Maschinelle Indexierung am Beispiel der DNB – Analyse und Entwicklungsmöglichkeiten

Heidrun Wiesenmüller, Hochschule der Medien Stuttgart

Zusammenfassung:
Der Beitrag untersucht die Ergebnisse des bei der Deutschen Nationalbibliothek (DNB) eingesetzten Verfahrens zur automatischen Vergabe von Schlagwörtern. Seit 2017 kommt dieses auch bei Printausgaben der Reihen B und H der Deutschen Nationalbibliografie zum Einsatz. Die zentralen Problembereiche werden dargestellt und an Beispielen illustriert – beispielsweise dass nicht alle im Inhaltsverzeichnis vorkommenden Wörter tatsächlich thematische Aspekte ausdrücken und dass die Software sehr häufig Körperschaften und andere „Named entities" nicht erkennt. Die maschinell generierten Ergebnisse sind derzeit sehr unbefriedigend. Es werden Überlegungen für mögliche Verbesserungen und sinnvolle Strategien angestellt.

Summary:
The paper discusses the results of the German National Library's process for machine indexing with subject headings. Since 2017, this has also been applied to print publications from the series B and H of the German National Bibliography. The most problematical areas are described and illustrated with real examples, e.g., that not all words appearing in the table of contents are meant to describe topics, and that the software very often does not recognize corporate bodies and other named entities. At present, the machine generated results are highly unsatisfactory. A number of possible improvements are discussed as well as a sensible overall strategy.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H4S141-153
Autorenidentifikation: Wiesenmüller, Heidrun: GND 122087801
ORCID: http://orcid.org/0000-0002-9817-5292
Schlagwörter: Inhaltserschließung; Automatische Indexierung; Deutsche Nationalbibliothek

1. Maschinelle Indexierung bei der DNB

Für Aufsehen sorgte vor einiger Zeit die Ankündigung der Deutschen Nationalbibliothek (DNB), die Sacherschließung in den nächsten Jahren weitestgehend auf automatische Methoden umstellen zu wollen.[1] Im Zuge dieser Entwicklung kommt seit 2017 ein Verfahren zur automatischen Vergabe von Schlagwörtern, das bei Netzpublikationen schon seit 2014 angewendet wird, auch bei Printausgaben der Reihen B und H der Deutschen Nationalbibliografie zum Einsatz. Es beruht auf einer kommerziellen Software der Firma Averbis und ist kein „lernendes" Verfahren.[2]

1 Vgl. Wiesenmüller, Heidrun: Das neue Sacherschließungskonzept der DNB in der FAZ, Blog Basiswissen RDA, 02.08.2017, <https://www.basiswissen-rda.de/neues-sacherschliessungskonzept-faz/>, Stand: 21.07.2018.

2 Zum angewendeten Verfahren vgl. Mödden, Elisabeth; Schöning-Walter, Christa; Uhlmann, Sandro: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek. Breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse, in: BuB 70 (1), 2018, S. 30-35. Online:

Neben den Titeldaten sowie – bei Online-Ressourcen – Teilen des Volltexts stellen die Inhaltsverzeichnisse die Grundlage für die Ermittlung der Schlagwörter dar. Diese werden bei Printpublikationen ohnehin eingescannt und es erscheint plausibel, dass sie den Inhalt einer Ressource gut repräsentieren. Jedoch hat, wie noch zu zeigen ist, die Textgattung „Inhaltsverzeichnis" einige Besonderheiten, die zu Problemen bei der maschinellen Indexierung führen.

Die zu bearbeitenden Texte werden zunächst mit computerlinguistischen Methoden bearbeitet: Zum einen werden die sinntragenden Bestandteile identifiziert (z.B. werden Artikel grundsätzlich ignoriert), zum anderen soll die Vielfalt der sprachlichen Formen reduziert werden – insbesondere durch die Zerlegung von Komposita und die Rückführung auf Stammformen. Beispielsweise wird die Phrase „entzündliche Erkrankungen des Herzmuskels" in die vier Segmente „entzuend", „krank", „herz" und „muskel" umgewandelt. Danach werden die ermittelten Zeichenketten mit der Gemeinsamen Normdatei (GND) abgeglichen. Diese wird ebenfalls computerlinguistisch „vorbehandelt" und fungiert als Wörterbuch. Genutzt wird dabei nur der Teilbestand „s", d.h. alle Datensätze, die für die Verwendung als Schlagwort zugelassen sind bzw. schon einmal als Schlagwort verwendet wurden.[3]

Eine besondere Schwierigkeit stellen Fälle dar, in denen gefundene Zeichenketten bzw. Kombinationen davon auf mehrere Konzepte in der GND gemappt werden könnten. Beispielsweise taucht „Leistung" in mehreren GND-Datensätzen auf, u.a. als „Leistung", „Leistung <Elektrotechnik>", „Leistung <Physik>" und „Leistung <Recht>". Mit verschiedenen Routinen wird dann versucht, die wahrscheinlichste Zuordnung zu identifizieren. Schließlich werden die ermittelten GND-Schlagwörter gewichtet. In der derzeitigen Implementierung für gedruckte Publikationen werden einem Titel maximal sieben Schlagwörter zugeordnet.[4] Die maschinell ermittelten Schlagwörter werden klar als solche gekennzeichnet. Sie erscheinen nicht nur im Katalog der DNB, sondern werden auch in den Datendiensten ausgeliefert.

Für den Vortrag auf dem Berliner Bibliothekartag, der diesem Beitrag zugrunde liegt, wurde eine größere Anzahl von Beispielen – überwiegend aus der Reihe B – gesichtet und analysiert. Das Ziel war es dabei, typische Probleme zu erkennen und daraus Schlussfolgerungen zu ziehen. Alle Überlegungen beziehen sich zunächst also nur auf ein konkretes Verfahren und nicht auf maschinelle Methoden der Inhaltserschließung insgesamt. Dennoch zeigen sie einige grundsätzliche Schwierigkeiten auf, die bei der Weiterentwicklung der maschinellen Indexierung beachtet werden sollten.

