Beatrice S. Weers

Automatisierte Inhaltserschließung an der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften

BEATRICE S. WEERS

Automatisierte Inhaltserschließung an der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften

HTWK Hochschule für Technik, Wirtschaft und Kultur Leipzig

Über die Autorin:
Beatrice Weers studierte Bibliotheks- und Informationswissenschaft an der Hochschule für Technik, Wirtschaft und Kultur Leipzig und schloss dort sowohl ihren Bachelor- als auch Masterstudiengang erfolgreich ab. Während ihres Studiums war sie als wissenschaftliche Hilfskraft an der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften tätig. Ihr Forschungsschwerpunkt liegt seit mehreren Jahren auf der maschinell unterstützten Inhaltserschließung sowie auf Change-Management-Prozessen in Teams. In ihrer Arbeit verbindet sie technisches Verständnis mit einem klaren Blick für organisatorische Weiterentwicklung.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie. Detaillierte bibliografische Daten sind im Internet unter http://dnb.de abrufbar.

Der Text dieses Werks ist unter der Creative-Commons-Lizenz CC BY 4.0 DE veröffentlicht. Den Vertragstext der Lizenz finden Sie unter https://creativecommons.org/licenses/by/4.0/. Die Abbildungen sind von dieser Lizenz ausgenommen, hier liegt das Urheberrecht beim jeweiligen Rechteinhaber.

Die Online-Version dieser Publikation ist abrufbar unter
http://doi.org/10.33968/9783966270786-00

© 2025 Beatrice S. Weers

Herausgeber
Projektgruppe Open-Access-Hochschulverlag
Hochschule für Technik, Wirtschaft und Kultur Leipzig
Eichendorffstraße 14
04277 Leipzig, Deutschland

ISBN (Hardcover): 978-3-96627-079-3
ISBN (PDF): 978-3-96627-078-6
ISBN (Epub): 978-3-96627-080-9

Kurzbeschreibung

Die vorliegende Bachelorarbeit befasst sich mit dem Thema der automatisierten Inhaltserschließung an der Bibliothek des Max Planck Instituts für Mathematik in den Naturwissenschaften in Leipzig. Ziel ist es, zu ermitteln, wie gut das eigens von der Bibliothek entwickelte Programm für die automatisierte Inhaltserschließung im Vergleich zur intellektuellen Erschließung funktioniert.

Dazu wurden einige theoretische Grundlagen beleuchtet und andere aktuelle Projekte in Deutschland vorgestellt.

Durch eine Untersuchung, bei der 97 Publikationen sowohl vom Programm als auch von drei Testpersonen mit Notationen versehen wurden, ergab sich eine 70-prozentige Übereinstimmung des Programms mit jeweils mindestens einer der Testpersonen.

Zusätzlich wurden Verbesserungsmöglichkeiten für das Programm aufgezeigt.

Inhaltsverzeichnis

1 Einleitung ... **11**

2 Fachterminologie ... **15**

3 Aktueller Forschungsstand in Deutschland ... **17**
3.1 Digitaler Assistent DA-3 ... 17
3.2 Annif ... 19
3.3 DNB – Deutsche Nationalbibliothek ... 21
3.4 ZBW – Leibniz-Informationszentrum Wirtschaft ... 29

4 Vorstellung der Bibliothek des MPI MIS ... **33**
4.1 Nutzung der Library of Congress Classification ... 34
4.2 Aufbau der Library of Congress Classification ... 35

5 Programmbeschreibung ... **37**
5.1 Fachterminologie ... 37
5.1.1 Algorithmus ... 37
5.1.2 Java ... 37
5.2 Datenbank des MPI MIS ... 38
5.3 Funktionsweise des Programms ... 38

6 Untersuchung ... **43**
6.1 Testpersonen ... 43
6.2 Untersuchungsgegenstand ... 44
6.3 Durchführung ... 45
6.4 Auswertung der Untersuchung ... 45

7 Optimierungsmöglichkeiten ... **53**
7.1 Vergrößerung des Vokabulars ... 53
7.2 Einbindung weiterer Metadaten ... 53
7.3 Normalisierung der Termini ... 54
7.4 Verwendung von Annif ... 55

8 Fazit ... **57**

Literatur- und Quellenverzeichnis ... 59
Anhang A Untersuchungsergebnisse ... 64
Anhang B Ausgewählte Untersuchungsergebnisse ... 71
Anhang C Quasi-Code des Programms ... 74

Abbildungsverzeichnis

Abb. 1 Visualisierung des Programmablaufs 42
Abb. 2 Übereinstimmung mit dem Programm in Prozent 46
Abb. 3 Anzahl der Übereinstimmungen mit dem Programm in Prozent 47
Abb. 4 Übereinstimmungen der Personen mit dem Programm in Prozent 48
Abb. 5 Absolute Übereinstimmungen der Personen miteinander und mit dem Programm 49
Abb. 6 Absolute Häufigkeit des Vorkommens von Titeln mit der Anzahl an Stichworten 50

Abkürzungsverzeichnis

BK	Basis Klassifikation
BnF	Bibliothèque nationale de France
CBS	Zentrales Katalogisierungssystem
DA-3	Digitaler Assistent 3
DNB	Deutsche Nationalbibliothek
Ema	Erschließungsmaschine
GND	Gemeinsame Normdatei
IBS BW	Integriertes Bibliothekssystem Baden-Württemberg
IDN	Identifikationsnummer
JVM	Java Virtual Machine
KI	Künstliche Intelligenz
LCC	Library of Congress Classification
LoC	Library of Congress
MARC	Machine Readable Cataloging
MLLM	Mauilike Lexical Matching
MPI MIS	Max-Planck-Institut für Mathematik in den Naturwissenschaften
MSC	Mathematical Subject Classification
OBV	Österreichischer Bibliotheksverbund
OCR	Optical Character Recognition
PACS	Physics and Astronomy Classification Scheme
RVK	Regensburger Verbundklassifikation
SBB	Staatsbibliothek zu Berlin
STW	Standard-Thesaurus Wirtschaft
SVM	Support Vector Machine
TOC	Table of Contents
UB	Universitätsbibliothek
WORA	Write once, run anywhere
ZBW	Zentralbibliothek für Wirtschaftswissenschaften

1 Einleitung

In einer Zeit, in der die Automatisierung von Arbeitsprozessen zunehmend an Bedeutung gewinnt, steht insbesondere die Frage im Mittelpunkt, wie Arbeitsabläufe nicht nur unterstützt, sondern vollständig automatisiert werden können. Auch im Bibliothekswesen spiegelt sich dieser Trend wider, unter anderem rückt dabei die Frage der automatisierten Inhaltserschließung in den Fokus. Diese Entwicklung zeigt sich auch in der aktuellen wissenschaftlichen Literatur, die ein wachsendes Interesse an der Automatisierung und der Rolle künstlicher Intelligenz (KI) aufzeigt. Es finden sich viele Projekte und Ideen, die in Zusammenarbeit mit Software und Firmen explizit auf automatisierte Inhaltserschließung ausgerichtet sind. Insbesondere die Bewegung hin zur vollständigen Automatisierung, ohne dabei Zwischenschritte der Halbautomatisierung oder computergestützten Prozessen zu berücksichtigen, stellt eine interessante Tendenz dar.

Die Relevanz der Thematik wird durch die intensive Auseinandersetzung einiger führender Institutionen mit Projekten im Gebiet der automatischen Sacherschließung unterstrichen, wobei insbesondere die Deutsche Nationalbibliothek (DNB) und die Zentralbibliothek für Wirtschaftswissenschaften (ZBW) hervorstechen. Diese Einrichtungen haben durch ihre Publikationen und Forschungsbeiträge wertvolle Einblicke in die technischen Abläufe und Entwicklungen ihrer Projekte ermöglicht. In diesem Zusammenhang sind vor allem die Software Annif und der Digitale Assistent DA-3 von besonderem Interesse. Im Rahmen dieser Arbeit wurde eine umfangreiche Recherche durchgeführt, die gängige Kataloge, Fachdatenbanken sowie Vortrags- und Tagungsverzeichnisse umfasste und am 4. März 2024 abgeschlossen wurde.

Im Zentrum der Arbeit steht die Evaluierung eines eigens entwickelten Programms, welches konzipiert wurde, um die Effizienz und Genauigkeit der Sacherschließung innerhalb der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften (MPI MIS) zu steigern. Grund für die Themenwahl ist, dass ich selbst dort seit drei Jahren als studentische Hilfskraft tätig bin. In dieser Zeit habe ich mich intensiv in die dortigen Abläufe, unter anderem auch der automatischen Inhaltserschließung einarbeiten können. Einige Bestandteile meiner Arbeit basieren daher auf abstrahiertem Wissen, welches ich im Rahmen meiner Beschäftigung erworben habe.

Das primäre Untersuchungsziel ist, zu analysieren, inwiefern die algorithmische Verarbeitung von Stichwörtern im Sachtitel der Monografien mit der traditionellen, manuellen Inhaltserschließung korreliert, ohne dabei auf bereits existierende von Verlagen vergebene Schlagwörter zurückzugreifen. Dieser Ansatz reflektiert die innovativen Bestrebungen der Bibliothek, sich von der herkömmlichen Praxis der Übernahme von Schlagwörtern, die bereits eine gewisse inhaltliche Erschließung implizieren, zu distanzieren. Die von Verlagen vergebenen Schlagwörter sind oft zu

unspezifisch und thematisch generisch gehalten, um als Orientierung in der Klassifikationsvergabe zu helfen. Der Fokus liegt nun auf einer originären Erschließung basierend auf den Sachtiteln der Werke, was eine unmittelbare und unverfälschte Analyse des Inhalts verspricht.

Die Bedeutung der Prüfung ergibt sich aus der zunehmenden Herausforderung, qualifiziertes Fachpersonal zu rekrutieren – klassischerweise BibliothekarInnen, welche über einen Abschluss im Fachgebiet der jeweiligen Bibliothek verfügen.[1] Traditionell obliegt es dieser spezialisierten Gruppe, eine präzise und fundierte Inhaltserschließung der Fachpublikationen sicherzustellen.[2] Die Realität in vielen Spezialbibliotheken, insbesondere in One-Person-Libraries, zeigt jedoch, dass der zeitliche und personelle Rahmen oft nicht ausreicht, um eine derart detaillierte und tiefgehende inhaltliche Analyse jedes einzelnen Werks zu gewährleisten.[3] Vor diesem Hintergrund erscheint die Evaluation des Programms, das eine grobe, jedoch vor allem präzise und konsistente Einordnung der Werke in die entsprechenden Sachgruppen vornehmen soll, als ein logischer und innovativer Schritt.

Die Untersuchung der Forschungsfrage wurde in mehreren Teilschritten durchgeführt. Zusätzlich zur Überprüfung, wie oft die maschinell und die intellektuell vergebenen Notationen insgesamt identisch sind, wurde ausgewertet, mit welcher Person die größte Übereinstimmung vorliegt. Hierfür wurden im Anschluss mögliche Gründe betrachtet. Konkret wurde weiterhin untersucht, aus welchen Gründen es bei bestimmten Titeln nur wenige oder keine Übereinstimmungen gibt und ob die Stichwortmenge des Titels einen ausschlaggebenden Einfluss darauf besitzt.

Die Ergebnisse sind vor allem für das MPI MIS von Bedeutung, da das entsprechende Programm nur dort eingesetzt wird. Die Verbindung von bibliothekarischem Fachwissen mit maschineller Unterstützung zielt darauf ab, die traditionelle, von Personen durchgeführte Klassifizierung auf ein automatisiertes Programm, welches mit einem speziellen Algorithmus läuft, zu übertragen. Das Ergebnis dieser Bemühungen ist der hier vorgestellte Prototyp zur automatisierten Inhaltserschließung, welcher nun zur Revision bereit ist. Im Zuge dessen wurde eine statistische Untersuchung durchgeführt, deren Ziel es war, herauszufinden, ob der Ansatz, der vorgestellten automatisierten Inhaltserschließung, weiterverfolgt werden sollte. Die systematische Analyse und Bewertung des Programms im direkten Vergleich mit der manuellen Inhaltserschließung durch Fachpersonal mit unterschiedlichem Vorwissen ist hier der Schlüsselfaktor.

Die Erkenntnisse aus anderen Institutionen, die ihre Arbeitsprozesse ebenfalls maschinell unterstützen wollen und die Ergebnisse der durchgeführten Untersuchung könnten dabei wertvolle Hinweise auf Verbesserungsmöglichkeiten für das MPI MIS bieten.

---

1 Vgl. Tréfás, David: Das Fachreferat: vom Universalgelehrten zur Schwarmintelligenz, 2018, S. 865 f.

2 Vgl. Ebenda, S. 865 f.

3 Diesen Eindruck konnte ich durch meine Teilnahme an Bibliothekstagungen der MPG gewinnen, auf welchen ich mich mit vielen in OPLs tätigen KollegInnen unterhalten habe.

Zuerst wird dafür die allgemeine Situation der Bibliothek dargelegt, einschließlich einer Erläuterung der verwendeten Klassifikationssysteme. Anschließend erfolgt eine Erklärung in die relevanten Begrifflichkeiten und eine kurze Einführung zu Annif sowie DA-3. Darüber hinaus wird ein detaillierter Blick auf die aktuellen Entwicklungen in der Deutschen Nationalbibliothek und der Zentralbibliothek für Wirtschaftswissenschaften bezüglich ihrer Projekte zur automatischen Inhaltserschließung geworfen. Es folgen eine genaue Beschreibung des am MPI MIS entwickelten Programms und möglichen Ansätzen zur Optimierung. Abschließend werden die Methodik der statistischen Untersuchung erläutert, die Ergebnisse ausgewertet und ein Fazit gezogen.

2 Fachterminologie

Indexierung spielt eine zentrale Rolle in der Organisation und Erschließung von Informationen. Sie umfasst die Vergabe von inhaltswiderspiegelnden Ausdrücken, die unter anderem das Auffinden von Dokumenten bei Stichwortsuchen erleichtern. Diese Ausdrücke, auch Indexterme genannt, können Merkmale wie Notationen, Schlagwörter oder direkt aus dem Text extrahierte Stichwörter umfassen. Die Praxis der Indexierung lässt sich in drei Hauptkategorien unterteilen: intellektuelle, computerunterstützte und automatische Indexierung.[4]

Bei der *intellektuellen Indexierung*, einem Teilbereich der klassischen Inhaltserschließung, handelt es sich um eine detaillierte Inhaltsanalyse von Dokumenten. Auf der Grundlage dieser Analyse werden Indexterme gebildet, die eine begriffliche Abbildung des Inhalts darstellen. Diese Form der Indexierung zeichnet sich durch eine hohe Qualität aus, da sie ein tiefes Verständnis des Inhalts voraussetzt. Als nachteilig kann angesehen werden, dass das qualitativ hochwertige Ergebnis nur durch den „Einsatz menschlicher Arbeitskraft“[5] hervorgebracht werden kann.[6]

Die *automatisierte Indexierung* hingegen basiert auf der Extraktion von Indextermen direkt aus dem Textdokument ohne menschliches Zutun. Moderne Technologien und Algorithmen, die von linguistischen bis hin zu statistischen Verfahren reichen, werden eingesetzt, um relevante Begriffe zu identifizieren und zu verknüpfen. Die Automatisierung der Indexierung hat sich aufgrund der fortschreitenden Entwicklungen in der Informations- und Kommunikationstechnologie erheblich weiterentwickelt und wird ständig optimiert, um die Genauigkeit und Relevanz der Indizierung zu verbessern.[7]

Die *computerunterstützte Indexierung* stellt einen Mittelweg zwischen der manuellen und der vollautomatischen Methode dar. Sie kombiniert Elemente beider Verfahren, wobei die automatisch erzeugten Ergebnisse in der Regel einer manuellen Nachkontrolle unterzogen werden. Dieser Ansatz ermöglicht es, die Effizienz der automatisierten Verfahren mit der Präzision der intellektuellen Indexierung zu verbinden, wobei die Endkontrolle durch den Menschen eine wichtige Qualitätssicherung darstellt.[8]

Jede dieser Indexierungsmethoden hat ihre spezifischen Vorzüge und Herausforderungen. Während die intellektuelle Indexierung für ihre hohe Qualität bekannt ist, bietet die automatisierte Indexierung eine sehr hohe Effizienz bei der Verarbeitung großer Mengen von Daten. Die computerunterstützte Indexierung bildet ein

4 Vgl. Lepsky, Klaus: B 3 Automatisches Indexieren, 2023, S. 171
5 Ebenda, S. 171
6 Vgl. Ebenda, S. 171
7 Vgl. Ebenda, S. 171
8 Vgl. Ebenda, S. 171

Gleichgewicht zwischen diesen beiden Extremen, indem sie die Vorteile der Automatisierung mit der menschlichen Überprüfung kombiniert. Die Wahl der geeigneten Indexierungsmethode hängt letztlich von den spezifischen Anforderungen und Ressourcen der jeweiligen Institution oder des Projekts ab.

3 Aktueller Forschungsstand in Deutschland

3.1 Digitaler Assistent DA-3

Der Digitale Assistent DA-3 ist ein fortschrittliches webbasiertes Tool, der Firma Eurospider[9] das speziell zur Unterstützung der „intellektuellen verbalen und klassifikatorischen Inhaltserschließung“[10] entwickelt wurde. Dieses Instrument bietet zwei Hauptfunktionen: Zum einen generiert es Schlagwortvorschläge und Notationen aus einer Vielzahl unterschiedlicher Quellen, um die Bearbeitung von Titelaufnahmen zu ergänzen. Zum anderen ermöglicht es die Gewährleistung einer vollständigen intellektuellen Erschließung.[11]

Im Rahmen einer Zusammenarbeit von Eurospider Information Technology[12], einem Konsortium des Integrierten Bibliothekssystems Baden-Württemberg (IBS BW), der Staatsbibliothek zu Berlin (SBB) sowie den Verbundzentralen VZG und BSZ, wurde ein dreijähriges Projekt umgesetzt, das den Einsatz von DA-3 auf der Bibliotheksservice-Plattform K10 Plus vorsah. Der 2016 eingeführte Vorläufer dieses Projekts, DA-2, erfreute sich bereits großer Beliebtheit, vor allem aufgrund seiner einfachen Handhabung. In einer Zeit, in der BibliothekarInnen aufgrund von Personalabbau und einem exponentiellen Anstieg der Datenmengen und Ressourcenvielfalt nach Arbeitshilfen suchen, erwies sich DA-2 schnell als effektives Tool.[13]

Einer der herausragenden Aspekte von DA-3 sei, nach Meinung des Österreichischen Bibliotheksverbundes (OBV), seine unkomplizierte Vergabe von Normdaten. Dieses Feature erleichtere nicht nur die Standardisierung und Vernetzung von Metadaten, sondern fördere auch die Interoperabilität innerhalb des Bibliotheks- und Informationssystems. Durch die Automatisierung dieses Prozesses würden nicht nur Fehlerquellen minimiert, sondern auch eine konsistente Anwendung von Normdaten und Klassifikationen gewährleistet.[14] Die Leistungsfähigkeit von DA-3 als Verschlagwortungstool wurde besonders hervorgehoben. Seine Performance sei nicht nur in der Geschwindigkeit der Datenverarbeitung erkennbar, sondern auch in der Präzision der Ergebnisse, die geliefert werden.[15]

---

9 Vgl. Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen, 2022, S. 1

10 Beckmann, Regine: Der Digitale Assistent DA-3: eine Plattform für die Inhaltserschließung, 2023, S. 1

11 Vgl. Ebenda, S. 2

12 Vgl. Eurospider Information Technology AG: Online Workshop „Computerunterstützte Inhaltserschließung“ am 14./15. November 2023, 2023

13 Vgl. Ebenda, S. 2 f.

14 Vgl. Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen, 2022, S. 1

15 Vgl. Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen, 2022, S. 1 f.

Die SBB zeigte aufgrund ihrer vielfältigen Anforderungen (Forschungsliteratur aus nahezu allen Fachbereichen und in verschiedenen Publikationssprachen, ein hoher wissenschaftlicher Anspruch sowie diverse Sonderabteilungen und Fachinformationsdienste) als eine der ersten Institutionen ernsthaftes Interesse an DA. Tatsächlich eingesetzt wurde die Technologie erst nach einer umfassenden Anpassung und Überarbeitung im Rahmen der Neuprogrammierung des Tools.[16]

Der Zusammenschluss der Verbunddatenbanken GBV und BSZ, die zusammen über 200 Millionen Titelnachweise verfügen, markierte einen entscheidenden Schritt zur Verbesserung des DA-Projekts. Die Schaffung eines einheitlichen Regelwerks ermöglichte die Etablierung von DA-3 als zentrales Arbeitsinstrument für die intellektuelle Sacherschließung.[17]

DA-3 fungiert als Webapplikation, die eine direkte SRU-Schnittstelle zum Verbundkatalog bietet und den Zugriff auf externe Quellen für bereits vorhandene Sacherschließungsdaten ermöglicht.[18] Die modulare Struktur des Assistenten macht ihn benutzerabhängig und individuell konfigurierbar, was den AnwenderInnen erlaubt, das Tool an spezifische Bedürfnisse anzupassen. Die Applikation integriert Normdaten wie die Gemeinsame Normdatei (GND)[19], den Standard-Thesaurus Wirtschaft (STW)[20], die Regensburger Verbundklassifikation (RVK)[21], die Basis Klassifikation (BK) sowie spezifische Klassifikationen wie die des Index Theologicus[22].

