus
Zahl der relevanten Treffer (bereinigt)	5

Erweiterte Suche	
Transferbegriffe	Frauenhaus, Artefakt, Sponsoring, Qualitätssicherung, Kärnten
Zahl der Treffer	17
Zahl der zusätzlichen Treffer (bereinigt)	12
Zahl der relevanten Treffer insges.	9
Zahl der zusätzlichen relev. Treffer	4
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	33%

Semantisches Feld: Frauenhäuser, Frauenhausaufenthalt, Mädchenhaus, „Gewalt und Frau“, Gewalterfahrung, Gewalttätigkeit, „Opfer und Frau“, „Opferhilfe und Frau“, Frauenmißhandlung, „Mißhandlung und Frau“, Frauennotwohnung, violence

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Frauenhaus	5	-	-
Qualitätssicherung	8	6 (5E, 1K)	1 (1E)
Sponsoring	7	5 (5E)	3 (3E)
Kärnten	4	2 (1E, 1K)	0
Artefakt	2	0	0

Die 4 zusätzlichen Suchbegriffe führen zu 12 weiteren Treffern, von denen jedoch lediglich 4 relevant sind. Bezogen auf die Ausgangsmenge von 5 relevanten Treffern, die mit dem Suchbegriff „Frauenhaus“ gefunden wurden, beträgt der Zugewinn allerdings 80%.

Der Anteil der nicht-relevanten Treffer an den neugefundenen weiteren Treffern beträgt 67% (8 von 12).

2 der 4 zusätzlichen Transferbegriffe führen zu den 4 neugefundenen relevanten Treffern. Beide Begriffe kommen in der Domäne relativ häufig (8 bzw. 7mal) vor. Als unergiebig erweisen sich die seltener vorkommenden Transferbegriffe „Kärnten“ und „Artefakt“.

Die Einbeziehung der 4 zusätzlichen Suchbegriffe erbringt eine (relativ) deutlich bessere Zahl relevanter Treffer als die Suche mit „Frauenhaus“ alleine. Allerdings hätte eine Suche mit „Frauenhaus Qualitätssicherung Sponsoring“ ausgereicht, um alle 9 relevanten Treffer zu finden. Durch die Einbeziehung der zusätzlichen Suchbegriffe werden 8 nicht-relevante Treffer gefunden. Die precision der Recherche ist mit 53% (9 relevante von 17) relativ niedrig.

5.3.3.1.3 Ergebnisse aus der Domäne Industriesoziologie

5.3.3.1.3.1 Erstes Thema: Bewerbung (im Sinne von Stellenbewerbung)

Einfache Suche	
Suchbegriff	Bewerbung
Zahl der relevanten Treffer	7

Erweiterte Suche	
Transferbegriffe	Bewerbung, Sprachkompetenz, Randgruppe, Validität
Zahl der Treffer	18
Zahl der zusätzlichen Treffer	11
Zahl der relevanten Treffer insges.	13
Zahl der zusätzlichen relev. Treffer	6
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	55%

Semantisches Feld: Bewerber, Bewerbungsgespräch, Vorstellungsgespräch, Bewerbungsbogen, Bewerbungsschreiben, Bewerbungsunterlagen, Spontanbewerbung, Initiativbewerbung, Stellenbewerbung, Stellenbesetzung, Stellenausschreibung, Personalauswahl(-gespräch), Personaleinstellung, Personalbeschaffung, Auswahlverfahren, „assessment center“, Anforderungsprofil

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Bewerbung	7	-	-
Validität	10	8 (8E)	6 (6E)
Randgruppe	4	3 (3E)	0
Sprachkompetenz	1	0	0

Die 3 zusätzlichen Suchbegriffe führen zu 11 weiteren Treffern, von denen 6 relevant sind. Bezogen auf die Ausgangsmenge von 7 relevanten Treffern, die mit dem Suchbegriff „Bewerbung“ gefunden wurden, beträgt der Zugewinn 86%.

Der Anteil der relevanten Treffer an den neugefundenen weiteren Treffern beträgt 55% (6 von 11).

Lediglich einer (Validität) der 3 zusätzlichen Transferbegriffe führt zu den 6 neugefundenen relevanten Treffern. Der Begriff kommt in der Domäne relativ häufig (10 mal) vor. Als unergiebig erweisen sich die beiden anderen seltener vorkommenden Transferbegriffe „Randgruppe“ und „Sprachkompetenz“.

Die Einbeziehung der 3 zusätzlichen Suchbegriffe erbringt eine (relativ) deutlich bessere Zahl relevanter Treffer als die Suche mit „Bewerbung“ alleine. Allerdings hätte eine Suche mit „Bewerbung Validität“ ausgereicht, um alle 13 relevanten Treffer zu finden. Durch die Einbeziehung der zusätzlichen Suchbegriffe werden 5 nicht-relevante Treffer gefunden. Die precision liegt bei 72% (13 relevante von 18).

Eine Besonderheit dieser Recherche besteht darin, dass bei 5 der 6 neugefundenen relevanten Treffer die Wörter „Bewerbungen“, „Bewerbungsunterlagen“, „Bewerbungskosten“ und „Bewerbungsschreiben“ im Text vorkommen. Hätte man die einfache Suche mit dem rechts-truncierten Suchbegriff „Bewerbung%“ durchgeführt, wären diese 5 relevanten Treffer auch bei der einfachen Suche gefunden worden. Es wäre zu untersuchen, unter welchen Bedingungen eine traditionelle Freitextsuche mit truncierten Suchbegriffen die Einbeziehung von Transferbegriffen ergänzen kann.

5.3.3.1.3.2 Zweites Thema: Leiharbeit

Einfache Suche	
Suchbegriff	Leiharbeit
Zahl der relevanten Treffer	10

Erweiterte Suche	
Transferbegriffe	Leiharbeit, Arbeitsphysiologie, Organisationsmodell, Risikoabschätzung
Zahl der Treffer	20
Zahl der zusätzlichen Treffer	10
Zahl der relevanten Treffer insges.	12
Zahl der zusätzlichen relev. Treffer	2
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	20%

Semantisches Feld: Arbeitnehmerüberlassung, Arbeitskräfteüberlassung, Leiharbeitnehmer, Randbelegschaft, Leiharbeiter/in, Leiharbeitsplatz, Leiharbeitsmarkt, Leiharbeitsverhältnis, Leihfirma, Arbeitsflexibilität, „Arbeitsmarkt und Flexibilität“

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Leiharbeit	10	-	-
Organisationsmodell	9	9 (9E)	2 (2E)
Risikoabschätzung (bereinigt)	2	1 (1E)	0
Arbeitsphysiologie	1	0	0

Die 3 zusätzlichen Suchbegriffe führen zu 10 weiteren Treffern, von denen lediglich 2 relevant sind. Bezogen auf die Ausgangsmenge von 10 relevanten

Treffern, die mit dem Suchbegriff „Leiharbeit“ gefunden wurden, beträgt der Zugewinn nur 20%.

Der Anteil der relevanten Treffer an den neugefundenen weiteren Treffern beträgt 20% (2 von 10).

Lediglich einer (Organisationsmodell) der 3 zusätzlichen Transferbegriffe führt zu den beiden neugefundenen relevanten Treffern. Der Begriff kommt in der Domäne relativ häufig (9 mal) vor. Als unergiebig erweisen sich die beiden anderen seltener vorkommenden Transferbegriffe „Risikoabschätzung“ und „Arbeitsphysiologie“.

Die Einbeziehung der 3 zusätzlichen Suchbegriffe erbringt eine nur geringfügig bessere Zahl relevanter Treffer als die Suche mit „Leiharbeit“ alleine. Eine Suche mit „Leiharbeit Organisationsmodell“ hätte ausgereicht, um alle 12 relevanten Treffer zu finden. Durch die Einbeziehung der zusätzlichen Suchbegriffe werden 8 nicht-relevante Treffer gefunden. 7 davon gehen auf das Konto des Suchbegriffs „Organisationsmodell“ Die precision der Recherche liegt bei 60% (12 relevante von 20).