<https://b-u-b.de/wp-content/uploads/2018-01a.pdf>, Stand: 21.07.2018; Uhlmann, Sandro: Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND), in: Dialog mit Bibliotheken 2013/2, S. 26-36, <http://d-nb.info/1048376788/34>, Stand: 21.07.2018; Mödden, Elisabeth: Die maschinelle Erschließung der Deutschen Nationalbibliothek (Vortragsfolien), Workshop „Computerunterstützte Inhaltserschließung" am 8./9. Mai 2017 in der UB Stuttgart, <https://blog.ub.uni-stuttgart.de/veranstaltungen/workshop-computerunterstuetzte-inhaltserschliessung/>, Stand: 21.07.2018.

3 Ein GND-Datensatz kann nur dann als Schlagwort verwendet werden, wenn er im entsprechenden Feld mit dem Code „s" für den Teilbestand Sacherschließung versehen ist. Bei Sachschlagwörtern (Sachbegriffen) ist dies natürlich stets der Fall. Viele Datensätze für Personen und Körperschaften haben aber nur den Code „f" aus der Formalerschließung (z.B. weil eine Person bisher nur als Autor oder Herausgeber aufgetreten ist). Wird ein solcher Datensatz als Schlagwort benötigt, müssen zunächst bestimmte Felder (z.B. GND-Systematik und Quelle) und der Code „s" ergänzt werden. Bei intellektueller Erschließung stellt dies kein Problem dar. Hingegen kann der Averbis-Algorithmus die nur mit „f" gekennzeichneten Datensätze grundsätzlich nicht nutzen

4 Es gibt mehrere Konfigurationen für unterschiedliche Dokumenttypen; teilweise werden dabei auch mehr als sieben Schlagwörter ausgegeben.

Im Folgenden werden zentrale Problembereiche dargestellt und an Beispielen illustriert. Öfter wird dasselbe Beispiel an mehreren Stellen verwendet, um unterschiedliche Aspekte zu veranschaulichen. Angegeben werden jeweils Haupttitel und relevante Titelzusätze (in der bei der DNB erfassten Form) sowie die maschinell vergebenen Schlagwörter; außerdem der Link zum DNB-Katalogisat, wo auch die Inhaltsverzeichnisse abgerufen werden können. Die Angaben entsprechen dem Stand vom Juni 2018. In einigen Fällen wurden damals angezeigte maschinelle Schlagwörter mittlerweile gelöscht. Ob dies eine Folge der Präsentation dieser Beispiele auf dem Bibliothekartag war oder ob die Veränderungen in einem anderen Kontext stehen, ist der Verfasserin nicht bekannt. Die Vortragsfolien enthalten zahlreiche Screenshots der Katalogisate zum damaligen Stand sowie Ausschnitte aus den Inhaltsverzeichnissen.[5]

2. Verhältnis von Wörtern und Themen

Eine wichtige Erkenntnis bei der Analyse der Beispiele war, dass keineswegs alle Wörter in den ausgewerteten Quellen auch wirklich für Themen stehen, d.h. inhaltliche Aspekte des Werks ausdrücken. Als Beispiel sei die folgende Erschließung betrachtet: Unternehmen ; Guben ; Land ; Dokument ; Reprivatisierung ; Spree ; Brandenburg.[6] Geht es hier um die Reprivatisierung von Unternehmen – in der Stadt Guben, an der Spree oder in Brandenburg? Tatsächlich handelt es sich um eine Festschrift für eine in Guben ansässige Spedition: „200 Jahre Wilhelm Wilke Guben : 1817-2017 : Bilder & Dokumente einer wechselvollen Firmengeschichte“.

Das Schlagwort „Dokument“ ergibt sich aus dem Titelzusatz („Bilder & Dokumente“) – hier wird aber kein inhaltlicher, sondern ein formaler Aspekt benannt. Das Wort „Reprivatisierung“ kommt im Inhaltsverzeichnis zweimal vor, dient dort allerdings nur der Untergliederung der Firmengeschichte in den Zeitraum vor und nach diesem Ereignis.[7] Auch die Schlagwörter „Land“, „Spree“ und „Brandenburg“ haben nichts mit dem Thema zu tun: Sie basieren auf Grußwort-Angaben („Grußwort vom Ministerpräsidenten des Landes Brandenburg“ bzw. „Grußwort vom Landrat des Landkreises Spree-Neiße“). Von den sieben ermittelten Schlagwörtern verbleiben also nur „Guben“ und „Unternehmen“ als thematisch irgendwie relevant. Jedoch geht es nicht allgemein um Firmen in dieser Stadt, sondern nur um eine einzige – deren Name jedoch nicht als Schlagwort erkannt wurde. In dem bei der DNB angewendeten Verfahren wird außerdem die zeitliche Dimension grundsätzlich nicht berücksichtigt, weshalb auch ein passendes Zeitschlagwort fehlt.

Wörter, die nur der Strukturierung des Texts dienen, kommen in Inhaltsverzeichnissen besonders häufig vor – z.B. „Einleitung“, „Fragestellung“, „Ziele“, „Material und Methoden“, „Lebenslauf“ oder „Danksagung“. Bei gängigen Formulierungen lässt sich manches abfangen, aber dennoch kommt es immer wieder zu Fehlern. So erhielt die medizinische Dissertation „Entwicklung der Anzahl

5 Wiesenmüller, Heidrun: Maschinelle Indexierung am Beispiel der DNB. Analyse und Entwicklungsmöglichkeiten (Vortragsfolien), BIB OPUS-Publikationsserver, 04.06.2018, <https://nbn-resolving.org/urn:nbn:de:0290-opus4-36346>.

6 Katalogisat: <http://d-nb.info/1136451161>.

7 Die entsprechenden Kapitel heißen „Die Jahre von 1903 bis zur Reprivatisierung 1990“ und „Zur Entwicklung der Spedition seit der Reprivatisierung 1990“.

schulterendoprothetischer Eingriffe in Deutschland von 2005 bis 2012“ folgende Schlagwörter: Prothese ; Indikation ; Deutschland ; Zeithintergrund ; Ergebnis ; Diskussion.[8] Die drei letzten beruhen auf rein formalen Überschriften („Historischer Hintergrund“, „Ergebnisse“ und „Diskussion“), während der thematische Kern – Operationen, bei denen künstliche Schultergelenke eingesetzt werden – nicht erfasst wurde.

