Entwicklung einer Metadaten-Harvest-Struktur für Dissertationen aus dem informationswissenschaftlichen Spektrum

Bachelorarbeit zur Erlangung des Bachelor-Grades
Bachelor of Science im Studiengang Data and Information Science
an der Fakultät für für Informations- und Kommunikationswissenschaften
der Technischen Hochschule Köln

vorgelegt von: Leon Paul Mondrian Munz

eingereicht bei: Prof. Dr. Philipp Schaer

Zweitgutachter/in: Björn Engelmann

Köln, 11.01.2022

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

Kurzfassung/Abstract

Ziel der vorliegenden Arbeit ist die Generierung einer möglichst vollständigen Sammlung von Metadaten, referenzierend auf Dissertationsschriften, die dem erweiterten Themenspektrum der Informationswissenschaft entsprechen. Hierzu stellen sich die folgenden Fragen: Welche Disziplinen können als relevant für das erweiterte Themenspektrum der Informationswissenschaft betrachtet werden? Existiert eine Vollständige Übersicht über die Landschaft der deutschen Hochschulschriftenserver? Wie muss das System aufgebaut werden um die Metadaten, die auf Dissertationsschriften referenzieren, zu beziehen, zu selektieren und zu homogenisieren?

Um diese Fragen zu beantworten wird erarbeitet, aus welchen Disziplinen sich das erweiterte informationswissenschaftliche Themengebiet zusammensetzt. Ebenso wird eine Vollständige Liste aller deutschen Hochschulen und der identifizierten Hochschulschriftenserver angefertigt. Diese erarbeiteten Hochschulschriftenserver werden mittels eines Softwareentwurfs über das Open Archives Initiative Protocol for Metadata Harvesting abgefragt. Die erhaltenen Metadaten werden homogenisiert abgelegt. Weiter findet eine Schlagwortsuche nach programmatisch erstellten Schlagworten innerhalb der Disziplinen statt, die der Informationswissenschaft als nahestehend ermittelt wurden. Aus der Aufgabenstellung resultierend ergab sich, dass die Methoden und Erkenntnisse der Informatik und der Bibliothekswissenschaft als relevant für das erweiterte Themenspektrum der Informationswissenschaft betrachtet werden können. Durch den Harvesting Prozess konnten zwei Sammlungen von Metadatensätzen erstellt werden. Eine Sammlung umfasst 378 Metadatensätze, die ausschließlich Dissertationen aus den Bibliotheks- und Informationswissenschaften beinhalten und eine weitere Sammlung besteht aus 3.698 Metadatensätzen, die dem erweiterten Themenspektrum der Informationswissenschaft entsprechen.

Kurzfassung/Abstract englisch

The aim of this work is to generate a complete collection of metadata referring to dissertation publications that correspond to the extended range of topics in information science. Thus, the following questions arise: Which disciplines can be considered relevant to the broader range of topics in information science? Does a complete overview of the landscape of German university repositories exist? How is the system to be structured in order to relate, select and homogenize the metadata referencing dissertation theses?

In order to answer these questions the disciplines that make up the broader field of information science will be identified. Likewise, a complete list of all German universities and the identified university repository servers will be compiled. These identified university repositories will be queried by means of a software design using the Open Archives Initiative Protocol for Metadata Harvesting. The metadata obtained is stored in a homogenized manner. Furthermore, a keyword search for programmatically created keywords within the disciplines that have been identified as being close to information science will take place. As a result of the task definition, the methods and findings of computer science and library science can be considered relevant for the extended range of topics in information science. Through the harvesting process, two collections of metadata records were created. One collection consisting of 378 metadata records that exclusively contain dissertations from library and information science and another collection consisting of 3,698 metadata records that correspond to the broader range of topics in information science.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Einleitung

Als zentrale Aufgabe des Hochschulverbands Informationswissenschaft gilt die Stärkung der informationswissenschaftlichen Forschung und Lehre in deutschsprachigen Ländern. Darunter fällt unter anderem auch die Förderung des Wissens- und Technologietransfers innerhalb der Wissenschaft sowie der Transfer informationswissenschaftlicher Forschungsergebnisse in die Öffentlichkeit (Hochschulverband Informationswissenschaft, o.D.). Zur Unterstützung dieser Aufgaben und um Trends der Disziplin beobachten zu können, kann eine aktuelle Sammlung von Dissertationen aus dem informationswissenschaftlichen Themenspektrum nützlich sein. Ebenso kann eine solche Sammlung relevanter Dissertationen eines Fachgebiets für Studenten und die Wissenschaftler gleichermaßen relevant sein. Üblicherweise werden Metadaten von Hochschulpublikationen auf Hochschulschriftenservern zur Verfügung gestellt. Aktuell existiert jedoch kein System, das die Metadaten von veröffentlichten Dissertationen aus dem erweiterten informationswissenschaftlichen Themengebiet bereitstellt.

Demensprechend soll, ausgehend von der erwähnten Problemstellung, ein System entwickelt werden, das auf der Grundlage von Metadaten-Harvesting eine einheitliche Sammlung aller auf Hochschulschriftenservern zur Verfügung gestellten Metadaten zu Dissertationen aus dem erweiterten informationswissenschaftlichen Themengebiet erstellt. Hierzu stellen sich die folgenden Fragen:

Welche Disziplinen können als relevant für das erweiterte Themengebiet der Informationswissenschaft betrachtet werden?

Existiert eine Vollständige Übersicht über die Landschaft der deutschen Hochschulschriftenserver?

Wie muss das System aufgebaut werden um die Metadaten, die auf Dissertationsschriften referenzieren, zu beziehen, nach Wissensgebiet zu selektieren und zu homogenisieren?

Die Arbeit gliedert sich in fünf übergeordnete Teile. Der erste Teil erläutert die für diese Arbeit relevanten Grundlagen. Der zweite übergeordnete Teil beschreibt das praktische Vorgehen zur Lösung der Aufgabenstellung. Hier wird die Methodik zur Erarbeitung einer Liste der deutschen Hochschulschriftenserver und die zugrunde liegende Datenheterogenität einiger Datenanbieter betrachtet. Der Teil schließt mit der Beschreibung der Softwarearchitektur, die verwendet wird, um die Metadaten zu beziehen und in einer einheitlichen Struktur abzulegen. Innerhalb des dritten Teils wird die Evaluation der verschiedenen Module des Softwarentwurfs anhand einer deskriptiven Analyse der bezogenen Metadaten und Testsets präsentiert. Der vierte Teil präsentiert die erhaltenen Ergebnisse und diskutiert diese um darauf folgend

mit einem Fazit in Anbetracht der Ergebnisse und einem entsprechenden Ausblick abzuschließen.

Der gesamte Quellcode des in dieser Arbeit erstellen Softwareentwurfs und die erzeugten Daten die für die Analysen der Ergebnisse verwendet wurden liegen der gedruckten Version dieser Bachelorthesis als DVD bei. Ebenso stehen der Quellcode und die erzeugten Daten unter der DOI: 10.5281/zenodo.6038417 oder der URL: https://zenodo.org/record/6038417#.YgYiPJYxmUk zur Verfügung.

1 Grundlagen und Begriffe

1.1 Promotionspraxis in Deutschland

Da das primäre Ziel dieser Arbeit das Sammeln und Selektieren von Dissertationsschriften ist, wird im folgenden Kapitel die Promotion und deren Rahmenbedingungen in der Praxis erläutert. Als Promotion wird der Prozess zum Erlangen des akademischen Doktorgrades bezeichnet. Hierbei muss selbstständig ein wissenschaftliches Projekt durchgeführt und die Ergebnisse innerhalb einer Dissertation festgehalten werden. Die Dauer dieses Projekts ist individuell und unterscheidet sich in den jeweiligen Disziplinen (*Bundesbericht Wissenschaftlicher Nachwuchs*, 2021, S.138).

1.1.1 Die Rolle der Promotion

In der Wissenschaft allgemein spielt die Promotion eine tragende Rolle. So stellt sie die Qualifikation für Nachwuchswissenschaftler dar und durch sie wird ein großer Anteil der gesamten Forschungsleistung erbracht (Wissenschaftsrat, 2009, S.7). Ebenso ist sie zwingende Voraussetzung für die Einstellung als Professor*in an einer Hochschule (*Bundesbericht Wissenschaftlicher Nachwuchs*, 2021, S.126).

Es bestehen Grundsätze für die Veröffentlichung von Dissertationen in Deutschland. Diese werden von der Kultusministerkonferenz (KMK) herausgegeben. Hier wird festgehalten, dass die oder der Doktorand*in verpflichtet ist, eine schriftliche Dissertation anzufertigen und die Ergebnisse unentgeltlich der Öffentlichkeit durch Vervielfältigung und Veröffentlichung zugänglich zu machen. Darüber hinaus muss die Verbreitung sichergestellt werden. Dies kann unter anderem durch eine Publikation in einer wissenschaftlichen Fachzeitschrift geschehen oder durch die digitale Bereitstellung auf einem von der Hochschule betriebenen Server (Kultusministerkonferenz, 1997, S.2).

1.1.2 Promotionsrecht

Das institutionelle Promotionsrecht bestimmt, welche Institution Promotionen durchführen darf und wird vom Staat geregelt und verliehen (*Bundesbericht Wissenschaftlicher Nachwuchs*, 2021, S.128). Dies spiegelt sich auf Länderebene in den Landeshochschulgesetzen wider. Hier wird festgelegt, unter welchen Bedingungen welche Institution das Promotionsrecht erhalten kann. Hierbei unterscheiden sich die einzelnen Landeshochschulgesetze in Umfang und den Regelungen der festgelegten Mindeststandards für die Qualifizierung voneinander. Die Institute von Hochschulen mit Promotionsrecht stellen sogenannte Promotionsordnungen aus. Sie formulieren Rahmenbedingungen und dienen primär der Qualitätssicherung. Zur Orientierung für die Institute hat die KMK Empfehlungen zur Qualitätssicherung in Promotionsverfahren veröffentlicht. Die formulierten Leitlinien betonen unter anderem insbe-

sondere die Verantwortung um die Qualität der Promotionen, die Transparenz der Zugangsvoraussetzung für Doktorand*innen, die Verpflichtung zur Betreuung und die Bewertung (*Bundesbericht Wissenschaftlicher Nachwuchs*, 2021, S.127).

Traditionell obliegt es den Universitäten Promotionen durchzuführen. Jedoch ist es seit einiger Zeit möglich, dass vereinzelte forschungsstarke Fachbereiche an Fachhochschulen unter bestimmten Voraussetzungen das Promotionsrecht erhalten können (*Bundesbericht Wissenschaftlicher Nachwuchs*, 2021, S.128). Dem liegt der Wandel der Hochschullandschaft und damit ebenso der des Promotionsverfahrens zugrunde. Primär wurde dieser Wandel durch den Bologna-Prozess angestoßen (Engelfried & Ibisch, 2016, S.10). Dieser kennzeichnet sich durch die gegenseitige Anerkennung von Studienleistungen und Studienabschlüssen innerhalb 45 europäischer Staaten. Um sich als europäischer Staat für die Teilnahme zu qualifizieren, müssen das europäische Kulturabkommen und die gemeinsamen Ziele des Europarats anerkannt werden. Um die gegenseitige Anerkennung zu realisieren, musste die Transparenz und Vergleichbarkeit von Abschlüssen geschaffen werden. Dies geschah durch das gestufte Graduierungssystem des Bachelor- und Masterabschlusses (Kultusminister Konferenz, o.D.). Dadurch konnten Fachhochschulen (FH), beziehungsweise Hochschulen für angewandte Wissenschaften (HAW), Abschlüsse anbieten, die denen der Universitäten gleichbedeutend waren (Engelfried & Ibisch, 2016, S.10).

Weiter ist es möglich eine Promotion durch eine nicht berechtigte Institution in Kooperation mit einer berechtigten Institution durchzuführen. Aus studierender Sicht berechtigt generell seit dem Beschluss der KMK 1999 in allen Bundesländern ein Masterabschluss, unabhängig von der Institution in dem dieser erlangt wurde, zur Promotion. Ferner kann auch mittels Eignungsfeststellungsverfahren ein Bachelorabschluss ausreichend sein (Meurer, 2018, S.6).

1.1.3 Die Rolle der DNB

Wie in Abschnitt 1.1.1 beschrieben besteht eine Veröffentlichungspflicht für Dissertationsschriften. In den jeweiligen Promotionsordnungen der Universitäten wird spezifisch festgelegt, wie diese wahrgenommen werden muss. Da die DNB seit 1913 im Rahmen der Pflichtablieferungsverordnung gesetzlich verpflichtet ist von allen in Deutschland veröffentlichten Medienwerken Kopien zu Sammeln, müssen auch Dissertationen hier abgeliefert werden. Sollte eine Dissertation sowohl online als auch physisch publiziert werden, muss die digitale Form an die DNB abgeliefert werden. Wird nur eine Physische Form publiziert, müssen zwei Exemplare dieser der DNB bereitgestellt werden (Deutsche Nationalbibliothek, o.D. b). Aus diesem Auftrag ging 1998 das Projekt *DissOnline* hervor, indem sich mit der Aufbereitung und Archivierung deutscher Hochschulschriften befasst wurde. Das Projekt *DissOnline* wurde 2012 abgeschlossen und in den Gesamtkatalog der DNB integriert. Dabei werden allerdings bis heute institutionelle Repositorien automatisiert abgefragt und

entsprechende Dissertationsschriften von den Hochschulen bezogen. Mit dem Stand November 2020 verzeichnet der Bestand nach eigenen Angaben über 284.000 Dissertationsschriften (Deutsche Nationalbibliothek, o.D. a).

1.2 Informationswissenschaft

Ziel dieses Kapitels ist nicht die klare Abgrenzung der Informationswissenschaft von anderen Disziplinen oder die exakte Definition der Informationswissenschaft. Vielmehr soll betrachtet werden wie sich die Informationswissenschaft selbst wahrnimmt und wo ihre Aufgabengebiete liegen. Hierzu muss im Folgenden eine Betrachtung des Selbstverständnisses und eine kurze Erläuterung der Entstehung der Disziplin vorgenommen werden.

1.2.1 Die vier Phasen der Informationswissenschaft

Die Entstehung der vergleichsweise jungen Disziplin der Informationswissenschaft ist laut Kuhlen von vier maßgeblichen Phasen geprägt. Dabei beginnt die Entwicklung der Informationswissenschaft mit der ersten Phase Ende des 19. Jahrhunderts. So habe die Informationswissenschaft hier, *"[...] auf die Klassifikationsanstrengungen der Wissenschaft im 19. Jahrhundert [...]"* reagiert (Kuhlen, 2013, S.14) mit dem Ziel wissenschaftliche Publikationen besser verfügbar zu machen, da es durch den wissenschaftlichen Fortschritt zu einer starken Zunahme dieser kam (Kuhlen, 2013, S.14). Das Auswerten, Bereitstellen, Organisieren und Präsentieren von Information und damit auch Wissen bilden hier die Kernaufgaben der Informationswissenschaft. Die zweite Phase verortet sich zwischen den Jahren 1950 und 1974 und wurde geprägt von technologischem Fortschritt. Somit wurde die Informationswissenschaft einem Wandel unterzogen von der rein bibliografischen Erschließung hin zu einer technisch orientierten inhaltlichen Erschließung mit Fokus auf der automatischen Indexierung (Kuhlen, 2013, S.12). Als Auslöser für die stärker ausgeprägte und schnell wachsende Informationsarbeit gilt der sogenannte *Sputnik-Schock* im Jahr 1957. Dieser führte zur Entwicklung erster Online-Datenbanken und damit zu einem Ausbau der Informationsinfrastruktur (Kuhlen, 2013, S.15). Mitte der 70er Jahre bis 2000 verzeichnet sich die dritte Phase. Sie war geprägt durch eine *„System-orientierte technologische Sicht auf Information Retrieval [...] und maschinelle Informationsverarbeitung;“* (Kuhlen, 2013, S.13). Hierbei fand erneut eine Verschiebung des Schwerpunkts statt, die eher praktisch orientiert war. Es wurden vermehrt Methoden aus der Informatik und der Arbeit mit künstlicher Intelligenz genutzt, die zu einer weiterwachsenden technischen Sicht auf die Informationswissenschaft geführt haben (Kuhlen, 2013, S.13). Dem zugrunde lagen die technologischen Fortschritte der Forschung mit Bezug zur künstlichen Intelligenz (Kuhlen, 2013, S.15). Die vierte und bis heute andauernde Phase der Informationswissenschaft begann Ende des 20. Jahrhunderts. Laut Kuhlen

fand in dieser Phase eine Rückbesinnung auf die sozial- und geisteswissenschaftlichen Wurzeln der Disziplin statt. Grund hierfür war die vermehrte Vernetzung von Informationssystemen und die damit einhergehende allgemeine Verfügbarkeit von Informationen. Durch die offene Verfügbarmachung von Wissensbeständen durch das World Wide Web (WWW) ergbit sich ebenfalls ein Schwerpunkt des Information Retrieval, vornehmlich im Kontext von Suchmaschinen (Kuhlen, 2013, S.13).

1.2.2 Definitionen der Informationswissenschaft

Bei der Betrachtung der vier Disziplinphasen fällt die häufige Verschiebung des Hauptfokus innerhalb eines kürzeren Zeitraums auf, was bei einer jungen Disziplin nicht verwunderlich ist. Diese Tatsache führt zu unterschiedlichen Selbstwahrnehmungen des Fachgebietes, zudem führt sie in der Literatur zu Diskussionen darüber, wie eine vollumfängliche Definition der Informationswissenschaft aussehen könnte (Bawden & Robinson, 2012, S.5).

Werden einige unterschiedliche Studiengangsbeschreibungen der Hochschulen betrachtet, die einen Studiengang der Informationswissenschaft anbieten, fällt auf, dass keine Verbindlichen Definitionen der Informationswissenschaft aufgeführt werden und jeweils andere Schwerpunkte gelegt sind. So beschreibt die Universität Regensburg das Aufgabenfeld der Informationswissenschaft als das Befassen mit Informationssystemen sowie unterstützenden Informationsprozessen unter Berücksichtigung aller Aspekte, die das Informationsgeschehen beeinflussen. Dabei greife sie auf Theoriebestände der Teildisziplinen der Sozial- und Geisteswissenschaften sowie Natur und Ingenieurswissenschaften zurück (Baderschneider, 2021). Die Hochschule Darmstadt stellt die technischen Aspekte in den Vordergrund. In der Beschreibung des Studiengangs Information Science wird aufgeführt, dass sich die Aufgabenbereiche aus dem Zusammenführen von Wissen aus großen Datenmengen und der Bereitstellung für Wirtschaft und Gesellschaft bilden (Hochschule Darmstad, o.D.). Bei dem angebotenen Studiengang Informationswissenschaften der Hochschule der Medien in Stuttgart kann zwischen den Schwerpunkten Bibliotheks-, Kultur- und Bildungsmanagement sowie Daten und Informationsmanagement gewählt werden. Hier liegt laut Beschreibung der Fokus auf dem Management von Information (Roos, o.D.).

