universität wien

MASTER THESIS

Titel der Master Thesis / Title of the Master‘s Thesis

„Vergabe von DDC-Sachgruppen mittels eines Schlagwort-Thesaurus“

verfasst von / submitted by

Dipl. Tonmeister Sebastian Gabler

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (Library and Information Studies) (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / Postgraduate programme code as it appears on the student record sheet:	UA 992 600
Universitätslehrgang lt. Studienblatt / Postgraduate programme as it appears on the student record sheet:	Library and Information Studies (MSc)
Betreut von / Supervisor:	Dr. Christoph Steiner

Zusammenfassung

Vorgestellt wird die Konstruktion eines thematisch geordneten Thesaurus auf Basis der Sachschlagwörter der Gemeinsamen Normdatei (GND) unter Nutzung der darin enthaltenen DDC-Notationen. Oberste Ordnungsebene dieses Thesaurus werden die DDC-Sachgruppen der Deutschen Nationalbibliothek. Die Konstruktion des Thesaurus erfolgt regelbasiert unter der Nutzung von Linked Data Prinzipien in einem SPARQL Prozessor.

Der Thesaurus dient der automatisierten Gewinnung von Metadaten aus wissenschaftlichen Publikationen mittels eines computerlinguistischen Extraktors. Hierzu werden digitale Volltexte verarbeitet. Dieser ermittelt die gefundenen Schlagwörter über Vergleich der Zeichenfolgen Benennungen im Thesaurus, ordnet die Treffer nach Relevanz im Text und gibt die zugeordneten Sachgruppen rangordnend zurück. Die grundlegende Annahme dabei ist, dass die gesuchte Sachgruppe unter den oberen Rängen zurückgegeben wird.

In einem dreistufigen Verfahren wird die Leistungsfähigkeit des Verfahrens validiert. Hierzu wird zunächst anhand von Metadaten und Erkenntnissen einer Kurzautopsie ein Goldstandard aus Dokumenten erstellt, die im Online-Katalog der DNB abrufbar sind. Die Dokumente verteilen sich über 14 der Sachgruppen mit einer Losgröße von jeweils 50 Dokumenten. Sämtliche Dokumente werden mit dem Extraktor erschlossen und die Ergebnisse der Kategorisierung dokumentiert. Schließlich wird die sich daraus ergebende Retrievalleistung sowohl für eine harte (binäre) Kategorisierung als auch eine rangordnende Rückgabe der Sachgruppen beurteilt.

Abstract

The construction of a thematically ordered thesaurus based on the subject headings of the Gemeinsame Normdatei (GND) using the DDC-Notations contained therein is presented. The DDC subject groups of the German National Library will be the top level of this thesaurus. The construction of the thesaurus is rule-based using Linked Data principles in a SPARQL processor.

The thesaurus is used for the automated extraction of metadata from scientific publications by means of a computational linguistic extractor. For this purpose, digital full texts are processed. The extractor determines the keywords found by comparing the character strings from the full text with concept labels, ranks the hits according to their relevance in the text, and returns the assigned subject categories in rank order. The basic assumption is that the correct subject group is returned among the upper ranks.

In a three-step procedure, the performance of the method is validated. For this purpose, a gold standard is first created from documents retrievable from the DNB online catalog using metadata and findings from a short autopsy. The documents are distributed over 14 of the subject categories with a batch size of 50 documents each. All documents are indexed with the extractor and the results are documented. Finally, the information retrieval performance is assessed for both hard (binary) categorization and rank-ordered return of the subject categories.

Inhaltsverzeichnis

Abkürzungsverzeichnis

BASE	Bielefeld Academic Search Engine
BNF	Bibliothèque nationale de France
DDC	Dewey-Dezimalklassifikation
DFG	Deutsche Forschungsgemeinschaft
DNB	Deutsche Nationalbibliothek
GND	Gemeinsame Normdatei
HTML	Hyper Text Markup Language
http	Hypertext Transfer Protocol
IRI	Internationalized Resource Identifier
IDN	Identifikator der DNB
ISKO	International Society of Knowledge Organization
ISO	International Standards Organization
KI	Künstliche Intelligenz
KOS	Knowledge Organisation System
LCSH	Library of Congress Subject Headings
LoC	Library of Congress
MACS	Multilingual Access to Subjects
MARC	Machine-Readable Cataloging
ML	maschinelles Lernen
MRR	mean reciprocal rank
OCLC	Online Computer Library Center
OWL	Web Ontology Language
PDF	Portable Document Format
PETRUS	Prozessunterstützende Software für die digitale DNB
POS-Tagger	Part-of-Speech-Tagger
RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RDA	Resource Description and Access

RDF	Resource Description Framework
RSWK	Regeln für den Schlagwortkatalog
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Protocol and RDF Query Language
STW	Standard Thesaurus Wirtschaft
SVM	Support Vector Machine
SWD	Schlagwortnormdatei
TF/IDF	Term Frequenz/Inverse Dokument Frequenz
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VIAF	Virtual International Authority File
W3C	World Wide Web Consortium
XML	Extensible Markup Language
ZBW	Leibniz-Informationszentrum Wirtschaft

Konventionen

Die vorliegende Arbeit beschäftigt sich mit Webtechnologien und enthält somit zahlreiche Hyperlinks als Referenzen. Wenn die Quelle selber einen Permalink anbietet, wurde dieser genutzt. Ansonsten wird der vollständige, zum Zeitpunkt des Aufrufs gültige Link als Quelle vermerkt. Auf URL-Kürzungsdienste wurde verzichtet, da diese eventuell nicht dauerhaft zu Verfügung stehen. Die Zugriffe erfolgten alle im Zeitraum der Erstellung der Arbeit (Mai 2021-Oktober 2021) und wurden vor Abgabe vollständig erneut überprüft. Daher gilt für das Zugriffsdatum das Abgabedatum der Arbeit, falls nicht anders angegeben.

Zitationen sind in der Form (>Verfasser<, >Jahr<) angegeben und verweisen auf eine ausführliche Bibliografie im Anhang. Für Monografien sind zusätzlich Seitenzahlen im Format „p(p) ###“ angegeben. Beispiel: (Gödert & Hubrich, 2014, pp. 27 -53)

Angaben zu bestimmten RDF-Typen oder Prädikaten erfolgen auch im Fließtext unter der üblichen Kurzbezeichnung des Namensraums der jeweiligen Ontologie anhand des SPARQL-PREFIX, siehe Anhang 2. Beispiel: „skos:Concept“.
Auf eine Auszeichnung der Syntax der Codebeispiele im Anhang wurde zugunsten der Druckbarkeit in Graustufen verzichtet.

1. Einführung

In den letzten 15 Jahren kommen im Bereich der Wissenschaftlichen- und Nationalbibliotheken vermehrt automatisierte Erschließungsverfahren zum Einsatz. Insbesondere durch umfangreiche Retrodigitalisierungsprojekte sowie die Erweiterung des Sammlungsauftrags auf rein elektronische Ressourcen, hat sich die Notwendigkeit ergeben diese technischen Möglichkeiten vermehrt zu nutzen. Dies geschieht sowohl um Metadaten automatisch zu gewinnen, aber auch um den intellektuellen Erschließungsvorgang zu unterstützen.

Meist finden hierbei KI-basierte Methoden, insbesondere maschinelles Lernen (ML), Verwendung (vgl. Golub, 2019). Für diese besteht das sogenannte „Kaltstartproblem“, d.h. der Maschine muss das Problem beigebracht werden; in der Regel geschieht das durch ein annotiertes Referenzkorpus. Ein solches steht oft nicht für alle Fachgebiete zu Verfügung (siehe hierzu auch: Kapitel 2.5). Ein weiterer Nachteil von ML-Methoden ist ihre prinzipielle Intransparenz: Die Leistungsfähigkeit wird alleine anhand der von einer *Blackbox* getroffenen Entscheidungen beurteilt. Wie die Maschine zu ihren Entscheidungen kommt, und ob sie eventuell einen Fehler gemacht hat, ist zumeist kaum nachvollziehbar.

Ein weiteres Verfahren ist die Nutzung einer Beschreibungslogik, insbesondere in Form eines Thesaurus, in Verbindung mit einer statistisch gewichteten Indexierung des Volltextes. Auch hier sind umfangreiche Vorarbeiten zu erledigen, insbesondere die Erstellung des Thesaurus. Dies mag dazu beitragen, dass diese Vorgehensweise in der Literatur seltener beschrieben wird (vgl. Golub, 2019, Kapitel 3.3.3).

Allerdings ist das Verfahren sehr leistungsfähig und weist komplementäre Eigenschaften zu den KI-basierten Methoden auf. Insbesondere sind die von einer Beschreibungslogik unterstützten oder getroffenen Entscheidungen immer nachvollziehbar.

Für diese Arbeit fokussiert sich die Erschließung auf die Vergabe von DDC-Sachgruppen. Diese haben gegenüber Kurznotationen der obersten drei Dewey-Klassen den Vorzug das Publikationsaufkommen fachlich in etwa 100 Kategorien grob einzuordnen. Sie sind sowohl von ihrer Anzahl her übersichtlich als auch von Benutzer_Innen verstehbar, da sie sich auf diesen meist vertraute Themengebiete beziehen.

Genaue, vollständige und konsistente Metadaten in dieser Hinsicht sind also besonders geeignet, einen Mehrwert für Autor_Innen, Erschließer_Innen und Benutzer_Innen, sowie auch für zur feineren Erschließung eingesetzte Maschinen zu schaffen.

1.1 Verbale Beschreibung und Klassifikation

Traditionell wird in der bibliothekarischen inhaltlichen Erschließung zwischen verbaler Erschließung (der Indexierung mittels natürlicher Sprache) und der klassifikatorischen Erschließung von Dokumenten (der Beschreibung mittels einer kunstsprachlichen Notation) unterschieden.

Hinsichtlich einer späteren Auffindbarkeit von Dokumenten oder Fakten dienen beide Verfahren demselben Ziel, nämlich der Beschreibung mit Indextermen. Diese erlauben es in der Folge aus einer zunächst unübersichtlichen Menge an Informationsquellen diejenigen Einträge aufzufinden oder nachzuweisen, mit denen ein konkretes Informationsbedürfnis befriedigt wird.

Jüngere technische Entwicklungen lassen die Trennung verbaler und klassifikatorischer Ansätze zunehmend verschwimmen. Die Funktion einer Klassifikation für die Aufstellung (also die räumliche Auffindbarkeit) ist letztlich aus bibliothekarischer Sicht für die Organisation elektronischer Ressourcen obsolet. Ihre Funktion zur thematischen Ordnung jedoch hat an Relevanz erheblich gewonnen.

Da eine Klassifikation diese Funktion unabhängig von der Repräsentation eines Themas in natürlicher Sprache erfüllen kann, ist sie für die automatisierte Informationsverarbeitung besonders attraktiv. Der Ansatz ist so erfolgreich, dass wesentliche Normdaten inzwischen weitgehend mit der weitest verbreiteten Universalklassifikation, der Dewey-Dezimalklassifikation (DDC), annotiert sind; unter anderem sind das seit den 1990er Jahren die Library of Congress Subject Headings (LCSH) (Chan, 2000), Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU)[1] und die Gemeinsame Normdatei (GND) der deutschsprachigen Nationalbibliotheken[2].
Die Nutzung der DDC geht inzwischen weit über den bibliothekarischen Bereich hinaus, wie das universelle und inzwischen für die Anwendung künstlicher Intelligenz wesentliche Linked Data Projekt WikiData[3] belegt (Voß, et al., 2014). Notationen dienen auch hier als Ankerpunkte für die Konkordanz verschiedensprachiger Normdatensätze.

Etliche Wissenschaftsdatenbanken und Metakataloge bieten die thematische Suche über DDC, im deutschsprachigen Raum z.B. die Bielefeld Academic Search Engine (BASE)[4]:

---

[1] vgl. z.B. https://data.bnf.fr/en/16590035/justice_transitionnelle/rdf.xml
[2] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page
[4] https://www.base-search.net/Browse/Dewey

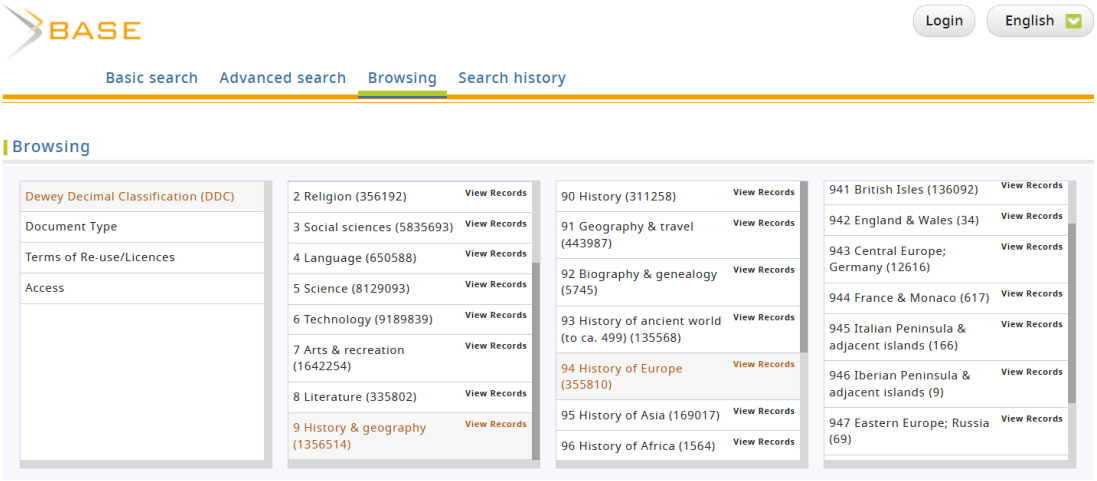


Abbildung 1: DDC Browsing in BASE

In dieser sind auch etwa 200.000 Hochschulschriften von 32 Institutionen aus Österreich verzeichnet, die jedoch oft ohne DDC-Notationen verbleiben. Von den größeren Zulieferern scheinen lediglich das IIASA, sowie der E-Theses Server der Uni Wien systematisch DDC-Kurznotationen zu vergeben.

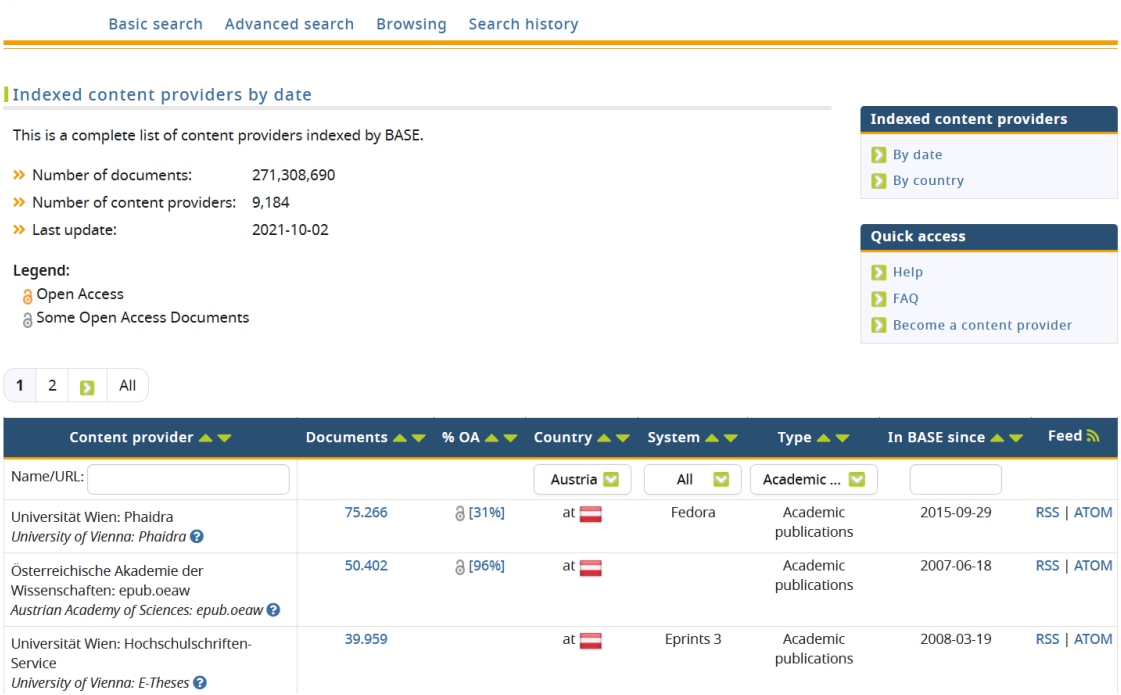


Abbildung 2: Akademische Publikationen aus Österreich in BASE, absteigend nach Dokumentenanzahl

Die Auffindbarkeit österreichischer Publikationen würde sich wohl durch eine systematische Vergabe von DDC-Sachgruppen oder DDC Kurznotationen deutlich verbessern lassen.

Mit der Suchmaschine WebDewey können die Bestände der DNB, mehrerer deutscher Verbände sowie der Staats- und Universitätsbibliothek Göttingen und der FU Berlin thematisch nach DDC-Notationen durchgeblättert werden.

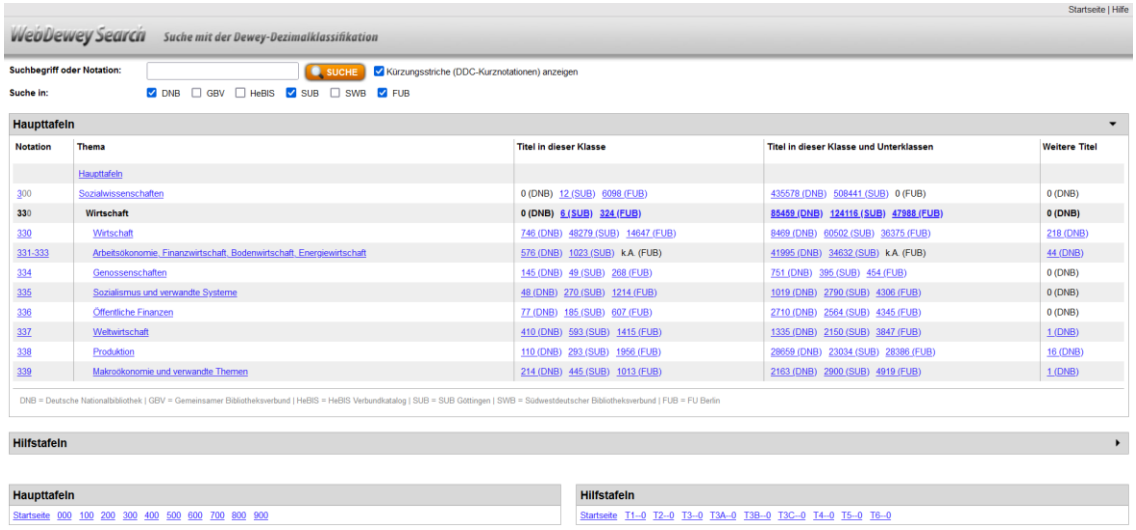


Abbildung 3: Suche in WebDewey

Klassifikationen leisten sprachübergreifend die thematische Sortierung von Wissen. Daher machen sie Bedeutung auch für Agenten nutzbar, die das beschriebene Dokument in seiner Bedeutung nicht vollständig erfassen können. Dies gilt sowohl für mangelnde Sprach- und Fachkenntnisse von Menschen, aber vor allem auch für Maschinen.

1.2 Information Retrieval und Mehrdeutigkeit von Sprache

Die Menge an (elektronischen) Informationsquellen nimmt so rapide zu, dass wir vor immer größeren Herausforderungen stehen, aus diesen Quellen Antworten auf unsere Frage zu finden. Um die Informationsflut bewältigen zu können, bedienen wir uns hierfür Maschinen, insbesondere solchen für das sog. Information Retrieval (vgl. Gödert, et al., 2011, p. 247 ff).

Die derzeit meist verwendete Maschinenart für das Information Retrieval ist die Volltextsuche, bei der ein Stichwort oder auch eine Folge von Stichwörtern mit dem Inhalt eines vorher erzeugten Index verglichen wird. Diese Aufgabe erledigen Maschinen inzwischen äußerst schnell und erfolgreich. Auch wenn Informationssuchende diese Möglichkeit in der Regel als äußerst hilfreich wahrnehmen, erlangt die Maschine hierfür keinerlei Wissen über die Bedeutung der Zeichenketten, ganz zu schweigen von Wissen über die Thematik eines Dokumentes oder die Absicht hinter der Suchanfrage.

Die Bedeutungsebene von Wörtern ist den meisten Maschinen bislang nicht zugänglich. Welche Folgen dies für die maschinelle Indexierung für die bibliothekarische Erschließung haben

kann, hat Heidrun Wiesenmüller in einem Vortrag am Deutschen Bibliothekarstag 2018 in folgenden Stichworten zusammengefasst:

1. *Nicht alle Wörter stehen für Themen*
2. *Nicht alle Themen kann man an Wörtern ablesen*
3. *Falsche oder fehlende Erkennung von Konzepten*
4. *Gewichtung und Auswahl von Schlagwörtern*
   (Wiesenmüller, 2018)

Maschinen können natürliche Sprache zunächst nicht verstehen. Sie können jedoch Sprache als Daten sehr effizient nach verschiedenen Prinzipien sortieren. Das gilt in besonderem Maße für kontrollierte Vokabulare. Die aktuelle Norm und Empfehlung der International Standards Organization (ISO) zur Entwicklung von Thesauri ISO-24965-2011 führt hierzu insbesondere folgende Ordnungsprinzipien auf:

i.) Alphabetischer Index,
ii.) hierarchische Ordnung der Schlagwörter, sowie der
iii.) systematische Index.

Der systematische Zugriff auf ein Vokabular wird dabei durch eine thematische Klassifikation ermöglicht (ISO, 2011, p. 66 f.).

Dabei stehen Systematik und Hierarchie der Schlagwörter zumeist orthogonal zueinander: Systematik macht taxonomisch sortierte Schlagwörter nach thematischen Aspekten zugänglich.

Auch wenn früh erkannt wurde, dass kontrollierte Vokabulare für eine automatische Indexierung und Retrieval erhebliche Vorteile gegenüber einem Volltextindex besitzen - insbesondere indem sie einen Volltextindex mit zusätzlichen Metadaten anreichern und mit verknüpften Klassifikationen einen thematischen Zugriff ermöglichen - verbleiben dennoch zahlreiche Herausforderungen bei der maschinellen Verarbeitung natürlicher Sprache.

Weitere Fallstricke können dadurch entstehen, dass Dokumente unter Umständen nicht durchgehend natürlichsprachig sind. Ein illustrierendes Beispiel dafür liefert Wiesenmüller mit dem wiederholten Vorkommen der Benennung „Etage“ im Inhaltsverzeichnis eines Museumsführers, der sich zudem auf die Gebäudestruktur des Museumsbaus bezieht, aber nicht auf eine vermeintliche Thematik „Geschoss (Bauwesen)“ (vgl. Wiesenmüller, 2018).

1.3 Disambiguierung der Begriffe

Zuvorderst stellt sich allerdings weiterhin das Problem der Mehrdeutigkeit von Sprache: Der Begriff „Knie“ bezeichnet gemeinhin ein Körperteil. Die Zeichenkette hat allerdings auch als ähnlich geformtes Stück eines Abflussrohres Eingang in das Fachvokabular der Sanitärinstallation gefunden.

Eine im deutschsprachigen Raum in verschiedenen Linien tätige Zirkus-Dynastie mit dem Familiennamen Knie hat über Jahrzehnte hinweg mehrere gleichnamige Unternehmen gegründet.

Auch als Straßenname in München-Pasing wurde das Wort für die Benennung eines entsprechend geformten Straßenzuges mitsamt einer gleichnamigen Straßenbahnhaltestelle verwendet.

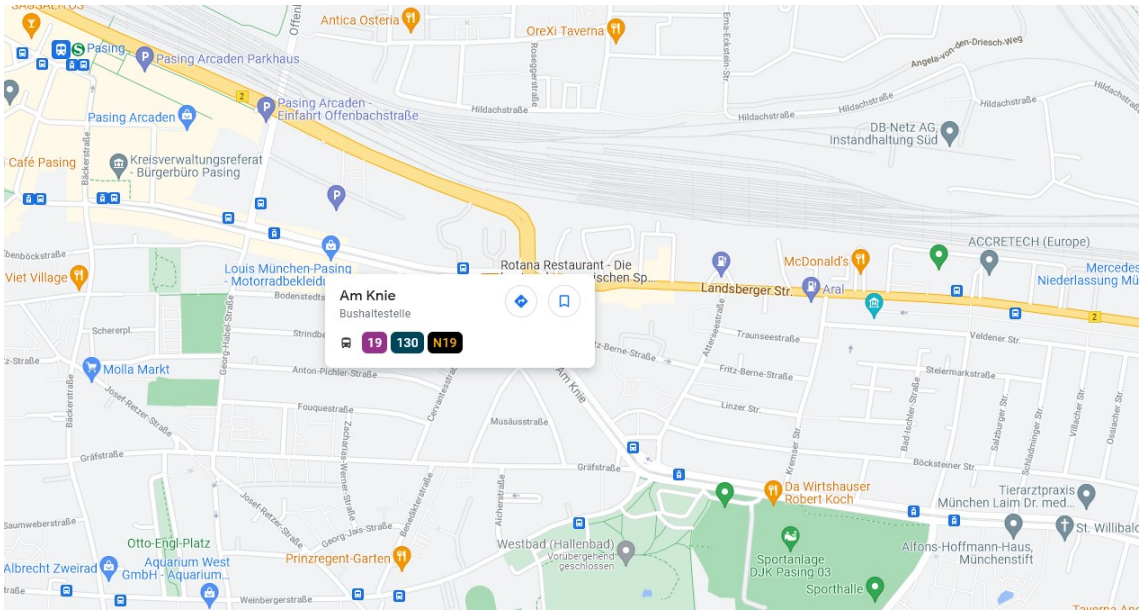


Abbildung 4: Straßenzug „Am Knie“, Google Maps[5]

Diese Menge an Fakten[6] können wir in einer Enzyklopädie nachlesen, oder wir erwerben sie im Lebenslauf als Hintergrundwissen.

Das Bestehen unterschiedlicher Bedeutungen homographer Begriffe können wir Maschinen inzwischen beibringen, zum Beispiel indem wir Konzepte und ihren Kontext in einer maschinenlesbaren Form modellieren, insbesondere in Form von Thesauri, die dann auch die

[5] https://www.google.com/maps/@48.1444038,11.4685601,16z
[6] vgl. https://de.wikipedia.org/w/index.php?title=Knie&oldid=190761493

verschiedenen Bedeutungen des Wortes „Knie“ umfassen könnten. Hiermit sind Bedeutungsunterschiede formal deklarierbar. Dies ist ein erster, notwendiger Schritt zur Disambiguierung.

1.4 Von der Terminologie zum Thesaurus

Die Verwendung von Thesauri für Information Retrieval wurde zuletzt 2011 von der International Organization for Standardization (ISO) formal im Standard ISO-24965-2011 beschrieben. Neben der Beschreibung der Aufgaben und Anforderungen von Vokabularen, die zur Suche von Informationsressourcen, unabhängig von den verwendeten Medien (Text, Ton, Stand- oder Bewegtbild, physisches Objekt oder Multimedia), einschließlich Wissensdatenbanken und Portalen, bibliografischen Datenbanken, Text-, Museums- oder Multimediasammlungen und den darin enthaltenen Objekten verwendet werden, hat das Gremium Datenmodelle und Austauschformate für diese Vokabulare spezifiziert (ISO, 2011).

Wesentliche Beiträge hierzu hat das World Wide Web Consortium (W3C) mit dem Simple Knowledge Organization System (SKOS)[7] geleistet. SKOS ist eine Anwendung des Resource Description Framework (RDF), einer universellen Methode zur Beschreibung beliebiger Dinge im Web.[8] SKOS Konzepte (Begriffe) tragen die Benennungen der Dinge (bevorzugte Benennung und Synonyme) und ermöglichen somit die Verbindung des Thesaurus mit unstrukturierten Texten.

Durch SKOS werden Konzepte die ein definiertes Wissensgebiet umfassen mit Internationalized Resource Identifiers (IRIs) identifiziert, mit Zeichenketten in einer oder mehreren natürlichen Sprachen benannt, sowie mit verschiedenen Arten von Notizen dokumentiert, in informellen Hierarchien und Assoziationsnetzen semantisch miteinander verknüpft und zu Konzeptschemata aggregiert.

Besonderen Stellenwert erhalten hierbei die IRIs, deren Funktion über eine reine Identifikation weit hinausgeht. IRIs-tragende Webressourcen können über standardisierte Übertragungs- und Verzeichnisprotokolle direkt aufgerufen und verarbeitet werden, wodurch natürliche Sprache und Maschinensprache miteinander verbunden werden.[9] Die hierarchischen und

[7] vgl. https://www.w3.org/TR/skos-primer/
[8] vgl. https://de.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=209155194
[9] vgl. https://www.w3c.de/Press/uri-iri-pressrelease.html

assoziativen Beziehungen dienen der Modellierung von Wissen und erlauben in gewissem Umfang aus den Begriffen und ihrem Kontext Schlüsse zu ziehen.

Im Bereich der bibliothekarischen Erschließung wurde von diesen Möglichkeiten umfassend Gebrauch gemacht: Mit den Library of Congress Subject Headings, dem französischen RAMEAU und der Dewey-Dezimalklassifikation sind die wesentlichen Schlagwort- und Klassifikationssysteme als SKOS Thesauri verfügbar.

Durch diese Vorgehensweise ist es zumindest schon einmal möglich, die modellierten Konzepte anhand ihrer Benennungen in einem Text aufzufinden. Auch erfolgt eine ansatzweise Modellierung der Bedeutung, insbesondere die Zuordnung unterschiedlicher Benennungen für dieselbe Sache (Synonymiekontrolle), und die Schaffung unterschiedlicher Konzepte für homonyme und polyseme Benennungen als Keimzelle für eine mögliche Disambiguierung.

1.5 Syntaktische Erschließung und Suche mit spezifischer Semantik

Um Maschinen darüber hinaus umfassendes Sprach- und Hintergrundwissen beizubringen, sind weitere Schritte notwendig. Für diese werden verschiedene Methoden der Wissensrepräsentation, der Computerlinguistik und der Künstlichen Intelligenz verwendet. Im Bereich der DNB gab es hierzu in den letzten Jahren sowohl konkret umgesetzte Projekte, als auch in die Zukunft weisende, konzeptionelle Ansätze.

Besonders einflussreich war und ist dabei der schon erwähnte Kölner Forscher Winfried Gödert. In seinem Artikel „Ein Ontologie-basiertes Modell für Indexierung und Retrieval“ beschreibt er die auch von Wiesenmüller aufgeworfene Problematik, dass nur eine koextensive Inhaltsrepräsentation[10] zu präzisen Suchergebnissen führen kann, sowie dass dabei manche Aspekte bei der Suche nicht im Indexat aufscheinen werden, da sie keine Themen des Dokumentes sind.

Gödert weist ferner darauf hin, dass weder eine postkoordinierende Suche, noch ein rein gleichordnendes Indexieren allein das gewünschte Ergebnis erzielen können. Die Expressivität einer Schlagwortkette ist hierfür nicht ausreichend. Notwendig sind vielmehr syntaktisches Indexieren und die Nutzung syntaktischer Operatoren bei der Suche. Diese Operatoren müssen spezifisch sein, um die nötige Differenzierung leisten zu können.

[10] Also, eine den Inhalt vollumfassend erfassende Indexierung.

Hierzu nutzt Gödert in mehreren Publikationen wiederholt diese Beispiele:

1. *„Gesucht werden Dokumente über den Wandertrieb **von** Singvögeln.“*
2. *„Gesucht werden Dokumente über Singvögel **mit** Wandertrieb“*
   [Unterstreichung der Themen von mir]
   Siehe auch: (Gödert & Hubrich, 2014, p. 149 ff.) sowie (Gödert, 2014).

Im ersten Fall wäre für eine koextensive Inhaltsrepräsentation das Vorkommen beider Themen, sowie deren Beziehung im Dokument zu indexieren. Im zweiten Fall ist Wandertrieb kein Thema des Dokuments, sondern ein Aspekt des Themas Singvögel.

Diese Aufgabe ist für eine Maschine einigermaßen anspruchsvoll, denn hierzu ist eine umfassende Wissensrepräsentation notwendig: während der zweite Fall wohl mit der GND in Verbindung mit der Zoologischen Taxonomie gelöst werden könnte, ist für den ersten Fall nicht nur erforderlich einer Maschine beizubringen, dass der Wandertrieb von Vögeln etwas anderes ist, als beispielsweise die Wanderung von Rentieren; es ist darüber hinaus nötig, dass Maschinen diese einigermaßen spezielle Unterscheidung selbsttätig bei Analyse von Zeichenketten treffen. Dies über Inferenzen einer Beschreibungslogik lösen zu wollen, erfordert doch einigen Aufwand den vor allem für einen Universalthesaurus bisher noch niemand zu leisten vermochte.

Dennoch haben die Ansätze von Gödert wesentlich dazu beigetragen, die aktuellen Entwicklungen bei der Erstellung und Vernetzung von Thesauri im Bereich der deutschsprachigen Bibliotheksverbünde voranzutreiben. Einer der größten Verdienste ist, dass es nunmehr möglich ist, über spezifische Beziehungen zu einer Universalklassifikation thematisch auf einen Großteil der GND-Schlagwörter zuzugreifen, sowie der Umstand, dass wir nunmehr in Form der GND-Ontologie eine Beschreibungslogik zur Auswertung von Schlagwortketten besitzen (siehe Kapitel 3.2.4).

Diese Arbeiten setzen die konzeptionellen Studien in Göderts Publikationen in nutzbare Beschreibungslogik um (Gödert, et al., 2011, p. 197 ff.). Sie sind nicht zuletzt die mittelbare Voraussetzung für die Entwicklung aktueller Erschließungsmaschinen der deutschsprachigen Bibliotheken, wie wir in Kapitel 2.6 sehen werden.

1.6 Thesaurusbasierte Bestimmung der DDC-Sachgruppen

Das hier vorgestellte Verfahren beruht auf der Indexierung eines Textes in natürlicher Sprache gegen den Wortschatz eines kontrollierten Vokabulars. Darauf basierend erfolgt eine

Kategorisierung des Textes entlang einer hierarchischen Zuordnung der im Wortschatz enthaltenen Benennungen eines thematisch geordneten Klassifikationsthesaurus.