Bei einem Führer durch eine Kasseler Gemäldegalerie unter dem Aspekt der Provenienz („Provenienzgeschichten : Gemäldegalerie Alte Meister Schloss Wilhelmshöhe“) findet sich: Geschoss <Bauwesen> ; Meister ; Gemäldegalerie ; Spanisch.[9] Das besonders eigentümliche bautechnische Schlagwort geht – wie ein Blick ins Inhaltsverzeichnis zeigt – ebenfalls auf strukturierende Informationen zurück. Denn der Führer ist nach Stockwerken gegliedert („1. Etage: Altdeutsche, Italienische, Französische und Spanische Meister“ etc.), und „Etage“ ist eine Verweisung von „Geschoss <Bauwerk>“.

Einige weitere Beispiele illustrieren unterschiedliche Spielarten des Problems, dass nicht alle Wörter für Themen stehen: Eine Konferenzschrift „Deutsch-polnische Erinnerungsorte : 10. Deutsch-Polnischer Kommunalpolitischer Kongress der Landsmannschaft Ostpreußen“ erhielt die Schlagwörter: Kongress ; Allenstein ; Hotel.[10] Letztere rühren daher, dass die Tagung in einem Hotel in Allenstein stattfand („10.-11. Oktober 2015 Hotel Warminski Allenstein“). Hier verbleibt kein einziges Schlagwort, das irgendetwas mit dem Thema der Konferenz zu tun hätte.

Ein Werk über eine sächsische Herzogin des 18. Jahrhunderts mit dem Haupttitel „Voller Esprit und Wissensdurst“ wurde u.a. mit „Esprit“ und „Wissbegier“ verschlagwortet.[11] Die entsprechenden Titelwörter charakterisieren freilich nur die thematisierte Person, sind aber nicht selbst Thema.

Sehr häufig erscheinen außerdem Namen von Verfasserinnen und Verfassern u.ä. – die natürlich ebenfalls keine Themen sind – unter den maschinell vergebenen Schlagwörtern. So wurde ein Buch über nationalsozialistische Verbrechen u.a. mit „Scholz, Olaf“ verschlagwortet.[12] Dies beruht auf einem Grußwort, das Olaf Scholz als Erster Bürgermeister von Hamburg für den Band geschrieben hat. Als weiteres Beispiel sei ein im elektronischen Volltext vorliegender Aufsatz von Gerhart von Graevenitz genannt, bei dem der Autor zum Schlagwort geworden ist.[13]

8 Katalogisat: <http://d-nb.info/1134898630>.
9 Katalogisat: <http://d-nb.info/1136337318>.
10 Katalogisat: <http://d-nb.info/1128671891>.
11 „Voller Esprit und Wissensdurst : Herzogin Luise Dorothea von Sachsen-Gotha-Altenburg (1710-1767) : mit einer kommentierten Edition ihres Nachlassinventars“. Schlagwörter: Luise Dorothea, Sachsen-Gotha-Altenburg, Herzogin ; Herzogin ; Esprit ; Inventar ; Nachlass ; Wissbegier ; Eberle, Martin ; Netzwerk. Katalogisat: <http://d-nb.info/1137880813>.
12 „Transport in den Tod : von Hamburg-Langenhorn in die Tötungsanstalt Brandenburg : Lebensbilder von 136 jüdischen Patientinnen und Patienten“. Schlagwörter: Landes-Pflegeanstalt Brandenburg a.H. ; Hamburg-Langenhorn ; Patientin ; Rheinische Kliniken Langenfeld ; Scholz, Olaf ; Euthanasie <Nationalsozialismus> ; Hamburg. Katalogisat: <http://d-nb.info/1147658048>.
13 Katalogisat: <http://d-nb.info/1080824693>. Für weitere Angaben zu diesem Beispiel s.u. mit Fußnote 26.

Nicht selten wird dabei auch noch die falsche Person zugeordnet. In einer Festschrift für den Verein Deutscher Studenten in Karlsruhe[14] tauchen im Inhaltsverzeichnis elf Namen von Beiträgern auf. Einer davon, Christian Roth, findet sich als Schlagwort wieder. Tatsächlich gibt es in der GND einen für die Sacherschließung zugelassenen Datensatz für eine Person dieses Namens – doch handelt es sich dabei nicht um den Autor, sondern um einen 1934 gestorbenen bayerischen Staatsminister. Häufig zu beobachten ist auch das Extrahieren falscher Schlagwörter aus den Nachnamen von Personen. Beispielsweise führte der Name des Autors Norbert Mersch zum geografischen Schlagwort „Mersch" (ein Ort in Luxemburg).[15]

Während einerseits nicht alle vorkommenden Wörter für Themen stehen, lassen sich umgekehrt die tatsächlichen Themen nicht immer an einzelnen Wörtern ablesen. Vielmehr wird zum Verständnis häufig bestimmtes „Einordnungswissen" benötigt – eine Kombination von Wissen über die Welt und sprachlichem Wissen. Ein gutes Beispiel bietet der gerade schon mit Blick auf das Schlagwort „Mersch" betrachtete Band mit dem Titel „Der Legionär als Leistungssportler : die Leistung römischer Soldaten auf dem Prüfstand". Beim Blick ins Inhaltsverzeichnis und Anlesen des Bands wird für menschliche Leserinnen und Leser der Kontext rasch klar: Einerseits geht es um Soldaten im Römischen Reich, worauf u.a. lateinische Wörter hindeuten, andererseits um experimentelle Archäologie. So beschreibt etwa der Beitrag „Belastungsprofil eines Legionärs – eine sportwissenschaftliche Annäherung" ein Experiment, bei dem eine Testperson in voller Montur eines römischen Soldaten einen Stufentest auf einem Laufband absolvierte. Eine treffende Erschließung, anhand derer man sich eine gute Vorstellung vom Inhalt machen kann, wäre: Römisches Reich ; Soldat ; Körperliche Belastung ; Experiment ; Archäologie.[16] Man vergleiche dies mit den maschinell erstellten Schlagwörtern: Leistung <Recht> ; Soldat ; Leistungssportler ; Lastganglinie ; Prüfstand ; Mersch. Mit Ausnahme des Schlagworts „Soldat" (mit Verweisungsform „Legionär") geht dieses Indexat völlig am Thema vorbei und führt in gänzlich falsche Richtungen; eine Vorstellung über den tatsächlichen Inhalt lässt sich daraus nicht ableiten.

Ähnliches lässt sich beim bereits zitierten Gemäldegalerieführer „Provenienzgeschichten" beobachten, wo u.a. die Schlagwörter „Meister" und „Spanisch" ermittelt wurden.[17] Auch hier fehlte der Maschine das Einordnungswissen, um zu verstehen, dass die Formulierung „Spanische Meister" sich auf Gemälde bedeutender spanischer Maler bezieht.

3. Falsche oder fehlende Erkennung von Konzepten

Die häufig auftretenden Probleme bei der Disambiguierung von Konzepten wurden bereits kurz angesprochen. So wurde im schon besprochenen Beispiel aus der experimentellen Archäologie das Wort „Leistung" aus dem Titelzusätz fälschlich dem GND-Datensatz „Leistung <Recht>" zugeordnet.[18]

14 Katalogisat: <http://d-nb.info/1131098242>. Für weitere Angaben zu diesem Beispiel s.u. mit Fußnote 21.
15 Katalogisat: <http://d-nb.info/1049547829>. Das Schlagwort „Mersch" wurde mittlerweile entfernt. Für weitere Angaben zu diesem Beispiel s.u.
16 Vgl. die intellektuell vergebenen Schlagwortfolgen für diesen Titel in den Verbundkatalogen des GBV und hbz. „Experimentelle Archäologie" ist gemäß einem Hinweissatz in der GND zu zerlegen in „Experiment" und „Archäologie".
17 S.o. mit Fußnote 9.
18 Dieses Schlagwort wurde mittlerweile entfernt.

Ähnliches ist bei der Publikation „Leitfaden für die Markierung von Wanderwegen“ zu beobachten, bei der sich die Maschine für „Markierung <Chemie>“ entschied.[19] Dies ist besonders verblüffend, da es beim hier korrekten Schlagwort „Markierung“ sogar einen Verwendungshinweis gibt, der das Wort „Wanderweg“ enthält („Verknüpfe z.B. mit Wanderweg“).

Weniger bekannt sind die Schwierigkeiten, die sich aufgrund der computerlinguistischen Vorbearbeitung ergeben können. Eine online vorliegende Publikation mit dem Titel „Kläranlagen in der Energiewende : Faulung optimieren & Flexibilität wagen“ wurde u.a. mit „Müßiggang“ verschlagwortet.[20] Dies beruht offenbar auf der Rückführung von „Faulung“ in „faul“, was dann wiederum auf „Faulenzen“ (eine Verweisung von „Müßiggang“) gemappt wurde. Trotz gleichen Wortstamms ist aber „Faulung“ nicht dasselbe wie „Faulenzen“. Eine interessante Variante zeigt die schon erwähnte Festschrift „60 Jahre VDSt Karlsruhe : die Geschichte des Vereins Deutscher Studenten Karlsruhe von 1957 bis 2017“, für die u.a. das Schlagwort „Geschichtsverein“ ausgegeben wurde.[21] Die Formulierung „Geschichte des Vereins“ dürfte die Segmente „geschicht“ und „verein“ ergeben haben, mit denen dann beim Abgleich ein falscher Treffer „Geschichtsverein“ erzielt wurde.

Die vielleicht größte Problematik in diesem Bereich stellen jedoch die sogenannten „Named entities“ dar. Namen von Körperschaften und Personen oder Titel von Werken werden vielfach nicht erkannt, auch wenn sie in der GND enthalten sind. Oft führt dies zu unbefriedigenden, manchmal aber auch zu völlig irreführenden Schlagwörtern. Beim Führer „Provenienzgeschichten“[22] wurde etwa der Name der Gemäldegalerie nicht erkannt, obwohl die im Titelzusatz vorkommende Form exakt einer Verweisung in der GND entspricht und der Datensatz auch als Schlagwort zugelassen ist. Immerhin wurde das Sachschlagwort „Gemäldegalerie“ ausgegeben. Ungleich schwerwiegender ist etwa die Extraktion der Schlagwörter „Geologe“ und „Norden“ aus dem Körperschaftsnamen „Arbeitsgemeinschaft Norddeutscher Geologen“.[23]

Auch bei Personennamen funktioniert die Erkennung unerwartet schlecht. Als Beispiel sei der Band „’68 : Stichworte Marburg A-Z“ genannt, der – teilweise aus spezifisch Marburger Sicht – das Jahr 1968 behandelt und wie ein Lexikon aufgemacht ist.[24] Maschinell wurde dafür u.a. „Luther, Martin“ vergeben und man rätselt, was der Reformator mit den Achtundsechzigern zu tun haben könnte. Das Personenschlagwort basiert auf der Überschrift „King, Martin Luther (1929-1968)“. Obwohl der

19 Schlagwörter: Wanderweg ; Beschilderung ; Unfallversicherung ; Rundwanderweg ; Markierung <Chemie>. Katalogisat: <http://d-nb.info/1151148180>. Das Schlagwort „Markierung <Chemie>“ wurde mittlerweile entfernt.

20 Schlagwörter: Kläranlage ; Müßiggang ; Biegsamkeit ; Biogas ; Erneuerbare Energien ; Energiepolitik ; Klärschlammstabilisierung ; Schlammbehandlung. Katalogisat: <http://d-nb.info/1149512849>. Das Schlagwort „Müßiggang“ wurde mittlerweile entfernt.

21 Schlagwörter: Geschichtsverein ; Karlsruhe ; Roth, Christian ; Heim ; Vorstand. Katalogisat: <http://d-nb.info/1131098242>. Das Schlagwort „Geschichtsverein“ wurde mittlerweile entfernt.

22 S.o. mit Fußnote 9.

23 „80. Tagung der Arbeitsgemeinschaft Norddeutscher Geologen, 6.-9. Juni 2017 in Rendsburg : Tagungsband und Exkursionsführer“. Schlagwörter: Geologe ; Norden ; Salzstock ; Hydrogeologie ; Grube ; Bauer, Sebastian ; Sediment ; Exkursion ; Jasmund ; Sandstein. Katalogisat: <http://d-nb.info/113649944X>.

24 Schlagwörter: Marburg ; Luther, Martin ; Prager Frühling ; Maiunruhen <1968> ; Verführer. Katalogisat: <http://d-nb.info/1156209951>. Das Schlagwort „Luther, Martin“ wurde mittlerweile entfernt.

Name von Martin Luther King hier vollständig (sogar in invertierter Form) und inkl. der Lebensdaten aufgeführt wird, ist es dem Algorithmus nicht gelungen, den korrekten Datensatz zuzuordnen.

Bei Literatur über Herrscher kommt es öfter vor, dass nur der Fürstentitel als Schlagwort übernommen wird, nicht aber der Personenname. Beispielsweise wurde bei der Publikation „Landgraf Carl (1654-1730) : fürstliches Planen und Handeln zwischen Innovation und Tradition" als einziges Schlagwort „Landgraf" ausgegeben.[25] Dabei hätte es anhand der Lebensdaten und dem an mehreren Stellen im Inhaltsverzeichnis genannten Herrschaftsgebiet eigentlich möglich sein müssen, die richtige Person (Karl, Hessen-Kassel, Landgraf von, 1654-1730) zu identifizieren.

Ähnlich ist die Situation bei den Werktiteln, was anhand des bereits erwähnten Aufsatzes von Gerhart von Graevenitz illustriert werden kann: „Das Ich am Ende : Strukturen der Ich-Erzählung in Apuleius' Goldenem Esel und Grimmelshausens Simplicissimus Teutsch".[26] Natürlich gibt es sowohl für die „Metamorphoses" des Apuleius (mit Verweisungsform „Der goldene Esel") als auch für den „Simplizissimus" des Hans Jakob Christoffel von Grimmelshausen Normdatensätze in der GND, die auch als Schlagwörter zugelassen sind. Dennoch wurde keins der Werke als Schlagwort ermittelt; stattdessen wird uns das Sachschlagwort „Esel" präsentiert. Zumindest wurde mit „Apuleius, Madaurensis" einer der beiden Autoren richtig erkannt. Das ebenfalls vergebene Schlagwort „Grimmelshausen" bezieht sich hingegen nicht auf die Person, sondern auf den gleichnamigen Ort.

Natürlich gibt es auch viele „Named entities", die keinen Datensatz in der GND besitzen. Typischerweise erkennt der Algorithmus diese Besonderheit nicht und behandelt solche Eigennamen nicht anders als normale Wörter – mit fatalen Konsequenzen. Ein Beispiel dafür ist ein Führer durch einen Teil des Stuttgarter Stadtviertels Degerloch, der als „Haigst" bekannt ist: „Der Haigst – ein Spaziergang durch Geschichte und Gegenwart".[27] Dafür ermittelte die Software u.a. die Schlagwörter „Santiago de Chile" (aus der Überschrift „Der Santiago-de-Chile-Platz") und „Meistersang" (aus der Überschrift „Die Meistersingerstraße").

Schließlich sei auf die Problematik des uneigentlichen Wortgebrauchs, also der Verwendung von Wörtern in einem übertragenen Sinne (Metapher), hingewiesen. Die Maschine kann dies nicht erkennen und nimmt alles wörtlich. Folglich erhält ein Gesundheitsratgeber mit dem Titel „Mein Kompass : Wegweiser zur Gesundheit für Dich" das Schlagwort „Kompass",[28] und die Broschüre „Geldanlage für Faule"[29] wird wiederum mit „Müßiggang" versehen.

25 Katalogisat: <http://d-nb.info/1147882894>.

26 Schlagwörter: Ich-Erzählung ; Apuleius, Madaurensis ; Das Andere ; Esel ; Grimmelshausen ; Rede ; Autobiografische Erzählung ; Graevenitz, Gerhart von ; Ich-Form ; Autobiografie ; Sprecher ; Dialog <Literaturgattung>. Katalogisat: <http://d-nb.info/1080824693>.

27 Schlagwörter: Spaziergang ; Lehen ; Westen ; Berg ; Santiago de Chile ; Meistersang ; Grenzstein ; Maler. Katalogisat: <http://d-nb.info/1119549752>.

28 Schlagwörter: Gesundheit ; Kompass ; Therapeut ; Tastsinn ; Humor ; Geborgenheit ; Gesundheitsstörung. Katalogisat: <http://d-nb.info/1144316731>.

29 Schlagwörter: Kapitalanlage ; Müßiggang ; Altersversorgung ; Fonds ; Eigenheim. Katalogisat: <http://d-nb.info/1070846937>.

4. Gewichtung und Auswahl der Schlagwörter

Auch bei der Gewichtung und Auswahl der Schlagwörter kommt der Algorithmus immer wieder zu wenig befriedigenden Ergebnissen. Wie bei statistischen Verfahren üblich, spielt die Häufigkeit des Vorkommens einer Zeichenkette dabei eine große Rolle. Manchmal funktioniert dies sehr gut, etwa bei der medizinischen Dissertation „Akupunktur als Behandlungsmöglichkeit bei Depression in der Schwangerschaft : systematische Übersichtsarbeit“.[30] Die ermittelten Schlagwörter sind: Akupunktur ; Depression ; Schwangerschaft. Diese drei Wörter bzw. ihre Ableitungen erscheinen mit 13, 9 und 6 Vorkommen signifikant häufig im Inhaltsverzeichnis.

