Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN

Robert Strötgen und **Stefan Kokkelink**

Informationszentrum Sozialwissenschaften (IZ)
Lennéstr. 30
53113 Bonn
Tel. 0228/2281-110
Fax 0228/2281-120
stroetgen@bonn.iz-soz.de

Universität Osnabrück
Fachbereich Mathematik/Informatik
Albrechtstraße 28
49069 Osnabrück
Tel. 0541/969-2735
Fax 0541/969-2770
skokkeli@mathematik.uni-osnabrueck.de

Abstract: Die Sonderfördermaßnahme CARMEN (Content Analysis, Retrieval and Metadata: Effective Networking) zielt im Rahmen des vom BMB+F geförderten Programms GLOBAL INFO darauf ab, in der heutigen dezentralen Informationswelt geeignete Informationssysteme für die verteilten Datenbestände in Bibliotheken, Fachinformationszentren und im Internet zu schaffen. Diese Zusammenführung ist weniger technisch als inhaltlich und konzeptuell problematisch. Heterogenität tritt beispielsweise auf, wenn unterschiedliche Datenbestände zur Inhaltserschließung verschiedene Thesauri oder Klassifikationen benutzen, wenn Metadaten unterschiedlich oder überhaupt nicht erfasst werden oder wenn intellektuell aufgearbeitete Quellen mit in der Regel vollständig unerschlossenen Internetdokumenten zusammentreffen. Im Projekt CARMEN wird dieses Problem mit mehreren Methoden angegangen: Über deduktiv-heuristische Verfahren werden Metadaten automatisch aus Dokumenten generiert, außerdem lassen sich mit statistisch-quantitativen Methoden die unterschiedlichen Verwendungen von Termen in den verschiedenen Beständen aufeinander abbilden, und intellektuell erstellte Crosskonkordanzen schaffen sichere Übergänge von einer Dokumentationssprache in eine andere. Für die Extraktion von Metadaten gemäß Dublin Core (v.a. Autor, Titel, Institution, Abstract, Schlagworte) werden anhand typischer Dokumente (Dissertationen aus Math-Net im PostScript-Format und verschiedenste HTML-Dateien von WWW-Servern deutscher sozialwissenschaftlicher Institutionen) Heuristiken entwickelt. Die jeweilige Wahrscheinlichkeit, dass die so gewonnenen Metadaten korrekt und vertrauenswürdig sind, wird über Gewichte den einzelnen Daten zugeordnet. Die Heuristiken werden iterativ in ein Extraktionswerkzeug implementiert, getestet und verbessert, um die Zuverlässigkeit der Verfahren zu erhöhen. Derzeit werden an der Universität Osnabrück und im InformationsZentrum Sozialwissenschaften Bonn anhand mathematischer und sozialwissenschaftlicher Datenbestände erste Prototypen derartiger Transfermodule erstellt.

1 Projekt Carmen

Die Sonderfördermaßnahme CARMEN („Content Analysis, Retrieval and Metadata: Effective Networking“) zielt im Rahmen des vom BMB+F geförderten Programms GLOBAL INFO darauf ab, in der heutigen dezentralen Informationswelt geeignete Informationssysteme für die verteilten Datenbestände in Bibliotheken, Fachinformationszentren und im Internet zu schaffen. Diese Zusammenführung ist weniger technisch als inhaltlich und konzeptuell problematisch. Heterogenität tritt beispielsweise auf, wenn unterschiedliche Datenbestände verschiedene Thesauri oder Klassifikationen benutzen, wenn Metadaten unterschiedlich oder

überhaupt nicht erfasst werden oder wenn intellektuell aufgearbeitete Quellen mit in der Regel vollständig unerschlossenen Internetdokumenten zusammentreffen.
Im Projekt CARMEN wird diese Aufgabe in den drei Arbeitsgruppen „Metadaten“, „Retrieval“ und „Heterogenität“ bearbeitet. Zur letztgenannten Arbeitsgruppe gehört auch das Arbeitspaket „Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren“, deren Arbeiten hier vorgestellt werden. Beteiligt sind der Fachbereich Mathematik/Informatik der Universität Osnabrück sowie das InformationsZentrum Sozialwissenschaften Bonn.