5.3.3.1.4 Ergebnisse aus der Domäne Migration

5.3.3.1.4.1 Erstes Thema: Fremdbild (im Sinne von „soziales Stereotyp“)

Einfache Suche	
Suchbegriff	Fremdbild
Zahl der relevanten Treffer	0

Erweiterte Suche	
Transferbegriffe	Fremdbild, Preisverleihung, Nutzer, Diskurs, Messung, Operationalisierung, Wissenstransfer, Gegenwart, Infrastruktureinrichtung, Gestaltung, Informationsangebot
Zahl der Treffer (bereinigt)	16
Zahl der zusätzlichen Treffer (bereinigt)	16
Zahl der relevanten Treffer insges.	2
Zahl der zusätzlichen relev. Treffer	2
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	13%

Semantisches Feld: Stereotyp, Stereotypisierung, Heterostereotyp, Klischee, Image, Fehlwahrnehmung, Vorurteil, Rollenbild, Fremdinszenierung, Feindbild, „Fremdenfeindlichkeit und Wahrnehmung“, „Ausländerfeindlichkeit und Wahrnehmung“

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Fremdbild	0	-	-
Gestaltung	8	8 (7E, 1K)	0
Diskurs	4	4 (1E, 3K)	1 (1K)
Messung	3	3 (2E,1K)	1 (1E)
Wissenstransfer	3	1 (1E, 2K)	1 (1K)
Operationalisierung	2	2 (1E, 1K)	0
Gegenwart	2	2(1E, 1K))	0
Nutzer	1	1 (1K)	0
Preisverleihung	1	1 (1K)	0
Informationsangebot	1	1 (1K)	0
Infrastruktureinrichtung	1	1 (1K)	0

Im Vergleich zu den anderen 5 Recherchen weist diese Recherche eine Besonderheit auf: Der Suchbegriff „Fremdbild“ kommt in der Domäne „Migration“ überhaupt nicht vor und führt bei der einfachen Suche deshalb zu dem Suchergebnis Null. Die 10 zusätzlichen Suchbegriffe führen dann zu 16 Treffern, von denen jedoch lediglich 2 relevant sind.

Der Anteil der relevanten Treffer an den neugefundenen Treffern beträgt 13% (2 von 16). Die Transferbegriffe sind weitgehend unergiebig.

3 Transferbegriffe führen (in einem Fall in Kombination) zu den beiden neugefundenen relevanten Treffern. Die 3 Begriffe kommen in der Domäne jeweils eher selten (4 bzw. 3mal) vor. Als unergiebig erweisen sich die anderen selten vorkommenden Transferbegriffe, aber auch der relativ häufig vorkommende Begriff „Gestaltung“

Die Einbeziehung der 10 zusätzlichen Suchbegriffe erbringt nur eine kleine Zahl relevanter Treffer. Auch wenn man sagen kann „2 Treffer ist besser als nichts“, erscheint die Einbeziehung der Transferbegriffe für den Nutzer nicht besonders hilfreich. Durch die zusätzlichen Suchbegriffe werden 14 nicht-

relevante Treffer gefunden. Die precision der Recherche ist mit 13% (2 relevante von 16) sehr schlecht.

Dieses unbefriedigende Ergebnis hängt wahrscheinlich mit einer Besonderheit der Testdomäne „Migration“ zusammen. Bei mehreren Testrecherchen zeigte sich, dass ein erheblicher Teil der gefundenen Dokumente aus ein- und demselben Institut (efms) stammen. Es handelt sich dabei vorwiegend um monatlich erscheinende Presseschauen zum Thema „Migration, Flüchtlinge, Asyl“. Bei einem anderen Teil der Dokumente handelt es sich um Selbstdarstellungen des Instituts, Berufsbiographien der Mitarbeiter, Institutssatzungen und dgl. In diesen Dokumenten kommen die eher (sozial-)psychologischen Fragestellungen, die mit „Fremdbild“ und „Stereotyp“ zusammenhängen, nicht vor. Das Institut beschäftigt sich offenbar eher mit politischen und rechtlichen Fragestellungen.

5.3.3.1.4.2 Zweites Thema: Ausländerrecht

Einfache Suche	
Suchbegriff	Ausländerrecht
Zahl der relevanten Treffer	12

Erweiterte Suche	
Transferbegriffe	Ausländerrecht, Rechtswissenschaft, Forschungsschwerpunkt, Politikwissenschaft, Angestellte
Zahl der Treffer	20
Zahl der zusätzlichen Treffer	8
Zahl der relevanten Treffer insges.	14
Zahl der zusätzlichen relev. Treffer	2
Anteil der zusätzlichen relev. Treffer an den zusätzlichen Treffern	25%

Semantisches Feld: Ausländergesetz, Ausländerpolitik, Einwanderung(-srecht), Aufenthaltsrecht, Migration(-spolitik), Asylrecht, Asylpolitik, Asylverfahren, „sichere Drittstaaten“, „sicherer Herkunftsstaat“, Flüchtling(-srecht), Einbürgerung, „Zuzug und Aussiedler“.

Analyse Transferbegriffe	*Vorkommen in der Domäne*	*Zusätzliche Treffer Einzeln / Kombination*	*davon: Zusätzliche relevante Treffer*
Ausländerrecht	12	-	-
Angestellte	6	5 (4E, 1K)	0
Politikwissenschaft	3	2 (1E, 1K)	0
Rechtswissenschaft	2	1 (1E)	1 (1E)
Forschungsschwerpunkt	1	1 (1E)	1 (1E)

Die 4 zusätzlichen Suchbegriffe führen zu 8 weiteren Treffern, von denen lediglich 2 relevant sind. Bezogen auf die Ausgangsmenge von 12 relevanten Treffern, die mit dem Suchbegriff „Ausländerrecht“ gefunden wurden, beträgt der Zugewinn nur 17%.

Der Anteil der relevanten Treffer an den neugefundenen weiteren Treffern beträgt 25% (2 von 8).

Die beiden Transferbegriffe „Rechtswissenschaft“ und „Forschungsschwerpunkt“ führen mit jeweils einem Treffer zu den neugefundenen relevanten Treffern. Beide Begriffe kommen in der Domäne selten (2 bzw. 1mal) vor. Als unergiebig erweisen sich die beiden anderen Transferbegriffe „Angestellte“ und „Politikwissenschaft“. „Angestellte“ kommt immerhin 6 mal in der Domäne vor, erweist sich aber als unergiebig, weil bei den angezeigten Dokumenten Angestellte eines Instituts vorgestellt werden, die sich überwiegend nicht mit Ausländerrecht beschäftigen.

Die Einbeziehung der 4 zusätzlichen Suchbegriffe erbringt eine nur geringfügig bessere Zahl relevanter Treffer als die Suche mit „Ausländerrecht“ alleine. Durch die Einbeziehung der zusätzlichen Suchbegriffe werden 6 nichtrelevante Treffer gefunden. Die precision der Recherche liegt bei 70% (14 relevante von 20).

Auch hier muss auf die im vorherigen Beispiel erwähnte Besonderheit der Testdomäne „Migration“ (sehr viele Dokumente aus einem einzigen Institut) hingewiesen werden, die eine Interpretation der Ergebnisse im Hinblick auf die Ergiebigkeit der Transferbegriffe sehr erschwert.

5.3.3.1.5 Zusammenfassung und Schlussfolgerungen

Die Frage, ob die Einbeziehung zusätzlicher Transferbegriffe in die Freitextrecherche einen Zuwachs an relevanten Dokumenten erbringt, kann positiv beantwortet werden: die erweiterte Suche enthält in allen 6 Beispielfällen mehr relevante Treffer als die einfache Suche.

Allerdings sind die absoluten Zahlen teilweise sehr gering: in 3 der 6 Beispielfälle wurden lediglich 2 zusätzliche relevante Treffer gefunden. Da aus methodischen Gründen für die Beispielrecherchen kein Anker (Zahl aller relevanten Dokumente pro Anfrage in der Testdatenbank) ermittelt werden konnte, kann nicht entschieden werden, ob beispielsweise 2 zusätzlich gefundene Treffer einen erheblichen oder einen unerheblichen Zuwachs darstellen.

Die Qualität der Suchergebnisse soll statt dessen an dem Anteil der durch die Transferbegriffe zusätzlich gefundenen relevanten Dokumente an der Gesamtzahl der durch die Transferbegriffe zusätzlich gefundenen Dokumente gemessen werden. Dieser Wert ist aus Nutzersicht in der vorliegenden Recherchekonstellation von erheblicher Bedeutung, weil die Ergebnisliste auf sehr unterschiedliche und auch sehr unterschiedlich umfangreiche Internetdokumente verweist, die der Nutzer jeweils auf ihre Relevanz für seine Fragestellung beurteilen muss.

Betrachtet man die 6 Recherchen unter diesem Gesichtspunkt, fällt auf, dass die Ergebnisse sehr unterschiedlich sind: der Anteil der zusätzlich gefundenen relevanten Treffer an der Gesamtzahl der zusätzlich gefundenen Treffer liegt bei den Recherchen „Bewerbung“ und „Dominanz“ mit 55% bzw. 50% am höchsten. In beiden Fällen bekommt der Nutzer eine nennenswerte Zahl zusätzlicher Dokumente angeboten (11 bzw. 14), deren Durchsicht sich lohnt, da ca. die Hälfte relevant sind. Andererseits ist die Zahl der angebotenen Dokumente auch nicht zu groß, so dass die Relevanzbeurteilung mit einem vertretbaren Zeitaufwand erfolgreich durchgeführt werden kann. Die Einzelanalyse beider Recherchen zeigt, dass nicht die absolute Zahl der zusätzlichen Transferbegriffe, sondern die Häufigkeit ihres Vorkommens in der Domäne von entscheidender Bedeutung ist. Wie die Analyse der anderen Recherchen zeigt, ist die Häufigkeit aber wohl eine notwendige, aber keineswegs eine hinreichende Bedingung für die Ertragskraft eines Transferbegriffs.

