Art und Verteilung von PDF-Formaten auf wissenschaftlichen deutschen Repositorien mit Hinblick auf Langzeitarchivierung

Types and distribution of PDF-formats on scientific German repositories with regard to digital preservation

Bachelorarbeit

im Studiengang
Bibliotheks- und Informationsmanagement

vorgelegt von

Dennis Müller

am 02. Februar 2015
an der Hochschule der Medien Stuttgart

Erstprüfer/in:	Prof. Markus Hennies
Zweitprüfer/in:	Prof. Magnus Pfeffer

Eidesstattliche Versicherung

Name:	Müller	Vorname:	Dennis
		Studiengang:	BI7

Hiermit versichere ich, Dennis Müller, an Eides statt, dass ich die vorliegende Bachelorarbeit mit dem Titel „Art und Verteilung von PDF-Formaten auf wissenschaftlichen deutschen Repositorien mit Hinblick auf Langzeitarchivierung“ selbständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der eidesstattlichen Versicherung und prüfungsrechtlichen Folgen (§ 26 Abs. 2 Bachelor-SPO bzw. § 19 Abs. 2 Master-SPO der Hochschule der Medien Stuttgart) sowie die strafrechtlichen Folgen (siehe unten) einer unrichtigen oder unvollständigen eidesstattlichen Versicherung zur Kenntnis genommen.

Auszug aus dem Strafgesetzbuch (StGB)

§ 156 StGB Falsche Versicherung an Eides Statt

Wer von einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

Ort, Datum

Unterschrift

Kurzfassung

Langzeitarchivierung ist schon seit langer Zeit ein wichtiges Thema in der Bibliothekswelt und wird dies naturgemäß auch in Zukunft sein. Allerdings stellen der technische Fortschritt und die Entwicklungen in der Informationstechnologie die Bibliotheken vor neue An- und Herausforderungen. Die PDF/A-Dateiformate sind mittlerweile internationaler Standard für die digitale Langzeitarchivierung und müssen daher beherrscht und unterstützt werden. Ziel dieser Arbeit ist zu untersuchen, inwiefern ausgewählte wissenschaftliche Bibliotheken und Forschungseinrichtungen in Deutschland diesen Anforderungen gerecht werden und ihre Repositorien, wie auch Ihre Hilfestellungen für Publizierende angepasst haben. Dazu werden Metadaten und die dazugehörigen Dokumente von den Publikationsservern per Harvesting zusammengetragen und über eine Validationssoftware analysiert. Anteile, zeitlicher Verlauf und Fehler der Dateien bzw. Formate werden dabei untersucht.

Schlagwörter: digitale Langzeitarchivierung, Portable Document Format, PDF, PDF/A, Validierung

Abstract

Long-term preservation has been and naturally will be an important topic in the world of libraries. But technical progress and advancements in information technology impose new challenges and requirements to libraries. Meanwhile the PDF/A file formats have become international standard in digital preservation and therefore must be mastered and supported. The object of this paper is to examine, how selected German scientific libraries and research facilities meet the requirements and how they have adjusted their repositories as well as their support for authors. For this purpose metadata and the respective documents are collected from the publication servers via harvesting and analyzed by validation software. Amount, progression and errors of the files respectively the file formats are examined.

Keywords: digital preservation, portable document format, pdf, pdf/a, validation

Inhaltsverzeichnis

EIDESSTATTLICHE VERSICHERUNG **2**

KURZFASSUNG **3**

ABSTRACT **3**

INHALTSVERZEICHNIS **4**

ABBILDUNGSVERZEICHNIS **6**

TABELLENVERZEICHNIS **6**

ABKÜRZUNGSVERZEICHNIS **7**

1 EINLEITUNG **8**

2 ALLGEMEINES ZUR LANGZEITARCHIVIERUNG **9**

2.1 ZWECK UND PROBLEMSTELLUNG 9
2.2 ENTWICKLUNG 10
2.3 VERFAHREN 11
2.3.1 Migration und Emulation *11*
2.3.2 OAIS Referenz Modell *12*
2.3.3 Repositorien *15*
2.3.4 Dateiformate *16*
2.3.4.1 Auswahlkriterien 16
2.3.4.2 TXT und ODF 17
2.3.4.3 PDF 17
2.3.4.4 PDF/A 18

3 UNTERSUCHUNG **20**

3.1 REPOSITORIEN 20
3.1.1 Auswahlkriterien der Repositorien *20*
3.1.2 Richtlinien und Hilfestellungen zur Langzeitarchivierung *20*
3.2 HARVESTING 21
3.2.1 OAI-PMH *21*
3.2.1.1 Begriffe 22
3.2.1.2 Anfragen 24
3.2.2 Verwendete Software und Verfahrensweise *26*
3.3 VALIDIERUNG 28
3.3.1 Hintergrund *28*
3.3.2 Verwendete Software und Verfahrensweise *29*

4 AUSWERTUNG **31**

4.1 AUSSORTIERUNG VON REPOSITORIEN 31
4.2 AUSWERTUNG DER EINZELNEN REPOSITORIEN 32

4.2.1 Albert-Ludwigs-Universität Freiburg 34
4.2.2 Deutsches Institut für Internationale Pädagogische Forschung (DIPF) 36
4.2.3 Fachhochschule Köln 38
4.2.4 Hochschule Konstanz Technik, Wirtschaft und Gestaltung (HTWG) 38
4.2.5 Forschungszentrum Jülich 39
4.2.6 Johannes Gutenberg-Universität Mainz 40
4.2.7 Justus-Liebig-Universität Gießen 42
4.2.8 Philipps Universität Marburg 43
4.2.9 Ruhr-Universität Bochum 45
4.2.10 Schloss Dagstuhl Leibniz-Zentrum für Informatik 46
4.2.11 Technische Universität Chemnitz 47
4.2.12 Technische Universität Kaiserslautern 48
4.2.13 Universität des Saarlandes 48
4.2.14 Universität Duisburg-Essen 49
4.2.15 Universität Potsdam 50
4.2.16 Universität Siegen 52
4.2.17 Universität Stuttgart 52
4.2.18 Universität Ulm 53
4.2.19 Virtuelle Fachbibliothek Fachinformation für Politikwissenschaft, Verwaltungswissenschaft und Kommunalwissenschaften (ViFaPol) 54
4.2.20 Virtuelle Fachbibliothek Psychologie (PsyDok) 54
4.2.21 Westfälische Wilhelms-Universität Münster 55
4.2.22 Hochschule Heilbronn 57
4.2.23 Humboldt-Universität zu Berlin 57
4.2.24 Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden 58
4.3 GESAMTAUSWERTUNG 59
5 FAZIT & AUSBLICK 65
LITERATURVERZEICHNIS 66
ANHANG A: PERL-SKRIPT FÜR HARVESTING UND VALIDIERUNG AM BEISPIEL DER UNIVERSITÄT FREIBURG 76

