# Extrahierung bibliographischer Daten aus dem Internet

Paul Ortyl[1], Stefan Pfingstl[2]
[1] LS Informatik für Ingenieure und Naturwissenschaftler,
Universität Karlsruhe (TH), D-76128 Karlsruhe
ortyl@ira.uni-karlsruhe.de
[2] LS für Effiziente Algorithmen, TU München, D-85748 Garching
stefan.pfingstl@in.tum.de

**Abstract:** Im Projekt FIS-I, das vom Bundesministerium für Bildung und Forschung (BMBF) gefördert wird, soll der Zugriff auf Informatik-Literatur zentralisiert werden. Die Projektpartner Universtät Karlsruhe (*Collection of Computer Science Bibliographies*[1]) und TU München(*LEABiB*[2]) stellen hierfür die bibliographischen Daten bereit. In diesem Beitrag werden die praktischen Erfahrungen vorgestellt, die bei der Erfassung und Bearbeitung von bibliographischen Daten gesammelt wurden. Erstens wird betont, dass die genaue Einhaltung von Standards (u. a. OAI-PMH) wesentliche Grundlage für die Interoperabilität ist. Dadurch kann die Datenqualität erhöht und überflüssige und fehlerträchtige Anpassungsarbeit erspart werden. Zweitens werden die Probleme bei der Datenerfassung mittels Wrapper aufgezeigt.

## 1 Open Archives Initiative

Die *Open Archives Initiative*[3], die um die Jahreswende 1999/2000 gegründet wurde, beschäftigt sich mit der Verbreitung und Vernetzung von Dateien über vorhandene Zeitschriften und Vordrucke (*eprints*). Es wurde ein Protokoll [OAI03] (*Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)) entworfen und implementiert, welches die einfache Sammlung von Metadaten ermöglicht.

Die *Collection of Computer Science Bibliographies* [Ac04] als Dienstleistungsanbieter (und Datenanbieter für das `io-port.net` Projekt) integriert unter anderem die Daten, die durch OAI-PMH zur Verfügung stehen, in den internen, durchsuchbaren Datenbestand. Trotz des möglichst genau spezifizierten Protokolls (OAI-PMH) und bekannten Datenaustauschformats (*Dublin Core* (DC) [DCMI]) gibt es in der Praxis viele im weiteren Text beschriebene Probleme, die nicht immer einfach und sauber gelöst werden können.

[1] `http://liinwww.ira.uka.de/bibliography`
[2] `http://wwwmayr.in.tum.de/leabib`
[3] `http://www.openarchives.org`

### 1.1 OAI-PMH Server

Jede OAI-Datenquelle bietet ihre Daten durch OAI-PMH über HTTP-Server an. Um die Daten zu erfassen, braucht man eine genaue URL für jeden einzelnen Server. Datenanbieter OAI-PMH-konformer Server können sich registrieren[4], aber bisher hat das nur ein kleiner Teil von ihnen getan. Seit Anfang 2004 hat sich die Situation verbessert, da mehrere Server — wie im Protokoll vorgesehen — ihre „Freunde" als zusätzliche Liste von bekannten URLs anzeigen. Es gibt in der Zwischenzeit auch Webseiten, die alle bisher gefundenen Server auflisten und bewerten. Es kommt leider relativ häufig vor, dass eine Institution auf ihrer Webseite behauptet, dass sie einen OAI-PMH-konformen Server zur Verfügung stellt, ohne eine passende URL anzugeben. Dann kann man nur bekannte URL-Muster durchprobieren, was dem eigentlichen Ziel der Informationsverbreitung nicht förderlich ist.

Manche Datenanbieter halten sich nicht genau genug an das OAI-PMH. Das Protokoll verlangt, dass alle Daten in XML mit UTF-8 Sonderzeichenkodierung übertragen werden. In der Praxis prüfen manche Serverimplementationen allerdings nicht, ob die Daten, die von der Datenbank mittels OAI-PMH angeboten werden, richtig kodiert sind. Diese Dateien müssen deshalb erst repariert werden, bevor sie mit dem XML-Parser weiter bearbeitet werden können. Zum einen werden Daten mit ISO-8859-X kodiert. In anderen Fällen werden die Sonderzeichen von CP1252 nicht richtig umkodiert, sondern lediglich eins zu eins auf UTF-8 abgebildet (was nur bei ISO-8859-1 klappt). Und schließlich werden schon in UTF-8 codierte Daten als in ISO-8859-1 befindlich aufgefasst und erneut nach UTF-8 umgewandelt. Diese Kodierungsfehler lassen sich nicht in allen Fällen maschinell wieder rückgängig machen. Es gibt nur ein paar Heuristiken für häufig vorkommende Fehler.

Andere XML-Strukturverletzungen passieren dadurch, dass nicht standardkonforme zusätzliche `<text>` Knoten in jedem Textelement erscheinen. Dies macht ein aufwändigeres XML-Strukturentdeckungsverfahren erforderlich, statt Felderinhalte aus dem XML-Datenstrom direkt zu lesen.

