Weiterentwicklung des virtuellen Bibliotheksregals *MyShelf* mit *Semantic Web*-Technologie: Erste Erfahrungen mit informationswissenschaftlichen Inhalten

Ralph Kölle, Thomas Mandl, René Schneider, Robert Strötgen

{koelle|mandl|rschneid|stroetgen}@ uni-hildesheim.de
Universität Hildesheim, IFAS
Marienburger Platz 22
31141 Hildesheim
Tel.: 05121/883-847
Fax: 05121/883-802

Abstract: Die wachsenden Datenmengen im Internet und das sich verändernde Informationsverhalten von Benutzern stellen eine große Herausforderung für Bibliotheken und andere Informationsanbieter dar. Das virtuelle Bibliotheksregal *MyShelf* erlaubt Benutzern einen flexiblen Zugriff durch das Wechseln zwischen verschiedenen Ontologien (*Ontology Switching*).

Dieser Beitrag stellt die Weiterentwicklung des virtuellen Bibliotheksregals *MyShelf* mit Techniken des *Semantic Web* (RDF, OWL) im Rahmen eines Projektseminars im Studienschwerpunkt Angewandte Informationswissenschaft vor und diskutiert die dadurch entstandenen Chancen für die Integration weiterer Wissensquellen und weiterer Mehrwerte. In Kooperationen mit dem FIZ Karlsruhe und der Universitätsbibliothek Hildesheim werden dabei bibliografische Daten und Patente einbezogen.

1 *Ontology Switching* und das virtuelle Bibliotheksregal *MyShelf*

Ontology Switching (Mandl & Womser-Hacker 2002) greift das Problem der Heterogenität auf zwei Ebenen auf. Zum einen erfordert es zunächst *Knowledge Engineering* zur Behandlung der semantischen Heterogenität. (Hellweg et al. 2001, Strötgen & Kokkelink 2001) Aus der Perspektive des Benutzers ermöglicht es dann den flexiblen und integrativen Zugriff per Browsing auf heterogene Kollektionen.

Die Wissensgrundlage für *Ontology Switching* bilden Beziehungen zwischen Ontologie-Einträgen oder Dokumenten, welche die Basis-Systeme so nicht zur Verfügung stellen. Wenn etwa ein Dokument nur in einer Ontologie oder Beschreibungssprache repräsentiert ist, dann wird zunächst eine Repräsentation in einer oder mehreren anderen, im System vertretenen Ontologien erstellt. Diese Transfer-Beziehungen können intellektuell, halb-automatisch oder voll-automatisch erstellt werden. Wenn zum Beispiel ein Buch nur in einem Bibliothekssystem verschlagwortet wurde, dann muss zunächst eine Verschlagwortung in den anderen Bibliothekssystemen erstellt oder abgeleitet werden. Die Metapher für das *Ontology Switching* ist das virtuelle Bibliotheksregal, das seine Dokumente je nach Perspektive des Benutzers (also der gewählten Ontologie) neu anordnet.

Somit kann jeder Benutzer die Perspektive seiner Disziplin einnehmen. In einem betrieblichen Umfeld kann etwa jeder Mitarbeiter die Dokumente aus der Sicht seiner Abteilung und deren Anforderungen organisieren lassen.

Ontology Switching führt zu folgenden Mehrwerten in Informationssystemen:

- Eine Benutzungsoberfläche für Browsing dient für mehrere Ontologien und Kollektionen.
- Die Reichweite einzelner Ontologien wird erhöht.
- Der Benutzer kann die Perspektive wechseln.
- Die Initiative für den Perspektivenwechsel liegt beim Benutzer und resultiert nicht aus Einschränkungen des Systems.
- Bereits getroffene Auswahl-Zustände bleiben auch beim Perspektivenwechsel erhalten.
- Durch die Darstellung mehrerer Ontologien in einem Informationssystem wird Transparenz in einem semantisch heterogenen Umfeld geschaffen.

Das Konzept des *Ontology Switching* wurde erstmals im System *MyShelf* implementiert und als ein virtuelles Bibliotheksregal für informationswissenschaftliche Inhalte realisiert.