Abbildungsverzeichnis

Abbildung 1: OAIS Prozessmodell ... *13*
Abbildung 2: OAI-PMH Datenmodell ... *23*
Abbildung 3: Flow Control in OAI-PMH ... *25*
Abbildung 4: Anteil der PDF-Formate an der Universität Freiburg ... *35*
Abbildung 5: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Freiburg ... *35*
Abbildung 6: Anteil der PDF-Formate am Dt. Institut für Internationale Pädagogische Forschung ... *36*
Abbildung 7: Art und letztes Änderungsdatum von PDF-Dateien am Dt. Institut für Internationale Pädagogische Forschung ... *37*
Abbildung 8: Art und letztes Änderungsdatum von PDF-Dateien am Forschungszentrum Jülich ... *39*
Abbildung 9: Anteil der PDF-Formate an der Universität Main ... *40*
Abbildung 10: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Mainz ... *41*
Abbildung 11: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Gießen ... *42*
Abbildung 12: Anteil der PDF-Formate an der Universität Marburg ... *44*
Abbildung 13: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Marburg ... *44*
Abbildung 14: Anteil der PDF-Formate an der Universität Bochum ... *45*
Abbildung 15: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Bochum ... *46*
Abbildung 16: Art und letztes Änderungsdatum von PDF-Dateien am Schloss Dagstuhl Leibniz-Zentrum für Informatik ... *47*
Abbildung 17: Anteil der PDF-Formate an der Universität Potsdam ... *51*
Abbildung 18: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Potsdam ... *51*
Abbildung 19: Art und letztes Änderungsdatum von PDF-Dateien bei der Virtuellen Fachbibliothek Psychologie (PsyDok) ... *55*
Abbildung 20: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Münster ... *56*
Abbildung 21: Anteil der PDF-Formate insgesamt ... *60*
Abbildung 22: Art und letztes Änderungsdatum von PDF-Dateien insgesamt ... *61*

Tabellenverzeichnis

Tabelle 1: Fehlermeldungen des 3-Heights-PDF-Validators ... *34*
Tabelle 2: Art und letztes Änderungsdatum von PDF-Dateien insgesamt ... *62*
Tabelle 3: Häufigkeit der Fehlermeldungen insgesamt ... *64*

Abkürzungsverzeichnis

AIP	Archival Information Package
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
DINI	Deutsche Initiative für Netzwerkinformationen
DIP	Dissemination Information Package
HTTP	Hyper Text Transfer Protocol
ICC	International Color Consortium
ISO	International Organization for Standardization
JHOVE	JSTOR/Harvard Object Validation Environment
LZA	Langzeitarchivierung
OA	Open Access
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
ODF	Open Document Format
Open DOAR	Directory of Open Access Repositories
PDF	Portable Document Format
PDF/A	Portable Document Format for Archives
SIP	Submission Information Package
TXT	Dateinamenserweiterung für reine Textdateien
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
UTF-8	8-Bit Universal Character Set Transformation Format
XML	Extensible Markup Language
XMP	Extended Metadata Platform

1 Einleitung

Das Thema Langzeitarchivierung beschäftigt die Bibliothekswelt schon seit langer Zeit und wird dies naturgemäß auch in Zukunft tun. Allerdings stellen der technische Fortschritt und die Entwicklungen in der Informationstechnologie die Bibliotheken vor neue An- und Herausforderungen. Dazu zählt auch die Unterstützung entsprechend konzipierter Dokumenttypen. Während aktuell zumeist das Management und die Archivierung von Forschungsdaten Diskussionsgegenstand sind, fokussiert sich diese Arbeit auf die bereits seit über zehn Jahre laufenden Publikations- und Dokumentenserver, die hauptsächlich Textdokumente im Bestand haben. Dabei sind die Archivformate des Portable Document Format (PDF) von besonderer Bedeutung. Die PDF for Archive (PDF/A) Dateiformate sind mittlerweile internationaler Standard für die digitale Langzeitarchivierung und müssen daher beherrscht und unterstützt werden.

Ziel dieser Arbeit ist zu untersuchen, inwiefern ausgewählte wissenschaftliche Bibliotheken und Forschungseinrichtungen in Deutschland diesen Anforderungen gerecht werden und ihre Repositorien, wie auch Ihre Hilfestellungen für Publizierende angepasst haben und damit die Verbreitung der PDF/A-Formate fördern. Dazu werden Metadaten und die dazugehörigen Dokumente von den Publikationsservern per Harvesting zusammengetragen und analysiert. Dabei wird darauf eingegangen, wie weit verbreitet die verschiedenen PDF-Versionen sind, welche Fehler die Dateien aufweisen und wie diesen ggf. entgegengewirkt werden kann.

Zunächst werden aber die Grundlagen der digitalen Langzeitarchivierung erläutert, um einen Überblick über die grundlegenden Problemstellungen und Lösungsansätze zu geben. Dies beinhaltet diverse Strategien bezüglich der Verfügbarkeit digitaler Objekte sowie das Open Archival Information System Referenz Modell als Rahmensystem langzeitarchivierungsorientierter Einrichtungen. Anschließend wird auf Dokumentenserver und deren Anforderungen bezüglich Langzeitarchivierung eingegangen, im Besonderen auf Open Access Repositorien. Eine dieser Anforderungen ist eine Schnittstelle für das Open Archive Initiative Protocol for Metadata Harvesting, welches ebenfalls näher erläutert wird, da es Grundlage für das für diese Arbeit praktizierte Harvesting ist.

Zudem werden diverse Textdateiformate unter der Perspektive der Langzeitarchivierung betrachtet. Dabei wird näher auf PDF- und PDF/A-Dateien und ihre Eigenschaften eingegangen. Im Zuge der Analyse der PDF-Dateien wird das Thema der Validierung behandelt. Abschließend werden die gewonnenen Daten ausgewertet und auf die angesprochenen Fragestellungen untersucht.

2 Allgemeines zur Langzeitarchivierung

2.1 Zweck und Problemstellung

Langzeitarchivierung (LZA) meint die dauerhafte Aufbewahrung von Dokumenten und die Sicherstellung des Zugangs und der Benutzbarkeit.[1] Da Archivierung eigentlich ohnehin für eine unbestimmt lange Zeit vorgesehen ist, erscheint der Begriff Langzeit-Archivierung zunächst redundant. Im Kontext digitaler Objekte, also Objekte, die aus einer Reihe Bitsequenzen bestehen[2], ist diese Doppelung aber wieder sinnvoll, um die digitale Archivierung von einem simplen Backup abzugrenzen. Aufgrund des technologischen Fortschritts ergeben sich in gewissen zeitlichen Abständen größere Probleme beim Öffnen älterer Dokumente. Im Gegensatz zu einem Backup zielt die digitale LZA darauf ab, die entsprechenden Problemschwellen und technischen Sprünge zu überdauern.

Einerseits kann es vorkommen, dass das Dateiformat des digitalen Objekts nicht mehr ausgelesen werden kann, weil es mittlerweile veraltet oder beschädigt ist. So gibt es möglicherweise keine Software mehr, die einen proprietären Textverarbeitungsdateityp (wie etwa Vorläufer von Microsoft Word) öffnen kann bzw. kein Betriebssystem, auf dem eine solche Software noch ausgeführt werden könnte.

Andererseits kann das Speichermedium selbst, auf dem die Datei gesichert ist, seine Lebensdauer überschritten haben. So können Daten etwa durch Entmagnetisierung von Magnetbändern verloren gehen oder wegen nicht mehr vorhandener oder zumindest unzureichend verbreiteter Lesegeräte nicht mehr ausgelesen werden. Als Beispiel hierfür kann die Verdrängung der Disketten und deren Laufwerke als Standardspeicher durch CD-ROM und USB-Speicher herangezogen werden.