Am anderen Ende der Qualitätsskala liegen die 3 Recherchen mit einem Anteil der zusätzlich gefundenen relevanten Treffer an der Gesamtzahl der zusätzlich gefundenen Treffer von 13% bis 25%. In allen 3 Fällen werden jeweils 2 zusätzliche relevante Treffer gefunden. Dafür müssen aber 8, 10 oder 16 Dokumente durchgesehen werden. Aus Nutzersicht dürfte dies, vom Ertrag

her betrachtet, eine eher ärgerliche Angelegenheit sein. Die einfache Suche mit dem Ausgangs-Suchbegriff hätte genügt.

Zusammenfassend kann man festhalten, dass 2 der 6 Recherchen mit der durch die Transferbegriffe erweiterten Suche ein positives Resultat erbringen. Eine weitere Recherche („Frauenhaus“) liefert ein zumindest akzeptables Resultat, vor allem im Hinblick auf den (von der einfachen Suche aus gesehen) relativ hohen Zuwachs an relevanten Treffern (von 5 auf 9). Allerdings sind nur 4 der 12 zusätzlichen Treffer relevant. Die restlichen 3 Recherchen erbringen unbefriedigende Ergebnisse. Allerdings müssen bei der Beurteilung der beiden Recherche-Ergebnisse in der Domäne „Migration“ die Restriktionen im Carmen-Testkorpus berücksichtigt werden. In dieser Domäne sollten bei zukünftigen Tests andere bzw. zusätzliche Dokumente aus anderen Internetquellen bereitgestellt werden.

5.4 Transfer-Service

5.4.1 Transfer-Architektur

Folgende Anforderungen wurden an die Architektur gestellt:

- Bestandsspezifische Übersetzung von Anfragen in einer Dokumentationssprache in eine andere
- Nutzung statistischer und intellektueller Transfers
- Einbau von Ressourcen in vorhandenen Datenquellen (JDBC-Datenbank, SIS-TMS)
- Konfigurierbarkeit und Wiederverwendbarkeit

Ziel der Transfers ist die Manipulation einer Anfrage auf einen bestimmten Dokumentbestand hin. Eine Anfrage wird mit einer Dokumentationssprache (z.B. einem Thesaurus) formuliert, die ggf. für einen der abzufragenden Dokumentbestände für die Indexierung benutzt wurde. Für diesen Bestand kann die Anfrage unverändert bleiben. Andere abzufragende Bestände, die diese Dokumentationssprache nicht benutzen, machen eine Veränderung der Anfrage nötig, es findet dabei für den jeweiligen Dokumentbestand eine Übersetzung der Anfrage in eine andere Dokumentationssprache statt, mit der dieser Bestand erschlossen wurde (siehe Abbildung 10).

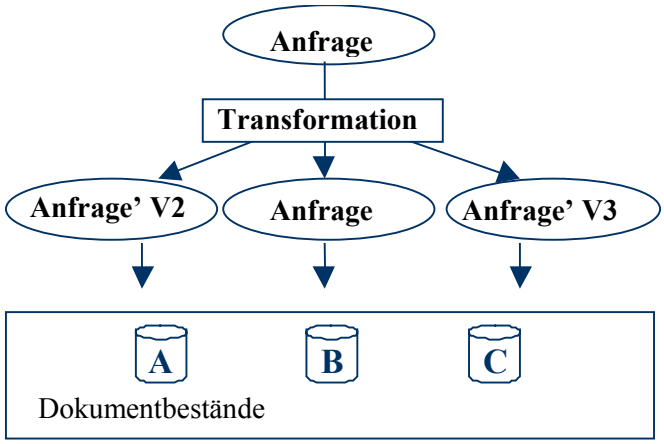

Abbildung 10: Anfrage-Transformation

Die Transfers (oder Übersetzungen) von Anfragen wird durch Transfer-Module erledigt, die in Java implementiert wurden. Für die verschiedenen Arten von Transfers wurden allgemeine abstrakte Klassen (`TransferModule` für die allgemeine Transferlogik, `AtomicTransfer` für die einzelnen Arten von Transfers und Datenspeichern) mit spezifischen Implementierungen bereitgestellt (siehe Abbildung 11).

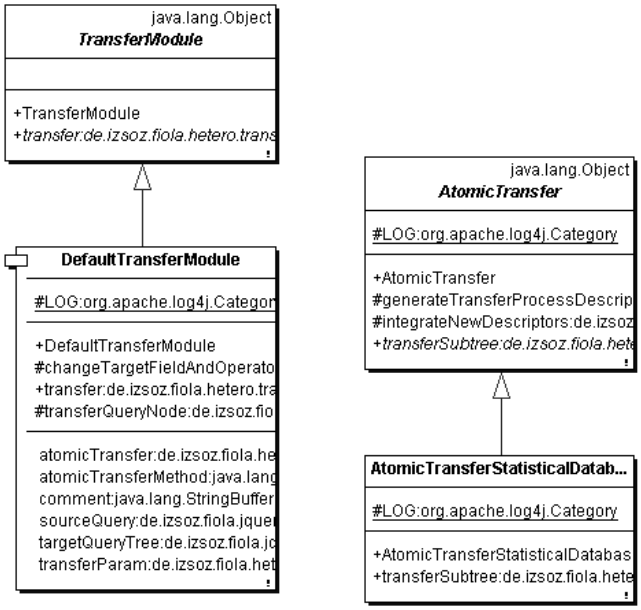

Abbildung 11: Transfer-Module und Atomare Transfers

Die allgemeine Transfer-Logik steuert die Abarbeitung von Transfer-Anfragen. Für Bestände und Dokumentationssprachen werden (möglicherweise mehrere alternative) Transfer-Arten (intellektuell, statistisch) bereitgestellt, denen Prioritäten zugeordnet sind. Führt ein Transfer-Versuch zu kei-

nem Erfolg, wird ein weiterer Versuch mit der nächstniedrigeren Priorität durchgeführt, bis entweder ein erfolgreicher Transfer durchgeführt werden konnte oder kein weiterer Transfer für die vorliegenden Bestände und Dokumentationssprachen vorliegen.

Zu der allgemeinen Transfer-Logik gehört weiterhin das Einmischen der neu gefundenen Deskriptoren in die Anfrage oder die Ersetzung der Ausgangs-Anfrage durch eine ganz neue Anfrage. Erweiternde Anfragen werden genutzt, wenn die Ausgangs-Dokumentationssprache im Zielbestand vorhanden ist, eine Übersetzung in eine andere Dokumentationssprache aber zusätzlich erfolgen soll. Ist die Ziel-Dokumentationssprache unkontrolliert, werden auch die Ausgangs-Deskriptoren in der Ziel-Anfrage beibehalten (Term-Erweiterung), bei einem kontrollierten Vokabular fallen diese weg, wenn sie nicht ausdrücklich durch Transferbeziehungen vorgesehen sind (Term-Ersetzung).

Weiterhin sind evtl. Element-/Feld-Namen und Operatoren für den Zielbestand zu verändern, sofern diese durch die Übersetzung in eine andere Dokumentationssprache betroffen sind.

Schließlich wird für jeden Transfer eine Beschreibung erzeugt, die u.a. die Art des genutzten Transfers und die Zahl der neuen Deskriptoren benennt. Beschreibungstexte werden detailliert jedem einzelnen Knoten im Anfragebaum und zusammenfassend der gesamten Ziel-Anfrage beigefügt.

Die Transfers lassen sich auf verschiedene Weise parametrisieren. Für die einzelnen Arten von Transfers sowie für Transfers im Allgemeinen sind Parameter erforderlich, die (technisch) Feldnamen, Datenspeicher und ähnliches bestimmen und (inhaltlich) auf die Bearbeitung des Transfers Einfluss nehmen. Die Parameter werden für den Transfer in einer Hierarchie spezieller Parameter-Klassen repräsentiert (siehe Abbildung 12). Die noch im Zwischenbericht beschriebene Klassenhierarchie einer Vielzahl spezialisierter Transfer-Klassen wurde zugunsten einer geringeren Zahl flexibler Transfer-Klassen aufgegeben, die über geeignete Parameter angepasst werden können. Der Name der den Transfer ausführenden Klasse (einer Kind-Klasse von `AtomicTransfer`) ist einer der Parameter. Während des Transfers wird diese Klasse über den Java-ClassLoader instanziert.