2 Heterogenitätsbehandlung

Heterogenität wird im Projekt CARMEN nicht als technische, sondern als semantische betrachtet. Sie tritt beispielsweise auf, wenn unterschiedliche Datenbestände zur Inhaltserschließung verschiedene Thesauri oder Klassifikationen benutzen, wenn Metadaten unterschiedlich oder überhaupt nicht erfasst werden oder wenn intellektuell aufgearbeitete Quellen mit in der Regel vollständig unerschlossenen Internetdokumenten zusammentreffen.
Dieses Problem wird mit mehreren Methoden angegangen: Über deduktiv-heuristische Verfahren werden Metadaten automatisch aus Dokumenten generiert, außerdem lassen sich mit statistisch-quantitativen Methoden die unterschiedlichen Verwendungen von Termen in den verschiedenen Beständen aufeinander abbilden, und intellektuell erstellte Crosskonkordanzen schaffen sichere Übergänge von einer Dokumentationssprache in eine andere.
Während die im Folgenden genauer beschriebene Extraktion von Metadaten während des Sammelns der Dokumente erfolgen soll und diese zusätzlichen Informationen dem Dokument dauerhaft zugeordnet werden, setzen die anderen Ansätze während des Retrievals an. Statistisch-quantitativ (Termersetzungs- und -erweiterungsregeln zwischen Freitexttermen und kontrolliertem Vokabular über Kookkurrenzanalyse) wie intellektuell erstellte semantische Relationen zwischen Termen ermöglichen es, eine Benutzeranfrage bestmöglich an die einzelnen Bestände und deren inhaltliche Erschließung anzupassen und dabei für jeden der Bestände spezifisch zu formulieren. Diese Transfermodule werden in die CARMEN-Suchmaschine HyREX eingebunden.

3 Erstellung einer Test-Datenbank

Als Grundlage für die Analyse der Heterogenität von Internetquellen im Bereich **Sozialwissenschaften** und die Evaluierung der Transfermodule und der Komponenten zur Metadatengenerierung wurde ein Beispielkorpus mit ca. 4000 Internetquellen aufgebaut. Die Quellen wurden dem Clearinghouse (SocioGuide) der Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen (GESIS) entnommen, das eine Sammlung von ausgewählten Internet-Adressen auf dem Gebiet der Sozialwissenschaften ist.
Inhaltlich wurde der Korpus auf drei Themenbereiche des GIRT[1]-Scopes, d.h. auf die Themen „Frauenforschung“, „Migration“ und „Industrie- und Betriebssoziologie“ eingeschränkt. Die Themenauswahl für den Carmen-Korpus bietet somit die Möglichkeit vergleichender Retrievaltests. Es wurden nur Internetquellen aufgenommen, die eine inhaltliche Beschreibung einer wissenschaftlichen Aktivität (z.B. eine Projektbeschreibung) darstellten. 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 Dokumenttypen und Dateiformate wurden keine Einschränkungen festgelegt, d.h. der Korpus enthält sowohl Projektbeschreibungen und Literatur(nachweise)

[1] German Indexing and Retrieval Testdatabase (cf. Elisabeth Frisch; Michael Kluck (1997): Pretest zum Projekt German Indexing and Retrieval Testdatabase (GIRT) unter Anwendung der Retrievalsysteme Messenger und freeWAISsf. (IZ-Arbeitsbericht Nr. 10) Bonn.

als z.B. auch Beschreibungen der Arbeitsschwerpunkte von sozialwissenschaftlichen Institutionen, kommentierte Vorlesungsverzeichnisse, Inhaltsangaben von Zeitschriften, Themenschwerpunkte von Konferenzen, Publikationslisten u.ä. Es sind überwiegend HTML-Dokumente im Korpus vertreten, aber auch RTF-, PDF- und Word-Dokumente. Die Internetdokumente wurden im Dateisystem und in einer Oracle-Datenbank abgespeichert.

