Automatische Generierung von DDC-Notationen für Hochschulveröffentlichungen

---

Bachelorarbeit

vorgelegt von Maike Sommer

Hochschule Hannover

Fakultät III – Medien, Information und Design

Abteilung Information und Kommunikation

Studiengang Informationsmanagement

Matrikelnummer 1088390

Erstprüfer: Prof. Dr. Christian Wartena

Zweitprüferin: Prof. Dr. Jutta Bertram

Hannover, den 22. Mai 2012

Abstract

Das Thema dieser Bachelorarbeit ist die automatische Generierung von Notationen der Dewey-Dezimalklassifikation für Metadaten. Die Metadaten sind im Dublin-Core-Format und stammen vom Server für wissenschaftliche Schriften der Hochschule Hannover.

Zu Beginn erfolgt eine allgemeine Einführung über die Methoden und Hauptanwendungsbereiche des automatischen Klassifizierens. Danach werden die Dewey-Dezimalklassifikation und der Prozess der Metadatengewinnung beschrieben. Der theoretische Teil endet mit der Beschreibung von zwei Projekten. In dem ersten Projekt wurde ebenfalls versucht Metadaten mit Notationen der Dewey-Dezimalklassifikation anzureichern. Das Ergebnis des zweiten Projekts ist eine Konkordanz zwischen der Schlagwortnormdatei und der Dewey-Dezimalklassifikation. Diese Konkordanz wurde im praktischen Teil dieser Arbeit dazu benutzt um automatisch Notationen der Dewey-Dezimalklassifikation zu vergeben.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Apolda	Automated Processing of Ontologies with Lexical Denotations for Annotation
BASE	Bielefeld Academic Search Engine
CREOLE	Collection of REusable Objects for Language Engineering
DC	Dublin Core
DDC	Dewey-Dezimalklassifikation
DFG	Deutsche Forschungsgemeinschaft
DNB	Deutsche Nationalbibliothek
DOI	Digital Object Identifier
EPC	(Decimal Classification) Editorial Policy Committee
GATE	General Architecture for Text Engineering
GKD	Gemeinsame Körperschaftsdatei
HTML	Hyper Text Markup Language
http	Hypertext Transfer Protocol
ID	Identifikator
JAPE	Java Annotaion Patterns Engine
KOS	Knowledge Organisation System
KVK	Karlsruher Virtueller Katalog
LCSH	Library of Congress Subject Headings
LGPL	Lesser General Public License
LoC	Library of Congress
MACS	Multilingual Access to Subjects

OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting
OCLC	Online Computer Library Center
OWL	Web Ontology Language
PDF	Portable Document Format
PICA	Project of Integrated Catalogue Automation
PND	Personennamendatei
POS-Tagger	Part-of-Speech-Tagger
RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RDF	Resource Description Framework
RSWK	Regeln für den Schlagwortkatalog
RTF	Rich Text Format
SerWisS	Server für wissenschaftliche Schriften der Hochschule Hannover
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organisation System
STTS	Stuttgart-Tübingen-Tagset
SVM	Support Vector Machines
SWD	Schlagwortnormdatei
TF/IDF	Term Frequenz/Inverse Dokument Frequenz
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Abbildungs- und Tabellenverzeichnis

Einleitung

In den letzten Jahren hat die Zahl von elektronischen Dokumenten immer mehr zugenommen. Auf Grundlage dieser Entwicklung lässt sich für die Zukunft ein weiteres Ansteigen der Datenflut prognostizieren. Im World Wide Web ist die Suchmaschine Google das „Maß der Dinge“ wenn es um die Strukturierung und Aufbereitung weltweit verfügbarer elektronischer Dokumente geht. Daneben gibt es aber noch weitere Suchmaschinen wie Yahoo! oder Bing. Solche Unternehmen haben lange vor den Bibliotheken damit angefangen, automatische Erschließungsmethoden anzuwenden um digitale Informationen verfügbar zu machen. Aber auch in den Bibliotheken vollzieht sich langsam ein Wandel in Richtung Suchmaschinentechnologie. Es wird zum Beispiel daran gearbeitet die vielen verschiedenen Kataloge und Fachdatenbanken, auf die eine Bibliothek Zugriff hat, unter einer möglichst einheitlichen Suchoberfläche zu präsentieren. Bei diesem Prozess spielt automatische Inhaltserschließung eine sehr wichtige Rolle, denn um Dokumente einheitlich präsentieren zu können, müssen sie auch möglichst einheitlich erschlossen sein. Ein anderes wichtiges Argument für die automatische Inhaltserschließung, bei der Bewältigung der Informationsflut, ist der Zeitaspekt. In der Informationsgesellschaft des 21. Jahrhunderts ist es wichtig die elektronischen Dokumente möglichst zeitnah zur Verfügung zu stellen (vgl. Groß, Faden 2010, S. 1120).

> Angesichts der steigenden Anzahl elektronischer Inhalte und vor dem Hintergrund stagnierender bzw. knapper werdender personeller Ressourcen in der Sacherschließung schafft keine Bibliothek bzw. kein Bibliotheksverbund es mehr, weder aktuell noch zukünftig, alle digitalen Daten zu erfassen, zu strukturieren und zueinander in Beziehung zu setzen (Groß, Faden 2010, S. 1120).

Das automatische Klassifizieren, als ein Teilaspekt automatischer Inhaltserschließung, ist sehr interessant für Bibliotheken, denn dadurch bieten sich viele Möglichkeiten die elektronischen Ressourcen zu präsentieren. Den Bibliotheksnutzern können zum Beispiel Möglichkeiten zum thematischen Browsing geboten werden.

Fragestellung

Diese Bachelorarbeit beschäftigt sich mit der Möglichkeit einer automatischen Generierung von Notationen der Dewey-Dezimalklassifikation (im Folgenden: DDC) für Hochschulveröffentlichungen. Konkret handelt es sich dabei um Dokumente aus dem Server für wissenschaftliche Schriften der Hochschule Hannover (im Folgenden: SerWisS). Für das automatische Klassifizieren sollen aber nicht die Volltextdokumente,

sondern nur deren Metadaten verwendet werden. Die meisten Metadaten wie Verfasser, Erscheinungsort, Sprache etc. sind für diesen Zweck vollkommen irrelevant. Daher wird die Klassifikationsentscheidung nur auf Grundlage von Titel, Schlagwörtern und Zusammenfassung getroffen.

Der Versuch zur automatischen Generierung von DDC-Notationen wird mit der Open Source Software GATE durchgeführt. Das Text-Mining-Programm ist sehr flexibel und kann auf die unterschiedlichsten Bedürfnisse angepasst werden. Über Plugins können verschiedene Tools, die Texte annotieren, geladen werden. Diesem Baukastenprinzip ist es auch zu verdanken, dass eine Ontologie der Deutschen Nationalbibliothek für die Generierung der DDC-Notationen verwendet werden konnte. Es handelt sich hierbei um die Ergebnisse des CrissCross-Projektes, bei dem eine Konkordanz zwischen der Schlagwortnormdatei und der DDC erstellt wurde.

Ziel der Arbeit ist es herauszufinden, ob die vorhandenen Metadaten ausreichen um die Dokumente von SerWisS automatisch einer DDC-Notation zuzuordnen. Da die Metadaten bereits nach DDC klassifiziert wurden, ist es einfach eine anschließende Erfolgskontrolle durchzuführen.

GATE ist eine frei verfügbare Software. Im Gegensatz zu vielen anderen Methoden werden für die hier vorgestellte Methode, keine Trainingsdaten benötigt. Daher zeigt die Arbeit auf, was schon mit einfachen Mitteln im Bereich des automatischen Klassifizierens erreicht werden kann.

Forschungsstand

Zum Thema automatische Inhaltserschließung gibt es einige zusammenfassende Darstellungen auf Deutsch. Zu nennen wären da beispielsweise „Grundlagen automatischer Indexierung“ von Holger Nohr (2003) und „Text Mining: Wissensrohstoff Text“ von Gerhard Heyer, Uwe Quasthoff und Thomas Wittig (2008). Speziell zum Thema „automatisches Klassifizieren“ gibt es ein sehr gutes Buch von Otto Oberhauser (2005). Nach einer kleinen Einführung in das Thema wird ein Überblick über alle relevanten Klassifizierungsprojekte von Anfang der Neunziger bis ca. 2004 gegeben. Das Projekt „Automatische Anreicherung von OAI-Metadaten“ von der Universitätsbibliothek Bielefeld, welches in dieser Bachelorarbeit vorgestellt wird, ist in dem Werk von Oberhauser noch nicht enthalten. Es wurde erst 2011 beendet. Durchführung und Ergebnisse des Projekts werden in zwei englischsprachigen Aufsätzen von Mathias Lösch

(2011) und Ulli Waltinger (2011) beschrieben. Wie in den meisten aktuellen Projekten, so wurde auch für dieses Projekt ein maschinelles Lernverfahren gewählt. Einen sehr guten Überblick über maschinelle Lernverfahren bei der automatischen Klassifizierung gibt der Aufsatz von Fabrizio Sebastiani (2002). Für den praktischen Versuch in dieser Bachelorarbeit wurde kein maschinelles Lernverfahren angewendet. Die hier vorgestellte Herangehensweise wurde erst durch das 2011 erfolgreich abgeschlossene CrissCross-Projekt möglich. In der Literatur ist zum Zeitpunkt der Erstellung dieser Bachelorarbeit nichts über die Verwendung der CrissCross-Ergebnisse zum automatischen Klassifizieren gefunden worden.

Theoretischer Bezugsrahmen und Methodisches Vorgehen

Die Bachelorarbeit kann in das theoretische Umfeld des Text Mining, bzw. der automatischen Sprachverarbeitung eingeordnet werden. Eine Abgrenzung des automatischen Klassifizierens zu anderen Bereichen des Text Mining und eine Erörterung der verschiedenen Methoden erfolgt in den ersten beiden Kapiteln dieser Arbeit.

Die Arbeit besteht aus einem Theorieteil und der Dokumentation einer praktischen Aufgabe. Der Theorieteil ist dazu da um auf die praktische Aufgabe vorzubereiten und einen breiteren Kontext für die abschließende Beurteilung der Ergebnisse zu schaffen. Die praktische Aufgabe wurde mit einer Text Mining Software durchgeführt, daher ähnelt die Beschreibung des „Versuchsaufbaus“ einer Softwaredokumentation.

Aufbau der Arbeit

Die Bachelorarbeit ist in acht Kapitel untergliedert. Im ersten Kapitel erfolgt eine allgemeine Einführung in das Thema automatische Inhaltserschließung, bzw. Text Mining. In Kapitel zwei werden die unterschiedlichen Methoden des automatischen Klassifizierens vorgestellt. Sinn und Zweck, bzw. die Hauptanwendungsbereiche des automatischen Klassifizierens werden in Kapitel drei besprochen. Da für das automatische Klassifizieren nur Metadaten verwendet werden sollen, erfolgt eine Beschreibung des Metadatenformats in Kapitel vier. Die Zielklassifikation für das automatische Klassifizieren ist die Dewey-Dezimalklassifikation. Sie wird in Kapitel fünf beschrieben. Anschließend werden, in Kapitel sechs, zwei Projekte vorgestellt. Das Erste ist gewissermaßen ein Referenzprojekt, weil es ebenfalls eine automatische DDC-Klassifikation zum Ziel hatte. Das zweite Projekt (CrissCross) ist wichtig, weil dessen Ergebnis die praktische Aufgabe überhaupt erst möglich gemacht hat. In Kapitel sieben wird das Ergebnis dieses Projekts genauer

beschrieben. Danach erfolgt im letzten Kapitel eine Beschreibung der Text Mining Software GATE und ihrer Verwendungsweise für die praktische Aufgabe. Das Kapitel endet mit der Beschreibung der Ergebnisse. Die Arbeit schließt mit einer Zusammenfassung der Kernaussagen und einem Ausblick auf weiterführende Fragestellungen ab.

1 Automatische Inhaltserschließung, aka Text Mining

Unter dem Oberbegriff ***Text Mining*** werden alle Verfahren zur automatischen Wissensgewinnung aus natürlichsprachigen Texten zusammengefasst. Dies ist abzugrenzen von dem verwandten Begriff ***Data Mining***, der alle Verfahren und Methoden zur Wissensgewinnung aus Datenbanken beinhaltet. Während die Daten beim Data Mining bereits in strukturierter Form vorliegen und durch geeignete Abfragen „geholt“ werden können, müssen sie beim Text Mining zuvor noch interpretiert werden (vgl. Heyer, Quasthoff, Wittig 2008, S. 3-5).

Ein Unterscheidungsmerkmal von Text-Mining-Verfahren ist deren Ergebnis. Beim ***automatischen Indexieren*** wird zwischen Additions- und Extraktionsverfahren unterschieden. Beim *Extraktionsverfahren* werden die zentralen Terme aus einem Text als Stichwörter erkannt und extrahiert. Die so gewonnen Stichwörter können darüber hinaus noch mit einem Thesaurus verknüpft werden. Durch die Begriffsrelationen im Thesaurus entsteht zusätzliches Wissen. Es wurde quasi durch eine externe Wissensquelle hinzuaddiert (*Additionsverfahren*) (vgl. Nohr 2003, S. 20). Nun ist das Ergebnis nicht mehr nur ein Stichwort, sondern ein Deskriptor oder ein Schlagwort. Das Additionsverfahren kann auch als ein Teil des ***automatischen Klassifizierens*** angesehen werden, denn es ist nicht gesagt, dass die externe Wissensrepräsentation unbedingt ein Thesaurus sein muss. Es könnte sich genauso gut um ein Klassifikationssystem handeln und dann wäre das Ergebnis eine Notation. So gesehen ist das Additionsverfahren die Schnittstelle zwischen dem automatischen Indexieren und dem automatischen Klassifizieren. Davon abzugrenzen ist das ***automatische Clustern.*** Hierbei wird kein fertiges Klassifikationssystem benutzt. Die Kategorien werden erst beim Clustern anhand der vorliegenden Dokumente erstellt. Anders ausgedrückt, Texte mit ähnlichen Bedeutungen werden einander zugeordnet (vgl. Sebastiani 2002, S. 2).

2 Methoden des automatischen Klassifizierens

Dieses Kapitel beschäftigt sich zunächst ganz allgemein mit den Methoden des automatischen Klassifizierens von Texten. In der englischsprachigen Literatur werden hier die Bezeichnungen „Text Categorization“, „Text Classification“ oder auch „Topic Spotting“ synonym verwendet. Damit ist die Zuordnung von natürlichsprachigen Texten zu einem vordefinierten Set von thematischen Kategorien gemeint (vgl. Sebastiani 2002, S. 1). Im Bibliotheksbereich wird die Zuordnung von Klassen zu einem Dokument, bzw. natürlichsprachigen Texten mehrheitlich als „Klassieren“ bezeichnet. „Klassifizieren“ hingegen meint den Prozess der Klassenbildung (Greiner 1978, zit. nach Bertram 2005, S. 150-151). Das englische „classifying“ wird in der Bedeutung von „Klassieren“ verwendet. Die Literatur, welche sich mit der automatischen Inhaltserschließung von natürlichsprachigen Texten befasst, ist überwiegend englischsprachig. So wird auch im Deutschen häufig vom „automatischen Klassifizieren“ gesprochen, obwohl streng genommen das „automatische Klassieren“ gemeint ist. Aus diesem Grund erschien es zunächst sinnvoll sich an die bibliothekarische Perspektive zu halten und in dieser Arbeit immer streng zwischen „Klassifizieren“ und „Klassieren“ zu unterschieden. Doch selbst in der deutschen Übersetzung der Dewey-Dezimalklassifikation werden die Begriffe „Klassieren“ und „Klassifizieren“ inzwischen synonym verwendet. Vielmehr noch ist der Begriff „Klassifizieren“ hier der Gebräuchlichere. Im Inhaltsverzeichnis findet sich die Überschrift „Klassifizieren mit der Dewey-Dezimalklassifikation“ (vgl. Dewey, Mitchell 2005, S. vii) und im Glossar wird von „Klassieren“ (übersetzt mit „to class“) auf „Klassifizieren“ (übersetzt mit „to classify“) verwiesen. Hier wird „Klassifizieren“ als die Tätigkeit, „einem einzelnen Werk eine Notation zu[zu]ordnen“ definiert (Dewey, Mitchell 2005, S. lxxxiii). Aus den oben genannten Gründen wurde beschlossen sich an der Perspektive der deutschen Nationalbibliothek zu orientieren und vom „automatischen Klassifizieren“ zu sprechen.

Einen sehr guten Überblick über die verschiedenen Herangehensweisen zum automatischen Klassifizieren gibt der Aufsatz von Fabrizio Sebastiani (2002) mit dem Titel „Machine Learning in Automated Text Categorization“. Dabei wird zuerst immer zwischen ***maschinellen Lernverfahren*** und ***regelbasierten Verfahren*** unterschieden. Bei regelbasierten Verfahren werden von Experten Regeln aufgestellt, nach denen Texte annotiert werden. Der Nachteil dieser Verfahren ist, dass sie sehr aufwendig, d.h. kosten- und zeitintensiv sind, da das Expertenwissen zunächst in einen Algorithmus

umgeschrieben werden muss. Der so konstruierte „Klassifizierer“ kann dann auf die Dokumente angewendet werden. Der Vorteil dieser Methoden ist, dass keine Trainingsdokumente benötigt werden. Der Algorithmus kann sofort auf die zu klassifizierenden Texte angewendet werden. Maschinelle Lernverfahren benötigen im Gegensatz dazu hunderte von Trainingsdokumenten. Trainingsdokumente sind Texte, die bereits intellektuell klassifiziert wurden. Das heißt ein menschlicher Experte, in wissenschaftlichen Bibliotheken sind dies häufig Fachreferenten, hat jedes Dokument bereits einer Klasse zugeordnet. Diese Trainingsdokumente werden bei maschinellen Lernverfahren dazu benutzt automatisch einen „Klassifizierer“ zu erzeugen, in dem das Programm durch die vorklassifizierten Dokumente die Charakteristika einer Klasse lernt. „Lernen“ bedeutet es werden stochastische und statistische Methoden auf den Text angewandt. Das hat den Vorteil der Zeitersparnis, da kein Expertenwissen benötigt wird um den „Klassifizierer“ zu erzeugen. Des Weiteren kann ein maschinelles Lernverfahren ganz leicht auf ein anderes Klassifikationssystem angewendet werden. Dazu werden nur die entsprechenden Trainingsdokumente benötigt (vgl. Sebastiani 2002, S. 1-2).

Bei regelbasierten Verfahren wird auch manchmal von ***computerlinguistischen Verfahren*** gesprochen. Allerdings sind diese Begriffe nicht ganz deckungsgleich. Zu den computerlinguistischen Verfahren gehören neben den Regelbasierten- auch die ***wörterbuchbasierten Verfahren***. Diese Verfahren benutzen elektronische Wörterbücher oder Listen, um bestimmte Terme in einem Text zu referenzieren. Eine denkbare Anwendungsmöglichkeit wäre beispielsweise die Verwendung von Listen mit Namen von Firmen und Institutionen, um zu erkennen, ob diese in einem Text genannt werden[1] (vgl. Nohr 2003, S. 30).

Wenn kein Wörterbuch, sondern ein Thesaurus oder ein Klassifikationssystem verwendet wird, um Terme in einem Text zu referenzieren, dann wird von einem ***begriffsorientierten Verfahren*** gesprochen. Es handelt sich hierbei um ein Additionsverfahren. Der Vorteil liegt darin, dass nun auch die Bedeutungen von Wörtern erkannt werden. Ein solches Verfahren würde zum Beispiel erkennen, dass „Qualitätsmanagement“ und die gängige Abkürzung „QM“ dieselbe Bedeutung haben. Sie repräsentieren den gleichen Begriff (vgl. Nohr 2003, S. 79).

[1]Der ANNIE Gazetteer, des Text-Mining-Programms GATE, wäre ein praktisches Beispiel für ein wörterbuchbasiertes Verfahren.

In der Praxis werden die verschieden Verfahren zur automatischen Sprachverarbeitung meistens miteinander kombiniert. Allein für diese Bachelorarbeit wurden schon regelbasierte, wörterbuchbasierte und begriffsorientierte Verfahren angewendet. [2] Als Informationsgrundlage für das begriffsorientierte Verfahren in dieser Arbeit diente der Thesaurus der Deutschen Nationalbibliothek.

2.1 Eine Definition des Klassifizierens

Sebastiani (2002, S. 2-3) beschreibt das Klassifizieren von Texten ganz formal mit einer Funktion:

$$\phi : D \times C \rightarrow \{T, F\}$$

Im Kern geht es darum jedem Paar D und C einen Booleschen Wert (True oder False) zuzuordnen. Dabei steht D für die Menge aller Dokumente $D = \{d_1, \dots, d_n\}$ und C für alle Kategorien eines Klassifikationssystems $C = \{c_1, \dots, c_{|c|}\}$. Also ist $(d_j, c_i) \in D \times C$. Wenn das Dokument d_j der Kategorie c_j zugeordnet werden kann, folgt daraus das T zutrifft, also die Aussage Wahr (True) ist. Andernfalls trifft F zu. Die Aussage ist falsch (False).

Dabei ist zu berücksichtigen, dass die Entscheidung einem Dokument eine Klasse zuzuordnen nicht immer eindeutig ist. Es kommt zum Beispiel recht häufig vor, dass zwei verschiedene Personen das gleiche Dokument in zwei unterschiedliche Klassen einsortieren. Dieses Phänomen wird auch als Inter-Indexiererinkonsistenz bezeichnet. In der Praxis wird versucht, dem mit detailierten Regeln entgegenzutreten. Trotzdem ist es manchmal schwierig zu entscheiden, ob ein Artikel über den Besuch von Angela Merkel zum WM-Eröffnungsspiel bei Fußball oder bei Politik einsortiert werden sollte. Aber es kann auch sein, dass dieselbe Person das gleiche Dokument zu verschiedenen Zeitpunkten in unterschiedliche Klassen einsortiert. Dieses Phänomen wird als Intra-Indexierinkonsitenz bezeichnet (Lancaster 1998, S.68 zit. nach Bertram 2005, S. 73). Das Qualitätskriterium Indexierkonsistenz gibt also darüber Aufschluss inwieweit die semantische Bedeutung eines Textes zu unterschiedlichen Zeitpunkten und von unterschiedlichen Personen gleich erkannt, bzw. gleich interpretiert wird. Untersuchungen belegen, dass die Indexierkonsistenz in der Praxis teilweise bei unter 50 Prozent liegt (vgl.

[2] Obwohl ich das in der Literatur nicht explizit so gefunden habe, würde ich persönlich dazu tendieren regelbasierte, wörterbuchbasierte und begriffsorientierte Verfahren unter dem Oberbegriff computerlinguistische Verfahren zusammenzufassen.

Nohr 2003, S. 28). Dies muss auch bei der Beurteilung der Ergebnisse des automatischen Klassifizierens Berücksichtigung finden.

2.2 Einfach- oder Mehrfachklassifizierung

Einfachklassifizierung („single-label-categorization“) bedeutet, dass ein Dokument genau einer Klasse („category“) zugeordnet wird. Ein Spezialfall der Einfachklassifizierung ist die binäre Klassifizierung („binary categorization“). Das bedeutet ein Dokument kann entweder einer Klasse oder ihrem Gegenteil zugeordnet werden. Bei der Mehrfachklassifizierung („multi-label-categorization“) kann ein Dokument gleich mehreren Klassen zugeordnet werden. Von 0 bis zu sämtlichen Klassen, die in einem Klassifikationssystem vorhanden sind, ist alles möglich. Die binäre Klassifizierung kommt viel häufiger vor, als die Mehrfachklassifizierung. Das liegt auch daran, dass ein Algorithmus für die binäre Klassifizierung auch für die Mehrfachklassifizierung benutzt werden kann. Dazu muss die Klassifikationsentscheidung der Mehrfachklassifizierung nur für die Gesamtheit aller Klassen, auf binäre Entscheidungen reduziert werden.[3] Umgekehrt funktioniert das nicht. Ein Algorithmus für Mehrfachklassifizierung kann nicht für Einfachklassifizierung benutzt werden (vgl. Sebastiani 2002, S. 3-4).