Probleme bereiten fehlerhaft implementierte Server, die bei bestimmten Anfragen nur Freitextfehlermeldungen der darunterliegenden Skriptspracheninterpreter statt den vorgeschriebenen OAI-PMH Fehlermeldungen zurückliefern und damit die Interoperabilität erheblich erschweren.

Die meisten der oben genannten Probleme könnten durch Verständigung der Betreiber per E-Mail gelöst werden, allerdings bleiben E-Mails häufig unbeantwortet.

### 1.2 Dublin Core

Das OAI-PMH Protokoll verlangt, dass Daten im XML-Format sind. Die interne Struktur ist nicht auf ein Format begrenzt, die Daten können in mehreren parallelen Formaten kodiert sein. Um die Interoperabilität zu sichern, muss eines dieser Formate allerdings dem

[4] z.B. unter `http://www.openarchives.org/Register/BrowseSites.pl`

*Dublin Core* entsprechen.

Die meisten OAI-Quellen liefert ihre Daten nur in *DC* Struktur. *Dublin Core* wurde nicht exklusiv für bibliographische Daten, sondern für allgemeine Metadaten, einschließlich Katalogisierung von Museumsammlungen, entwickelt. Leider nutzen viele Datenlieferanten nicht mal diese eingeschränkten Möglichkeiten zur Spezifikation ihrer Daten [DCMI03] und vermindern dadurch die Nützlichkeit ihrer Beiträge.

Ein altbekanntes Problem bei großen Datensammlungen sind Duplikate. Sehr wichtige Felder bei Deduplizierungsmethoden sind Autoren und Titel. Im Feld `<dc:creator>` erscheinen außer Autornamen auch Institutionen, Adressen, E-Mail-Adressen, Geburtsdaten oder akademische und berufliche Titel. Man braucht komplexe heuristische Methoden, um daraus die Autorennamen zu extrahieren und eventuell sogar die restlichen nutzbaren Daten in die richtigen Felder einzusetzen. Trennzeichen wie Komma und Semikolon werden ohne Unterschied benutzt, dadurch ist es nicht eindeutig ob sie die verschiedene Informationen, Autorennamen oder Vorname und Nachname trennen. Das Feld `<dc:description>` wird häufig für alle Informationen außer Autor, Titel oder Jahr benutzt. Die automatische Auftrennung dieser Daten ist schwierig und nur für einfache und häufig vorkommende Felderinhalte erfolgreich.

Es gibt Zusammenfassungen und seltener Titel, die noch HTML oder LaTeX Formatierungskommandos insbesondere für mathematische Formeln enthalten. Diese sind nach mehrmaligen Umwandlungen oft nicht mehr inhaltlich korrekt und können dem Endbenutzer nicht mehr richtig präsentiert werden.

Die bibliographischen Einträge sind manchmal klassifiziert, um den in OAI-PMH vorgesehenen Zugriff auf ausgewählte Teile des Datenbestands zu ermöglichen. Fast jede OAI-Datenquelle benutzt selbst für die verbreiteten Klassifikationen seine eigenen Bezeichnungen. Um informatikrelevante Einträge zu erkennen, müssen heuristische Methoden verwendet werden.

Die Entwicklung von OAI-PMH und die ständig wachsende Anzahl von vorhandenen OAI-Datenquellen ist sicherlich ein sehr großer Fortschritt bei der Verbreitung semantikreicher bibliographischer Daten, aber um alle Vorteile aus diesem großen dezentralisierten Datenbestand ausschöpfen zu können, sollten die Datenproduzenten und -lieferanten noch mehr auf Datenqualität und Interoperabilität achten.

## 2 Datenextrahierung aus Internetseiten

Zur schnelleren Erfassung der bibliographischen Daten wurde an der TU München ein Tool (WrapGen und DataGen) entwickelt, das es ermöglicht, aus bestehenden Inhaltsverzeichnissen im Internet gültige und korrekte bibliographische Datensätze zu extrahieren. Dieses Tool generiert anhand von Beispielen einen Wrapper zur Extrahierung der gewünschten Daten. Die Implementierung orientiert sich am Algorithmus STALKER [MMK01]. Die Abweichungen sind im Folgenden beschrieben.

## 2.1 Wrapper

Wrapper [Ku00], [MMK01] sind Funktionen, die aus einem Text die relevanten Daten mittels Regeln extrahieren. Für die Extraktion bibliographischer Daten aus HTML-Seiten eignen sich HLRT-Wrapper am Besten, da diese der Struktur einer Inhaltsangabe im Internet am ähnlichsten sind. HLRT-Wrapper besitzen jeweils eine linke (L) und rechte Regel (R) für die zu extrahierenden Daten, und zwei Regeln, die den zu durchsuchenden Seitenbereich (H, T) definieren.

