Elisabeth Mödden, Katrin Tomanek

Maschinelle Sachgruppenvergabe für Netzpublikationen

Vom Projekt PETRUS in die Praxis

Die Aufnahme elektronischer Publikationen in bibliothekarische Sammlungen hat zu stark anwachsenden Dokumentmengen geführt, die allein durch Formen intellektueller Bearbeitung nicht zu bewältigen sind. Der Einbezug maschineller Verfahren ist daher für die Zukunft der Bibliotheken von großer Bedeutung. In der Deutschen Nationalbibliothek (DNB) hat man auf diese Herausforderung mit dem Projekt PETRUS (Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek) reagiert. Über drei Jahre hinweg wurden verschiedene Aspekte der automatischen Erschließung in differenzierten Anwendungsszenarien erprobt.[1)] Ein Schwerpunkt ist die maschinelle Sachgruppenvergabe. Mit Projektabschluss wird sie nun in die Erschließungspraxis der DNB eingebunden und soll künftig kontinuierlich weiterentwickelt werden. Seit Beginn des Jahres 2012 wird zunächst den deutsch- und englischsprachigen Netzpublikationen, die bei der DNB abgeliefert werden, maschinell eine DDC-Sachgruppe zugewiesen. Netzpublikationen, die zwischen 2010 und 2012 abgeliefert wurden, sollen rückwirkend eine DDC-Sachgruppe erhalten.

Kontinuierliche Weiterentwicklung der maschinellen Sachgruppenvergabe

Die DDC-Sachgruppen stellen ein System zur thematischen Ordnung von Titeldaten dar, auf das die DNB zur Gliederung der Deutschen Nationalbibliografie und des Neuerscheinungsdienstes zurückgreift. Jeder Veröffentlichung, die die Abteilung Inhaltserschließung durchläuft, werden eine Hauptsachgruppe und bis zu zwei Nebensachgruppen zugewiesen. Sie basieren in ihrer Grundstruktur auf der zweiten Ebene der Dewey-Dezimalklassifikation (DDC), den so genannten hundert Klassen der zweiten Ebene. Die DNB setzt die Sachgruppen seit dem Bibliografiejahrgang 2004 ein. Mittlerweile wurden geringfügige Modifikationen vorgenommen, insofern zusätzliche Sachgruppen gebildet wurden. Die Sortierung der Deutschen Nationalbibliografie nach Sachgruppen wurde auch beibehalten, nachdem die gedruckte Ausgabe der Bibliografie im Jahr 2010 von der PDF-Ausgabe abgelöst wurde. Die Verwendung der Sachgruppen beschränkt sich nicht auf die Deutsche Nationalbibliografie; auch die Österreichische Bibliografie und das Schweizer Buch greifen darauf zurück. Außerdem bedienen sich Buchhandel und Verlage der Systematik.

Inhaltserschließung mithilfe der DDC-Sachgruppen

Während die Zuweisung von Sachgruppen bei konventionellen Publikationen seit 2004 durchweg auf intellektuellem Wege erfolgte, hat sich die Erschließungspraxis bei Netzpublikationen mittlerweile grundlegend geändert. Bis 2010 waren Netzpublikationen in der Deutschen Nationalbibliografie auf unterschiedliche Reihen verteilt und die dazugehörigen Sachgruppen wurden von Fachreferenten intellektuell vergeben. Um der gestiegenen Bedeutung von Netzpublikationen Rechnung zu tragen, werden diese innerhalb der Deutschen Nationalbibliografie seit 2010 in einer eigenen Reihe, der Reihe O, erfasst. Mit Einführung dieser neuen Bibliografiereihe wurde die intellektuelle Erschließung von Netzpublikationen eingestellt, d. h. es werden keine DDC-Sachgruppen mehr durch Mitarbeiter der DNB vergeben.[2)] Teilweise nutzen die Ablieferer von Netzpublikationen die bestehende Möglichkeit, selbst Sachgruppen zu vergeben.

Verzeichnung von Netzpublikationen in der Bibliografiereihe O

Mit Beginn des Jahres 2012 startete die maschinelle Sachgruppenvergabe. Zunächst wird den täglich neu eintreffenden Netzpublikationen eine DDC-Sachgruppe mit einem Konfidenzwert zugeordnet. Beide Angaben – DDC-Sachgruppe und Konfidenzwert – werden direkt in den bibliografischen Datensatz übernommen. Die Konfidenzwerte erlauben Rückschlüsse auf die Vertrauenswürdigkeit der maschinell erzeugten Sachgruppen und werden für die Steuerung des weiteren Geschäftsgangs eingesetzt. Unterschreitet der Konfidenzwert einen festgelegten Schwellenwert, wird zusätzlich ein Statusindikator in den Datensatz geschrieben. Die Netzpublikationen, bei denen dieser Fall vorliegt, können dann im weiteren Geschäftsprozess gezielt nachbearbeitet werden.