2.3 Klassen- oder Dokumentenzentrierte Methoden

Es gibt zwei verschiedene Möglichkeiten einen „Klassifizierer“ zu benutzen:

1. Dokumentenzentriert: Einem Dokument werden eine oder mehrere Klassen zugeordnet. (DPC – „document-pivoted categorization“)
2. Klassenzentriert: Einer Klasse werden alle zu ihr passenden Dokumente zugeordnet. (CPC – „ category-pivoted categorization“)

Die Klassenzentrierte Methode bietet sich an, wenn eine neue Klasse zu einem bereits existierenden Klassifikationssystem, hinzugefügt werden soll. Dann müssen alle bereits klassifizierten Dokumente noch einmal neu bewertet werden. Die Dokumentenzentrierte Methode wird viel häufiger benutzt, da sie angewendet werden kann wenn Dokumente zu unterschiedliche Zeitpunkten klassifiziert werden müssen. In Bibliotheken ist diese Situation beispielsweise alltäglich. Aber auch das Filtern von E-Mails wäre so ein Fall (vgl. Sebastiani 2002, S. 4).

[3] Das funktioniert allerdings nur, wenn die Kategorien stochastisch unabhängig voneinander sind.

2.4 „Hartes“ oder Rangordnendes Klassifizieren

Eine vollständige Automatisierung des Klassifizierungsprozesses benötigt eine „harte“ T oder F Entscheidung für jedes Dokument-Klasse-Paar (d_j, c_i). Es kann aber auch ausreichen, den Prozess des Klassifizierens nur teilweise zu automatisieren. In diesem Fall würde der „Klassifizierer“ ein Ranking der Klassen in absteigender Reihenfolge nach dem Grad ihrer Passgenauigkeit zu einem Dokument ausgeben. Oder bei der klassenzentrierten Methode, den Grad der Zugehörigkeit der Dokumente zu einer Klasse. Nun könnte ein menschlicher Experte, z.B. ein Fachreferent, die ultimative Klassifikationsentscheidung treffen. Dieses semiautomatische, „interaktive“ Klassifizieren eignet sich besonders dann, wenn den Ergebnissen eines automatischen „Klassifizierers“ noch nicht vollständig vertraut werden kann. Dies könnte zum Beispiel der Fall sein wenn die Qualität der Trainingsdaten gering ist. Das heißt die Trainingsdaten repräsentieren nicht die Gesamtheit der zu klassifizierenden Dokumente (vgl. Sebastiani 2002, S. 4-5).

3 Hauptanwendungsbereiche des automatischen Klassifizierens

Im Folgenden werden fünf Anwendungsbereiche für das automatische Klassifizieren vorgestellt.

3.1 Automatisches Indexieren

In den 60er und 70er Jahren, den Anfängen der Forschung zum automatischen Klassifizieren, wurde hauptsächlich über das automatische Indexieren publiziert. Genauer gesagt über das automatische Indexieren mit Hilfe eines kontrollierten Vokabulars. Wenn die Einträge im Thesaurus als Kategorien angesehen werden, kann das automatische Indexieren von Dokumenten als ein Teil des automatischen Klassifizierens betrachtet werden (vgl. Sebastiani 2002, S. 5).

3.2 Organisation von Dokumenten

Im Arbeitsalltag lassen sich viele Probleme der Dokumentenorganisation durch Klassifikationssysteme lösen. Ein großer Zeitungsverlag würde beispielsweise von einem solchen System profitieren. So könnten die Anzeigen, die täglich das E-Mail-Konto füllen, schon automatisch in die entsprechenden Rubriken (Stellenanzeigen, Automarkt, Immobilienanzeigen, Partnerschaft, etc.) vorsortiert werden (vgl. Sebastiani 2002, S. 6).

Eine andere Anwendungsmöglichkeit wäre das Klassifizieren von Patientenakten in Krankenhäusern nach Diagnosen.

3.3 Textfiltern nach Themengebieten

Ein typischer Anwendungsbereich für das Textfiltern nach Themengebieten sind Newsfeeds, die von Nachrichtenagenturen an Zeitschriftenverlage gesandt werden. Das Textfiltern kann als eine Art der Einfachklassifizierung („single-label-categorization“) angesehen werden. Die einkommenden Texte werden in die zwei Kategorien relevant und irrelevant einsortiert. Im zweiten Schritt könnten dann die relevanten Dokumente weiter in thematische Kategorien einsortiert werden. Ein Filtersystem kann sowohl auf der Seite der News-Ersteller („adaptive filtering“), als auch auf der Seite der News-Empfänger („routing“, „batch filtering“) eingerichtet werden. In den Informationswissenschaften hat das Filtern von Dokumenten eine lange Tradition, die zurück in die 60er Jahre reicht. Der Begriff „selective dissemination of information“ bezeichnet die Vorselektion wissenschaftlicher Neuveröffentlichungen nach Relevanz für den Empfänger. Ein anderes aktuelles Beispiel für thematisches Textfilter ist die Einrichtung einer eigenen personalisierten Online-Zeitung (vgl. Sebastiani 2002, S. 6-7).

3.4 Homonym- und Polysemkontrolle

Das Problem der „Wortbedeutungsdisambiguierung“ (engl. „word sense disambiguation“) kann auch als Klassifikationsproblem betrachtet werden. Die Kontexte der mehrdeutigen Wörter werden einfach als Dokumente angesehen und die verschiedenen Bedeutungen als Klassen. Dann kann dieses Problem mit einer dokumentenzentrierten Methode („document-pivoted categorization“) der Einfachklassifizierung („single-label-categorization“) gelöst werden (vgl. Sebastiani 2002, S. 7).

3.5 Hierarchisches Klassifizieren

Beim hierarchischen Klassifizieren von elektronischen Webdokumenten gibt es zwei Besonderheiten zu beachten. Zum Einen wäre da die Hypertextuelle Struktur der Webdokumente. Dadurch kann die Relevanz eines Schlüsselwortes, welches im Titel vorkommt, höher gewichtet werden als die eines Schlüsselwortes, welches nur im Text vorkommt. Weblinks sind eine weitere wichtige Informationsquelle bei der Beurteilung der Relevanz von Webdokumenten. Wenn von verschiedenen Webseiten oft auf eine Seite verlinkt wurde, kann davon ausgegangen werden, dass diese Seite relevant ist. Zum

Anderen wäre da noch die hierarchische Struktur der Klassifikationssysteme. Eine Lösungsmöglichkeit ist, das Klassifikationsproblem einfach, ähnlich einem Entscheidungsbaum, in eine Reihe kleinerer Klassifikationsprobleme zu untergliedern (vgl. Sebastiani 2002, S. 7). Für wissenschaftliche Bibliotheken ist es von besonderem Interesse, die elektronischen Dokumente schon direkt nach dem Harvesting aus den unterschiedlichen Repositorien zu klassifizieren. In diesem Fall würde das Klassifizieren von elektronischen Dokumenten direkt mit dem thematischen Textfiltern zusammenfallen (vgl. Oberhauser 2005, S. 21). Dies ist auch das Thema des von der Deutschen Forschungsgemeinschaft (im Folgenden: DFG) geförderten Projektes mit dem Kurztitel: *Automatische Anreicherung von OAI-Metadaten.* Ziel dieses Projektes war es, die Metadaten, welche mit dem Open Archives Initiative – Protocol for Metadata Harvesting – (im Folgenden: OAI-PMH) aus den verschiedenen Repositorien „geharvestet" wurden, automatisch mit DDC-Notationen anzureichern. Doch bevor eine detailliertere Beschreibung des Projekts erfolgen kann, soll hier zunächst damit begonnen werden, das OAI-PMH und den in diesem Zusammenhang entstandenen Neologismus „geharvestet" näher zu erläutern.

4 Open Archives Initiative – Protocol for Metadata Harvesting[4]

Die Open Archives Initiative (OAI) hat sich zum Ziel gesetzt die Interoperabilität bei der Verbreitung von elektronischen Dokumenten zu verbessern. Ihren Ursprung hat die OAI in der E-Print-Community. Aber bei diesem Projekt (Hauptentwicklungszeit 1999-2001) war das Ziel von Anfang an ein Interoperabilitäts-Framework für *alle* Anbieter elektronischer wissenschaftlicher Literatur zu entwickeln. Als Ergebnis wurde das Protocol for Metadata Harvesting (PMH) vorgestellt. Das PMH ist eine offene Schnittstelle zum Austausch von Metadaten (vgl. Lagoze, Van de Sompel 2001, S. 1-3). Als Metadatenformat wurde Dublin Core (im Folgenden: DC) festgelegt, weil es sehr gebräuchlich und einfach einsetzbar ist. Des Weitereren sind alle 15 DC-Elemente optional. Dies unterstreicht noch einmal die Intention ein möglichst offenes, variables und einfach einsetzbares Framework zu entwickeln, damit es häufig zum Einsatz kommt.[5] Die Daten werden beim OAI-PMH im XML-Format ausgegeben, bzw. ausgetauscht.

[4] Vgl. http://www.openarchives.org/OAI/openarchivesprotocol.htm (letzter Abruf am 25.03.2012)
[5] Das kann auch kritisch gesehen werden, denn dadurch sind die DC-Felder oft uneinheitlich ausgefüllt.

Abbildung 1: Beispiel für ein OAI Record

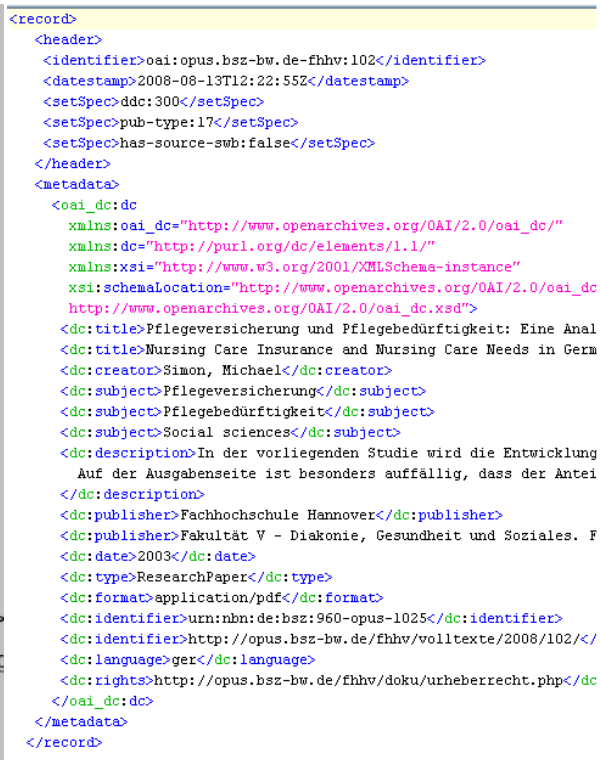


Abbildung 2: Screenshot eines OAI Records

Jedes Metadatenpaket wird in einem „Record“ (vgl. dazu Abb. 1 und Abb. 2) zusammengefasst. Und jedes Record repräsentiert ein Dokument in einer Datenbank. Ein Record besteht wiederum aus drei Containern: <header>, <metadata> und <about>. Die Informationen im <header> sind für den Harvesting-Prozess unentbehrlich. In ihm ist der „identifier“ enthalten. Innerhalb eines Repositories muss er einzigartig sein. Er fungiert als Schlüssel um die Metadaten zu extrahieren. Es handelt sich hierbei nicht um den Identifier für das Dokument selber, z.B. einen Persistent Identifier, wie den Digital Object Identifier (im Folgenden: DOI). Genau genommen ist der „identifier“ die eindeutige Repräsentation für ein „item“. Ein „item“ kann Metadaten in mehreren Formaten, die ein einzelnes Dokument repräsentieren, enthalten. Des Weiteren gibt es einen „datestamp“. Er enthält entweder das Datum der Erstellung, der Löschung oder der letzten Veränderung. Und es kann mehrere „setSpec“-Elemente geben. Ein Set ist ein optionales Konstrukt um „items“ zu gruppieren. Der <metadata>-Container enthält die Metadaten im unqualifizierten Dublin-Core-Format. Andere Metadaten-Formate sind, wie schon im Zusammenhang mit dem „item“ erwähnt, optional. Zu guter Letzt wäre da noch der <about>-Container. Auch er ist optional und wird zum Beispiel bei den SerWisS-Metadaten (rechte Seite) gar nicht verwendet. Er könnte benutzt werden, um über die Benutzungsrechte aufzuklären. Es gibt keine Vorgaben, in welches Schema diese Informationen zu verpacken sind. Darüber kann jede Institution selbst entscheiden.

Der Austausch der Metadaten erfolgt über das sogenannte Harvesting (engl. Ernte). In einer Harvesting-Session können die Metadaten von einem Data-Provider sozusagen „abgeerntet“ werden. Dies geschieht normalerweise durch sogenannte Service-Provider, wie wissenschaftliche Suchmaschinen, mit entsprechender Software. Aber Metadaten können auch mit einem einfachen Web-Browser, beispielsweise dem Firefox, „geerntet“ werden. Da die Datenmenge in den meisten Fällen zu groß ist, um sie in einer Abfrage zu harvesten und das meistens auch gar nicht gewünscht ist, können die Daten in kleinere Teilmengen zerlegt werden. Dies wird als „Selective Harvesting“ bezeichnet. Dabei gibt es zwei Möglichkeiten: Das daten-basierte und das set-basierte Harvesting. Beim daten-basierten Harvesting können die Metadaten durch den „datestamp“ auf einen bestimmten Zeitraum eingegrenzt werden. Das eignet sich besonders dann, wenn ein Service-Provider immer nur die neusten Metadaten, die ab einem bestimmten Zeitpunkt hinzugefügt wurden, benötigt. Beim set-basierten Harvesting können z.B. alle Metadaten angefordert werden, die über das „setSpec“-Element einer bestimmten Gruppe zugeordnet wurden. Wenn es doch gewünscht sein sollte eine größere Menge oder alle „Records“ zu harvesten, so wird dieses große Datenpaket in mehrere kleine Datenpakete partitioniert. Wie groß diese Datenpakete sind entscheidet das System. Das können 50 oder auch 100 Records sein. Jedenfalls gibt das System, wenn noch nicht alle Records zu einer Anfrage mit ausgegeben wurden, einen „resumptionToken“ mit einer ID aus. Mit dieser ID können dann die nächsten 50 oder 100 Records, je nach Partitionierung des Systems, angefordert werden, solange bis die Datenmenge vollständig ist. Jede Anfrage an einen Data-Provider erfolgt über eine URL im Web-Browser. Diese setzt sich zusammen aus der Basis-URL und der entsprechenden Anfrage. Die Basis-URL ist die Adresse von dem HTTP-Server, welcher als Data-Provider agiert. Für die Anfrage werden die GET- und POST-Methoden des http benutzt (vgl. Lagoze, Van de Sompel 2001, S. 4-6). Die folgende Anfrage gibt zum Beispiel das Record eines „items“ mit dem Identifier *oai:opus.bsz-bw.de-fhhv:2* im DC-Format aus:

http://opus.bsz-bw.de/fhhv/oai2/oai2.php?verb=GetRecord&identifier=oai:opus.bsz-bw.de-fhhv:2&metadataPrefix=oai_dc

Der Data-Provider ist in diesem Fall der Hochschulschriftenserver der Hochschule Hannover (SerWisS).

5 Dewey-Dezimalklassifikation (DDC)

Das Kernthema dieser Arbeit ist die automatische Anreicherung von Metadaten mit Notationen der Dewey-Dezimalklassifikation (DDC). Im Folgenden soll deshalb ein genauerer Blick auf dieses Klassifikationssystem geworfen werden.

Die DDC ist eine Universalklassifikation. Das heißt sie beschränkt sich nicht nur auf ein Fachgebiet, sondern soll das gesamte Wissen der Welt abbilden. Sie ist nach ihrem Erfinder Melvil Dewey benannt, welcher dieses Klassifikationssystem erstmals 1876 der Öffentlichkeit präsentierte. Er hatte es zuvor, innerhalb von nur drei Jahren, entwickelt. Heutzutage ist die DDC das weltweit am weitesten verbreitete Klassifikationssystem. Es wird in über 135 Ländern eingesetzt. In mehr als 60 Ländern wird die DDC sogar für die Nationalbibliographien verwendet. Inzwischen wurde sie sogar in mehr als 30 Sprachen übersetzt und seit 2005 gibt es auch eine Deutsche Ausgabe der DDC. Neben der großen Verbreitung hat die DDC noch eine weitere Stärke. Sie wird von der amerikanischen Nationalbibliothek, der Library of Congress (im Folgenden: LoC) gepflegt und bleibt so aktuell. Durch die Ansiedlung der Dewey-Abteilung bei der dortigen Klassifikationsabteilung, wo jährlich über 110 000 Notationen vergeben werden, können neue Entwicklungen in der Literatur frühzeitig erkannt werden. Die Redakteure der Dewey-Abteilung geben ihre Verbesserungsvorschläge an das *Decimal Classification Editorial Policy Committee* (im Folgenden: EPC) weiter. Das EPC ist ein aus zehn Mitgliedern der Bibliothekslandschaft (Öffentlich-, Spezial-, und Hochschulbibliotheken) bestehendes Gremium. Es ist ein beratendes Gremium, welches die Anwenderinteressen bei Expansions- und Revisionsentscheidungen der DDC vertritt. Für die Veröffentlichung, als Kurz- und vollständige Ausgabe sowie als Web-Dewey, ist das *Online Computer Library Center Inc.* verantwortlich (im Folgenden: OCLC) (vgl. Dewey, Mitchell 2005a, S. xlvii-xlviii). Seit 2006 bietet die Deutsche Nationalbibliothek auch einen eigenen, deutschen DDC-Web-Service mit dem Namen MelvilClass[6] an (vgl. Dewey, Mitchell 2005a, S. xxvii).

5.1 Klassenbildung

Die DDC wird Dezimalklassifikation (lat. decem für *zehn*) genannt, weil sie aus zehn Hauptklassen besteht, die wiederum jeweils zehn Unterklassen haben, welche wiederum zehn Unterklassen haben. Auf jeder Hierarchieebene kommen pro Klasse zehn neue

[6] Vgl. http://melvil.d-nb.de/melvilsearch?bs=dnb-portal (letzter Abruf am 03.04.2012)

Unterklassen dazu. Daraus folgt, dass es bereits auf der dritten Hierarchieebene 1.000 verschiedene Klassenbenennungen gibt (vgl. Dewey, Mitchell 2005a, S. xlviii-xlix). Insgesamt hat die 22. Ausgabe der DDC ca. 50.000 Klassen (Gödert 2002, S. 399 zit. nach Jacobs, Mengel, Müller 2010, S. 236). Die zehn Hauptklassen sind nach Fachgebieten oder Forschungsbereichen eingeteilt. In der deutschen DDC-Version sind das die Folgenden (vgl. Dewey, Mitchell 2005a, S. xlviii-xlix):

000	Informatik, Informationswissenschaft, allgemeine Werke
100	Philosophie und Psychologie
200	Religion
300	Sozialwissenschaften
400	Sprache
500	Naturwissenschaften und Mathematik
600	Technik, Medizin, angewandte Wissenschaften
700	Künste und Unterhaltung
800	Literatur
900	Geschichte und Geografie

Da die DDC der Konvention folgt, dass keine Notation weniger als drei Stellen haben darf, werden die fehlenden arabischen Ziffern auf der ersten Ebene durch Nullen aufgefüllt. Ansonsten gilt, dass jede Ziffer eine Ebene kennzeichnet. Die erste Ziffer, kennzeichnet die erste Ebene, die zweite Ziffer die zweite Ebene und die dritte Ziffer steht für die dritte Ebene. Nach den ersten drei Hierarchieebenen folgt ein *Dewey-Punkt.* Er hat vor allem optische Gründe und soll die Monotonie der Ziffern durchbrechen. Nach dem Punkt kann die Klassifizierung mit arabischen Ziffern bis zum gewünschten Spezifitätsgrad fortgesetzt werden. Dabei gibt es nur eines zu beachten: Rechts des Dewey-Punktes darf eine Notation niemals auf null enden. Folgendes Beispiel soll die Notationshierarchie verdeutlichen:

500	Naturwissenschaften und Mathematik
500	Allgemeine Naturwissenschaftliche Werke
510	Mathematik
520	Astronomie
530	Physik
530	Allgemeine Werke über Physik
531	Klassische Mechanik
531.2	Festkörperstatik
531.3	Festkörperdynamik

Die Hierarchieebene wird durch die Länge der Notation ausgedrückt. Ob zwei Klassen sich auf derselben Hierarchieebene befinden, kann an der Länge der Notation erkannt werden.

Dies wird als *notationelle Hierarchie* bezeichnet (vgl. Mitchell 2001, S. 213). Die DDC folgt dem Prinzip der *hierarchischen Vererbung*. Das bedeutet alle Regeln und Anmerkungen, welche für die obere Hierarchieebene gelten, gelten auch für alle ihr nachgeordneten Klassen (vgl. Dewey, Mitchell 2005a, S. l-li). Wenn ein Teil eines Themas oder ein Konzept nicht vollständig durch eine Hierarchieebene ausgedrückt werden kann, gibt es in der DDC Verweise. (vgl. Mitchell 2001, S. 214) In der deutschen Übersetzung der DDC würde das folgendermaßen aussehen (vgl. Dewey, Mitchell 2005b, S. 511):

338.2 Bergbau
Für die Verarbeitung von mineralischen Rohstoffen und Energiestoffen siehe 338.47

Joan S. Mitchell (2001, S. 214-218) beschreibt zwei Arten von Hierarchierelationen, welche mit der DDC ausgedrückt werden können: (1) *generische und partitive Hierarchierelation* sowie (2) *Polyhierarchische Relationen*. Die generische Beziehung (Eine-Art-von-Beziehung) wird häufig durch Verweise ausgedrückt. Die partitive Beziehung (Ein-Teil-von-Beziehung) kann sowohl durch Verweise, als auch durch notationelle Hierarchie ausgedrückt werden. Manchmal passt dasselbe Konzept in mehr als eine DDC-Klasse. Eine solche, polyhierarchische Relation, wird auch durch Verweise ausgedrückt.

An den Hierarchierelationen wird deutlich, wie ähnlich die Strukturen von DDC und Thesaurus sind. Die Hierarchierelationen wurden bereits ausführlich beschrieben. Es fehlen noch die *Äquivalenz-* und *Assoziationsrelationen*. Auch sie können in der DDC ausgedrückt werden. Äquivalenzrelationen werden in der DDC immer in runden Klammern hinter den Begriff geschrieben. Das Wort in Klammern könnte ein Synonym, ein äquivalentes Adjektiv für eine Person, eine Sprache oder ein Gebiet, sowie ein Akronym oder die ausgesprochene Version davon sein. Ansonsten werden die Äquivalenzrelationen, genau wie alle anderen Relationen, durch Verweise ausgedrückt. Bei der Assoziationsrelation wird zum Beispiel durch die Klassifiziere-in-Anweisung auf verwandte Begriffe verwiesen (vgl. Dewey, Mitchell 2005a, S. lxii):

791.43 Film
klassifiziere fotografische Aspekte beim Film in 778.53;
klassifiziere Fernsehspielfilme, Videoaufzeichnungen von Filmen in 791.45

Um eine möglichst präzise Notation für ein Werk zu finden, bietet die DDC des Weiteren die Möglichkeit zur Facettenklassifikation. Allerdings nur dann, wenn bei der

Grundnotation durch entsprechende Anweisungen darauf hingewiesen wird. Für die Notationssynthese gibt es vier Möglichkeiten: (1) den Standardschlüssel der Hilfstafel 1; (2) die Hilfstafeln 2-6; (3) andere Teile der Haupttafeln; (4) spezielle Anhängetafeln in den Haupttafeln. Der Standardschlüssel 1 kann auf jedes Thema oder Fachgebiet angewendet werden. Er steht für die äußere Form oder die Art der Darstellung, z.B. Zeitschrift, Verzeichnis, Wörterbuch, Biografie, historischen Behandlung oder wissenschaftliche Untersuchung. Die Hilfstafeln 2-6 enthalten die folgenden Schlüssel:

Hilfstafel 2	für Zeitabschnitte, Personen und geographische Gebiete
Hilfstafel 3	für Künste, Literaturen, einzelne literarische Gattungen
Hilfstafel 4	für einzelne Sprachen oder Sprachfamilien
Hilfstafel 5	für ethnische und nationale Gruppen
Hilfstafel 6	Für Sprachen als Grundlage für die Notationssynthese

Beim Anhängen mehrerer Facetten an eine Notation muss die „Citation Order“ berücksichtigt werden. Wenn in den Haupttafeln die Möglichkeit zur Facettenklassifikation eröffnet wird, wird dort normalerweise auch genau die Reihenfolge beschrieben, in der die einzelnen Facetten an die Notation angehängt werden können. Das folgende Beispiel soll dies illustrieren (vgl. Lösch, 2009 und Dewey, Mitchell 2005a, S.215):

„Partizipatorische Demokratie in Frankreich“ **323.**0420944
300 Sozialwissenschaften
320 Politikwissenschaft
323 Grundrechte und politische Rechte
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen
Hilfstafel 1: Standardschlüssel 09: Historische, geographische, personenbezogene Behandlung
Hilfstafel 2: Standardschlüssel 44: Frankreich und Monaco

Wenn es nicht vorgesehen ist, mehrere Facetten an eine Notation anzuhängen, muss sich an die „Vorzugsreihenfolge“ gehalten werden. Wenn zum Beispiel für „arbeitslose Bibliothekare“ eine Notation festgelegt werden sollte, gäbe es zwei Möglichkeiten, da es nicht vorgesehen ist beide Aspekte auszudrücken. Die „Vorzugsreihenfolge“ legt hier fest, dass der Beruf des Bibliothekars (305.9092) gegenüber dem Status der Arbeitslosigkeit (305.90694) zu bevorzugen ist (vgl. Dewey, Mitchell 2005a, S. lxiv-lxx).