Die LZA zielt aber nicht nur darauf ab, einzelne Bitfolgen (digitale Objekte) zu archivieren. Das eigentliche Ziel ist die Erhaltung des konzeptuellen Objektes, also der Einheit, die für den Menschen begreifbar ist. Dazu wird nicht nur die Bitfolge benötigt und ein Dateiformat, das diese charakterisiert (logisches Objekt), sondern auch die dazugehörigen Programme, die die Bitfolge in einer für den Menschen rezipierbaren Form darstellen.[3]

Wichtig bei der LZA ist neben der Aufbewahrung auch die Möglichkeit des Auffindens und des Zugangs. Dies erfordert die Erschließung der digitalen Objekte, sowie Anweisungen zum Auslesen, also das Vergeben von Metadaten. Als sozusagen kleinster gemeinsamer Nenner hat

[1] Vgl. Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K. (2010): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 18

[2] Vgl. The Consultive Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS), S. 1.11

[3] Vgl. Funk, S. E. (2010): Digitale Objekte und Formate. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.3 ff.

sich das Dublin Core Metadaten Set[4] etabliert. Deutsche Open Access Repositorien verpflichten sich beispielsweise im Rahmen des DINI-Zertifikats (s.u.) zumindest das Metadatenset Dublin Core Simple[5] zu unterstützen, welches alle grundlegenden Aspekte eines Objekts abbilden bzw. beschreiben kann.

Unter juristischen Gesichtspunkten wirft die LZA weitere Probleme auf.[6] Die Sicherstellung der Verfügbarkeit, sowie die gegebenenfalls zwingenden Veränderungen von Dokumenten verlangen nach der Zustimmung des Urhebers. Ähnliches gilt auch für die Rechte an Abspielumgebungen beim Ansatz der Emulation.

2.2 Entwicklung

Die Aufbewahrung elektronischer Dokumente beschäftigt die Bibliotheks- und Archivwelt spätestens seit der massenhaften Verbreitung der digitalen Textverarbeitung in Forschung und Wirtschaft. Manche sprechen von der Beschäftigung mit dem Thema bereits seit den 1960er Jahren.[7] Hauptsächlich an der Entwicklung von Lösungen der damit einhergehenden Fragestellungen waren und sind die jeweiligen Nationalbibliotheken und -archive, sowie Universitäten verschiedener Länder beteiligt. In den USA sind speziell das National Digital Information Infrastructure and Preservation Program, die National Archives and Records Administration (und deren Arbeit am Electronic Records Archive) und die Research Libraries Group zu nennen.

Bedeutende Fortschritte wurden in den 1990er Jahren gemacht. Mit der Entwicklung des Open Archival Information System Reference Modells (OAIS) entstand ein funktionales, detailliertes Modell für digitale Archive, das seitdem weite Verbreitung und Akzeptanz fand. Zudem wurde im gleichen Jahrzehnt das PDF-Format veröffentlicht, welches für die Archivierung digitaler Textdokumente zu besonderer Bedeutung gelangte. Sowohl auf OAIS, als auch auf PDF wird in den folgenden Kapiteln näher eingegangen.

In Deutschland engagiert sich besonders das 2003 gegründete Kompetenznetzwerk „nestor“ im Bereich der digitalen LZA.[8] Ziele von nestor sind der Austausch von Informationen unter Experten und damit auch die Entwicklung von Standards, Leitlinien und Veröffentlichung von Hilfestellungen für Bibliotheken, Archive und Museen.[9] Ebenso dient der Verbund als Bindeglied zu ähnlichen Einrichtungen anderer Länder. Zu den aktiven Teilnehmern zählen u.a. das Bibliotheks-

[4] Vgl. Dublin Core Metadata Initiative (2015): Dublin Core Metadata Element Set, Version 1.1

[5] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2013): DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste 2013, S. 47

[6] Vgl. Borghoff, U. M. (2003): Langzeitarchivierung, S. 21 ff.

[7] Vgl. Hirtle, P. B. (o.J.): The History and Current State of Digital Preservation in the United States, S. 124 ff.

[8] Vgl. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (2013): nestor - Über uns

[9] Vgl. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (2012): nestor - Publikationen

service-Zentrum Baden Württemberg, die Bayrische Staatsbibliothek, das PDF/A Competence Center, die Deutsche Nationalbibliothek und das Bundesarchiv.[10]

2.3 Verfahren

2.3.1 Migration und Emulation

Als Antworten auf die angesprochenen Problemstellungen bei der LZA haben sich die Konzepte der Migration und der Emulation herauskristallisiert.[11]

Verhältnismäßig simpel strukturierte Objekte, wie etwa Bilddateien, können durch Migration von alten Dateiformaten in neue umgewandelt werden und somit beim nächsten technologischen Sprung mitgezogen werden.[12] Allerdings bedeutet diese Methode zwangsläufig eine Veränderung des Objekts, was wiederum zur Folge hat, dass die Authentizität des Objekts damit nicht mehr gewährleistet ist. Zudem ist Migration ein Ansatz, der immer und immer wieder, nämlich bei der Einführung der nächsten Formatversion, für alle Objekte durchgeführt werden muss.

Für komplexe Dateien wie beispielsweise die Photoshop-Datei eines Bildes ist dieses Verfahren aber zu grob und daher risikoreicher, da Bearbeitungsinformationen verloren gehen können. Deshalb bietet es sich hierbei an, die zum Auslesen notwendige Hard- und Softwareumgebung zu emulieren. Emulation bedeutet, dass auf einem Computer die bestimmte Hard- und Softwarespezifikation simuliert wird. So kann die für das digitale Objekt notwendige Umgebung erzeugt werden, um es darin abzuspielen bzw. zu lesen.[13] Dabei bleibt das Objekt selbst unverändert, also authentisch. Bei der Emulation können entweder nur die Hardware, die Hardware mit Betriebssystem oder Hardware, Betriebssystem und Anwendungssoftware zusammen emuliert werden. Als am meisten praktikabel hat sich die reine Hardwareemulation erwiesen, da zum einen die Schnittstellen der meisten Hardwareplattformen offen liegen[14] und es zum anderen ein bedeutender Mehraufwand ist, sämtliche Kombinationen von Hardware, Betriebssystem und Anwendungssoftware abzudecken.

Zwar ist die Emulation komplexer als die Migration, doch bietet sie den Vorteil, dass nicht bei jedem technischen Sprung erneut alle Objekte bearbeitet werden müssen. Sie kommt der Erhaltung des konzeptuellen Objektes und damit dem eigentlichen Ziel der LZA näher als die Migration. Allerdings kann es im Laufe der Zeit notwendig werden, Emulationssoftware zu aktualisieren bzw. zu migrieren.

---

[10] Vgl. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (2013): nestor - Über uns

[11] Vgl. Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K. (2010): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S.19

[12] Vgl. Borghoff, U. M. (2003): Langzeitarchivierung, S. 37 f.,

[13] Vgl. ebd., S. 95 ff.

[14] Vgl. Suchodoletz, D. v. (2009): Emulationen für Archive

Das Betreiben von Hardware-Museen, in denen – im Idealfall möglichst alle – alten, echten Rechnermodelle stehen, ist allein schon wegen des Wartungs-, Platz- und Kostenaufwandes schlichtweg nicht praktikabel. Zudem ist dadurch nur bedingt Zugang gewährleistet.[15]

2.3.2 OAIS Referenz Modell

Das Open Archival Information System Referenz Modell wurde ab 1995[16] vom Consultative Committee for Space Data Systems entwickelt und gilt heute als Standardmodell für digitale LZA. Es wurde von der International Organization for Standardization (ISO) in Auftrag gegeben und 2003 als offizielle ISO-Norm akzeptiert, welche 2012 aktualisiert wurde[17].