Maschinelle Sachgruppenvergabe seit Beginn des Jahres 2012

Untersuchungen zur Machbarkeit

Die Grundlage für dieses Verfahren bilden die Ergebnisse und Erkenntnisse, die mit dem Projekt PETRUS erreicht wurden. Die maschinelle DDC-Sachgruppenvergabe für Netzpublikationen als ein Teilprojekt gliederte sich in eine Vorbereitungs-, eine Evaluierungs- und eine Entwicklungsphase.

Im September 2009 wurde eine erste europaweite Ausschreibung durchgeführt, um Softwaresysteme zu identifizieren, die für die Realisierung der Projektziele besonders geeignet erschienen. Vier Systeme bildeten im Jahr 2010 die Grundlage der Evaluierung softwaregestützter Methoden und Verfahren für die Einordnung der Netzpublikationen in die Systematik der DDC-Sachgruppen. Hierfür wurde eine Testumgebung eingerichtet, die sowohl von DNB-Mitarbeitern in Leipzig und Frankfurt am Main als auch von Mitarbeitern der Firmen über einen Remote Access Service direkt vom Arbeitsplatz aus bedient werden konnte. Im Rahmen der Tests wurde ermittelt, dass sich Machine-Learning-Verfahren aus dem Bereich der Support Vector Machines (SVMs) gut dafür eignen, textbasierte Dokumente, wie sie üblicherweise in Bibliotheken gesammelt werden, nach Sachgruppen zu klassifizieren. Bei einem solchen Verfahren lernt das System durch intellektuell erschlossene Publikationen. Im konkreten Fall standen für Training und Test etwa 45.000 digitale Volltexte zur Verfügung. Bei diesen handelte es sich überwiegend um Online-Hochschulschriften. Schwierigkeiten bereitete das extreme Ungleichgewicht der Sachgruppen: 90 % der Publikationen verteilten sich auf lediglich 20 der insgesamt 100 Sachgruppen. Nur dieser Anteil der Sachgruppen erfüllte mit 50 oder mehr Beispielobjekten die Mindestvoraussetzungen für das Training. Demgegenüber stellte die Sachgruppe für Medizin mit 20.000 Beispielobjekten, überwiegend Dissertationen, nahezu die Hälfte aller Dokumente. Das Erschließungssystem konnte also bei Weitem nicht für alle Sachgruppen trainiert werden. Deshalb wurden digitalisierte Inhaltsverzeichnisse von Printpublikationen als eine weitere Trainings- und Testbasis hinzugenommen. In der Testphase 2010 konnten etwa 120.000 gescannte Inhaltsverzeichnisse gedruckter Monografien mit einbezogen werden. Auf diese Weise konnten 81 Sachgruppen berücksichtigt werden, die jeweils mindestens 70 Beispielobjekte aufwiesen.

Trotz der dargestellten Schwierigkeiten überzeugten die Ergebnisse der Evaluierungsphase, da ohne besondere Anpassungen immer dann, wenn Trainings- und Testkorpus vom Objekttyp gut zueinander passten, etwa 80 % der Dokumente richtig eingeordnet wurden. Das führte zu der Entscheidung, mit diesen Methoden ein Verfahren für den Produktivbetrieb zu entwickeln. Im Zuge eines zweiten europaweiten Ausschreibungsverfahrens wurde die Averbis Extraction Platform der Averbis GmbH in Freiburg im Januar 2011 als Erschließungssoftware ausgewählt und für die automatische Erschließung in der DNB lizenziert.

Überzeugende Ergebnisse der Evaluierungsphase

Rückblick

Wie funktioniert die maschinelle Klassifizierung nach Sachgruppen?

Die Erschließungssoftware muss zunächst die Aufgabe lösen, aus maschinenlesbaren Texten Informationen zu extrahieren und dabei die Informationseinheiten zu identifizieren, die im Text eine hohe inhaltliche Relevanz besitzen. Die Averbis Extraction Platform stellt dafür eine Vielzahl linguistischer Vorverarbeitungsschritte zur Verfügung, mit denen verschiedene sprachliche Analyseebenen (Satz, Wort, Wortart etc.) erfasst werden können. Darauf aufbauend kann dann mit verschiedenen maschinellen Lernverfahren die Klassifizierung durchgeführt werden.