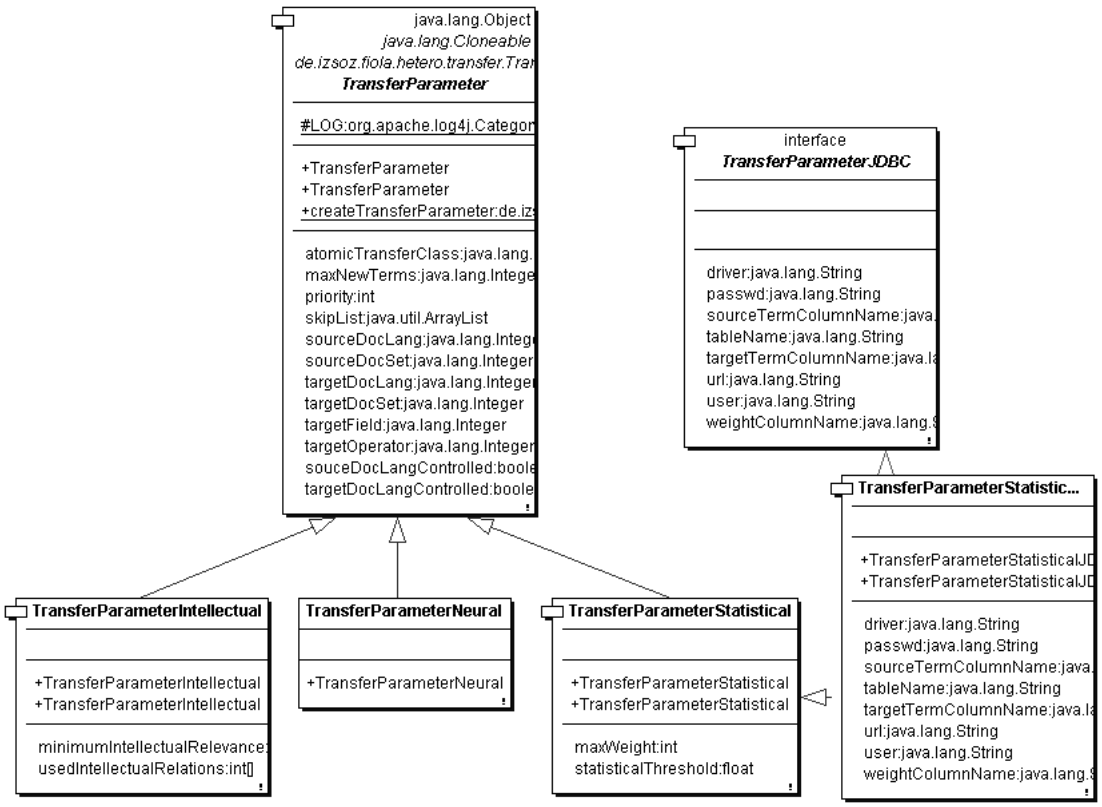

Abbildung 12: Transfer-Parametrisierung

Je nach Art des Transfers sind hier z.B. Schwellenwerte für die Berücksichtigung statistischer Term-Term-Relationen oder Äquivalenzarten und Relevanzen intellektueller Crosskonkordanzen manipulierbar. Für relationale Datenbanken lassen sich die JDBC-Verbindungsdaten und die Tabellen- und Spaltennamen konfigurieren, für SIS-TMS-Datenquellen die entsprechenden Verbindungsparameter. Allen Transfers gemeinsam sind beispielsweise optionale Listen von Deskriptoren, die nicht übersetzt werden sollen und eine Zahl höchstens zu verwendender neuer Ziel-Deskriptoren.

Die Parametrisierung erfolgt mit geeigneten Standard-Werten auf dem Transfer-Server, so dass die Durchführung von Transfers ohne Nutzung (und Kenntnis) der Parameter möglich ist. Der Betreiber eines HyREX-Servers, der die Transfer-Dienste nutzt, kann die voreingestellten inhaltlichen Parameter durch eigene Standard-Vorgaben überschreiben, die technischen Parameter (Datenquellen) sind nicht veränderbar. Schließlich ist vorgesehen, dass End-Benutzer selber auf die Parameter der Transfers Einfluss nehmen können, sofern sie dies wünschen. In einem iterativen Prozess soll es möglich sein, durch Manipulation der Transfer-Parameter die Wirkung des Transfer-Dienstes optimal an die Bedürfnisse der Benutzer anzupassen. Da im Projekt CARMEN

allerdings keine Benutzungsschnittstelle bereitgestellt werden kann, die die Wirkung der Transfers dem Benutzer anschaulich anzeigen und diesem die Möglichkeit der Beeinflussung geben kann, ist diese Art der Parametrisierung derzeit nicht möglich.

Die serverseitige Parametrisierung der Transfer-Dienste erfolgt über eine flexible XML-Konfiguration. Über eine DTD ist definiert, welche Parameter auf welche Art gesetzt werden können (siehe Abbildung 13).

