Christa Schöning-Walter

Automatische Erschließung bei der Deutschen Nationalbibliothek

Themen-Überblick

- Das Projekt PETRUS:
  Projektziel, Leitlinien, Ergebniserwartung, Organisation
- Arbeitsschwerpunkte:
  - Szenario 3 – Datenübernahme aus Parallelausgaben
  - Szenario 2 – Normdatenverknüpfung
  - Szenario 1 – automatische Sachgruppenvergabe
  - Szenario 4 – automatische Beschlagwortung
- Ausblick

Das Projekt PETRUS

„Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek“ (2009 – 2011)

Ziel:
Einführung softwaregestützter Verfahren für

- die Extraktion/Generierung formaler Erschließungsdaten
- die Verknüpfung mit den Normdateien
- die klassifikatorische und verbale Erschließung

Erprobung in einem automatisierten Geschäftsprozess zur Erschließung der Netzpublikationen

Leitlinien für PETRUS

- sinnvolle Verknüpfung konventioneller und maschineller Verfahren bei der Formalerschließung und Inhaltserschließung
- automatisierte Verfahren als Basisform der Verarbeitung für alle maschinenlesbaren Objekte
- perspektivisch soll auch die Beitragsebene mit in die Erschließung einbezogen werden: Zeitschriften-, Zeitungs-, Sammelbandartikel etc.
- die Regelwerke und bibliografischen Dienstleistungen bleiben im Grundsatz unangetastet
- das Finden von Informationen rückt in den Vordergrund:
  - schnelle Verfügbarkeit der Erschließungsdaten
  - Schaffung neuer Zugänge/Sucheinstiege für den Nutzer

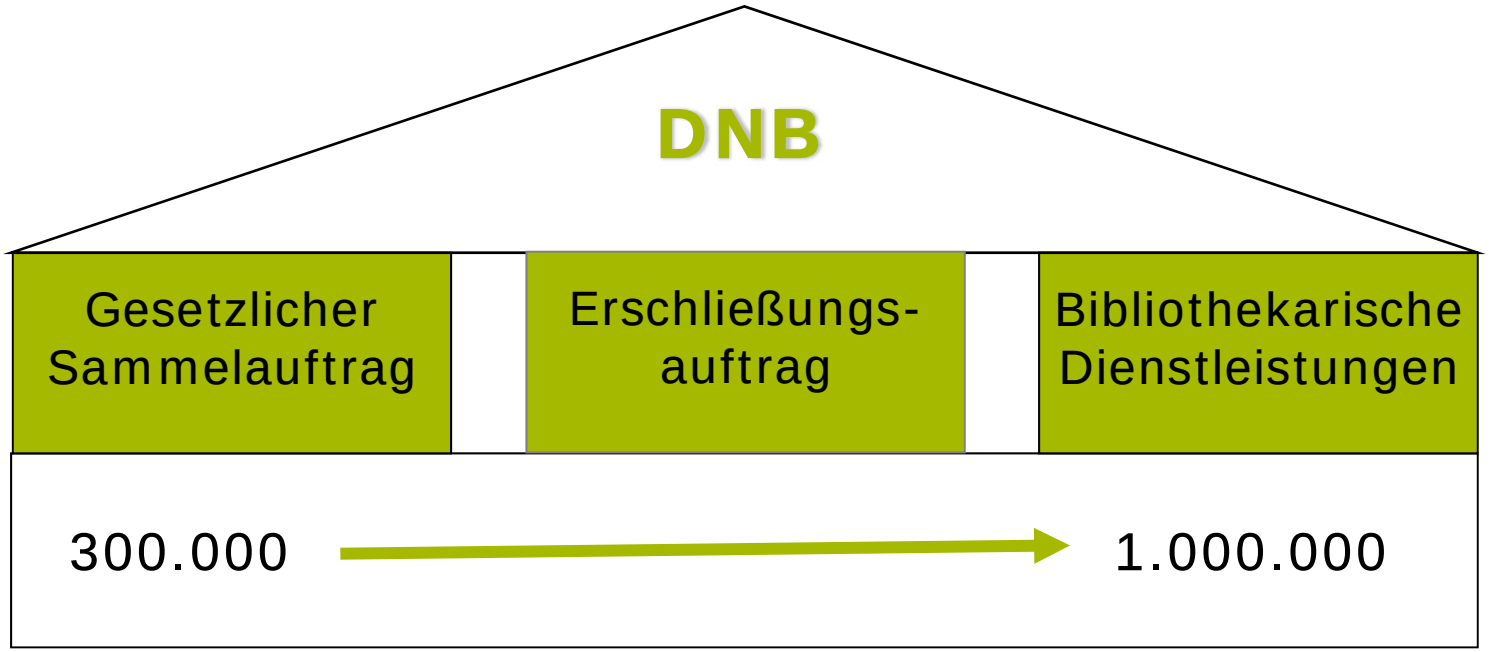
Ergebniserwartung

Entstehen sollen

- ein modular zusammengestelltes, nachregelbares Gesamtsystem mit Differenzierungsstufen für verschiedene Objekttypen zur Unterstützung neuer Geschäftsprozesse und Erschließungsverfahren,
- sowie ein Bewertungssystem mit definierten Qualitätskriterien für die Messung von Qualitätsstufen als Handlungs- und Entscheidungsbasis und zur Steuerung und Rückkopplung der Prozesse.

Dies wird in Szenarien entwickelt, erprobt und eingeführt.

Warum erforderlich?



PETRUS ist angelegt als standortübergreifendes Projekt …

… mit einer abteilungsübergreifenden Organisationsstruktur

Szenario 3:
Datenübernahme aus Parallelausgaben

Aufgabe:

- automatische Erkennung von parallelen Print- und Onlineausgaben (PICA Match&Merge)
- maschinelle Übernahme bereits vorhandener Inhaltserschließungsdaten und Normdatenverknüpfungen

Ergebnis:

- identische Erschließungsdaten in beiden Datensätzen
- Kennzeichnung der Parallelausgaben

Szenario 2: Normdatenverknüpfung

Aufgaben:

- Normdatensätze in der Personennamendatei (PND) automatisch angelegen und mit den Titeldaten verknüpfen
- Datenanalyse- und Statistikfunktionen zur Steuerung des Geschäftsgangs bereitgestellen

1. Phase:

- Übernahme aller Personennamen in den Titeldatensatz
- Verknüpfung mit einem existierenden Datensatz in der PND oder Anlegen eines neuen PND-Satzes

Szenario 1:
Automatische Sachgruppenvergabe

Aufgabe:

Maschinell lesbare Publikationen (Netzpublikationen) sollen mit automatischen Verfahren in die richtige Hauptsachgruppe eingeordnet werden.

Angewendet wird die Systematik der DNB-Sachgruppen mit ihren ca. 100 Klassen.

Qualitätsziel:

Die automatische Zuordnung soll in mindestens 80 % der Fälle korrekt sein.

Vergleichsmaß ist die intellektuell vergebene Sachgruppe.

Haupt-Fragestellungen der laufenden Testphase mit 4 Testsystemen

- Welche statistischen Kategorisierungsverfahren eignen sich besonders gut für die Einordnung von Publikationen in die Systematik der DNB-Sachgruppen (SVM, kNN etc.)?
- Können die Klassifikatoren auch mit den gescannten Inhaltsverzeichnissen, mitgelieferten Klappentexten etc. trainiert werden – also mit der wachsenden Menge maschinenlesbarer Texte aus der Kataloganreicherung?
- Ist das Qualitätsziel (80 % richtige Kategorisierungen) mit maschinellen Methoden überhaupt erreichbar? Ist es günstig, bei der automatischen Erschließung evtl. bis zu drei DNB-Sachgruppen zuzuweisen?

Erprobung lernender Systeme

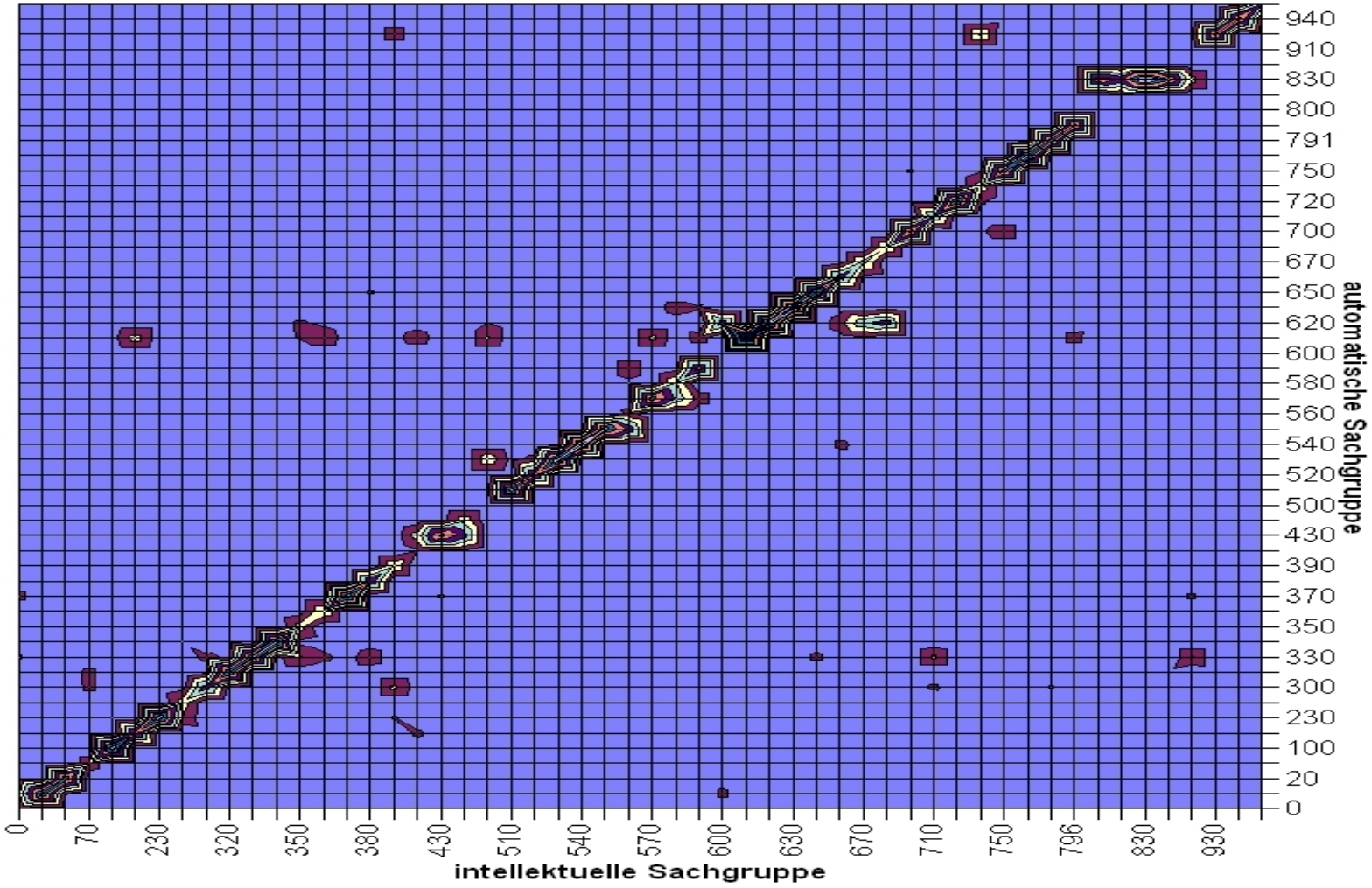
Testfälle für das maschinelle Training der Klassifikatoren mit intellektuell erschlossenen Objekten:

- Training und Test mit digitalen Volltexten (ca. 45.000 Online-Hochschulschriften und Monografien)
- Training und Test mit Daten aus der Kataloganreicherung (ca. 120.000 gescannte Inhaltsverzeichnisse gedruckter Monografien)
- Training mit den gescannten Inhaltsverzeichnissen; Test mit den digitalen Volltexten
- Praxistests mit Publikationen aus der Reihe O (Netzpublikationen, die ab Jahresbeginn 2010 gesammelt und bisher noch nicht erschlossen wurden)

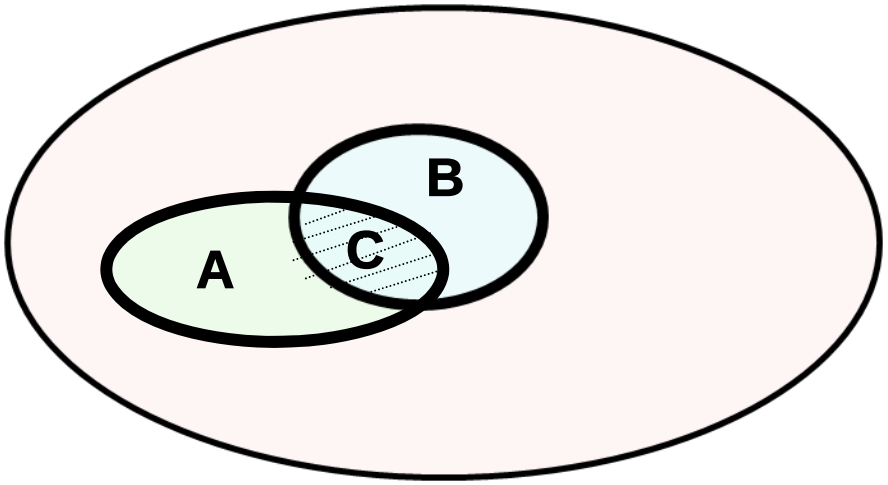
Herausforderungen und Schwierigkeiten

- Die Verteilung der Publikationen auf die Sachgruppen ist unausgewogen.
  - Testfall Volltexte: 90 % der Objekte verteilen sich auf 20 Sachgruppen
  - Testfall Inhaltsverzeichnisse: 70 % der Objekte verteilen sich auf 20 Sachgruppen
- Manche Sachgruppen lassen sich inhaltlich nur schwer abgrenzen.
  - Beispiele: Technische Chemie (660), Soziale Probleme, Sozialarbeit (360), Industrielle Fertigung (670), Literatur, Rhetorik, Literaturwissenschaft (800)
- Für manche Sachgruppen existiert kaum Trainingsmaterial.
  - Beispiel: neue Sachgruppe 333.7 (Natürliche Ressourcen, Energie und Umwelt)
- Die Datenaufbereitung ist aufwändig und schwierig.
  - Beispiele: Entpacken von zip-Ordnern, Formatkonvertierung, Umkodierung, Sprachenidentifizierung etc.

Gegenüberstellung von intellektueller und automatischer Erschließung



Qualitätsmaße zur Beurteilung der Erschließungsqualität



A: Objekte einer Sachgruppe nach intellektueller Erschließung

B: Objekte, die dieser Sachgruppe bei der automatischen Erschließung zugeordnet wurden

$$\text{Recall} = \frac{C}{A}$$

$$\text{Precision} = \frac{C}{B}$$

F-Measure (Harmonisches Mittel) =

$$\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Bisherige Erkenntnisse

- Die Testsysteme erreichen Ergebnisse von durchschnittlich 65 % bis 75 % richtig zugewiesener Hauptsachgruppen.

  Damit scheinen die Möglichkeiten der statistischen Verfahren weitgehend ausgeschöpft zu sein. Für den weiteren Optimierungserfolg sind voraussichtlich andere Aspekte ausschlaggebend.

- Das Training mit gescannten Inhaltsverzeichnissen liefert Erschließungsergebnisse ähnlicher Qualität wie das Training mit digitalen Volltexten. Gescannte Inhaltsverzeichnisse können also für die Modellbildung mit herangezogen werden.

- Die Vergabe von mehr als 2 Sachgruppen hat sich als nicht sinnvoll erwiesen. Auch bei einer automatischen Klassifizierung wird die Vergabe von lediglich einer Hauptsachgruppe als Regelfall angestrebt.

Szenario 4:
Automatische Beschlagwortung

Aufgaben:

- automatische Vergabe von Schlagwörtern mit Normdateien (Schlagwortnormdatei, Personennamendatei) als kontrolliertes Vokabular
- zusätzliche Vergabe freier Schlagwörter

Die Zuordnung von Schlagwörtern soll begrenzt werden

- entweder über den Konfidenzwert,
- oder durch Festlegung einer festen Anzahl für die Schlagwörter.