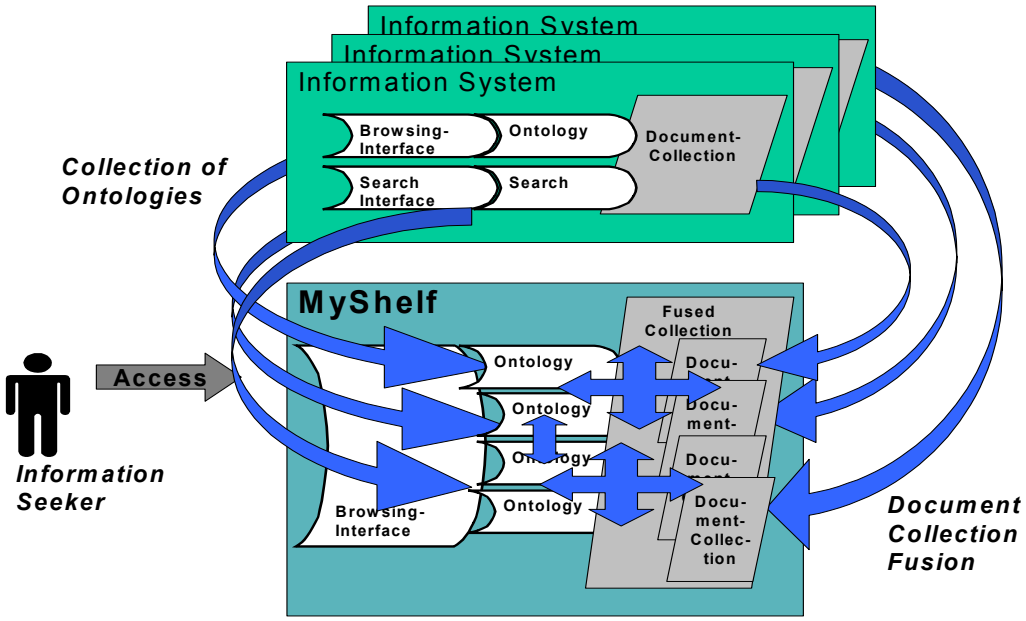

Abb. 1: Architektur *Ontology Switching* in *MyShelf*

In Kooperation mit der Universitätsbibliothek Hildesheim entstand *MyShelf* als Eigenentwicklung der Informationswissenschaft an der Universität Hildesheim (Hanke 2002, Heinz 2003). Die Informationswissenschaft an der Universität Hildesheim trägt zum Curriculum des Magisterstudiengangs Internationales Informationsmanagement[1] sowie zum BA-Studiengang Informationsmanagement/Informationstechnologie[2] bei. In beiden Studiengängen spielen andere Fächer wie Sprachwissenschaft, Interkulturelle Kommunikation oder BWL eine große Rolle. Diese Situation hat dazu geführt, dass keine eigene Signatur für die Informationswissenschaft existiert und relevante Literatur an sehr unterschiedlichen Orten und bei verschiedensten Fächern aufgestellt wird. Eine virtuelle Signatur Informationswissenschaft, welche die relevante Literatur an einem Ort zusammenfasst, ist daher dringend erforderlich, um die Literaturversorgung für die Studierenden zu verbessern.

Zur Erschließung des relevanten Buchbestandes in der UB Hildesheim erfolgte ein Abgleich mit dem Literaturangebot anderer Bibliotheken. Dabei boten sich Standorte informationswissenschaftlicher Studiengänge an, die Bibliotheken der Universitäten Konstanz

[1] http://www.uni-hildesheim.de/~angsprwi/iim.html
[2] http://www.imit.uni-hildesheim.de/

und Regensburg wurden ausgewählt. In einem Abgleich wurden 6000 Titel ermittelt, die in *MyShelf* momentan verfügbar sind. In einem halbautomatischen Verfahren wurden die Repräsentationen für diese Bücher in allen Bibliothekskatalogen erstellt (Hanke et al. 2002). Zusätzlich wurde eine eigene, für die Informationswissenschaft in Hildesheim optimierte Klassifikation entwickelt (HArmonized NomenKlature information sciencE, HANKE).

MyShelf wurde in HTML realisiert und kann mit einem Webbrowser benutzt werden.[3] Eine Evaluierung ergab, dass die Benutzer die Idee des *Ontology Switching* sehr positiv aufnehmen (Heinz 2003).

2 *Semantic Web* und Ontologien

Das aktuelle *World Wide Web* (WWW) ist für Maschinen nur sehr bedingt interpretierbar. Dokumente im WWW sind durch Hyperlinks verbunden, aber die Semantik dieser Verbindung (z.B. Belege für eine Aussage, andere Versionen eines Dokuments oder Informationen über den Autor) ist normalerweise nicht explizit und kann in der Regel nur durch einen menschlichen Nutzer erkannt werden. Folglich sind Internet-Suchmaschinen kaum in der Lage, Informationen auf ihren Kontext hin zu differenzieren. Für einen Benutzer beispielsweise, der nach Dokumenten eines bestimmten Autors sucht, werden diese nicht von denen unterschieden, in denen der Autor behandelter Gegenstand ist.

An dieser Stelle setzt die Idee des *Semantic Web*[4], einer Erweiterung des WWW, an. Ressourcen, d.h. Informationsobjekte jeder Art, können mit Metadaten versehen werden. Beziehungen zwischen verlinkten Ressourcen können explizit gemacht werden, und Ressourcen selbst lassen sich typisieren (z.B. Software, Bild, Aufsatz oder Person).

Zentrale Bestandteile des *Semantic Web* sind daher Metadaten und Vokabulare (Taxonomien oder Ontologien) die zur Beschreibung verwendet werden. Ein zentraler Baustein ist das *Resource Description Framework* (RDF), das die Beschreibung von Ressourcen auf einer sehr allgemeinen Ebene ermöglicht und in der *eXtensible Markup Language* (XML) ausgedrückt werden kann. Dabei werden alle Informationen über die Ressource in Tripel der Struktur Subjekt, Prädikat und Objekt zergliedert, wobei jedes Objekt auch wiederum eine Ressource sein kann. So entsteht ein Graph, in dem sich z.B. alle Dokumente zusammenführen lassen, die einen gemeinsamen Autor haben. Unterschiedliche Metadatenstandards wie beispielsweise das *Dublin Core Element Set* lassen sich mit RDF anwenden und kombinieren. (Obrst et al. 2003, Hjelm 2001)

Ein weiterer Bestandteil sind Ontologien, die die Menge der verwendeten Begriffe einschränken und deren Bedeutung konsensual festlegen. Ein Wort oder Begriff wird dabei von seinem Konzept unterschieden, um z.B. Homonyme differenzieren zu können. Ontologien können dabei mehr oder weniger stark formalisiert sein, auch „klassische“ Thesauri und Klassifikationen mit ihren semantischen Relationen zwischen Klassen und Begriffen lassen sich darunter fassen. Für das *Semantic Web* werden Ontologien in der Regel in den älteren Ontologie-Sprachen DAML+OIL, in RDF oder seit einiger Zeit auch in der *Web Ontologie Language* (OWL) ausgedrückt. (Davies et al. 2003)

Die eingangs angesprochenen neuen Möglichkeiten bei der Suche nach Dokumenten im *Semantic Web* stellen nur eines der Potenziale dar. So lassen sich beispielsweise Ressourcen von Benutzern annotieren, Agenten können gezielt Informationen sammeln und interpretieren, logisches Schließen zur Erkennung von Widersprüchen ist möglich, Vertrauensnetze können gebildet werden.

---

[3] http://web1.bib.uni-hildesheim.de/edocs/2003/363197524/meta/
[4] http://www.w3.org/2001/sw/

Eines der wesentlichen Probleme, das mit dem *Semantic Web* angegangen wird, ist das der semantischen Heterogenität. Die Nähe zum virtuellen Bibliotheksregals *MyShelf* ergibt sich hier vor allem durch die Integration von Dokumentenbeständen und den Einsatz von Ontologien zur benutzerorientierten Erschließung der Dokumente. Eine Weiterentwicklung von *MyShelf* soll vorhandene Potenziale nutzbar machen und gleichzeitig den beteiligten Studierenden die Möglichkeit geben, diese am konkreten Beispiel auszuloten.