Die Textanalyse erfolgt in Teilschritten:

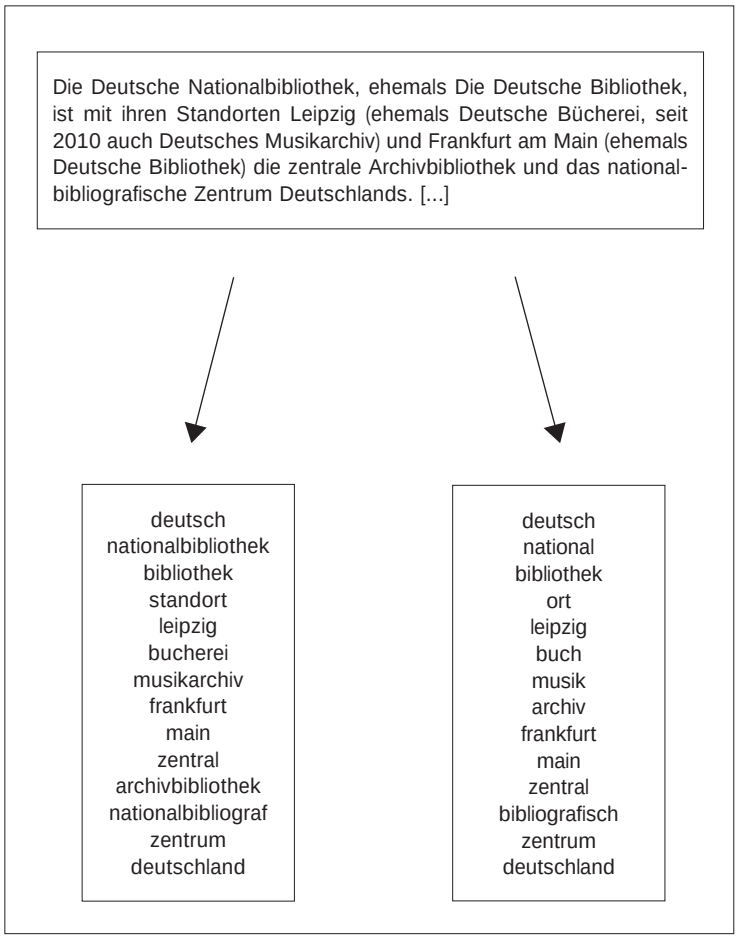
- Einlesen der Dokumente: Je nach Format wird die Syntax des Originaldokuments analysiert und relevante Textelemente werden identifiziert. Außerdem werden Metadaten aus der bibliografischen Datenbank (bei der DDC-Sachgruppenvergabe z. B. Informationen über Autoren, Titel, Verlage etc.) mit eingebunden.
- Linguistische Vorprozessierung: Satz- und Wortgrenzen sowie Wortarten werden erkannt, Nominalphrasen und Stoppwörter werden identifiziert.
- Morphologische Analyse: Auf dieser Ebene wird eine Stammformbildung und Kompositazerlegung durchgeführt.
- Semantische Analyse: hierbei werden Referenzen auf Konzepte einer Terminologie im Text

erkannt. Als Terminologie wird dabei die Schlagwortnormdatei (SWD) verwendet.

Ein so bearbeitetes Dokument muss anschließend in eine spezielle Darstellungsform überführt werden, die von dem Averbis-Klassifikator weiterverarbeitet werden kann.

Erstellung von Merkmalsvektoren

Für jedes Dokument wird zunächst ein so genannter Roh-Merkmalsvektor erstellt, der die Ergebnisse der morphologischen und semantischen Analyse enthält. Im einfachsten Fall werden alle im Text gefundenen Stammformen als Merkmale in einen solchen Vektor übertragen. Alternativ können statt Stammformen auch Segmente (Einzelwortbestandteile aus der Kompositazerlegung) oder die bei der semantischen Analyse erkannten Konzepte für die Bildung des Merkmalvektors verwendet werden. Die Verwendung von Segmenten oder Konzepten als Merkmale bietet den Vorteil, dass dadurch eine höhere Abstraktionsebene erreicht wird.