107 DDC-Sachgruppen bilden seit 2004 die fachliche Unterteilung der Deutschen Nationalbibliografie (Alex, 2014). Ferner dienen die DDC-Sachgruppen als Facetten in der Suchmaschine der DNB. Die österreichische Bibliografie hat für die Reihe A diese Systematik ebenfalls übernommen.[11]

104 dieser Sachgruppen umfassen eine oder mehrere Klassen der Dewey-Dezimalklassifikation (DDC). In der Regel entsprechen diese Gruppen den Notationen der obersten drei Dewey-Klassen. Etwa zehn der Sachgruppen entsprechen Notationen der Ebenen 4 oder 5. Darüber hinaus bestehen drei Sachgruppen, die mit lateinischen Buchstaben notiert sind.[12] Diese Gruppen sind nicht mit der Dewey-Klassifikation verbunden und spielen im Verlauf dieser Arbeit keine weitere Rolle.

Mit der Zuweisung von DDC-Sachgruppen ist eine grobe Sortierung des Publikationsaufkommens in Fachgebiete verbunden. Zugleich bilden die damit verbundenen Notationen einen über den Bestand hinaus nutzbaren Sucheinstieg für jede nach DDC unterteilte Suchmaschine. Für den Erschließungsprozess bietet eine zuverlässige und konsistente Zuweisung von Sachgruppen eine grobe Orientierung, die einen tieferen Erschließungsvorgang effektiv unterstützt

1.6.1 Indexierung gegen einen Thesaurus

Bei der Indexierung gegen einen Thesaurus erfolgt ein Vergleich der Zeichenketten zwischen Wörtern oder Phrasen der zu indexierenden Dokumenten und den Begriffen des kontrollierten Vokabulars. Die Vorverarbeitung umfasst in der Regel das Entfernen von Stoppwörtern sowie eine linguistische Normalisierung der Lexeme. Da bei Stemming zusätzlich Ambiguitätsprobleme der Grundformen der Wörter auftreten, ist in diesem Anwendungsfall eine (wörterbuchbasierte) Lemmatisierung vorzuziehen.
Die Benennungen des Thesaurus werden aus dem Dokument extrahiert. Anschließend werden ihnen anhand verschiedener Heuristiken Gewichte zugewiesen. Eine ausführliche Beschreibung der für diese Arbeit konkret verwendeten Methodiken findet sich in Kapitel 6.

[11] Österreichische Hochschulschriften sind nach der international wenig verbreiteten Basisklassifikation erschlossen.
[12] B – Belletristik, K – Kinder- und Jugendliteratur und S – Schulbücher

Das hier verwendete Verfahren ist primär die Annotation eines Volltextes angelegt. Wir schaffen also quasi einen automatischen Agenten, der ein Merkmal über ein gesamtes Dokument extrahiert. Diese Funktion ist in der Erschließung vom Menschen schon länger nicht mehr leistbar. Daher ist das Verfahren auch in dieser Hinsicht komplementär gedacht.

1.6.2 Formale Anforderungen an einen Klassifikationsthesaurus

Die gewählte Methode der Kategorisierung basiert auf der Annahme, dass DDC-Sachgruppen für ein Dokument relevant sind, wenn darin bestimmte Benennungen vorkommen die eine oder mehrere DDC-Sachgruppen als transitive Oberbegriffe besitzen.

Für die Erstellung der Beschreibungslogik ist damit eine rekursionsfreie, hierarchische Struktur erforderlich, im einfachsten Fall also eine thematisch aufgebaute Taxonomie. Hierbei handelt es sich um eine *containment hierarchy*[13].

Für diese gilt:

$$x \subsetneq y$$

Dies bedeutet: *y* enthält *x*, wobei (*x, y*) voneinander verschiedene Entitäten sind. Aus dieser Eigenschaft können wir auch schließen, dass jeder Begriff einen Pfad in den Wurzelknoten besitzen muss. Taxonomien sind reine *containment hierarchies*. Die DDC-Klassifikation ist ein Beispiel für eine solche Taxonomie, da sie im Bereich der Haupttafeln monohierarchisch ist. Für Aufstellungssystematiken ist diese Eigenschaft generell gegeben: kann ein Buch doch immer nur einen Standort haben.

Die Unterscheidbarkeit der Entitäten der Beschreibungslogik ist primär durch ihre IDs erfüllt. Entitäten eines Thesaurus beschreiben unterscheidbare Dinge. Die Benennungen der Entitäten können aber durchaus überlappen. Dies kann sowohl durch Ambiguität in den Benennungen (Homographie), oder auch durch Mehrfachbezüge, also Polyhierarchie, geschehen.

Polyhierarchie ist für einen Thesaurus charakteristisch. Bezogen auf die Indexierung von Dokumenten haben wir also jedenfalls mit Mehrfachbezügen in die jeweiligen Oberbegriffe zu rechnen. Polyhierarchische Thesauri sind azyklische, gerichtete Graphen.[14]

---

[13] vgl. https://en.wikipedia.org/w/index.php?title=Hierarchy&oldid=1033323365#Containment_hierarchy

[14] vgl. https://en.wikipedia.org/w/index.php?title=Directed_acyclic_graph&oldid=1025990146

Für diese gilt:

$$x \subsetneq (y, z)$$

Also: *z* und *y* enthalten x; dabei sind (*x, y, z*) voneinander unterscheidbare Entitäten.

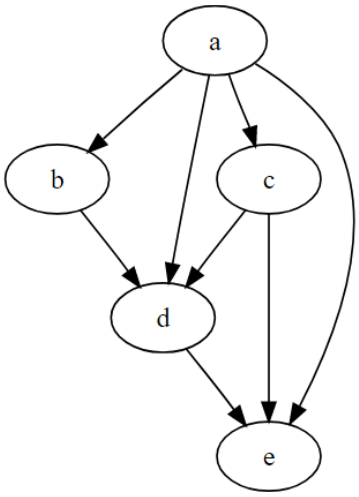


Abbildung 5: Azyklischer, gerichteter Graph, Lizenz Public Domain, https://commons.wikimedia.org/w/index.php?curid=2611155

Auch hier gilt, dass wir immer einen Pfad in den Wurzelknoten besitzen. Das letzte Element vor dem Wurzelknoten stellt die oberste Ebene dar, auf der wir Dokumente kategorisieren können. Der Thesaurus muss also so aufgebaut werden, dass die Oberbegriffe der höchsten Klasse den gesuchten Kategorien entsprechen. Diese Zuordnung kann direkt oder transitiv erfolgen. Transitiv bedeutet, dass aus

> *a ist Oberbegriff von b ist Oberbegriff von c ist Oberbegriff von e*
> *auch*
> *a ist Oberbegriff von e*

geschlossen werden kann (Siehe Abbildung 5).

SKOS- Vokabulare bieten die erforderlichen Möglichkeiten, eine solche Struktur formal zu beschreiben.

Die *containment hierarchy* ist übrigens ein Spezialfall der *nested hierarchy*. Für diese gilt:

$$x \subset y$$

Dies bedeutet: *y* enthält *x*, wobei *x,y* nicht notwendigerweise unterscheidbar sind. Beispiele für *nested hierarchy* wären insbesondere rekursive Pfade, also Schleifen. Diese sind für die gestellte Aufgabe untauglich, da wir damit keinen eindeutigen Pfad in den Wurzelknoten besitzen. Bei Zirkelbezügen laufen wir möglicherweise im Kreis.

2 Relevante Projekte der deutschsprachigen Bibliotheksverbünde

Seit etwa 25 Jahren gibt es Bestrebungen, die Normdaten über die Grenzen von Kultur- und Sprachräumen hinweg zu vernetzen. Webtechnologien sind als technische Voraussetzung dafür bereits länger etabliert.

Umfangreiche Konkordanz- und Erschließungsprojekte nutzen das kontrollierte Vokabular der GND und nutzen die Möglichkeiten der Linked Data Cloud, um die GND mit internationalen Normdatenprojekten, der freien Enzyklopädie Wikipedia, sowie Metakatalogprojekten wie dem Worldcat zu verbinden. Hierüber wird in der Folge ein Überblick gegeben, um den bibliothekarischen Kontext des Vorhabens darzustellen.

2.1 Deutschsprachige Normdaten in Zeiten des Web

Die Entwicklung von Webtechnologien für Daten hatte auch maßgeblichen Einfluss auf die jüngste Entwicklung der Normdaten im deutschen Sprachraum: Die Nationalbibliotheken, Bibliotheksverbünde, und weitere Einrichtungen im Dokumentations- und Museumswesen in Deutschland, Österreich und der Schweiz arbeiteten seit 2009 an einem gemeinsamen Normdatensatz, der Gemeinsamen Normdatei (GND). Mit der Veröffentlichung 2012 vereinte diese die bestehenden Normdaten und stellte sie auf eine mit Webtechnologien nutzbare Plattform.[15] Von Anfang an wurden diese Daten im RDF-Format angeboten, was mit einer grundlegenden Änderung des darunter liegenden Entitätenmodells einherging. Grund hierfür war die Zusammenführung der bisher getrennt geführten Normdateien GKD, PND und SWD sowie der DMA-EST-Datei in einem gemeinsamen Datenformat.

Ein weiterer Auslöser war auch die sich damals abzeichnende Ablösung der Regeln für die alphabethische Katalogisierung (RAK-WB und RAK-Musik) durch das nunmehr eingeführte Regelwerk Resource Description and Access (RDA). Ebenfalls über Webtechnologien erfolgt die Einbindung der GND in das internationale Virtual International Authority File (VIAF).[16]

Die weltweiten Verbindungen der Normdaten lässt sich mit dem Service https://lod-cloud.net interaktiv visualisieren. In Abbildung 6 sehen wir die direkten Verbindungen der LCSH mit anderen Diensten.

[15] https://gnd.network/
[16] http://viaf.org/

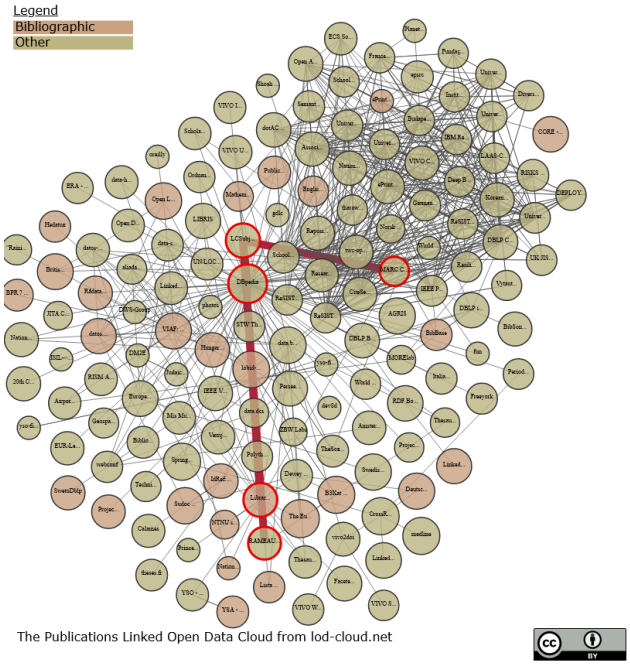


Abbildung 6: Publications-Subcloud, der Knoten LCSH mit seinen direkten Verbindungen ist aktiviert [17]

Insbesondere zeigt die Verbindung zu DBPedia (einer Linked Data Version der gesamten Wikipedia-Enzyklopädie), wie vernetzte Normdaten bereits jetzt über den bibliothekarischen Bereich hinaus Wissen organisieren. Augenfällig ist auch die Verbindung zum Knoten der BNF, sowie über das Service lobid jene in die deutschen Bibliotheksverbünde.

2.2 Sprachübergreifende Wissenserschließung: MACS

Die Verknüpfung der deutschsprachigen Normdaten mit denen der BNF und der LoC war eine der ersten größeren Konkordanz-Projekte. Im Zuge des Projektes *MACS: Multilingual Access to Subjects* wurden ab 1997 die Sachschlagwörter der englischen, französischen und deutschen Sprachbereiche verknüpft.[18] Die geschaffenen Verbindungen sind mittlerweile in den GND Normdaten enthalten und können über Webservices aufgerufen werden.

2.3 Konkordanz der Sachschlagwörter mit der DDC

Für diese Arbeit besonders herauszuheben ist ein weiteres Konkordanz-Projekt: CrissCross.[19,20] Dieses von der Hochschule Köln in Zusammenarbeit mit der DNB durchgeführte Projekt der

[17] https://lod-cloud.net/clouds/publications-lod.svg
[18] https://www.dnb.de/DE/Professionell/Metadatendienste/Metadaten/Voclink/voclink_node.html
[19] https://ixtrieve.fh-koeln.de/crisscross/
[20] Anmerkung: Ein Nebenprodukt von CrissCross war die Modellierung der MACS-Daten in RDF.

Deutschen Forschungsgemeinschaft (DFG) erstellte zwischen 2006 und 2011 unidirektionale Verbindungen zwischen Sachschlagwörtern der Schlagwortnormdatei (SWD) als Quell- und Notationen der Dewey-Dezimalklassifikation (DDC) als Zielsystem.
Wesentliche Aufgabe des Projekts war die Anreicherung des relativen Index der deutschen Ausgabe der Dewey-Dezimalklassifikation (DDC Deutsch) mit Einträgen aus der Schlagwortnormdatei (SWD). Die Konkordanz für zunächst etwa 150.000 Begriffe erfolgte intellektuell und möglichst spezifisch, wozu auch synthetische Notationen in der DDC angelegt wurden (Gödert & Hubrich, 2014, p. 250 ff.). Damit verfügen die GND-Sachschlagwörter in wesentlichen Bereichen über Notationen der international weitest verbreiteten Universalklassifikation.

Darüber hinaus ermöglicht diese Konkordanz einen Zugriff auf die GND-Sachschlagwörter entlang eines alternativen Hierarchiebaumes, nämlich nach dem Aufbau der DDC. Diese Möglichkeit bildet die wesentliche Grundlage für die in dieser Arbeit genutzte Beschreibungslogik.

2.4 Automatische Indexierung von Onlinepublikationen

Das zwischen 2009 und 2011 laufende Programm „Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek“ (PETRUS) implementierte unter anderem die automatische Vergabe von Schlagworten aus dem Wortschatz der GND. Hierbei kamen umfangreiche statistische und computerlinguistische Verfahren zum Einsatz, insbesondere auch eine morphosemantische Behandlung von Komposita und Disambiguierung. Die Auswertung wurde 2013 für 14 der 104 DDC-Sachgruppen in einem Fachbeitrag der *Zeitschrift Dialog mit Bibliotheken* publiziert. Die Werte für Recall für diese Sachgruppen befinden sich demzufolge überwiegend zwischen 65% und 75%; die Werte für Precision liegen zwischen 38% für die Sachgruppe Informatik und 62% für die Sachgruppe Wirtschaft (Uhlmann, 2013).
Die Herausforderungen des Vorhabens liegen insbesondere in der Universalität des GND- Vokabulars. Auch war die Disambiguierung von homographen Sachbegriffen, Personenbegriffen und Geografika letztlich nicht ausreichend.[21] Diese Kritik übt auch Wiesenmüller in ihrem oben angeführten Vortrag in etlichen Beispielen (Wiesenmüller, 2018).

[21] Uhlmann stellte in einer früheren Präsentation von 2011 die zusätzliche Auswertung der beschreibenden Zusätze bzw. die Verwendung von NER zur besseren Unterscheidung verschiedener Entitätsarten in Aussicht. Die Schlussauswertung lässt die Frage offen, ob diese Maßnahmen inzwischen zur Anwendung kamen. Die publizierten Kennzahlen sind jedenfalls 2013 erheblich besser als die 2011 präsentierten.

2.5 ML-basierte Vergabe von DDC-Sachgruppen und Kurznotationen

Ein weiteres in Kontext mit PETRUS stehendes Vorhaben implementiert die automatisierte Vergabe von DDC-Sachgruppen und DDC-Kurznotationen mittels Support-Vector Machine (SVM), einem ML-basierten Klassifikator.[22]

Zusammengefasst werden bei diesem Verfahren unbekannte Dokumente automatisiert einer Klasse zugeordnet. Die hierfür maßgeblichen Eigenschaften lernt die SVM jeweils aus einem pro Sachgruppe angelegten Referenzkorpus von bereits zugeordneten Dokumenten. Das Verfahren eignet sich prinzipiell für die Zuordnung eines Sachgebietes zu einem Dokument, wobei die konkrete Implementierung die Möglichkeit vorgesehen hat, die Sachgruppen nur in besonders eindeutigen Fällen rein automatisch zu vergeben, sonst aber die jeweils drei Sachgruppen mit den höchsten Konfidenzwerten auszugeben, so dass diese intellektuell überprüft und ergänzt werden können.

Eine SVM benötigt überwachtes Lernen, also insbesondere den intellektuellen Input vorklassifizierter Dokumente, sowie eine Auswertung der Trainingsdurchläufe. Die Ergebnisse eines Klassifikators hängen in erster Linie von der Abdeckung des Untersuchungsgegenstandes in den Trainingsdokumenten ab.

Laut den Autorinnen lag die besondere Herausforderung dabei in der Ungleichverteilung der verfügbaren Publikationen: Ein Fünftel der 104 DDC-Sachgruppen umfassen 90% des Publikationsaufkommens, wobei alleine 20.0000 der 45.0000 Publikationen in das Fachgebiet Medizin fielen. Für etliche der verbleibenden Sachgruppen konnten zunächst nicht einmal 50 Dokumente zum Training gefunden werden, so dass hier ersatzweise digitalisierte Inhaltsverzeichnisse gedruckter Monografien hinzugezogen wurden. Die Autoren geben eine Qualitätsziel von 0,7 für das F1- Maß zu Beginn des Produktionsbetriebs an; dieses Ziel war offenbar zum Zeitpunkt der Veröffentlichung des Berichtes noch nicht erreicht (Mödden & Tomanek, 2012). Das Verfahren ist mittlerweile im produktiven Einsatz.

Aufbauend auf diese Vorgehensweise hat die DNB ein Verfahren zur maschinellen Vergabe von DDC-Kurznotationen entwickelt. Dies ist insbesondere für Sachgruppen mit hohen Publikationsaufkommen interessant. Dies kam zunächst für die Reihe O, B und H der

[22] vgl. https://en.wikipedia.org/w/index.php?title=Linear_classifier&oldid=1030098943

Nationalbibliografie für die Sachgruppe 610 – Medizin zur Anwendung und in der Folge auf weitere Sachgruppen ausgeweitet wurde.

Laut einer beim *Dritten Stuttgarter Workshop Computerunterstützte Inhaltserschließung* 2019 publizierten Auswertung beträgt das F1-Maß in 16 untersuchten Sachgruppen zwischen 66% für den Sachgruppe 300 – Sozialwissenschaften und 83% für Sachgruppe 600 – Technik. Im Median liegt die Übereinstimmung bei 67%. Das F1-Maß der Kurznotationen für die Sachgruppe Medizin im Produktivbetrieb lag bei ebenfalls 67% (Busse, 2019).

2.6 PETRUS- Nachfolgeprojekte

Zurzeit sind im Bereich der DNB zahlreiche Nachfolgeprojekte für PETRUS in Arbeit. Dies ist zunächst dem geschuldet, dass die damals zugekaufte Softwareplattform Averbis Extraction Platform abgelöst werden soll, da sie vom Hersteller nicht mehr wie erwartet weiterentwickelt wird. Darüber hinaus möchte man die Erkenntnisse aus PETRUS in die Weiterentwicklung der Verfahren und den aktuellen Stand der Technik einfließen lassen.

2.6.1 Neue Computerlinguistik

Derzeitiger Favorit bei der Ablöse der NLP Plattform ist die open source Plattform Annif[23], ein Projekt der Finnischen Nationalbibliothek. Wie schon Averbis, kombiniert Annif lexikalische und statistische Verfahren mit Künstlicher Intelligenz. In einer frühen Entscheidungsphase des Projekts „Erschließungsmaschine (EMa)“ hat sich herausgestellt, dass diese Plattform zu signifikant besseren Vorschlägen als das Bestandssystem führt (Uhlmann, 2020).

2.6.2 Kollaborative Erschließung

Zuletzt sei noch ein derzeit hoch gehandeltes Projekt aus dem Umfeld der ETH Zürich erwähnt, der Digitale Assistent DA-3.[24] Hierbei handelt es sich um eine Plattform, die sich mit dem vorhandenen Katalogsystem des Anwenders, Verbundkatalogen, sowie Normdaten (GND, LCSH, STW) verbindet. Abbildung 7 visualisiert die prinzipiellen Abläufe:

[23] vgl. http://annif.org/
[24] vgl. https://www.eurospider.com/images/Produkte/DA-3-factsheet.html

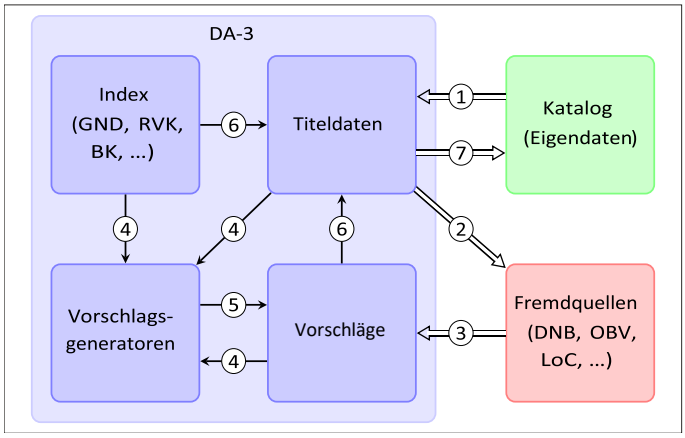


Abbildung 7: Abläufe DA-3; Grafik: Eurospider

DA-3 unterstützt den Anwender mit Vorschlägen aus Index- und Fremddaten. Er erlaubt sowohl die Übernahme von Einzelschlagworten als auch von gesamten Schlagwortketten. Im Bereich der Normdaten nutzt DA-3 insbesondere die Vorarbeiten der Projekte MACS und CrissCross zur Verknüpfung mit LCSH, RAMEAU, sowie die DDC als Rückgrat der sternförmig vernetzten Datensätze. Der Client erlaubt eine Übernahme der Erschließungsdaten in den lokalen Katalog des Anwenders (vgl. Schritt 7 in Abbildung 7).

2.7 Bewertung und Zusammenfassung

Alle Projekte haben gemeinsam, dass sie die intellektuelle Erschließung unterstützen oder Eingriff von Menschen jedenfalls erlauben. Die jüngeren Vorhaben zeichnen sich dadurch aus, dass sie kollaborative Ansätze in den Vordergrund stellen. Man setzt inzwischen auch auf eine modulare Architektur die evolutionär erweitert und entwickelt werden kann.

Demgegenüber steht allerdings eine bestehende Praxis, insbesondere Netzpublikationen und elektronische Ressourcen nurmehr automatisch zu erschließen. Immerhin war die automatisierte Klassifikation und Indexierung der Reihen H und O der Deutschen Nationalbibliografie ausgemachtes Ziel des PETRUS-Programms (Schöning-Walter, 2010). Diesem Ansatz wird angesichts der Informationsflut in diesem Bereich auch kaum noch widersprochen. Automatisch vergebene Schlagwörter kommen auch bei weiteren großen Retrodigitalisierungsprojekten zum Einsatz, so auch beim Projekt AustriaN Newspapers Online (ANNO).[25]

[25] vgl. https://de.wikipedia.org/w/index.php?title=ANNO_%E2%80%93_AustriaN_Newspapers_Online&oldid=214740939#Benutzung

Allerdings hat schon die Ankündigung[26], automatisierte Erschließungsdaten künftig für alle Medienarten (also auch für gedruckte Dokumente) anzuwenden zu einer kontroversen Debatte geführt.

Über einen viel beachteten Beitrag des Generaldirektors der Bayerischen Staatsbibliothek, Klaus Ceynowa in der FAZ hat diese Kontroverse selbst in der breiten Öffentlichkeit Aufmerksamkeit gefunden. Die wesentlichen Bedenken der Kritiker lauteten, ob eine angestrebte Vereinheitlichung der Erschließungsstandards zu einer Nivellierung nach unten führen würde, und ob der neue Gedanke der Erschließung als zyklischem Prozess von den Bibliotheken prozedural und methodisch leistbar sei (Ceynowa, 2017).

Bisher scheinen die Bedenken sich nicht zu bewahrheiten. Die aufgeführten Projekte streben im Wesentlichen eine Vereinheitlichung der Erschließungsstandards im Sinne einer Vernetzung der bestehenden Verfahren, sowie einen erhöhten Kollaborationsgrad an. Die Tatsache, dass Bücher eventuell (zuerst?) von einer Maschine gelesen werden (also, eine Zuspitzung der Überschrift von Ceynowas Beitrag), ist weniger ein durch maschinell unterstützte Erschließungsarbeit erzeugtes Problem, als vielmehr ihre notwendige Voraussetzung.

Kollaborative Ansätze führen unabhängig von der Beteiligung maschineller Agenten dazu, dass ein Produkt, hier also ein Indexat, nicht mehr als abgeschlossen gelten kann.
An diesem Grundsatz agiler Methoden können auch Bibliotheken nicht vorbeigehen. Die Frage ob eine Erschließungsarbeit gut genug ist, beantwortet sich allenfalls darin, ob das nächste Informationsbedürfnis damit befriedigt werden wird. Dies jedoch ist ein sich ständig veränderndes Ziel, das alle mit der Organisation von Wissen befassten Personen und Institutionen fortlaufend herausfordert.

Transparenz des Erschließungsvorganges ist ein Wert, den sämtliche maßgeblich Beteiligten immer wieder als besonders wichtig herausstellen. Wir können davon ausgehen, dass die Provenienz der Werte eines Datensatzes in der Regel gegeben ist, da hierfür im Regelwerk schon immer umfangreich gesorgt wurde. Wenn also eine Inhaltserschließung vom Verleger übernommen wurde, werden wir das auch in den Daten erkennen können. Falsche oder wenig

[26] vgl. https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/konzeptWeiterentwicklungInhaltserschliessung.html

hilfreiche Indexierungen können korrigiert werden. Maschinen sind dabei eventuell ein Teil der Lösung.

Herausfordernd ist die Frage der Transparenz allerdings für Methoden, die letztlich in einer *Blackbox* geschehen, so wie das eigentlich für alle derzeitigen ML-Verfahren zutrifft.

Für die in dieser Arbeit verwendete Methode gilt dies allerdings nicht: da sämtliche Entscheidungen auf Grundlage einer nachvollziehbaren Beschreibungslogik und einfacher Aggregationsverfahren getroffen werden, sind die Resultate immer nachvollziehbar – ob sie nun richtig oder falsch waren.

3 Nutzung der GND-Sachschlagwörter als Indexierungssprache

Die GND-Sachschlagwörter (Begriffe, GND-SH) dienen der Erschließung bibliographischer und archivalischer Ressourcen nach inhaltlichen Kriterien, insbesondere der Indexierung. Die Begriffe beschreiben die Themen, mit denen die Ressource, also ein Dokument, ein audiovisuelles Medium, oder ein Gegenstand verbunden ist (RSWK, 2017, p. 35f.).

Zwar enthalten die GND-SH in der Regel keine Formschlagwörter, dennoch kann ein als Formschlagwort verwendeter Begriff polysem im Vokabular vorkommen, so wie der Begriff „Buch“ in seiner Verwendung als literarisches Motiv. Das Vokabular enthält weitere als Formschlagwörter verwendete Allgemeinbegriffe, wie z.B. „Zeitung“, „Zeitschrift“, „Kapitel“ oder „Dokument“.
Ähnliches gilt für Homographe von Geografika und Eigennamen, die durchaus gelegentlich thematische Aspekte einer Ressource darstellen. Unter anderem kommen diese Individualbegriffe in Schiffs- oder Flugzeugnamen vor.

In der Folge soll erläutert werden, inwiefern die GND-SH als Indexierungssprache zur automatisierten Erschließung eingesetzt werden kann. Zudem wird erklärt werden, warum die ursprüngliche taxonomische Struktur des Vokabulars nicht zur thematischen Kategorisierung geeignet erscheint.

3.1 Entstehung, Bestandsaufnahme und Regelwerke

Die Herkunft der GND-SH ist die vormalige Schlagwortnormdatei, SWD[27], welche ihre Anfänge in den 1980er Jahren hatte. Sie stellt das letztlich erfolgreiche Ergebnis der Bemühungen dar,

[27] https://de.wikipedia.org/w/index.php?title=Schlagwortnormdatei&oldid=211715482

das Vokabular zur verbalen Sacherschließung im deutschsprachigen Raum zu vereinheitlichen (Cappelaro, 2003).

Der Thesaurus wurde nach den „Regeln für die Schlagwortkatalogisierung“ (RSWK)[28] kooperativ erstellt und wird laufend weiterentwickelt. Die RSWK stellen ebenfalls das wesentliche Regelwerk für seine Verwendung dar. Mit der Veröffentlichung 2012 sind die SWD in der Gemeinsamen Normdatei (GND) aufgegangen. Dies bedeutete einen Umbau des Entitätenmodells auf ein neues den Standards des Semantic Web entsprechendes System, sowie im Jahr 2015 einen Umstieg auf das Regelwerk RDA (RSWK, 2017, p. 3). Auch die Abtrennung von Formschlagwörtern, Personen und Geografika wurde in diesem Zusammenhang vollendet.

Der Thesaurus umfasst 213.779 unterscheidbare Begriffe[29,30], die sämtliche in der deutschsprachigen Literatur behandelten Sachgebiete umfassen. Somit lässt es sich ohne Einschränkung als kontrolliertes Vokabular mit Anspruch der universalen Themenabdeckung bezeichnen. Ob es sich auch um einen Universalthesaurus handelt, war in den Anfängen der SWD umstritten (Gödert, 1990).

Diese Kontroverse bezieht sich in erster Linie auf die anfänglich schwache hierarchische Struktur des Vokabulars, die sich in der jüngeren Entwicklung erheblich verbessert hat, wie wir weiter unten sehen werden. Jedenfalls bezeichnet Esther Scheven in ihrem Aufsatz über die GND und den aktuellen ISO- Standard für Thesauri die GND unumwunden als Thesaurus (Scheven, 2017).

3.2 Formale Struktur des Thesaurus

Die Deutsche Nationalbibliothek (DNB) stellt diese Information in verschiedener Form zu Verfügung. Das Interesse für dieses Vorhaben richtet sich insbesondere auf das Format der RDF-Serialisierung.[31]

Es soll angemerkt werden, dass es sich bei einer Serialisierung[32] nicht um Linked Data im eigentlichen Sinne eines Webservice handelt,[33] sondern im Grunde um ein herkömmliches

---

[28] Regeln für die Schlagwortkatalogisierung, 4., vollständig überarbeitete Auflage 2017

[29] Eigene Zählung über SPARQL bezogen auf die unten angeführte Quelle; man beachte die Abweichung der Zählung insbesondere zum oben angemerkten Wikipedia-Artikel, der von 600.000 Vorzugsbezeichnungen bezogen auf das Entitätenmodell und den Entwicklungsstand von 2003.

[30] Siehe hierzu auch Folie 12 in „Einführung in das Projekt PETRUS“, (Schöning-Walter, 2011)

[31] Download im Turtle-Format, siehe https://data.dnb.de/opendata/authorities-sachbegriff_lds.ttl.gz

[32] vgl. https://www.w3.org/TR/rdf11-primer/#section-graph-syntax

[33] vgl. https://www.w3.org/standards/semanticweb/data#summary

Textdokument. Lediglich der innere Aufbau des Dokuments folgt dem RDF-Standard. Daneben stellt die DNB die GND auch als Webservice zu Verfügung. Auch wenn beide Formen dieselbe Information tragen können, sind sie dennoch funktional unterschiedlich: Die Serialisierung ist vor der Nutzung als maschinenlesbare Datenquelle erst in ein entsprechendes Tool zu importieren. Dies hat den Vorteil, dass die durchgeführten Operationen ad hoc festgelegt und auf einer lokalen Quelle mit geringen Wartezeiten durchgeführt werden können. Dieses Vorgehen eignet sich zu Aufbau und Betrieb eines Extraktors, da hier die Daten aus Latenzgründen in einer lokalen Datenbank vorgehalten werden.

Der Thesaurus ist mit den Regeln der Web Ontology Language (OWL, OWL2)[34] modelliert, und bildet damit eine formale Beschreibungslogik[35]. Somit sind die Informationen innerhalb der tatsächlich modellierten Expressivität direkt für eine maschinelle Verarbeitung nutzbar. In der Folge geben wir eine Übersicht über die wichtigsten strukturbildenden Elemente in Referenz zur GND-Ontologie[36]; die URIs der Elemente sind jeweils als Fußnoten belegt. Erläuterungen erfolgen in der Regel anhand der deutschsprachigen Annotationen; wenn hilfreich, werden die englischsprachigen Literale dazugefügt.

3.2.1 Entitätstypen

Der Thesaurus setzt sich aus Entitäten des Typs „Schlagwort“[37] zusammen. Die Klasse „Schlagwort“ ist vom W3C- Standard skos:Concept[38] abgeleitet; beide sind über das Axiom owl:equivalentClass in Beziehung gesetzt. Dies bedeutet, dass sie formal unabhängig sind, aber über eine Vielzahl an ähnlichen Eigenschaften verfügen und somit gleichwertig verwendbar sind. Eine solche Nähe ist nicht zufällig, denn die GND ist über SKOS mit anderen in SKOS publizierten Schlagwortsystemen, insbesondere den LCSH[39] und RAMEAU[40] verbunden.

Ein Schlagwort verfügt über einen (eindeutigen) Internationalized Resource Identifier (IRI)[41] in Form einer URL (Uniform Resource Locator), die Vorzugsbezeichnung ist mit „Bevorzugter

[34] vgl. https://www.w3.org/TR/owl2-overview/#Introduction
[35] vgl. https://en.wikipedia.org/w/index.php?title=Description_logic&oldid=1016320325
[36] vgl. https://d-nb.info/standards/elementset/gnd
[37] https://d-nb.info/standards/elementset/gnd#SubjectHeading
[38] https://www.w3.org/2009/08/skos-reference/skos.html#Concept
[39] https://id.loc.gov/authorities/subjects.html
[40] https://rameau.bnf.fr/informations/rameauenbref
[41] https://datatracker.ietf.org/doc/html/rfc3987

Name des Schlagworts“[42], Synonyme sind mit „Varianter Name des Schlagworts“[43] bezeichnet. Beide Benennungsarten sind als Literale[44] ausgeprägt.