Ein weiteres Kernelement von DA-3 ist die Nutzung von Fremddatenvorschlägen und „approximativen Übersetzungen“[23] für die Erschließung, wobei Fremddatenquellen wie B3Kat, swissbib und die Kataloge der DNB, des OBV, der Library of Congress (LoC) und der Bibliothèque nationale de France (BnF) herangezogen werden.[24]

Thomas Murphy von Eurospider beschreibt die Funktionsweise der Ähnlichkeitssuche im DA-3 durch die Differenzierung in vier spezifische Nutzungsszenarien, die die Flexibilität und Effizienz dieses Instruments in der bibliothekarischen Praxis unterstreichen. Wenn einer der Fälle nicht zielführend ist, kommt es zur Ausführung des Nächstfolgenden.[25]

Im ersten Fall wird die Ähnlichkeitssuche dazu genutzt, innerhalb eines Katalogs nach anderen Exemplaren dieses Titels zu suchen. Diese Funktion ist besonders nützlich, wenn für ein Exemplar bereits Schlagwörter vergeben wurden, die dann

16 Vgl. Beckmann, Regine: Der Digitale Assistent DA-3: eine Plattform für die Inhaltserschließung, 2023, S. 5 f.
17 Vgl. Ebenda, S. 6
18 Vgl. Ebenda, S. 7
19 Vgl. Ebenda, S. 8
20 Vgl. Ebenda, S. 9
21 Vgl. Ebenda, S. 10
22 Vgl. Ebenda, S. 6
23 Ebenda, S. 11
24 Vgl. Ebenda, S. 11
25 Vgl. Schmucker, Susanne: Fortschritte der Automatisierung und Kollaboration in der Sacherschließung: Bericht vom 6. Workshop „Computerunterstützte Inhaltserschließung“, 2023, S. 3

einfach auf andere Exemplare desselben Titels übertragen werden können, was den Erschließungsprozess erheblich beschleunigt und vereinfacht.

Der zweite Fall tritt ein, wenn die Stichwörter des Titels nicht in der GND verzeichnet sind, jedoch bereits aus anderen Klassifikationssystemen wie der DDC, dem STW, der BK oder der RVK vorliegen. In dieser Situation ermöglicht die Ähnlichkeitssuche eine effiziente Generierung einer Schlagwortliste basierend auf der vorhandenen Datenlage, wodurch der Erschließungsprozess auch ohne GND-Einträge optimiert wird.[26]

Der dritte Fall beschreibt ein Szenario, in dem Sacherschließende die Ähnlichkeitssuche nutzen, um nach Titeln mit ähnlicher Erschließung zu recherchieren. Dies dient der Überprüfung und möglichen Ergänzung ihrer vorläufigen Erschließung. Das System ist hierbei in der Lage, verschiedene Ausgaben desselben Titels sowie Titel in anderen Sprachen anzuzeigen, was die Qualität und Vollständigkeit der Erschließung unterstützt.[27]

Im vierten und letzten Fall wird nach Titeln zu einem spezifischen Thema sowie nach passenden GND-Schlagwörtern gesucht. Diese Funktion erlaubt es, die Suche nach der Überprüfung der ersten Suchergebnisse zu verfeinern, um noch relevantere Titel und zusätzliche passende Schlagwörter zu identifizieren, was die Präzision der bibliothekarischen Recherche und Erschließung weiter erhöht.[28]

Über diese spezifischen Anwendungsfälle hinaus eröffnen die beschriebenen Funktionen interessante Perspektiven für die Zukunft, besonders im Hinblick auf die Analyse von Forschungsdaten und die Einbindung in Discovery-Systeme sowie die Weiterentwicklung bibliothekarischer Such- und Erschließungstools.[29]

Trotz dieser fortschrittlichen Funktionen obliegt die Qualitätssicherung den AnwenderInnen selbst, die die generierten Vorschläge beurteilen und gegebenenfalls anpassen müssen.[30]

3.2 Annif

Annif ist eine Open-Source Software für die automatisierte Klassifizierung und Schlagwort-Indexierung. Open-Source bedeutet, dass das Projekt auf GitHub für die Öffentlichkeit einsehbar, kommentierbar und zum Download verfügbar ist. Neben der direkten Verfügbarkeit auf GitHub kann Annif auch aus dem PyPI-Repository heruntergeladen oder als Docker-Container, ein kleines Betriebssystem, wo isoliert der Code läuft, genutzt werden, was seine Zugänglichkeit und Anwendbarkeit in verschiedenen Umgebungen und Systemen sicherstellt.[31]

---

26 Vgl. Ebenda, S. 3
27 Vgl. Ebenda, S. 3
28 Vgl. Ebenda, S. 3
29 Vgl. Ebenda, S. 3
30 Vgl. DA-3: Der Digitale Assistent DA-3 ist eine Plattform zur Unterstützung der inhaltlichen Erschließung in Bibliotheken, 2022
31 Vgl. National Library of Finland: Annif, 2024

Annif wurde ursprünglich von der finnischen Nationalbibliothek initiiert, mit Osma Suominen als Hauptentwickler, der den ersten Prototypen im Jahr 2017 erstellte.[32] Bis zum Jahr 2019 integrierte Annif in seinen Backends vier zentrale Algorithmen, die allesamt auf Open-Source-Bibliotheken basieren.[33] Diese Algorithmen wurden implementiert, um eine breite Palette von Erschließungsaufgaben zu unterstützen. Jeder dieser Algorithmen bringt spezifische Stärken in den Prozess der automatisierten Schlagwort- und Klassifikationszuweisung ein und erfordert eine initiale Trainingsphase mit Metadaten oder Volltextdokumenten, wie es für maschinelles Lernen[34] üblich ist.[35]

Im Jahr 2022 wurden die neuen Algorithmen MLLM (Maui-like Lexical Matching), Parabel und Bonsai in Annif integriert. Diese Erweiterungen reflektieren das Bestreben des Entwicklerteams, die Leistungsfähigkeit von Annif kontinuierlich zu verbessern und an die sich wandelnden Anforderungen der bibliothekarischen Erschließungspraxis anzupassen. Das Projekt Annif wird somit keineswegs als abgeschlossen betrachtet, vielmehr findet eine stetige Verbesserung und Anpassung der Algorithmen statt.[36]

Ein prägnantes Beispiel für die Notwendigkeit dieser kontinuierlichen Anpassung ist die Herausforderung der Normalisierung des Vokabulars. Die Algorithmen müssen in der Lage sein, Wörter auf ihren Wortstamm zu reduzieren, um grammatikalische Varianten zu vereinheitlichen und Fehler zu vermeiden.[37] So wird beispielsweise das Wort „Bäume“ auf den Wortstamm „Baum“ reduziert.

Inzwischen hat sich Annif zu einem umfangreichen Tool entwickelt, das von verschiedenen Institutionen, darunter das ZBW – Leibniz-Informationszentrum Wirtschaft und die Niederländische Nationalbibliothek, sowohl eingesetzt als auch weiterentwickelt wird.[38]

Einige der wichtigsten Vorteile von Annif sind seine Flexibilität und Anpassungsfähigkeit an bestehende Erschließungsverfahren sowie die Fähigkeit zur Verarbeitung von Texten in verschiedenen Sprachen.[39] Diese Multilingualität macht Annif zu einem vielseitigen Werkzeug, das unabhängig von der Sprache des zu

32 Vgl. Suominen, Osma: Annif: DIY automated subject indexing using multiple algorithms, 2019, S. 5

33 Vgl. Ebenda, S. 7–9

34 Eine genauere Einführung in den Bereich des maschinellen Lernens würde tiefergehende Kenntnisse im Bereich der Informatik voraussetzen und wird daher in dieser Arbeit nicht vorgenommen. Weitere Informationen zu maschinellem Lernen im bibliothekarischen Kontext sind unter anderem bei E. Mecking zu finden.

35 Vgl. Suominen, Osma: Annif: DIY automated subject indexing using multiple algorithms, 2019, S. 7

36 Vgl. Suominen, Osma: Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing automated subject indexing, 2022

37 Vgl. Suominen, Osma: Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing automated subject indexing, 2022

38 Vgl. National Library of Finland: Annif, 2024

39 Vgl. Mecking, Elizabeth: Automatisierte Vergabe von Notationen der Regensburger Verbundklassifikation RVK mithilfe des Toolkits Annif, 2023, S. 10

erschließenden Materials eingesetzt werden kann, sofern die entsprechenden Sprachtechnologietools für die jeweilige Sprache verfügbar sind.[40]

Der finnischen Nationalbibliothek zufolge ermöglicht Annif die Verwendung jedes Indexierungsvokabulars[41], was die Anwendbarkeit in unterschiedlichen fachlichen Kontexten und für verschiedene Erschließungsaufgaben ermöglicht. Die modulare Erweiterbarkeit von Annif ist ein weiterer entscheidender Vorteil, denn so können neue Funktionen und Algorithmen leicht integriert und einfach individualisiert werden.[42]

3.3 DNB – Deutsche Nationalbibliothek

Die Nationalbibliografie der Deutschen Nationalbibliothek zielt darauf ab, eine umfassende Verzeichnung aller seit 1913 erschienenen deutschen und deutschsprachigen Ressourcen zu erstellen. Ebenfalls enthalten sind „im Ausland [erscheinende] Germanica und Übersetzungen deutschsprachiger Werke [und] zwischen 1933 und 1945 [erschienene] Werke deutschsprachiger Emigranten“[43]. Die Ressourcen werden systematisch auf unterschiedliche Reihen verteilt und die bibliografischen Beschreibungen zugänglich gemacht. Die innere Unterteilung basiert auf den Sachgruppen der Dewey-Dezimalklassifikation. Jede der sieben Reihen verzeichnet unterschiedliche Ressourcen[44]:

Reihe A: Monografien und Periodika des Verlagsbuchhandels
Reihe B: Monografien und Periodika außerhalb des Verlagsbuchhandels
Reihe C: Karten
Reihe H: Hochschulschriften
Reihe M: Musikalien
Reihe O: Online-Publikationen
Reihe T: Musiktonträgerverzeichnis

Die Entwicklung, welche die DNB im Bereich der automatischen Inhaltserschließung im letzten Jahrzehnt durchlaufen hat, soll im Folgenden chronologisch umrissen werden. Die DNB initiierte Anfang 2010 einen signifikanten Wandel in der Erschließung von Online-Publikationen, insbesondere mit der Einführung der Reihe O in der Deutschen Nationalbibliografie[45]. Dies markierte den Übergang von manuellen Verfahren der Inhaltserschließung zu automatisierten Methoden, ein Fortschritt, welcher durch das *Petrus-Projekt*, einen der Versuche für die automatisierte Sacherschließung, realisiert werden konnte.[46]

---

40 Vgl. National Library of Finland: Annif, 2024
41 Vgl. Ebenda, 2024
42 Vgl. Mödden, Elisabeth: Annif gemeinsamer Werkzeugkasten, 2020, Folie 9–11
43 Deutsche Nationalbibliothek: Deutsche Nationalbibliografie, 2023
44 Vgl. Deutsche Nationalbibliothek: Deutsche Nationalbibliografie, 2023
45 Vgl. Ebenda
46 Vgl. Schöning-Walter, Christa: PETRUS – Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek, S. 15

Im Laufe der Zeit wurde dieser Ansatz weiterentwickelt und verfeinert: 2012 begann die maschinelle Vergabe von Sachgruppen[47], gefolgt von der automatisierten Schlagwortvergabe für deutschsprachige Netzpublikationen unter Verwendung der GND im Jahr 2014. Die Einführung der Publikationsverarbeitung im System zur Vergabe von DDC-Kurznotationen für medizinische Publikationen der Reihe O, markierte einen neuen Erfolg. Weiterhin fand im Jahr 2015 eine erfolgreiche Ausweitung auf andere Fachbereiche statt.[48]

Die Grundlage dieser maschinellen Klassifikation bildet wie bei Annif das maschinelle Lernen.[49] Ein spezielles Verfahren, das in diesem Kontext Anwendung findet, ist die *Support Vector Machine* (SVM). SVM ist ein überwachtes Lernmodell, das für Klassifikations- und Regressionsaufgaben eingesetzt wird. Es funktioniert, indem es den Raum der Merkmale so aufteilt, dass die verschiedenen Kategorien durch eine klare Lücke, die so weit wie möglich ist, getrennt werden.[50]

Für die Trainingsphase und somit die kontinuierliche Verbesserung des Systems werden bibliografische Metadaten wie Titel, Volltexte und Scans von Inhaltsverzeichnissen als Textmaterial zum Einspielen genutzt. Die Analyse der Daten erfolgt durch den Einsatz linguistischer und statistischer Methoden zur Identifikation sprachlicher Merkmale. In dieser Phase erstellt das System anhand intellektuell erschlossener Publikationen ein Referenzmodell für alle Klassen. Es ist entscheidend, dass jede Klasse über eine ausreichende Anzahl an Beispielen verfügt und dass diese Beispiele in Bezug auf Qualität und Quantität annähernd gleichwertig sind, um effektive Lernergebnisse zu gewährleisten.[51]

Wie hoch die Übereinstimmung einer neuen Publikation mit den bereits existierenden ist, wird mit einem statistischen Maß ermittelt. Die ähnlichsten Sachgruppen und Kurznotationen werden der neuen Publikation mit einem Konfidenzwert zugeordnet, wodurch täglich etwa 4000 englische und deutsche Publikationen automatisch erschlossen werden können.[52] Die Vergabe von Schlagworten erfolgt ausschließlich auf linguistischer Basis, wobei im Deutschen das GND-Vokabular und im Englischen die Library of Congress Subject Headings[53] verwendet werden. Des Weiteren findet eine Kombination mit Nicht-Sachbegriffen der GND wie Personen, Geografika, Körperschaften, Kongresse und Werke ebenfalls Anwendung.[54]

---

47 Vgl. Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek, 2021, S. 61

48 Vgl. Ebenda, S. 61 f.

49 Vgl. Mecking, Elizabeth: Automatisierte Vergabe von Notationen der Regensburger Verbundklassifikation RVK mithilfe des Toolkits Annif, 2023, S. 7 f.

50 Vgl. Kanade, Vijay: What is a support vector machine?: working, types, and examples, 2024

51 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

52 Vgl. Ebenda

53 Vgl. Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek, 2021, S. 61

54 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

Seit 2014 werden deutschsprachige Online-Publikationen in der Reihe O automatisch erschlossen, Zeitschriftentitel und Belletristik sind dabei ausgenommen. Außerdem werden Online-Zeitschriftenartikel und Print-Publikationen der Reihe H, die mithilfe von gescannten Inhaltsverzeichnissen durch das Table of Contents (TOC) Verfahren bearbeitet wurden, ebenfalls automatisiert erschlossen.[55] Durch die kontinuierliche Anpassung und Erweiterung der Prozesse der DNB wird dies seit 2018 auch für die englischsprachigen Online-Dissertationen (verzeichnet in Reihe O) angewandt.[56]

Im Rahmen der Konzepte zur Erschließung physischer Medien wurde unter anderem die Anwendung von Optical Character Recogniton (OCR) eingeführt. Genutzt wird OCR beispielsweise zur Verarbeitung digitalisierter Inhaltsverzeichnisse, damit im Anschluss die maschinelle Erzeugung von Schlagworten und Notationen erfolgen kann.[57] Diese Fortschritte ermöglichten es der DNB, zwei wichtige Anpassungen in ihrer Erschließungsstrategie vorzunehmen. Zum einen wurde die manuelle Inhaltserschließung von Kinder- und Jugendliteratur sowie belletristischen Werken eingestellt, zum anderen wurde die detaillierte Klassifizierung von Publikationen aus den Reihen B und H mit vollständigen DDC-Notationen zugunsten der Vergabe von Schlagwörtern aus der GND beendet.[58]

Das Erschließungssystem, das für die Verarbeitung und Klassifizierung von bibliographischen Daten verwendet wird, gliedert sich in drei grundlegende Bereiche: Steuerung, Funktion (d. h. Verarbeitungsketten) und Datenhaltung[59]. Ein wesentliches Instrument in diesem Prozess ist die Averbis Extraction Plattform, entwickelt vom Freiburger Unternehmen Averbis[60]. Diese Plattform ermöglicht es, täglich zu einer festgelegten Zeit mit der automatischen Erschließung zu beginnen. Dabei werden Identifikationsnummern (IDNs) an das Steuerungssystem übergeben und die zu erschließenden digital vorliegenden Volltexte oder Inhaltsverzeichnisse aus dem Repositorium sowie bibliografische Metadaten aus der Datenbank (CBS) geholt. Diese Informationen werden in Textdateien umgewandelt, damit eine Bestimmung der Sprache erfolgen kann. Die anschließende Klassifizierung und Verschlagwortung der Texte erfolgt durch die Averbis-Plattform, die Ergebnisse werden über das CBS-System an den jeweiligen Titeldatensatz angehängt.[61]

---

55 Vgl. Ebenda

56 Vgl. Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek, 2021, S. 62

57 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

58 Vgl. Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek , 2021, S. 62

59 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

60 Vgl. Mödden, Elizabeth: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek: breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse, 2018, S. 32

61 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

Die Datengrundlage für diese Prozesse bildet unter anderem *Big Data*, darunter zählen Inhaltsverzeichnisse, Online-Publikationen, Metadaten, Fremddaten, das GND-Vokabular, verschiedene Thesauri, usw. Aufgrund des starken Anstiegs von Online- im Vergleich zu Printpublikationen sowie der unterschiedlichen Anforderungen verschiedener Gebiete an die Klassifikationsvergabe und Verschlagwortung, müssen die Verfahren zur Textaufarbeitung und verarbeitung entsprechend angepasst werden. Diese Anpassung ist in der Averbis Software konfigurierbar.[62]

Dabei stellt die Heterogenität der Datenbasis eine Herausforderung dar. Beispielsweise sind Inhalte wie Danksagungen und Vorworte in Dissertationen meist irrelevant für die Erschließung, weshalb die Entfernung der ersten 20.000 Zeichen eines Textdokumentes lange eine gute Lösung darstellte. Voraussetzung dafür war jedoch eine homogene Struktur aller vorliegenden Texte. Angesichts der zunehmenden Heterogenität von bspw. Dissertationen, war die Einführung einer Textstrukturerkennung von Vorteil, die allerdings erst 2020 implementiert werden konnte.[63]

Weitere Schwierigkeiten bei der Verarbeitung sind Sprach- sowie OCR-Erkennungsfehler, die durch Ansätze aus dem Bereich des Deep Learning behoben werden sollen. Dafür ist bereits ein Projekt angedacht. Texte, die zu wenig Daten für eine entsprechende Analyse hergeben, sind ebenfalls äußerst herausfordernd. Das passiert, wenn zum Beispiel das Inhaltsverzeichnis den Inhalt der Publikation nicht aussagekräftig widerspiegelt. Doch auch Publikationen mit zu hohem heterogenem Inhalt erweisen sich als schwierig und erfordern nahezu immer eine intellektuelle Nachkontrolle.[64]

Sandro Uhlmann, ein Mitarbeitender der DNB, identifiziert die Ambiguität von Begriffen als größtes Problem der automatisierten Verschlagwortung. Er veranschaulicht dies an folgendem Beispiel: der Begriff *stars* aus der Publikation *Empirical essays on the role of stars in collaborative organizations* wurde fälschlicherweise mit Astronomie assoziiert (*Sternen*), anstatt mit *celebrities* bzw. *Berühmtheiten*.[65]

Die Mehrdeutigkeit von Begriffen erfordert komplexe Disambiguierungsverfahren.[66] Die Disambiguierung ist eng mit der Arbeit an Wörterbüchern verbunden. Dabei werden Sachbegriffe maschinell in ihre kleinsten Bestandteile, die Segmente, zerlegt. Das Ziel ist es, möglichst viele grammatikalische Varianten zu berücksichtigen, um die Genauigkeit der Verschlagwortung zu erhöhen. Diese Aufspaltung in

62 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

63 Vgl. Ebenda

64 Vgl. Ebenda

65 Vgl. Franke-Maier, Michael: Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung: Bericht zur Fachtagung Netzwerk maschinelle Erschließung an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019, 2020, S. 10

66 Vgl. Mödden, Elizabeth: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek: breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse, 2018, S. 32

Segmente und die anschließende grammatikalische Analyse sind essenziell, um die Mehrdeutigkeiten, die Sprache naturgemäß mit sich bringt, aufzulösen.[67]

Averbis hatte hierfür ein mehrstufiges Verfahren entwickelt. Aufgrund der Einstellung von Averbis seitens der Firma ist es ausschließlich möglich, die bereits gesammelten Daten und Erfahrungen zu nutzen. Alternativ in Betracht gezogenen Ansätze wie Wörterbuchanpassungen oder Erschließungskombinationen durch KI konnten bisher nicht umgesetzt werden.[68]