Aus dem Bereich **Mathematik** stehen verschiedenartige Testdaten zur Verfügung. Der Preprintdienst MPRESS[2] 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 stehen Daten zur Verfügung, die aus mathematischen Online Journalen von einem Harvester gesammelt wurden. Als dritter und sicherlich heterogenster Datenpool bieten 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-Datensätze enthalten auch die URLs der Originaldokumente; diese können zur Analyse dann jeweils lokal gespeichert werden.
Da zur Extraktion von Metadaten aus HTML-Files auf die Erfahrungen, die im Bereich Sozialwissenschaften gesammelt werden, aufgebaut werden soll, 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-Datensätzen dieser Dokumente die URLs extrahiert. Das Script holt dann über das http-Protokoll die Originaldokumente und speichert sie im Filesystem ab.
Es wurde eine Datenanalyse im Bereich Physik durchgeführt. Dabei ergab sich, dass die Struktur der Daten im Bereich Physik sich nicht wesentlich von der in der Mathematik unterscheidet. Aus diesem Grund wurde zunächst auf Testdaten aus der Physik verzichtet.

4 Erste Analyse der vorhandenen Heterogenität

Die Analyse der für **sozialwissenschaftliche Dokumente** vorzufindenden Heterogenität stützte sich auf das oben beschriebene Testkorpus. Diese Dokumente wurden bezüglich des Vorkommens von Metadaten untersucht. Nach einer ersten Durchsicht wurden Dokumententitel, Autor, Institution, Schlüsselwörter und Abstract als grundsätzlich extrahierbar identifiziert.
Die Heterogenität der Dokumenttypen machte hier eine Unterscheidung zwischen Autoren und Institutionen als Urheber eines Dokuments (in Projektbeschreibungen, Aufsätzen u.Ä.) einerseits und referenzierten Autoren und Institutionen (in Projektlisten, Literaturverzeichnissen u.Ä.) nötig.
Bei der Analyse wurde deutlich, dass die Unterschiede der Codierung von Metadaten innerhalb einer WWW-Site einer Institution oder eines Autors relativ einheitlich verwendet wurden, während vor allem zwischen den verschiedenen Sites große Unterschiede bestehen. In einer Stichprobe von etwa 100 Dokumenten wurden daher eine oder mehrere HTML-Dateien pro Site ausgewählt und für diese analysiert, wie im HTML-Quelltext diese Metadaten gefunden werden können. Dabei wurde insbesondere Folgendes festgestellt:

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

- Höchstens 5-10% der untersuchten sozialwissenschaftlichen Internetquellen weisen Meta-Tags auf. Es werden ausschließlich die Tags „Author“, „Description“ und „Keywords“ gefunden, Meta-Tags nach Dublin Core wurden nirgends verwendet.
- 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 vorgefundenen sozialwissenschaftlichen 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 – durchaus 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

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

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. Die Heterogenität der Dokumente kann durch diese Extraktion von Metadaten nicht vollständig behandelt werden, die Erfolgsquote wird auf ca. ein Drittel geschätzt. Für die übrigen Fälle ist eine Behandlung über die quantitativ-statistischen Verfahren unumgänglich.

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 nur eine 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, wurde in AP11 mit der Entwicklung von Verfahren zur Extraktion von Metadaten aus Postscript-Dokumenten begonnen.