Wird eine der frühen Definitionen der Informationswissenschaft von Borko aus dem Jahr 1968 betrachtet, beschreibt er diese als: *„Information science ist that discipline that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability.“* (Borko, 1968). Nach seiner Definition befasst sich die Informationswissenschaft als Disziplin kognitiv mit Informationen, jedoch auch mit der Verarbeitung dieser. Saracevic beschreibt wesentliche Merkmale, die kennzeichnend für die Entwicklung und Existenz für die Informationswissenschaft sind. Er stellt fest, dass der Informationswissenschaft in jedem Fall eine interdisziplinäre ver-

ortung innewohnt, die Beziehungen zu den verschiedenen Disziplinen seien jedoch nicht abgeschlossen und befänden sich im Wandel. Weiter stellt er heraus, dass die Informationswissenschaft untrennbar mit der Informationstechnologie verbunden ist und die Entwicklung der Informationswissenschaft immer durch einen technologischen Imperativ erzwungen und beeinflusst wird (Saracevic, 1999).

Aus den vielfältigen Definitionen der Aufgabenbereiche des Fachgebiets geht hervor, dass sich die Informationswissenschaft im Kern mit Informationen und den dahinterstehenden kognitiven und technischen Belangen befasst. Um dies zu bewerkstelligen sind Methoden und Erkenntnisse aus anderen Disziplinen unabdingbar. So beschreibt auch Kuhlen die Informationswissenschaft als multidisziplinär (Kuhlen, 2013, S.12). Dass das Themenspektrum der Informationswissenschaft Methoden und Forschung aus anderen Disziplinen impliziert, geht aus dem Blick auf die beschriebenen vier Phasen eindeutig hervor und bildet ebenso den Konsens in allen aufgeführten Definitionen, Aufgabenbereichen und Merkmalen der Informationswissenschaft. Eine klare Abtrennung der Informationswissenschaft als komplett eigenständige Disziplin ist also nicht möglich. Sie befindet sich stetig im Wandel und ist gebunden an informationstechnologische Fortschritte und gesellschaftliche Bedürfnisse.

1.3 Open Access

Da das automatisierte Sammeln von Metadaten zu Dissertationsschriften voraussetzt, dass diese offen zugänglich zur Verfügung gestellt werden, wird im Folgenden eine Betrachtung der Entstehung sowie der Bedeutung des Begriffs Open Access für die Wissenschaft vorgenommen.

1.3.1 Der Ursprung der Open Access Bewegung

Ihren Ursprung hat die Open Access-Bewegung in den Science-Technology-Medicine (STM) - Fächern. Die angehörigen Wissenschaftler*innen dieser Fächer teilten ihre Erkenntnisse vornehmlich innerhalb eines eher geschlossenen Kreises via E-Mails oder über die akademische Infrastruktur. Im Jahr 1991 gründete Paul Ginsberg vom Los Almos National Laboratory einen Preprintserver für physikalische Forschungsberichte, der seit 1998 unter dem Namen arXiv[1] bekannt ist. Dieser Server und die zugrunde liegende Software boten erstmals die Möglichkeit, Preprints der Physik zentral und für die interessierte Öffentlichkeit frei zugänglich zu archivieren. Das Projekt erfuhr in der wissenschaftlichen Gemeinschaft weltweit durchweg positive Resonanz. Die zentralisierte Archivierung von Preprint Dokumenten verbreitete sich dadurch auch vermehrt auf andere Disziplinen (Ginsparg, 1994) und bildeten somit eine wichtige Grundlage für die Selbstarchivierung. Als Basis für diese Entwicklung gilt die sogenannte Zeitschriftenkrise. Diese entstand durch stetig sinkende Budgets

[1] https://arxiv.org/

der Universitätsbibliotheken und den steigenden Preisen für den Erwerb von wissenschaftlichen Zeitschriften. Dies hat zur Folge, dass Zeitschriften abbestellt werden, wodurch die Verlage versuchten, durch eine weitere Erhöhung der Preise die Verluste auszugleichen (Nicholas et al., 2010). Das Resultat ist ein erschwerter Zugang zu wissenschaftlicher Literatur. Eine Alternative bietet der freie Zugang zu wissenschaftlichen Arbeiten im Sinne des Open-Access Paradigmas.

1.3.2 Definition von Open Access

Die Definition von Open Access bezieht sich laut dem vom Bundesministerium für Bildung und Forschung geförderten open-access.network[2] auf die freie Verfügbarkeit von wissenschaftlichen Dokumenten (open-access.network, 2021c). Der erste formulierte Anspruch an den freien Zugang zu wissenschaftlichen Arbeiten fußt auf dem 2002 veröffentlichten Aufruf der *Budapest Open Access Initiative*. In diesem Aufruf werden ein freier Zugang und die freie Nutzung von wissenschaftlicher Literatur über das Internet gefordert, sollten die Autoren diese Literatur unentgeldlich zur Verfügung stellen (Budapest Open Access Initiative, 2002). Hierdurch wurde der Diskurs entgegen dem traditionellen Publikationssystem maßgeblich geprägt. Das traditionelle System sieht vor, dass der Zugriff auf wissenschaftliche Dokumente gegen eine Gebühr gwährleistet wird, Autorenvergütungen sind jedoch auch hier unüblich. Der Aufruf von 2002 wurde in anderen Erklärungen, wie dem *Bethesda Statement* von 2003 und der *Berliner Erklärung* über offenen Zugang zu wissenschaftlichem Wissen von 2003 weiterentwickelt. In der sogenannten *Berliner Erklärung* wurde der Umfang von Open Access Publikationen erweitert und wie folgt beschrieben:

> „*Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.*“(Max Planck Gesellschaft, 2003)

Zu der erweiterten Beschreibung des in die Definition fallenden Materials, ergänzt die Berliner Erklärung zusätzlich auch die Nutzungsrechte dessen. So dürfen beispielsweise abgeleitete Werke oder Übersetzungen unter Einhaltung einer Zitatpflicht erstellt werden. Aus diesen Entwicklungen entfernte sich die Transformation der Open-Access Bewegung von einer selbstorganisierten Praxis einzelner Fachbereiche hin zu einem weltweit etablierten Paradigma der wissenschaftlichen Praxis.

Bei der Veröffentlichung von wissenschaftlichen Dokumenten innerhalb des Open-Access Rahmens wird primär zwischen dem *Goldenen-* und dem *Grünen Weg* unterschieden. Bei dem *Goldenen Weg* wird die verfasste Arbeit über einen Verlag innerhalb einer Open-Access-Fachzeitschrift herausgegeben. Die Nutzungsrechte werden

[2] https://open-access.network/startseite

hierbei innerhalb eines Publikationsvertrags mit dem Verlag abgestimmt. Bei der Veröffentlichung über einen solchen Verlag fallen sogenannte APCs (Article Processing Charges) oder Veröffentlichungsgebühren an. Diese Gebühren werden von einer/einem oder mehreren beteiligten Autor*innen oder der Institution der Autor*innen entrichtet. Die Arbeiten durchlaufen auf dem Weg zur Veröffentlichung dieselben Qualitätssicherungsprozesse wie bei den traditionellen Closed-Access Publikationen. Der *Grüne Weg*, auch bekannt als *Self-archiving*, versteht sich als Zweitveröffentlichung eines wissenschaftlichen Werks, sprich die Veröffentlichung nach dem Erscheinen in einer Fachzeitschrift. Diese Werke werden als Postprint bezeichnet, da sie den Qualitätssicherungsprozess in Form eines Peer-Review bereits durchlaufen haben und zur Veröffentlichung angenommen wurden. Der *Grüne Weg* eignet sich jedoch auch für Preprints, also Arbeiten die noch keinem Peer-Review unterzogen wurden. Diese Arbeiten werden dann häufig auf privaten Webseiten, disziplinären- oder institutionellen Repositorien der Öffentlichkeit kostenlos zur Verfügung gestellt(open-access.network, 2021a).

1.4 Hochschulschriftenserver

Auf Repositorien werden wissenschaftliche Dokumente digital archiviert und frei zugänglich gemacht. Sie garantieren die langfristige Speicherung von Dokumenten, die bei einer anderen Art der Archivierung, wie beispielsweise der Archivierung auf privaten Servern oder Webseiten, eventuell nicht gegeben ist. Laut der Definition von open-access.network wird generell zwischen disziplinären und institutionellen Repositorien unterschieden. Disziplinäre Repositorien finden institutionsübergreifend Verwendung und stehen Wissenschaftler*innen eines bestimmten Fachgebiets, unabhängig von ihrer institutionellen Anbindung, für die Publikation und Archivierung ihrer Veröffentlichungen zur Verfügung. Institutionelle Repositorien dagegen werden als Dokumentenserver, die von Institutionen betrieben und verwaltet werden, definiert (open-access.network, 2021b). Beide Repositorien-Typen sind in der Nutzung für die Autoren in der Regel kostenlos. Das allgemeine Grundprinzip besteht darin, eine Sammlung elektronischer Dokumente zu verwalten und der Öffentlichkeit zugänglich zu machen. Institutionelle Repositorien bilden einen essentiellen Teil der Informationsinfrastruktur einer Hochschule. Angesichts der freien Verfügbarkeit und Langzeitarchivierung von wissenschaftlichen Dokumenten fallen diese unter die Definition der Berliner Erklärung. Durch den Ansatz der Selbstarchivierung bilden sie ebenso einen Grundpfeiler der Open-Access Bewegung, insbesondere im Hinblick auf den im vorherigen Kapitel beschriebenen *Grünen Weg*.

1.4.1 Begriffsdefinition

Da die Hochschulbibliotheken der meisten größeren Institutionen oft mehrere institutionelle Repositorien betreiben, findet nochmal eine Differenzierung des Repositorien-Typs über dessen Inhalte statt. So werden sogenannte Medienserver für die Bereitstellung von digitalen und multimedialen Inhalten verwendet. Publikationsserver werden hingegen für die Archivierung der Post- und Preprints der Institution genutzt und Hochschulschriftenserver für die Veröffentlichung von Abschlussarbeiten der Studierenden (Dobratz & Müller, 2009). In der Praxis lässt sich diese Differenzierung jedoch oft nicht wiederfinden. So werden die Definitionen vermehrt durcheinandergeworfen und für jede der drei Unterscheidungen für Institutionelle Repositorien gültig gemacht. Oft werden die Begriffe Publikationsserver, Dokumentenserver, eDok-Server oder Archivserver allgemein für Institutionelle Repositorien herangezogen. Dies spiegelt sich auch in den uneinheitlichen Inhalten der verschiedenen Repositorien der Institutionen wider. Da sich diese Arbeit primär mit universitären Qualifikationsarbeiten beschäftigt, wird hier im weiteren Verlauf der Begriff "Hochschulschriftenserver"verwendet. Auch wenn die Inhalte einiger Server gleichzeitig andere Dokumententypen enthalten.

1.4.2 Deutsche Initiative für Netzwerkinformation

Die Deutsche Initiative für Netzwerkinformation (DINI) ermöglicht seit 2004 eine Zertifizierung für Dokumenten- und Publikationsserver. In diesem Zertifikat werden Mindeststandards beschrieben, die bei dem Aufbau eines institutionellen Repositoriums, also auch Hochschulschriftenservern, Orientierung bieten. Als Übergeordnete Ziele des DINI-Zertifikats wird die Verbesserung der Publikationsinfrastruktur für das elektronische Publizieren und die Stärkung von Open-Access-basierten Publikationsformen genannt (Deutsche Initiative für Netzwerkinformation, 2021a). Mit Hilfe dieses Zertifikats wird der Versuch unternommen, die heterogene Landschaft der Open-Access-Publikationsdienste weiter zu vereinheitlichen. So stehen entsprechende Dienstleistungen für Autor*innen und Herausgeber*innen, die langfristige Speicherung und die öffentliche Bereitstellung der Publikationen im Fokus. Weiter stellt die DINI in Zusammenarbeit mit der Bielefeld Academic Search Enginge (BASE) eine Liste von zugänglichen Repositorien zur Verfügung. In dieser wird der Name, der Typ des Publikationsdienstes, das Bundesland, die Plattform, die Anzahl der Dokumente inklusive des OA-Anteils und die Angabe, ob das jeweilige Repositorium ein von der DINI ausgestelltes Zertifikat besitzt, angegeben. Die Aktualität wird jedoch nicht angegeben (Deutsche Initiative für Netzwerkinformation, 2021b).

Software	Häufigkeit
Opus 4	90
Eprints 3	24
DSpace XOAI	18
Opus	14
MyCoRe	13
Qucosa	10
DSpace	10

Tabelle 1 zeigt die 10 am häufigsten vertretenen Softwareprodukte für die Realisierung von Repositorien für Hochschulpublikationen. Diese wurden aus einer von der Deutschen Initiative für Netzwerkinformation e.V. geführten Liste berechnet.

1.4.3 Technische Umsetzung

Technisch werden diese Repositorien meist durch für diesen Zweck entwickelte Softwareprodukte umgesetzt. Bei einer Betrachtung der von der DINI gepflegten Liste von deutschen Publikationsdiensten, unter Anwendung des Filters Hochschulpublikationen, zeichnet sich in Tabelle 1 ein eindeutiges Bild der am häufigsten verwendeten Softwareprodukte ab.

Das Softwareprodukt mit aktuell größter Relevanz in Deutschland ist OPUS und wurde 1998 von der Universität Stuttgart entwickelt. OPUS unterstützt die wichtigsten Schnittstellen für institutionelle Repositorien, wie etwa das *Open Archives Protocol for Metadata Harvesting* (OAI-PMH), auf das im nächsten Kapitel näher eingegangen wird. Ebenso werden die gängigen Standards unterstützt, wie die automatisierte Vergabe eines *Uniform Resource Name* (URN). Die Dokumente können nach Metadaten oder Volltexten durchsucht werden. Außerdem bietet OPUS die erhöhte Sichtbarkeit von Dokumenten, da diese auch über gängige Web-Suchmaschinen gefunden werden können. Es gilt demnach als robustes System, dessen Entwicklung abgeschlossen ist.

Das laut der DINI-Liste zweithäufigst genutzte Produkt, Eprints 3, wurde im Jahr 2000 von der University of Southampton entwickelt. Hier wird ebenfalls eine Schnittstelle für das OAI-PMH bereitgestellt, zudem verfügt das Produkt standardmäßig über fünf verschiedene Metadatenformate in der Ausgabe. Ferner werden, wie bei OPUS, standartisierte Funktionen wie die URN-Vergabe bereitgestellt. Eine Besonderheit bildet sich in der Möglichkeit mehrere Archive durch eine Installation zu

realisieren. Allerdings ist Eprints 3 eher für den englischen Sprachraum optimiert und bedarf für die Verwendung im deutschsprachigen Raum einiger Anpassungen.

Das dritthäufigst genutzte Produkt ist DSpace. Es wurde 2004 von dem Massachusetts Institute of Technology (MIT) in Zusammenarbeit mit Hewlett-Packard (HP) entwickelt. Auch hier werden die gängigen Standards, wie bei den bereits genannten Softwareprodukten, unterstützt. Ein Unterschied zu anderen Anbietern ist der offen zur Verfügung stehende Quellcode, der so an individuelle Bedürfnisse der Institutionen angepasst werden kann. Ebenso wird ein gutes Nutzermanagement bereitgestellt. Laut Dobratz und Müller sei die Technologie zukunftsträchtiger als die anderen genannten Beispiele, da sie auf der Java-Technologie beruhe (Dobratz & Müller, 2009).

1.4.4 Überblick deutscher Repositorien

Um einen umfassenden Blick auf die Breite der deutschen Repositorien zu bekommen, können unterschiedliche Kanäle einbezogen werden. Einen ersten Anhaltspunkt bietet die eingangs erwähnte Liste der DINI. Weiter bietet der Verbund Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV) einen Überblick über europäische Instanzen von OPUS, jedoch sind hier hauptsächlich deutsche Server zu finden (Kooperativer Bibliotheksverbund Berlin-Brandenburg, 2022). Zusätzlich gibt es noch das *Registry of Open Access Repositorys* (ROAR), das von der University of Southampton betrieben wird, von der auch die Software Eprints stammt. Auffallend ist hier, dass der Versuch unternommen wird, einen Überblick über die Aktivität der jeweiligen Server und die Anzahl der darauf liegenden Dokumente zu geben. Dies funktioniert allerdings nur bedingt, da die Angaben der Aktivität mit dem Jahr 2016 enden. Verantwortlich hierfür ist ein Fehler in der Datenbank, an dessen Behebung laut Website gearbeitet werde (Stand 2022). Auf der Plattform ROAR werden Repositorien weltweit gelistet (Registry of Open Access Repositories, 2022). Einen weiteren Anhaltspunkt bietet die Website der Open Archives Initiative. Auf dieser können, ähnlich wie bei ROAR, Repositorien weltweit betrachtet werden. Allerdings ist hier die Registrierung als Data Provider bei der Open Archives Initiative Voraussetzung um ein Repositorium dort in die Liste einzutragen. Für diese Registrierung müssen bestimmte Standards erfüllt sein (Open Archives Initiative, 2001a). Die Hintergründe der Open Archives Initiative werden im folgenden Kapitel näher erläutert.

1.5 Open Archives Initiative Protocol for Metadata Harvesting

Das Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) definiert ein Konzept für das Harvesting von Metadatensätzen aus Repositorien. Es

bietet Datenanbietern eine niedrigschwellige Möglichkeit, ihre Metadaten nach vorgegebenen Standards der Öffentlichkeit zur Verfügung zu stellen und Dienstanbietern die Möglichkeit, diese Daten abzufragen, zu sammeln und zur Verfügung zu stellen. Das OAI-PMH hat sich zu einer wichtigen Grundlage für die Interoperabilität zwischen vernetzten Informationssystemen entwickelt und verdankt seine Akzeptanz größtenteils der Einfachheit und den vergleichsweise geringen Kosten der Implementierung.

1.5.1 Open Archives Initiative

Entwickler des Protokolls ist die *Open Archive Initiative* (OAI), die ihren Ursprung in der E-Print-Community hat. Die Gründung dieser Initiative entstand aus dem wachsenden Bedarf einer Interoperabilitätslösung für den Zugang zu heterogenen Repositorien. Basis dafür ist das Konzept offener Archive. Der Kerngedanke dahinter verfolgt das übergeordnete Ziel den Austausch, die Veröffentlichung und die Archivierung von Metadaten durch interoperable Repositorien und niedrige Zugangsbarrieren zu ermöglichen. Seitdem fördert und entwickelt die OAI einen Interoperabilitätsrahmen und dazugehörige Standards. In der Aufgabenbeschreibung der OAI wird vermerkt: *„The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content.“* (Open Archives Initiative, 2001b).

1.5.2 Funktionsweise des Protokolls

Innerhalb des OAI-PMH-Frameworks partizipieren zwei Gruppen. Zum Einen die Datenanbieter (Data Provider), deren Zuständigkeit in der Verwaltung von Systemen liegt. Sie kümmern sich um die Hinterlegung und Veröffentlichung von Ressourcen in einem Repositorium und nutzen das OAI-PMH zur Bereitstellung von Metadaten. Zum Anderen die Dienstanbieter (Service Provider), die die zur Verfügung gestellten Metadaten von einem oder mehreren Dataprovidern sammeln, um einen oder mehrere Dienste für alle gesammelten Daten anzubieten (Open Archives Initiative, 2001c).

Ein bekannter Serviceprovider ist beispielsweise die BASE[3], die von der Universitätsbibliothek Bielefeld betrieben wird. Sie gilt als eine der weltweit größten Suchmaschinen für wissenschaftliche Dokumente und umfasst einen Index mit mehr als 270 Millionen Dokumenten von über 9.000 Dataprovidern (Stand 2022) (Universitätsbibliothek Bielefeld, 2022).

Das OAI-PMH basiert auf dem *Hypertext Transfer Protocol* (HTTP). Alle Anfragen, Zugangskontrollen, Komprimierungen und Fehlercodes folgen dessen Standards.