Bevor mit dem Klassifizieren begonnen werden kann, muss jede Bibliothek noch die grundsätzliche Entscheidung zwischen *feinem* und *grobem Klassifizieren* treffen (vgl. Dewey, Mitchell 2005a, S. lxxvii). Das obige Bespiel mit der partizipatorischen Demokratie in Frankreich wurde *fein klassifiziert*. Das heißt es wurde bis zur

letztmöglichen Ebene der Notation klassifiziert. Eine kleine Bibliothek würde dieses Werk vielleicht eher *grob klassifizieren* und die Notation 323 „Grundrechte und politische Rechte“ vergeben. Beim automatischen klassifizieren ist schon eine gute grobe Klassifizierung ein Erfolg. Im Abschließenden Experiment dieser Bachelorarbeit wurden Notationen bis zur zweiten und dritten Hierarchieebene vergeben. Ganz allgemein kann gesagt werden, dass beim automatischen klassifizieren auf der ersten und zweiten Hierarchieebene schon ganz gute Ergebnisse erzielt wurden. Alles was darüber hinaus geht, sollte bis dato besser einem intellektuellen Klassifizierer überlassen werden.

5.2 Konkordanzen

Im Jahr 1998 wurde von der Expertengruppe Klassifikation beschlossen, die DDC in der Deutschen Nationalbibliothek anzuwenden. Es wurde die Empfehlung ausgegeben, die DDC von nun an im deutschen Sprachraum zu verwenden. Damit ging der Beschluss einher, eine deutsche Version der DDC zu erstellen. Diese Entwicklung ist der Erkenntnis geschuldet, dass es mit der zunehmenden Verbreitung von Onlinekatalogen auch einen wachsenden Bedarf nach globalen Recherchemöglichkeiten gibt. Es wird für Bibliotheken immer wichtiger, ihren Nutzern Zugang zu verschiedenen Daten aus den unterschiedlichsten Repositorien unter einer möglichst einheitlichen Suchoberfläche anzubieten. Dazu müssen die inhaltsbeschreibenden Daten, wie Schlagwörter und Klassifikationssysteme, international austauschbar sein. Die DDC ist das Klassifikationssystem mit den besten Voraussetzungen um dieses Ziel möglichst schnell zu erreichen (vgl. Konferenz für Regelwerksfragen 2000, S. 45):

- Denn sie ist die einzige Klassifikation mit einer weltweiten Verbreitung.
- Bei der Klassifikation elektronischer Dokumente im Internet ist sie global führend.
- Dadurch gibt es bereits ein großes internationales Fremddatenangebot, welches durch die Anwendung der DDC im deutschen Sprachraum auch hier verfügbar wird.
- Hinter der DDC stehen mit der LoC und OCLC tragfähige Organisationen, die sich um die laufende Aktualisierung und Weiterentwicklung kümmern.
- Weil sie bereits in über 30 Sprachen übersetzt wurde, könnte die DDC als Grundlage für den Aufbau eines multilingualen Normenvokabulars dienen.
- Die DDC eignet sich aufgrund ihrer Struktur als „Dachsystematik“ für Konkordanzen mit anderen Systematiken.

Eine Konkordanz ist eine Art Verweis von einem Klassifikationssystem auf ein anderes Klassifikationssystem. Es kann auch gleichzeitig auf mehrere andere Klassifikationssysteme verwiesen werden. Dabei soll nicht die gesamte Struktur der Klassifikation aufeinander abgebildet werden, sondern eine Verbindung zwischen zwei Klassen geschaffen werden. Klassifikationen können generell nicht 1:1 aufeinander abgebildet werden. Das liegt unter anderem an der unterschiedlichen Gliederungstiefe. Grundsätzlich ist es einfacher eine fein gegliederte Klassifikation auf eine grob gegliederte Klassifikation abzubilden. Die unterschiedliche Untergliederung von Systemstellen spielt hierbei nicht so eine große Rolle, da von einer sehr „tiefen“ Hierarchieebene auf eine weniger „tiefe“ Hierarchieebene verwiesen wird. Der Nachteil ist, dass eine solche Konkordanz einen Informationsverlust zur Folge hat. Dadurch vergrößert sich die Treffermenge beim Retrieval und die Precision verschlechtert sich. Andersherum ist es sehr viel schwieriger. Bei der Abbildung einer groben Klassifikation auf eine fein gegliederte, können nur die Oberbegriffe der feinen Klassifikation auf die grobe Klassifikation abgebildet werden. Durch automatische Erschließungsverfahren können die Daten verbessert werden. So kann diesem Problem ein wenig Abhilfe verschafft werden. Bei der Erstellung einer Konkordanz ist es, aufgrund der eben geschilderten Problematiken, wichtig, die Art der Beziehungen zwischen zwei (oder mehr) Klassen abzubilden:

- Beziehung 1 : 1 (Synonyme Begriffe, parallele Notation)
- Oberbegriff : Unterbegriff
- Unterbegriff : Oberbegriff
- Verwandte Begriffe
- mit eigener Notation versehener Begriff : (nur) durch Schlüsselung recherchierbarer Begriff
- Maß für die Übereinstimmung

  (Konferenz für Regelwerksfragen 2000, S. 43)

> Ziel einer Konkordanz ist, eine integrierte Suche nach sachlichen Gesichtspunkten in verteilten Datenbeständen mit unterschiedlichen inhaltlichen Schwerpunkten zu ermöglichen (Konferenz für Regelwerksfragen 2000, S. 41).

Es spielt dabei keine Rolle, ob es sich um Bestände von Fachbibliotheken oder Universalbibliotheken handelt. Wichtig ist nur, dass diese Daten für die fächerübergreifende Suche zugänglich gemacht werden sollen. Es ist auch nicht so, dass Konkordanzen nur zwischen Klassifikationssystemen erstellt werden können. Thesauri

bieten sich zum Beispiel aufgrund ihrer ähnlichen Struktur sehr für Konkordanzen mit Klassifikationssystemen an. Auf dem Gebiet der Konkordanz-Erstellung wurden bereits einige Projekte durchgeführt. Zu nennen wäre da beispielsweise das Projekt Multilingual Access to Subjects (im Folgenden: MACS). Es wurde von der Deutschen Nationalbibliothek in Kooperation mit der Britisch Library London, der Schweizerischen Landesbibliothek Bern und der Bibliothèque Nationale Paris durchgeführt. In diesem Projekt wurde eine Konkordanz zwischen der Schlagwortnormdatei (im Folgenden: SWD), den Library of Congress Subject Headings (im Folgenden: LCSH) und dem Répertoire d'autorité-matière encyclopédique et alphabétique unifié (im Folgenden: RAMEAU) erstellt (vgl. Konferenz für Regelwerksfragen 2000, S. 40-44). In einem anderen Projekt mit dem Titel CrissCross wurde eine Konkordanz zwischen der SWD und der DDC erstellt. Dieses Projekt wird im weiteren Verlauf dieser Arbeit vorgestellt. Das ist wichtig, weil das Ergebnis dieses Projektes, die Grundlage für den hier vorgestellten Versuch einer automatischen DDC-Klassifizierung bildet.

6 DFG-Projekte

In diesem Kapitel sollen zwei Projekte vorgestellt werden, die inhaltlich in engem Zusammenhang mit dem in dieser Arbeit vorgestellten Versuch zum automatischen Klassifizieren stehen. Beide Projekte sind sehr aktuell. Noch bis 2011 wurden sie von der Deutschen Forschungsgemeinschaft gefördert. Das erste Projekt „Automatische Anreicherung von OAI-Metadaten" hat sehr viele Gemeinsamkeiten mit dem Klassifikationsversuch der vorliegenden Arbeit. Zunächst werden auch in dem Projekt nur OAI-Records zur automatischen Klassifikation verwendet. Darüber hinaus ist die Zielklassifikation dieselbe. Beide Male sollen Notationen der DDC vergeben werden. In Vorgehensweise und Projektumfang gibt es jedoch deutliche Unterschiede. Letzteres ist natürlich dem Umstand geschuldet, dass es sich in der vorliegenden Arbeit nur um eine neun wöchige Bachelorarbeit handelt. Bei der Vorgehensweise wurde vom DFG-Projekt ein maschinelles Lernverfahren gewählt. Demgegenüber kann bei dem eignen Klassifikationsversuch eher von einem computerlinguistischen Verfahren gesprochen werden. Konkret wurden regelbasierte, wörterbuchbasierte und begriffsorientierte Verfahren angewendet.

Das zweite Projekt „CrissCross" bildet sozusagen die Grundlage für den eignen Klassifikationsversuch. In dem CrissCross-Projekt wurde eine Konkordanz zwischen der SWD und der DDC erstellt. Das besondere daran ist, dass die DDC-Notationen in die

SWD integriert wurden. Die um DDC-Notationen erweiterten SWD-Sachschlagwörter stehen zur Nachnutzung zur Verfügung und können bei der Deutschen Nationalbibliothek (im Folgenden: DNB) unter der Creative-Commons-Zero-Lizenz als RDF/XML-Datei heruntergeladen werden.[7] Diese RDF/XML-Datei wurde für den eigenen Klassifikationsversuch modifiziert und als Ontologie verwendet. Genauere Beschreibungen der RDF/XML-Datei und ihrer Modifikationen erfolgen im nächsten Kapitel, doch zunächst sollen die beiden DFG-Projekte beschrieben werden, welche als Inspiration für diese Arbeit dienten.

6.1 Automatische Anreicherung von OAI-Metadaten

Der vollständige Titel dieses Projektes lautet: ***„Automatische Anreicherung von OAI-Metadaten mit Hilfe computerlinguistischer Verfahren und Entwicklung von Services für die inhaltsorientierte Vernetzung von Repositorien.“*** Es wurde von der Deutschen Forschungsgemeinschaft im Zeitraum vom 1. Oktober 2009 bis zum 30. September 2011 gefördert.[8] Das Projekt war eine Kooperation zwischen der Universitätsbibliothek Bielefeld, dem Institut für Informatik an der Universität Leipzig und dem Text Technology Lab an der Universität Frankfurt. Die OAI-Metadaten sollten mit DDC-Notationen angereichert werden, um semantische Suchen in der wissenschaftlichen Suchmaschine Bielefeld Academic Search Engine (im Folgenden: BASE) zu ermöglichen. [9] Durch dieses Projekt hat sich die Sacherschließung in BASE deutlich verbessert. Die Zahl der mit DDC-Notationen angereicherten Dokumente konnte im Projektzeitraum von 429.496 auf 1.753.712 verdreifacht werden.[10] Diese Verbesserung ist jetzt auch für die Nutzer deutlich sichtbar. Neben der normalen Suche ist auch ein DDC-Browsing bis zur dritten Hierarchieebene möglich.[11] Des Weiteren wurde, mit den im Projekt entwickelten Klassifikatoren, eine Seite ins Netz gestellt, welche die automatische Klassifikation im Web-Browser ermöglicht. Hier können Texte, PDFs und Webseiten automatisch klassifiziert werden. Als Ergebnis werden Vorschläge für DDC-Notationen bis zur dritten Hierarchieebene ausgegeben.[12]

[7] Vgl. https://wiki.d-nb.de/display/LDS/Dokumentation+des+Linked+Data+Services+der+DNB (letzter Abruf am 18.04.2012)
[8] Vgl. http://www.ub.uni-bielefeld.de/biblio/projects/oai_projekt.htm (letzter Abruf am 22.03.2012).
[9] Vgl. http://www.ub.uni-bielefeld.de/wiki/AutoOAI (letzter Abruf am 22.03.2012)
[10] Vgl. http://www.ub.uni-bielefeld.de/wiki/OAIMErgebnisse (letzter Abruf am 22.03.2012)
[11] Vgl. http://baselab.base-search.net/Browse/Dewey (letzter Abruf am 04.04.2012)
[12] Vgl. http://act-dl.base-search.net/ (letzer Abruf am 04.04.2012)

6.1.1 Erstellung der Trainingsdokumente

Da ein maschinelles Lernverfahren angewandt werden sollte, mussten zunächst Trainingsdokumente erstellt werden. Beim automatischen Klassifizieren wird in diesem Zusammenhang auch von Korpus gesprochen. Ein Korpus ist eine Sammlung von Textdokumenten. In diesem Fall sind es Volltexte und OAI-DC-Metadaten aus der BASE-Datenbank. Die Texte sind in deutscher und englischer Sprache.

Zu dem Zeitpunkt, als das Korpus erstellt wurde, hatte BASE etwa 26 Millionen bibliographische Records (OAI-DC-Metadaten) aus über 1.700 Repository-Servern. Um den Prozess der Korpuserstellung so gut wie möglich zu automatisieren, wurde ein Programm geschrieben. Dieses Programm stellte zwei Bedingungen, die jedes Record erfüllen musste, um in das Korpus inkludiert zu werden:

1) Der dazugehörige Volltext muss im Open Access veröffentlicht, d.h. frei im Internet verfügbar sein. Die URL zum Downloaden muss im Record stehen. Außerdem muss der Volltext im PDF-Format vorliegen, oder in einem anderen maschinenlesbaren Format.
2) Die richtige DDC-Notation muss ermittelt werden können.

Der erste Punkt ist abhängig von der Veröffentlichungspolitik der Repositorien oder Autoren. Der zweite Punkt, die richtige Dewey-Nummer, musste auf verschiedenen Wegen ermittelt werden. Einige Repositorien erschließen ihre Dokumente bereits mit der DDC. Hier konnten die Records direkt in das Korpus importiert werden. In vielen Fällen benutzen die Repositorien aber andere Wissensrepräsentations- bzw. Klassifikationssysteme. Da es ein viel zu großer Aufwand gewesen wäre, für jedes Record und den dazugehörigen Volltext intellektuell eine DDC-Notation zu generieren, wurde dieser Prozess teilautomatisiert. Zunächst wurden manuell Konkordanzen zwischen der DDC und den anderen Wissensrepräsentationssystemen erstellt. Anschließend konnte die Zuordnung von DDC-Notationen zu OAI-Records automatisch erfolgen. Ein Programm hat das vorhandene Notationssystem durch die Analyse der Notationsstruktur automatisch erkannt und die dazu passende Konkordanz-Tabelle ausgewählt. Aus dieser Tabelle wurde dann eine zu der vorhandenen Notation äquivalente DDC-Notation ausgewählt. In einigen Repositorien wurden teilweise gleich mehrere Notationssysteme benutzt.

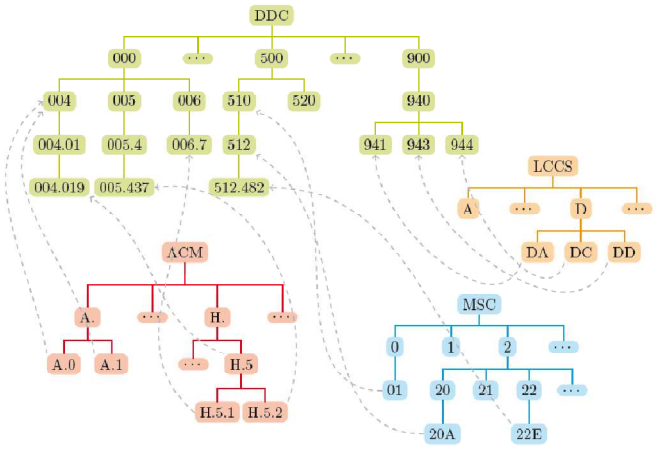


Abbildung 3: Konkordanzen zur Ermittlung der fehlenden DDC-Notationen

Das Problem mit den fehlenden DDC-Notationen konnte durch die Konkordanzen gelöst werden. Um ein funktionsfähiges Trainingskorpus zu erhalten, waren aber noch einige weitere Schritte notwendig. Zunächst mussten die heruntergeladenen PDF-Files in eine „Plain-Text-Version“ (engl. für Klartext) umgewandelt werden. Diese Version wurde dann durch ein Sprachenerkennungsprogramm geschickt. Dies war unumgänglich, da nicht jedes Repository das DC-Sprachenfeld richtig verwendet. Abschließend wurde die „Plain-Text-Version“ mit einem 32-Stelligen Hexadezimal-Code, einem sogenannten MD5-Hash-Code, eindeutig identifizierbar gemacht. Dieser Hash-Code wurde zusammen mit der DDC-Notation im Record abgespeichert. Dadurch wird verhindert, dass es beim Hinzufügen neuer Dokumente zu Dubletten kommt. Des Weiteren fungiert der MD5-Hash-Code als ID, denn die Volltext-Version und die OAI-DC-Records werden getrennt abgespeichert.

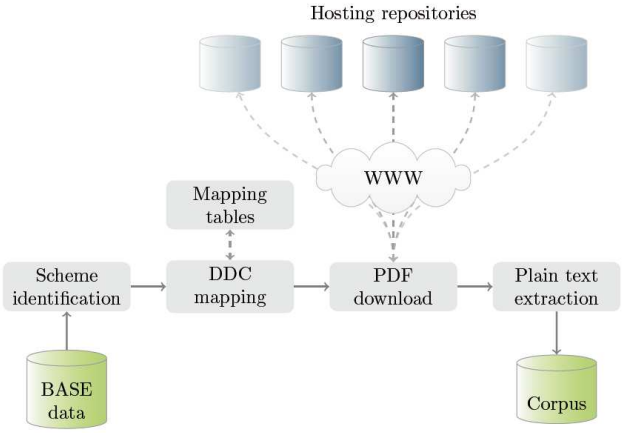


Abbildung 4: Erstellung der Trainingsdokumente

6.1.2 Zusammensetzung der Trainingsdokumente

Am Ende bestand das Korpus aus 52.905 englischsprachigen und 37.228 deutschsprachigen Dokumenten. Es stellte sich als ziemlich schwierig heraus, für jede DDC-Klasse dieselbe Menge an positiven Trainingsbeispielen zu bekommen (vgl. dazu Tabelle 1). Dies ist vor allem der Tatsache geschuldet, dass es in den Naturwissenschaften eine viel längere Open Access Tradition gibt als beispielsweise in den Geisteswissenschaften. Aber es gibt auch Inkonsistenzen innerhalb der Struktur der DDC. Dadurch, dass die DDC schon vor über hundert Jahren entwickelt wurde, musste sie im Laufe der Zeit öfters angepasst werden. Eine neue Klassendefinition konnte jedoch immer nur anhand der aktuellen Literaturlage erfolgen. Literatur, die zu einem späteren Zeitpunkt verfasst wurde, passt dann vielleicht nicht mehr zu hundert Prozent in diese DDC-Klasse. Diese Inkonsistenzen können auch ein Grund dafür sein, dass einige Klassen mit weniger Beispieldokumenten gefüllt sind als andere. Als Hauptgrund ist natürlich die unterschiedliche Open Access Tradition in den einzelnen Fachgebieten zu nennen.

Tabelle 1: Zusammensetzung der Trainingsdokumente

DDC	Englisch	German
000 Informatik, Informationswissenschaft, allgemeine Werke	6847	3778
100 Philosophie und Psychologie	3536	2169
200 Religion	1123	1973
300 Sozialwissenschaften	10948	8075
400 Sprache	1682	1297
500 Naturwissenschaften und Mathematik	23989	6969
600 Technik, Medizin, angewandte Wissenschaften	6669	5874
700 Künste und Unterhaltung	1280	3823
800 Literatur	740	2063
900 Geschichte und Geografie	2226	2863

Damit die Differenzen zwischen den einzelnen Klassen beherrschbar blieben, wurden Obergrenzen festgelegt. Eine Klasse auf der ersten Hierarchieebene der DDC sollte maximal durch 10.000 Dokumente repräsentiert werden. Auf der zweiten Ebene wurde das Limit auf 1.000 Dokumente und auf der dritten Ebene auf 100 Dokumente festgelegt. Sämtliche Dokumente wurden in mindestens eine der 10 Hauptklassen der ersten Ebene einsortiert. Darüber hinaus konnte jedes Dokument auch DDC-Notation der zweiten- und dritten Hierarchieebene erhalten. Die Dokumente wurden mehreren Klassen gleichzeitig zugeordnet, weswegen manche Klassen trotz Begrenzung weiter wuchsen. Abbildung 5 zeigt, dass es trotz dieser Bemühungen nicht gelang die Klassen auszubalancieren (vgl. Lösch, Waltinger, Horstmann, Mehler 2011, S. 1-7).

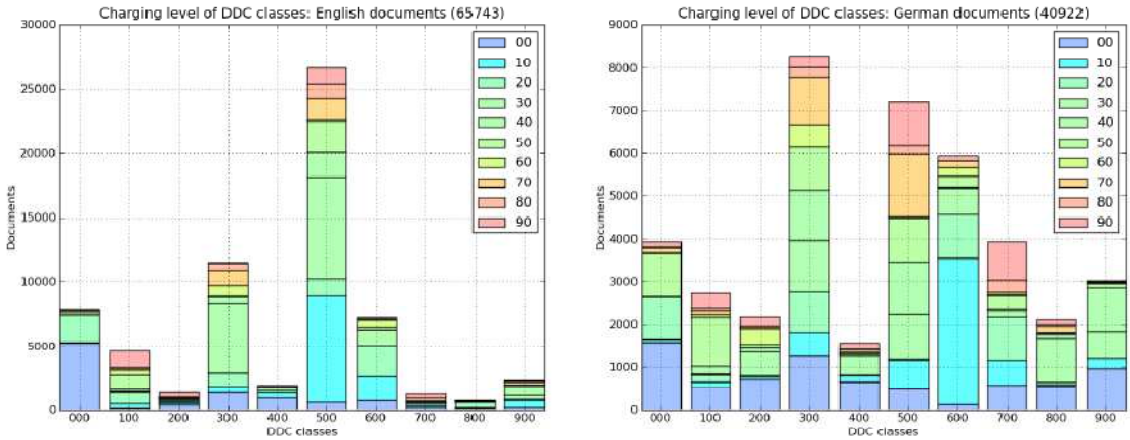


Abbildung 5: Verteilung der Trainingsdokumente auf die DDC-Klassen

Idealerweise sollten die Klassen bei Trainingsdokumenten jedoch ausgeglichen sein. Dies ist hier nicht gelungen, deswegen musste sichergestellt werden, dass wenigstens die zweite Anforderung erfüllt wird. Die Trainingsdokumente müssen so verschieden wie möglich sein. Es darf nur minimale Überschneidungen geben. Die Wörter, welche eine Klasse repräsentieren, sollten möglichst einzigartig sein und nicht, oder nur selten, in den anderen Klassen vorkommen (vgl. Konchady 2008, S. 278). Das liegt in den Methoden zum automatischen Klassifizieren begründet. Die meisten Programme benutzen wortbasierte Methoden. Bei diesen Methoden wird ein Text durch einen Vektor von Wortgewichten repräsentiert. Oft wird hier in der Literatur auch von Termgewichtung gesprochen. Ein Term ist sozusagen der Oberbegriff und beinhaltet Wörter und Phrasen. Eine Sammlung von Termen wird auch als Feature bezeichnet. Meistens bezeichnet ein Term jedoch ein Wort, weshalb hier auch von wortbasierten Methoden gesprochen werden soll (vgl. Sebastiani 2002, S. 10). Um sicher zu gehen, dass diese Bedingung von den Trainingsdokumenten erfüllt wird, wurde ein χ^2-Test durchgeführt. Für jedes Wort (dt. für Term) *t* sollte ermittelt werden, ob es statistisch signifikant genug ist um eine Klasse *c* zu beschreiben. Statistische Signifikanz gibt darüber Auskunft, ob die beschreibenden Wörter für eine Klasse zufällig sind oder nicht. Wenn t und c voneinander unabhängig sind, *t* also nur ein zufälliges Wort zur Beschreibung der Klasse *c* ist, dann ist der Wert von $\chi^2_{t,c} = 0$.

$$\chi^2_{t,c} = \frac{N(AD - BC)^2}{(A+B)(A+C)(B+D)(C+D)}$$

A = die Zahl der Dokumente aus der Klasse *c*, die den Term *t* enthalten
C = die Zahl der Dokumente von Klasse *c* die nicht den Term *t* enthalten
B = die Zahl der Dokumente von jeder anderen Klasse $d \neq c$, die *t* enthalten
D = die Zahl der Dokumente von jeder anderen Klasse $d \neq c$, die *t* nicht enthalten
N = die Gesamtzahl der Dokumente im Corpus

Der χ^2-Test wurde mit den OAI-DC-Metadaten des englischsprachigen Korpus durchgeführt. Zuvor wurden alle Stoppwörter und alle nicht lexikalischen Ausdrücke eliminiert. Die Begriffe wurden in Kleinbuchstaben geschrieben und auf ihre Stammform reduziert. Abbildung 7 zeigt die fünf Wörter, welche beim χ^2-Test jeweils am Besten für ihre Klasse abgeschnitten haben. Bemerkenswert ist, dass die Klasse 300 (Sozialwissenschaften) durch sehr viele Begriffe aus dem Bereich Ökonomie repräsentiert wird. Ansonsten sind die Wörter sehr unterschiedlich. Daraus kann die Schlussfolgerung gezogen werden, dass sich die Trainingsdokumente sehr gut als Repräsentanten ihrer Klasse eignen (vgl. Lösch, Waltinger, Horstmann, Mehler 2011, S. 7-8).