Diese Situation ist jedoch für Inhaltsverzeichnisse eher untypisch. Vielfach kommen alle Wörter nur ein- bis zweimal vor, sodass dies kaum als Basis für ein Ranking verwendet werden kann. Nicht selten taucht überdies das zentrale Thema im Inhaltsverzeichnis nicht oder kaum mehr auf, weil es schon im Titel benannt wurde und nun sozusagen vorausgesetzt wird. In den Kapitelüberschriften werden dann oft nur Unteraspekte benannt. Umgekehrt kommen manchmal auch unwichtige Wörter mehrfach im Inhaltsverzeichnis vor. Beispielsweise enthält die bereits zitierte Festschrift für den Verein Deutscher Studenten in Karlsruhe[31] drei Anhänge, in denen das Wort „Vorstände“ vorkommt – prompt wurde das Schlagwort „Vorstand“ ausgegeben. Eine Gewichtung durch Auszählen von Wörtern ist daher auf der Basis von Inhaltsverzeichnissen grundsätzlich problematisch. Zu überlegen wäre vielleicht, ob nicht der Umfang der Kapitel in die Gewichtung mit einbezogen werden müsste – aber auch dies würde nur einen Teil der Probleme lösen.

Beim derzeitigen Verfahren der DNB werden im Ergebnis häufig nicht die substanziell behandelten Themen abgebildet, sondern eine willkürlich erscheinende Auswahl von Einzelaspekten. Betrachten wir dazu nochmals den lexikonartig aufgemachten Band „’68 : Stichworte Marburg A-Z“.[32] Er enthält ca. 60 Einträge im Umfang von maximal zwei Seiten; darunter befinden sich alleine 14 Personen. In der Verschlagwortung berücksichtigt wurden gerade einmal vier dieser Überschriften – mit den Schlagwörtern „Luther, Martin“, „Prager Frühling“, „Maiunruhen <1968>“ und „Verführer“, von denen zwei auch noch falsch sind.[33] Warum ausgerechnet diese Aspekte ausgewählt wurden und keine anderen, ist sachlich nicht zu begründen – man kann hier eigentlich nur von Zufälligkeit sprechen.

Dabei muss man sich klarmachen, dass nicht nur falsche Schlagwörter, sondern auch Schlagwörter für Themen, die nur am Rande behandelt werden, zu Ballast in der Recherche führen. Im gerade besprochenen Beispiel mag es noch akzeptabel sein, wenn jemand, der nach dem Prager Frühling sucht, auch diesen Treffer erhält – auch wenn es gewiss besser geeignete Literatur gibt, um sich über dieses Thema zu informieren. Aber wer möchte beispielsweise bei einer Recherche nach Nordic Walking die Festschrift „125 Jahre Schwäbischer Albverein Ortsgruppe Crailsheim : 1892-2017“ erhalten?

30 Katalogisat: <http://d-nb.info/1140020501>.
31 S.o. mit Fußnote 21.
32 S.o. mit Fußnote 24.
33 Das falsche Schlagwort „Luther, Martin“ wurde weiter oben diskutiert. „Verführer“ basiert auf dem in einer Überschrift genannten Werktitel „Geheime Verführer“ (im Original „The hidden persuaders“, ein Werk über die Tricks der Werbeindustrie).

Die für diese Publikation ermittelten Schlagwörter sind: Ortsverein ; Crailsheim ; Familiengruppe ; Jugendgruppe ; Turnhalle ; Nordic Walking.[34] Basis für das Schlagwort „Nordic Walking" ist eine von knapp 30 Überschriften in einem Heft mit gerade einmal 51 Seiten Umfang; die einschlägigen Aktivitäten des Vereins werden auf etwa einer Seite beschrieben. Während der wichtigste Aspekt – der Schwäbische Albverein – in der Erschließung fehlt, stellen die maschinell vergebenen Schlagwörter eine beliebig wirkende Auswahl von Einzelaspekten dar: Die Jugendgruppe und die Familiengruppe des Vereins haben es in die Schlagwörter geschafft, nicht aber die Seniorengruppe. Und von den diversen Aktivitätsbereichen des Vereins wird nur das Nordic Walking aufgeführt.[35]

5. Überlegungen und Anregungen

Die nähere Beschäftigung mit dem bei der DNB eingesetzten Verfahren zeigt, dass dieses – zumindest in der derzeitigen Implementierung – keine ausreichenden Ergebnisse erbringt. Häufig wird angenommen, dass die Qualität maschineller Erschließungsmethoden bei etwa 80 % läge – von einem solchen Wert ist die Averbis-Software weit entfernt. Man erhält den Eindruck, dass es sich eher um eine „bessere Stichwortsuche" handelt als um eine echte inhaltliche Erschließung. Vorteile gegenüber einer normalen Stichwortsuche ergeben sich immer dann, wenn ein GND-Schlagwort korrekt zugeordnet wird und folglich auch Verweisungsformen berücksichtigt werden.

Eine Ursache für die schwachen Leistungen könnte die Eindimensionalität der Software sein, die sich ausschließlich auf Computerlinguistik und Statistik stützt. Weder „lernt" das System in irgendeiner Weise dazu, noch kann es über die ausgewerteten Texte hinaus weitere Informationen berücksichtigen. Angesichts der extrem komplexen Aufgabe wäre jedoch ein mehrdimensionaler Ansatz sicher erfolgversprechender. Wenn überhaupt, dann sollte das Averbis-Verfahren deshalb nur als ein Baustein in einer umfassenderen Methodik zur Anwendung kommen.

In der konkreten Ausgestaltung des Verfahrens empfindet die Verfasserin die Begrenzung auf einige wenige Schlagwörter als besonders problematisch. Denn dadurch sieht es so aus, als würde es sich um eine Erschließung der gewohnten Art handeln – nur eben maschinell erzeugt. Damit wird auch die Erwartung geweckt, dass hier nur „wichtige" Themen stehen würden, wie man es von der intellektuellen Verschlagwortung kennt. Stünden hier stattdessen z.B. 30 oder 40 Schlagwörter, so wäre von vornherein klar, dass diese Art der Erschließung einen anderen Charakter hat als eine intellektuell erstellte. Dies könnte auch den Eindruck von Beliebigkeit bei der Auswahl der mit einem Schlagwort wiedergegebenen Aspekte verringern. Allerdings würden dann natürlich auch noch mehr ungeeignete Schlagwörter produziert werden, die die Suche behindern.

Angesichts des erheblichen Anteils an „schlechten" Schlagwörtern, die der Algorithmus ausgibt, sollten in jedem Fall die Rechercheoptionen auf die veränderte Datenbasis abgestimmt werden. So könnten Titel mit intellektuell vergebenen Schlagwörtern höher gerankt werden als solche mit maschinell

34 Katalogisat: <http://d-nb.info/1135937346>.

35 Das Schlagwort „Turnhalle" beruht übrigens wieder auf einer nicht erkannten Named entity („Die 75 Jahr Feier im Turnhallensaal Altenmünster").

ermittelten. Auch sollte es möglich sein, die maschinellen Schlagwörter ganz von der Recherche auszuschließen, wenn sich dadurch ansonsten zu viel Ballast ergibt.

Als eines der Hauptprobleme im DNB-Verfahren wurde die mangelhafte Erkennung von „Named entities" identifiziert. Ein erster Schritt zur Verbesserung könnte ein der linguistischen Segmentierung vorgeschalteter Vergleichslauf sein, bei dem die (weitgehend unveränderten) Phrasen mit den in der GND enthaltenen Vorzugs- und Verweisungsformen abgeglichen werden, um z.B. Namen von Körperschaften zu identifizieren.

Naheliegend wäre es außerdem, die in der Formalerschließung erfassten Entitäten sozusagen als Hintergrundinformation für die Schlagwortvergabe zu nutzen: So sollten Personen, die als Verfasser/innen, Herausgeber/innen etc. erfasst sind, prinzipiell nicht als Schlagwort ausgegeben werden (mit möglichen Ausnahmen, wenn Formangaben wie „Tagebuch" oder „Autobiografie" vergeben wurden). Anders ist es bei den in der Formalerschließung erfassten Körperschaften: Sofern diese auch in Überschriften vorkommen, ist die Chance hoch, dass sie auch ein Thema des Werks sind.

Natürlich werden z.B. bei Aufsatzbänden nicht alle Beiträgerinnen und Beiträger in der Formalerschließung erfasst. Hier könnte eine vorgeschaltete Layout-Analyse der Inhaltsverzeichnisse helfen. Da deren Aufbau bestimmten Regeln und Mustern folgt, müsste es möglich sein, die Namen von Autorinnen und Autoren mit einiger Sicherheit von den Überschriften zu unterscheiden und nur letztere für die Auswertung zu verwenden.

Um auch die Erkennung von „Named entities" zu ermöglichen, die nicht in der GND enthalten sind, sind grundsätzlich zwei Wege denkbar: Einerseits könnten weitere Quellen mit einbezogen werden wie z.B. die DBpedia oder Google Maps, andererseits könnten zumindest für häufige Fälle Regeln formuliert werden. Eine solche Regel könnte beispielsweise besagen, dass Wörter, die auf „-straße" oder „-platz" enden, nicht zerlegt werden dürfen.

Nach Ansicht der Verfasserin wären außerdem umfangreiche Plausibilitätsprüfungen nötig, um möglichst viele falsche Schlagwörter im Vorfeld auszufiltern. Eine wichtige Voraussetzung dafür ist eine verlässliche Ermittlung des fachlichen Kontexts, in dem die jeweilige Publikation steht. Um hier nicht zu Zirkelschlüssen zu kommen, sollte dies nicht oder zumindest nicht nur auf maschinellen Methoden beruhen. Viele Informationen könnten dafür genutzt werden – beispielsweise die in der GND erfassten Berufe der beteiligten Personen oder die Fachgebiete, in denen sie bisher publiziert haben. Für letzteres könnte man klassifikatorische Erschließungen aus den Verbünden nutzen. Auf einer entsprechenden Datenbasis müsste es möglich sein, wenig wahrscheinliche Schlagwörter zu identifizieren. Ein Beispiel dafür ist das wohl auf einem Fehler bei der Kombination von Segmenten beruhende Schlagwort „Lastganglinie",[36] das für den bereits diskutierten Titel „Der Legionär als Leistungssportler" vergeben wurde. Wie die GND-Systemstelle zeigt, handelt es sich dabei um ein

36 Dieses Schlagwort beruht wahrscheinlich auf dem Wort „Belastungsprofil" im Titel des Beitrags „Belastungsprofil eines Legionärs". Eine Segmentierung ergibt „last" und „profil"; dies passt zu „Lastprofil <Elektrotechnik>" als Verweisungsform von „Lastganglinie".

Schlagwort aus der Elektrotechnik. Wie wahrscheinlich ist nun ein solches Schlagwort bei einem Werk, dessen Herausgeber Archäologe ist?

Auch der Vergleich zwischen intellektueller und maschineller Indexierung kann hilfreich sein, um Irrtümer des Algorithmus zu entdecken. Wird beispielsweise ein Schlagwort maschinell sehr viel häufiger vergeben als intellektuell, so spricht dies für einen systematischen Fehler. Solche Vergleiche sind freilich nur möglich, wenn dauerhaft eine ausreichende Menge an Titeln intellektuell erschlossen wird.

6. Fazit

Die derzeitige maschinelle Indexierung der DNB zeigt sehr deutliche Begrenzungen. Sie ist nur ein erster Schritt auf einem wohl noch sehr langen Weg hin zu akzeptablen Systemen für die automatische Schlagwortvergabe. Wenn das Verfahren auf der derzeitigen Entwicklungsstufe überhaupt zum Einsatz kommen kann, dann sollte dies auf Bereiche beschränkt bleiben, für die eine intellektuelle Erschließung – auch im arbeitsteiligen Verfahren – gänzlich unrealistisch ist.