Im Unterschied zu SKOS sind die Literale der GND nicht typisiert, d.h. die verwendete Sprache ist nicht explizit ausgezeichnet. In der Regel sind die Benennungen in der GND der deutschen Sprache entnommen. Es kommen aber durchaus anderssprachige Literale vor, auch solche in anderen Schriftsystemen.

3.2.2 Beziehungsarten

Äquivalenzbeziehungen modelliert der Thesaurus mit dem Attribut „In Beziehung stehendes Schlagwort“[45]. Die überwiegend in der GND verwendete Hierarchiebeziehung ist „Oberbegriff allgemein“[46]. Diese grundlegenden Beziehungsarten haben ebenfalls ein explizit modelliertes SKOS- Äquivalent, skos:related und skos:broader. Somit ergibt sich, dass die hierarchischen Beziehungsarten der GND invers (also, eine gegenläufige Beziehungsart „Unterbegriff allgemein“ besteht ebenfalls), sowie transitiv sind (vgl. Kapitel 1.6.2).

Wie Gödert betont, gilt die transitive Eigenschaft jedoch nur, wenn die Art der Beziehung über die Klassen hinweg gleichförmig bleibt, insbesondere also nicht der Aspekt der Beziehung gewechselt wird. Diese Konstellation ist für ein Vokabular das mit zahlreichen Komposita arbeitet allerdings relativ häufig, so auch Gödert in seinem Aufsatz „Ein Ontologie-basiertes Modell für Indexierung und Retrieval“ (Gödert, 2014).

Da allgemeine Hierarchiebeziehungen nicht erlauben, einen solchen Perspektivwechsel formal zu kodieren, ist eine Differenzierung der Beziehungstypen wünschenswert. Eine Solche ist in der aktuellen ISO-Norm auch vorgesehen. Hierfür existiert eine Erweiterung des SKOS-Standards, ISO-Thes[47]. Die auf den Thesaurus bezogenen Konstrukte der GND-Ontologie stellen im Wesentlichen proprietäre Implementierungen dieser Standards dar.[48]

---

[42] https://d-nb.info/standards/elementset/gnd#preferredNameForTheSubjectHeading
[43] https://d-nb.info/standards/elementset/gnd#variantNameForTheSubjectHeading
[44] Literale bilden die formalen Endpunkte eines Graphen. Sie bilden selber keine Entitäten aus, und können nicht weiter beschrieben werden.
[45] https://d-nb.info/standards/elementset/gnd#relatedSubjectHeading
[46] https://d-nb.info/standards/elementset/gnd#broaderTermGeneral
[47] vgl. http://pub.tenforce.com/schemas/iso25964/skos-thes
[48] Pragmatisch war dies für diese Arbeit ein Vorteil, da SKOS so mühelos auf die bestehenden Daten modelliert werden konnte.

Gemäß ISO-25964 werden Hierarchien in der GND daher in drei weiteren, spezifischeren Typen ausgeprägt: Als Abstraktionsbeziehungen, Ganzes-Teil-Beziehungen und Instanzbeziehungen (ISO, 2011).

„Oberbegriff generisch“[49] wird verwendet, wenn ein Unterbegriff über alle Eigenschaften des Oberbegriffs verfügt und zusätzlich ein Unterscheidungsmerkmal ausprägt. „Oberbegriff instantiell“[50] verwendet man zur Verknüpfung von Individualnamen mit dem übergeordneten Allgemeinbegriff. „Oberbegriff partitiv“[51] verknüpft einen Teilbegriff mit dem übergeordneten Verbandsbegriff.

Die zugehörigen Unterbegriffe sind als inverse Beziehung in der Ontologie definiert, aber im Thesaurus nicht instanziell ausgeprägt, so dass die inverse Relation gegebenenfalls anhand der Ontologie zu inferieren ist. Die Äquivalenzbeziehung ist symmetrisch ausgeprägt, so dass immer beide Richtungen explizit vorhanden sind.

Darüber hinaus verfügt die GND-Ontologie noch über eine Hierarchiebeziehung „Oberbegriff mehrgliedrig“[52]. Dieser Beziehungstyp erlaubt die postkoordinierende Modellierung linguistisch komplexer Sachverhalte.[53] Sie ist derzeit für GND Schlagworte allerdings nicht in Verwendung, was durch eine SPARQL-Abfrage zu verifizieren war.

Die Definition spezifischer, hierarchischer Beziehungen stellt einen wesentlichen Unterschied der GND-Ontologie zum SKOS-Standard dar, der eben nur einen Beziehungstyp hierfür kennt. Konzeptionell würde dies eine verbesserte Expressivität erlauben. In der Praxis kann sich diese besondere Expressivität dann ausprägen, wenn diese Typen konsequent im Thesaurus angewendet werden, was allerdings für die GND-SH zu bezweifeln ist.

Abbildung 8 stellt die Visualisierung des Begriffs „Bankrecht“ mitsamt seinen Unterbegriffen dar. Augenfällig wird, dass viermal der allgemeine- und lediglich einmal der partitive Beziehungstyp gewählt wurde.

---

[49] https://d-nb.info/standards/elementset/gnd#broaderTermGeneric
[50] https://d-nb.info/standards/elementset/gnd#broaderTermGeneric
[51] https://d-nb.info/standards/elementset/gnd#broaderTermPartitive
[52] https://d-nb.info/standards/elementset/gnd#broaderTermWithMoreThanOneElement
[53] Nicht zu verwechseln mit thematisch komplexen Schlagwörtern.

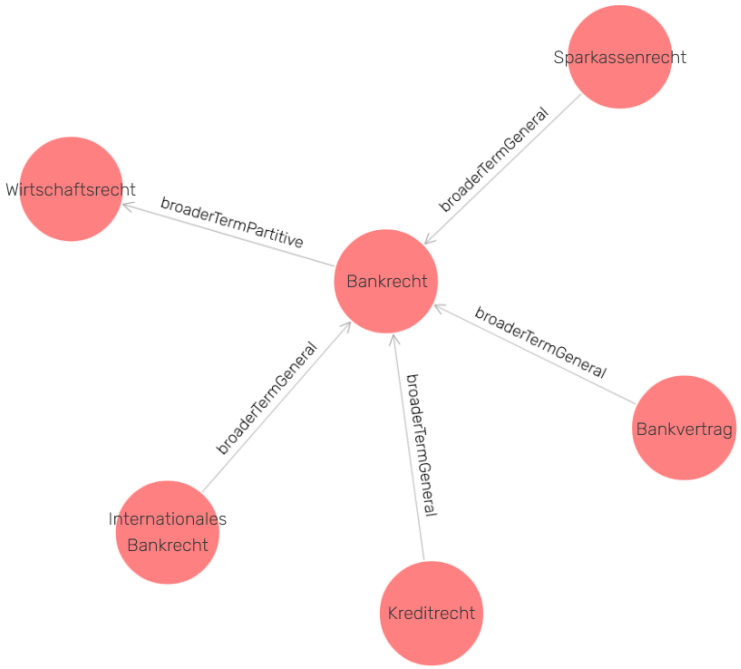


Abbildung 8 Der Begriff Bankrecht mit seinen direkten Unterbegriffen

„Internationales Bankrecht“ ist klar ein generischer Unterbegriff von Bankrecht. In welcher Hinsicht der Bankvertrag hierarchisch eingeordnet wurde, erschließt sich zunächst einmal überhaupt nicht. Für die übrigen, allgemeinen Unterbegriffe hat man zumindest zu konstatieren, dass diese nach heterogenen Merkmalen eingeordnet wurden, mithin unterschiedliche Aspekte zum Tragen kommen. Das Beispiel zeigt also exemplarisch, dass überwiegend Beziehungen es Typs „Oberbegriff Allgemein“ anstelle der spezifischen Beziehungstypen eingesetzt werden.

Über den Einzelfall hinaus lässt sich eine Nutzbarkeit spezifischer Beziehungen für die GND quantitativ falsifizieren: Über den gesamten Thesaurus hinweg hat die allgemeine Beziehungsart einen Anteil von beinahe zwei Dritteln (64,5%, Auswertung über SPARQL). Dies ist insbesondere deshalb bemerkenswert, da sich laut Wikipedia- Dokumentation drei Viertel der Schlagwörter auf Individualbegriffe beziehen[54], also über einen abstrakten Oberbegriff verfügen sollten. Die zugehörige Beziehungsart macht jedoch nur knapp 20% der hierarchischen Attribute aus.

[54] https://de.wikipedia.org/w/index.php?title=Schlagwortnormdatei&oldid=211715482

Selbst wenn man berücksichtigt, dass die Individualbegriffe untereinander generische oder partitive Beziehungen eingehen könnten, somit also nur an Position des obersten Individualbegriffs eine Instanzbeziehung bestünde, ist die häufige Verwendung des allgemeinen Beziehungstyps dennoch allein aus diesem Gesichtspunkt schon nicht mehr plausibel.

3.2.3 Polyseme Begriffe in der GND

Nicht immer kann man aus Begriffen der GND ein Thema eindeutig ermitteln, wie ein prägnantes Beispiel verdeutlichen soll. Betrachten wir den Begriff „Bearbeitung“.

Der Begriff verfügt über DDC-Notationen aus den Sachgruppen 700, 780 und 800; zudem ist er drei unterschiedlichen GND-Sachkategorien (12.1a, 14.4, 31.8a) zugeordnet.

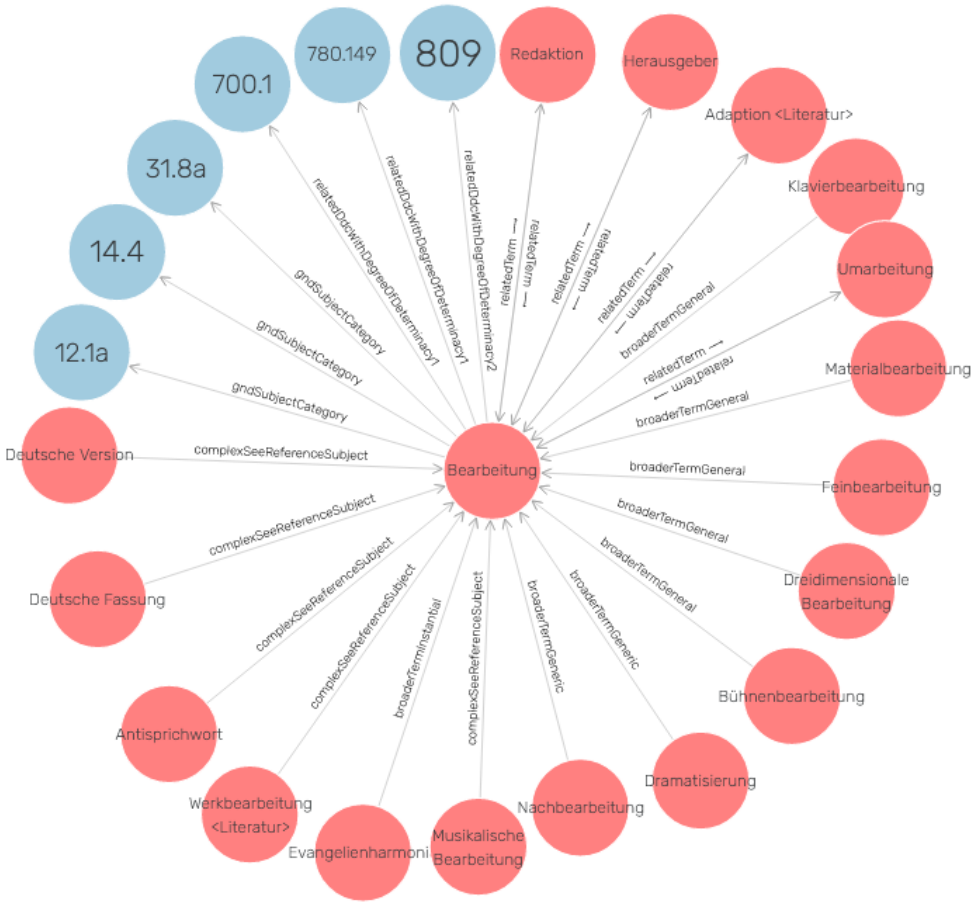


Abbildung 9: Begriff "Bearbeitung" mit Kontext

Die Benennung beschreibt sowohl die Bearbeitung eines industriellen Werkstücks, eines Kunstwerks, oder auch die Bearbeitung von Werken der Literatur oder Werken der Tonkunst. Jedoch handelt es sich sowohl von der Tätigkeit her, als auch aus Sicht der dabei hergestellten Produkte um vier unterschiedliche Begriffe, mithin also um unterschiedliche Themen polysemer

Benennung. Da der Verwendungshinweis einige der unterschiedlichen Bedeutungen ausdrücklich beschreibt, ist dies wohl kein Versehen.

Nach den Regeln der SWSK sollten die unterschiedlichen Bedeutungen gemäß §2,9 als getrennte Begriffe angesetzt und ggf. mit identifizierenden Zusätzen versehen werden (RSWK, 2017).

Des Weiteren sind Unterbegriffe aus allen vier Themenbereichen verbunden, denen ebenfalls jeweils Sachgruppen und Sachkategorien zugeordnet sind. Eine Unterscheidung über die Beziehungsarten der Themen der Unterbegriffe ist ohnehin ausgeschlossen, da diese mehrheitlich den allgemeinen Typ ausprägen, wie oben herausgearbeitet wurde.

Der Eintrag für den Begriff „Bearbeitung“ im DNB Katalog (siehe Fußnote oben) verweist folglich auch auf mehrere Dokumente, die auf die erwähnten heterogenen Themenbereiche bezogen sind. Somit belegt das Beispiel die praktische Relevanz dieser Problematik für die Erschließungsarbeit.

Die fehlenden Ansetzungen der unterschiedlichen Begriffe haben weitreichende Konsequenzen für die Indexierungssprache, die in der Folge kurz umrissen werden sollen:

- Eine thesaurusbasierte (bzw. jegliche) Disambiguierung bei der Indexierung ist für derart angesetzte Begriffe unmöglich
- Die Unterbegriffe sind über transitiven Schluss ebenfalls eventuell falschen Themengebieten zugeordnet.
- Dieselben Probleme bestehen bei einer Verwendung des Schlagwortes als Sucheinstieg

Eine ähnliche Problematik besteht beim Begriff „Pyramide“[55] , der sowohl das Bauwerk, als auch den mathematischen Körper beschreibt. Weitere Beispiele können gefunden werden.[56]

3.2.4 Präkombinierte Begriffe und postkoordinierende Elemente

Eine wesentliche Fragestellung bezüglich der Führung eines Thesaurus liegt in der Frage, ob präkombinierte Schlagwörter oder postkoordinierte Elemente angesetzt werden sollen. GND-SH enthält sowohl präkombinierte Schlagwörter (auch Schlagwortketten, insbesondere als Synonyme), als auch postkoordinierend verwendete Elemente.

[55] vgl. https://d-nb.info/gnd/4047908-0
[56] vgl. https://d-nb.info/gnd/4059317-4 - „Partikel“ (Physik, Verfahrenstechnik)

Die Stärken postkoordierender Recherche betonen § 20 RSWK, aber auch Gödert (Gödert, 1990), (Gödert & Hubrich, 2014). Einen wesentlichen Beitrag hierzu leistet die in den RSWK vorgesehene Zerlegungskontrolle in § 8; die umfassenden Möglichkeiten für die Ansetzung präkombinierter Schlagwörter finden wir in § 305 (RSWK, 2017).

Für eine maschinelle Indexierung aber auch für eine maschinenunterstützte Recherche stellen Schlagwortketten allerdings eher eine Herausforderung dar. Diese waren zunächst rein intellektuell auswertbar. Sie stammten weniger aus der systematischen Terminologiearbeit, als aus der Bearbeitung realer Bücher und waren daher in ihrer Semantik nicht explizit ausgezeichnet (Gödert, 1990).

Zerlegung und Postkoordination leisten hier ja nur dann einen Beitrag zu besseren Suchergebnissen, insofern die Beschreibungslogik diese Vorgänge automatisieren kann. Sowohl Zerlegung als auch Postkoordination erfordern allerdings nicht nur implizites Hintergrundwissen, sondern auch bibliothekarisches Spezialwissen. Beides sind Umstände, die einer Automatisierung entgegenstehen.

Betrachten wir hierzu ein explizites Beispiel aus §8, RSWK und dessen momentane tatsächliche Umsetzung im Thesaurus:

> *SWW s Geometrie; s Mathematikunterricht*
> *BF s Geometrieunterricht*

(RSWK, 2017, p. 54)

„Geometrieunterricht“ würde demnach nicht gesondert angesetzt, da dies kein (eigenständiges) Unterrichtsfach sei. Der Begriff ist allerdings durchaus üblich, wenn man sein Vorkommen in realen Dokumenten betrachtet. So führt er über den Volltextindex zu guten Suchtreffern im Onlinekatalog der DNB, da er unter anderem im Titel einschlägiger Fachbücher vorkommt. Als Schlagwortkette ist er allerdings nur intellektuell zu verarbeiten.

Diese hier ist sogar eine besonders „harte Nuss“ für eine Maschine, da die Komponente „Geometrie“ in der GND ambig ist (Mathematik, Motiv) und die Komponente „Mathematikunterricht“ selbst ein präkoordiniertes komplexes Schlagwort darstellt.

Für den Fall des Geometrieunterricht finden wir die interessante Konstellation in den GND-Daten, dass nunmehr hierfür entgegen dem RSWK Beispiel neben der Schlagwortkette auch

das Kompositum als Begriff zu finden ist. Beide verweisen auf dieselben Komponenten, sind aber ansonsten unverbunden:[57]

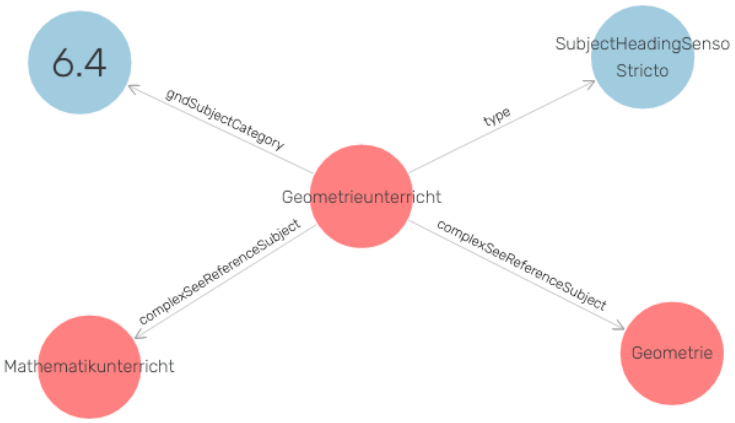


Abbildung 10: Präkombiniertes Schlagwort „Geometrieunterricht“.

Der Thesaurus verfügt mit Einführung der GND hierfür über den Beziehungstyp „Relationierter Deskriptor“[58], der für die GND-SH rund 4.000-mal vorhanden ist. Die Funktion dieses Typs ist eindrücklicher beschrieben, wenn man seine englische Benennung betrachtet. „Complex see reference – subject“ sowie Domain/Range anhand der Definition in der Ontologie heranzieht: Bei der Ansetzung eines komplexen Schlagwortes („Has Range“) kann man somit die mit ihm verbundenen Aspekte (also, wenigstens zwei) zum Ausdruck bringen. Diese Aspekte können wiederum ein Schlagwort, ein Werk, ein Geografikum, ein Personenschlagwort, oder eine Konferenz sein („Has Domain“).

[57] vgl. http://d-nb.info/gnd/7506331-1 und https://d-nb.info/gnd/7506329-3
[58] https://d-nb.info/standards/elementset/gnd#complexSeeReferenceSubject

Verdeutlicht wird das durch eine Visualisierung des Begriffs „Mathematikunterricht“ in den Daten:

Abbildung 11: Der Begriff Mathematikunterricht mit den verknüpften komplexen Schlagwörtern.

Jeweils paarweise sind bedeutungsgleiche Konzepte als Komposita und Schlagwortketten angesetzt; d.h. der Thesaurus besitzt dann zwei Entitäten für dieselbe Sache.

Für eine automatische Indexierung von Dokumenten natürlicher Sprache, sowie für den gewohnten Sucheinstieg für durchschnittliche Benutzende, eignet sich allerdings nur die Form des präkombinierten Schlagwortes; genauer gesagt: diejenigen Benennungen ohne weitere kunstsprachliche Zusätze, da nur diese Form im Wortschatz der Dokumente vorkommt. Die Schlagwortkette bleibt der intellektuellen Erschließung und dem Retrieval für Fachleute überlassen.

Durch diese Verwendung der Semantik leistet die GND nunmehr prinzipiell beide Funktionen, nämlich die Abbildung in der traditionellen Indexierungssprache der Bibliothekar_Innen, als auch die Terminologie in natürlicher Sprache, die für automatische Indexierung erforderlich ist. Insbesondere verbessert die Semantik die Postkoordination funktional:

Bei der Schlagwortfolge

> *SWW s Fluss; s Einzugsgebiet*
> *BF s Flussgebiet*

müssen partielle Treffer für den Bestandteil „Einzugsgebiet“ wirksam unterbunden werden, was rein aufgrund der Schlagwortfolge nicht gegeben wäre.

Der Status Quo zeigt leider auch, dass diese Möglichkeit im Thesaurus noch nicht umfassend genutzt wird. Etwas weniger als 4.700 derartige Schlagwörter und wohl noch einmal deutlich weniger Konzepte unterschiedlicher Bedeutung sind so ausgezeichnet, wie das obige Beispiel verdeutlicht.[59]

Es ist eher nicht ratsam, ein Retrieval-Prinzip auf ein im niedrigen einstelligen Prozentsatz im Thesaurus vorkommendes Konstrukt aufzubauen. Allerdings ist hier ein möglicherweise wegweisender struktureller Ansatz hin zu einer semantisch-syntaktisch nutzbaren Beschreibungslogik gemacht.

Zusammengefasst ist zu sagen: Die semantische Struktur der SWD/GND hat sich seit der Kritik von Gödert, et al. vor 30 Jahren qualitativ erheblich verbessert.[60] Für ein vollständig automatisiertes, syntaktisches Indexieren scheinen die strukturellen und tatsächlich modellierten Möglichkeiten des Thesaurus allerdings qualitativ und quantitativ noch nicht ausreichend. Syntaktisches Indexieren mit der GND ist somit vorerst weiterhin in erster Linie die Domäne ausgebildeter Bibliothekare. Diesen Umstand fasst Gödert in seinem Aufsatz von 2014 prägnant zusammen:

> *„Die in der jüngeren Vergangenheit entwickelten Methoden mit der besten Eignung zur Disambiguierung im Indexierungsergebnis und dem höchsten Potenzial zur Bildung präziser Treffermengen, Kombinationen aus Facettenansätzen mit syntaktischer Indexierung, sind wegen ihres intellektuellen Aufwandes am wenigsten Bestandteil der aktuellen Informationssysteme geworden.“ (Gödert, 2014)*

[59] Die Zahl der Konzepte liegt jedenfalls unter 2.350, da manche Ansetzungen auch drei oder mehr Komponenten enthalten können

[60] Anmerkung; Das Regelwerk wurde auch anderweitig angepasst Zum Beispiel wurde auch das Pleonasmus – Verbot in Schlagwortfolgen aufgehoben (§ 13) und damit der Spielraum einer sinnvollen, syntaktischen Erschließung erweitert.

Genutzt und erprobt für eine automatische Erschließung ist allerdings das gleichordnende Indexieren, d.h. das Auffinden und gleichrangige Nebeneinanderstellen einzelner, unverbundener Schlagwörter.[61] Die aktuelle Ausgabe der RSWK sieht diesen Ansatz durchweg optimistisch:

> *„Es wird davon ausgegangen, dass auch eine gleichordnende Indexierung Kombinationen mehrerer Schlagwörter erzeugt, die einen verstehbaren Kontext ergeben und mit den Schlagwortfolgen weitgehend interoperabel sein sollen.“ (RSWK, 2017)*

Schließlich lässt sich konstatieren, dass die GND derzeit die Möglichkeiten des zugrunde liegenden SKOS-Standards in der Modellierung noch nicht voll ausnützt, ganz zu schweigen von einer substantiellen Nutzung der selbst geschaffenen Fähigkeiten erweiterter Expressivität. Für eine Verwendung als Klassifikationsthesaurus ergibt sich die Erkenntnis, dass die taxonomische Struktur hierfür alleine nicht ausreicht, da sie eben nach begrifflichen und nicht nach thematischen Aspekten geordnet ist. Diese ist also erst zu schaffen (siehe Kapitel 4).

3.3 Formale Analyse der GND-Sachschlagwörter

Die 213.779 Begriffe der GND-SH sind ca. 40 Klassen (GND Entitätstypen) zugeordnet. Hierbei sind Mehrfachzuordnungen möglich. Die Zielsetzung war, den Thesaurus so vollständig wie möglich zu verwenden. Daher wurden keine Entitätstypen a priori ausgeschieden. Allerdings ist eine differenzierte Behandlung der Entitätstypen sinnvoll, um zu einem gut nutzbaren Vokabular zu kommen. Schlagwörter im engeren Sinne der GND sind nur die Instanzen, die einer der 13 Unterklassen der Klasse „Schlagwort“[62] zugewiesen sind. Diese sind in Tabelle 1 mit „*“ gekennzeichnet.

Die Verteilung auf die Entitätstypen stellt sich wie folgt dar.

Entitätstyp	**Anzahl**
gndo:SubjectHeadingSensoStricto*	137412
gndo:NomenclatureInBiologyOrChemistry*	31432
gndo:SubjectHeading	8756
gndo:SoftwareProduct*	8370
gndo:ProductNameOrBrandName*	6284

[61] Für diese genügen im Übrigen flache Schlagwortlisten.
[62] https://d-nb.info/standards/elementset/gnd#SubjectHeading

gndo:Language	5795
gndo:HistoricSingleEventOrEra*	5238
gndo:EthnographicName*	4353
gndo:CharactersOrMorphemes*	3180
gndo:MeansOfTransportWithIndividualName*	1460
gndo:CorporateBody	741
gndo:GroupOfPersons*	584
gndo:BuildingOrMemorial	264
gndo:Work	113
gndo:ProjectOrProgram	112
gndo:ConferenceOrEvent	54
gndo:LiteraryOrLegendaryCharacter	25
gndo:Manuscript	24
gndo:MusicalCorporateBody	18
gndo:TerritorialCorporateBodyOrAdministrativeUnit	12
gndo:WayBorderOrLine	11
gndo:Spirits	10
gndo:PlaceOrGeographicName	8
gndo:Collection	7
gndo:SeriesOfConferenceOrEvent	7
gndo:OrganOfCorporateBody	6
gndo:ExtraterrestrialTerritory	5
gndo:NaturalGeographicUnit	4
gndo:AdministrativeUnit	2
gndo:DifferentiatedPerson	2
gndo:FictiveTerm*	2
gndo:Gods	2
gndo:Country	1
gndo:FictiveCorporateBody	1

gndo:FictivePlace	1
gndo:MusicalWork	1
gndo:NameOfSmallGeographicUnitLyingWithinAnotherGeographicUnit	1
gndo:ReligiousCorporateBody	1
gndo:ReligiousTerritory	1

Tabelle 1:Entitätstypen GND-SH nach Anzahl

Der leichte Überhang zugewiesener Typen gegenüber der Summe der Begriffe (214.300 gegen 213.779) ist durch Mehrfachzuordnungen von Entitätstypen zu erklären.

Nach einem ersten Extraktionsversuch zeigte sich allerdings, dass der Entitätstyp „Buchstaben oder Morpheme“[63] eine besondere Betrachtung erforderlich macht:
Der für diese Arbeit verwendete Extraktor filtert Stoppwörter grundsätzlich, außer sie sind im kontrollierten Vokabular enthalten. In der Gruppe der Morpheme befinden sich etliche typische Stoppwörter, insbesondere auch Konjunktionen sind zahlreich enthalten. Auch sonstige Morpheme kann der Extraktor nicht sinnvoll verarbeiten.

Daher wurden Instanzen, die lediglich über eine Zuweisung dieses Typs verfügen, bei der Erstellung des Klassifikationsthesaurus nicht verarbeitet, obwohl dieser eine Unterklasse von „Schlagwort“ ist.

Bei den übrigen Klassen wurde keine Einschränkung getroffen. Insbesondere wurde auch nicht geprüft, ob die Zuweisungen von Typen außerhalb des strikten Schlagwörter-Kanons rein aufgrund von Mehrfachzuweisungen vorliegen. Das war vor allem daher irrelevant, da a priori nicht absehbar ist, ob es sich dabei um brauchbares Vokabular handelt.

4 Systematischer Zugriff auf die GND-Sachschlagwörter

Die GND-Sachschlagwörter bilden nur eine schwache Hierarchie aus. Insbesondere fehlen ordnende Oberbegriffe, so dass der Thesaurus über 40.000 Begriffe ohne weiteren Oberbegriff aufweist. Hierdurch sind die Zugriffsmöglichkeiten erheblich erschwert, wenn man den Thesaurus zunächst einmal für sich betrachtet.

[63] https://d-nb.info/standards/elementset/gnd#CharactersOrMorphemes

Angesichts des Umfangs des Vokabulars ist keine der in der ISO-Norm vorgesehenen Zugriffsmöglichkeiten geeignet, um einen Überblick über das Gesamtvokabular zu gewinnen.[64]

Für einen systematischen Zugriff stellen die GND Daten zwei Möglichkeiten zu Verfügung: Grundsätzlich sind alle Sachschlagwörter mit Notationen der GND Systematik versehen. Diese Systematik besteht bereits seit Beginn der SWD und wird fortlaufend gepflegt. Für die ca. 134.000 Begriffe, die im CrissCross-Projekt mit DDC-Notationen angereichert wurden und weiterhin in der GND enthalten sind, besteht darüber hinaus grundsätzlich die Möglichkeit, über die Systematik der Dewey-Dezimalklassifikation auf den Thesaurus zuzugreifen.

Vergleichen wir die thematische Einordnung anhand der GND und DDC-Notationen, so fällt auf, dass die DDC-Notationen den Begriff vollständiger und differenzierter beschreiben, als die der GND-Systematik. Das liegt einerseits daran, dass die DDC selber wesentlich granularer ist, als die GND-Systematik. Auch bilden die GND-Notationen nur einen Beziehungstyp aus, die DDC-Notationen in der GND aber deren vier. Anhand des Regelwerkes können wir allerdings auch konstatieren, dass die GND-Systematik eine koextensive Begriffsbeschreibung gar nicht beabsichtigt: Das Ideal dieser Systematik ist die Vergabe einer einzelnen Notation; eine rein aspektorientierte Notation ist in der Regel unzulässig (Deutsche Nationalbibliothek, 2011, p. 3).

4.1 Systematisches Browsing

Die Katalogsuche der DNB beschränkt sich im Wesentlichen auf den Einzelzugriff auf einen Normdatensatz. Hierbei handelt es sich um eine Volltextsuche über die Benennungen. Die folgende Abbildung stellt das Suchergebnis nach dem Sachbegriff „Bearbeitung“ dar.[65]

[64] Einzeldarstellung, alphabetischer Index, hierarchischer Zugriff, systematischer Zugriff und Visualisierung (ISO, 2011, p. 61)

[65] https://portal.dnb.de/opac/showShortList?currentResultId=Bearbeitung%26any%26subjects#inhalt

Abbildung 12: Darstellung des Thesaurus im Katalog der DNB

Das Suchergebnis schließt auch verwandte Begriffe ein; auf diesem Weg hat das oben an erster Stelle dargestellte Ergebnis „Adaption“ Eingang gefunden. Das Ergebnis ist alphabethisch, numerisch, sowie chronologisch sortierbar. Bei einer Normdatensuche in dieser Oberfläche ist es leider gerade nicht möglich eine systematische Vorauswahl zu treffen, während die Titelsuche dieses Prinzip durchgängig anbietet.

lobid-gnd, ein Dienst des Landes Nordrhein-Westfalen, ermöglicht eine facettierte Suche über die GND entlang der GND Systematik.

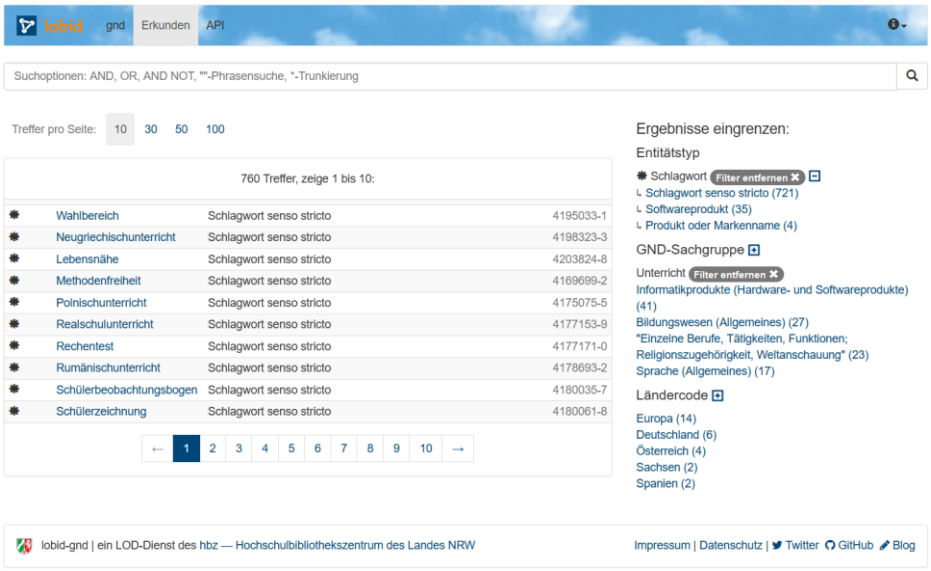


Abbildung 13:Facettierte Suche in lobid-gnd[66]

Die Firma Eurospider, ein spin-off der ETH Zürich, bietet ebenfalls ein alternatives Webinterface für die GND an das auch einen systematischen Zugriff ermöglicht. Neben der GND-Systematik und den DDC-Notationen erlaubt es auch den Zugriff über Untergliederung, Sprachen und geografische Regionen.

