Ulrike Junger, Ute Schwens

Die inhaltliche Erschließung des schriftlichen kulturellen Erbes auf dem Weg in die Zukunft

Automatische Vergabe von Schlagwörtern in der Deutschen Nationalbibliothek

Hintergrund

Wir leben im 21. Jahrhundert, und vieles, was vor hundert und noch vor fünfzig Jahren als Science Fiction abgetan worden wäre, ist mittlerweile Realität. Raumsonden fliegen zum Mars, machen dort Experimente und liefern Daten zur Erde zurück. Roboter werden für Routineaufgaben eingesetzt, zum Beispiel in der Industrie oder in der Medizin. Digitalisierung, künstliche Intelligenz und automatisierte Verfahren sind kaum mehr aus unserem Alltag wegzudenken. Grundlage vieler Prozesse sind lernende Algorithmen.

Neue Möglichkeiten durch Digitalisierung

Die fortschreitende digitale Transformation ist global und umfasst alle Lebens- und Arbeitsbereiche: Wirtschaft, Gesellschaft und Politik. Sie eröffnet neue Möglichkeiten, von denen auch Bibliotheken profitieren. Der starke Anstieg digitaler Publikationen, die einen wichtigen und prozentual immer größer werdenden Teil des Kulturerbes darstellen, sollte für Bibliotheken Anlass sein, diese Möglichkeiten aktiv aufzugreifen und einzusetzen. Die Auswertbarkeit digitaler Inhalte, beispielsweise durch Text- and Data-Mining (TDM), und die Entwicklung technischer Verfahren, mittels derer Inhalte miteinander vernetzt und semantisch in Beziehung gesetzt werden können, bieten Raum, auch bibliothekarische Erschließungsverfahren neu zu denken.

Prozesse bei der Erschließung verbessern

Daher beschäftigt sich die Deutsche Nationalbibliothek (DNB) seit einigen Jahren mit der Frage, wie sich die Prozesse bei der Erschließung von Medienwerken verbessern und maschinell unterstützen lassen.[1, 2] Sie steht dabei im regelmäßigen kollegialen Austausch mit anderen Bibliotheken, die sich ebenfalls aktiv mit dieser Fragestellung befassen, sowie mit europäischen Nationalbibliotheken, die ihrerseits Interesse an dem Thema und den Erfahrungen der DNB haben. Als Nationalbibliothek mit umfangreichen Beständen an digitalen Publikationen hat die DNB auch Expertise bei der digitalen Langzeitarchivierung aufgebaut und ist im Netzwerk ihrer Partner als kompetente Gesprächspartnerin geschätzt.

Funktion und Status der verbalen Inhaltserschließung in der Deutschen Nationalbibliothek

Inhaltserschließung – ob verbal oder klassifikatorisch – ermöglicht es, umfangreiche heterogene Datenbestände zu strukturieren und erlaubt dadurch sachliche und themenbezogene Recherchen. Auch wenn sie angesichts der Möglichkeiten von Volltextsuchen und Suchmaschinentechnologie manchmal in Frage gestellt wird, leistet sie einen wichtigen Beitrag, wenn es darum geht, den Zugang zu Medienwerken und dem in ihnen gespeicherten Wissen zu erlangen und zu verbessern. Das zeigen auch kommerzielle Verkaufsplattformen, wie zum Beispiel Amazon oder eBay, die ganz selbstverständlich thematische Suchen als Option anbieten. Bisher ist die intellektuelle verbale Inhaltserschließung in der DNB auf einen Teil des gesamten Neuzugangs an Medienwerken beschränkt: Mit Schlagwörtern aus dem Vokabular der Gemeinsamen Normdatei (GND) werden – mit einigen Einschränkungen – gedruckte Bücher und Zeitschriftentitel der Reihe A der Deutschen Nationalbibliografie (Monografien und Periodika des Verlagsbuchhandels) erschlossen.[3] 2016 waren dies rund 74.000 Medienwerke.[4]

Netzpublikationen hingegen wurden seit 2010 zunächst nicht inhaltlich erschlossen. Stattdessen nutzte die DNB für die Nachweise im Katalog die

Metadaten der Ablieferer nach und baute maschinelle Verfahren auf, um die Datensätze unter anderem mit inhaltserschließenden Daten anzureichern. Digitale Hochschulschriften und wissenschaftliche Verlagspublikationen erhalten mittlerweile auf diese Weise DDC-Sachgruppen und Schlagwörter.[5, 6] Die Verfahren werden laufend weiterentwickelt und optimiert sowie jetzt schrittweise auf andere, auch gedruckte, Publikationen ausgeweitet.

Weiterentwicklung erprobter Verfahren

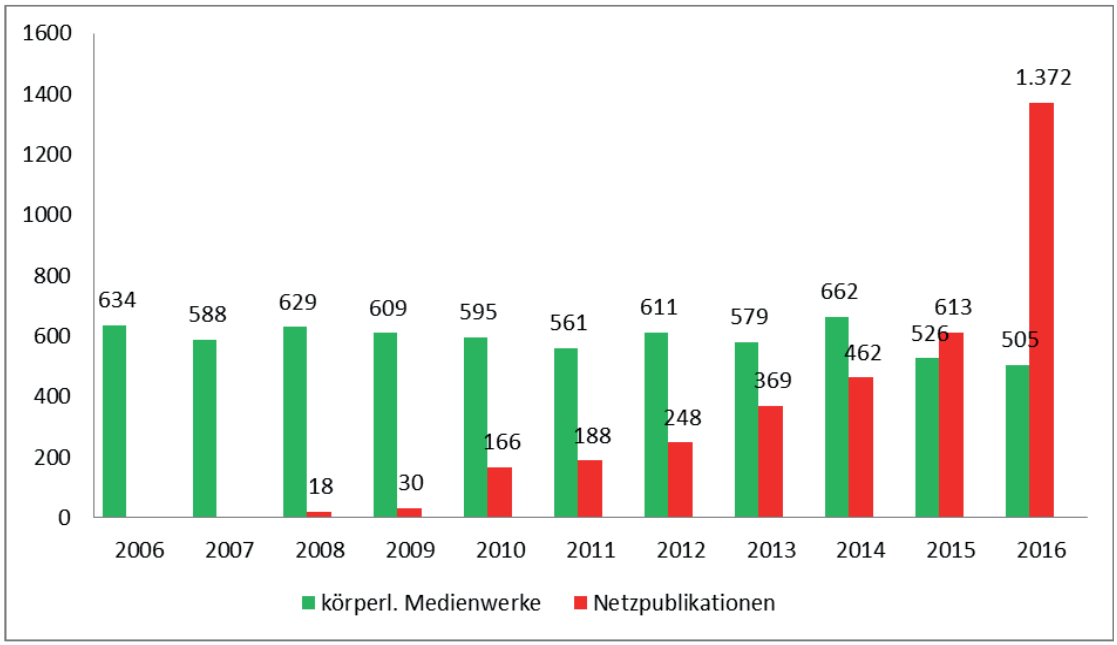
Der Zugang an Netzpublikationen weist derzeit enorme Steigerungsraten auf: So hat er sich 2016 im Vergleich zum Vorjahr auf 1,3 Millionen verdoppelt, für die kommenden fünf Jahre erwartet die DNB eine Versechsfachung des heutigen Bestandes auf dann 26 Millionen Netzpublikationen. Andererseits bleibt der Zugang an gedruckten und anderen körperlichen Medienwerken relativ konstant. Dies bedeutet, dass der Anteil intellektuell erstellter Inhaltserschließung in Relation zu maschinell erstellten Daten perspektivisch immer kleiner werden wird.

Prozentual gesehen nimmt der Anteil intellektuell erstellter Daten ab

Inhaltserschließung für alle Publikationen

Angesichts der enormen Flut an digitalen Publikationen und des Wunsches der Nutzerinnen und Nutzer, alle Werke auch inhaltlich suchen und finden zu können, ist es aus Sicht der DNB der einzig gangbare Weg, Verfahren zu entwickeln und Prozessketten aufzubauen, die möglichst vielen Publikationen über maschinelle Verfahren gute, das heißt im Sinne der Auffindbarkeit nützliche Inhaltserschließungsdaten mitgeben. Dabei beabsichtigt sie, perspektivisch alle eingehenden Publikationen einheitlich inhaltlich zu erschließen und mit thematischen Zugriffspunkten auszustatten. Gleichzeitig sollen die Ergebnisse im Hinblick auf das Retrieval maschineller und intellektueller Erschließungsdaten miteinander kompatibel sein. Deshalb werden maschinelle Verfahren zur Optimierung der Erschließungsprozesse und zur Erzeugung inhaltserschließender Daten auch für körperliche Medienwerke weiterentwickelt und auf diese ausgeweitet. Die DNB ist davon überzeugt, dass die umfassende Erfüllung des gesetzlichen Auftrags zur nationalbibliografischen Verzeichnung nur mit Hilfe solcher Verfahren gelingen kann.

Einheitlichkeit und Kompatibilität



Entwicklung des Zugangs in der DNB (in Tausend) seit 2006

Betrachtet man vor diesem Hintergrund die Gesamtheit der maschinenlesbar vorliegenden Katalogdaten, so zeigt sich, dass der Anteil der bibliografischen Nachweise mit verbaler Inhaltserschließung derzeit nur bei rund 20 Prozent liegt.

Durch den Einsatz maschineller Inhaltserschließungsverfahren als ein Anwendungsfall von TDM soll künftig eine qualitativ gute Inhaltserschließung aller Publikationen erfolgen. Dabei sollen die Verfahren erst dann und nur in den Fachgebieten eingesetzt werden, in denen gute Ergebnisse erzielt werden können, und wenn die DNB davon überzeugt ist, dass die so erzeugten Daten ein entsprechendes Qualitätsniveau erreicht haben und die Fehlerquote gering ist. Die DNB baut daher ein Qualitätsmanagement auf, das auch die intellektuelle Kontrolle von Verfahren und Ergebnissen beinhaltet, und wird das Vorgehen für Datenbezieher sowie Nutzerinnen und Nutzer transparent machen.

Qualitätsmanagement und Transparenz

Selbst wenn eine intellektuelle Erschließung im Einzelfall ein besseres Ergebnis liefern könnte als die automatische, so stellt doch die Inhaltserschließung aller Medienwerke einen Gewinn für die Recherche der Nutzerinnen und Nutzer dar. Thematische Recherchen werden insgesamt verbessert, weil ein viel größerer Anteil an Publikationen erschlossen werden kann. Es wird möglich sein, ein Desiderat vieler wissenschaftlich und insbesondere naturwissenschaftlich arbeitender Benutzerinnen und Benutzer zu erfüllen, nämlich die thematische Suche auch nach Zeitschriftenartikeln und Conference-Proceedings.

Thematische Suche über alle Publikationen möglich

Ein weiterer zentraler Vorteil der Anwendung maschineller Verfahren besteht darin, dass man Erschließungsvorgänge als dynamische Prozesse betrachten kann, bei denen die Daten immer wieder verändert und verbessert werden, und sie nicht mehr als einmalig und abgeschlossen angesehen werden müssen. Es ist also möglich, einen bereits bearbeiteten Bestand nach einer Optimierung der maschinellen Verfahren erneut zu erschließen und damit besser zugänglich zu machen. Alle Daten werden den anderen Bibliotheken und weiteren Daten-Nutzern frei und kostenlos zur Verfügung gestellt. Die technischen Voraussetzungen für die Verarbeitung der Datenströme bei einem dynamischen Verfahren müssen in Zusammenarbeit mit den Daten-Nutzern erarbeitet werden. Notwendige Anpassungen der Schnittstellen und der Infrastruktur wird die DNB wie bisher mit den Bibliotheksverbünden in den gemeinsamen Gremien besprechen.

Erschließung als dynamischer Vorgang

Die Gemeinsame Normdatei als Grundpfeiler der verbalen Inhaltserschließung

Sowohl die intellektuelle als auch die maschinelle Inhaltserschließung erfolgen mit dem Vokabular der Gemeinsamen Normdatei (GND). Eine intellektuelle Erschließung wird künftig weiterhin für Publikationen derjenigen Fachgebiete vorgenommen, für die automatische Verfahren (noch) keine zufriedenstellenden Ergebnisse liefern. Eine wichtige Voraussetzung für gute Ergebnisse bei der maschinellen Erschließung ist die Aktualität des eingesetzten Vokabulars. Deshalb wird ein GND-Pflegetool entwickelt, das im Laufe des Prozesses der maschinellen Erschließung Schlagwortkandidaten für die GND vorschlägt, die dann in die GND eingearbeitet werden. Dafür ist die Expertise der Bibliothekarinnen und Bibliothekare auch weiterhin unerlässlich. Die intellektuelle Erschließungsarbeit wird sich künftig aber je nach fachlicher Anforderung und der Verfügbarkeit geeigneter maschineller Verfahren verändern. Tätigkeitsschwerpunkte werden sich von der Einzelfallbearbeitung perspektivisch in Richtung einer verstärkten Normdatenpflege und zum Qualitäts- und Datenmanagement verschieben.

Neue Schwerpunkte bei der intellektuellen Erschließung

Nächste Schritte der Umsetzung

Da die Methoden und Verfahren der maschinellen Inhaltserschließung heute weiter entwickelt sind als zu Beginn ihrer Überlegungen dazu, macht die DNB nun einen nächsten Schritt in deren Anwendung und erschließt seit dem 1. September 2017 auch gedruckte Medienwerke maschinell. Die deutsch- und englischsprachigen Publikationen der Reihen B und H werden mit DDC-Sachgruppen erschlossen, darüber hinaus mit maschinell erzeugten Schlagwörtern aus der GND. Die entsprechenden Daten werden voraussichtlich ab Mitte September 2017 bereitgestellt und sind mit einem Herkunftskennzeichen codiert.[7]

Damit setzt die DNB den Weg zu einer – perspektivisch – alle Publikationen umfassenden inhaltlichen Erschließung ihres Zugangs fort, unabhängig

davon, ob ein Werk digital oder gedruckt vorliegt. Sie wird die maschinellen Erschließungsverfahren dem technischen Fortschritt folgend ständig anpassen und sie durch intellektuelle Qualitätskontrollen und Steuerung intensiv begleiten, um für möglichst alle Publikationen eine gute, das bedeutet, für Search and Retrieval nützliche, Inhaltserschließung zu ermöglichen.

Anmerkungen

1 Dazu wurde von 2009 bis 2011 das Projekt PETRUS (Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek) als Einstieg in die Nutzung maschineller Erschließungsverfahren durchgeführt: <http://www.dnb.de/DE/Wir/Projekte/Archiv/petrus.html> (zuletzt besucht am 15. August 2017)

2 Renate Gömpel, Ulrike Junger, Elisabeth Niggemann: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 20–22; online unter: <http://nbn-resolving.de/urn:nbn:de:101-2011012858>

3 Die Medienwerke der Reihen A werden außerdem klassifikatorisch mit Notationen der Dewey Dezimalklassifikation (DDC), die Medienwerke der Reihen B (Monografien und Periodika außerhalb des Verlagsbuchhandels) und H (Hochschulschriften) ausschließlich klassifikatorisch erschlossen. Darüber hinaus werden alle Publikationen mit DDC-Sachgruppen versehen; diese dienen zur fachlichen Gliederung der Deutschen Nationalbibliografie.

4 Deutsche Nationalbibliothek. Jahresbericht 2016, S. 27; online unter: <http://nbn-resolving.de/urn:nbn:de:101-2017061405>

5 Für rund 37.000 Netzpublikationen mit Erscheinungsjahr 2016 wurden maschinell Schlagwörter vergeben.

6 Ausführlich beschrieben sind die Ausgangslage und das Erschließungskonzept der Deutschen Nationalbibliothek zum Beispiel in Ulrike Junger: Quo vadis Inhaltserschließung der Deutschen Nationalbibliothek. Herausforderungen und Perspektiven. o-bib 2015/1; online unter: <https://doi.org/10.5282/o-bib/2015H1S15-26>

7 Die bisherige klassifikatorische Tiefenerschließung mit der Dewey-Dezimalklassifikation (DDC) wird für die Reihen B und H ab 1. September 2017 eingestellt. Sie soll durch DDC-Kurznotationen ersetzt werden.