In Anbetracht dieser Einschränkungen und der stetig wachsenden Anforderungen an die bibliographische Erschließung wurde der Beginn eines neuen Projekts namens EMa (Erschließungs-Maschine) im Jahr 2019 beschlossen, das zunächst bis Ende 2022 laufen sollte. Ziel war die Entwicklung eines neuen Systems für die automatisierte Inhaltserschließung, das bei gleichzeitiger Funktionsübernahme des Vorläufers an Flexibilität und Anpassungsfähigkeit gewinnt.[69] Beispiele für die Funktionen, die erhalten bleiben sollten, sind unter anderem die Vergabe der DDC-Sachgruppen bzw. Kurznotationen für deutsch- und englischsprachige Publikationen sowie die maschinelle Verschlagwortung mit GND-Vokabular für deutschsprachige Publikationen[70]. EMa sollte nicht nur die bisherigen Terminologien, Notationen und Codierungen verwenden können, sondern auch die Integration neuer Termini erleichtern. Das neue Erschließungsverfahren sollte auf modularen Elementen beruhen, womit der Austausch von Funktionen und die flexible Kombination unterschiedlicher Verfahren garantiert wäre. Letztendlich war durch das neue Projekt eine Zusammenarbeit der DNB mit internen, aber auch externen Systemen geplant. Die Automatisierungsprojekte führten zu einer Neuausrichtung der Aufgaben für das Bibliothekspersonal, die nun die Qualität der Ausgaben des Erschließungssystems überwachen, Averbis bei der Weiterentwicklung unterstützten und Testläufe durchführen.[71]

Im Jahr 2017 kam es zu einer öffentlichen heftigen Diskussion[72] bezüglich der automatisierten Inhaltserschließung der DNB. Grund war eine interne Analyse, die zeigte, dass die Ergebnisse der automatischen Inhaltserschließung, speziell die algorithmisch generierten Schlagworte, nur zu 50 % als hilfreich erachtet wurden, während die restlichen 50 % als weniger hilfreich eingestuft wurden. Das löste eine allgemeine Sorge aus, insbesondere angesichts der Rolle der DNB als zentrale

67 Vgl. Franke-Maier, Michael: Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung: Bericht zur Fachtagung Netzwerk maschinelle Erschließung an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019, 2020, S. 10

68 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

69 Vgl. Busse, Frank: Erschließungsmaschine gestartet, 2022

70 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

71 Vgl. Ebenda

72 Vgl. Ceynowa, Klaus: Deutsche Nationalbibliothek: in Frankfurt lesen jetzt zuerst Maschinen, 2017

Einrichtung für die Informationsbereitstellung in Deutschland. Die Frage, die daraus resultierte war, ob die DNB ihrer essenziellen Aufgabe der qualitativen Inhaltserschließung unter diesen Umständen nachkommt. Bisherige Algorithmen waren hauptsächlich darauf ausgerichtet, Publikationstitel zu analysieren, die reich an Text und aussagekräftig bezüglich ihres Inhalts waren. Diese Vorgehensweise führte zu einer adäquaten Verschlagwortung und Auffindbarkeit eben dieser Publikationen.[73]

Die offene Frage, die sich dabei ergab, war, wie sich der Algorithmus gegenüber geisteswissenschaftlichen Texten verhält oder solchen, die Allegorien, Ironie und andere stilistische Mittel enthalten. Diese Literatur ist von Algorithmen nur schwer zu erfassen und aus diesem Grund wurden Publikationen der Verlagsreihe A (Monografien und Periodika des Verlagsbuchhandels) weiterhin intellektuell erschlossen.[74]

Die Problematik und gleichzeitig die Notwendigkeit der automatischen Inhaltserschließung wurden durch Ute Schwens, die Direktorin der DNB in Frankfurt, unterstrichen. Sie bestätigte, dass in den letzten Jahren ein deutlicher Anstieg von Netzpublikationen im Vergleich zu gedruckten Medien zu verzeichnen war.[75]

> *„2016 verzeichnete die DNB einen Zugang von 1,37 Millionen Netzpublikationen und von 522.000 körperlichen Medienwerken; bei nur ca. 100.000 Titeln handelte es sich um Bucherscheinungen aus dem Verlagsbuchhandel, von denen ca. 74.000 verbal erschlossen wurden.“*[76]

Der Versuch alle Publikationen intellektuell zu erschließen stand somit im direkten Widerspruch zu dieser Entwicklung, insbesondere unter Berücksichtigung des sich verkleinernden Personalstamms.[77]

Die DNB entschied sich, trotz der primären Ausrichtung auf die automatisierte Sacherschließung, für die Reihe A weiterhin eine intellektuelle Inhaltserschließung zu nutzen, bis ein maschinelles Verfahren adäquaten Ersatz bietet. Dies betrifft insbesondere die „wissenschaftliche Literatur in sogenannten buchaffinen Fächern“[78] wie geistes- und sozialwissenschaftliche Literatur und die Ingenieurswissenschaften.[79]

Im Jahr 2019 traf die DNB die Entscheidung, die maschinelle Verschlagwortung für die Reihe B vorerst auszusetzen. Der Grund dafür war, dass der Anteil sozialwissenschaftlicher Publikationen ein zu hohes Maß an weniger spezifischem Vokabular aufweist und ein hoher Anteil an inhaltstragenden Individualbegriffen wie Körperschaften in „nichtwissenschaftlichen Publikationen“ (z. B. Ortsführern), zu

---

73 Vgl. Börsenblatt: Was Maschinen können und was nicht: automatisierte Inhaltserschließung in der Deutschen Nationalbibliothek, 2017

74 Vgl. Ebenda

75 Vgl. Ebenda

76 Ebenda

77 Vgl. Ebenda

78 Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek, 2021, S. 63

79 Vgl. Ebenda, S. 63 f.

einer hohen Quote an Erschließungsfehlern führte. Die maschinelle Erschließung der Reihe H wurde jedoch fortgesetzt, da die Ergebnisse als akzeptabel angesehen wurden.[80]

Im April 2022 wurde EMa in Betrieb genommen. Sie ist insbesondere darauf ausgerichtet die automatische Vergabe von Sucheinstiegen für deutschsprachige E-Books, elektronische Zeitschriftenartikel und gedruckte Hochschulschriften zu realisieren.[81] Als Schlüsselelement dieser Automatisierung dient dabei die Vergabe von DDC-Sachgruppen, die Zuweisung relevanter Schlagwörter aus dem GND-Vokabular und die Bestimmung von DDC-Kurznotationen für Sachgruppen, insbesondere im Bereich der Medizin, sowohl für deutsch- als auch englischsprachige Publikationen.

Die Architektur der Erschließungsmaschine ist modulartig aufgebaut, wodurch eine flexible Anpassung und Erweiterung des Systems ermöglicht wird. Grundlage und Kern von EMa bildet die im Kapitel 3.2 beschriebene Software Annif. Als integraler Bestandteil von EMa dient Annif der effizienten Vergabe von Schlagwörtern sowie der Zuweisung von Sachgruppen mit Kurznotationen der DDC.[82]

> *„Die vollautomatische Vergabe von Kurznotationen wird schon eingesetzt, Sachgruppen der DDC-Printpublikationen der Reihen B und H sind teilautomatisiert. Die Reihe O und Artikel werden mit Annif vollautomatisiert erschlossen, einschließlich DDC-Kurznotation und DDC-Sachgruppen.“*[83]

Dies verdeutlicht die zunehmende Integration und das Vertrauen in die Leistungsfähigkeit von Annif im Kontext der automatisierten bibliothekarischen Erschließung.[84]

Die Basis für das Training der Algorithmen von Annif sind die Trainingsdokumente der DNB, die es ermöglichen das Modell kontinuierlich zu verfeinern und zu verbessern. Der Prozess der Erschließung selbst erfährt durch die Einführung von Annif keine grundlegende Veränderung, außer dass die Übergabe der Textformate nicht mehr an Averbis, sondern direkt an Annif erfolgt.[85] Die genauen Abläufe innerhalb dieses Services bzw. der Algorithmen von Annif sind dabei nicht Gegenstand dieser Arbeit.

Die Qualitätssicherung gestaltet sich als herausfordernd, da es keinen allgemein gültigen Qualitätsbegriff für die Sacherschließung gibt. Die Qualität maschineller Inhaltserschließung wird anhand von Vergleichen mit Erschließungsergebnissen unterschiedlicher Einrichtungen, manuellen Erschließungs-Ergebnissen, Expertenbewertungen sowie den gängigen Gütemaßen aus dem Information Retrieval

80 Vgl. Ebenda, S. 64

81 Vgl. Busse, Frank: Erschließungsmaschine gestartet, 2022

82 Vgl. Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek, 2020

83 Ebenda

84 Vgl. Mecking, Elizabeth: Automatisierte Vergabe von Notationen der Regensburger Verbundklassifikation RVK mithilfe des Toolkits Annif, 2023, S. 10 f.

85 Vgl. Busse, Frank: Erschließungsmaschine gestartet, 2022

wie *recall*[86], *precision*[87] und dem daraus abgeleiteten *F-Maß*[88] bewertet.[89] Auch stochastisch ermittelte Konfidenzwerte spielen eine wichtige Rolle in der Bewertung der maschinellen Erschließungsqualität.[90]

Die maschinelle Erschließung zielt darauf ab thematische Zugriffspunkte zu bieten, um Dokumente auffindbar zu machen, wobei die Retrieval-Aspekte von großer Bedeutung sind. Die Nützlichkeit des vergebenen Schlagwortbegriffs wird durch eine maschinell verwertbare 4-stufige Skala erfasst, die von „sehr nützlich" bis „falsch" reicht. Diese Skalierung spiegelt die subjektive Einschätzung der Nützlichkeit eines Schlagworts wider und hilft bei der Feinabstimmung der Erschließungsprozesse.[91] Die Balance zwischen *precision* und *recall* ist entscheidend für die Effektivität der Erschließung. Das *F-Maß*, dient der DNB als zentrales Gütemaß zur Bewertung ihrer Erschließungsergebnisse.[92]

Die Modularität der Erschließungsmaschine erlaubt eine kontinuierliche Verbesserung und Anpassung an neue Anforderungen und Entwicklungen. Ein prägnantes Beispiel für die Integration digitaler Hilfsmittel ist der Einsatz des Digitalen Assistenten DA-3. Frank Scholze, der Generaldirektor der DNB, unterstrich im Jahr 2023 die Relevanz solcher technologischen Hilfsmittel und betonte die Notwendigkeit, auf die Unterstützung durch Tools wie DA-3 zu setzen. Der Digitale Assistent bietet eine umfassende Hilfestellung durch die Bereitstellung von Metadaten, die Pflege von Normdaten sowie die Unterstützung sowohl der intellektuellen als auch der maschinellen Erschließung.[93] Besonders wenn man die Herausforderung bedenkt, dass viele Dokumente sowohl in elektronischer als auch in gedruckter Form eingehen.

In der Regel sind elektronische Publikationen bereits im Vorfeld maschinell formal erschlossen, was eine erhebliche Arbeitserleichterung darstellt. DA-3 wirkt hier unterstützend, indem es die maschinell erschlossenen Daten in spezifischen Feldern innerhalb der Metadatensätze ablegt. Diese Unterfelder enthalten Informationen über den Prozess der Erschließung, die Herkunft der Daten sowie einen Konfidenzwert, der die Zuverlässigkeit der automatisierten Erschließung angibt. Durch die Integration zusätzlicher Unterfelder ist es weiterhin möglich, sich intellektuelle Relevanzbewertungen durch das System, anzeigen zu lassen. Um die Qualität und Verlässlichkeit der maschinell generierten Erschließungsinformationen zu gewährleisten, wird eine stichprobenartige Überprüfung durch FachreferentInnen vorgenommen.[94]

---

86 Frage nach der Vollständigkeit der abgerufenen Instanzen

87 Anteil der relevanten Instanzen unter den Abgerufenen

88 Vorhersageleistung, berechnet aus precision und recall

89 Vgl. Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek, 2021, S. 64 f.

90 Vgl. Ebenda, S. 65

91 Vgl. Ebenda, S. 65 f.

92 Vgl. Ebenda, S. 66

93 Vgl. Schmucker, Susanne: Fortschritte der Automatisierung und Kollaboration in der Sacherschließung: Bericht vom 6. Workshop „Computerunterstützte Inhaltserschließung", 2023, S. 1

94 Vgl. Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen, 2022, S. 5

Diese vielfältigen Funktionen des Digitalen Assistenten fördern eine effiziente und effektive Zusammenarbeit zwischen Menschen und Computer, wobei insbesondere die KI eine immer bedeutendere Rolle einnimmt.[95]

Dies wird unter anderem durch ein neues Forschungsprojekt der DNB seit 2021 unterstützt, das unter dem Titel „Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit Künstlicher Intelligenz" läuft. Dieses Projekt wird durch das Nationale Projekt der Bundesregierung für Kultur und Medien im Rahmen der KI-Strategie gefördert und ist bis zum 31. März 2025 angesetzt. Es zielt darauf ab die inhaltliche Erschließung von Publikationen mittels KI weiter zu optimieren.[96]

3.4 ZBW – Leibniz-Informationszentrum Wirtschaft

Seit 2016 verfolgt das ZBW – Leibniz-Informationszentrum Wirtschaft das Ziel, eine automatisierte bzw. computergestützte Inhaltserschließung für die individuellen Bedürfnisse der Institution zu entwickeln. Dieser Ansatz zielt darauf ab, die Effizienz und Präzision der Inhaltserschließung im Allgemeinen zu steigern.[97]

Die Datenbasis für diese Automatisierung bildeten Dokumente und Metadaten, die vorrangig durch den eigens entwickelten Standard-Thesaurus Wirtschaft (STW) intellektuell erschlossen wurden. Sowohl die inhaltliche Erschließung als auch die darauffolgende Eintragung und Verknüpfung der Metadaten mit dem Datensatz waren von besonderer Bedeutung. Aufgrund des nationalen Sammelauftrags im Bereich Wirtschaft und der Nutzung des STW ist die Fremddatennutzung für die jährlich 100.000 neuen Ressourcen sehr begrenzt.[98]

Angesichts dieser Herausforderung sowie der begrenzten Kapazität der ReferentInnen wurden von diesen neuen Ressourcen nur etwa 35.000, ca. 35 %, pro Jahr, intellektuell erschlossen. Daraus entwickelte sich das klare Ziel eine komplette Erschließung zu erreichen, um eine bessere Vergleichbarkeit der Daten gewährleisten zu können.[99]

Eine weitere Intention war es, wie von Moritz Fürneisen aus dem ehemaligen Projektteam formuliert, „so viele Abläufe wie möglich im Erschließungsworkflow zu automatisieren und dabei die Qualität der von der ZBW generierten Metadaten zu erhalten."[100] Die Rahmenbedingungen für die Automatisierung der

95 Vgl. Schmucker, Susanne: Fortschritte der Automatisierung und Kollaboration in der Sacherschließung: Bericht vom 6. Workshop „Computerunterstützte Inhaltserschließung", 2023, S. 1

96 Vgl. Deutsche Nationalbibliothek: Automatisches Erschließungssystem: Projekt Inhaltliche Erschließung mit KI, 2022

97 Vgl. Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht, 2023, S. 2

98 Vgl. Kasprzik, Anna: Automatisierte Inhaltserschließung auf der Basis von Machine-Learning-Methoden als Produktivverfahren an der ZBW, 2020

99 Vgl. Ebenda

100 Franke-Maier, Michael: Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung: Bericht zur Fachtagung Netzwerk maschinelle Erschließung an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019, 2020, S. 10–11

Inhaltserschließung seien bei der ZBW im Vergleich zur DNB deutlich günstiger. Ein wesentlicher Vorteil sei, dass die von der ZBW erschlossenen Publikationen aus einem klar begrenzten Fachgebiet stammen. Oft sind bereits Autoren-Keywords vorhanden, die eine solide Basis für die Erschließung bieten. Zudem gibt es im STW der ZBW nur wenige primär geografische Begriffe, was die Komplexität der Verschlagwortung reduziert.[101]

Während der ersten Versuchsphase entstanden im Rahmen des Projekts *AutoIndex* mehrere Prototypen. Diese waren jedoch nicht als Funktion einsetzbar, weil der entstandene Code nicht den erforderlichen Standards entsprach und keine adäquate IT-Infrastruktur für das nahtlose Aufarbeiten der Medatenverarbeitungsprozesse vorhanden war. Danach wurde die Prämisse verfolgt, dass für die spezifischen Anforderungen der ZBW eine maßgeschneiderte Lösung entwickelt werden musste. Auf Grundlage vorangegangener Ziele und Erkenntnisse wurde daher das Projekt AutoSE angestoßen.[102]

Bis zum Jahr 2020 konnte eine erste Softwarearchitektur entwickelt werden, die erste Lösungsansätze bot, um bestehende Nachweise bzw. Informationssysteme abzugleichen und zu verknüpfen. Sie dient als direkte Verbindung zur Bibliothek des Zentrums für Wirtschaftsforschung, wobei ein enger Austausch und eine intensive Zusammenarbeit mit dem Personal der intellektuellen Inhaltserschließung unerlässlich sind. Das bibliotheksbezogene Fachwissen spielt eine entscheidende Rolle in der Entwicklungsumgebung und der Funktionalität der entwickelten Programme.[103]

Einer der Ansätze von AutoSE war der Versuch, Annif als Kernmodul zu verwenden. Annif wird speziell für die Datenbank EconBiz genutzt, einem Suchportal der ZBW. Die dort enthaltenen Dokumente werden durch eine inhaltliche Erschließung aufbereitet, die maßgeblich auf dem STW basiert.[104] Das Tool spielt eine zentrale Rolle im Prozess der Inhaltserschließung, indem es Vorschläge für die Verschlagwortung der Dokumente unterbreitet. Diese Vorschläge basieren auf der Analyse der Textinhalte sowie der Identifizierung relevanter Begriffe, die mit dem im STW definierten kontrollierten Vokabular korrespondieren. Bislang beschränkt sich die Anwendung auf die Nutzung kontrollierter Vokabeln als Basis für die manuelle und automatisierte Erschließung.[105]

Die modulare Architektur von AutoSE erlaubt die Integration und Kombination mehrerer Machine-Learning-Module, einschließlich die von Eigenentwicklungen und Ansätzen aus dem Bereich des Deep Learning wie Transfermodelle für

101 Vgl. Ebenda, S. 11

102 Vgl. Kasprzik, Anna: Automatisierte Inhaltserschließung auf der Basis von Machine-Learning-Methoden als Produktivverfahren an der ZBW, 2020

103 Vgl. Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht, 2023, S. 3

104 Vgl. Kasprzik, Anna: Automatisierte Inhaltserschließung auf der Basis von Machine-Learning-Methoden als Produktivverfahren an der ZBW, 2020

105 Vgl. Mecking, Elizabeth: Automatisierte Vergabe von Notationen der Regensburger Verbundklassifikation RVK mithilfe des Toolkits Annif, 2023, S. 10–14

multilinguale Schlagwortvergaben.[106] Zudem ermöglicht es die Software Rohdaten ins Backend einzuspeisen und die Ergebnisse zu analysieren. In der aktuellen Implementierung von AutoSE dient Annif vornehmlich dazu, die eigenen Backend-Systeme (AutoSECore) zu integrieren und zu testen. Annif wird dabei als externe library verwendet.[107]

Trotz des erfolgreichen Einsatzes von Annif bestehen Unsicherheiten bezüglich seiner langfristigen Verwendung innerhalb von AutoSE. Dies liegt daran, dass die spezifischen Anforderungen der ZBW in einigen Fällen über die grundsätzlichen Funktionalitäten von Annif hinausgehen, dessen Anspruch eher auf einer niederschwelligen Anwendung liegt. Diese Diskrepanz zwischen den spezialisierten Bedürfnissen der ZBW und den allgemeineren Funktionen von Annif unterstreicht die Notwendigkeit einer fortwährenden Entwicklung und Anpassung der automatisierten Inhaltserschließungssysteme, um den spezifischen Anforderungen gerecht zu werden und die Effizienz sowie die Qualität der Metadatenverarbeitung kontinuierlich zu verbessern.[108]

Ein wesentliches Feature von AutoSE ist die stündliche Aktualisierung der EconBiz-Datenbank mit automatisierter Verschlagwortung für neue oder geänderte Metadatensätze, die bestimmte Kriterien erfüllen. Derzeit ist sie auf englischsprachige Metadaten beschränkt, da die Trainingsdaten ausschließlich auf englischsprachigen Modellen basieren, wobei Titel und ggf. Autoren-Keywords berücksichtigt werden. Die Nutzung von Abstracts steht zur Diskussion, ist aber durch rechtliche Einschränkungen bei lizenzierten Ressourcen kompliziert. Seit Juli 2021 ist es möglich, Vorschläge direkt zu übernehmen, was bis September 2022 zur automatisierten Verschlagwortung von über 1,4 Millionen Ressourcen führte. Zusätzlich werden weitere Schlagwörter durch eine Schnittstelle, den Digitalen Assistenten (DA-3), als Vorschläge für eine eventuelle zusätzliche intellektuelle Inhaltserschließung angeboten.[109]

Die Qualitätssicherung von AutoSE besteht aus drei elementaren Ebenen: erstens Information Retrieval mit den bekannten Kennzahlen *recall*, *precision* und dem *F-Maß*, zweitens dem Prinzip des *Human in the Loop*, welches eine intelligente Zusammenarbeit von maschinellen und menschlichen Aktionen vorsieht und drittens regelmäßigen Reviews durch Personen, die in der ZBW-Inhaltserschließung tätig sind.[110]

Die Herausforderung besteht darin, die Balance zwischen der Automatisierung der Erschließungsprozesse und der Sicherstellung der Qualität der Metadaten zu finden. Die stündliche Aktualisierung der EconBiz-Datenbasis und die Fokussierung

106 Vgl. Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht, 2023, S. 4

107 Vgl. Kasprzik, Anna: Automatisierte Inhaltserschließung auf der Basis von Machine-Learning-Methoden als Produktivverfahren an der ZBW, 2020