Tabelle 2: Die besten Begriffe zur Repräsentation ihrer DDC-Klasse

DDC	Top χ^2-scored terms
000	librari, comput, user, scienc, journal
100	psycholog, cognit, philosophi, mind, conscius
200	religion, church, religi, christian, theolog
300	market, polici, countri, wage, firm
400	languag, linguist, english, semant, syntact
500	physic, mathemat, energi, librari, polici
600	engin, biolog, cell, machineri, fibr
700	music, art, design, architectur, theatr
800	literari, australian, fiction, poetri, literatur
900	archaelog, histori, songster, geographi, war

6.1.3 Der Prozess des automatischen Klassifizierens

Das Projekt zur ***automatischen Anreicherung von OAI-Metadaten*** hatte im Wesentlichen zwei Ziele (Waltinger, Mehler, Lösch, Horstmann 2011, S.33):

1) Die Dokumente sollten nur anhand ihrer Metadaten klassifiziert werden. Genau genommen wurden für die Klassifizierung sogar nur die Metadatenfelder *title*, *description* und *subject* benutzt.
2) Nicht nur die für die oberste, sondern für die ersten drei Hierarchieebenen der DDC sollten Notationen vergeben werden.

Bei der Wahl für einen automatischen Klassifizierer ist die Entscheidung für *Support Vector Machines* (im Folgenden: SVM) gefallen. In einem vorigen Versuch hatte sich gezeigt, dass die SVMs am besten mit dem Dublin-Core-Schema funktionieren. Eine SVM unterteilt die positiven und negativen Trainingsdokumente so, dass ein möglichst breiter Rand zwischen ihnen entsteht. In Abbildung 6 werden die positiven und negativen Beispiele durch Pluszeichen und Kreise repräsentiert. Jeweils zwei Kreise und zwei

Pluszeichen sind von einem viereckigen Kästchen umschlossen. Diese Trainingsdokumente werden auch *Support Vectors* genannt, weil sie gewissermaßen die Entscheidungsgrundlage bilden. Die Support Vectors sind immer die Trainingsdokumente mit dem größtmöglichen Zwischenraum. Eine Entscheidung über die Unterteilung zweier Klassen wird also nur anhand einiger weniger Trainingsdokumente getroffen. Die mittlere Gerade σ_i markiert den größten Zwischenraum zwischen den positiven und negativen Trainingsdokumenten. Sie kennzeichnet die größte Distanz zwischen zwei Klassen (vgl. Sebastiani 2002. S. 30).

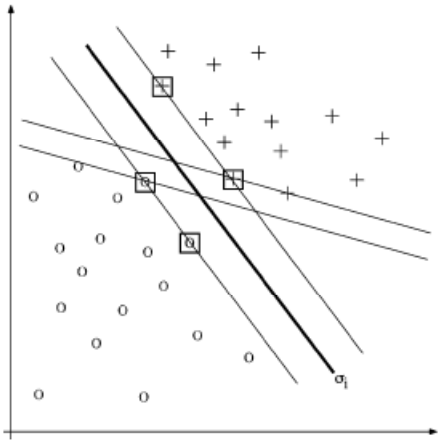


Abbildung 6: Arbeitsweise einer Support Vector Machine

Für jede DDC-Klasse wurde eine eigene SVM trainiert. Und zwar immer so, dass eine Klasse gegenüber allen anderen Klassen derselben Hierarchieebene abgegrenzt wird.

6.1.4 Ergebnisse

Bei dem Versuch Volltextdokumente nur mit deren OAI-Records zu klassifizieren, waren die Ergebnisse für die erste Hierarchieebene sehr gut. Für die englischen Records konnte ein *F1-Measure* von 0,81 erzielt werden. Für die deutschen Records lag das F1-Measure mit 0,79 etwas darunter.

Tabelle 3: Ergebnisse für die erste Hierarchieebene

Ergebnis für die *deutschen* OAI-Records

DDC	Precision	Recall	F1-Measure
000	0.948	0.878	0.915
100	0.906	0.815	0.925
200	0.903	0.720	0.888
300	0.852	0.691	0.871
400	0.828	0.621	0.896
500	0.868	0.819	0.922
600	0.856	0.764	0.770
700	0.812	0.631	0.299
800	0.805	0.620	0.775
900	0.878	0.745	0.355
Overall	**0.866**	**0.730**	**0.791**

Ergebnis für die *englischen* OAI-Records

DDC	Precision	Recall	F1-Measure
000	0.932	0.887	0.909
100	0.959	0.911	0.934
200	0.954	0.757	0.844
300	0.951	0.890	0.919
400	0.943	0.858	0.899
500	0.827	0.808	0.817
600	0.807	0.735	0.768
700	0.887	0.496	0.636
800	0.833	0.492	0.619
900	0.911	0.642	0.753
Overall	**0.900**	**0.747**	**0.810**

Das F1-Measure ist ein Maß für die Effektivität beim automatischen Klassifizieren. Es ist das harmonische Mittel von Precision und Recall. Genau genommen entspricht das F1-measure der F_β-Funktion, bei der Precision und Recall gleich gewichtet sind, d.h. $\beta = 1$. Die F_β-Funktion lautet (vgl. Sebastiani 2002, S. 36):

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

Precision und Recall haben in Verbindung mit dem automatischen Klassifizieren eine etwas andere Bedeutung als im klassischen Information Retrieval. *Precision* ist die Wahrscheinlichkeit dafür, dass die Entscheidung, einer Klasse ein Dokument zuzuordnen, richtig ist. Precision drückt den Grad der Zuverlässigkeit aus. Hier wird die Frage beantwortet, wie hoch die Wahrscheinlichkeit ist, dass ein automatischer Klassifizierer die richtige Klassifikationsentscheidung trifft. *Recall* ist die Wahrscheinlichkeit dafür, dass ein Dokument, welches einer bestimmten Klasse angehört, auch wirklich in diese klassifiziert wird. Recall drückt den Grad Vollständigkeit aus. Hier wird die Frage beantwortet, wie viele der Dokumente, die in eine bestimmte Klasse gehören, „gefunden“ werden. In Formeln ausgedrückt stellt sich das Ganze wie folgt da (vgl. Sebastiani 2002, S. 33):

Klasse c		Expertenurteil	
		Ja	**Nein**
Automatischer Klassifizierer	**Ja**	TP = True Positive	FP = False Positive
	Nein	FN = False Negative	TN = True Negative
Precision = TP / (TP + FP)		*Recall* = TP / (TP + FN)	

Auf der Zweiten Hierarchieebene konnte für die deutschen Records immerhin noch ein F1-Measure von 0,74 erreicht werden. Auf der dritten Ebene lässt die Klassifiziereffektivität dann deutlich nach. Hier konnte nur noch ein F1-Measure von 0,611 erreicht werden.

Tabelle 4: Gesamtergebnis für alle drei Hierarchieebenen

Gesamtergebnis: *deutsche* OAI-Records

Ebene	No	Precision	Recall	F1-Measure
1	10	0.866	0.730	0.791
2	31	0.841	0.682	0.744
3	87	0.763	0.545	0.611

Gesamtergebnis: *englische* OAI-Records

Ebene	No	Precision	Recall	F1-Measure
1	10	0.900	0.747	0.810
2	39	0.784	0.556	0.631
3	39	0.769	0.545	0.616

Obwohl es in der gesamten DDC bis zur dritten Hierarchieeben über 1.000 verschiedene Klassen gibt, konnte nur ein Bruchteil von ihnen benutzt werden. Bei den deutschsprachigen Records waren es 128 Klassen. Bei den englischsprachigen Records waren es sogar noch weniger. Hier wurden insgesamt nur 88 Klassen benutzt. Das liegt vor allem daran, dass es nicht ausreichend positive Trainingsbeispiele für alle Klassen gab. Für jede DDC-Klasse müsste es mindestens 15 Trainingsdokumente geben. Aber schon auf der zweiten Hierarchieebene konnten nicht mal die Hälfte aller Klassen mit ausreichend Trainingsbeispielen versorgt werden. Wenn es gelänge für eine größere Menge DDC-Klassen Trainingsbeispiele zu sammeln, dann könnten auch auf der zweiten und dritten Hierarchieebene bessere Ergebnisse zu erzielt werden (vgl. Waltinger, Mehler, Lösch, Horstmann 2011, S. 35-38). Trotzdem funktioniert der Klassifizierer des DFG-Projekts sehr gut, wie ein Vergleich mit den Ergebnissen aus dem eigenen Experiment, im letzten Kapitel dieser Bachelorarbeit, zeigen wird.

6.2 CrissCross[13]

Das von der DFG im Zeitraum von 2008 bis 2011 geförderte Projekt CrissCross wurde von der Fachhochschule Köln in Zusammenarbeit mit der Deutschen Nationalbibliothek durchgeführt. Ziel des Projektes war es einen multilingualen Zugang zu heterogen erschlossenen Informationsressourcen zu schaffen. Das bedeutet es soll möglich werden mit einer Suchanfrage, auf einer Suchoberfläche, gleichzeitig, mehrere unterschiedlich erschlossene Metadaten abzusuchen. Zwischen den verschiedenen Wissensorganisationssystemen (dt. für Knowledge Organisation System, abgekürzt: KOS) gibt es eine Vielzahl sprachlicher und struktureller Unterschiede. Jedes Wissensorganisationssystem repräsentiert das Wissen aus einem anderen Blickwinkel. Für die in den Wissensorganisationssystemen enthaltenen Begrifflichkeiten und Strukturen ergibt sich so eine perspektivische Vielfalt. Um diese zu überwinden werden Verknüpfungen zwischen den Begriffen unterschiedlicher Wissensorganisationssysteme hergestellt (vgl. Hubrich 2009a, S. 3). In CrissCross wurden Verknüpfungen zwischen drei Wissensrepräsentationssystemen hergestellt.

Das Projekt baut auf den Ergebnissen von zwei Vorgängerprojekten auf, welche bereits im Zusammenhang mit der DDC erwähnt wurden: DDC Deutsch und MACS. In MACS wurden die SWD-Schlagwörter, die englische Schlagwortsprache LCSH und die französische Schlagwortsprache RAMEAU gleichwertig miteinander verbunden. Dadurch wurden für CrissCross die Voraussetzungen geschaffen, um relativ leicht 70.000 SWD-Sachschlagwörter mit den LCSH und RAMEAU zu verbinden. Die wesentliche Aufgabe im Projekt CrissCross bestand jedoch darin 160.000 Sachschlagwörter der deutschen Schlagwortnormdatei mit Notationen der DDC zu verknüpfen (vgl. Hubrich 2008, S. 50).

6.2.1 Schlagwortnormdatei (SWD)

Die Schlagwortnormdatei ist eine universelle Indexierungssprache und wird nach den *Regeln für den Schlagwortkatalog* (im Folgenden: RSWK) und den *Praxisregeln zu den RSWK und den SWD* erstellt. Sie wird seit 1988 in den deutschsprachigen Ländern angewendet. Die SWD besteht aus ca. 550.000 Sachschlagwörtern, welche sich auf 37 Sachgruppen verteilen.[14] Darüber hinaus gibt es noch Personenschlagwörter,

[13] Vgl. http://linux2.fbi.fh-koeln.de/crisscross/ (letzter Abruf am: 18.04.2012)
[14] Vgl. http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/inhaltserschliessung/swd_syst.pdf;jse

Geographische/Ethnographische Schlagwörter, Zeitschlagwörter und Formschlagwörter. (vgl. Konferenz für Regelwerksfragen 1998, S. V-XII). Diese spielen allerdings für das Projekt CrissCross keine Rolle. Zwischen den Sachschlagwörtern gibt es Äquivalenz, hierarchische, und assoziative Beziehungen, allerdings sind diese sehr viel schlechter ausgeprägt als in der DDC. Laut einer unveröffentlichten Studie aus dem Jahr 2004, die im Auftrag der Fachhochschule Köln und der Deutschen Nationalbibliothek erstellt wurde, haben fast 87% der Sachschlagwörter keine assoziativen Beziehungen. 34% der Sachschlagwörter haben sogar weder hierarchische, noch assoziative Beziehungen (vgl. Jacobs, Mengel, Müller 2010, S. 237). Der Bedeutungsumfang eines Schlagwortes ergibt sich im Wesentlichen aus der Benennung. Weitere Informationen über den Kontext eines Schlagwortes können aus der Quelle, nach der das Schlagwort ursprünglich angesetzt wurde, oder aus der vergebenen Systematiknummer bezogen werden (vgl. Hubrich 2010a, S. 84). Darüber hinaus können der Verwendungshinweis, die Schlagwortrelationen oder die Verwendung des Schlagwortes im KVK (Karlsruher Virtueller Katalog) sowie in der WinIBW als Quellen zur Ermittlung des Bedeutungsumfangs herangezogen werden. Im Leitfaden zur Vergabe von DDC-Notationen an SWD-Schlagwörtern (2010, S. 1-2) wird sehr deutlich darauf hingewiesen, dass die mit einem Schlagwort verschlagworteten Werke nicht als Grundlage für das Mapping genommen werden dürfen. Es wird ausdrücklich betont, dass nur die durch die Schlagwörter repräsentierten Themen relevant sind.

6.2.2 „One-to-Many-Mapping“ und „Deep-Level-Mapping“

Da Begriffe bei der SWD, im Gegensatz zur DDC, in fast keinem übergeordneten Zusammenhang stehen, können sie verschieden interpretiert werden. Die meisten SWD-Schlagwörter sind Polysem und können in verschiedene DDC-Klassen gemappt werden. Dies geschieht fast immer und wird auch als „One-to-Many-Mapping“ bezeichnet. Durch die bereits beschriebenen großen strukturellen Unterschiede zwischen SWD und DDC ist es auch nicht möglich eine symmetrische Verknüpfung in beide Richtungen zu erzeugen. Das Mapping erfolgt immer nur in eine Richtung und zwar von der SWD zur DDC (vgl. Hubrich 2010b, S. 236). Dabei werden die DDC-Notationen für ein Schlagwort einfach in das Feld 816, des Pica-Normdatensatzes integriert. Pica ist das Katalogisierungssystem von OCLC, welches in vielen deutschen Bibliotheksverbünden und in der Deutschen Nationalbibliothek genutzt wird. Die Arbeitsumgebung von Pica nennt sich WinIBW (vgl. Hubrich 2008, S. 50). Wie bereits erwähnt, können einem Schlagwortnormdatensatz gleich

---

ssionid=A4933F9F72E3F66745DE78D7EBEE74E9.prod-worker5?__blob=publicationFile (letzter Abruf am: 16.04.2012)

mehrere DDC-Notationen zugewiesen werden. Das Sachschlagwort *Kaffee* kann beispielsweise aus drei verschiedenen Blickwinkeln betrachtet werden. Daraus ergeben sich demzufolge auch drei verschiedene Notationen:

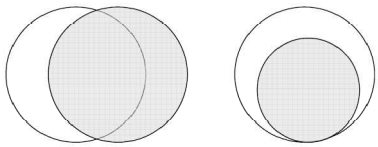
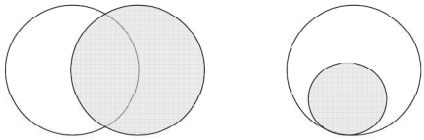
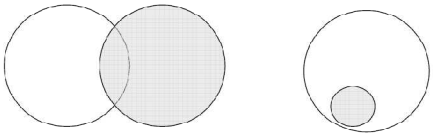
1) Die Notation *583.93 Gentianales (Enzianartige)*. Gentianales gehören zur Gruppe der bedecktsamigen Pflanzen in das Fachgebiet „Botanik“.
2) Die Notation *633.73 Kaffee*. Diese Klasse ist der Disziplin „Landwirtschaft“ zugeordnet.
3) Die Notation *641.3373 Kaffee*. Der Oberbegriff lautet hier „Hauswirtschaft & Familie“.

Die Bedeutung eines SWD-Schlagwortes ist oft sehr spezifisch. Um dem gerecht zu werden, werden auch die DDC-Notation so spezifisch wie möglich gestaltet. Dieses sogenannte „Deep-Level-Mapping“ führt zum Beispiel zu der synthetischen Notation *633.732* für das Schlagwort *Kaffeezüchtung*. Die Notation setzt sich zusammen aus zwei Klassen: *633.73 Kaffee* und *631.52 Produktion vermehrbarer Organismen und neuer Sorten* (vgl. Jacobs, Mengel, Müller 2010, S. 237).

6.2.3 Determiniertheitsgrade („Differenziertes Mapping“)

Die zugeordneten Notationen entsprechen aber nicht immer hundert prozentig einem SWD-Sachschlagwort. Um hier eine differenzierte Betrachtung und Suche zu ermöglichen wurden vier Determiniertheitsgrade (abgekürzt D) eingeführt. Dieses Grundprinzip wird vom CrissCross-Projektteam als „Differenziertes Mapping“ bezeichnet. Differenziertes Mapping gibt Auskunft darüber, inwieweit ein Sachschlagwort mit einer DDC-Notation deckungsgleich ist. Um die Aktualität und Gültigkeit eines Mappings nachprüfbar zu gestalten, wurde zusätzlich ein Datumsstempel eingeführt. Die vier Determiniertheitsgrade definieren sich wie folgt (vgl. Projektmitarbeiter des DFG-Projekts CrissCross 2010, S. 1-7):

D4	Bei Determiniertheitsgrad 4 (D4) stimmen Bedeutungsumfang von SWD-Schlagwort und DDC-Klasse vollständig überein. Sie sind äquivalent zueinander. Ein Schlagwort kann immer nur einer DDC-Notation mit einem D4 zugeordnet sein. Umgekehrt kann es in einer DDC-Klasse immer nur ein Schlagwort mit D4 geben.

D3	Determiniertheitsgrad 3 (D3) bedeutet eine wesentliche Übereinstimmung zwischen DDC-Klasse und SWD-Schlagwort. Das ist immer dann der Fall, wenn ein Schlagwort mehr als die Hälfte des Themas, dass durch eine DDC-Klasse repräsentiert wird, abdeckt. Oft gilt das für Themen, welche in den *Hier-auch-Anweisungen* der DDC stehen. 
D2	Der Determiniertheitsgrad 2 (D2) wird vergeben, wenn ein SWD-Schlagwort Teilaspekte einer DDC-Klasse wiederspiegelt, oder einen geringeren Bedeutungsumfang als die Klasse hat. Das können zum Beispiel Schlagwörter sein, die Unterbegriffe einer DDC-Klasse repräsentieren. Auch Themen, die in *Einschließlich-Hinweisen* der DDC stehen gehören dazu. 
D1	Wenn DDC-Klasse und SWD-Schlagwort nur eine geringe thematische Schnittmenge haben, wird der Determiniertheitsgrad 1 (D1) vergeben. Es handelt sich hierbei um Schlagwörter, deren Bedeutungsumfang deutlich von dem der DDC-Klasse abweicht oder sehr viel geringer ist, als dieser. 

6.2.4 Anwendung der CrissCross-Daten

Die Mappings in CrissCross sind hauptsächlich dazu da, um Retrievalprozesse in heterogenen Informationsräumen zu optimieren. Dabei sind drei Anwendungsszenarien vorstellbar:

Verbesserung des Zugangs zu DDC-Klassen und DDC-Indexierten Dokumenten:
Um über einen Begriff zu einer DDC-Klasse zu kommen, spielt das Register eine wichtige Rolle. Durch das Mapping von der SWD zur DDC wurde das Vokabular des DDC-Registers gewissermaßen um 160.000 Sachschlagwörter erweitert. Das Schlagwort *Drehkran* wurde beispielsweise mit einem Determinierheitsgrad 2 (D2) in die Klasse

621.873 Krane gemappt. Ein Bibliotheksbenutzer könnte so von dem Schlagwort *Drehkran* zur DDC-Klasse *621.873 Krane* weitergeleitet werden. Da bei diesem Mapping nur der Determiniertheitsgrad 2 vergeben wurde, dürften viele der gefunden Dokumente nur eine geringe Relevanz aufweisen. Die Besten Suchergebnisse sind mit einer Kombination aus [SWD-Schlagwort]: *Drehkran* AND [DDC-Klasse]: *621.873* zu erwarten (vgl. Jacobs, Mengel, Müller 2010, S. 238).

Strukturierung von Suchergebnissen:
Die Mappings zwischen SWD und DDC können sehr gut benutzt werden, um Suchergebnisse zu strukturieren. Bei einer DDC-Suchanfrage sind meistens nicht alle Ergebnisse relevant. Oft sucht der Nutzer nur ein Teil eines Themas, dass durch eine DDC-Klasse repräsentiert wird. Hier könnte ein Ranking nach Determiniertheitsgraden das Ergebnis verfeinern. Zum Beispiel könnten alle Dokumente, die mit einem SWD-Schlagwort indexiert sind, welches zu dieser DDC-Klasse mit D3 oder D4 gemappt wurde, höher gerankt werden. Des Weiteren können die Suchergebnisse anhand der Schlagwörter und der mit ihnen verbundenen Determiniertheitsgrade gruppiert werden. Allerdings muss die Reihenfolge der Schlagwörter nicht immer unbedingt die Reihenfolge sein, in der sie relevant für den Nutzer sind. Um dem Nutzer eine Möglichkeit zu geben, dass Suchergebnis auf seine eigenen Bedürfnisse einzuschränken, könnte er aus einer Liste von Schlagwörtern, die für ihn relevanten, auswählen.

Andersherum kann auch das Ergebnis einer Schlagwortsuche mit den SWD-DDC-Mappings strukturiert werden. Wenn eine Suchanfrage zu viele Treffer liefert, kann die Precision verbessert werden, indem das Suchergebnis auf bestimmte DDC-Klassen eingegrenzt wird. Die Auswahl der DDC-Klassen kann durch den Nutzer getroffen oder automatisiert werden. Bei sehr großen Treffermengen bietet es sich vielleicht an das Suchergebnis von vorneherein automatisch auf DDC-Klassen mit den höchsten Determiniertheitsgraden zu beschränken. Wenn eine Suchanfrage zu wenige Treffer liefert, gibt es drei Varianten den Recall über die DDC-Mappings zu vergrößern. Eine Möglichkeit wäre eine zusätzliche Suche nach den gemappten DDC-Klassen. Auch die End-Trunkierung der gemappten DDC-Klassen würde den Recall vergrößern. Die dritte Variante wäre eine Stufe höher in der DDC-Hierarchie zu steigen. Dabei gibt es allerdings zu bedenken, dass dieser Schritt auch einen Verlust an Precision nach sich zieht (vgl. Jacobs, Mengel, Müller 2010, S. 238-239).

