Christian Beyer, Daniela Trunk

Automatische Verfahren für die Formalerschließung im Projekt PETRUS

Die Deutsche Nationalbibliothek (DNB) erstellt für Netzpublikationen seit Anfang 2010 keine intellektuell per Autopsie und nach gültigem Regelwerk (RAK-WB bzw. RAK-NBM) erschlossen Titeldaten mehr.[1)] Da das stetig wachsende Kontingent an Netzpublikationen durch intellektuelle Bearbeitung nicht mehr zu bewältigen ist, wurde mit dem Projekt PETRUS damit begonnen, die Erschließungsprozesse für Netzpublikationen zu automatisieren.[2)]

Eines der Verfahren aus dem Projekt ist die automatische Verknüpfung von Personennamen in Titeldatensätzen mit der Personennamendatei (PND). Durch die Übernahme von Fremddaten soll der eigene intellektuelle Erschließungsaufwand minimiert und auf gezieltes Überprüfen zur Qualitätssicherung beschränkt werden. Eine intellektuelle Bearbeitung erfolgt nur dann, wenn einer Namensform mehrere Titel automatisch zugeordnet wurden.

Ein weiteres Verfahren ist der automatische Abgleich zur Erkennung von parallelen Print- und Onlineausgaben. Hierbei werden Datensätze maschinell mit den Datensätzen von parallel erschienenen Online- und Printpublikationen verknüpft und materialartunabhängige Informationen wie Inhaltserschließungsdaten und Normdatenverknüpfungen wechselseitig ausgetauscht. So können intellektuell erfasste Informationen auch an Datensätze übertragen werden, die nach dem Erschließungskonzept nicht mehr bearbeitet werden.

Automatische PND-Verknüpfungen in Titeldaten

Im Hinblick auf die Erschließung von Personen hat die Intensivierung und Förderung der Fremddatenübernahme dazu geführt, dass seit April 2010 alle Personennamen aus Metadaten zu Netzpublikationen sowie alle von der MVB Marketing- und Verlagsservice des Buchhandels GmbH für körperliche Medienwerke gelieferten Personennamen dauerhaft in den Titeldatensatz übernommen werden, auch solche, die nach Regelwerk eigentlich nicht zu erfassen wären. Bis zu dieser Entscheidung wurden nicht regelgerechte Personenangaben in der Katalogisierung manuell entfernt.

Ausgangspunkt: Fremddatenübernahme

Für Netzpublikationen werden bis zu drei Verfasser in die PICA3-Felder 3000-3002 umgesetzt, alle weiteren Personennamen werden inklusive ihrer Funktionsbezeichnung in das wiederholbare Feld 3019 übernommen. In Datensätzen zu körperlichen Medienwerken werden die Personenangaben, die in der Katalogisierung keine Haupt- oder Nebeneintragung erhalten haben, ebenfalls im PICA3-Feld 3019[3)] aufgehoben.

Die aus Fremddaten übernommenen Personennamen liegen lediglich als Textphrase vor. Um der gewünschten Erschließungsqualität zu entsprechen, sollen die Namen eindeutig Personen zugeordnet werden, da das Ziel der Formalerschließung darin besteht, ein Werk anhand seiner formalen Merkmale so zu beschreiben, dass es möglichst zweifelsfrei und eindeutig identifiziert werden kann. Wenn für die Suche im Katalog verlässliches Finden ermöglicht werden soll, um beispielsweise alle Werke eines bestimmten Verfassers nachzuweisen, müssen Personenangaben eindeutig der individuellen Person zugeordnet sein. Die Zuordnung der Titeldaten erfolgt über eine Verknüpfung zur PND.

Warum werden Personennamen erschlossen?