### 2.1.1 Heuristik zur Regelauswahl

In der Lernphase werden alle gültigen Regeln für die Extrahierung eines Feldes erzeugt und bewertet. Die erzeugten Regeln bestehen aus folgenden Teilen:

- HTML-Tags, wobei HTML-Tags mit Optionen nur als $<$*TAGNAME* (z.B.: $<$*img* ) betrachtet werden
- folgende Sonderzeichen: `. : , ; - + _ # * ( )`

Die gültigen Regeln werden anschließend bewertet. Regeln, die nur aus HTML-Tags bestehen, werden höher bewertet als solche, die Sonderzeichen enthalten.

### 2.1.2 Automatische Nachbearbeitung der Daten

Da die Internetseiten verschiedene HTML-Konstrukte enthalten, die Speicherung der Daten aber in BibTEX erfolgen soll, ist es nötig, die mit dem Wrapper extrahierten Daten aufzubereiten. Bei der automatischen Extrahierung der Daten treten folgende Fehler auf:

**Autoren-Namen:** Die Autoren-Namen sollen in einer einheitlichen Schreibweise erfasst werden. Die verschiedenen Verlage verwenden aber unterschiedliche Arten, um die einzelnen Autoren einer Publikation anzugeben und zu trennen (z.B. `, <br> \n`).

**Mathematische Formeln:** In den HTML-Seiten werden mathematische Formeln meist als Grafiken eingebunden. Diese Grafiken müssen durch entsprechende Skripte entfernt und als LATEX-Code neu eingefügt werden. Diese Ersetzung kann nur für einen Teil der Formeln automatisch erfolgen.

**Sonderzeichen:** Sonderzeichen in HTML oder als Grafiken gilt es, nach LATEX zu übersetzen (z.B. ä nach \"a).

**Weitere HTML-Konstrukte:** Die restlichen HTML-Konstrukte müssen nach LATEX übersetzt (z.B. Text nach {\bf Text}) oder entfernt werden.

**Vordefinierte Schreibweisen:** Für bestimmte Felder gelten vordefinierte Schreibweisen. Die Daten aus den HTML-Seiten müssen an diese Schreibweisen angepasst werden (z.B. Seitenzahlen: '3–5' anstatt '3– –5').

### 2.2 Manuelle Datenkorrektur

Eine fehlerfreie automatische Erfassung ist mit Hilfe von Wrappern leider nicht möglich. Aus diesem Grund ist die manuelle Korrektur der Daten immer noch nötig. Hierbei soll der Benutzer vom verwendeten Tool bestmöglichst unterstützt werden. Diese kann auf mehrere Weisen geschehen:

**Vordefinierte Wertelisten:** Journal-Namen, Serien-Titel, Verlage, ... können mit den bisherigen Daten oder vordefinierten Wertelisten abgeglichen werden. Damit wird vermieden, dass z.B. identische Journale auf verschiedene Weisen geschrieben oder abgekürzt werden.

**Rechtschreibprüfung:** In den Titeln und Abstracts können mit Hilfe einer Rechtschreibprüfung Rechtschreibfehler korrigiert werden.

**Autoren-Datenbank:** Mittels einer Autoren-Datenbank ist es möglich, schneller einen Autoren eindeutig zu identifizieren. Hilfsmittel sind dabei die Autoren–Mitautoren–Beziehungen, Klassifikationen, ...

## 3 Zusammenfassung

Die Einhaltung von Standards (OAI-PMH) ist eine wesentliche Grundlage für die Interoperabilität. Dadurch kann die Datenqualität erhöht und überflüssige und fehlerträchtige Anpassungsarbeit erspart werden. Bei der manuellen Erfassung bibliographischer Daten konnte durch Einsatz von Wrappern die Erfassungszeit um ca. 70% verkürzt werden. Ein neuer Editor, der im Rahmen des Projektes FIS-I an der TU München entwickelt wird, wird diese Zeit abermals verringern, wobei aber die Qualität der Daten deutlich steigt.

## Literatur


[Ac04] Achilles, A.-C. The Collection of Computer Science Bibliographies. 1994–2004. `http://liinwww.ira.uka.de/bibliography/`.

[DCMI] Dublin Core Metadata Initiative. `http://www.dublincore.org/`.

[DCMI03] Dublin Core Metadata Initiative. Dcmi metadata terms. November 2003. `http://www.dublincore.org/documents/dcmi-terms/`.

[Ku00] Kushmerick, N.: Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*. 118(1-2):15–68. 2000.

[MMK01] Muslea, I., Minton, S., und Knoblock, C. A.: Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*. 4(1/2):93–114. 2001.

[OAI03] Open Archives Initiative. The Open Archives Initiative Protocol for Metadata Harvesting. February 2003. `http://www.openarchives.org/OAI/openarchivesprotocol.html`.