[3] https://www.base-search.net/

Anfrageargumente (Requests) werden als GET- oder POST- Parameter formuliert. Die Antworten des OAI-PMH basieren auf der Syntax der Extensible Markup Language (XML). Es wird jedes in XML codierte Metadatenformat von dem Protokoll unterstützt, jedoch wird *Dublin Core* (DC) als für die Interoperabilität grundlegendes Mindestformat vorausgesetzt. So wird sichergestellt, dass Metadaten aus vielen heterogenen Repositorien einheitlich abgefragt und gesammelt werden können. Datenanbietern wird die Möglichkeit geboten, ihre Metadaten in Sets zu organisieren, um Serviceprovidern das selektive Harvesting zu ermöglichen. Als Set wird ein optionales Konstrukt für die Gruppierung von Objekten in einem Repositorium beschrieben. Die Organisation der Sets kann flach oder hierarchisch vorgenommen werden. Bei einer flachen Organisation handelt es sich um eine einfache Liste, wohingegen eine hierarchische Organisation mehrere Hierarchien mit unterschiedlichen und unabhängigen Top-Level-Knoten erlaubt.

Jeder Metadatensatz eines Repositoriums wird in einem sogenannten *Item* zusammengefasst. Ein solches *Item* repräsentiert jeweils ein Dokument in einer Datenbank und kann je nach Spezifikationen des Repositoriums in verschiedenen XML-Formaten abgefragt werden. Jede Abfrage von Items muss unter der Angabe eines XML-Formats stattfinden. Die dadurch erhaltene Antwort wird als Record bezeichnet. Ein Record besteht aus den drei übergeordneten Elementen, *header*, *metadata* und *about* . Dabei enthält das *header* Element den spezifischen Identifier eines jeden Items, dieser ist nur auf der Itemebene innerhalb des OAI-PMH Frameworks gültig, ebenso wird hier ein Datestamp angegeben, der das Erstellungs-, Lösch- oder Änderungsdatum des Items enthält. Weiter findet sich ein oder mehrere *setSpec* Elemente innerhalb des Header Containers. Sie beschreiben die Zugehörigkeiten des Items zu den vom Repositorienbetreiber festgelegten Sets. Das Element *Metadata* beinhaltet das Record selbst im ensprechend angefragten Format oder zumindest, wie oben beschrieben, im DC-Format. Der *about* Container ist optional zu verwenden und kann Informationen über Nutzungsbedingungen oder Herkunft der Metadaten enthalten.

Die Abfrage der Metadaten bei einem Data Provider erfolgt über das sogenannte *Harvesten* durch einen Service Provider. Für Service Provider unterstützt das OAI-PMH sechs Anfragetypen, auch Verben genannt. Diese Verben können durch sechs mögliche Argumente spezifiziert werden. Hierbei wird je nach verwendetem Verb unterschieden, ob ein Argument optional, erforderlich oder exklusiv in die Anfrage aufgenommen werden kann, bzw. muss. Optionale Argumente können in die Anfrage aufgenommen werden, erforderliche Argumente müssen in die Anfrage aufgenommen werden und exklusive Argumente müssen als alleinstehendes Argument mit dem Verb in die Anfrage aufgenommen werden. Jede Anfrage findet über eine URL statt, die sich aus der Basis-URL und der Anfrage zusammensetzt, wie in Abbildung 1 zu sehen ist.

Die Metadaten können hier entweder komplett oder selektiv von dem Datapro-

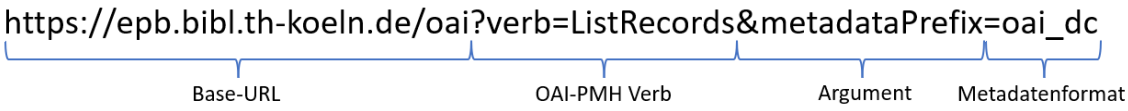

Abbildung 1 zeigt die beispielhafte Zusammensetzung einer URL zur Anfrage an einen Server der das OAI-PMH unterstützt.

vider abgefragt werden. Bei dem selektiven Harvesten gibt es zwei Möglichkeiten: Entweder kann, wie bereits beschrieben, über einzelne Sets selektiert werden oder aber eine Eingrenzung datenbasiert über den Datestamp auf einen bestimmten Zeitraum erfolgen. Dies ist vor allem nützlich, wenn im Vorfeld schon eine Datenabfrage stattgefunden hat und die Dokumente ab dem Standpunkt der letzten Abfrage aktualisiert werden sollen. Die Antworten, also die Summe der abgefragten Records, wird ab einer vom Repositorienbetreiber festgelegten Menge in mehrere Teilmengen partitioniert. Sollten bei einer Anfrage aufgrund der Größe der abgefragten Records nicht alle enthalten sein, gibt das angefragte System einen Resumption Token inklusive einer ID mit jeder inkompletten Antwort aus. Diese ID wird dann vom Service Provider in der nächsten Anfrage angegeben, um die nächste Teilmenge der angefragten Records zu erhalten, bis die Menge vollständig ist (Open Archives Initiative, 2001c).

1.6 Metadaten

Die Definition für den Begriff Metadaten ist komplex, eine allgemeingültige Begriffsbestimmung dafür besteht daher nicht. Metadaten finden in unterschiedlichen Kontexten Anwendung und übernehmen dabei verschiedene Funktionen.

1.6.1 Definition

Kramer beschreibt, dass die grundlegende Definition des Begriffs Metadaten immer auf dem jeweiligen Nutzungsszenario beruhen sollte und es demnach ermöglichen müsse, die spezifischen Definitionen einzelner Fachrichtungen miteinzubeziehen (Kramer, 2019, S.15).

Das baden-württembergische Begleit- und Weiterentwicklungsprojekt für Forschungsdatenmanagement unterscheidet zwischen vier übergeordneten Typen von Metadaten. Zum Einen administrative Metadaten, mit Referenz auf Dateitypen, Standorte, Zugriffsrechte und Lizenzen, die bei der Verwendung und Erhaltung der Daten unterstützen. Zudem Prozessmetadaten, die ausgeführte Aktionen, sowie die dafür beim Datenentstehungsprozess angewandten Methoden und Hilfsmittel beschreiben. Weiterhin inhaltsbezogene bzw. deskriptive Metadaten, die sich in den verschiedenen Disziplinen unterscheiden können und zusätzliche Informationen zu Inhalt und Entstehung beinhalten. Schließlich noch bibliografische Metadaten, die

auf Autoren, Titel und Beschreibungen von Dokumenten Bezug nehmen. (Universität Konstanz, 2022).

Heery beschreibt, dass Metadaten in ihrer Funktion andere Daten referenzieren. Sie müssen sich also immer auf andere Informationen beziehen, die in einer vom Metadatensatz getrennten Form existieren können (Heery, 1996). Zu den Informationen, die durch Metadaten beschrieben werden können, gehören beispielsweise Titel, Autor*innen, Abstract, Identifier, Ort und Zeitraum. Das deckt sich mit der Definition der bibliografischen Metadaten des baden-württembergischen Begleit- und Weiterentwicklungsprojekts für Forschungsdatenmanagement.

Da sich diese Arbeit ausschließlich mit Metadaten und deren dahinterstehenden Ressourcen befasst, die zumindest auch als digitale Objekte vorliegen, wird hier die Definition von Heery, beziehungsweise die der bibliografischen Metadaten genutzt.

1.6.2 Verwendung von Metadaten

Aus Nutzersicht werden die Informationen, die durch Metadaten repräsentiert werden, genutzt um das Finden relevanter Ressourcen zu erleichtern. Gleichzeitig dienen sie zur Organisation und Identifikation elektronischer Ressourcen und zur Archivierung, wie beispielsweise im Fall der Repositorien. Sie bieten ebenso die Möglichkeit der maschinellen Verarbeitung. Um standardisierte Verfahren, die mit Metadaten agieren, entwickeln zu können, existieren Metadatenschemata, die jeweils eine möglichst einheitliche Darstellung zum Ziel haben.

Eine Großzahl wissenschaftlicher Disziplinen verfügen über festgelegte, spezifische Schemata für die Darstellung von Metadaten. Diese legen fest, welche Informationen ein Metadatensatz beinhalten soll und wie diese dargestellt werden. Eine Übersicht der Standards nach Disziplinen geordnet, findet sich auf der Webseite des Digital Curation Centers[4]. Für bibliografische Beschreibungen wird häufig das DataCite[5] Schema genutzt, was ebenfalls Verwendung bei der Registrierung von den sogenannten Dokument Object Identifier (DOI) findet.

1.6.3 Dublin Core

Im Abschnitt 1.5.1 Funktionsweise des Protokolls wurde bereits das Schema Dublin Core (DC) erwähnt. Da es für diese Arbeit eine besondere Relevanz hat, wird es im Folgenden genauer erläutert. Bei DC handelt es sich um einen Standard, der durch die Dublin Core Metadata Initiative 1995 herausgegeben wurde. Ziel dieser Initiative war es, ein Metadatenschema zu entwerfen, das eine allgemein verständliche Semantik beinhaltet. Dieses sollte einen internationalen Geltungsbereich erlangen, um die Interoperabilität zwischen Sammlungen und Indexierungssystemen zu ermöglichen.

[4] https://www.dcc.ac.uk/guidance/standards/metadata
[5] https://schema.datacite.org/

Ebenso standen die Einfachheit und Erweiterbarkeit des Standards im Vordergrund.

Der Dublin Core Standard besteht aus den sogenannten *Core elements*. Hierbei gilt jedes dieser Elemente als optional und wiederholbar, zudem können sie in beliebiger Reihenfolge angegeben werden. Sie lassen sich in die Gruppen *Content*, *Intelectual Property* und *Instantiation* einteilen. Dabei referenziert die Gruppe Content auf die inhaltlichen Eigenheiten eines Dokuments wie Titel, Thema, Beschreibung oder die Herkunft. Die zweite Gruppe Intelectual Property gibt Auskunft über den Herausgeber, die Mitwirkenden und die zugrunde liegenden Rechte eines Objekts. In der letzten Gruppe Instantiation werden Datumswerte, das Format des Dokuments sowie dessen Sprache und die dazugehörigen Identifier angegeben (Weibel et al., 1998). Das Metadatenschema DC kann unter anderen durch eine XML Struktur, die im folgenden näher beschrieben wird dargestellt werden.

1.6.4 Extensible Markup Language

Die Abkürzung XML steht für Extensible Markup Language was so viel bedeutet wie Erweiterbare Auszeichnungssprache, sie wurde 1998 basierend auf der Standard Generalized Markup Language (SGML) von der Arbeitsgruppe *XML Working Group* entwickelt, die von dem World Wide Web Consortium (W3C) organisiert wurde. Die Entwicklung dieser Auszeichnungssprache unterlag mehreren Designzielen, so sollte sie über das Internet nutzbar und mit SGML kompatibel sein. Weiter sollte die Programmatische Verarbeitung von XML-Dokumenten einfach und die Dokumente menschenlesbar sein. XML Entwürfe sollten einfach und schnell erstellt und formal prägnant sein.

In der Praxis zeichnet sich XML durch eine hierarchische Struktur aus, die es ermöglicht dargestellte Daten in Beziehungen zu setzen. So besteht ein XML Dokument immer aus Elementen, die sich ineinander verschachteln lassen. Diese werden durch öffnende und schließende Tags gekennzeichnet. Das auf der höchsten hierarchischen Ebene vorkommende Element wird als *root node* bezeichnet und beinhaltet sämtliche anderen vorkommenden Elemente. Weiter können XML-Elemente beliebig viele Attribute in sich tragen, die erweiternde Informationen über den Inhalt eines Elements spezifizieren (The World Wide Web Consortium, 2013).

1.7 Dewey Dezimalklassifizierung

Als Erfinder und Namensgeber der Dewey Dezimalklassifikation (DDC) gilt der US-amerikanische Bibliothekar Melvil Dewey (1851-1931). Ursprünglich wurde das Klassifikationssystem entwickelt um Bücher bei einer Erweiterung des bibliothekarischen Bestands nicht mehrmals reklassifizieren zu müssen (Chan et al., 2006, S.18). Heute gilt die Klassifizierungsmethode für die Sacherschließung als eine der meistgenutzten Methoden weltweit (Chan et al., 2006, S.25).

Erstmals veröffentlicht wurde das Klassifikationsschema im Jahr 1876. Die Entwicklung wurde von Dewey selbst bis zur 12. Ausgabe begleitet. Mit der 13. Ausgabe erschien die Erste nach seinem Tod. Bis heute wird das System weitergeführt, aktuell in der 2011 veröffentlichten 23. Ausgabe (Heidrun et al., 2018, S.65). Die DDC liegt in einer kurzen Ausgabe vor, die für kleinere Bibliotheken ohne einen schnell wachsenden Bestand gedacht ist, und in einer vollständigen Ausgabe. Seit 1993 werden auch digitale Produkte der DDC entwickelt.

Die Pflege und Entwicklung der DDC obliegt der in Washington, D.C. ansässigen Library of Congress (LoC). Diese arbeitet eng mit dem Rechteinhaber der DDC, dem Online Library Center (OCLC), zusammen und entwickelt die DDC kontinuierlich weiter. Um neue Entwicklungen in der Literatur frühzeitig erkennen zu können, werden seit 1930 neue Notationen durch das in Dewey Editorial Office (DEO) vergeben, in dem Klassifikationsspezialisten aus den verschiedenen Fachrichtungen arbeiten. In den letzten Jahren wurden auf diesem Weg jährlich über 110.000 Notationen für die bibliografischen Bestände der LoC vergeben. Für die Veröffentlichung der verschiedenen Versionen und den darauf entwickelten Produkten ist das OCLC zuständig(Chan et al., 2006, S.22). Die Aktualisierungen zwischen den einzelnen Druckausgaben findet über den Online-Dienst Web-Dewey statt (Chan et al., 2006, S.25).

1.7.1 DDC Übersetzungen

Da das Klassifikationssystem in seinem Ursprung für den angloamerikanischen Raum entwickelt wurde, existieren mittlerweile Übersetzungen für mehr als 30 Sprachen und es findet Anwendung in über 135 Ländern. Seit Oktober 2005 existiert die erste deutsche Druckausgabe(Chan et al., 2006, S.25). Die deutsche Übersetzung wurde in Kooperation von der Deutschen Nationalbibliothek und der Technischen Hochschule Köln (damals Fachhochschule Köln) im Rahmen des Projekts *DDC deutsch* entwickelt. Sie basiert auf der vollständigen 2003 veröffentlichten 22. Version des DDC unter der Anpassung an deutsche Rahmenbedingungen (Heidrun et al., 2018, S.81). So wurden im Zuge des Projekts DDC deutsch zusätzliche Klassen erstellt, bei denen es sich vor allem um Erweiterungen für speziell deutsche und mitteleuropäische Aspekte handelt (Heidrun et al., 2018, S.82). Als deutsches Gegenstück zum englischen WebDewey und zur Verbesserung des Umgangs mit dem deutschen DDC wurde das Klassifizierungstool MelvilClass 2006 von der DNB entwickelt, dieses wurde 2012 von WebDewey Deutsch mit einem verbesserten Funktionsumfang abgelöst(Heidrun et al., 2018, S.84). WebDewey Deutsch bietet durch seine ständige Aktualisierung bis heute immer den aktuellen Stand der deutschen DDC Übersetzung.

1.7.2 Aufbau der Notation

Nicht nur der Erfinder der DDC ist namensgebend für diesen, sondern auch die Art der Klassifizierung. Es handelt es sich bei dem DDC um eine Dezimalklassifikation, da sie aus 10 Hauptklassen besteht, die wiederum jeweils 10 Unterklassen beinhalten, die sich wieder in jeweils 10 Unterklassen aufteilen (u.s.w). Dadurch ermöglicht die dezimale Notation in der hierarchischen Struktur eine erweiterbare Abbildung aller wissenschaftlicher Disziplinen. Hierbei werden die 1.000 Klassen auf den ersten drei Ebenen als DDC-Übersichten bezeichnet und bieten einen Überblick über die Grundstruktur. Grundlegend gilt die Organisation nach Fachgebiet auf der obersten Ebene und dann nach den jeweils dazugehörigen Themen bis hin zum spezifischen. Hierbei ist anzumerken, dass Aussagen, die zu einer Oberklasse gemacht werden, auch immer für die Untergeordneten gültig sein müssen (Chan et al., 2006, S.26). Bei der Notation gelten bestimmte Regeln. So muss jede Notation aus mindestens drei Ziffern bestehen, auf den obersten Ebenen werden zur Auffüllung auf drei Ziffern Nullen verwendet. Nach der dritten Ziffer folgt der Dewey-Punkt, dieser hat keine Bedeutung, sondern dient nur der optischen Gliederung. Weiter endet eine Notation rechts des Dewey-Punktes niemals auf null (Chan et al., 2006, S.27, ff.).

1.8 Methoden der natürlichen Sprachverarbeitung

Das Feld der *Natürlichen Sprachverarbeitung* beschäftigt sich damit, wie natürliche Sprache in Wort oder Schrift von Computern interpretiert und analysiert werden kann. Dabei gilt es als besondere Herausforderung natürliche Sprache, die als unstrukturierte Daten bezeichnet wird, maschinell interpretierbar zu machen. Um dies zu bewerkstelligen finden unterschiedliche Verfahren aus der Informatik und dem maschinellen Lernen Anwendung (Chowdhury, 2003). Die Methoden, die im Praktischen teil dieser Arbeit Anwendung finden, werden zur thematischen Einordnung im folgenden vorgestellt.

Ein grundlegender Schritt bei der Verarbeitung von natürlicher Sprache ist die sogenannte *Tokenization*. Bei dieser wird ein Text in mehrere sogenannte Token aufgeteilt. Dabei Kann sich ein Token aus einem einzelnen Wort, aus Teilworten oder Satzzeichen bilden. Eine der häufig verwendeten Arten, um Tokens aus Texten zu generieren, ist die Verwendung von Leerzeichen als Signal, dass ein Token beginnt und endet. Für den Abstraktionsprozess der maschinellen Verarbeitung bedeutet das, dass der Eingabetext in Untereinheiten aufgeteilt wird (Grefenstette, 1999, S.117). Die so erzeugten Token können verwendet werden um ein Vokabular aus dem zugrundeliegenden Korpus zu erstellen. Mit diesem Vokabular kann beispielsweise die Term Frequenz bestimmt werden. Diese gibt an, wie häufig ein Term innerhalb eines Dokuments auftritt.

Unter dem sogenannten *Part-of-Speech Tagging* versteht man die automatisierte

Ermittlung der Wortart eines Wortes innerhalb eines Satzes. Hierbei wird jedem Wort ein Label mit der entsprechenden Wortart zugeordnet. So kann das Wissen über die Wortart Informationen über dessen Bedeutung innerhalb eines Satzes vermitteln (Charniak, 1997). Beispielsweise kann so das Thema eines Satzes oft schon durch die Identifizierung der vorkommenden Substantive extrahiert werden.