108 Vgl. Kasprzik, Anna: Ebenda

109 Vgl. Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht, 2023, S. 5

110 Vgl. Kasprzik, Anna: Ebenda, S. 6–9

auf englischsprachige Publikationen zeigen das Bestreben der ZBW, Dokumente mit größtmöglicher Relevanz und Aktualität bereitzustellen.[111]

Im Jahr 2023 präsentierte die ZBW die ersten Ergebnisse ihrer Qualitätssicherungsmaßnahmen. Bis zu diesem Zeitpunkt wurden bereits rund 1,4 Millionen Metadatensätze mithilfe von AutoSE automatisiert verschlagwortet.[112]

Eine der grundlegenden Methoden zur Überprüfung der Qualität dieser automatisierten Verschlagwortung bestand im Abgleich der AutoSE-Vorschläge mit der EconBiz-Datenbasis, das heißt den vorab entnommenen Trainingsdaten. Darüber hinaus führte das ZBW im Jahr 2022 ein spezielles Bewertungstool innerhalb des Digitalen Assistenten DA-3 ein, das spezifisch für die Beurteilung der AutoSE-Vorschläge konzipiert wurde. Dieses Tool bietet den Nutzenden die Möglichkeit, die Qualität der automatisierten Schlagwortvorschläge zu bewerten, wodurch ein direktes Feedback in den Prozess der Qualitätssicherung integriert wird. Diese interaktive Komponente stärkt die Rolle der Nutzenden im Qualitätsmanagement und fördert eine kontinuierliche Verbesserung der automatisierten Erschließungsprozesse.[113]

Die Ergebnisse und das Feedback bestätigen, dass automatisierte Inhaltserschließung eine zentrale und dauerhafte Aufgabe darstellt, die vielseitige und ausreichende Personalressourcen erfordert, einschließlich angewandter Forschung, Softwareentwicklung und IT-Administration unter fachlicher Anleitung. Partnerschaften und Kooperationen sowie die enge Zusammenarbeit aller Abteilungen sind entscheidend für den Erfolg und tragen zur gegenseitigen Verständigung und Verständnisentwicklung für das jeweilige Fachgebiet bei. AutoSE gilt daher als Praxisbeispiel eines sehr erfolgreichen Projekts der automatisierten Inhaltserschließung, das zeigt, wie durch innovative Ansätze und interdisziplinäre Zusammenarbeit die Herausforderungen der modernen Informationsgesellschaft bewältigt werden können.[114]

Gleichzeitig blickt die ZBW mit einer Mischung aus Sorge und Hoffnung auf die Entwicklungen im Bereich der KI und der Deep Learning-Strategien. Diese Technologien bieten zwar ein enormes Potenzial zur Verbesserung und Weiterentwicklung der automatisierten Inhaltserschließung, setzen jedoch annotierte Trainingsdaten und eine stetige Weiterentwicklung der Algorithmen voraus. Ohne diese Grundlagen ist eine effektive Integration von KI-Systemen in den Erschließungsworkflow schwierig.[115]

---

111 Vgl. Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen, 2022, S. 17

112 Vgl. Schmucker, Susanne: Fortschritte der Automatisierung und Kollaboration in der Sacherschließung: Bericht vom 6. Workshop „Computerunterstützte Inhaltserschließung“, 2023, S. 2

113 Vgl. Ebenda, S. 2

114 Vgl. Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht, 2023, S. 11 f.

115 Vgl. Franke-Maier, Michael: Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung: Bericht zur Fachtagung Netzwerk maschinelle Erschließung an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019, 2020, S. 11

4 Vorstellung der Bibliothek des MPI MIS

Der Medienbestand des MPI MIS umfasst derzeit über 66.000 gedruckte Werke, mehr als 150.000 elektronische Bücher und etwa 63.000 lizenzierte Fachzeitschriften und stellt somit eine umfangreiche und vielfältige Sammlung dar.[116] Innerhalb dieser Ressourcen bilden die gedruckten Monographien eine besonders relevante Kategorie, vor allem aufgrund ihrer intensiven Nutzung im Freihandbereich der Bibliothek. Nutzende dieses Kernbestandes sind sowohl die Mitglieder des Instituts selbst als auch externe Personen. Der Sammlungsbeginn für die heute als Eberhard-Zeidler-Bibliothek bekannte Bibliothek erfolgte zeitnah zur Institutsgründung im Jahr 1995. In dieser Anfangsphase lag der Sammlungsschwerpunkt vornehmlich auf neuen Publikationen aus den Disziplinen der Mathematik und den mathematischen Naturwissenschaften. Das Erwerbungsprinzip folgte der Maxime „Just in Case", mit dem Ziel, eine möglichst umfassende Verfügbarkeit wissenschaftlicher Literatur ab dem Gründungsjahr zu gewährleisten, um so den potenziellen Bedürfnissen der Institutsmitglieder jederzeit gerecht werden zu können.

Die Erwerbungspolitik der Bibliothek fokussierte sich primär auf Werke aus den Vereinigten Staaten, dem Vereinigten Königreich und Deutschland, um so die Publikationen der führenden Verlagshäuser im Bereich der Fachliteratur zu erfassen. Zu diesen zählen unter anderem renommierte Namen wie Springer, Oxford University Press, Cambridge University Press, Kluwer, World Scientific, Princeton University Press, Wiley, Birkhäuser sowie die wissenschaftlichen Gesellschaften AMS, EMS und LMS. Konferenzbände unterliegen seit jeher einer differenzierten Erwerbungspolitik, die nicht dem Anspruch auf Vollständigkeit aller fachlich relevanten Monographien folgt. Die Bibliothek wendet hier den „Just in Time"-Ansatz an, wonach Konferenzberichte nur auf explizite Anfrage eines Institutsmitglieds oder bei herausragender fachlicher Bedeutung der jeweiligen Konferenz beschafft werden. Für Werke, die vor dem Gründungsjahr 1995 publiziert wurden, verfolgte die Bibliothek eine selektive Erwerbungsstrategie. In einer gemeinschaftlichen Initiative erstellten die Direktoriumsmitglieder und Forschende des Instituts eine kuratierte Auswahl von Titeln aus vergangenen Jahrzehnten, die als unverzichtbar für die Institutsbibliothek erachtet wurden. Diese Zusammenstellung wurde in den Folgejahren durch gezielte Anfragen und Akquisitionen kontinuierlich ergänzt.[117]

Angesichts des internationalen Nutzerkreises der Bibliothek dominieren englischsprachige Medien den Bestand. Neben Englisch finden sich Werke in einer Reihe weiterer Sprachen unter anderem: Deutsch, Französisch, Russisch und

116 Vgl. Administrative Department: Eberhard Zeidler Library, 2024
117 Vgl. Administrative Department: Eberhard Zeidler Library: [Intranet]

Italienisch. Das spiegelt die globale Ausrichtung und die internationale Relevanz der Institutsbibliothek wider. Die Systematisierung des Bestandes orientiert sich an der Library of Congress Classification (LCC) der Library of Congress (LoC), einem vor allem in internationalen wissenschaftlichen Bibliotheken weit verbreiteten Schema.

Die Verantwortung für die Sacherschließung obliegt dem Bibliotheksleiter, der diese Aufgaben als Teil seiner Vollzeitstelle wahrnimmt. Neben dem Bibliotheksleiter besteht der Personalstamm aus zwei Teilzeitkräften, einer projektbezogenen Teilzeitkraft, einer studentischen Hilfskraft und zwei Aushilfen.

4.1 Nutzung der Library of Congress Classification

Das Notationssystem der LoC zählt zu den weltweit am verbreitetsten Klassifikationsschemata und genießt insbesondere im angloamerikanischen Raum eine herausragende Stellung.[118] Die LoC, welche über die Jahrzehnte als eine der prägendsten Bibliotheken der Vereinigten Staaten von Amerika etabliert wurde, stellt eine zentrale Referenzquelle für bibliothekarische Klassifikationen dar.[119] Darum greifen zahlreiche Bibliotheken weltweit, darunter auch die Bibliothek des MPI MIS, auf die Katalogisierungssysteme und -daten der LoC zurück. Die eigenen Bestände werden dabei insbesondere im Bereich der Monografien mit präzisen und umfassenden Katalogdaten angereichert. Trotz der internationalen Verbreitung und Anerkennung der LCC bleibt ihre Anwendung im deutschen Bibliothekswesen eher die Ausnahme. Aufgrund der außerordentlichen Spezifik und Präzision im Fachbereich Mathematik ist sie jedoch für spezialisierte Forschungsbibliotheken von besonderem Interesse.

Angesichts der Tatsache, dass die Nutzungsbasis der Bibliothek des MPI MIS nahezu ausschließlich aus WissenschaftlerInnen besteht, die oft nur für die Dauer ihrer Forschungsprojekte am Institut verweilen, ist eine international anerkannte Systematik von essenzieller Bedeutung. Die Entscheidung, die LCC als Grundlage für die Organisation des Bibliotheksbestandes zu wählen, trägt somit der hohen Fluktuation innerhalb der wissenschaftlichen Gemeinschaft Rechnung und ermöglicht es den Forschenden, sich rasch und effizient in der bibliothekarischen Sammlung zurechtzufinden.

Diese Orientierung an internationalen Standards spiegelt sich auch in der Bibliothek des Max-Planck-Instituts für Mathematik in Bonn wider, welche ebenfalls die LCC für die Strukturierung ihres Bestandes heranzieht. Durch die Nutzung dieser weitverbreiteten Klassifikation fördern die Max-Planck-Institute nicht nur die internationale Kompatibilität ihrer bibliothekarischen Dienste, sondern unterstützen auch die WissenschaftlerInnen in ihrer transdisziplinären und grenzüberschreitenden Forschungsarbeit, indem sie einen kohärenten und leicht zugängliche Zugriff auf relevante literarische Ressourcen bieten.

118 Vgl. Erschließungswiki: Library of Congress Classification, 2016
119 Vgl. The Library of Congress: Library of Congress Classification, 2023

4.2 Aufbau der Library of Congress Classification

Die LCC ist für ihren Ansatz bekannt, der sich stärker an praktischen Bedürfnissen und weniger an einer strikten Hierarchie orientiert. Diese pragmatische Herangehensweise an die Klassifikation, die sich durch Flexibilität und Anpassungsfähigkeit an die diversen und sich entwickelnden Wissensfelder auszeichnet, könnte als „pragmatische Klassifikation" oder „anwendungsorientierte Klassifikation" bezeichnet werden.[120] Diese Terminologie hebt hervor, dass das System primär darauf abzielt, die Auffindbarkeit und Zugänglichkeit von Materialien zu erhöhen, indem es sich an den konkreten Nutzungsweisen und Bedürfnissen der Bibliotheksnutzenden orientiert, anstatt eine rigide oder ausschließlich theoretisch begründete hierarchische Struktur zu verfolgen.

Die LCC stellt ein deutlich von anderen großen bibliothekarischen Klassifikationssystemen, wie der DDC, abweichendes Notationsschema dar. Die pragmatische Ausrichtung der LCC hat sie zu einem modellhaften Vorbild für andere Klassifikationssysteme gemacht, darunter auch für die im deutschsprachigen Raum besser bekannte RVK. Diese pragmatische Herangehensweise erleichtert insbesondere die Fremddatenübernahme deutscher Publikationen. Diese werden in der LoC, im Gegensatz zur Universitätsbibliothek Regensburg (UB Regensburg), in der zeitlichen Reihenfolge erst zweitrangig erschlossen, weshalb oftmals die RVK-Notation vor der LCC-Notation vorliegt. Das LCC-basierte System des MPI MIS kann die RVK-Notationen entsprechend nutzen und weiterverarbeiten. Diese Praxis der Fremddatenübernahme gewährleistet eine konsistente und effiziente Katalogisierung auch bei Publikationen, die ursprünglich nach anderen Systemen klassifiziert wurden.

Ein charakteristisches Beispiel für die Struktur der LCC ist die Notation QA 279, wobei der Buchstabe *Q* die Hauptklasse *Wissenschaft* symbolisiert. Angesichts der Spezialisierung der Institutsbibliothek auf wissenschaftliche Disziplinen sind Monografien aus anderen Hauptklassen selten. Die Unterklassifizierung innerhalb der Hauptklasse erfolgt durch einen weiteren Großbuchstaben, wie *A* für Mathematik, was die grobe Einordnung in Fachbereiche ermöglicht (QA für Mathematik, QB für Astronomie, QC für Physik usw.). Eine weitere Spezifizierung wird durch eine ein- bis dreistellige Zahl erreicht, die die Thematik innerhalb des Fachbereichs näher bestimmt, wobei auch Dezimalstellen zur Feingliederung hinzugefügt werden können (z. B. QA 501 für allgemeine Werke über Darstellende Geometrie).

Trotz dieser detaillierten Strukturierung, die eine präzise Klassifikation ermöglicht, sieht sich das MPI MIS aufgrund seiner vergleichsweise spezifischen Sammlung der Herausforderung gegenüber, vielen unter- oder sogar unbesetzten Klassen zu begegnen. Um diese Problematik anzugehen, hat die Bibliothek eine Anpassung der klassischen LCC vorgenommen, indem aufeinanderfolgende Klassen mit geringer Bestandsdichte zusammengelegt wurden.[121] Die Entscheidung für diese Zusammenlegung orientierte sich an der Faustregel, dass eine Klasse erst dann

120 Vgl Administrative Department: Eberhard Zeidler Library: [Intranet]

121 Eberhard Zeidler Library: Call Number List: slightly modified Library of Congress System

eigenständig ist, wenn sie mindestens 20 Exemplare enthält bzw. ihr zugeordnet wurden. Dabei wurde stets darauf geachtet thematisch unzusammenhängende Klassen nicht zu weit zu vergröbern, sodass eine zu starke Vermischung der Themengebiete verhindert werden konnte. Diese angepasste, signifikant vereinfachte Klassifikation bildet auch die Grundlage für das im Rahmen dieser Arbeit untersuchte Programm. Weiterhin spiegeln sich die Notationen in der Signatur wider, welche an die RVK angelehnt ist.

5 Programmbeschreibung

5.1 Fachterminologie

5.1.1 Algorithmus

Ein Algorithmus ist in der Informatik und Mathematik als eine präzise Sequenz, die darauf abzielt, ein spezifisches Problem zu lösen oder eine bestimmte Aufgabe zu vollführen.[122] Die charakteristischen Eigenschaften eines Algorithmus umfassen seine Eindeutigkeit, in dem Sinne, dass jede Anweisung klar definiert und frei von Mehrdeutigkeiten sein muss, sowie seine Endlichkeit, die besagt, dass der Algorithmus nach einer begrenzten Anzahl von Schritten zu einem Abschluss kommen muss. Des Weiteren ist ein Algorithmus durch seine Fähigkeit gekennzeichnet, Eingabewerte zu verarbeiten und daraufhin eine Ausgabe oder ein Ergebnis zu produzieren.[123]

Hierbei ist es essenziell, zwischen herkömmlichen algorithmischen Prozessen und KI zu differenzieren. Während Algorithmen klar definierte, schrittweise Anweisungsfolgen zur Problemlösung darstellen, zeichnet sich KI durch die Fähigkeit aus, aus Daten zu lernen, Muster zu erkennen und auf dieser Basis Entscheidungen zu treffen, die über die ursprünglichen Programmierungen hinausgehen.[124]

Im Folgenden wird mit einem selbst entwickelten Algorithmus gearbeitet, dieser führt nur vorher fest definierte Aktionen aus und lernt in keinster Weise dazu bzw. ist nicht fähig Muster von selbst zu erkennen. Weder das Programm noch der Algorithmus selbst greifen auf KI zurück.

5.1.2 Java

Java ist eine weit verbreitete Programmiersprache, die Anfang der 1990er Jahre von Sun Microsystems entwickelt wurde.[125] Das primäre Ziel war, Plattformunabhängigkeit und Netzwerkfähigkeit in der Softwareentwicklung zu fördern.[126] Die Sprache folgt dem Paradigma *Write Once, Run Anywhere* (WORA), was bedeutet, dass wenn Java-Code einmal geschrieben, auf jeder Plattform ausführbar ist, die über eine Java Virtual Machine (JVM) verfügt. Diese Eigenschaft macht Java besonders attraktiv für die Entwicklung von Internet- und Unternehmensanwendungen. Darüber

122 Vgl. Gumm, Heinz-Peter: Programmierung, Algorithmen und Datenstrukturen, 2016, S. 75
123 Vgl. Ebenda, S. 77 ff.
124 Vgl. Bünte, Claudia: Künstliche Intelligenz: die Zukunft des Marketing: ein praktischer Leitfaden für Marketing-Manager, 2018, S. 5
125 Vgl. Java TM: Was ist Java-Technologie, und wozu brauche ich sie?, 2024
126 Vgl. Goll, Joachim: Java als erste Programmiersprache: Grundkurs für Hochschulen, 2016, S. 70 ff.

hinaus wird eine umfangreiche Standardbibliothek geboten, die eine Vielzahl von Hilfsfunktionen für Netzwerkoperationen, Dateizugriffe und Benutzeroberflächen bereitstellt.[127]

5.2 Datenbank des MPI MIS

Ein essentieller Aspekt moderner Bibliothekssysteme ist die Implementierung digitaler Kataloge, basierend auf Datenbanken, welche Titelaufnahmen umfassend verzeichnen und zugänglich machen. Diese Titelaufnahmen werden am MPI MIS in MARC-Datensätzen (Machine-Readable Cataloging) gespeichert, weil dieses Datenformat speziell für die Bedürfnisse bibliothekarischer Katalogisierung entwickelt wurde. MARC-Datensätze ermöglichen eine detaillierte Beschreibung bibliografischer Elemente, sodass jede Titelaufnahme präzise abgebildet und individuell abgefragt werden kann.

Um die Verwaltung und Pflege der Datenbank zu optimieren, sind die Titelaufnahmen in einzelnen Datenbanktabellen organisiert. Diese enthalten jeweils verschiedene Tabellen mit entsprechenden Entitäten (bspw. Monografie, Print-Publikation, Ebook etc.). Diese Strukturierung erlaubt es der Bibliothek nicht nur, spezifische Datensätze effizient zu lokalisieren und abzurufen, sondern bietet auch die Flexibilität, Einträge manuell zu bearbeiten oder zu aktualisieren.

5.3 Funktionsweise des Programms

Die Bibliothek des MPI MIS arbeitet mit einem Verfahren der Fremd- und Bestelldatenübernahme durch den Einsatz individuell entwickelter, langjährig bewährter Softwareprogramme. Diese ermöglichen die Integration von Publikationsdaten in die Datenbanken durch Datenübernahme oder mittels OCR-Erkennung, wobei die Informationen in MARC-Datensätzen mit entsprechenden Unterfeldern abgelegt werden. Diese Erfassung bildet die Datenbasis für das Programm zur Selbstnotation[128], ein Tool, welches automatisch Klassifizierungsnotationen für bislang nicht klassifizierte Publikationen generiert.

Im Prozess der inhaltlichen Erschließung einer Monografie wird zunächst überprüft, ob bereits eine Notation vergeben wurde. Sollte die Monographie bereits kategorisiert worden sein, wird die Notation aus den Fremddaten übernommen. Diese stammt entweder aus der UB Regensburg oder der LoC. Weitere Bibliothekskataloge werden nicht abgefragt.

Auch wenn im späteren Verlauf das Programm gestartet wird, werden falls vorhanden beide Notationen als Zusatz bzw. als Rückversicherung ausgegebenen. Die Zuordnung zu den Sachgruppen der Bibliothek erfolgt in einer vereinfachten Form, da nicht alle spezifischen Untergruppen vorhanden sind.

---

127 Vgl. IBM Deutschland GmbH: Was ist Java?, 2024

128 Eine Darstellung des Programmcodes in Form von Quasi-Code ist in Anhang C einzusehen. Sie dient dem allgemeinen Verständnis, ist aber für eine Replikation ungeeignet

Die ermittelte RVK-Notation wird im MARC-Feld 900 eingetragen. Da die UB Regensburg die Titel tendenziell schneller einarbeitet als die LoC, wird dieses Verfahren primär für Publikationen aus Deutschland angewendet. Der Fokus dieser Arbeit liegt auf englischsprachigen Veröffentlichungen aus englischsprachigen Verlagen, weshalb die Übernahme und Vergabe von RVK-Notationen nicht weiter betrachtet wird.

Falls sowohl eine LCC-, als auch eine RVK-Notation vorhanden sind, wird die LCC bevorzugt behandelt und in Feld 50 eingetragen. Eine anschließende Anpassung an die vergröberten Sachgruppen der eigenen Bibliothek findet automatisiert vor Ort statt. Sollte die spezifische LCC-Sachgruppe nicht exakt übereinstimmen, erfolgt in der Regel die Zuordnung zur Nächsthöheren.

Falls keine der Bibliotheken den Titel bereits erschlossen hat, ist die Bibliothek zur Kategorisierung auf die Expertise des Fachreferenten angewiesen. Wenn keine fachliche Einschätzung eingeholt werden kann, z. B. aufgrund von längerfristiger Abwesenheit, wird die Notation durch die anderen Mitarbeitenden der Bibliothek vergeben. Dies erfordert eine gute Kenntnis des eigenen Bestandes, die Zuordnung ist umso schneller und intuitiver, je geübter die verantwortlichen BibliothekarInnen sind. Um ihnen die Klassifizierung noch weiter zu erleichtern, nutzt das MPI MIS das intern entwickelte Programm zur automatisierten Inhaltserschließung.