Das OAIS Modell besteht aus einem Informationsmodell und einem Prozessmodell.[18] Es gibt keine konkreten Vorgaben zur technischen Umsetzung oder Spezifikationen der digitalen Objekte, wodurch das OAIS Modell allgemein bleibt und den jeweiligen Archiven Gestaltungsfreiraum gewährt.

Das Informationsmodell unterscheidet Informationen von Daten. Letztere können die unterschiedlichsten physischen wie digitalen Objekte sein, die allerdings noch keine Bedeutung haben. Bedeutung gewinnen sie erst durch die Interpretation des Rezipienten, wodurch sie zu Informationen werden.

Damit der Rezipient die Daten richtig interpretieren kann, benötigt er eine Grundlage an Vorwissen (Knowledge Base). Besitzt der Rezipient nicht die notwendige Wissensgrundlage (z.B. das Beherrschen einer Programmiersprache), kann diese durch die Repräsentationsinformation (Representation Information, z.B. ein Lehrbuch der Programmiersprache) erworben werden.[19] Dies erfordert wiederum auch eine Wissensgrundlage, etwa das Beherrschen der englischen Sprache, um die Erklärung der Programmiersprache lesen zu können. Durch die Interpretation mit Hilfe von Wissensgrundlage und Repräsentationsinformation wird also aus dem physischen oder digitalen Datenobjekt ein Informationsobjekt.

In einem Archiv gibt es nun wiederum verschiedene Informationspakete (Information Packages). Ein solches besteht grundlegend aus der Inhaltsinformation (Content Information) und der Erhaltungsmetadaten (Preservation Description Information). Die Inhaltsinformation besteht aus dem eigentlichen Informationsobjekt und seiner Repräsentationsinformation. Die Erhaltungsmetadaten machen Angaben zu Herkunft, Kontext (zu Informationen außerdem des Informationspakets), Identifier (w.z.B. eine ISBN), Beständigkeit (z.B. anhand einer Prüfsumme) und Zugriffsrechten der Inhaltsinformation.[20]

---

[15] Vgl. Borghoff, U. M. (2003): Langzeitarchivierung, S. 16 ff.

[16] Vgl. ebd., S. 26

[17] Vgl. International Organization for Standardization (o.J.): ISO 14721:2012 - Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model

[18] Vgl. Borghoff, U. M. (2003): Langzeitarchivierung, S. 26 ff.

[19] Vgl. The Consultive Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS), S. 2.4

[20] Vgl. ebd., S.2.6 f.

Die Erhaltungsmetadaten und die Inhaltsinformation werden zu einer logischen Einheit zusammengefasst und können entsprechend mit Verpackungsinformation (Packaging Information) versehen werden. Diese beschreibt den Zusammenhang von Erhaltungsmetadaten und Inhaltsinformation mit Bezug auf den physischen Speicher, auf dem sie liegen, z.B. durch die Beschreibung der Verzeichnisstruktur einer CD-ROM, um zu zeigen, welche der Informationen auf der CD-ROM nun Inhaltsinformation und welche Erhaltungsmetadaten sind.

Um dieses gesamte Informationspaket zu erschließen, wird es mit beschreibender Information (Descriptive Information), also Metadaten versehen, um später das Auffinden des Pakets zu ermöglichen.

Das Prozessmodell des OAIS[21] [22] kennt verschiedene Arten von Informationspaketen. Dies rührt von den drei verschiedenen Akteuren beim Betrieb eines Archivs, nämlich dem Management, dem Produzenten und dem Endnutzer, sowie von den diversen, damit einhergehenden Prozessen.[23]

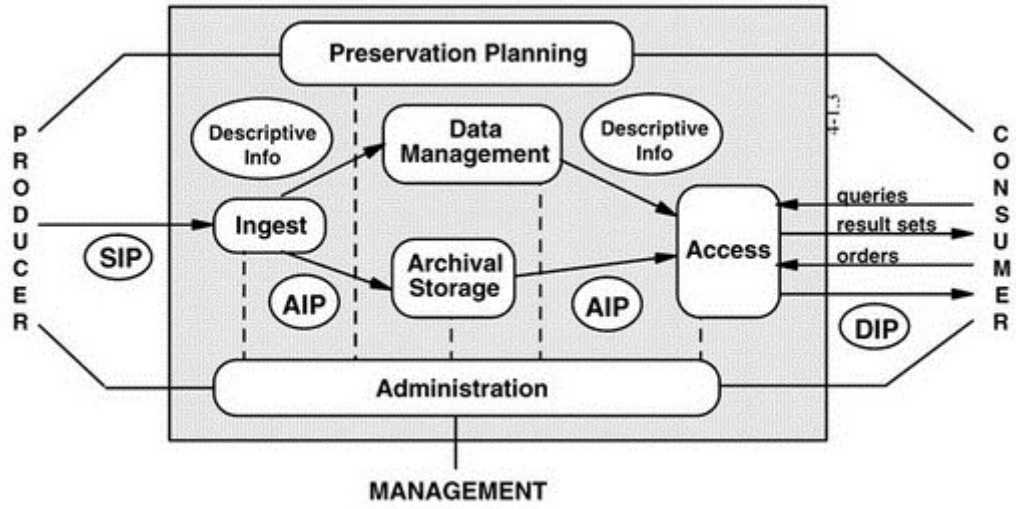

Abbildung 1: OAIS Prozessmodell[24]

Der Prozess beginnt damit, dass der Produzent (extern oder ggf. der Administrationsprozess) dem Archiv ein Submission Information Package (SIP, Übergabeinformationspaket) liefert. Seit der Ergänzung des OAIS im Jahr 2009 gibt es hier die für den Produzenten verpflichtend mitzuliefernde „Transformational Information Property“ (transformationelle Informationseigentschaft).

[21] Vgl. ebd. S. 4.1 ff.

[22] Vgl. nestor-Arbeitsgruppe OAIS-Übersetzung / Terminologie (2013): Referenzmodell für ein Offenes Archiv-Informationssystem

[23] Aus Gründen der Übersichtlichkeit und der Nachvollziehbarkeit anhand der Abbildung der OAIS-Prozesse werden hier vornehmlich die englischen Begriffe und Abkürzungen verwendet.

[24] Vgl. Schweizerische Nationalbibliothek (2011): Das OAIS-Modell

Dies ist Information darüber, welche Eigenschaften des digitalen Objekts für wie lange erhalten werden müssen, um die Authentizität zu bewahren.[25]

Der Ingest Prozess (Übernahmeprozess) nimmt dieses Packet an und bereitet es für die folgenden Prozesse vor. Er prüft die Qualität und Vollständigkeit der SIPs und erzeugt den Standards des jeweiligen Archivs entsprechend ein Archival Information Package (AIP, Archivinformationspaket). Dieses wird durch Descriptive Information (Erschließungsinformation) erschlossen. Das AIP wird an den Archival Storage Prozess (Archivspeicher) weitergegeben, die Descriptive Information an den Data Management Prozess (Datenmanagement) des Archivs.