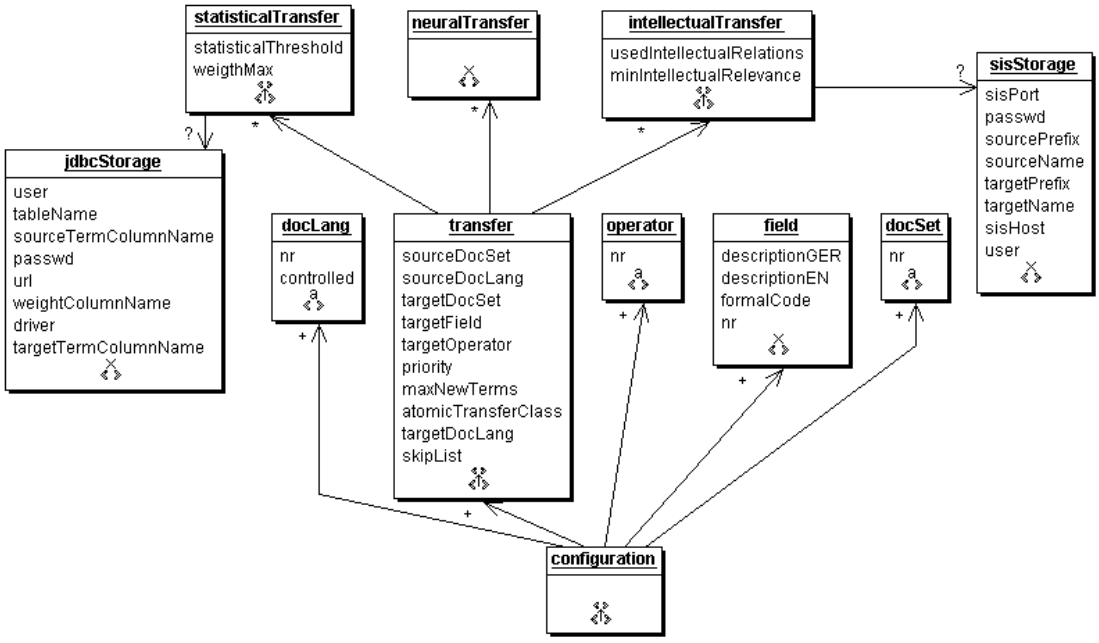

Abbildung 13: DTD für Transfer-Konfiguration

Die Konfiguration ermöglicht eine einfache Anpassung der Transfers sowie die Setzung von Standard-Werten. Zusätzliche Transfers zwischen neuen Dokumentbeständen und Dokumentationssprachen können bereitgestellt werden, ohne dass dafür der Quelltext des Transfer-Servers geändert werden müsste. Ein neue Eintrag in der XML-Konfigurationsdatei bzw. das Hinzufügen eines neuen `transfer`-Elements reicht aus, um auf neue Transferbeziehungen in einer vorhandenen Datenquelle zugreifen zu können. Abbildung 14 zeigt einen Ausschnitt einer Beispiel-Konfigurationsdatei, in der Dokumentbestände, Dokumentationssprachen, Operatoren und Feldnamen sowie Transfers konfiguriert werden. In diesem Beispiel wird auf eine Crosskonkordanz zwischen dem IZ-Thesaurus und der Schlagwortnormdatei zugegriffen, die in einer SIS-TMS-Datenbank vorgehalten wird.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE configuration (View Source for full doctype...)>
- <configuration>
    <docSet nr="1">MathNet</docSet>
    <docSet nr="2">PhysNet</docSet>
    <docSet nr="3">SOLIS</docSet>
    <docSet nr="4">FORIS</docSet>
    <docSet nr="6">Die Deutsche Bibliothek</docSet>
    <docSet nr="5">SoWi Internetquellen</docSet>
    <docLang nr="0" controlled="false">Freitext</docLang>
    <docLang nr="1" controlled="true">MSC</docLang>
    <docLang nr="2" controlled="true">PACS</docLang>
    <docLang nr="3" controlled="true">IZ Klassifikation</docLang>
    <docLang nr="10" controlled="true">IZ Thesaurus</docLang>
    <docLang nr="11" controlled="true">SWD</docLang>
    <operator nr="1">$te$</operator>
    <field descriptionGER="Schlagwort" descriptionEN="Keyword" formalCode="kw" nr="1" />
  - <transfer sourceDocSet="3" sourceDocLang="10" targetDocSet="6" targetDocLang="11" skipList="Test-Term1|Test-Term2|Test-
      Term3" atomicTransferClass="de.izsoz.fiola.hetero.transfer.tms.AtomicTransferTMS">
    - <intellectualTransfer priority="1" minIntellectualRelevance="LOW"
        usedIntellectualRelations="EXACT|NARROWER|BROADER|INEXACT">
        <sisStorage sisHost="repository.bonn.iz-soz.de" sisPort="1201" sourceName="IZ" sourcePrefix="DeTerm`"
          targetName="SWD" targetPrefix="DeTerm`" />
      </intellectualTransfer>
    </transfer>
  - <transfer sourceDocSet="6" sourceDocLang="11" targetDocSet="3" targetDocLang="10" skipList=""
      atomicTransferClass="de.izsoz.fiola.hetero.transfer.tms.AtomicTransferTMS">
    - <intellectualTransfer priority="1" sisHost="soe.bonn.iz-soz.de" sisPort="1203" minIntellectualRelevance="LOW"
        usedIntellectualRelations="EXACT|NARROWER|BROADER|INEXACT">
        <sisStorage sisHost="repository.bonn.iz-soz.de" sisPort="1201" sourceName="SWD" sourcePrefix="DeTerm`"
          targetName="IZ" targetPrefix="DeTerm`" />
      </intellectualTransfer>
```

Abbildung 14: Beispiel für Transfer-Konfiguration (Ausschnitt)

Für die Verarbeitung von Anfragen durch die Transfer-Module wird die Ausgangs-Anfrage in eine interne Anfrage-Repräsentation überführt. Dabei wird auf eine am IZ im Rahmen der Projektgruppe FIOLA (Framework for Interest Oriented Library Architectures) entwickelte Baum-Repräsentation (JQueryTree) zurückgegriffen, die zwischen Vergleichs-Knoten und bool'schen Knoten unterscheidet. Die vorhandene Anfrage-Repräsentation wurde hier durch spezifische Kind-Klassen erweitert, die die besonderen Anforderungen der HyREX-Anfragesprache „Xirql" berücksichtigen und außerdem inhaltliche Transfer-Parameter übermitteln können (siehe Abbildung 15). Während des Transfer-Prozesses werden Anfrage-Knoten durch neue Knoten ersetzt oder ggf. erweitert.

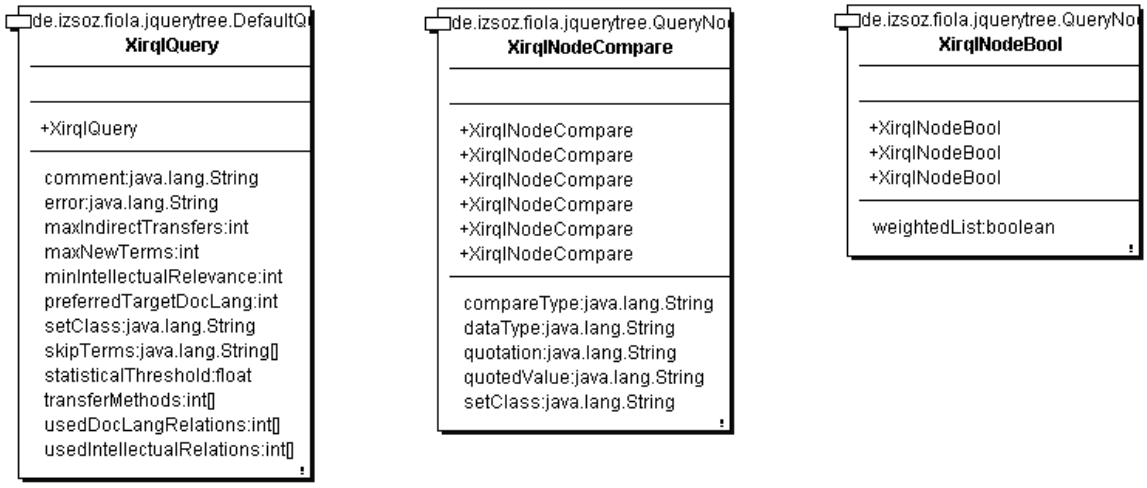

Abbildung 15: Klassen zur Repräsentation von Xirql-Teilanfragen

Ein großer Vorteil der internen Anfrage-Repräsentation ist die einfache Erweiterbarkeit der Transfer-Module. Eine Anfrage wird vor dem Transfer durch einen `QueryParser` in die interne Repräsentation überführt und nach dem Transfer durch einen `QueryPrinter` in die Ausgangs-Anfragesprache zurück übersetzt (siehe Abbildung 16). Um außer Xirql weitere Anfragesprachen als Xirql nutzen zu können, müssen also lediglich entsprechende `QueryParser` und `QueryPrinter` bereitgestellt werden.

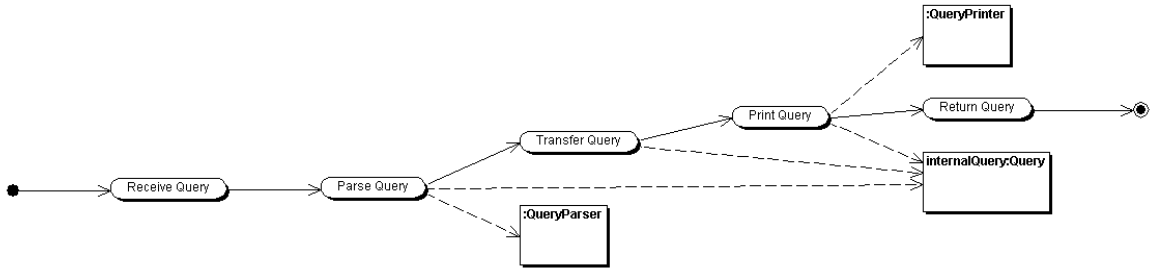

Abbildung 16: Anbindung von Anfrage-Sprachen über QueryParser und QueryPrinter

Als Beispiele für eine solche einfache Erweiterbarkeit wurden bereits mit geringem Aufwand einfache Transfer-Dienste implementiert, die eine Übersetzung von SWD-Deskriptoren in Deskriptoren aus dem Thesaurus Sozialwissenschaften leisten und dabei andere Anfragesprachen nutzen.

5.4.1.1 Anbindung an die Retrievalkomponente

Der Transfer von Anfragen in die Retrievalkomponente HyREX anzubinden. Gemeinsam mit CAP7 wurde eine Lösung entwickelt, die eine Integration ermöglicht. Die Xirql-Anfrage wird von HyREX prozessiert und dafür zunächst auf transferierbare Xirql-Teilanfragen hin analysiert (nicht für alle Felder stehen Transfers zur Verfügung). Diese werden dann per http-Request an den Transfer-Server übermittelt, der die Xirql-Anfrage parst, den eigentlichen Transfer durchführt und die nach Xirql zurückübersetzte Anfrage als http-Response an HyREX zurück übermittelt. Dort wird die transferierte Teilanfrage von HyREX wieder in die Xirql-Gesamtanfrage eingebaut (siehe Abbildung 17). Diese Lösung ist flexibel und vor allem plattformübergreifend (HyREX ist in Perl, die Transfer-Module in Java implementiert) und ermöglich eine verteilte Architektur. HyREX-Server an beliebigen Standorten können so einen zentralen oder auch verschiedene Transfer-Server nutzen. Das Datenvolumen der übertragenen Anfragen ist gering, so dass es zu keiner erheblichen Verzögerung kommt.

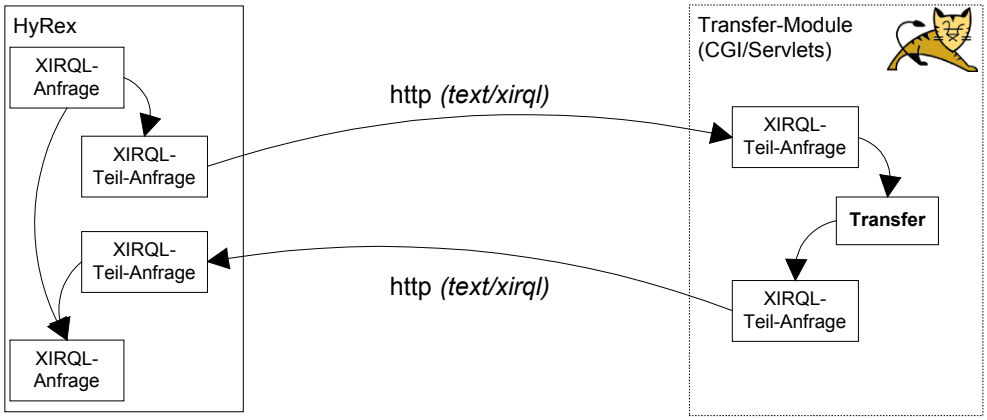

Abbildung 17: Anbindung der Transfer-Module an HyRex über http

Der Transfer-Server läuft in einem Apache-http-Server[20] und dem Java-Servlet-Server Tomcat aus dem Apache-Projekt Jakarta[21]. Es könnten aber auch andere http- und Servlet-Server eingesetzt werden.

