# Retrievalexperimente mit bibliothekarischen Daten – Historischer Überblick und aktueller Forschungsstand

Bachelorarbeit zur Erlangung des Bachelor-Grades
*Bachelor of Arts* im Studiengang Bibliothekswissenschaft
an der Fakultät für Informationswissenschaft
der Technischen Hochschule Köln

vorgelegt von: Oliver Christian Meyer

eingereicht bei: Prof. Dr. Klaus Lepsky
Zweitgutachter/in: Prof. Dr. Philipp Schaer

Köln, 14.07.2022

# Kurzfassung/Abstract

Die Retrievalforschung in der Bibliothekswissenschaft hat in den letzten Jahrzehnten beachtliche Fortschritte gemacht. Automatische Indexierungsmethoden werden immer häufiger angewendet, obwohl dieses Thema in der Bibliothekswelt kontrovers diskutiert wird. Die Ergebnisse maschineller Erschließungsarbeit werden anhand von Retrievaltests festgehalten. Der Gegenstand dieser Arbeit ist die Darstellung von Retrievalexperimenten mit bibliothekarischen Daten. Zu Beginn werden die Grundlagen solcher Retrievaltests sowie das Cranfield-Paradigma erläutert. Es folgt eine Vorstellung verschiedener wissenschaftlicher Projekte aus diesem Forschungsfeld in chronologischer Reihenfolge. Wenn Verbindungen oder Einflussnahmen zwischen den einzelnen Projekten bestehen, werden diese herausgestellt. Im besonderen Umfang wird das Retrievalprojekt GELIC der TH Köln beschrieben, an dem der Autor dieser Arbeit beteiligt war. Obwohl es isolierte Retrievalprojekte gibt, lässt sich aus methodischer Sicht eine Verbindung von den frühesten Experimenten zu den heutigen Retrievalexperimenten herstellen. Diese Entwicklung ist noch nicht abgeschlossen.

Research regarding document retrieval in the field of library science has made remarkable progress over the past decades. The use of automatic indexing procedures keeps increasing, even though this is still a matter of controversy. The results of automatic indexing efforts are observed through retrieval tests. The objective of this thesis is to illustrate a number of retrieval experiments using library data. The paper will begin by laying out the basic requirements of retrieval tests and the components of the Cranfield Paradigm. As a next step, a number of scientific projects from this field of research will be introduced in chronological order. If connections or influences exist between individual projects, these will be pointed out. Special emphasis will be given to the retrieval project GELIC, which is in development at the Technische Hochschule Köln. The author of this thesis collaborated in this project. Even though isolated retrieval projects do exist, there are methodological links between the earliest and contemporary retrieval experiments. This development is still ongoing.

Schlagwörter: Retrievaltest, Information Retrieval, Cranfield Paradigma, MILOS, GELIC

# Inhalt

# 1. Einführung

In den Informations- und Bibliothekswissenschaften werden seit fast 70 Jahren verschiedene Aspekte von Bibliothekssystemen mit Retrievaltests überprüft. Mit diesen wird analysiert, wie gut ein Bibliothekssystem, anhand von Suchanfragen, für die NutzerInnen relevante Dokumente aus den im System repräsentierten Medien wiederfinden und anzeigen kann. Im Rahmen solcher Vorhaben werden neu implementierte Funktionen getestet oder es werden verschiedene Retrievalmethoden miteinander verglichen. Im Mittelpunkt dieser Experimente steht häufig die sachliche Recherche, bei der NutzerInnen mit einem Thema konfrontiert werden, mit dem sie noch nicht vertraut sind. Sie wissen noch nicht genau oder gar nicht, wie sie Anfragen am besten formulieren, oder welche konkreten Werke ihr derzeitiges Informationsbedürfnis befriedigen können.[1] Eine gute Retrievalleistung ist daher eine Voraussetzung für erfolgreiche Recherchen.

1953 wurden in den USA und Großbritannien die ersten, äußerst wichtigen Grundlagenarbeiten für Retrievaltests durchgeführt, bei denen verschiedene Methoden der Indexierung miteinander verglichen wurden.[2] Nach der Entwicklung von Bibliotheks-OPACs sollten durch Retrievalexperimente die Möglichkeiten und Grenzen der neuen digitalen Kataloge untersucht werden. Viele weitere Retrievaltests fanden im Rahmen von Forschungsvorhaben mit dem Ziel statt, halb- oder vollautomatische Indexierungskomponenten in Bibliothekssystemen zu integrieren.

Speziell die an der Cranfield-University durchgeführten Experimente sind für die Retrievalforschung von großer Bedeutung, da ihre Grundbausteine, das Cranfield-Paradigma, bis heute in Retrievaltests Anwendung finden.

## 1.1 Probleme der intellektuellen sachlichen Erschließung

Um die sachliche Suche in einem Bibliothekskatalog zu ermöglichen, müssen im System vermerkte Dokumente/Medien entsprechend erschlossen bzw. indexiert sein. Dies geschieht in Form von Indexierungssprachen und Klassifizierungen, die den thematischen Inhalt eines Werkes möglichst genau repräsentieren sollen. Dadurch werden Dokumente

[1] Nohr, Holger: Grundlagen der automatischen Indexierung. Ein Lehrbuch. 3., überarb. Aufl. Berlin: Logos-Verl., 2005. S.22-23.

[2] Borlund, Pia: Interactive Information Retrieval: An Introduction. In: Journal of Information Science Theory and Pratice 1 (3), 2013. S.14.

mittels hinzugefügter Metainformationen im Zuge des Information Retrieval wiederauffindbar gemacht.[3]

In der klassischen intellektuellen Erschließung begutachten BibliotheksmitarbeiterInnen zu erschließende Werke persönlich und vergeben nach einem Regelwerk[4] Schlagwörter und Notationen, die der Titelaufnahme hinzugefügt werden. Leider ist diese Form der Erschließung mit einem hohen zeitlichen und personellen Aufwand verbunden, z.B. durch den Aufbau und die Pflege von entsprechenden Terminologien. Außerdem besteht das Risiko menschlicher Inkonsistenz während des Indexierungsvorgangs und damit eine erhöhte Fehleranfälligkeit.[5]

Kaum eine bibliothekarische Einrichtung verfügt über die Ressourcen, um alle eingearbeiteten Dokumente intellektuell mit passenden Schlagwörtern in ausreichender Zahl zu versehen.[6] Bei entsprechenden Untersuchungen in den 1990er Jahren stellte sich heraus, dass selbst große Einrichtungen wie die ULB Düsseldorf lediglich ca. ein Drittel ihres Bestandes sachlich erschlossen hatten.[7] Eine vollständige händische Erschließung durch die immer größer werdende Zahl von Online-Publikationen galt im Bericht der Deutschen Nationalbibliothek (DNB) „Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek“ von 2010 bereits als nicht mehr machbar, zumal die Zahl physischer Neuerscheinungen nicht geringer wird.[8] Alternativen zur intellektuellen Erschließung werden daher dringend benötigt.

## 1.2 Automatische Indexierung als Alternative zur intellektuellen Erschließung

Aus diesem Grund gibt es immer mehr Bestrebungen, die sachliche Erschließungsarbeit zu automatisieren. Bei der automatischen Indexierung handelt es sich um Computer-gestützte Verfahren, die Dokumente analysieren und zum Zweck des Information Retrieval Titelaufnahmen Indexterme oder Deskriptoren hinzufügen, damit die so indexierten Titel

[3] Nohr, S.26-27.

[4] Z.B. die Regeln für die Schlagwortkatalogisierung (RSWK) und der Dewey Decimal Classification (DDC).

[5] Ebd., S.31.

[6] Ebd., S.31-32

[7] Lepsky, Klaus: Automatisierung in der Sacherschließung: Maschinelles Indexieren von Titeldaten. In: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen. Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft 63. Hrsg. von Sabine Wefers. Frankfurt: Klostermann 1996. S. 223.

[8] Gömpel, Renate; Junger, Ulrike; Niggemman, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken 20 (1), 2010. S.20.

recherchierbar werden.[9] Die DNB nutzt seit 2010 maschinelle Methoden zur Unterstützung der manuellen Erschließungsarbeit[10] und hat bereits mehrere automatische Indexierungslösungen erprobt. Dieses Vorhaben wird zum Zeitpunkt des Schreibprozesses dieser Arbeit fortgesetzt.[11]

Eine Anzahl von wissenschaftlichen Projekten haben sich seit der Entwicklung von digitalen Bibliothekskatalogen mit der Auswirkung automatischer Indexierungsmethoden auf die Retrievalleistung von Bibliothekssystemen beschäftigt. Diese und andere Retrievalexperimente bilden den Gegenstand dieser Arbeit.

## 1.3 Erkenntnisinteresse

In dieser Arbeit soll ein historischer Überblick über die Anwendung von Retrievaltests unter Verwendung von bibliothekarischen Daten gegeben werden. Der Zeitrahmen reicht von den 50er-Jahren des vorigen Jahrhunderts bis in die 2020er Jahre.

Die Motivation hinter diesem Vorhaben ergibt sich aus dem Umstand, dass trotz der wichtigen Grundlagenforschung in den 1950ern und aussagekräftiger nachfolgender Tests in anderen Jahrzehnten die Zahl solcher Experimente im Bibliothekssektor eher überschaubar geblieben ist. Es wäre wünschenswert, wenn Bibliotheken von diesem Werkzeug häufiger Gebrauch machen würden, um die Retrievalleistung ihrer Systeme zu testen und aus den Ergebnissen Grundlagen für Verbesserungen der Such- und Retrievalfunktionen abzuleiten.

Aus der oben erwähnten Überschaubarkeit bibliothekarischer Retrievaltests ergibt sich zudem ein Interesse, diese verschiedenen Experimente miteinander in Verbindung zu setzen. Es soll gezeigt werden, welche von diesen Projekten direkten Einfluss auf andere Retrievalexperimente ausüben. Retrievalprojekte, die weitgehend unbeeinflusst von anderen Projekten sind, werden ebenfalls vorgestellt.

[9] Nohr, s. 27.

[10] Deutsche Nationalbibliothek: Grundzüge und erste Schritte der künftigen inhaltlichen Erschließung von Publikationen in der Deutschen Nationalbibliothek. Stand Mai 2017. https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/konzeptWeiterentwicklungInhaltserschliessung.pdf?__blob=publicationFile&v=4 (Letzter Zugriff: 10.07.2022).

[11] Busse, Frank; Grote, Claudia; Jacobs, Jan Helge et al.: Erschließungsmaschine gestartet. DNB-Blog. 4.5.2022 https://blog.dnb.de/erschliessungsmaschine-gestartet/ (Letzter Zugriff: 10.07.2022).

## 1.4 Aufbau der Arbeit

Zu Beginn werden einige Grundlagen zu Retrievaltests im Allgemeinen erläutert, und zwar wozu solche Tests dienen und welche Erkenntnisse damit typischerweise gewonnen werden sollen. In diesem Zusammenhang werden Retrievalexperimente in bibliothekarischen und informationswissenschaftlichen Kontexten kurz miteinander verglichen, um Besonderheiten von Retrievaltests mit bibliothekarischen Daten herauszustellen. Im Anschluss daran wird das Cranfield-Paradigma im Detail beschrieben, da sein Einfluss bis heute in der Retrievalforschung sichtbar ist.

Die historische Darstellung der Retrievaltests ist chronologisch angeordnet. Jedes Experiment wird einzeln anhand der folgenden Aspekte vorgestellt:

- Motivation bzw. Forschungsinteresse
- Details des experimentellen Setup
- Datengrundlage
- Beschaffenheit der Elemente des Cranfield-Paradigmas im jeweiligen Projekt
- Relevanzkriterien
- Darstellung von Ergebnissen und Interpretation durch die Experimentsteilnehmer
- Einfluss durch oder auf andere Retrievalexperimente

Die oben erwähnten Bemühungen zum Thema automatische Indexierung werden dabei eine besondere Rolle spielen. Es werden jedoch auch Projekte mit anderen Forschungsbestrebungen vorgestellt.

In besonderem Umfang soll das Projekt **GELIC** (German Library Indexing Collection) beschrieben werden, das seit 2017 von Dozenten der Technischen Hochschule Köln in Zusammenarbeit mit Studierenden des ehemaligen Studiengangs Bibliothekswissenschaft durchgeführt wurde. Das Ziel dieses Projekts besteht darin, mit Hilfe von Open-Source-Software wie Solr und Trec_eval auf der Basis von Daten der DNB eine wiederverwendbare Testkollektion zu erstellen, mit der u.a. die Qualität der verschiedenen Erschließungsmethoden der DNB (automatisch vs. intellektuell) untersucht werden soll.[12]

[12] Munkelt, Johanna: Erstellung einer DNB-Retrieval-Testkollektion. Bachelor-Arbeit. Köln: Technische Hochschule Köln, 2018. S.2-3.

Zum Abschluss der Arbeit wird ein Resümee gezogen und ein Ausblick auf aktuelle Entwicklungen in der dt. Bibliothekslandschaft zum Thema automatische Indexierung und Retrievalforschung gegeben.

# 2 Zur Theorie von Retrievaltests

## 2.1 Motivationen für Retrievaltests

Retrievaltests können durchgeführt werden, um eine Anzahl möglicher Ziele zu erreichen. Bei den Cranfield- und ASTIA-Uniterm-Experimenten z.B. ging es darum, eine Anzahl von Retrievalsprachen bezüglich ihrer Leistungsfähigkeit miteinander zu vergleichen.[13] Andere Tests wiederum arbeiten im Kontext eines einzelnen Bibliothekssystems, indem beispielsweise eine neue Indexierungskomponente erprobt wird. Manchmal werden in einem vergleichenden Experimentsaufbau bewährte Retrievalkomponenten und neu entwickelte Nachfolgekandidaten auf Vergleichsbasis getestet. Das hat Auswirkungen auf die künftige Entwicklung eines Bibliothekssystems, wenn die Retrievalkomponente gegebenenfalls weitgehend ausgetauscht und beispielsweise von klassischer intellektueller Sacherschließung zu einer semi- oder vollautomatischen Lösung gewechselt wird. Nachfolgende Retrievaltests können dazu verwendet werden, die Ergebnisse vorangegangener Experimente zu reproduzieren und ggfs. zu widerlegen oder zu bestätigen.

## 2.2 Grundlagen von Retrievalexperimenten

Vor dem Beginn eines Retrievalexperiments müssen bestimmte Aspekte geklärt sein, damit das Projekt nicht selbstzweckhaft durchgeführt wird. Der angestrebte Erkenntnisgewinn ist entscheidend für die Ausrichtung des gesamten Vorgangs. Die folgenden Fragen sind zu klären:

- Welche neuen Erkenntnisse sollen durch den Test gewonnen werden?
- Hat bereits eine andere Projektgruppe ein ähnliches Experiment durchgeführt?
- Wenn ja, welche Ergebnisse und Erkenntnisse wurden dabei erzielt?

[13] Sachse, Elisabeth; Liebig, Martina; Gödert, Wilfried: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II Projekt. Kölner Arbeitspapiere zur zur Bibliotheks- und Informationswissenschaft. Band 14. Köln: Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1998. S.7.

- Welche Informationen lassen sich durch die Lektüre von Fachliteratur zu Information Retrieval gewinnen? [14]

Zweitens müssen die Bedingungen, unter denen das Experiment durchgeführt wird, festgelegt werden. Retrievalexperimente können unter Labor-Bedingungen oder in einem laufenden Bibliothekssystem durchgeführt werden. Bei Tests mit bibliothekarischen Daten bedeutet dies entweder den Aufbau einer Testdatenbank mit aus einer anderen Quelle importierten Dokumenten/Titeldaten oder die Nutzung eines vorher existierenden Bibliothekskatalogs zu Testzwecken unter den Bedingungen, die sowohl BibliotheksmitarbeiterInnen als auch NutzerInnen aus dem Alltag kennen.[15]

Beide Varianten bieten Vor- und Nachteile und beeinflussen die Durchführung des Experiments. Unter Laborbedingungen hat die Projektgruppe eine größere Kontrolle über den praktischen Ablauf und die Variablen des Experiments (TeilnehmerInnen, die Datengrundlage, Suchbedingungen). Allerdings tendieren Experimente unter Laborbedingungen dazu, kostspieliger und arbeitsintensiver zu sein, da Elemente wie die Testdatenbank erst erstellt werden müssen.[16]

Bei Experimenten in einem laufenden System geht es häufig um die Beobachtung des Systems unter „realen“ Bedingungen. Bei solchen Experimenten hat die Projektgruppe wenig bis keine Kontrolle über die teilnehmenden NutzerInnen oder die Datengrundlage. Zudem lassen sich die gewonnenen Erkenntnisse nicht immer generalisieren, bzw. auf andere IR-Systeme anwenden, aber sie geben die reale Benutzung des Systems wieder.[17]

### 2.2.1 Retrievaltests in Informationssystemen vs. Bibliothekssysteme

Retrievaltests können in allen IR-Systemen durchgeführt werden, um die Retrievalleistung zu untersuchen. Bei Projekten mit bibliothekarischen Daten entstehen jedoch gewisse Unterschiede im Vergleich zu beispielsweise Retrievaleffektivitätsstudien bei Internetsuchmaschinen.

Bei Studien mit Suchmaschinen ist die Art der Anfrage ein wichtiger Faktor. In entsprechenden Arbeiten wird häufig zwischen informational, navigational und transactional queries unterschieden. Informational queries bzw. Informationsanfragen entsprechen in

[14] Teague-Sutcliffe, Jean: The Pragmatics of Information Retrieval Experimentation Revisited. In: Information Processing & Management 28 (4), 1992. S.468.

[15] Teague-Sutcliffe., S.469.

[16] Ebd.

[17] Ebd.

ihrer Absicht der sachlichen Recherche in einem Bibliothekskatalog, d.h. es geht um die Suche nach Dokumenten zu einem bestimmten Thema, um das Informationsbedürfnis der NutzerInnen zu befriedigen. Navigational queries bzw. Navigationsanfragen repräsentieren die Suche nach einer bereits bekannten Website, und der Erfolg ist darin zu sehen, ob die Suchmaschine die gesuchte Seite finden kann oder nicht. Transactional queries bzw. Transaktionale Anfragen schließlich sind mit der Navigation zu einer Website verbunden, wo eine Transaktion wie der Download einer Datei o.ä. erfolgen soll. Diese sollten laut Dirk Lewandowski unbedingt getrennt beobachtet werden.[18] Bei Retrievaltests in Bibliotheken ist eine solche Unterscheidung der Anfragetypen meistens nicht gegeben, da die sachliche Recherche (das Gegenstück zu den informational queries) oft Gegenstand der Untersuchungen ist. Allerdings gibt es mit der known-item-search ein bibliothekarisches Gegenstück zur navigational query, d.h. die Suche nach einem bestimmten Dokument, wie in Harald Kaluzas Retrievalexperiment (s.43 ff.).

Ein weiterer Unterschied liegt in den Dokumentenmengen. Die Samples bei Internet-Suchmaschinen sind normalerweise deutlich kleiner als bibliothekarische Dokumentenkollektionen. Manche Retrievaleffektivitätsstudien stellen lediglich 50 Anfragen an die zu untersuchende Suchmaschine. Ein für dieses Forschungsfeld großes Sample von Anfragen findet sich bei Lewandowski, der mit je 1000 Informations- und Navigationsfragen die Suchmaschinen von Google und Microsoft (Bing) miteinander verglich.[19] Währenddessen enthalten Dokumentenkorpora in einem bibliothekarischen Retrievaltest oft deutlich mehr Dokumente, damit repräsentative Ergebnisse entstehen. Solche Dokumentenmengen können ohne weiteres Hunderttausende oder Millionen Dokumente oder Datensätze enthalten, wie z.B. die TREC-Testkollektionen, die absichtlich so konzipiert wurden, um die Nützlichkeit großer Dokumentenmengen bei Retrievalexperimenten aufzuzeigen.[20] Bei Retrievaltests mit Suchmaschinen ist der Einsatz von Testkollektionen generell nicht sinnvoll, da sich die Ergebnisse daraus nicht auf die Masse von Dokumenten anwenden lassen, auf die eine Internet-Suchmaschine zugreifen kann.[21]

[18] Lewandowski, Dirk: Evaluating the retrieval effectiveness of Web search engines using a representative query sample. In: Journal of the Association for Information Science and Technology 66 (9), 2014. Preprint-Ausgabe. S. 2. https://arxiv.org/ftp/arxiv/papers/1511/1511.05817.pdf (Letzter Zugriff: 10.07.2022)

[19] Ebd., S.5.

[20] Vorhees, Ellen M.: Continuing Information Retrieval's Tradition of Experimentation. In: Communications of the ACM 50 (11), 2007. S.52-53.

[21] Lewandowski, S.12.