Die intellektuelle Ansetzung der Personennamen in der PND erfolgt gemäß den RAK-WB und der PND-Redaktionsanleitung. Des Weiteren gelten die Praxisregeln zu den RAK-WB.[4)] In der PND gibt es sowohl individualisierte Datensätze für jeweils genau eine Person (Satzart Tp) als auch nicht-individualisierte Datensätze für Namen, hinter denen sich mehrere Personen mit demselben Namen verbergen können (Satzart Tn). »Tp-Sätze ermöglichen durch die Angabe von Lebensdaten, Beruf oder Wirkungsbereich und weiterer individualisierender Merkmale die Identifizierung einer Person und bei ihrer Verwendung im Bibliothekskatalog die exakte Titelzuordnung, die es dem Benutzer ermöglicht,

die Literatur von oder auch über eine Person zu finden.«[5]

Nach Möglichkeit soll eine Individualisierung vorgenommen werden, wenn die Personalressourcen der Bibliothek dies zulassen bzw. wenn die Vorlage ausreichend Informationen für eine Individualisierung bietet.[6]

Bei der intellektuellen Formalerschließung werden zur Individualisierung der Personendatensätze die Angaben aus der Vorlage ermittelt, z. B. aus den biografischen Angaben auf dem Buchumschlag. Ein automatisches Erschließungsverfahren könnte ähnlich funktionieren, wenn in Fremddaten die individualisierenden Angaben zu Personen, wie Geburtsjahr oder Beruf, strukturiert in eigenen Feldern aufgelistet und maschinell interpretierbar wären. Über Metadaten werden allerdings nur zu einem sehr geringen Teil individualisierende Angaben strukturiert an die DNB geliefert. Freitextangaben, die teilweise auch Informationen zu Personen enthalten, sind zwar häufiger vorhanden, können automatisch aber nur mit hohem Aufwand und unsicherem Ergebnis ausgewertet werden.

Zunächst Implementierung eines einfachen Verfahrens auf Tn-Satz-Ebene

Im Projekt PETRUS wurde daher zunächst ein einfaches automatisches Erschließungsverfahren für Namen implementiert, bei dem individualisierende Angaben nicht berücksichtigt werden. Dies entspricht zwar nicht dem gewünschten Erschließungsstandard, ist aber trotzdem ein erster wirkungsvoller Automatisierungsschritt, da eine große Anzahl an PND-Sätzen, zu denen ein DNB-Bestand vorliegt, nur mit einem einzigen Titeldatensatz verknüpft sind. Ein Tn-Satz entspricht in diesen Fällen einer Quasi-Individualisierung.[7] Erst wenn zwei oder mehr Titel einer Namensform automatisch zugeordnet wurden, wird der Datensatz intellektuell um individualisierende Angaben angereichert. Damit steht ein Instrument zur Verfügung, das zur Steuerung der Geschäftsprozesse eingesetzt werden kann: Nicht mehr jeder Name muss intellektuell mit der PND verknüpft werden, sondern nur solche, zu denen mehrere Titeldatensätze vorliegen.

Die automatische Verknüpfung zu Tn-Sätzen erfolgt für alle aus Fremddaten übernommenen Namen in Titeldatensätzen, die als Textphrase vorliegen. Bei Netzpublikationen sind das alle Namen, bei körperlichen Medienwerken alle Namen, die nach Regelwerk intellektuell nicht erschlossen werden, aber aus Fremddaten in den Titeldatensatz übernommen wurden.[8] Die Namensphrasen in den Titeldaten werden mit den Ansetzungsformen der PND-Einträge abgeglichen und automatisch in Verknüpfungen zur PND umgewandelt, wenn ein Tn-Satz der Phrase entspricht. Ist kein Tn-Satz in der PND vorhanden, wird automatisch ein Tn-Satz mit dem Katalogisierungslevel 7 (»maschinell aus Metadaten erstellt«) angelegt und zu diesem verknüpft. Die automatisch erzeugten Relationen werden in den Verfasserfeldern bzw. im Feld für weitere beteiligte Personen (PICA3-Eingabeformat 3000-3002 und 3019) mit »|m|« codiert.

Wie funktioniert der Abgleich?