Der Archival Storage Prozess speichert das AIP auf verschiedenen Speichermedien (z.B. auf CD-Rom und Server oder auf verschiedenen physischen oder virtuellen Serverinstanzen) und achtet dabei auf Unversehrtheit der Objekte und Aktualisierung der Speicher. Er verwaltet die hierarchische Struktur des Speichers, führt Fehlersuchen durch und erstellt Methoden zur Wiederherstellung der Daten im Falle eines Datenverlustes.

Der Prozess des Data Management wiederum verwaltet die Descriptive Information, bereitet sie auf und stellt sie beispielsweise in einem Katalog zur Recherche zur Verfügung. Zusätzlich ist der Prozess auch für die administrativen, also archivinternen Daten zuständig. Das Data Management ist zuständig für die Datenbankpflege und die Bearbeitung von Datenbankabfragen.

Diese Abfragen kommen von Seiten des Endnutzers über den Access Prozess (Zugang) herein, beispielsweise über einen Katalog. Der Access Prozess ist also die kommunikative Schnittstelle zwischen Archiv und Nutzer und betreibt daher Nutzerverwaltung, Zugriffsverwaltung (z.B. bei geschützten Dokumenten), sowie die Koordination der Dienstleistung. Dazu zählt neben der Abwicklung der Suchanfragen mit Hilfe des Data Management auch die Zusammenstellung und Auslieferung der Dissemination Information Packages (DIP, Auslieferungsinformationspaket) mit Hilfe des Archival Storage. Die Zusammensetzung der DIPs muss bei jeder Anfrage überprüft und ein entsprechendes DIP zusammengestellt werden. Zwei Nutzer erhalten beispielsweise verschiedene DIPs des gleichen digitalen Objekts, wenn sie unterschiedliche Nutzungsrechte dafür besitzen.[26]

Über all diesen Prozessen steht der Administrationsprozess. Er ist für den Betrieb des gesamten Archivs verantwortlich und kümmert sich etwa um vertragliche Bedingungen mit Produzenten, um die Formulierung und Einhaltung der Archivstandards (und ggf. Veranlassungen von Änderungen), um die Repräsentation nach außen und zum Träger durch Berichte und um die Migration und Aktualisierung der Archivinhalte.

Unterstützt wird die Administration durch den Prozess des Preservation Planning (Erhaltungsplanung). Dieser berät die Administration auf Grundlage ihrer Evaluierungen und Beobachtungen des eigenen Archivs, der Entwicklungen im Bereich der digitalen Langzeitarchivierung und

[25] Vgl. Brübach, N. (2010): Die Überarbeitung und Ergänzung des OAIS. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 4.14 f.

[26] Vgl. Brübach, N. (2010): Das Referenzmodell OAIS. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S.4.7

des technischen Kontextes und Fortschritts. Auf dieser Basis empfiehlt er der Administration die Veranlassung von Aktualisierungen, Migrationen, Emulationen, Entwicklungen neuer Standards und Leitlinien und sonstigen Änderungen im Archiv.

2.3.3 Repositorien

Repositorien sind Publikations- bzw. Dokumentenserver, welche hauptsächlich an Universitäten, wissenschaftlichen Bibliotheken oder Forschungseinrichtungen betrieben werden und durch die wissenschaftliche Materialien archiviert und zugänglich gemacht werden.[27] Sie können Teil von Open-Access-Strategien und daher für die Nutzer kostenfrei sein.

Es gibt fachliche Repositorien, welche institutionsübergreifend betrieben werden und sich auf bestimmte Themengebiete konzentrieren, und institutionelle Repositorien, welche von einer Einrichtung, wie etwa einer Universität, betrieben werden und multidisziplinär aufgestellt sind. Allerdings lassen sich sämtliche Publikationsserver noch durch ihre Granularität in der Auswahl der veröffentlichten Publikationen unterscheiden. Manche Repositorien veröffentlichen nur Abschlussarbeiten, wohingegen andere auch noch unveröffentlichte und ungeprüfte Zeitschriftenaufsätze (Preprints), Lehrmaterialien oder Arbeitspapiere bereitstellen.

Es gibt mittlerweile einige als Open Source Software verfügbare Softwarepakete zum Aufbau und Betrieb eines Repositoriums (Repository Systeme), die damit den Open Access Gedanken weiter unterstützen. Sie bieten verschiedene Gewichtungen bezüglich ihrer Funktionen. Diese reichen von Unterstützung der Produzenten durch Ingest-Schnittstellen über Kontextdaten, Nutzungsanalyse, Unterstützung von LZA bis hin zu Retrievalfähigkeiten und Suchmechanismen.[28]

Mittlerweile gibt es diverse Angebote zur Metasuche in Repositorien, die allesamt über das „Open Archives Initiative – Protocoll for Metadata Harvesting“ (OAI-PMH) funktionieren. Das Protokoll basiert auf HTTP (Hypertext Transfer Protocol) und XML (Extensible Markup Language) und kann daher plattformunabhängig von einem OAI-Data-Provider Metadaten abfragen und übertragen. Auf OAI-PMH wird später detaillierter eingegangen.

Für die Qualitätssicherung deutscher Repositorien setzt sich die Deutschen Initiative für Netzwerkinformation e.V. (DINI) ein. Das von ihr entwickelte „DINI-Zertifikat Open Access Repositorien und Publikationsdienste“ [29] [30] trägt durch seine standardisierten Mindestanforderungen als Bewertungsgrundlage zur Qualitätssicherung von deutschen Repositorien bei. Dazu zählt neben der bereits angesprochenen Unterstützung von Dublin Core auch die Einrichtung einer OAI-PMH-Schnittstelle. Zudem verpflichten sich die Repositorien durch das Zertifikat auch beispielsweise dazu, die Dokumente mindestens fünf Jahre vorzuhalten.[31] Die DINI selbst emp-

[27] Vgl. Informationsplattform Open Access (2013): Repositorien

[28] Vgl. Aschenbrenner, A. (2010): Repository Systeme. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S.11.3-11.6

[29] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2014): DINI-Zertifikat

[30] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2013): DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste 2013

[31] Vgl. ebd., S.30 f.

fiehlt, die Langzeitverfügbarkeit sicherzustellen und dazu passende Formate wie PDF, ODT oder TXT zu verwenden.

2.3.4 Dateiformate

Das OAIS Modell gibt keine Spezifikation des Typs des zu archivierenden, digitalen Objekts vor. Prinzipiell kann jeder Objekttyp archiviert werden (Bildarchive, Musikarchive, Videoarchive, etc.). Da es in dieser Arbeit um wissenschaftliche Arbeiten geht, wird nur auf Textformate näher eingegangen. Generell sind Dateiformate auf ihre Archivfähigkeit zu untersuchen, um sich auf Standards einigen zu können. Dabei müssen bestimmte Faktoren beachtet werden.[32] [33] Wegen der schieren Anzahl an Formaten werden Dateiformatregister (File Format Registries, z.B. PRONOM) geführt, welche verschiedene Dateiformate identifizieren und charakterisieren.[34]

2.3.4.1 Auswahlkriterien

Das Format sollte kein proprietäres sein, da hierbei eine Abhängigkeit zum betreffenden Unternehmen bezüglich der Weiterentwicklung entstehen kann. Wichtig ist vor allem, dass die Spezifikationen des Dateiformats dokumentiert und idealerweise frei zugänglich sind. So können bei offenen Formaten durch Einsicht der Formatspezifikationen selbst Weiterentwicklungen initiiert werden, wie etwa durch das Schreiben eines Leseprogramms, das auch auf dem neusten Betriebssystem funktioniert, sodass Informationen weiterhin aus Dateien ausgelesen werden können.