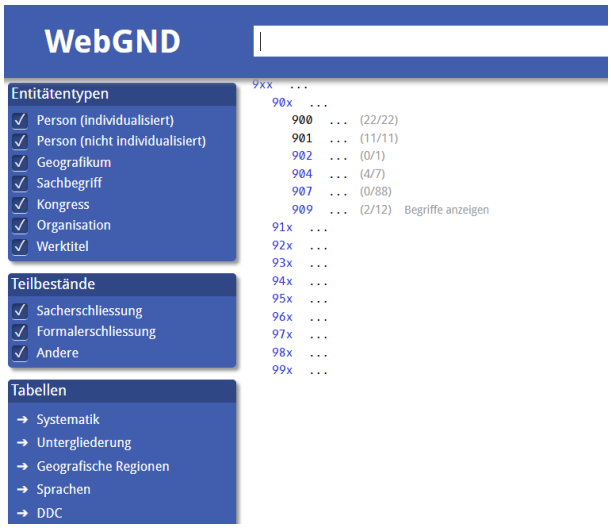


Abbildung 14: WebGND[67]

[66] https://lobid.org/gnd/search?filter=%2B%28type%3ASubjectHeading%29+%2B%28gndSubjectCategory.id%3A%22https%3A%2F%2Fd-nb.info%2Fstandards%2Fvocab%2Fgnd%2Fgnd-sc%236.4%22%29

[67] vgl. http://gnd.eurospider.com/s?table=ddc&node=90

Es ist somit in gewissem Maße möglich, sich einen Überblick über die GND zu verschaffen. Allerdings lässt die Benutzerfreundlichkeit etwas zu wünschen übrig. Wie aus Abbildung 14 ersichtlich, wird nur auf manchen Klassen die Anzahl der enthaltenen Begriffe ausgegeben. Überdies erweckt die weiterhin aktive Facettenanzeige den Eindruck, dass man die Ausgabe eventuell nach dem Entitätstyp filtern könnte. Anstelle dessen bewirkt diese Aktion allerdings eine Rückkehr zur leeren Suchmaske.

Die vorhandenen, öffentlichen Services sind eher als Nachschlagewerke für die gelegentliche Nutzung geeignet. Nur maßgeschneiderte Arbeitsoberflächen, wie der DA-3, erfüllen die Anforderungen professioneller Erschließungsarbeit.

4.2 Nutzung einer Systematik als upper ontology

Insbesondere in der Folge der Fertigstellung des zweiten Teils der ISO-Thesaurus Spezifikation, ISO-25962-2:2013 Interoperability with other vocabularies, beschäftigen sich Bibliothekare / Bibliothekarinnen und Informationsarchitekt_Innen mit der Transformation der GND-SH nach SKOS. Eine Solche ist mit der Anlehnung der GND-Ontologie an SKOS sowie der Bereitstellung von systematischen SKOS-Oberbegriffen ohnehin vorgezeichnet.

Die ISO-Norm liefert einen umfassenden Katalog an Methoden solcher Interoperabilität, einschließlich differenzierter Mappings, generischer Mappings, oder auch dem Verschmelzen mit einer *upper ontology* (ISO, 2013, p. 85 ff.). Solche Transformationen materialisieren zum Beispiel die Notationen der GND-Systematik als Hierarchie, um damit einen systematischen Zugriff auf das Vokabular zu ermöglichen.

4.2.1 Zugriff über GND Systematik

Schon für den Vorgängerdatensatz SWD wurde eine Systematik zu Verfügung gestellt, um thematisch auf den Thesaurus zugreifen zu können. Diese wird mit der GND weiter gepflegt. Ein Regelwerk zu Notation wird als „Leitfaden GND Systematik“ von der DNB zu Verfügung gestellt (Deutsche Nationalbibliothek, 2011).

Die GND Systematik unterteilt die Entitäten in 37 thematische Fachgebiete. Die Systemstelle 00 sammelt hierbei unspezifische Allgemeinwörter. Insgesamt weist die Systematik 478 Notationsstellen, Sammelstellen und Überschriften auf. Die jeweilige Unterteilung erfolgt nach den unterschiedlichen Gepflogenheiten der jeweiligen Fachgebiete. Zusätzlich sind jeweils Systemstellen für Personen vorgesehen, welche sich auf unterschiedlichen Klassenebenen befinden.

Das Vokabular der GND-Sachgruppen (englisch: subject categories) wird von der DNB als SKOS Vokabular zu Verfügung gestellt.[68] Bei genauerer Analyse stellt sich heraus, dass es sich dabei um eine vollständige Repräsentation der GND Systematik handelt. Am augenfälligsten wird das durch die IRIs, bei denen die Endstelle aus der zugehörigen Notation besteht.[69]

Bemerkenswert ist, dass das Vokabular derzeit in Form einer flachen Liste veröffentlicht wird, was weder der zweckmäßigen Struktur eines SKOS-Thesaurus entspricht, noch die hierarchische Ordnung der GND Systematik abbildet.

Dazu soll angemerkt sein, dass laut Leitfaden nicht alle Systemstellen gleich behandelt werden. Manche Systemstellen sind als Überschriften angesetzt; die Notation tragenden Systemstellen sind ihnen untergeordnet. Als Beispiel mag die Gliederung der Notation 3 (Religion) dienen: Sie selbst ist als Überschrift angelegt, enthält aber auch den Bereich 3.2 - 3.6* (Christentum), der wiederum das Fachgebiet der Christlichen Theologie in der dort üblichen Gliederung ordnet (Deutsche Nationalbibliothek, 2011, p. 12).

Sammelstellen sind zumeist Nebennotationen für bestimmte Fachbegriffe, deren fachlicher Bezug zum Auffinden möglicherweise nicht ausreicht. Beispiele hierfür: 2.3 (Presse) für Periodika aller Art, sowie 33.3 (Mode, Kleidung) für Kleidung jeglicher Art (Deutsche Nationalbibliothek, 2011, p. 4).

Überschriften und Sammelstellen können ihre Funktion in einer Beschreibungslogik nur dann erfüllen, wenn Pfade zu den darunter liegenden Notationsstellen bestehen. Nicht zuletzt deshalb erscheint es naheliegend und zweckmäßig, diese Ordnung im Thesaurus über Unterbegriffe (skos:narrower) herzustellen, zudem sie aus der alphanumerischen Notation ohnehin offensichtlich ist. Es entsteht so eine monohierarchische taxonomische Struktur entlang der Aspekte der GND-Systematik mit einer Tiefe von bis zu vier Ebenen.

[68] vgl. https://d-nb.info/standards/vocab/gnd/gnd-sc.html
[69] Beispiel: „Glas, Keramik, Steine und Erden“ - https://d-nb.info/standards/vocab/gnd/gnd-sc#31.15

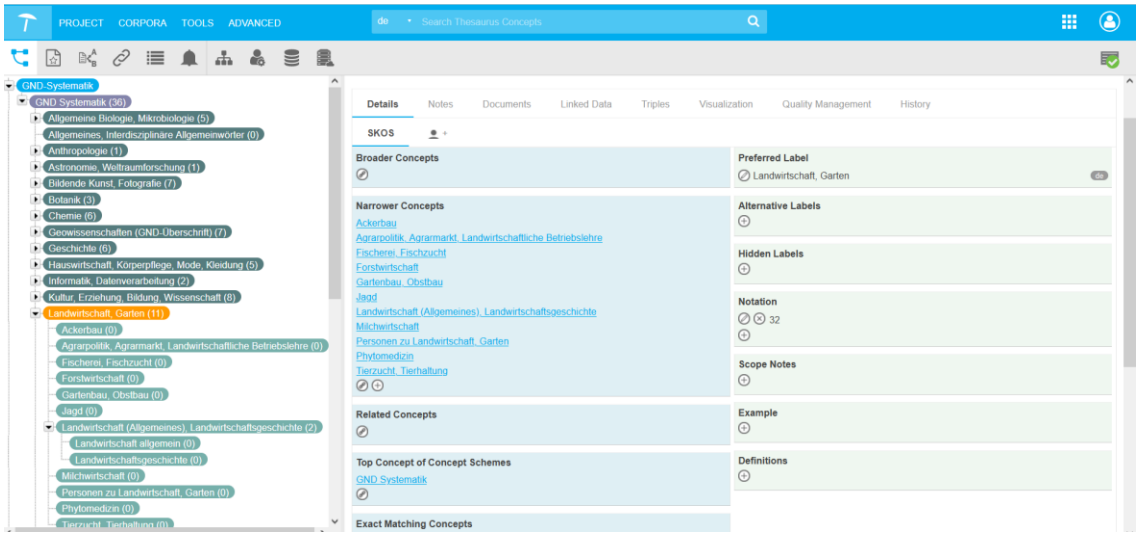


Abbildung 15: GND Systematik, hierarchisch sortiert

Hinsichtlich der Eignung für einen Klassifikationsthesaurus wäre für die Sammelstellen eine gesonderte Behandlung erforderlich; ggf. sind Maßnahmen zu treffen, dass nicht nach ihnen kategorisiert wird.

In der GND-SH selber sind die Verknüpfungen zur Systematik über einen spezifischen Beziehungstyp ausgedrückt, „GND-Sachgruppe“[70]. Die Range der Verknüpfung zeigt auf eine Sachgruppe[71], also ein SKOS-Konzept im obigen Datensatz. Die Gegenrichtung ist in den Daten nicht modelliert; die Sachgruppen befinden sich nicht im Namensraum der GND-Ontologie. Die Information ist also nicht direkt im Thesaurus enthalten, sondern gegebenenfalls nur über Linked Data Prinzipien, insbesondere über Verknüpfung zweier Ontologien auswertbar. Für eine Visualisierung des Kontexts aber insbesondere auch für eine maschinelle Verarbeitung macht es Sinn, diese Daten zu verschmelzen und die Beziehungen zu materialisieren.

4.2.1.1 GND-SKOS-Transformation der ZBW

Ein gut dokumentierter und öffentlich zugänglicher Ansatz hierzu erfolgte 2016 bei der ZBW in Hamburg.[72] Ausgehend von der Frage mit den GND-Sachkategorien ein systematischer Zugriff auf die GND-SH zu schaffen wäre, hat Joachim Neubert einen Ausschnitt des Vokabulars, die „Schlagwörter senso stricto“[73], als SKOS-Thesaurus konstruiert, sowie diese mit den GND-

[70] https://d-nb.info/standards/elementset/gnd#gndSubjectCategory
[71] https://d-nb.info/standards/vocab/gnd/gnd-sc.html#GndSubjectCategoryValue
[72] SKOS-Transformation von Teilen der SWD nach SKOS von Neubert, ZBW (2016) siehe: https://github.com/zbw/swdskos/blob/master/sparql/construct_as_skos.rq
[73] https://d-nb.info/standards/elementset/gnd#SubjectHeadingSensoStricto

Sachkategorien als skos:topConcept (Oberbegriffe) verbunden, wobei diese eine neue, oberste Hierarchieebene des Thesaurus ausbilden. Dieser Ansatz lässt sich einfach auf alle Entitätstypen der GND-SH erweitern. Konzeptionell handelt es sich hierbei am Ehesten um eine Verschmelzung.

Da jeder Begriff zumindest eine Zuordnung zu einer Sachkategorie hat, möglicherweise jedoch einen weiteren, (transitiven) Oberbegriff, werden durch diesen Ansatz mehrere Ordnungsprinzipien überlagert und im Beziehungstyp skos:broader zusammengeführt.

Für das oben angeführte Beispiel „Bearbeitung“ ergibt sich somit zuweilen die Konstellation einer schlichtweg falschen Polyhierarchie, wie in der Baumdarstellung in Abbildung 16 visualisiert:

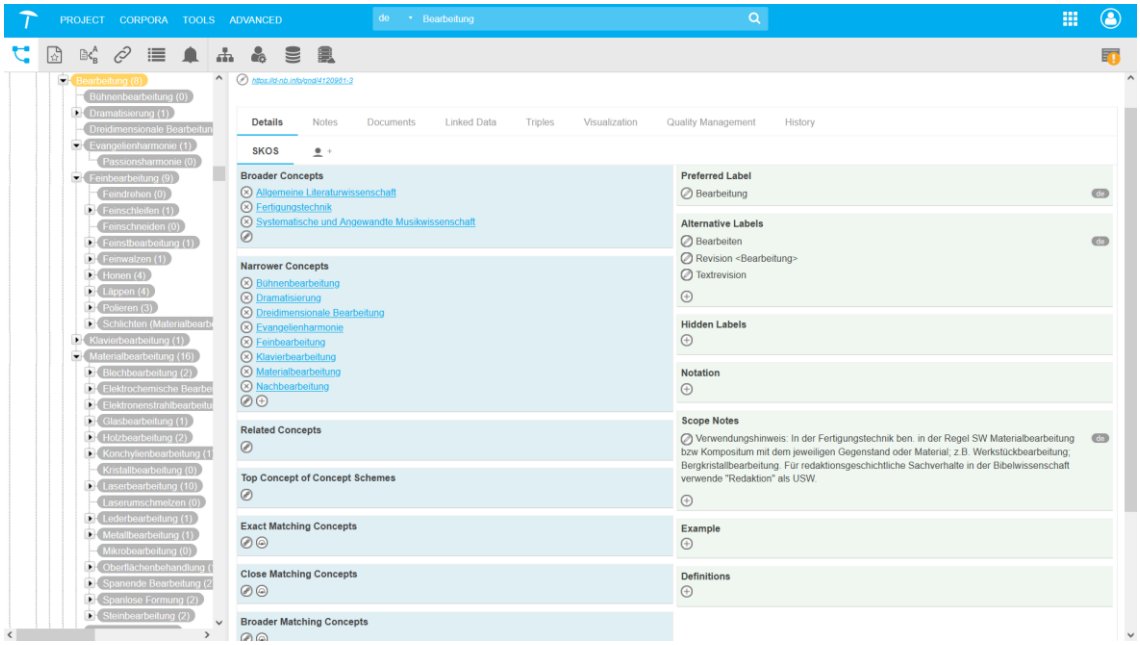


Abbildung 16: SKOS Transformation der GND nach Neubert (Ausschnitt)

Sucht man die Systematik zum Begriff „Laserbearbeitung“ über dessen transitive Oberbegriffe, landet man eben auch in der Literatur- und Musikwissenschaft:

> *Laserbearbeitung -> Materialbearbeitung ->Bearbeitung -> (SC 12.1a (Allgemeine Literaturwissenschaft) | SC 31.8a (Fertigungstechnik) | SC 14.4 (Systematische und Angewandte Musikwissenschaft))*

Ein weiteres Beispiel hierfür wäre das Schlagwort „Judenstern“ [74]. Der Begriff trägt die DDC-Notation 305.8924 (Personengruppen). Der transitive Oberbegriff „Abzeichen“ besitzt die Notationen 355.1342 (Militär), beziehungsweise 929.81 (Orden und Ehrenzeichen). Die GND-Sachkategorien lauten 16.2 (Quellen und historische Hilfswissenschaften), 6.1a (Kultur, Künste allgemein), sowie 8.5 (Militär). Über den transitiven Schluss erhielte das betreffende Sachschlagwort somit keinen Pfad in die gewünschte DDC-Sachgruppe 300, wohl aber Zuweisungen in thematisch irreführende Sachgruppen wie Militär und Heraldik.

Abbildung 17: Begriff "Abzeichen"

Da sich unschwer weitere Stichproben analogen Musters finden lassen, ist dieser Ansatz problematisch und für einen Klassifikationsthesaurus jedenfalls erst einmal zu verwerfen.

Hieraus zu schließen, dass die Konstruktion einer durchgängigen SKOS-Hierarchie ein neues Problem geschaffen hätte, würde sicherlich zu weit gehen. Die strukturellen Probleme dieser Beispiele waren in der GND bereits vollständig angelegt, wie zuvor schon erörtert. Würde man hierarchische Darstellung nach der GND-Ontologie vornehmen, wäre das Ergebnis im Grunde dasselbe.

[74] http://d-nb.info/gnd/1037696549

Den systematischen Bezug als Oberbegriff auf Ebene jedes einzelnen Blatt-Konzeptes zu manifestieren, ist ebenfalls kein sinnvoller Ansatz. Die Folge hiervon wären unzählige Nebenbezüge, deren transitive Schlüsse auf dieselbe Problematik wie oben hinauslaufen werden. Die oben konstatierten Perspektivenwechsel im Hierarchiebaum schließen ebenfalls auch aus, die Systematik nur auf oberster Ebene der taxonomischen Hierarchie auszuwerten.

Für einen Klassifikationsthesaurus ist es also erforderlich, diese Problematik so zu lösen, dass ein rein systematischer Bezug hergestellt wird. Die einfachste Möglichkeit besteht dabei darin, die taxonomische Hierarchie im Thesaurus hierbei gar nicht zu nutzen, sondern zur Kategorisierung von Texten ausschließlich die Hierarchie der Systematik auszuwerten.

4.2.2 Zugriff über DDC-Notationen

Die im DFG-Projekt CrissCross geschaffene Konkordanz zwischen Sachschlagwörtern und DDC Klassen ermöglicht einen alternativen thematischen Zugriff auf den Thesaurus und die damit beschlagworteten Dokumente. Die Wahl der DDC als Zielsystem ist zudem aus Sicht der kollaborativen Erschließung relevant, da die DDC international sehr weit verbreitet ist.

Die CrissCross Mappings sind sowohl als one-to-many Mapping (1:n Mapping) unidirektional ausgeführt, als auch im Grad der Determiniertheit differenziert. Die differenzierten Mehrfachnotationen dienen der koexistenten thematischen Verknüpfung eines Schlagwortes mit der entsprechenden Notation.

Mehrfachnotationen ermöglichen die umfassende Bedeutungsbeschreibung eines Begriffs und sind daher wünschenswert. Sie sind auch die hinreichende Voraussetzung für differenzierte Zuordnung, denn es würde nicht viel Sinn machen, eine nur einmal vorkommende Beziehung zwischen zwei Konzepten groß zu differenzieren.

Das differenzierte Mapping des CrissCross Projekts gibt Auskunft darüber, inwieweit ein Sachschlagwort mit einer DDC-Notation deckungsgleich ist. Hierzu sind in der GND-Ontologie folgende Prädikate definiert:

- *In Beziehung stehende Dewey-Dezimalklassifikation mit Determiniertheitsgrad 4*[75]
  „Die GND-Entität stimmt in seinem Bedeutungsumfang vollständig mit dem Bedeutungsumfang des in der DDC-Klassenbenennung

[75] https://d-nb.info/standards/elementset/gnd#relatedDdcWithDegreeOfDeterminacy4

hervorgehobenen Themas überein, d.h. es weist auch denselben fachlichen Kontext wie die DDC-Klasse auf.“

- *In Beziehung stehende Dewey-Dezimalklassifikation mit Determiniertheitsgrad 3*[76]
  „Die GND-Entität stimmt in seinem Bedeutungsumfang vollständig oder weitgehend mit dem Bedeutungsumfang eines Themas überein, das wesentliche Übereinstimmung mit der DDC-Klasse aufweist. In der DDC-Terminologie liegt eine wesentliche Übereinstimmung vor, wenn ein Thema nahezu koextensiv mit einer DDC-Klasse ist oder mehr als die Hälfte des Inhalts einer Klassenbenennung abdeckt.“

- *In Beziehung stehende Dewey-Dezimalklassifikation mit Determiniertheitsgrad 2*[77]
  „Die GND-Entität stimmt in seinem Bedeutungsumfang vollständig oder weitgehend mit dem Bedeutungsumfang eines Themas überein, das sinngemäß Teil einer DDC-Klasse ist, aber einen geringeren inhaltlichen Umfang aufweist als der durch die Notation repräsentierte Begriff.“

- *In Beziehung stehende Dewey-Dezimalklassifikation mit Determiniertheitsgrad 1*[78]
  „Der Bedeutungsumfang der GND-Entität und der Bedeutungsumfang der DDC-Klasse haben eine geringe Schnittmenge.“

Die Determiniertheitsgrade werden abgekürzt mit D4 … D1 bezeichnet.

Bei der Übersetzung in einen Klassifikationsthesaurus führen die Mehrfachnotationen dann zu Polyhierarchie, wenn die verknüpfte Notation transitiv in unterschiedliche Sachgruppen zeigt. Mehrfache Notationen in dieselbe Sachgruppe haben keine besonderen Auswirkungen.

4.2.3 DDC als Linked Data

Im Sommer 2012 wurde DDC 23 als SKOS Thesaurus publiziert, wie ein Blogbeitrag des damaligen OCLC Mitarbeiters Michael Panzer verkündet.[79] Der SPARQL Endpunkt von

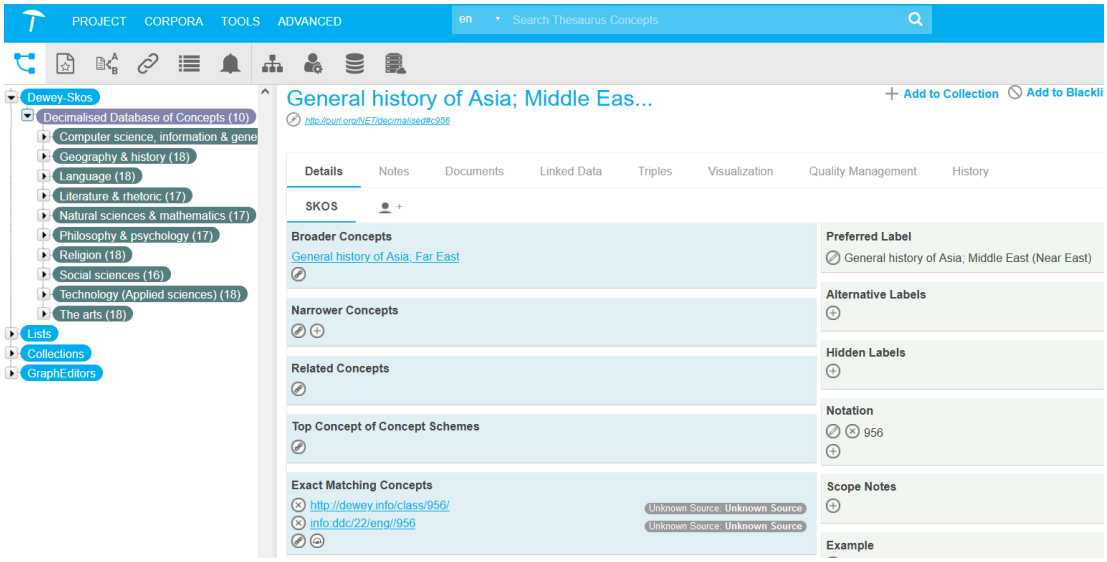
[76] https://d-nb.info/standards/elementset/gnd#relatedDdcWithDegreeOfDeterminacy3
[77] https://d-nb.info/standards/elementset/gnd#relatedDdcWithDegreeOfDeterminacy2
[78] https://d-nb.info/standards/elementset/gnd#relatedDdcWithDegreeOfDeterminacy1
[79] vgl. https://ddc.typepad.com/025431/2012/06/ddc-23-released-as-linked-data-at-deweyinfo.html

http://dewey.info ist derzeit nicht erreichbar, so dass sich zu Zeit keine direkte Möglichkeit ergibt, die Daten programmatisch abzurufen.[80]

Die Obsoleszenz der Datenquelle ist dennoch praktisch irrelevant, zumindest was den Zweck dieser Arbeit angeht. Die Daten lassen sich nämlich zumindest ausschnittsweise unschwer synthetisieren, da die Struktur zumindest für die Haupttafeln monohierarchisch, somit also die Position jeder Klasse allein durch ihre Notation offensichtlich ist. Alternativ kann man die DDC-Klassifikation auch aus frei zugänglichen Linked Data Quellen, wie z.B. WikiData gewinnen.

Durch die Dezimalstruktur ergibt sich zwangsläufig eine sehr übersichtliche Baumstruktur. Ab der zweiten Klasse besitzt jede Klasse bis zu neun Unterklassen, was eine sehr kompakte und übersichtliche Struktur ergibt. Detailinformation, wie z.B. die URIs des Concept Scheme lassen sich aus Publikationen gewinnen, wenn man diese für eine eigene Modellierung benötigt (Mitchell & Panzer, 2013).

Eine frei verfügbare „Kopie“ der obersten 1000 Klassen ist als „Decimalised Database of Concepts“ [81] frei erhältlich. Die Konzepte besitzen ein eigenes IRI-Schema, jedoch ist jede Klasse als skos:exactMatch annotiert und mit den URIs der Ausgaben DDC-14 und DDC-23 versehen.



[80] Der englische Artikel zur DDC sagt hierzu am 10.September 2021:„An experimental version of Dewey in RDF was previously available at dewey.info beginning in 2009, https://old.datahub.io/dataset/dewey_decimal_classification but has not been available since 2015. https://www.oclc.org/developer/news/2015/dewey-down.en.html [diese Seite ist nicht auffindbar], siehe: https://en.wikipedia.org/w/index.php?title=Dewey_Decimal_Classification&oldid=1035578162

[81] vgl. https://ontologi.es/decimalised/decimalised.html

Abbildung 18: Decimalised Database of Concepts

Bezüglich der Lizensierung besteht eine interessante Situation: Die RDF-Daten stehen unter einer Creative Commons Lizenz (CC BY-NC-ND), was die nicht-kommerzielle Weitergabe ohne Veränderung erlaubt. Die Dewey URIs sind in zahllosen Datensätzen enthalten, die selber unter freizügigeren Lizenzen stehen, insbesondere die GND selber (CC-0). Für geschäftliche Zwecke ist die Klassifikation bei OCLC zu lizensieren. Es ist bedauerlich, dass OCLC diese duale Lizenz nicht aktiver unterstützt.

Für die Beschreibungslogik dieser Arbeit wird nur eine geringe Anzahl an Konzepten aus dem DDC Thesaurus benötigt, nämlich diejenigen, die den DDC-Sachgruppen entsprechen. Diese also wurden schlicht aus der Tabelle im DDC-Leitfaden ausgelesen, (Alex, 2014), in ein Projekt der des Thesaurus-Editors PoolParty Semantic Suite importiert, und dabei in RDF konvertiert. Hierbei wird programmatisch ein flacher SKOS Thesaurus mit allen 104 Sachgruppen erzeugt.[82] Es wird dabei jeweils ein SKOS-Concept mit base-IRI „dewey.info/class/, gefolgt von der Notationsstelle, angelegt. Beispiel: http://dewey.info/class/610 . Die URIs entsprechen also den Originaldaten der DDC-23.
Als Benennung der Konzepte wird die Notation, gefolgt vom Namen der Sachgruppe verwendet, Beispiel: „610 – Medizin“. Damit erreicht man die gewünschte Sortierung nach den Notationen.

4.3 Konkordanz von GND-Systematik und DDC-Sachgruppen

Lediglich 133.871 Sachschlagwörter in der GND verfügen momentan über DDC-Notationen. Damit lassen sich ca. 79.000 Begriffe nicht unmittelbar in die DDC-Struktur einsortieren. Notationen für GND-SC sind jedoch für alle Sachschlagwörter vorhanden.[83]

Da die globale thematische Abdeckung beider Systematiken als gleichwertig angenommen werden darf liegt es nahe, die fehlenden Zuordnungen der Sachschlagwörter in die Sachgruppen über eine Konkordanz beider Systematiken zu ergänzen.[84]

---

[82] https://help.poolparty.biz/en/user-guide-for-knowledge-engineers/basic-features/import,-export-and-reporting-with-poolparty/poolparty-excel---csv-tabular-import---export---overview/the-poolparty-excel-format.html
[83] Interessanterweise hat sich eine Ausnahme gefunden. Diese möge die Regel bestätigen.
[84] Hierzu erfolgte eine separate Transformation nach denselben Prinzipien wie für die DDC-Sachgruppen, aber mit GND-SC als Oberbegriffen.

In erster Iteration wurden dazu die beiden Systematiken in eigene SKOS- Concept Schemes im selben Thesaurus einsortiert, und dann alle GND-Notationen zusätzlich als Unterbegriffe der Sachgruppen verknüpft. Die Zuordnung erfolgt so granular wie möglich. Der Baum entspannt sich insgesamt ohne Zirkelschlüsse, und ohne zusätzliche Polyhierarchie.

Operativ ist diese Arbeit über einen grafischen SKOS-Editor einigermaßen schnell erledigt, die Einsortierung aller ca. 400 Notationsstellen samt aller Schlagwörter der GND dauert etwa einen halben Tag.

4.3.1 Problemstellungen der Konkordanz

Beim Durcharbeiten der beiden Leitfäden haben sich allerdings folgende Problemstellungen hinsichtlich einer GND-SC / DDC Konkordanz herausgebildet, die dann zum Teil einer Umsetzung im Weg standen:

i. Behandlung der GND-SC Sammelstellen und Überschriften
   Eine Konkordanz wird nur auf Ebene der notationstragenden GND-Systematikstellen hergestellt. Überschriften werden für das Mapping nicht herangezogen. Dies ist durch die Verknüpfung auf unterster GND-Notationsebene erfüllt.
ii. Umgang mit Allgemeinwörtern (z.B. GND- Systematik 00): Diese können aus dem Annotationsthesaurus herausgehalten werden, wenn die GND-Systematikstelle 00 unverknüpft bleibt. Eine weitere Anreicherung mit Allgemeinwörtern ist eher nicht erwünscht.
iii. Heterogene Ordnungsprinzipien

Heterogene Ordnungsprinzipien zwischen beiden Systematiken betreffen insbesondere die Fachgebiete 800 – Literatur, sowie 900 – Geschichte. In beiden Fällen verfügen die DDC-Sachgruppen über partielle sprachliche bzw. geografische Unterteilungen. Die GND-Systematik folgt jedoch hier, wie auch in den anderen Sachgebieten, einer inhaltlichen Unterteilung.

Insbesondere trifft das die GND-Systemstelle 16.5: Die enthaltenen Instanzen wären zum Teil auf die DDC-Notationen 940 – 990 jeweils aufzuteilen. Auch die GND Systemstelle 12.3 (Literaturgattungen) enthält zahlreiche Instanzen, die selektiv in Sachgruppen 810-891.8 notiert werden sollten. Geowissenschaften selber sind Notationsstelle, somit gehören die zugehörigen Instanzen nach DDC 550 (Geowissenschaften). Geografie gehört hingegen laut DDC nach Sachgruppe 910. (außer: Physikalische Geografie, die wieder zu DDC-Sachgruppe 550 gehört.)

Insgesamt sind diese Widersprüche nur mit einigem Aufwand auflösbar, so dass von einer umfänglichen Umsetzung abgesehen werden musste, und die vorläufige Teilumsetzung wieder aus methodischen Gründen rückgängig gemacht wurde.

4.3.2 Punktuelle Erweiterungen mittels Konkordanz

Für eine punktuelle Erweiterung des Vokabulars eignet sich die Konkordanz allerdings durchaus. Insbesondere für spärlich besetzte DDC-Sachgruppen können die Ergebnisse der Kategorisierung so schnell verbessert werden. Praktisch umgesetzt wurde dies in diesem Projekt für die Sachgruppe 370, die nach DDC nur über etwa 1.600 Schlagwörter verfügt. Verknüpft man die GND-Systemstellen 6.2, 6.3, 6.4 und 6.6 mit der Sachgruppe, verdoppelt sich das dort referenzierte Vokabular in etwa.

Ähnliche Möglichkeiten ergeben sich prinzipiell für die Notationsstelle 32.3 – Ackerbau und 32.9 – Jagd in der Sachgruppe 630 – Landwirtschaft, sowie die Notationsstelle 31.3a und 720 (beide Architektur). „Einzelne Sportarten“, 34.3, sind ebenfalls ein Indikator für die DDC-Sachgruppe Sport.

4.4 Qualitätssicherung von Thesauri

Probleme wie in Kapitel 4.2.1.1 beschrieben sind in den Quelldaten schon angelegt, die werden aber erst über die Untersuchung des systematischen Zugangs, insbesondere in einer SKOS-Darstellung deutlicher in Erscheinung treten. Das hat einerseits mit dem vereinfachenden Ansatz zu tun, der SKOS mit sich bringt. Nicht zuletzt aber liegt das daran, dass es für SKOS standardisierte Qualitätsprüfungswerkzeuge gibt, insbesondere die qSKOS Bibliothek von Christian Mader. Hierbei handelt es sich um ein umfangreiches, vom W3C empfohlenes Werkzeug zur systematischen Analyse von SKOS-Thesauri.

Mader behandelt hier unter anderem Terminologie-, Dokumentations- und strukturelle Probleme von SKOS-Thesauri.[85] Über den Umweg der SKOS-Transformation ist es somit möglich, unter anderem sieben hierarchische Zirkelbezüge in den GND-SH zu finden (siehe: Report in Anhang 1: qSKOS Report).[86]
Das untenstehende Beispiel besteht bereits explizit in den GND-Daten, da hier zwei Konzepte reziprok Oberbegriffe unterschiedlichen Typs zueinander sind. Unbeachtet dessen ist die Analyse und Auflösung von Zirkelschlüssen in einem derartig großen Vokabular einigermaßen aufwändig.

[85] Eine Übersicht der Tests bietet. https://github.com/cmader/qSKOS/wiki/Quality-Issues
[86] Grundlage hierbei war die Neubert-Transformation.

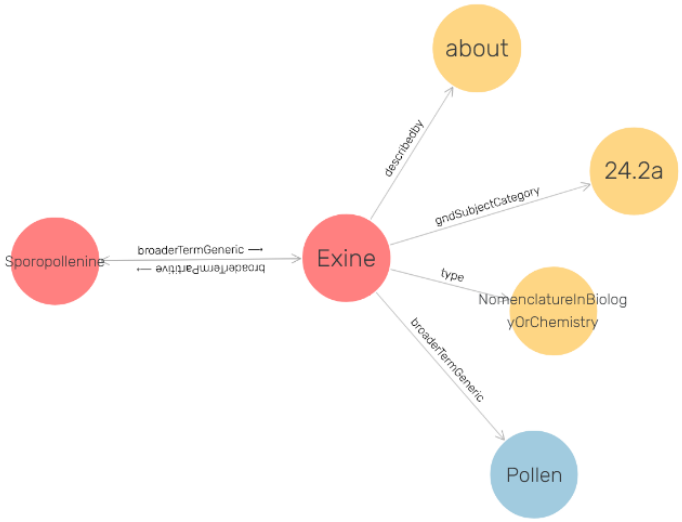


Abbildung 19: Zirkelschluss über gegenseitige Oberbegriffe

Wie in Anhang 1 ersichtlich, spannen sich Zirkelschlüsse in den GND-SH über bis zu vier Hierarchieebenen.

qSKOS weist noch weitere Probleme in der GND nach, insbesondere beinahe 2.000 Fälle von sich zwei- oder mehrfach überlappenden Benennungen in den GND-SH. Dies deutet auf gewisse Herausforderungen bezüglich der Disambiguierung hin.[87] Suominen und Mader haben dieses Szenario unter anderem auch in RAMEAU in 5.539, sowie der DDC in 40.729 Fällen gefunden. Quantitativ ist die Herausforderung in der GND damit relativ harmlos, allerdings hinsichtlich möglicher Disambiguierungsprobleme dennoch im Auge zu behalten (Suominen & Mader, 2013).