Merkmalsvektoren für die Wortstamm-Ebene (links) und die Komposita-Ebene (rechts) für einen Beispieltext (oben)

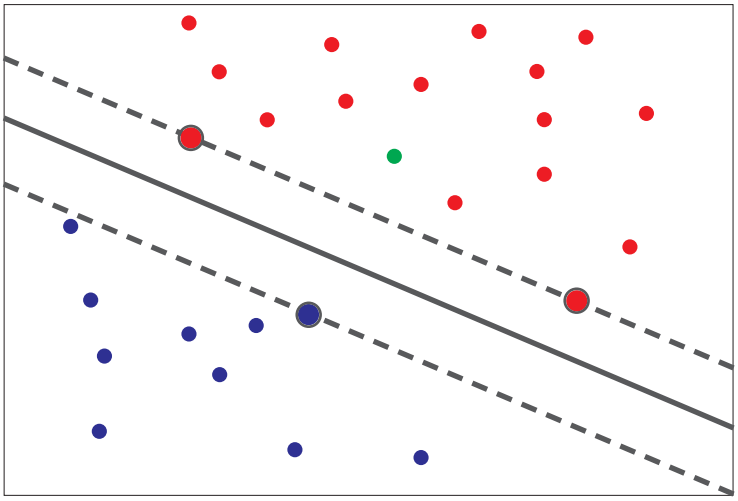
Die Roh-Merkmalsräume sind i. d. R. sehr groß und enthalten viele Merkmale, die keinen oder nur einen geringen Beitrag zur Klassifizierung neuer Dokumente leisten. Durch Eingrenzung der Merkmalsräume werden die Roh-Merkmalsvektoren so transformiert, dass nur solche Merkmale beibehalten werden, die möglichst viel zur richtigen Klassifizierung neuer Dokumente beitragen.

Gewichtung der Merkmale im zweiten Schritt

In einem zweiten Transformationsschritt der Merkmalsvektoren können die Merkmale anschließend gewichtet werden. Dabei werden Faktoren wie die Merkmalshäufigkeit im Dokument und die Merkmalshäufigkeit in Bezug auf den gesamten Merkmalsraum berücksichtigt. I. d. R. werden die Merkmalsgewichte am Ende des Gewichtungsprozesses schließlich einem Normalisierungsschritt unterzogen, um den Einfluss von Ausreißern zu minimieren.

Einsatz von Support Vector Machines

Der Averbis-Klassifikator stellt verschiedene Ansätze zur maschinellen Klassifikation bereit. Diese umfassen Entscheidungsbaumverfahren, probabilistische Verfahren (z. B. Naive Bayes oder Logistische Regression) und auch so genannte Large-Margin Classifier (beispielsweise die schon erwähnten Support Vector Machines). Für die Sachgruppenvergabe bei der DNB werden ausschließlich SVMs eingesetzt, da sich diese für den gegebenen Anwendungsfall (große Dokumentenkollektionen und vor allem große Merksmalsräume bei Volltexten) in der Evaluierungsphase als besonders geeignet erwiesen haben. Kernelement der SVMs ist eine Trennebene im hochdimensionalen Merkmalsraum, die die Klassen voneinander trennt. Diese Trennebene muss im Trainingsschritt aus gegebenen Trainingsdaten gelernt werden. Trainingsdaten sind Dokumente, zu denen Informationen zu der Klassenzugehörigkeit (hier die DDC-Sachgruppe) vorliegen. Anhand dieser Trennebene kann bei der Klassifikation neuer Dokumente entschieden werden, in welche Klasse eine zu klassifizierende Publikation einzuordnen ist. Soll ein neues Dokument klassifiziert werden, muss überprüft werden, auf welcher »Seite« der Trennebene das Dokument im Merkmalsraum liegt. Aus dem Abstand zur Trennebene lässt sich zusätzlich noch ablesen, wie sicher sich der Klassifikator ist. Je weiter ein Dokument im Merkmalsraum von der Trennlinie entfernt ist, desto sicherer ist die korrekte Klassifizierung dieses Dokumentes.



Diese Abb. zeigt beispielhaft einen Vektorraum (vereinfacht in einem 2-dimensionalen Merkmalsraum dargestellt): blaue Punkte sind Trainingsbeispiele der Klasse A, rote Punkte sind Trainingsbeispiele der Klasse B. Die dicke Linie stellt die gelernte Trennebene dar. Der grüne Punkt stellt ein neu zu klassifizierendes Dokument dar; da dieser Punkt oberhalb der Trennebene liegt, fällt er in die Klasse B

Anhand dessen wird der bereits erwähnte Konfidenzwert berechnet.