Die Initiierung des Programms zur Selbstnotation erfolgt durch die Eingabe des Barcodes. Zuerst wird der zugehörige Sachtitel im Hintergrund aufgerufen und anschließend in seine konstituierenden Stichwörter getrennt.

Diese Zerlegung berücksichtigt Leerzeichen als Trennzeichen, wohingegen Sonderzeichen wie Bindestriche, die typischerweise Komposita verbinden, bewusst als Einheit betrachtet werden. Dieser Ansatz unterscheidet sich von der in anderen Katalogisierungskontexten üblichen Praxis, wo solche Komposita in ihre Bestandteile aufgespalten werden. Ziel des hier beschriebenen Programms ist es jedoch, eine möglichst exakte Notation zu erhalten.

Dafür ist es elementar, die intendierte Bedeutung der Komposita beizubehalten, was schwieriger ist, wenn die Wörter getrennt voneinander verarbeitet werden. Veranschaulicht werden kann dieser Prozess am Beispiel des Kompositums *IT-Systeme*. Normalerweise würde eine Zerteilung in die Bestandteile *IT* und *Systeme* erfolgen, damit sowohl eine Suche nach den einzelnen Wörtern als auch eine Suche nach dem Kompositum zu einer vollständigen Trefferliste führt. In der Bibliothek des MPI MIS wird ein Kompositum vielmehr verschmolzen: *IT-Systeme* wird zusammengeführt zu *ITSysteme*. Die Auffindbarkeit im Katalog erfolgt demnach nur durch eine Suche nach *IT-Systeme* oder *ITSysteme*. Der genaue Verarbeitungsprozess ist bedeutend komplexer, ist aber für das Thema dieser Arbeit nicht näher relevant und wird daher vernachlässigt.

Anschließend wird eine Tabelle mit dem Barcode der Monographie und der Anzahl der Wörter im Sachtitel generiert. Diese Information veranschaulicht die Komplexität des Titels und unterstreicht die Bedeutung jedes einzelnen Wortes, insbesondere in kürzeren Titeln. Der nächste Schritt besteht im Abgleich aller Stichwörter des Titels mit einer vorgefertigten Liste, um relevante von irrelevanten

Stichwörtern zu trennen. Hierbei spielen sogenannte Stoppwörter eine zentrale Rolle, da sie aufgrund ihrer häufigen Verwendung und der geringen inhaltlichen Aussagekraft aus der weiteren Analyse ausgeschlossen werden. Beispiele hierfür sind Artikel (*the; a; an …*), Präpositionen (*at; of; on …*) und Jahreszahlen. Diese Liste wird noch manuell erstellt und muss stetig bearbeitet sowie erweitert werden. Die identifizierten relevanten Stichwörter dienen als Grundlage für die weitere Analyse im Hintergrund des Systems, diese wird durch den eigens für das Programm entwickelten Algorithmus durchgeführt. Er bildet den Kern des gesamten Systems und kommt hier zum Tragen. Daraufhin erfolgt eine Suche in der bibliothekarischen Datenbank nach Titeln, die alle extrahierten relevanten Stichwörter enthalten. Die Reihenfolge dieser Worte ist dabei nicht von Bedeutung. Bei einem Treffer wird dessen zugehörige Sachgruppe ausgegeben und das Programm beendet. Kann kein exakter Treffer ermittelt werden, startet der eigentliche Algorithmus den das Programm beherbergt.

Wird der Algorithmus initiiert startet ein Prozess, in dem schrittweise jeweils ein Stichwort aus der Liste der relevanten Stichwörter des entsprechenden Titels entfernt wird und alle möglichen Kombinationen der verbleibenden Stichwörter mit den vorhandenen Datensätzen abgeglichen werden. Dieser iterative Prozess setzt sich so lange fort, bis entweder ein passender Datensatz gefunden wird oder alle möglichen Stichwortkombinationen ausgeschöpft sind. Sollte es nach Abschluss aller Durchläufe noch immer keine Übereinstimmungen geben, verringert sich die initiale relevante Stichwortmenge um ein Wort und die Durchläufe beginnen wieder.

Ein Durchlauf mit nur einem relevanten Stichwort wird vom Programm im Allgemeinen ausgeschlossen, da die Wahrscheinlichkeit von zu vielen Treffern, die keine Aussagekraft haben, schlichtweg zu hoch wäre. Um den Algorithmus so einfach wie möglich zu veranschaulichen, sei folgendes Beispiel gegeben: Nach dem Durchlaufen der Stoppwortliste sind die relevanten Stichworte des zu klassifizierenden Titels A, B, C und D. Bevor der wirkliche Algorithmus startet, wird zuerst nach Übereinstimmungen für alle vier Stichworte in einer Folge gesucht: A, B, C, D. Es wird keine Schnittmenge gefunden, das heißt der Algorithmus startet. Beim ersten Durchlauf wird nun nach A, B, C gesucht – D entfällt zuerst, da das letzte Stichwort zuerst weggelassen wird. Kommt es zu keiner Übereinstimmung, startet der zweite Durchlauf, bei dem nach A, B, D gesucht wird – C entfällt. Gibt es wieder keine Treffer wird im Dritten Durchlauf nach A, C, D gesucht – B wird weggelassen. Bei keinem Ergebnis startet der letzte Durchlauf wo nach B, C, D gesucht wird und A entfällt. In jedem Durchlauf wird ein Stichwort, von hinten beginnend, entfernt. Es wäre zu komplex, die tatsächliche Wortbedeutung und ihre Relevanz zu bestimmen, dafür ist ein intelligentes System (KI-System) notwendig.

Es ergeben sich nun zwei Möglichkeiten: Entweder es erfolgt ein Treffer aus den bisherigen Datensätzen und es gibt eine Übereinstimmung oder nicht. Wenn die bisherigen Durchläufe ergebnislos bleiben, wird das letzte Wort aus der Stichwortmenge entfernt, im entsprechenden Beispiel handelt es sich dabei um Wort D. Die relevante Stichwortmenge ist nun A, B und C. Im ersten Durchlauf entfällt nun zunächst C und es wird nach Übereinstimmungen für A und B gesucht. Im folgenden

Durchlauf wird nach einer Schnittmenge für die Kombination aus A und C gesucht. Als letztes nach B und C, wenn A entfällt. Bringt das wieder keine Treffermenge werden nun neue Durchläufe eingeleitet, in welchen C vollständig weggelassen wird. Die initiale Stichwortmenge besteht nun aus den Wörtern A, B, D. Sollte der Algorithmus einen oder mehrere übereinstimmende Datensätze identifizieren, wird die Sachgruppe des oder der gefundenen Titel(s) ausgegeben. Bei nur einem Treffer wird die Sachgruppe des Titels ausgegeben, welcher durch die Stichwortsuche des Algorithmus ermittelt wurde. Im Fall von mehreren Treffern ist das weitere Vorgehen davon abhängig, ob sich ihre Sachgruppen unterscheiden oder nicht. Wenn alle Treffer die gleiche Notation haben, wird diese ausgegeben. Bei mehreren Treffern mit verschiedenen Notationen erfolgt die Auswahl der Sachgruppe basierend auf der Anzahl der übereinstimmenden relevanten Stichwörter und der Länge des Titels. Je mehr relevante Stichworte im Titel übereinstimmen, während der Titel an sich möglichst kurz ist, desto besser. Das Verhältnis der relevanten Stichwörter (Anzahl) zur Gesamtanzahl der Wörter im Titel ist entscheidend.

Angenommen, die Treffermenge bestehe aus zwei Titeln, welche beide dieselbe Anzahl an relevanten Stichwörtern enthalten, wird somit der Titel ausschlaggebend für die Notationsvergabe, welcher insgesamt weniger Wörter beinhaltet.

Gibt es keine Treffer nach allen möglichen Stichwortkombinationen bricht der Algorithmus ab, es kann keine Sachgruppe durch das Programm ermittelt werden. Die Ausgabe bleibt demnach leer. Das kann vorkommen, wenn der Titel sehr kurz ist oder zu viele Worte aus der manuell erstellten Stoppwortliste enthält. Da im Rahmen der Durchläufe aber suksessiv relevante Stichwörter entfernt werden, kommt dieser Fall wenig vor und hat in der Untersuchung, welche für diese Arbeit vorgenommen wurde, keinen Bestand.

In idealen Fällen führt der Algorithmus zur Identifizierung einer passenden Sachgruppe, die entweder mit den übernommenen Fremddaten übereinstimmt oder, bei Abweichungen, einer Überprüfung durch FachreferentInnen unterzogen wird, um die adäquateste Zuordnung zu gewährleisten. Sollte keine Sachgruppe ermittelt werden können, werden mögliche Gründe hierfür untersucht (wie Titellänge, Sprache oder Rechtschreibfehler etc.) und gegebenenfalls manuelle Anpassungen auf Basis der Fremddaten vorgenommen.

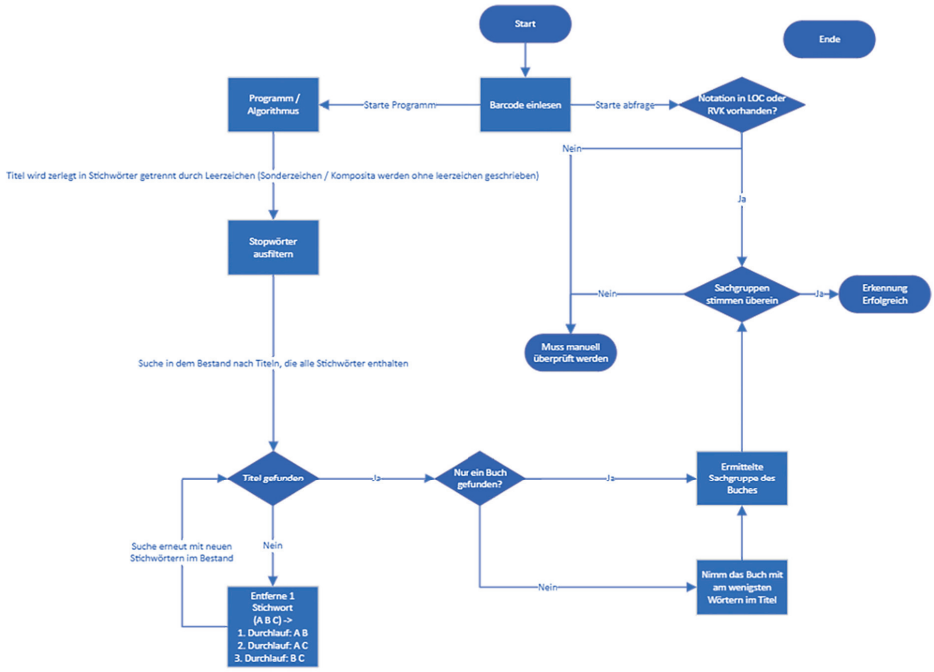

Abb. 1 Visualisierung des Programmablaufs

6 Untersuchung

Die vorliegende Untersuchung zielt darauf ab, die Funktionsweise und Effektivität des vorgestellten Programms und Algorithmus zu analysieren und diese mit der manuellen, intellektuellen Inhaltserschließung zu vergleichen. Dabei konzentriert sich die Studie ausschließlich auf die Vergabe von Notationen, während die Indexierung von Schlagwörtern aufgrund der aktuellen Entwicklungsstufe des Programms noch nicht in Betracht gezogen wird.

Eine zentrale Fragestellung ist, inwieweit die vom Algorithmus generierten Notationen mit den durch menschliche Expertise ermittelten Ergebnissen konform gehen. Erwartet werden differenzierte Ergebnisse hinsichtlich der Übereinstimmung zwischen computerunterstützter und intellektueller Erschließung. Des Weiteren werden die Unterschiede und Gemeinsamkeiten zwischen den Ergebnissen des Programms und denen der befragten Personen eingehend betrachtet. Hierbei ist von Interesse, ob bestimmte Muster oder Tendenzen erkennbar sind, die auf systematische Stärken oder Schwächen in der algorithmischen Vorgehensweise hinweisen. Eine klassische Fehleranalyse, beispielsweise in Form einer Glockenkurve, erweist sich in diesem Kontext als ungeeignet, da der entsprechende Standard für „Richtigkeit“ (Übereinstimmung mit LCC oder RVK) durch die modifizierten Sachgruppen schwierig eingehalten werden kann. Stattdessen wird ein nuancierter Ansatz verfolgt, der die inhärente Subjektivität und Variabilität berücksichtigt, was eine differenziertere Betrachtung der „Fehler“ ermöglicht.

Ein weiterer Untersuchungspunkt sind die Schwierigkeiten, die sowohl für die Personen als auch für das Programm bei der Notationsvergabe auftreten. Letztendlich sollen auch Aussagen zur Frage nach der Qualitätssicherung thematisiert werden, insbesondere hinsichtlich der Durchführbarkeit und Effizienz dieser Prozesse im Kontext zur klassischen manuellen Inhaltserschließung. Es soll eruiert werden, ob die Nutzung des Programms eine Steigerung der Effizienz darstellt.

6.1 Testpersonen

Für die durchgeführte Untersuchung werden hier die Vorkenntnisse und das Fachwissen der verschiedenen befragten Personen kurz zusammengefasst. Sie repräsentieren unterschiedliche akademische und berufliche Hintergründe, was für die Aussagekraft der Ergebnisse entscheidend ist.

Person A verfügt über einen akademischen Abschluss in Mathematik und besitzt rund fünf Jahre Berufserfahrung in diesem Fachbereich. Sie verfügt über ausgeprägte Kenntnisse im Fachbereich Mathematik und besitzt ein gutes Verständnis für den strukturellen Aufbau der Bibliothek des MPI MIS. Das spezifische Wissen bezüglich bibliothekarischer Arbeitsprozesse, insbesondere in den Bereichen der Klassifizierung und Notationsvergabe, ist wenig bis gar nicht vorhanden. Es gibt

keine Erfahrung mit bibliotheksspezifischen Softwareanwendungen sowie den entwickelten Programmen der Bibliothek. Person B hat ein bibliothekarisches Studium absolviert und verfügt über etwa 3 Jahre Berufserfahrung in diesem Bereich. Sie besitzt umfassende Kenntnisse über die inhaltliche und formale Erschließung von Medien. Fachspezifische Kenntnisse in den Bereichen Mathematik und Physik sind nicht bzw. kaum vorhanden (Wissenserwerb über Arbeit in Spezialbibliothek). Person B verfügt über gute Kenntnisse in der Anwendung hauseigener Bibliothekssoftware und hat sehr gute Kenntnisse des Bibliotheksaufbaus. Person C besetzt ein Fachreferat im Fachbereich Physik und besitzt mehr als 20 Jahren Berufserfahrung. Diese Person hat herausragende Kenntnisse im Fachbereich Physik und sehr gute im Bereich Mathematik. Außerdem verfügt sie über ein tiefgreifendes Verständnis für die eigens für die Bibliothek entwickelten Softwareanwendungen. Weiterhin ist die Person sehr erfahren in der bibliothekarischen Erschließung und Notationsvergabe.

Das in der Untersuchung verwendete Programm wurde, wie bereits erwähnt, intern von der Bibliothek entwickelt und basiert auf einem selbst erarbeiteten Algorithmus. Die EntwicklerInnen des Programms verfügen über keinen formalen Abschluss in Informatik oder einem verwandten Feld. Das Programm zeichnet sich durch eine minimalistische Herangehensweise aus, bei der bewusst auf umfangreiche externe Frameworks und librarys verzichtet wird, um eine möglichst unabhängige und einfache Lösung zu schaffen.

6.2 Untersuchungsgegenstand

Die vorliegende Untersuchung bezieht sich ausschließlich auf Neuerwerbungen der Bibliothek im Zeitraum von Anfang Juni bis Ende Oktober 2023. Der Fokus liegt dabei spezifisch auf Monografien, um eine konsistente Basis für die Analyse zu gewährleisten. Die Auswahl von Neuerwerbungen als Untersuchungsgegenstand ist strategisch motiviert, da diese oft Neuerscheinungen darstellen, für die zum Zeitpunkt der lokalen Erschließung noch keine standardisierten Notationen der LoC oder UB Regensburg vorliegen. Dieser Umstand ermöglicht eine unvoreingenommene Betrachtung der Effektivität und Präzision der Notationsvergabe sowohl durch die Testenden als auch durch das Klassifizierungsprogramm. Bei der Sprache der Publikationen handelt es sich um ein weiteres Auswahlkriterium. Es wird sich auf englischsprachige Exemplare beschränkt, da das Programm aktuell ausschließlich mit englischsprachigem Vokabular operiert. Diese Beschränkung spiegelt jedoch auch die Realität des Erwerbungsprozesses im MPI MIS wider, in welchem englischsprachige Publikationen einen signifikanten Anteil der Neuerwerbungen ausmachen. Die Entscheidung, sich im Programm auf englischsprachige Werke zu konzentrieren, reflektiert die Rolle des Englischen als dominierende Wissenschaftssprache, insbesondere an einer wissenschaftlichen Institution wie einem MPI. Serien und Konferenzbände wurden von der Untersuchung ausgeschlossen. Diese Publikationsformate unterliegen speziellen Erschließungspraktiken und werden separat klassifiziert.

6.3 Durchführung

Der Test für die Untersuchung wurde von drei Personen mit unterschiedlichem Fach- und Bibliothekswissen sowie variierenden Kenntnissen zu bibliothekseigenen Programmen durchgeführt. Der Fokus lag auf der Klassifizierung der 97 Neuerwerbungen im Bereich englischsprachiger Monografien.[129] Die Auswahl von 97 Publikationen dient dazu, eine statistisch relevante Datenmenge zu bilden. Diese ermöglicht es signifikante Muster und Tendenzen in den Notationsvergaben zu identifizieren, damit fundierte Schlüsse über die Effizienz und Genauigkeit der verschiedenen Klassifizierungsansätze gezogen werden können. Die Teilnehmenden erhielten unabhängig voneinander die ausgewählten Monografien und eine Notationstafel, welche eine Übersicht über die von der Bibliothek aufgestellten LCC-Sachgruppen zeigt[130]. Diese stark modifizierten Sachgruppen dienten als Leitfaden für die Klassifizierungsentscheidungen. Person A erhielt zusätzlich eine kurze Einführung über die Relevanz von Klassifizierungen, um ein grundlegendes Verständnis für die Aufgabe der Notationsvergabe zu schaffen.

Eine Rahmenzeit in welcher die Vergabe beendet sein sollte, gab es nicht. Grund dafür war zum einen, um die Qualität der Ergebnisse nicht durch zeitlichen Druck zu beeinflussen, zum anderen sollten die Testenden die Bearbeitung problemlos in ihren Arbeitsalltag integrieren können.

Die Nutzung bibliotheksspezifischer Ressourcen (Kataloge, lizenzierte Fachdatenbanken etc.) wurde gestattet, um eine realitätsnahe Arbeitsumgebung zu simulieren. Eine Absprache untereinander war nicht gewünscht. Die vergebenen Notationen wurden von allen Teilnehmenden in einer Tabelle festgehalten. Bis zu diesem Punkt erfolgte noch kein Eintrag in die Datenbank und somit den Katalog. Parallel dazu wurde das Programm zur automatisierten Inhaltserschließung mit den Barcodes der 97 Monografien gefüttert, die Ergebnisse wurden ebenfalls in einer separaten Tabelle verzeichnet. Diese enthielt neben dem Barcode auch den Titel des jeweiligen Werks, um eine klare Zuordnung zu ermöglichen. Schlussendlich wurden alle Daten, sowohl die manuell vergebenen Notationen der Personen A, B und C als auch die vom Programm generierten Klassifizierungen, in einer umfassenden Tabelle zusammengefasst. Diese enthält eine detaillierte Auflistung mit Nummerierung, Barcode, Titel und den Klassifizierungsergebnissen jeder einzelnen Person sowie des Programms. Die Ergebnisse der Untersuchung sind in Anhang A einzusehen.

6.4 Auswertung der Untersuchung

Die durchgeführte Untersuchung ergab, dass in 70 % der Fälle eine Übereinstimmung des computergestützten Programms mit mindestens einer der testenden Personen besteht. In 30 % war keine Übereinstimmung festzustellen. Veranschaulicht wird dieses Gesamtergebnis im Kreisdiagramm der Abbildung 2.

---

129 Die Untersuchung fand im November 2023 statt, zu diesem Zeitpunkt wurden noch keine Notationen von der LoC vergeben.

130 [Eberhard Zeidler Library]: Call Number List: slightly modified Library of Congress System

Dieses Ergebnis kann für einen Prototypen der ersten Generation als durchaus positiv bewertet werden. Trotzdem ist es essentiell tiefergehende Analysen durchzuführen, um die exakte Übereinstimmungsrate der individuellen Klassifizierungen der Personen, basierend auf ihrem jeweiligen Fachwissen, mit den Ergebnissen des Programms zu ermitteln. Besondere Bedeutung kommt zudem den 30 % der Fälle zu, in denen keine Übereinstimmung erzielt wurde. Es ist wichtig genauer zu untersuchen, wo genau diese Diskrepanzen auftraten, um mögliche Fehlerquellen im Programm zu identifizieren.

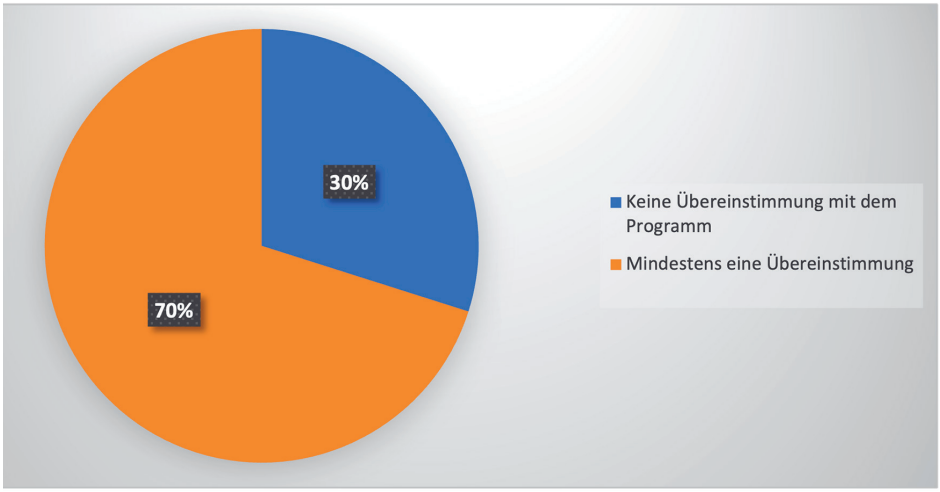

Abb. 2 Übereinstimmung mit dem Programm in Prozent

In einer weiterführenden Analyse wurde das Gesamtergebnis verfeinert, um die Übereinstimmungen zwischen den drei Testpersonen und dem Programm genauer zu beleuchten (vgl. Abb. 3). Dieses Diagramm verdeutlichte erneut, dass in 30 % der Fälle keine Übereinstimmung festgestellt werden konnte. Zudem bietet es eine klarere Darstellung der 70 % der Fälle, in denen mindestens eine Testperson mit dem Programm übereinstimmte.

Die detaillierte Aufschlüsselung dieser 70 % zeigte, dass zu 25 % genau eine der Testpersonen ein mit dem Programm übereinstimmendes Ergebnis erzielte. In weiteren 18 % der Fälle fand sich eine Übereinstimmung zwischen zwei der Testpersonen und dem Programm. Bemerkenswert ist, dass in 27 % der Fälle alle drei Testpersonen und das Programm identische Notationszuordnungen vergaben.

Die Ergebnisse zeigen auf, dass die Anzahl der Publikationen, bei denen keine Übereinstimmung zwischen den Testpersonen und Programm festgestellt wurde, in etwa so groß ist wie die Anzahl der Publikationen, bei denen alle Testpersonen mit dem Programm übereinstimmten.

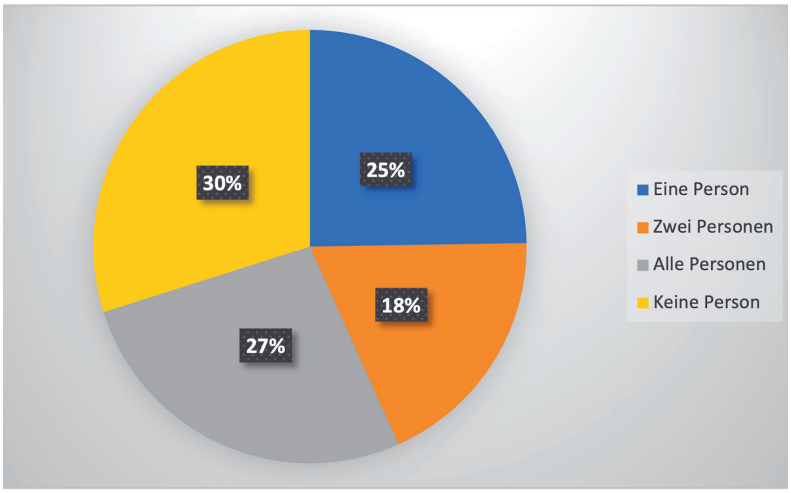

Abb. 3 Anzahl der Übereinstimmungen mit dem Programm in Prozent

Es zeigte sich, dass in 30 % der Fälle, das heißt 29 Publikationen, keine Übereinstimmungen bei der Klassifizierung zwischen den Personen und dem Programm vorlagen. Diese ausgewählten Untersuchungsergebnisse sind in der Tabelle, in Anhang B einzusehen. Von diesen 29 Monografien gab es in 21 Fällen auch bei den von den Testenden vergebenen Notationen keine Übereinstimmung untereinander. Es wurden jeweils drei verschiedene Notationen vergeben. Bei den übrigen acht Publikationen gab es hingegen bei mindestens zwei der drei Testenden eine Übereinstimmung untereinander, was diese Fälle besonders hervorhebt.

Eines dieser Beispiele (Nummer 8: *Physics of Fluids*) ist eine Publikation, bei der alle Testpersonen, trotz ihres unterschiedlichen Fachwissens, einstimmig dieselbe Notationsvergabe tätigten, nur das Programm wich ab. Dies könnte zufällig sein, ist aber dennoch von besonderem Interesse. Die Anzahl der Titelwörter und die relevanten Stichwörter *physics* und *fluids*, welche generisch erscheinen, lieferten keine weiteren Aufschlüsse. Es lässt sich wohl sagen, dass die eindeutige Notation QC durch das Stichwort *physics* klar war.

Weitere Fälle, in denen mindestens zwei Personen miteinander, aber nicht mit dem Programm übereinstimmten, betreffen sieben Publikationen (Nummern: 36, 43, 52, 64, 67, 68, 88). In vier dieser Fälle (36, 64, 68, 88) stimmten die Personen B und C überein, während in den übrigen drei Fällen (43, 52, 67) eine Übereinstimmung zwischen den Personen A und B war. Viele der in der Tabelle dargestellten Monografien zeigen, dass die Notationsvergaben der verschiedenen Personen und des Programms oft sehr ähnlich sind, mit nur geringfügigen Abweichungen in einzelnen Sachgruppen. Dies könnte auf kleinere thematische Differenzen oder Unklarheiten in der inhaltlichen Aufarbeitung der Ressourcen zurückzuführen sein. Eine ausführliche intellektuelle Erschließung wirkt diesen Fehlern in der Regel entgegen, erfordert jedoch ausreichend Zeit und geschultes Personal.

Zudem gab es Fälle in denen die vergebenen Notationen aller Testpersonen und dem Programm stark divergierten (9, 10, 19, 24, 29, 32, 33, 35, 45, 48, 49, 70). Ein möglicher Grund hierfür könnte die erzwungene Beschränkung der Vergabe einer einzigen Sachgruppe sein, wodurch sich die Personen, wie das Programm, entscheiden mussten.

Des Weiteren wurde speziell untersucht, wie oft die einzelnen Personen mit dem Programm übereingestimmt haben (vgl. Abb. 4).

Person A, die einen rein mathematischen Hintergrund besitzt, wies mit 26 % die geringste Übereinstimmung mit dem Programm auf. Person B, die einen rein bibliothekarischen Hintergrund und eine große Vertrautheit mit der Institutsbibliothek aufweist, stimmte mit 42 % am häufigsten mit dem Programm überein. Die Übereinstimmungsrate von Person C, welche über fachliches Wissen und gute Bibliothekskenntnisse verfügt, lag bei 32 %.

Interessant ist es hier ebenfalls den zeitlichen Aufwand mit einzubeziehen, den die einzelnen Personen für die Klassifizierung der 97 Monografien benötigten. Person A benötigte etwa 1,5 bis 2 Stunden, was auf eine sehr effiziente, methodische Herangehensweise hindeutet. Person C gab an, etwa 4 Stunden gebraucht zu haben. Person B, die die höchste Übereinstimmung mit dem Programm erzielte, wendete 10–15 Minuten pro Publikation auf, was in einer Gesamtzeit von etwa 24 Stunden und 15 Minuten resultiert.

Die Korrelation zwischen dem Grad der Übereinstimmung mit dem Programm und dem zeitlichen Aufwand deutet darauf hin, dass eine höhere Übereinstimmungsrate mit einer zeitaufwendigeren Vorgehensweise einhergeht. Es ist jedoch zu beachten, dass Person A und C, bedingt durch ihren hohen Arbeitsaufwand, die Notationen so schnell wie möglich vergeben haben, was ihre geringere Übereinstimmung mit dem Programm teilweise erklären könnte.

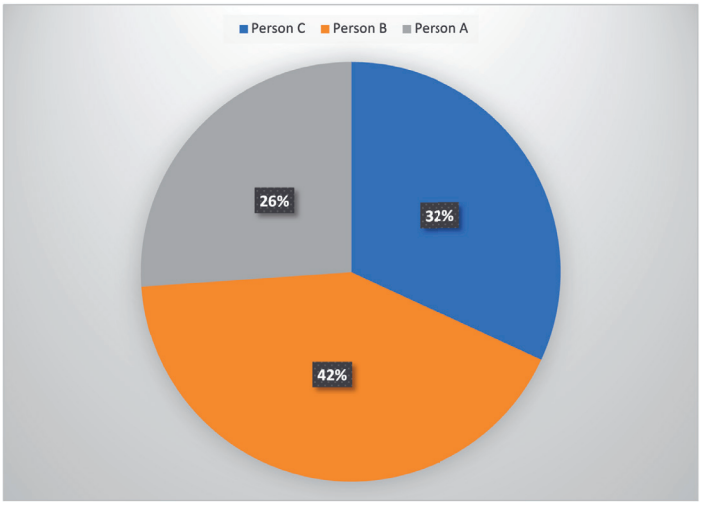

Abb. 4 Übereinstimmungen der Personen mit dem Programm in Prozent

Ebenfalls interessant ist die Betrachtung, wie oft die von den Testenden vergebenen Notationen miteinander übereinstimmen (Abb. 5). Aufschlussreich ist nun, inwiefern das individuelle Vorwissen der Personen zu gegenseitig kongruenten Klassifizierungen führte. Auffällig ist, dass Person A und Person B mit 36 Notationen die meisten Übereinstimmungen aufwiesen, was bei Person A möglicherweise auf ihr gemeinsames fachliches Vorwissen in den Bereichen Mathematik zurückzuführen ist.

Person C wiederum zeigte mit sowohl Person A als auch Person B eine gleich hohe Anzahl von 30 identischen Notationen.

Weiterhin ist erkennbar, dass bei einer Übereinstimmung zwischen den Personen A und C auch das Programm immer dieselbe Notation vergab. Dies könnte ein Indiz für die Zuverlässigkeit des Programms und die Eignung des verwendeten Vokabulars zur Klassifizierung sein, da wie bereits erwähnt nur Person A und Person C über den fachlichen Hintergrund verfügen.

Bei einer Verknüpfung dieser Schlussfolgerungen mit den Ergebnissen des Programms lassen sich verschiedene Vermutungen ableiten. Die hohe Überstimmung der Notationen von Person B mit dem Programm (58 Publikationen) könnte darauf hindeuten, dass das Programm den klassischen Ansatz der intellektuellen bibliothekarischen Inhaltserschließung verfolgt. Es kann weiterhin vermutet werden, dass das Programm durch die vorherige Verarbeitung von Datensätzen zu einzelnen Sachgruppen ein gewisses „Grundverständnis" für die individuelle Aufstellungssystematik der Institutsbibliothek entwickelt hat. Hierbei ist mit „Grundverständnis" die Fähigkeit eines mit Metadaten trainierten Algorithmus gemeint, Muster zu erkennen und Entscheidungen zu treffen, obwohl das Programm ohne KI agiert und somit nicht im eigentlichen Sinne lernen kann. Stattdessen wird das Vorhandensein der Titelstichwörter mit denen in der Metadatenbasis abgeglichen.

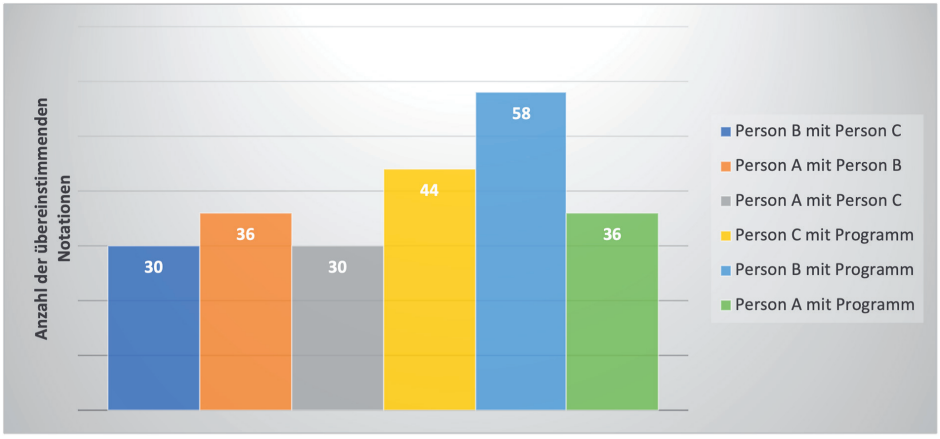

Abb. 5 Absolute Übereinstimmungen der Personen miteinander und mit dem Programm

Die durchgeführte statistische Untersuchung lieferte nicht nur Aufschluss über die Einigkeit der Sachgruppenvergabe durch Testpersonen und Programm, sondern versuchte auch zu verdeutlichen, inwiefern die Stichwortmenge im Publikationstitel von Bedeutung für die Funktionsweise und Effektivität des Algorithmus ist. In der zugehörigen Grafik (Abb. 6) wurde die absolute Häufigkeit des Vorkommens von Titeln mit der Anzahl an Stichworten dargestellt. Durch die Integration der absoluten Häufigkeitswerte kann ein vollständiges und aussagekräftiges Bild der Daten gewonnen werden, welches die Beziehung zwischen der Wortanzahl im Titel und der Genauigkeit der Notationsvergaben durch das Programm klar herausstellt.

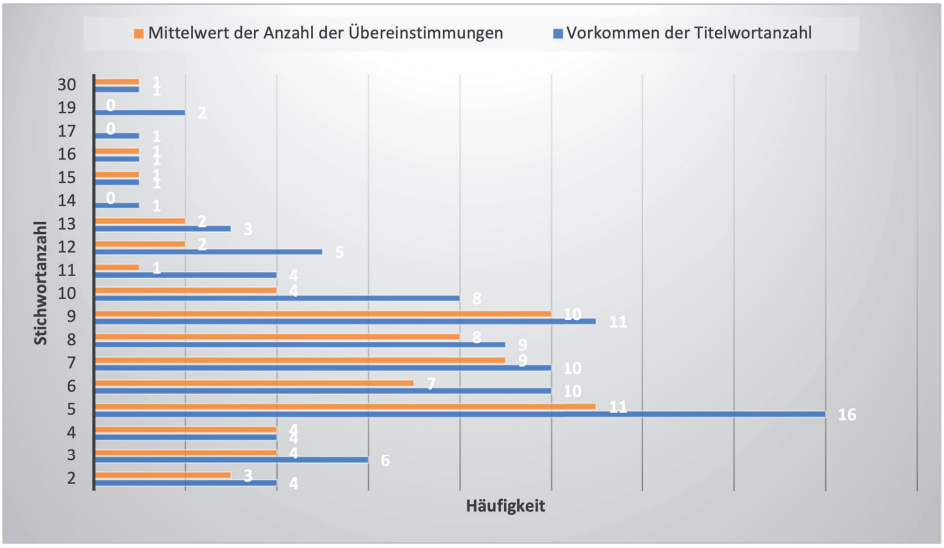

Abb. 6 Absolute Häufigkeit des Vorkommens von Titeln mit der Anzahl an Stichworten

Die Analyse der Daten lässt keine eindeutige Schlussfolgerung zu, ob die Menge der Titelstichwörter einer Publikation einen direkten Einfluss auf die Präzision der Klassifikation durch das Programm hat. Das könnte wiederum darauf hindeuten, dass eine grundlegende Annahme bei der Entwicklung des Algorithmus für das Programm möglicherweise revidiert werden muss. Derzeit fehlen dafür allerdings aussagekräftige Vergleichswerte für Publikationen mit sehr vielen Stichwörtern im Titel, was eine umfassende Bewertung der Situation erschwert. Aus diesem Grund konnte auch nicht rechnerisch ermittelt werden, ob eine empirische Unabhängigkeit vorliegt.

Eine detailliertere Untersuchung mit einer erweiterten Titelauswahl würde belegen, ob eine signifikante Erweiterung der Stoppwortliste sowie eine Verfeinerung des Fachvokabulars bzw. Erweiterung auf Wortgruppen notwendig sind, um die Genauigkeit der Notationsvergabe zu verbessern.

Obwohl die Diagramme interessante erste Einblicke gewähren, erweisen sie sich für die Zwecke dieser Arbeit als nicht aussagekräftig genug, um weitere fundierte Schlussfolgerungen zu ziehen.

Zusammenfassend lässt sich konstatieren, dass die Bewertung der Effektivität des Programms aufgrund der stark variierenden Ergebnisse eine große Herausforderung darstellt, was allerdings bereits im Vorfeld zu erwarten war. Trotz dieser Schwierigkeit nähert sich die Leistung des Programms in einem bemerkenswerten Maße der menschlichen Datenausgabe an. Dies ist besonders hervorzuheben, da der Ansatz für das Programm bewusst simpel gehalten wurde und es sich dabei um einen Prototyp der ersten Generation handelt. Um definitiv feststellen zu können, inwieweit die Wortanzahl im Titel oder die Menge an relevanten Stichworten für die Präzision der Notationsvergabe ausschlaggebend ist, ist eine separate, umfassendere Studie mit einer größeren Untersuchungsmenge erforderlich. Mithilfe der durchgeführten Untersuchung lässt es sich nicht konkret beweisen, dennoch könnte eine Verbesserung der Funktionalität durch eine stärkere Präzisierung des Fachvokabulars erzielt werden, um eine Zuordnung zu mehreren Themengebieten zu vermeiden. Des Weiteren könnten die Einbindung von normierten Termini und ein Verfahren zur Disambiguierung dazu beitragen, eine konsistentere Klassifikation in die vorgesehenen Sachgruppen zu gewährleisten (bspw. Physik: QC).

Die vorliegenden Ergebnisse zeigen, dass Vorkenntnisse im mathematischen Fachbereich oder die Erfahrung in diesem Feld keinen signifikanten Einfluss auf die Ergebnisse hatten. Wird die klassische intellektuelle Erschließung, die viel Zeit in Anspruch nimmt, als Maßstab für eine allgemeingültig *gute* Inhaltserschließung herangezogen, so ist die Übereinstimmung mit dem Programm eine sehr interessante Beobachtung. Das Programm ist in der Lage, ähnliche Aufgaben in deutlich kürzerer Zeit zu bewältigen, wenn auch die Qualität nicht dem gleichen Standard entspricht. Diese Beobachtung unterstreicht das Potenzial automatisierter Klassifikationssysteme, insbesondere im Hinblick auf die Effizienzsteigerung bibliothekarischer Prozesse. Diese ist bisher jedoch noch nicht absehbar, da das erzielte Ergebnis noch zu weit von den durch die Personen vergebenen Notationen abwich und somit nicht dem Qualitätsanspruch genügt.

7 Optimierungsmöglichkeiten

Damit die Präzision und Effektivität der ersten Generation des Programms gesteigert werden können, stehen einige Verfahren und Überlegungen zur Diskussion.

7.1 Vergrößerung des Vokabulars

Eine Möglichkeit wäre das verwendete Fachvokabular innerhalb des Programms zu erweitern. Derzeit basiert der Algorithmus auf einem spezifischen Wörterbuch, das aus den relevanten Stichwörtern jedes untersuchten Titels und den hauseigenen Sachgruppen der Bibliothek besteht. Diese Stichwörter dienen als Grundlage für den Abgleich und die anschließende Klassifizierung der Publikationen. Eine Erweiterung dieses Vokabulars durch spezifischere Fachtermini und Wortgruppen könnte die Genauigkeit der Zuordnungen signifikant verbessern. Eine potentielle Lösung wären die Integration der Mathematical Subject Classification (MSC) für Mathematik[131] und der Physics and Astronomy Classification Scheme (PACS) für Physik.[132]

Ein großes Problem der aktuellen Erschließungspraxis besteht darin, dass zahlreiche Publikationen nicht eindeutig dem Fachbereich oder dem interdisziplinären Bereich der mathematischen Physik zuordnen lassen. Diese Unschärfe führt zu Fehlern bei der eindeutigen Zuweisung der Publikationen in die jeweiligen Fachbereiche. Die Einbindung der MSC- und PACS-Listen könnte für eine inhaltlich präzisere Auswahl sorgen. Die klarere Trennung der Fachbereiche (Mathematik und Physik) würde einen entscheidenden Fortschritt zeigen. Die Listen (MSC und PACS) sind öffentlich als PDF-Dateien zugänglich, was ihre Integration in den bestehenden Code erleichtert. Allerdings ist es wichtig, die Aktualität dieser Klassifikationsschemata zu überprüfen, da die verfügbaren freien Versionen aus den Jahren 2020 bzw. 2008 stammen. Trotzdem würde die Einbindung eine wertvolle Erweiterung des vorhandenen Vokabulars darstellen.

Aufgrund der erwarteten Vorteile ist die Wahrscheinlichkeit hoch, dass diese Anpassung als nächster Schritt in der Weiterentwicklung umgesetzt wird.