In natürlichen Sprachen existieren in der Regel unterschiedliche grammatikalische Wortformen. Oft kann es zur weiteren Verarbeitung hilfreich sein diese Wortformen auf ihre Grundform zu normalisieren. Hierbei kann zwischen zwei Methoden die für diese Arbeit relevant sind unterschieden werden, dem sogenannten *Stemming* und der *Lemmatisierung*. Ein sogenannter Stemmer ist ein Algorithmus, der den Wortstamm eines Begriffs ermittelt. Innerhalb dieser Arbeit wird für die Verarbeitung englischer Worte der *Porters stemming algorithm* verwendet. Dieser folgt dabei keinem linguistischen Regelwerk zur Erzeugung von Wortstämmen, daher erzeugt er genau genommen auch nicht den Wortstamm eines Begriffs. Die Konzeption des Stemmers geht davon aus, dass die Suffixe in der englischen Sprache aus der Kombination kleinerer und einfacherer Suffixe besteht und folgt dabei einem Regelwerk bestehend aus 60 Regeln zur Abtrennung von Suffixen (Jivani, 2011). Der Prozess der Lemmatisierung hingegen wird als das Zurückführen eines Wortes in dessen lexikalische Form betrachtet.

2 Praktischer Teil

Im praktischen Teil folgt die Erläuterung der Umsetzung des Projekts auf Basis der zugrundeliegenden Theorien, beginnend bei der Abgrenzung der für die Informationswissenschaft relevanten Disziplinen über die Erstellung einer möglichst vollständigen Liste von Hochschulschriftenservern bis hin zu der detaillierten Beschreibung des Softwareentwurfs.

2.1 Abgrenzung des Informationswissenschaftlichen Themenspektrums

Auf Basis der unterschiedlichen Definitionen der Informationswissenschaft und der Verortung des multidisziplinären Selbstverständnisses in Kapitel 1.2 geht hervor, dass Methoden und Themen aus Disziplinen, die sich mit Informationen und deren Verarbeitung befassen, als relevant für die Informationswissenschaft betrachtet werden können. Da im Grunde jede wissenschaftliche Disziplin im weiteren Sinne mit Informationen arbeitet, wird hier betrachtet, welche Disziplinen sich im Kern mit der Verarbeitung von Informationen auseinandersetzen.

Mit Blick auf die in Kapitel 1.2 beschriebene erste Phase der Informationswissenschaft, die sich maßgeblich mit der Erschließung und Klassifikation von Wis-

senschaftlichen Dokumenten befasst hat, lässt sich hier eine Parallele zur Bibliothekswissenschaft ziehen. So beschreibt Umstätter die Bibliothekswissenschaft als wesentlichen Teil der Informationswissenschaft. Sie sei der Teil, der sich mit der publizierten Information der Menschheit befasse. Weiter beschreibt er, dass sich die Bibliothekswissenschaft *"[...] mit den Einrichtungen befasst, die unter archivarischen, ökonomischen und synoptischen Gesichtspunkten publizierte Information für die Benutzer sammelt, ordnet und verfügbar macht."*(Umstätter, 2009). Daraus geht hervor, dass sich die Bibliothekswissenschaft mit ähnlichen bis gleichen Themengebieten befasst wie die Informationswissenschaft, nur eben mit dem Hauptfokus auf Bibliotheken. Einen weiteren Anhaltspunkt für die Beziehung der beiden Disziplinen ist, dass innerhalb der in Kapitel 1.7 erläuterten DDC die Disziplinen unter der Notation *020 - Bibliotheks- und Informationswissenschaften* zusammengefasst werden.

Ebenfalls in Kapitel 1.2 angedeutet finden Verfahren aus der Informatik in der Informationswissenschaft Anwendung. Dass diese eine enge Beziehung zur Information selbst hat, lässt sich schon an dem Begriff erahnen, der sich von Information ableitet. Da das Feld der Informatik sehr groß ist, besteht sie aus mehreren Teilgebieten. Jedoch versteht sie sich im Kern als Disziplin der maschinellen Informationsverarbeitung (Gumm & Sommer, 2012, S.1). Sie ist also auf die Konzeption und Realisierung von Algorithmen, die Daten verarbeiten und übertragen, spezialisiert. Daten repräsentieren hierbei die maschinenlesbare Form von Informationen (Gumm & Sommer, 2012, S.4). Damit bildet sie den technischen Grundpfeiler der heutigen Arbeit und Forschung mit Informationen.

Somit lässt sich festhalten, dass auch mit Hinblick auf die Entstehung der Informationswissenschaft, eine enge Verbindung zwischen den Methoden und Themen der Bibliothekswissenschaft und der Informationswissenschaft besteht. Ebenfalls besteht diese Verbindung zwischen den Erkenntnissen der Datenverarbeitung die durch die Disziplin der Informatik gewonnen werden. Aus dieser Abgrenzung der relevanten Themengebiete ergeben sich für die weitere Selektion der Daten durch die DDC die DDC Notationen: *020 - Bibliotheks- und Informationswissenschaften*, *004 - Informatik*, *005 - Computerprogrammierung, Computerprogramme & Daten* und *006 - Spezielle Computerverfahren.*

2.2 Liste der Hochschulschriftenserver

Da Promotionen wie in Abschnitt 1.1.2 beschrieben mittlerweile auch von vereinzelten Fachhochschulen sowie in Zusammenarbeit mit Institutionen die kein Promotionsrecht haben durchgeführt werden können ist es sinnvoll alle deutschen Hochschulen in die Abfrage nach Metadaten zu Dissertationsschriften einzubeziehen um eine möglichst vollständige Sammlung zu erzeugen. Um diese Hochschulschriftenserver

automatisiert abfragen zu können, wurde eine Liste dieser in Form einer CSV-Datei erstellt. Diese Datei beinhaltet sieben Spalten, in der Spalte *Hochschule* ist der jeweilige Name der Hochschule vermerkt. Die Spalte *OAI_PMH_URL* enthält die identifizierte Adresse der OAI-PMH-Schnittstelle. In der Spalte *OAI_PMH_SET* wurden, wenn vorhanden, die vom Data Provider zur Verfügung gestellten Sets eingetragen, in dem die Dissertationen abgelegt sind. Wenn kein Set für Dissertationen zur Verfügung steht, bleibt das Feld leer. Die Spalte *Land* enthält das zweistellige Kürzel des Bundeslandes der jeweiligen Hochschule. In der Spalte *Form* wird angegeben, ob es sich um eine Universität, Fachhochschule, Pädagogische-Hochschule, Duale-Hochschule oder Fernhochschule handelt. Die Spalte *Promotions-recht* gibt Auskunft darüber, ob eine Hochschule das Recht hat, den akademischen Grad Doktor/Doktorin zu verleihen. Hierbei ist es wichtig, dass diese CSV-Datei manuell erweitert und aktualisiert werden kann, da sich die Schnittstellenadressen oder OAI-PMH-Sets der Hochschulschriftenserver verändern können. Weiter bietet dieser Aufbau die Möglichkeit die Datei an andere Bedingungen anzupassen, so können beispielsweise die Sets verändert werden, wenn ein Harvestingprozess mit einem Schwerpunkt auf anderen Dokumententypen ausgeführt werden soll.

Zur Erstellung einer möglichst vollständigen Sammlung von Hochschulschriftenservern aus dem deutschen Raum wurden insgesamt vier Quellen verwendet. Um zu Beginn einen Überblick über die Landschaft der deutschen Hochschulen zu bekommen, wurde die Tabelle *Liste der Hochschulen in Deutschland*[6] von Wikipedia übernommen. Aufbauend auf dieser Tabelle wurden die einzelnen Hochschulen mit den in Abschnitt 1.4.4 erläuterten Datenbanken abgeglichen und die entsprechenden OAI-PMH-URLs jeweils zugeordnet. Konnte in den Datenbanken kein Verweis zu einer OAI-PMH-Schnittstelle einer Hochschule gefunden werden, wurde versucht, diese händisch zu recherchieren. Im anschließenden Schritt wurde jede Schnittstelle aufgerufen, um zu überprüfen, ob der jeweilige Data Provider ein Set explizit für Dissertationsschriften zur Verfügung stellt. Wenn das der Fall war, wurde das Set jeweils zugeordnet.

Betrachtet man indes die Serverliste, gibt es einige Auffälligkeiten, die es zu untersuchen gilt. Es finden sich insgesamt 38 Hochschulen mit einem konfessionellen Träger in der Liste. Davon teilen sich 17 Bibliotheken dieser Hochschulen aus Deutschland, Österreich und der Schweiz den sogenannten "Kirchlichen Dokumenten Server"[7] (Kidoks). Organisatorisch setzt sich diese Kooperation aus der Arbeitsgemeinschaft Katholisch-Theologischer Bibliotheken (AKThB) und des Verbands kirchlich-wissenschaftlicher Bibliotheken (VkwB) zusammen. Ebenso teilen sich die Hochschulen des privaten Trägers SRH-Holding (ehemals Stiftung Rehabilitation Heidelberg) einen gemeinsamen Hochschulschriftenserver. Die beiden genannten Bei-

[6] https://de.wikipedia.org/wiki/Liste_der_Hochschulen_in_Deutschland#cite_note-4
[7] https://kidoks.bsz-bw.de/home

spiele der kooperativen Server werden in der späteren Analyse jeweils nur als einzelne Server betrachtet.

2.2.1 Wikipedia Tabelle

Wird die Tabelle *Liste der Hochschulen in Deutschland* von Wikipedia betrachtet ergibt sich eine Aufzählung aller deutschen Hochschulen, die auf den Daten der Stiftung zur Förderung der Hochschulrektorenkonferenz[8] (HRK) beruht (Stand 2022). Bei einem Vergleich der Gesamtzahlen der deutschen Hochschulen mit den Gesamtzahlen des Statistischen Bundesamts von 2020/2021 fällt eine Differenz von einer Hochschule auf. Ebenso zeigen sich Unterschiede in der Angabe der unterschiedlichen Hochschultypen. Das Statistische Bundesamt zählt sechs Pädagogische Hochschulen, in der Tabelle der HRK wird keine Differenzierung vorgenommen und die Pädagogischen Hochschulen werden als Universitäten eingeordnet. Auch die Anzahl der einzelnen Hochschultypen unterscheiden sich zwischen den zwei Datenquellen. So gibt das Statistische Bundesamt die Anzahl 108 anstelle von 120 Universitäten an, 16 anstelle von 34 Theologischen Hochschulen beziehungsweise Hochschulen mit konfessionellem Träger, 30 anstelle von 34 Verwaltungshochschulen und 52 anstelle von 57 Kunsthochschulen an (Statistische Bundesamt, 2021).

Die Differenzen der beiden Datenquellen können für das primäre Ziel dieser Arbeit jedoch vernachlässigt werden, da die Hochschulschriftenserver händisch recherchiert wurden und auch ausschließlich jene Hochschulen als relevant für diese Arbeit betrachtet werden, die einen Hochschulschriftenserver zur Verfügung stellen.

Weiter besteht die entnommene Tabelle aus sieben Spalten. Hierbei werden die meisten Spalten, wie in Abschnitt 2.2 beschrieben, übernommen. Die einzige Ausnahme bildet die Spalte *Gründungsjahr*, hier besteht keine Relevanz für diese Arbeit.

2.2.2 Liste der DINI

Die Liste der DINI umfasst insgesamt 474 Repositorien, dabei werden die für diese Arbeit relevanten Repositorientypen in Abschlussarbeiten und Hochschulpublikationen differenziert. Unter der Anwendung des Filters *Abschlussarbeiten* finden sich insgesamt 12 Repositorien. Bei näherer Betrachtung fällt jedoch auf, dass die Hälfte davon auf die Plattform DissOnline der DNB verweist (Stand 2022) und somit keinen eigenständigen Hochschulschriftenserver repräsentiert. Unter Anwendung des Filters *Hochschulpublikationen* finden sich insgesamt 213 Repositorien. Die hier angegebenen Server repräsentieren die verschiedenen Repositorientypen der jeweiligen Hochschulen. So finden sich fachspezifische Server, wie das Repositorium für den Fachbereich Informatik der Universität Hamburg oder auch der Server von EconStor der deutschen Nationalbibliothek für Wirtschaftswissenschaften. Dies verdeutlicht

[8] https://www.hochschulkompass.de/service/impressum.html

abermals die in Abschnitt 1.4.1 erwähnte Problematik der dokumentspezifischen Bezeichnung der Repositorien. Weiter finden sich einige Eigenheiten in den angegebenen Repositorien. So sind teilweise veraltete Uniform Resource Locators (URL) hinterlegt, wie am Beispiel der angegebenen OAI-PMH-URL Universität Bamberg[9] zu sehen ist. In diesen Fällen konnte nur über eine Manipulation der Basis-URL auf die aktuelle OAI-Schnittstelle zugegriffen werden. Ebenso werden hier teilweise Server gelistet, die als OAI-Schnittstellen-URL eine Verlinkung auf die Sammlung DissOnline der DNB verweisen(Stand 2022)(Deutsche Initiative für Netzwerkinformation, 2021b).

2.2.3 Liste des KOBV

Die Liste des KOBV enthält insgesamt 116 Repositorien. Auch hier handelt es sich nicht ausschließlich um Hochschulschriftenserver. Weiterhin finden sich bei der KOBV vereinzelt Server, die sich aktuell noch im Aufbau befinden und somit nicht in die finale Liste mit aufgenommen werden konnten. Hier werden noch keine vorläufigen URLs zur Verfügung gestellt (Stand 2022). Zudem finden sich hier OPUS-Instanzen, die über keine offene OAI-PMH Schnittstelle verfügen, wie am Beispiel der Hochschulschriftenserver der École supérieure de commerce de Paris Europe (ESCP) und der Hochschule für bildende Künste Braunschweig zu sehen ist (Kooperativer Bibliotheksverbund Berlin-Brandenburg, 2022).

2.2.4 Registry of Open Access Repositories

Das Registry of Open Access Repositories listet unterschiedliche Repositorien weltweit. Unter Anwendung des Filters *Germany* finden sich insgesamt 261 Repositorien. Wendet man daraufhin den Filter *e-Theses* an, verbleiben 19 Hochschulschriftenserver. Ein weiterer Anlaufpunkt ist der Filter *Research Institutional or Departmental*, hier werden insgesamt 182 Repositorien angezeigt, die zum Teil der Filterbezeichnung entsprechen, jedoch finden sich auch hier Hochschulschriftenserver. Ebenso finden sich Veraltete Verlinkungen zu den Repositorien wie am Beispiel des Repositoriums der Technischen Universität Braunschweig zu sehen (Stand 2022) (Registry of Open Access Repositories, 2022).

2.3 Heterogenität der Daten

Durch die Abfrage einer Vielzahl verschiedener Data Provider ergibt sich eine Heterogenität der erhaltenen Daten. Das spiegelt sich nicht nur in der Vollständigkeit der Metadaten einzelner Records wider, sondern auch in der in Kapitel 1.6.3 beschriebenen freien und optionalen Nutzung der DC-Elemente.

[9] https://fis.uni-bamberg.de/

2.3.1 Datenstruktur der DNB Datenbank

Die erste Differenzierung kann zwischen Hochschulschriftenservern und der Datenbank der DNB vorgenommen werden. Die Nutzung der Elemente innerhalb der Metadaten-Records der DNB ist einheitlich, unterscheidet sich allerdings teilweise von den Hochschulschriftenservern. So wird das DC-Element *dc:type* bei den meisten Institutionen verwendet um den Typ des Dokuments zu beschreiben, also beispielsweise Dissertation, Bachelor- oder Masterthesis. Die Datenbank der DNB hingegen nutzt dieses Feld für die Angabe ob es sich um ein online verfügbares Dokument handelt und übermittelt in diesem Element nur die Information *Online-Ressource*. Weiter wird das DC-Element *dc:description* von allen Hochschulschriftenservern verwendet um eine Kurzbeschreibung des Inhalts eines Dokuments abzulegen. Die DNB hingegen nutzt dieses Element um den Dokumententyp, die Stadt der Hochschule, die Hochschule sowie das Jahr der Veröffentlichung zu spezifizieren. Ebenso werden bei der DNB als einzigem Data Provider innerhalb der *dc:creator* und *dc:title* Elemente teilweise zusätzlich zu den Autor*innen auch die Gutachter*innen angegeben wie in Abbildung 2 zu sehen.

```
<dc:title>Cross-lingual question answering / Bogdan Eugen Sacaleanu. Betreuer: Hans Uszkoreit</dc:title>
  <dc:creator>Sacaleanu, Bogdan [Verfasser]</dc:creator>
  <dc:creator>Uszkoreit, Hans [Akademischer Betreuer]</dc:creator>
  <dc:publisher>Saarbrücken : Saarländische Universitäts- und Landesbibliothek</dc:publisher>
  <dc:date>2012</dc:date>
  <dc:language>eng</dc:language>
  <dc:identifier xsi:type="tel:URN">urn:nbn:de:bsz:291-scidok-47820</dc:identifier>
  <dc:identifier xsi:type="tel:URL">http://nbn-resolving.de/urn:nbn:de:bsz:291-scidok-47820</dc:identifier>
  <dc:identifier xsi:type="tel:URL">http://d-nb.info/1052292585/34</dc:identifier>
  <dc:identifier xsi:type="tel:URL">http://scidok.sulb.uni-saarland.de/volltexte/2012/4782/</dc:identifier>
  <dc:identifier xsi:type="dnb:IDN">1052292585</dc:identifier>
  <dc:subject>004 Informatik</dc:subject>
  <dc:description>Saarbrücken, Universität des Saarlandes, Diss., 2011</dc:description>
  <dc:rights>lizenzfrei</dc:rights>
  <dc:type>Online-Ressource</dc:type>
  <dc:relation>http://d-nb.info/1023235293</dc:relation>
```

Abbildung 2 zeigt ein Beispiel der XML-Datenstruktur und Verwendung der DC-Elemente bei Records der Datenbank der DNB.

2.3.2 Datenstrukturen der Hochschulschriftenserver

Die Nutzung der DC-Elemente und der darin abgelegten Spezifikationen sind zwar bei den einzelnen Hochschulschriftenservern in der Regel einheitlich, unterscheiden sich jedoch zwischen diesen teilweise stark voneinander. Ebenso unterscheidet sich die Anzahl der einzelnen Elemente. Bei der Universität des Saarlandes beispielsweise existieren in manchen Records mehrere *dc:date* Elemente, bei denen eines genutzt wird um das Publikationsdatum unter Angabe von Jahr-Monat-Tag abzubilden. Ein weiteres besteht um nur das Jahr zu spezifizieren und ein drittes um ein mögliches Ablaufdatum des Embargos zu bestimmen. Ebenfalls existieren hier teilweise auch Dopplungen des Elements, wie in Abbildung 3 zu sehen ist. Weiter finden sich in dem DC-Element *dc:type* unterschiedliche Angaben für den Dokumententyp der Dis-

sertationsschrift. Hier reicht das Spektrum von *Dissertation* zu *Doctoral Thesis* und *DoctoralThesis* bis hin zu *info:eu-repo/semantics/doctoralthesis*. Bei der Universität Regensburg werden Dissertationen als *Hochschulschrift* oder *Hochschulschrift der Universität Regensburg* abgelegt.

```
<dc:date>embargoEnd/2023-09-09</dc:date>
<dc:date>2021-09-27T12:32:32Z</dc:date>
<dc:date>2021-09-27T12:32:32Z</dc:date>
<dc:date>2021</dc:date>
<dc:type>doc-type:doctoralThesis</dc:type>
<dc:identifier>urn:nbn:de:bsz:291--ds-346987</dc:identifier>
<dc:identifier>hdl:20.500.11880/31804</dc:identifier>
<dc:identifier>http://dx.doi.org/10.22028/D291-34698</dc:identifier>
<dc:language>eng</dc:language>
<dc:rights>embargoedAccess</dc:rights>
<dc:rights>Alle Ressourcen in diesem Repository sind urheberrechtlich
    geschützt</dc:rights>
<dc:publisher>Saarländische Universitäts- und Landesbibliothek</dc:publisher>
```

Abbildung 3 zeigt ein Beispiel der XML-Datenstruktur und Verwendung der DC-Elemente bei Records der Datenbank der DNB.