Automatisch erzeugte Verknüpfung im Titeldatensatz:

3000 |m|!1012002802!Warning, Martina

Automatisch erstellter Tn-Satz:
005 Tn7
011 /f
012 /v
100 Warning, Martina

In der DNB werden nicht individualisierte PND-Sätze, denen automatisch mehrere Titel zugeordnet wurden, intellektuell überarbeitet, d. h., sie werden individualisiert und die Relationen überprüft. Alle anderen PND-Anwender sollen Tn7-Sätze unbearbeitet im PND-Bestand belassen und auf Hinweise zu Dubletten oder Individualisierungsmöglichkeiten verzichten; es sei denn, es wird ein Tn-Satz dieses Namens benötigt. In diesem Fall darf ein Tn7-Satz von jedem PND-Teilnehmer auf Level 3 angehoben und nachgenutzt werden.[9]

Ideen zur Entwicklung eines automatischen Erschließungsverfahrens für Tp-Sätze

Für die Zukunft soll ein Verfahren entwickelt werden, das die automatische Personenerschließung auf Tp-Satz-Ebene auch ohne intellektuelles Eingreifen ermöglicht und die vielen neu hinzukommenden Tn-Sätze[10] bereinigt. Zunächst wird eine Variante des im zweiten Abschnitt beschriebenen Abgleichs zur Erkennung von parallelen Ausgaben getestet. Titeldatensätze sollen auf Ähnlichkeit überprüft werden, um bereits vorhandene Tp-Ver-

knüpfungen aus intellektuell erschlossenen Datensätzen übernehmen zu können.

Das Match- und Merge-Verfahren könnte nach dem Vorbild des VIAF (Virtual International Authority File)[11] weiterentwickelt werden, indem für Normsätze individualisierende Merkmale aus dem semantischen Umfeld berücksichtigt werden. Die Tn- und Tp-Sätze, zu denen in der DNB ein Bestand vorliegt, würden dabei um Elemente aus den Titeldaten angereichert und anschließend über einen automatischen Abgleich zusammengeführt werden. Dazu müssten die Tn-Sätze zunächst um die Anzahl der verknüpften Titeldatensätze vervielfältigt und beim folgenden Abgleich, bei dem die individualisierenden Merkmale berücksichtigt würden, ab einem bestimmten Schwellenwert zusammengeführt werden. Für den Bestand in der DNB würde es dann – zumindest virtuell – keine Tn-Sätze mehr geben, da alle Normdatensätze automatisch zu individualisierten Datensätzen angereichert sein würden.

Automatisches Abgleichverfahren zur Erkennung von parallelen Print- und Onlineausgaben

Die Verknüpfung von Datensätzen paralleler Ausgaben ermöglicht dem Nutzer einen schnellen Zugriff von der Print- auf die Online-Publikation. Darüber hinaus führt die gegenseitige Übernahme von Erschließungsdaten zur Vereinheitlichung der Sucheinstiege und zu höheren Trefferquoten bei Suchabfragen im Katalog. Damit wird nicht nur dem Bestandsschutz gedruckter Ausgaben Rechnung getragen, sondern auch die Benutzerfreundlichkeit im Portal und das Finden der verfügbaren Netzpublikationen verbessert.

Daten und Werkzeuge

Die Datengrundlage für den regelmäßig stattfindenden Abgleich bildet der gesamte ILTIS-Bestand an Print- und Online-Monografien. Dabei spielt die regelwerksgemäße Differenzierung zwischen Sekundärausgaben (im vorliegenden Fall Digitalisate) und »echten« Parallelausgaben keine Rolle. Sobald eine Online-Ausgabe und die zugehörige Printausgabe vorliegen, werden die Datensätze durch das automatische Verfahren verknüpft und Daten können ggf. untereinander ausgetauscht werden. Die Zeitschriftendatensätze sind vom Verfahren ausgenommen, da diese nach dem Zeitschriftenregelwerk ZETA bereits parallele Verknüpfungen zwischen Ausgaben in unterschiedlichen Materialarten (darunter auch Online-Ausgaben) enthalten.