Einsatz linearer SVMs zur Sachgruppenvergabe

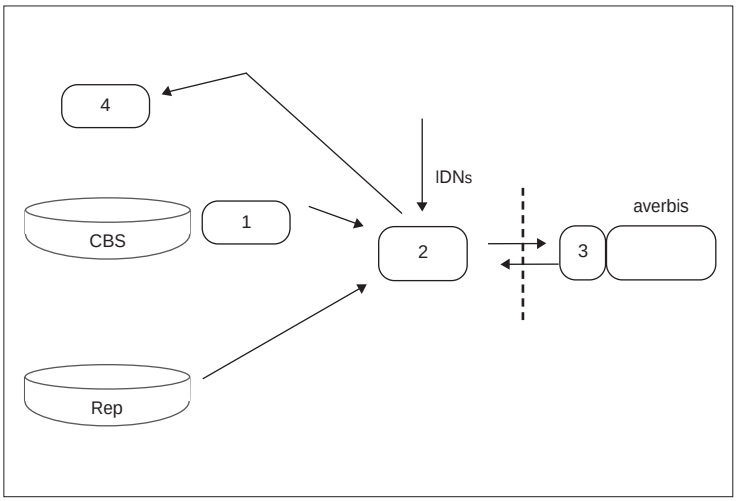
Für die DDC-Sachgruppenvergabe bei der DNB werden lineare SVMs eingesetzt, da diese sehr schnell sind, was sich vor allem auf die für das Training benötigte Zeit positiv auswirkt. Außerdem haben die Vorstudien in der Evaluierungsphase gezeigt, dass lineare SVMs für Textklassifikationsprobleme mit großen Merkmalsräumen gut geeignet sind. Die Trainingszeit für eine Dokumentenkollektion, die aus rund 225.000 Volltexten und Inhaltsverzeichnissen besteht, beträgt beispielsweise nur ca. 5 Minuten, wenn pro Dokument jeweils die ersten 40.000 Zeichen berücksichtigt werden.

Übergang in den Produktionsbetrieb

Um die maschinelle Sachgruppenvergabe durchführen zu können, muss die Averbis Extraction Platform in die Geschäftsprozesse und in die Systeminfrastruktur der DNB eingebettet werden. Die Erprobung neuer Entwicklungen findet in einer Testumgebung statt, die Abnahme erfolgt in einer Approvalumgebung. Die produktive maschinelle Erschließung wird schließlich in der Produktionsumgebung durchgeführt, in der auch andere Verarbeitungsprozesse ablaufen.

Vorbereitung des Produktionsbetriebes

Die Testumgebung aus der Evaluierungsphase diente im Jahr 2011 zum Testen der Anpassungen und Entwicklungen. Parallel dazu wurde der Produktionsbetrieb vorbereitet. Dafür mussten organisatorische und technische Maßnahmen geplant und umgesetzt werden. Für die maschinelle DDC-Sachgruppenvergabe wurde ein neuer Geschäftsprozess entwickelt und in die bestehenden Geschäftsprozesse integriert. Die Metadatenfelder im Katalogisierungsclient WinIBW wurden um Unterfelder erweitert. Von Mitarbeitern der DNB wurde ein so genannter »DNBPetrusService« entwickelt. Dieser koppelt die als Webservice umgebaute Averbis Extraction Platform an, die somit flexibel als Erschließungsmodul in die Systeminfrastruktur der DNB eingebettet werden kann. Die Abb. beschreibt den prinzipiellen Ablauf des Produktionsbetriebs.



Schematische Darstellung der Abläufe im Produktionsbetrieb

Im laufenden Betrieb startet der Prozess der maschinellen DDC-Sachgruppenvergabe täglich zu einer festgelegten Zeit. Die Identnummern (IDN) der neuen Netzpublikationen werden über das Erfassungsdatum, den Publikationstyp und weitere Kriterien selektiert und an den »DNBPetrusService« (2) übergeben. Die Übergabe bewirkt, dass der »DNBPetrusService« sich die zu erschließenden Texte mittels ARAS (Abstract Repository Access Service) aus dem Repository (Rep) und die zugehörigen Metadaten über eine lesende Schnittstelle (1) aus der bibliografischen Datenbank (CBS) holt.