Da die Daten in dieser heterogenen Struktur nicht oder nur sehr schwer analysierbar sind und auch bei der späteren Verwendung als ungeeignet betrachtet werden können, erfordert dies eine Homogenisierung der Attribute.

2.4 Architektur des Softwareentwurfs

Im folgenden Kapitel wird die grundlegende Architektur des Softwareentwurfs vorgestellt. Die Architektur legt fest, wie das System aufgebaut ist und wie die einzelnen Komponenten ineinandergreifen. Der Entwurf wurde modular aufgebaut, um die Übersicht zu verbessern und Anpassbarkeit im Falle einer Wiederverwendung zu vereinfachen. So können einzelne Komponenten unabhängig voneinander an spezifische Anforderungen angepasst und getestet werden. Der Entwurf wurde mittels eines Testsets, bestehend aus 13 relevanten Hochschulschriftenservern von Institutionen, die einen Studiengang im Bereich der Informationswissenschaft anbieten, sowie aus der Datenbank der DNB erstellt um ihn dann auf alle ermittelten Hochschulschriftenserver anzuwenden.

Das Flow-Chart des kompletten Programmablaufs ist in Abbildung 4 zu sehen. In erster Instanz erfolgt eine automatisierte Abfrage von Metadaten mittels des OAI-PMH. Bei dieser Abfrage werden nach Verfügbarkeit nur OAI-PMH-Sets bezogen, die explizit Dissertationsschriften enthalten. Da die Metadaten der unterschiedlichen Dataprovidern eine heterogene Struktur aufweisen, müssen diese im Anschluss homogenisiert und teilweise bereinigt werden. Vereinzelt sind die via OAI-PMH zur Verfügung gestellten Metadaten zudem unvollständig. Die entsprechend fehlenden Informationen finden sich jedoch teilweise auf den Webseiten der Institutionen und

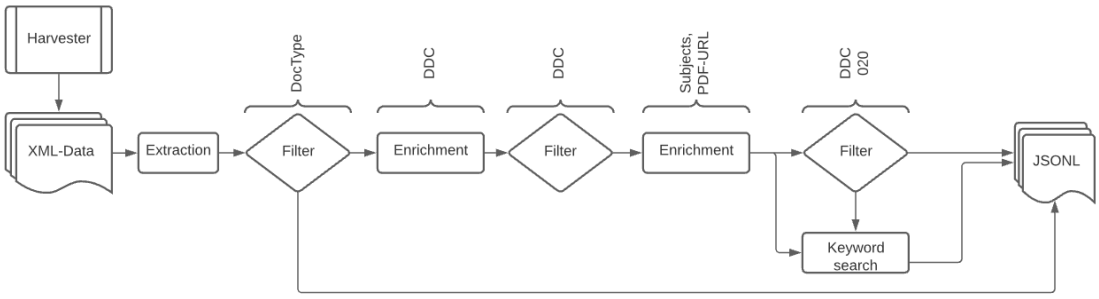

Abbildung 4 zeigt das Flusslaufdiagramm des des Softwareentwurfs.

können dadurch ergänzt werden. Da eine Abfrage aller Metadaten-Records stattfindet, wenn kein Set vom jeweiligen Data Provider zur Verfügung gestellt wurde, ist es notwendig, die erhaltenen Daten nach dem entsprechenden Dokumententypen zu selektieren. Ebenfalls muss eine Selektion nach den wissenschaftlichen Disziplinen erfolgen, um möglichst nur Metadaten von relevanten Dissertationsschriften zu identifizieren. Zusätzlich zur reinen Abgrenzung der Disziplinen findet noch eine Schlagwortsuche in den Metadaten-Records der Disziplinen statt, die in Kapitel 2.1 als der Informationswissenschaft nahestehend ermittelt wurden. Abschließend können die verarbeiteten Daten als JSONlines Datei extrahiert werden. Darüber hinaus werden zwischen den jeweiligen Prozessschritten Sicherungsdateien angelegt, die im Falle eines System- oder Programmabsturzes sicherstellen, dass die bisherigen Verarbeitungsschritte nicht verloren gehen.

2.4.1 Harvester

An erster Stelle des gesamten Prozesses steht das Harvesting-Modul. Das Modul ist für den gesamten Prozess der Anfragen an die Data-Provider und der Speicherung der erhaltenen Antworten zuständig. Es werden insgesamt zwei Typen von Dataprovidern angefragt, die identifizierten Hochschulschriftenserver und die Datenbank der DNB.

Direkt zu Beginn liest das Modul die in Kapitel 2.2 beschriebene CSV-Datei der Hochschulschriftenserver ein und überprüft, ob eine OAI-PMH-URL und ein Set zu der jeweiligen Hochschule zur Verfügung gestellt werden. Diese Informationen werden dann an die nächste Funktion weitergegeben und es wird überprüft, ob es sich bei dem Data Provider um einen Hochschulschriftenserver oder die Datenbank der DNB handelt. Da eine Anfrage an den kompletten Inhalt eines Sets der DNB in einem Fehler mündet, wie in Abbildung 5 zu sehen ist, muss die Methode des selektiven harvestings nach Datumswerten erfolgen.

So werden in der ersten Anfrage nur Items mit einem Datestamp, der bis zum 01.01.2020 reicht, angefragt. Die zweite Anfrage komplettiert das Set mit Items, die

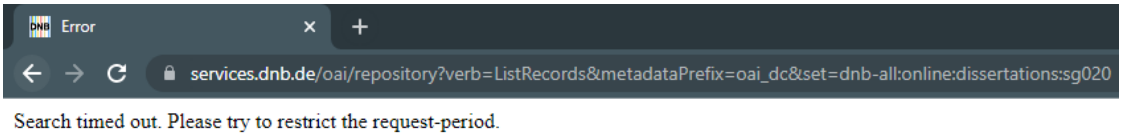

Abbildung 5 zeigt ein Beispiel der Fehlermeldung bei der Anfrage an den Server der DNB die nicht zeitlich begrenzt ist.

ab dem 01.01.2020 datiert sind. Abgesehen von dieser Eigenheit gleichen sich die Anfragen an die beiden Data Provider Typen. So wird immer eine Anfrage unter Verwendung des OAI-PMH-Verbs *ListRecords* gestellt mit dem angehängten Argument *metadataPrefix=oai_ dc*, das das angefragte Metadatenschema definiert. Sollte ein Set zu dem jeweiligen Data Provider zur Verfügung stehen, wird dieses unter Angabe des Arguments *set* angefragt. Diese Anfragen resultieren in XML codierten Antworten der Data Provider, die die entsprechenden Records in dem angefragten Metadatenschema beinhalten. Wie in Kapitel 1.5.2 beschrieben, wird die Antwort bei einer zu großen Datenmenge in Teilmengen aufgeteilt. In der Praxis hat sich gezeigt, dass die Antworten meist zwischen 100 und 500 Records enthalten. Um zu überprüfen, ob die Menge vollständig ist, parst das Modul die jeweils letzte Antwort und sucht nach dem XML-Tag *<resumptionToken>*. Ist dieser Tag vorhanden, signalisiert dies dem Modul, dass die Abfrage noch nicht abgeschlossen ist. In diesem Fall entnimmt das Modul die ID des Tags und stellt im Anschluss eine Wiederaufnahmeanfrage, bis die Gesamtmenge vollständig ist. Im letzten Schritt werden die erhaltenen Antworten lokal unter Angabe der Institution, des Sets und der Nummerierung im Dateinamen abgelegt.

2.4.2 Datenextraktion

Die im Harvesting-Prozess erzeugten Dateien müssen für die weitere Verarbeitung in eine von Python verarbeitbare Datenstruktur überführt und homogenisiert werden, diese Aufgabe übernimmt das Extraktionsmodul.

Im ersten Schritt parst das Modul die erzeugten XML-Dateien und liest die Inhalte der DC-Elemente aus. Die so extrahierten Attribute der einzelnen Records werden als Values in einem sogenannten nested Dictionary abgelegt. Jeweils ein Metadaten Record wird als Dictionary repräsentiert. Hierbei bilden die extrahierten Attribute die Values des Dictionarys. Diese werden wiederum in einem übergeordneten Dictionary als dessen Values abgelegt, die Keys des übergeordneten Dictionarys bestehen aus einer bei der Iteration vergebenen Nummerierung. Der Vorteil dieser komplexen Datenstruktur liegt in den Key-Value-Paaren. So kann durch die Nummerierung des übergeordneten Dictionarys auf die darunterliegenden einzelnen Records zugegriffen werden. Zudem kann durch die entsprechenden Keys auf deren einzelne Attribute, die die bibliometrischen Informationen enthalten können, zugegriffen werden. Wie

sich die extrahierten Attribute bei der Homogenisierung aus den DC-Elementen ergeben, ist im Detail in Abbildung 6 zu sehen. Hierbei können die Elemente *creator*, *type*, *rights*, *lang* und *format* ohne weitere Verarbeitungsschritte übernommen werden.

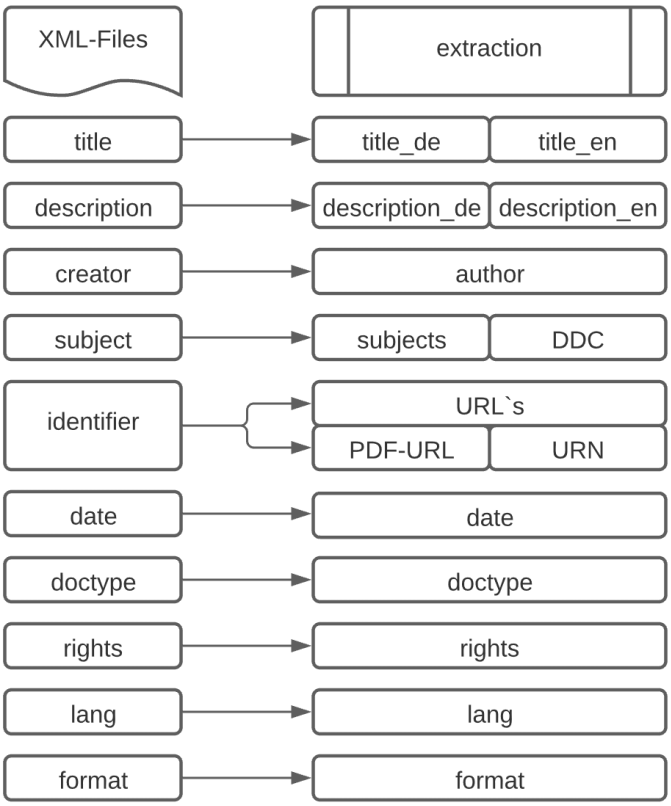

Abbildung 6 zeigt wie die Elemente der erhaltenen XML Dateien in die Datenstruktur des Softwareentwurfs übernommen werden.

Teilweise beinhalten die Records Titel und Kurzbeschreibung in deutscher und englischer Sprache, diese werden jeweils in einem DC-Element repräsentiert. In manchen Fällen wird unter Angabe des XML-Attributs *xml:lang* die entsprechende Sprache angegeben, jedoch stimmt diese Angabe nicht in allen Fällen mit der tatsächlichen Sprache überein. Wenn das XML-Attribut *xml:lang* nicht vorhanden ist, wird hier eine eigenständige Klassifikation vorgenommen. Mit Hilfe der Python Bibliothek *Langdetext* wird die Sprache der Titel oder Kurzbeschreibungen ermittelt. Hierbei wird zwischen den Sprachen deutsch und englisch differenziert und da *Langdetect* nicht immer die korrekte Sprache erkennt, werden die Titel oder Beschreibungen im Zweifel als der englischen Sprache entsprechend abgelegt um keine der Inhalte zu verlieren. Die so ermittelten Titel oder Kurzbeschreibungen werden entsprechend ihrer Sprache abgelegt.

Innerhalb des DC-Elements *subject* befinden sich größtenteils zweierlei relevante Informationen. So werden dort, dem Dokument entsprechend, themenspezifische Stichworte und ein oder mehrere DDC Notationen abgelegt. Die Stichworte werden,

wie in Abbildung 6 gezeigt, als Attribute der *subjects* übernommen. Die DDC-Codes werden mit Hilfe eines regulären Ausdrucks von etwaigen, nicht numerischen Strings bereinigt und als Value des Key *ddc* übernommen. Das DC-Element *identifier* kann drei relevante Informationen enthalten. Hier wird differenziert zwischen URN, PDF-URL und weiteren URLs. Auch diese Informationen werden mittels regulärer Ausdrücke identifiziert und getrennt voneinander in den jeweils entsprechenden Keys als Values übernommen.

Da die gesamte Datenextraktion unter Berücksichtigung aller erhaltenen Antworten von allen Dataprovidern stattfindet, können bei der Extraktion Redundanzen in Form von Dubletten entstehen. Um diese im letzten Schritt vor der Speicherung des Dictionary zu entfernen, wird in erster Instanz eine Prüfung doppelt vorkommender URNs vorgenommen. Sollte keine URN zur Verfügung stehen, werden die deutschen sowie die englischen Titel zur Überprüfung verwendet und die redundanten Metadatensets entfernt.

2.4.3 Datenanreicherung

Da wie in Kapitel 2.3 beschrieben nicht immer alle Attribute der erhaltenen Records vollständig vorliegen, nimmt das Modul der Datenanreicherung eine Komplettierung mittels Web Scraping-Verfahren vor. Da eine Anpassung an die individuellen Spezifikationen der unterschiedlichen Webseiten mit sehr hohem Aufwand verbunden wäre, beschränkt sich dieser Prozess auf die besonders relevanten Attribute *ddc*, *pdf_url* und *subjects*. Hierbei gilt der DDC-Code als besonders relevant, da er zur späteren Selektion nach wissenschaftlichen Disziplinen benötigt wird. Die URL, die auf das vollständige Dokument in Form einer PDF-Datei verweist, kann später verwendet werden, um die Dokumente im Volltext zu beziehen. Die Schlagworte werden extrahiert um eine weitere themenspezifische Einordnung der Metadatensets zu erhalten.

Im ersten Schritt der Datenanreicherung überprüft das Modul die eingangs erwähnten Attribute und identifiziert fehlende Einträge. Um die Anreicherung dieser vorzunehmen, löst das Modul die in den Records hinterlegte URN auf. Dazu wird sie an die Base-URL des *URN:NBN Resolver für Deutschland und Schweiz*[10] angehängt, um eine URL zu erzeugen, die auf die Webpräsentz des jeweiligen Dokuments verweist. Für den Fall, dass keine URN zur Verfügung stehen sollte, entnimmt das Modul die in dem Key urls hinterlegten URLs und versucht über diese eine Anfrage zu erzeugen. Die Anfragen an die Webserver werden mit der Python Bibliothek *Requests*[11] realisiert, um diese im Anschluss mit der Python Bibliothek *Beautiful Soup*[12] zu parsen. Da die Webpräsentz der Dokumente, die auf den Hochschulschriftenservern liegen, meist von der darunter liegenden Softwareprodukten realisiert

[10] https://nbn-resolving.org/
[11] https://docs.python-requests.org/en/latest/
[12] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

werden, gleicht sich deren Struktur in einigen Fällen. So liegt oft eine HTML-Tabelle vor, deren HTML-Tags und Inhalte mithilfe von Regulären Ausdrücken identifiziert werden können. Auf diese Weise kann eine einheitliche Abfragemethode für einen Teil der Webrepräsentationen der Metadaten vorgenommen und diese in den Metadatensätzen ergänzt werden.

2.4.4 Datenfilterung

In Kapitel 2.2 wurde beschrieben, dass nicht alle Data Provider OAI-PMH-Sets für Dissertationsschriften zur Verfügung stellen. Das hat zur Folge, dass die erhaltenen unterschiedlichen Dokumententypen nach Dissertationsschriften gefiltert werden müssen. Ebenso muss eine Selektion nach wissenschaftlichen Disziplinen erfolgen. Diese Aufgaben übernimmt das Modul der Datenfilterung.

Für die Filterung des Dokumententyps findet im ersten Schritt eine textuelle Normalisierung der Bezeichnungen für Dissertationsschriften statt. Alle in Kapitel 2.3.2 aufgeführten Synonyme, die in den Records vorhanden sind, werden unter dem am häufigsten verwandten Begriff *doctoralThesis* vereinheitlicht. Dadurch kann die Selektion der relevanten Records mit entsprechendem Dokumententypen durch ein einfaches String-Matching erfolgen. Das Selektionsverfahren der wissenschaftlichen Disziplinen erfolgt ebenfalls über ein String-Matching. Hierbei wird in erster Instanz überprüft, ob das jeweilige Record mindestens einen DDC-Code enthält. Wenn das zutrifft, wird im nächsten Schritt überprüft, ob einer der enthaltenen DDC-Codes den in Kapitel 2.1 als relevanten identifizierten Disziplinen entspricht. Wenn auch das zutrifft, werden die entsprechenden Records übernommen.

2.4.5 Schlagwortsuche

Die Sammlung von Metadaten-Records, die durch die vorherig erläuterten Module erzeugt wurde, beinhaltet Metadaten zu Dissertationsschriften, die explizit den Bibliotheks- und Informationswissenschaften zugeordnet wurden, und Metadaten zu Dissertationen aus den in Kapitel 1.2 als relevant identifizierten nahestehenden Wissensgebieten. Da nicht jede Publikation der nahestehenden Fächer als relevant für die Informationswissenschaft eingestuft werden kann, wird versucht mithilfe der im folgenden vorgestellten Methode eine relevante engere Auswahl zu erstellen.

Hierbei werden Schlagworte aus den deutschen und englischen Titeln der Records, die mindestens mit der DDC Notation *020 - Bibliotheks- und Informationswissenschaften* ausgezeichnet wurden extrahiert. Dazu werden die Titel einem der Sprache entsprechenden Preprocessing unterzogen. Hierbei werden im ersten Schritt alle Sonderzeichen und die Stoppworte entprechend der Sprache entfernt, dies geschieht unter Anwendung der durch die Python Bibliothek *NLTK*[13] zur verfügung

[13] https://www.nltk.org/

gestellten Stoppwortlisten. Ebenso wird die Groß- und Kleinschreibung der einzelnen Wörter zu ausschließlich Kleinbuchstaben vereinheitlicht. Daraufhin werden die Titel tokenisiert und entsprechend ihrer Sprache weiterverarbeitet. Aus den deutschen Titeln werden mittels des *The Hanover Tagger*[14] die Substantive der Zeichenketten extrahiert, da diesen innerhalb der Titel die größte Aussagekraft über den spezifischen Inhalt des Dokuments innewohnt. Um auch durch unterschiedliche Wortformen einzelner Substantive später Treffer bei dem Matching-Verfahren erzeugen zu können, werden diese durch eine Lemmatisierung auf ihre Grundform reduziert. Aus den englischen Titeln hingegen werden nach der Tokenisierung durch die Python Bibliothek *NLTK* die Nomen extrahiert, diese werden unter Anwendung des *Porter Stemming algorithm* gestemmt und somit auf ihren Wortstamm reduziert. Da die auf diese Weise erzeugten Tokens noch keine hohe Aussagekraft speziell für die Disziplin beinhalten, wird deren Relevanz mithilfe der Termfrequenz bestimmt unter der Annahme, dass Begriffe, die häufig in Titeln relevanter Dokumente verwendet werden, eine besondere Aussagekraft für die Disziplin beinhalten.