### 2.2.2 Kennzahlen zur Messung von Retrievalergebnissen

Sowohl in informationswissenschaftlichen als auch bibliothekarischen Retrievalexperimenten können die Ergebnisse mit denselben Kennzahlen gemessen werden, die ihrerseits auf das Experiment Cranfield II zurückgehen.[22]

- **Recall** setzt die Zahl gefundener relevanter Dokumente in der Treffermenge in ein Verhältnis zu allen als relevant eingestuften Dokumenten in der gesamten Dokumentenkollektion.
- **Precision** misst die Zahl gefundener relevanter Dokumente innerhalb der Treffermenge als Verhältnis von relevanten und nicht relevanten Dokumenten.
- Recall und Precision-Werte sind zwischen 0 und 1 angesetzt und lassen sich in Prozentzahlen ausdrücken. [23]

Cyril Cleverdon stellte während des Experiments Cranfield I eine inverse Beziehung zwischen „der Fähigkeit eines Systems, relevante Dokumente zu finden“[24] (aka dem Recall) und „der Fähigkeit eines Systems, nicht relevante Dokumente zurückzuhalten“ (aka der Precision) fest.[25] Demnach bedeutet ein hoher Recall meistens eine niedrigere Precision, oder bzw. eine hohe Precision einen niedrigen Recall. Dass der Unterschied zwischen beiden Werten jedoch nicht groß sein muss, zeigen z.B. die MILOS-Experimente.[26]

Den Recall zuverlässig zu messen ist oft nur bei sehr kleinen Dokumentenmengen möglich. In unveränderlichen Kollektionen mit bibliothekarischen Dokumenten wäre die Ermittlung theoretisch möglich, sofern die Dokumentenmenge klein genug ist und genug Personal vorhanden ist, um die nötigen Relevanzurteile zu fällen. Im Normalfall ist dies jedoch aus Zeit- und Personalmangel nicht praktikabel, da selbst bei kleinen Testkollektionen viele Tausend Relevanzurteile gefällt werden müssten.[27] Retrievaleffektivitätsstudien zu Internet-Suchmaschinen verzichten oft gänzlich auf die Messung des Recall. Wenn im Internet gesucht wird und eine unüberschaubar hohe Anzahl von Treffern vorliegt, ist es unmöglich einen Recall-Wert auch nur zu schätzen.[28]

[22] Sachse, Liebig, Gödert: S.9.
[23] Nohr, S.154-156.
[24] Sachse, Liebig, Gödert, S.8.
[25] Ebd.
[26] Nohr, S.156.
[27] Ebd., S.158.
[28] Lewandowski, S.12.

## 2.3 Das Cranfield-Paradigma

Während der Cranfield-Experimente in den 1950ern wurden Grundlagen gelegt, die den Aufbau und die Durchführung vieler nachfolgender Retrievalexperimente bis heute beeinflussen. Dieses Paradigma beschreibt drei Komponenten für den Aufbau einer sog. Testkollektion.

### 2.3.1 Dokumentenkorpus

Die erste Komponente ist ein Dokumentenkorpus bzw. eine Dokumentenkollektion, d.h. eine Ansammlung von Dokumenten, in der während des Experiments gesucht werden soll. Die Größe des Korpus kann zwischen einer geringen Dokumentenanzahl (z.B. wenige Hundert) und kompletten Datenbanken mit Millionen Dokumenten liegen. Wenn die Größe des Korpus festgelegt wird, muss beachtet werden, welche Datenmengen mit den TeilnehmerInnen des Experiments und der eingesetzten Software zuverlässig untersucht werden können. Eine zu große Datenmenge kann die Interpretation der Ergebnisse erschweren, während eine zu kleine Kollektion zwar überschaubar ist, aber möglicherweise keine aussagekräftigen Resultate liefert.[29]

### 2.3.2 Topics

Bei Topics handelt es sich um die Formulierung von Informationsbedürfnissen, die im Laufe des Experiments an den Korpus gestellt werden. Diese Topics müssen dabei in eine Anfrageform übersetzt werden, die die im Experiment genutzte Retrievalsoftware verarbeiten kann.[30]

Zur genauen Definition können bei der Erarbeitung von Topics sogenannte descriptions und narratives hinzugefügt werden. Die description umfasst eine kurze Beschreibung des Topics und seines thematischen Gegenstands. Das narrative stellt eine Vertiefung der description dar, die definiert wann ein Dokument bei Anfragen zu diesem Topic relevant ist und wann nicht, damit keine themenfremden Dokumente unter ein Topic fallen.[31]

Topics sollten realistische Sachverhalte und Formulierungen abbilden, indem z.B. Anfragen aus realen Bibliothekssystemen analysiert werden. Clough und Sanderson empfehlen, mindestens 50 verschiedene Topics zu entwickeln, die eine Reihe verschiedener

[29] Munkelt, S.16-17.
[30] Ebd., S.18-19.
[31] Ebd.

Sachthemen umfassen und die heterogen formuliert sind, d.h. es sollten sowohl Topics mit nur einem Begriff als auch solche mit Phrasenformulierungen vorhanden sein.[32]

### 2.3.3 Relevanzurteile

Bevor mit den Anfragen an das System begonnen wird, muss geklärt werden, welche Dokumente innerhalb des Korpus für Suchen mit einem bestimmten Topic relevant sind. Ann O'Brien definiert Relevanz als „die Angemessenheit der gefundenen Dokumente im für die Anfrage und den Nutzer [meine Übersetzung]"[33].

In der Retrievalforschung wird zwischen subjektiver und objektiver Relevanz unterschieden. Nohr definiert Dokumente dann als subjektiv relevant, wenn sie ein aktuelles Informationsbedürfnis befriedigen. Dagegen gelten Dokumente als „objektiv" relevant, wenn auch Dokumente, die den Suchenden bereits vor der Recherche bekannt waren, als relevant angesehen werden.[34] O'Brien dagegen definiert objektive bzw. logische Relevanz im Kontext aller Dokumente, die der Aussage der Suchfrage so genau wie möglich entsprechen. Bei der subjektiven Relevanz findet sich auch bei ihr der Aspekt der aktuellen Nützlichkeit, die stark von den persönlichen Eigenschaften der Suchenden abhängig ist.[35]

Die Anwesenheit der NutzerInnen in der Definition verdeutlicht ein großes Problem, denn Relevanz ist kein vollständig objektiver Begriff, sondern kann in Bezug auf den Erfolg eines Suchvorgangs unterschiedlich beurteilt werden.[36] Laut O'Brien wurden Dokumente von manchen ForscherInnen beispielsweise nur dann als relevant angesehen, wenn der Suchbegriff im gefundenen Dokument vorkommt, z.b. im Titel.[37] Ein Dokument kann für eine/n ExperimentsteilnehmerIn für ein bestimmtes Topic als relevant gelten, während andere Personen diesem Urteil widersprechen könnten. Ebenso sind Relevanzeinschätzungen nicht zwingend konstant. Individuelle Informationsbedürfnisse sind veränderlich. Ein Dokument, das zu einem Zeitpunkt als relevant angesehen wird, kann zu einer anderen Gelegenheit als nicht mehr relevant gelten, wenn sich der individuelle Wissenstand

[32] Clough, Paul; Sanderson, Mark: Evaluating the performance of information retrieval systems using test collections. In: Information Research 18 (2), 2013. http://informationr.net/ir/18-2/paper582.html#.Yr90P4TP1PY (Letzter Zugriff: 01.07.2022).

[33] O'Brien, Ann: Relevance as an aid to evaluation in OPACs. In: Journal of Information Science 16 (4), 1990. S. 265.

[34] Sachse, Liebig, Gödert, S.29.

[35] O'Brien, S.266-267.

[36] Nohr, S. 152.

[37] O'Brien, S. 266.

zu einem Topic zwischenzeitlich geändert hat.[38] Aus diesem Grund werden die Ergebnisse von Retrievaltests häufig kritisiert, weil sowohl die Methode zur Relevanzermittlung als auch die tatsächlichen Relevanzurteile sehr unterschiedlich bewertet werden können und es keine objektive Grundlage für eine Übereinstimmung gibt.[39]

Das Erstellen von Relevanzurteilen birgt Herausforderungen, die sich nicht immer vollständig lösen lassen. Bei bibliothekarischen Dokumentenkollektionen, die aus vielen Tausend Dokumenten bestehen können, ist es in der Regel aus Personal- und Zeitgründen nicht möglich, die Relevanz jedes vorhandenen Dokuments für jedes Topic zu ermitteln. Es gibt Hilfsmittel wie das sogenannte Pooling-Verfahren, bei dem die ExperimentsteilnehmerInnen kooperativ Suchanfragen aus allen Topics an die Testkollektion stellen, die Ergebnislisten zusammentragen und aus diesen für jedes Topic relevante oder irrelevante Dokumente bestimmen. Dieser Ansatz hilft dabei, zumindest annähernd genaue Relevanzurteile für jedes Topic zu erstellen, sodass mit den Ergebnissen gearbeitet werden kann. Streng genommen handelt es sich hierbei um Schätzungen, aber diese stellen oft die einzige Grundlage für aussagekräftige Ergebnisse dar.[40] Die Kennzahlen Precision und Recall werden in Verbindung mit den Relevanzurteilen verwendet, um Retrievalergebnisse aussagekräftig darzustellen.

Außerdem muss entschieden werden, wie die Relevanz gemessen wird. Hier besteht i.A. die Auswahl zwischen einer binären Einschätzung von relevant und nicht relevant oder einer Relevanzskala, die von nicht relevant über eine Anzahl von Abstufungen bis zu relevant reicht.[41]

# 3 Historischer Überblick: Retrievaltests mit Bibliotheksdaten

Im Folgenden werden so viele Retrievalexperimente mit Bibliotheksdaten vorgestellt, wie es in den Rahmenbedingungen dieser Arbeit möglich ist. Viele dieser Projekte stammen aus dem deutschsprachigen Raum. Obwohl die Ursprünge der Retrievaltests im englischsprachigen Raum liegen, wurde in Deutschland ab den 1990ern mehrere zum Teil aufeinander aufbauende Projekte durchgeführt.

[38] O‘Brien, S.268.
[39] Nohr, S.152.
[40] Munkelt, S. 20.
[41] O’Brien, S.266.

Diese Darstellung erhebt keinen Anspruch auf Vollständigkeit, einerseits aus Zeit- und Platzgründen, andererseits weil nicht alle bibliothekarischen Retrievalexperimentgruppen ihre Ergebnisse nachvollziehbar publizieren.

Jedes Projekt wird wie folgt dargestellt: es werden die Motivation bzw. das Forschungsinteresse kurz erläutert, die jeweiligen Elemente des Cranfield-Paradigmas (Kollektionsgröße, Zahl der Anfragen/Topics, Relevanzbeurteilung) genannt und das Endergebnis sowie die Interpretation desselben durch die durchführenden ProjektteilnehmerInnen aufgezeigt. Zudem sollen Verbindungen wie Zitationen oder direkte Einflüsse zwischen den einzelnen Projekten aufgezeigt werden, sofern diese vorhanden sind.

## 3.1 Erste Retrievalexperimente in Großbritannien und in den USA (1953 – 1969)

### 3.1.1 ASTIA-Uniterm-Experiment

1953 wurde an der Armed Services Technical Information Agency (ASTIA) des US-amerikanischen Verteidigungsministeriums eine Evaluation des Uniterm-Systems der Firma Documentation Inc. (DI) im Vergleich mit damals konventionellen Ansätzen der Indexierung und ASTIAs eigenem System durchgeführt. Nach dem UNITERM-System sollten per Extraktionsverfahren Dokumente durch einzelne aus dem Titel oder einem Abstract entnommene Begriffe automatisch repräsentiert werden Zwei Projektgruppen führten voneinander unabhängig Indexierungsexperimente durch: einmal MitarbeiterInnen von ASTIA mit der internen ASTIA subject heading list (basierend auf intellektueller sachlicher Erschließung) und einmal Angestellte der Firma Documentation Inc. mit dem Uniterm-System. Die Datengrundlage war der komplette ASTIA-Dokumentenbestand, also ca. 15.000 Dokumente .[42] Es wurden 98 Anfragen an die Kollektion gestellt, um die Ergebnisse beider Indexierungsarten zu vergleichen.[43]

Das Kriterium für die Effektivität des jeweiligen Systems lag in der Relevanz der gefundenen Dokumente. Die Relevanzurteile wurden von den Teilnehmern beider Teams separat erstellt. Daraus entstand Uneinigkeit darüber, welche Dokumente für welche Anfragen relevant waren. Die ASTIA-Gruppe fand 2200 von ihnen als relevant bezeichnete Dokumente, während die DI-Mitarbeiter 1560 als relevant eingestufte Dokumente fanden. 580 Dokumente wurden dabei von beiden Teams gefunden. Von den insgesamt 3760

[42] Zitiert bei Sachse, Liebig, Gödert, s.7.

[43] Gray, Dwight E.: Report on the Reference Test of the Conventional and Uniterm Systems. Washington, D.C.: ASTIA Reference Center, 1954. S. 24.

gefundenen Dokumenten wurden 1577 von einem Team für relevant, vom anderen jedoch für irrelevant gehalten. Eine endgültige Einigung fand nicht statt, womit dieses erste Retrievalexperiment die erste kontroverse Diskussion über Relevanzurteile enthält.[44]

### 3.1.2 Cranfield-Uniterm-Experiment

Im selben Jahr wurde in Großbritannien ein weiteres UNITERM-Experiment am Cranfield College of Aeronautics durchgeführt. Auch hier wurde das UNITERM-System mit einem traditionellen Indexierungssystem, diesmal basierend auf der Universal Decimal Classification (UDC), verglichen.

Im Vergleich zum ASTIA-Experiment war die Datengrundlage wesentlich kleiner und strenger kontrolliert. Es wurden 200 Dokumente aus dem Bestand der Bibliothek des College für den Test ausgewählt. Diese stammten aus dem thematischen Bereich der Aeronautik. 40 Anfragen an die Dokumentenkollektion wurden erarbeitet, die auf sogenannten Source-Dokumenten basierten. Das Kriterium für die Effektivität eines Systems, und damit einhergehend für die Relevanz des Ergebnisses, bestand ausschließlich darin, durch eine Anfrage das dazu gehörige Source-Dokument zu finden. Wenn dies nicht gelang, galt das Suchergebnis als nicht relevant. UNITERM erzielte die besseren Ergebnisse, da unter diesem System 85% der Source-Dokumente gefunden wurden. Die Anfragen mit Hilfe der UDC wiesen dagegen nur 50% der Source-Dokumente nach.[45]

Später wurde dieses Experiment dafür kritisiert, sich auf ein Kriterium, nämlich das Retrieval der Source-Dokumente, als einziges Erfolgskriterium zu beschränken. Die zusätzlich gefundenen Dokumente, ob relevant oder nicht, wurden ignoriert. Zudem begünstigte das eingesetzte Verfahren Begriff-basierte Ansätze wie UNITERM.[46]

### 3.1.3 Cranfield I (1957)

4 Jahre nach dem ersten Uniterm-Experiment führte Cyril Cleverdon ein weiteres Retrievalexperiment am Cranfield-College durch. Bei diesem wurden vier Indexierungsmethoden miteinander verglichen statt zwei. Zusätzlich zu UDC und Uniterm wurden untersucht: der Alphabetical Subject Catalogue, der Themengebiete (subject headings) in Phrasenform alphabetisch anordnet sowie das Faceted Classification Scheme, bei dem

[44] Zitiert bei Borlund, S. 14.
[45] Zitiert bei Sachse, Liebig, Gödert, S.7-8.
[46] Ebd., S.8.

Sachgebiete miteinander kombiniert werden können, um neue komplexe Kategorien zu erzeugen.[47]

Die grundsätzliche Durchführung entsprach weitgehend dem Vorgänger-Experiment, machte allerdings von einer wesentlich größeren Datengrundlage Gebrauch. Diesmal wurden 18000 Dokumente aus dem Themengebiet „Luftfahrt-Ingenieurswesen" als Korpus ausgewählt. Die oben beschriebene Methode der Source-Dokumente wurde beibehalten und für 1200 Anfragen an den Korpus verwendet. Diese wurden von Experten in den vier Systemen ausgearbeitet und durchgeführt. Wieder galten nur die Suchanfragen als erfolgreich, welche die vorher festgelegten Source-Dokumente finden konnten. Unerfolgreiche Suchvorgänge wurden analysiert, um eine Erklärung für die Retrieval-Fehler zu finden. Als mögliche Problemzonen wurden die Anfragen-Formulierung, der Suchvorgang selbst, die Indexierung oder das gesamte Bibliothekssystem angenommen.[48]

Bei den Ergebnissen ergaben sich die folgenden Trefferquoten:

- Uniterm 82%
- Alphabetical Subject Catalogue: 81,5%
- UDC-Klassifizierung 75,6%
- Faceted Classification Scheme 73,8%[49].

Die Unterschiede zwischen den vier Systemen fielen nicht allzu hoch aus. Als Grund für das schwache Abschneiden des FCS wurden Probleme bei der ausgewählten Klassifikations-Repräsentation im Karteikartenkatalog identifiziert. Eine alternative Repräsentation wurde ausprobiert, mit der FCS sogar die Ergebnisse von Uniterm übertraf. Daraus schlossen die Projektteilnehmer, dass nicht die Prinzipien der einzelnen Systeme entscheidend für ihre Retrievalqualität sind, sondern deren Implementierung.[50] Sachse, Liebig und Gödert sehen darin jedoch Anzeichen für die Überlegenheit automatischer Indexierungsmethoden über klassische intellektuelle Arbeitweisen.[51] Wie bereits beim Uniterm-Experiment vier Jahre zuvor wurde die ausnahmslose Ausrichtung des Retrievals

[47] Robertson, Stephen: On the history of evaluation in IR. In: Journal of Information Science 34 (4), 2008. S.441.

[48] Zitiert bei Borlund, S. 14-15.

[49] Zitiert bei Sachse, Liebig, Gödert, S.8.

[50] Zitiert bei Robertson, S.441.

[51] Uniterm kann als solche angesehen werden, da die Erschließung mit Hilfe einer Datenverarbeitungsmaschine durchgeführt wurde. Siehe Sachse, Liebig, Gödert, S.8.

auf Source-Dokumente kritisiert, u.a. weil es im echten Bibliotheksalltag normalerweise keine solchen Dokumente gibt.[52]

Um dieser Kritik zu begegnen, wurden zwei Folgetests durchgeführt. Im ersten Test erhielten fachverwandte externe Bibliotheken und Informationseinrichtungen ein Sample mit 100 der in Cranfield-I benutzten Anfragen. Anhand dieser sollten Literaturlisten angefertigt und an das Cranfield-College zurückgeschickt werden. Aus 88 Listen wurden 359 Titel ermittelt und bzgl. Ihrer Relevanz zu den Source-Dokumenten und Test-Anfragen analysiert, in einer Abstufung von „genauso relevant wie das Source-Dokument“ bis zu „nicht relevant“. Das Ergebnis bestand aus 41 Fragen, die 120 Dokumente mit zumindest geringer Relevanz fanden. Eine Wiederholung des Testsetups mit diesen 41 Fragen erzielte die folgenden Trefferquoten:

- Uniterm 75%
- Alphabetic Subject Headings 75%
- UDC 74%
- FCS 60%[53]

Auffallend ist der Abfall der FCS gegenüber den anderen drei untersuchten Systemen, die fast identisch gute Ergebnisse lieferten.

Der zweite Folgetest fand intern statt. 79 Suchfragen wurden per Zufallsprinzip ausgewählt und an das System gestellt. Daraus ergaben sich 759 Dokumente, die hinsichtlich ihrer Relevanz geprüft werden. Das Ziel dieses Tests bestand darin, die Fähigkeit des Systems, nicht-relevante Dokumente zurückzuhalten, herauszustellen. Die ACS erzielten mit 12,5% das beste Ergebnis, gefolgt von Uniterm mit 12%, FCS mit 7,5% und Cranfields intern genutzter UDC mit 7%. Cleverdon schloss aus diesen Ergebnissen, dass zwischen den Fähigkeiten eines Systems relevante Dokumente zu finden und zugleich irrelevante Dokumente zurückzuhalten, eine inverse Beziehung existiert. Dies entspricht dem typischen Verhältnis zwischen Recall und Precision in der Retrievalforschung.[54]

### 3.1.4 Cranfield II (1966)

9 Jahre nach Cranfield I startete eine Forschungsgruppe um Cyril Cleverdon ein weiteres Retrievalexperiment. Dieses basierte auf der Annahme, dass Indexierungssprachen aus

[52] Borlund, S. 15.
[53] Zitiert bei Sachse, Liebig, Gödert, S.8-9.
[54] Ebd.

der Kombination verschiedener Indexierungsarten bestehen. Das Ziel des Projekts bestand darin, eine Reihe verschiedener Indexierungsarten bzgl. ihrer Systemperformance zu testen und den Einfluss derjenigen Indexierungsarten festzustellen, die entweder die Precision oder den Recall erhöhen. Diese Retrievalmaße wurden während der Cranfield-Experimente erstmals so bezeichnet.[55]

Für den Test wurden 33 Indexierungssprachen aus verschiedenen Terminologien und Strukturen entwickelt. Diese unterschieden sich durch die Nutzung von Einzelwörtern (8 Sprachen), Komposita bzw. einfachen Konzepten (15 Sprachen), Abstracts und Titeln (4 Sprachen) oder kontrolliertem Vokabular (6 Sprachen).[56]

Cranfield II nutzte eine kleinere Testkollektion als Cranfield I. Die Datengrundlage für das Experiment bildeten 1400 Dokumente aus dem Bereich Luftfahrt. Es wurden 211 Suchfragen erarbeitet. Diese basierten auf den Grundfragen, die die Verfasser der Dokumente in der Testkollektion zum Verfassen der Texte formuliert hatten. Vor der praktischen Durchführung waren für jede Suchfrage relevante Dokumente identifiziert worden. Zunächst führten teilnehmende Studierende Suchen im System durch und gaben ihre Urteile bezüglich der Relevanz der gefundenen Dokumente ab. Die MitarbeiterInnen, die die Anfragen ausgearbeitet hatten, überprüften diese Ergebnisse auf Fehler. Danach wurden die Suchergebnisse den AutorInnen der Dokumente zur Beurteilung vorgelegt. Die Retrievalperformance der eingesetzten Indexierungssysteme wurde ausschließlich danach bewertet, wie zuverlässig sie diese vorher ausgewählten Titel auffinden konnten.[57]

Die besten Ergebnisse wurden durch acht Einzelwort-Indexierungssprachen erzielt. Diese erreichten Recall-Werte zwischen 65,82 und 64,05 % während dreizehn der fünfzehn Konzept-basierten Sprachen die schwächsten Ergebnisse zeigten, mit einem Recall zwischen 57,41 und 44,64 %.[58] Die Retrieval-Performance sank, wenn zusätzlich Begriffsgruppen oder -klassen über Synonyme oder Wortformen hinaus gebildet wurden. Die zufriedenstellendsten Ergebnisse wurden mit Einzelbegriffen in natürlicher Sprache, teilweise unter Zuhilfenahme von Wortformen und Synonymen erzielt. Dies bedeutete, dass Indexiersprachen nicht zwingend komplex oder kompliziert sein müssen, wenn auf natürlicher Sprache basierende Systeme genauso gute oder sogar bessere Ergebnisse

[55] Sachse, Liebig, Gödert, S.9.