5 Metadaten-Extraktion

5.1 Erstellung von Heuristiken und Extraktoren

Für die Extraktion von Metadaten aus **sozialwissenschaftlichen Internetquellen** wurden die Auswertungen aus dem vorhergehenden Arbeitsschritt zugrunde gelegt. Nicht einzelne Heuristiken für einzelne Sites, sondern eine gemeinsame Heuristik über den gesamten Testkorpus wurde gesucht.
Die deduktiven Heuristiken liefern aufgrund der heterogenen und oft auch fehlerhaften Codierung der Metadaten keine sicheren Ergebnisse, sondern Wahrscheinlichkeiten, mit denen ein Datum korrekt identifiziert wurde. Diese Wahrscheinlichkeit wird durch Gewichte von 0 (sehr schlecht) bis 1 (sehr gut) dem Datum zugeordnet.
Für die unterschiedlichen Metadaten lassen sich verschieden zuverlässige Heuristiken finden. Während der Titel eines Dokuments relativ gut identifizierbar ist, sind die Schlüsselwörter und Abstracts deutlich schlechter zu finden. Zu den bisherigen Ergebnissen mehr in Kapitel 5.2.

Eine Voraussetzung für die Extraktion von Informationen aus HTML Dokumenten ist die Fähigkeit die Struktur von HTML Dokumenten zu analysieren. Leider stehen einer effizienten Auswertung von HTML Dokumenten folgende Punkte im Wege:

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

Im Umfeld von XML sieht die Situation wesentlich besser aus. Es existiert eine Fülle von standardisierter Technologie, 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*: Eine Anfragesprache für XML Dokumente zur Extraktion von Informationen.

Aufgrund der Verwandschaft von HTML und XML liegt es nahe vorhandene XML Technologien für die Analyse von HTML Dokumenten zu nutzen. Dazu sind im wesentlichen 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. (Dies definiert momentan wohl leider den Korrektheitsbegriff für HTML Dokumente im World Wide Web.) Unter dieser Prämisse ist ein geringer Informationsverlust unumgänglich.

Die Konvertierung von HTML Dokumenten in wohlgeformte XML Dokumente erfolgt in zwei Schritten.

1. Mit Hilfe eines HTML Parsers wird versucht, das HTML Dokument auf eine Baumstruktur abzubilden. Bei diesem Prozess werden verschiedene Heuristiken zur Bereinigung der Daten angewandt. (Beispiel: Überlappende Tags hello world werden aufgelöst.)
2. Der HTML Baum wird in einen DOM Baum konvertiert. Hierbei werden weitere XML-spezifische Bedingungen an die Baumstruktur von XML Dokumenten berücksichtigt. (Beispiel: XML Attribute müssen mit einem Buchstaben beginnen.).

Aus den konvertierten HTML Dokumenten lassen sich nun mit Hilfe von XPath-Anfragen Informationen extrahieren, die als Ausgangspunkt für die Heuristiken zur Metadatenextraktion dienen. Die Definition dieser Heuristiken läßt sich somit grob in folgende Schritte unterteilen.

1. Spezifikation von XPath-Anfragen zur Extraktion relevanter Daten.
2. Definition einer Heuristik, die auf den extrahierten Daten operiert und als Resultat eine mit Wahrscheinlichkeiten gewichtete Liste von Ergebnissen liefert.

Über verschiedene Formulierungs- und Formalisierungsschritte wurde eine algorithmische halbformale Sprache gefunden, über die Heuristiken ausgetauscht werden können. Für die Extraktion eines Titels sieht die Heuristik in dieser Sprache beispielsweise so aus ([x] gibt das Gewicht eines Datums an):