Prinzipiell werden (fast) alle im Portal angezeigten monografischen Print- und Online-Datensätze in den Abgleich mit einbezogen. Dies betrifft sowohl die von der DNB intellektuell bearbeiteten als auch (noch) unbearbeitete (Fremd-)Datensätze, die durch potenzielle automatische Verknüpfung und Datenaustausch eine frühestmögliche Aufwertung erfahren sollen. Dennoch darf die Erkennung paralleler Ausgaben möglichst nicht durch eine abweichende Datensatzstruktur bzw. durch unterschiedliche Schreibweisen oder Rechtschreibfehler verhindert werden. Deshalb wurden im Anschluss an die Festlegung der Kriterien für den Abgleich mit der PICA-Match-und-Merge-Software Tests durchgeführt, um eine höchstmögliche Qualität des Verfahrens zu erreichen.

Erster und zweiter Abgleich

Zuerst wird nach einem passenden Pendant gesucht. In Datensätzen, die potenziell ein paralleles Datensatzpaar darstellen, werden bestimmte Datenfelder auf ihre Übereinstimmung hin überprüft. Dadurch wird sichergestellt, dass es sich bei beiden Datensätzen nicht nur um dasselbe Werk handelt (Titel, Personen, Körperschaften), sondern dass auch die übrigen Kriterien für eine parallele Ausgabe erfüllt sind (Sprache, Auflage, Verlag). Die Software berechnet für jede der genannten Kategorien Ähnlichkeitswerte, die jeweils mit einer festgelegten Gewichtung zur Bildung eines sogenannten Matchwertes herangezogen werden. Ist der Matchwert hoch genug, gilt das Datensatzpaar als gefunden.

Felder, die in einem Datensatz nicht besetzt sind, können nicht abgeglichen und gewichtet werden und haben demzufolge auf die Berechnung des Matchwertes keinen Einfluss. So kann es sein, dass aufgrund fehlender Feldinhalte die Übereinstimmungswerte anderer, vorhandener Felder höher ins Gewicht fallen und den Matchwert stärker als geplant beeinflussen. Genauso ist es möglich, dass eine sehr hohe Ähnlichkeit in stark gewichteten Feldern zur Bestimmung eines parallelen Datensatzpaars führt, auch wenn gleichzeitig Inhalte in einem einzelnen geringer gewichteten (aber nicht unwichtigen!) Feld zu 100 % voneinander abweichen.

Da ein fehlerfreies Zusammenspiel der Gewichtungen ohnehin nicht in allen Fällen gewährleistet werden kann, und auch die Implementierung von Ausschlusskriterien im ersten Abgleich nicht umzusetzen ist, werden die gefundenen Datensatzpaare in einen zweiten Abgleich geschickt. Dadurch werden sie erneut auf ihre Parallelität hin untersucht, indem die Angaben zu Ausgabebezeichnung und Verlag einem strengeren Abgleich unterzogen werden. Außerdem wird sichergestellt, dass nicht eine echte Hochschulschrift und eine spätere Verlagsausgabe derselben parallel miteinander verknüpft werden, da beide Versionen eigene parallele Online- oder Druckausgaben haben können. Schwierigkeiten macht in einigen Fällen der Abgleich der Verlagsangaben, da vor allem der Verlagsname häufig in unterschiedlicher Ansetzung vorliegt und der Verlagsort hinzugezogen werden muss.

Wurden die zwei Datensätze schließlich als paralleles Datensatzpaar erkannt, werden beide über das für horizontale Verknüpfungen vorgesehene Feld als Druck- und Online-Ausgabe miteinander verknüpft.

Parallele Verknüpfung und Codierung

Darüber hinaus werden beide Datensätze mit dem Code »pb« und zusätzlich mit einem kurzen, erklärenden Kommentar im Bemerkungsfeld versehen.