7.2 Einbindung weiterer Metadaten

Eine weitere Möglichkeit zur Optimierung betrifft die Verbesserung und Diversifizierung der genutzten Metadaten. Die Diskussionen und Entwicklungsarbeiten dieses Ansatzes sind bereits in vollem Gange, der Fortschritt wird allerdings durch die parallele Bewältigung des Tagesgeschäfts ausgebremst. Der Algorithmus arbeitet aktuell vorrangig durch Stichwörter, die aus den Titeln und Untertiteln der zu

131 Mathematical Reviews: MSC2020-Mathematics Subject Classification System, 2024
132 AIP Publishing: PACS 2010 Alphabetical Index: alphabetical Index, 2010

klassifizierenden Ressourcen extrahiert werden. Eine signifikante Verbesserung könnte jedoch durch die Erweiterung der Stichwortbasis erreicht werden. Wenn relevante Begriffe aus Inhaltsverzeichnissen, Indizes oder sogar aus dem gesamten Volltext der Werke extrahiert werden, könnte das nicht nur die Menge der relevanten Stichwörter erhöhen, sondern auch den inhaltlichen Abgleich mit ähnlichen Publikationen erleichtern und präzisieren. Das setzt allerdings die Verfügbarkeit von Digitalisaten und eine äußerst genaue OCR-Technologie voraus. Selbst wenn die Texte digital vorliegen, bleibt die Herausforderung bestehen, dass das Programm zuverlässig erkennen muss, wo beispielsweise ein Inhaltsverzeichnis beginnt und endet. Die zunehmende Heterogenität von Publikationsformaten erschwert diese Aufgabe zusätzlich und macht eine verlässliche automatisierte Verarbeitung schwierig.

Ein Ansatz ist die Nutzung von Volltexten als Datenbasis. Dies hängt jedoch von der digitalen Verfügbarkeit der Texte ab. Bei vorhandenen Digitalisaten könnte eine Tokenisierung erfolgen, das heißt, dass jedes Wort, das eine bestimmte Häufigkeitsschwelle überschreitet, in die Bewertung einfließt.[133] Eine sehr spezifische Anpassung der Stoppwortlisten ist dafür unabdingbar. Trotz des hohen Potenzials dieser Idee ist die Umsetzung aufgrund von Ressourcenmangel und fehlender IT-Infrastruktur derzeit nicht realisierbar.

Die Nutzung von Registern oder Glossaren als zusätzliche oder alternative Datenquelle stellt eine weitere Möglichkeit dar. Register sind in der Fachliteratur meist homogen aufgebaut und enthalten sowohl Fachtermini als auch Fachwortgruppen, welche für die Klassifizierung relevant sein können. Allerdings besteht hier die Schwierigkeit, dass nicht eindeutig bestimmbar ist, wie relevant ein einzelner Term im Index für die Gesamtpublikation ist, was die Fehleranfälligkeit erhöhen könnte. Als vielversprechendste Idee wird aktuell die Nutzung von Inhaltsverzeichnissen angesehen. In der Fachliteratur beschreiben diese in der Regel präzise den Inhalt der Publikation. Sie könnten durch die Angabe von Seitenzahlen pro Kapitel Anhaltspunkte liefern, wie groß die Relevanz eines einzelnen Themas im Gesamtwerk ist. Die größte Problematik bei diesem Ansatz besteht in der selbstständigen Erkennung des Inhaltsverzeichnisses durch die Software, da die Struktur und der Umfang von Inhaltsverzeichnissen stark variieren können. Eine zuverlässige Lösung für diese Problematik ist entscheidend für die Realisierbarkeit dieses Ansatzes. Entsprechende Lösungen sind intern bereits in der Entwicklung.

7.3 Normalisierung der Termini

Eine allgemeine Möglichkeit zur Optimierung des Vokabulars, das für den Abgleich verwendet wird, ist die Normalisierung der Termini. Dabei handelt es sich eine grammatikalische Anpassung, bei der Wörter auf ihren Wortstamm reduziert werden, sodass beispielsweise *stars* zu *star* konvertiert wird. Die Umsetzung dieser Normalisierung sollte konsistent über das gesamte genutzte Vokabular erfolgen,

133 Vgl. Capterra Inc.: Tokenisierung

sowohl in der Wortliste, die für den Abgleich mit neuen Titeln herangezogen wird, als auch bei den Stichwörtern, die direkt aus den zu klassifizierenden Titeln extrahiert werden. Durch die Reduktion der Termini auf ihre Wortstämme wird die Komplexität des Abgleichprozesses signifikant verringert und die Wahrscheinlichkeit erhöht, dass relevante Übereinstimmungen zwischen den Stichwörtern der Titel und dem vorhandenen Vokabular gefunden werden.

Die Praxis der Termini-Normalisierung wurde auch 2022 von Annif vorgenommen. Ähnliche Methoden der Normalisierung einzuführen kann entweder die Entwicklung eigener Lösungen bedeuten oder die Adaption bestehender Frameworks nach sich ziehen (vgl. Kapitel 3.2).

7.4 Verwendung von Annif

Ein anderer Ansatz könnte die Integration von Annif selbst darstellen. Die Kompatibilität von Annif mit den bereits vorhandenen, eigens entwickelten Programmen und den betriebsinternen Prozessen bedarf jedoch einer genaueren Prüfung. Eine Herausforderung, die sich dabei ergeben könnte, ähnelt einem Problem, mit dem sich die ZBW konfrontiert sah: die Verarbeitung sehr fachspezifischer Termini und das Fehlen ausreichender Vergleichsdaten für eine präzise Klassifizierung.

Die mögliche Kooperation mit Annif würde unweigerlich zusätzliche Ressourcen erfordern, einschließlich Personal und Metadaten. Ob die erforderlichen Ressourcen in naher Zukunft bereitgestellt werden können, ist zum jetzigen Zeitpunkt unklar. Dennoch könnte eine Integration von Annif signifikante Verbesserungen in der Genauigkeit und Effizienz der Inhaltserschließung mit sich bringen.

8 Fazit

Damit die Forschungsfrage, *wie gut das Programm zur automatisierten inhaltlichen Erschließung mittels eines Algorithmus in der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften funktioniert,* umfassend beantwortet werden konnte, war eine gründliche Auseinandersetzung mit den spezifischen Gegebenheiten der Bibliothek unerlässlich. Dafür war es unter anderem wichtig eine allgemeine Übersicht zu erlangen, inwieweit das Thema der computergestützten bzw. automatischen Erschließung in bibliothekarischen Fachkreisen an Bedeutung gewonnen hat. Das zunehmende Publikationsaufkommen, vor allem durch die voranschreitende Digitalisierung bei gleichzeitig schrumpfendem Personalstamm, verdeutlicht die besondere Relevanz dieser erforschten Verfahren. Insbesondere wissenschaftliche Bibliotheken stehen vor der Herausforderung, die inhaltliche Erschließung von Publikationen, die traditionell in den Aufgabenbereich von FachreferentInnen fällt, weiterhin effizient und qualitativ hochwertig zu bewältigen. Aufgrund des Umfangs der zu erschließenden Publikationen und weiterer, zunehmend auf das Personal übergehender Aufgaben, stößt die intellektuelle Erschließung an ihre Grenzen.

Angesichts dieser Problematik verfolgen nicht nur große Einrichtungen wie die ZBW oder die DNB, sondern auch kleinere wissenschaftliche Spezialbibliotheken wie die des MPI MIS Ansätze zur maschinellen Unterstützung der Inhaltserschließung. Hier wurde im Speziellen ein eigens entwickeltes, vergleichsweise einfaches Programm implementiert, das ohne komplexe Softwarelösungen wie Annif oder DA-3 auskommt. Es extrahiert relevante Stichwörter aus dem Sachtitel einer zu erschließenden Publikation. Ein Algorithmus verarbeitet diese entsprechend Daten weiter, um die Notationsvergabe durchzuführen. Die Effizienz und somit die Qualität des Programms zu bewerten, gestaltete sich als äußerst schwierig, da es durch die stark modifizierte Klassifikation der Bibliothek keinen einheitlichen Maßstab für die „Richtigkeit" einer Notation im eigentlichen Sinne gibt. Beim Untersuchen verschiedene Kenngrößen ließ sich allerdings feststellen, dass das Programm in 70 % der Fälle Notationen vergibt, die mit denen einer intellektuellen Inhaltserschließung übereinstimmte. Die größte Übereinstimmung lag bei den Notationen vor, die von der bibliothekarisch ausgebildeten Person vergebenen wurden. Nichtsdestotrotz wies das Programm in 30 % der Fälle keine Übereinstimmungen auf, die vergebenen Sachgruppen gaben nicht das Thema der zu erschließenden Publikationen wieder. Daraus resultiert die Einschätzung, dass das Programm derzeit noch nicht als vollwertiger Ersatz für manuelle Inhaltserschließung geeignet ist.

Es muss dabei kritisch betrachtet werden, dass die Titelauswahl anhand des festgelegten Zeitraums Juni bis Oktober 2023 randomisiert festgelegt wurde. Im Rahmen der Studie hat sich gezeigt, dass der Einfluss der Stichwortmenge ein weiteres interessantes Kriterium ist und in einem nächsten Test mit bedacht werden sollte. Weiterhin wäre es empfehlenswert, die hier ermittelten Notationen zu einem spä-

teren Zeitpunkt mit der intellektuellen Erschließung der LoC abzugleichen. Durch diese Gegenüberstellung könnte ermittelt werden, ob die aktuell vergebenen Notationen tatsächlich den Inhalt der Publikationen widerspiegeln. Auch wenn die durch die Testenden vorgenommenen Ergebnisse bereits einen hohen Qualitätsanspruch erfüllen, können Fehler bei der Erschließung nicht ausgeschlossen werden. Es ist wichtig zu erwähnen, dass die durchgeführte Untersuchung keinen Anspruch auf Vollständigkeit besitzt und nicht repräsentativ die Arbeitsweisen der automatisierten Inhaltserschließung darstellt.

Die Arbeit legt jedoch nahe, dass es verschiedene Möglichkeiten gibt das Programm weiterzuentwickeln. Der vielversprechendste Ansatz resultiert dabei aus einer Kombination der Anpassung des Fachvokabulars und einer vergrößerten Metadatenbasis pro Publikation. Zudem könnte die Integration etablierter Frameworks oder Softwarelösungen wie Annif oder DA-3 zu einer Verbesserung der Ergebnisse führen. Angesichts der hohen Relevanz der Thematik auch in großen bibliothekarischen Einrichtungen erscheint es umso wichtiger, sich frühzeitig mit potenziellen Lösungsansätzen auseinanderzusetzen.

Zusammenfassend ist erkennbar, dass der verfolgte Ansatz der Bibliothek des MPI MIS ein potenzieller Weg zur Bewältigung der Herausforderungen in der modernen Inhaltserschließung darstellen könnte. Trotz der Notwendigkeit weiterer Verbesserungen und Anpassungen war die erzielte vergleichsweise hohe Übereinstimmungsquote mit einem so trivialen Ansatz in der Programmierung ein interessanter Erkenntnisgewinn. Die Forschung und Entwicklung in diesem Bereich bieten auch zukünftig vielversprechende Ansätze, um die Effizienz und Qualität bibliothekarischer Arbeit vor dem Hintergrund wachsender Herausforderungen zu steigern.

Literatur- und Quellenverzeichnis

AIP Publishing: PACS 2010 Alphabetical Index: alphabetical Index / AIP Publishing. – Melville, NY: AIP Publishing, 2010. – 1 Online-Ressource. – https://publishing.aip.org/wp-content/uploads/2019/01/PACS_2010_Alpha.pdf. – Gesehen am 03.04.2024

Beckmann, Regine: Der Digitale Assistent DA-3: eine Plattform für die Inhaltserschließung / Regine Beckmann, Imma Hinrichs, Melanie Janßen, Gérard Milmeister, Peter Schäuble. – 1 Online-Ressource (Seite 1–20). – Gesehen am 03.04.2024 // In: o-bib / Herausgeber: VDB. – ISSN 23639814. – https://doi.org/10.5282/o-bib/2019H3S1–20

Bee, Guido: 5. Workshop Computerunterstützte Inhaltserschließung: Bericht und thematischer Überblick zu den Beiträgen / Guido Bee, Veronika Plößnig, Christoph Steiner. – 1 Online-Ressource (Seite 1–8). – Gesehen am 03.04.2024 // In: o-bib / Herausgeber: VDB. – ISSN 2363-9814. – https://doi.org/10.5282/o-bib/5788

Börsenblatt: Was Maschinen können und was nicht: automatisierte Inhaltserschließung in der Deutschen Nationalbibliothek / von Börsenblatt. – 1 Online-Ressource. – Gesehen am 03.04.2024 // In: Börsenblatt / Herausgeber: Börsenverein des Deutschen Buchhandels e. V. – ISSN 1611-4280. – https://www.boersenblatt.net/archiv/1361205.html

Bünte, Claudia: Künstliche Intelligenz: die Zukunft des Marketing: ein praktischer Leitfaden für Marketing-Manager / Claudia Bünte. – Wiesbaden: Springer Gabler, 2018. – 1 Online-Ressource (42 Seiten). – (essentials). – https://doi.org/10.1007/978-3-658-23319-8_2. – Gesehen am 03.04.2024. – ISBN 978-3-658-23319-8

Busse, Frank: Erschließungsmaschine gestartet / von Frank Busse, Claudia Grote, Jan-Helge Jacobs, Matthias Nagelschmidt, Christoph Poley, Sandro Uhlmann, Kirsten Tröller, Nico Wagner. – 1 Online-Ressource. – Gesehen am 03.04.2024 // In: Blog / DNB. – ISSN 27513238. – https://blog.dnb.de/erschliessungsmaschine-gestartet/

Capterra Inc.: Tokenisierung / Capterra Inc. – München: Capterra Inc, [o. J.]. – Online-Ressource. – https://www.capterra.com.de/glossary/415/tokenization. – Gesehen am 03.04.2024

Ceynowa, Klaus: Deutsche Nationalbibliothek: in Frankfurt lesen jetzt zuerst Maschinen / von Klaus Ceynowa. – 1 Online-Ressource. – gesehen am 03.04.2024 // In: Frankfurter Allgemeine Zeitung FAZ.NET / Herausgegeben von Gerald Braunberger, Jürgen Kaube, Carsten Knop, Berthold Kohler. – ISSN 0174-4909. – https://www.faz.net/aktuell/feuilleton/buecher/maschinen-lesen-buecher-deutsche-nationalbibliothek-setzt-auf-technik-15128954.html

DA-3: Der Digitale Assistent DA-3 ist eine Plattform zur Unterstützung der inhaltlichen Erschließung in Bibliotheken / Bibliotheksservice-Zentrum Baden-Württemberg (BSZ). – Konstanz, Bibliotheksservice-Zentrum, 2022–. – Online-Ressource. – https://www.da-3.de/. – Gesehen am 03.04.2024

Deutsche Nationalbibliothek: Automatisches Erschließungssystem: Projekt inhaltliche Erschließung mit KI / Deutsche Nationalbibliothek. – Frankfurt am Main: Deutsche Nationalbibliothek, 2022–. – Online-Ressource. – https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html. – Gesehen am 03.04.2024

Deutsche Nationalbibliothek: Deutsche Nationalbibliografie / Deutsche Nationalbibliothek. – Frankfurt am Main: Deutsche Nationalbibliothek, 2023–. – Online-Ressource. – https://www.dnb.de/DE/Professionell/Metadatendienste/Metadaten/Nationalbibliografie/nationalbibliografie_node.html. – Gesehen am 03.04.2024

Eberhard Zeidler Library: Administrative Department: Eberhard Zeidler Library: [Intranet] / MPI für Mathematik in den Naturwissenschaften. – Leipzig: Max-Planck-Gesellschaft, 2024–. – Online-Ressource. – Gesehen am 28.03.2024
Im Besitz der Verfasserin

Eberhard Zeidler Library: Call Number List: slightly modified Library of Congress System / Eberhard Zeidler Library. – [o. J.]. – Leipzig: MPI für Mathematik in den Naturwissenschaften–. – Online-Ressource. – https://fileswww.mis.mpg.de/mpitypo3/Institute/Services/Library/Library-Call-Number-List.pdf. – Gesehen am 03.04.2024

Erschließungswiki: Library of Congress Classification / von Studierenden der HföD, Fachbereich AuB, Fachrichtung Bibliothekswesen. – WikiMedia: WikiMedia default, 2016–. – Online-Ressource. – https://wiki.fachbereich-aub.de/wiki/index.php/Library_of_Congress_Classification. – Gesehen am 03.04.2024

Eurospider Information Technology AG: Online Workshop „Computerunterstützte Inhaltserschließung" am 14./15. November 2023 / by Eurospider Information Technology AG. – Zürich: Eurospider Information Technology AG, 2023–. – Online-Ressource. – https://www.eurospider.com/de/relevancy-produkt/digitaler-assistent-da-3. – Gesehen am 03.04.2024

Franke-Maier, Michael: Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung: Bericht zur Fachtagung Netzwerk maschinelle Erschließung an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019 / Michael Franke-Maier, Cyrus Beck, Anna Kasprzik, Jan Frederik Maas, Sarah Pielmeier, Heidrun Wiesenmüller. – 1 Online-Ressource (Seite 1–12). – Gesehen am 03.04.2024 // In: obib / Herausgeber: VDB. – ISSN 23639814. – https://www.o-bib.de/bib/article/view/5565

Goll, Joachim: Java als erste Programmiersprache: Grundkurs für Hochschulen / Joachim Goll, Cornelia Heinisch. – 8., überarbeitete Auflage. – Wiesbaden: Springer Vieweg: 2016. – 1 Online-Ressource (XIV, 903 Seiten). – https://doi.org/10.1007/978-3-658-12118-1_3. – Gesehen am 03.04.2014. – ISBN 978-3-658-12118-1

Gumm, Heinz-Peter: Programmierung, Algorithmen und Datenstrukturen / Heinz-Peter Gumm, Manfred Sommer. – Berlin/Boston: De Gruyter Oldenbourg, 2016. – 1 Online-Ressource (450 Seiten). – (Informatik ; Band 1). – https://doi.org/10.1515/9783110442267. – Gesehen am 03.04.2024. – ISBN 9783110442267

IBM Deutschland GmbH: Was ist Java? / IBM. – Ehningen: IBM Deutschland GmbH, 2024–. – Online-Ressource. – https://www.ibm.com/de-de/topics/java. – Gesehen am 03.04.2024

Java TM: Was ist Java-Technologie, und wozu brauche ich sie? / Java TM. – Redwood Shores: Oracle, 2024–. – Online-Ressource. – https://www.java.com/de/download/help/whatis_java.html. – Gesehen am 03.04.2024

Junger, Ulrike: Neue Wege und Qualitäten: die Inhaltserschließungspolitik der Deutschen Nationalbibliothek / Ulrike Junger und Frank Scholze. – 1 Online-Ressource (Seite 55–70). – Gesehen am 03.04.2024 // In: Qualität in der Inhaltserschließung / Herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann. – (Bibliotheks- und Informationspraxis ; Band 70). – Berlin: De Gruyter Saur, 2021. – ISBN 9783110691597. – https://doi.org/10.1515/9783110691597-004

Kanade, Vijay: What is a support vector machine?: working, types, and examples / Vijay Kanade. –Last updated: September 29, 2022. – Spiceworks Inc., 2024-. – Online-Ressource. – https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/. – Gesehen am 03.04.2024

Kasprzik, Anna: Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW: ein Status- und Erfahrungsbericht / Anna Kasprzik. – 1 Online-Ressource (Seite 1–13). – Gesehen am 03.04.2024 // In: obib / Herausgeber: VDB. – ISSN 23639814. – https://doi.org/10.5282/o-bib/5903

Kasprzik, Anna: Automatisierte Inhaltserschließung auf der Basis von Machine-Learning-Methoden als Produktivverfahren an der ZBW / Anna Kasprzik. – Hannover: Technische Informationsbibliothek (TIB), 2020. – 1 Online-Ressource (35:09 Minuten). – (Schweitzer Forum: Digitale Bibliotheken 2020 ; 4). – Gesehen am 03.04.2024. – https://doi.org/10.5446/50207

Lepsky, Klaus: B 3 Automatisches Indexieren / Klaus Lepsky. – 1 Online-Ressource (Seite 171–182). – Gesehen am 03.04.2024 // In: Grundlagen der Informationswissenschaft / Rainer Kuhlen. – 7. Auflage. – Berlin: De Gruyter Saur, 2023. – ISBN 9783110769043. – https://doi.org/10.1515/9783110769043-015

The Library of Congress: Library of Congress Classification / The Library of Congress. – Last updated: 12/19/2023. – Washington DC: Library of Congress, 2023–. – Online-Ressource. – https://www.loc.gov/catdir/cpso/lcc.html. – Gesehen am 03.04.2024

Mathematical Reviews: MSC2020-Mathematics Subject Classification System / Associate Editors of Mathematical Reviews and zbMATH. – zbMATH, 2024. – 1 Online-Ressource (224 Seiten). – https://zbmath.org/static/msc2020.pdf. – Gesehen am 03.04.2024

Mecking, Elizabeth: Automatisierte Vergabe von Notationen der Regensburger Verbundklassifikation RVK mithilfe des Toolkits Annif / Elizabeth Mecking. – Wildau: Technische Hochschule Wildau Fachbereich Wirtschaft, Informatik, Recht, 2023. – 1 Online-Ressource (74 Seiten). – https://opus4.kobv.de/opus4-th-wildau/files/1845/Masterarbeit_Mecking_2023_final.pdf. – Gesehen am 03.04.2024. – Masterarbeit, Technische Hochschule Wildau, 2023

MPI für Mathematik in den Naturwissenschaften: Administrative Department: Eberhard Zeidler Library / MPI für Mathematik in den Naturwissenschaften. – Leipzig: Max-Planck-Gesellschaft, 2024–. – Online-Ressource. – https://www.mis.mpg.de/de/services/library. – Gesehen am 03.04.2024