Extraktion relevanter Daten mittels XPath-Anfragen:

```
1.  Finde den ersten <title> Tag im HTML Dokument (XPath: //title)

2.  Sei H1=(x_1,..,x_n) die Knotenliste der Anfrage //h1.
    Fasse maximale Folgen x_l,..,x_k mit (x_i ist Vorgänger
    von x_i+1) zu einem Eintrag zusammen. (Somit ist H1 eine Liste
    von Einträgen, so dass kein x_i mehr direkter Vorgänger von
    von x_i+1 ist).

    Führe dasselbe für h2,..,h6 durch. Das Gesamtergebis ist eine
    Matrix H=
    (1, string1)
    (1, string2)
    (2, string3)
    usw. Hierbei entspricht der erste Eintrag der Größe der
    Überschrift. Sei H1 der erste(!) Eintrag (1,H1) in H.
    Analog für H2,..,H6. Sei HMAX die größte dieser Überschriften.

    Suche nach allen Kombinationen //p/b , //b/p usw.
    Sei also s_1,..,s_n die Knotenliste, die aus der Anfrage
    //*[self::p/parent::strong or self::p/parent::b or
    [...]
    self::i/parent::td or
    ]
    resultiert. Die Liste ist nach der Reihenfolge des Vorkommens
    im Dokument geordnet. Finde das nun das größte k, so dass
    für alle s_i aus  {s_1,..,s_k} gilt:
   1. name(s_i)=name(s_(i+1))
   2. name(parent(s_i))=name(parent(s_i+1))
   3. parent(s_i) ist Vorgaenger  von parent(s_(i+1)).
   Fasse diese s_1,..,s_k zu einem String S zusammen.
```