Konnten die beiden Datensätze im zweiten Abgleich nicht als parallele Ausgaben bestätigt werden (weil es sich um verschiedene Auflagen, unterschiedliche Verlagsausgaben oder eine »echte« Hochschulprüfungsarbeit und deren Verlagsausgabe handelt), findet keine Verknüpfung statt. Dennoch ist auch für diese Datensätze der anschließend beschriebene Datenaustausch möglich, da beide Datensätze zwar nicht parallel zueinander sind, aufgrund des ersten Abgleichs jedoch sichergestellt ist, dass es sich um unterschiedliche Ausgaben ein- und desselben Werkes handelt. Dies bedeutet, dass z. B. in beiden Datensätzen verzeichnete, übereinstimmende Personennamen identisch sind und in der Regel auch die Inhaltserschließungsdaten eines Datensatzes in dem Pendant ihre Gültigkeit besitzen. Deshalb werden diese Datensatzpaare mit dem Code »pn« gekennzeichnet und erhalten ebenfalls einen erklärenden Kommentar im Bemerkungsfeld.

Durch die Vergabe der beiden Codes können alle Datensätze, die das weitere Verfahren durchlaufen, problemlos selektiert werden.

Austausch von PND- und GKD-Verknüpfungen sowie von Inhaltserschließungsdaten

Bei allen im ersten Abgleich gefundenen Datensatzpaaren – unabhängig davon, ob sie nach dem zweiten Abgleich parallelverknüpft wurden oder nicht – werden die Normdatenverknüpfungen und Inhaltserschließungsdaten miteinander verglichen.

Dabei wird erkannt, ob in einem Datensatz eine Verknüpfung zu einem Personen- oder Körperschaftsnormdatensatz vorhanden ist, während im zugehörigen Pendant für dieselbe Person oder Körperschaft nur eine Textphrase steht. In diesen Fällen wird die Verknüpfung zum Normdatensatz in den parallelen Datensatz übernommen. Bei den Personenfeldern ist der Abgleich der verbalen Einträge etwas weniger streng als in den Körperschaftsfeldern, um ggf. die häufig etwas abweichende Ansetzung (z. B. durch abgekürzte Vornamen) auszugleichen.

Außerdem können fremderstellte Metadaten für Netzpublikationen – sofern diese selbst noch keine (auch keine verbalen!) Körperschaftsinformationen enthalten – mit den Körperschaftsangaben aus einem im zweiten Abgleich als parallele Ausgabe bestätigten Print-Datensatz angereichert werden.

Auch bei den Inhaltserschließungsdaten findet ein Austausch zwischen den beiden Datensätzen statt, falls sich die Druckausgabe zu dem Zeitpunkt nicht im Bearbeitungsstatus befindet. RSWK-Ketten bzw. -Folgen, DDC-Notationen und alle weiteren Inhaltserschließungsinformationen werden wechselseitig übernommen, sofern sie im jeweiligen Pendant nicht bereits zu finden sind. Schlagwörter, die die Materialart beschreiben, sind vom Austausch ausgeschlossen.

Abgleich ist für Netz- und Printpublikationen implementiert

Nach Abschluss der Testphase wurden zwei automatische Geschäftsprozesse implementiert. Im ersten Prozess wird im 24-Stunden-Rhythmus für alle Netzpublikationen, die über die eingeführten Ablieferungsverfahren als Fremddaten in das System der DNB importiert werden, nach einem passenden parallelen Print-Datensatz gesucht. Die Verknüpfungsquote für dieses Verfahren liegt nach vier Monaten Laufzeit derzeit bei etwa 25 %. Darüber hinaus haben alle Netzpublikationen, die bereits vor der Produktivnahme dieser Prozesse im Katalog der DNB verzeichnet waren, einmalig rückwirkend das Verfahren durchlaufen. Dabei konnte für mehr als 100.000 Netzpublikationen (rund 40 %) eine Verknüpfung zur parallelen Print-Ausgabe realisiert werden.