[56] Cleverdon, Cyril u. Keen, Michael: ASLIB Cranfield Research Project. Factors Determining the Performance of Indexing Systems. Cranfield: Aslib, 1966. S.7.

[57] Zitiert bei Sachse, Liebig, Gödert S.9-10.

[58] Cleverdon , S.44.

erzeugen können.[59] Cleverdon und sein Team waren von diesem Ergebnis selbst überrascht und zweifelten zunächst die Richtigkeit dieses Ausgangs an, konnten aber keinen Gegenbeweis erbringen.[60] Cleverdons Hypothese nach Cranfield I über die inverse Beziehung zwischen Recall und Precision wurden durch die Ergebnisse von Cranfield II bestätigt.[61]

Erneut wurde die Relevanzbeurteilung dieses Projekts kritisiert. Die Studierenden hätten Dokumente, die für bestimmte Suchfrage relevant hätten sein können, nicht berücksichtigt, was die Ergebnisse negativ beeinflusst haben könnte. Außerdem wurde erneut in Frage gestellt, ob die Labor-Bedingungen des Experiments auf den bibliothekarischen Alltag angewendet werden könnten.[62]

Die Cranfield-Experimente stellen die Grundlage für empirische Retrievalforschung unter Laborbedingungen dar, bei denen alle Variablen unter starker Kontrolle der ProjektteilnehmerInnen stehen und Erkenntnisse über Retrievalsysteme im Allgemeinen Priorität besitzen. Alle Komponenten des Cranfield-Paradigmas sind in Cranfield II vorhanden: das Auswählen von Dokumenten für einen Testkorpus, die Ausarbeitung von Suchfragen an das System, der Prozess der Relevanzbewertung ohne Rückgriff auf Source-Dokumente, sowie das Messen der Ergebnisse anhand der Kennzahlen Recall und Precision.[63]

Die Aspekte der Interaktivität und NutzerInnenorientierung spielen bei den Cranfield-Retrievalexperimenten keine wesentliche Rolle.[64] An diesem Punkt setzte ein weiteres zeitgenössisches Retrievalprojekt an.

### 3.1.5 Medlars (1966, Abschlussbericht von 1968)

Das MEDLARS-Experiment wurde im gleichen Jahr wie Cranfield-2 durchgeführt. Die ausführende Einrichtung war die National Library of Medicine in Washington, D.C. Im Gegensatz zu den Cranfield-Experimenten handelte es sich beim MEDLARS-Experiment um die Evaluation des bereits bestehenden Systems MEDLARS (Medical Literature Analysis and Retrieval System) und der Suche nach Wegen, um die Retrievalperformance zu

[59] Rasmussen, Edie: Evaluation in Information Retrieval. In: „The MIR/MDL Evaluation Project White Paper Collection“. Edition #3. Workshop on the Evaluation of Music Information Retrieval Systems. Toronto, 2003. S.45.

[60] Cleverdon, S. 252.

[61] Sachse, Liebig, Gödert, S.10.

[62] Ebd.

[63] Robertson, S.44.

[64] Borlund, S.15.

verbessern. Dieses wurde nicht unter Labor-, sondern unter den praktischen Bedingungen des Bibliotheksalltags untersucht[65].

Die genauen Zielsetzungen lauteten wie folgt:

- die Suchanforderungen von MEDLARS-NutzerInnen zu typisieren und zu analysieren
- Effektivität und Effizienz der Erfüllung der NutzerInnenanforderungen zu bewerten
- Identifikation von Faktoren, die die System-Performance negativ beeinflussen
- Methoden aufzuzeigen, wie die Anforderungen der NutzerInnen effizienter und wirtschaftlicher erfüllt werden können[66]

Das MEDLARS-Experiment ist ein Zeitgenosse des Cranfield II Experiments, aber es bestehen sehr große Unterschiede im Experiment-Design und den Zielsetzungen des Projekts. Cyril Cleverdon gehörte zu einem auf Grundlage des Experiments gegründeten Komitee, welches das Design und die Ausführung des Projekts beobachten und beratend tätig sein sollte. F.W. Lancaster erwähnt im Abschlussbericht des Projekts Cleverdons Mithilfe in den Design- und Analyse-Phasen des Projekts mit besonderem Lob[67]. Zudem werden die in den Cranfield-Experimenten erstmals etablierten Kennzahlen Recall und Precision zur Darstellung der Ergebnisse in diesem Experiment genutzt.[68]

Das MEDLARS-Experiment nutzte die komplette Datenbank der National Library of Medicine mit ca. 700.000 Einträgen. 302 NutzerInnenanfragen wurden gesammelt und auf Grundlage des Bestandes bearbeitet.[69] Um diese Anfragen zu erhalten, kollaborierte die Bibliothek mit einer Reihe von Organisation, die regelmäßig vom MEDLARS-Bestand Gebrauch machten.[70]

Relevanzbeurteilungen wurden zunächst durch BibliotheksmitarbeiterInnen anhand der kompletten Texte vorgenommen. Diese Beurteilungen wurden im Anschluss durch die NutzerInnen, die die Informationsanfragen gestellt hatten, überprüft und ihrerseits bewertet. Die Relevanz wurde in abgestufter Form nach dem folgenden Schema bewertet:

[65] Borlund., S.16.

[66] Lancaster, F.W.: Evaluation of the MEDLARS Demand Search Service. January 1968. National Library of Medicine. Bethesda, MD, 1968. S. 8.

[67] Lancaster (1968), S.1.

[68] Ebd., S.16.

[69] Zitiert bei Borlund, S.17.

[70] Lancaster (1968), S.33.

„major“ „minor“ und „no value“. Alle teilnehmenden NutzerInnen wurden außerdem gebeten, einen Auszug des Medlars-Suchoutputs auf Relevanz zu überprüfen, zusammen mit ausgewählten Dokumenten aus anderen Quellen.[71]

Die Recall-Basis wurde anhand einer Anzahl als relevant eingestufter Dokumente vorgenommen, die vom System basierend auf den NutzerInnenanfragen ausgegeben worden waren. Hierbei handelte es sich um Dokumente, die den NutzerInnen bereits vorher bekannt waren oder über die Bibliotheks-Mitarbeiter schon vor dem Experiment Kenntnis hatten. Recall und Precision wurden anhand des Erschöpfungsgrades und der Spezifität der NutzerInnenanfragen nach dem kontrollierten Eingabeformular von MEDLARS berechnet.[72]

Als Durchschnittsergebnisse für alle bearbeiteten Anfragen ergaben sich Werte von ca. 58% für den Recall sowie ca. 50% für die Precision. Lancaster weist jedoch darauf hin, dass diese Werte im Zusammenhang der einzelnen Ergebnisse gesehen werden müssen, bei denen die Messzahlen z.T. weit über bzw. unter diesen Durchschnittszahlen liegen. Er betonte außerdem, dass die Durchschnittswerte zwar niedrig anmuteten, bezweifelte aber, dass ein anderes System bei einer so gründlichen Überprüfung wie bei MEDLARS ähnlich gute Ergebnisse erzielen würde.[73]

## 3.2 Projekte aus dem englischsprachigen Raum in den 1980ern und 1990ern

In diesem Kapitel werden zwei weitere Retrievaltests aus dem englischsprachigen Raum vorgestellt. Diese fanden in den späten 80er Jahren, bzw. in den frühen 90er Jahren statt und dienen als frühe Beispiele für bibliothekarische Retrievaltests mit für die NutzerInnen zugänglichen OPACs.

Das ADFA-Experiment von 1988 und F.W. Lancasters Retrievalprojekt an der Bibliothek der Universität von Illinois im Jahr1991 verbindet, dass beide die OPACs ihrer jeweiligen Einrichtungen hinsichtlich der Retrievalleistung bei der sachlichen Recherche und Auswirkungen der Anreicherung von Katalogdatensätzen mit zusätzlichem Wortmaterial untersuchten.

[71] Zitiert bei Borlund, S.17.
[72] Lancaster (1968), S.16-19
[73] Ebd., S.185.

Diese beiden Projekte werden hier zudem vorgestellt, weil sie wichtige Erkenntnisse für die Retrievalexperimente lieferten, die in der Folge im deutschsprachigen Raum durchgeführt wurden.

### 3.2.1 Das ADFA-Experiment (1988)

In der zweiten Hälfte der 1980er wurde an der Bibliothek der Australian Defence Force Academy (ADFA) in Canberra ein Projekt durchgeführt, um die sachliche Recherche im relativ neuen lokalen OPAC (Urica Library System) durch das Hinzufügen von mehr Sucheinstiegen zu verbessern. Dieser OPAC ermöglichte sachliche Recherchen zunächst nur per Schlagwort in den MARC 650er-Feldern oder über kontrollierte Library of Congress subject headings.[74] Das Personal der ADFA-Bibliothek identifizierte verschiedene Probleme mit der Benutzung des Urica-Systems, u.a. bei der sachlichen Recherche. Das richtige subject heading zu finden, das zum eigenen Informationsbedürfnis passt, war eine Herausforderung, die durch die Einführung einer Schlagwortkomponente nicht zufriedenstellend gelöst werden konnte. Eine Freitextsuche war nicht möglich, und die Schlagwörter konnten nur den LoCSHs entnommen werden. Bibliothekar Lynn Hard und sein Team suchten nach Lösungen, um die sachliche Recherche im Urica-System zu erweitern und einfacher zu gestalten. Sie wählten die Methode nach Pauline Atherton, die darin besteht, ergänzende Begriffe aus Inhaltsverzeichnissen und Indizes der zu erschließenden Werke zu extrahieren.[75]

Der Plan sah vor, dem Katalogeintrag jedes Buches, das diesem Prozess unterzogen wurde, im MARC-Feld 653 durchschnittlich 21 Multiwortbegriffe (zw. 20 und 25) hinzuzufügen, die über die Schlagwortsuche zu einer besseren Wiederauffindbarkeit der Werke führen sollten. Die dazu erforderlichen Daten wurden den jeweiligen Inhaltsverzeichnissen oder Indices entnommen. Der Prozess begann im Dezember 1986. Das Projekt erhielt den Namen Enhance Subject Projekt (ESP), und die den MARC-Datensätzen hinzugefügten Begriffe wurden als ESP terms betitelt.[76]

[74] Byrne, Alex; Micco, Mary: Improving OPAC Subject Access: The ADFA Experiment. In: College & Research Libraries 49 (5), 1988. S.432.

[75] Ebd., S.432-433.

[76] Ebd., S.434.

Das Ziel dieses Vorhabens bestand darin, den durchschnittlichen Recall im System deutlich zu steigern und die Precision bei sachlichen Recherchen zu verbessern ohne die Kosten zur Informationsbearbeitung und – aufbewahrung unnötig zu erhöhen.[77]

In einer ersten praktischen Projektphase durchliefen 6139 der insgesamt ca. 160.000 Bände[78] aus dem Bestand der ADFA-Bibliothek diesen Prozess. Die Zahl der keywords im Urica-OPAC stieg dabei von ursprünglich ca. 30.000 auf über 101.000 durch hinzugefügte Titelstichwörter und ESP terms.[79]

Diese waren den Bibliotheks-NutzerInnen zugänglich, die positives Feedback zu dieser neuen Möglichkeit der sachlichen Recherche abgaben. Besonders hervor gehoben wurde die Tatsache, dass die NutzerInnen dank der zusätzlichen Begriffe die Relevanz eines bestimmten Werkes für ihre Informationsbedürfnis besser einschätzen konnten, da die ESP terms auf der Terminologie der AutorInnen basieren und Zugang zu Begriffen auf Kapitelebene gewähren.[80]

Durch eine quantitative Analyse sollte ermittelt werden, ob die ESP terms tatsächlich den Recall erhöhten oder eher Ballast für das System darstellten. Dazu wurden Suchen mit insgesamt 200 Schlagwörtern (Titelstichwörter, LoC subject headings und ESP terms) durchgeführt. Bei einem Testlauf mit Werken aus dem Bereich Computerwissenschaften wiesen die ESP terms einen Recall von ca. 72% auf, gegenüber einer Retrievalrate von ca. 10% bei den Titelstichwörtern und ca. 15% über die LoC subject headings. Effektiv bedeutete dies, dass das Hinzufügen von Schlagwörtern aus Inhaltsverzeichnissen und Indizes die Retrievalleistung um ca. 300% verbesserte. Der Recall war durch die ESP terms unzweifelhaft erhöht worden. Im nächsten Schritt sollte sich erweisen, ob dies auch für die Precision galt, mit besonderer Sicht auf die Zahl irrelevanter gefundener Dokumente.[81]

Experten wählten 31 Begriffe aus dem Bereich Computerwissenschaften aus, mit denen in natürlicher Sprache gesucht wurde. 13 von diesen Begriffen korrespondierten mit den LOC subject headings, 18 (58%) taten dies nicht. Die Experten wurden im Anschluss gebeten, jedes auch nur annähernd relevante Dokument als positives Ergebnis zu werten. Bei Suchen mit LoCshs wurden nur 12,5% irrelevante Dokumente gefunden, während

[77] Byrne u. Micco, S.434
[78] Ebd., S.432.
[79] Ebd., S.435.
[80] Ebd., S.434.
[81] Ebd., S.436.

Titelstichwörter und ESP terms kombiniert nur 16,8% der relevanten Dokumente fanden.[82] Die 18 Begriffe ohne Bezug zu den LoCshs fanden 318 Dokumente, von denen 220 bzw. 69% als relevant bewertet wurden. Diese hohe Precision wurde auf die Anwesenheit der Titelstichwörter zurückgeführt.

Diese quantitative Studie zeigte deutlich, dass der Einsatz von Begriffen aus Inhaltsverzeichnissen und Indizes bei der sachlichen Recherche eine machbare und kosteneffektive Methode darstellen, um die Zahl von möglichen Sucheinstiegen und die Retrievalleistung deutlich zu erhöhen, ohne dass der Ballast-Anteil in den Ergebnislisten allzu stark ansteigt. Zugleich wurde auch die Nützlichkeit der LoCshs unterstrichen, die daher nicht abgeschafft, sondern nach Auffassung der ADFA-Bibliotheksmitarbeiter verbessert werden sollten.[83]

### 3.2.2 F.W. Lancasters Retrievalexperiment in der Bibliothek der University of Illinois

F.W. Lancaster, der am MEDLARS-Experiment beteiligt gewesen war und mehrere Kolleginnen untersuchten anhand des OPACs der Bibliothek der University of Illinois, ob der neue digitale Bibliothekskatalog tatsächlich eine Verbesserung in der Retrievalleistung gegenüber dem vorherigen Karteikartenkatalog darstellte. Vor allem bei der sachlichen Recherche böten sich bei einem digitalen Katalog die Möglichkeit wesentlich bessere Ergebnisse bei der sachlichen Suche zu erzielen, z.B. durch die Anreicherung von Datensätzen mit zusätzlichen Sucheinstiegen. Lancaster et al. beziehen sich dabei explizit auch auf das Experiment von Byrne und Micco (s.o.).[84]

Sie gingen davon aus, dass der Erfolg einer sachlichen Recherche sich nur in feinen Abstufungen messen lässt. Demnach kann eine sachliche Suche in einem Bibliothekskatalog laut Lancaster et al. nur dann als erfolgreich angesehen werden, wenn Katalognutzer die Materialien ausfindig machen, die ihr Informationsbedürfnis aufgrund ihrer Aktualität, Vollständigkeit oder Autorität genau wiedergeben und eins oder mehrere Items aus den Suchergebnissen auswählen und entleihen.[85]

Lancaster et al. formulierten die Zielsetzung, mit Hilfe einer Retrievalstudie die Wahrscheinlichkeit zu messen, dass erfahrene KatalognutzerInnen in einem Bibliothekssystem

[82] Byrne u. Micco, S.436-438.

[83] Ebd., S.440.

[84] Lancaster, F.W; Harkness Connell, Cora; Bishop, Nancy et al: Identifying Barriers to Effective Subject Access in Library Catalogs. In: Library Resources and Technical Services 35 (4), 1991. S.377.

[85] Ebd., S.379.

die angemessensten Materialien für ihre Informationsbedürfnisse finden. Falls das nicht gelänge, sollten die nötigen Veränderungen am System identifiziert werden, um den Erfolg in Zukunft sicher zu stellen.[86]

Die ProjektmitarbeiterInnen erstellten auf der Grundlage von Spezialbibliografien insgesamt 51 Topics aus Leselisten mit relevanten Werken zu verschiedenen Sachthemen wie „noise hazards to humans“ oder „political music“.[87] Dies geschah in der Annahme, dass die Literatur in diesen Listen wahrscheinlich im Katalog einer Forschungsbibliothek vorhanden sein würde. Ein thematischer Schwerpunkt lag in Literatur aus den Sozialwissenschaften.[88]

Für jede dieser 51 Listen wurden die folgenden Schritte unternommen:

1. Zeitschriftenaufsätze wurden eliminiert, da diese in den frühen 90ern noch nicht in Bibliothekskatalogen erschlossen wurden.

2. Zu dem betreffenden Thema wurde im kompletten Katalog der University of Illinois aka Full Bibliographic Record (FBR) eine Suche durchgeführt, der zu dieser Zeit ca. 4,5 Mio. bibliographische Aufzeichnungen umfasste.

3. Für Materialien der Bibliografie, die nicht über eine sachliche Recherche aufzufinden waren, wurden im FBR AutorInnen- oder Titelsuchen durchgeführt.

4. Wenn als relevant eingestufte Titel nicht bei der sachlichen Suche gefunden wurden, wurde analysiert, woran dies liegen könnte und wie die Suchstrategie geändert werden müsste. Viele dieser nicht gefundenen Titel konnten erst lokalisiert werden, nachdem ihre bibliographischen Aufzeichnungen um Inhaltsverzeichnisse, Indizes oder Volltexte ergänzt worden waren.[89]

Aus den 51 Leselisten waren 607 Titel im FBR vorhanden. 327 von diesen wurden während des Experiments gefunden, was einem Recall von 53,9 % entspricht. Der durchschnittliche Recall für alle 51 Suchvorgänge lag bei 59,4 %. Diese Ergebnisse enthielten 8 Suchen mit einem Recall von 100% und zwei Nullsummen-Ergebnisse. Dies erscheint oberflächlich betrachtet als anständiges Ergebnis[90]. Die Autoren der Studie waren damit jedoch nicht zufrieden, da die Suchen im FBR von Studierenden der

[86] Lancaster (1991)., S.379.
[87] Die komplette Liste ist bei Lancaster (1991) auf S.380-381 zu finden.
[88] Ebd., S.379.
[89] Ebd., S.379 u. 382.
[90] Ebd., S.382

Bibliothekswissenschaft mit Expertenkenntnissen des Systems durchgeführt wurden. Dies geschah mit Blick auf den höchstmöglichen Recall ohne Berücksichtigung der Precision. Beim Thema „feminist methodology“ z.B. wurden 90% Recall erreicht, aber die Ergebnismenge belief sich auf 1200 Dokumente, die zum größten Teil für dieses Suche irrelevant waren. Mit spezifischeren Suchbegriffen dagegen wäre der Recall deutlich geringer ausgefallen. Außerdem wurden diese Suchen unter Laborbedingungen durchgeführt. Unter „realen“ Bedingungen wären diese Ergebnisse nicht zustande gekommen, da es als unwahrscheinlich angesehen wurde, dass sich BibliotheksnutzerInnen hunderte von Dokumenten ansehen, um eine kleine Zahl relevanter Titel zu finden. Laut der Studie könnten NutzerInnen ohne Vorkenntnisse einen solch hohen Recall nicht erreichen.[91]

Die AutorInnen gaben darüber hinaus zu bedenken, dass die 59% Recall nur einen Teil der Literatur abbilden, da Zeitschriftenartikel im Vorfeld schon nicht berücksichtigt wurden. Sie sahen es darüber hinaus als nicht sinnvoll an, andere Teile der bibliografischen Aufzeichnungen jenseits der subject headings in die Suche einzubeziehen, da dies den Recall nicht wesentlich erhöht hätte. Die AutorInnen erklärten dies mit der Tatsache, dass die zugewiesenen subject headings bereits nahe an der Terminologie der Titel lägen und daraus kein Zugewinn an Informationen entstanden sei. Selbst unter Einbezug von Titelstichwörtern und allen irgendwie für eine Suchanfrage relevanten subject headings, wäre kein höherer Recall als 63,9 % möglich gewesen[92]. Lancaster et al. sahen wenig Chancen, mit den zeitgenössischen bibliographischen Aufzeichnungen und Katalogisierungspraktiken die Situation deutlich zu verbessern. Eine typische Katalogaufzeichnung hätte zu wenig Zugangspunkte, als dass ein akzeptabler Recall durch den Einsatz von Begriffskombinationen erreicht werden könnte. Es bestehe ein großer Unterschied zwischen einer Aufzeichnung, der 2-3 subject headings zugewiesen wurden, und einer mit 10-12 Deskriptoren und einem Abstract. Als Möglichkeit der Verbesserung nennen die AutorInnen weitreichender strukturierte subject headings, sodass Verbindungen zwischen ihnen hergestellt und tiefergehende subheadings genutzt werden. So könnten die Themen von Dokumenten präziser abgebildet werden als mit dem bisherigen Ansatz.[93]

Am Ende der Studie kommen Lancaster et al. zu folgendem Ergebnis: von ExpertInnen vorgeschlagenes Lesematerial zu einem Thema besteht aus wichtigen Dokumenten, die bei einer sachlichen Recherche in einem Bibliothekskatalog gefunden werden sollten.