Einzelne Arbeitsschritte

Dabei werden die Textobjekte vom PDF- in das UTF8-Format umgewandelt. Anhand festgelegter Schwellenwerte für Stoppwörter, Größe, Wortanzahl usw. wird der Text analysiert. Ist mindestens ein Kriterium nicht erfüllt, kann keine Sachgruppe maschinell vergeben werden. Es folgt eine Fehlermeldung. Bei Dokumenten, die aus mehreren Dateien bestehen, wird die hinsichtlich der Wortanzahl größte Datei für die weitere Verarbeitung übergeben. Im nächsten Schritt ermittelt der Language Guesser – ein vorgeschaltetes Modul aus der Averbis Extraction Platform – die Sprache des Textes und erstellt als Ergebnis für jedes Dokument eine Sprachen-Rangliste. Die Sprache mit dem höchsten Rang wird als Dokumentsprache gewählt. Nach der Übergabe an den Averbis-Webservice (3) werden deutschsprachige und englischsprachige Dokumente jeweils einer eigenen Konfiguration der Averbis Extraction Platform zugeführt. Die zwei Konfigurationen wurden durch ein separates Training erstellt, die eine für deutschsprachige, die andere für englischsprachige monografische Netzpublikationen.

Konfidenzwerte

Die Averbis Extraction Platform weist den Netzpublikationen Sachgruppen mit den berechneten Konfidenzwerten zu. Liegt der höchste dieser Konfidenzwerte für einen Titel unter einem vorgegebenen Schwellenwert, wird zusätzlich ein Statusindikator hinzugefügt, damit diese Netzpublikationen intellektuell nachbearbeitet werden können. Die drei Sachgruppen mit den höchsten Konfidenzwerten und der Statusindikator werden über den Averbis-Webservice an den »DNBPetrusService« zurückgegeben. Dieser fügt über eine schreibende Schnittstelle (4) die Sachgruppe mit dem höchsten Konfidenzwert und den Statusindikator als Metadaten für diesen Titel in den bibliografischen Datensatz (CBS) ein.

Dafür war eine Anpassung der Metadatenfelder im internen PICA+ Datenformat und dessen Repräsentationsformat PICA3 im Katalogisierungsclient WinIBW notwendig. In Zukunft gibt es für die Sachgruppen folgende Erfassungsarten bei Netzpublikationen der Reihe O:

- Eine Sachgruppe kann bei der Ablieferung erstellt werden. Dieser Fall liegt vor, wenn sie entweder durch den Ablieferer im Webformular ausgewählt oder über Konkordanzen zu anderen mitgelieferten Systematiken erstellt wurde.
- Falls eine parallele Druckausgabe vorhanden ist, können bis zu drei Sachgruppen durch einen maschinellen Abgleich aus dieser Ausgabe übernommen werden.[3)]
- Eine Sachgruppe kann durch die maschinelle Sachgruppenvergabe vergeben werden.
- Bis zu drei Sachgruppen können im Zuge der Qualitätsmanagementprozesse durch intellektuelle Erschließung vergeben werden.

Die differenzierte Dokumentation aller im Erschließungsprozess entstehenden Sachgruppen mit Erfassungsart und Herkunft ermöglicht es, mit einem geringen Aufwand die Qualität der maschinell vergebenen Sachgruppen zu prüfen und zu verbessern. Dies dient im Produktionsprozess zur Überwachung des Klassifikators, zur Qualitätssicherung und zur Qualitätssteigerung. Das PICA3-Feld 5050 bzw. PICA+-Feld 045E wurde für die verschiedenen Erfassungsarten erweitert. Um die bisherige Struktur des Feldes nicht zu verändern, wurden die neuen Unterfelder an den bisherigen Inhalt angeschlossen. Die neuen Unterfelder sind in der folgenden Tabelle dargestellt.

Qualitätskontrolle

Das Unterfeld Erfassungsart $E bei Netzpublikationen der Reihe O gibt jeweils Auskunft, ob die Sachgruppen maschinell, durch Import, durch Datenübernahme aus einer parallelen Ausgabe oder intellektuell vergeben wurden. Das Unterfeld mit dem Konfidenzwert $K kann zwischen 1,000 (signalisiert eine hohe Konfidenz) und 0,000 (signalisiert eine niedrige Konfidenz) liegen. Das Unterfeld $H gibt Auskunft über die Herkunft der DDC-Sachgruppe. Das Datumsunterfeld $D zeigt das Übernahmedatum beim Import für übernommene Sachgruppen aus der parallelen Ausgabe, das Datum der maschinellen Erstellung für maschinell vergebene Sachgruppen und das Datum der Ablieferung für abgelieferte Sachgruppen an.