Der zweite Prozess wird durch die Printdatensätze initiiert, die intellektuell erschlossen und für die Anzeige in der Deutschen Nationalbibliografie freigegeben wurden. Sollten die Printdatensätze zu diesem Zeitpunkt bereits die Verknüpfung zur parallelen Online-Ausgabe besitzen, wird die Suche nach einer solchen übersprungen. In diesem Fall werden lediglich erneut Inhaltserschließungsdaten und Normdatenverknüpfungen miteinander abgeglichen. Dadurch wird zum einen gewährleistet, dass intellektuell erstellte Informationen auch nachträglich an die Datensätze paralleler Ausgaben übertragen werden können (falls sich der Printsatz zum Zeitpunkt des ersten Prozesses noch im Bearbeitungsstatus befand). Zum anderen erhalten parallele Datensätze durch dieses zweite Verfahren »in entgegengesetzter Richtung« auch dann die Chance zu automatischer Verknüpfung und Datenaustausch, wenn die Printausgabe erst nach der Online-Ausgabe in der DNB eintrifft.

Fazit

In der Anfangsphase, in der alle verknüpften Datensätze kontrolliert wurden, lag die Quote der ermittelten Fehler bei etwa 2 %. Die meisten Fehler sind auf Unterschiede bei den Erschließungsstandards zurückzuführen. Im Geschäftsgang der DNB werden die verknüpften Datensätze stichprobenhaft überprüft und ggf. nachgebessert. Fallen bei der normalen Bearbeitung unpassende oder falsche Verknüpfungen auf, werden die Datensätze ebenfalls wieder bereinigt. Eventuelle systematische Fehler werden gesammelt und dokumentiert, um Rückschlüsse auf die Qualität der automatischen Erschließung ziehen und evtl. notwendige Korrekturen anstoßen zu können. Technisch laufen beide Prozesse seit ihrer Inbetriebnahme stabil.

Anmerkungen

1 Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 20 - 22.
<http://www.d-nb.de/service/pdf/dialog_2010_1_veraenderungen_erschliessungskonzept.pdf>

2 Schöning-Walter, Christa: Automatische Erschließungsverfahren für Netzpublikationen. In: Dialog mit Bibliotheken, 23 (2011) 1, S. 31 - 36.
<http://files.d-nb.de/pdf/petrus/petrus_dialog_2011_1.pdf>

3 Das Feld 3019 ist im Katalog der DNB suchbar, wird aber derzeit nicht an Datendienstbezieher ausgeliefert.

4 vgl. PND-Redaktionsanleitung, Teil 1: Zuständigkeiten, Aufgaben und Befugnisse in der PND, Abschnitt »Satzarten«. Stand: November 2010.
<http://www.d-nb.de/standardisierung/pdf/pnd_1.pdf.>
Link zu den Praxisregeln zu RAK-WB § 311:
<http://www.d-nb.de/standardisierung/pdf/praxisregel_individualisierung_311.pdf>

5 ebd.

6 vgl. PND-Redaktionsanleitung, Teil 3: Individualisierungsrichtlinie für die PND, Abschnitt »Gemeinsame Individualisierungsrichtlinie der PND-Anwender«. Stand: Mai 2010
<http://www.d-nb.de/standardisierung/pdf/pnd_3.pdf>

7 Stand 29.07.2009: 3.411.376 PND-Sätze, davon 1.431.523 nur mit einem Titel verknüpft (ca. 42 %).

8 Das Verfahren wird für Netzpublikationen direkt nach dem Import der Metadaten eingesetzt, körperliche Medienwerke durchlaufen zunächst die intellektuelle Katalogisierung.

9 Zu diesem Abschnitt vgl. PND-Mitteilung Nummer 10, S. 3.
<http://www.d-nb.de/standardisierung/pdf/nr_10_info_pnd.pdf>

10 Bisher waren in der PND rund die Hälfte der Personennamen individualisiert, aber durch die neu entstehenden Tn7-Sätze und durch Einspielungen regionaler Normdatenbestände (Tn6-Sätze), steigt die Anzahl der nicht individualisierten Sätze an. Stand am 06.07.2011: 2.055.919 Tp-Sätze und 2.928.286 Tn-Sätze, davon 6.520 Tn7-Sätze und 1.286.074 Tn6-Sätze.

11 Link zum Virtual International Authority File: <http://viaf.org/>