Wo immer aber eine intellektuelle Erschließung verfügbar ist, sollte zunächst diese nachgenutzt werden. Zwar gibt es auch dabei Mängel, doch sind diese üblicherweise weit weniger schwerwiegend, da Menschen andere Fehler machen als Maschinen. Bibliothekarinnen und Bibliothekare wenden vielleicht die „Regeln für den Schlagwortkatalog“ (RSWK) nicht ganz korrekt an – der Algorithmus wendet das Regelwerk hingegen gar nicht an. Und menschliche Erschließerinnen und Erschließer sind zwar nicht immer konsistent bei der Vergabe der Schlagwörter, werden aber nur in ganz seltenen Fällen ein völlig irreführendes Schlagwort vergeben.

Bemerkenswerterweise nutzt die DNB derzeit nicht einmal ihre eigene intellektuelle Erschließung ausreichend nach. Beispielsweise wird bei Fällen, in denen dasselbe Werk parallel in- und außerhalb des Buchhandels erscheint (z.B. Verlags- und Museumsausgabe bei Ausstellungskatalogen), dieses für die Reihe A intellektuell verschlagwortet und für die Reihe B maschinell. Mit einem vorgeschalteten Clustering aller Ausgaben eines Werks ließe sich dies vermeiden.

Darüber hinaus sollte die DNB die intellektuelle Sacherschließung aus den Verbünden nachnutzen. Denn auch Titel aus der Reihe B werden durchaus häufig inhaltlich erschlossen, insbesondere von den regionalen Pflichtexemplarbibliotheken. Mit einem vorschlagsbasierten Werkzeug wie dem „Digitalen Assistenten“ kann die Übernahme von vorhandener Sacherschließung aus unterschiedlichen Quellen – und teilweise sogar aus unterschiedlichen Erschließungssystemen – zudem sehr effizient erfolgen.[37]

Auch in der zum Berliner Bibliothekartag veröffentlichten „Stellungnahme zur Entwicklung der Inhaltserschließung im D-A-CH-Raum“ einer Initiativgruppe des Standardisierungsausschusses

37 Zum Digitalen Assistenten vgl. Hinrichs, Imma; Milmeister, Gérard; Schäuble, Peter u.a.: Computerunterstützte Sacherschließung mit dem Digitalen Assistenten (DA-2), in: o-bib 3 (4), 2017, S. 156-185, <https://doi.org/10.5282/o-bib/2016H4S156-185>; Beckmann, Regine; Hinrichs, Imma: Alles unter einer Haube: Die nächste Generation des Digitalen Assistenten DA-3 (Vortragsfolien), BIB OPUS-Publikationsserver, 04.06.2018, <https://nbn-resolving.org/urn:nbn:de:0290-opus4-36355>.

wird die Rolle der Nachnutzung von Sacherschließungsinformationen betont: „Generell erscheint je nach Material, Inhalt, Bedarf und Datenlage ein fallspezifischer Methodenmix aus intellektueller Erschließung, maschinellen Verfahren und Fremddatenübernahme sinnvoll. Gerade die konsequente Nachnutzung bereits existierender inhaltserschließender Daten, insbesondere unter den bibliothekarischen Partnern im D-A-CH-Raum, ist in diesem Zusammenhang besonders wichtig.“[38]

Bei konsequentem und „intelligentem“ Poolen der an vielen Stellen vorhandenen intellektuellen Erschließung wird sich vielleicht zeigen, dass die Lücken im System gar nicht so groß sind wie gedacht.

Literaturverzeichnis

- Beckmann, Regine; Hinrichs, Imma: Alles unter einer Haube: Die nächste Generation des Digitalen Assistenten DA-3 (Vortragsfolien), BIB OPUS-Publikationsserver, 04.06.2018, <https://nbn-resolving.org/urn:nbn:de:0290-opus4-36355>.

- Hinrichs, Imma; Milmeister, Gérard; Schäuble, Peter u.a.: Computerunterstützte Sacherschließung mit dem Digitalen Assistenten (DA-2), in: o-bib 3 (4), 2017, S. 156-185, <https://doi.org/10.5282/o-bib/2016H4S156-185>.

- Mödden, Elisabeth: Die maschinelle Erschließung der Deutschen Nationalbibliothek (Vortragsfolien), Workshop „Computerunterstützte Inhaltserschließung“ am 8./9. Mai 2017 in der UB Stuttgart, <https://blog.ub.uni-stuttgart.de/veranstaltungen/workshop-computerunterstuetzte-inhaltserschliessung/>, Stand: 21.07.2018.

- Mödden, Elisabeth; Schöning-Walter, Christa; Uhlmann, Sandro: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek. Breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse, in: BuB 70 (1), 2018, S. 30-35. Online: <https://b-u-b.de/wp-content/uploads/2018-01a.pdf>, Stand: 21.07.2018.

- Standardisierungsausschuss (Initiativgruppe): Stellungnahme zur Entwicklung der Inhaltserschließung im D-A-CH-Raum, Deutsche Nationalbibliothek, 01.06.2018, <http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/protokolle/stellungnahmeIEDachRaum.html>, Stand: 21.07.2018.

- Uhlmann, Sandro: Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND), in: Dialog mit Bibliotheken 2013/2, S. 26-36, <http://d-nb.info/1048376788/34>, Stand: 21.07.2018.

38 Standardisierungsausschuss (Initiativgruppe): Stellungnahme zur Entwicklung der Inhaltserschließung im D-A-CH-Raum, Deutsche Nationalbibliothek, 01.06.2018, <http://www.dnb.de/SharedDocs/Downloads/DE/DNB/standardisierung/protokolle/stellungnahmeIEDachRaum.html>, Stand: 21.07.2018.

- Wiesenmüller, Heidrun: Maschinelle Indexierung am Beispiel der DNB. Analyse und Entwicklungsmöglichkeiten (Vortragsfolien), BIB OPUS-Publikationsserver, 04.06.2018, <https://nbn-resolving.org/urn:nbn:de:0290-opus4-36346>.

- Wiesenmüller, Heidrun: Das neue Sacherschließungskonzept der DNB in der FAZ, Blog Basiswissen RDA, 02.08.2017, <https://www.basiswissen-rda.de/neues-sacherschliessungskonzept-faz/>, Stand: 21.07.2018.