Um möglichst nur spezifische Terme zu extrahieren, werden die fünf am häufigsten vorkommenden Terme verwendet. Da sich in diesen ebenfalls sehr generelle Begriffe befinden die in unterschiedlichsten Kontexten verwendet werden, wurden sie zusätzlich manuell durch eine sogenannte Blacklist eingegrenzt. Diese Liste beinhaltet Begriffe, die nicht in die Sammlung von Termen mit aufgenommen werden sollen. Um mit der erzeugten Liste in den Titeln der Metadaten aus nahestehenden Disziplinen ein Matching-Verfahren vorzunehmen, wird das erläuterte Preprocessing verfahren ebenfalls auf die Titel angewendet in denen mittels der extrahierten Schlagworte gesucht wird. Wird ein Term in einem Titel gefunden, wird dieser der Treffermenge hinzugefügt.

2.4.6 Datenausgabe

Die verarbeiteten Daten können als JSON Lines Datei exportiert werden. Hierbei werden die Daten separiert abgelegt, so besteht eine Sammlung ausschließlich aus Metadaten zu Dissertationsschriften die mindestens mit der DDC Notation *020 - Bibliotheks- und Informationswissenschaften* klassifiziert wurden, eine Sammlung, die aus der Schlagwortsuche innerhalb der nahestehenden Disziplinen resultiert und eine komplette Sammlung aller extrahierter Dissertationen. Weiter werden zwischen den Verarbeitungsschritten durch die jeweiligen Module des Softwareentwurfs Pickel Dateien erstellt. Von diesen enthält eines die unverarbeiteten Metadatensets, ein weiteres beinhaltet die verbleibenden Records nach der Filterung des Dokumententyps und der DDC Notation und eines umfasst die komplett gefilterten und angereicherten Daten.

[14] https://github.com/wartaal/HanTa

3 Evaluation des Softwareentwurfs

Im Folgenden wird der Softwareentwurf hinsichtlich der erhaltenen Metadaten und anhand eines Testsets evaluiert, dabei findet auch eine deskriptive Analyse der erhaltenen Daten Statt. Ebenso wird überprüft, wie effektiv das Anreichern der Attribute durch das entsprechende Modul und die Schlagwortsuche in den Records der nahestehenden Disziplinen funktioniert hat. Dabei ist zu beachten, dass die Auswertung der Daten auf einer Abfrage der Data Provider vom 04.02.2022 beruht und sich bei einer erneuten Abfrage verändern kann.

3.1 Metadaten Harvesting und Extraktion

Durch den in Kapitel 2.4.1 beschriebenen Harvesting Prozess wurden insgesamt 9.866 XML Dateien erzeugt. Aus diesen konnten nach Entfernung der Duplikate 252.190 einzigartige Metadatensets unterschiedlicher Dokumententypen von 94 verschiedenen Dataprovidern extrahiert werden.

Durch die Selektion nach dem für diese Arbeit relevanten Dokumententypen der Dissertationsschrift ergibt sich eine Gesamtmenge von 143.073 Metadatensets, referenzierend auf Dissertationsschriften, die von 73 unterschiedlichen Dataprovidern stammen. Ausgehend von dieser Gesamtmenge wurden davon 139.276 Metadatensets von deutschen Hochschulschriftenservern bezogen und 3.797 von der Datenbank der DNB. Abbildung 7 zeigt die detaillierte Verteilung der bezogenen Metadaten nach entsprechenden Dataprovidern. Hierbei liegt der gerundete Mittelwert der extrahierten Metadatensets pro Data Provider bei 1.960 Records. Der größte Anteil stammt von der Freien Universität Berlin.

Die Vollständigkeit der extrahierten Attribute ist in Tabelle 3 im Detail abgebildet. Das Attribut *doc_type* wird, wie in Kapitel 2.4.4 beschrieben, verwendet um die Dokumente zu selektieren. Das Attribut *source* hingegen wird während der Verarbeitung der Daten automatisiert vergeben, daher finden sich diese vollständig wider. Da Metadatensets sowohl einen englischen als auch einen deutschen Titel enthalten können, treten bei den absoluten Häufigkeiten Überschneidungen auf. Insgesamt verfügen 61.543 Metadatensets über die Angabe des Titels in beiden Sprachen. Weiter finden sich insgesamt acht Metadatensätze, die ohne die Angabe von Autor*innen übermittelt wurden. Dieser Fall lässt sich nicht ausschließlich auf einen einzelnen Data Provider zurückführen. So entstammen fünf Records ohne Autor*innen von der Universität Paderborn und jeweils ein Record ohne Autor*innen von den Institutionen: Johann Wolfgang-Goethe-Universität Frankfurt, Otto-Friedrich-Universität Bamberg und der Technischen Universität Braunschweig. Ferner enthalten 82.985 Metadatensätze keine URL, die zu einer PDF-Datei verlinkt, die den Volltext der entsprechenden Dissertation bereitstellt.

Um einen Überblick über die Zeitliche Einordnung der Vertretenen Records zu

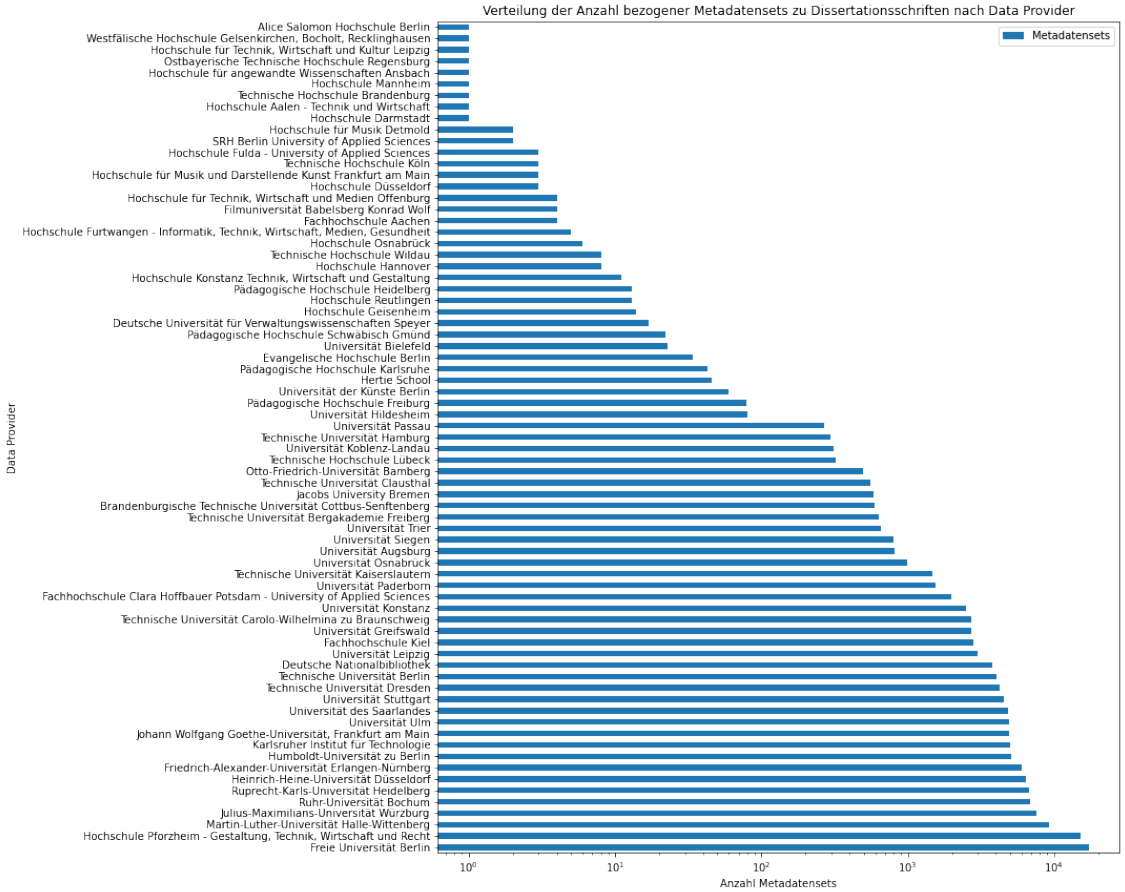

Abbildung 7 zeigt die logarithmisch dargestellte Verteilung der Anzahl von Metadaten zu Dissertationen die von den jeweiligen Dataprovidern bezogen werden konnten.

bekommen, werden die Publikationsdaten betrachtet. Die Datierungen der 143.073 angegebenen Publikationsdaten decken einen Zeitraum zwischen den Jahren 1584 und 2022 ab. Hierbei finden sich 5.107 Metadatensets, die innerhalb der Jahre 1584 und 2000 veröffentlicht wurden sowie 560 Records, die aus dem Jahr 2022 stammen. Abbildung 8 zeigt den Verlauf der absoluten Häufigkeit von extrahierten Metadaten zu Dissertationen zwischen den Jahren 2000 und 2021. Die grafische Darstellung verdeutlicht, dass sich die Anzahl der Publikationen binnen 11 Jahren fast verzehnfacht hat. Die meisten der vorliegenden Metadatensets zu Dissertationen wurden mit einer Anzahl von 9.205 im Jahr 2019 veröffentlicht. Der größte Zuwachs an Publikationen verzeichnet sich zwischen den Jahren 2006 und 2007 mit 1.712 Publikationen mehr als im Vorjahr. Zwischen den Jahren 2015 und 2021 verhält sich die Anzahl der Publikationen kontinuierlich mit einer Ausnahme im publikationsstärksten Jahr 2019.

Die Verteilungen der verschiedenen Themengebiete lässt sich durch die Betrachtung der in 128.344 Metadatensätzen vorhandenen unterschiedlichen DDC Notationen ermitteln. Dabei ist zu beachten, dass bei der Vergabe insgesamt 114.188 Einzel- und 14.156 Mehrfachnotationen zugeordnet wurden. Hierbei bildet die Vergabe einer Einzelnotation die Verortung des Themas der Arbeit spezifisch ab. Bei

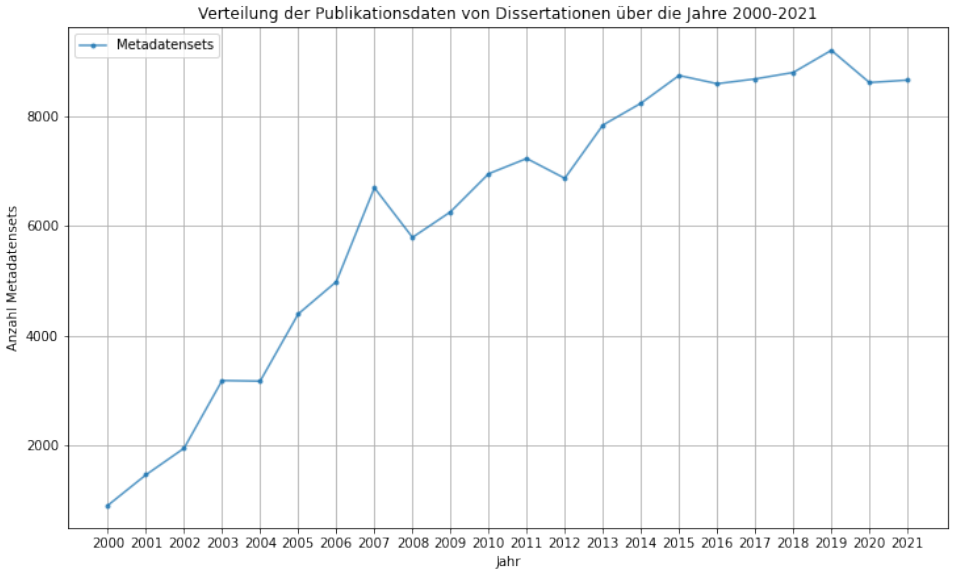

Abbildung 8 zeigt die Verteilung der Publikationszeiträume von Dissertationsschriften zwischen den Jahren 2000-2021.

den Mehrfachnotationen hingegen zeichnet sich kein Einheitliches Bild der Vergabe durch die Data Provider ab. So werden diese entweder genutzt um die übergeordnete DDC Klasse der eigentlichen spezifischen Notation anzugeben oder um eine Multidisziplinarität der Arbeit zu kennzeichnen.

Um einen Gesamtüberblick, über die vertretenen Wissensgebiete und deren Häufigkeit zu bekommen, wurden alle 683 in Einzel- oder Mehrfachnotationen vorkommenden unterschiedlichen DDC Notationen der Metadatensets extrahiert und unter ihren Hauptklassen in Abbildung 9 zusammengefasst. Bei der Betrachtung fällt auf, dass sich Metadatensets aus Unterklassen aller der zehn Hauptklassen in den Daten wiederfinden. Ebenfalls zeigt sich, dass ein überdurchschnittlich großer Anteil der verschiedenen DDC Notationen aus den Hauptklassen *600 – Technik* und *500 – Naturwissenschaften* entstammt. Die Absolute Häufigkeit der vergebenen DDC Notationen zeigt, dass mit 32.989 aus der Klasse *610 - Medizin und Gesundheit* und mit 14.677 aus der Klasse *570 - Biowissenschaften; Biologie* diese Notationen am häufigsten vertreten sind. Der Geringste Anteil an vertretenen DDC Notationen entstammt aus der Hauptklasse *200 - Religion* mit insgesamt 655 vergebenen Notationen.

Ein besondere Stellung nehmen die Daten, die durch die Datenbank der DNB bezogen wurden ein. Da diese schon durch die Rollen der DNB als Service Provider von unterschiedlichen Dataprovidern bezogen und homogenisiert wurden. Ebenso wurden die Daten hier selektiv nach DDC Notationen abgefragt. Insgesamt wurden wie eingangs erwähnt 3.797 Metadatensets zu Dissertationen durch die DNB bezogen. Davon entsprechen insgesamt 2.551 Metadatensets mindestens einer der

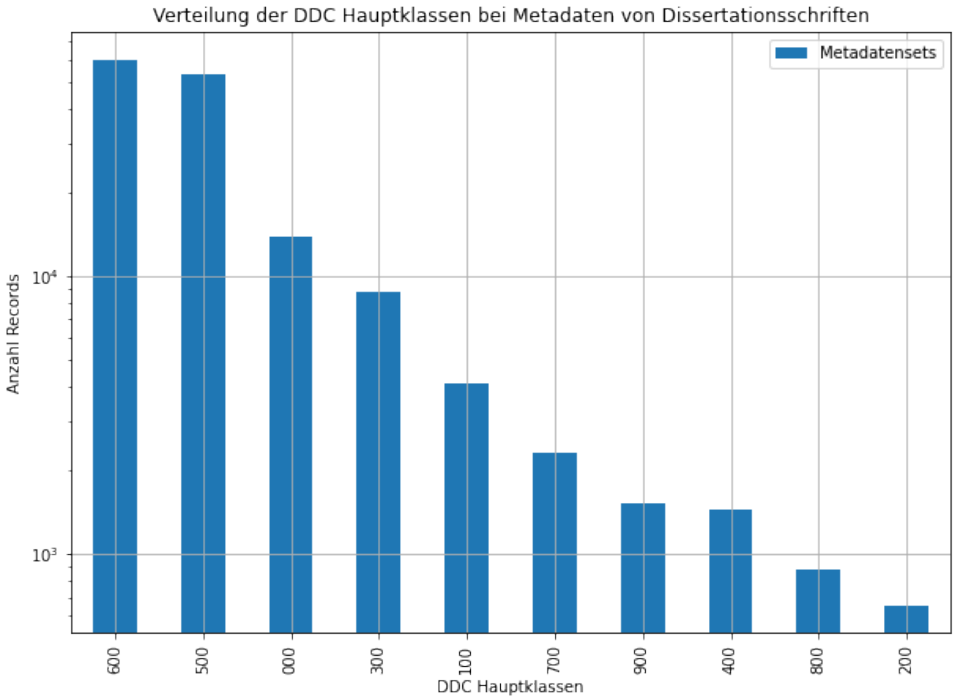

Abbildung 9 zeigt die logarithmisch dargestellte Verteilung der 683 zu Hauptklassen zusammengefassten DDC Notationen innerhalb aller bezogenen Metadatensets.

Angefragten DDC Notationen. Ebenfalls wurden 47 Metadatensätze gänzlich ohne die Angabe einer mindestens dreistelligen DDC Notation übermittelt. Der Großteil der erhaltenen Daten mit 2.462 Records trägt mindestens die DDC Notation *004 - Informatik* und 147 Records entsprechen mindestens der DDC Notation *020 - Bibliotheks- und Informationswissenschaften*. Hierbei können aufgrund der Mehrfachnotationen Dopplungen entstehen.

Um die erzeugte Sammlung aus Metadaten zu Dissertationsschriften auf Vollständigkeit zu Überprüfen, bietet der in Kapitel 1.1.3 erwähnte Gesamtbestand von Dissertationen der DNB einen Anhaltspunkt. Hier wird mit Stand November 2020 angegeben, dass sich 284.000 Dissertationsschriften in der Datenbank befinden. Die hier erzeugte Sammlung entspricht damit, inklusive der Einbeziehung der Records, die durch die DNB bezogen wurden, 50,38% des kompletten Bestandes der DNB.

Weiter wurde zur Überprüfung der Vollständigkeit ein Testset bestehend aus 50 URNs relevanter Dissertationen von insgesamt 34 unterschiedlichen Institutionen und aus unterschiedlicher Jahren erstellt. Diese wurden durch die BASE und die Zeitschrift Datenbank Spektrum[15] zusammengetragen. Bei einem Abgleich zeigte sich, dass 41 der Dissertationen in der Sammlung enthalten sind.

[15]https://www.springer.com/journal/13222

3.2 Datenanreicherung

Durch die in Abschnitt 2.4.3 erläuterte Methode der Datenanreicherung wurde versucht fehlende Metadaten der Records zu ergänzen. Hierbei bezog sich die Anreicherung auf die Attribute *ddc*, *subjects* und *pdf_ url*. Da das Anreichern fehlender Metadaten durch die einzelnen Anfragen an Webserver besonders zeitintensiv ist, wurde dieser Prozess, wie in Abbildung 4 zu sehen, an unterschiedlichen Stellen des Softwareentwurfs durchgeführt um möglichst nur Metadatensets anzureichern, die für die weitere Selektion relevant sind. Daher erfolgt das Anreichern der DDC Notationen auf die Metadatensets, die auf Dokumente des Typs der Dissertationsschrift referenzieren. Somit bildet sich die Gesamtmenge hier aus 143.073 Records. Von dieser Menge ausgehend sind 128.344 mit mindestens einer DDC Notation Klassifiziert. Von den 14.729 Records, die ohne eine oder mehrere DDC Notationen bezogen wurden, konnten 6.450 durch eine DDC Notation ergänzt werden.

Das Anreichern der Attribute *subjects* und *pdf_ url* erfolgt nach der Anreicherung und Selektion der Records, die mindestens einer der relevanten DDC Notationen entsprechen. Daher bildet sich die Gesamtmenge hier aus 7.697 Records. Von dieser Gesamtmenge beinhalten 5.022 Records die Angabe eines oder mehrerer *Subjects*. Durch die Anreicherung konnten insgesamt 487 Records durch Subjects ergänzt werden. Weiter verfügten vor der Anreicherung 1.977 Records über eine PDF-URL, diese konnte in 945 Fällen angereichert werden.