Das Format sollte zudem möglichst verbreitet sein, was auch eine geringe Abhängigkeit von Hard- und Software einschließt, und eine geringe Komplexität in Aufbau und Umgang aufweisen, da mit der Komplexität das benötigte Wissen zum Auslesen der Datei steigt. Dieses Wissen muss selbst wiederum erhalten bleiben. Ein LZA-geeignetes Format sollte zudem möglichst robust, d.h. widerstandsfähig gegenüber versehentlichen Veränderungen sein. So sollte die Korrumpierung einzelner Bits nicht die gesamte Darstellung des Dokuments verhindern, sondern beispielsweise nur zu einem kleinen Darstellungsfehler führen, um den ggf. unvermeidbaren Schaden zumindest zu begrenzen.

Schutzmechanismen wie Kopierschutz oder Zugriffskontrollen per Passwort sind unerwünscht, da sie – jedenfalls in Deutschland – von Rechts wegen her nicht ohne weiteres umgangen werden dürfen[35]. Solche Kryptofunktionen schränken nicht nur die Verwendung des Dokuments ein, sondern sie verhindern auch die unkomplizierte Migration oder Konvertierung des Dokuments in ein benötigtes Dateiformat. Zumindest werden diese Vorgänge durch den Umstand des Einholens der Erlaubnis seitens des Urhebers mit einem nicht zu rechtfertigenden Mehraufwand ver-

[32] Vgl. Berthold, H. (2014): Die Eignung des Dateiformates PDF für die Langzeitarchivierung, S.4

[33] Vgl. Ludwig, J. (2010): Auswahlkriterien. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.10 f.

[34] Vgl. Aschenbrenner, A.; Wollschläger, T. (2010): File Format Registries. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.19 ff.

[35] Vgl. Upmeier, A. (2010): Rechtliche Aspekte. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S.16.6 f.

sehen. Das Dateiformat muss sich zudem selbst gut dokumentieren und beschreiben, d.h. Metadaten mitliefern, sodass seitens des Archivs nicht auf externe Quellen zurückgegriffen werden muss.

Die Speicherung von Textdokumenten in Bildformaten empfiehlt sich nur dann, wenn das äußere Erscheinungsbild wesentlich relevant und ebenfalls Gegenstand der Archivierung ist.[36] Dies ist bei wissenschaftlichen Arbeiten aber in der Regel nicht der Fall. Hier liegt das Hauptaugenmerk auf der Archivierung der Inhalte und der Textinformation, also Struktur des Dokuments, sowie Zeichensatz und ggf. Schriftart. Idealerweise handelt es sich beim Zeichensatz um ein auf Unicode basierendes Format[37] wie etwa UTF-8 (8-Bit Universal Character Set Transformation Format). Damit lassen sich sämtliche Schrift- und Sonderzeichen (z.B. Umlaute) darstellen.

Damit auch die optische Darstellung des Archivobjekts authentisch bleibt, ist es notwendig, die verwendete Schriftart (auch Font genannt) abbilden zu können. Schriftarten gehören meist zum Betriebssystem, weshalb es vorkommen kann, dass eine Schriftart, die auf einem bestimmten Betriebssystem zur Formatierung des Textes ausgewählt wurde oder schlichtweg eher ausgefallen ist, nicht auf einem anderen Betriebssystem dargestellt werden kann.

2.3.4.2 TXT und ODF

Die bereits erwähnten Formate TXT und ODF sind für die Langzeitarchivierung nur ansatzweise geeignet. Das TXT-Format erfüllt zwar einige der geforderten Eigenschaften, ist allerdings in seiner Funktionalität sehr stark eingeschränkt. So gibt es weder die Möglichkeit zu Formatierungen (z.B. Linksbündiger Text, Blocksatz, 1. Überschrift, 2. Überschrift, etc.), noch können Grafiken eingefügt werden.

Das Open Document Format hingegen bietet mehr Möglichkeiten. Allerdings ist dies nicht ein Format, sondern eine Sammlung. Für Textdateien gibt es etwa ODT-Format (Open Document Text), für Präsentationen, Vektorgrafiken und Tabellen wiederum jeweils eigene Formate. Bis auf Grafiken lassen sich aber all diese Dateitypen auch als problemlos PDF- bzw PDF/A-Dateien darstellen, was die Verwendung von ODF zu Archivzwecken unnötig kompliziert erscheinen lässt. Zudem ist die technische Struktur von ODF zu komplex und unbeständig, sodass verschiedene Programme eine Datei unterschiedlich darstellen.

2.3.4.3 PDF

PDF steht für Portable Document Format und ist ein Austauschformat für Textdateien. Es ermöglicht die Beschreibung von Seiten anhand deren Inhalt, Struktur und Metadaten. Die erste Version erschien 1993 von der Softwarefirma Adobe. Bereits die 2001 erschienene Variante PDF/X-1a wurde zu einem ISO-Standard für Textaustauschformate. Später wurden die verschiedenen Varianten des PDF/A (PDF for Archive) Formats ebenfalls von der ISO als Stan-

[36] Vgl. Huth, K. (2010): Textdokumente. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 17.4

[37] Vgl. ebd., S. 17.6

dards für die LZA anerkannt. Die aktuelle PDF Version ist 1.7[38], eine Version 2.0 ist gegenwärtig in Arbeit[39].

Generell ermöglicht PDF – im Gegensatz zu ODF und TXT – durch die Verwendung geeigneter Leseprogramme verschiedene Skalierungen und Auflösungen, sowie die gezielte Ansteuerung bestimmter Seiten. Zudem können Elemente, wie Graphiken, Audio, Video, 3D-Animationen oder schlicht Schriftarten, die nicht standardmäßig auf jedem Rechner vorhanden sind, in das Dokument eingebettet werden, sodass die Darstellung korrekt bleibt.[40] Dies ist auch für Sonderzeichen in mathematischen oder sonstigen Formeln nützlich. Fernerhin gibt es für jede Softwareplattform mindestens ein kostenfreies Programm, das PDF-Dateien anzeigen kann.

Daneben bietet das Format eine Volltextsuche, sowie die Möglichkeit, Text zu extrahieren. Zusätzlich können auch Bilder verarbeitet werden. Wenn dies Bilder von Texten sind, können diese mit externer Software für Optical Character Recognition (OCR) analysiert und zu kopierbarem Text gewandelt werden.[41]

Problematisch sind allerdings die Möglichkeiten zur Verschlüsselung und Einschränkung der Nutzung durch sog. Kryptofunktionen. PDF-Dateien können codiert, mit Passwörtern oder mit weiteren Maßnahmen des Digital Rights Managements versehen werden. Dies können etwa Druckeinschränkungen oder Kopierschutz für Textausschnitte bzw. ganze Dateien sein. Durch die Umstellung auf PDF/A kann die dahingehende Überprüfung und ggf. Modifizierung von PDF-Dateien entfallen.