3 Maschinelles Lernen

Wie die Mehrzahl der Arbeitsgebiete innerhalb der Informationswissenschaft bietet sich das Thema „Semantic Web und Ontologien“ für einen Einsatz von Methoden des Maschinellen Lernens an. Insbesondere auf dem Gebiet des (semi-)automatischen Erwerbs von Ontologien zeigt sich aktuell eine äußerst rege Forschungstätigkeit, die sich teils in der Reaktivierung bewährter Verfahren, teils in neuen Ansätzen äußert (Doan et al. 2003, Noy & Musen 2003, Maedche & Staab 2001). Das Hauptziel dieser Verfahren besteht in der Regel in der Überwindung des sog. ‚domain acquisition bottleneck‘ und der damit verbundenen Reduktion der ‚Handarbeit’ während der Datenerfassung und -aufbereitung. In diesem Zusammenhang kristallisieren sich dabei zwei Hauptanwendungsfelder heraus: einmal die (semi-) automatische Erstellung von Beschreibungsstrukturen für Rohdaten, zum anderen das Matching bzw. Mapping bereits bestehender ontologischer Strukturen.

Das Hauptkriterium für die Auswahl eines geeigneten Paradigmas bzw. eines konkreten Verfahrens besteht hierbei insbesondere in der Struktur (d.h. der Quantität bzw. Qualität) der zur Verfügung stehenden Daten, die zu Trainings- und Testzwecken bereitstehen. Erst daran schließt sich die Wahl eines entsprechenden Algorithmus an, der zu den erwünschten Ergebnissen führen soll.

Als grundsätzliche Einschränkung ist gleichfalls zu beachten, dass Ontologien aufgrund ihrer besonderen Struktur (Ontologien sind Wissensrepräsentationen in Form von operationalen Semantiken.) eine besondere Herausforderung für den automatischen Wissenserwerb darstellen und somit im Rahmen eines einsemestrigen Projektseminars ausschließlich Teilprobleme bearbeitet werden können, die innerhalb dieses eingeschränkten Zeitrahmens zu bewältigen sind.

4 Weiterentwicklung von *MyShelf* mit *Semantic Web*-Technologien

Zu der dargestellten Thematik fand im Wintersemester 2003/04 im Studiengang Internationales Informationsmanagement an der Universität Hildesheim, Schwerpunkt Angewandte Informationswissenschaft, ein Projektseminar im Hauptstudium statt. Fortgeschrittene Studierende hatten hier Gelegenheit, sich selbstständig einzuarbeiten und Teilaspekte in eigener Verantwortung zu bearbeiten. Sieben Arbeitsgruppen hatten als Folge dessen ihre Arbeiten zu koordinieren[5], um das Ziel einer Erweiterung des virtuellen Bibliotheksregals *MyShelf* mit Techniken des *Semantic Web* zu erreichen.[6]

Die zentrale Idee war eine Erweiterung des bisherigen *MyShelf*-Systems, in dem nur Literaturreferenzen verschiedener Unibibliotheken enthalten waren, um Patentdaten. Dadurch soll vor allem eine integrierte Recherche ermöglicht werden. Mit diesen beiden Domänen ergibt sich eine formale und semantische Heterogenität, die im Kontext des Projektseminars

[5] Zur Koordinierung der Arbeitsgruppen und zum Austausch von Daten und Informationen wurde der CSCW-Server der Fraunhofer-Gesellschaft (http://bscw.fit.fraunhofer.de/) eingesetzt.
[6] Weitere Informationen sind unter http://www.uni-hildesheim.de/~semweb/ verfügbar.

als Problemstellung gewollt ist. Durch diesen Ansatz wurde innerhalb der informationswissenschaftlichen Ausrichtung ein Schwerpunkt auf die Informationstechnologie gelegt, da hier am ehesten Patentierungen zu finden sind. Für die Biblitotheksdaten wurden als Klassifikationen die Sachgruppen der UB Hildesheim und die Pica-Basisklassifikation vorgefunden, für Patente die Internationale Patentklassifikation[7] (IPC). Die HANKE-Klassifikation, die in *MyShelf* schon genutzt und mit Bibliotheksdaten verknüpft war, sollte eine Brücke zu den Patentdaten herstellen.

Außer den schon genannten Magisterarbeiten von Peter Hanke (Hanke 2002) und Sabine Heinz (Heinz 2003) konnte auch auf Vorarbeiten vorangegangener Projektseminare zurückgegriffen werden, insbesondere auf das Projektseminar „Virtualisierung von Bibliotheksvorgängen“ im Wintersemester 2002/03.

Die Bearbeitung der Aufgabe lässt sich in drei Schwerpunkte aufteilen, die ihre Entsprechung in den gebildeten Arbeitsgruppen findet:

1. Modellierung der RDF-Daten und Konvertierung der OPAC- und Patentdaten
2. Evaluierung und Einsatz von RDF-Werkzeugen für Retrieval, Validierung, Visualisierung und andere Zwecke
3. Einsatz der HANKE-Klassifikation als OWL-Ontologie und als Brücke zwischen den Beständen mittels Methoden des *Knowledge Engeneering* und Maschinellen Lernens

Zwei Arbeitsgruppen beschäftigten sich mit der Modellierung und Konvertierung der RDF-Daten. Während sich eine Arbeitsgruppe damit beschäftigte, eine Suchmaschine für RDF-Daten bereitzustellen, evaluierte eine andere Arbeitsgruppe weitere RDF-Werkzeuge. Es wurden weiterhin zwei Arbeitsgruppen gebildet, die sich konkret mit den Einsatzmöglichkeiten von Verfahren zum Maschinellen Lernen auseinander setzten und geeignete Teilaufgaben bearbeiteten: Eine dieser Gruppen beschäftigte sich mit der semi-automatischen Domänenerweiterung, konkret mit der Klassifikation von Neuzugängen innerhalb einer virtuellen Repräsentation von verstreuten Buch- und Medienbeständen (Hanke 2003). Die zweite Gruppe setzte sich zum Ziel, Patentdaten anhand dieser informationswissenschaftlichen Bibliotheksklassen zu annotieren. Darüber hinaus wurde eine Arbeitsgruppe gebildet, die sich mit der Erstellung von Ontologien in OWL beschäftigte, mit dem Teilziel eine geeignete Datengrundlage für den semi-automatischen Ontologieerwerb zu erstellen.