Hinsichtlich eines Klassifikationsthesaurus sind die strukturellen Probleme eindeutig höher zu priorisieren als die terminologischen, da hierfür ein azyklischer, gerichteter Graph erforderlich ist. Auch müssen mehrdeutige Pfade minimiert und falsche Pfade vermieden werden.

Die Schlussfolgerung aus den Betrachtungen der GND Daten und der Neubert-Transformation sind daher eindeutig. Hinsichtlich des vertretbaren Aufwandes und der Qualität der systematischen Ordnung verbleibt nur die Lösung, auf eine Auswertung der transitiven Begriffshierarchien bei der Klassifikation zu verzichten, sowie die Auswertung der Klassifikationsbezüge

[87] vgl. https://github.com/cmader/qSKOS/wiki/Quality-Issues#overlapping-labels

jeweils auf die CrissCross- oder GND-SC-Mappings zu beschränken. Diese Lösung ist jedenfalls prinzipiell frei von Zirkelbezügen und ungewollt vererbten thematischen Bezügen.

Eine Qualitätssicherung war auch nach dem Import der GND-Systematik-Transformation in den bestehenden, nach CrissCross strukturierten Thesaurus nach den oben beschriebenen Konkordanzregeln durchzuführen. Diese waren jedoch, dank des einfachen Aufbaus der GND-Systematik, ohne Befund.

Da qSKOS in den von mir benutzten Taxonomieeditor PoolParty Semantic Suite eingebunden ist, erfolgten diese und noch weitere Konsistenzprüfungen jeweils beim Import der Daten in das Webservice.[88]

5 Erzeugung des Klassifikationsthesaurus

Ein Klassifikationsthesaurus muss die Aspekte der Begriffe vollumfänglich und so differenziert wie möglich wiedergeben. Für die gewählte Methode sollen Hierarchie und thematischer Bezug zusammenfallen. Das gelingt anhand der DDC- Notationen in der GND, sowie entlang von transitiven Schlüssen in DDC-Sachgruppen als Oberbegriffe.

Das Verfahren hierzu orientiert sich am Ansatz von Neubert, der der GND für den systematischen Zugriff eine *upper ontology* hinzuzufügt (siehe hierzu Kapitel 4.2.1.1). Allerdings verzichten wir darauf, mehrere Ordnungsprinzipien in einer SKOS-Hierarchie zu verschmelzen und nutzen nur die DDC-Notationen und die Hierarchie der DDC zur Erzeugung der gewünschten Topologie. Damit erwarten wir eine Struktur, die auf der obersten Ebene genau 104 top concepts als Kategorien besitzt, aber darunter vollständig flach ist.

5.1 Polyhierarchische Systematik

Betrachten wir hierzu als Beispiel den Kontext des Begriffes „Flughafen“[89] in allen möglichen Hierarchiebäumen, also taxonomischen Oberbegriffen, den zugeordneten GND-Sachkategorien, sowie den zugeordneten DDC-Sachgruppen. Abbildung 20 zeigt die transitiven Oberbegriffe von Flughafen nach der DDC-Transformation.

[88] vgl. https://help.poolparty.biz/en/user-guide-for-knowledge-engineers/advanced-features/poolparty-data-validator.html#access-the-poolparty-data-validator
[89] vgl. https://d-nb.info/gnd/4154752-4/

Abbildung 20: Transitive Oberbegriffe von „Flughafen“ nach DDC-Sachgruppen

Wir können also eine im transitiven Schluss bestehende Zuordnung des Begriffs zu insgesamt sieben Sachgruppen unterscheiden. Da der Begriff lebensnah allen diesen Bereichen zuzurechnen ist, sind diese Beziehungen allesamt sinnvoll.

Eine erweiterte Darstellung nach beiden Systematiken unter Einschluss der (transitiven) Oberbegriffe leistet Tabelle 2:

OB	**OB transitiv (Klassen)**	**GND-SC**	**GND-SC transitiv (Sachgruppen des Oberbegriffes)**	**Related DDC (Determiniertheitsgrad)**	**DDC-Sachgruppe transitiv (Transitive Sachguppe des transitiven Oberbegriffs**
Verkehrsbau	Infrastruktur (6)	10.6a	10.6a – Telekommunikation und Verkehr, Fremdenverkehr	387.736 (4)	380- Handel, Kommunikation, Verkehr
				629.136 (3)	620 - Ingenieurswissenschaften
				623.66 (2)	620 – Ingenieurswissenschaften
			(10.7b Raumordnung, Stadtplanung, Landschaftsgestaltung	711.78 (2)	710 – Raumplanung & Landschaftsarchitektur
		31.3a -	31.3a - Architektur	725.39 (2)	720 – Architektur
				343.0997 (2)	340 - Recht
				354.79 (2)	350 – öffentliche Verwaltung
				647.9639 (2)	640 - Hauswirtschaft
			(10.2dc – Wirtschaftsstruktur)	(338.01 (1))	330 - Wirtschaft

Tabelle 2: Umfassender hierarchischer Kontext des Begriffs "Flughafen"

Dieser Begriff eignet sich insbesondere deshalb als Beispiel, da er mit acht DDC-Notationen die wohl thematisch höchste Komplexität im gesamten Thesaurus aufweist.[90] Gelingt es hierfür eine schlüssige Lösung zu finden, so ist diese auch für die weniger komplex notierten Begriffe tauglich.

Taxonomisch handelt es sich beim Begriff „Flughafen“ um einen Unterbegriff von „Verkehrsbau“ der über insgesamt sechs Klassen den das top concept „Infrastruktur“ besitzt. Der begriffliche Kontext ist durchgängig monohierarchisch. Thematisch fällt dieser Pfad mit der stärksten DDC- Notation zusammen, deren transitiver Oberbegriff auch gut zur GND-Sachkategorie 380 passt.

Die DDC-Sachgruppe ermittelt man, indem man den skos:broader Beziehungen bis zur entsprechenden Klasse folgt. Anschaulich ist das über WebDewey möglich, für eine programmatische Verarbeitung ist der Weg über RDF-Daten besser geeignet.

Haupttafeln

Notation	Thema
	Haupttafeln
300	Sozialwissenschaften
380	Handel, Kommunikation & Verkehr
383-388	Kommunikation und Verkehr
387	Schifffahrt, Luft-, Weltraumverkehr
387.7	Luftverkehr
387.73	Luftfahrzeuge und Einrichtungen
387.736	**Flughäfen**
387.7362	Einrichtungen
387.7364	Teilbereiche und Dienste

Abbildung 21: Suchergebnis über DDC-Notation (Web Dewey)

Allerdings besitzt der Begriff sieben weitere DDC-Notationen, die in insgesamt sieben DDC-Sachgruppen zeigen und somit alle zu seiner koextensiven Beschreibung beitragen. Ab dem Determiniertheitsgrad 2 handelt es sich dabei um Nebenaspekte derselben Sache (Gödert & Hubrich, 2014). Dennoch sagen auch diese Notationen aus, dass wir mit diesen Sachgruppen in Dokumenten mit dem Begriff „Flughafen“ jedenfalls zu rechnen haben.

Betrachtet man die (transitiven) Sachgruppen des transitiven Oberbegriffs, fällt auf, dass trotz eines monohierarchischen und aspektisch unproblematischen Pfades der Themenbezug

[90] Siehe hierzu auch Kapitel 5.4.2.2

wechselt: der transitive Oberbegriff fällt in eine andere DDC-Sachgruppe und auch in eine andere GND-Sachkategorie. Hierzu muss man nicht einmal alle Ebenen durchschreiten: schon der direkte Oberbegriff „Verkehrsbau“ zeigt nur noch in die Sachgruppe 720, somit in ein fremdes Themengebiet. Somit bestätigt auch dieses Beispiel erneut, dass eine Vermischung der Ordnungsprinzipien für die Erzeugung eines solchen Thesaurus tunlichst zu vermeiden ist.

5.2 Vorgehensweise und Eigenschaften der Beschreibungslogik

Die gewählte Vorgehensweise zeichnet sich darüber hinaus dadurch aus, dass alle bestehenden Entitäten und Beziehungen der GND-Quelle erhalten bleiben. Die Transformation fügt diese Daten musterbasiert neue Prädikate in Form von Literalen und Beziehungen hinzu; ebenfalls hinzugefügt werden die SKOS-Daten (Entitäten) der *upper ontology*.[91]

Prinzipiell gibt es zwei Möglichkeiten, diese Daten zu konstruieren:

i. Man verknüpft die Konzepte an der konkreten Notationsstelle mit den DDC RDF Daten. Hierzu benötigt man einen vollständigen Abzug der SKOS-DDC. Synthetische Notationen müssen ggf. nachgeschaffen werden. In der Folge müsste man überlegen, ob die Benennungen der DDC Klassen zwischen den verknüpften Sachschlagwörtern und den DDC-Sachgruppen zum Vokabular gehören.

ii. Man verknüpft die Konzepte direkt und materialisiert die transitive Beziehung über 1…n Klassen in einer einzigen skos:broader Instanz. Dazu benötigt man nur die DDC-Sachgruppen als SKOS-Daten.

Im Ergebnis unterscheiden sich beide Thesauri in der Tiefe der modellierten Klassen. Beim zweiten Ansatz beträgt die Tiefe genau 2 Klassen, beim ersten Ansatz hängt es von der Klasse der verknüpften Notationsstelle ab.

Da Ansatz 2 deutlich einfacher umzusetzen und zudem für den von mir verwendeten Extraktor ohne Nachbearbeitung direkt nutzbar ist, fiel die Entscheidung hierauf.

5.3 Technische Umsetzung

Die technische Umsetzung der Thesauruserzeugung erfolgt in einem SPARQL Prozessor nach dem SPARQL 1.1 Standard. Die Wahl fiel dabei aus praktischen Gründen auf eine lokal betriebene Instanz des Produktes GraphDB des bulgarischen Herstellers Ontotext.[92] Dabei handelt es sich um eine Graph-Datenbank mit webbasiertem Editor, der sogenannten Workbench.

---

[91] Auch hierin besteht ein Unterschied zur Neubert-Transformation, der einen Teil der Daten in einen neuen Graphen kopiert.

[92] https://graphdb.ontotext.com/

Abbildung 22 zeigt den offenen Editor mit einer Auswahl von Tabs mit für die Konstruktion gespeicherten Abfragen.[93]

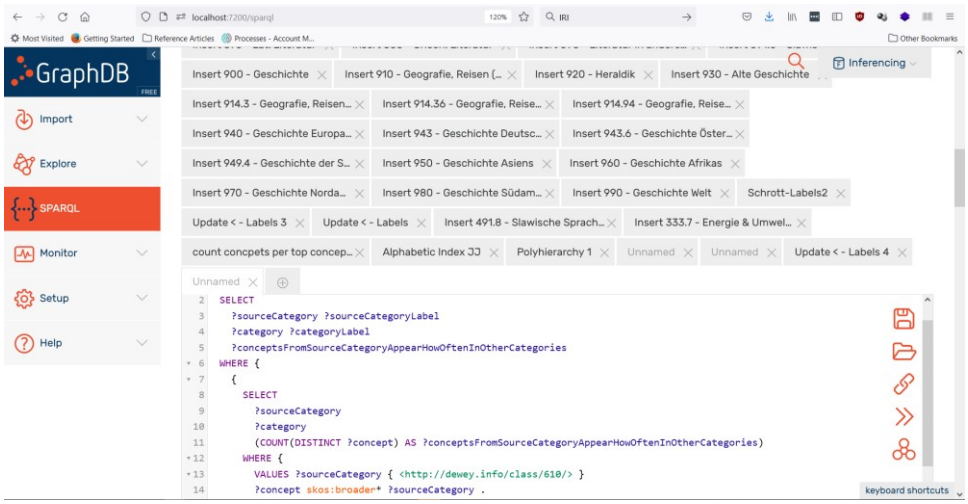


Abbildung 22:GraphDB Workbench

5.3.1 Import des GND Dump

In ein leeres repository der Graphdatenbank werden die Rohdaten der GND-SH, sowie die DDC-Sachgruppen importiert. Hierzu wurden die selbst erzeugten und von der DNB bereitgestellten Dateien über RDF Import in die Datenbank geladen. Damit werden diese Daten verschmolzen. Hiermit sind alle gewünschten Entitäten in einen Graphen geladen und die Ontologien verschmolzen; zu erledigen ist noch ihre Verknüpfung mit den gewünschten Beziehungen.

5.3.2 Berechnung der Zahlenbereiche für die DDC-Sachgruppen

Für die Konstruktion des Klassifizierungsthesaurus sind SKOS-Hierarchiepaare (broader / narrower) von jedem Begriff auf die übergeordnete DDC-Sachgruppe anzulegen. Diese erfolgen gemäß der in den Daten vorhandenen CrissCross mappings. Um die Zuordnungen programmatisch zu vollziehen macht man sich die Konstruktionsprinzipien der DDC zunutze, insbesondere dass man mit den Notationen wie mit Zahlen rechnen kann. Hierzu verwende ich eine SPARQL-Engine, welche die für eine Zuordnung benötigten Ungleichungen lösen kann.

Um mit den DDC-Notationen zu rechnen, kann man diese als reelle Zahlen auffassen. Hierbei betrachte man den Dewey-Punkt einfach als Dezimalpunkt. In der Praxis handelt es sich um rationale Zahlen, da die Anzahl der Nachkommastellen endlich ist. Um den Wertebereich mit

[93] Die Workbench erlaubt mit der Funktion Visual Graph ebenfalls die Darstellung des Kontextes einzelner Ressourcen, die für etliche Abbildungen in dieser Arbeit genutzt wurde.

dem der gesuchten DDC-Sachgruppe zu vergleichen, sind die Werte der Notation in den Datentyp xsd:decimal zu konvertieren.

Das Regelwerk für die Gültigkeitsbereiche ergibt sich aus dem Leitfaden für die Vergabe der DDC-Sachgruppen (Alex, 2014).

Prüfen wir also die Ungleichung für ein real vorkommendes Beispiel mit der Notation 620.112[94] gegen einen Wertebereich der Sachgruppe 620, erhalten wir:

620 < 620.112 < 621.3 (wahr)

Meist umfassen diese komplette Dekaden (z.B. 610 – Medizin). Es kann jedoch auch vorkommen, dass Bereiche unterbrochen, verkettet, oder sogar verschachtelt sind. Hierfür muss man dann mehrere, nicht überlappende Filterbereiche für eine Sachgruppe anlegen.

Ein Beispiel für die selektive Zuordnung der Notationen aus dem Bereich 620 – 630 in die verschränkten Bereiche für die Sachgruppen

i. 620 – Ingenieurswissenschaften
ii. 621.3 -Elektrotechnik, Elektronik
iii. 624-Ingenieurbau und Umwelttechnik

in die Sachgruppe 620 findet sich in untenstehendem SPARQL-Schnipsel:

```
# 620 - Ingenieurswissenschaften
              FILTER ((?d >= 620 && ?d <621.3) || (?d >= 621.4 && ?d < 621.46)
|| ?d >= 621.47 && ?d < 622) || (?d >= 623 && ?d < 624) || (?d >= 625.19 && ?d
< 625.3) || (?d >= 629 && ?d < 629.8) || (?d >= 629.9 && ?d <630))
              BIND
(xsd:decimal(STRBEFORE(STRAFTER(STR(?o),"http://dewey.info/class/"), "/"))AS
?d)
```

Hierbei erfolgt im Binding (BIND) eine Extraktion der im CrissCross Mapping hinterlegten Notation als Dezimalzahl. Damit wird die Variable *?d* befüllt. Die FILTER -Klausel führt die Lösung der gegebenen Ungleichungen für jede DDC-Notation im Thesaurus durch. Ist eine Lösung für die im Binding ermittelten Werte für die Notation für den definierten Bereich wahr, wird dem Schlagwort die entsprechende DDC-Sachgruppe als Oberbegriff zugewiesen.

[94] https://d-nb.info/gnd/4455656-1

Dies entspricht exakt der natürlichsprachigen Anweisung des Leitfadens:

```
620 Ingenieurwissenschaften und Maschinenbau 620, 621 (außer 621.3 und
621.46), 623, 625.19,625.2, 629 (außer 629.8)
621.3 Elektrotechnik, Elektronik 621.3, 621.46, 629.8
624 Ingenieurbau und Umwelttechnik 622, 624-628 (außer 625.19 und
625.2)
```

(Alex, 2014)

5.3.3 Konstruktion des Thesaurus über SPARQL

Zur Erreichung der gewünschten Eigenschaften, insbesondere des schlichten Hinzufügens von RDF-Typen und Prädikaten, wird SPARQL UPDATE INSERT[95] verwendet.

Es ist zwar prinzipiell möglich, diese Transformation für alle Sachgruppen über ein einziges Skript zu erledigen. Aus praktischen Gründen habe ich mich dafür entschieden, für die Zuordnungen jeder Sachgruppe separate Skripte zu schreiben und nacheinander auszuführen. Hierbei sind die GND-Entitätstypen sowie die CrissCross Beziehungen über jeweilige VALUES Klauseln an Variablen gebunden.

Ein kommentiertes Codebeispiel in Anhang 2 zeigt die Konstruktion für die Sachgruppe 891.8 – Slawische Literatur.

5.3.4 Behandlung der CrissCross-Mehrfachzuordnungen

Die Determiniertheitsgrade 2-4 der CrissCross Mappings erscheinen zur Ausprägung Klassenhierarchie tauglich, da diese wenigstens einen Teilumfang einer Notation abbilden, die somit als wenigstens zugehörig interpretierbar sind. Für die Beziehungen mit Determiniertheitsgrad 1 ist dies nicht unmittelbar gegeben, wie Jessica Hubrich erläutert:

> *„Repräsentiert ein Schlagwort ein Thema, das quer zur DDC-Hierarchie steht und/oder einen deutlich amorpheren Bedeutungsumfang als die zugeordnete DDC-Klasse aufweist, so besteht nur eine geringe Schnittmenge zwischen SWD-Schlagwort und DDC-Klasse, die durch den Determiniertheitsgrad 1 festgehalten wird.“ (Hubrich, 2008, p. 52)*

Insbesondere der Querstand zur DDC-Hierarchie ist ein wesentlicher Ausschlussgrund für die Nutzung dieser Prädikate, da ja genau die Logik der DDC-Hierarchie zur Kategorisierung genutzt wird.

Bei der Übersetzung in SKOS-Hierarchien kommt es unweigerlich zu einer Nivellierung der Determiniertheitsgrade. Dies bedeutet, dass der Grad der Verbundenheit mit einer DDC-

[95] https://www.w3.org/TR/2013/REC-sparql11-update-20130321/#insertData

Sachgruppe nicht unmittelbar aus der Beziehung erkennbar ist. Die Information ist allerdings nach wie vor aus dem GND- Prädikat des Begriffes auswertbar, sofern man das wünscht, da die ursprünglichen RDF- Eigenschaften ja nach wie vor in den Daten vorhanden sind.

Für die Frage, ob diese Nivellierung eine große Rolle bei der Kategorisierung spielt, sind folgende Gesichtspunkte maßgeblich:

Klassenzentrisch werten wir bei der Kategorisierung nicht unmittelbare Beziehung zur DDC aus, sondern die Beziehung zum transitiven und weiter gefassten Oberbegriff. Mehrfache Zuordnungen von jeweils D3 und D4 kommen praktisch nicht vor, da für diese Grade bereits eine Übereinstimmung der koextensiven Bedeutung eines Schlagwortes mit einer Dewey-Klasse festgestellt wurde.Zieht man alle Instanzen von Kombinationen der Grade D4-D2 in Betracht, so kommen hier Mehrfachzuordnungen durchaus vor, siehe erneut das Beispiel Flughafen in Kapitel 5.1:

Dieses Schlagwort verfügt über insgesamt acht DDC-Notationen, davon jeweils eine Beziehung mit dem Determiniertheitsgrad 4, eine mit dem Determiniertheitsgrad 3 sowie über sechs weitere Beziehungen mit dem Determiniertheitsgrad 2.
Klassenzentrisch betrachtet sind die Bezüge des Begriffs in die Sachgruppen 380 und 620 für den Begriff „Flughafen“ sicher höher zu bewerten, als die Übrigen. Allerdings wird es schwerfallen, das Beziehungsgewicht einer Zuordnung nach Sachgruppe 380 mit D4 von dem eines Beziehungspaares D3+D2 nach Sachgruppe 620 zu differenzieren. Für die fünf weiteren Zuweisungen mit D2 untereinander können wir nur konstatieren, dass jede von diesen gleich relevant wäre.

Aus klassenzentrischer Sicht ist es also nicht unmittelbar nicht ersichtlich, welche Auswirkungen die Notation in mehrere Sachgruppen an sich, oder der Determiniertheitsgrad einer einzelnen Beziehung auf die Zuverlässigkeit einer Kategorisierung haben wird.

Aus dokumentenzentrischer Sicht ist zu sagen, dass wir bei der Kategorisierung niemals Einzelbeziehungen auswerten, sondern ein Aggregat aller im Text vorkommenden Benennungen und deren Relevanz. Wie für alle lexikalischen Indexierungsverfahren, kommt es dabei letztlich auf den Kontext im Dokument an, zu welcher Sachgruppe ein Begriff zu zählen ist. Daher scheint die vorgenommene Nivellierung der Beziehungsarten jedenfalls unbedenklich.

Es ist also für diese Anwendung das Bestehen der Beziehung an sich relevanter als ihr Gewicht. Dies gilt hier jedenfalls für die Grade D4-D2, nicht jedoch für Beziehungen nach Grad D1,

welche tatsächlich dazu geeignet sind, die Ergebnisse zu verwässern. So auch die Erkenntnis der Untersuchungen von Maike Sommer (vgl. Sommer, 2012, p. 53).

5.3.5 Sonderfall: DDC-Notationen in Hilfstafeln

Einige von CrissCross verwendete Notationen zeigen in die Hilfstafeln. Da diese Notationen keine auf die Haupttafeln bezogene Hierarchie ausbilden, sind sie zur Verknüpfung nach von DDC-Sachgruppen nicht nutzbar. Auf eine Modellierung dieser Beziehungen wurde daher verzichtet.

Beispiel: http://dewey.info/class/4--11/ diese zeigt nach WebDewey auf die Hilfstafel „T4—11-Schriftsystem“. Somit wird für ein solches Prädikat keine Oberbegriff-Beziehung angelegt. Praktisch geschieht das damit, dass der Ausdruck „4--11“ die Ungleichung im SPARQL-Skript nicht löst, womit eine Zuordnung programmatisch unterbleibt.

5.4 Analyse des Klassifikationsthesaurus

Nach Konstruktion des Thesaurus lassen sich 123.578 Begriffe zählen.[96] Dies bedeutet, dass etwa 42% oder 90.211 der 213.779 GND- Sachschlagwörter kein CrissCross-Mapping D2-D4 aufweisen, oder nicht zum ausgeschlossenen Entitätstyp Morpheme gehören.

Bemessen an den Begriffen, die überhaupt über ein CrissCross Mapping verfügen (133.872) erreichen wir eine Abdeckung von 92%. In der Differenzgruppe von 10.398 Begriffen finden sich somit sowohl diejenigen, welche lediglich eine Zuordnung in eine DDC-Nebentafel haben, lediglich ein Morphem sind, oder lediglich eine oder mehrere D1-Zuordnungen besitzen.

5.4.1 Verteilung der Begriffe auf die einzelnen Sachgruppen

Das in Abbildung 23 umgesetzte Pareto-Diagramm zeigt die Verteilung der Sachschlagwörter über die einzelnen Sachgruppen; die Pareto-Linie zeigt den Anteil der von links aufsummierten Zuweisungen am Gesamtbestand des Vokabulars, einschließlich ihrer Mehrfachzuweisungen.

[96] SKOS Konzepte, die Oberbegriffe in Sachgruppen besitzen.

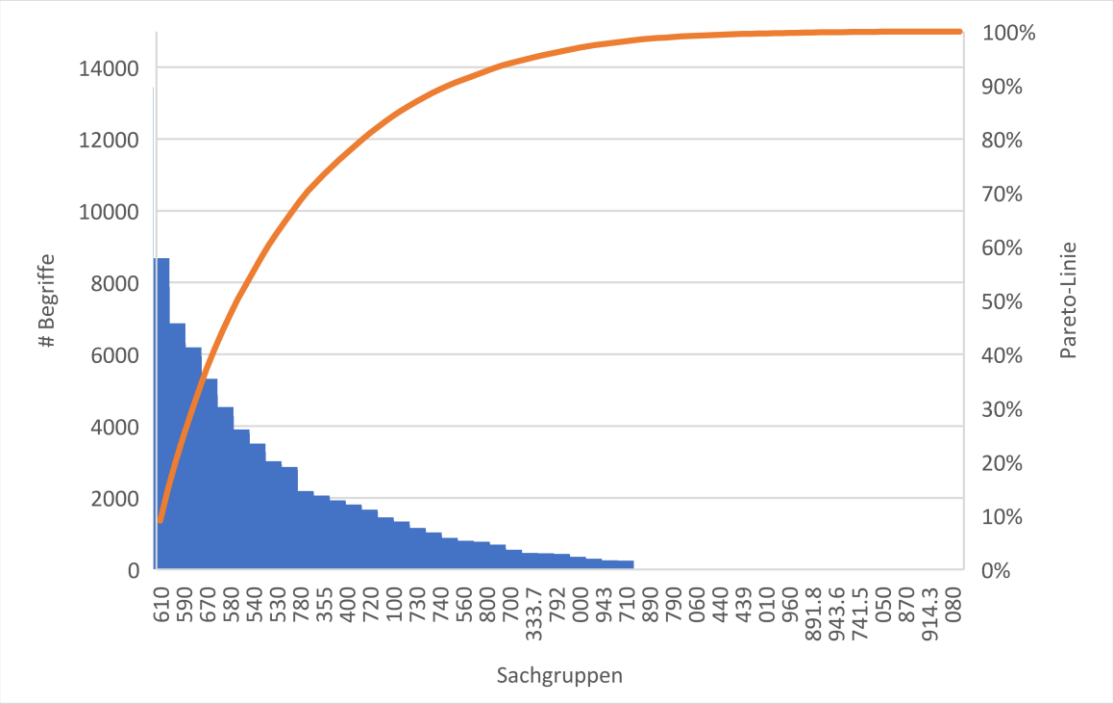


Abbildung 23: Verteilung Begriffe auf DDC-Sachgruppen

Die Sachgruppen 080 – Allg. Sammelwerke und 914.34 - Geografie, Reisen (Österreich) erhalten über die Zuordnung keine Schlagwörter und sind in der Auswertung nicht enthalten. Besonders umfangreich ist die Sachgruppe 610 – Medizin mit 13.449 Einträgen, die alleine fast 11% aller Begriffe zuordnet; die nächst schwächere Sachgruppe 570 – Biologie verfügt bereits über ein Drittel weniger Einträge.[97]

Im Mittel enthalten die Sachgruppen 451 Begriffe. Der Median fällt mit der Sachgruppe 792 – Theater, Tanz zusammen, welche 455 Schlagwörter enthält. In dieser Sachgruppe wird zugleich die 96%-Marke der Pareto-Linie für die zugeordneten Schlagworte überschritten.
Anders herum gesagt lässt sich somit feststellen, dass der unteren Hälfte der Sachgruppen lediglich 4% der Begriffe zugeordnet sind. Die untersten drei Dezilen (31 Sachgruppen) weisen im Schnitt lediglich 29 Begriffe pro Sachgruppe auf. Die Schlagwörter sind also eher ungleichmäßig verteilt, wobei ein nahezu hyperbelförmiger Abfall hin zu den spärlich besetzten Sachgruppen gegeben ist.

[97] Hinweis: die Beschriftungen der horizontalen Achse können in dieser Darstellung nicht vollständig erfolgen – Sachgruppe 570 findet sich im zweiten Balken von links.

Aus dieser Verteilung sind einige Erwartungen für die Qualität der Kategorisierung durch den Extraktor abzuleiten.

Generell ist eine Ungleichverteilung der Sachschlagwörter über alle Kategorien einem guten Ergebnis eher abträglich. Für ein gutes Ergebnis sollte jede Sachgruppe über ein für sein Fachgebiet vollständiges Vokabular verfügen. Diese sollten auch in ihrem Umfang keine allzu große Varianz aufweisen. Der konkret sinnvolle Umfang für eine Sachgruppe hängt in erster Linie von den spezifischen Gegebenheiten des Fachs, aber auch vom Zuschnitt des jeweiligen Fachgebietes ab. Auch in dieser Hinsicht sind die Sachgruppen eher inhomogen gestaltet. Während z.B. die Sachgruppen in Geschichte, Sprache und Literatur eher stark fragmentiert sind, umfasst die Sachgruppe 630 also: Landwirtschaft (darin Tierproduktion und Ackerbau), Forstwirtschaft, Veterinärmedizin, Fischerei und Jagd. Sie beschreibt all dies mit lediglich ca. 6.200 Begriffen.

5.4.2 Polyhierarchie des Klassifikationsthesaurus

Aus Sicht eines kompakten, vollständigen Wortschatzes ist Polyhierarchie durchaus wünschenswert. Es entspricht unseren Denkmustern, dass bestimmte Dinge sich nicht durch eine einzelne Zuordnung umfassend beschreiben lassen. Polyhierarchien entstehen hier durch Übersetzung der CrissCross-Mehrfachzuordnung der Schlagwörter. Für die Zuordnung eines Themas zu einer Kategorie stellt Polyhierarchie eine Herausforderung dar.

Je nachdem, ob man diesen Umstand aus Sicht der Schlagwörter oder aus Sicht der Oberbegriffe betrachtet, ergeben sich unterschiedliche Gesichtspunkte bei der qualitativen Beurteilung der Auswirkungen für die Kategorisierung von Dokumenten.

5.4.2.1 Polyhierarchie aus Sicht des gesamten Thesaurus

21.419 (14%) der Schlagwörter haben im verwendeten Thesaurus mehr als einen Oberbegriff. Teilt man die Summe der zugewiesenen Beziehungen aller Sachgruppen durch die Anzahl der Begriffe, ergibt sich für den durchschnittlichen Grad an Polyhierarchie *p*:

$p = 147.954 / 123.578 \approx 1{,}19$

Somit ist jeder Begriff im Schnitt etwa 1,19 Oberbegriffen zugewiesen. Für eine Kategorisierung bedeutet dies, dass mit einer gewissen Unschärfe zu rechnen ist, zu welcher Sachgruppe ein Begriff im konkreten Zusammenhang gehört.

5.4.2.2 Quantitative Verteilung auf die Schlagwörter

Polyhierarchie ist nicht gleichmäßig über den Thesaurus verteilt. Wie schon konstatiert, verfügen 14% der Begriffe über zwei oder mehr Beziehungen in unterschiedliche Sachgruppen. Die quantitative Verteilung dieser Beziehungen stellt sich wie folgt dar:

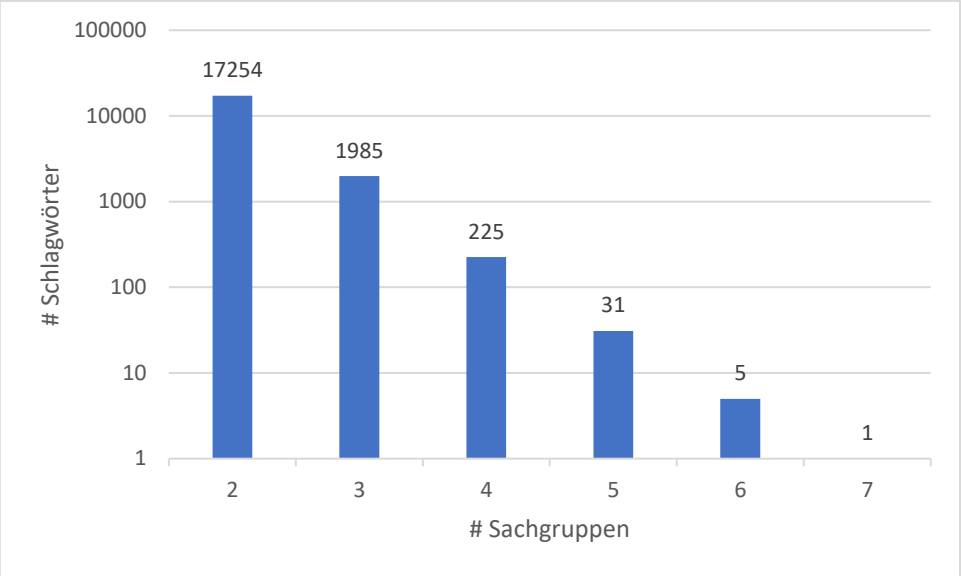


Abbildung 24: Verteilung polyhierarchischer Zuweisungen

Die Anzahl der Schlagwörter ist in Abbildung 24 logarithmisch aufgetragen. Die Polyhierarchie nimmt annähernd logarithmisch mit der Anzahl der zugeordneten Sachgruppen ab. Somit weisen nur ca. 1,8% der Begriffe mehr als zwei Oberbegriffe auf.

Mehrfachzuordnungen bedeuten, dass die Aussagekraft der Zuordnung bei der Kategorisierung geringer ist; in einer ersten Näherung wird man davon sprechen können, dass sich dies umgekehrt proportional zur Anzahl der Zuordnungen verhält. Für das in Kapitel 5.1 aufgeführte Beispiel „Flughafen“ betrüge das Gewicht also 1/7 eines nur einmal zugeordneten Schlagworts. Dieser Umstand sollte bei der Relevanzbewertung der Ergebnisse berücksichtigt werden.

5.4.2.3 Überlappung des Vokabulars der Sachgruppen

Da die Oberbegriffe im verwendeten Thesaurus Themengebiete darstellen, stellen die darunter angeordneten Schlagwörter (Themen) das Vokabular des jeweiligen Themengebietes dar. Das Vokabular führt dann zu guten Klassifikationsergebnissen, wenn es

i.) die Domäne vollständig abbildet und

ii.) gegenüber den anderen Fachgebieten abgeschlossen ist.

Da bestimmte Begriffe nun einmal in verschiedenen Fachgebieten relevant sind, sind diese beiden Bedingungen eventuell gegenläufig: Will man Mehrdeutigkeiten kategorisch ausschließen, ist das Vokabular eines anderen Sachgebiets unvollständig; ein vollständiges Vokabular kann also niemals gegenüber allen anderen Fachgebieten abgeschlossen sein.