[91] Lancaster (1991), S.384.
[92] Ebd., S.385.
[93] Ebd., S.386.

BibliotheksnutzerInnen sollten diese vorzugsweise vor anderen Dokumenten im Katalog finden können. Tatsächlich ließen Bibliothekskataloge, inklusive der damals neuen OPACs, nur oberflächliche Themensuchen zu. Sie ließen NutzerInnen dadurch im Stich, dass sie nur Zugang zu einem kleinen Teil der relevanten Literatur gewähren, der nicht einmal unbedingt die besten verfügbaren Werke enthalten muss.[94]

Entgegen dem allgemeinen Glauben hätte der Übergang vom Karteikartenkatalog zum OPAC laut Lancaster sachliche Recherchen nicht signifikant verbessert oder erleichtert. Stattdessen könnte dank der viel größeren Mengen von Dokumenten in den Katalogen die Lage eher verschlechtert werden. Die Ergebnisse der Studie machten sichtbar, dass signifikante Verbesserungen innerhalb der Einschränkungen existierender thematischer Katalogisierungspraktiken nicht möglich sei und daher die Konsultation von ExpertInnen, bzw. von Fachbibliografien und Referenzwerken bessere Erfolge verspreche als die Suche in Datenbanken oder Bibliothekskatalogen.[95]

Am Ende der Studie steht die sehr pessimistische Perspektive, dass Kataloge mit Zugang zu mehreren Millionen Dokumenten nie mehr als krude Werkzeuge für die sachliche Suche sein könnten.[96] Die folgenden Beispiele werden zeigen, dass dem nicht unbedingt so sein muss, auch wenn der Prozess, sachliche Recherchen in Bibliothekskatalogen zu verbessern, immer noch andauert.

## 3.3 Bibliothekarische Retrievalexperimente in Deutschland in den 1990ern und frühen 2000er Jahren

Die bis hierhin vorgestellten Retrievalexperimente fanden im englischsprachigen Raum statt. Es sollte bis in die 1990er dauern, bis in Deutschland Retrievaltests in Bibliotheken durchgeführt wurden. Diese waren im Dokumentationsbereich bereits hinreichend etabliert, doch im Bibliotheksbereich fehlte es an theoretischem wie praktischem Grundwissen. Daraus resultierte ein Mangel an empirischen Daten zu Erschließung und Retrieval.[97] Dies sorgte für Unsicherheit hinsichtlich der Retrievalleistung von OPACs in dt. Bibliotheken. Klaus Lepskys Retrievalexperimente MILOS I, MILOS II und KASCADE, die zwischen 1994 und 2000 durchgeführt wurden, verbesserten diese Situation. Die MILOS-

[94] Lancaster (1991), S.387-388.

[95] Ebd., S.388.

[96] Ebd., S.389.

[97] Lepsky, K., J. Siepmann u. A. Zimmermann: Automatische Indexierung für Online-Kataloge – Ergebnisse eines Retrievaltests. In: Zeitschrift für Bibliothekswesen und Bibliographie 43 (1), Januar/Februar 1996. S. 47-48.

Experimente erwiesen sich als einflussreich und regten die Durchführung weiterer Retrievaltests mit anderen Forschungsansätzen an. Vor MILOS I führte Lepsky eine Machbarkeitsstudie an der ULB Düsseldorf durch, um die Möglichkeit eines Retrievaltests auszuloten. Dies sollte als erster Schritt einer Kooperation von InformationswissenschaftlerInnen und BibliothekarInnen dienen.[98] Bei der Vorbereitung der Studie wurden u.a. auch die oben vorgestellten Projekte von F.W. Lancaster[99] und Byrne u. Micco bzgl. der Anreicherung von Titelaufnahmen mit Titelstichwörtern, Inhaltsverzeichnissen und Indices zitiert.[100]

### 3.3.1 MILOS I (1994)

Ab den 1990ern widmeten sich Klaus Lepsky und andere Vertreter aus den Informations- und Bibliothekswissenschaften einer Problematik in dt. Bibliotheken, die bis heute existiert. Zu wenige Titel sind in Bibliothekskatalogen intellektuell sachlich erschlossen. Als Beispiel diente Lepsky die ULB Düsseldorf mit seinerzeit ca. 35% sachlich erschlossenen Medien. Zudem wurden nicht genug Schlagwörter pro Titel vergeben, sodass sich bei sachlichen Recherchen geringe Treffermengen oder Nulltreffermengen ergeben.[101] Dies sorgt dafür, dass Stichwortretrieval hinzugezogen werden muss. Dies verursacht weitere Probleme, da Titelstichwörter häufig in flektierten Formen vorliegen oder Teile von Komposita sind.[102]

Um sich dieses Problems anzunehmen, wurde das DFG-geförderte Projekt MILOS I (Maschinelle Indexierung zur verbesserten Literaturerschließung in Online-Systemen) gestartet. MILOS I fand über das Jahr 1994 statt. Die Fachrichtung Informationswissenschaft der Universität des Saarlandes und die ULB Düsseldorf nahmen daran teil. Zum Einsatz kam die Software IDX, entwickelt von Harald Zimmermann, als System zur automatischen Indexierung von Stich- und Schlagwörtern für die sachliche Erschließung. Für das

[98] Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen. Heft 18. Köln: Greven Verl., 1994. S.4.

[99] Ebd., S.8.

[100] Ebd., S.20.

[101] Lepsky, K.: Automatisierung in der Sacherschließung: Maschinelles Indexieren von Titeldaten. S. 223.

[102] Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Zukunft der Sacherschließung im OPAC. Schriften der Universitäts- und Landesbibliothek Düsseldorf. Band 25. Hrsg. von Elisabeth Niggemann. 1996. S. 15.

Experiment wurde IDX modifiziert, um in Bibliotheken zum Einsatz kommen zu können. IDX basiert auf Wörterbüchern, die regelmäßiger Pflege bedürfen.[103]

Die Datengrundlage für das Experiment war der gesamte maschinenlesbare Titelbestand der USB Düsseldorf. Diese 800.000 Titel wurden in IDX indexiert. Die Überprüfung der Verbesserungen, die durch automatische Indexierung erzielt werden sollten, erfolgte anhand von 40.000 Titeldaten in einer BISMAS-Datenbank, die differenziertes Suchen über Freitexte und Indices gestattete. Eine fachliche Selektion fand hierbei nicht statt. Deskriptoren, die aus dem Indexierungsvorgang entstanden waren, wurden direkt den Titeln zugespielt.[104] Drei Indices für Suchanfragen wurden in BISMAS angelegt, da die Software diese für Suchen benötigt: Index 1 enthielt nur Titelstichwörter, Index 2 Titelstichwörter und Deskriptoren aus der automatischen Indexierung und Index 3 beinhaltete Titelstichwörter und die lokalen verstichworteten Schlagwörter der ULB Düsseldorf.[105]

50 Topics wurden erarbeitet, die an die tägliche Suchpraxis angelehnt und auf die Retrievalproblematik der automatischen Indexierung abgestimmt waren. Die Berücksichtigung verschiedener Fachgebiete und die Komplexität der Suchbegriffe (Mehrwortbegriffe und Komposita) spielten eine wichtige Rolle. Einfache Suchbegriffe bildeten eine Ausnahme. Alle 50 Fragen kamen bei insgesamt 4 Suchen zum Einsatz, mit 200 Suchvorgängen insgesamt:

- Freitextretrieval mit Trunkierung
- Suche im reinen Titelstichwortindex
- Suche in Titelstichwortindex und Indexierungsergebnissen
- Suche in Titelstichwortindex mit Schlagwörtern[106]

Aus den Trefferzahlen wurden Recall und Precision sowie der Einheitswert nach van Rijsbergen ermittelt, der die beiden Messzahlen zueinander in Beziehung setzt und die Gewichtung der Ergebnisse ermöglicht. Aufgrund niedriger Recall-Werte wurde diese Kennzahl mit dem Faktor 2 gewichtet, also doppelt so hoch bewertet wie die Precision.[107]

[103] Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. S.16.

[104] Ebd., S.23-24.

[105] Ebd., S. 27.

[106] Ebd.

[107] Ebd., S. 26.

Als Zielzahl des Retrievaltests wurde die Zahl der Dokumente festgelegt, die mit allen verfügbaren Mitteln in der BISMAS-Datenbank gefunden werden konnten.[108] Das Einspielen von Deskriptoren, die durch automatische Verfahren aus Titelstichwörtern und verstichworteten Schlagwörtern extrahiert worden waren, hatte eine deutliche Verbesserung der Suchergebnisse zur Folge. Die Indizes Titelstichwörter und Indexierungsergebnisse sowie Titelstichwörter und Schlagwörter erzielten vergleichbar gute Ergebnisse, mit derselben Precision bei höherem Recall.[109]

#### 3.3.1.1 Ergebnisse

Insgesamt wurden in der Datenbank 876 für die Suchfragen relevante Dokumente gefunden. Die Einzelwerte lagen zwischen 1 und 244 Dokumenten. Im Mittel ergeben sich daraus 17,52 relevante Dokumente pro Frage.

- Die Suche in Index 1 (nur Titelstichwörter) erbrachte 136 Dokumente, von denen 109 relevant waren. Der durchschn. Recall lag bei 14%, während die Precision mit 59% deutlich höher lag. Für die sachliche Recherche im untersuchten OPAC war die Titelstichwortsuche allein nicht geeignet.
- Im Index 2 (Titelstichwörter und Indexierungsergebnisse) wurden 568 Dokumente gefunden, von denen 391 als relevant gezählt wurden. Der Recall lag im Mittel bei 51%, die Precision bei 83%.
- Im dritten Index (Titelstichwörter und Schlagwörter) wurden 302 Dokumente gefunden, von denen 270 relevant waren. Der Recall betrug im Mittel 39%, die Precision lag wie bei Index 2 bei 83%.[110]

Die Ergebnisse zeigten, dass eine Steigerung des Recalls nicht zwingend mit niedriger Precision einher geht. [111]

#### 3.3.1.2 Schlussfolgerungen

Verfahren zur automatischen Indexierung können zuverlässige Ergebnisse liefern. Allerdings gab Lepsky zu bedenken, dass die automatische Indexierung die intellektuelle Methode noch nicht ersetzen, aber ergänzen konnte.[112] Dies wurde ausdrücklich als

[108] Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. S.26-27

[109] Lepsky, Automatisierung in der Sacherschließung, S.227.

[110] Lepsky, Siepmann, Zimmermann, S.54-55.

[111] Lepsky, Automatisierung in der Sacherschließung, S.227.

[112] Ebd., S.227-28

erstrebenswert betitelt wie auch die Vermeidung von doppelter Erschließungsarbeit. Die Ressourcen einer Bibliothek, um zwei Indexierungsmethoden zu erhalten, sind begrenzt und müssen daher gut koordiniert werden.[113] Als primäre Zielsetzung für die automatische Indexierung gibt Lepsky die Verbesserung des Stichwortretrievals in Titelstichwortdatenbanken an: „Je höher die die Quote nicht intellektuell erschlossener Titel in solchen Datenbänken ist, desto wichtiger ist die Verbesserung des Zugangs auf Titelstichwortbasis“.[114]

Nachdem MILOS I primär auf die grammatikalischen Funktionen fokussiert war, stand die Möglichkeit einer semantischen Indexierungskomponente im Raum. Dies wurde mit dem Nachfolge-Experiment MILOS II untersucht.[115]

### 3.3.2 MILOS II (1995-96)

Nach Abschluss des Projekts MILOS I begann nach weniger als einem Jahr die Arbeit am direkten Nachfolgeprojekt. MILOS II war größer angelegt als der Vorgänger. Die wissenschaftliche Bearbeitung erfolgte durch Mitglieder des Fachbereichs Bibliotheks/Informationswesen der FH Köln (heute TH Köln) in Kooperation mit dem Fachbereich Informationswissenschaften der Universität des Saarlandes.

Das Projekt begann im November 1995 und endete im August 1996. Die IDX-Software zur automatischen Indexierung aus MILOS I wurde „durch Einbindung von Thesaurusrelationen der Schlagwortnormdatei (SWD) in das Wörterbuchkonzept um semantische Funktionalitäten erweitert“.[116]

#### 3.3.2.1 Retrievaltestkomponenten

Die Datengrundlage für das Experiment belief sich auf 190.000 Titel aus dem DB-Katalog (heute DNB), die zwischen 1990 und 1995 erschienen waren. Es handelte sich dabei um Literatur aus allen Sachgruppen. Belletristik-Titel, Kinder- und Jugendliteratur sowie Kalender wurden nicht für den Test verwendet. Die Titeldaten wurden durch Deskriptoren erschlossen, die über die automatische Indexierung durch die MILOS-Sofware entstanden waren. Jegliche Bearbeitung des Textmaterials erfolgte im Abgleich mit den elektronischen Wörterbüchern in IDX.[117]

[113] Lepsky, Automatisierung in der Sacherschließung., S.230
[114] Ebd., S.228.
[115] Lepsky (1996), Ergebnisse des DFG-Projekts MILOS I, S.23.
[116] Sachse, Liebig, Gödert, S.6.
[117] Ebd., S.16-17

Für MILOS II wurden 100 Suchanfragen eingesetzt. 50 davon waren speziell für dieses Projekt erarbeitet worden, während die übrigen 50 aus MILOS I übernommen worden waren. Wiederum sollten diese Anfragen der alltäglichen Arbeit mit dem OPAC entnommen sein. Sie deckten Sachrecherchen verschiedener Komplexität ab, von einfachen Sachverhalten mit einem oder mehreren Suchbegriffen hin zu Fragen mit Mehrwortverbindungen. [118]

Die Relevanzbeurteilung wurde binär und großzügig vorgenommen. Alle Titel, die nicht von vornherein als irrelevant für eine Anfrage erschienen, wurden als relevant gewertet, da ein Interesse vermutet werden konnte, sich das Originaldokument genauer anzusehen.[119]

Die Anwendung der Kennzahlen Recall und Precision erwies sich bei MILOS II als problematisch. Zur Berechnung des Recall müsste die Menge relevanter Dokumente für spezielle Fragen in der Datenmenge bekannt sein. Dafür war die Testkollektion deutlich zu groß. Die Projektgruppe verzichtete darauf, annähernde Werte zu schätzen, da nicht sicherzustellen war, dass dies für alle 100 Suchanfragen gleich gut gelingen würde. Eine Errechnung des Recall-Wertes war also nicht möglich. Die quantitative Analyse setzte daher die Angabe der Werte der gefundenen Titel in Beziehung zu den relevanten Dokumenten in der Treffermenge, um so die Precision zu messen.[120]

Die automatische Indexierung lief in zwei Schritten ab. Im ersten Indexierungslauf liefen die Sprachselektion der Titel und Unterteilung zwischen der deutschen und anderen unterstützten Sprachen ab, sowie eine Rechtschreibkontrolle, Wörterbuchpflege und der Indexierungsvorgang an sich. In einem zweiten Lauf wurden Fehler identifiziert und korrigiert. Darauf aufbauend wurden alle neu gebildeten Wörter erkannt und den passenden Datensätzen in den Kategorien IDX (Titel) und IDX (RSWK) hinzugefügt.[121]

Ähnlich zu MILOS I wurden die folgenden 5 Register als Suchumgebungen erstellt. Dies geschah bei MILOS II in der Bibliothekssoftware Allegro, um differenzierte Suchen zu ermöglichen, denn IDX selbst besitzt keine Datenbank- und Retrievalfunktionen. So

[118] Sachse, Liebig, Gödert, S.20-21.
[119] Ebd., S.30-31.
[120] Ebd., S.29.
[121] Ebd., S.18.

ließen sich die Indexierungsdaten an vorhandene Titeldaten anbinden, und die individuelle Generierung von Suchregistern war ebenfalls möglich.[122]

- Titelstichwörter
- Verstichwortete RSWK-Ketten
- Indexierungsergebnisse
- RSWK-Ketten ohne Stichwörter
- Basic Index mit Titelstichwörtern, RSWK-Ketten und Indexierungsergebnissen[123]

Auf dieser Grundlage wurde ein Retrievaltest durchgeführt, um die Qualität der Indexierungsergebnisse in Recherchesituationen mit Hilfe der vorbereiten Suchfragen zu bewerten. Als Qualitätskriterium galt der Umfang der Treffermenge und ihre inhaltliche Präzision. MILOS II hatte dabei nicht das Ziel, einen Vergleich zwischen den RSWK-Daten und den automatischen Indexierungsergebnissen hinsichtlich der Überlegenheit einer Methode gegenüber der anderen anzustellen. Es ging um die Frage, ob automatische Indexierung sachliche Retrievalvorgänge verbessern konnte, wenn Daten aus intellektueller Indexierung dabei genutzt werden.[124]

#### 3.3.2.2 Ergebnisse

Index 3 (Indexierungsergebnisse) erzielte gegenüber den Titelstichwörtern eine dreimal so hohe Zahl relevanter Treffer, brachte allerdings auch mehr Ballast in die Ergebnismengen. Die Retrievalergebnisse waren in diesem Register stark vom Titelmaterial abhängig, aber dies war nicht der einzige Faktor. Der Ballastanteil wurde mit den Formulierungen der Suchanfragen begründet.[125] Die verstichworteten RSWK-Ketten fanden doppelt so viele relevante Dokumente gegenüber dem reinen Titelstichwortindex. Die Suche mit unbehandelten Titelstichwörtern erbrachte keine akzeptablen Ergebnisse, da diese rein vom Titelmaterial abhingen.[126]

[122] Sachse, Liebig, Gödert., S.24.

[123] Ebd., S.25

[124] Gödert, Wilfried; Liebig, Martina: Maschinelle Indexierung auf dem Prüfstand. Ergebnisse eines Retrievaltests zum MILOS II Projekt. In: Bibliotheksdienst 31 (1), 1997. S.60.

[125] Sachse, Liebig, Gödert, S.32.

[126] Ebd., S.33-34.

Die Suche über die RSWK-Ketten ohne Stichwörter erbrachte keine zufriedenstellenden Ergebnisse. Der Grund dafür wurde in den engen Ansetzungsregeln für diese Schlagwörter gesehen, die die NutzerInnwn nicht vollständig kennen. Dank exakten Übereinstimmungen in den Ergebnismengen kam teilweise allerdings eine höhere Precision als bei der automatischen Indexierung zustande.[127]

Diesen Zugewinn an relevanten Titeln erklären sich die ProjektteilnehmerInnen nicht rein durch Ballast-reiche Treffermengen. Die Precision-Werte lagen bei 0,82 für die Titelstichwörter, 0,75 für die automatischen Indexierungsergebnisse trotz des Ballastanteils (also nur geringfügig niedriger gegen dem reinen Titelstichwortregister bei dreimal so vielen gefundenen relevanten Titeln!) und 0,95 für die verstichworteten RSWK-Ketten. Nur 3 von 100 Suchfragen mit Daten aus automatischer Indexierung erzielten Nulltrefferergebnisse, gegenüber 15 bei den Titelstichwortdaten und 30 bei den RSWK-Ketten. Dies lässt den Schluss zu, dass maschinelle Indexierungsmethoden einen wichtigen Beitrag zur Eliminierung von ergebnislosen sachlichen Recherchen liefern können.[128]

#### 3.3.2.3 Schlussfolgerungen

Automatische Indexierung ermöglicht bessere Retrievalmöglichkeiten dank des für das Retrieval zur Verfügung stehenden Vokabulars. Die Ergebnisse zeigen durchgehend einen höheren Recall bei zugleich nur geringfügig niedrigerer Precision. Die unterschiedlichen Ergebnismengen lassen erkennen, dass bei Suchen mit IDX-Stichwörtern oft mehr relevante Dokumente gefunden werden als mit Hilfe der intellektuell erstellten RSWK-Schlagwörter.[129]

Die automatische bzw. maschinelle Indexierung kann die Erschließung bibliographischer Daten und den Aufbau sachlicher Abfragekomponenten bedeutend ergänzen. Diese Form der Indexierung lässt sich theoretisch unbegrenzt wiederholen und nachträglich verbessern. Sie löst allerdings nicht alle Probleme des sachlichen Retrievals.[130]

Automatische Indexierung, die die intellektuelle Erschließung unterstützt, ist so lange einer rein intellektuellen Erschließung überlegen, bis Retrievalsysteme mit gleich guten Retrievaleigenschaften aufbauend auf intellektueller Indexierung entwickelt werden.[131]

[127] Sachse, Liebig, Gödert, S.35-36.
[128] Gödert u. Liebig, S.63.
[129] Sachse, Liebig, Gödert, S.37.
[130] Gödert u. Liebig, S.65.
[131] Ebd, S.66.

Die MILOS-Experimente waren die ersten Projekte ihrer Art in Deutschland und führten zu einer Reihe von weiteren Retrievalprojekten, die sich daran orientierten. Die Nachnutzung des Systems war leider nur begrenzt. Die ULB Düsseldorf verwendete IDX eine Zeit lang, bis ein Wechsel zur Software Aleph stattfand. Außerdem setzte die Bibliothek der Friedrich-Ebert-Stiftung in Bonn IDX zur inhaltlichen Erschließung von Zeitschriftenaufsätzen ein.[132]

### 3.3.3 Projekt KASCADE (2000)

Das Projekt KASCADE baute auf den Erkenntnissen aus MILOS I und II auf und fügte IDX weitere Funktionen hinzu, um die Leistung der automatischen Indexierung zu verbessern.