Zusätzlich wird das Feld 5050 bzw. 045E wiederholbar. So können in einem Titeldatensatz mehrere Sachgruppen verschiedener Erfassungsarten und unterschiedlicher Herkunft vorhanden sein. Was vielleicht unübersichtlich erscheint, bietet sehr gute Möglichkeiten, um Abfragen und maschinelle Auswertungen für das Qualitätsmanagement durchzuführen. Durch den Abgleich von intellektuell und maschinell vergebenen Sachgruppen kann im Produktionsbetrieb die Qualität des Klassifikators

PICA3	PICA+	Unterfeld	Trennzeichen	Bezeichnung
5050	045E			Sachgruppen der Deutschen Nationalbibliografie
		$E	"$E"	Erfassungsart der DDC-SG $Ei intellektuell vergeben $Ep aus Parallelausgabe übernommen $Em maschinell vergeben $Ea abgeliefert Neuer Suchschlüssel: efa
		$H	"$H"	Herkunft der DDC-Sachgruppe dnb durch die DNB maschinell erzeugte Sachgruppe mrc Ablieferung im Format MARC 21 onx Ablieferung im Format ONIX xmp Ablieferung im Format XmetadissPlus wbf Ablieferung über das Webformular Neuer Suchschlüssel: her
		$K	"$K"	Konfidenzwert der maschinell erstellten DDC-Sachgruppe zwischen 1,000 und 0,000. Neuer Suchschlüssel: kfw
		$D	"$D"	Datum (JJJJ-MM-TT) der maschinellen Erstellung der DDC-Sachgruppe/Übernahme beim Import. Neuer Suchschlüssel: sda
5051		$a		Statusindikator: Kennzeichnung mit qs, wenn die maschinell erstellte Sachgruppe unter einem definierten Schwellwert liegt. Neuer Suchschlüssel: sgf

Tabelle der Unterfelder im Pica+-Feld 045E

überwacht und optimiert werden. Auf die gleiche Weise kann das Mapping abgelieferter Systematiken auf die DDC-Sachgruppen bewertet und optimiert werden. Es soll auch geprüft werden, ob die abgelieferten Sachgruppen den maschinellen Erschließungsprozess bei schwierig zu vergebenden Sachgruppen künftig unterstützen können.

Vor diesem Hintergrund ist es sinnvoll, zunächst alle verschiedenartig erzeugten Sachgruppen im Datensatz zu verzeichnen. Löschungen können ggf. zu einem späteren Zeitpunkt gezielt durchgeführt werden, falls sich dies als sinnvoll erweist. Ein Auszug aus einem Datensatz für einen Titel der Reihe O ist in der folgenden Abb. beispielhaft dargestellt. Für die Anzeige im Portal, die Indexierung über die Suchmaschine und für die Datendienste wird zunächst immer nur die höchstrangige Sachgruppe verwendet.

```
...
3010 !130464511!Hartinger, Anselm [Hrsg.]

4000 Vergnügte Pleißenstadt:
Bach in Leipzig / Anselm Hartinger
...
5050 780$Ei$D2011-02-10

5050 780$Ep$D2011-03-05

5050 610$Em$K0,643$D2012-01-19

5050 330$Ea$D2010-03-04

5051 qs
...
```

Auszug aus einem Datensatz für einen Titel im PICA3-Format. Die Reihenfolge gibt die Rangfolge wieder.

Die Datenauslieferung erfolgt im MAB- und im MARC 21-Format. An der MAB-Schnittstelle werden keine Veränderungen mehr vorgenommen, da dieses Format von MARC 21 abgelöst werden wird. Wenn mehrere DDC-Sachgruppen vorhanden sind, wird die höchstrangige übernommen. Die Kennzeichnungen von Erfassungsart, Herkunft, Konfidenzwert sowie Datum werden nicht übernommen. Für MARC 21 wird die DNB eine Erweiterung des Formats vorschlagen, bei der künftig alle Unterfelder beim Datenaustausch berücksichtigt werden. Die Umsetzung der Formatänderung erfordert zunächst eine Abstimmung auf nationaler und eventuell auf internationaler Ebene. Eine Lösung soll möglichst im Laufe des Jahres 2012 implementiert werden. Dann werden mit der DDC-Sachgruppe auch alle Unterfelder über MARC 21 an die Datenbezieher transportiert. Ferner können dann auch alle vorhandenen DDC-Sachgruppen ausgeliefert werden.

Qualitätsmanagement als neue Daueraufgabe

Ein wichtiger Aspekt des Produktionsbetriebes ist das Qualitätsmanagement. Es dient der Überwachung, Sicherung und Verbesserung der Qualität. Durch einen kontinuierlichen maschinellen Vergleich der Sachgruppen aus maschineller und intellektueller Erschließung soll die Qualität der maschinellen Prozesse fortlaufend beobachtet werden, um bei Bedarf nachzusteuern. Die intellektuell erstellten DDC-Sachgruppen für den Abgleich entstehen bei der Übernahme aus parallelen Ausgaben, durch Mitarbeiter der Inhaltserschließung bei der Überprüfung willkürlich ausgewählter Stichproben und bei der Nachbearbeitung von Netzpublikationen, wenn der Konfidenzwert unter dem Schwellenwert liegt.