2.3.4.4 PDF/A

PDF/A steht für PDF for Archive und ist eine Produktfamilie von PDF. Sie ist der ISO Standard für die LZA von Textdokumenten[42], der in internationaler Kooperation von Unternehmen, Behörden und Experten aus dem Bibliotheks- und Archivbereich entwickelt wurde.[43] Dementsprechend wird PDF/A den oben genannten Anforderungen wie Plattformunabhängigkeit, Robustheit, etc. gerecht. Es wird nicht nur im bibliothekarisch-dokumentarischen Bereich eingesetzt, sondern findet auch im Rechnungswesen oder in der Industrie Anwendung.[44] Über PDF/A informiert vor allem das 2006 gegründete PDF/A Competence Center[45], welches Teil der PDF Association[46] ist.

---

[38] Vgl. International Organization for Standardization (2014): ISO 32000-1:2008 - Document management

[39] Vgl. International Organization for Standardization (2014): ISO/DIS 32000-2 - Document management

[40] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 6

[41] Vgl. Berthold, H. (2014): Die Eignung des Dateiformates PDF für die Langzeitarchivierung, S.8

[42] Vgl. Brübach, N. (2010): Das Referenzmodell OAIS. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S.4.4

[43] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 7

[44] Vgl. Ergül, A.; Böhm, A.; Schmidt, E.; Hissen, S.; Sariklis, T. (2012): Erfolgsfaktoren für die Durchsetzung von PDF/A als weltweiter Standard für elektronische Langzeitarchivierung, S. 363 f.

[45] Vgl. Association for Digital Document Standards e.V. (o.J.): The PDF/A Competence Center

[46] Vgl. Association for Digital Document Standards e.V. (o.J.): PDF Association

PDF/A unterscheidet sich von normalem PDF dadurch, dass es einige Maßnahmen nicht erlaubt, die der LZA nicht zuträglich sind. Dazu zählen beispielsweise die Verwendung von Kryptofunktionen oder das Einbetten von Elementen, die auf externe Software zurückgreifen müssen, wie etwa Animationen.[47] Zudem stellt der Standard Anforderungen wie die Angabe von Metadaten im XMP-Format (Extended Metadata Platform)[48] oder Farbangaben als ICC Profil (International Color Consortium); auch die Seitengröße wird limitiert. Es gibt verschiedene Versionen[49] von PDF/A, die sich allerdings ergänzen und nicht einander ersetzen. Die Wahl des bevorzugten Formats obliegt dem jeweiligen Archiv auf Grundlage seiner Rahmenbedingungen.

PDF/A-1[50] aus dem Jahr 2005 basiert auf PDF 1.4 und verlangt – wie bereits erwähnt – Farbangaben als ICC und Metadaten als XMP Elemente, sowie die Einbettung von Graphiken, Schriftarten, etc. in das Dokument. Verboten sind Zugriffsbeschränkungen durch Passwörter, transparente Elemente (durchsichtige Markierungen oder Bilder), bestimmte Kompressionsarten (u.a. JPEG2000), PDF-Ebenen (z.B. eine für den Hintergrund, eine für den Inhalt, eine für den Rahmen, etc.) und interaktive Elemente, wie etwa Aktionen mit JavaScript. Digitale Signaturen, die für rechtlich bindende Dokumente wichtig sein können, und das Benutzen von Hyperlinks werden unterstützt.

PDF/A-2[51] aus dem Jahr 2011 basiert auf PDF 1.7 und erlaubt im Gegensatz zur ersten Version die Kompression mit JPEG2000, Ebenen und transparente Elemente. Hinzu kommen neue Schriftartformate, eine Erweiterung der Signaturfunktion, sowie eine Containerfunktion. Diese erlaubt es, andere PDF/A-Dateien in eine PDF/A-2-Datei einzubetten.

PDF/A-3[52] aus dem Jahr 2012 ergänzt die Einbettungsfunktion seines Vorgängers. Nun können beliebige Dateien eingebettet werden, wie beispielsweise XML-Dateien oder die Textverarbeitungsdatei, mit der das Dokument erstellt wurde. Wie Archive mit dieser Datei wiederum umgehen sollen, ist nicht beschrieben.

Erweitert wird die Klassifizierung durch drei Konformitätsstufen.[53] Stufe A (Accessible) ist erfüllt, wenn alle Anforderungen des Standards erfüllt sind und die inhaltliche Struktur und Leseabfolge des Inhalts adäquat abgebildet werden. Zudem muss der Text durchsuchbar sein aus dem Dokument herauskopiert werden können. Stufe B (Basic) verlangt statt der strukturellen die visuelle Reproduzierbarkeit der Inhalte. Eine Variante davon ist Stufe U (Unicode), die verlangt, dass der gesamte Text im Schriftzeichenstandard Unicode vorhanden ist.

---

[47] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 6

[48] Vgl. Rosenthol, L. (2011): PDF/A Metadaten XMP, RDF & Dublin Core

[49] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 8

[50] Vgl. International Organization for Standardization (2005): ISO 19005-1:2005 - Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)

[51] Vgl. International Organization for Standardization (2011): ISO 19005-2:2011 - Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2)

[52] Vgl. International Organization for Standardization (2012): ISO 19005-3:2012 - Document management -- Electronic document file format for long-term preservation -- Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)

[53] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 8

3 Untersuchung

3.1 Repositorien

3.1.1 Auswahlkriterien der Repositorien

Da im Rahmen dieser Arbeit nur begrenzt Repositorien bearbeitet werden können, wird anhand der folgenden Auswahlkriterien eingeschränkt. Auf die manuelle Suche und Überprüfung von OAI-PMH-fähigen Repositorien etwa anhand des DINI-OAI-Validators[54] wegen des unangemessenen Aufwandes verzichtet.

Die erste Einschränkung ist die geographische Einordnung des Repositoriums bzw. seiner tragenden Institution. Diese Arbeit wird sich ausschließlich mit deutschen Repositorien befassen. Zwingend notwendig ist natürlich eine OAI-PMH-Schnittstelle, über die problemlos mit dem Server kommuniziert werden kann. Zusätzlich sollten die Repositorien das Harvesting der PDF-Dateien ohne ausdrückliche Einwilligung erlauben. Repositorien, die das DINI-Zertifikat „Open-Access-Repositorien und Publikationsdienste“ erworben haben, erfüllen all diese Kriterien bereits. Daher wird auf die entsprechende Liste[55] bei der DINI zurückgegriffen.

Entscheidend ist auch eine möglichst lange Laufzeit, um einen historischen Verlauf erkennen zu können. Daher werden Repositorien in die Untersuchung einbezogen, welche möglichst früh das DINI-Zertifikat erworben haben. Die frühsten Vergaben des Dini-Zertifikats sind auf das Jahr 2004 datiert. Wegen diverser technischer Probleme konnte nicht ausschließlich auf die im Jahr 2004 zertifizierten Repositorien zurückgegriffen werden. Die individuelle Lösung erforderte einen unverhältnismäßigen zeitlichen Aufwand.

3.1.2 Richtlinien und Hilfestellungen zur Langzeitarchivierung

Durch den Erwerb des DINI-Zertifikates verpflichten sich die Repositorien bereits zur Einhaltung der entsprechenden Richtlinien. Diese Richtlinien, auch als Leitlinien oder Policies bezeichnet, umfassen u.a. die Garantieerklärung für die Mindestdauer der Verfügbarkeit der Dokumente. Diese muss mindestens fünf Jahre betragen, kann aber ansonsten von jedem Repositorium selbst gesetzt werden.[56] Langzeitverfügbarkeit wird von der DINI empfohlen.