Begriffsexploration:

Durch den Einzug neuer Semantic-Web-Technologien in das Bibliothekswesen sind explorative Suchen zu einem Trend geworden. Explorative Suchen bieten dem Informationssuchenden eine Hilfestellung zur schnelleren Orientierung in einem neuen Themengebiet. Dadurch kann er zielgerichteter suchen und kommt schneller zu besseren Ergebnissen (vgl. Jacobs, Mengel, Müller 2010, S. 239). Im Prinzip meint Begriffsexploration mittels CrissCross die Erweiterung der Schlagwortrelationen durch das DDC-Relationsgefüge. Durch CrissCross wurden die Schlagwortnormdatensätze um einige zusätzliche DDC-Notationen ergänzt. Dies wird auch *integrierte begriffliche Interoperabilität* genannt, da die Verknüpfungen nicht durch Verlinkungen zu anderen, externen Datenbanken entstanden sind, sondern integrale Bestandteile der Dokumentationssprache geworden sind. Die DDC-Notationen haben strukturbildende Wirkung auf die SWD. Dadurch werden Schlagwörter zueinander in Beziehung gesetzt, die vorher nicht zueinander in Beziehung standen. Es entstehen Schlagwortcluster mit denen sich Informationssuchende einen besseren Überblick über das begriffliche Umfeld ihrer Suche verschaffen können. Des Weiteren können anhand der Notationslänge auch hierarchische Beziehungen ermittelt werden. Abbildung 7 zeigt wie aus zuvor unzusammenhängend nebeneinander stehenden Schlagwörtern ein Relationsgefüge entsteht.

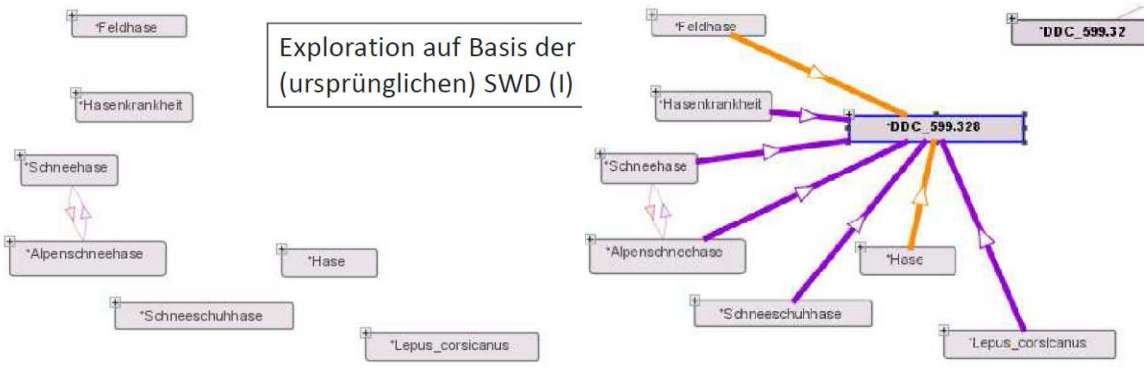


Abbildung 7: Clustering von Schlagwörtern durch CrissCross-Mappings

Aber die DDC führt nicht nur zu neuen Verbindungen zwischen Schlagwörtern, sondern auch zu einer Spezifizierung bereits bestehender Relationen. In der SWD ist das Relationsgefüge weit weniger ausgeprägt als in der DDC. Das führt dazu, dass es viele Begriffe gibt, die mit ein und derselben Relation verknüpft sind. Abbildung 8 zeigt wie die Unterbegriffe von „Jagd“ durch die DDC-Notation thematisch sortiert werden können. So lassen sich die neun Unterbegriffe von „Jagd“ nach „Jagdmethoden“, „Jagd mit Tieren“

und „Jagd auf einzelne Tierarten“ differenzieren. Eine schnellere Einschätzung der Relevanz der Unterbegriffe im Hinblick auf das Suchinteresse wird möglich (vgl. Hubrich 2009a, S. 6-8 und Hubrich 2009b).

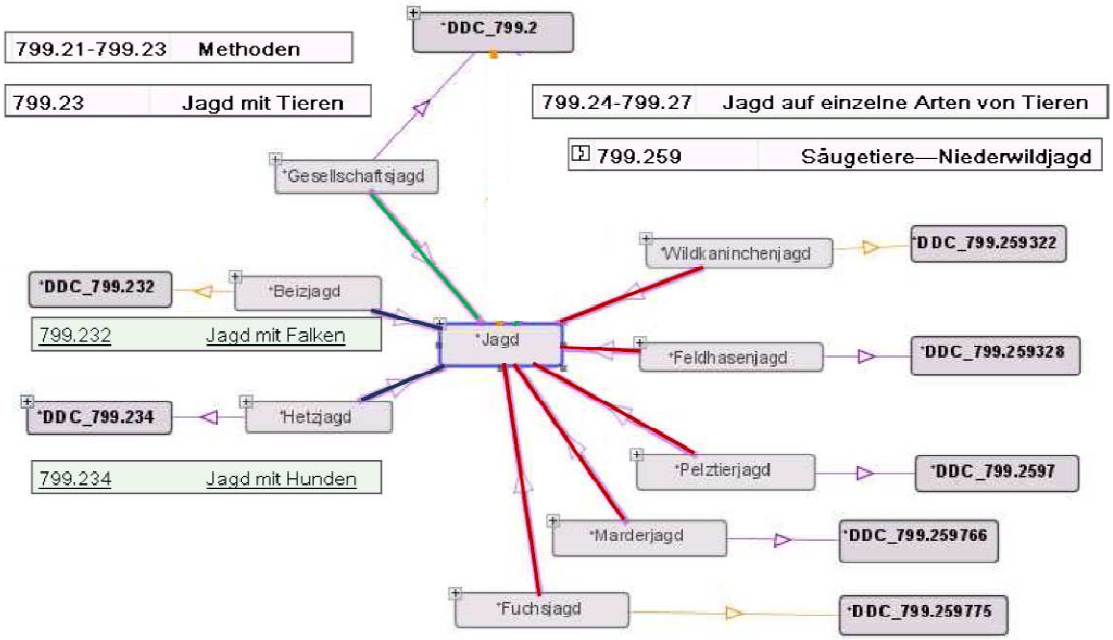


Abbildung 8: Differenzierung von SWD-Relationen durch CrissCross-Mappings

7 Das CrissCross-Mapping als Semantic-Web-Ontologie

Im Zuge der Linked-Data-Bewegung wurde von der Deutschen Nationalbibliothek (im Folgenden: DNB) eine Semantic-Web-konforme Version der CrissCross-Mappings erstellt. Ziel der Linked-Data-Bewegung ist es, Wissensrepräsentationssysteme, wie Thesauri, Lexika, Klassifikationssysteme, bibliographische Daten, Normdaten, o.ä. im Internet frei zur Verfügung zu stellen. Dadurch können sie besser zwischen den verschiedenen Institutionen im Bereich des Bibliotheks- und Informationswesens ausgetauscht und nachgenutzt werden. Die Deutsche Nationalbibliothek hat langfristig das Ziel sämtliche ihrer Wissensrepräsentationssysteme der „Linked-Data-Cloud“ zur Verfügung zu stellen. Bisher wurden ca. 1,3 Millionen Körperschaften aus der Gemeinsamen Körperschaftsdatei (im Folgenden: GKD), ca. 2 Millionen individualisierte Personen aus der Personennamendatei (im Folgenden: PND) und ca. 188.000 Sachschlagwörter aus der SWD in eine Semantic-Web-konforme Darstellungsweise gebracht. Des Weiteren sind auch 51.748 DDC-Klassen und 110 DDC-Sachgruppen in diese Datenbasis integriert. Resource Description Framework (im Folgenden: RDF) ist das Standard-Format um Daten in eine Semantic-Web-konforme Darstellungsweise umzuwandeln. Doch bevor eine RDF-

Ontologie erstellt werden kann, müssen die Datenstrukturen der DNB-Normdaten durch geeignete Vokabulare abgebildet werden. Dieser Prozess wird ebenfalls als Mapping bezeichnet. In diesem Fall bedeutet es, die Pica+-Felder (von der DNB verwendetes Internformat) werden durch Ausdrücke einer Web-Ontologie-Sprache ersetzt.

Für die Modellierung der SWD schienen zwei Vokabulare in syntaktischer Hinsicht am vielversprechendsten: Dublin Core und Simple Knowledge Organisation System (im Folgenden: SKOS). Darüber hinaus wurden von der DNB eigene Elemente mit dem Namensraum *gnd* (Gemeinsame Normdatei) erschaffen. Dies war nötig da sich herausgestellt hat, dass nicht alle Charakteristika der SWD durch SKOS und Dublin Core abzubilden waren. Die *gnd-Elemente* sind noch nicht registriert und können daher von der DNB noch in ihrer Verwendungsweise angepasst werden. Eine Registrierung soll allerdings in Zukunft erfolgen (vgl. Deutsche Nationalbibliothek 2012, S. 1-11). Die nachfolgende Tabelle zeigt, welche Elemente zur Beschreibung der Sachschlagwörter angewandt werden (vgl. Deutsche Nationalbibliothek 2012, S. 23-24). Dort ist zu sehen, dass das SKOS-Vokabular für die Repräsentation vieler Pica-Felder genommen werden konnte.

Tabelle 5: Repräsentation der Pica-Felder durch das SKOS-Vokabular

PICA3	PICA+	Ind.	Feldinhalt	RDF-Element	Bemerkung
021	007Q	$0	**SWD-Nummer**	dcterms:identifier	(DE-588c)…
026	007G	$0	**Identifikationsnummern umgelenkter Datensätze** (nur bei Hinweissätzen)	gnd:invalidIdentifier ForTheSubject	(DE-888c)…
601	041G	$s	**Nicht Deskriptor**	rdfs:label	
606	0410	$0	**Zu verknüpfende Deskriptoren**	gnd:useIndsteadSWD	Link auf URI des SWD-Satzes mittels Expansion der NID aus $0
800	041A	$s	**Hauptschlagwort**	skos:prefLabel	Mit Sprachattribut „de“
808			**Erläuterungen zum Schlagwort**		
	046A	$S	„|b|“ Indikator für Definition	skos:definition	Mit Sprachattribut „de“
	046A	$S	„|c|“ Indikator für Benutzungshinweise	skos:scopeNote	Mit Sprachattribut „de“

	046A	$a	Text gemäß Indikator		Übernahme des Inhalts gemäß Indikator in Literal des zugehörigen Elements
811	042B	$a	**Ländercode nach DIN EN 23166 (ISO 3166)**	gnd:countryCodeFor TheSubject	
812	042C	$a	**Sprachencode nach ISO/TC46/SC4-N350**	dcterms:language	
830	041F	$s	**Äquivalente Bezeichnung**	skos:altLabel	Mit Sprachattribut „de“
845	041S		**Übergeordneter Begriff zu Individualbezeichnung**	skos:broader	Mit Sprachattribut „de“ mehrgliedriger Oberbegriff
850	039C	$s	**Übergeordnetes Schlagwort**	skos:broader	Mit Sprachattribut „de“
860	039D	$s	**Verwandtes Schlagwort**	skos:related	Mit Sprachattribut „de“

Die Hauptschlagwörter werden durch *skos:prefLabel*, die äquivalenten Bezeichnungen durch *skos:altLabel*, verwandte Schlagwörter durch *skos:related* und übergeordnete Schlagwörter durch *skos:broader* ausgedrückt. Neben übergeordneten Schlagwörtern wird *skos:broader* auch bei übergeordneten Begriffen für Individualnamen benutzt. Das sind mehrgliedrige Oberbegriffe, die sich neben Sachschlagwörtern unter anderem auch aus Personenschlagwörtern oder Körperschaftsschlagwörtern zusammensetzen können. Für die Definition eines Schlagwortes wird *skos:definition* und für die Benutzungshinweise wird *skos:scopeNote* verwendet. Der Sprachencode wird durch *dcterms:language* repräsentiert. Er steht nicht für die Sprache des Schlagwortes, sondern für eine sprachenspezifische Besonderheit im Zusammenhang mit dem Schlagwort. Die von der DNB eigens entwickelten GND-Elemente werden für Ländercodes (*gnd:countryCodeForTheSubject*) und Identifikationsnummern umgelenkter Datensätze (*gnd:invalidIdentifierForTheSubject*) benutzt. Das von der DNB entwickelte RDF-Element *gnd:useInsteadSWD* für Deskriptoren, die via URI mit einem anderen Deskriptor verknüpft werden sollen, wird in der aktuellen RDF/XML-Datei noch nicht verwendet. Der gültige Identifier, die SWD-Nummer, wird mit *dcterms:identifier* ausgewiesen.

Wie bereits im Zusammenhang mit dem Projekt CrissCross erwähnt wurde, sind auch Verknüpfungen zwischen der SWD und den Schlagwortsprachen Library of Congress Subject Headings (LCSH) und Répertoire d'autorité-matière encyclopédique et

alphabétique unifié (RAMEAU) hergestellt worden. Diese Verknüpfungen konnten ebenfalls durch SKOS repräsentiert werden. Sowohl die LCSH, als auch die RAMEAU werden durch das Element *skos:closeMatch* dargestellt (vgl. Deutsche Nationalbibliothek 2012, S. 24-25).

Die Kernaufgabe des CrissCross-Projektes bestand jedoch darin unidirektionale Verbindungen zwischen SWD-Sachschlagwörtern und DDC-Notationen zu schaffen. Um die CrissCross-Mappings im Linked-Data-Service zur Verfügung zu stellen, war es nötig sie in der Semantic-Web-Ontologie durch ein entsprechendes Vokabular zu repräsentieren. Problematisch war die Tatsache, dass SKOS immer eine bidirektionale Beziehung impliziert. Es musste also ein Weg gefunden werden die durch SKOS repräsentieren Sachschlagwörter mit den DDC-Notationen zu verknüpfen, ohne dabei das Prinzip der Unidirektionalität zu verletzen. Um dieses Problem zu lösen, wurde das *Coordinated Concept* erfunden.

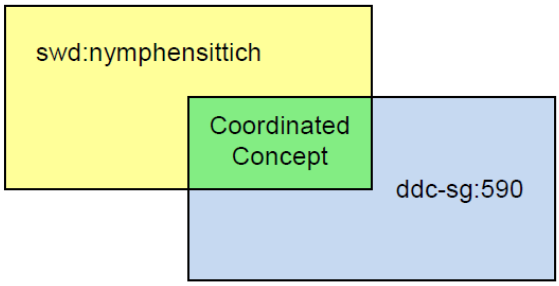


Abbildung 9: Schematische Darstellung des Coordinated Concept

```
<skos:closeMatch rdf:resource="http://id.loc.gov/authorities/sh85067435#conc
<skos:altLabel xml:lang="de">Politische Konflikte</skos:altLabel>
<skos:broader rdf:resource="http://d-nb.info/gnd/4032081-9"/>
<skos:related rdf:resource="http://d-nb.info/gnd/4046539-1"/>
<dnb:hasCoordinatedConcept-of>
  <dnb:CoordinatedConcept>
    <dnb:coordination-of rdf:resource="http://d-nb.info/ddc-sg/320"/>
    <dnb:coordination-of rdf:resource="http://d-nb.info/gnd/4115589-0"/>
    <dnb:det2 rdf:resource="http://d-nb.info/ddc/class/320.9"/>
  </dnb:CoordinatedConcept>
</dnb:hasCoordinatedConcept-of>
<dcterms:identifier>(DE-588)041155890</dcterms:identifier>
<skos:prefLabel xml:lang="de">Politischer Konflikt</skos:prefLabel>
</skos:Concept>
```

Abbildung 10: Praktische Umsetzung des Coordinated Concepts

Das Coordinated Concept besteht aus den Teilkonzepten *dnb:coordination-of* und den Determiniertheitsgraden (*dnb:det1* bis *dnb:det4*). Für jede DDC-Klasse, die mit einem SWD-Schlagwort verknüpft wurde, wird ein eigenes koordiniertes Konzept angelegt. Es besteht immer aus zwei Teilkonzepten und dem Determiniertheitsgrad. Das Teilkonzept

dnb:coordination-of wurde von *skos:broader* abgeleitet um einen weitreichenden Interpretationsspielraum zu gewährleisten. Auch die vier Determiniertheitsgrade wurden von bereits bestehenden SKOS-Vokabularien abgeleitet. Die Determiniertheitsgrade 4 und 3 wurden von *skos:closeMatch* abgeleitet, Determiniertheitsgrad 2 von *skos:broadMatch* und Determiniertheistgrad 1 wurde von *skos:relatedMatch* abgeleitet. Neben der Verknüpfung von SWD-Schlagwörtern mit DDC-Klassen, sollte auch eine Crawling-Funktion impliziert werden. Dies wurde durch das zusätzliche Element *dnb:hasCoordinatedConcept-of* ermöglicht (vgl. Deutsche Nationalbibliothek 2012, S. 27-30).

Wie bereits erwähnt wurde, kann die eben beschriebene RDF/XML-Ontologie von der Linked-Data-Service-Seite der DNB unter der „Creativ-Commons-Zero-Lizenz“ frei heruntergeladen werden.[15] Dadurch ergeben sich viele Möglichkeiten die Datei weiterzuverwenden und in andere Anwendungen zu integrieren. Eine Möglichkeit ist die des automatischen Klassifizierens von OAI-Metadaten, die nachfolgend genauer untersucht wird.

8 Automatisches Klassifizieren mit der CrissCross-Ontologie

Das letzte Kapitel dieser Bachelorarbeit beschreibt einen Versuch die RDF/XML-Ontologie zu modifizieren und in das Text-Mining-Programm GATE zu integrieren. Hier schließt sich auch der Kreis zum DFG-Projekt „Automatische Anreicherung von OAI-Metadaten“, denn die DDC-Notationen sollen nur anhand der Dublin-Core-Felder *description*, *title* und *subject* generiert werden.

Dazu wird zuerst die Textminig-Software *GATE* beschrieben. Anschließend wird das „Ontologie-Werkzeug“ *Apolda* erklärt. Apolda ist ein kleines Programm, welches in GATE integriert wird. Es fungiert sozusagen als Schnittstelle um die CrissCross-Ontologie für GATE nutzbar zu machen. Neben Apolda werden in GATE noch sechs weitere „kleine Programme“ verwendet um die DDC-Notationen zu generieren. Die Abfolge, nach der die Programme die Metadaten bearbeiten, wird in einer Pipeline festgelegt. Der Abschnitt *Korpus-Pipeline* beschreibt dieses Vorgehen. Damit Apolda und zwei weitere Programme mit der CrissCross-Ontologie arbeiten können, mussten ein paar kleine Veränderungen vorgenommen werden. Dies wird zusammenfassend in dem Abschnitt *Modifikationen an*

[15] Vgl. https://wiki.d-nb.de/display/LDS/Dokumentation+des+Linked+Data+Services+der+DNB (letzter Abruf am: 21.04.2012)

der CrissCross-Ontologie besprochen. Nach einer kurzen Beschreibung des *Korpus*, erfolgt eine ausführliche Darstellung der *Ergebnisse*.

8.1 GATE

GATE ist ein Akronym und bedeutet aufgelöst *General Architecture for Text Engineering*. Das Programm vom Departement of Computer Science der University of Sheffield gibt es bereits seit über 15 Jahren. Es hat inzwischen weltweite Verbreitung gefunden, was wohl auch der Tatsache geschuldet ist, dass es sich um eine Open Source Software handelt. Es kann sowohl für große Sprachverarbeitungsprojekte mit großen Datenmengen, als auch für kleine Studentenprojekte verwendet werden.[16] GATE stellt eine graphische Entwicklungsumgebung mit einigen Basiskomponenten zur Verfügung. Diese können, je nach Bedarf, variabel angepasst oder erweitert werden. Darüber hinaus gibt es viele Möglichkeiten neue Komponenten in das Programm zu integrieren. Das besondere an GATE ist die Trennung von Datensicherung, Datenvisualisierung, dem Laden von Komponenten und Dokumenten und dem tatsächlichen Prozess der automatischen Sprachverarbeitung. Alle Komponenten von Gate können einem der folgenden drei Typen zugeordnet werden:

- **LanguageResources** (LRs): Dazu gehören Thesauri, Lexika, Ontologien und Text-Dokumente
- **ProcessingResources** (PRs): Dazu gehören alle Komponenten zur automatischen Sprachverarbeitung, die auf einem Algorithmus basieren. Z.B. Tokenizer
- **VisualResources** (VRs): Das sind graphische Benutzeroberflächen. Dazu gehören alle Komponenten zum Visualisieren und Editieren.

Diese Dreiteilung hat den Vorteil, dass die verschieden Komponenten unabhängig voneinander weiterentwickelt werden können. So kann beispielsweise ein Linguist einen Thesaurus unabhängig von dem Fachwissen eines Programmierers weiterentwickeln. Es können auch verschiedene Visualisierungsvarianten entwickelt werden, ohne dass dabei das Sprachverarbeitungsprogramm verändert werden muss.

Alle Programme zur automatischen Sprachverarbeitung werden in GATE unter dem Oberbegriff CREOLE (a Collection of REusable Objects for Language Engineering) zusammengefasst. Nach der Installation stehen in GATE bereits eine ganze Reihe von

[16] Vgl. http://gate.ac.uk/overview.html (letzter Abruf am: 24.04.2012)

(Standard-)Programmen zur Verfügung. Jedes dieser Programme kann über die „Plugin Management Console“ geladen werden. Neben den bereits vorhandenen Programmen können auch eigene Programme in GATE geladen werden. Damit diese von GATE erkannt werden können, müssen die Initialisierungsparameter in einer XML-Datei (creole.xml) abgespeichert werden. Ein Programm, welches über die „Plugin Management Console“ geladen wurde, kann mehrere *ProcessingResources* enthalten. Eine ProcessingResource kann als ein kleines Miniprogramm betrachtet werden, dass nur „einen Schritt“ im Prozess der Automatischen Sprachverarbeitung ausführt. Mit „einem Schritt“ ist zum Beispiel die Untergliederung eines Textes in Wörter, Leerzeichen und Interpunktionen gemeint. Dieser „Schritt“ gehört zu den Basics der Automatischen Sprachverarbeitung und wird von einem *Tokeniser* erledigt. Eine weitere ProcessingResource ist ein *Sentence Splitter*. Er unterteilt, wie der Name richtig vermuten lässt, Texte in Sätze. Welche ProcessingResources, in welcher Reihenfolge benutzt werden sollen, kann individuell und variabel festgelegt werden. Dabei gibt es nur zu beachten, dass einige PRs aufeinander aufbauen, wodurch sich automatisch eine festgelegte Reihenfolge ergibt. Zur Festlegung der Reihenfolge gibt es in GATE eine *Applikation* mit der eine *Pipeline* erstellt werden kann. In einer Pipeline geschieht genau das eben beschriebene. Aus allen, in GATE geladenen PRs, können diejenigen ausgewählt werden, welche zum Annotieren des Textes in Frage kommen. Nach der Auswahl der PRs erfolgt hier auch die Bestimmung der Reihenfolge in der die PRs einen Text annotieren sollen. Es gibt normale *Pipelines,* die auf einen einzigen Text angewandt werden und *Korpus-Pipelines,* die auf eine Sammlung von Texten (Korpus) angewandt werden. Abbildung 11 visualisiert die zuvor beschriebenen GATE-Komponenten. In der linken Spalte befinden sich die *Applications, LanguageResources, ProcessingResources* und die *Datastores*. Die Datastores dienen der dauerhaften Speicherung von annotierten Texten. In der rechten Spalte ist eine *Korpus-Pipeline* mit den ausgewählten ProcessingResources zu sehen. Diese Pipeline wurde auch für den in dieser Arbeit vorgestellten Versuch zum automatischen Klassifizieren benutzt und wird in einem eigenen Abschnitt näher erläutert.

Ein Text wird in GATE immer als *LanguageResource* hochgeladen. Dabei werden die folgenden Formate von GATE unterstützt: XML, RTF, HTML, SGML, E-Mail und „Plain-Text“ (vgl. Cunningham, Maynard, Bontcheva, Tablan 2002, S. 1-3). Wenn jetzt beispielsweise ein XML-Dokument in GATE geladen wird, dann werden die vorhanden Markups einfach vom Text separiert. Der Text bleibt nun immer „sauber“, d.h. ohne Markups. Sämtliche Annotationen, die später durch die Automatische Sprachverarbeitung

dazu kommen, werden auch vom Text getrennt behandelt. Bildlich gesprochen wird durch jede ProcessingResource eine neue Schicht Annotationen hinzugefügt. Diese Technik wird *Stand-off-Markup* oder *Stand-off-Annotation* genannt. GATE bietet die Möglichkeit die annotierten Texte später im XML-Format zu exportieren. XML ist als Austauschformat unschlagbar, denn egal welches spezifische Format die Annotationen in einem System auch haben mögen, sie können immer durch XSLT-Transformationen umgewandelt werden (vgl. Wilcock 2009, S.14-18). Auch Ontologien, Thesauri und Lexika werden in GATE als *LanguageResource* behandelt. Um eine Ontologie in GATE zu laden, wird das OWLIM2-Ontology-Tool benötigt. Es ist standardmäßig in GATE vorhanden und kann über die „Plugin Management Console“ aktiviert werden. Über die OWLIM-Ontology-Schnittstelle wurde auch die „CrissCross-Ontologie“ in GATE geladen. Die Integration der „CrissCross-Ontologie“ in die Korpus-Pipeline erfolgte durch *Apolda*.