3.3 Schlagwortsuche

In Kapitel 2.4.5 wird die Methode vorgestellt, die verwendet wird um möglichst relevante Records aus den in Kapitel 1.2 als der Informationswissenschaft nahestehend identifizierten Disziplinen in die Ergebnismenge einzubeziehen. Bei dieser Methode werden anhand der Termhäufigkeit möglichst relevante Terme aus den Titeln der Records, die der DDC Notation *020 - Bibliotheks- & Informationswissenschaften* entsprechen, extrahiert. Auf den in Kapitel 3.1 vorgestellten Datensatz ergab sich die in Tabelle 2 dargestellte Liste aus Schlagworten.

Durch die Schlagwortsuche wurden insgesamt 3.616 Metadatensätze identifiziert, die mindestens einen der in Tabelle 2 aufgelisteten Terme im englischen oder deutschen Titel enthalten. Aus dieser geht hervor, dass die häufigsten Treffer durch den Begriff *data* erzeugt wurden und die beiden geringsten Treffermengen durch die Terme *bibliothek* und *librari* entstanden sind. Um das System exemplarisch zu evaluieren wurde ein Testset, bestehend aus 50 relevanten Records aus den Metadaten, erstellt und überprüft, ob diese in der Treffermenge enthalten sind. Hierbei konnten 39 der 50 relevanten Dokumente in der Treffermenge identifiziert werden.

Keywords_de	Treffer	Keywords_en	Treffer
bibliothek	7	inform	192
information	96	data	614
untersuchung	40	search	66
internet	41	librari	5
lernen	25	manag	130

Tabelle 2 zeigt die fünf nach Sprache extrahierten Terme der Schlagwortsuche und die Treffermenge, die bei der Schlagwortsuche erzielt wurde. Hierbei wurden die deutschen Terme lemmatisiert und die englischen gestemmt.

4 Ergebnisse

Im folgenden Kapitel werden die Ergebnisse dieser Arbeit beschrieben. Diese unterteilen sich in die erarbeitete Liste aus Hochschulschriftenservern von deutschen Hochschulen und dem selektierten Datensatz der Metadaten zu Dissertationen, die dem erweiterten Themenspektrum der Informationswissenschaft entsprechen.

4.1 Liste der Hochschulschriftenserver

Um die Abfrage deutscher Hochschulschriftenserver via dem OAI-PMH zu ermöglichen, wurde eine Liste aus deutschen Hochschulen erstellt. Diese Liste wurde nach Verfügbarkeit durch die entsprechende OAI-PMH-URL sowie entsprechende OAI-PMH-Sets, die Dissertationsschriften beinhalten, erweitert. Im Folgenden wird die Verteilung der Hochschulen und die Häufigkeit der identifizierten Hochschulschriftenserver betrachtet.

4.1.1 Hochschulen in Deutschland

Insgesamt finden sich in der erstellten liste 423 Hochschulen aus allen deutschen Bundesländern, die sich in drei Typen von Trägerschaften aufgliedern. Der größte Teil der deutschen Hochschulen bezieht sich mit einer Anzahl von 273 auf eine Staatliche Trägerschaft. Weiterhin finden sich 112 Hochschulen mit privater und 38 Hochschulen mit kirchlicher Trägerschaft. Bei Betrachtung der Hochschulen pro Bundesland in Abbildung 10 wird deutlich, dass in Baden-Württemberg, Nordrhein-Westfalen und Bayern die meisten Hochschulen angesiedelt sind. Ebenso befinden sich in diesen Bundesländern die meisten Hochschulen, die ein Promotionsrecht besitzen.

Die Hochschulen teilen sich ebenfalls in unterschiedliche Formen auf. Die häufigste Form sind mit 205 Hochschulen die Fachhochschulen (FH) oder Hochschulen für

angewandte Wissenschaften (HAW). Am zweithäufigsten finden sich 120 Universitäten. Weiter existieren 57 Künstlerische Hochschulen, 34 Verwaltungshochschulen und sieben Hochschulen des eigenen Typs in Deutschland. Ausgehend von der Gesamtzahl der 423 Hochschulen besitzen 156 das in Kapitel 1.1 erläuterte Promotionsrecht. Davon sind 110 Universitäten, 40 Künstlerische Hochschulen, 4 Fachhochschulen, eine Verwaltungshochschule und eine Hochschule eigenen Typs mit Promotionsrecht.

4.1.2 Hochschulen mit Hochschulschriftenserver

Insgesamt konnten von den 423 Hochschulen die OAI-PMH-URL der entsprechenden Schnittstelle von 117 Hochschulschriftenservern der jeweiligen Hochschulen identifiziert werden. Die identifizierten Server teilen sich auf 110 staatlich und sechs privat finanzierte Hochschulen auf. Lediglich ein Hochschulschriftenserver, der dem in Kapitel 2.2 erwähnten Verbund AKThB angehört, konnte von den Hochschulen mit konfessionellen Trägern identifiziert werden. Von den Hochschulen mit OAI-PMH Schnittstelle besitzen 69 ein Promotionsrecht. Bei Betrachtung der Verteilung der identifizierten Hochschulschriftenserver nach Bundesland in Abbildung 10 fällt auf, das jeweils nur von einem kleinen Teil der Hochschulen der Bundesländer ein Hochschulschriftenserver mit den angewandten Methoden ermittelt werden konnte. Zum Großteil verhält sich die Anzahl der identifizierten Server proportional zur Anzahl der Hochschulen im jeweiligen Bundesland.

Neben den Schnittstellen-URLs wurden nach Verfügbarkeit auch OAI-PMH-Set-Namen, die ausschließlich Dissertationsschriften beinhalten, in der Liste hinterlegt. Bei Betrachtung insgesamt 89 Set-Namen mit 16 vorkommenden Ausprägungen, fällt auf, dass bei 65 Servern *doc-type:doctoralThesis* verwendet wird. Jedoch existieren auch 10 Versionen des selbigen Set-Namens, die komplett kleingeschrieben sind. Ebenso finden sich sehr individuelle Bezeichnungen, wie *74797065733D6469737365 72746174696F6E*, die auf das Set mit den Dissertationen verweisen.

4.2 Beschreibung der Selektierten Daten

Die final selektierten Daten bestehen aus den Metadatensätzen die mindestens mit der DDC Notation *020 - Bibliotheks- und Informationswissenschaften* ausgezeichnet sind und den durch die Schlagwortsuche ermittelten Records aus den DDC Notationen der relevanten Wissensgebiete. Diese beiden Sammlungen werden im folgenden zusammen betrachtet. Insgesamt umfasst die erzeugte Sammlung 3.930 Metadatensets die das erweiterte Themenspektrum der Informationswissenschaft abdecken. Dabei teilt sich der Datensatz in 378 Metadatensätze die durch die entsprechenden Institutionen mit der DDC Notation für Bibliotheks- und Informationswissenschaften klassifiziert wurden und 3.698 Metadatensätzen die durch die Schlagwortsuche ermittelt wurden. Dabei decken die Daten einen Zeitraum der Publikation zwischen

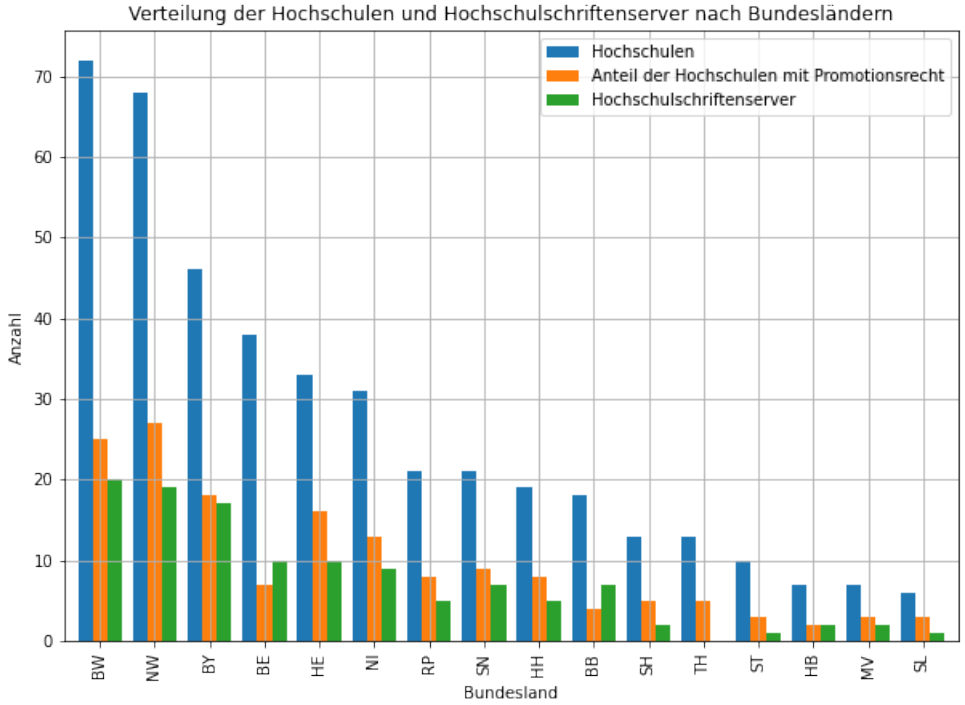

Abbildung 10 zeigt die Verteilung der Hochschulen in Deutschland, den Anteil der Hochschulen mit Promotionsrecht und den Anteil der identifizierten Hochschulschriftenserver nach Bundesland.

den Jahren 1969 und 2022 ab. Hierbei entstammen die meisten Metadatensätze aus dem Jahr 2021 mit einer Anzahl von 350, der geringste Anteil mit einem Datensatz entstammt aus dem Jahr 1969.

Die Vollständigkeit der Metadatensets ist in Tabelle 3 detailliert aufgeführt. Wie in Kapitel 3.1 bereits erwähnt finden sich aufgrund der vorhergehenden Verarbeitung und Selektion die Attribute *source*, *doc_type*, und *ddc* vollständig vor. Ebenso können auch hier bei der Betrachtung der beiden absoluten Häufigkeiten der Titelattribute und der *description* Attribute Dopplungen auftreten, da sich insgesamt 1.185 Datensätze finden, die sowohl einen englischen als auch einen deutschen Titel aufweisen und 1.654 Datensätze die sowohl eine deutsche und englische Kurzbeschreibung beinhalten. Weiter enthalten 2.287 der bezogenen Metadatensets keine Angabe einer URL die zu einer PDF-Datei die den Volltext der entsprechenden Dissertation beinhaltet verlinkt. Dafür enthalten alle Metadatensets mindestens eine URL, eine URN, die Angabe eines Autor*in und eine Angabe des Publikationsdatums.

Insgesamt entstammen die selektierten Daten von 40 unterschiedlichen Dataprovidern, in Abbildung 11 wird die Anzahl der bezogenen Metadatensets nach den entsprechenden Dataprovidern von denen Daten in die Selektion einbezogen wurden gezeigt. Bei der Betrachtung wird deutlich, dass ein Großteil der selektierten Metadaten mit einer absoluten Häufigkeit von 1.146 Metadatensets von der Datenbank

Attribut	Anzahl bei Dissertationen	Anzahl selektierter Datensatz
title_de	104.458	1.649
title_en	100.158	3.466
author	143.065	3.930
subjects	133.980	2.912
date	143.073	3.930
doc_type	143.073	3.930
language	143.010	3.892
rights	113.121	3.375
ddc	128.344	3.930
pdf_url	60.088	1.643
urn	143.072	3.930
urls	143.073	3.930
rec_format	114.000	1.936
description_de	143.073	1.917
description_en	143.073	2.353
source	143.073	3.930

Tabelle 3 zeigt die Attribute der Datensätze und deren Absolute Häufigkeit bei den Metadaten die auf Dissertationsschriften referenzieren und bei den Metadaten die dem selektierten Datensatz entsprechen. Hierbei wird bei den Metadaten die auf Dissertationen referenzieren von einer Gesamtmenge von 143.073 ausgegangen und bei dem selektierten Datensatz von einer Gesamtmenge von 3.930 Metadatensätzen.

der DNB bezogen wurden. Ebenfalls ist zu sehen, dass der Großteil der bezogenen Metadatensets die der DDC Notation 020 entsprechen von der DNB und den Hochschulschriftenservern der Humboldt-Universität zu Berlin und der Hochschule Pforzheim entstammen.

Die Verteilung der DDC Notationen untergliedert sich in 3.031 Metadatensets die eine Einzel- und 899 Metadatensets die eine Mehrfachnotation enthalten. Hierbei sind die am häufigsten vertretene Einzelnotation die das Thema der Arbeit spezifisch einordnen mit einer absoluten Häufigkeit von 2.808 Metadatensets *004 - Informa-*

Abbildung 11 zeigt die logarithmisch dargestellte Verteilung der Metadatensets insgesamt und der Metadatensets entsprechend der DDC Notation 020 nach Data Provider in dem selektierten Datensatz.

tik und mit 160 Metadatensets *020 - Bibliotheks- Informationswissenschaften*. Da wie in Kapitel 3.1 bereits erwähnt die Mehrfachnotationen auch verwendet werden um die Übergeordnete DDC Klasse der eigentlichen spezifischen DDC Notation zu kennzeichnen, werden in Abbildung 12 nur die Mehrfachnotation betrachtet die nicht auf eine der Hauptklassen referenzieren. Bei Betrachtung der häufigsten DDC Notationskombinationen zeigt sich, dass hier die Kombination Bibliotheks- und Informationswissenschaften und Informatik am häufigsten vertreten ist. Es finden sich aber auch häufige Kombinationen von *004 - Informatik* und *650 - Management und unterstützende Tätigkeiten* oder Informatik und *570 - Biowissenschaften; Biologie*.

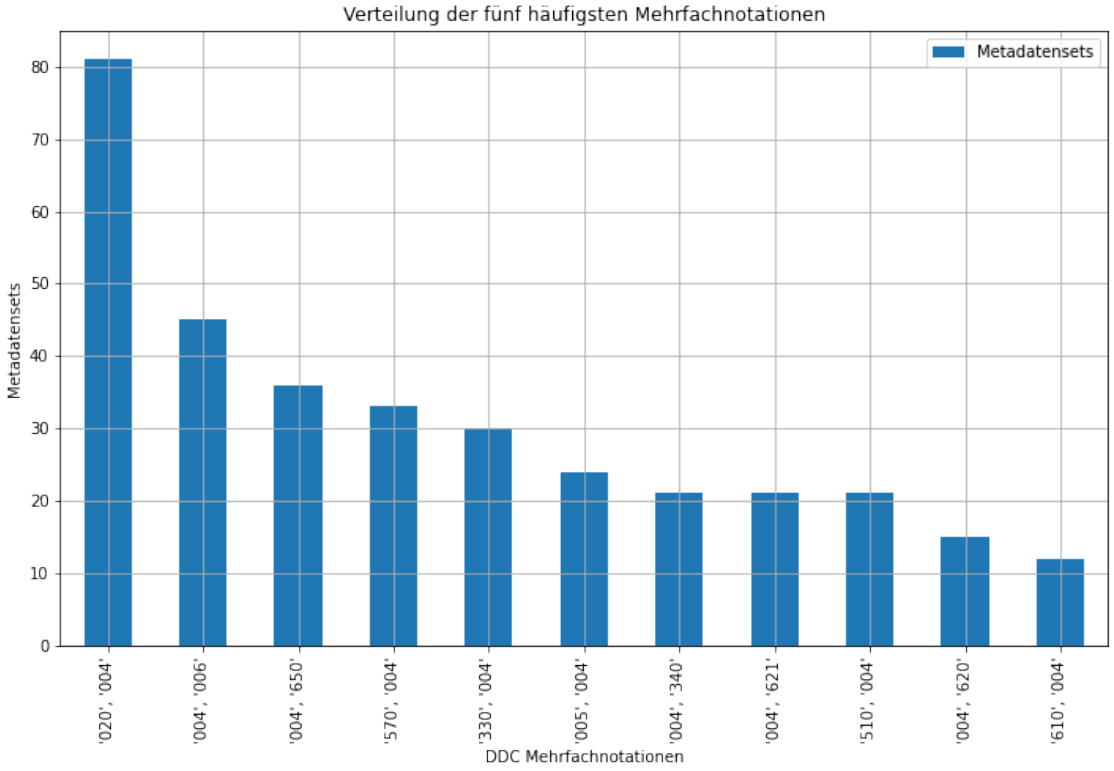

Abbildung 12 zeigt die Verteilung der fünf am häufigsten vertretenen DDC Mehrfachnotationen in dem selektierten Datensatz.

4.3 Diskussion der Ergebnisse

Die Ergebnisse der ermittelten Hochschulschriftenserver spiegeln wider, dass 117 Hochschulschriftenserver von 423 Hochschulen mit unterschiedlicher Trägerschaft identifiziert werden konnten. Die Anzahl der identifizierten Server, im Vergleich zu allen identifizierten deutschen Hochschulen, stellte sich als unerwartet gering heraus. Vor allem in Anbetracht der Publikationspflicht von Dissertationen und dem wachsenden Trend zu Open Access ist es fragwürdig, dass auf Basis der Tabelle von Wikipedia nur von 69 von insgesamt 156 Hochschulen mit Promotionsrecht ein Hochschulschriftenserver Identifiziert werden konnte. Ebenso ist die Sichtbar- beziehungsweise Auffindbarkeit der OAI-PMH-Schnittstellen auf den Webseiten der entsprechenden Institutionen nur selten gegeben.

Die finalen Datensätze bestehen aus Metadatensets, die mindestens der DDC Notation für die Bibliotheks- und Informationswissenschaften entsprechen und den durch die Schlagwortsuche als relevant ermittelten. Hierbei sollte davon ausgegangen werden können, dass die 378 Metadatensets, die der DDC Notation der Bibliotheks- und Informationswissenschaften entsprechen, eindeutig als relevant zu betrachtet sind. Jedoch ist die Relevanz der Dokumente hier abhängig von der vorherigen Klassifikation. So findet sich mit der DDC Notation 020 beispielsweise ein Metadatensatz mit dem Titel *Fettreiche Ernährung induziert Stoffwechselstörung in der Fruchtfliege*

Drosophila melanogaster. Dieser wurde auf der Repositorien Webseite[16] auch mit der Notation für Bibliotheks- und Informationswissenschaften ausgezeichnet, behandelt jedoch augenscheinlich kein für die Bibliotheks- und Informationswissenschaften relevantes Thema. Das zeigt, dass selbst durch die Selektion der Wissensgebiete durch die DDC keine vollumfängliche Relevanz der Dokumente gewährleistet werden kann.