Auch diese Überlappung fällt je nach Sachgruppe unterschiedlich aus. Besonders aufgefallen ist dabei die Sachgruppe 630 mit über 3.000 polyhierarchischen Prädikaten in andere

Sachgruppen (rund 50% der Konzepte). Auch die Sachgruppe mit den meisten Schlagwörtern, 610, überlappt mit 4.758 Beziehungen, also etwa einem Viertel ihrer Konzepte mit anderen Sachgruppen. Jedoch verteilt sich die Gesamtzahl dieser Mehrfachzuweisungen auf 40 Sachgruppen; lediglich ihre Überlappung mit Sachgruppe 570 fällt mit 2.747 Beziehungen signifikant aus.

Bemerkenswert ist insbesondere eine hohe Überlappung der Sachgruppen 300 und 150. Hier überlappen 486 Beziehungen paarweise, mithin ein Viertel der Konzepte aus Sachgruppe 150.

Da die Kategorisierung aufgrund der aggregierten Treffer aller Sachgruppen erfolgt, können große Mengen überlappender Begriffe mehrerer Sachgruppen eine Auswirkung auf das Ergebnis haben. Im Einzelfall hängt dies von der Relevanz der Begriffe im Dokument ab. Auch ist diesem Phänomen in gewissen Grenzen durch eine geeignete Gewichtung von polyhierarchisch zugeordneten Begriffen bei der Rangfolge der Ergebnisse zu begegnen.

5.5 Nachbearbeitung des Thesaurus

Die wesentlichen Nachbearbeitungen belaufen sich auf eine Behandlung problematischer Literale für die Benennungen, sowie die Erweiterung des Vokabulars durch Anwendung der Konkordanz mit den GND-Sachkategorien. Das unmittelbare Ziel hierbei ist, die Auffindbarkeit von Benennungen im Volltext bestmöglich sicherzustellen, sowie mögliche Nebenwirkungen von Sonderzeichen im RDF Datensatz zu verhindern.

5.5.1 Bereinigung der Literale

Aufgrund bestehender und historischer Ansetzungsregeln sind etliche Benennungen der GND Schlagworte nicht rein natürlichsprachig. Insbesondere wurden folgende Fälle identifiziert, die für die Indexierung jedenfalls problematisch erscheinen:

1. Nichtsortierzeichen in den Benennungen „@“. Hierfür sind in etwa 1.350 Fälle zu finden.
2. Benennungen enthalten identifizierende Zusätze in „<>“. Hierfür können 28.301 Fälle für bevorzugte Benennungen, sowie 18.652 Fälle für Synonyme gefunden werden.[98]
3. Durch „/“ zerlegte Benennungen, also Schlagwortketten (ca. 10.000)

[98] SELECT ?dis (STRBEFORE (STR(?schrott), "<") as ?newLabel)
WHERE {
?x gndo:variantNameForTheSubjectHeading ?dis
FILTER regex(?dis, "<")
}

4. Benennungen, in denen sowohl Zusätze als auch Zerlegungen vorkommen
5. Vorzugsbezeichnung und Synonym sind ident
6. Das Literal enthält das Muster „!!!GESPERRT!!!“
7. Benennungen in anderen Sprachen und Schriftsystemen
8. Durchgerutschte Morpheme, die zugleich subject heading senso stricto sind.
9. Entfernen von mehrfachen Whitespaces in Benennungen (programmatisch gelöscht)

5.5.1.1 Kunstsprachliche Zusätze

Kunstsprachliche Zusätze in Benennungen haben Nebenwirkungen bei der maschinellen Indexierung und sind deshalb für das gewünschte Resultat unerwünscht. Der Extraktor findet beim Vorhandensein solcher Zusätze eine Übereinstimmung nur für den Fall, in dem eine Benennung exakt so im Text vorkommt. Davon ist nicht auszugehen. Allerdings gilt das auch für die morphologische Verarbeitung und, insbesondere im Deutschen, für die Komposition.

Herkömmliche diesbezügliche Verfahren beruhen auf Wörterbüchern, in denen weder Schlagwortketten noch Zusätze enthalten sind. Die Bedeutung von Schlagwortketten in Benennungen ist auf rein morphologischer Ebene nicht lösbar (Gödert, 1990). Identifizierende Zusätze lösen also nicht das linguistische Problem der Disambiguierung. Beide Zusätze wirken sich lediglich negativ auf den Recall aus.

Die Nicht-Sortierzeichen werden laut mehrerer RSWK-§§ in den Benennungen kodiert (RSWK, 2017). Da sie so keine Funktion für den Extraktor besitzen und den Recall verringern, wurden sie systematisch aus Benennungen entfernt.

Identifizierende Zusätze sind gemäß RSWK in den bevorzugten Benennungen zu erwarten; vgl. § 10, RSWK. Diese sind der homonymen Bezeichnung angehängt. Die Tatsache, dass die Kodierung dieser Zusätze in den RDF Daten ausgerechnet in Spitzklammern erfolgt, ist für RDF und XML allerdings eher misslich, da diese dort anderweitig zur Auszeichnung verwendet werden. Erwartet wären Homonymenzusätze nach RSWK 2017 jedenfalls in runden Klammern. Ebenfalls laut Regelwerk nicht zu erwarten sind Zusätze am Synonym anstelle der Vorzugsbezeichnung (RSWK, 2017, p. 55 ff.).

Dieses Beispiel zeigt:

SW s Wilde Leute
BF Wilde Leute <Motiv>[99]

Inhaltlich finden sich bei den identifizierenden Zusätzen sowohl Jahreszahlen, als auch Zusätze in natürlicher Sprache. Als Bestandteil der Benennung sind diese Zusätze für einen Extraktor nur mit einer zusätzlichen, einigermaßen komplexen Heuristik nutzbar. Dies weist über den Umfang der gegenständlichen Arbeit weit hinaus und konnte deshalb unterbleiben.

Zunächst wurde versucht, die Zusätze in Spitzklammern einfach zu entfernen. Die Erwartung war, dass dies den Recall verbessert. Allerdings haben erste Testläufe ergeben, dass die Precision unter solchen Benennungen leidet. Beispielhaft seien hierfür die Begriffe „A <Buchstabe, Motiv>“[100] , bzw. „Hamburg <Schiff, Schiffsname>“[101] genannt.
Andere Versuche des Indexierens mit der GND behandeln das Thema nicht direkt, obwohl davon auszugehen ist, dass jedes Verfahren das auf einem Indexer beruht, davon wesentlich betroffen ist, insbesondere auch die Bachelor-Arbeit „Automatische Generierung von DDC-Notationen für Hochschulveröffentlichungen“ (Sommer, 2012). Bemerkenswert scheint auch, dass dieses Thema in einem der jüngeren Konzeptpapiere der DNB zur Weiterentwicklung der automatischen Erschließung zum ersten Mal explizit angesprochen wird.

Jedenfalls bin ich zu der Überzeugung gelangt, dass der Schaden durch den Verlust an Precision sowie der Aufwand für eine allfällige Disambiguierung den Nutzen des höheren Recall jedenfalls überwiegt. Daher wurde lediglich eine automatische Konvertierung der Spitzklammern in Rundklammern durchgeführt, sowie im Einzelfall die Zusätze natürlichsprachig aufgelöst (in der Regel über zusätzliche skos:altLabel).

5.5.1.2 Redundante Benennungen

Der Fall von redundanten Benennungen im selben Begriff tritt ohne weitere Transformationen nur in Einzelfällen auf. Hier wurde derart vorgegangen, dass die Dublette im skos:altLabel entfernt wird. Dieses ist ein übliches Verfahren für SKOS-Thesauri. Beim Entfernen von Zusätzen steigt die Anzahl an redundanten Benennungen drastisch an (siehe: „Wilde Leute“). Das Verfahren zur Bereinigung ist die Löschung des altLabel.

---

[99] https://d-nb.info/gnd/4189887-4/ (Wilde Leute)
[100] http://d-nb.info/gnd/1113036230
[101] http://d-nb.info/gnd/4329899-0

5.5.1.3 Gesperrte Begriffe

Mit der Zeichenfolge „!!!GESPERRT!!!“ beginnende Schlagwörter zeigen an, dass nach diesen nicht zu indexieren ist. Auf Begriffsebene funktioniert dies unmittelbar, da auch hier ein entsprechendes Vorkommen im Text auszuschließen sein wird. Eine umfassende Sperrung muss jedoch auch die Synonyme einschließen, da sonst über die Äquivalenzbeziehungen falsche Treffer generiert werden. Daher werden diese Konzepte mit einem *blocklist-flag* versehen.

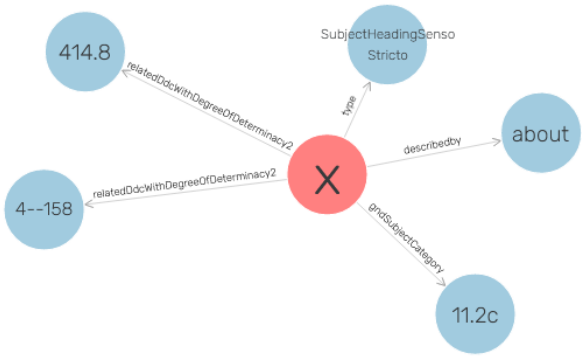


Abbildung 25: Schlagwort „X“

Ebenso zu sperren waren einige Morpheme, die überraschend über ein DDC-Mapping sowie den Entitätstyp subjectHeadingSensoStricto verfügten. Exemplarisch zeichnet sich das Schlagwort „x“ aus, dass sich dann schnell als häufigstes „Thema“ im Extraktor fand – wohl über die typische Verwendung in Tabellen einiger Hochschulschriften.

5.5.1.4 Benennungen in anderen Sprachen

Gelegentlich kommen Benennungen in anderen Sprachen oder aber auch Schriftsystemen vor. Diese sind nach § 308 RWSK zu erwarten; die Herausforderung besteht allerdings darin, diese systematisch in den Daten zu erfassen, da die GND für Literale keine language tags[102] verwendet. Da SKOS diese allerdings erfordert, wurden alle Benennungen mit dem language tag „@de“ versehen, d.h. als deutscher Wortschatz. Dies hat den Vorteil, dass sie, so lange sie in einem deutschen Text exakt vorkommen, ggf. gefunden werden können und den Recall verbessern.

[102] http://www.w3.org/1999/02/22-rdf-syntax-ns#langString

Zusammenfassend ist festzustellen, dass die vorkommenden Problemstellungen eine nicht unerhebliche Beeinträchtigung von Recall und Precision bedeuten können. Eine entsprechende Behandlung erfolgte also zurecht.

5.5.2 Erweiterungen über Konkordanz

Wie schon in Kapitel 4.3 erläutert, wurde die geschaffene Konkordanz dazu genutzt, das punktuell spärliche Vokabular zu verstärken. Praktisch durchgeführt wurde dies lediglich für die Sachgruppe 370 – Erziehungswissenschaften, da hierfür frühe Extraktionsversuche zunächst wirklich dürftige Recall-Werte gezeigt hatten. Hierzu wurden die aus einer zusätzlich erstellten Transformation nach GND-SC vorhandenen Teilbäume als RDF-Daten in den Thesaurus importiert, und als Unterbegriffe mit der Sachgruppe 370 verknüpft. Hierbei erhält der Thesaurus in etwa 1.600 Begriffe zusätzliches, spezifisches Fachvokabular aus dem Bereich der Erziehungswissenschaften.

6 Text Mining

Für das Text Mining wurde das Service PoolParty Extractor verwendet.
PoolParty Extractor ist als Web Service über eine RESTful API ansprechbar. Die Rückgabe erfolgt als JSON Dokument. Eine Dokumentation ist unter https://help.poolparty.biz öffentlich verfügbar.[103]

PoolParty Extractor bietet verschiedene Verfahren an, um Text zu klassifizieren, nach Termen und Konzepten zu indexieren, und Dokumente zu kategorisieren. Hierzu gehören unter anderem eine Named Entity Recognition (NER), ein Regular Expression Prozessor, ein KI- basierter Klassifikator, sowie der sogenannte *Main Extractor*, ein Lexem-basierter Tokenizer.
Hierbei handelt es sich um die für dieses Verfahren verwendete Hauptfunktion. Der folgende Abschnitt erläutert seine Funktionsweise.

Der *Main Extractor* extrahiert Einzelterme und Phrasen und vergleicht diese mit dem Wortschatz des Thesaurus. Entspricht ein Lexem dem einer Benennung eines Konzeptes im Thesaurus, führt dies zu einem Treffer. Es erfolgt sowohl eine Ausgabe der ermittelten Konzepte, als auch der sonstigen Terme in absteigender Relevanz. Die Verarbeitung erfolgt multilingual; es ist sowohl möglich die zu verwendenden Sprachen als API-Parameter anzugeben, als auch

[103] vgl. https://help.poolparty.biz/en/developer-guide/enterprise-server-apis/entity-extractor-apis/information-extraction-services/concept-extraction-service.html#web-service-method--extract-from-file

einen Text unbekannter Sprache an den Dienst zu übergeben, um die Sprache zu ermitteln. Die grammatikalische Verarbeitung beruht auf einer sprachabhängigen Lemmatisierung.

PoolParty Extractor verarbeitet Text, HTML, sowie sämtliche von Apache Tika[104] unterstützten Dokumentenformate, die digitalen Text enthalten. Die Vorverarbeitung schließt eine Spracherkennung, sowie eine Entfernung von Stoppwörtern und Behandlung von Sonderzeichen ein.

Abbildung 26 erläutert den Ablauf der Indexierung.

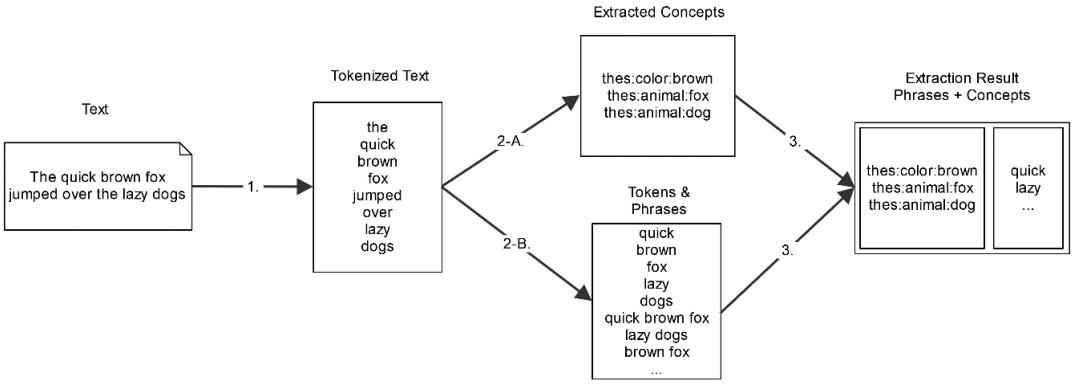


Abbildung 26: PoolParty Main Extractor

Nach der Vorverarbeitung werden folgende Schritte durchlaufen:

1. Tokenisierung:
   Der Text wird für die Verarbeitung in eine normalisierte Form umgewandelt, d.h. in eine Liste von Token
2. Concept Matching:
   Auf der Grundlage der gefundenen Tokens werden die möglichen Konzeptübereinstimmungen aus dem Index abgerufen. Im Allgemeinen wird beim Abgleich die Groß- und Kleinschreibung nicht berücksichtigt. Handelt es sich bei einem Token oder einer Konzeptbezeichnung um ein Ein-Wort-Akronym oder sind die Konzeptbezeichnungen als exakte Übereinstimmungen gekennzeichnet, erfolgt der Abgleich unter Berücksichtigung der Groß- und Kleinschreibung.
3. Term- und Phrasen Erkennung
4. Scoring:
   Die Relevanzbewertung von Begriffen und Konzepten basiert auf der Häufigkeit, der Position und der Länge der Phrase
5. Kategorisierung: (optional)
   Die Treffer im Thesaurus werden anhand der transitiven Oberbegriffe in Position der

[104] Apache Tika ist ein in Java geschriebenes Framework zur Erkennung und Analyse von Inhalten, das von der Apache Software Foundation verwaltet wird. Es erkennt und extrahiert Metadaten und Text aus über tausend verschiedenen Dateitypen. Siehe https://tika.apache.org/

obersten Klasse aggregiert. Diese werden entweder anhand der Anzahl pro Klasse oder anhand der aggregierten Relevanzen bewertet.

6. Ausgabe:
   Anhand der Aufrufparameter erfolgt die Ausgabe von Termen, Konzepten und Kategorien.

6.1.1 Gewichtung für Terme und Konzepte

Die Relevanz eines bestimmten Konzepts oder Terms wird anhand von Frequenz und Position im Dokument bewertet. Alle Bewertungen für Konzepte und Terme werden getrennt zwischen 0… 100 normalisiert.

Das Gewicht der Positionen ermittelt sich wie folgt:

$$\sum_{p_0}^{p_n} (W - 0.75p^2)/W^2$$

(Mit *p* = Position des Terms und *W* = Gesamtzahl der Wörter im Text.)

Somit fällt das Positionsgewicht des Auftretens eines Treffers mit der Lauflänge des Textes. Am Anfang auftretende Treffer werden höher gewichtet; die Position im letzten Wort des Dokuments besitzt noch ¼ des Positionsgewichts des ersten Wortes.

Phrasen werden zusätzlich mit zunehmender Anzahl Wörtern bewertet. Der Faktor beträgt ½ (n); die Benennung „Field programmable gate array“ erhält also das Gewicht 2.

Alternative Scoring-Verfahren, einschließlich TF/IDF sind möglich. Sie erfordern allerdings ein Referenzkorpus, welches für diese Arbeit nicht erstellt wurde. Einen Überblick über die vollständigen Scoring-Optionen von PoolParty Extractor findet sich in der API-Dokumentation.[105]

6.1.2 Kategorisierung

Die Gewinnung der Kategorisierung ist in das Service integriert. Hierbei ordnet der Extraktor die im Text gefundenen Treffer für Konzepte anhand ihres Kontextes im Thesaurus, wobei die jeweiligen skos:TopConcept als Kategorien ausgegeben werden. Diese werden in ihrer Rangfolge des ermittelten Gewichtes ausgegeben.
Damit unterscheidet sich das Verfahren vom typischen Verhalten ML-basierter Klassifikatoren, welche in der Regel ein binäres Ergebnis (true/false) liefern und allenfalls noch einen Konfidenzwert angeben.

---

[105] https://help.poolparty.biz/en/developer-guide/enterprise-server-apis/entity-extractor-apis/information-extraction-services/concept-extraction-service.html#web-service-method--extract-from-file

6.1.3 Gewichtung der Kategorien

Die Relevanz der gefundenen Kategorien kann in PoolParty auf zwei Arten ermittelt werden:

i. Die Anzahl der gefundenen Konzepte pro Kategorie wird durch die Anzahl der gefundenen Kategorien geteilt. Dies ist das Standardverfahren.

ii. Die Summe der Scores der pro Kategorie gefundenen Konzepte wird durch die Summe der Scores aller gefundenen Konzepte geteilt; dabei werden Gewichte von Konzepten die mehreren Kategorien zugeordnet sind, auch mehrfach gezählt. Diese Methode wird durch den API-Parameter „categorizationWithPpxBoost = true“ aktiviert.

Während das erste Verfahren vom bloßen Vorkommen eines Schlagworts einer bestimmten Kategorie abhängt, hat das zweite Verfahren den Vorzug, sowohl die Relevanz der Treffer zu berücksichtigen, als auch die geringere Relevanz von Treffern polyhierarchischer Konzepte zu modellieren, da jedes Vorkommen jeder Zuordnung sich auf die relativen Gewichte der Kategorien anteilig verteilt. Für die Kategorisierung wurde daher die zweite Methode verwendet. Siehe hierzu auch Kapitel 5.4.2.2. Für beide Verfahren sind die Scores der Kategorien zwischen 0 und 1 normalisiert.

6.1.4 Linguistische Behandlung

Die linguistische Behandlung bei der Klassifikation in PoolParty Extractor beschränkt sich im Grunde auf Erkennung der Sprache[106], Tokenisierung, eine Filterung sprachbezogener Stoppwörter, eine wörterbuchbasierte Lemmatisierung, und eine thesaurusbasierte Disambiguierung. Insbesondere erfolgt kein part-of-speech (POS)-Tagging. Diese Methode ist prinzipiell für die zuverlässige Erkennung einzelner Schlagwörter als Themen prinzipiell hilfreich, da sie erlaubt, lediglich thematisch tragende Nominalsätze zu indexieren. Da bei diesem Vorhaben sehr lange Texte indexiert wurden, ist dies für eine aggregierte Auswertung aller Treffer nicht unmittelbar notwendig.

Es ist möglich, die grammatikalische Behandlung auf Konzeptebene abzuschalten. „Exakt Match“ führt dazu, dass das Schlagwort nur dann gefunden wird, wenn es im Text exakt so wie in einer seiner Benennungen vorkommt.[107] Diese Option wurde fallweise zur Ausfilterung von Überschneidungen von Individualbegriffen mit Allgemeinwörtern verwendet, falls diese durch

[106] Eine mehrsprachige Indexierung ist möglich, war hier aber wegen der fehlenden Mehrsprachigkeit des Thesaurus nicht relevant.
[107] vgl. https://help.poolparty.biz/en/poolparty---release-notes/release-notes---poolparty-7-0.html#exact-matching-in-entity-extraction

Lemmatisierung entstanden sind, oder wenn sich diese durch gemischte Groß- und Kleinschreibung ergeben konnte.[108]

Dieser relativ einfache Ansatz der linguistischen Verarbeitung ist für das hier durchgeführte Verfahren völlig ausreichend. Auch wenn es hierdurch vereinzelt zu Fehlzuweisungen kommen mag, sind diese über alle Schlagwörter hinweg nicht gravierend genug, um das Ergebnis der Kategorisierung in relevantem Ausmaß zu verfälschen.
Diese Aussage gilt insbesondere im Zusammenhang mit der Lauflänge der untersuchten Dokumente: Diese betrug wenigstens 40 Seiten, im Mittel zwischen 150 – 200 Seiten, und überstieg in manchen Fällen 1.000 Seiten.

Bei der Klassifizierung kürzerer Dokumente ist es recht wahrscheinlich, dass eine umfangreichere Vorverarbeitung, insbesondere POS-Tagging, merkbare Auswirkungen auf das Ergebnis haben wird.

Wesentlich ist jedoch die Disambiguierung anhand des Thesaurus. PoolParty nutzt hierzu die vorkommenden homographen Benennungen, sowie deren Distanz im Graphen. Selbst mit der weitgehend flachen Struktur des hier verwendeten Thesaurus kommt es dadurch noch zu minimalen, jedenfalls aber immer noch rangordnenden Unterschieden anhand der jeweiligen Oberbegriffe.

7 Validierung der Methodik.

Die Validierung orientiert sich an einem von Koraljka Golub vorgeschlagenen Modell, das für ein vergleichbares Projekt der Schwedischen Nationalbibliothek entwickelt wurde. Golub stellt ein hierzu ein mehrstufiges Verfahren vor, dass die übliche Vorgehensweise formalisiert.

i.) Bewertung der Indexierung durch Vergleich mit einem Goldstandard;

ii.) Bewertung der Qualität der computergestützten Indexierung im Kontext eines Indexierungs-Workflows

iii.) Bewertung der Indexierungsqualität durch Analyse der Retrievalleistung. (Golub, 2016)

Hierbei ist die Untersuchung auf die Frage beschränkt, in welche Hauptsachgruppe das Dokument fällt. Auch wenn es möglich wäre, bis zu drei Sachgruppen für ein Dokument zu vergeben

[108] Die Grundeinstellung ist case-insensitive, Benennungen in GROßBUCHSTABEN werden immer als Eigennamen behandelt und nicht gebeugt.

(Beyer & Trunk, 2011), gehen wir hier vereinfachend davon aus, dass es nur eine korrekte Vorhersage pro Dokument gibt.

Dieses Verfahren ist bei der Verwendung binärer Klassifikatoren somit direkt umsetzbar. Für eine rangordnende Klassifikation ist es ggf. zu erweitern. Da die üblichen Metriken Precision und F1-Maß nur für eine harte Kategorisierung oder Klassifikation unmittelbar interpretierbar sind, wurde die Standardmethodik für die hier gewünschte rangordnende Ausgabe durch ein hierfür einschlägiges Maß, den *mean reciprocal rank*, ergänzt.

7.1 Erstellung eines Goldstandards

Für die Erstellung eines Referenzkorpus ist eine bestimmte Menge an Dokumenten mit bekannten Eigenschaften heranzuziehen. Der konkrete Untersuchungsgegenstand sind Hochschulschriften, insbesondere Dissertationen und Habilitationsschriften, die als digitaler Text vorliegen. Aufgrund der in Kapitel 2 dokumentierten, umfangreichen Erschließungsarbeiten, sowie dem expliziten Sammlungsauftrag war die Vermutung naheliegend, dass solche in ausreichender Anzahl aus dem Online-Katalog der DNB bezogen werden konnten. Zur Auswahl der Dokumente kommen folgende Kriterien in Betracht:

Eigenschaft	**Muss-Kriterium**	**Validierung**	**Erwartung erfüllt?**	**Kommentar**
Dokument frei verfügbar	Muss	Halbautomatisch	Ja	Metadaten in Ausnahmen falsch
Digitaler Text	Muss	Manuell oder Fehlermeldung des Extraktors	Nein	Dokumente mit Bilddateien gefunden, in Jus große Mengen an Retrodigitalisierungen
Dokumentensprache Deutsch	Muss	Metadaten, Autopsie	Nein	Metadaten zum Teil falsch, manuelle Vorprüfung effizienter
Sachgruppe annotiert	Muss	Filter	Nein	Siehe Text unten.
Sachgruppe eindeutig	Muss	Metadaten, Autopsie	Ja	Mehrfachzuweisungen sind auszusondern
Dokumentenart Hochschulschrift, Dissertation	Soll	Metadaten	Ja	In Einzelfällen Masterarbeiten
Einordnung des Themas im Themengebiet	Kann	Manuell, Metadaten Zufall	Ja	Stratifikation über das gesamte Themengebiet erwünscht, bei großen Mengen nur begrenzt leistbar

Tabelle 3: Auswahlkriterien Dokumente

Die DNB-Suchmaschine bietet eine Suche nach DDC-Sachgruppen an.[109] Die ursprüngliche Absicht war also, das Korpus programmatisch durch zufällige Auswahl aus dem Suchergebnis aus dem Katalog der DNB aufzubauen. Dies stellte sich als undurchführbar heraus. Die wesentlichen Gründe hierfür waren:

- Der Umfang der Suchergebnisse nach den obigen Kriterien war unerwartet klein; in manchen Sachgruppen betrug sie weniger als 100 Dokumente (einschließlich irrelevanter Resultate)
- Die Menge der anhand der Metadaten im DNB Katalog nicht-relevanten Ergebnissen war durchwegs sehr hoch; für manche Sachgruppen-Filter betrug diese Rate deutlich mehr als 50% der Treffer.

Manche Herausforderungen waren vorhersehbar. Insbesondere war vorher klar, dass das Filter auch Dokumente zurückgeben würde, bei denen nur eine der annotierten Sachgruppen dem Suchwert entspricht. Auch bestanden keine Illusionen bezüglich der Spracherkennung – eine deutsche Zusammenfassung führt eventuell zu Fehlerfassungen. Auch war bereits bekannt, dass im Katalog zwischen Scans und digitalem Text nicht unterschieden wird.

Das schwerwiegendste und auch in diesem Ausmaß unerwartete Problem war die Ausgabe von Ergebnissen, die nicht den Suchkriterien entsprachen, insbesondere Treffer ohne Sachgruppenzuweisung, Treffer erfasst nach anderer Systematik als der DDC, aber eben auch Treffer mit Einfachzuweisungen anderer DDC-Sachgruppen, als im Suchfilter angegeben. Dazu kamen weitere Qualitätsprobleme, wie falsche Sprachangaben, fehlende oder unzugängliche Dokumente und offenkundige Inkonsistenzen zwischen zugewiesener Sachgruppe und DDC Kurznotation.

Angesichts der sich daraus ergebenden Unmöglichkeit, eine halbwegs zuverlässige Grundgesamtheit zu ermitteln, wurde die Arbeit mit Stichproben methodisch verworfen. Hinzu kam die Einsicht, dass die zu untersuchende Problematik ohnehin keiner üblichen stochastischen Verteilung folgt.

Dokumentenzentrisch betrifft dies das Publikationsaufkommen. Dies belegen im Grunde sämtliche Auswertungen aus dem PETRUS Programm (vgl. Kapitel 2.5), die ständige bibliothekarische Erfahrung, sowie insbesondere die Forschungsarbeit von Golub, die dann zu der hier verwendeten Vorgehensweise führte. Aus klassenzentrischer Sicht ergab sich zudem im Laufe der

[109] Eine entsprechende Suche ist durch Ausführung des folgenden http-Aufrufs in einem Browser nachvollziehbar, hier für Sachgruppe 330: https://portal.dnb.de/opac/moveDown?currentResultId=woe+all+%22*diss%22%26any%26online%26**sg330**&categoryId=onlinefree

Arbeit die Erkenntnis, dass die Schlagwörter nicht nur ungleich auf die Kategorien verteilt sind, sondern auch, dass das vorhandene Vokabular in manchen Fachgebieten diese nur unzureichend repräsentiert. Siehe hierzu Kapitel 5.4.1.
Ohne Vokabular ist eine korrekte Zuordnung eben wenig aussichtsreich.

Daher wurde die Entscheidung getroffen, die Dokumente für den benötigten Goldstandard manuell auszuwählen und sich bei den ausgewerteten Sachgruppen an die Gruppe anzulehnen, die für die automatische Beschlagwortung in PETRUS ausgewertet wurden (Uhlmann, 2013).

Die Anzahl der Dokumente pro Klasse wurde auf 50 festgelegt. Dies entspricht im Wesentlichen den Standards im PETRUS Projekt; dort betrug die Stichprobengröße 30. Ein Projekt der Schwedischen Nationalbibliothek beschränkte die Untersuchung ebenfalls auf ausreichend dokumentierte Gebiete und arbeitete mit ähnlichen Stichprobengrößen (Golub, 2016). Auch Sommer kam am Ende auf ein Korpus von lediglich 35 ausgewerteten Dokumenten (Sommer, 2012, p. 52).

Letztlich wurden insgesamt 700 Dokumente aus 14 Sachgruppen untersucht, wobei die von PETRUS getroffene Auswahl der Sachgruppen weitgehend übernommen wurde. Tabelle 4 liefert hierzu eine Übersicht.

Sachgruppe	# Dokumente PETRUS	# Dokumente Thesaurus	Bemerkung
004 – Informatik	30	50	
020 – Bibliothekswissenschaft	0	50	Ersatz für SG 900
100 – Philosophie	30	50	
150 – Psychologie	30	50	
320 – Politik	30	50	
330 – Wirtschaft	30	50	
340 – Recht	30	50	
370 – Erziehung	30	50	Wortschatz durch Konkordanz ergänzt
510 – Mathematik	30	50	
530 – Physik	30	50	
610 – Medizin	30	50	
620 – Ingenieurw.	30	50	
630 – Landwirtsch.,Veterinärmed.	30	50	
780 – Musik	0	50	Ersatz für SG 830
830 – Deutsche Literatur	30	0	Kein Wortschatz, keine Dokumente
900 – Geschichte	30	0	Kein Wortschatz, keine Dokumente

Tabelle 4: Untersuchte Sachgruppen

Diese Sachgruppen decken das des relevante Publikationsaufkommens in hohem Maße ab, insbesondere da Medizin, Ingenieurswissenschaften, Wirtschaft und Pädagogik im Satz enthalten sind.

In zwei Fällen musste von der Auswahl bei Uhlmann abgewichen werden: Für die Sachgruppen 900 und 830 wäre es nur unter einer wesentlichen Abänderung der Kriterien möglich gewesen, entsprechende Dokumente zu finden. Hier wirkte sich die hohe Ungenauigkeit der Filterung in der DNB Suchmaschine massiv aus. Zudem gehören beide Sachgruppen zu Bereichen der DDC, die hochgradig differenziert, aber deshalb auch stark fragmentiert sind. Es wäre wohl mit einigem Aufwand und dem Fachwissen der Referent_Innen für Literatur und Geschichte möglich gewesen, aus den umliegenden Sachgruppen relevante Dokumente zu identifizieren.

Für die geringe Anzahl von anhand der DNB- Metadaten für Sachgruppen 830 und 900 als relevant identifizierbaren Dokumenten konnte festgestellt werden, dass diese vom Thesaurus nur unzureichend beschrieben werden. Das Vokabular in Sachgruppe 830 besteht im Wesentlichen aus Individualbegriffen von Literaturpreisen und -stipendien. Das Vokabular in 900 besteht aus

weniger als 100 Begriffen, die zudem als nicht besonders spezifisch für das Fach zu erachten sind.

Die hohe Fragmentierung der Sachgruppen im Bereich der Geschichte und Literatur erscheint somit sowohl hinsichtlich des Vokabulars als auch hinsichtlich des Publikationsaufkommens etwas fragwürdig. Eine systematische, anhand des Vokabulars evidenzbasierte Untersuchung für einen praktikablen Zuschnitt der Sachgruppen wäre sicherlich wünschenswert, ist aber nicht Gegenstand dieser Arbeit.

Es wäre denkbar und sogar aussichtsreich gewesen, die Begriffe jeweils unter den DDC-Klassen 800 und 900 nach oben zu aggregieren; dies hätte allerdings einen methodischen Bruch bedeutet.

Aufgrund meiner eigenen Kompetenzen bin ich somit auf die Sachgruppen 020 - Bibliothekswissenschaft und 780 - Musik ausgewichen. Auch wenn hier zum Teil auf die BASE- Datenbank zurückzugreifen war, um die erforderliche Anzahl an relevanten Dokumenten in Bibliotheks- und Musikwissenschaft aufzufüllen, erschien das Vorhaben, jeweils 50 relevante Dokumente zu identifizieren und gegen einen ausreichend spezifischen Wortschatz zu indexieren, aussichtsreich und umsetzbar.
Die Kategorisierung erfolgte mit einem für Sachgruppe 370 durch Konkordanz erweiterten Wortschatz, siehe Kapitel 5.5.2.

Aufgrund der schon in ersten Probeläufen identifizierten Herausforderungen, erfolgte zur Auswahl eine Kurzautopsie für jedes Dokument laut folgenden Kriterien:

i. Überprüfung der Metadaten: Korrekte Sachgruppe, sowie keine offensichtlichen Widersprüche zu weiteren Notationen
ii. Keine offenkundigen Widersprüche des aus dem Titel ersichtlichen Themas zu den Metadaten
iii. Ausschluss identifizierbarer Dubletten
iv. Ausschluss bei offenkundigen Widersprüchen anhand des angestrebten akademischen Grades oder der Fakultät; dies war insbesondere in den Wirtschaftswissenschaften und den Ingenieurswissenschaften hilfreich.
v. Prüfung des Dokuments auf vorhandenen digitalen Volltext und dessen Sprache

7.2 Indexierungsworkflow

Die Dokumente werden über API-Aufruf an das PoolParty Webservice geschickt. Hierbei wurde das Dokument als Body eines POST-Aufrufs übergeben. Der Thesaurus enthält dabei alle 104 Sachgruppen, von denen 14 ausgewertet werden, wie im vorigen Kapitel beschrieben. Tabelle

5 beschreibt die in der Folge verwendeten Parameter des Aufrufs. Die Hintergründe hierzu sind in Kapitel 6 erläutert.

Parameter	Wert	Erläuterung
projectId		Angabe des Thesaurus, über den indexiert wird. Der Wert enthält die UID des Projektes.
language	de	Extraktionssprache
categorize	true	Rangordnende Ausgaben der TopConcepts als Kategorien, deren zugeordnete Begriffe im Text gefunden wurden.
categorization-WithPpxBoost	true	Berücksichtigt Position und Polyhierarchie der Begriffe. Siehe Kap. 6.1.3
numberOfTerms	0	Unterdrückt die Indexierung freier Schlagwörter
numberOfCon-cepts	5	Ausgabe der 5 relevantesten Begriffe des Dokuments zu informativen Zwecken, z.B. hier Entdeckung problematischer Begriffe.
disambiguate	true	Disambiguiert nach den für das Projekt hinterlegten Regeln.

Tabelle 5: API-Parameter für Indexierungsworkflow

Für jedes Dokument im Korpus wird ein JSON- Dokument ausgegeben, das die ermittelten Metadaten enthält. Ein entsprechendes Resultat stellt sich exemplarisch wie folgt dar:

```
{
  ▾ document: {
      ▸ concepts: […],
      ▾ categories: [
          ▾ {
              prefLabel: "320 - Politik",
              uri: http://dewey.info/class/320/,
              score: 0.17058011049723756,
            ▸ categoryConceptResults: […]
            },
          ▾ {
              prefLabel: "300 - Sozialwissenschaften, Soziologie, Anthropologie",
              uri: http://dewey.info/class/300/,
              score: 0.12810773480662985,
            ▸ categoryConceptResults: […]
            },
          ▾ {
              prefLabel: "510 - Mathematik",
              uri: http://dewey.info/class/510/,
              score: 0.06975138121546962,
            ▾ categoryConceptResults: [
                ▾ {
                    uri: https://d-nb.info/gnd/4038613-2,
                    prefLabel: "Menge",
                    score: 2.0
                  },
                ▾ {
                    uri: https://d-nb.info/gnd/4300606-1,
                    prefLabel: "Schiefe Wahrscheinlichkeitsverteilung",
                    score: 1.0
                  },
                ▾ {
                    uri: https://d-nb.info/gnd/4078859-3,
                    prefLabel: "Versuchsplanung",
                    score: 1.0
                  },
                ▾ {
                    uri: https://d-nb.info/gnd/4137007-7,
                    prefLabel: "Wahrscheinlichkeit",
                    score: 3.0
                  },
                ▾ {
                    uri: https://d-nb.info/gnd/4056995-0,
                    prefLabel: "Statistik",
```


Abbildung 27: Ergebnis für Dokument IDN 999901796

Das JSON Dokument würde in der hier vorgesehenen Darstellung über 230 Druckseiten einnehmen. In der eingeklappten Darstellung in Abbildung 27 sind die Kategorien der ersten drei Ränge, sowie einige Schlagwörter der dritten Kategorie mit ihren jeweiligen Scores sichtbar.

In der Sektion „concepts“ finden sich die extrahierten Einzelschlagwörter im Rang sortiert. Es wurden zusätzlich zu den Kategorien fünf Schlagwörter extrahiert. Dies diente als zusätzliche Information, zudem konnte so schnell herausgefunden werden, ob im Thesaurus noch Anomalien bezüglich einzelner Begriffe bestehen.

Für die Auswertung wurden die relevanten Metadaten in eine Tabellenkalkulation übernommen. Der gesamte Ablauf erfolgte manuell, wobei die Arbeitsschritte mit der ohnehin manuellen Auswahl des Goldstandards verschränkt wurden. Für einen Prototypen ist dies eine angemessene Vorgehensweise; eine Automation für einen Produktionsbetrieb ist mit üblichen Methoden unschwer durchführbar.

7.3 Messung der Retrievalleistung

Die Qualität der Indexierung lässt sich indirekt an der damit verbundenen Retrievalleistung ablesen. Üblicherweise werden hierbei Recall (die Vollständigkeit) des Suchergebnisses und Precision (die Genauigkeit) des Suchergebnisses ermittelt.

Das ideale Suchergebnis würde bedeuten, dass alle Dokumente genau diese Sachgruppe im höchsten Rang erhalten, das dem Erwartungswert entspricht. Erwartungswert und Ergebnis können wir also wie folgt vergleichen:

$$Recall\ der\ Kategorisierung = \frac{Anzahl\ der\ korrekt\ kategorisierten\ Dokumente}{Anzahl\ der\ Dokumente\ im\ Goldstandard}$$

$$Precision\ der\ Kategorisierung = \frac{Anzahl\ der\ korrekt\ kategorisierten\ Dokumente}{Anzahl\ kategorisierte\ Dokumente}$$

Diese Definitionen folgen den Vorschlägen von Golub (Golub, 2016, p. 8).

Betrachten wir die Ergebnisse im ersten Rang, erhalten wir eine harte Kategorisierung. Hierfür sind Recall und Precision anhand der üblichen Vorgehensweisen direkt abzulesen.[110]

Schließlich sind auch Precision und Recall in ein Verhältnis zu setzen.

Der Recall im Information Retrieval kann im Allgemeinen ganz einfach dadurch verbessert werden, indem man die Spezifizität der Anfrage herabsetzt. Ebenso sinkt mit einer Erhöhung der Genauigkeit in der Regel der Recall. Um die Auswirkungen dieser gegenläufigen Eigenschaften für das untersuchte System darzustellen, gibt man zusätzlich das sog. F1-Maß, also das Harmonische Mittel von Precision und Recall, an. Dieses lautet nach üblicher Definition:

$$F_1 = 2\frac{Precision\ \mathrm{x}\ Recall}{Precision + Recall}$$

Im Idealfall nimmt das F1-Maß den Wert 1 an: Die ideale Suchmaschine findet also alle relevanten Dokumente und schließt alle irrelevanten Dokumente aus. Für den konkreten Anwendungsfall bestehen hierzu sowohl klassenzentrische als auch dokumentenzentrische Einflussgrößen. Diese wurden in Kapitel 5.4 im Wesentlichen erörtert.

Neben der Leistung bei einer harten Kategorisierung, ist es hier ebenfalls interessant die Leistung einer rangordnenden Kategorisierung zu untersuchen. Im Grunde führt ist die harte

[110] vgl. https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=1044361798

Kategorisierung ein Sonderfall des angewendeten Verfahrens: Die Sachgruppen werden in absteigender Rangfolge der Treffer zurückgegeben; mithin bis zu 102 Ränge. Wertet man nur die Ergebnisse auf Rang 1, entspricht dies der Ausgabe eines binären Klassifikators.

Beabsichtigt man die vollen Eigenschaften des Verfahrens, insbesondere für eine Verwendung in einem halbautomatischen Workflow zu nutzen, sind jedoch auch die Ergebnisse der folgenden Ränge relevant. Dabei werden vom Extraktor die Ergebnisse der obersten Ränge *k* ausgegeben. Es erfolgt eine abschließende intellektuelle oder automatische Entscheidung über die zutreffende Sachgruppe.

Hierbei können wir betrachten, ob in den folgenden Rängen ebenfalls korrekte Resultate zurückgegeben werden, bzw. kann gezählt werden, ob die gesuchte Sachgruppe zurückgegeben wurde. Dadurch erhalten wir den Recall für diese Ränge.

Zusätzlich ist es für dieses Szenario interessant, auf welchem durchschnittlichen Rang das korrekte Ergebnis einer Sachgruppe zurückgegeben wurde. Hierfür wird der *mean reciprocal rank* (MRR) der erwarteten Sachgruppe angegeben.[111]

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank\ (i)}$$

MMR ist das arithmetische Mittel des Kehrwerts der aufsummierten Ränge der korrekten Vorhersagen. Die Kardinalität |Q| entspricht hierbei der Anzahl der Anfragen, also hier der Anzahl der Dokumente. MRR kann nach Sachgruppen segmentiert angegeben werden. Damit enthält man den durchschnittlichen Rang des richtigen Ergebnisses pro Sachgruppe.

8 Auswertung und Diskussion der Ergebnisse

Die Auswertung erfolgte für alle 700 Dokumente, also 50 Dokumente in 14 untersuchten Sachgruppen laut Kapitel 7.1.

Für jedes Dokument wurde jeweils das Resultat für die harte Kategorisierung sowie die rangordnend zurückgegebenen Ergebnisse bis hin zur gesuchten Sachgruppe untersucht. Dazu wurden die im Indexierungsworkflow ermittelten Metadaten in eine Tabelle übertragen, programmatisch mit den Erwartungswerten verglichen, sortiert und gruppiert ausgewertet.

[111] https://en.wikipedia.org/w/index.php?title=Mean_reciprocal_rank&oldid=1028668019

Die Gültigkeit der Auswertung ergibt sich innerhalb der untersuchten Sammlung. Dieser Umstand ist für Bibliotheksbestände charakteristisch; ein expliziter Hinweis dazu findet sich in Abschnitt 4 des Artikels „Automatic subject indexing of text“ in der Fachenzyklopädie der International Society of Knowledge Organization (ISKO) (Golub, 2019). Die Ergebnisse sind somit auf unbekannte Dokumente übertragbar, insofern diese den Dokumenten im Referenzkorpus vergleichbar sind.

8.1 Verhalten bei harter Kategorisierung

Zur Beurteilung des Verhaltens bei harter Kategorisierung untersuchen wir F1-Maß, Recall und Precision. Hierbei untersuchen wir die quantitative Leistungsfähigkeit für das Information Retrieval von Dokumenten. Dabei besitzen Dokumente und vorhergesagte Kategorien dieselbe Kardinalität, also bei der gegebenen Grundgesamtheit 50 Dokumente pro Kategorie. Es wird für jedes Dokument eine Kategorie vorhergesagt: diejenige, die vom Extraktor im höchsten Rang ausgegeben wird.

8.1.1 F1 Maß

Das F1 Maß gibt Auskunft für die globale Leistung für das Information Retrieval.

Über alle Sachgruppen hinweg erreicht die Indexierung im Mittel ein F1-Maß von 74%.

Die Spitzengruppe bilden die Sachgruppen 100, 340, 530, 610 und 780, bei denen die 80%-Schwelle erreicht oder überschritten wird.

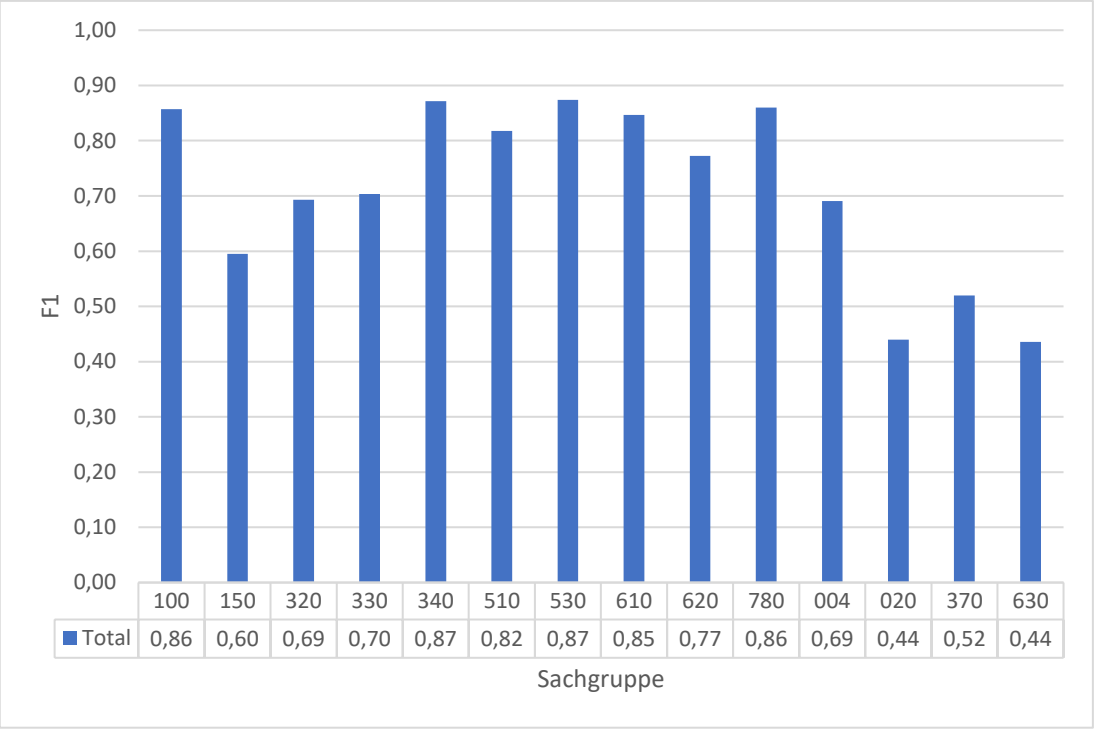


Abbildung 28:F1-Maß pro Sachgruppe

Die Sachgruppen 020, 370, 630 und 150 erreichen Werte deutlich unter 70%, wobei 020 und 630 die Schlusslichter bilden. Recall und Precision geben Auskunft darüber, wie diese Werte zustande kommen.

8.1.2 Recall

Der Recall gibt in diesem Fall an, welcher Anteil der Dokumente im Referenzkorpus auf Rang 1 korrekt kategorisiert wurde.

Einen Hinweis auf die generelle Leistungsfähigkeit in diesem Szenario gibt der Median. Dieser beträgt 77%.

Über die untersuchten Sachgruppen weisen diese Werte eine Schwankungsbreite zwischen 44% für die Sachgruppen 020 und 630 und 96%, bzw. 94% für die Sachgruppen 100 und 610 auf. Die Sachgruppe 370 kommt trotz eines aufgefetteten Vokabulars auf einen Recall von 52%.

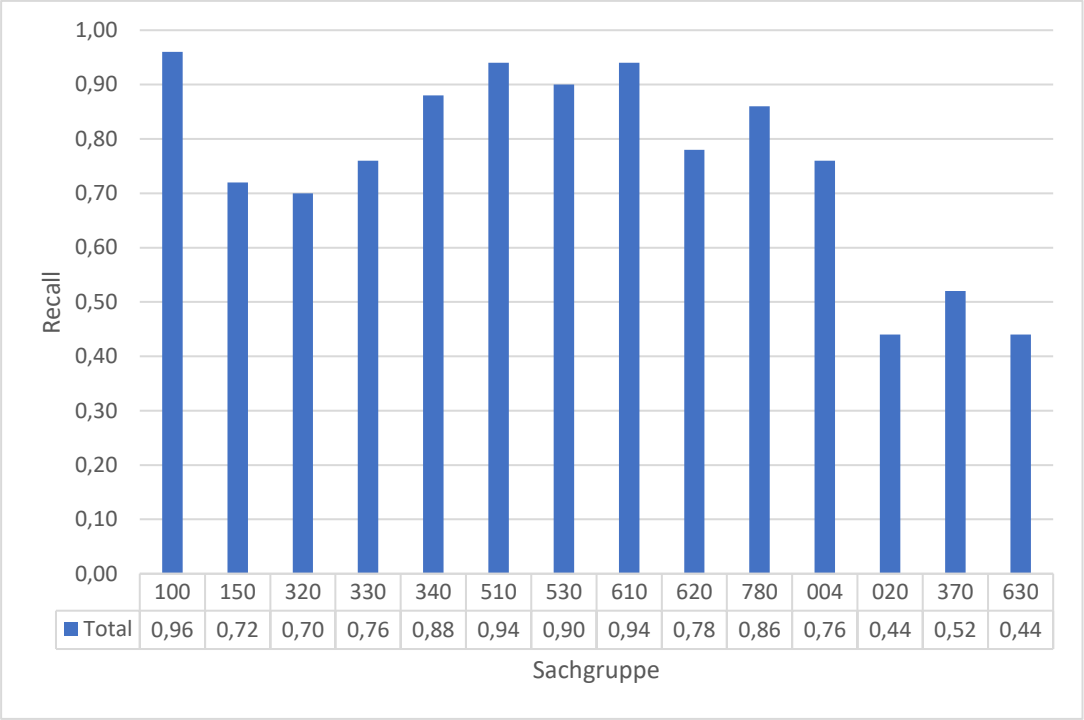


Abbildung 29:Precision pro Sachgruppe

8.1.3 Precision

Precision gibt uns zusätzlich Auskunft darüber, welche Dokumente anderer Sachgruppen fälschlicherweise als die gesuchte Sachgruppe kategorisiert wurden.

Der Medianwert für die untersuchten Sachgruppen liegt bei 70%. Die Spitzengruppe bilden hier die Sachgruppen 340, 530 und 780 mit 86% bzw. 85%. Das Schlusslicht bilden die Sachgruppen 020 und 630 mit 44% und 43%.

Bemerkenswert ist auch das Ergebnis für die Sachgruppe 150, die in diesem Wert deutlich hinter den Recall zurückfällt, mithin zu einer signifikanten Zahl an falschen Vorhersagen führt. Wenn auch in geringerem Ausmaß, trifft dies auf Sachgruppe 610 ebenfalls zu.

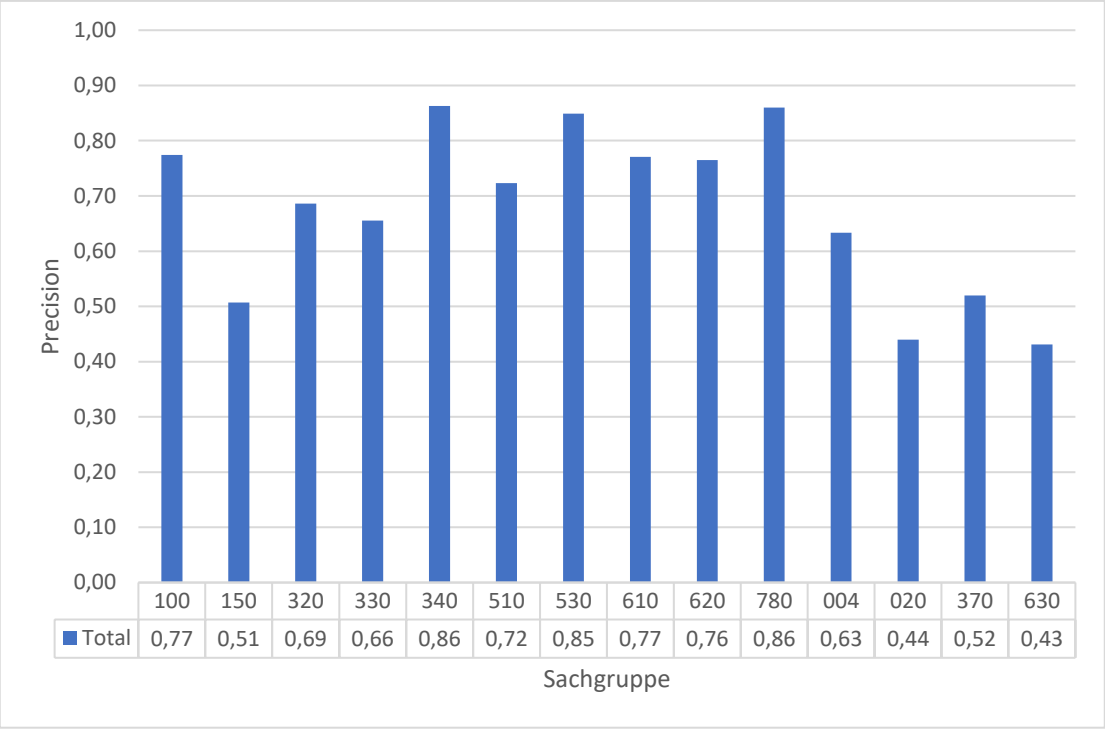


Abbildung 30:Precision pro Sachgruppe

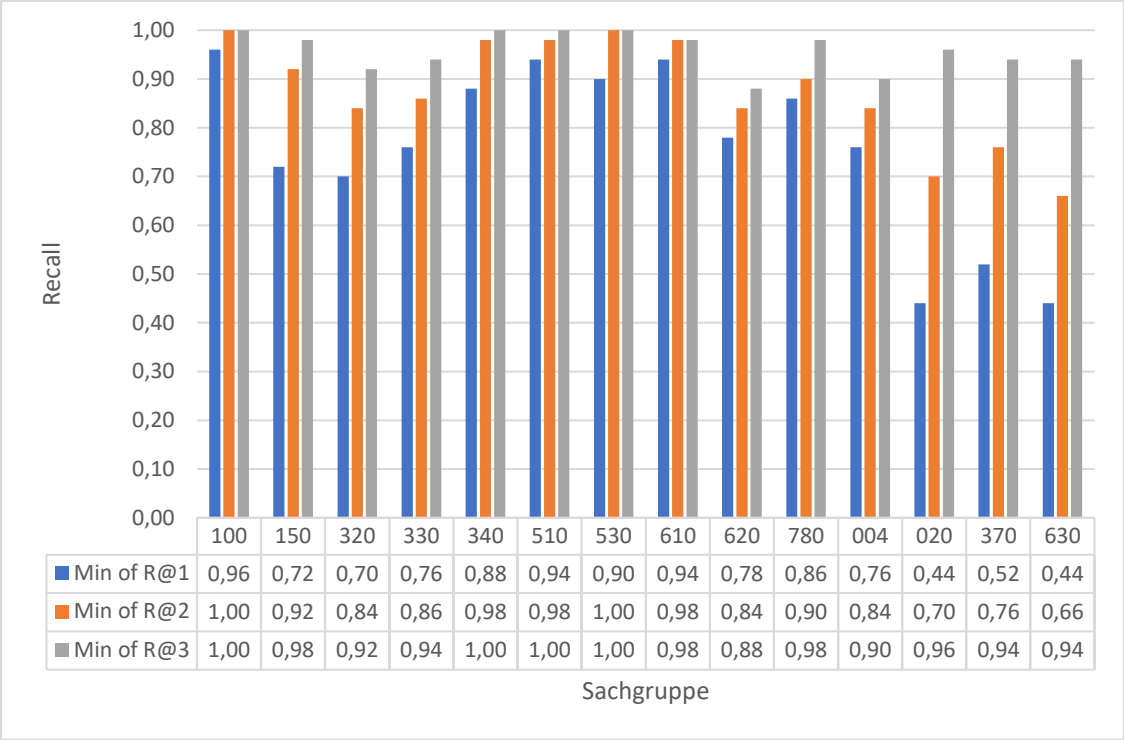
8.2 Untersuchung rangordnenden Kategorisierung

Ein wesentlicher Unterschied bei der Untersuchung der rangordnenden Kategorisierung besteht darin, dass hierbei mehrere Kategorien pro Dokument vorhergesagt werden. Wir treffen daher in diesem Szenario nicht mehr Aussagen über Dokumente, wie im klassischen Information Retrieval, sondern über die vorhergesagten Kategorien.

Dabei untersuchen wir in der Folge den Recall über die obersten drei Ränge, sowie den Wert für den mittleren Rang, auf dem das erwartete Ergebnis ausgegeben wird.

8.2.1 Recall über mehrere Ränge

Bei einer rangordnenden Kategorisierung gibt Recall eine Auskunft über die Vollständigkeit über mehrere Ränge. Die Auswertung ist hier für die obersten drei Ränge erfolgt.



	100	150	320	330	340	510	530	610	620	780	004	020	370	630
Min of R@1	0,96	0,72	0,70	0,76	0,88	0,94	0,90	0,94	0,78	0,86	0,76	0,44	0,52	0,44
Min of R@2	1,00	0,92	0,84	0,86	0,98	0,98	1,00	0,98	0,84	0,90	0,84	0,70	0,76	0,66
Min of R@3	1,00	0,98	0,92	0,94	1,00	1,00	1,00	0,98	0,88	0,98	0,90	0,96	0,94	0,94

Abbildung 31:Kumulierter Recall der Ränge 1-3

Wie zu erwarten, steigt der Recall mit zunehmendem Rang in allen Sachgruppen an. Der Median für Rang 1 von 77% steigt in Rang 2 bereits auf 88%. Schließlich werden im Median 97% aller Dokumente über die ersten drei Ränge hinweg richtig kategorisiert. Der Mittelwert über alle 700 Dokumente beträgt 96%.

Insbesondere die Sachgruppen 020, 370 und 630, die bei harter Kategorisierung unterdurchschnittlich gut vorhergesagt werden, entwickeln sich bis Rang 3 dynamisch. Vier Sachgruppen erreichen in Rang 3 einen Recall von 100%.

Hervorzuheben sind noch die Sachgruppen 620 und 004, die auch im dritten Rang einen Recall von lediglich 88% bzw. 90% aufweisen.

8.2.2 Mean Reciprocal Rank

MRR gibt Auskunft über die Genauigkeit der rangordnenden Klassifikation. Dabei werden alle Ränge ausgewertet, bis zu dem Rang, an dem die erwartete Kategorie ausgegeben wurde.

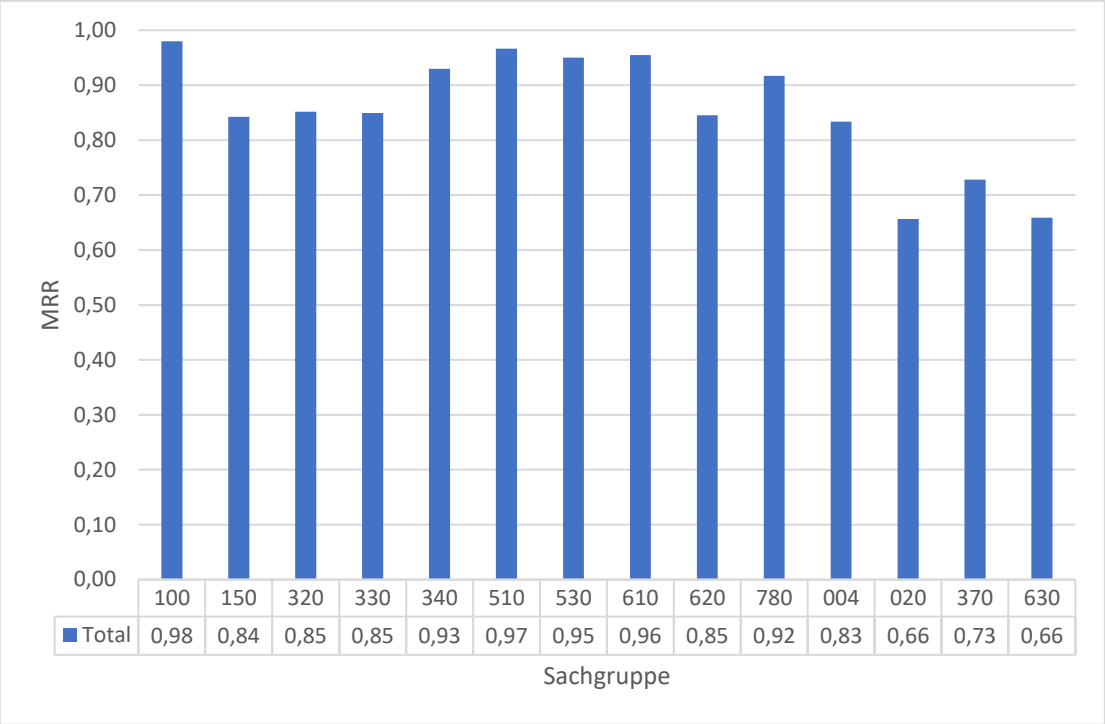


Abbildung 32: Mean reciprocal rank pro Sachgruppe

Die Kategorisierung ist also umso genauer, je näher der Wert sich 1 annähert.

Bis auf drei Sachgruppen überschreiten alle dabei die Marke von 0,8.

Ein Spitzenwert ergibt sich für die Sachgruppe 100, gefolgt von Sachgruppe 510 – Mathematik.

Die Sachgruppen 020, 370 und 630 bilden hier wiederum das Schlusslicht.

Wie sich aus der vorhergehenden Darstellung von MRR und Recall ergibt, werden 672 von 700 Dokumenten (96%) in den ersten drei Rängen korrekt vorhergesagt. Damit findet sich bei 28 Dokumenten die korrekte Sachgruppe erst auf einem der folgenden Ränge.

Gefunden wird die korrekte Kategorie allerdings in allen Fällen: Das Dokument mit der IDN 1008006637 erhält die erwartete Sachgruppe 004 erst auf Rang 11; weitere fünf Dokumente erhalten die korrekte Voraussage auf Rang 8 von 102. Diese sechs Dokumente bilden die Schlussgruppe; die verbleibenden 22 Dokumente weisen Rang 4-7 für das korrekte Resultat auf.

8.3 Diskussion der Ergebnisse

Die nach dem dreistufigen Verfahren vorgenommene Leistungsmessung für die Kategorisierung ist vor dem Hintergrund der zuvor analysierten Struktur des Thesaurus einzuordnen.

Zunächst ist festzustellen, dass Ergebnisse für eine lediglich prototypisch erstellte Maschine bereits beachtlich sind. Mehr oder weniger aus dem Stand werden für viele der Sachgruppen

F1-Maße erzielt, die andere Verfahren erst nach aufwändigen Optimierungen erreicht haben. Hierbei ist insbesondere die Schwelle von 70% zu nennen, die in der automatischen Vergabe von Sachgruppen für das KI-Verfahren in PETRUS genannt wurden (vgl. Kapitel 2.5).

Besonders hervorzuheben ist zudem, dass die Bereiche guter und weniger guter Leistungsfähigkeit der verwendeten Methode unmittelbar aus der Beschaffenheit der darunterliegenden Beschreibungslogik zu begründen sind. Sie sind damit nachvollziehbar und auch gezielt beeinflussbar, also auch zielgerichtet optimierbar. Bei einem ML-Verfahren verbleibt hier oft nur *trial and error*.

Fachgebiete, die ein spezifisches und umfassendes Vokabular aufweisen, erzielen aus dem Stand hohe Recall-Werte. Das ist insbesondere für Medizin, Mathematik und Physik gegeben. Die Sachgruppe 150 – Psychologie weist ebenfalls gute Recall-Werte auf, allerdings ist ihr Vokabular weder spezifisch, noch besonders groß. Gemessen daran, dass die Sachgruppe nur auf Rang 27 nach Anzahl der zugeordneten Schlagwörter (Rang 28, wenn man die erweiterte Sachgruppe 370 neu einordnet) steht, führt sie zu relativ vielen *false positives*. Dies ist, neben der Beschaffenheit des Vokabulars, mit der Überlappung zu anderen Sachgruppen begründbar. Siehe hierzu insbesondere Kapitel 5.4.2.3.

Die Sachgruppe 780 – Musik ist ein gutes Beispiel für einen zwar spezifischen aber nicht ganz umfassenden Wortschatz: Die Recall- Werte sind mit 2.864 zugeordneten Schlagwörtern ordentlich, aber für eine harte Kategorisierung noch verbesserbar. Hervorragend sind allerdings die Precision-Werte, da es hier kaum zu *false positives* kommt.

Beispiele mit einem unzureichenden und überlappenden Vokabular finden wir bezüglich der Sachgruppen 020 und 630. Für die Sachgruppe 370 trifft das selbst nach der Auffettung über Konkordanz zur GND-Systematik noch zu. Diese Sachgruppen weisen im harten Szenario niedrige Recall-Werte auf, die mit einem unzureichenden Vokabular erklärbar sind. Etwas problematisch erscheint auch die Sachgruppe 004 – Informatik, die ebenfalls nur unterdurchschnittliche Werte aufweist. Insbesondere Arbeiten der Wirtschaftsinformatik werden häufig als 330 – Wirtschaft kategorisiert. Hier bestehen offenbar zusätzlich Unklarheiten beim Zuschnitt des Fachs, siehe auch das Beispiel zur Wirtschaftsinformatik in Kapitel 9.1.5.

Die Sachgruppen 020 und 004 überlappen zudem recht häufig in ihren Ergebnissen, was sich dadurch erklären lässt, dass Sachgruppe Informatik 59 der nur 744 Schlagwörter mit den Bibliothekswissenschaften gemeinsam hat.

Eine besonders interessante Situation besteht hinsichtlich Sachgruppe 630: Diese ist, wie schon zuvor erläutert, extrem weit gefasst. Obwohl sie auf Rang 6 der zugeordneten Schlagwörter liegt, sind die Recall-Werte nicht besonders gut. Die schlechten Ergebnisse bei der Precision sind zwar durch Überlappung mit anderen Sachgruppen erklärbar, nicht aber die schlechten Recall-Werte.

Bei der Indexierung fiel allerdings auf, dass in 20% der Dokumente dieser Sachgruppe das relevanteste Schlagwort in einer fachlich verwandten Kategorie, aber nicht in Sachgruppe 630 verortet war. Das betraf insbesondere Begriffe aus der Medizin, den Biowissenschaften und der Zoologie. Besonders davon betroffen waren Dokumente aus der Veterinärmedizin, die der Thesaurus zurzeit nicht zuverlässig zuordnen kann. Es wäre also naheliegend diese Schlagwörter, insbesondere Individualbegriffe aus den Biowissenschaften und der Humanmedizin, ebenfalls in Sachgruppe 630 zu notieren, eine Erweiterung des Vokabulars mit Hilfe der in Kapitel 4.3.2 beschriebenen Methode zu versuchen, oder besonders mehrdeutige Begriffe in die *blocklist* zu geben.