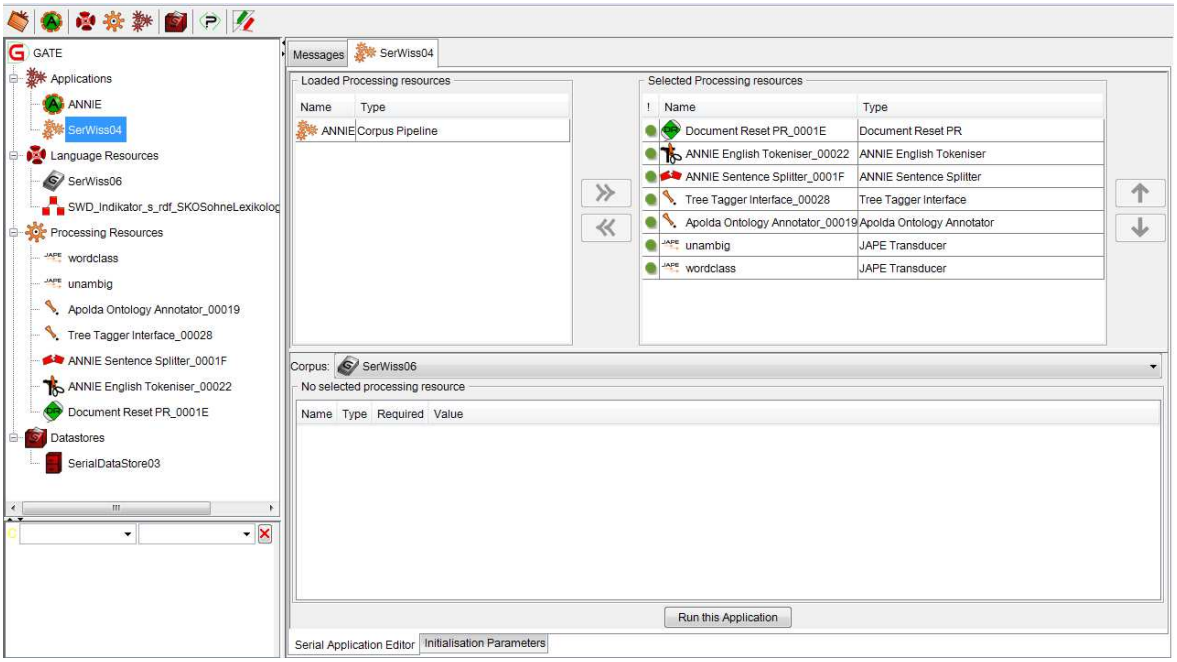


Abbildung 11: Screenshot von GATE mit der Korpus-Pipeline

8.2 Apolda[17]

Apolda ist ebenfalls ein Akronym und bedeutet *Automated Processing of Ontologies with Lexical Denotations for Annotation*. Es ist eine ProcessingResource, die als Plugin in GATE geladen wird. Genau wie GATE ist auch Apolda im Internet frei verfügbar. Es wird unter der *GNU Lesser General Public License* (LGPL) veröffentlicht. Apolda annotiert

[17] Vgl. http://apolda.sourceforge.net/ (letzter Abruf am: 25.04.2012)

Texte auf Basis einer Ontologie. Damit eine Ontologie zusammen mit Apolda funktioniert, müssen zwei Bedingungen erfüllt sein:

1) Die sprachliche Repräsentation eines Ontologie-Konzeptes muss in der Ontologie enthalten sein. In dieser Beziehung eignet sich die CrissCross-Ontologie perfekt für Apolda, denn die SWD-Sachschlagwörter sind quasi die textuelle Repräsentation der DDC-Klassen.
2) Die textuelle Repräsentation der Ontologie muss als OWL-Annotation-Property codiert werden (*owl:AnnotationProperty*).

Bei der Initialisierung der Ontologie wird festgelegt, welche *AnnotationProperties* für die Annotation des Dokuments genutzt werden sollen. Dafür stehen in Apolda zwei Parameter zur Verfügung: *prefRepresentation* und *altRepresentation*. Bei der Initialisierung der CrissCross-Ontologie wurde festgelegt, dass die textuelle Repräsentation für prefRepresentation das *skos:prefLabel* sein sollte. Die textuelle Repräsentation für altRepresentation sollte *skos:altLabel* sein. Es hätte aber auch genauso gut andersherum sein können, denn Apolda unterscheidet nicht zwischen pref- und altRepresentation. Darüber hinaus gibt es mit *language* noch einen dritten Initialisierungsparameter. Dieser Parameter ist allerdings nur von Bedeutung, wenn es in einer Ontologie mehrere textuelle Repräsentationen in verschiedenen Sprachen gibt. In diesem Fall könnte ein OWL-Sprachen-Attribut ausgewählt werden. Alle anderen Sprachen würden dann ignoriert werden. Für den Klassifikationsversuch in dieser Bachelorarbeit spielt der *language*-Initialisierungsparameter keine Rolle, da es mit der SWD nur eine deutschsprachige textuelle Repräsentation gibt.

Apolda annotiert Texte auf der Grundlage eines Vergleichs der textuellen Repräsentation mit den Wörtern im Text. Dies kann nur funktionieren, wenn der Text zuvor von einem *Tokeniser* in Wörter, Leerzeichen und Interpunktionen untergliedert wurde. Im Fall dieser Bachelorarbeit hat ein Abgleich zwischen den Sachschlagwörtern der SWD und den *Token* der OAI-Metadaten stattgefunden. Es ist auch möglich einen Text mit einem *Lemmatizer* zu bearbeiten und einen Vergleich mit *Lemmata* durchzuführen. Ein Lemma ist die Grund- oder Zitierform eines Wortes, z.B. Infinitiv, Nominativ oder Singular. Lemmata werden als Eigenschaften von Token abgespeichert. Apolda vergleicht immer nur Token mit der textuellen Repräsentation einer Ontologie. Bei einer textuellen Repräsentation, die aus mehr als zwei Token besteht, werden die zwei Token unabhängig voneinander betrachtet. Z.B. würde mit der textuellen Repräsentation „exhibiting painter“ in einem Text auch der

Ausdruck „exhibiting painters“ annotiert werden. Denn dieser Ausdruck setzt sich aus den lemmatisierten Token *exhibit* und *painter* zusammen. Das erste Wort trifft auf der String-Ebene und das zweite Wort auf der Lemma-Ebene zu. Ein Abgleich auf der Lemma-Ebene funktioniert allerdings nur, wenn der Laufzeitparameter *LemmaFeature* gesetzt ist. Dieser Laufzeitparameter überprüft, ob bei den Eigenschaften eines Tokens auch ein Lemma mit abgespeichert ist. Wenn das Lemma mit der textuellen Repräsentation in der Ontologie übereinstimmt, bekommt das Token eine Annotation.

Apolda annotiert Texte ohne dabei zu disambiguieren. Ein Wort oder Token kann so viele Annotationen bekommen, wie Apolda textuelle Repräsentationen in der Ontologie findet. Es ist sogar möglich, dass zwei Annotationen sich teilweise überlappen. Von diesem Prinzip gibt es nur eine Ausnahme: Wenn in einem Text der Ausdruck „Leonardo da Vinci“ steht und es zwei textuelle Repräsentationen für dieses Konzept gibt, nämlich „Leonardo“ und „Leonardo da Vinci“, dann vergibt Apolda nur eine Annotation. Das liegt daran, dass der String „Leonardo“ ein Teil von „Leonardo da Vinci“ ist. In allen anderen Fällen würde Apolda beide Wörter annotieren. Für jede gefundene textuelle Repräsentation kreiert Apolda die Annotation *Mention* (vgl. Wartena, Brussee, Gazendam, Huijsen 2007, S. 291). Beim Annotieren mit der CrissCross-Ontologie gab es bei manchen Ausdrücken teilweise mehr als zwanzig Annotationen. Um diese Ambiguitäten aufzulösen waren zwei weitere ProcessingRessources in der Korpus-Pipeline nötig. Die Funktionsweise dieser PRs wird im nächsten Abschnitt, im Zusammenhang mit dem Aufbau und der Zusammensetzung der Korpus-Pipeline, erläutert.

8.3 Korpus-Pipeline

Die für „das automatische Klassifizieren mit der CrissCross-Ontologie“ verwendete Korpus-Pipeline setzt sich aus den folgenden sieben ProcessingResources zusammen:

1) Document Reset PR
2) ANNIE English Tokeniser
3) ANNIE Sentence Splitter
4) Tree Tagger Interface
5) Apolda Ontology Annotator
6) Unambig (JAPE Transducer)
7) Wordclass (JAPE Transducer)

8.3.1 Document Reset PR

Diese ProcessingResource sollte am Beginn einer jeden Pipeline stehen, da sie sämtliche Annotationen auf null zurücksetzt. Das ist wichtig um Fehler zu vermeiden, falls ein Text schon mal mit GATE oder einem anderen Textverarbeitungsprogramm bearbeitet wurde. Es ist auch möglich das Zurücksetzen nur auf bestimmte Annotationstypen zu beschränken. Die Original Textauszeichnungen (Original Markups) bleiben beim zurücksetzen immer erhalten, allerdings werden sie von den neuen Annotationen separiert. Wenn die Original Textauszeichnungen nicht erwünscht sind, dann können sie gelöscht werden, indem der Parameter *KeepOriginalMarkupsAS* auf „false“ gesetzt wird (vgl. Cunningham, et al. 2012, S. 118-119).

8.3.2 ANNIE English Tokeniser

Ein *Tokeniser* unterteilt den Text in sogenannte „Token“. Das können Wörter (Word), Nummern (Number), Symbole (Symbol), Interpunktionen (Punctuation) oder Leerzeichen (Space Token) sein. Der Satz *„Michael gewinnt 700 €.“* würde vom Tokeniser wie folgt annotiert:

Michael		gewinnt		700		€	.
token	Space token	token	Space token	token	Space token	token	token

Ein Wort wird als eine ununterbrochene Folge von Groß- und/oder Kleinbuchstaben, sogar inklusive Bindestriche, definiert. Ein „Word-Token“ besitzt die Eigenschaft *orth*. Dieser Eigenschaft können die vier Werte *upperInitial* (der erste Buchstabe ist groß, der Rest ist klein), *allCaps* (alles Großbuchstaben), *lowerCase* (alles Kleinbuchstaben), *mixedCaps* (eine Mischung aus Groß- und Kleinbuchstaben) zugeteilt werden. Eine Nummer ist als irgendeine Abfolge von Zahlen definiert. Bei Symbolen wird zwischen Währungssymbolen (z.B. €, $, £) und anderen Symbolen (z.B. §, &, %) unterschieden. Jede Interpunktion ist ein eigener Token. Es gibt drei Arten von Interpunktionen: Start-Interpunktionen, End-Interpunktionen und andere Interpunktionen. Zu den Start- und End-Interpunktionen gehören zum Beispiel auch Anführungszeichen („ “). Es gibt zwei Unterscheidungsmöglichkeiten bei „Space-Token“. Entweder sie markieren Zwischenräume zwischen Wörtern (kind=space) oder sie strukturieren den Text (kind=control) (Cunningham, et al. 2012, S. 119-121). In diesem Fall ist es überhaupt nicht schlimm, dass ein englischer Tokeniser auf deutsche Texte angewandt wurde. Wichtig ist

nur die Untergliederung der OAI-Metadaten in Token und SpaceToken. Diese werden nämlich durch das *Tree Tagger Interface* mit Lemmata angereichert.

8.3.3 ANNIE Sentence Splitter

Der *Sentence Splitter* unterteilt Texte in Sätze. Um das Satzende von Punkten in Abkürzungen unterscheiden zu können, benutzt der Splitter eine Liste mit Abkürzungen. Es werden sowohl Annotationen für den ganzen Satz („Sentence"), als auch für die Punkte, die den Satz beenden („Split"), vergeben. Ein Sentence Splitter ist eine notwendige Vorstufe für den Tagger (vgl. Cunningham, et al. 2012, S. 123-124).

8.3.4 Tree Tagger Interface

Der *Tree Tagger* ist ein *Lemmatizer* und ein *Part-of-Speech-Tagger* (im Folgenden: POS-Tagger) in einem. Beim POS-Tagging wird jedem Wort innerhalb eines Satzes eine grammatikalische Markierung zugeordnet. Diese grammatikalischen Markierungen sind nicht identisch mit den zehn Wortarten (Substantiv, Verb, Adjektiv, etc.) der deutschen Grammatik, da beim POS-Tagging differenziertere Entscheidungen getroffen werden müssen. Daher gibt es viel mehr grammatikalische Markierungen, als es Wortarten gibt. Eine Sammlung solcher grammatikalischer Markierungen wird als *Tagset* bezeichnet. Ein sehr bekanntes englischsprachiges Tagset ist das *Penn Treebank Tagset*. Für die deutsche Sprache gibt es das *Stuttgart-Tübingen-Tagset*[18] (im Folgenden: STTS). Es besteht aus 55 Tags. Das folgende Beispiel zeigt einen mit STTS getaggten Satz. Die Tags stehen dabei in eckigen Klammern (vgl. Heyer, Quasthoff, Wittig 2008, S.126-129):

Kaffeetassen [NN], [$,] T-Shirts [NN], [$,] Videobänder [NN*] und [KON] Bierdeckel [NN] lassen [VVFIN] sich [PRF] als [APPR] Erinnerung [NN] an [APPR] das [ART] berüchtigte [ADJA] Gefängnis [NN] von [APPR] Alcatraz [NE] mit [APPR] nach [APPR] Hause [NN] nehmen [VVINF]. [$.]

Es gibt sowohl regelbasierte als auch stochastische Tagger. Der Tree Tagger ist ein stochastischer Tagger, basierend auf dem Markov-Modell. Allerdings wurde er um einen binären Entscheidungsbaum erweitert, weswegen er auch Tree Tagger genannt wird (vgl. Schmid 1994, S.44). Er kann auf viele verschiedene Sprachen angewandt werden. Für die deutsche Version wird das STTS-Tagset benutzt[19]. Auf das POS-Tagging soll an dieser Stelle nicht weiter eingegangen werden, da der Tree Tagger dem Apolda Ontology Annotator nur wegen der *Lemmatisierung* vorgeschaltet wurde. Bei der Lemmatisierung wird ein Wort auf seine lexikalische Grundform zurückgeführt. Das *Lemma* von „tanzte"

[18] Vgl. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/stts_guide.pdf (letzter Abruf am: 01.05.2012)
[19] Vgl. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (letzer Abruf am: 01.05.2012)

wäre beispielsweise „tanzen“. Die größte Schwierigkeit liegt darin, dass die deutsche Sprache reich an Morphologie ist. Während es im Englischen oft reicht, die Wortanfänge und Wortendungen zu streichen, geht das im deutschen nicht so einfach. Zum Beispiel ist das Lemma von Mäuse > Maus. Hier verändert sich sogar die Grundform. Manchmal kann die richtige Grundform auch nur aus dem Kontext erschlossen werden. Das Lemma von „Buchten“ könnte das Verb „buchen“ oder aber „die Bucht“ am Meer sein. Um die richtige Wortform zu finden, wird mit einem Lexikon gearbeitet. Hier wird die richtige Wortform quasi „nachgeschlagen“ (vgl. Kappler 2006, S. 3-4). Die *Lemmata,* als Ergebnis des Tree Tagger Interfaces, vergrößern die Wahrscheinlichkeit, ein Sachschlagwort im Text zu finden.

8.3.5 Apolda Ontology Annotator

Der Apolda Ontology Annotator ist gewissermaßen das Herzstück dieser Pipeline. Daher wurde er auch schon in einem vorangegangenen Abschnitt ausführlich beschrieben. Alle PRs vor Apolda sind nur dazu da, die Texte vorzubereiten, und alle PRs danach sind nur dazu da, das Ergebnis nachzubearbeiten. Beim Nacharbeiten waren im Wesentlichen zwei Probleme zu lösen. Der Betreuer dieser Bachelorarbeit, Herr Prof. Dr. Wartena, löste sie durch zwei selbst geschriebene PRs (Unambig und Wordclass) am Ende der Korpus-Pipeline. Doch bevor diese beschrieben werden können, soll zunächst das Ergebnis von Apolda etwas genauer betrachtet werden.

Wie bereits erwähnt wurde, annotiert Apolda ohne dabei zu disambiguieren. Das Wort „Stress“ beispielsweise wurde fünf Mal annotiert. Es kam ein Mal als Hauptschlagwort (*prefLabel*) und vier Mal als äquivalente Bezeichnung (*altLabel*) in der CrissCross-Ontologie vor. Bei den Konzepten, wo „Stress“ nur als äquivalente Bezeichnung auftaucht, waren die Hauptschlagwörter „Stressreaktion“, „Schulstress“, „Stressresistenz“ und „Stressbewältigung“. Diese Begriffe sind sehr spezifisch und das Ziel war eine DDC-Notation bis maximal zur dritten Hierarchieebene zu erreichen. Daher war eine erste Überlegung, nur die Hauptschlagwörter als textuelle Repräsentationen für die DDC zu verwenden. Bei dieser Herangehensweise gehen allerdings unter Umständen wichtige Sachschlagwörter verloren, wie zum Beispiel „Quality Management“. Dieser Ausdruck kommt in der gesamten CrissCross-Ontologie nur ein Mal vor und zwar als äquivalente Bezeichnung (altLabel). „Quality Management“ ist ein tragender Ausdruck und wurde, obwohl englischsprachig, als Schlagwort für einen deutschsprachigen Text vergeben. Mit dem Ausdruck sind zwei DDC-Notationen verknüpft: Mit dem Determiniertheitsgrad 2 die

Notation 350 und mit dem Determiniertheitsgrad 3 die Notation 650. Da am Ende die Notationen ausgezählt werden und die, die am Meisten vorkommt, „gewinnt“, könnte das weglassen der äquivalenten Bezeichnungen (altLabel) zu einer Verschiebung des Ergebnisses hin zu einer anderen Notation führen. Eine bessere Lösung für das Problem mit den Ambiguitäten wurde durch die PR **Unambig** gefunden.

Eine Annotation von Apolda hat drei Eigenschaften (engl. feature): *identifier*, *class* und *ontology*. Der *identifier* hat als Wert den Namen des repräsentierten Konzeptes. Die Klasse hat als Wert die Klasse des repräsentierten Konzeptes. Da die Konzepte in der CrissCross-Ontologie keine Namen haben, gibt es nur die Eigenschaft *class*. Die Eigenschaft *ontology* gibt einfach den Namen an, unter dem die Ontologie in Apolda hochgeladen wurde.[20] Wichtig ist eigentlich nur die Eigenschaft *class*. Sie hat als Wert die URL des Skos-Konzeptes, welche durch das Sachschlagwort repräsentiert wird (z.B. class=http://d-nb.info/gnd/4043774-7). Diese Information ist jedoch nicht ausreichend, daher wurde eine siebte PR namens **Wordclass** geschrieben, welche eine Annotation namens DDC mit der DDC-Notation als Eigenschaft vergibt.

8.3.6 Unambig (JAPE Transducer)

Die PR *Unambig* wurde von Herrn Prof. Dr. Wartena mit der JAPE-Grammatik geschrieben und für diese Bachelorarbeit zur Verfügung gestellt. JAPE steht für *Java Annotation Patterns Engine* und wurde extra entwickelt, um es GATE-Nutzern zu ermöglichen eigene kleine PRs zu schreiben, ohne dass sie dafür gleich ein eigenes Plugin kreieren müssen. Für das Hochladen und kompilieren der Grammatiken steht mit dem JAPE-Transducer bereits ein eigens für diesen Zweck kreiertes GATE-Plugin zur Verfügung. Eine JAPE-Grammatik baut auf den Annotationen von anderen PRs auf. Sie besteht aus Phasen und jede Phase besteht aus einer Reihe von Annotationsregeln, die nacheinander abgearbeitet werden. Die Annotationsregeln suchen nach Mustern in bereits bestehenden Annotationen und fügen diesen neue Annotationen hinzu. Eine Annotationsregel besteht aus einer linken und einer rechten Seite. Auf der linken Seite wird mit einem regulären Ausdruck das Muster beschrieben, nachdem gesucht werden soll. Auf der rechten Seite stehen Regeln, nach denen eine neue Annotation kreiert werden soll. Hier kann auch Java-Code benutzt werden, um Annotationen zu kreieren. Des Weiteren kann mit JAPE auf Ontologien zurückgegriffen werden, wenn diese als LanguageResource

[20] Vgl. http://apolda.sourceforge.net/, (letzter Abruf am: 05.05.2012)

in GATE geladen sind (vgl. Cunningham, et al. 2012, S. 189-191). Genau dies wurde sich auch in der „Unambig.jape-Datei“ zu Nutze gemacht.

Die PR *Unambig* macht im Grunde genommen genau das, was der Name vermuten lässt: Sie löst die Ambiguitäten auf. Dies funktioniert in drei Schritten:

1) Zunächst wird festgestellt, ob es mehr als eine Annotation für ein Wort gibt. (Apolda vergibt die Annotation *Mention*)
2) Wenn ein Schlagwort mehrfach annotiert wurde, dann wird in der CrissCross-Ontologie nach dem Hauptschlagwort (*PrefLabel)* als Textuelle Repräsentation gesucht.
3) Im letzten Schritt wird eine neue Annotation *Schlagwort* hinzugefügt. Und zwar für jede Annotation, die entweder nicht ambig ist, oder für die ein Hauptschlagwort (*PrefLabel*) gefunden wurde. So bekommen „Stress“ und „Quality Management“ jeweils nur eine Annotation *Schlagwort*.

8.3.7 Wordclass (JAPE Transducer)

Die PR *Wordclass* wurde dankenswerterweise ebenfalls von Herrn Prof. Dr. Wartena zur Verfügung gestellt. Diese PR sucht zu jedem Schlagwort die dazugehörigen DDC-Notationen heraus. Wordclass hat als Input die Annotation *Schlagwort* und als Output die Annotation *DDC*. *DDC* hat als Eigenschaft keine URL mehr, sondern die gewünschte DDC-Notation. Um dahin zu kommen, mussten wieder Informationen aus der CrissCross-Ontologie neu verarbeitet werden. Die DDC-Notationen sind in der CrissCross-Ontologie im *CoordinatedConcept* verpackt und müssen von der PR Wordclass nur noch „geholt“ werden.

Ansonsten stellt sich das Ergebnis der Korpus-Pipeline wie in Abbildung 12 dar. In der rechten Spalte ist zu erkennen, dass die Annotation DDC angehakt ist. In der mittleren Spalte ist der Text, bzw. die Metadaten, mit den annotierten Schlagwörtern zu sehen. Darunter stehen die Annotationen in einer Liste mit Start- und End-Tag und den Eigenschaftswerten (engl. feature). Durch die Determiniertheitsgrade haben die Schlagwörter oft mehrere Annotationen, bzw. DDC-Notationen. Die weitere Auswertung der Ergebnisse erfolgte manuell und zwar durch auszählen der Eigenschaftswerte, bzw. DDC-Notationen mit Hilfe von Excel.

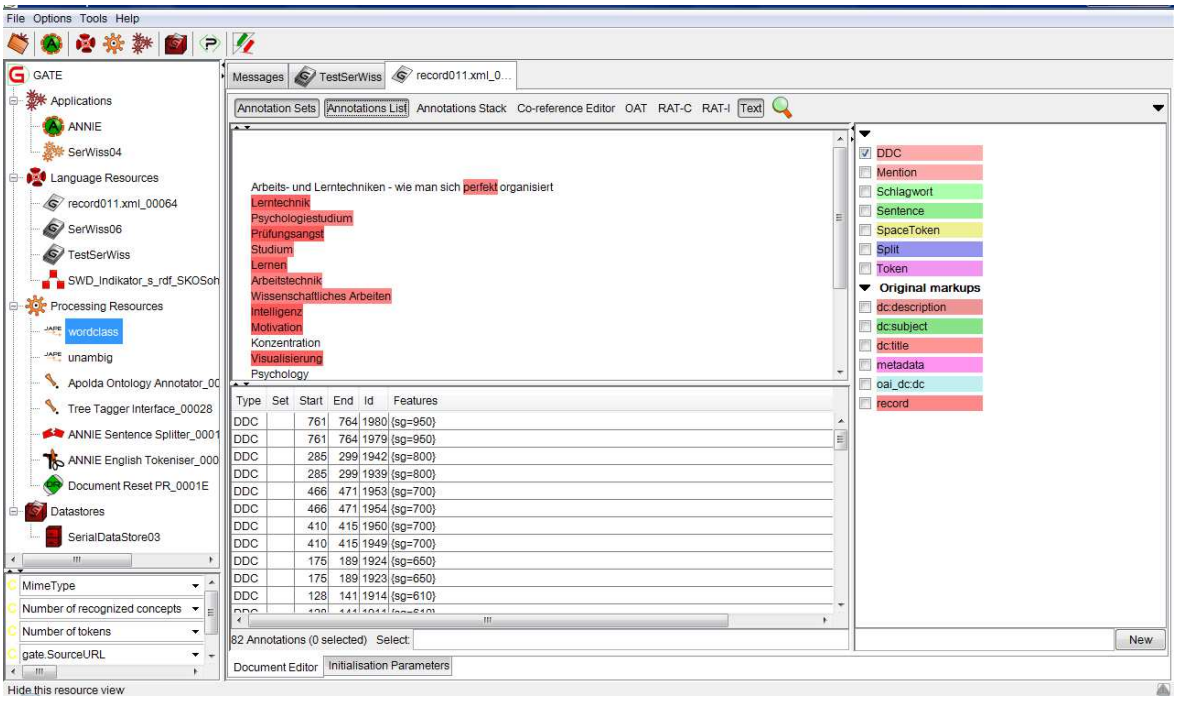


Abbildung 12: Screenshot von GATE mit einem annotierten Text

8.4 Modifikationen an der CrissCross-Ontologie

In diesem Abschnitt erfolgt eine zusammenfassende Beschreibung aller nötigen Modifikationen an der CrissCross-Ontologie.

Bei der unbearbeiteten RDF/XML-Ontologie war jedes einzelne SKOS-Konzept von einem *RDF-Root-Element* mit einer XML-Namensraum-Deklaration umschlossen. Damit die GATE-Programme mit der Ontologie arbeiten können dürfen sämtliche Konzepte nur von einem RDF-Root-Element umschlossen sein. Deswegen wurden alle *rdf:RDF*-Elemente, ausgenommen das erste (<Start-Tag>) und letzte (</End-Tag>) Element, aus der RDF/XML-Ontologie gelöscht.

Bei einem ersten Versuch mit einem Test-Korpus stellte sich heraus, dass Wörter wie „mit“, „ein“, „an“, „und“ annotiert wurden. Für diese Wörter wurde die DDC-Notation 435 „Grammatik des Deutschen vergeben“. Diese sogenannten *Stoppwörter* kommen häufig im Text vor, sind aber vollkommen irrelevant und verfälschen das Ergebnis sogar. Aus diesem Grund wurden alle Konzepte aus der Ontologie aussortiert, welche in die DDC-Klasse 435 Lexikologie gehörten. Davon ausgenommen wurden nur die Schlagwörter „rational“, „irrational“ und „Gloria“, weil sie noch weiteren DDC-Klassen zugeteilt sind. Auch das Hauptschlagwort „Grammis“ wurde nicht aussortiert, weil es sich hierbei nicht um ein Stoppwort, sondern um das „Grammatische Informationssystem des IDS“ handelt. Es gab

noch zwei weitere Schlagwörter, die aus der Ontologie entfernt werden mussten, weil sie immer falsch annotiert wurden: Die Schlagwörter IN für „intelligentes Netz“ und IM für „Informeller Mitarbeiter“. Darüber hinaus gab es drei Konzepte mit dem „**?**“ als Hauptschlagwort. Auch sie wurden entfernt um irreführende Annotationen zu vermeiden.

8.5 Korpus

Genau wie bei dem DFG-Projekt *Automatische Anreicherung von OAI-Metadaten* wurden auch für dieses Experiment OAI-DC-Metadaten verwendet. Allerdings wurden sie in diesem Fall nicht von BASE, sondern vom Server für Wissenschaftliche Schriften der Hochschule Hannover (SerWisS) geharvestet.

Nach dem Harvesting wurde jedes Record in eine eigene XML-Datei überführt. Anschließend wurden alle Dublin-Core-Metadaten-Felder bis auf *title*, *subject* und *description* entfernt. Die Felder *title* für Titel und *subject* für Schlagwort sind eindeutig. Aber das Feld *description* kann von Abstracts, Zusammenfassungen, Inhaltsverzeichnissen bis Textausschnitten alles enthalten. Sogar eine graphische Repräsentation des Inhalts wäre möglich.[21] Die Description-Felder im SerWisS-Korpus enthielten nur Abstracts und Inhaltsverzeichnisse. Alle Records, in denen das Description-Feld leer war, wurden aus dem Korpus aussortiert. Englischsprachige Titel und Abstracts wurden ebenfalls aus den Records entfernt, um falsche Annotationen, durch Wörter die zwar gleich geschrieben werden, aber im englischen und deutschen verschiedene Bedeutungen haben, zu vermeiden. Die Schlagwörter wurden alle beibehalten, da sie das Ergebnis eines intellektuellen Erschließungsvorgangs sind. Am Ende bestand das Korpus aus 35 Records.

Eine quantitative Analyse des Korpus kam zu dem folgenden Ergebnis: Die Zahl der Schlagwörter je Record variiert sehr stark. Sie liegt zwischen 2 und 12. Im Durchschnitt hat jedes Record 6,8 Schlagwörter. Auch die Länge des Description-Feldes ist sehr unterschiedlich. Es beherbergt zwischen 7 und 387 Wörtern. Allerdings sind das nur die Extreme. Bei etwas über 50 Prozent des Korpus liegt die Länge des Description-Feldes zwischen 50 und 200 Wörtern. Im Durchschnitt hat das Description-Feld eine Länge von 118 Wörtern.

Zusätzlich, zu dem eben beschriebenen Korpus, wurden noch zwei weitere Korpora erstellt. Inhaltlich waren es immer die gleichen Records. Nur die Zusammensetzung der OAI-DC-Metadaten variierte bei den drei Korpora:

[21] Vgl. http://dublincore.org/documents/dces/ (letzter Abruf am: 05.05.2012)

1) SerWisS-Korpus 1: Records mit *title*-, *subject*- und *description*-Feld.
2) SerWisS-Korpus 2: Records mit *title*- und *subject*-Feld
3) SerWisS-Korpus 3: Records mit *title*- und *description*-Feld

8.6 Ergebnisse

Der Versuch zum automatischen Klassifizieren von OAI-Metadaten des Hochschulschriftenservers SerWisS brachte gute Ergebnisse zustande. Innerhalb der Gruppe der ersten drei vorgeschlagenen DDC-Notationen, war zu 80 Prozent die richtige DDC-Notation dabei. Als Vergleichsmaß zur Beurteilung der „Richtigkeit“ einer DDC-Notation wurden die Ergebnisse des intellektuellen Klassifizierens herangezogen. Da das Ergebnis durch Auszählung sämtlicher vorgeschlagener DDC-Notationen zustande gekommen ist, hatten oft mehrere Notationen dieselbe Punktzahl. Für eine differenziertere Betrachtung wurden fünf Kategorien gebildet:

Kategorie A	1.Platz: vorgeschlagene DDC-Notation mit der höchsten Punktzahl ist „richtig“
Kategorie B	1. Platz (geteilt): Es gibt zwei Vorschläge mit der gleichen Punktzahl auf dem ersten Platz.
Kategorie C	2.Platz für die „richtige“ DDC-Notation
Kategorie D	3.Platz für die „richtige“ DDC-Notation
Kategorie E	DDC-Notation kommt nicht in den „Top-Drei“ vor

Es kam auf den Platzierungen zwei und drei ebenfalls vor, dass mehrere Notationen die gleiche Punktzahl erreicht haben. Dies erschien allerdings nicht so relevant wie auf dem ersten Platz und wurde, zugunsten einer etwas übersichtlicheren Darstellung, in dieser Einteilung nicht berücksichtigt. Die Art der Präsentation macht auch deutlich, dass das Ergebnis ganz gut für eine rangordnende Klassifizierung geeignet ist. Für eine „harte“ Klassifizierung ist es jedoch völlig ungeeignet. Die Ursache für die vielen Notationen mit der gleichen Punktzahl liegt in den Determiniertheitsgraden. Meistens ist ein Schlagwort mit zwei bis drei weiteren Notationen verknüpft. Eine Idee war daher, einfach den Determiniertheitsgrad 1 aus der CrissCross-Ontologie heraus zu löschen und die Korpora erneut zu annotieren. Dies brachte eine leichte Verbesserung, führte aber nicht, wie gehofft, zu einem eindeutigeren Ergebnis mit weniger Mehrfachplatzierungen. Bei diesem zweiten Versuch wurden nur SerWisS-Korpus 1 und SerWisS-Korpus 2 ausgewertet. Es hatte sich bereits im ersten Versuch gezeigt, dass die Methode die schlechtesten Ergebnisse hervorbringt, wenn nur das Dublin-Core-Feld *title*- und *description* als Datengrundlage

genutzt werden. Dies ist auch nicht weiter verwunderlich, da hier DDC-Notationen mithilfe der Schlagwortnormdatei der DNB vergeben werden.

Tabelle 6: Ergebnisse des Experiments

Mit Determiniertheitsgrad 1 title, subject und Description			
Ergebnis	Records	Prozent	kumuliert
A	8	22,86%	22,86%
B	5	14,29%	37,14%
C	13	37,14%	74,29%
D	3	8,57%	82,86%
E	6	17,14%	100,00%
Summe	35	100,00%	

Mit Determiniertheitsgrad 1 Nur title und subject			
Ergebnis	Records	Prozent	kumuliert
A	12	34,29%	34,29%
B	8	22,86%	57,14%
C	8	22,86%	80,00%
D	1	2,86%	82,86%
E	6	17,14%	100,00%
Summe	35	100,00%	

Mit Determiniertheitsgrad 1 Nur title und description			
Ergebnis	Records	Prozent	kumuliert
A	5	14,29%	14,29%
B	5	14,29%	28,57%
C	12	34,29%	62,86%
D	2	5,71%	68,57%
E	11	31,43%	100,00%
Summe	35	100,00%	

Ohne Determiniertheitsgrad 1 titel, subject und description			
Ergebnis	Records	Prozent	kumuliert
A	9	25,71%	25,71%
B	4	11,43%	37,14%
C	12	34,29%	71,43%
D	5	14,29%	85,71%
E	5	14,29%	100,00%
Summe	35	100,00%	

Ohne Determiniertheitsgrad 1 Nur title und subject			
Ergebnis	Records	Prozent	kumuliert
A	11	31,43%	31,43%
B	7	20,00%	51,43%
C	12	34,29%	85,71%
D	2	5,71%	91,43%
E	3	8,57%	100,00%
Summe	35	100,00%	

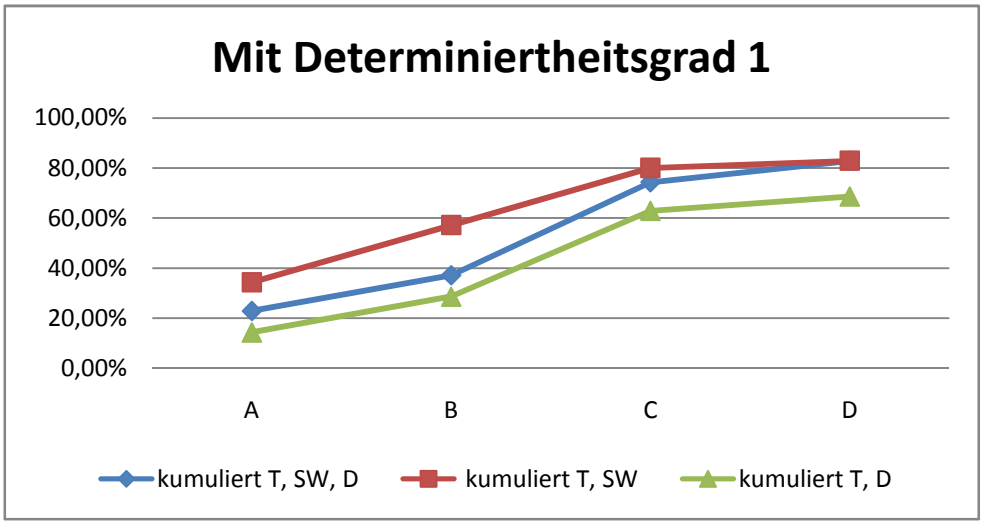


Abbildung 13: Ergebnisse des Experiments mit Det1

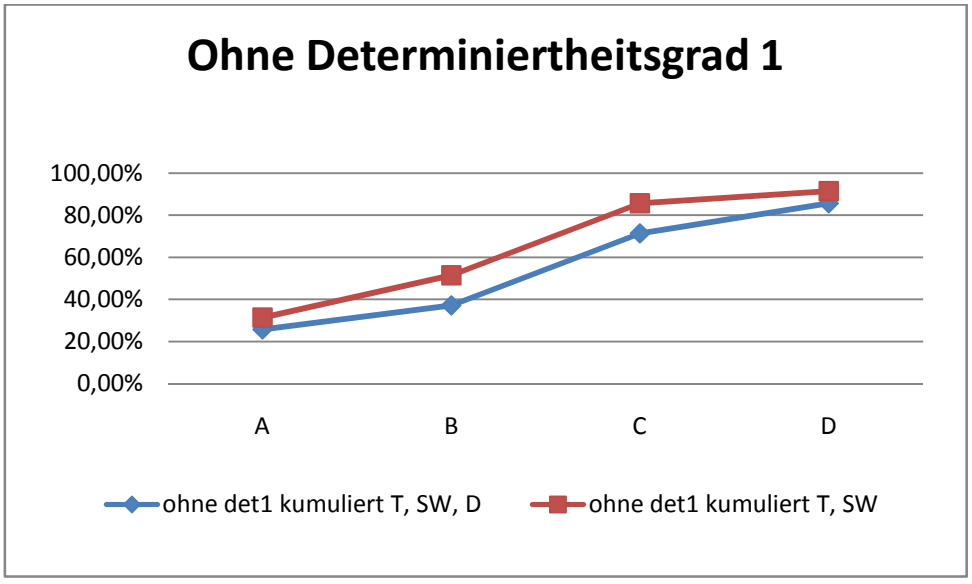


Abbildung 14: Ergebnisse des Experiments ohne Det1

Bei der Beschreibung des Korpus wurde bereits erwähnt, dass die Menge der Schlagwörter und die Menge der Wörter im Description-Feld von Record zu Record sehr unterschiedlich sind. Daher wurde untersucht, ob es einen Zusammenhang zwischen der quantitativen Zusammensetzung des Korpus und der „Richtigkeit“ einer vorgeschlagenen DDC-Notation gibt. Bei der Menge der Wörter im Description-Feld konnte kein Zusammenhang festgestellt werden. Es gibt sowohl in der Kategorie A Records mit wenigen Wörtern im Description-Feld, als auch in den Kategorien B, C, D und E (vgl. dazu Abb. 15). Dies ist nur eine weitere Bestätigung für die Vermutung, dass die Wörter im Description-Feld teilweise zu unspezifisch für die Herangehensweise mit Schlagwörtern sind. Sie können sogar zu einer Verschlechterung des Ergebnisses beitragen, wenn in einem Abstract oder

Inhaltsverzeichnis viele Wörter stehen, die zwar Schlagwörter sind, aber mit DDC-Notationen verknüpft sind, die in eine völlig andere Fachrichtung weisen.

Wie bereits vermutet, besteht ein Zusammenhang zwischen der Menge der Schlagwörter und der Richtigkeit der DDC-Notation. Es gibt zwar in jeder Kategorie (A-E) Records mit nur vier oder fünf Schlagwörtern, aber im Durchschnitt deutet sich ein leichter Trend an. Je mehr Schlagwörter in einem Text vorhanden sind, umso größer ist die Wahrscheinlichkeit, dass die vorgeschlagene DDC-Notation „richtig“ ist (vgl. dazu Abb. 16).

Trotzdem kann im Fall der Schlagwörter wirklich nur von einer leichten Andeutung gesprochen werden. Es muss konstatiert werden, dass nur ein sehr geringer Zusammenhang zwischen der quantitativen Zusammensetzung der Records und der Klassifizierqualität besteht.

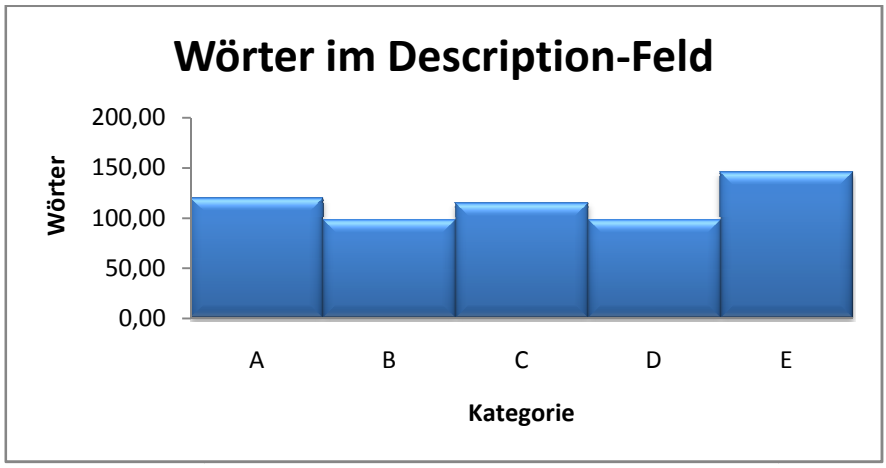


Abbildung 15: Durchschnittliche Zahl der Wörter im Description-Feld je Kategorie

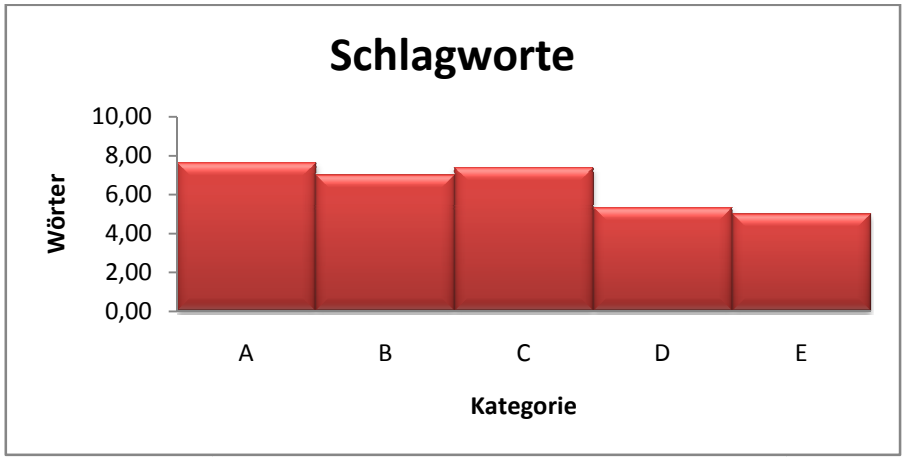


Abbildung 16: Durchschnittliche Zahl der Schlagwörter je Kategorie

Die Einteilung der Ergebnisse in die Kategorien A, B, C, D und E ist eine eigene, die sich aus der ersten Betrachtung ergeben hat. Normalerweise wird die Qualität eines Programms zur automatischen Sprachverarbeitung mit Precision (Genauigkeit) und Recall (Trefferquote) gemessen. Diese Darstellungsweise ist in diesem Fall wenig aussagekräftig, da insgesamt nur 35 Records analysiert wurden. Die DDC-Notationen 080 und 230 kamen jeweils nur einmal in dem gesamten Korpus vor. Da sie von dem Programm nicht richtig erkannt, bzw. falsch annotiert wurden, ist ihr Wert für Precision und Recall gleich Null. Am häufigsten kamen die Notationen DDC 020 (10 Mal) und DDC 150 (8 Mal) vor. Daher sind diese Werte noch am verlässlichsten. Von 8 Records aus dem Bereich Psychologie (DDC 150) wurden 5 richtig erkannt. Das ist schon mal ein ziemlich guter Recall. Die Precision ist hier noch besser. Die Notation DDC 150 wurde nur einmal falsch vergeben. Die anderen DDC-Notationen kamen vier (DDC 000/DDC 004), fünf (DDC 300) und sechs (DDC 330) Mal vor. Die Notationen aus dem Bereich Sozialwissenschaften (DDC 300) und Wirtschaft (DDC 330) wurden auch oft an anderen Stellen vergeben, daher die schlechte Precision.

Tabelle 7: Ergebnisse des Experiments gemessen in Precision und Recall

Für title, subject, description mit Det1			
DDC	**Precision**	**Recall**	**F1-Measure**
000/004	0,500	0,500	0,500
020	1,000	0,300	0,460
080	0,000	0,000	0,000
150	0,830	0,630	0,710
230	0,000	0,000	0,000
300	0,250	0,200	0,220
330	0,290	0,330	0,310

Für title und subject mit Det1			
DDC	**Precision**	**Recall**	**F1-Measure**
000/004	0,500	0,500	0,500
020	1,000	0,600	0,750
080	0,000	0,000	0,000
150	1,000	0,750	0,860
230	0,000	0,000	0,000
300	0,250	0,400	0,310
330	0,500	0,500	0,500

Verglichen mit den Ergebnissen des DFG-Projektes „Automatische Anreicherung von OAI-Metadaten“ für die zweite und dritte Hierarchieebene sind diese Werte gar nicht mal so schlecht (vgl. dazu Tabelle 8).

Tabelle 8: Ergebnisse des DFG-Projekts für die zweite und dritte Hierarchieebene

Ergebnisse für die zweite Hierarchieebene

DDC	Precision	Recall	F1-Measure
300	0.795	0.619	0.696
310	0.633	0.432	0.513
320	0.904	0.761	0.826
330	0.936	0.893	0.915
340	0.930	0.847	0.886
350	0.947	0.882	0.913
360	0.845	0.722	0.779
370	0.885	0.816	0.849
380	0.881	0.642	0.743
390	0.860	0.711	0.779
Overall	**0.862**	**0.732**	**0.790**

Ergebnisse für die dritte Hierarchieebene

DDC	Precision	Recall	F1-Measure
610	0.910	0.836	0.871
611	0.833	0.217	0.345
612	0.167	0.026	0.045
613	0.500	0.139	0.218
614	-	-	-
615	0.533	0.127	0.205
616	0.817	0.606	0.696
617	0.857	0.481	0.617
618	-	-	-
619	-	-	-
Overall	**0.659**	**0.347**	**0.428**

Dieser Vergleich ist zugegebenermaßen nicht besonders aussagekräftig, was die tatsächliche Leistungsfähigkeit der beiden Klassifikatoren betrifft. Glücklicherweise haben die Beteiligten des DFG-Projektes ihren Klassifikator ins Netz gestellt.[22] Dadurch wird ein direkter Vergleich möglich. Alle Notationen, die richtig erkannt wurden, sind in der Tabelle 9 grün markiert. Die Notationen der Kategorie B wurden gelb markiert. Sie zählen aber als eine richtige Notation.