Die Anfragen werden per XML übertragen, bei der Anfrage als urlencodeter String im get-Parameter und bei der Antwort als String im http-Body. Zu-

[20] http://httpd.apache.org/

[21] http://jakarta.apache.org/tomcat/

sammen mit CAP7 wurde eine DTD für den Austausch der Anfragen entwickelt, die die für die transferierbaren Anfragen nötigen Ausschnitte aus Xirql abdeckt und alle Transfer-Parameter übermitteln kann. Diese entsprechen weitgehend den in Kapitel 5.4.1 beschriebenen inhaltlichen Parametern und ermöglichen einerseits, dem Betreiber eines HyREX-Servers, eigene Standard-Werte für Parameter zu setzen, und andererseits, von Benutzern über die Benutzungsschnittstelle gesetzte Parameter zu übermitteln.

Darüber hinaus wurde die Möglichkeit geschaffen, neben der zentrale Konfiguration für Feldnamen und Operatoren auch eine dezentrale Konfiguration durch den Server-Betreiber zu erstellen. Dies ermöglicht dem Betreiber, Änderungen an der Datenstruktur der indexierten Daten selbstständig auf den Transfer hin abzubilden, ohne dafür den Betreiber des Transfer-Servers um Anpassungen bitten zu müssen.

Treten während des Transfers Fehler auf (z.B. nicht parsbare Anfrage, Ausfall einer Datenbank), so wird eine http-Fehlermeldung zurückgegeben (Error 304, „Internal Server Error“) und zusätzlich ein XML-Dokument mit einem fehlerbeschreibenden Text. Ein detaillierter und flexibel anpassbarer Logging-Mechanismus, der auf Log4J[22] aufsetzt, ermöglicht ein genaues Nachverfolgen von Fehlern.

Während des Transfers werden den einzelnen Deskriptoren Gewichte zugeordnet, die je nach Art des Transfers unterschiedlich ermittelt werden. Diese Gewichte liegen für die interne Anfrage-Repräsentation zwischen 0 und 100. Bei Xirql addieren sich die Gewichte aller Deskriptoren einer gewichteten (Teil-)Anfrage (`wsum`-Element) auf 1. Die nötige Umrechnung und Anpassung der Deskriptoren-Gewichte für Xirql wird sichergestellt.

Die Anbindung der Retrievalkomponente an den Transfer-Server wurde beispielhaft in den Demo-Systemen von CAP6 implementiert. Dort lassen sich die Bestände „sozialwissenschaftliche Internetquellen“[23] und „GIRT“[24] unter Ausnutzung von Transfers zwischen Freitext-Termen und dem Thesaurus Sozialwissenschaften prozessieren.

---

[22] http://jakarta.apache.org/log4j/

[23] http://ines.mathematik.uni-osnabrueck.de:43212/

[24] http://ines.mathematik.uni-osnabrueck.de:43213/

Einen Sonderfall stellt das Demosystem „Math-Broker & MAJOUR Search“[25] (Elib) dar, bei dem Transferbeziehungen nicht zwischen verschiedenen Beständen, sondern innerhalb eines einzelnen Bestandes, der teilweise mit MSC und teilweise mit PACS erschlossen wurde, genutzt werden. Bei einer Suche nach einer MSC-Notation werden automatisch auch Dokumente mit äquivalenten PACS-Notationen gefunden – und umgekehrt.

Bisher ungelöst ist das Problem des Einbaus transferierter Xirql-Teilanfragen für XML-Dokumente mit sehr komplexer Struktur. Bisher können die veränderten Anfrageknoten in die Gesamtanfrage eingebaut werden, solange die zu durchsuchenden Elemente Nachbarn der ursprünglichen Knoten sind. Dies ist bei allen in CARMEN vorliegenden Beständen der Fall.

5.4.1.2 Einbindung von Transferbeziehungen

Die erstellten Transferbeziehungen zwischen Dokumentationssprachen bzw. zwischen Freitexttermen und einer Dokumentationssprache (siehe Kapitel 5.3.2) wurden in einer Oracle-Datenbank abgespeichert. Die Transfer-Module zur Nutzung dieser Transferbeziehungen setzen auf diese Datenbank auf und sprechen diese über die passenden JDBC-Treiber an. Die Art der JDBC-Datenbank ist in der XML-Konfiguration wählbar und kann, sofern die passenden Treiber vorliegen, durch einen einfachen Eintrag auf andere DBMS umgestellt werden.

In der XML-Konfiguration werden neben den JDBC-Verbindungsdaten auch die Namen der Tabellen und Spalten bezeichnet, so dass sich über die Konfiguration sehr einfach Transferbeziehungen aus weiteren Datenbanken einbinden lassen, ohne dass diese einem vorgegebenen Namensschema folgen müssen.

Für die Ausgangsdeskriptoren werden Transfer-Beziehungen case-insensitive gesucht, so dass die Groß-Klein-Schreibung für die Anfrage keine Rolle spielt.

Es lassen sich Schwellenwerte angegeben, die das Mindestgewicht eines Deskriptors bezeichnen, der für den Transfer genutzt werden soll. Die von Jester ermittelten Gewichte werden normalisiert und den neu ermittelten Deskriptoren beigefügt und später für Xirql in einem `wsum`-Element zusammengefasst.

---

[25] http://ines.mathematik.uni-osnabrueck.de:43211/

Beim Transfer von Freitext-Termen zu einer Dokumentationssprache liegt der Sonderfall vor, dass keine bestimmte Flektion des Ausgangsterms erwartet werden kann wie bei einem kontrollierten Vokabular. Daher werden Freitext-Terme vor der Datenbank-Abfrage linguistisch vorbehandelt. Dafür wird der Analyzer der Java-Suchmaschine Lucene[26] eingesetzt. Für deutschsprachige Terme liegt ein eigener Stemmer vor, der sehr zuverlässig Substantive und Adjektive sowie vertretbar Verben behandelt. Damit werden die Freitext-Terme in die in der Datenbank gespeicherte Wortform gebracht. Für englischsprachige Terme wird ein Porter-Stemmer benutzt, der auch für die Erstellung der statistischen Transfer-Beziehungen eingesetzt wurde und so ebenfalls die in der Datenbank gespeicherte Flektion bereitstellt.

5.4.1.3 Einbindung von Crosskonkordanzen

Im CAP12 wurden intellektuell Crosskonkordanzen erstellt, die von den Transfermodulen im Retrieval eingebunden werden können. Bei der Konzeption der Transfermodule wurde daher frühzeitig auf eine gute Integration der Ergebnisse von CAP12 geachtet. Es bot sich an, UAP14 vorzuziehen und die Anpassungsarbeiten für CAP12 vorzunehmen, da dies gleichzeitig ein effizienteres Arbeiten für CAP12 bedeutete.

Für die Erstellung, Speicherung und Vorhaltung der Crosskonkordanzen aus CAP12 wurde das System SIS-TMS ausgewählt, das Thesauri und die Interthesaurusbeziehungen in einem semantischen Netzwerk speichert. Von dem Produzenten ICS-FORTH wurde ein generisches Thesaurusmodell mitgeliefert, das intensiv geprüft und entsprechend den Anforderungen aus CAP12 an die einzuspeisenden Thesauri und die gewünschten Arten von Interthesaurusbeziehungen angepasst wurde. Außerdem wurde das Modell so angepasst, dass auch Klassifikationen und zugehörige Crosskonkordanzen verarbeitet werden können.

