

Sandro Uhlmann

# Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND)

»Betrachtet man (...) die ungeheure, stets wachsende Büchermasse, so möchte einem wohl bange werden, besonders um der Herrn Bibliothekare und Literatoren willen, deren Pflicht es ist, diese Masse zu kennen und zu ordnen. Wie wird sie am Ende in den Köpfen der Bibliothekare und noch mehr, wie wird sie in den Bibliotheken Platz finden?«

(anonym): Betrachtungen über Bücher und Büchervermehrung. In: Literarisches Conversations-Blatt, Nr. 273 vom 27. November 1822.

## Motivation

Die Deutsche Nationalbibliothek (DNB) hat 1996 begonnen Netzpublikationen, insbesondere Online-Dissertationen, später auch E-Books und elektronische Zeitschriften zu sammeln, zu erschließen und zu archivieren<sup>1)</sup>. Mit dem Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek (DNBG)<sup>2)</sup> im Jahr 2006 wurde diese Aufgabe zum Bestandteil ihres gesetzlichen Auftrags. Zugleich haben sich seit dem Aufkommen des Internets<sup>3)</sup> die Möglichkeiten der nahezu unbegrenzten Speicherung, Verbreitung und Vernetzung von digitalen Daten und Informationen um ein Vielfaches gesteigert – und damit ist auch deren Bedeutung im Verhältnis zu den gedruckten Informationsquellen stark gestiegen. Für die DNB ist die Sammlung und Verzeichnung von elektronischen Materialien mehr als nur ein zeitgemäßer Beitrag zur Bewahrung des digitalen Kultur- und Wissenschaftserbes im deutschsprachigen Raum. Angesichts stetig wachsender Veröffentlichungsmengen in elektronischer Form und einer gleichbleibend großen Anzahl gedruckter Publikationen ist dies eine gewaltige Herausforderung, organisatorisch wie auch technologisch. Der Massenbetrieb an

sich, aber auch die Diversität der digitalen Formate erfordern neue Lösungen im Hinblick auf die Gestaltung der Geschäftsprozesse zu Sammlung und Langzeitarchivierung, aber auch für das Ordnen der Inhalte, um ein effizientes Suchen und Finden im umfangreichen Bestand der DNB<sup>4)</sup> zu ermöglichen.

Die bibliothekarische Praxis der intellektuellen verbalen und klassifikatorischen Inhaltserschließung – in der DNB durch Schlagwörter der GND, Notationen aus der Dewey-Dezimalklassifikation (DDC) und die auf der DDC basierenden DNB-Sachgruppen – ist ebenfalls einer notwendigen Veränderung unterworfen, um der skizzierten Herausforderung in all ihren Facetten gerecht werden zu können. Dabei ist das Ziel unverändert, die Dokumente such- und auffindbar zu machen. Um dieses Ziel zu erreichen, muss die bisherige Erschließungsarbeit erweitert werden: Sie muss z. B. durch technologische Hilfsmittel ergänzt werden, die automatisiert oder teilautomatisiert inhaltsbeschreibende Metadaten liefern. Formate und Regelwerke müssen optimiert und angepasst werden, um einerseits eine einfache (maschinelle) Lesbarkeit und Austauschbarkeit von Daten zu ermöglichen, und andererseits den intellektuellen Aufwand für die Erstellung und Pflege der genutzten Indexierungsterminologien auf ein vertretbares Maß zu begrenzen<sup>5)</sup>. Die intellektuell erstellte Inhaltserschließung ist weiterhin von hohem Nutzen, daneben treten aber andere Methoden und Verfahren, um die wachsenden Anforderungen zu bewältigen. Für den Nutzer ist letztlich entscheidend, ob die Suchanfragen zu adäquaten Retrievalergebnissen führen<sup>6)</sup>. Die DNB unterstützt den Zugriff der Nutzer auf ihre Bestände durch eine Kombination von bibliografischer und inhaltlicher Erschließung, Kataloganreicherung

Neue Lösungsansätze für neue Herausforderungen

Erweiterung der bisherigen Erschließungsarbeit durch technologische Hilfsmittel

Sammlung von Netzpublikationen seit 2006 gesetzlicher Auftrag

Nutzer erwarten den Suchanfragen entsprechende Retrieval-ergebnisse

z. B. mit Inhaltsverzeichnissen, Digitalisierung von Volltexten und den Einsatz von Suchmaschinentheorie. Für die verschiedenen Publikationsformen entsteht – in Abhängigkeit von formalen, technischen und spezifisch inhaltlichen Eigenschaften – ein differenziertes Erschließungskonzept (auch Schalenmodell<sup>7)</sup>, dem unterschiedliche Verzeichnungsstufen und -qualitäten zugrunde liegen<sup>8)</sup>. Korrespondierend wurden im Jahr 2010 mehrere Veränderungen am Erschließungskonzept vorgenommen, u. a. auch die Einführung der Reihe O – Online-Publikationen der Deutschen Nationalbibliografie für Netzpublikationen, die seither nicht mehr intellektuell erschlossen werden<sup>9)</sup>.

## Das Szenario »Automatische Beschlagwortung mit kontrolliertem Vokabular« im Projekt PETRUS

Im Jahr 2009 hat die DNB damit begonnen, für die inhaltliche Erschließung der Netzpublikationen und deren Aufbereitung für ein Suchen und Finden neue technologische Möglichkeiten im Rahmen des Projektes PETRUS<sup>10)</sup> zu evaluieren.

Evaluierung neuer technologischer Möglichkeiten im Projekt PETRUS

Das Ziel eines Teilprojektes<sup>11)</sup> war die automatische Anreicherung von deutschsprachigen Netzpublikationen mit Schlagwörtern eines kontrollierten Vokabulars auf der Grundlage digital vorhandener Volltexte, Abstracts oder Objekte der Kataloganreicherung wie beispielsweise Inhaltsverzeichnisse. Zu Projektbeginn wurde auch die Möglichkeit einer ausschließlich freien Indexierung – also der Ermittlung aller relevanten Stichwörter eines Textes aus dem Text selbst und nicht aus einer verbindlichen Liste an Schlagwörtern – in Betracht gezogen. Der Mehrwert einer Beschlagwortung mit kontrolliertem Vokabular z. B. die Einschränkung der Vielfalt von Begriffsbenennungen durch Synonyme, die Möglichkeit gleichlautende Begriffe und Namen durch Homonymenzusätze unterscheiden zu können oder auch die Relationierung von Begriffen durch Ober- und Unterbegriffe, kann dem Ergebnis einer freien Schlagwortvergabe eindeutig vorgezogen werden<sup>12)</sup>. Freie Schlagwörter könnten aber künftig zusätzlich genutzt werden, beispielsweise als Indikator für neu anzusetzende Schlagwörter.

Als kontrolliertes Vokabular wird primär die GND (zu Beginn des Projektes noch die Schlagwortnormdatei SWD und die Personennamendatei PND) eingesetzt. Die Möglichkeit, andere – insbesondere fremdsprachige – Vokabulare einzusetzen, wird perspektivisch mitgedacht. Zu Projektbeginn musste zunächst grundsätzlich geprüft werden, ob die Normdaten mit ihrem spezifischen Format überhaupt in eine Software zur automatischen Beschlagwortung integrierbar sind und wie sich der Wortschatz linguistisch verarbeiten lässt. Über eine europaweite Ausschreibung wurden im Jahr 2010 geeignete Softwaresysteme als Kandidaten für das Experiment gesucht und letztlich Testlizenzen für zwei Softwareprodukte erworben, die auf unterschiedlichen computerlinguistischen Verfahren basierten. Vor dem Hintergrund einer späteren Integration in die Systemarchitektur der DNB wurden auch bestimmte technische Eigenschaften, insbesondere ein hochgradig offener, modularer und transparenter Systemaufbau, verlangt. Die beiden Systemhersteller haben die Tests bei der DNB ein Jahr lang aktiv begleitet. Auf der Basis der gewonnenen Erkenntnisse wurde 2011 ein zweites Ausschreibungsverfahren durchgeführt, mit dem Ziel konkrete softwaregestützte Erschließungsverfahren aufzubauen und in den Produktivbetrieb bei der DNB zu überführen. Den Zuschlag hat die Averbis Extraction Platform der Averbis GmbH<sup>13)</sup> aus Freiburg im Breisgau erhalten. In enger Zusammenarbeit wird seitdem auf eine erste Konfiguration für die automatische Beschlagwortung und einen produktiven Workflow hingearbeitet.

GND als Terminologie für computerlinguistische Verarbeitung

## Die grundlegende Funktionalität der Averbis Extraction Platform<sup>14)</sup>

Die Averbis Extraction Platform bietet Komponenten zur Textanalyse, mit denen Terme aus elektronischen Dokumenten extrahiert werden können. Sie identifiziert einzelne Begriffe eines Textes (v. a. Nominalphrasen), die eine berechenbare Relevanz besitzen. In der Software ist eine Vielzahl linguistischer Vorverarbeitungsschritte implementiert, um verschiedene sprachliche Analyseebenen (Satz, Wort, Wortart etc.) zu erfassen. Die extrahierten Nominalphrasen können anschließend mit dem

Analyse und Extraktion einzelner Begriffe eines Textes

Averbis Concept Mapper auf beliebige Wörterbücher oder Terminologien abgebildet werden.

Bei der Beschlagwortung werden verschiedene Verarbeitungsschritte durchlaufen (s. Abb. 1). Das Einlesen der Daten erfolgt mit dem »Collection Reader«. Es schließt sich die Zerlegung und Umwandlung der unterschiedlichen Strukturinformationen in ein für die Weiterverarbeitung brauchbares Format durch einen Parser und die linguistische Verarbeitung an. Hierauf aufbauend erfolgt der eigentliche Prozess der Schlagwortvergabe, innerhalb dessen Schlagwortkandidaten aus einer hinterlegten Terminologie erkannt und in den Texten annotiert werden. Aus den Kandidaten werden schließlich mittels verschiedener Filtermechanismen diejenigen Schlagwörter ermittelt, die die höchste Relevanz besitzen. Zum Abschluss werden die Ergebnisse durch den Ausgabe-Generator herausgeschrieben.

Ermittlung relevanter Schlagwörter durch Filtermechanismen

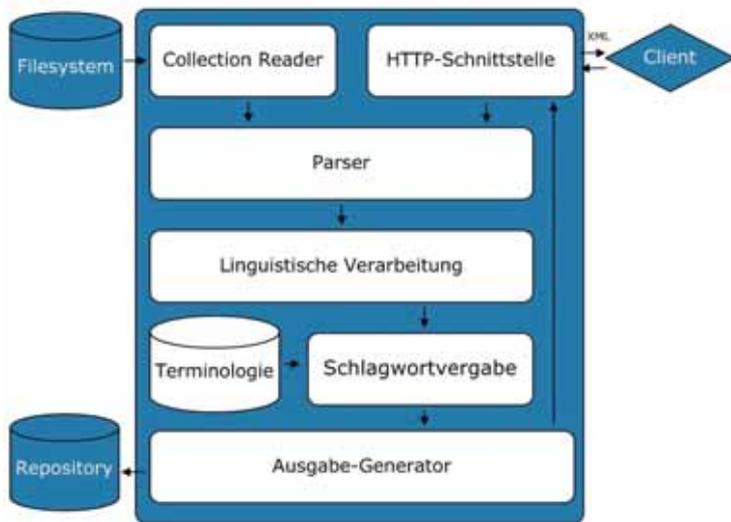


Abb. 1: Architektur der Beschlagwortungskomponente der Averbis Extraction Plattform (Quelle: Averbis GmbH)

Die Tokenisierung zerlegt die Eingabesequenzen in einzelne Tokens. Tokens sind dabei nicht nur Einzelwörter, sondern auch Satzzeichen wie Punkt und Komma. Wie bei der Satzerkennung besteht die Schwierigkeit der Aufgabe darin, dass Tokens auch Zahlen (»B2B«) oder andere Zeichen, wie Punkte (»Dr.«) oder Gedankenstriche (»E-Mail«), enthalten können. Sind einzelne Tokens erkannt, so werden diese mittels Part-of-Speech-Tagging (Wortarterkennung) mit Informationen, wie Nomen, Artikel, Verb etc., versehen. Eine charakteristische Abfolge von Tokens mit bestimmten Wortarten kann anschließend bei der Phrasenerkennung in Phrasen bzw. Chunks zusammengefasst werden. Der Chunker erkennt Phrasen, wie Nominal-, Präpositional- oder Verbalphrasen, in einem Text. Für die Beschlagwortung eines Textes kann es beispielsweise vorteilhaft sein, nur Nominalphrasen zu untersuchen und dabei Verbalphrasen zu ignorieren. Verschiedene morphologische Varianten eines Wortes können mit einem regelbasierten Stemmer auf ihren gemeinsamen Wortstamm zurückgeführt werden (Stammformbildung). Für Eigennamen (Personen und Geografika) kommt zudem ein lexikonbasierter Lemmatizer zum Einsatz, der die Genitiv-Formen auf die entsprechende Grundform zurückführt. Aus »Deutschlands« wird beispielsweise entsprechend »Deutschland« und aus »Hippolytus'« entsprechend »Hippolytus« (Lemmatisierung). Die im Deutschen üblichen Komposita (z. B. »Kindertanzenlehrerin«) zu entschlüsseln, ist eine der Aufgaben der morphosemantischen Indexierung. Bei diesem Verfahren werden Texte in verschiedenen Schritten sprachlich analysiert und normalisiert. Relevante Passagen (Segmente) – seien es Wortteile, Wörter oder Wortgruppen – werden erkannt, und Wörter mit der gleichen Bedeutung werden über semantische Gruppen (sogenannte Morphem-Identifizierer) inhaltlich miteinander vernetzt. Dabei vereinheitlicht die morphosemantische Indexierung sprachliche Varianten gleichbedeutender Ausdrücke.

Mehrstufige linguistische Verarbeitung

## Linguistische Verarbeitung

Bei der linguistischen Verarbeitung durchlaufen die zu erschließenden Dokumente eine Reihe modular einsetzbarer Komponenten zur Sprachanalyse (s. Abb. 2). Bei der Satzerkennung annotiert der Sentence-Detector Markierungen von Satzzeichen zu Satzzeichen, bei der anschließenden Worterken-

## Konzepterkennung

Zur Abbildung von linguistisch verarbeiteten Textstrings auf eine Terminologie dient der Averbis



Abb. 2: Übersicht der linguistischen Verarbeitungsschritte der Averbis Extraction Platform (Quelle: Averbis GmbH)

Abbildung einer Terminologie auf relevante Textpassagen

Concept Mapper, ein konfigurierbarer, lexikonbasierter Annotator, der pro Konzept (= Schlagwortdatensatz) auch Synonyme und weitere Attribute berücksichtigen kann. Der Abgleich mit der im Wörterbuch enthaltenen Terminologie kann auf Dokumentenebene, auf zusammenhängenden oder getrennten Textblöcken und Phrasen durchgeführt werden, wodurch auch sprachliche Konstruktionen mit Bindestrichen erkannt werden. Zudem kann der Abgleich sowohl in der Reihenfolge der einzelnen Wörter als auch unabhängig davon erfolgen und auf unterschiedlichen Ebenen der linguistischen Verarbeitung stattfinden, etwa auf Wort-, Wortstamm-, Lemma- oder Segment-Ebene.

### Lexikalische Ressource GND

Als Terminologie für die Beschlagwortung deutschsprachiger Publikationen wurde die GND<sup>15)</sup> im Format GND-MarcXML in die Averbis Terminology Platform<sup>16)</sup> eingelesen. Diese stellt einen umfassenden Zugriff auf den Wortschatz von kontrollierten Vokabularen (Thesauri, Taxonomien, Terminologien, Ontologien) bereit. Auf der Grundlage einer detaillierten fachlichen Spezifikation wurden aus

Integration der GND in die Averbis Terminology Platform

der GND die Satzarten Tp (individualisierte Personen), Ts (Sachschlagwörter), Tg (Geografika) sowie Ts1e (Hinweissätze) in die Software integriert. In einem nächsten Schritt sollen auch die Satzarten Tb (Körperschaften), Tf (Kongresse) und Tu (Werke, oft auch »Werktitel« genannt) aufgenommen werden. Dabei wurden jeweils nur die Datensätze mit Katalogisierungslevel 1 und aus dem Teilbestand s (Sacherschließung) der GND berücksichtigt<sup>17)</sup>. Dies sind in Zahlen bislang folgende Größenordnungen<sup>18)</sup>:

- Tp** – Person (individualisiert) 345.350 Datensätze
- Tg** - Geografikum 198.788 Datensätze
- Ts** - Sachbegriff 182.957 Datensätze (ohne Hinweissätze)
- Ts1e** – Hinweissatz 4.747 Datensätze

Diese Datensätze wurden in die interne Terminologiedatenbank eingespielt und können mit der

Überführung der Datensätze in Wörterbücher

Averbis Terminology Platform vom Anwender in Wörterbücher überführt werden. Die Wörterbücher stehen dann der Averbis Extraction Platform zur Verfügung. Dabei wird der Text gegen alle bevorzugten Benennungen und Synonyme im Wörterbuch abgeglichen. Bei einem Treffer werden die bevorzugte Benennung eines Begriffes, die dazugehörige Identifikationsnummer des Datensatzes (IDN) und der ermittelte Konfidenzwert ausgegeben.

Schwerpunkt beim Einbau der GND war es, alle maschinell interpretierbaren Textstrings und Relationen eines GND-Datensatzes für die automatische Beschlagwortung nutzbar zu machen. Es wurden neben der IDN, der bevorzugten Benennung und den Synonymen daher z. B. auch Oberbegriffe, verwandte Begriffe, Berufsbezeichnungen bei Personen als auch Codierungen, wie die GND-Systematik, die Ländercodes, die DDC-Notationen und die Entitätencodes, in die Averbis Terminology Platform aufgenommen. Diese Informationen werden im Prozess der Konzepterkennung verarbeitet, interpretiert und u. a. für die Disambiguierung genutzt.

Nutzung der GND-Relationen für die Beschlagwortung

## Disambiguierung

Beim Abgleich der im Text ermittelten Terme mit dem Wortschatz der GND kann es zu Ambiguitäten kommen, d. h. es werden zu einer Textstelle mehrere gleichlautende Begriffe (bevorzugte Benennung oder Synonyme im Schlagwortdatensatz) gefunden. Diese Begriffe sind dann »ambig« (mehrdeutig). Der Begriff »Bank« kann sich sowohl auf ein Kreditinstitut als auch auf ein Möbelstück beziehen. In der GND befinden sich sehr viele ambige Begriffe sowohl innerhalb als auch zwischen den Satzarten Sachschlagwörter, Geografika und Personen. Ambiguität tritt zudem auch auf, wenn Synonyme einer linguistischen Vorverarbeitung – wie beispielsweise der Stammformbildung oder der morphosemantischen Indexierung – unterzogen und dadurch auf eine generalisierte Form zurückgeführt werden. Daher ist eine Disambiguierung im Anschluss an die Konzepterkennung unabdingbar. Sie wird immer dann eingesetzt, wenn an einer Textstelle mehrere Treffer (d. h. verschiedene bevorzugte Benennungen oder Synony-

Behandlung mehrdeutiger Begriffe

me) gefunden wurden. Das Disambiguierungsverfahren durchläuft mehrere Stufen und bricht ab, wenn in einer Stufe jegliche Ambiguität für eine Textstelle aufgelöst werden konnte. Das Verfahren funktioniert aktuell auf einem verlässlichen Niveau, wobei eine vollständige, d. h. immer eindeutige, Zuweisung eines jeden Terms zu seiner semantischen Herkunft, mit technologischen Mitteln nicht erreichbar ist.

## Wörterbuchpflege

Die Averbis Terminology Platform dient zum Export von Terminologien in ein Wörterbuch der Averbis Extraction Platform und ermöglicht auch eine komfortable Navigation in Terminologien. Neben der Exportfunktion bietet das Modul auch Möglichkeiten, um bevorzugte Benennungen, Synonyme, komplette Schlagwortdatensätze oder auch ganze Teilbäume einer Terminologie vom Export auszuschließen, also problematische Terme aus dem Wörterbuch zu entfernen. Außerdem ist es möglich, Terme so zu markieren, dass sie nur bei exaktem Auftreten im zu erschließenden Text erkannt werden. Hierzu bietet die Averbis Terminology Platform verschiedene Modi: im Standardfall sind die Schlagwörter auf DEFAULT gesetzt, im IGNORE-Modus kann ein gesamtes Schlagwort oder auch nur ein Synonym eines Schlagwortes nicht in das erstellte Wörterbuch übernommen werden, und im Modus EXACT wird verhindert, dass das Schlagwort oder ein Synonym des Schlagwortes einer linguistischen Verarbeitung unterzogen wird. Das umfangreiche Vokabular der GND zwingt dazu, von den Möglichkeiten des IGNORE-Modus Gebrauch zu machen, um gezielte Einschränkungen vorzunehmen und somit Fehlidentifikationen zu vermeiden. Mit dem Modus EXACT sollen v. a. Überidentifikationen bei der morphosemantischen Indexierung verhindert werden.

Modifikation der Wörterbuchterme

## Update-Verfahren für GND-Daten

Um neue GND-Datensätze, aber auch Korrekturen, Zusammenführungen, Aufspaltungen oder Löschungen von Datensätzen regelmäßig in die

Schnittstelle für maschinelles Update der Terminologie

Averbis Terminology Platform zu überführen, wurde eine Schnittstelle für ein maschinelles Update der Terminologie geschaffen. Genutzt werden die regulären Datendienste der DNB, die derzeit wöchentlich einen Änderungsdienst und halbjährlich einen Gesamtabzug der GND umfassen. Da eine sehr zeitnahe Synchronisierung des Wörterbuchs angestrebt wird, sollen die Neuerungen aus dem Änderungsdienst künftig in einem noch festzulegenden Rhythmus (jeweils wöchentlich oder mehrere wöchentliche Abzüge gebündelt) automatisch eingespielt werden. Ein sogenannter GND-Reader liest alle neuen und geänderten Datensätze in die Averbis Terminology Platform ein und eliminiert alle als »gelöscht« gekennzeichneten Daten (inkrementelles Update). Anhand von Zeitstempeln wird sichergestellt, dass die Wörterbuchpflege kontrolliert und nachhaltig betrieben werden kann.

### Evaluierung

Die Qualitätsmessung wird anhand von Stichproben durchgeführt und beruht auf einer intellektuellen

Bewertung der inhaltlichen Übereinstimmung zwischen den automatisch vergebenen Schlagwörtern und dem Thema des Dokuments. Durchgeführt wird die Bewertung durch die jeweils für die Fachgebiete zuständigen Mitarbeiter der Abteilung Inhaltserschließung der DNB. Die Bewerter erhalten in einer Auswertungsdatenbank den Autor, den Titel und einen Link zum elektronischen Volltext sowie eine Liste der automatisch vergebenen GND-Schlagwörter pro Objekt (s. Abb. 3).

Qualitätsmessung beruht auf intellektueller Bewertung

Für jedes Dokument werden die einzelnen Schlagwörter bewertet. Dabei wird jedem Schlagwort auf einer 4-Punkte-Skala ein Wert zugewiesen. Folgende Kategorien sind möglich:

- Sehr nützlich
- das einzelne Schlagwort beschreibt einen wichtigen Aspekt des Textes in genauer Übereinstimmung;

- Nützlich
- das einzelne Schlagwort beschreibt einen wichtigen Aspekt des Textes aus einer weiteren (oder auch engeren) Perspektive;

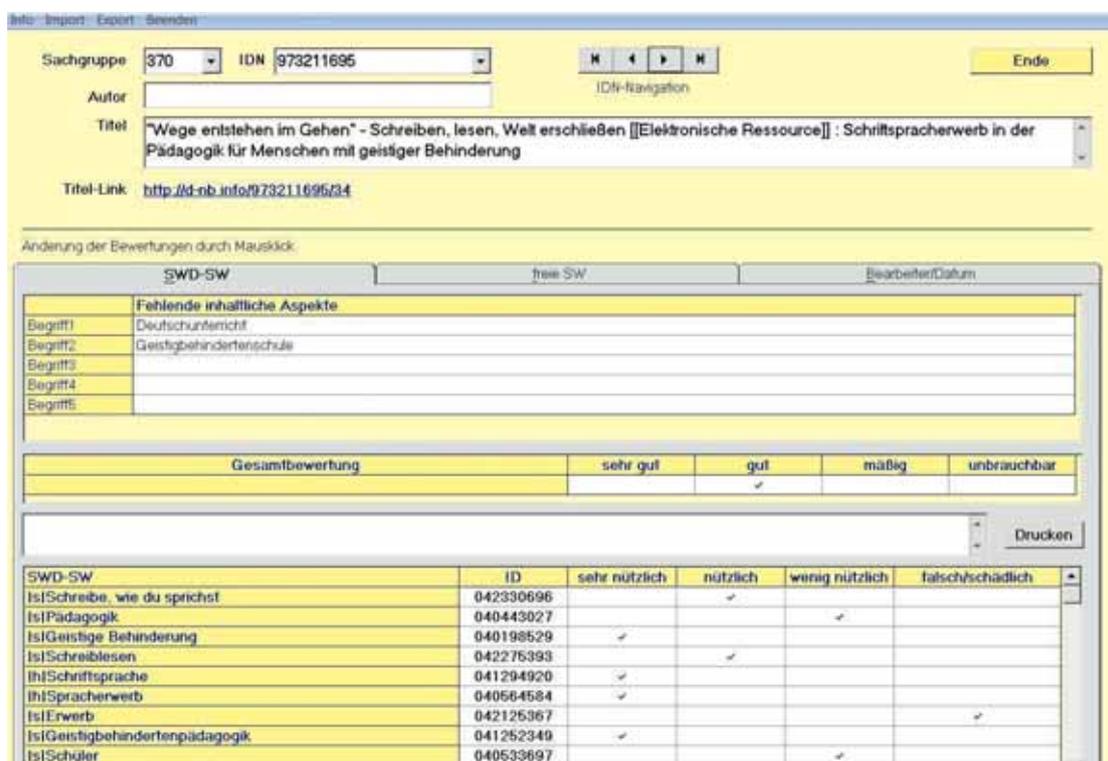


Abb. 3: Ansicht der Auswertungsdatenbank, hier mit einem Dokument der Sachgruppe 370 mit bewertetem Indexat

Wenig nützlich

- das einzelne Schlagwort beschreibt einen wichtigen Aspekt des Textes nicht ausreichend, ist aber auch nicht völlig unzutreffend oder falsch;

Falsch

- das einzelne Schlagwort beschreibt keinen wichtigen Aspekt des Textes und ist falsch.

Der Test zielt auf den Grad der inhaltlichen Übereinstimmung zwischen dem Schlagwort und dem Text, also ob das Thema der Publikation durch die einzelnen GND-Schlagwörter richtig und sinnvoll beschrieben wird. Um einer der drei nützlichen Stufen zugeordnet zu werden, muss das Schlagwort auf der begrifflichen Ebene zu einem Thema des Textes gehören. Wenn es geeignet ist, einen Aspekt vollständig zu beschreiben, ist es sehr nützlich. Wenn es eine nur unwesentliche Übereinstimmung zeigt, ist es wenig nützlich. Es wird auch eine Gesamtbewertung des Indexates auf einer 4-Punkte-Skala durchgeführt. Die Bewerter sollen außerdem im Indexat als fehlend erachtete Begriffe angeben, um die Vollständigkeit des Indexates pro Dokument messbar zu machen. Ein vollständiges Indexat enthält demnach die Summe aller nicht als falsch bewerteten und aller fehlenden Schlagwörter. Die zur Bewertung von Treffermengen im Retrieval verbreiteten Maße Precision und Recall können

Beurteilung der inhaltlichen Übereinstimmung zwischen den Schlagwörtern und dem Thema eines Textes

auch für die intellektuellen Bewertungen verwendet werden. Die Precision misst die Nützlichkeit, d. h. welcher Anteil der gefundenen Schlagwörter tatsächlich relevant ist. Der Recall beschreibt die Vollständigkeit, also wie viele relevante Schlagwörter eines Dokumentes gefunden wurden. Der festgelegte Standard für die korrekte Beschlagwortung ist das oben definierte vollständige Indexat. Dabei gehen die vier Kategorien der Bewertung mit einem annähernd linearen Relevanzfaktor in die Berechnung ein (sehr nützlich = 1.0 ; nützlich = 0.7 ; wenig nützlich = 0.3 und falsch = 0.0 sowie für fehlende Schlagwörter ebenfalls Faktor 1.0)<sup>19</sup>. Die Kennzahlen für Precision und Recall können Werte zwischen 0 und 1 annehmen, wobei das optimale Indexat einen Precision-/Recall-Wert von jeweils 1 aufweisen müsste. Anders ausgedrückt: das Optimum wird erreicht bei größtmöglicher Precision und maximalem Recall.

Precision und Recall

## Tests und Ergebnisse

Die aktuellsten Ergebnisse stammen aus einem Test der Softwareversion 2.0 im Frühjahr 2013. Durchgeführt wurde eine automatische Beschlagwortung von deutschen Volltexten (i. d. R. Hochschulschriften) mit Sachschlagwörtern, Geografika, Hinweissätzen sowie den individualisierten Personen-

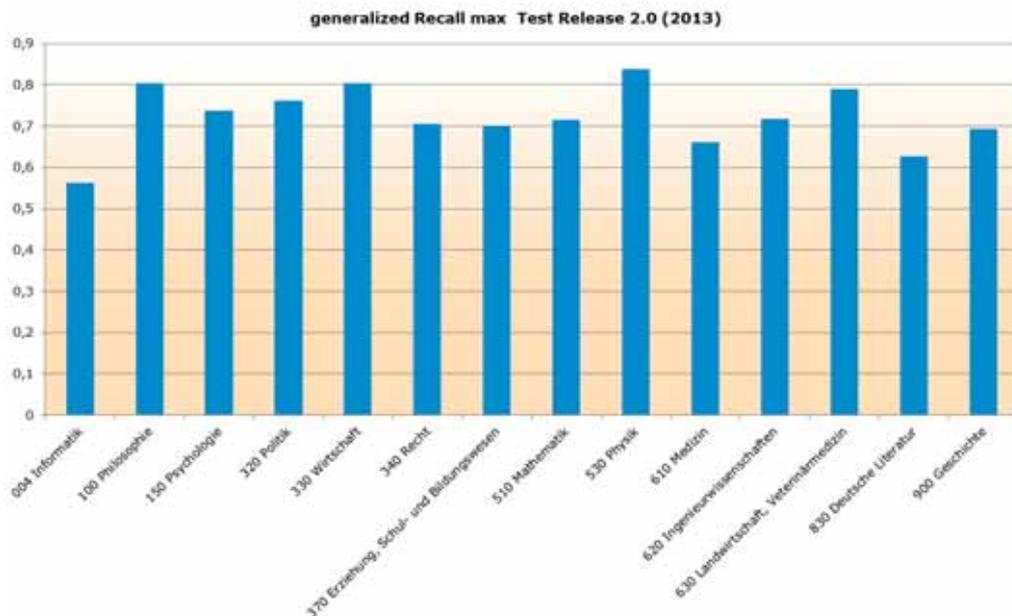


Diagramm 1: Recall Test Release 2.0 (2013)

namen der GND. Der Test umfasste 4.336 Objekte aus 14 Sachgruppen, dabei wurde pro Dokument eine feste Anzahl von 10 Schlagwörtern ausgegeben. Pro Sachgruppe wurde eine Stichprobe von 30 Dokumenten beurteilt.

Die statistische Auswertung zeigt einen Recall-Wert pro Sachgruppe (siehe Diagramm 1), der sich überwiegend in einem Wertebereich von 0,65 bis 0,75 bewegt (mit Ausreißern von 0,56 in der Sachgruppe 004 Informatik bis zu 0,84 in der Sachgruppe 530 Physik).

Die Precision liegt zwischen 0,38 in der Sachgruppe 004 Informatik und 0,62 in der Sachgruppe 330 Wirtschaft. Der Großteil der Sachgruppen pendelt zwischen 0,45 bis 0,55 (siehe Diagramm 2).

Zurzeit kommt es aufgrund der fehlenden GND-Satzarten Körperschaften, Kongresse und Werktitel noch vermehrt zu Fehlidentifikationen. Diese Satzarten sollen bis Herbst 2013 in die Software integriert werden. Außerdem wird zurzeit geprüft, ob durch Einführung einer Schwelle für den Konfidenzwert eine bessere Precision ohne große Verluste beim Recall erreicht werden kann.

## Übergang in den Routinebetrieb und Qualitätssicherung

Neben der Softwareanpassung und den Tests wurden die notwendigen Vorkehrungen für den Pro-

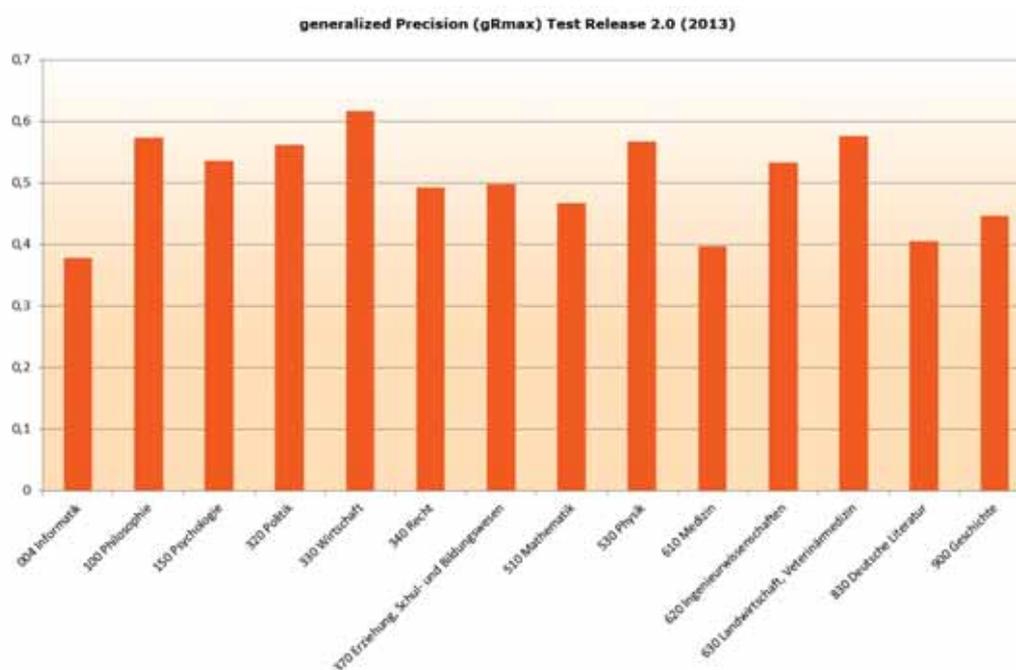


Diagramm 2: Precision Test Release 2.0 (2013)

Die Tests zeigen über alle Sachgruppen hinweg ein positives Ergebnis in Bezug auf den Anteil der als nützlich und sehr nützlich bewerteten Schlagwörter pro Indexat. Der Anteil der wenig nützlichen Schlagwörter ist ebenfalls zu betrachten, sie sind bei einer Recherche eher Ballast als Hilfe. Noch problematischer sind falsche Schlagwörter: Sie werden trotz kontinuierlicher Optimierung der Verfahren auch künftig nicht vollständig zu vermeiden sein, da ein automatisches Indexierungssystem niemals in der Lage sein wird, ausschließlich korrekte Schlagwörter zu vergeben.

Testergebnisse

duktionsbetrieb getroffen. Der Ablaufprozess musste dafür organisatorisch geplant und technisch umgesetzt sowie in die Gesamtgeschäftsprozesse für die Bearbeitung der Netzpublikationen integriert werden. Der Prozess der automatischen Beschlagwortung (siehe Schema in Abbildung 4) startet täglich zu einer festgelegten Zeit, indem die IDNs neu importierter Netzpublikationen über das Erfassungsdatum, den Publikationstyp und weitere Kriterien selektiert und an einen Webservice, den sogenannten Petrus-Service (1), übergeben werden. Dieser holt die zu erschließenden Texte aus dem

Repository (3) und die zugehörigen Metadaten (2) aus der bibliografischen Datenbank (CBS). Dabei werden die Netzpublikationen in plain text umgewandelt und nach UTF8 formatiert. Der Language Guesser – ein vorgeschaltetes Modul aus der Averbis Extraction Platform – erkennt anschließend die Sprache des Textes und erstellt als Ergebnis für jedes Dokument eine Sprachen-Rangliste. Die Sprache mit dem höchsten Rang wird als Dokument-sprache gewählt. Nach der Übergabe an den Averbis-Webservice (4) werden die deutschsprachigen Dokumente der jeweiligen Konfiguration in der Erschließungssoftware zugeführt. Es können derzeit sechs verschiedene Konfigurationen parallel betrieben werden, eine Erweiterung ist möglich. Die Erschließungssoftware prozessiert die elektronischen Dokumente wie oben beschrieben und der Averbis-Webservice gibt als Ergebnis eine Liste der automatisch aus dem Text ermittelten GND-Schlagwörter an den Petrus-Service zurück (5). Über eine schreibende Schnittstelle (6) werden die Ergebnisse anhand der IDN in den bibliografischen Datensatz des Titels im CBS geschrieben.

Ablaufprozess im Produktionsbetrieb

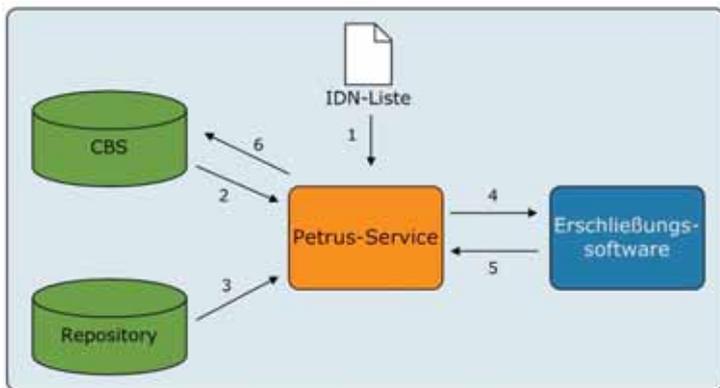


Abb. 4: Technischer Ablauf der automatischen Beschlagwortung im Produktionsbetrieb

Die automatisch vergebenen Schlagwörter werden in eigens dafür vorgesehene Felder (5540 in Pica3 bzw. 044H in Pica+) geschrieben. Erfasst werden IDN, bevorzugte Benennung und Konfidenzwert des Schlagwortes sowie das Datum der Einspielung. Damit ist der Prozess der automatischen Beschlagwortung beendet.