Die Arbeiten und Ergebnisse werden im Folgenden dargestellt. Insgesamt ist zu berücksichtigen, dass innerhalb der Zeit und mit den Möglichkeiten der Studierenden bei dieser komplexen Aufgabenstellung keine fertigen Lösungen, sondern eher Lösungsansätze und prototypische Implementierungen zu erwarten sind.[8]

Das Projekt wurde durch Kooperationen mit der Universitätsbibliothek Hildesheim und dem Fachinformationszentrum Karlsruhe unterstützt, die vor allem Datenabzüge aus dem OPAC und der Patentdatenbank LPATDPA zur Verfügung stellten und für Fragen und freundliche Unterstützung zur Verfügung standen.[9]

[7] http://www.wipo.int/classifications/en/

[8] Da die Arbeiten zum Zeitpunkt der Schriftfassung dieses Beitrags noch nicht vollständig abgeschlossen waren, bleiben an einzelnen Punkten noch Leerstellen, die im Vortrag gefüllt werden.

[9] Dank an dieser Stelle vor allem an Herrn Michael Schwantner vom FIZ Karlsruhe und Herrn Benjamin Ahlborn von der UB Hildesheim.

4.1 Erstellung von RDF-Konvertern

Die Metadaten der beiden Datenbestände LPATDPA und OPAC der UB Hildesheim liegen in unterschiedlichen Formaten und mit unterschiedlichen Strukturen vor.

Ziel des Arbeitspakets „Erstellung von RDF-Konvertern“ war die Konvertierung der Metadaten nach RDF, wobei die prinzipielle Entscheidung zu treffen war, welche Metadaten-Standards dabei angewendet werden sollen und wie diese dann korrekt in RDF umgesetzt werden können.

Die RDF-Daten, die aus diesem Arbeitspaket hervorgehen, werden v.a. von den Arbeitspaketen „Exploration von RDF-Tools“ (vgl. Kap.4.2) und „Retrieval auf RDF-Daten“ (vgl. Kap. 4.3) weiter verwendet.

Das Arbeitspaket lässt sich aufteilen in die Bereiche „Parsen vorhandener Daten“ und „Generierung der RDF-Daten“. Beiden Paketen liegt eine entsprechende (objektorientierte) Modellierung zugrunde, dessen Kern die Schnittstelle zwischen beiden Bereichen beschreibt.

4.1.1 Parsen vorhandener Daten

Ziel dieser Arbeitgruppe war die Programmierung eines Parsers (in Java), der einerseits die vorhandenen Literatur-Datenbestände der UB Hildesheim, andererseits die Patent-Datenbanken des FIZ Karlsruhe in die vorgegebene Schnittstelle zur Generierung von RDF-Daten konvertiert. Die Daten der Bibliothek lagen in einem proprietären Pica-Format des OPAC-Systems vor, die Daten der Datenbank PATDPA im Messenger-Ausgabeformat.

Nach der Auswahl geeigneter Tools und Bibliotheken und der Modellierung der Schnittstelle zur Arbeitsgruppe „Generierung von RDF-Daten“ war die Grundsatzentscheidung zu treffen, ob ein Parser-Generator (yacc, ANTLR) eingesetzt werden soll. Man entschied sich aufgrund des Einarbeitungsaufwands dagegen und verarbeitete die Eingabedaten mit einem einfachen zeilenorientierten Java-Programm.

Die Schnittstelle besteht im Wesentlichen aus zwei Klassen, einer `Document`-Klasse, deren Objekte die Bücher oder Patente repräsentieren und mittels geeigneter get-Methoden Metadaten liefert. Die zweite Klasse ist der Container für `Document`-Objekte und ist mit `DocumentStore` bezeichnet.

Darüber hinaus waren die Klassifikationen entsprechend einzulesen. Für die HANKE-Klassifikationen, die nicht direkt im OPAC enthalten sind, waren die Zuordnungen zu Werken der Bibliothek aus zusätzlichen Tabellen auszulesen. Für die Zuordnung von Notationen zu Bezeichnungen und weiteren Informationen einer Klasse wurde eine zusätzliche Schnittstelle erzeugt.

Durch diese Schnittstellen war es fortan möglich, dass beide Gruppen unabhängig voneinander ihren Teil der Software weiterentwickelten.

4.1.2 Generierung von RDF-Daten

Ziel dieser Arbeitsgruppe war es, Metadaten der beiden Datenbanken nach RDF zu konvertieren. Dazu war zunächst die Einarbeitung in die unterschiedlichen verwendeten Standards notwendig, um anschließend die Daten modellieren zu können. Nach einer semantischen Analyse der Formate konnte die Entscheidung für einen gemeinsamen Metadatenstandard zur Konvertierung in RDF vorgenommen werden. Anhand der semantischen Analyse konnten erste RDF-Beispieldatensätze manuell erzeugt und mit RDF-Validatoren auf ihre Korrektheit hin überprüft werden. Ziel war es, ein Java-Programm zu schreiben, das die RDF-Struktur als DOM-Baum repräsentiert, und das die Daten aus beiden

Datenbeständen zu XML-Dateien generiert, so dass Retrieval-Maschinen darauf integriert zugreifen können.

Die Metadaten der Patentdatenbank des FIZ und der Literaturdatenbank der UB liegen in unterschiedlichen Formaten und Strukturen vor. Das FIZ verwendet einen speziellen Patentcode zur Beschreibung der Patente und die Universitätsbibliothek PICA+. Beide Ressourcentypen, sowohl das Patentdatenformat, als auch das Bibliotheksdatenformat, stellen unterschiedliche Ansprüche an die Metadaten, die sie beschreiben müssen, so dass jeweils eine andere Auswahl von Elementen bei diesen beiden unterschiedlichen Ressourcentypen zu beachten ist.