Die MILOS-Experimente hatten gezeigt, wie sich die Mängel der intellektuellen Erschließung in OPACs beheben lassen, nämlich durch die Anreicherung der Aufnahmen durch zusätzliches, „informationsdichtes“[133] Begriffsmaterial. Hier wird wieder explizit auf die Experimente von Lancaster sowie Byrne & Micco verwiesen.[134] Es hatte sich allerdings ergeben, dass die Textbasis von Titelaufnahmen eine nicht ausreichende Grundlage bietet, da sich abhängig vom vorhandenen Material große qualitative Unterschiede für die automatische Indexierung herausbilden können. Daraus ergab sich das Vorhaben, das MILOS-Verfahren zur Verarbeitung von Sprachdaten auf der Basis von umfangreicheren Textmengen auszubauen. Inhaltsverzeichnisse, Indices und ggfs. Abstracts sollten genutzt werden, um bibliothekarische Datensätze anzureichern. Darauf aufbauend folgte die automatische Indexierung. diesmal mit einer statistischen Komponente zur Gewichtung der so gewonnenen Begriffe, damit nicht zu viele irrelevante Begriffe als Deskriptoren verwendet wurden.[135] Diese Komponente wurde SELIX (selektive automatische Indexierung) genannt und speziell für KASCADE entwickelt. Sie berechnet per statistischem Verfahren Termgewichte für alle aus der Indexierung gewonnenen Begriffe und sorgt so

[132] Grummann, Martin: Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? In: Bibliothek 24 (3), 2000. S.302.

[133] Lohmann, Hartmut: KASCADE: Dokumentanreicherung und automatische Inhaltserschließung. Projektbericht und Ergebnisse des Retrievaltests. Schriften der Universitäts- und Landesbibliothek Düsseldorf 31. Düsseldorf: ULB Düsseldorf, 2000. S. 17.

[134] Ebd.

[135] Lepsky, Klaus; Zimmermann, Harald H.: Katalogerweiterung durch Scanning und automatische Dokumenterschließung. Ergebnisse des DFG-Projekts KASCADE. In: Zeitschrift für Bibliothekswesen und Bibliographie 47 (4), 2000. S.1-2 der Onlineausgabe ohne Seitenzahlen. https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/25554/1/2000a.pdf (Letzter Zugriff: 12.07.2022)

dafür, dass nicht alle neuen Deskriptoren unkontrolliert den Titelsätzen zugespielt werden.[136]

#### 3.3.3.1 Der KASCADE-Retrievaltest

Die Daten für das Experiment entstammten dem Jura-Fachbestand der ULB Düsseldorf mit einem Umfang von 30.000 Titeln. Die inhaltsrelevanten Daten von 3000 Titeln wurden durch ein langwieriges Scanningverfahren gewonnen und in die Testumgebung importiert. Abstracts und Indices erwiesen sich als untauglich und wurden ausgeklammert. Nur das Sprachmaterial aus den Inhaltsverzeichnissen wurde für die Anreicherung der Datensätze verwendet.[137]

Statt Allegro wurde die Freitext-Retrieval-Software askSAM für den Retrievaltest genutzt.[138] In Kooperation mit MitarbeiterInnen der juristischen Fakultät der Heinrich-Heine-Universität Düsseldorf wurden 73 Suchanfragen formuliert.[139] Von diesen wurden 13 vorab eliminiert, da sie aufgrund von Nulltrefferergebnissen unbrauchbar waren. Die Ergebnisse wurden durch juristische Fachleute überprüft. Insgesamt wurden mit allen 60 Fragen 1421 Dokumente gefunden, von denen 873 als relevant eingestuft wurden.[140]

Analog zu MILOS II wurden verschiedene Indices mit unterschiedlichem Wortmaterial für die Anreicherung angelegt:

- IDXTITEL mit in IDX automatisch indexierten Titel- und Schlagwörtern. Gefunden wurden 59 Titel mit 57 relevanten Treffern. Der Recall lag mit 0,06 inakzeptabel niedrig, während die Precision mit 0,98 sehr hoch ausfiel. IDXTITEL lieferte niemals die meisten, sondern fast immer die wenigsten relevanten Treffer. Die reine Suche über Titelstichwörter und Schlagwörter ist wenig erfolgversprechend, selbst wenn sie automatisch indexiert sind.[141]
- FREITEXT mit Titel- und Schlagwörtern und nichtindexierten Volltexten der Inhaltsverzeichnisse. Hier wurden 689 Dokumente gefunden. 457 davon waren relevant. Mit einem Recall von 0,54 und einer Precision von 0,75 ist schon durch

[136] Lohmann, S.18.
[137] Ebd., S.19
[138] Ebd., S.63.
[139] Ebd., S.60.
[140] Ebd., S.70-71
[141] Ebd., S.75.

die Anwesenheit unbehandelter Inhaltsverzeichnistexte eine deutliche Verbesserung erkennbar. Allerdings entstanden 7 Nulltrefferergebnisse.[142]

- IDXVOLLTEXT mit Titel- und Schlagwörtern sowie Inhaltsverzeichnissen, die alle in IDX automatisch indexiert wurden. Es wurden 1358 Titel gefunden. Von diesen waren 824 relevant. Dies entspricht einem Recall von 0,92 und einer Precision von 0,69. Dieses Ergebnis bestätigt die Erkenntnisse, die aus MILOS II gewonnen wurden. Verglichen mit dem FREITEXT-Index wurde hier ein deutlich höherer Recall erreicht, während die Precision nur leicht niedriger ausfiel. Mit der automatischen Indexierung von IDX lässt sich die Retrievalleistung qualitativ steigern, ohne zu große Ballastmengen zu erzeugen.[143]

#### 3.3.3.2 Erprobung eines Cut-Off-Verfahrens

Zusätzlich wurden drei Indices namens SELIX30, SELIX100 und SELIX200 erstellt, die mit SELIX‘ automatischer Indexierung gewonnene Deskriptoren aus Titeln, Schlagwörtern und Inhaltsverzeichnissen enthielten[144]. Die Zahlen in den Indextiteln kennzeichnen den jeweiligen Cut-Off-Wert, der die Zahl der verwendeten Deskriptoren bei jeder angereicherten Titelaufnahme entsprechend auf die 30, 100 oder 200 höchstgewichteten Terme begrenzte.[145] Dieses Vorgehen entstand aus der Annahme, dass die zahlreichen, aus verschiedenen Verfahren gewonnenen Begriffe nicht alle gleich relevant sein können. Zudem wurde angenommen, dass durch eine zu große Zahl von Deskriptoren die Precision absinkt, da zu viele nicht relevante Dokumente in die Ergebnismengen geraten.[146]

Die Ergebnisse von Suchdurchläufen mit diesen Cut-Off-Werten gestalteten sich wie folgt:

- SELIX30: Recall 0,25, Precision 0,75. Gefunden wurden 333 Treffer, 240 davon relevant. Außerdem wurden hiermit 26 Nulltreffergebnisse erzeugt.
- SELIX100: Recall 0,58, Precision 0,73. Gefunden wurden 848 Treffer, 556 davon relevant.

[142] Lohmann, S.75-76.
[143] Ebd., S. 76.
[144] Ebd., S.64
[145] Ebd., S.60
[146] Ebd., S.17.

- Selix200: Recall 0,77, Precision 0,70. Gefunden wurden 1148 Treffer, 714 davon relevant.[147]

Dieses Ergebnis überraschte die Projektgruppe. Speziell der niedrige Recall von SELIX30 entsprach nicht den Erwartungen. Dieser lag deutlich unter den Ergebnissen des Index IDXVOLLTEXT und war als Retrievalergebnis nicht akzeptabel. Außerdem wurde der Precision-Wert aufgrund von zahlreichen Nulltrefferergebnissen aus nur 34 Suchergebnissen errechnet. SELIX 30 basierte auf der Kombination von Titelanreicherung wie oben angegeben, automatischer Indexierung nach IDX und dem Einsatz von SELIX. Das Problem bei SELIX30 wurde beim Cut-Off-Wert gesehen, der offenbar zu niedrig angesetzt war. Entgegen den vorherigen Annahmen verbesserte sich der Recall mit höher angesetzten Cut-Off-Werten und wäre ohne Begrenzung wahrscheinlich noch besser ausgefallen. Dafür sank auf diese Weise die Precision ab und war im Vergleich zu FREITEXT und IDXVOLLTEXT zu niedrig.[148]

#### 3.3.3.3 Schlussfolgerungen

Die Erwartungen an das Projekt wurden größtenteils, aber nicht ganz erfüllt. Die aus MILOS II übernommen modularen Indexierungsverfahren wurden erfolgreich verbessert und erzielten überlegene Retrievalergebnisse. Das SELIX-Verfahren arbeitete stabil und lieferte zuverlässige Ergebnisse. Aber: entgegen der Annahme, dass ein Cut-Off-Verfahren die Retrievalleistung verbessern könnte, erzielte ein reines MILOS-Verfahren zur automatischen Indexierung der erweiterten Titeldaten die im Schnitt besten Ergebnisse, da durch den Cut-Off-Wert relevante Deskriptoren ausgeschlossen wurden. Eine deutliche Verbesserung der Precision-Werte wog die zu niedrigen Recall-Werte nicht auf. Die KASCADE-Erschließung nach SELIX erwies sich dann als überlegen, wenn besonderes komplexe Suchfragen bearbeitet wurden. Die Ergebnisse erlaubten einerseits Rückschlüsse für den Verlauf verschiedener Indexierungsmethoden. Andererseits führten die Erfahrungen aus dem Test zu weiteren Verbesserungen an der Gewichtungskomponente in der Software.[149]

[147] Lohmann, S.76.

[148] Ebd., S.77.

[149] Lepsky u. Zimmermann, S.7-8 der Online-Ausgabe ohne Seitenzahlen.

## 3.4 Weitere bibliothekarische Retrievalexperimente im deutschsprachigen Raum

Die MILOS- und KASCADE-Experimente zeigten nicht nur die Nützlichkeit automatischer Indexierungsmethoden zur Verbesserung der Retrievalqualität in Bibliothekssystemen. Sie waren auch Inspiration für eine Anzahl weiterer Retrievalprojekte in oder in Zusammenarbeit mit Bibliotheken im deutschsprachigen Raum in den späten 1990ern und frühen 2000er Jahren. Zwei von diesen und andere zeitgenössische Retrievalprojekte werden auf den folgenden Seiten vorgestellt.

### 3.4.1 Projekt OSIRIS (Osnabrück Intelligent Research Information System)

Hinter dem Namen OSIRIS stand ein Projekt zur Weiterentwicklung des OPACs der Universität Osnabrück. Dieser funktionierte ursprünglich über ein Textparser-Interface und war nicht so leicht zu benutzen, wie es für einen Publikumskatalog wünschenswert gewesen wäre. Zudem war der vorhandene Bestand nicht vollständig sachlich erschlossen, und es war nicht abzusehen, dass sich dies je ändern würde. Für sachliche Recherchen stand nur die Titelstichwort-Suche zur Verfügung, und das vorhandene Klassifikationssystem nach der GHBS-Systematik bleib weitgehend ungenutzt. OSIRIS sollte diese Missstände ausräumen[150] und darüber hinaus ein in anderen Bibliotheken nutzbares System darstellen.[151]

Das Projektziel sah vor, mit Hilfe eines sogenannten „intelligenten User-Interfaces“ [152] die Formal- und Sachrecherche im vorhandenen Daten- und Informationsbestand nutzerfreundlicher zu gestalten. Die Sachrecherche sollte auf natürlich-sprachlicher Basis erfolgen, in der Form des zu vervollständigenden Satzes „Ich suche Literatur zum Thema…“.[153] Als Grundlage dafür wurde ein neues OSIRIS-Vokabular aus den Inhalten der OPAC-Datenbanken erstellt. Für den Bereich der Formalerschließung lagen ausreichend Daten vor. Da Sacherschließungsmerkmale nicht ausreichend vorkamen, waren weitere Maßnahmen wie die Ableitung RSWK-konformer Suchbegriffe aus der GHBS-Systematik nötig.[154]

[150] Recker, Ingrid; Ronthaler, Marc; Zillmann, Hartmut: OSIRIS. Osnabrück Intelligent Research Information System – ein Hyper-Base Front End System für OPACs. In: Bibliotheksdienst 30 (5), 1996. S.833-34.

[151] Ebd., S.848.

[152] Ebd., S.834.

[153] Ebd., S.836-837.

[154] Ebd., S.838-839.

Ein primäres Ziel von OSIRIS war die Vermeidung von sowohl Nulltreffer- als auch zu großen Ergebnismengen. Dies sollte durch die Option erreicht werden, die Klassenbezeichnungen selbst durchsuchen und die passenden auswählen zu können. OSIRIS‘ besondere Funktion sollte darin liegen, von NutzerInnen verwendete Suchbegriffe zu erkennen und für zukünftige Wiederverwendung in den Wortindex aufzunehmen. Dies wurde als „selbstlernende Sachrecherche“ bezeichnet.[155]

Die Suche über den Textparser wurde fehlertolerant gestaltet. Bei eventuellen Rechtschreibfehlern kam eine Rückfragefunktion zum Einsatz. Mit Hilfe einer Morphologiekomponente wurden flektierte Formen von Suchbegriffen erkannt und korrekt zugeordnet.[156]

Die technische Realisierung sollte Internet-konform und kompatibel mit anderen Systematiken und Bibliothekssystemen sein, die ähnliche Probleme wie der OPAC der Universität Osnabrück haben. [157]

Mit Hilfe eines Retrievaltests sollte OSIRIS‘ Leistungsfähigkeit dargestellt werden. Die Untersuchung fand auf der Grundlage von 15 Suchanfragen statt. Einige davon wurden aus den MILOS-Tests übernommen, andere wurden selbst formuliert.[158] Die Grundlage für den Test bildete der gesamte maschinenlesbare katalogisierte Bestand der USB Osnabrück mit ca. 600.000 Datensätzen. Als Relevanzkriterium galt die Zugehörigkeit gefundener Titel zu einer relevanten Klasse der OSIRIS-Systematik. Die Relevanzurteile wurden von Hartmut Zillmann erstellt. Diese stützten sich auf die Urteile von Fachreferenten, die die Zugehörigkeit eines Titels zu einer Klasse während des Katalogisierungsvorgangs bestimmten.[159]

Bei der Präsentation der Ergebnisse sticht die Aussage hervor, dass OSIRIS im Vergleich zum vorhandenen OPAC bei sachlichen Recherchen im selben Bestand ca. 11mal so viele Treffer liefere.[160] Diese Ergebnisse wurden leider nicht detailliert präsentiert und lassen sich daher nicht wissenschaftlich nachvollziehen.

[155] Recker, Ronthaler, Zillmann, S.842.

[156] Ebd., S.844-845.

[157] Ebd., S.848.

[158] Ronthaler, Marc: Dialog-Schnittstellen in Online-Informationssystemen: Notwendigkeit, Leistungsfähigkeit und Entwicklungsmöglichkeiten am Beispiel des OSIRIS-Systems. Dissertation. Osnabrück, Oktober 2000. S.75.

[159] Ebd., S.76.

[160] Ebd., S.78.

Ronthaler stellte einen Vergleich zwischen OSIRIS und anderen rezenten oder zugleich stattfindenden Retrievalvorhaben an und verglich die Errungenschaften des Systems, an dem er mitgearbeitet hatte, mit den Ergebnissen dieser Projekte. Darunter waren MILOS I und II[161] und KASCADE[162] zu finden. Beiden Projekten wurde zwar eine ähnliche Ausrichtung zugesagt (die Verbesserung der Retrievalleistung in Bibliotheks-OPACs), aber es wurde auf große Unterschiede in der Methodik hingewiesen. Ronthaler schätzte OSIRIS u.a. aufgrund der Arbeiten an der Dialogschnittstelle mit den erweiterten natürlichsprachlichen Eingabemöglichkeiten im Ansatz als "breiter" ein als MILOS.[163]

Mittelbach und Probst kamen in einer unabhängigen Untersuchung des Projekts zu dem Schluss, dass die Ansätze, die OSIRIS beim Nutzerinterface und der lernfähigen Datenbank gezeigt hatte, einen guten Weg aufzeigten und bedauerten, dass das System kaum Nachnutzer fand und daher nicht mehr weiterentwickelt wurde.[164]

### 3.4.2 Retrievalexperiment mit EKZ-Daten: Eignung des MILOS-Ansatzes bei Sachliteratur in öffentlichen Bibliotheken (2000)

Dieses Projekt von Martin Grummann war ausdrücklich von MILOS inspiriert und nahm als Anlass dieselbe Kritik an der sachlichen Erschließungspraxis in Bibliotheks-OPACs. Im Mittelpunkt stand die Frage, ob eine automatische Indexierung nach dem MILOS-Prinzip eine sinnvolle Maßnahme darstellt, um die sachliche Recherche in den OPACs von Öffentlichen Bibliotheken zu verbessern. Ein Retrievaltest sollte diese Frage klären[165]. Grummann konzentrierte sich auf Sachtitel in den Katalogen öffentlicher Bibliotheken, die uneindeutig formulierte Titel enthalten können, sodass Suchen mit Titelstichwörtern weniger erfolgsversprechend sind.[166]

Normalerweise erfolgt die intellektuelle Erschließung von Sachtiteln in ÖBs anhand der RSWK-Regeln. Diese sah Grummann jedoch als ungeeignetes Mittel an, da sie zu wenig Sucheinstiege für die Recherche böten und Nutzer vorher wissen müssten, welche Schlagwörter zu ihrem Informationsbedürfnis passen. In vielen Fällen übernehmen ÖBs

[161] Ronthaler, S.79-82.

[162] Ebd., S.84-87.

[163] Ebd., S.82.

[164] Mittelbach, Jens; Probst, Michaela: Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung. Strategien für das Bibliothekssystem der Freien Universität Berlin. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft. Heft 183. Berlin: Insitut für Bibliotheks- und Informationswissenschaft der Freien Universität Berlin, 2006. S. 43-44.

[165] Grummann, S.297-298.

[166] Ebd., S.303.

Fremddaten für die Erschließung, z.B. von der EKZ. Doch diese Daten haben Lücken bei der Erschließung, sodass Neuzugänge unvollständig erschlossen sein können.[167]

In Vorbereitung des Retrievaltests wurde eine repräsentative Testdatenbank aufgebaut. Die Grundlage bildeten die Titeldaten über Sachliteratur für Erwachsene auf der CD-ROM ekz.aktuell vom Oktober 1998. Insgesamt enthielt die Testdatenbank 46.769 Datensätze, von denen 73% sachlich per RSWK und formal per RAK erschlossen und 38% mit Annotationen ausgestattet waren, die zusätzliches Wortmaterial liefern konnten.[168]

30 Suchanfragen und insgesamt 64 Suchformulierungen wurden aus einer Sammlung von Informationsanfragen entwickelt, die zwischen dem 15. und 19. Februar 1999 an der Infotheke der Stadtbibliothek Oldenburg gestellt worden waren. Diese sollten typische Suchverhaltensweisen von Bibliotheksnutzern nachbilden und positive und negative Auswirkungen der überprüften Erschließungsmethoden aufzeigen.[169]

IDX/MILOS wurde für diesen Test verwendet, da seine Verwendung gut dokumentiert war und es keine schwierige Aufgabe darstellte, den maschinenlesbaren Bestand einer ÖB in dieser Software automatisch zu indexieren.[170] Da IDX über keine Retrievalfunktion verfügt, wurden wie bei MILOS II in Allegro sechs Register angelegt:

1. Unbehandelte Titelstichwörter

2. Indexierte Titelstichwörter

3. RSWK-Schlagwörter

4. Titelstichwörter und Schlagwörter

5.Titelstichwörter und Schlagwörter in unbehandelter und indexierter Form (Basic Index)

6. Indexierte Titelstichwörter und Annotationstexte[171]

Als Kennzahlen wurden Recall und Precision verwendet, sowie das Einheitsmaß nach van Rijsbergen mit dem Gewichtungsfaktor 1, wodurch beide Werte gleich gewichtet werden. Grummann stellte hierbei den Anspruch, dass nur bei ausreichender Höhe beider Werte ein Ergebnis als gut zu bezeichnen sei.[172]

[167] Grummann, S.299.
[168] Ebd., S.303.
[169] Ebd., S.304.
[170] Ebd., S.300.
[171] Ebd., S.305.
[172] Ebd., S.305.