In der Kataloganzeige werden automatisch beschlagwortete Titel künftig mit dem Hinweis »Automatisch aus dem Text ermittelte Schlagwörter« versehen. Zudem soll bei der Katalogsuche

und Trefferanzeige danach unterschieden werden können, ob die Schlagwörter intellektuell oder maschinell erstellt wurden. Ebenso wird bei der Auslieferung der maschinell erstellten Schlagwörter an die Datendienstbezieher eine Kennzeichnung am Datensatz erfolgen.

Für die Qualitätssicherung wird ein Stichprobenverfahren auf der Grundlage des beschriebenen Testverfahrens implementiert. Über einen längeren Zeitraum kumuliert, sollen auf diese Weise Kennzahlen zur Güte des Erschließungsverfahrens im Produktivbetrieb gewonnen und systematische Fehler identifiziert werden. Eine wichtige Erkenntnis aus den Qualitätstests ist z. B., welche Schlagwörter häufig als »falsch« beurteilt wurden. Oft ist ein Blick in die Protokolle zur linguistischen Verarbeitung ausreichend zur Beurteilung, ob und wie ein Term künftig im Wörterbuch behandelt werden sollte. Durch die Arbeit mit dem Wörterbuch ergeben sich auch wertvolle Hinweise für die Pflege der GND.

## Ausblick

Bis Ende 2013 laufen noch die Vorbereitungen zur Inbetriebnahme der automatischen Beschlagwortung, insbesondere Tests der Geschäftsprozesse und weiterer technischer Routinen. Ab Jahresbeginn 2014 soll eine erste Konfiguration zur Beschlagwortung deutschsprachiger elektronischer Hochschulschriften gestartet werden. Weitere Konfigurationen für andere Objektgruppen sollen folgen.

Die automatische Beschlagwortung bietet die Chance, Publikationen, die sonst gar nicht oder nur sehr grob thematisch erschlossen sind, mit verbalen Sucheinstiegen zu versehen und damit ihre Auffindbarkeit im Retrieval zu erhöhen. Die Grenzen einer automatischen Beschlagwortung beginnen dort, wo an die inhaltliche Erschließung der Anspruch gestellt wird, eine möglichst eindeutige, d. h. spezifische und nicht redundante, Essenz eines Textes zu formulieren. Dieses ureigene Geschäft der Dokumentare, Archivare, Bibliothekare und anderer Information Professionals bringt auch im Zeitalter der (elektronischen) Massen von Medien ein hervorragendes Ergebnis an inhaltlicher Erschließung<sup>20</sup>. Es muss daher gar nicht der

Stichprobenverfahren als Basis für die Qualitätssicherung

Ab 2014 automatische Beschlagwortung deutschsprachiger elektronischer Hochschulschriften

Versuch unternommen werden, beide Erschließungsformen mit demselben Maßstab zu messen, auch wenn das Resultat beider Verfahren im hier beschriebenen Szenario – ein aus GND-Vokabular bestehendes Indexat – dies scheinbar suggeriert<sup>21)</sup>. Automatische Beschlagwortung ist immer abhängig von dem zugrunde liegenden Text und der zur Beschreibung genutzten Terminologie, also von den Begriffen, die vorhanden sind oder eben auch nicht. Das Erkennen von inhaltlichen Zusammenhängen ist auch maschinell noch durchaus möglich, beispielsweise auf der Basis von Verknüpfungen, Kookkurrenzen oder der sachlichen Zuordnung einzelner Terme, aber eine Abstraktion des Inhaltes eines Textes dagegen nicht. Auch muss für automatische Verfahren eine bestimmte Fehlerquote in Kauf genommen, und ein gewisser Kontrollverlust akzeptiert werden, denn die Verarbeitung von großen Dokument- und Datenmengen lässt lediglich eine Stichprobenkontrolle zu. Dennoch

stellt der Einsatz automatischer Beschlagwortung auf jeden Fall einen Gewinn für das Suchen und Finden dar.

Neben den eingangs bereits erläuterten geänderten Bedingungen im Medienmarkt und stagnierenden oder rückgängigen Personalkapazitäten, wandelt sich auch zunehmend das Selbstverständnis der Informations- und Dokumentationsinstitutionen bezüglich ihrer Aufgaben und deren öffentlicher Wahrnehmung<sup>22)</sup>. Das Internet hat die Produktion, Zirkulation und Bewahrung von digitalen Daten, Informationen und Wissen massiv verändert. Die Ära der Wissensverknappung ist einer schrankenlosen Informationsfülle gewichen. Das macht neue Filter erforderlich, die alten Strategien der Informationsreduzierung greifen angesichts des Ozeans vernetzten Wissens nur noch bedingt<sup>23)</sup>. Automatische Verfahren als eine Möglichkeit, die Erschließung von Medien zu erweitern, werden daher auch in der DNB künftig verstärkt zum Einsatz kommen.

## Anmerkungen

- 1 Schwens, Ute; Wiechmann, Brigitte: Netzpublikationen in der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 21 (2009) 1, S. 10 - 13.
- 2 Online unter: <<http://www.gesetze-im-internet.de/dnbg/index.html>>  
(Letzter Zugriff auf alle im Artikel angegebenen Online-Quellen: 04.08.2013)
- 3 Berners-Lee, Tim: The world wide web: past, present and future. Online unter:  
<<http://www.w3.org/People/Berners-Lee/1996/ppf.html>>
- 4 DNB Bestandszahlen siehe Jahresbericht 2012, S. 40 ff. Online unter: <<http://files.dnb.de/jahresbericht2012/>>
- 5 Siehe auch die interessanten, weil (selbst)kritischen, Überlegungen verschiedener Fachkollegen zu Formaten, Regelwerken und Erschließungsarbeit im Sammelband: Radical cataloging: essays at the front. Ed. by K. R. Roberto. - Jefferson & London: McFarland, 2008.
- 6 Vgl. Rowlands, Ian [u.a.]: The Google generation: the information behaviour of the researcher of the future. In: Aslib Proceedings, 60 (2008) 4, S. 290 - 310. Online catalogs: What Users and Librarians want: an OCLC report (2009).  
Online unter: <<https://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf>>
- 7 Ein von Jürgen Krause bereits 1996 vorgeschlagenes »Schalenmodell der Informationserschließung« geht von verschiedenen Niveaus der Datenrelevanz und der daher notwendigen Inhaltsererschließung aus. Der Kern enthält dabei die Literatur mit der höchsten Relevanz. Er wird möglichst tief und qualitativ hochwertig erschlossen. Innerhalb der folgenden Schalen lockern sich die Relevanzbedingungen und parallel dazu die Anforderungen an die Qualität der Inhaltsererschließung. Wie viele Schalen angesetzt werden und welche Merkmale sie definieren, richtet sich nach den Gegebenheiten eines Fachgebietes und/oder eines Sammelauftrages in Abhängigkeit von den Anforderungen der jeweiligen Nutzergruppen. Siehe Krause, Jürgen: Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung: Schalenmodell. Bonn, 1996 (IZ-Arbeitsbericht ; 6).  
Online unter: <[http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis\\_reihen/iz\\_arbeitsberichte/ab6.pdf](http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/iz_arbeitsberichte/ab6.pdf)>
- 8 Auch die IFLA Richtlinien für Inhaltsererschließung in Nationalbibliografien von 2012 gehen davon aus, dass die Entscheidung für verschiedene Qualitätslevel letztlich auf der Basis des Zusammenspiels vieler Faktoren geschieht: neben den Benutzerwünschen, den

- Objekttypen und der den Erschließungsinstrumenten (Schlagwortsprache, Klassifikation) innewohnenden Qualität sind dies z. B. Budget, Personalressourcen und Publikationsaufkommen. Siehe *Guidelines for subject access in national bibliographies*. Ed. by Yvonne Jahns. Berlin: De Gruyter Saur, 2012. IFLA Series on Bibliographic Control; Nr 45. - S. 37 ff.
- 9 Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: *Dialog mit Bibliotheken*, 22 (2010) 1, S. 20 - 22.
- 10 Schöning-Walter, Christa: PETRUS - Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. In: *Dialog mit Bibliotheken*, 22 (2010) 1, S. 15 - 19.
- 11 Das PETRUS-Projekt dauerte von 2009 bis 2011. Mehrere Erschließungsverfahren aus den einzelnen Teilprojekten sind mittlerweile in den Routinebetrieb übergegangen. Siehe Beyer, Christian; Trunk, Daniela: Automatische Verfahren für die Formalerschließung im Projekt PETRUS. In: *Dialog mit Bibliotheken*, 23 (2011) 2, S. 5 - 10. Mödden, Elisabeth; Tomanek, Katrin: Maschinelle Sachgruppenvergabe für Netzpublikationen. In: *Dialog mit Bibliotheken*, 24 (2012) 1, S. 17 - 24.
- 12 Zum Vergleich Volltextindexierung versus (intellektuelle) Indexierung mit kontrolliertem Vokabular s. bspw. Savoy, Jacques: *Bibliographic database access using free-text and controlled vocabulary: an evaluation*. In: *Information Processing and Management*, 41 (2004), S. 873 - 890. Beall, Jeffrey: *The weakness of full-text searching*. In: *The Journal of Academic Librarianship*, 34 (2008) 5, S. 438 - 444.
- 13 <<http://www.averbis.de>>
- 14 Zu den nachfolgenden technischen Details siehe Benutzerhandbuch Averbis Extraction Platform Version 2.0.3 (März 2013).
- 15 Siehe <[http://www.dnb.de/DE/Standardisierung/GND/gnd\\_node.html](http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html)>
- 16 <<http://termbrowser.de>>
- 17 Bei Katalogisierungslevel 1 sind alle für den jeweiligen GND-Satztyp erforderlichen Datenelemente vorhanden. Sie sind regelkonform ermittelt, dem GND-Standard entsprechend angesetzt und redaktionell geprüft. Bei Teilbestand s = Sacherschließung handelt es sich um den Teil der GND-Datensätze, der für die Inhaltserschließung genutzt wird.
- 18 Stand: 21.01.2013.
- 19 Precision und Recall sind in der üblichen Anwendung auf binäre Relevanzbewertungen bezogen. Kekäläinen und Järvelin haben eine Erweiterung auf gestufte Relevanzbewertungen eingeführt, die im DNB-Kontext bezogen auf Schlagwörter genutzt wird. Siehe Järvelin, Kalervo; Kekäläinen, Jaana: *Using graded relevance assessments in IR evaluation*. In: *Journal of the American Society for Information Science and Technology*, 53 (2002), 13. - S. 1120 - 1129.
- 20 Zum Nutzen und zum Verstehen intellektueller verbaler Erschließung für und im Rechercheprozess s. u. a. Gross, T.; Taylor, A. G.: *What have we got to lose? The effect of controlled vocabulary on keyword searching results*. In: *College & Research Libraries*. - 66 (2005) 3, S. 212 - 230. Taylor, A. G.: *On the subject of subjects*. In: *The Journal of Academic Librarianship*. 21 (1995) 6, S. 484 - 491. Salaba, Athena: *End-user understanding of indexing language information*. In: *Cataloging & Classification Quarterly*. - 47 (2009) 1, S. 23 - 51.
- 21 Pro und Contra beider Verfahren sind pointiert, aber mit Klarsicht dargestellt im Kapitel »Automatic indexing versus manual indexing«. In: De Keyser, Pierre: *Indexing: from thesauri to the semantic web*. - Oxford: Chandos, 2012. S. 39 - 63.
- 22 Siehe z. B. Bonte, Achim; Ceynowa, Klaus: *Bibliothek und Internet: die Identitätskrise einer Institution im digitalen Informationszeitalter*. In: *Lettre International*. - (2013) 100, S. 115 - 117. Darnton, Robert: *The library in the new age*. - In: *The New York Review of Books*, vom 12.06.2008, S. 72 - 80. Internationale Perspektiven finden sich im Sammelband: *Conversations with catalogers in the 21st century / ed. by Elaine E. Sanchez*. - *Libraries Unlimited Library Management Series*. Santa Barbara: Libraries Unlimited, 2011.
- 23 Siehe z. B. die Analysen und Ideen von Weinberger, David: *Too big to know*. - Bern: Huber, 2013.