Zunächst sollte aus einer Reihe möglicher Metadatenschemata eines ausgewählt werden, das sowohl zum Ressourcentyp „Patentschrift" als auch zum Ressourcentyp „Buch" passt. Beide Typen stellen unterschiedliche Anforderungen an die Elemente, die ein Metadatenstandard zur Verfügung stellt. Nur zwei Metadatenstandards kamen daher in die engere Auswahl:

- MODS (*Metadata Object Description Schema*)[10] und
- „*Dublin Core*" (DCMI *Metadata Terms*)[11]

MODS ist XML-basiert und wird meist für digitale Bibliotheken verwendet. Es gibt 19 Top-Level-Elements und noch zahlreiche untergeordnete Sub-Elements. Alle Elemente außer titleInfo sind fakultativ. Die Elemente und Subelemente sind beliebig wiederholbar (außer recordInfo).

Dublin Core (DC) ist ein relativ einfacher Metadatenstandard, der zur domänenübergreifenden Suche dient. Das DC *Metadata Element Set* besteht aus 15 Elementen, die jeweils noch durch *Qualifier* erweitert werden können.

Die Struktur von *Dublin Core* lässt sich wesentlich einfacher in die Tripel-Struktur von RDF umsetzen als die komplexe und verschachtelte Struktur von MODS. Aufgrund der einfacheren Handhabbarkeit entschied man sich daher für den DC-Standard.

Für die Umsetzung der beiden Datenbestände in *Dublin Core* in Anlehnung an die Empfehlungen der DCMI[12] waren einige Erweiterungen erforderlich. So wurden in einem RDF-Schema[13] die beteiligten Klassifikationen als Codierungs-Schemata definiert und weitere Felder wie z.B. für die Patente der Erfinder als Element-Verfeinerung von `DC.Creator` erzeugt.

Die semantische Analyse sollte im Folgenden relevante Metadaten und deren Codierungsform (PICA+ in der Bibliotheksdatenbank) hervorbringen. Sind die wesentlichen Metadaten und deren ursprüngliche Codierung bekannt, lassen sie sich später gezielt und eindeutig aus den Testdatensätzen herausparsen.

Nach der Modellierung der Schnittstelle und der semantischen Analyse war der nächste Schritt, ein Java-Programm zu schreiben, das aus den geparsten Daten der ersten Arbeitsgruppe (vgl. Kap. 4.1.1) RDF-Daten generierte. Dazu bediente man sich der XML-Bibliothek Apache Xerces, um die RDF-Dokumente als DOM-Baum zu erzeugen und dann in einer XML-Datei zu serialisieren.

---

[10] http://www.loc.gov/standards/mods/
[11] http://dublincore.org/documents/dcmi-terms/
[12] http://dublincore.org/documents/2002/07/31/dcmes-xml/
[13] http://www.uni-hildesheim.de/~semweb/submodel.rdf

4.2 Exploration von RDF-Tools

Im Arbeitspaket „Erstellung von RDF-Konvertern“ wurden RDF-Daten generiert, die Arbeitspakete „Ontologien in OWL“ und „Ontologie-Erwerb“ lieferten Ontologien in OWL. Ziel dieses Arbeitspakets war es, vorhandene RDF-Werkzeuge zu sichten und auszuprobieren und dabei zu erkunden, welcher Mehrwert sich durch den Einsatz von RDF gegenüber den Ausgangsdatenbanken erzeugen lässt.

4.2.1 RDF-Validierung

Evaluiert wurde der „Validating RDF Parser“[14] (VRP) des *Institute of Computer Science of the Foundation for Research and Technology*. Es ist eines der Werkzeuge der ICS-FORTH RDF-Suite, die neben dem VRP auch aus der Schema-Specific Database (RSSDB) und einem Interpreter für die RDF Query Language (RQL) besteht.

VRP ist ein Tool, um RDF-Schemata und -Ressourcenbeschreibungen zu analysieren, zu überprüfen und zu verarbeiten. Es kann mit RDF-Daten umgehen, die in XML- oder in HTML-Dateien eingebettet bzw. in Unicode dargestellt sind. Der Parser analysiert die Syntax der RDF-Daten gemäß der aktuellen „RDF Model & Syntax Specification“ des W3C. Der Validierer prüft dagegen, ob die geparsten RDF-Statements sowohl in der Ressourcenbeschreibung als auch in dem zugehörigen RDF-Schema den Vorgaben aus der „RDF Schema Specification“ entsprechen.

4.2.2 RDF-Visualisierung

Weiterhin wurde „ezOWL“[15], ein Visualisierungs-Plugin für den Ontologie-Editor Protégé[16] des *Department of Internet Computing* des *Electronics and Telecommunications Research Institute* in Korea, evaluiert. Die OWL-Daten werden als Graph visualisiert und können editiert werden.

Als ein weiteres Visualisierungs-Plugin für Protégé wurde „Jambalaya“[17], das von der *CHISEL Software Engineering* Gruppe der Universität von Victoria in Kanada entwickelt wird, evaluiert. Es wird auch als hierarchischer Ontologie-Browser bezeichnet und soll interaktives Editieren existierender Daten ermöglichen. Die Visualisierung erfolgt hier nicht als Graph, sondern mittels Technik SHriMP (Simple Hierarchical Multi-Perspective), einer Visualisierungstechnik zur Untersuchung komplexer Daten.

Schließlich wurde „IsaViz“[18] evaluiert. IsaViz ist eine visuelle Umgebung, um RDF-Daten als Graph darzustellen, RDF-Dateien zu erstellen und zu ändern. Es wurde von Emmanuel Pietriga, Wissenschaftler beim W3C, entwickelt. Die Benutzeroberfläche erlaubt das Zoomen und Navigieren durch den Graph. RDF/XML, Notation 3 und N-Triple können importiert werden. Diese Typen können auch wieder exportiert werden, zusätzlich kann man die Daten auch als SVG oder PNG exportieren.

Zum Zeitpunkt der Schriftfassung wird darüber hinaus noch das Werkzeug „MR³“[19] getestet.

---

[14] http://athena.ics.forth.gr:9090/RDF/VRP/
[15] http://iweb.etri.re.kr/ezowl/
[16] http://protege.stanford.edu/
[17] http://shrimp.cs.uvic.ca/jambalaya/jambalaya_intro.shtml
[18] http://www.w3.org/2001/11/IsaViz/
[19] http://mmm.semanticweb.org/mr3/

4.3 Retrieval auf RDF-Daten

In diesem Arbeitspaket soll unter Einsatz vorhandener RDF-Tools (die Auswahl fand in Kooperation mit dem Arbeitspaket „Exploration von RDF-Tools“ statt) ein Retrieval-System aufgesetzt werden, mit dem die generierten RDF-Daten für Recherchen genutzt werden können. Dabei sollen möglichst auch die OWL-Ontologien der Arbeitspakete „Ontologien in OWL“ und „Ontologie-Erwerb“ integriert werden, z.B. für automatische Anfrageerweiterung.