Zur Relevanzbewertung sah Grummann die Kriterien aus MILOS II als nicht ausreichend an, da mit uneindeutigen Titelformulierungen zu rechnen war, die Missverständnisse verursachen hätten können. Ein Titel galt genau dann als relevant, wenn eine thematische Relevanz im Kontext der Suchanfrage aus dem Titel, den verwendeten Schlagwörtern und der Rezension auf der ekz-CD sichtbar war. In einem Testlauf wurden alle Suchfragen mit den CD-Daten getestet und die Titel in den Ergebnislisten auf ihre Relevanz hin überprüft. Dadurch entstand eine Liste relevanter Titel vor dem Beginn des Retrievaltests. Es wurde binär zwischen relevant und nicht relevant unterschieden.[173]

Die Ergebnisse des Tests stützten die Schlussfolgerungen, die aus MILOS I und II gezogen worden waren. Von den 6 Registern erzielte die Suche im Basic Index mit sowohl unbehandelten als auch automatisch indexierten Stich- und Schlagwörtern die besten Resultate und erfüllte Grummanns oben genannte Kriterien mit einem Recall von 74% und einer Precision von 78%. Zudem wurde hier nur ein einziges Nulltreffergebnis festgestellt. Die Indexierung von Schlagwörtern kann also auch mit dem Titelmaterial von populär-wissenschaftlicher Sachliteratur für bessere Retrievalergebnisse sorgen.[174]

Grummann zieht ein ähnliches Schlussfazit wie Lepsky bei den MILOS-Experimenten. Durch automatische Indexierung und den Einsatz von zusätzlichem Begriffsmaterial können Retrievalergebnisse deutlich verbessert werden, aber ihr bester Einsatz findet im Tandem mit der intellektuellen Erschließung statt. Außerdem zeigt Grummanns Projekt, dass Suchergebnisse mit frei vergebenen Deskriptoren, die nicht einem etablierten Regelwerk entstammen, ebenfalls verbessert werden können.[175]

### 3.4.3 Retrievalexperiment nach MILOS-Vorbild im ÖBV (2003)

Dieses Retrievalexperiment wurde mit Daten des Verbundkatalogs des Österreichischen Bibliotheksverbundes (ÖBV) 2003 in Kooperation mit der FH Köln durchgeführt. Das Ziel bestand darin, die Auswirkungen einer gründlicheren Erschließung der vorhandenen Datensätze per automatischer Indexierung zu untersuchen. Der Verbundkatalog des ÖBV verfügte zum Zeitpunkt des Projekts über ca. 3,76 Mio. Titel- und ca. 6,78 Mio. Exemplardatensätze, von denen nur ca. 42,5 % nach RSWK und SWD sachlich

[173] Grummann, S.305-306.
[174] Ebd., S.307-308.
[175] Ebd., S.316.

erschlossen waren. Es bestand die Hoffnung, dass sich die Erschließungsquote und die Retrievalquote im Verbundkatalog durch automatische Indexierung deutlich erhöhen ließen.[176]

Dieses Experiment nahm sowohl direkten Bezug auf die MILOS-Experimente wie auch auf Martin Grummanns Projekt aus dem vorherigen Kapitel. Die übereinstimmenden Schlussfolgerungen zu dem Beitrag, den automatische Indexierungsmethoden für die Retrievalqualität in einem OPAC leisten, wurden ebenfalls erwähnt.[177] Auf der Grundlage dieser positiven Ergebnisse hatte die Zentrale des ÖBV ein Interesse daran, das MILOS-Verfahren anhand eines OPACs zu erproben und die Retrievalergebnisse vor und nach der Anreicherung des Basic Index (BI) mit automatisch erzeugten Schlagwörtern in einer „Alle-Felder-Suche“ zu vergleichen. Die Aussage aus den MILOS-Experimenten, dass die automatische Indexierung für einen höheren Recall ohne signifikante Einbußen bei der Precision sorgen könnte, sollte ebenfalls überprüft werden.[178]

Der Retrievaltest wurde mit Hilfe von IDX/MILOS anhand einer für den ÖBV-OPAC repräsentativen Stichprobe von ca. 100.000 Datensätzen durchgeführt. Nach der Löschung von Dubletten und anderen Bereinigungsmaßnahmen blieb ein Sample von 106.904 Aufnahmen. Die meisten von diesen verfügten über einen Hauptsachtitel, Zusatz zum HST und RSWK-Schlagwortketten, die das nötige Begriffsmaterial für den Indexierungsvorgang liefern würden.[179]

Das Sample wurde aus dem Verbundkatalog extrahiert und in einen Aleph-OPAC übertragen. Die Datensätze wurden auf die relevanten MAB-Kategorien reduziert. Dadurch verkleinerte sich die Testkollektion um knapp 10.000 Datensätze. Dieses neue Sample wurde an die FH Köln übermittelt und in einer Allegro-Umgebung automatisch per IDX/MILOS indexiert. Die Größe der Kollektion verringerte sich auf 72.006 Aufnahmen. Die Datensätze waren um zwei Kategorien erweitert worden, die aus der automatischen Indexierung hervorgegangen waren: Indexate auf der Basis von Titel/Untertitel (HST und Zusatz) und Indexate auf Basis der Schlagwörter. Diese in IDX/MILOS indexierte Kollektion wurde an den ÖBV übermittelt und wieder in die Aleph-500-Umgebung

[176] Oberhauser, Otto; Labner, Josef: OPAC-Erweiterung durch automatische Indexierung. Empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. In: ABI-Technik 23 (4), 2003. S.306-307.

[177] Ebd., S.306.

[178] Ebd., S.307.

[179] Ebd.

importiert. Es wurde ein zweiter BI (Alle Felder inkl. Neue Indexate) aufgebaut, mit allen Mehrwortgruppen in Form von Einzelstichwörtern.[180]

Die 100 Suchanfragen aus MILOS II wurden für dieses Experiment verwendet. Auch bei der Messung der Retrievalergebnisse war MILOS das Vorbild. Wie dort wurde aufgrund des Umfangs der Testkollektion auf die Überprüfung der Recall-Werte verzichtet. Die Relevanzurteile waren binär (relevant oder nicht relevant) und wie bei MILOS II wurde jeder für eine Suchanfrage annähernd relevante Titel als relevant bezeichnet.[181]

Insgesamt wurden drei Retrievaltests durchgeführt. 1. Ein Vergleich zwischen den Basic Indices mit und ohne Indexate. Der BI mit Indexaten fand mehr relevante Titel pro Frage bei nur einer leicht gesunkenen Precision mit weniger Nulltrefferergebnissen.[182] 2. Eine Suche im Korpus um zwischen Datensätzen mit und ohne Schlagwörtern zu differenzieren. Nach der automatischen Indexierung wurden mehr Titel mit und ohne Schlagwörter gefunden, inklusive unbeschlagwortete Titel von hoher Relevanz, doch die Precision litt darunter (56,7%). [183] 3. Eine Überprüfung, ob die verwendeten Suchformulierungen, u.a. mit dem booleschen Operator ODER, die Effekte der automatischen Indexierung verringerten. Anpassungen in der Formulierung der Suchanfragen erbrachten jedoch keine signifikanten Änderungen.[184]

Die Autoren waren mit den Ergebnissen zufrieden und sahen das Potential für zukünftige Verbesserungen im ÖBV-Verbundkatalog. Über alle Suchvorgänge hinweg war im Durchschnitt ein Zuwachs von etwa einem Drittel mehr an gefundenen relevanten Titeln sichtbar. Von durchschnittlich 3,45 gefundenen Titeln waren fast 2 relevant. Zugleich sank die Zahl von Nulltreffergebnissen um ein Drittel. Nach der automatischen Indexierung wurden mehr nicht per RSWK-Kette verschlagwortete Titel gefunden, die ebenso relevant für die Suchanfragen waren wie Titel mit Schlagwörtern. Die Autoren zogen das Fazit, dass die Studie im ÖBV-OPAC die Ergebnisse der MILOS-Tests bestätigte. Oberhauser und Labner befürworteten daher die Anwendung auf den gesamten Katalog des ÖBV und weitere Verfeinerungen und Erweiterungen des Ansatzes für die Zukunft.[185]

[180] Oberhauser u. Labner, S.307-308.
[181] Ebd., S.309.
[182] Ebd., S.310.
[183] Ebd., S.310-311.
[184] Ebd., S.311-312.
[185] Ebd., S.312-313.

### 3.4.4 Retrievalexperiment im MALIS-Studiengang der FH Köln: Google Scholar und Ebsco Discovery Service (2013)

Die Suchmaschine Google Scholar bietet kostenlosen Zugang zu wissenschaftlichen Artikeln aus zahlreichen Forschungsfeldern und ist für alle, die schon einmal Googles Suchmaschinentechnologie benutzt haben, unkompliziert zu nutzen. Diese Erwartungen werden oft auch an die Retrievalkomponenten von Bibliothekssystemen gestellt.[186]

In Bibliotheken erhalten NutzerInnen Zugang zu sehr heterogenen Informationsquellen, die eigene Retrievalstrategien und Einarbeitungszeit erfordern. Sie bevorzugen im Allgemeinen die einfache Bedienbarkeit von Websuchmaschinen. Ab ca. 2009 entstanden die ersten kommerziellen Discovery Services, die alle Informationsressourcen einer Bibliothek hinter einem Interface vereinen und diese NutzerInnen per Suchmaschinentechnologie verfügbar machen. [187] Die USB Köln nutzt ein USB-Webportal, das seit 2009 alle Informationsangebote der Bibliothek unter einem Front-End vereinigt. [188] Der EBSCO Discovery Service wurde dem im Juli 2011 als weitere Suchoption hinzugefügt.[189]

Harald Kaluza interessierte die Fragestellung, ob eine Bibliothek zwingend eine kostenpflichtige Lösung wählen muss, oder ob es alternativ ausreichen würde, einen Link zu GS im Webauftritt der Einrichtung zu platzieren, um so leichteren Zugang zu wissenschaftlichen Artikeln anzubieten. Die Antwort darauf sollte ein Retrievaltest liefern, mit dem anhand von 500 vorher ausgewählten interdisziplinären Zeitschriftenartikeln die freie Volltextverfügbarkeit in Google Scholar und der EBSCO-Disocvery-Service-Installation der USB Köln nachgewiesen werden sollte.[190]

Kaluza bezieht sich nicht explizit auf vorherige Retrievaltests. Seine Arbeitsweise ähnelt den Uniterm-Experimenten und Cranfield I mehr als späteren Retrievalprojekten, die sich der sachlichen Recherche widmeten. Alle 500 Artikel wurden in beiden Services manuell gesucht. Die Suche in GS fand außerhalb des IP-Bereiches wissenschaftlicher Institutionen statt. Als Erfolgskriterium galt primär, ob ein Artikel nachgewiesen wurde oder nicht. Dieser galt in GS als nicht gefunden, wenn das Suchergebnis nur eine Zitation angab.

[186] Kaluza, Harald: Google Scholar versus EBSCO Discovery Service: Ein vergleichender Retrieval-Test. In: b.i.t.online innovativ Bd. 44. Malis Praxisprojekte 2013. Projektberichte aus dem berufsbegleitenden Master-Studiengang Bibliotheks- und Informationswissenschaft der Fachhochschule Köln. S.60.

[187] Ebd., S.60-61.

[188] Kostädt, Peter: Einführung eines Discovery Service in der in der Universitäts- und Stadtbibliothek Köln. In: Ein Bibliothekar mit Informationskompetenz. Festschrift zum 60. Geburtstag von Dr. Rolf Thiele. Köln: Universitäts- und Stadtbibliothek Köln, 2012. S.78.

[189] Ebd., S.85.

[190] Kaluza, S.63.

Unterschieden wurde außerdem zwischen Volltext-Zugriffen und lediglich Nachweisen bzw. Links zu kostenpflichtigen externen Angeboten.[191]

In EBSCO wurde ebenfalls außerhalb des IP-Bereiches von Institutionen gesucht, aber mit Hilfe einer Bibliotheksnutzerkennung der ULB Köln. Dadurch gab es keinen Zugriff auf kostenpflichtige Inhalte. Hinweise auf Volltextzugriffe waren jedoch in den Einzeltrefferanzeigen ggfs zu sehen. Hier wurden die Ergebnisse in drei Kategorien unterteilt: Nachweise darüber, ob der Artikel gefunden wurde oder nicht, Volltextzugriff per EBSCO oder über eine andere Quelle.[192]

EBSCO wies 91,8% der gesuchten Artikel nach. Dies lag deutlich über den 70,8% von Google Scholar. 96% der in GS nachgewiesenen Artikel wurden auch in EDS gefunden. 115 der 500 Artikel (23%) wurden nur in EDS nachgewiesen, während 11 Artikel (2,2%) nur in GS gefunden worden.[193]

Insgesamt gesehen erreichte EDS eine wesentlich breitere Abdeckung. Dafür aber ist eine Authentifizierung seitens des Herstellers sowie die Berechtigung als Bibliotheksnutzer erforderlich. Der Anteil von Volltexten ist in EDS gegenüber GS deutlich höher. Dies gilt auch für den Anteil von exklusiv in EDS verfügbaren Volltexten (41,2% gegenüber 4,4% in GS).[194]

Aufgrund der Ergebnisse kam Kaluza zu dem Schluss, dass Google Scholar den Zugriff auf Fachdatenbanken und Discovery Services mit mehreren spezialisierten Informationsquellen nicht ersetzen, aber ergänzen kann. Die Abdeckung von Inhalten lag in EDS deutlich höher, obwohl sich dieser Vorteil nur mit entsprechender Zugangsberechtigung darstellen ließ. Kaluza kam daher zu dem Schluss, dass sich die Anschaffung eines Discovery Service für eine wissenschaftliche Bibliothek lohnt.[195]

# 4 Projekt GELIC – Retrievalexperiment der TH Köln

Im Folgenden wird das Retrievalexperiment GELIC (German Library Indexing Collection) vorgestellt, das ab 2017 unter der Leitung von Prof. Dr. Klaus Lepsky und Prof. Dr. Philipp Schaer an der TH Köln durchgeführt wird. Die Vorstellung von GELIC wird in

[191] Kaluza, S.70.
[192] Ebd., S.71.
[193] Ebd., S.72-74.
[194] Ebd., S.74-75.
[195] Ebd. S.76-77.

einem größeren Detailgrad als bei den vorherigen Projekten erfolgen und den zeitlichen Ablauf des Projekts darstellen. GELIC ist nicht nur der Projektname, sondern bezeichnet die für Retrievaltests verwendbare Testkollektion, die durch die Projektarbeit etabliert werden soll. Sie soll den Prinzipien ähneln, die in der TREC-Community gelten.[196]

Studierende des ehemaligen Studiengangs Bibliothekswissenschaft aus mehreren Jahrgängen unterstützten die Durchführung des Projekts. Der Autor dieser Arbeit nahm an der Projektarbeit ab dem Sommersemester 2018 bis zum Anfang des Jahres 2020 teil. Lepsky und Schaer veröffentlichten in Zusammenarbeit mit der ehemaligen studentischen Projektteilnehmerin Johanna Munkelt den Artikel „Towards an IR Test Collection for the German National Library“ als Vorstellung des Projekts in der internationalen Fachwelt. Im Dezember 2019 wurde GELIC im Rahmen des Studierenden-Workshop für informationswissenschaftliche Forschung (SWIF) 2019 in Potsdam einem Publikum aus Wissenschaftlern und Studierenden aus den Bibliotheks- und Informationswissenschaften vorgestellt.

## 4.1 Projektanlass

Im Jahr 2010 suchte die Deutsche Nationalbibliothek (DNB) nach Möglichkeiten, die sachliche Erschließungsarbeit mittels automatischer Arbeitsabläufe zu erleichtern. Der erste Schritt der DNB bestand darin, die Reihe O (Online-Publikationen) aufgrund der immer schneller anwachsenden Zahl von zu erschließenden Publikationen nur noch automatisch zu indexieren.[197] Im September 2017 begann die DNB damit, die Reihen B (Monografien und Periodika außerhalb des Verlagsbuchhandels) und H (Hochschulschriften) ausschließlich automatisch zu erschließen. Die DNB argumentierte für den Wechsel zu maschinellen Methoden mit der fortschreitenden Modernisierung im Bibliothekssektor und mit der zunehmenden Heterogenität der sachlichen Erschließung für verschiedene Arten von Dokumenten. Durch die automatische Indexierung, so die DNB, könne der Zugang zu diesen Medien erleichtert und beschleunigt werden. Zudem könne die Erschließung vieler verschiedener Medientypen aneinander angeglichen werden.[198]

[196] Munkelt, Johanna; Schaer, Philipp; Lepsky, Klaus: Towards an IR Test Collection for the German National Library. Proceedings of LWA 2018, 2018. S.276

[197] Gömpel, Renate; Junger, Ulrike; Niggemman, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken 20 (1), 2010. S.20-21.

[198] Deutsche Nationalbibliothek: Grundzüge und erste Schritte der künftigen inhaltlichen Erschließung von Publikationen in der Deutschen Nationalbibliothek. Onlinedokument. https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/konzeptWeiterentwicklungInhaltserschliessung.pdf?__blob=publicationFile&v=4 (Letzter Zugriff: 11.07.2022).

Dieser Schritt der DNB führte zu einer kontroversen Diskussion, die sogar im Feuilleton der Frankfurter Allgemeinen Zeitung nachzulesen war. Klaus Ceynowa, seit 2015 Generaldirektor der Bayerischen Staatsbibliothek, äußerte sich in einem Gastartikel kritisch zur Entscheidung der DNB, in Zukunft verstärkt automatisch zu indexieren. Die Qualität automatischer Indexate sei der intellektuellen Erschließung unterlegen und führe unweigerlich zu Katalogisaten minderer Qualität.[199]

Weitere Stimmen aus der dt. Bibliotheksszene brachten sich in die Diskussion ein, die sowohl für als auch gegen die Pläne der DNB argumentierten.[200]

## 4.2 Erkenntnisinteresse seitens der TH Köln

Aus der oben erwähnten Kontroverse ergibt sich die Frage, ob eine automatisch ablaufende Inhaltserschließung die Qualitätsansprüche von professionellen BibliothekarInnen erfüllen kann. Eine mögliche Antwort darauf liegt in Retrievalexperimenten, die die Qualität der Erschließungsarbeit der DNB anhand ihrer Retrievalleistung untersuchen. Eine geeignete Testkollektion zur Untersuchung dieses Sachverhalts gab es jedoch noch nicht, als die Diskussion um die Qualität automatischer Indexierung begann. Das Projekt GELIC wurde etabliert, um solch eine Testkollektion aufzubauen und damit aussagekräftige Retrievaltests durchzuführen, in denen die intellektuellen und automatischen Sacherschließungs-Methoden der DNB miteinander verglichen werden.[201]

## 4.3 Experiment-Setup

Wie in den meisten Retrievalexperimenten, die in den vorherigen Kapiteln vorgestellt wurden, wurde das Experiment um die Komponenten des Cranfield-Paradigmas aufgebaut, wie sie auch in der TREC-Konferenz Anwendung finden:

- eine Dokumentenkollektion, die eine ausreichende und von Ballast bereinigte Datenmenge aus dem DNB-Katalog enthalten soll. Die so repräsentierten Dokumente sollen ohne Schwerpunkt thematisch so breit gefächert sein möglich.

[199] Ceynowa, Klaus: In Frankfurt lesen jetzt zuerst Maschinen. Frankfurter Allg. Zeitung. 31.07.2017. Online unter: https://www.faz.net/aktuell/feuilleton/buecher/maschinen-lesen-buecher-deutsche-nationalbibliothek-setzt-auf-technik-15128954.html (Registrierung erforderlich; letzter Zugriff: 11.07.2022).

[200] Wiesenmüller, Heidrun: Das neue Sacherschließungskonzept der DNB in der FAZ. Basiswissen RDA. August 2017. Das neue Sacherschließungskonzept der DNB in der FAZ - Basiswissen RDA (basiswissen-rda.de) (Letzter Zugriff: 11.07.2022).

[201] Munkelt, Schaer, Lepsky, S.277.

- eine Anzahl von realen Informationsbedürfnissen (Topics) mit ausformulierten Erläuterungen (description und narrative), die die thematische Breite der Dokumente im Korpus wiedergeben soll.
- eine Anzahl von Relevanzurteilen, die für die Ergebnismengen in der Kollektion für die jeweiligen Topics gefällt wurden. Hier musste eine Methode zur Fällung der Urteile gefunden werden, die für eine kleine Projektgruppe in akzeptablem Tempo durchgeführt werden konnte.[202]

## 4.4 Der Dokumentenkorpus

Der Dokumentenkorpus wurde in Form einer kostenlosen Datenselektion dem DNB-Gesamtkatalog entnommen.[203] Die DNB stellte für GELIC eine Sammlung von 200.000 Datensätzen zur Verfügung. Diese umfasst ausschließlich Sachliteratur und verzichtet auf Titel aus dem Bereich Belletristik oder auf Kalender, da diese sich für Retrievalexperimente zur sachlichen Recherche nicht eignen. Über 100 Sachgruppen sind darin vertreten.[204] In der ersten Version der Kollektion waren 131.538 Datensätze intellektuell und 68.462 Datensätze (ca. 35%) automatisch erschlossen.[205] Die Datensätze enthielten mehrere Felder mit Schlagwort-Daten. Unterschieden wurde zwischen intellektuell vergebenen GND-Schlagwörtern im Feld 510, automatisch vergebenen Schlagwörtern im Feld 530 und Schlagwörtern aus dem Verlagsbuchhandel im Feld 520.[206]

Der Korpus lag zunächst im Austauschformat Pica+ vor. Dieses war mit den während des Experiments verwendeten Tools nicht kompatibel und musste in ein Format konvertiert werden, mit dem die während des Projekts verwendete Suchmaschinen-Software arbeiten konnte. Nicht benötigte Informationen wurden herausgefiltert. Nur inhaltstragende Felder blieben übrig.

## 4.5 Verwendete Software

Für die Durchführung des Projekts wurden verschiedene Tools verwendet. Besonders wichtig war eine Suchmaschine, in der Suchanfragen an den DNB-Korpus gestellt werden konnten. Die Projektleitung entschied sich für die Suchmaschine Solr/Lucene.

[202] Munkelt, Schaer, Lepsky, S.276.

[203] Böckmann, Ina. Aufbau einer Pipeline zur Transformation von Titeldaten der DNB. Bachelorarbeit, TH Köln 2020. S.35.

[204] Munkelt, S.25-26.

[205] Munkelt, Schaer, Lepsky: S.277.

[206] Munkelt, S.31-32.

Solr:

- ist Open-Source und kann daher den Zwecken des Projekts angepasst werden
- ist eine Suchmaschine und keine Datenbank, eignet sich jedoch für die Verarbeitung unstrukturierter Daten wie in der Testkollektion
- ist dazu fähig, Relevanzrankings per Messung der Termhäufigkeit anhand gesuchter Keywords und Phrasen durchzuführen
- ist dazu fähig, Anfragen mit booleschen Operatoren (AND, OR, NOT) zu verarbeiten

Für die Indexierung in Solr ist eine managed-schema-Datei nötig, die festlegt, wie mit den eingegeben Daten umzugehen ist. Diese sorgt für die Vorverarbeitung der Daten nach den Anforderungen des Projekts, identifiziert Terme und Tokens (einzelne Text-Fragmente, die zur Verarbeitung getrennt werden) und verarbeitet diese durch linguistische Module der Suchmaschine. Auf diese Weise kann Solr u.a. mitgeteilt werden, wie viele Instanzen eines inhaltstragenden Feldes innerhalb eines Datensatzes vorkommen dürfen. Demnach kann ein Datensatz beispielsweise die Felder title_txt_de und title_s (der Hauptsachtitel in Text- und in Stringform) jeweils nur einmal besitzen, während Schlagwortfelder multi-valued sind und mehrmals im selben Datensatz vorkommen können.