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]=<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 && S vorhanden) {
     /* d.h. <title> vorhanden UND es existiert ein Eintrag
      //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
    }
 }
```

Die Heuristiken wurden implementiert, um einen direkten Einsatz in der Retrievalkomponente zu ermöglichen. In einem iterativen Prozess wurden (und werden noch weiterhin) 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 Relevanzbewertungen der Dokumente angestellt und die extrahierten Metadaten mit den im Dokument vorhandenen Daten verglichen. Zu unterscheiden sind hier falsche oder ungültige Metadaten, die extrahiert wurden und eine Korrektur der Heuristik erfordern, von richtigen, gewünschten Metadaten, die im Dokument vorhanden, aber nicht extrahiert wurden und daher eine Erweiterung der Heuristik nötig machen.
Bisher wurden Heuristiken für die Extraktion der Metadaten Titel, Keywords und Abstract erstellt, für die referenzierten Titel, Autoren und Institutionen steht eine Erstellung noch aus.

Im Bereich **Mathematik** wurden zunächst Heuristiken entwickelt, die Metadaten wie Abstract, Schlüsselwörter oder MSC Klassifikation aus mathematischen Orginalarbeiten im Postscript Format extrahieren lassen, falls diese von den Autoren im Volltext als solche kenntlich gemacht wurden.
Ausgangspunkt für die Behandlung von Postscript Dokumenten ist das von der New Zealand Digital Library[3] entwickelte Programm `prescript`[4], das Postscript Dokumente in Text- oder HTML-Dokumente konvertiert.

[3] Nähere Informationen siehe unter http://www.nzdl.org/.
[4] Siehe auch http://www.nzdl.org/html/prescript.html.

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. 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 des CARMEN-Arbeitspakets „A Document Referencing and Linking System“. In der von diesem Arbeitspaket 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` erstellt, die eine Navigation in den generierten HTML Dokumenten erlaubt und somit die Implementierung von speziellen Heuristiken ermöglicht. Für den Bereich der Mathematik wurde die Klasse `MathHeuristics` erstellt, die Methoden zur Extraktion von Abstract, Schlüsselwörter und MSC Klassifikationen aus den erzeugten HTML Dokumenten anbietet. Als Beispiel sei hier die Heuristik für die MSC-Klassifikation aufgeführt:

> *MSC-Klassifikation*. Suche in den ersten drei Seiten nach Paragraphen, die mindestens zwei der Wörter "ams","msc","classification","subject" oder "mathematics" enthalten. Suche in diesen 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).

Derzeit wird an Heuristiken zur Extraktion von Literaturangaben gearbeitet. Literaturangaben können in Dublin Core als DC.Relation mit dem Attribute "references" kodiert werden. Durch diese Relationen zwischen einzelnen Dokumenten entsteht in einem Index von Dokumenten ein Netz, das weiter untersucht werden sollte. Es kann z.B. verwendet werden, um zu beurteilen, ob Dokumente ähnliche Inhalte haben. Dies kann durch Entfernungen als Anzahl der Referenzen, die zwischen den Dokumenten liegen bewertet werden.