Mödden, Elisabeth: Annif gemeinsamer Werkzeugkasten / Elisabeth Mödden, Christoph Poley. – Frankfurt am Main: Deutsche Nationalbibliothek, 2020. – 1 Online-Ressource (15 Folien). – https://wiki.dnb.de/download/attachments/181751388/4-1_Gemeinsamer%20Werkzeugkasten_2020-12-04_final.pdf?version=1&modificationDate=1607353027000&api=v2. – Gesehen am 03.04.2024

Mödden, Elizabeth: Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek: breiter Sammelauftrag stellt hohe Anforderungen an die Algorithmen zur statistischen und linguistischen Analyse / Elizabeth Mödden, Christa Schöning-Walter, Sandro Uhlmann. – 1 Online-Ressource (Seite 30–35). – Gesehen am 03.04.2024 // In: BuB – Forum Bibliothek und Information / BIB. – ISSN 18691137. – https://nbn-resolving.org/urn:nbn:de:0290-opus4-160844

Mödden, Elisabeth: KI im Einsatz für die inhaltliche Erschließung – ein Erfahrungsbericht aus der Deutschen Nationalbibliothek / Elisabeth Mödden. – Hannover: Technische Informationsbibliothek (TIB), 2020. – 1 Online-Ressource (22:37 Minuten). – (#vBIB20; 90). – Gesehen am 03.04.2024. – https://doi.org/10.5446/36428

National Library of Finland: Annif / National Library of Finland. – Helsinki: National Library of Finland, 2024–. – Online-Ressource. – https://www.kansalliskirjasto.fi/en. – Gesehen am 03.04.2024

National Library of Finland: Annif: Tool for automated subject indexing and classification / National Library of Finland. – Helsinki: National Library of Finland, 20202023–. – Online-Ressource. – https://annif.org/. – Gesehen am 03.04.2024

Schmucker, Susanne: Fortschritte der Automatisierung und Kollaboration in der Sacherschließung: Bericht vom 6. Workshop „Computerunterstützte Inhaltserschließung“ / Susanne Schmucker, Cornelia Schöntube. – 1 Online-Ressource (Seite 1–11). – Gesehen am 03.04.2024 // In: o-bib / Herausgeber: VDB. – ISSN 23639814. – https://doi.org/10.5282/o-bib/5928

Schöning-Walter, Christa: Einführung in das Projekt PETRUS: PETRUS-Workshop am 21. und 22. März 2011 / Christa Schöning-Walter. – Frankfurt am Main: Deutsche Nationalbibliothek, 2011. – 1 Online-Ressource (18 Folien). – https://files.dnb.de/pdf/petrus/einfuehrung_petrus_schoening-walter.pdf. – Gesehen am 03.04.2024

Schöning-Walter, Christa: PETRUS – Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek / Christa Schöning-Walter. – 1 Online-Ressource (Seite 15–19). – Gesehen am 03.04.2024 // In: Dialog mit Bibliotheken / Deutsche Nationalbibliothek. – ISSN 09361138. – https://files.dnb.de/pdf/petrus/dialog_2010_1_petrus.pdf

Suominen, Osma: Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing/ by Osma Suominen, Ilkka Koskenniemi. – 1 Online Ressource. – Gesehen am 03.04.2024 // In: code{4}lib Journal, Issue 54 / code{4}lib Journal – ISSN 19405758. – https://journal.code4lib.org/articles/16719

Suominen, Osma: Annif: DIY automated subject indexing using multiple algorithms / Osma Suominen. – 1 Online-Ressource (Seite 1–25). – Gesehen am 03.04.2024 // In: Liber Quarterly / LIBER. – ISSN 2213056X. – https://doi.org/10.18352/lq.10285

Tréfás, David: Das Fachreferat: vom Universalgelehrten zur Schwarmintelligenz / David Tréfás. – 1 Online-Ressource (Seite 864–874). – Gesehen am 03.04.2024 // In: Bibliotheksdienst – ISSN 21949646. – https://doi.org/10.1515/bd-2018-0103

Anhang

Anhang A Untersuchungsergebnisse

Nr.	Bücher	Titel	Person A	Person B	Person C	Programm
1	00699570	Algebraic Foundations for Applied Topology and Data Analysis	QA 612	QA 611	QA 612	QA 612
2	00699587	Mathematical Principles of Topological and Geometric Data Analysis	QA 611	QH 324	QA 440	QA 611
3	00692887	The Nature of X-Rays and Their Interactions with Matter	QC 474	QC 474	QC 474	QC 474
4	00700009	Weak Convergence and Empirical Processes With Applications to Statistics	QA 276	QA 274	QA 274	QA 274
5	00699884	Understanding the Path from Classical to Quantum Mechanics	QC 174.17	QC 173.96	QC 174	QC 174.12
6	00700085	Group Theory Finite Discrete Groups and Applications	QA 171	QA 184	QA 150	QA 171
7	00699914	Boundary Value Problems and Hardy Spaces for Elliptic Systems with Block Structure	QA 379	QA 323	QA 377	QA 377
8	00699921	Physics of Fluids	QC 150	QC 150	QC 150	QC 120
9	00699808	Bifurcation Analysis of Fluid Flows	QA 374	QA 911	QC 20.7.17	QA 440
10	00699754	Maurer-Cartan methods in deformation theory: the twisting procedure	QA 801	QA 169	QA 251	QC 20.7.H
11	00699860	Complex Function Theory	QA 331.7	QA 329	QA 330	QA 331
12	00699983	Introducing String Diagrams The Art of Category Theory	QA 8	QA 169	QA 169	QA 169
13	00699839	Introduction to nonlinear control: stability, control design, and estimation	QA 427	QA 402.5	QA 403	QA 402
14	00699815	What determines an alge-braic variety?	QA 564	QA 564	QA 564	QA 564

15	00699785	How to Derive a Formula 2. Further Analytical Skills and Methods for Physical Scientists	QA 401	QC 20	QC 19	QA 40
16	00699907	Discrete-Time Semi-Markov Random Evolutions and Their Applications	QA 274	QA 274	QA 276	QA 274
17	00699792	Topology and Geometry of Intersections of Ellipsoids in R^π Series of Comprehensive Studies in Mathematics	QA 612	QA 440	QA 564	QA 440
18	00700146	Prescribing Scalar Curvature in Conformal Geometry	QA 801	QA 440	QA 440	QA 440
19	00699938	Adversarial Learning and Secure AI	QA 76.87	Q 190	QA 78	Q 295
20	00699853	Partial Differential Equations	QA 377	QA 374	QA 372	QA 374
21	00699822	Inverse problems and data assimilation	QA 371.5.D37	QA 378	QA 278	QA 378
22	00700160	Maximal subellipticity	QA 329	QA 377	QA 378	QA 377
23	00699761	Physics of binary star evolution: from stars to x-ray binaries and gravitational wave sources	QC 801	QB 980	QB 400	QB 495
24	00698603	SAS Essentials: Mastering SAS for data analytics.	QA 276	QA 276	QA 276	QA 276
25	00699549	An Introduction to Statistical Learning with Applications in Python	QA 276	QA 276	QA 276	QA 276
26	00699488	Statistical Inference Based on Kernel Distribution Function Estimators	QA 276	QA 276	QA 274	QA 276
27	00700016	ANOVA with Dependent Errors	QA 278	QA 279	QA 276	QA 279
28	00699679	Sparse graphical modeling for high dimensional data: a paradigm of c	QA 276	QA 279	QA 166	QA 402
29	00699457	Bayesian filtering and smoothing	QA 276	QA 279	QA 274	QA 279

30	00699358	Time series for data scientists: data management, description, model	QA 279	QA 279	QA 276	QA 279
31	00699709	Generalized Notions of Continued Fractions Ergodicity and Number Theoretic applications	QA 313	QA 281	QA 211	QA 295
32	00700092	Algebra and number theory: a selection of highlights	QA 150	QA 281	QA 251	QA 295
33	00698634	Six pillars of calculus: Business edition	QA 303	QA 303	QA 300	QA 303
34	00699693	Constructive analysis of semicircular elements: from orthogonal projections to semicircular elements	QA 326	QA 329	QA 320	QA 9
35	00698771	Conductive Homogeneity of Compact Metric Spaces and Construction of p	QA 320	QA 329	QA 329	QA 183
36	00699716	More Explorations in Complex Functions	QA 331.7	QA 331	QA 350	QC 173
37	00699556	A Friendly Approach to Complex Analysis	QA 331.7	QA 331	QA 350	QA 808
38	00699105	Metrical Almost Periodicity and Applications to Integro-Differential E	QA 371	QA 431	QA 351	QA 351
39	00699242	Numerical Analysis of Ordinary and Delay Differential Equations	QA 378	QA 297	QA 371	QA 371
40	00699181	Kernel Determination Problems in Hyperbolic Integro-Differential Equat	QA 351	QA 377	QA 431	QA 377
41	00699464	Introduction to Harmonic Analysis	QA 403	QA 403.S24	QA 403	QA 403
42	00699228	Introduction to Methods for Nonlinear Optimization	QA 427	QA 427	QA 430	QA 321
43	00699334	Essays in geometrydedicated to Norbert A.Campo	QA 440	QA 440	QA 440	QA 440
44	00698559	A Variational Theory of Convolution-Type Functionals	QA 374	QA 564	QA 377	QA 601
45	00699440	The Calabi Problem for Fano Threefolds	QA 613	QA 564	QA 564	QA 564

46	00698672	Algebraic Geometry between Tradition and Future An Italian Perspective	QA 564	QA 564	QA 564	QA 564
47	00698566	Twisted Isospectrality, Homological Wideness, and Isometry A Sample of Algebraic Methods in Isospectrality	QA 150	QA 179	QA 613	QA 564
48	00698702	Linear Fractional Transformations: An Illustrated Introduction	QA 440	QA 911	QA 564	QA 601
49	00698641	Introduction to Smooth Ergodic Theory	QA 313	QA 611	QA 611	QA 611
50	00698740	Minimal Surfaces through Nevanlinna Theory	QA 641	QA 641	QA 640	QA 641
51	00699198	Analytical Mechanics Classical, Lagrangian and Hamiltonian Mechanics, Stability Theory, Special Relativity	QA 801	QA 801	QA 800	QC 173
52	00699341	Mechanics of Fluids	QA 911	QA 801	QA 800	QA 801
53	00699990	Elementary Classical Mechanics	QC 120	QA 801	QC 120	QA 801
54	00699976	Elementary Classical Mechanics Problems and Solutions	QC 120	QA 801	QC 120	QA 801
55	00699129	Spectral flow: a functional analytic and index-theoretic approach /	QA 320	QA 911	QA 613	QA 911
56	00698610	From Kinetic Theory to Turbulence Modeling: The Legacy of Carlo Cercig	QC 174.9	QC 19.2	QC 19	QH 323
57	00698719	Computational Physics With Worked Out Examples in FORTRAN and MATLAB	QC 21	QC 51	QC 50	QC 51
58	00699433	Quantum Mechanics – Axiomatic Approach and Understanding Through Mathematics	QC 173.96	QC 174.17	QC 173	QC 174.17

59	00699891	Quantum Field Theory and Functional Integrals An Introduction to Feynman Path Integrals and the Foundations of Axiomatic Field Theory	QC 174.45 A1	QC 174.2	QC 174	QC 174.45
60	00698375	The Limiting Absorption Principle for Massless Dirac Operators, Properties of Spectral Shift Functions, and an Application to the Witten Index of Non-Fredholm Operators	QA 329	QC 174.45	QC 174.5	QC 174.5
61	00699365	Statistical mechanics: theory and molecular simulation	QC 174.7	QC 174.7	QC 174.7	QC 174.7
62	00699389	A student's guide to the Ising model	QC 750	QC 174.85	QC 174.7	QC 173.7
63	00699235	Field Theoretic Method in Phase Transformations	QC 120	QC 174.9	QC 174.9	QC 175
64	00700108	Atoms, Radiation, and Radiation Protection	QC 770	QC 474	QC 750	QC 770
65	00699150	An Introduction to the Mathematical Fluid Dynamics of Oceanic and Atmospheric Flows	QC 801	QC 801	QC 150	QC 809
66	00699471	Complete Guide to the Fragment Molecular Orbital Method in GAMESS From One Atom to a Million, at your Service	QC 770	QD 450	QD 450	QD 461
67	00699204	Patterns and Interfaces in Dissipative Dynamics Revised and Extended, Now also Covering Patterns of Active Matter	QC 150	Q 172	Q 350	Q 172
68	00699273	European Congress of Mathematics. Portoroz, 20–26 June 2021	QA 36	QA 9	QA 7	QA 1
69	00700115	Artificial Intelligence Foundations of Computational Agents	QA 76.87	QA 76	QA 78	QA 76

70	00698443	Combinatorics, Modeling, Elementary Number Theory From Basic to Advanced	QA 164	QA 164	QA 164	QA 164
71	00699747	Enumerative Combinatorics	QA 164	QA 164	QA 164	QA 164
72	00698580	Star-Critical Ramsey Numbers for Graphs	QA 166	QA 166	QA 166	QA 166
73	00699112	Nonlocal Continuum Limits of p-Laplacian Problems on Graphs	QA 166	QA 166	QA 166	QA 166
74	00699402	Discrete Quantum Walks on Graphs and Digraphs	QA 166	QA 166	QA 166	QA 166
75	00698627	Domination in Graphs: Core Concepts	QA 166	QA 166	QA 166	QA 166
76	00699778	Algebraic groups and number theory	QA 241	QA 171	QA 247	QA 247
77	00699662	Euclidean Buildings	QA 440	QA 171	QA 171	QA 171
78	00698696	Linear Algebra A Minimal Polynomial Approach to Eigen Theory	QA 184	QA 184	QA 184	QA 184
79	00699501	Linear Algebra in Action	QA 184	QA 184	QA 184	QA 184
80	00698436	Matrix mathematics: A second course in linear algebra.	QA 184	QA 184	QA 184	QA 184
81	00699280	Purely Arithmetic PDEs Over a p-Adic Field: (Se(B-Characters and (Se(B-Modular Forms	QA 374	QA 241	QA 241	QA 241
82	00699297	An Introduction to the Circle Method	QA 36	QA 241	QA 241	QA 241
83	00699426	Number theory through the eyes of Sophie Germain: an inquiry course	QA 241	QA 241	QA 241	QA 241
84	00699082	Analytic Number Theory for Beginners	QA 241	QA 241	QA 241	QA 241
85	00699174	Field Arithmetic	QA 150	QA 247	QA 246	QA 241
86	00698856	Iwasawa Theory and Its Perspective	QA 171	QA 247	QA 247	QA 273
87	00698832	Algebra and Coding Theory	QA 150	QA 267	QA 252	QA 252
88	00700078	Game Theory A Nontechnical Introduction to the Analysis of Strategy	QA 269	QA 269	QA 269	QA 269

89	00698665	Concentration and Gaussian Approximation for Randomized Sums	QA 274	QA 273	QA 273	QA 273
90	00699310	Stochastic calculus of variations: for jump processes	QA 274	QA 274	QA 274	QA 274
91	00699532	Stochastic Mechanics The Unification of Quantum Mechanics with Brownian Motion	QA 274	QA 274	QC 20.7.S8	QA 274
92	00698733	Random Walks on Infinite Groups	QA 274	QA 274	QA 274	QA 274
93	00698481	Bayesian Non-linear Statistical Inverse Problems	QA 276	QA 274	QA 274	QA 274
94	00690386	Stochastic Partial Differential Equations with Additive Gaussian Noise	QA 374	QA 274	QA 274	QA 274
95	00699969	Non-Gaussian Selfsimilar Stochastic Processes	QA 274	QA 274	QA 274	QA 274
96	00693881	The geometry of cubic hypersurfaces	QA 564	QA 564	QA 564	QA 564
97	00692997	Rectifiability: a surey	QA 615	QA 615	QA 615	QA 615

Anhang B Ausgewählte Untersuchungsergebnisse

Nr.	Bücher	Titel	Person A	Person B	Person C	Programm	Anzahl Wörter
8	699921	Physics of Fluids	QC 150	QC 150	QC 150	QC 120	3
9	699808	Bifurcation Analysis of Fluid Flows	QA 374	QA 911	QC 20.7.17	QA 440	5
10	699754	Maurer-Cartan methods in deformation theory: the twisting procedure	QA 801	QA 169	QA 251	QC 20.7.H	11
11	699860	Complex Function Theory	QA 331.7	QA 329	QA 330	QA 331	3
13	699839	Introduction to nonlinear control: stability, control design, and estimation	QA 427	QA 402.5	QA 403	QA 402	10
15	699785	How to Derive a Formula 2. Further Analytical Skills and Methods for Physical Scientists	QA 401	QC 20	QC 19	QA 402	14
19	699938	Adversarial Learning and Secure AI	QA 76.87	Q 190	QA 78	Q 295	5
24	699761	Physics of binary star evolution: from stars to x-ray binaries and gravitational wave sources	QC 801	QB 980	QB 400	QB 495	17
29	699679	Sparse graphical modeling for high dimensional data: a paradigm of c	QA 276	QA 279	QA 166	QA 402	12
32	699709	Generalized Notions of Continued Fractions Ergodicity and Number Theoretic applications	QA 313	QA 281	QA 211	QA 295	10

33	700092	Algebra and number theory: a selection of highlights	QA 150	QA 281	QA 251	QA 295	9
35	699693	Constructive analysis of semicircular elements: from orthogonal projections to semicircular elements	QA 326	QA 329	QA 320	QA 9	11
36	698771	Conductive Homogeneity of Compact Metric Spaces and Construction of p	QA 320	QA 329	QA 329	QA 183	10
37	699716	More Explorations in Complex Functions	QA 331.7	QA 331	QA 350	QC 173	5
38	699556	A Friendly Approach to Complex Analysis	QA 331.7	QA 331	QA 350	QA 808	6
43	699228	Introduction to Methods for Nonlinear Optimization	QA 427	QA 427	QA 430	QA 321	6
45	698559	A Variational Theory of Convolution-Type Functionals	QA 374	QA 564	QA 377	QA 601	7
48	698566	Twisted Isospectrality, Homological Wideness, and Isometry A Sample of Algebraic Methods in Isospectrality	QA 150	QA 179	QA 613	QA 564	13
49	698702	Linear Fractional Transformations – An Illustrated Introduction	QA 440	QA 911	QA 564	QA 601	5
52	699198	Analytical Mechanics Classical, Lagrangian and Hamiltonian Mechanics, Stability Theory, Special Relativity	QA 801	QA 801	QA 800	QC 173	11
57	698610	From Kinetic Theory to Turbulence Modeling: The Legacy of Carlo Cercig	QC 174.9	QC 19.2	QC 19	QH 323	12

60	699891	Quantum Field Theory and Functional Integrals An Introduction to Feynman Path Integrals and the Foundations of Axiomatic Field Theory	QC 174.45 A1	QC 174.2	QC 174	QC 174.45	19
63	699389	A student‘s guide to the Ising model	QC 750	QC 174.85	QC 174.7	QC 173.7	8
64	699235	Field Theoretic Method in Phase Transformations	QC 120	QC 174.9	QC 174.9	QC 175	6
67	699150	An Introduction to the Mathematical Fluid Dynamics of Oceanic and Atmospheric Flows	QC 801	QC 801	QC 150	QC 809	12
68	699471	Complete Guide to the Fragment Molecular Orbital Method in GAMESS From One Atom to a Million, at your Service	QC 770	QD 450	QD 450	QD 461	19
70	699273	European Congress of Mathematics. Portoroz, 20–26 June 2021	QA 36	QA 9	QA 7	QA 1	10
87	699174	Field Arithmetic	QA 150	QA 247	QA 246	QA 241	2
88	698856	Iwasawa Theory and Its Perspective	QA 171	QA 247	QA 247	QA 273	5

Anhang C Quasi-Code des Programms

```
definiere barcode = barcodeEinlesen()

definiere titel = barcodeZuTitel(barcode)

definiere stichwortListe = titel.zertrenneInStichwörter()

definiere stichwortListeOhneStopwörter = stichwortListe.entferneStopwörter()

definiere stichwortKombinationen = liste.neu()

iteriere von stichwortListeOhneStopwörter.anzahlWörter() bis inklusive 2
als aktuelleStichwortLänge:

    stichwortKombinationen.fügeHinzu(alleWortKombinationenMitLänge(aktu-
    elleStichwortLänge))

iteriere über stichwortKombinationen bis gefunden als aktuelleStichwörter:

    prüfe bibliothek.enthält(aktuelleStichwörter):

        falls genau ein titel gefunden wird:

            definiere SACHGRUPPE = gefundenesBuch->SACHGRUPPE

        falls mehrere titel gefunden werden:

            definiere SACHGRUPPE = gefundeneBücher.BuchMitAmWenigstenWör-
            terImTitel()->SACHGRUPPE

        falls keins gefunden wird:

            wiederholePrüfungMitNächstenStichwort()

definiere locSachgruppe = frageSachgruppeVonLOCundRKVab(barcode)

prüfe locSachgruppe->vorhanden:

    falls vorhanden:

        prüfe ob locSachgruppe gleich SACHGRUPPE:

            falls übereinstimmung:

                gib „Erkennung erfolgreich“ aus

            falls nicht übereinstimmt:

                manuelleÜberprüfungAnfordern()

    falls nicht vorhanden:

        gibSachgruppeAus(SACHGRUPPE)

        manuelleÜberprüfungAnfordern()
```