Trotz der eher durchwachsenen Ergebnisse ist das Verhalten bezüglich Sachgruppe 370 als Erfolg zu werten. Das Fachgebiet weist die besondere Herausforderung auf, Didaktik für andere Fachgebiete zu leisten. Daher ist eine hohe dokumentenzentrische Überlappung des verwendeten Vokabulars mit anderen Fachgebieten letztlich unvermeidbar. Umso wichtiger ist eine vollständige Abbildung des spezifischen Vokabulars der Domäne.
Die ersten Versuche mit den ursprünglich 1.679 Begriffen führten zu unbrauchbaren Resultaten. Die Erweiterung über die Konkordanz führt immerhin dazu, dass 94% der relevanten Dokumente innerhalb der ersten drei Ränge korrekt klassifiziert wurden.

Die eher bescheidenen Precision-Werte für Psychologie und Medizin zeigen auch, dass ein großer Umfang des Vokabulars allein nicht immer zu einer besonders genauen Kategorisierung führt. Beide Sachgruppen liegen hinsichtlich des Umfangs auf Rang 1 und Rang 28 bezogen auf den gesamten Thesaurus, sowie Rang 1 und 7 bezogen auf die Größe des Vokabulars der ausgewerteten Sachgruppen. Ein *blocklisting* der besonders häufig zu *false positives* führenden Begriffe aus diesen Sachgruppen würde für die Leistung des Gesamtsystems einen effektiven Beitrag leisten. Dies ist insbesondere für die Sachgruppen mit starker Besetzung wichtig: Eine Verringerung der *false positives* für eine Sachgruppe verbessert zugleich die Leistung des Systems für die anderen davon betroffenen Sachgruppen.

Die Auswertungen des MRR zeigen, dass eine Ausgabe der obersten drei Ränge ausreicht, um in der überwiegenden Mehrzahl der Fälle und für alle ausgewerteten Sachgruppen aus den präsentierten Werten die korrekte Auswahl zu treffen. Auch in dieser Hinsicht ist das Verfahren also der ML-basierten Methode mindestens ebenbürtig (vgl. Kapitel 2.5).

9 Übertragung auf andere Szenarien, Sammlungen und Sprachräume

Nach der erfolgreichen Validierung des Verfahrens gegen einen Goldstandard, bietet es sich an, auszuprobieren, wie sich die Maschine in anderen, praxisrelevanten Szenarien verhält. Dazu wurde exemplarisch erprobt, wie die Kategorisierung sich gegen unbekannte, weil möglicherweise falsch erschlossene Dokumente verhält, und wie sich das Verfahren auf Dokumente unterschiedlicher fachlicher Tradition des selben Sprachraums übertragen lässt. Abschließend sind noch konzeptionelle Gedanken zu einer möglichen Übertragung in einen anderen Sprachraum aufgeführt.

9.1 Prüfung wahrscheinlich falsch kategorisierter Dokumente

Die Beobachtungen bei der Arbeit an der Zusammenstellung einer *ground truth* brachten es mit sich, über eine doch recht umfangreiche Anzahl an Dokumenten zu stolpern, in denen die vorhandenen Notationen nach einer Kurzautopsie fragwürdig schienen, oder offenkundig falsch waren.

Da keine eingehende Untersuchung des MARC-Datensatzes vorgenommen wurde, kann nicht ansatzweise beantwortet werden, wie es zu diesen Fehlern in den DNB Metadaten kam. Allerdings lag es nahe, in diesen Fällen den Thesaurus um eine zweite Meinung zu konsultieren, um zusätzliche Hinweise auf eine korrekte Erschließung zu erhalten.

Die in der Folge aufgeführten fünf Beispiele zeigen, dass eine thesaurusbasierte Kategorisierung einen wertvollen Beitrag zur Bereinigung eventuell fehlerhafter Erschließungsdaten liefern kann. Die Arbeiten sind jeweils unter ihrer DNB ID vermerkt und sind allesamt im Volltext zugänglich.

Die Auswertungen entsprechen grundsätzlich dem Vorgehen beim Goldstandard. Für eine genauere Einschätzung der vom Extraktor getroffenen Vorhersagen sind allerdings hier nicht nur die Ränge, sondern auch die Relevanzwerte der Sachgruppen angegeben. Das relevanteste Schlagwort ist mit Frequenz seines Auftretens zusätzlich angegeben.

9.1.1 IDN 1007482478

Die Arbeit „Untersuchungen zum Einfluss des Leukozyten-Inhibitions-Moduls auf die posthämorrhagische Inflammation und Gewebehypoxie im Tiermodell“ ist am 18.09.2021 bei der DNB unter Sachgruppe 510 – Mathematik verzeichnet.[112]

Titel und Metadaten deuten darauf hin, dass diese Zuweisung falsch ist. Die Arbeit trägt die DDC Kurznotation 617.2 (maschinell ermittelte DDC-Kurznotation). Diese zeigt nach der Hierarchie der DDC nach Sachgruppe 610 – Medizin.

Die Indexierung mit dem Thesaurus ergibt folgendes Resultat:

Sachgruppe 1	**Sachgruppe 2**	**Sachgruppe 3**	**Relevantes Schlagwort**
610 – Medizin: 0.26	570 – Bio.: 0.20	540 – Chemie: 0.06	Stunde[113] 77

Die im Katalog verzeichnete Sachgruppe findet sich auf Rang 15 mit einer Relevanz von 0.012. Wahrscheinlich handelt es sich bei der Angabe der Sachgruppe um einen Tippfehler. Für solche Fälle könnte die hier vorgestellte Methode ohne Weiteres zu einer automatisierten Konsistenzprüfung herangezogen werden.

9.1.2 IDN 1017972591

Die Arbeit „Persistente organische Spurenstoffe in Kompost und Rückständen der Biomassevergärung : Belastungsituation, Abbau und Bewertung“ ist am 18.09.2021 bei der DNB unter Sachgruppe 620 – Ingenieurwissenschaften verzeichnet.[114]

Titel und Metadaten deuten auch hier auf eine Falschzuweisung hin. Die im Katalog verzeichneten Schlagwörter „Polycyclische Aromaten ; Polychlorierte Biphenyle ; Persistenter organischer Schadstoff ; Weichmacher“ gehören zu den Fachgebieten Chemie bzw. Technische Chemie.

Die Indexierung mit dem Thesaurus ergibt folgendes Resultat:

[112] https://portal.dnb.de/opac/simpleSearch?query=idn%3D1007482478&cqlMode=true
[113] https://d-nb.info/gnd/4799320-0
[114] https://portal.dnb.de/opac/simpleSearch?query=idn%3D1017972591&cqlMode=true

Sachgruppe 1	Sachgruppe 2	Sachgruppe 3	Relevantes Schlagwort
540 – Chemie: 0.10	530 – Physik: 0.09	660 - Technische Chemie: 0.07	Temperatur[115]: 98

Die von der DNB vermerkte Sachgruppe 620 folgt an 6. Stelle mit einer Relevanz von 0.04. Aufgrund der ermittelten Relevanz und der übrigen Metadaten war diese Zuweisung eindeutig falsch. Die verbundene Institution, die Fakultät für Chemie der Uni Stuttgart, legt nahe, dass das Ergebnis des Klassifikators hingegen korrekt ist.

Ob das Dokument zu Sachgruppe 540 oder 660 gehört, wäre einer weiteren Entscheidungsfindung zu überlassen; hierzu liegen die Abstände der Ränge zu nahe beieinander.

9.1.3 IDN 107574377X

Die Arbeit „Altersabhängige Effekte aggressiv bewerteter Musik auf die autonome Arousal-Antwort“ ist am 18.09.2021 bei der DNB unter Sachgruppe 780 – Musik verzeichnet.[116]

Die Kurzautopsie legt nahe, dass es sich wohl eher um eine Dissertation in Medizin handelt. Untersucht werden altersabhängige Unterschiede bestimmter Stoffwechselparameter. Musik dient hier schlicht als Mittel zu deren Beeinflussung.

Die Indexierung mit dem Thesaurus ergibt folgendes Resultat:

Sachgruppe 1	Sachgruppe 2	Sachgruppe 3	Relevantes Schlagwort
610 – Medizin: 0.14	780 – Musik: 0.10	570 – Biol.: 0.09	Musik[117]: 98

Die Aussage des Klassifikators ist als sehr zuverlässig zu werten. Der Abstand der Relevanzwerte ist deutlich. Eine Überlappung der Vokabulare zwischen Medizin und Musik besteht kaum.

9.1.4 IDN 990346145

Die Arbeit „Philosophische Überlegungen zum Luftsicherheitsgesetz“ ist am 18.09.2021 bei der DNB unter Sachgruppe 340 – Recht verzeichnet.[118]

[115] https://d-nb.info/gnd/4059427-0
[116] https://portal.dnb.de/opac/simpleSearch?query=idn%3D107574377X&cqlMode=true
[117] https://d-nb.info/gnd/4040802-4
[118] https://portal.dnb.de/opac/simpleSearch?query=idn%3D990346145&cqlMode=true

Eine Kurzautopsie zeigt, dass das Dokument den Vermerk „Philosophie“ trägt und an der Philosophischen Fakultät der Universität Münster verfasst wurde. Die Arbeit befasst sich mit der Frage, ob man aus moralischer Sicht töten dürfe, um zu retten.

Die Indexierung mit dem Thesaurus ergibt folgendes Resultat:

Sachgruppe 1	Sachgruppe 2	Sachgruppe 3	Relevantes Schlagwort
100 – Philosophie: 0.132	340 – Recht: 0.126	300 – Soz.: 0.06	Gesetz[119] 72

Der enge Abstand der Relevanz beider Kategorien zeigt die hohe inhaltliche Überlappung beider Fachgebiete in der Arbeit. Gestützt wird die Rangfolge allerdings von spezifischen, relevanten Schlagwörtern der Sachgruppe 100: Leben (85), Entscheidung (33), Tod (29), Argument (18), Wert (16), Freiheit (15).
Meine Einschätzung ist, dass Fachreferent_Innen aus Jus und Philosophie diese Arbeit gerne beide ihren Benutzer_Innen zu Verfügung stellen möchten, die Hauptsachgruppe eher 100 lauten wird. Das intellektuelle Urteil ist für eine korrekte Kategorisierung hier jedenfalls gefragt; eine automatisierte Entscheidung ist in einem solchen Fall eher nicht sinnvoll.

9.1.5 IDN 1021499684

Die Arbeit „Branchenspezifische IT-Innovationssysteme : von der Analyse zur Intervention ; am Beispiel des IT-Innovationssystems für Krankenhäuser in Deutschland“ ist am 18.09.2021 bei der DNB unter der Sachgruppe 360 – Soziale Probleme , Sozialdienste, Versicherungen verzeichnet.[120]

Schlagwörter und Titel legen nahe, dass diese Zuweisung falsch ist. Es handelt sich um eine Arbeit aus der Wirtschaftsinformatik; die Anwendung der Lösung im Krankenhaus ist nur ein Nebenaspekt. (Siehe hierzu auch das Beispiel Göderts: Zugvögel, die Singvögel sind, Kapitel 1.5.)

Die Indexierung mit dem Thesaurus ergibt folgendes Resultat:

Sachgruppe 1	Sachgruppe 2	Sachgruppe 3	Relevantes Schlagwort
330 – Wirtschaft: 0.11	650 – Management: 0.10	300 – Soz.: 0.08	Innovation[121] 412

[119] https://d-nb.info/gnd/4020660-9
[120] https://portal.dnb.de/opac/simpleSearch?query=idn%3D1021499684&cqlMode=true
[121] https://d-nb.info/gnd/4027089-0

Das Ergebnis des Klassifikators kann die Fragestellung letztlich nicht beantworten. Zwar kann die von der DNB zugewiesene Sachgruppe mit einiger Sicherheit falsifiziert werden – sie findet sich auf Rang 6 mit einem Score von 0.05.
Das Beispiel zeigt allerdings deutlich, dass Arbeiten der Wirtschaftsinformatik sich anhand dieses Thesaurus nicht zuverlässig klassifizieren lassen.
Der Begriff Wirtschaftsinformatik fällt laut DDC-Notation unter die Sachgruppe 330.[122] Demnach wäre das Ergebnis des Klassifikators sogar korrekt.
Demgegenüber steht die Praxis vieler Universitäten, die Wirtschaftsinformatik der Informatik zuzuordnen. Dies deckt sich mit den unbefriedigenden Ergebnissen für Dokumente in diesem Themenbereich, die im DNB Katalog unter 004 eingeordnet waren. Damit der Thesaurus dies entsprechend wiedergibt, wäre das Vokabular von der GND Redaktion entsprechend nachzuarbeiten.

9.2 Indexierung österreichischer Hochschulschriften

Die Tatsache, dass es im Bereich der DNB eine Anzahl nach Dewey vorklassifizierter Dokumente gibt, in Österreich aber im Bereich der Hochschulschriften nach Basisklassifikation notiert wird, führt zu der Frage, ob das verwendete Verfahren auf österreichische Verhältnisse übertragbar ist.
Ein Fachgebiet, das über ein ausgeprägt eigenständiges österreichisches Fachvokabular verfügt, ist 340 – Recht. Daher wurde der Versuch unternommen, anhand exemplarischer, zufällig ausgewählter Dokumente zu prüfen, ob die Kategorisierung mit solchen Dokumenten ebenfalls funktioniert.

Die Ergebnisse sind in Tabelle 6 zusammengefasst, angegeben ist die Sachgruppe auf Rang 1, sowie der Rang der gesuchten Sachgruppe 340 für das Dokument vom Extraktor:

[122] http://d-nb.info/gnd/4112736-5

Quelle für Dokument	Sachgebiet	SG Rang 1	Ges. SG Rang
http://othes.univie.ac.at/51154/	Mietrecht	340	1
http://othes.univie.ac.at/64095/	Vertragsrecht	340	1
http://othes.univie.ac.at/19260/	Familienrecht	300	3
http://othes.univie.ac.at/6037/	Erbrecht	340	1
http://othes.univie.ac.at/60353/	Arbeitsrecht	330	2

Tabelle 6: Österreichische Hochschulschriften Jus

Offenbar unterscheiden sich die Ergebnisse für diese Auswahl nicht wesentlich von denen für juristische Hochschulschriften aus Deutschland; ein Blick in die relevanten Schlagwörter zeigt auch, dass das Zugangsvokabular für den spezifischen Sprachgebrauch in Österreich im Thesaurus ausreichend enthalten ist. Ein Beispiel hierfür ist der Kollektivvertrag[123], der weitläufig dem Tarifvertrag[124] in Deutschland entspricht. Die Unterschiede in der Rechtslage rechtfertigen allerdings eine eigenständige Ansetzung durchaus.

Dieser exemplarische Ausblick kann natürlich nicht belegen, dass alle Ergebnisse aus dem bei der DNB ermittelten Goldstandard nach Österreich übertragbar sind. Für wissenschaftliche Publikationen anderer Disziplinen unterscheidet sich das Fachvokabular allerdings eher kaum von dem in der Bundesrepublik. Jus stellt also eine Art „Stresstest“ dar. Damit sollte einer Nutzung des Verfahrens für österreichische Hochschulschriften in dieser Hinsicht wenig entgegenstehen.

9.3 Übertragung in andere Sprachräume

Es wäre naheliegend, den hier verfolgten Ansatz auf andere Sprachräume auszuweiten. Insbesondere wäre, da in vielen Disziplinen die Publikationstätigkeit zunehmend nur noch in Englisch erfolgt, eine Übertragung auf ein englischsprachiges Schlagwortsystem attraktiv.

Konzeptionell ist hierbei allerdings zu beachten, dass eine direkte Verwendung eines ausländischen Schlagwortsystems auch dessen kulturellen Hintergrund mit sich bringt. Dieser kann sich von den hiesigen Gepflogenheiten durchaus unterscheiden.

[123] https://d-nb.info/gnd/4351721-3
[124] http://d-nb.info/gnd/4117170-6

Gödert führt dazu aus:

> *„As indexing languages usually describe extracts of reality from a (culture-) specific point of view, adjusted to local needs and originally envisioned applications, entities of different indexing languages seldom represent exactly the same concept." (Gödert & Hubrich, 2014, p. 112 ff.)*

Im zitierten Abschnitt erklären die Autoren, dass die Entitäten die den Begriff Gesetzgebungsverfahren in den LCSH und der GND (SWD) beschreiben, durchaus unterschiedlich sind, und dass nicht zu erwarten ist, im Zielsystem immer ein äquivalentes Gegenstück zu finden. Für eine Notation nach DDC ist auf jeden Fall die lokale Ausgabe zu verwenden, um die Menge der Variablen gering zu halten.

Grundsätzlich kommen drei unterschiedliche Ansätze in Frage, die nach ihrer Praktikabilität und Anwendbarkeit auf die lokalen Gegebenheiten zu bewerten wären.

i.) Konstruktion auf Basis eines Schlagwortsystems in einer anderen Sprache unter Nutzung der lokalen DDC-Ausgabe

ii.) Übersetzung des lokalen Vokabulars in eine andere Sprache

iii.) Nutzung von Linked Data, insbesondere Nutzung von VIAF, MACS oder Wikidata (siehe Kapitel 2) zur Erweiterung des Zugangsvokbulars.

10 Schlussbemerkung

Die vorliegende Arbeit zeigt den hohen praktischen Wert eines thematisch geordneten Thesaurus für die automatisierte Sacherschließung. Hierbei konnte auf umfangreiche Vorarbeiten in den Daten zurückgegriffen werden, insbesondere die DDC-Notationen der Schlagwörter aus dem CrissCross Projekt. Die dabei erzeugten spezifischen semantischen Verknüpfungen ermöglichen den Aufbau einer hochpräzisen Beschreibungslogik.

Dennoch ist das gewählte Verfahren letztlich eher gleichordnend. Es erreicht seine Leistungsfähigkeit über die Aggregation einer sehr großen Menge an Schlagwörtern, der treffsicheren computerlinguistischen Bewertung der Relevanz für das Dokument und derer systematischen Verortung anhand des Regelwerks der DDC und der DDC-Sachgruppen. Die Vision einer vollständig syntaktisch-ordnenden Erschließungsmaschine, wie sie Gödert konzeptionell vorstellt, scheint hingegen vorerst noch Zukunftsmusik. Die GND bildet die hierfür erforderliche Semantik bislang nicht ab.

Die gewählte Vorgehensweise unterscheidet sich konzeptionell und methodisch wesentlich von der ansonsten weit verbreiteten Nutzung von ML-basierten Klassifikatoren. Die beiden herausragenden Vorzüge sind:

i.) Trainingsdaten und Trainingsprozesse sind nicht notwendig

ii.) Das Verfahren ist bis auf das einzelne Schlagwort hinab transparent.

Schon der hier gezeigte Prototyp erreicht aus dem Stand eine mit optimierten ML-Verfahren vergleichbare Leistungsfähigkeit.

Die in Kapitel 9 durchgearbeiteten Beispiele zeigen, dass das Verfahren auch in Zweifelsfällen hilfreiche Beiträge zu einer effizienten und konsistenten Erschließung liefert. Es ist auch auf Sammlungen außerhalb des Goldstandards übertragbar, und kann auf andere Sprachräume erweitert werden.

Bemerkenswert ist jedoch auch, dass manche Herausforderungen dieser Methodik sich gar nicht so sehr von den gewohnten Ansätzen unterscheiden: Wie im PETRUS-Programm ist es auch hier nicht aus dem Stand möglich, alle Fachgebiete einheitlich zu erschließen; insbesondere fehlt in denselben Fachgebieten das Vokabular, in denen auch bei den ML-Verfahren die Referenzdokumente knapp waren. Bei den Geisteswissenschaften muss das System also zurzeit eher passen. In den Disziplinen mit hohem Publikationsaufkommen (insbesondere Medizin) besteht hier wie dort viel, manchmal zu viel an Information. Hier ging das mit einem eher

überbordend großen Vokabular einher, wie die *false positives* aus der Fachgruppe 610 belegen.

Aus einer Pareto-Sicht auf das Problem ist das Verfahren aber schon jetzt in der Lage, etwa 90% des relevanten wissenschaftlichen Publikationsaufkommens in ausreichender Qualität zu bearbeiten. Darin besteht der wesentliche Erfolg. Die prinzipielle Leistungsfähigkeit für eine genaue und vor allem konsistente Ermittlung von DDC-Sachgruppen konnte erfolgreich belegt werden.

Über Konkordanzen, aber auch über statistische Verfahren wie Korpusanalyse oder Analyse der Verteilung der hilfreichen / schädlichen Schlagwörter, kann das System schnell optimiert werden. Da die Beschreibungslogik aus GND-SH programmatisch erzeugt wird, ist sie leicht pfleg- und erweiterbar. Weitere Optimierungsmöglichkeiten, wie die Verwendung von Referenzkorpora, oder die Kombination mit anderen Informationsquellen (z.B. der Fakultät an der eine Arbeit entstanden ist), sind unkompliziert zu implementieren.

Insbesondere die hier angetroffenen Herausforderungen bei der Kategorisierung von Arbeiten aus der Informatik zeigen einerseits die Stärken des hier angewendeten Verfahrens und andererseits wie wichtig es ist, dass Regelwerk und Erschließungspraxis übereinstimmen.
Für die Beschlagwortung von Arbeiten aus der Wirtschaftsinformatik weist der Thesaurus in Sachgruppe 004 kein Vokabular auf, da die einschlägigen Begriffe laut DDC eben Unterbegriffe der Notation 330 sind. Diese Arbeiten werden dennoch häufig in Sachgruppe 004 verzeichnet, was offensichtlich in einem Spannungsverhältnis zum Regelwerk steht.
Einen solchen Widerspruch kann die Maschine zwar evidenzbasiert feststellen, sie kann ihn aber nicht anders lösen, als es dem Regelwerk entspricht auf dem sie aufgebaut wurde – also den nach den Regeln der DDC zugeordneten Schlagwörtern. Diese Konsistenz kann ein ML-basierter Klassifikator nicht garantieren: Er wird Dokumente aus der Wirtschaftsinformatik nach 004 klassifizieren, wenn man ihn vorher entsprechend mit Referenzdokumenten dazu trainiert hat. Ein thesaurusbasierter Extraktor wird dies erst dann tun, wenn auch das Regelwerk entsprechend angepasst ist – im einfachsten Fall also die Zuweisung der Notationen der Wirtschaftsinformatik zur Sachgruppe 004.[125]

[125] Ein entsprechender Präzedenzfall besteht übrigens für die Sachgruppe 100 (Alex, 2014).

Ein weiterer Vorzug ist die grundsätzlich rangordnende Klassifikation des Verfahrens, da hierdurch die Konfidenzwerte der Vorhersagen unmittelbar ablesbar sind und so die Ausgabe der relevanten Vorschläge an verschiedene Workflows sehr einfach angepasst werden kann. Somit eignet sich das Verfahren insbesondere auch in einer semiautomatischen Erschließung, beispielsweise auch als zusätzliches Hilfsmittel in einem digitalen Assistenten oder auch zur Unterstützung einer Erschließung durch die Autoren selber.

Bedauerlich ist, dass der Umfang der mit der DDC verknüpften Schlagwörter nur etwa 60% des Gesamtvokabulars der GND ausmacht. Wenigstens für die spezifischen Entitäten jeder Fachgruppe wäre es sehr wünschenswert, die fehlenden DDC-Notationen laufend nachzupflegen. Auch ist es für eine thesaurusbasierte Erschließung eher hilfreich, eine große Menge an präkombinierten Schlagwörtern im Wortschatz zu haben. Die hochrelevanten, spezifischen Schlagwörter waren fast ausnahmslos Komposita, während die postkoordierenden Elemente anfällig für Ambiguität oder im Thesaurus sogar zuweilen polysem angesetzt sind, wie die hier gezeigten Beispiele zeigen konnten. Die weitere terminologische Arbeit wird die Ergebnisse zweifelsohne verbessern.

Erneut hat sich in der Durchführung dieser Arbeit gezeigt, wie wichtig Sacherschließung für die Nutzungsfreundlichkeit von Bibliothekskatalogen ist. Die Zusammenstellung des Goldstandards verdeutlicht erneut, dass für die deutschsprachigen Hochschulschriften in dieser Hinsicht noch Potential nach oben besteht, auch wenn die Herausforderung zuweilen auch in einer korrekten Präsentation der vorhandenen Daten in der Suchmaschine zu bestehen scheint, wie die Arbeit am Goldstandard ergeben hat.

Hierzu können konsistente, vollständige und präzise Notationen einen wesentlichen Beitrag leisten. Sachgruppen geben hierbei einen groben Überblick über die Einordnung eines Dokumentes. Die hier vorgestellte Methodik kann dies in besonders guter Weise leisten.

Darüber hinaus wäre prinzipiell dieses Verfahren auch komplementär für die Ermittlung der DDC-Kurznotationen geeignet. Hierzu wäre lediglich die Konstruktion des Thesaurus anzupassen.

Anhang

Anhang 1: qSKOS Report

Zirkelbezüge:

```
*  Detailed coverage of each Quality Issue:

--- Cyclic Hierarchical Relations
Description: Finds concepts that are hierarchically related to each other
Detailed information: https://github.com/cmader/qSKOS/wiki/Quality-Issues#cyc-
lic-hierarchical-relations
count: 7
Cycle 1, size: 2
Cycle 2, size: 2
Cycle 3, size: 4
Cycle 4, size: 3
Cycle 5, size: 1
Cycle 6, size: 1
Cycle 7, size: 1
Cycle 1, size: 2
       https://d-nb.info/gnd/1212388364
       https://d-nb.info/gnd/4381441-4
Cycle 2, size: 2
       https://d-nb.info/gnd/4322766-1
       https://d-nb.info/gnd/4140228-5
Cycle 3, size: 4
       https://d-nb.info/gnd/4017396-3
       http://zbw.eu/stw/descriptor/19453-3
       https://d-nb.info/gnd/4017398-7
       https://d-nb.info/gnd/4129966-8
Cycle 4, size: 3
       http://zbw.eu/stw/descriptor/18301-3
       https://d-nb.info/gnd/4159290-6
       https://d-nb.info/gnd/4224520-5
Cycle 5, size: 1
       https://d-nb.info/gnd/4200437-8
Cycle 6, size: 1
       https://d-nb.info/gnd/4200101-8
Cycle 7, size: 1
       https://d-nb.info/gnd/4100125-4
```

Anhang 2: SPARQL Codebeispiel

Konstruktion für Sachgruppe 891.8 - Slawische Literatur:

```
PREFIX gndo: <https://d-nb.info/standards/elementset/gnd#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
INSERT {
  ?concept a skos:Concept ;

    # labels
    skos:prefLabel ?prefLabel;
    skos:altLabel ?altLabel;

    # link to broader
    skos:broader <https://dewey.info/class/891.8/> .

    # missing reverse relation
    <https://dewey.info/class/891.8/> skos:narrower ?concept .

   }
WHERE {
    ?concept a ?type;
    gndo:preferredNameForTheSubjectHeading ?prefPlain .
    ?concept ?p ?o;

#891.8 - Slawische Literatur
    FILTER (?d >= 891.8 && ?d < 891.9 )

    BIND(xsd:decimal(strbefore(strafter(str(?o),"http://dewey.info/class/"),
"/"))as ?d)
    BIND (strlang(?prefPlain, "de") as ?prefLabel)

OPTIONAL {
    ?concept gndo:variantNameForTheSubjectHeading ?altPlain
     BIND(strlang(?altPlain, "de") as ?altLabel)
  }

  #  values
    VALUES (?type)

{(gndo:SubjectHeadingSensoStricto)(gndo:NomenclatureInBiologyOrChemistry)(gndo
:SubjectHeading)(gndo:Language)
(gndo:SoftwareProduct)(gndo:ProductNameOrBrandName) (gndo:EthnographicName)
(gndo:HistoricSingleEventOrEra) (gndo:MeansOfTransportWithIndividualName)
(gndo:GroupOfPersons) (gndo:CorporateBody) (gndo:ProjectOrProgram) (gndo:Work)
(gndo:BuildingOrMemorial) (gndo:ExtraterrestrialTerritory)
(gndo:MusicalCorporateBody) (gndo:ConferenceOrEvent) (gndo:WayBorderOrLine)
(gndo:LiteraryOrLegendaryCharacter) (gndo:Collection)
(gndo:PlaceOrGeographicName) (gndo:SeriesOfConferenceOrEvent)
(gndo:TerritorialCorporateBodyOrAdministrativeUnit) (gndo:FictiveTerm)
(gndo:NaturalGeographicUnit) (gndo:Manuscript) (gndo:OrganOfCorporateBody)
(gndo:DifferentiatedPerson) (gndo:ReligiousTerritory)
(gndo:NameOfSmallGeographicUnitLyingWithinAnotherGeographicUnit) (gndo:Gods)
(gndo:ReligiousCorporateBody) }

    VALUES (?p)
    { (gndo:relatedDdcWithDegreeOfDeterminacy2)
(gndo:relatedDdcWithDegreeOfDeterminacy3)(gndo:relatedDdcWithDegreeOfDetermina
cy4)}
}
```

Abbildungsverzeichnis

Tabellenverzeichnis

Bibliografie

Alex, H., 2014. *DDC-Sachgruppen der deutschsprachigen Nationalbibliografien : Deutsche Nationalbibliografie, Das Schweizer Buch, Österreichische Bibliografie : Leitfaden zu ihrer Vergabe,* Frankfurt: Deutsche Nationalbibliothek.

Beyer, C. & Trunk, D., 2011. Automatische Verfahren für die Formalerschließung im Projekt PETRUS. *Dialog mit Bibliotheken,* 23(2), pp. 5 - 10.

Busse, F., 2019. *DDC-Kurznotationen : Entwicklung & Maschinelle Klassifikation.* Stuttgart, s.n.

Cappelaro, C., 2003. *Die Schlagwortnormdatei - ein zentrales Hilfsmittel der verbalen Sacherschließung.* [Online]
Available at: http://www.ib.hu-berlin.de/texte/hausarbeiten/capellaro/swd-capellaro.htm
[Zugriff am 20 07 2021].

Ceynowa, K., 2017. *In Frankfurt lesen jetzt zuerst Maschinen.* [Online]
Available at: https://www.faz.net/aktuell/feuilleton/buecher/maschinen-lesen-buecher-deutsche-nationalbibliothek-setzt-auf-technik-15128954-p2.html
[Zugriff am 15 04 2021].

Chan, L., 2000. *Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources.* [Online]
Available at: https://www.loc.gov/catdir/bibcontrol/chan_paper.html
[Zugriff am 09 05 2021].

Deutsche Nationalbibliothek, 2011. *Leitfaden GND Systematik.* [Online]
Available at: http://d-nb.info/1018626042
[Zugriff am 15 05 2021].

Gödert, W., 1990. Zur semantischen Struktur der Schlagwortnormdatei (SWD). Ein Beitrag zur Problematik des induktiven Aufbaus kontrollierten Vokabulars. *Libri (København),* Band 40 (3), pp. 228 - 241.

Gödert, W., 2014. Ein Ontologie-basiertes Modell für Indexierung und Retrieval. *Information. Wissenschaft & Praxis,* 65(2), pp. 83 -98.

Gödert, W. & Hubrich, J. a. N. M., 2014. *Semantic Knowledge Representation for Information Retrieval.* Berlin/Boston: Walter de Gruyter GmbH.

Gödert, W., Lepsky, K. & Nagelschmidt, M., 2011. *Informationserschließung und Automatisches Indexieren : Ein Lehr- und Arbeitsbuch.* Berlin / Heidelberg: Springer.

Golub, K., 2016. Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries.. *International Information & Library Review,* 48(3), pp. 204 - 210.

Golub, K., 2019. Automatic subject indexing of text. In: B. H. a. C. Gnoli, Hrsg. *Encyclopedia of Knowledge Organization.* s.l.:ISKO.

Hubrich, J., 2008. Criss Cross: SWD-DDC-Mapping. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare,* 61(3), pp. 50 - 58.

ISO, 2011. *ISO 25964: Thesauri and interoperability with other vocabularies: Part 1: Thesauri for information retrieval.,* Geneva: International Organization for Standardization.

ISO, 2013. *ISO 25964-2:2013 Information and documentation: Thesauri and interoperability with other vocabularies, Part 2: Interoperability with other vocabularies,* s.l.: International Organization for Standardization.

Mitchell, J. & Panzer, M., 2013. Dewey linked data: making connections with old friends and new acquaintances. *Jlis.it,* 4(1), pp. 177 - 199.

Mödden, E. & Tomanek, K., 2012. Maschinelle Sachgruppenvergabe für Netzpublikationen. *Dialog mit Bibliotheken,* Band 1, pp. 17 - 24.

RSWK, 2017. *Regeln für die Schlagwortkatalogisierung (RSWK). 4., vollständig überarbeitete Auflage 2017.* [Online]
Available at: http://d-nb.info/1126513032/34
[Zugriff am 01 06 2021].

Scheven, E., 2017. Die neue Thesaurusnorm ISO 25964 und die GND. In: *Theorie, Semantik und Organisation von Wissen.* s.l.:Ergon Verlag, pp. 289 - 305.

Schöning-Walter, C., 2010. 2010. *Dialog mit Bibliotheken,* 22(1), pp. 15-19.

Schöning-Walter, C., 2011. *Einführung in das Projekt PETRUS.* [Online]
Available at: https://files.dnb.de/pdf/petrus/einfuehrung_petrus_schoening-walter.pdf
[Zugriff am 14 03 2021].

Sommer, M., 2012. *Automatische Generierung von DDC-Notationen für Hochschulveröffentlichungen.* [Online]
Available at: https://dx.doi.org/10.25968/opus-326
[Zugriff am 25 06 2021].

Souminen, O., 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries,* 29(1), pp. 1 - 25.

Suominen, O. & Mader, C., 2013. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics,* 3(1), pp. 47 - 73.

Uhlmann, S., 2013. Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND). *Dialog mit Bibliotheken,* 25(2), pp. 26 - 36.

Uhlmann, S., 2020. *Fachtagung Netzwerk maschinelle Verfahren in der Erschliessung.* [Online]
Available at: https://wiki.dnb.de/download/attachments/181751388/2-3_Automatische-Vergabe-von-GND-Schlagw%C3%B6rtern_Uhlmann_2020-12-03_final.pdf?version=1&modificationDate=1607352872000&api=v2
[Zugriff am 30 08 2021].

Voß, J. et al., 2014. *Normdaten in Wikidata.* [Online]
Available at: https://hshdb.github.io/normdaten-in-wikidata/normdaten-in-wikidata.html
[Zugriff am 22 07 2021].

Wiesenmüller, H., 2018. *Maschinelle Indexierung am Beispiel der DNB - Analyse und Entwicklungsmöglichkeiten}.* Berlin, s.n.