Ergebnis erster Recherchen war die Favorisierung des mittels JAVA implementierten eRQL-Prozessors eRqlEngine.[20] Die eRql-Engine benutzt eine neu entwickelte Abfragesprache für RDF namens eRQL (*easy RQL*). eRQL wurde auf Basis der bereits bestehenden Abfragesprache RQL entwickelt. Mittels dieser ist es möglich, einfache Ein-Wort-Abfragen zu stellen. eRqlEngine verfügt über eine grafische Benutzerschnittstelle und ermöglicht selbst für den Laien relativ einfach, eine Anfrage zu stellen. Die eRqlEngine reicht eRQL-Abfragen an den RQL-Prozessor (RqlEngine) weiter. RqlEngine selbst basiert auf dem RDF-Parser VRP (Validating RDF-Parser, vgl. Kap. 4.2.1), der jedoch durch andere RDF-Parser ersetzt werden kann.

Ursprüngliches Ziel der Arbeitsgruppe war die Implementierung eines eigenen Retrieval-Systems für RDF und OWL. Dazu bediente man sich schließlich des Frameworks Jena 2.0.[21] Jena 2.0 ist eine JAVA API für RDF und wurde von dem *Semantic Web*-Team von *Hewlett & Packard* entwickelt. Im Toolkit von Jena 2.0 ist der RDF-Parser ARP (*Another RDF Parser*) enthalten. Des Weiteren stellt es RDQL (*RDF Data Query Language*)-konforme Schnittstellen zur Verfügung, die letztlich von der Arbeitsgruppe benutzt wurden. Alles, was zu tun blieb, war die Entwicklung einer geeigneten Benutzeroberfläche und die Analyse der RDQL-Schnittstelle von Jena, um schließlich beides zu verknüpfen.

4.4 Zuordnung von Patendaten zu informationswissenschaftlichen Bibliotheksklassen

Im Konzept des *Semantic Web* erweitert die *Web Ontology Language* (OWL) die Möglichkeiten von RDF um Vokabulare mit formaler Semantik. Eines der Arbeitspakete sollte daher vorhandene Thesauri und Klassifikationen (beziehungsweise Teile davon) in die *Web Ontology Language* überführen, damit diese von weiteren Arbeitsgruppen (etwa den Arbeitsgruppen „Exploration von RDF-Tools“, „Retrieval auf RDF-Daten“ oder „Maschineller Ontologieerwerb“) eingesetzt werden konnten.

Damit die Daten möglichst ohne großen Mehraufwand weiter verarbeitet werden konnten, entschied sich die Arbeitsgruppe nach kurzer Zeit für das Format OWL Lite[22] für die Umsetzung der HANKE-Klassifikation.

Im ersten Arbeitsgang wurde die HANKE-Klassifikation in OWL umgesetzt. Anschließend wurde die Struktur der Klassifikation, die aus einer primären Klassenebene mit den von Hanke definierten Sachgebieten und darunter liegenden Klassen mit einer weiteren Schicht von Unterklassen besteht, nach OWL überführt und einer Plausibilitätsprüfung unterzogen. Da die OWL Spezifikation relativ neu ist, existieren derzeit wenige Tools zur Validierung, von denen die meisten jedoch frei verfügbar sind.

Anschließend wurden die Klassen der Kategorie „Informationstechnologie und EDV“ in OWL formal beschrieben und die vorhandene Klassifikation so intellektuell erweitert und im

---

[20] http://gorleben.dbis.informatik.uni-frankfurt.de/~tolle/RDF/eRQL/eRQLEngine/
[21] http://jena.sourceforge.net/
[22] Es gibt außerdem die OWL-Varianten OWL DL (für Description Logic) und OWL Full.

Sinne einer formalen Ontologie formalisiert. Dafür wurde Protégé-2000 Version 2.0.1 benutzt.

Die nächsten Schritte sahen eine Verknüpfung der Daten mit den Resultaten anderer Arbeitsgruppen vor. Dabei wurden die HANKE-Klassen der OWL-Datei mit den OPAC-Daten durch ein `rdf:resource`- oder `rdf:about`-Attribut verknüpft.