Bislang bereitet die Extraktion der Autoren und des Titels eines Dokumentes große Probleme, da das Konvertierungsprogramm `prescript` keine Informationen über Zeichensätze und Zeichengröße erhält. Im weiteren wird versucht diese Probleme durch Kombination mit einer Analyse der Literaturangaben lösen.

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`[5] gesehen. `Kea` extrahiert mit Hilfe probabilistischer Verfahren Schlüsselphrasen aus Texten. Anhand einer Trainingsmenge von Texten mit ausgewählten

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

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.

5.2 Test der Heuristiken und Extraktoren

Da derzeit noch kein lauffähiges System für Endbenutzer vorliegt, konnten Benutzertests mit Retrieval-Nutzern noch nicht durchgeführt werden.
Es wurden Tests der Metadaten-Extraktion aus HTML-Seiten mit Inhalten aus dem Bereich **Sozialwissenschaften** durchgeführt. 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.
Zur Vereinfachung der Tests wurde als Java-Programm eine Test-Suite entwickelt, das die Tests wesentlich vereinfacht. Die extrahierten Metadaten werden in eine Baustruktur 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.

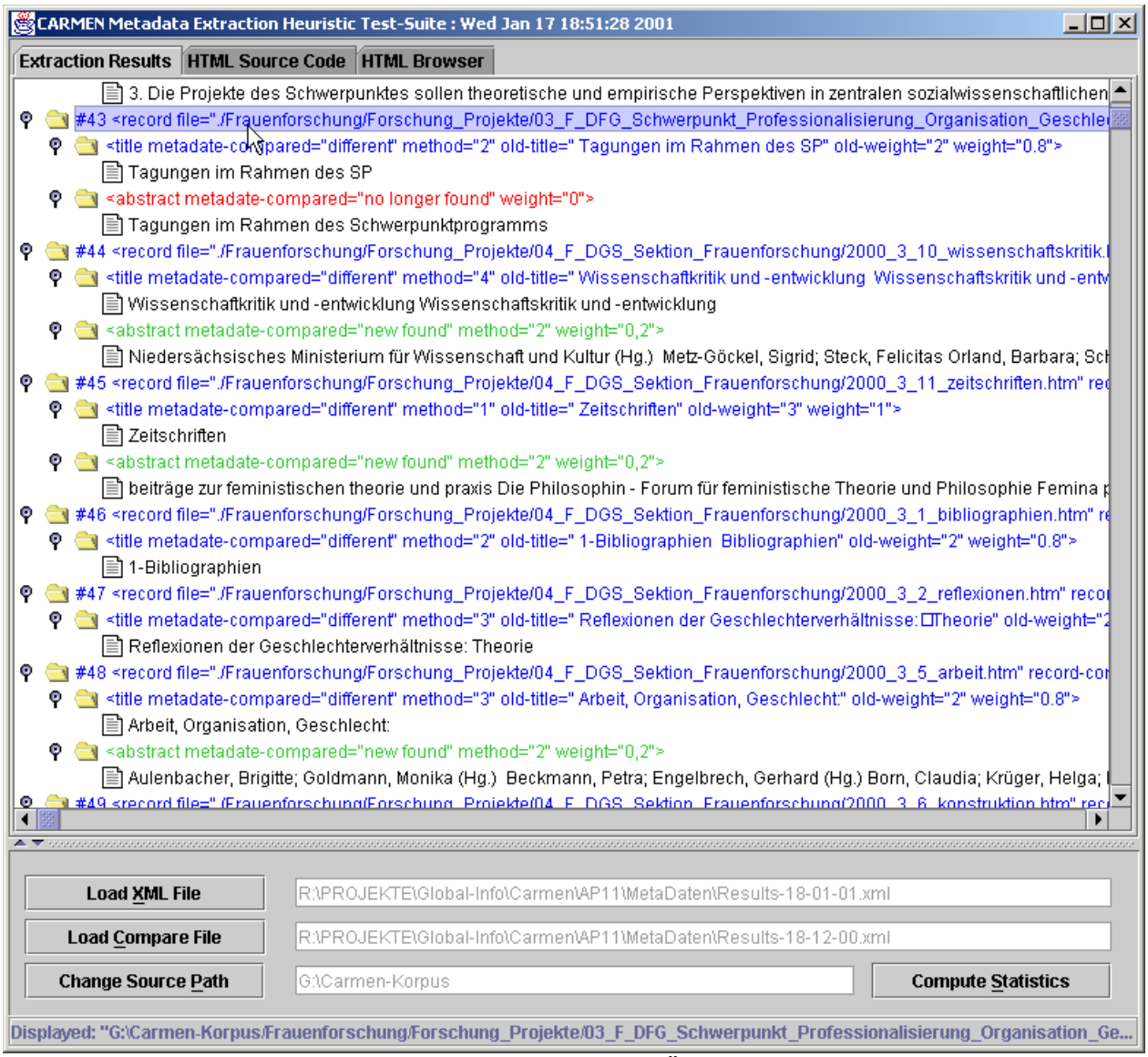

Abb. 1: Bildschirmkopie der Test-Suite für die Überprüfung der Heuristiken

Überprüft werden dabei derzeit die Vorkommen der Metadaten Titel, Autor, Institution, Keywords, Abstract. Erste Tests wurden auf einem kleineren Testkorpus von knapp 1.500 Dokumenten ausgeführt, weitere Test laufen über den gesamten Korpus von knapp 4.000 Dokumenten. Bei 3.715 HTML-Dokumenten im sozialwissenschaftlichen Test-Korpus konnten zuletzt aus 3.661 Dokumenten der Titel (davon 3.291 mit hohem Gewicht), aus 447 Dokumenten die Schlüsselwörter und aus 531 Dokumenten Abstracts gewonnen werden.

Aus dem Bereich **mathematische Preprints** wurden in einem ersten Versuch 37 Preprints untersucht, in denen folgenden Angaben vorhanden waren:

Abstracts	33 (89%)
Keywords	21 (56%)
MSC	20 (54%)

Es wurden 29 Abstracts korrekt extrahiert (89%), ein Abstract wurde überhaupt nicht und drei Abstracts wurden nicht vollständig erkannt. Es wurden alle Schlüsselwörter erkannt (100%). Es wurden alle MSC Klassifikationen erkannt (100%), allerdings wurde eine ungültige Klassifikation zuviel ausgegeben.

Statistisch aussagekräftigere Tests werden durchgeführt werden, wenn die Entwicklung der Heuristiken abgeschlossen ist. Die extrahierten Metadaten werden dann analog zum Vorgehen im Bereich Sozialwissenschaften mit Gewichtungen versehen. Sollten die bisherigen hoch signifikanten Ergebnisse sich nicht wesentlich ändern, können die extrahierten Metadaten mit dem höchsten Relevanzwert versehen werden.

6 Zusammenfassung

Mit den beschriebenen Arbeiten ist es gelungen, mit einigem Erfolg über nicht site-spezifische Heuristiken wichtige Meta-Daten aus strukturierten und unstrukturierten Internet-Dokumenten zu gewinnen. Die nächsten Schritte sind in einer weiteren Verbesserung der Heuristiken zur Metadaten-Extraktion zu sehen und deren Einbindung in die Gatherer-Komponente im CARMEN-Projekt (CARA) zu sehen.