Beobachtung der maschinellen Prozesse

Die Sachgruppen der bearbeiteten Netzpublikationen werden jeden Monat in einer eigens dafür ent-

wickelten Auswertungsdatenbank evaluiert. Es werden die Qualitätsmaße Recall, Precision und das daraus resultierende harmonische Mittel, das so genannte F-Measure, berechnet. Mit Beginn des Produktionsbetriebs soll mindestens ein F-Measure von 0,7 über alle Sachgruppen erreicht werden. Mittelfristig angestrebt ist, mindestens einen F-Measure von 0,7 für jede einzelne Sachgruppe zu erreichen. Die Konfigurationen werden von Zeit zu Zeit neu trainiert. Mit der selektiven intellektuellen Nachbearbeitung von Netzpublikationen im Zuge des Qualitätsmanagements wird versucht, den Trainingskorpus für die Sachgruppen, die bisher nicht gut trainiert werden konnten, im Laufe der Zeit schrittweise anzureichern. Aber auch die Sachgruppen, für die schon jetzt ausreichend Trainingsmaterial vorhanden ist, müssen immer wieder mit neueren Beispielen ergänzt werden, um das Vokabular aktuell zu halten. Die bisherigen Tests zeigen deutliche Unterschiede in der Erschließungsqualität von Sachgruppe zu Sachgruppe. Einige Schwierigkeiten können voraussichtlich durch diesen Zuwachs an Trainingsdaten mittelfristig überwunden werden. Manche Sachgruppen allerdings lassen sich grundsätzlich nur schwer mit rein mathematischen Methoden klassifizieren. Es soll deshalb auch geprüft werden, ob das statistische Klassifizierungsverfahren durch Regeln ergänzt werden kann, um die Ergebnisse weiter zu verbessern.

In der jetzt implementierten Ausbaustufe der DDC-Sachgruppenvergabe wird immer nur eine Sachgruppe als Hauptsachgruppe in den Datensatz für Netzpublikationen übernommen. In der nächsten Ausbaustufe soll die maschinelle Zuweisung – wenn möglich – datenabhängig realisiert werden. In Anbetracht einer immer größeren Zahl interdisziplinär ausgerichteter Publikationen wird ein Modus angestrebt, der für Grenzfälle die Vergabe von bis zu drei Sachgruppen zulässt, wenn die eindeutige Zuordnung einer Sachgruppe nicht möglich ist. Hierzu soll experimentell ermittelt werden, über welche Kriterien sich die Vergabe einer flexiblen Zahl von Sachgruppen steuern lässt.

Zukünftig Vergabe von bis zu drei Sachgruppen geplant

Mit der zunehmenden Vielfalt und Heterogenität der gesammelten Netzpublikationen wächst auch die Notwendigkeit, die Geschäftsprozesse nach Objektgruppen zu differenzieren. Dabei geht es einerseits um eine Unterscheidung technischer Merkmale, beispielsweise die Berücksichtigung neuer Formate wie Epubs. Andererseits soll auch auf eine Differenzierung nach inhaltlichen Merkmalen hingearbeitet werden, um beispielsweise schwer klassifizierbare Objekte wie die Belletristik-Titel zunächst herauszufiltern und dann gesondert weiterzuverarbeiten.

Differenzierung der Geschäftsprozesse nach Objektgruppen

Mit dem Abschluss des PETRUS-Projektes ist für die maschinelle DDC-Sachgruppenvergabe der Übergang in den Produktionsbetrieb auf einer ersten Stufe erreicht. Weitere Entwicklungen sollen sich anschließen, sobald erste Erfahrungen aus der Praxis vorliegen.

Anschrift von Dr. Katrin Tomanek, Senior Software Engineer:
Averbis GmbH, Tennenbacher Str. 11, 79106 Freiburg,
E-Mail: katrin.tomanek@averbis.com

Anmerkungen

1 Schöning-Walter, Christa: Automatische Erschließungsverfahren für Netzpublikationen. In: Dialog mit Bibliotheken, 23 (2011) 1, S. 31 – 36.

2 Gömpel, Renate u. a.: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 20 - 22.

3 Beyer, Christian; Trunk, Daniela: Automatische Verfahren für die Formalerschließung im Projekt PETRUS. In: Dialog mit Bibliotheken, 23 (2011) 2, S. 5 – 10.