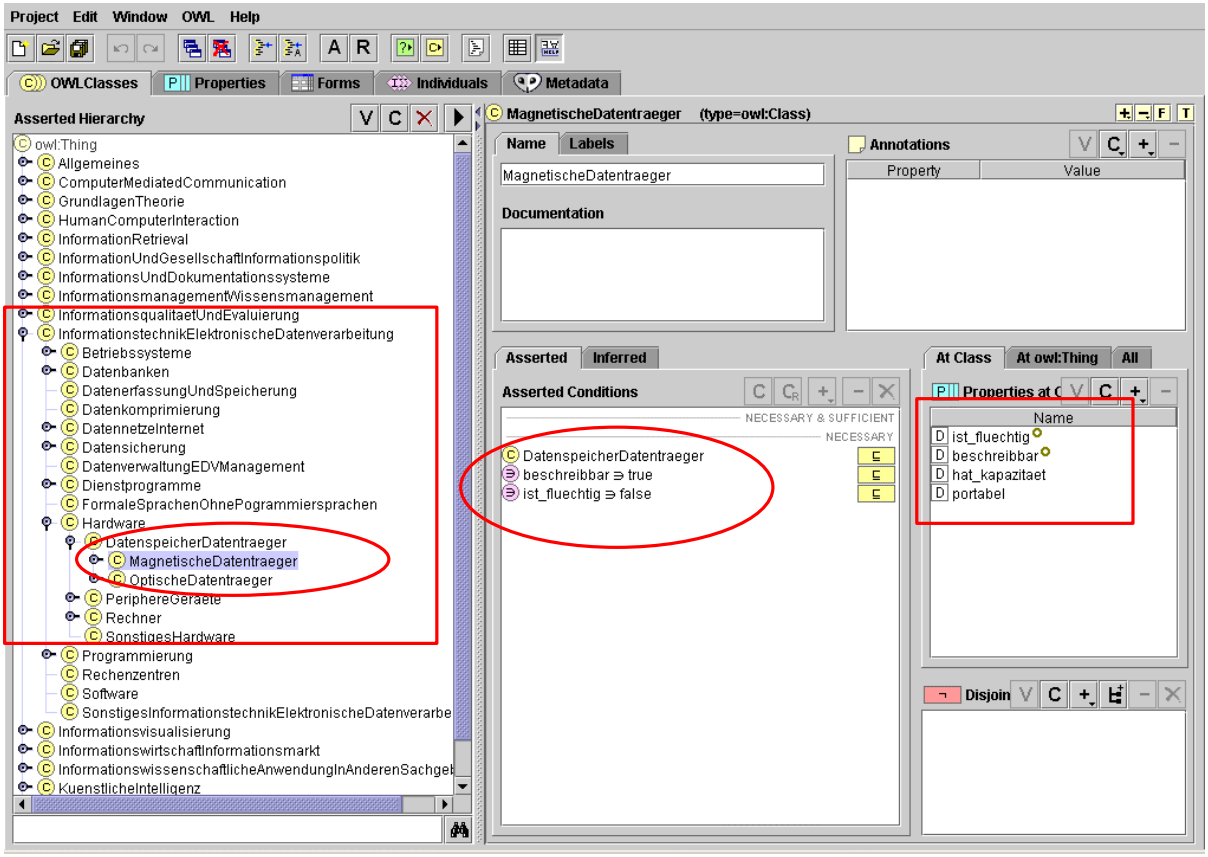

Abb. 2: Modellierung einer OWL-Klasse mit Protégé.

4.5 Klassifikation von Neuzugängen zu informationswissenschaftlichen Bibliotheksklassen

Die Aufgabe einer weiteren Arbeitsgruppe aus dem Umfeld des Maschinellen Lernens bestand in der Erstellung eines Systems, welches unter Einbeziehung der vorhandenen Literaturbestände die Klassifizierung nach HANKE maschinell lernt und damit die automatische Einordnung von Neuzugängen bzw. eine semi-automatische Domänenerweiterung ermöglicht.

Das bedeutet im Zusammenhang mit der Schaffung eines virtuellen Bibliotheksregals zunächst die automatische Klassifizierung der vorhandenen Buch- und Medienbestände durch die HANKE-Klassifikation. Metadaten der bereits manuell nach HANKE klassifizierten Bücher mussten derart aufbereitet werden, dass sie einem Klassifizierer des WEKA-Projekts[23] (Witten & Eibe 2001) als Trainingsdatensatz übergeben werden können. Das Programm musste demnach lernen, eine Beziehung zwischen den verschiedenen Attributwerten (z.B. einem bestimmten Schlagwort) und den jeweiligen HANKE-Klassen herzustellen. Auf Grundlage der daraus folgenden Ergebnisse sollte schließlich eine automatische Klassifikation von Neuzugängen ermöglicht werden.

[23] http://www.cs.waikato.ac.nz/~ml/weka/

Ein RDF-Dokument enthält die aus den OPAC-Daten generierten Metadaten über den informationswissenschaftlichen Buchbestand. Das System extrahierte zunächst die für eine Klassifizierung relevanten Daten wie Titel, Schlagwörter, Sachgruppe, Pica-Klasse und HANKE-Klasse. Um ein gutes Klassifizierungsergebnis zu erzielen, war es notwendig, die Metadaten, die außer auf deutsch auch auf englisch vorliegen, ins Deutsche zu übersetzen und einer morphologischen Analyse zu unterziehen. Schlagwörter und Titel wurden deswegen in einzelne Tokens zerlegt, Stopwörter gelöscht und Sonderzeichen sowie Leerzeichen entfernt. Die Übersetzung wurde durch eine maschinelle Anbindung an das Übersetzungstool „*Babelfish*" der Suchmaschine Altavista[24] bewerkstelligt.

Anschließend wurden die auf ihre Grundformen reduzierten Daten in Merkmalsvektoren des von WEKA benötigten Arff-Formats überführt. Diese Matrizen wurden dann durch WEKA bereitgestellten Algorithmen entsprechend bearbeitet, um die Ähnlichkeit bzw. Distanz der einzelnen Merkmalsvektoren zu bestimmen und durch Auswahl geeigneter Schwellenwerte eine maschinelle Zuweisung der Bücher zu den entsprechenden HANKE-Klassen zu bewerkstelligen.

Da bei Drucklegung des Artikels dieser letzte Arbeitsschritt noch nicht abgeschlossen war, liegen leider noch keine diesbezüglichen Ergebnisse vor.

4.6 Erstellung einer Ontologie in OWL aus informationswissenschaftlichen Bibliotheksklassen

In einer weiteren Arbeitsgruppe wurde versucht, Patentdaten aus der Datenbank LPATDPA anhand der HANKE-Klassifikation zu klassifizieren und mittels WEKA semi-automatisch in die bestehende Struktur einzubinden. Das Teilziel bestand darin, herauszufinden, inwiefern die universitätseigene informationswissenschaftliche Klassifikation ausreicht, um Patente, die für das Fach Informationswissenschaften von Interesse sind, einzugliedern und ob ein automatisches Erlernen für die Zuordnung der Patentdaten in die bestehenden Klassen bei gegebener Datengrundlage möglich ist.

Man kann davon ausgehen, dass sich diese beiden Themenbereiche zumindest an einigen Stellen überschneiden. Die HANKE-Klassifikation strukturiert die informationswissenschaftlich relevanten Klassen der Bibiliothek Hildesheim. Die gesuchte Dokumentmenge sind also die informationswissenschaftlich relevanten Patente aus LPATDPA.

Vorbereitend war eine manuell-intellektuelle Kategorisierung von Testdaten notwendig, um das Trainingskorpus mit wichtigen Attributen, nämlich den HANKE-Klassen, zu versehen, so dass das Tool zum maschinellen Lernen anhand dieser Kriterien die automatische Zuordnung lernen und anwenden kann.