Um die Sammlung der selektierten Daten zu erweitern wurde die Methode der Schlagwortsuche angewandt. Durch diese Methode werden alle Metadatensets in die Ergebnismenge mit einbezogen die eines der Schlagworte innerhalb des Titels enthalten. Die so ermittelten Metadatensets lassen sich in ihrer Relevanz innerhalb dieser Arbeit nicht optimal überprüfen, da keine belastbaren Test-Daten zur Evaluation bestehen. So konnte zwar anhand eines kleinen Testsets mit relevanten Dokumenten überprüft werden, wie viele der darin enthaltenen Metadatensets sich in der Ergebnismenge befinden und bei der Durchsicht der Titel ergibt sich ein gutes Bild der selektierten Daten, allerdings ist nicht belegbar, wie viele relevante und wie viele nicht relevante Metadatensätze sich in der Ergebnismenge befinden. Ebenso ist abzuwägen, ob beispielsweise eine Dissertation, die sich mit maschinellem Lernen vor dem Hintergrund einer medizinischer Fragestellung auseinandersetzt, als relevant betrachtet werden kann. Diese behandelt zwar das relevante Thema des maschinellen Lernens, jedoch mit Bezug zu einem nicht relevanten Themengebiet. Ein weiteres Beispiel wäre ein Metadatensatz mit dem Titel: *Aspekte zur Qualitätskontrolle für Hochdurchsatzsequenzierungsdaten in der medizinischen Genetik*. Auch hier ist abzuwägen, ob die Inhalte dieser Dissertation als relevant betrachtet werden können, nur weil sich das Schlagwort *data* aus der englischen Variante des Titels extrahieren ließ. Eine mögliche Argumentation wäre, dass in beiden Beispielen die Datenverarbeitung und Auswertung eine Rolle spielt und sie damit als relevant für das erweiterte Themenspektrum der Informationswissenschaft betrachtet werden können. Eine weitere Problematik ergibt sich aus der Beurteilung der Vollständigkeit, da auch hier keine belastbaren Zahlen existieren, wie viele für die Informationswissenschaft relevante Dissertationen aus dem deutschen Raum insgesamt existieren.

Positiv lässt sich die Aktualität der Daten hervorheben. So entstammt der größte Anteil der selektierten Metadaten aus dem Jahr 2021.

Es zeigt sich also, dass die erzielten Ergebnisse den Anforderungen der Aufgabenstellung entsprechen, jedoch nicht hinsichtlich ihrer Relevanz belegbar evaluiert werden konnten. Ebenso konnte keine Evaluation der Vollständigkeit der selektierten Metadaten vorgenommen werden.

[16] https://macau.uni-kiel.de/receive/diss_mods_00016705?lang=de

5 Fazit und Ausblick

Ziel dieser Arbeit war die Erstellung einer möglichst vollständigen und aktuellen Sammlung von Metadaten zu Dissertationsschriften, die dem erweiterten Themenspektrum der Informationswissenschaft entsprechen. Hierzu wurde im ersten Schritt eine Liste von Hochschulschriftenservern erarbeitet, die eine OAI-PMH Schnittstelle zur Verfügung stellen. Um darauf aufbauend einen Softwareentwurf zu entwickeln, der die identifizierten Data Provider nach Möglichkeit spezifisch nach Dissertationsschriften abfragt und die Metadatensets speichert. Die erhaltenen Daten sollten homogenisiert und selektiert abgelegt werden. Um dieses Ziel zu erreichen mussten bestimmte Voraussetzungen erfüllt werden. So war es nötig einen Überblick über die Promotionspraxis in Deutschland zu bekommen und die Rolle der Dissertation zu erörtern. Zudem musste ermittelt werden, wie sich das erweiterte informationswissenschaftliche Spektrum abgrenzen lässt und welche Disziplinen Themen und Methodiken präsentieren, die als relevant eingestuft werden können. Ebenso musste sich in die Funktionsweise des OAI-PMH eingearbeitet und eine Möglichkeit identifiziert werden, wie die erhaltenen Metadaten nach Wissensgebiete und Relevanz selektiert werden können.

Bei der Recherche mit dem Ziel, die Hochschulschriftenserver zu identifizieren, ist in erster Instanz aufgefallen, dass die unterschiedlichen Begriffsdefinitionen, die auf die jeweiligen Inhalte der Repositorien verweisen, sehr uneinheitlich verwendet werden. Dem zugrunde liegt die Tatsache, dass nicht jede Hochschule mehrere Repositorientypen getrennt nach Inhalten bereitstellt oder eine individuelle Aufteilung der Institutionellen Repositorien mit eigenen Bezeichnungen vorgenommen wird. Bei den Vorgestellten Quellen zur Suche nach entsprechenden Dataprovidern zeichnete sich zudem die uneinheitliche Benennung der Repositorien ab und erschwerte die Selektion relevanter Server. Weiter sind bei einer Recherche über die offiziellen Webseiten der Hochschulen die OAI-PMH Schnittstellen nicht einfach ausfindig zu machen. In einigen Fällen konnten diese nur durch eine Manipulation der URL der Webpräsenz des Repositoriums ermittelt werden. Ferner zeichnete sich ein stark heterogenes Bild der verwendeten Set-Namen für OAI-PMH-Sets ab, die Dissertationsschriften enthalten. Dies führt dazu, dass bei jedem Data Provider die Sets manuell abgefragt werden und nach dem entsprechenden Set und dessen Bezeichnung gesucht werden muss. Die Tatsache, dass sich diese Set-Namen auch wieder verändern können und damit die Qualität einer automatisierten Abfrage deutlich verschlechtert werden kann, erfordert, dass die Abfrage der einzelnen Repositorien in regelmäßigen Abständen überprüft werden muss. Da teilweise auch nummerische Daten oder unterschiedliche Set- Bezeichnungen vorkommen, wäre ein maschinelles Auslesen der Set-Namen von einem ähnlichen Problem betroffen. Die Tatsache,

dass nicht jeder der identifizierten Data Provider ein Set für Dissertationen zur Verfügung stellt, führt dazu, dass teilweise die gesamten Records innerhalb eines Repositoriums bezogen werden müssen. Von diesen wird, wenn überhaupt, nur ein Bruchteil genutzt. Das verlängert die gesamte Laufzeit des Softwareentwurfs und steht auch nicht im Sinne der Datensparsamkeit. Weiterhin führt die unterschiedliche Verwendung der DC-Elemente sowie die unterschiedlichen Bezeichnungen für Dissertationsschriften zu einem erheblichen Mehraufwand in der Verarbeitung und Homogenisierung der Daten. So musste jede Metadatenstruktur der unterschiedlichen Data Provider betrachtet werden, um etwaige Sonderfälle abdecken zu können und die Daten einheitlich abzulegen. Besonders positiv hervorzuheben ist die überdurchschnittlich häufige Verwendung der DDC zur Klassifizierung der Metadatensets zu Dissertationen. Dadurch konnte die Selektion innerhalb dieser Arbeit ohne die Anwendung rechenintensiver Algorithmen vorgenommen werden, die entsprechende Themengebiete aus den Kurz- oder Volltexten extrahieren.

Der Erfolg des OAI-PMH basiert auf der Einfachheit seiner Anwendung. Dadurch konnte der Harvesting Prozess für dieses Projekt mit der Einarbeitung in die Funktionsweisen des Protokolls und wenigen technischen Mitteln realisiert werden. Einzig die Datenbank der DNB erfordert hier ein spezielleres Vorgehen. Dabei konnte kein Verweis auf den offiziellen Seiten der DNB dazu gefunden werden, weshalb der Server eine Anfrage mit undefiniertem Zeitraum abbricht. Besonders, da nicht die Datenmenge Grund dafür zu sein scheint. Dafür spricht die Tatsache, dass sich der deutlich größere abgefragte Datensatz der Dissertationen aus der Disziplin der Informatik in den selben zwei Schritten abfragen lässt, wie der wesentlich kleinere Datensatz der Dissertationen aus den Disziplinen Bibliotheks- und Informationswissenschaften. Weiter wurden Records von der DNB übermittelt, die nicht den angefragten Disziplinen entsprachen. Dem zugrunde liegt vermutlich, dass die DNB die Daten selbst von Institutionellen Repositorien automatisiert bezieht. Wenn dort keine saubere Trennung vorliegt, gehen diese Datenbestände natürlich auch so in die Sammlung der DNB mit ein. Die Vollständigkeit der insgesamt erhaltenen Metadaten zu Dissertationsschriften ließ sich zwar anhand eines Vergleichs mit der Gesamtanzahl von Dissertationen die die Datenbank der DNB aufführt ansatzweise evaluieren, allerdings ist der angegebene Stand 2022. Zusätzlich ist keine Garantie gegeben, dass hier alle Dissertationen enthalten sind, die in Deutschland publiziert wurden. Ferner wurde die Vollständigkeit anhand eines Testsets überprüft. Allerdings wird hierdurch nur überprüft, ob eine Stichprobe von Metadaten zu Dissertationsschriften enthalten ist.

Der Softwareentwurf wurde auf Basis von 13 relevanten Dataprovidern erstellt. Daher besteht die Möglichkeit, dass bei der Abfrage aller Server, während der Datenextraktion oder bei der Normalisierung der Dokumententypen von Dissertationsschriften, einige Sonderfälle durch das Raster fallen und so nicht erfasst oder falsch

strukturiert werden. Die angewandte Methode zur Datenanreicherung konnte keine optimalen Ergebnisse erzielen. Grund hierfür bildet der generische und recht einfache Ansatz zur Extraktion der gewünschten Metadaten aus den geparsten HTML Strukturen. Einen höheren Zeitaufwand vorausgesetzt, könnten hier Scraping-verfahren angewandt werden, die speziell an die Bedingungen der einzelnen Repositorien-Webseiten angepasst sind, um so möglichst viele Daten zu ergänzen. Ebenso könnte die Methode der Schlagwortsuche weiter ausgefeilt werden um noch präzisere Ergebnisse zu erzielen.

Insgesamt entsprechen die Ergebnisse aus dieser Arbeit zwar der Aufgabenstellung, konnten jedoch nicht hinsichtlich ihrer Vollständigkeit und Relevanz überprüft werden, da keine belastbare Evaluation möglich war.

Der in dieser Arbeit entwickelte Softwareentwurf kann ebenfalls verwendet werden um Dissertationen aus anderen Disziplinen zusammenzutragen. Durch das Verändern der Set-Namen könnten andere Dokumententypen oder alle Dokumente der Repositorien bezogen werden. So bekäme man eine Übersicht über Inhalte und Anteile von Dokumenten, die auf deutschen Hochschulschriftenservern abgelegt wurden. Zudem bestünde die Möglichkeit durch das Abfragen der PDF-URLs eine Sammlung aus Volltexten zu erzeugen, die tiefgreifendere Analysen zuließen.

Literatur

Baderschneider, M. (2021). *Informationswissenschaft B.A. - Universität Regensburg* (Universität Regensburg, Hrsg.). Abgerufen am 10. Januar 2022 von https://www.uni-regensburg.de/studium/studienangebot/studiengaenge-a-z/informationswissenschaft-ba/index.html

Bawden, D. & Robinson, L. (2012). *Introduction to Information Science*. Facet.

Borko, H. (1968). Information science: What is it? *JASIST*, *9*(12), 3–5. https://doi.org/10.1002/asi.5090190103

Budapest Open Access Initiative. (2002). *Budapest Open Access Initiative - Erklärung* (Open Society Institute, Hrsg.). Abgerufen am 5. Dezember 2021 von http://www.budapestopenaccessinitiative.org/translations/german-translation

Bundesbericht Wissenschaftlicher Nachwuchs: Statistische Daten und Forschungsbefunde zu Promovierenden und Promovierten in Deutschland. (2021). wbv Media GmbH Co. KG. https://doi.org/https://doi.org/10.3278/6004603aw

Chan, M., Mitchell, J. & Alex, H. (2006). *Dewey-Dezimalklassifikation: Theorie und Praxis: Lehrbuch zur DDC 22*. De Gruyter Saur.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine*, *18*(4), 33. https://doi.org/10.1609/aimag.v18i4.1320

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. https://doi.org/https://doi.org/10.1002/aris.1440370103

Deutsche Initiative für Netzwerkinformation. (2021a). *DINI-Zertifikat 2019 für Open-Access-Publikationsdienste* (Deutsche Initiative für Netzwerkinformation e. V., Hrsg.). Abgerufen am 3. Januar 2022 von https://dini.de/dienste-projekte/dini-zertifikat/

Deutsche Initiative für Netzwerkinformation. (2021b). *Liste der Publikationsdienste* (Deutsche Initiative für Netzwerkinformation e. V., Hrsg.). Abgerufen am 3. Januar 2022 von https://dini.de/dienste-projekte/publikationsdienste/

Deutsche Nationalbibliothek. (o.D. a). *DissOnline und Online-Dissertationen* (Deutsche Nationalbibliothek, Hrsg.). Abgerufen am 5. Januar 2022 von https://www.dnb.de/DE/Professionell/Services/Dissonline/dissonline_node.html

Deutsche Nationalbibliothek. (o.D. b). *Unser Sammelauftrag* (Deutsche Nationalbibliothek, Hrsg.). Abgerufen am 3. Januar 2022 von https://www.dnb.de/DE/Professionell/Sammeln/sammeln_node.html

Dobratz, S. & Müller, U. (2009). Wie entsteht ein Institutional Repository? *cms-journal*, (32), 47–54. https://doi.org/10.18452/6570

Engelfried, C. & Ibisch, P. (2016). *Promovieren an und mit Hochschulen für Angewandte Wissenschaften: am Wendepunkt?* Verlag Barbara Budrich.

Ginsparg, P. (1994). First Steps Towards Electronic Research Communication. *Computers in Physics*, *8*(4), 390–396. https://doi.org/https://doi.org/10.1063/1.4823313

Grefenstette, G. (1999). Tokenization. In H. van Halteren (Hrsg.), *Syntactic Word-class Tagging* (S. 117–133). Springer Netherlands. https://doi.org/10.1007/978-94-015-9273-4_9

Gumm, H. & Sommer, M. (2012). *Einführung in die Informatik* (10. Aufl.). Oldenbourg Verlag.

Heery, R. (1996). Review of metadata format. *Program: electronic library and information systems*, *30*(4), 345–373. https://doi.org/https://doi.org/10.1108/eb047236

Heidrun, A., Bee, G. & Junger, U. (2018). *Klassifikationen in Bibliotheken: Theorie – Anwendung – Nutzen*. De Gruyter Saur.

Hochschule Darmstad. (o.D.). *Information Science (Bachelor of Science)* (Hochschule Darmstadt, Hrsg.). Abgerufen am 18. Januar 2022 von https://h-da.de/studium/studienangebot/studiengaenge/information-science-und-informatik/information-science-bsc

Hochschulverband Informationswissenschaft. (o.D.). *Über uns* (Institut für Bibliotheks- und Informationswissenschaft, Hrsg.). Abgerufen am 25. Januar 2022 von https://www.informationswissenschaft.org/ueber_uns/

Jivani, A. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930–1938.

Kooperativer Bibliotheksverbund Berlin-Brandenburg. (2022). *Bekannte OPUS 4-Instanzen in Europa* (Zuse-Institut Berlin, Hrsg.). Abgerufen am 3. Januar 2022 von https://www.kobv.de/entwicklung/software/opus-4/referenzen/

Kramer, F. F. (2019). *Ein allgemeiner Ansatz zur Metadaten-Verwaltung* (Diss.). University of Kiel, Germany. https://d-nb.info/1179399102

Kuhlen, R. (2013). *Grundlagen Der Praktischen Information Und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis* (R. Kuhlen, W. Semar & S. D., Hrsg.; 13. Aufl.). De Gruyter Saur.

Kultusminister Konferenz. (o.D.). *Der Bologna-Prozess* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). Abgerufen am 3. Januar 2022 von https://www.kmk.org/themen/hochschulen/internationale-hochschulangelegenheiten.html

Kultusministerkonferenz. (1997). *Grundsätze für die Veröffentlichung von Dissertationen* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). Abgerufen am 3. Januar 2022 von https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1977/1977_04_29-Grundsaetze-Veroeffentlichungen-Dissertationen.pdf

Max Planck Gesellschaft (Hrsg.). (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Abgerufen am 12. Januar 2022 von https://www.hdm-stuttgart.de/iw

Meurer, P. (2018). *Zugang von FH-/HAW-Absolventinnen und -Absolventen zur Promotion, kooperative Promotionen und Promotionsrecht* (Studien zum deutschen Innovationssystem Nr. 16-2018). Berlin, Expertenkommission Forschung und Innovation (EFI). http://hdl.handle.net/10419/175558

Nicholas, D., Rowlands, I., Jubb, M. & Jamali, H. (2010). The impact of the economic downturn on libraries: With special reference to university libraries. *The Journal of Academic Librarianship*, *36*(5), 376–382. https://doi.org/https://doi.org/10.1016/j.acalib.2010.06.001

Open Archives Initiative. (2001a). *OAI-PMH Registered Data Providers* (The Open Archives Initiative, Hrsg.). Abgerufen am 20. Dezember 2021 von http://www.openarchives.org/Register/BrowseSites

Open Archives Initiative. (2001b). *Open Archives Initiative Organization* (The Open Archives Initiative, Hrsg.). Abgerufen am 20. Dezember 2021 von https://www.openarchives.org/organization/

Open Archives Initiative. (2001c). *The Open Archives Initiative Protocol for Metadata Harvesting* (The Open Archives Initiative, Hrsg.). Abgerufen am 20. Dezember 2021 von http://www.openarchives.org/OAI/openarchivesprotocol.html

open-access.network. (2021a). *Grün und Gold* (open-access.network, Hrsg.). Abgerufen am 1. Januar 2022 von https://open-access.network/informieren/open-access-grundlagen/open-access-gruen-und-gold

open-access.network. (2021b). *Repositorien* (open-access.network, Hrsg.). Abgerufen am 1. Januar 2022 von https://open-access.network/informieren/open-access-grundlagen/repositorien

open-access.network. (2021c). *Was bedeutet Open Access?* (open-access.network, Hrsg.). Abgerufen am 1. Januar 2022 von https://open-access.network/informieren/open-access-grundlagen/was-bedeutet-open-access

Registry of Open Access Repositories. (2022). *Registry of Open Access Repositories* (University of Southampton, Hrsg.). Abgerufen am 3. Januar 2022 von http://roar.eprints.org/

Roos, A. (Hrsg.). (o.D.). *Studiengang Informationswissenschaften.* Abgerufen am 12. Januar 2022 von https://www.hdm-stuttgart.de/iw

Saracevic, T. (1999). Information Science. *JASIS*, *50*(12), 1051. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12\<1051::AID-ASI2\>3.0.CO;2-Z

Statistische Bundesamt (Hrsg.). (2021). *Hochschulen nach Hochschularten.* Abgerufen am 10. Februar 2022 von https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Tabellen/hochschulen-hochschularten.html

The World Wide Web Consortium. (2013). *Extensible Markup Language (XML) 1.0 (Fifth Edition)* (The World Wide Web Consortium, Hrsg.). Abgerufen am 10. Januar 2022 von https://www.w3.org/TR/REC-xml/#sec-intro

Umstätter, W. (2009). Bibliothekswissenschaft im Wandel, von den geordneten Büchern zur Wissensorganisation. *Bibliothek Forschung und Praxis*, *33*(3), 327–332. https://doi.org/doi:10.1515/bfup.2009.036

Universität Konstanz. (2022). *Metadaten und Metadatenstandards* (Baden-württembergisches Begleit- und Weiterentwicklungsprojekt für Forschungsdatenmanagement, Hrsg.). Abgerufen am 31. Januar 2022 von https://www.forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards/

Universitätsbibliothek Bielefeld. (2022). *Datenlieferanten: Nach Aktivierungsdatum* (U. Bielefeld, Hrsg.). Abgerufen am 6. Januar 2022 von https://www.base-search.net/about/de/about_sources_date.php

Weibel, S., Kunze, J. A., Lagoze, C. & Wolf, M. (1998). Dublin Core Metadata for Resource Discovery. *RFC*, *2413*, 1–8. https://doi.org/10.17487/RFC2413

Wissenschaftsrat. (2009). *Empfehlungen zur Vergabe des Promotionsrechts an nichtstaatliche Hochschulen* (G. des Wissenschaftsrats, Hrsg.). Abgerufen am 3. Januar 2022 von https://www.forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards/