Christa Schöning-Walter

PETRUS

Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek

Intellektuelle Erschließungsleistung kann quantitativ nicht mehr bewältigt werden

Der erweiterte Sammelauftrag der Deutschen Nationalbibliothek (DNB) und die stetig steigende Zahl der zu bearbeitenden Publikationen führen zu Anforderungen an die Erschließungsleistung, die auf herkömmliche Art und Weise – d. h. durch die intellektuelle Erstellung bibliografischer Metadaten für die Nationalbibliografie – nicht mehr erfüllt werden können.

Die DNB steht vor der Herausforderung, eine Brücke zwischen den traditionellen bibliothekarischen Erschließungsverfahren und den heutigen quantitativen und qualitativen Anforderungen zu schlagen. Anfang des Jahres 2009 wurde das Projekt PETRUS gestartet, um neue Methoden zu erproben und einzuführen. Ziel dieses Vorhabens ist es, die Erschließungsvorgänge durch eine sinnvolle Verknüpfung konventioneller und maschineller Verfahren effizienter zu gestalten und so einerseits den gesetzlichen Erschließungsauftrag auch bei dramatisch zunehmenden Publikationsmengen zu erfüllen und andererseits durch eine zeitnahe Erschließung dafür zu sorgen, dass alle bei der DNB gesammelten Objekte so schnell wie möglich in der Nationalbibliografie angezeigt werden. In Zukunft sollen softwaregestützte Verfahren der automatischen Indexierung und Klassifizierung sowie der Metadatenextraktion und Metadatengenerierung – wo immer möglich – die formale und die inhaltliche Erschließung unterstützen und dabei helfen, die intellektuelle Bearbeitung der Medien auf das Notwendige zu konzentrieren.

Ziel: Erschließungsvorgänge effizienter gestalten

Beginn mit den Onlinepublikationen

Die Dringlichkeit, neue Wege zu beschreiten, ist bei den elektronischen Medien besonders groß. Für die Bearbeitung von Onlinepublikationen will die DNB grundsätzlich nur noch automatisierte Verfahren einsetzen. Diese Leitlinie gilt nicht nur für das Sammeln der Veröffentlichungen (vorzugsweise über Harvestingverfahren) und für die Bereitstellung und Langzeitarchivierung, sondern ausdrücklich auch für die Erstellung der bibliografi-

PETRUS Kick-off-Veranstaltung im Februar 2009
Foto: Deutsche Nationalbibliothek / Stephan Jockel

schen Datensätze. Die folgenden Zahlen verdeutlichen die Notwendigkeit des Umstiegs: Waren es bisher etwa 300.000 Medieneinheiten, die pro Jahr bei der DNB zu erschließen waren, so ist zu erwarten, dass das Volumen schon in Kürze die Millionengrenze überschreiten wird. Dieses Anwachsen ist vor allem dadurch begründet, dass seit Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek (DNBG)[1)] im Jahr 2006 auch Netzpublikationen gesammelt werden.

Mit der Fertigstellung von Übertragungsschnittstellen und Ablieferungsverfahren für elektronische Publikationen nimmt deren Bestand jetzt deutlich zu. Um die notwendigen Freiräume für die geplante Entwicklung neuer Methoden und Arbeitsabläufe zu schaffen, führt die DNB seit Beginn dieses Jahres bei Netzpublikationen grundsätzlich keine intellektuelle Nachbearbeitung der bibliografischen Daten mehr durch, sondern setzt auf automatisierte Erschließungsverfahren als Basisform der Verarbeitung für alle maschinenlesbaren Objekte.

Ausschließlich automatisierte Verfahren zur Erschließung von Onlinepublikationen

Geplant ist eine iterative Vorgehensweise. Die DNB übernimmt die mitgelieferten Fremddaten zunächst ungeprüft in den bibliografischen Datensatz. Eine Nachbearbeitung erfolgt ggf. zu einem späteren Zeitpunkt mit geeigneten automatischen Erschließungsmethoden. Softwaregestützte Verfahren zur Extraktion oder Generierung von Daten für die Formal- und Inhaltserschließung werden in die Arbeitsabläufe der DNB integriert, sobald sie die festgelegten Qualitätsanforderungen erfüllen können. Die Performanz soll durch Weiterentwicklungen stetig verbessert werden. Das schließt auch das Training lernender Verfahren im laufenden Betrieb und den Austausch von Modulen durch bessere Lösungen mit ein. Onlinepublikationen werden zukünftig also von Zeit zu Zeit wiederholt Erschließungsprozesse durchlaufen und damit eine immer bessere Erschließungsstufe erreichen.

Onlinepublikationen werden wiederholt automatische Erschließungsprozesse durchlaufen

Die Systemsteuerung und Rückkopplung soll über definierte Qualitätsstufen erfolgen. Die Anforderungen an die Tiefe und Güte der Erschließung variieren von Objekttyp zu Objekttyp. Webseiten werden voraussichtlich anders erschlossen als Dissertationen oder Forschungsberichte, Verlagsveröffentlichungen anders als Mitteilungen von Behörden oder Instituten, wissenschaftliche Publikationen anders als die Belletristik. Von den automatischen Verfahren wird teilweise verlangt, dass sie Konfidenzwerte mitliefern, um diese mit den festgelegten Qualitätsanforderungen abgleichen zu können. Projektiert sind maschinengestützte Arbeitsabläufe, in denen die eingehenden Medien mit spezifischen – d. h. auf die Eigenschaften eines Medientyps abgestimmten – Erschließungsmodulen automatisiert bearbeitet werden. Intellektuelle Eingriffe sollen möglichst nur an wenigen definierten Schnittstellen erfolgen, z. B. bei Unsicherheiten, bei Qualitätschecks und bei der Steuerung und Kontrolle des Systems. In Stufen soll so ein nachregelbares Gesamtsystem entstehen.

Maschinengestützte Arbeitsabläufe müssen unterschiedliche Objekttypen berücksichtigen

Auswirkungen auf andere Dienste

Durch diese für die Onlinepublikationen geplanten Entwicklungen werden sich voraussichtlich auch für die traditionellen Medien neue Möglichkeiten der Verarbeitung eröffnen. Automatische Verfahren können im Prinzip immer dann angewendet werden, wenn zu einer gedruckten Publikation auch maschinenlesbare Daten zur Verfügung stehen. Surrogate in elektronischer Form entstehen z. B. im Zuge der vielfältigen Maßnahmen zur Kataloganreicherung (Titelblatt-Scans, gescannte Inhaltsverzeichnisse, Register, Klappentexte, Abstracts etc.). In der Nationalbibliografie sollen in zunehmendem Umfang auch auf der Beitragsebene (Zeitschriften-, Zeitungs-, Sammelbandartikel, Musik-Tracks etc.) Erschließungsdaten bereitgestellt werden. In Anbetracht der Mengenproblematik setzt auch dieses Anliegen voraus, dass die bibliografischen Daten dafür entweder bereits in ausreichender Qualität mit der Publikation mitgeliefert werden – oder dass sie bei der DNB automatisch extrahiert bzw. generiert werden können.

Zukünftig auch Erschließungsdaten auf Beitragsebene

Die bibliografischen Datensätze dienen dem Nutzer zum schnellen, genauen und direkten Zugang zum eigentlichen Objekt – und sind insofern lediglich Mittel zum Zweck. Erschließung muss den Bedürfnissen des Findens folgen. Für die Nationalbibliografie – gleichbedeutend mit dem Onlinekatalog der DNB – hat dies zur Konsequenz, dass das Angebot immer stärker auf die Retrievalbedürfnisse

der Nutzer ausgerichtet werden sollte. Der Katalog muss die Fragen des »wer«, »wo«, »wann« und »was« möglichst optimal bedienen. Ziel ist deshalb eine weitreichende Anreicherung der Katalogdaten mit inhaltstragenden Begriffen (z. B. auch freie Schlagwörter).

Onlinekatalog muss auf Retrievalbedürfnisse abgestimmt sein

Die bibliografischen Dienstleistungen der DNB dürfen durch den Umstieg auf automatische Erschließungsverfahren nicht beeinträchtigt werden. Die DNB hat im Jahr 2008 mehr als 80 Mio. bibliografische Einheiten (Titeldaten- und Normdatensätze) an andere Institutionen ausgeliefert. Damit die Erschließungsdaten auch weiterhin in Qualitätsdiensten (z. B. WorldCat, Europeana, lokale Bibliothekssysteme, Fachportale etc.) nachgenutzt werden können, müssen die bibliothekarischen Standards im Grundsatz unangetastet bleiben.

Anwendungsszenarien als Teilprojekte

Die Ausgangssituation des Projektes ist gekennzeichnet durch eine große Vielfalt der Medieneinheiten (Struktur und Größe der Objekte, Umfang der maschinell extrahierbaren Metadaten, unterschiedliche Verfügbarkeit maschinenlesbarer Daten etc.) und eine große Breite der Fachgebiete, aus denen die Objekte stammen. Die Erschließungsaufgabe umfasst letztlich die gesamte Spannbreite an Publikationsformen, die bei der DNB gesammelt werden.

Erschließung umfasst sämtliche Publikationsformen

In Anbetracht der Komplexität gliedert sich das Projekt PETRUS in inhaltlich überschaubare Anwendungsszenarien, die bestimmt werden durch ausgewählte Medientypen und an sie geknüpfte Erschließungsanforderungen. Die einzelnen Szenarien werden wie Teilprojekte realisiert. Dabei arbeiten die Fachabteilungen und die IT-Abteilung eng zusammen.

Ein wichtiges Projektziel ist es, für die maschinellen Verfahren auch das Potenzial des umfangreichen Datenbestandes und der Normdateien zu nutzen. D. h.: Die vorhandenen Daten sollen stärker »mitarbeiten«, z. B. beim Datenabgleich oder beim Trainieren lernender maschineller Verfahren. Die Normdateien sind als »Wissensbasen« hoher Qualität wichtige Bausteine des Gesamtkonzepts. Es ist zu vermuten, dass ihre schon jetzt große Bedeutung im Zuge der Einführung automatischer Erschließungsverfahren sogar noch weiter zunehmen wird.

Normdateien sollen genutzt werden

Szenario Normdatenrelationierung

Fremddaten sollen zukünftig in größerem Umfang als bisher zur Erstellung von Titeldatensätzen und zur Pflege und zum Aufbau von Normdateien genutzt werden. Strukturiert gelieferte Daten werden bei der Übernahme direkt in den bibliografischen Datensatz übernommen und dem Nutzer für die Suche zur Verfügung gestellt.

Bei der Übernahme von Personenmetadaten soll zukünftig auch der Datensatz in der Personennamendatei (PND)[2)] automatisch erstellt (bzw. gepflegt) und eine Verknüpfung zwischen Normdaten und Titeldaten hergestellt werden. Die auf diese Weise automatisch generierten Normdatensätze und Relationen sollen in Abhängigkeit von

ihrer Häufigkeit intellektuell nachgepflegt werden. Eine so genannte Individualisierung von Namenssätzen in der PND erfolgt nur dann, wenn eine vielfache Relationierung vorliegt. Diese Maßnahmen sind ein erster Schritt in die Richtung einer möglichst weitreichenden Verknüpfung von bibliografischen Daten und Normdaten im Sinne von Modellen wie FRBR (Functional Requirements for Bibliographic Records) und FRAD (Functional Requirements for Authority Records).