SIS-TMS lässt sich über eine eigene Datenmanipulationssprache („TELOS“) konfigurieren. Das erweiterte SIS-TMS-Modell wurde auf einem SIS-Server im InformationsZentrum Sozialwissenschaften eingespeichert. Über mittels Java erstellte Parser wurden die Thesauri sowie schon in Word-/Excel-Dateien erstellte Konkordanzen ebenfalls eingelesen und auf dem SIS-Server bereitgehalten.

---

[26] http://jakarta.apache.org/lucene/

Durch die enge Zusammenarbeit mit CAP12 ließen sich die Anforderungen bei der Einbindung in die Transfermodule frühzeitig berücksichtigen. Erweitert wurde das Thesaurusmodell vor allem um Gewichte für die Interthesaurusbeziehungen sowie um n:m-Beziehungen zwischen Termen. Durch die Einführung gruppierter Terme lassen sich komplexe Und-oder-Beziehungen zwischen mehreren Termen herstellen.

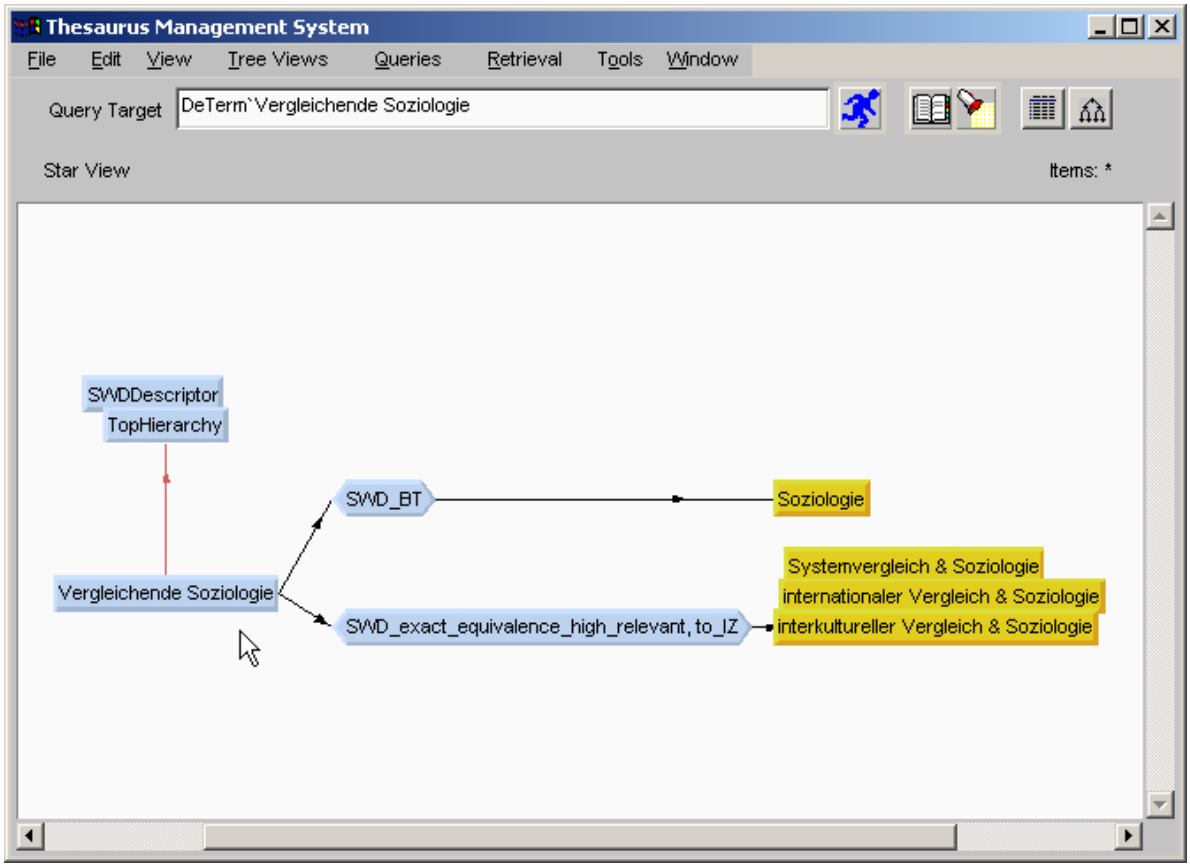

Abbildung 18: Bildschirmkopie SIS-TMS Thesaurus-Browser

Außer dem Zugriff über die TELOS-Schnittstelle zur Daten- und Strukturmanipulation wird SIS-TMS mit einer Java-Schnittstelle ausgeliefert. Mit diesem ausgesprochen unübersichtlichen API (sämtliche Methoden in einer einzelnen Java-Klasse) haben sich die Mitarbeiter vertraut gemacht. Zur besseren Nutzung und zur Vereinfachung der Integration in die Transfermodule wurde dieses API durch Java-Klassen besser strukturiert und für objektorientierte Programmierung nutzbar gemacht (im Paket `de.izsoz.fiola.hetero.transfer.tms`, Ausschnitt aus dem Klassenmodell siehe Abbildung 19). Dabei wurden Klassen erstellt, die Knoten und Kanten im semantischen Netz repräsentieren, außerdem wurden Methoden geschrieben, über die die Kanten bestimmter Art eines Knoten eingelesen werden können, so dass über diese Methoden im semantischen Netz navigiert

werden kann. Insbesondere stehen die Methoden bereit, um die Äquivalenz-Beziehungen zwischen verschiedenen Dokumentationssprachen abzufragen. Parallele Anfragen an die SIS-TMS-Datenbank sind allerdings derzeit über das SIS-TMS-API nicht möglich, was in einem Server mit parallelen Threads Probleme bereitet. Die Abfragen wurden über das Synchronized-Statement teilweise geschützt, aber der Navigation sind so enge Grenzen gesetzt.

Ein atomarer Transfer für die Verarbeitung von Crosskonkordanzen aus SIS-TMS in Transfer-Modulen wurde erstellt, der sich über die XML-Konfiguration (siehe Kapitel 5.4.1) des Transfers einfach einbinden läßt. So wird ein schneller Zugriff auf die in SIS-TMS gehaltenen Konkordanzbeziehungen zum Retrievalzeitpunkt sichergestellt.

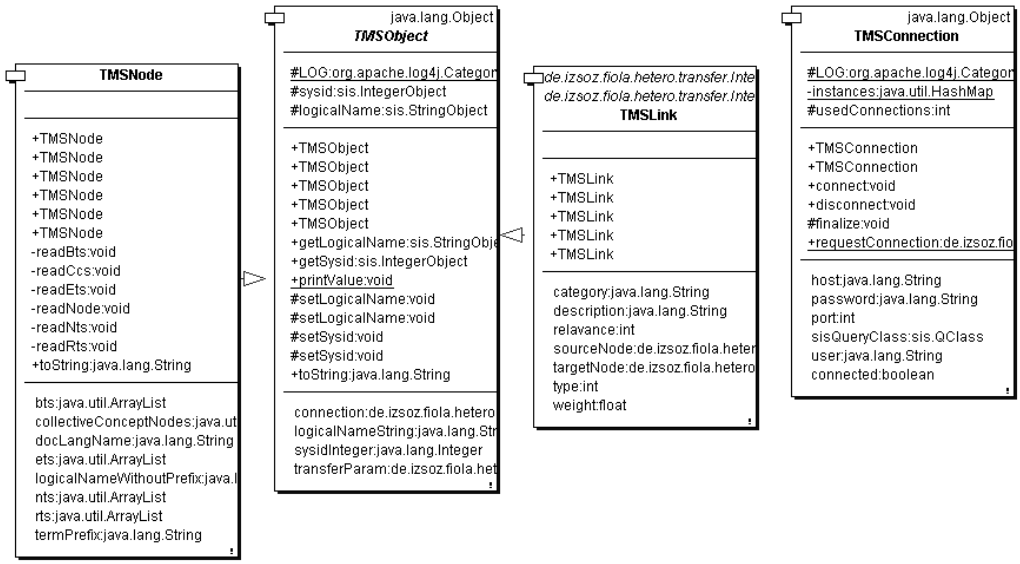

Abbildung 19: Klassen zur Abfrage von SIS-TMS-Datenbanken

Über einfache Heuristiken wurden die Arten von Äquivalenz-Relationen und die Relevanz-Gewichte (hoch, mittel, niedrig) in numerische Gewichte umgerechnet, die den Deskriptoren als Term-Gewicht mitgegeben und für die spätere Verarbeitung in der Retrievalkomponente genutzt werden.

6 Verwertbarkeit der Ergebnisse

Folgende Dienste liegen als Ergebnisse von CAP 11 zur Weiternutzung vor[27]:

[27] Bei Interesse an diesen Modulen bitte Mail an: soe@bonn.iz-soz.de.

a) Transfer-Service

Transfer von XIRQL-Anfragen aus HyRex auf der Grundlage von:

- Statistisch-quantitativ erzeugten Transferbeziehungen
  - IZ-Thesaurus <-> Internet-Freitext
  - MSC/PACS <-> Internet-Freitext
  - MSC <-> PACS (erstellt von CAP9)
  - IZ-Thesaurus <-> SWD (UAP 9 aus CAP12)
  - IZ-Klassifikation <-> BK (UAP 9 aus CAP12)
- Intellektuell erzeugte Crosskonkordanzen (erzeugt von CAP12)
  - IZ-Thesaurus <-> SWD
  - DIPF-Thesuarus <-> SWD
  - RVK <-> IZ-Klassifikation (Auszüge)
  - MSC <-> PACS
  - ETB -> EUN (aus ETB)
  - ETB -> EET (aus ETB)

Spezielle Transfer-Services sind neben dem CARMEN/XIRQL-Transfer bereits vorbereitet, z.B. der Transfer von DDB/SWD nach SOLIS/IZ-Thesaurus für den Übergang von der Deutschen Bibliothek an das GBI; mit geringem Anpassungsaufwand (v.a. für die jeweilige Anfragesprache) lassen sich zusätzliche Transfer-Dienste anbieten.

b) Transfer-Module

- Java-Klassen zum Einlesen und Ausgeben von Anfragen:
  Die Anfragen werden in eine interne Repräsentation überführt und über einen in XML konfigurierbaren Transferprozess über statistisch-quantitative und/oder intellektuelle Transferregeln verändert. Transferregeln können aus relationalen Datenbanken (über JDBC) oder aus SISTMS eingelesen werden.
- Transfer über http:
  Plattformunabhängig können die Transfermodule über eine anpassbare Schnittstelle abgefragt werden; der Server läuft mit den frei verfügbaren http- und Servlet-Servern Apache/TomCat oder entsprechenden anderen Servern.

- Transfer über Einbau der Transfer-Klassen in eine Java-Architektur:
  Die erstellten Klassen können auch ohne Server-Schnittstelle direkt von anderen Java-Programmen genutzt werden.
- Erweiterung um zusätzliche Anfrage-Sprachen:
  Hierzu müssen nur die Klassen zum Parsen und Schreiben der Anfragen ausgetauscht werden.
- Jester:
  Jester, ein Java-Werkzeug zum Erzeugen von Transfer-Regeln über statistisch-quantitative Analyse von Parallelkorpora, liegt zur Weiternutzung vor.

c) Metadatenextraktoren

- Die Metadatenextraktoren, die in der Zusammenarbeit zwischen Osnabrück und Bonn entstanden sind, sind in den CARMEN-Gatherer (CAP7) eingebaut worden. Die Perl-Module können auch einzeln geladen werden.[28]

7 Fazit

Erklärtes Ziel des CARMEN AP11 war es, vor dem Hintergrund nicht flächendeckend greifender Standardisierungs- und Normierungsbemühungen, eine Verbesserung der Recherchen in dezentralen, heterogenen (im Sinne von inhaltlich unterschiedlich erschlossenen) Informationsbeständen zu bewirken. Die Arbeitshypothese war, dass dieses Ziel durch einen mehrstufigen Maßnahmenkatalog zu erreichen ist:

1. das Erzeugen fehlender Metadaten aus den Dokumenten über deduktiv-heuristische Verfahren
2. die Behandlung der aus 1. verbleibenden Heterogenität durch Abbildung der unterschiedlichen Sacherschließungssysteme der zu durchsuchenden Dokumentbestände aufeinander mittels:
   a) Einsatz von statistisch-quantitativen Methoden