Für die Auswertung der einzelnen Evaluationsdurchläufe in Solr wurde das Tool Trec_eval ausgewählt. Dieses wird standardmäßig von der TREC-Konferenz für die Bewertung von Suchläufen in Retrievaltests verwendet. Es stellt die aus Solr exportierten Suchergebnisse in Form von Kennzahlen nachvollziehbar dar und erlaubt eine gemeinsame Bewertungsgrundlage. Zu den wichtigen Kennzahlen, die über Trec_eval ausgegeben werden, gehören num_ret (die Anzahl der gefundenen Dokumente in der Ergebnismenge) und num_rel (die Anzahl der relevanten Dokumente in der gesamten Dokumentenmenge). Als Kombination daraus ergibt sich der Wert num_rel_ret, der die Anzahl der gefundenen relevanten Dokumente abbildet. Aus diesen Werten werden Precision und Recall berechnet und angezeigt.

Außerdem kamen in den frühen Projektphasen diese Tools zum Einsatz:

- die Bibliothekssoftware Allegro
- Relevation, ein Tool zur Erstellung von Relevanzurteilen in Retrievalsystemen[207]

[207] Munkelt., S.39.

- die Datenbank- und Thesaurus-Software Midos 6 für die Konversion der DNB-Daten in ein Solr-kompatibles XML-Format
- der Texteditor Notepad++ für Anpassungen und Fehlerkorrekturen in den XML-Versionen der Testkollektion

Im weiteren Verlauf des Projekts wurden andere Software-Lösungen eingesetzt, sodass die oben erwähnten Tools nicht mehr benötigt wurden.

## 4.6 Entwicklung der Topics und Relevanzurteile

GELIC nutzt Topics, die von MILOS II übernommen wurden. Die erste Projektgruppe testete alle 100 MILOS II-Topics und filterte diejenigen heraus, die nur sehr wenige oder überhaupt keine Ergebnisse bei Suchen im Korpus erbrachten. Übrig blieben 50 Topics, die für die gesamte Projektdauer verwendet wurden. Jedes Topic besteht aus einer ID-Nummer, dem Titel, der description (einer kurzen Beschreibung des Suchgegenstands) und dem narrative, das definiert, welche Dokumentinhalte für das Informationsbedürfnis relevant, teilweise relevant oder nicht relevant sind.[208]

Wie bei jeder großen Dokumentenkollektion war es bei GELIC unmöglich, die Relevanz jedes Datensatzes für alle Topics festzustellen. Als Basis für die Relevanzurteile wurde daher für jedes Topic ein Poolingverfahren genutzt, wie es bei TREC angewendet wird. Vier TeilnehmerInnen aus der ersten Projektgruppe entwickelten Suchstrategien für jedes Topic, was insgesamt 200 Suchanfragen (4x50) entspricht. Die vier Ergebnismengen für jedes Topic wurden zusammengeführt und von Dubletten bereinigt.[209] Die TeilnehmerInnen nahmen in der Folge mit dem Tool Relevation Relevanzurteile für die übrig gebliebenen Titel vor. Ein dreistufiges Skalenniveau teilte die untersuchten Titeldaten in nicht relevante, teilweise relevante und relevante Titel. [210] Insgesamt wurden 6.984 Relevanzurteile vergeben. 909 Titel wurden als relevant, 1457 als teilweise relevant und 4616 als nicht relevant bewertet.[211]

Im weiteren Verlauf wurden auf der Basis aller Topics erste konkrete Suchanfragen formuliert, mit denen in Solr gesucht werden konnte. Diese Anfragen wurden festgehalten

[208] Munkelt, S.34-35.
[209] Ebd., S.38-39.
[210] Ebd., S.39.
[211] Munkelt, Schaer, Lepsky, S.279.

und über eine online zugängliche Excel-Tabelle allen ProjektteilnehmerInnen zugänglich gemacht.[212]

Spätere Untersuchungen zeigten jedoch, dass die Verteilung der Relevanz in den Topics unausgewogen war. Bei 8 der 50 Topics wurden weniger als 10 relevante Dokumente zugeordnet (z.B. Topic 45 Betriebspsychologie und Topic 19 Zimmerpflanzen). Ca. 72% der relevanten Dokumente waren maschinell erschlossen. Bei den intellektuell erschlossenen relevanten Dokumenten lag diese Zahl mit 56% deutlich niedriger. Zudem wurden bei 19 von 50 Topics weniger als 10 relevanten Dokumente intellektuell erschlossen.[213]

## 4.7 Projektverlauf und aufgetretene Probleme 2018-19

Die erste Gruppe von Studierenden hatte 2017 und 2018 an GELIC mitgearbeitet und mit den oben genannten Punkten die Grundlagen für das eigentliche Vorhaben des Projekts etabliert. Nun sollte mit den Retrievalexperimenten begonnen werden. Allerdings wechselte zum Sommersemester 2018 die Gruppe der studentischen MitarbeiterInnen. Die neuen TeilnehmerInnen konnten auf die Vorarbeit des vergangenen Jahres zurückgreifen, mussten allerdings die Projektgrundlagen und den Umgang mit den benötigten Tools von Grund auf neu erlernen. Bis alle Tools auf den Computern aller TeilnehmerInnen zuverlässig liefen und damit gearbeitet werden konnte, verstrich ein großer Teil der für das Projekt veranschlagten Zeit.

Der Ablauf und die angewandten Methoden während des Sommersemesters 2018 und der Fortführung des Projekts im folgenden Wintersemester wurden leider nicht vollständig dokumentiert. Die folgenden Schilderungen basieren auf persönlichen Notizen und Gesprächen unter den ProjektteilnehmerInnen und geben den Verlauf nach bestem Wissen wieder.

### 4.7.1 Überarbeiteter Dokumentenkorpus von der DNB

Während des Wintersemesters 2018/19 erhielt das GELIC-Team eine von der DNB überarbeitete Dokumentenkollektion, wiederum mit 200.000 Titeldaten. Diese enthielt eine weitaus größere Anzahl von automatisch erschlossenen Datensätzen (120.783) als die erste Version. Enthalten waren außerdem 132.777 intellektuell erschlossene Datensätze und 72.252, die doppelt erschlossen waren, also sowohl automatisch als auch

[212] Diese Tabelle ist online zu finden unter https://docs.google.com/spreadsheets/d/1940qrRalEE-DEZZmXWueUGM7_J9e8zEvBdD61_Qj7U9k/edit#gid=0 einsehbar. (Letzter Zugriff: 11.07.2022)

[213] Böckmann, S.3.

intellektuell. Die Projektgruppe erhoffte sich durch den höheren Anteil von automatisch und doppelt erschlossenen Datensätzen eine bessere Vergleichbarkeit der Retrievalergebnisse.[214] Ca. 10% der Datensätze sind allerdings weder intellektuell noch automatisch erschlossen und können daher nicht bei Schlagwortsuchen gefunden werden, was Retrievaltest-Ergebnisse beeinflussen kann. Der Anteil nicht erschlossener Dokumente wurde daher aus den Tests „herausgerechnet".[215]

### 4.7.2 Datenkonversion von Pica+ nach XML

Die nächste Aufgabe der Projektgruppe bestand darin, die Daten der Testkollektion vom Austauschformat Pica+ in ein Solr-kompatibles XML-Format zu konvertieren. Es wurden verschiedene Möglichkeiten erwogen, wie z.B. die Umwandlung in das Allegro-kompatible Format Pica3 oder in MAB-Diskette. Ein Durchbruch gelang erst, nachdem Frau Stephani Scholz vom Hochschulbibliothekszentrum Nordrhein-Westfalen die Testkollektion in das Allegro-kompatible Format MAB-Band umwandelte. Daraus ergaben sich allerdings Probleme bei der Nachvollziehbarkeit, da den ProjektteilnehmerInnen nun wichtige Information über den genauen Ablauf der Datenkonvertierung fehlten.

Die Projektleitung entschied sich dazu, auf dieser Grundlage weiterzuarbeiten. Die Testkollektion im Format MAB-Band wurde zunächst in Allegro importiert, wo sie für die weitere Verarbeitung mit dem Datenbank- und Thesaurusprogramm Midos 6 konvertiert wurde. In Midos wiederum fanden die Schritte für die finale Konvertierung in ein Solr-kompatibles XML-Format statt.

Der Indexierungsvorgang der Dokumentenkollektion in Solr lief nicht problemlos ab. Die Suchmaschine brach den Prozess wiederholt vorzeitig ab, sodass nur ein Teil der Datensätze in Solr übernommen wurde. Weitere Untersuchungen der XML-Datei der Kollektion per Notepad++ zeigten, dass dieses Problem durch ungewöhnlich formulierte Dokumententitel verursacht wurde, die Sonderzeichen wie Pipes im Hauptsachtitel enthielten. Ein Beispiel: Schimmel über Berlin: Literatur | Gespräche | Visual Art.

Beim Indexierungsvorgang interpretierte Solr die Pipes als Ende des Hauptsachtitels und versuchte, pro Pipe ein weiteres Feld für den Hauptsachtitel mit den so übrig gebliebenen Textinhalten anzulegen. Da aber in der managed-schema-Datei (s.o.) die Titelfelder (title_txt_de und title_s) nicht als multi-valued deklariert waren, also jeweils nur einmal

[214] Seegert, Christin: Evaluation von automatischer und intellektueller Verschlagwortung der Deutschen Nationalbibliothek anhand einer Retrieval-Testkollektion. Bachelorarbeit. TH Köln, 2019. S.7.

[215] Böckmann, S.2.

pro Datensatz vorkommen dürfen, erkannte Solr den Fehler und brach den Indexierungsvorgang an dieser Stelle ab. Die Projektgruppe ließ den Indexierungsvorgang immer wieder ablaufen, um Dokumente mit Pipes im HST zu identifizieren und die doppelten Titelfelder zu korrigieren, indem die Pipes durch Bindestriche ersetzt wurden.

Der gesamte Prozess, die DNB-Testkollektion in ein arbeitsfähiges Datenformat zu konvertieren, war sehr zeitaufwändig und erwies sich für die weitere Verwendung im Projekt als wenig geeignet[216]. Durch eine XSLT-Transformation konvertierte die Projektleitung die Daten in das Format, mit dem im folgenden Jahr weitergearbeitet wurde.[217]

## 4.8 Entwicklung von GELIC 2019/2020

Zum Sommersemester 2019 änderte sich erneut die Zusammensetzung der Projektgruppe. Um einer langen Einarbeitungszeit wie bei der vorherigen Gruppe vorzubeugen, wurden grundlegende Änderungen am Experimentsetup vorgenommen. Suchanfragen an Solr wurden ab sofort nicht mehr händisch und für jedes Topic einzeln eingegeben. Stattdessen wurden über ein WinPython-Skript sämtliche 50 Topics zugleich abgefragt, automatisch evaluiert und die Ergebnisse im Format von Trec_eval ausgegeben[218]. Dies geschah aus Gründen der Zeitersparnis, da die Etablierung einer funktionsfähigen Indexierung der Dokumentenkollektion auf den Computern aller TeilnehmerInnen mit viel Zeitaufwand verbunden war. Ebenso wäre es sehr zeitaufwändig gewesen, 50 Topics mit lediglich vier studentischen ProjektteilnehmerInnen abzufragen und auszuwerten

Die neue Projektgruppe nahm ihre Arbeit mit einer Überprüfung der Relevanzurteile auf, die zwei Jahre zuvor erarbeitet worden waren. 5 Topics wurden beispielhaft ausgewählt und es wurden von den TeilnehmerInnen in einer verkleinerten Form des Poolingverfahrens relevante Titel aus der Kollektion gesucht. Dabei entstanden keine signifikanten Abweichungen zu den Ergebnissen der ersten Projektgruppe, sodass eine Überarbeitung der Topics oder Relevanzurteile nicht sinnvoll erschien.

Bei den ersten, voll automatisch durchgeführten Evaluationsläufen zeigten sich neue Probleme. Nur 37 von 50 Topics generierten überhaupt Ergebnismengen. 13 Topics erzielten Nulltrefferergebnisse, die eigentlich unbedingt zu vermeiden sind. Eine Antwort auf die Frage, wie es zu diesen Ergebnissen kommen konnte, blieb zunächst aus. Unklar

[216] Böckmann, S.1.
[217] Ebd., S.6.
[218] Seegert, S.22.

blieb, warum die automatisch vergebenen Schlagwörter im Vergleich schlechter abschnitten, denn es kam mitunter zu großen Unterschieden in Recall und Precision zu Gunsten der intellektuell vergebenen GND-Schlagwörter, die isoliert betrachtet besser abschnitten Weitere Evaluationsläufe mit veränderten Parametern änderten nichts an der Menge der Nulltrefferergebnisse. Zum Teil wuchs ihre Zahl auf 15 von 50 Topics ohne Ergebnismengen.

Um diese inakzeptablen Ergebnisse zu erklären, entwickelte die Projektgruppe die Hypothese, dass die GND-Vorzugsbenennungen in der Testkollektion mit der Retrievalleistung interferierten. Eine automatische Verschlagwortung kann u.U. ungünstig gewählte Vorzugsbenennungen nicht kompensieren.

Auf dieser Annahme aufbauend entwickelte die Projektgruppe die Idee, Synonyme in den automatischen Abfrageprozess einzubeziehen. Die Frage lautete dabei, ob mit der Zuhilfenahme von Synonymen dieselben Dokumente gefunden wurden wie mit den Vorzugsbenennungen. Ein Beispiel: Hyperaktivität ist ein Synonym für die GND-Vorzugsbenennung Hyperkinese. In den Dokumententiteln steht oft Hyperaktivität anstatt Hyperkinese. Eine Suche in den Titelfeldern erbringt daher mehr relevante Ergebnisse als die Suche in den Schlagwortkategorien. Dies war eine mögliche Erklärung für die bis dahin schlechten Retrievalergebnisse.

Daraus ergab sich die Frage, wie diese neuen Ideen in das Experimentsetup einfließen könnten und welchen Mehrwert die Synonyme mitbringen. Über den Synonym-Graphfilter wurden Synonyme in die Solr-Schreibweise überführt und in einer eigenen Datei neben der Dokumentenkollektion abgespeichert, die während Evaluationsläufen abgefragt wurde. Weitere Überlegungen führten zu Experimenten mit Solrs Stemming-Funktionen, um zu untersuchen, ob sich auf diese Weise die Retrievalergebnisse verbessern ließen. Weitere Unklarheiten lagen im Einsatz der booleschen Operatoren. Neue Suchdurchläufe sollten klären, ob Suchanfragen mit AND bessere Ergebnisse lieferten als mit OR. Dies wurde auf die Topics selbst ausgedehnt, indem bei Phrasentopics die einzelnen Begriffe mit AND oder OR für Suchdurchläufe verbunden wurden. Dieser Versuch zeigte, dass die Synonyme mitunter keinen Einfluss auf die Ergebnisse hatten, entweder weil ein Topic bereits mit der jeweiligen Vorzugsbenennung formuliert worden war oder weil keine Vorzugsbenennung existierte. Recall und Precision verschlechterten sich oft, wenn die Narratives einbezogen wurden, da die Treffermengen zu groß wurden.

Als Endergebnis für die Arbeit im Wintersemester zeigte sich, dass der Einsatz von Synonymen oder Stemming kaum Einfluss auf das Gesamtergebnis hatte. Bei den Topics „Magersucht“ und „Tierexperimente“ war die Zuhilfenahme einer Synonymliste immerhin erfolgreich, bei allen anderen Topics jedoch nicht.

## 4.9 Neuaufsetzen der Dokumentenkollektion mit einer Pipeline-Lösung

Nach diesen wenig ermutigenden Ergebnissen beschrieb die Projektteilnehmerin Ina Böckmann ein komplettes Neuaufsetzen der Kollektion zur weiteren Verwendung mit einem Pipeline-Mechanismus, der die Kollektion automatisch per Schnittstellen-Verbindung zur DNB[219] aktualisiert und Erweiterungen und Verwendung durch zukünftige Projektgruppen ermöglicht. Zunächst ging es darum, einen neuen Korpus aufzubauen, der aktualisiert und gefiltert ist, aber noch nicht erweitert wurde.[220] Böckmann präferiert hierfür eine Lösung, die sich an der Programmiersprache Python orientiert, da zu erwarten ist, dass zukünftige ProjektteilnehmerInnen Kenntnisse in dieser Sprache mitbringen. Die Pipeline-Lösung muss so konzipiert sein, dass sie gut dokumentiert und strukturell verständlich aufgebaut ist.[221]

Böckmann schlägt eine Neu-Standardisierung der Datenfelder innerhalb der Kollektion vor und gibt an, welche Inhalte jedes Feld enthält und wie neue Feldbezeichnungen lauten sollen, sofern diese als nötig angesehen werden. Z.B. wird aus subject_gnd, der bisherigen Feldbezeichnung für intellektuell vergebene Schlagwörter, das Feld subject_int. Eine Neuschöpfung ist das Feld subject_other, das Schlagwörter anderer Herkunft als intellektuell, automatisch oder durch Verlage vergeben sammelt. Eine Orientierung an bekannten Regelwerken bei der Felderbenennung (DDC, RDA, MARC-Standards) wird befürwortet um bibliothekarisch ausgebildeten MitarbeiterInnen den Einstieg in die Nutzung der Pipeline zu erleichtern.[222]

Für eine möglichst verlustfreie Konvertierung der DNB-Daten wird das Format MARCXML angestrebt.[223] Ein unregelmäßiges Herunterladen des Gesamtabzugs der DNB wäre die unkomplizierteste Herangehensweise, um die DNB-Daten aktuell zu

[219] Böckmann, S.35.
[220] Ebd., S.6-7.
[221] Ebd., S.18.
[222] Ebd., S.19-21.
[223] Ebd., S.35.

halten. Wichtig hierbei ist es, die Rohdaten der DNB bei jeder Aktualisierung zu sichern, da sich der Prozess zur Erstellung der Testkollektion wahrscheinlich ändern wird.[224]

Die Pipeline soll vollautomatisch arbeiten, damit Varianzen durch manuelles Eingreifen verhindert werden und die Vergleichbarkeit der Daten beibehalten wird. Die Programmierung muss verständlich und wartungsfreundlich sein, damit bei einer Strukturveränderung der Rohdaten oder in Python selbst leicht Anpassungen vorgenommen werden können. Außerdem muss der Code so gut dokumentiert sein, dass nachfolgende Projektgruppen eine Grundlage haben, um sich in die Materie einzuarbeiten. 5 Module kommen bei der Datenbeschaffung zum Einsatz: Herunterladen, Filtern, Dekodieren, Transformieren und Kodieren der MARCXML-Daten der DNB in einen für GELIC nutzbaren Korpus.[225]

Beim Filtervorgang werden zunächst Titel anhand ihrer ID herausgefiltert, die in der bisherigen Testkollektion nicht vorkamen. Dies gilt auch für Titel, die weder automatisch noch intellektuell verschlagwortet sind. Mindestens eines der Schlagwortfelder muss über Inhalte verfügen, damit es für die Testkollektion zugelassen wird.[226]

Beim Dekodieren wird für jedes abgefragte inhaltstragende Feld die passende Verfahrensanweisung in Python ausgewählt und die Ergebnisse in ein oder zwei Variablen abgespeichert. Bei den Schlagwortfeldern wird nach vier Varianten unterschieden: intellektuell, automatisch, vom Verlag vergeben, aus Fremdquellen stammend.[227]

Manche Felder können nach dem Dekodieren direkt in die Testkollektion übernommen werden. Andere Daten müssen entsprechend verteilt, also transformiert werden. Dies gilt im Besonderen für die Schlagwörter. Die intellektuell und automatisch vergebenen Schlagwörter werden je den Feldern subject_int und subject_auto zugeteilt, und die restlichen Schlagwörter werden dem Feld subject_other zugeordnet.[228]

Der mit der Pipeline erzeugbare Korpus verfügt nach Abzug von 2848 Datensätzen, die nicht den Kriterien entsprechen, über 197.152 Titel. Ende 2020 befinden sich 63.169 automatisch erschlossene, 189.931 intellektuell erschlossene und 55.948 doppelt verschlagwortete Datensätze in der Kollektion. Die automatisch und doppelt erschlossenen

[224] Böckmann, S.36-37.
[225] Ebd., S.38.
[226] Ebd., S.43.
[227] Ebd., S. 47-48.
[228] Ebd., S.51-52.