PND-Datensatz soll automatisch erstellt werden

Als Datengrundlage für das Szenario dienen die von der Marketing- und Verlagsservice des Buchhandels GmbH (MVB) sowohl für konventionelle Publikationen als auch für Netzpublikationen gelieferten Personenmetadaten.

Szenario Parallelausgaben

Im Rahmen eines weiteren Szenarios erfolgt die automatische Verknüpfung von parallelen Print- und Onlineausgaben und eine maschinelle Übernahme ggf. bereits vorhandener Inhaltserschließungsdaten in den Titeldatensatz der entsprechenden Parallelausgaben.

Für Netzpublikationen sollen auf diese Weise die Daten der verbalen und klassifikatorischen Feinerschließung von der entsprechenden Druckausgabe übernommen werden.

Datenübernahme von Druckausgaben

Auch in diesem Szenario muss letztlich ein Abgleich von Zeichenketten durchgeführt werden. Dabei müssen auch Ähnlichkeitswerte berechnet werden können, damit z. B. Rechtschreibfehler und alternative Schreibweisen aufgefangen werden. Begonnen werden soll mit den Hochschulschriften (pro Jahr werden rund 10.000 Online-Hochschulschriften gesammelt, der Gesamtbestand umfasste zu Beginn dieses Jahres rund 82.000 Objekte). In nachfolgenden Schritten sollen weitere parallel vorkommende Erscheinungsformen einbezogen werden.

Szenario automatische Vergabe der DNB-Sachgruppen

Seit dem Bibliografiejahrgang 2004 werden die verschiedenen Reihen der Deutschen Nationalbibliografie nach 100 Sachgruppen gegliedert. Auch mit Einstellung der gedruckten Ausgabe zu Beginn dieses Jahres wird die Sortierung nach Sachgruppen in den wöchentlichen Verzeichnissen beibehalten. Die DNB vergibt für jeden Titel eine Hauptsachgruppe und zur Verbesserung des Retrievals nach klaren Regeln bis zu zwei Nebensachgruppen.

Die Sachgruppen ergänzen die verbale Inhaltserschließung und dienen der thematischen Ordnung der Titeldaten. Im Onlinekatalog ermöglichen sie die Realisierung von Filtern, Browsingfunktionen und Alert-Diensten und sind ein wichtiges Instrument, um große Treffermengen bei der Recherche einzugrenzen. Andere Bibliotheken nutzen die DNB-Sachgruppen als Selektionskriterium für Erwerbungszwecke.

Synergieeffekte

Ziel des Szenarios ist eine automatische Generierung der Sachgruppen. Es soll ein automatisches Klassifizierungsverfahren zum Einsatz kommen, das durch Anwenderaktionen oder Trainingskorpora lernt und schrittweise zu einer Vollautomatisierung führt. Bei der Klassifizierung werden die analysierten Objekte anhand von charakteristischen Merkmalen einer Kategorie (Sachgruppe) zugewiesen. Zum Trainieren der Verfahren können bereits erschlossene Materialien im Bestand der DNB benutzt werden. Mit Blick auf die Bedeutung der Sachgruppen wird eine hohe Qualität der Erschließungsergebnisse verlangt. Das Qualitätsziel ist eine mindestens 80 %-ige Übereinstimmung der automatisch erzeugten Kategorien im Vergleich mit einer Erschließung auf intellektueller Basis.

Szenario automatische Beschlagwortung

Im Rahmen von PETRUS sollen auch Methoden zur Unterstützung der verbalen Inhaltserschließung implementiert werden. Ein Ziel ist die automatische Anreicherung der Titeldaten deutschsprachiger Publikationen mit Schlagwörtern und freiem Vokabular. Die Schlagwortnormdatei (SWD)[3] ist zunächst die Grundlage für eine automatische Beschlagwortung mit einem kontrollierten Vokabular. Eventuell werden mittelfristig auch andere Thesauri (z. B. Standardthesaurus Wirtschaft, Fachthesaurus Technik etc.) eingebunden.

SWD als Basis für automatische Beschlagwortung

Neben den genormten Begriffen der verwendeten Thesauri sollen bei der automatischen Beschlagwortung auch freie Deskriptoren extrahiert werden, die aus der Publikation selbst stammen. Ein Ziel ist es, damit die Retrievalfunktionalität im Onlinekatalog zu verbessern. Ein weiteres Ziel ist es, mit neuen Begriffen die Pflege der SWD zu unterstützen.

Die Schlagwörter in der SWD basieren auf der natürlichen Sprache und unterliegen terminologischer Kontrolle. Sie sind nach den Regeln für den Schlagwortkatalog (RSWK) angesetzt. Für das PETRUS-Szenario sind insbesondere die in der SWD enthaltenen rund 160.000 Sachschlagwörter mit ihren verzeichneten Relationen (bevorzugte Bezeichnung, Synonyme, Homonyme und Polyseme, über- und untergeordnete Begriffe, verwandte Begriffe etc.) von Bedeutung. Ob es gelingt, die SWD in ihrer ganzen Komplexität erfolgreich in automatische Erschließungsverfahren einzubetten, sollen Untersuchungen in den nächsten Monaten zeigen.

Auch in diesem Szenario sollen Softwarewerkzeuge erprobt und einführt werden, die linguistische und statistische Verfahren sinnvoll miteinander kombinieren. Zur Ermittlung inhaltsrelevanter Deskriptoren werden meistens statistische Methoden angewendet. Die (stark flektierende) deutsche Sprache ist jedoch schwierig zu analysieren. Voraussichtlich sind deshalb auch linguistische Verarbeitungsschritte erforderlich, z. B. um eine Reduktion auf Grundformen zu erreichen.

Bei der Bewertung der Qualität einer automatischen Beschlagwortung können die Ergebnisse der intellektuellen Erschließung nicht so einfach wie bei der Sachgruppenvergabe als Maßstab herangezogen werden. Hier sind komplexere Auswertungsverfahren erforderlich.

Nächste Schritte

Die Realisierung der Szenarien soll ausschließlich mit erprobten Softwarewerkzeugen erfolgen. Die eigene Entwicklung von Erschließungswerkzeugen ist nicht vorgesehen. Existierende Technologien sollen im Anwendungskontext von PETRUS hinsichtlich ihrer Eignung untersucht werden. Dabei sollen auch verschiedene methodische Ansätze verglichen werden. Ein Schwerpunkt der Projektarbeiten ist deshalb die Auswahl und Evaluierung geeigneter Systeme und deren Integration in die eigene Arbeitsumgebung.

Keine Eigenentwicklung von Erschließungswerkzeugen für die Szenarien

Die automatischen Erschließungsverfahren müssen letztlich eng mit den bei der DNB schon vorhandenen Applikationen (Ablieferungsverfahren, zentrales Bibliothekssystem, Langzeitarchivierung, Portalanwendungen etc.) verknüpft und gut in die Systemarchitektur integriert werden können. Sobald geeignete Lösungen identifiziert sind, sollen diese so schnell wie möglich in die Arbeitsabläufe der DNB integriert werden.

Zur Unterstützung der Geschäftsprozesse entsteht auf diese Weise Schritt für Schritt ein modular zusammengestelltes und nachregelbares Erschließungssystem, das zwischen verschiedenen Medientypen und deren Erschließungsanforderungen differenzieren kann und dessen Ablauf über definierte Qualitätskriterien gesteuert wird.

Anmerkungen

1 <http://bundesrecht.juris.de/dnbg/index.html>

2 <http://www.d-nb.de/standardisierung/normdateien/pnd.htm>

3 <http://www.d-nb.de/standardisierung/normdateien/swd.htm>