Gestaltung der Testfälle für 2 Testsysteme

- Die erste Stufe des Thesaurus enthält (nur) die 160.000 Sachschlagwörter der SWD.
- Aktuell analysiert werden 16 ausgewählte Sachgruppen:
  004 – Informatik ; 100 – Philosophie ; 150 – Psychologie ;
  300 – Sozialwissenschaften, Soziologie ; 330 – Wirtschaft ;
  340 – Recht ; 370 – Erziehung, Schul- und Bildungswesen ;
  510 – Mathematik ; 530 – Physik ; 540 – Chemie ; 610 – Medizin ;
  620 – Ingenieurwissenschaften ; 650 – Management ;
  700 – Künste, Bildende Kunst ; 830 - Deutsche Literatur ;
  900 – Geschichte
- Die Bewertung erfolgt über Stichproben: intellektuelle Bewertung der inhaltlichen Übereinstimmung zwischen den automatisch vergebenen SWD-Schlagwörtern und dem Thema der Publikation.

Beispiel:
Die Schweiz in Europa: mittendrin, doch außen vor?

automatisch vergebene SWD-SW	ID	sehr nützlich	nützlich	wenig nützlich	falsch/schädlich
Schweiz	4713011-8	☐	☐	☐	☑
Volksabstimmung	4134790-0	☑	☐	☐	☐
Beitritt	4120988-6	☑	☐	☐	☐
Mitwirkung	4140386-1	☐	☑	☐	☐
Volksbegehren	4063810-8	☑	☐	☐	☐
Zeitungsartikel	4125430-2	☐	☐	☑	☐
Volksrechte	4188544-2	☐	☑	☐	☐
Eidgenossenschaft	4151157-8	☐	☑	☐	☐
Res	4552375-7	☐	☐	☐	☑
Marschroute	4750587-4	☐	☐	☐	☑
Monozyten-Makrophagen-System	4177897-2	☐	☐	☐	☑
direkteDemokratie	4134792-4	☑	☐	☐	☐
Demokratie	4011413-2	☐	☑	☐	☐
Schweizer	4643846-4	☐	☑	☐	☐
wirtschaftlicheIntegration	4066410-7	☐	☑	☐	☐
Projekt	4115645-6	☐	☐	☑	☐
europ?ischeIntegration	4071013-0	☑	☐	☐	☐
Parlament	4044685-2	☐	☑	☐	☐
daspolitische	4136977-4	☐	☐	☑	☐
Integration	4027238-2	☐	☐	☑	☐
Mitgliedschaft	4135885-5	☐	☐	☑	☐

Thesaurusaufbau

Im Text gefundene Begriffe und Synonyme werden über den Thesaurus auf die jeweilige Ansetzungsform zurückgeführt.

Zukünftig sollen auch berücksichtigt werden:

- Homonyme (Disambiguierung)
- Geografika/Ethnografika
- Personennamen
- Körperschaften
- etc.

Linguistische Vorverarbeitung, Indexierung und Ranking

- morphologische Verarbeitung: Part of Speech (PoS) Tagging, Komposita-Zerlegung, Reduktion auf Grundformen, Stoppworterkennung, Phrasenerkennung etc.
- alternativ: Zerlegung in Morpheme
- Häufigkeit des Terms innerhalb der Publikation (Termfrequenz)
- alternativ: Ranking über eine TF-IDF-Statistik (Termfrequenz/inverse Dokumentfrequenz)
- Auswertung der Position des Terms innerhalb der Publikation (z. B. im Titel)
- Auswertung des semantischen Kontextes über semantische Netze
- verschiedene Filter

Ausblick für 2011

- Beschaffung und Anpassung eines Softwaresystems für die automatische Klassifizierung und Beschlagwortung
- Realisierung der einzelnen Erschließungsverfahren als Web Services
- Planung und Realisierung der Qualitätsmanagement-Verfahren
- Verknüpfung der Erschließungsmodule zu einer automatisierten Prozesskette
- schrittweise Anreicherung des bibliografischen Datensatzes in aufeinander aufbauenden Erschließungsstufen
- Anwendung am Beispiel der Netzpublikationen

Kontakt

Deutsche Nationalbibliothek

Christa Schöning-Walter

Digitale Dienste

+69-1525-1014

c.schoening@d-nb.de

http://www.d-nb.de/