Tabelle 9: Vergleich des Experiments mit dem DFG-Projekt

		DFG-Projekt	Experiment	DFG-Projekt	Experiment
		T, SW, D	T, SW, D	T, SW	T, SW
Record	**Intellektuelle DDC-Klassifikation**	**Level 3**		**Level 3**	
record002	DDC 020	DDC 020	DDC 020 / DDC 370	DDC 020	DDC 370
record005	DDC 020	DDC 020	DDC 100	DDC 020	DDC 020 / DDC 300
record006	DDC 020	DDC 004	DDC 650	DDC 020	DDC 650
record008	DDC 020	DDC 020	DDC 000 / DDC 300	DDC 020	DDC 000 / DDC 020
record009	DDC 000	DDC 000	DDC 300	DDC 020	DDC 300 / DDC 370
record010	DDC 150	DDC 150	DDC 330	DDC 150	DDC 330
record011	DDC 150	DDC 150	DDC 150	DDC 150	DDC 150
record013	DDC 150	DDC 150	DDC 100 / DDC 150	DDC 150	DDC 150
record014	DDC 150	DDC 150	DDC 150	DDC 100	DDC 150 / DDC 360
record015	DDC 150	DDC 150	DDC 100	DDC 100	DDC 150

[22] Vgl. http://act-dl.base-search.net/ (letzer Abruf am 04.05.2012)

record016	DDC 150	DDC 150	DDC 650	DDC 150	DDC 650
record018	DDC 150	DDC 150	DDC 150	DDC 150	DDC 150
record020	DDC 330	DDC 300	DDC 100	DDC 300	DDC 800
record021	DDC 020	DDC 000	DDC 000	DDC 004	DDC 020 / DDC 300
record022	DDC 020	DDC 020	DDC 330	DDC 020	DDC 000 / DDC 020 / DDC 300
record025	DDC 020	DDC 020	DDC 100	DDC 020	DDC 300
record026	DDC 330	DDC 300	DDC 330	DDC 100	DDC 330
record029	DDC 004	DDC 150	DDC 000	DDC 150	DDC 000
record030	DDC 004	DDC 004	DDC 000 / DDC 340	DDC 004	DDC 000
record031	DDC 330	DDC 330	DDC 650	DDC 330	DDC 330 / DDC 650
record032	DDC 330	DDC 330	DDC 330 / DDC 650	DDC 330	DDC 330 / DDC 650
record048	DDC 080	DDC 020	DDC 230 / DDC 355	DDC 020	DDC 355
record049	DDC 300	DDC 360	DDC 360	DDC 360	DDC 360
record050	DDC 300	DDC 300	DDC 300	DDC 300	DDC 300
record054	DDC 230	DDC 230	DDC 330	DDC 830	DDC 330
record071	DDC 330	DDC 330	DDC 650	DDC 100	DDC 650
record076	DDC 004	DDC 150	DDC 620	DDC 370	DDC 620
record102	DDC 300	DDC 360	DDC 360	DDC 430	DDC 610 / DDC 360
record103	DDC 300	DDC 330	DDC 330	DDC 791	DDC 330
record104	DDC 330	DDC 330	DDC 650	DDC 150	DDC 650
record105	DDC 150	DDC 330	DDC 150	DDC 330	DDC 150
record130	DDC 020	DDC 700	DDC 020	DDC 700	DDC 020
record132	DDC 300	DDC 340	DDC 150	DDC 340	DDC 300
record133	DDC 020	DDC 000	DDC 020 / DDC 300	DDC 000	DDC 020
record136	DDC 020	DDC 020	DDC 330	DDC 020	DDC 020 / DDC 300 / DDC 330

Das Gesamtergebnis dieses direkten Vergleichs stellt sich wie folgt da:

Tabelle 10: Gesamtergebnis des Vergleichs zwischen DFG-Projekt und Experiment

	DFG-Projekt T, SW, D	Experiment T, SW, D	DFG-Projekt T, SW	Experiment T, SW
Richtig	21 (60%)	13 (37,14%)	16 (45,71%)	20 (57,14%)
Falsch	14 (40%)	22 (62,86%)	19 (54,29%)	15 (42,86%)

Mit 21 richtig erkannten DDC-Notationen erzielt der Klassifikator des DFG-Projekts deutlich bessere Ergebnisse als die eigene Korpus-Pipeline. Des Weiteren sind die Ergebnisse des Klassifikators eindeutig, während beim eigenen Experiment oft Mehrfachklassifikationen vergeben werden. Bemerkenswert ist das Ergebnis jedoch, wenn nur Titel und Schlagwörter für die Klassifikation zur Verfügung stehen. Dann kann dieses

Experiment sogar mit dem DFG-Projekt mithalten. Allerdings gibt es auch hier wieder zahlreiche Mehrfachklassifikationen.

Eine weitere Auffälligkeit sind die rot markierten Notationen in der Ergebnistabelle. Viermal wurden sowohl vom Klassifikator des DFG-Projektes als auch von der Korpus-Pipeline die gleichen „falschen“ Notationen vergeben. Bei Record021 wurde im automatischen Verfahren zwei Mal die Notation DDC 000 vergeben. Die vom intellektuellen Klassifizierer vergebene Notation DDC 020 (Bibliotheks- und Informationswissenschaften) ist hier allerdings besser, da es sich um einen Text handelt, der sich mit dem Berufsbild eines Spezialbibliothekars befasst. Record049 wurde bei der ersten Analyse in die Kategorie C einsortiert, das heißt die „richtige“ intellektuelle Notation DDC 300 (Sozialwissenschaften) war auf dem zweiten Platz. In allen vier Fällen schlagen die Text-Mining-Programme die Notation DDC 360 (Soziale Probleme und Sozialdienste; Verbände) vor. In diesem Fall ist diese Notation viel spezifischer und passender, denn es handelt sich um einen Text mit dem Titel: „Soziale Gruppenarbeit als Alternative zur geschlossenen Unterbringung strafmündiger Kinder“. Record102 und Record103 wurden in der ersten Einteilung in die Kategorie E einsortiert. Beide Male wurde die intellektuelle Klassifikation DDC 300 vergeben und beide Male kam sie nicht in den „Top-Drei-Klassifikationsvorschlägen“ vor. Record102 enthält Metadaten von einer Studie, die sich mit der Leistungsentwicklung der Pflegeversicherung in den Jahren 1997 bis 2001 beschäftigt. Die von beiden Text-Mining-Programmen vorgeschlagene Notation DDC 360 (Soziale Probleme und Sozialdienste; Verbände) passt hier also viel besser. In Record103 geht es um Stellenabbau beim Pflegepersonal im Krankenhaus. Auch in diesem Fall kann die automatisch vergebene Notation DDC 330 (Wirtschaft) als richtig angesehen werden. In dieser Klasse gibt es einen Unterpunkt 331 Arbeitsökonomie mit einem weiteren Unterpunkt 331.1 Arbeitskräfte und Arbeitsmarkt.

Aufgrund solcher Beispiele wurden in dieser Ergebnisbeschreibung die Wörter „richtig“ und „falsch“ im Zusammenhang mit der Beurteilung von Verfahren zur Automatischen Sprachverarbeitung in Anführungsstriche gesetzt. Intellektuelle Klassifikationen als Vergleichsmaßstab zu nehmen ist sicherlich die beste Möglichkeit die es gibt, dennoch sollte dabei nicht vergessen werden, dass Indexieren und Klassifizieren keine exakten Wissenschaften sind. Es gibt oft mehrere Möglichkeiten einem Dokument eine Klasse zuzuordnen und auch intellektuelle Klassifizierer sind sich dabei keineswegs immer einig. Das belegt auch eine Indexierkonsistenz von teilweise unter 50 Prozent (vgl. Nohr 2003, S.

28). Der Vorteil von automatischen Verfahren gegenüber Menschen ist, dass die automatischen Verfahren immer zu demselben Schluss kommen. Die Intra-Indexierkonsistenz liegt bei 100 Prozent. Der Nachteil ist, dass die automatischen Verfahren manchmal völlig daneben liegen, d.h. eine Notation vergeben, die überhaupt nicht im Zusammenhang mit dem Dokument steht. *Semiautomatische Verfahren* können die Stärken von beiden Herangehensweisen kombinieren und zu einer schnelleren Klassifikationsentscheidung führen. Am Besten eignen sich hierfür *Rangordnende Verfahren*. Ein automatisches Sprachverarbeitungsprogramm schlägt für ein Dokument eine Reihe von Klassen vor. Die Begrenzung, wie viele Klassen vorgeschlagen werden sollen, ist immer etwas unterschiedlich und hängt vom Programm und intellektuellen Klassifizierer ab. Bei dem Experiment in dieser Bachelorarbeit war innerhalb der ersten drei vorgeschlagenen Notationen zu 80 Prozent die „Richtige“ Notation dabei. Die ersten fünf Klassen würden also ausreichende Wahlmöglichkeiten bieten. Daher würde sich dieses Ergebnis ganz gut für ein Semiautomatisches Verfahren eignen.

Zusammenfassung und Ausblick

Die Bachelorarbeit hat gezeigt, dass es durch die CrissCross-Ontologie schon mit einfachen Mitteln möglich ist, ganz gute Klassifikationsvorschläge zu erreichen. Darüber hinaus machen die Ergebnisse deutlich, dass die Schlagwortnormdatei als textuelle Repräsentation für ein begriffsorientiertes Klassifikationsverfahren geradezu dazu prädestiniert ist, auf Metadaten angewendet zu werden. Die besten Ergebnisse werden erzielt, wenn nur die OAI-DC-Metadaten Titel und Schlagwort verwendet werden. Das ist nachteilig, da die Schlagwörter bereits das Ergebnis eines intellektuellen Erschließungsvorgangs sind. Außerdem ist das Ergebnis sehr stark davon abhängig, wie sorgfältig die Schlagwortvergabe (nach RSWK) erfolgt ist. Die Beweggründe für automatische Klassifikationsverfahren sind Kosten- und Zeitersparnis bei der Erschließung von elektronischen Dokumenten. Um hier einen wirklichen Vorteil zu erreichen, müsste auch der Prozess der Schlagwortvergabe automatisiert werden.

Für die Schlagwortnormdatei als textuelle Repräsentation spricht, dass sie ein umfangreiches Vokabular, mit eher engen Begriffen bereithält. Sie ist besonders für die Erschließung von allgemeinen Daten geeignet. Daher kann die SWD sich auch gut auf die Metadaten des Hochschulschriftenservers der Hochschule Hannover angewendet werden. Gegen die Schlagwortnormdatei sprechen vor allem die schlecht ausgeprägten Begriffsrelationen. Laut einer Studie haben 34% der Schlagwörter, weder hierarchische-

noch assoziative Beziehungen (vgl. Jacobs, Mengel, Müller 2010, S. 237). Aus diesem Grund ist es auch schwierig die SWD auf ein hierarchisches Klassifikationssystem wie die DDC abzubilden. Es gibt viele Schlagwörter, die eine teilweise oder manchmal nur geringe Übereinstimmung mit einer oder mehreren DDC-Notationen aufweisen. Nur selten stimmt ein Schlagwort vollständig mit einer DDC-Notation überein. Die Beteiligten des Kölner CrissCross-Projektes haben diese Schwierigkeiten bei der Umsetzung berücksichtigt. Nach dem Prinzip des „one-to-many-mapping“ wurde ein Schlagwort meistens in mehrere DDC-Klassen „gemappt“. Um eine differenziertere Darstellung der Übereinstimmung eines Schlagwortes mit einer DDC-Klasse zu ermöglichen, wurden den „Mappings“ Determiniertheitsgrade hinzugefügt. Durch die Ergebnisse des CrissCross-Projekts verbessert sich das Relationsgefüge in der SWD. Dies wird auch als *Begriffsexploration* bezeichnet. Zuvor unverbunden nebeneinander stehende Schlagwörter werden durch dieselbe DDC-Klasse in eine Äquivalenzrelation überführt und es entstehen Hierarchierelationen durch unterschiedlich „tiefe“ Mappings.

Diese sogenannte Begriffsexploration spielt auch für das automatische Klassifizieren mit der CrissCross-Ontologie eine Rolle, allerdings in einer etwas anderen Weise als beim Retrieval mit CrissCross.

```
<dnb:hasCoordinatedConcept-of>
  <dnb:CoordinatedConcept>
    <dnb:coordination-of rdf:resource="http://d-nb.info/ddc-sg/320"/>
    <dnb:coordination-of rdf:resource="http://d-nb.info/gnd/4115589-0"/>
    <dnb:det2 rdf:resource="http://d-nb.info/ddc/class/320.9"/>
  </dnb:CoordinatedConcept>
</dnb:hasCoordinatedConcept-of>
```

Abbildung 17: Beispiel für ein Coordinated Concept

In Abbildung 17 ist noch einmal ein Coordinated Concept zu sehen. Die DDC-Klassen, die den Determiniertheitsgraden zugeordnet sind (ddc/class/320.9), sind zu „tief“ klassifiziert, um sie auszuzählen. Es werden immer die Klassen auf den oberen Hierarchieebenen ausgezählt (ddc-sg/320). Dadurch können Mappings, die bei der „tiefen“ Klassifizierung eine unterschiedliche Notation haben auf der höheren Ebene dieselbe Notation bekommen. Diese Tatsache ist manchmal das „Zünglein an der Waage“, denn das kann zur Folge haben, dass bestimmte Schlagwörter ein höheres Gewicht erhalten als andere Schlagwörter. Das Schlagwort „E-Learning“ hat z.B. vier Mappings erhalten und viermal wurde daraus die Notation DDC 370. Das Schlagwort „Bibliothek“ wurde nur einmal gemappt. Dadurch wird es im Text auch nur einmal mit der Notation DDC 020 annotiert.

In dem betreffenden Record002 hat das zur Folge, dass sowohl die Notation DDC 020, als auch die Notation DDC 370 auf dem „ersten Platz“ landen, obwohl es viel mehr Schlagwörter wie Informationskompetenz, Informationsvermittlung, Bibliothek und information literacy mit der Notation DDC 020 im Text gibt. In diesem Fall wäre das Ergebnis unter Einbeziehung der Determiniertheitsgrade sogar immer noch das Gleiche, weil E-Learning alle vier Mal mit dem Determiniertheitsgrad 3 gemappt wurde.

Bei Record030 würde die Einbeziehung der Determiniertheitsgrade jedoch einen Unterschied machen. Die Klassifikation, die als richtig erachtet wurde ist, DDC 000 (Informatik, Informationswissenschaft, allgemeine Werke). Bei dem Versuch mit Titel, Schlagwort und Description-Feld erhielten die Notationen DDC 000 und DDC 340 (Recht) gleich viele Punkte. Bei dem Text handelt es sich um einen Vergleich von Data-Mining-Verfahren, daher ist die Klasse „Recht“ schlichtweg falsch. Zu diesem Notationsvorschlag ist es gekommen, weil in den OAI-DC-Metadaten dreimal das Schlagwort „Verfahren“ mit jeweils zwei Mappings in die DDC-Klasse 340 vorkommt. Die anderen wichtigen Schlagwörter („Data Mining“, „Computer“, „Informatik“, „Algorithmen“) wurden hier wieder nur jeweils einmal in die DDC-Klasse 000 gemappt. Allerdings sind beide Mappings bei „Verfahren“ nur mit dem Determiniertheitsgrad 1 erfolgt. „Data Mining“ wurde mit dem Determiniertheitsgrad 4, „Computer“ und „Informatik“ mit Determiniertheitsgraden 3 und „Algorithmen“ mit Determiniertheitsgrad 2 gemappt. Würden jetzt für Determiniertheitsgrad 4 = 4 Punkte, für Determiniertheitsgrad 3 = 3 Punkte, usw. verteilt, dann stände es am Ende 19:6 für die DDC-Klasse 000. Vorher stand es unentschieden.

In einem weiteren Experiment mit der CrissCross-Ontologie könnte untersucht werden, welchen Einfluss die Einbeziehung der Determiniertheitsgrade auf das Ergebnis hat. Möglicherweise würde dadurch das Ergebnis etwas deutlicher zugunsten einer Notation ausfallen.

Literaturverzeichnis

Bertram, Jutta (2005): Einführung in die inhaltliche Erschließung: Grundlagen, Methoden, Instrumente. Würzburg: Ergon (Content and communication, 2)

Cunningham, Hamish; Maynard, Diana; Bontcheva, Kalina; Tablan, Valentin (2002): GATE: an Architecture for Development of Robust HLT Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia USA, 07.-12.07.2002. S. 168-175. Online verfügbar unter http://gate.ac.uk/sale/acl02/acl-main.pdf, zuletzt geprüft am 24.04.2012

Cunningham, Hamish; et al. (2012): Developing Language Processing Components with GATE Version 7 (a User Guide): for GATE version 7.1-snapshot (development builds): (built April 24, 2012). Herausgegeben von der University of Sheffield, Department of Computer Science. Online verfügbar unter http://gate.ac.uk/sale/tao/tao.pdf, zuletzt geprüft am 30.04.2012

Deutsche Nationalbibliothek (2012): Linked Data Service der Deutschen Nationalbibliothek: Version 4.1. Stand: 29. Februar 2012. Online verfügbar unter http://files.d-nb.de/pdf/linked_data.pdf, zuletzt geprüft am 21.04.2012

Dewey, Melvil; Mitchell, Joan S. (Hg.) (2005a): Dewey Dezimalklassifikation und Register: DDC 22, Bd. 1. Dt. Ausg. München: Saur

Dewey, Melvil; Mitchell, Joan S. (Hg.) (2005b): Dewey Dezimalklassifikation und Register: DDC 22, Bd. 2. Dt. Ausg. München: Saur

Groß, Thomas; Faden, Manfred (2010): Automatische Indexierung elektronischer Dokumente an der Deutschen Zentralbibliothek für Wirtschaftswissenschaften. In: Bibliotheksdienst, Jg. 44, H. 12, S. 1120-1135. Online verfügbar unter http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte2010/Erschliessung011210.pdf, zuletzt geprüft am 12.05.2012

Heyer, Gerhard; Quasthoff, Uwe; Wittig, Thomas (2008): Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. 1. korrigierter Nachdr. Herdecke: W3L-Verl.

Hubrich, Jessica (2008): Criss Cross: SWD-DDC-Mapping. In: Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare, Jg. 61, H. 3, S. 50-58. Online verfügbar unter http://www.univie.ac.at/voeb/fileadmin/Dateien/Publikationen/VOB-Mitteilungen/vm6120083.pdf, zuletzt geprüft am 02.02.2012

Hubrich, Jessica (2009a): Vom Stringmatching zur Begriffsexploration: das Potential integrierter begrifflicher Interoperabilität. In: Tagungsband der 12. Tagung der Deutschen ISKO 2009 in Bonn. Würzburg: Ergon-Verl. Online verfügbar unter http://linux2.fbi.fh-koeln.de/crisscross/publikationen/Hubrich_IntegrierteBegrifflicheInteroperabilitaet.pdf, zuletzt geprüft am 16.04.2012

Hubrich, Jessica (2009b): Vom Stringmatching zur Begriffsexploration: das Potential integrierter begrifflicher Interoperabilität. Vortrag auf der 12. Tagung der Deutschen ISKO, „Wissen, Wissenschaft, Organisation“ Bonn, 20. Oktober 2009. Online verfügbar unter http://linux2.fbi.fh-koeln.de/crisscross/publikationen/hubrich_isko09.pdf, zuletzt geprüft am 16.04.2012

Hubrich, Jessica (2010a): Multilinguale Wissensorganisation im Zeitalter der Globalisierung: das Projekt CrissCross. In: Ohly, Peter H.; Sieglerschmidt, Jörg (Hg.): Wissensspeicher in digitalen Räumen. Nachhaltigkeit, Verfügbarkeit, semantische Interoperabilität. Proceedings der 11. Tagung der Deutschen Sektion der Internationalen Gesellschaft für Wissensorganisation Konstanz, 20. – 22. 02.2008. Würzburg: Ergon-Verl., S. 81-90

Hubrich, Jessica (2010b): Thematische Suche in heterogenen Informationsräumen. In: Bergner, Ute; Gömpel, Erhard (Hg.): The ne(x)t Generation, das Angebot der Bibliotheken: 30. Österreichischer Bibliothekartag Graz 15. – 18.09.2009. Graz-Feldkirch: Neugebauer (Schriften der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 7), S. 234-242

Jacobs, Jan-Helge; Mengel, Tina; Müller, Katrin (2010): Benefits of the CrissCross project for conceptual interoperability and retrieval. In: Gnoli, Claudio; Mazzocchi, Fulvio (Hg.): Paradigms and conceptual systems in knowledge organization: Proceedings of the Eleventh International ISKO Conference Rome, 23. – 26.02.2010 Würzburg: Ergon-Verl., S. 236-241

Kappler, Thomas (2006): Einführung in die Computerlinguistik. In: Witte, René; Mülle, Jutta (Hg.): Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten, S. 1-19. Online verfügbar unter http://digbib.ubka.uni-karlsruhe.de/volltexte/1000005161, zuletzt geprüft am 01.05.2012

Konchady, Manu: Building Search Applications: Lucene, LingPipe, and Gate. Oakton: Mustru Publ., 2008

Konferenz für Regelwerksfragen (Hg.) (1998): Regeln für den Schlagwortkatalog: RSWK. 3., überarb. und erw. Aufl. Berlin: Dt. Bibliotheksinstitut

Konferenz für Regelwerksfragen / Arbeitsgruppe Klassifikatorische Erschließung (Hg.) (2000): Einführung und Nutzung der Dewey Decimal Classification (DDC) im deutschen Sprachraum. Frankfurt am Main: Dt. Bibliothek

Lagoze, Carl; Van de Sompel, Herbert (2001): The Open Archives Initiative: building a low-barrier interoperability framework. In: Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries, S. 54-62. Online verfügbar unter http://www.openarchives.org/documents/jcdl2001-oai.pdf, zuletzt geprüft am 22.03.2012

Lösch, Mathias (2009): Automatische Klassifikation von OAI-Metadaten mit linguistischen Methoden. Vortrag im Kolloquium Wissensinfrastruktur an der UB Bielefeld, 30. Oktober 2009. Online verfügbar unter http://129.70.12.22//wikifarm/fields/ub_edv/uploads/Oeffentlich/auto_oai_slides.pdf, zuletzt geprüft am 16.04.2012

Lösch, Mathias; Waltinger, Ulli; Horstmann, Wolfram; Mehler, Alexander (2011): Building a DDC-annotated Corpus from OAI Metadata. In: Journal of Digital Information, Jg. 12, H. 2. Online verfügbar unter https://journals.tdl.org/jodi/article/viewPDFInterstitial/1765/1767, zuletzt geprüft am 14.03.2012.

Mitchell, Joan S. (2001): Relationships in the Dewey Decimal Classification System. In: Bean, C. A.; Green, R. (Hg.): Relationships in the organization of knowledge. Boston: Kluwer, S. 211–226. Online verfügbar unter http://polaris.gseis.ucla.edu/gleazer/462_readings/mitchell.PDF, zuletzt geprüft am 02.02.2012

Nohr, Holger (2003): Grundlagen der automatischen Indexierung: ein Lehrbuch. Berlin: Logos

Oberhauser, Otto (2005): Automatisches Klassifizieren: Entwicklungsstand, Methodik, Anwendungsbereiche. Frankfurt am Main: Lang (Europäische Hochschulschriften: Reihe XLI, Informatik ; 43)

Projektmitarbeiter des DFG-Projekts CrissCross (2010): Leitfaden zur Vergabe von DDC-Notationen an SWD-Schlagwörtern. Stand: 30. September 2010. Online verfügbar unter http://linux2.fbi.fh-koeln.de/crisscross/CrissCross_Endg_Grundlagenpapier_Sept2010.pdf, zuletzt geprüft am 17.04.2012

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, S. 44-49. Online verfügbar unter ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf, zuletzt geprüft am 30.04.2012

Sebastiani, Fabrizio (2002): Machine learning in automated text categorization. In: ACM Computing Surveys, Jg. 34, H. 1, S. 1-47. Online verfügbar unter http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf, zuletzt geprüft am 02.02.2012

Waltinger, Ulli; Mehler, Alexander; Lösch, Mathias; Horstmann, Wolfram (2011): Hierarchical Classification of OAI Metadata using the DDC Taxonomy. In: Bernardi, R.; Chambers, S.; Gottfried, B.; Segond, F.; Zaihrayeu, I. (Hg.): Advanced Language Technologies for Digital Libraries. Berlin: Springer (Lecture Notes in Computer Science, 6669), S. 29-40. Online verfügbar unter http://www.springerlink.com/content/x20257512g818377/, zuletzt geprüft am 02.02.2012

Wilcock, Graham (2009): Introduction to Linguistic Annotation and Text Annalytics. San Rafael, Calif.: Morgan & Claypool (Synthesis lectures on human language technologies, 3)

Wartena, Christian; Brussee, Rogier; Gazendam, Luit; Huijsen, Willem-Olaf (2007): Apolda: a practical tool for semantic annotation. In: 4th International Workshop on Text-base Information Retrieval (TIR-07), in conjunction with DEXA, Regensburg 03.-07.09.2007, S.288-292. Online verfügbar unter http://www.uni-weimar.de/medien/webis/research/events/tir-07/tir07-papers-final/wartena07-apolda-practical-tool-semantic-annotation.pdf, zuletzt geprüft am 25.04.2012

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die eingereichte Bachelorarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommen Stellen als solche kenntlich gemacht habe.

Hannover, den 22. Mai 2012

Maike Sommer