Umwandlung der Daten (Zeichen-Kodierung in BibTeX, Normalisierung der Autorennamen) ins LEABib-eigene SRC-Format.

Im Rahmen des Projektes FIS-I wurde der ***io-port.net*-Editor** zur Erfassung und Bearbeitung bibliographischer Daten entwickelt. Durch die Unterstützung verschiedener Formate kann der *io-port.net*-Editor flexibel eingesetzt werden. Er vereinfacht die einheitliche Erfassung und Korrektur der Daten durch vordefinierte Wertelisten, einer Autorendatenbank und weiterer Hilfsmittel. Korrekte LaTeX-Formeln können auf einfache Weise mit dem integrierten LaTeX-Formel-Editor eingefügt werden.

Datenkorrektur - Ein Java-Programm prüft die erfassten Daten mit Hilfe verschiedener Verfahren auf Richtigkeit durch Abgleichen mit vordefinierten Wertelisten, Test auf richtigen Zeichensatz (ASCII) und gültige LaTeX-Formatierungen und Klammerungen. Mittels des GUI können die gefundenen Fehler schnell korrigiert werden.

1.4 Heuristische Verfahren für die semantische Anreicherung unstrukturierter bibliographischer Daten

Im Projekt FIS-I wird der Zugriff auf Informatik-Literatur zentralisiert. Der Projektpartner Universität Karlsruhe – Collection of Computer Science Bibliographies (CCSB) ist einer der Datenlieferanten für das Projekt. Die bibliographischen Daten der CCSB sind sehr unterschiedlicher Qualität. Sie sind oft mehrmals konvertiert, aus verschiedenen Quellen maschinell extrahiert und nicht selten in einem Dateiformat, das die Daten nicht vollständig semantisch beschreibt. In dem Vortrag zeigen wir Beispiele von realen, bei uns aufgetretenen Problemen und deren Lösung, welche die semantische Qualität und damit die Nutzbarkeit der bei uns gesammelten bibliographischen Daten wesentlich verbessert.

Die Verfahren kann man sich auf drei Ebenen vorstellen. Die Daten müssen konvertiert, extrahiert und bereinigt werden. Der größte Teil der Konvertierung geht zur Zeit von XML (mit DublinCore Namensraum [WK98]) und (X)HTML Dateiformaten aus. Der Datenbestand, den wir aus XML/DublinCore in CCSB integrieren, beträgt ca. 780 000 Einträge – 35% der Gesamtanzahl (Stand: Juni 2005). Das DublinCore Schema ist aus Prinzip sehr allgemein und ungenau. Die existierenden Vorschriften für Qualified DublinCore [DCMI] und Richtlinien für die Kodierung der bibliographischen Einträge in DublinCore [Ap05] werden leider in der Praxis noch nicht verwendet und fast alle Einträge kommen in Unqualified DublinCore [DCES]. Das bedeutet, dass nicht festgestellt werden kann, was aus semantischer Sicht die benannten Knoten des XML-Baums eigentlich beinhalten. Die Abbildung von Knotennamen in einen anderen Bibliographiespezifischen Namensraum (z.B. BibTeX) macht Datenextrahierung notwendig. Die einzigen Felder, die potenziell direkt umgeschrieben werden könnten, sind DC.TITLE und DC.CREATOR. Fast alle anderen Informationen sind sehr oft zu DC.DESCRIPTION umgeleitet. Es gibt aber noch Felder DC.IDENTIFIER, DC.SOURCE, DC.LINK und DC.RELATION, die, falls vorhanden, normalerweise die notwendigen Informationen bieten, um den Eintrag brauchbar zu machen. Die obengenannten Felder werden inkonsistent benutzt. Das erfordert die Erkennung von Inhalten und Extrahierung aller relevan-

ter Informationen wie z.B. Zeitschriftenname, Band (Vol.), Ausgabe (No.), Seitenzahl, ISSN, ISBN, Verlag, Herausgeber. Auch das Datum muss aus vielen verschiedenen Schreibarten in „Jahr-Monat-Tag“ umgewandelt werden. Der Typ der Publikation kann manchmal nur aus der URL erraten werden. Man kann annehmen, dass die extrahierten Daten vollständig semantisch korrekt und in dem Extrahierungsprozess bereits bereinigt worden sind. Für die übrigen Felder, wie Autoren, Titel und Zusammenfassung muss dies erst durchgeführt werden. Das bedeutet Entfernung aller überflüssigen Textauszeichnungen (HTML und LaTeX Strukturen), Korrektur und Umschreibung von Sonderzeichen (UTF-8, HTML Entities, TeX Darstellungsart, falsch genutzter Mathematikmodus von TeX). Die Autorennamen, die als zusammengesetztes Textelement vorliegen, müssen getrennt werden und entschieden, welcher Teil der Teilkette der Vorname ist, welcher der Nachname und was nicht zum Namen gehört und entfernt werden soll.

Die obengenannten Probleme sind nur ein Teil von in der CCSB aufgetretenen Aufgaben, die im Projekt größtenteils gelöst wurden. Mehrere zusammengesetzte heuristische Regeln, mit regulären Ausdrücken als Werkzeug, haben die Datenextrahierung und -bereinigung möglich gemacht.

2 Projekt Semantische Methoden und Werkzeuge für Informationsportale (SemIPort)

2.1 Ontologie-basiertes Web Mining zum Aufbau großer Informationsportale

Die Erkennung und Extraktion relevanter Daten im Internet wird zunehmend durch den rapiden Zuwachs an Dokumenten erschwert. Bestehende Ansätze, denen aktuelle Suchmaschinen in der Regel folgen, entgegnen den anfallenden Datenmengen mit immer neuer Rechenleistung. Diese Vorgehensweise wird sich jedoch nicht beliebig fortsetzen lassen. In dem SemIPort Projekt wurde der fokussierte Web-Crawler METIS (http://ontoware.org/projects/metis) zur Identifikation und Extraktion kontextrelevanter Informationen aus dem Internet entwickelt, welcher Hintergrundwissen in Form von Ontologien verwendet.

Grundsätzlich wird zwischen mehreren Arten von Ontologien unterschieden. Zum einen wird eine **Web-Ontologie** modelliert. Diese beschreibt die Struktur und Eigenschaften von Dokumenten im Internet, sowie deren Verknüpfungen mittels sog. *Hyperlinks*. Sie repräsentiert außerdem *Hosts,* auf denen Internet-Dokumente gespeichert werden. In der **Domänen-Ontologie** wird die eigentliche Domäne beschrieben. Das dort gespeicherte Wissen stellt letztendlich das Ziel der fokussierten Suche dar. Zum Aufbau eines Informationsportals für wissenschaftliche Publikationen aus der Informatik beschreibt die Domänen-Ontologie z. B. Fachrichtungen, Eigenschaften von Publikationen und beschreibt u.a. Personen und Forscher.

Im Gegensatz zu Informationsextraktionsmechanismen, die eine Bewertung von Res-