   b) Einsatz von intellektuell erstellten Crosskonkordanzen (aus CARMEN AP12)

[28] http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP11/

Die Architektur, Module und Tests, die während der Laufzeit von CAP11 realisiert bzw. durchgeführt wurden, zeigen, dass eine Heterogenitätsreduktion und damit eine Rechercheverbesserung im geschilderten Sinne (s. Kap. 1) durchaus möglich ist, womit die Arbeitshypothese als bestätigt angesehen werden kann:

1. Die Erfahrungen im Bereich Metadatengenerierung zeigen, dass mit deduktiv-heuristischen Methoden sehr gute Ergebnisse zu erzielen sind, was die Qualität der extrahierten Metadaten anbelangt. Allerdings gelingt es nicht, alle für eine inhaltlich angemessene Recherche benötigten Metainformationen flächendeckend zu erzeugen. Hier gibt es je nach Domäne (schwach vs. Stark strukturiert), Dokumenttyp (ps- vs. html-Format) und Metadatum große Schwankungen (s. Kap. 5.2). Zwar erreichen wir in der Mathematik mit einem durchschnittlichen Deckungsgrad von ca. 87,5% einen relativ hohen Wert, die Probleme bei der Extraktion von Parametern wie „Autor“, „Titel“ etc. zeigen jedoch die Grenzen dieser Methode.

   Schwieriger stellt sich die Situation in schwach strukturierten Bereichen wie den Sozialwissenschaften dar, wo neben einem sehr geringen Ausgangsdeckungsgrad bzgl. intellektuell vergebener Metadaten (ca. 15% für die berücksichtigten Parameter „Title“, „Keywords“, „Abstract“) mit Ausnahme des Datums „Title“ nur relativ geringe Generierungsquoten zu erzielen waren.

   Die somit im Datenmaterial verbleibende Heterogenität muss daher mit zusätzlichen Methoden angegangen werden.

2. Quantitativ-statistische Methoden, die Transferbeziehungen zwischen Sacherschließungssystemen mittels Wort-Kookkurrenzen ermitteln, haben im Rahmen der CAP11-Nutzertests (für die Domäne Sozialwissenschaften) ihre Wirksamkeit bewiesen. Die Frage, ob die Einbeziehung zusätzlicher Transferbegriffe in die Freitextrecherche einen Zuwachs an relevanten Dokumenten erbringt, konnte grundsätzlich positiv beantwortet werden (siehe Kapitel 5.3.3).

Neben dieser Bestätigung der Arbeitshypothese sind in CAP11 eine Reihe von Modulen zur Heterogenitätsbehandlung entstanden, die modular aufgebaut sind und für andere Zwecke direkt weiterverwendet werden können (siehe Kapitel 6).

Offen blieb die Frage nach dem angemessenen Mischungsverhältnis zwischen quantitativ-statistischen Transfers und intellektuell erstellten Crosskonkor-

danzen. Im Rahmen von CAP9 fanden vergleichende Tests mit beiden Methoden zwischen MCS und PACS statt. Diese geben Hinweise darauf, dass die Statistik andere Beziehungen zwischen den Termen verschiedener Sacherschließungssysteme aufdeckt als dies bei intellektueller Erstellung der Fall ist (z.B. Problem-/Methodenrelation vs. Ober-/Unterbegriffsrelation; s. Kap. 5.3.2.3). Diese Frage näher zu untersuchen, bleibt Folgeprojekten vorbehalten.

8 Veröffentlichungen

Hellweg, Heiko; Krause, Jürgen; Mandl, Thomas; Marx, Jutta; Müller, Matthias N.O.; Mutschke, Peter; Strötgen, Robert (2001): Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 23).

Krause, Jürgen (2000): Integration von Ansätzen neuronaler Netzwerke in die Systemarchitektur von ViBSoz und CARMEN. Bonn: IZ Sozialwissenschaften. (IZ-Arbeitsbericht; Nr. 21).

Krause, Jürgen (2000): Sacherschließung in virtuellen Bibliotheken - Standardisierung von Heterogenität. S. 202-212. In: Rützel-Banz, Margit (Hrsg.): 89. Deutscher Bibliothekartag in Freiburg im Breisgau 1999: Grenzenlos in die Zukunft. Frankfurt am Main: Klostermann. (Zeitschrift für Bibliothekswesen und Bibliographie, Sonderheft; 77).

Krause, Jürgen (2001): Heterogenität und Integration: zur Weiterentwicklung von Recherche und Inhaltserschließung in der Fachinformation. S. 21-31. In: Schmidt, Ralph (Hrsg.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarkt; 23. Online-Tagung der DGI und 53. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Frankfurt am Main, 8. bis 10. Mai 2001; Proceedings. Frankfurt am Main: DGI. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 4).

Krause, Jürgen (2001): How to Integrate Different Text Data and Fact Information: A Conceptual Transfer Problem in Digital Libraries and its Connection to Agent Theory, News Agent-based Information Retrieval. S. 933-937. In: Smith, Michael J.; Salvendy, Gavriel; Harris, Don; Koubek, Richard J. (Hrsg.): Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality; Proceedings of the HCI International 2001; August 5-10, 2001, New Orleans, Louisiana, USA; Vol. 1. Mahwah: Erlbaum.

Krause, Jürgen (2001): Virtual Libraries, Library Content Analysis, Metadata and the Remaining Heterogeneity. S. 209-214. In: ICADL 2000: Challenging to Knowledge Exploring for New Millennium: the Proceedings of the 3rd International Conference of Asian Digital Library and the 3rd Conference on Digital Libraries, Korea, December 6 - 8, 2000, Seoul.

Krause, Jürgen; Marx, Jutta (2000): Vocabulary Switching and Automatic Metadata Extraction or How to Get Useful Information from a Digital Library. S. 133-134. In: Information Seeking, Searching and Querying in Digital Libraries: Proceedings of the First DELOS Network of Excellence Workshop. Zürich, Switzerland, December, 11-12, 2000. Zürich.

Krause, Jürgen; Schwänzl, Roland; Plümer, Judith (2000): Content Analysis, Retrieval and Metadata: Effective Networking for Mathematics, Physics and Social Sciences. In: Blasius, Jörg; Hox, Joop; Leeuw, Edith de; Schmidt, Peter (Hrsg.): Social Sciences Methodology in the New Millenium. CD-ROM Proceedings of the Fifth International Conference on Logic and Methodology, Cologne, October 3-6, 2000. Amsterdam: TT-Publikaties.

Strötgen, Robert; Kokkelink, Stefan (2001): Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. S. 56-66. In: Schmidt, Ralph (Hrsg.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarkt; 23. Online-Tagung der DGI und 53. Jahrestagung der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V., DGI, Frankfurt am Main, 8. bis 10. Mai 2001; Proceedings. Frankfurt am Main: DGI. (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis; 4).