Datensätze haben sich z.T. stark verringert (32% Anteil gegenüber vorher 61%), während der Anteil intellektuell erschlossener Titel zugenommen hat (mehr als 55.000 mehr, 96% statt wie vorher 68%). Böckmann führt diese Verschiebung u.a. darauf zurück, dass automatisch erzeugte Schlagwörter von den intellektuell vergebenen Termen in vielen Fällen nicht zu unterscheiden sind, da Angaben zur Metadatenherkunft fehlen. Es bleibt unklar, warum sich die Anzahl doppelt erschlossener Titel verringert hat.[229] Eine mögliche Erklärung mag in der Entscheidung der DNB liegen, die Reihe B nicht mehr automatisch zu erschließen, da die Qualität der automatischen Indexierung als unzureichend eingestuft wurde.[230]

Anschließend wurde mit der alten Version 2 der Testkollektion und dem durch die Pipeline erzeugten Korpus ein vergleichender Test durchgeführt. In drei Fällen erzielt die neue Version mehr Treffer, in einem die gleiche Anzahl und in zwei Fällen weniger als die vorherige Version. Böckmann gibt aber zu bedenken, dass z.B. bei den 67 Treffern der 4. Fragevariante Titel dabei sein könnten, die fälschlicherweise den intellektuellen Schlagwörtern zugeordnet wurden.[231]

Aufgrund der Tatsache, dass nicht immer zwischen intellektuell und automatisch erzeugten Schlagwörtern unterschieden werden kann, kann aktuell kein Korpus basierend auf den IDs der ersten Dokumentenkollektion mit der derzeitigen Datenlage der öffentlich abrufbaren Formate erstellt werden, der für Retrievaltests zum Vergleich der Retrievalqualität von intellektuell und automatisch erzeugten Schlagwörtern geeignet wäre.[232]

Böckmann sieht drei Möglichkeiten, wie das Projekt trotzdem vorwärtskommen kann:

- das Warten auf die Anreicherung von Altdaten. Dies hätte geringen Arbeitsaufwand zur Folge, aber es ist unsicher, ob die DNB dies jemals tun wird.
- Projekt GELIC könnte die DNB um regelmäßige Datenselektionen im Format Pica+ bitten, die in PICAXML umgewandelt und mit Hilfe der Pipeline transformiert und weiter genutzt werden könnten. Dadurch würde sich GELIC allerdings mehr von der DNB und externen Konversionstools abhängig machen, die den Konvertierungsprozess wieder aufwändiger gestalten würden. All dies wäre mit

[229] Böckmann, S.55.
[230] Ebd., S.57.
[231] Ebd., S.56.
[232] Ebd., S.58.

viel Aufwand verbunden, und die automatisch ablaufende Pipeline müsste in gutem Zustand gehalten werden.

- Das neue Aufsetzen eines Korpus mit Titeln, die ab 2018 erschlossen wurden, mit der Filterfunktion von gelic_mt. Diese sind vollständiger beschrieben und ermöglichen dadurch aussagekräftigere Analysen. Dies würde ein neues Poolingverfahren zum Fällen von Relevanzurteilen erfordern. Das Projekt hätte dadurch mehr Informationen zur Herkunft der Metadaten und könnte die Schlagwörter genauer einteilen. Durch Fremddatenübernahme erzeugte Schlagwörter ließen sich so leichter von den anderen Schlagwortarten unterscheiden. Diese Variante ist sehr zeit- und arbeitsaufwändig, bietet aber großes Potential.[233]

Soweit es dem Autor bekannt ist, repräsentiert das Konzept der Pipeline-Lösung den aktuellen Stand der Arbeit an GELIC. Dieses muss weiterhin als work-in-progress betrachtet werden, und zukünftige Projektgruppen werden darüber zu entscheiden haben, wie das Projekt fortgeführt werden kann.

# 5. Fazit und Ausblick

## 5.1 Fazit

Wie oben dargestellt, hat die Retrievalforschung eine lange Geschichte in der Bibliothekswissenschaft. In den 1950ern wurden mit den Uniterm- und Cranfield-Experimenten wichtige Grundlagen für Retrievalexperimente unter Labor-Bedingungen gelegt, wie sie bis heute stattfinden. Verwandte Projekte wie das MEDLARS-Experiment verfolgten andere Zielsetzungen und fanden im laufenden Bibliotheksbetrieb statt. Sie übernahmen manchmal allerdings Aspekte des Cranfield-Paradigmas wie die etablierten Kennzahlen Recall und Precision zur Darstellung der Ergebnisse.

Zwei Retrievalexperimente aus Australien (Byrne u. Micco 1988) und den USA (F.W. Lancaster 1991, der bereits bei MEDLARS involviert war) zeigten die Möglichkeiten auf, Retrievaltests im Rahmen des catalogue enrichment auf die damals neuen Bibliotheks-OPACs anzuwenden. Obwohl beide Experimente zu höchst unterschiedlichen Schlussfolgerungen bzgl. der Möglichen der sachlichen Recherche in OPACs kamen, beeinflussten sie die ersten Retrievalexperimente im deutschsprachigen Raum. Die MILOS-Experimente und das Nachfolgeprojekt KASCADE zeigten auf, dass die Möglichkeiten der

[233] Böckmann, S. 59.

automatischen Indexierung die Qualität des sachlichen Retrievals in OPACs signifikant verbessern können. MILOS im Besonderen erwies sich als vielbeachtetes Projekt, das Nachfolgeexperimente wie sowohl Martin Grummanns Anwendung der MILOS-Abläufe auf die sachliche Erschließung in den OPACs öffentlicher Bibliotheken als auch die Arbeit von Oberhauser und Labner bzgl. automatischer Indexierung nach MILOS-Vorbild im Verbundkatalog des ÖBV inspirierte. Beide Experimente bestätigten die positiven Ergebnisse der MILOS-Tests.

Ebenfalls vorgestellt wurden Projekte, die bewusst andere oder breitere Ansätze bei Retrievalexperimenten wählten. Dazu zählen die hier vorgestellten Projekte OSIRIS sowie Harald Kaluzas Vergleich zwischen Google Scholar und dem EBSCO Discovery Service bei einer Suche nach vorher ausgewählten Dokumenten, anstelle einer sachlichen Rechercheaufgabe. Aus Platzgründen konnten weitere interessante Retrievalprojekte wie das kollaborative Online-Erschließungsportal DANDELON[234] oder das LibRank-Experiment der HAW Hamburg und der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBIW)[235] zur Erforschung neuer Relevanzrankingverfahren hier leider nicht berücksichtigt werden.

Besondere Aufmerksamkeit wurde dem Ablauf des Retrievalprojekts GELIC zur Etablierung einer wiederverwendbaren Testkollektion gewidmet, mit der die intellektuellen und automatischen Erschließungsmethoden der DNB evaluiert werden sollen. Dieses Ziel ist noch nicht erfüllt, obwohl Ina Böckmanns Ansatz einer Pipeline-Lösung zur automatischen Aktualisierung und Indexierung eines Korpus aus DNB-Daten vielversprechend erscheint. Es wird interessant sein zu sehen, wie es mit GELIC weitergehen wird.

## 5.2. Ausblick

Währenddessen hat sich die Situation, die zum Start des GELIC-Projekts beigetragen hat, weiterentwickelt. Die DNB hat ihre Bestrebungen, automatische Indexierungsmechanismen in ihre Erschließungsarbeit einzugliedern, fortgesetzt. Dabei geht sie jedoch mittlerweile andere Wege. Die Zusammenarbeit mit der Firma Averbis, deren Software bisher die automatische Indexierungsarbeit übernahm, wurde mittlerweile beendet. Stattdessen

[234] Siehe dazu: Hauer, Manfred: intelligentCAPTURE und dandelon.com: Collaborative Catalogue Enrichment. In: Open Innovation. Neue Perspektiven im Kontext von Information und Wissen. Konstanz: UVK Verl., 2007. S. 403-412.

[235] Siehe dazu: Behnert, Christiane; Borst, Timo: Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen; Das DFG-Projekt LibRank. In: Bibliothek – Forschung Praxis 39 (3), 2015. S.384-393.

arbeiten die ZBIW und die DNB derzeit am Projekt EMa bzw. „Erschließungsmaschine". EMa soll u.a. Klassifikationen nach verschiedenen Systemen vornehmen und mit kontrolliertem Vokabular automatisch indexieren können sowie um neue Funktionen und Verfahren erweiterbar sein, um die Erschließungsergebnisse kontinuierlich zu verbessern.[236]

Im Kern von EMa liegt das Open-Source-Toolkit Annif, das an der Finnischen Nationalbibliothek in Entwicklung ist. Annif nutzt sprachunabhängige Text-Mining- und Machine-Learning-Verfahren für automatische Klassifikations- und Indexierungsaufgaben und ist nach Einschätzung von Sandro Uhlmann für den Einsatz in der DNB-Erschließungsarbeit geeignet.[237] Erste interne Retrievaltests mit Annif, bei denen das Toolkit 434 Datensätze um 1650 Schlagwörter anreicherte verliefen vielversprechend. Die vergebenen Schlagwörter wurden von DNB-Mitarbeitern intellektuell überprüft. 43% dieser Schlagwörter wurden als sehr nützlich, 26% als nützlich, 22% als weniger nützlich und 9% als falsch bewertet. Diese Ergebnisse übertreffen die der Averbis-Altsoftware in jeder Hinsicht.[238]

Die DNB zieht für ihre Arbeit mit automatischer Indexierung bisher eine positive Bilanz. Nach eigener Aussage möchte sie in Kooperation mit anderen Einrichtungen „[…] Hauptakteurin eines Kompetenznetzwerks für maschinelle Erschließungsverfahren im Bibliotheks- und Informationswesens […]"[239] werden. Jedoch hat die klassische intellektuelle Erschließung noch nicht ausgedient. Auch nach Jahrzehnten der Erforschung und Erprobung automatischer Indexierung werden beide Erschließungsmethoden weiterhin in der DNB gemeinsam eingesetzt.[240] Es wird interessant sein zu sehen, wie sich der Bereich der Retrievalforschung und der Einsatz automatischer Indexierungsmethoden in deutschen Bibliotheken weiter entwickeln wird.

[236] Uhlmann, Sandro: EMa – Erschließungsmaschine. Automatische Vergabe von GND-Schlagwörtern mit Annif – Ergebnisse einer Evaluation im DNB-Projekt EMa. Powerpoint-Präsentation, Deutsche Nationalbibliothek. 03.12.2020. https://wiki.dnb.de/download/attachments/181751388/2-3_Automatische-Vergabe-von-GND-Schlagw%C3%B6rtern_Uhlmann_2020-12-03_final.pdf?version=1&modificationDate=1607352872000&api=v2 (Letzter Zugriff: 11.07.2022) S. 4-5.

[237] Ebd. S.19.

[238] Ebd., S.17.

[239] Junger Ulrike; Scholze, Frank: Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek. In: Qualität in der Inhaltserschließung. Bibliotheks- und Informationspraxis, Band 70. Berlin/Boston: de Gruyter, 2021. S. 68.

[240] Mödden, Elisabeth; Karg, Helga: Automatisch generierte Vorschläge für die intellektuelle Inhaltserschließung in der DNB. Powerpoint-Präsentation, Deutsche Nationalbibliothek, 11.11.2021. https://wiki.dnb.de/download/attachments/217540739/Vorschlagskomponente_DNB.pdf?version=2&modificationDate=1636974853000&api=v2 (Letzter Zugriff: 11.07.2022).

# Literaturverzeichnis

Behnert, Christiane; Borst, Timo: Neue Formen der Relevanz-Sortierung in bibliothekarischen Informationssystemen. Das DFG-Projekt LibRank. In: Bibliothek – Forschung und Praxis 39 (3), 2015. S.384-393.

Böckmann, Ina. Aufbau einer Pipeline zur Transformation von Titeldaten der DNB. Unveröffentlichte Bachelorarbeit. Köln: TH Köln, 2020.

Borlund, Pia: Interactive Information Retrieval: An Introduction. In: Journal of Information Science Theory and Pratice 1 (3), 2013.

Busse, Frank; Grote, Claudia; Jacobs, Jan Helge et al.: Erschließungsmaschine gestartet. DNB-Blog. 4.5.2022 https://blog.dnb.de/erschliessungsmaschine-gestartet/ (Letzter Zugriff: 01.07.2022).

Byrne, Alex; Micco, Mary: Improving OPAC Subject Access: The ADFA Experiment. In: College & Research Libraries 49 (5), 1988. S.432-441.

Ceynowa, Klaus: In Frankfurt lesen jetzt zuerst Maschinen. Frankfurter Allg. Zeitung. 31.07.2017. Online unter: https://www.faz.net/aktuell/feuilleton/buecher/maschinen-lesen-buecher-deutsche-nationalbibliothek-setzt-auf-technik-15128954.html (Registrierung erforderlich; letzter Zugriff: 06.07.2022).

Cleverdon, Cyril; Keen, Michael: ASLIB Cranfield Research Project. Factors Determining the Performance of Indexing Systems. Cranfield: Aslib, 1966.

Clough, Paul; Sanderson, Mark: Evaluating the performance of information retrieval systems using test collections. In: Information Research 18 (2), 2013. http://informationr.net/ir/18-2/paper582.html#.Yr90P4TP1PY (Letzter Zugriff: 01.07.2022).

Deutsche Nationalbibliothek: Grundzüge und erste Schritte der künftigen inhaltlichen Erschließung von Publikationen in der Deutschen Nationalbibliothek. Stand Mai 2017. https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/konzeptWeiterentwicklungInhaltserschliessung.pdf?__blob=publicationFile&v=4 (Letzter Zugriff: 01.07.2022).

Gödert, Wilfried; Liebig, Martina: Maschinelle Indexierung auf dem Prüfstand. Ergebnisse eines Retrievaltests zum MILOS II Projekt. In: Bibliotheksdienst 31 (1), 1997. S.59-68.

Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken 20 (1), 2010. S.20-22.

Gray, Dwight E.: Report on the Reference Test of the Conventional and Uniterm Systems. Washington, D.C.: ASTIA Reference Center, 1954.

Grummann, Martin: Sind Verfahren zur maschinellen Indexierung für Literaturbestände Öffentlicher Bibliotheken geeignet? In: Bibliothek 24 (3), 2000. S. 297-318.

Hauer, Manfred: intelligentCAPTURE und dandelon.com: Collaborative Catalogue Enrichment. In: Open Innovation. Neue Perspektiven im Kontext von Information und Wissen. Hrsg. von Achim Oßwald, Maximilian Stempfhuber, Christian Wolff. Konstanz: UVK Verl., 2007. S. 403-412.

Junger Ulrike; Scholze, Frank: Neue Wege und Qualitäten – Die Inhaltserschließungspolitik der Deutschen Nationalbibliothek. In: Qualität in der Inhaltserschließung. Bibliotheks- und Informationspraxis, Band 70. Berlin/Boston: de Gruyter, 2021. S.55-70.

Kaluza, Harald: Google Scholar versus EBSCO Discovery Service: Ein vergleichender Retrieval-Test. In: b.i.t.online innovativ, Bd. 44. Malis Praxisprojekte 2013. Projektberichte aus dem berufsbegleitenden Master-Studiengang Bibliotheks- und Informationswissenschaft der Fachhochschule Köln. Hrsg. von Achim Oßwald. Wiesbaden: Dinges & Frick, 2013. S.59-79.

Kostädt, Peter: Einführung eines Discovery Service in der in der Universitäts- und Stadtbibliothek Köln. In: Ein Bibliothekar mit Informationskompetenz. Festschrift zum 60. Geburtstag von Dr. Rolf Thiele. Elektronische Schriftenreihe der der Universitäts- und Stadtbibliothek Köln, Band 5. Hrsg. von Wolfgang Schmitz, Katja Halassy, Irmgard Jordan-Schmidt. Köln: Universitäts- und Stadtbibliothek Köln, 2012. S.77-87

Lancaster, F.W.: Evaluation of the MEDLARS Demand Search Service. January 1968. National Library of Medicine. Bethesda, MD, 1968.

Lancaster, F.W; Harkness Connell, Cora; Bishop, Nancy et al: Identifying Barriers to Effective Subject Access in Library Catalogs. In: Library Resources and Technical Services 35 (4), 1991. S.377-391.

Lepsky, Klaus: Automatisierung in der Sacherschließung: Maschinelles Indexieren von Titeldaten. In: Die Herausforderung der Bibliotheken durch elektronische Medien und neue Organisationsformen. Zeitschrift für Bibliothekswesen und Bibliographie: Sonderheft 63. Hrsg. von Sabine Wefers. Frankfurt: Klostermann 1996, S. 223-233.

Lepsky, Klaus: Automatische Indexierung und bibliothekarische Inhaltserschließung: Ergebnisse des DFG-Projekts MILOS I. In: Zukunft der Sacherschließung im OPAC. Schriften der Universitäts- und Landesbibliothek Düsseldorf. Band 25. Hrsg. von Elisabeth Niggemann. Düsseldorf: Universitäts- und Landesbibliothek Düsseldorf, 1996. S.13-36.

Lepsky, Klaus: Maschinelle Indexierung von Titelaufnahmen zur Verbesserung der sachlichen Erschließung in Online-Publikumskatalogen. Dissertation. Kölner Arbeiten zum Bibliotheks- und Dokumentationswesen. Heft 18. Köln: Greven Verl., 1994.

Lepsky, Klaus.; Siepmann, Jörg; Zimmermann, Andrea: Automatische Indexierung für Online-Kataloge – Ergebnisses eines Retrievaltests. In: Zeitschrift für Bibliothekswesen und Bibliographie 43 (1), 1996. S.47-56.

Lepsky, Klaus; Zimmermann, Harald H.: Katalogerweiterung durch Scanning und automatische Dokumenterschließung. Ergebnisse des DFG-Projekts KASCADE. In: Zeitschrift für Bibliothekswesen und Bibliographie 47 (4), 2000. S.305-316. Online-Ausgabe ohne Seitenzahlen. https://publikationen.sulb.uni-saarland.de/bitstream/20.500.11880/25554/1/2000a.pdf (Letzter Zugriff: 12.07.2022)

Lewandowski, Dirk: Evaluating the retrieval effectiveness of Web search engines using a representative query sample. In: Journal of the Association for Information Science and Technology 66 (9), 2014. Preprint-Ausgabe ohne Seitenzahlen. https://arxiv.org/ftp/arxiv/papers/1511/1511.05817.pdf (Letzter Zugriff: 10.07.2022)

Lohmann, Hartmut: KASCADE: Dokumentanreicherung und automatische Inhaltserschließung. Projektbericht und Ergebnisse des Retrievaltests. Schriften der Universitäts- und Landesbibliothek Düsseldorf 31. Düsseldorf: Universitäts- und Landesbibliothek Düsseldorf, 2000.

Mittelbach, Jens; Probst, Michaela: Möglichkeiten und Grenzen maschineller Indexierung in der Sacherschließung. Strategien für das Bibliothekssystem der Freien Universität Berlin. Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft. Heft 183. Berlin: Institut für Bibliotheks- und Informationswissenschaft der Freien Universität Berlin, 2006.

Mödden, Elisabeth; Karg, Helga: Automatisch generierte Vorschläge für die intellektuelle Inhaltserschließung in der DNB. Powerpoint-Präsentation. Deutsche Nationalbibliothek, 11.11.2021. https://wiki.dnb.de/download/attachments/217540739/Vorschlagskomponente_DNB.pdf?version=2&modificationDate=1636974853000&api=v2

Munkelt, Johanna: Erstellung einer DNB-Retrieval-Testkollektion. Unveröffentlichte Bachelor-Arbeit. Köln, 2018.

Munkelt, Johanna; Schaer, Philipp; Lepsky, Klaus: Towards an IR Test Collection for the German National Library. In: Proceedings of the Conference „Lernen, Wissen, Daten, Analysen“, LWADA 2018. Hrsg. von Rainer Gemulla, Simone Paolo Ponzetto, Christian Bizer et al. Mannheim, 2018. S.275-280.

Nohr, Holger: Grundlagen der automatischen Indexierung. Ein Lehrbuch. 3., überarb. Aufl. Berlin: Logos-Verl., 2005.

O’Brien, Ann: Relevance as an aid to evaluation in OPACs. In: Journal of Information Science 16 (4), 1990. S.265-271.

Oberhauser, Otto; Labner, Josef: OPAC-Erweiterung durch automatische Indexierung. Empirische Untersuchung mit Daten aus dem Österreichischen Verbundkatalog. In: ABI-Technik 23 (4), 2003. S. 305-314.

Rasmussen, Edie: Evaluation in Information Retrieval. In: „The MIR/MDL Evaluation Project White Paper Collection“. Edition #3. Workshop on the Evaluation of Music Information Retrieval Systems. Toronto, 2003.

Recker, Ingrid; Ronthaler, Marc; Zillmann, Hartmut: OSIRIS. Osnabrück Intelligent Research Information System – ein Hyper-Base Front End System für OPACs. In: Bibliotheksdienst 30 (5), 1996. S.833-848.

Robertson, Stephen: On the history of evaluation in IR. In: Journal of Information Science 34 (4), 2008. S.439-456.

Ronthaler, Marc: Dialog-Schnittstellen in Online-Informationssystemen: Notwendigkeit, Leistungsfähigkeit und Entwicklungsmöglichkeiten am Beispiel des OSIRIS-Systems. Dissertation. Osnabrück: Universität Osnabrück, Oktober 2000.

Sachse, Elisabeth; Liebig, Martina; Gödert, Wilfried: Automatische Indexierung unter Einbeziehung semantischer Relationen: Ergebnisse des Retrievaltests zum MILOS II Projekt. Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft. Band 14. Köln: Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen, 1998.

Seegert, Christin: Evaluation von automatischer und intellektueller Verschlagwortung der Deutschen Nationalbibliothek anhand einer Retrieval-Testkollektion. Unveröffentlichte Bachelorarbeit. Köln: Technische Hochschule Köln, 2019.

Teague-Sutcliffe, Jean: The Pragmatics of Information Retrieval Experimentation Revisited. In: Information Processing & Management 28 (4), 1992. S.467-490.

Uhlmann, Sandro: EMa – Erschließungsmaschine. Automatische Vergabe von GND-Schlagwörtern mit Annif – Ergebnisse einer Evaluation im DNB-Projekt EMa. Powerpoint-Präsentation. Deutsche Nationalbibliothek. 03.12.2020. https://wiki.dnb.de/download/attachments/181751388/2-3_Automatische-Vergabe-von-GND-Schlagw%C3%B6rtern_Uhlmann_2020-12-03_final.pdf?version=1&modificationDate=1607352872000&api=v2 (Letzter Zugriff: 08.07.2022)

Vorhees, Ellen M.: Continuing Information Retrieval's Tradition of Experimentation. In: Communications of the ACM 50 (11), 2007. S.52-53.

Wiesenmüller, Heidrun: Das neue Sacherschließungskonzept der DNB in der FAZ. Basiswissen RDA. August 2017. Das neue Sacherschließungskonzept der DNB in der FAZ - Basiswissen RDA (basiswissen-rda.de) (Letzter Zugriff: 06.07.2022).

# Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum

Rechtsverbindliche Unterschrift