Nach einer Analyse des Aufbaus und der Strukturierung der IPC, nach welcher die Patente jeweils einer Hauptklasse und keiner bis mehreren ICS-Klassen zugeordnet sind, kristallisierte sich heraus, dass die Attribute TI (Titel), ICM (Hauptklasse), ICS (Nebenklassen), MCLM (Description) von Bedeutung sind. Diese Felder beinhalten Informationen, die für die Einordnung der Patente in die HANKE-Klassifikation wichtig sein könnten. Hierbei ist zu beachten, dass anstatt der deutschen Titel TI auch englische Titel vorkommen können, die eine Übersetzung ins Deutsche erfordern.

Analog zum in Kap. 4.5 beschriebenen Verfahren wurde das WEKA-Tool als geeignet für die Komponente des Maschinellen Lernens betrachtet, was eine Umwandlung der einfachen Patent-Textdateien in das für WEKA geeignete Arff-Format notwendig machte, deren

[24] http://babelfish.altavista.com/

Merkmalsvektoren die für die Beschreibung der Patente relevanten Attribute TI, ICM, ICS und MCLM in einer nominal-skalierten Ausprägung repräsentieren.

Für die Übersetzung englischsprachiger Titel wurde auf die in Kap. 4.5 beschriebene Arbeit zurückgegriffen und analog zu den Arbeitsschritten verschiedene Verfahren des WEKA-Tools getestet und interpretiert.

Da bei Drucklegung des Artikels dieser letzte Arbeitsschritt noch nicht abgeschlossen war, liegen noch keine diesbezüglichen Ergebnisse vor. Im Anschluss an die Trainingsphase wurden dann eine Reihe von Patentdaten entsprechend der gelernten Klassifikation der bibliotheksinternen Beschreibungsstruktur hinzugefügt. In diesem Zusammenhang wurden Überlegungen zur Evaluation der Ergebnisse angestellt, ausgehend von der Tatsache, dass es innerhalb der IPC-Klassifikation einige informationswissenschaftliche Cluster gibt, deren Patente eine größere Nähe zur HANKE-Klassifikation aufweisen müssten. Im Idealfall müssten es genau diese Patente sein, die problemlos in die virtuelle Klassifikation nach HANKE überführt werden könnten. Nach Abschluss der Klassifikation kann mittels dieses Gütekriteriums über die Leistungsfähigkeit des gewählten Algorithmus entschieden werden.

5 Zusammenfassung

Auch wenn noch einige Ergebnisse ausstehen, so ist doch bereits deutlich geworden, dass Techniken des *Semantic Web*, vor allem RDF und OWL, große Potenziale für die Integration heterogener Datenbestände und die Behandlung der damit verbundenen semantischen Probleme bieten. Es hat sich aber ebenso herausgestellt, dass sich einerseits viele entsprechende Werkzeuge in einem noch nicht ganz ausgereiften Zustand befinden und dass andererseits auch die Standardisierung, wie beispielsweise das Dublin Core Element Set in RDF zu codieren ist, noch keinen wirklich stabilen Zustand erreicht hat.

Insofern hat sich die Erweiterung des virtuellen Bibliotheksregals *MyShelf* weniger als erwartet auf ausgereifte und einsatzbereite Werkzeuge stützen können. Auch wenn die praktische Anwendung der Ergebnisse in Bibliotheken oder Fachinformationszentren aktuell noch auf zu wackeligen Füßen stände, lässt sich dennoch hoffen, dass die weitere Entwicklung des *Semantic Web* stabile Lösungen für die Behandlung semantischer Probleme in heterogenen Umgebungen liefern wird.

6 Literatur

Davies, John; Fensel, Dieter; Harmelen, Frank v. (2002): Towards the Semantic Web. Ontology-driven Knowledge Management. New York et al. : John Wiley & Sons.

Doan, AnHai; Madhavan, Jayant; Dhamankar, Robin; Domingos, Pedro; Halevy, Alon (2003): Learning to Match Ontologies on the Semantic Web. In VLDB Journal, Special Issue on the Semantic Web.

Fensel, Dieter; Hendler, James; Liebermann, Henry; Wahlster, Wolfgang (Hrsg.) (2003): Spinning the Semantic Web. Cambridge/London.

Hanke, Peter (2002): Neue Chancen und Möglichkeiten für Ordnungssystematiken durch Visualisierung: Anwendung am Beispiel der Erfassung und Klassifizierung des informationswissenschaftlichen Bücherbestandes der Universitätsbibliothek Hildesheim. Hildesheim, Mag.-Arb.

Hanke, Peter; Mandl, Thomas; Womser-Hacker, Christa (2002): Ein „Virtuelles Bibliotheksregal“ für die Informationswissenschaft als Anwendungsfall semantischer Heterogenität. In: Proceedings ISI 2002. S. 289–302.

Heinz, Sabine (2003): Realisierung und Evaluierung eines virtuellen Bibliotheksregals für die Informationswissenschaft an der Universitätsbibliothek Hildesheim. Hildesheim, Mag.-Arb.

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

Hjelm, Johan (2001): Creating the Semantic Web with Rdf. New York et al. : Wiley Publishing.

Mädche, Alexander; Staab, Steffen (2001): Learning Ontologies for the Semantic Web. In: Proceedings 2nd Intl. Workshop on the Semantic Web. Hongkong, China.

Mandl, Thomas; Womser-Hacker, Christa (2002): Virtual ontologies for browsing interfaces in digital libraries. In: Isaías, Pedro (Hrsg.): Proceedings of the 2nd International Workshop on New Developments in Digital Libraries (NDDL 2002). In conjunction with the 4th International Conference On Enterprise Information Systems (ICEIS). April 2, 2002. Ciudad Real, Spanien. S. 39–50.

Noy, N. F.; Musen, M. A. (2003). The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping. International Journal of Human-Computer Studies.

Obrst, Leo; Smith, Kevin T.; Daconta, Michael C. (2003): Semantic Web. A Guide to the Future of XML, Web Services, and Knowledge Management. New York et al.

Strötgen, Robert; Kokkelink, Stefan (2001): Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt CARMEN. In: Information research & content management; 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. Frankfurt am Main : DGI. S. 56–66.

Witten, Ian H.; Eibe, Frank (2001): Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen. München/Wien : Hanser.