Entsprechend der Richtlinien stellen die tragenden Einrichtungen der Repositorien Anforderungen an die eingereichten Dokumente. Die Sicherstellung, dass veröffentlichte Dateien in passenden Formaten vorliegen kann auf unterschiedliche Weise erfolgen. Einerseits kann der Betreiber des Repositoriums, beispielsweise die Bibliothek, einfach alle eingereichten Dateien an-

[54] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2014): DINI Check - OAI Validator

[55] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2014): Liste der Repositorien

[56] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2013): DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste 2013, S. 14

nehmen und selbst auf entsprechende Konformität prüfen. Andererseits kann die Bibliothek versuchen sicherzustellen, dass die Autoren ihre Werke bereits in den richtigen Formaten abliefern. Dafür sind Hilfestellungen wie Schulungen und Leitfäden zur LZA und Tools zur Validierung und Konvertierung von PDF- bzw. PDF/A-Dateien denkbar.

Fernerhin wird auch der Ansatz verfolgt, von den Autoren die Originaldatei zu verlangen. Dabei werden XML-basierte Formate wie ODF bevorzugt. Aus diesem Archivformat kann noch im Nachhinein eine Datei im Präsentationsformat PDF bzw. PDF/A erzeugt werden.

3.2 Harvesting

Harvesting (engl. „to harvest“, dt. “ernten”) bezeichnet Verfahren, die automatisiert größere Datenmengen von Webservern extrahieren und zusammentragen. Es können verschiedene Arten von Daten und Dateien auf diese Weise angehäuft werden; bevorzugt wird der Begriff aber im Kontext von Metadaten verwendet.[57]

Das Harvesting von Metadaten wurde vor dem Hintergrund der schlechten Auffindbarkeit von elektronischen Dokumenten wie Preprints und wissenschaftlichen Arbeiten durch deren Zerstreuung auf verschiedene institutionelle Repositorien zu einem wichtigen Thema. Um das Auffinden zu erleichtern sollten die Repositorien die Metadaten ihrer Dokumente ausstellen, sodass diese aggregiert und für Dienstleistungen wie etwa auf Z39.50 basierende Metasuchen verwendet werden konnten.[58]

Dazu wurden verschiedene Ansätze entwickelt und um die Jahrtausendwende beim sog. „Santa Fe Meeting“ diskutiert. Die mangelnde Interoperabilität der einzelnen Ansätze führte letztlich zur Gründung der Open Archives Initiative (OAI) und der Entwicklung eines übergreifenden Protokolls für das Harvesting von Metadaten, dem „Open Archives Initiative Protocol für Metadata Harvesting“ (OAI-PMH).

OAI-PMH kann sowohl für “Incremental Harvesting”, als auch “Selective Harvesting” eingesetzt und Repositories damit also entweder komplett oder gezielt durchsucht werden. Letzteres kann nach Zeitstempel und/oder Zugehörigkeit zu einer bestimmten Gruppe („Set“) der Metadatensätze erfolgen.[59]

3.2.1 OAI-PMH

Die Version 1.0 des OAI-PMH wurde im Jahr 2000 veröffentlicht und später auf 1.1 ergänzt; zwei Jahre später folgte die stabile Version 2.0, die seitdem betreut wird[60]. Das Protokoll ist wegen seiner Simplizität und Interoperabilität international verbreitet. Die meisten Repositorysysteme wie OPUS oder EPrints, aber auch einige andere Programme wie das Resource Dis-

[57] Vgl. Seminar für Sprachwissenschaft (2012): Glossar

[58] Vgl. Carpenter, L.; Heery, R. (2003): 2. History and development of OAI-PMH

[59] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2008): The Open Archives Initiative Protocol for Metadata Harvesting

[60] Vgl. Shreeves, S. L. (2006): Search Interoperability, OAI, and Metadata, S. 18

covery System VuFind oder das Open Journal System bringen standardmäßig bereits Schnittstellen für OAI-PMH mit. So können nicht nur Dokumentenserver von Forschungseinrichtungen und Archiven, sondern auch Kataloge und Online-Zeitschriften ihre Metadaten bereitstellen bzw. die Metadaten anderer einbinden.

OAI-PMH unterscheidet zwischen Datenanbietern (Data Provider) und Dienstleistungsanbietern (Service Provider).[61] Erstere sind die Inhaber der Repositorien, auf denen die digitalen Objekte und deren Metadaten liegen. Die Datenanbieter stellen ihrem Namen entsprechend die Metadaten bereit, sodass Dienstleistungsanbieter diese nutzen können. Ein Dienstleistungsanbieter verfügt über einen Harvester, also eine Client-Anwendung, die OAI-PMH handhabt. So kann der Dienstleistungsanbieter die bereitgestellten Daten auslesen und daraus eine Dienstleistung, beispielsweise eine Metasuche für ein bestimmtes Fachgebiet, erstellen.

Die Bielefeld Academic Search Engine (BASE) ist ein Beispiel für einen solchen Dienstleistungsanbieter[62], der eine gemeinsame Suchoberfläche für verschiedene Datenanbieter[63] bereitstellt. Diese erstrecken sich von WikiBooks über Universitätsbibliotheken bis hin zu Staats- und Landesbibliotheken.

Die Kommunikation zwischen Daten- und Dienstleistungsanbietern basiert auf HTTP und XML.[64] So wird eine standardisierte Kommunikation gewährleistet. Grundsätzlich lassen sich OAI-PMH anfragen also auch manuell in einen Browser eingeben, solange die Basis-URL (Base-URL, die Adresse des Repositoriums) bekannt ist und die Syntax von OAI-PMH beachtet wird. Die Verwendung von UTF-8 zur Darstellung von Schriftzeichen ist verpflichtend. Eine OAI-PMH Antwort muss einem bestimmten XML-Schema entsprechen, um gültig zu sein.[65]

3.2.1.1 Begriffe

Um Funktionsweise und Handhabung des Protokolls zu verstehen, müssen zunächst einige Begriffe voneinander abgegrenzt werden.[66]

„Resource“ meint das eigentliche, ursprüngliche Objekt, welches erschlossen werden muss. Dabei ist zunächst unerheblich, ob es ein physisches Objekt (z.B. Michelangelos „David“) oder ein digitales Objekt (z.B. eine wissenschaftliche Arbeit im PDF/A-Format) ist.

Zu dieser Ressource gibt es beim Repository einen Datensatz, das sog. Item. Dieses enthält die Metadaten der Ressource, wie beispielsweise Titel, Thema, Autor, Erstellungsdatum oder Dateiformat der wissenschaftlichen Arbeit. Des Weiteren besitzt das Item einen für das Reposi-

---

[61] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2008): The Open Archives Initiative Protocol for Metadata Harvesting

[62] Vgl. Universitätsbibliothek Bielefeld (2014): Suchmaschine BASE - Bielefeld Academic Search Engine | Über BASE

[63] Vgl. Universitätsbibliothek Bielefeld (2014): Suchmaschine BASE - Bielefeld Academic Search Engine | Die Quellen

[64] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2008): The Open Archives Initiative Protocol for Metadata Harvesting

[65] Vgl. Open Archives Initiative (2008): OAI-PMH XML Schema

[66] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2008): The Open Archives Initiative Protocol for Metadata Harvesting

torium eindeutigen OAI-PMH Identifier. Der Identifier bezeichnet nicht die Ressource, sondern das Item. Er muss der URI-Syntax (Uniform Resource Identifier) entsprechen.

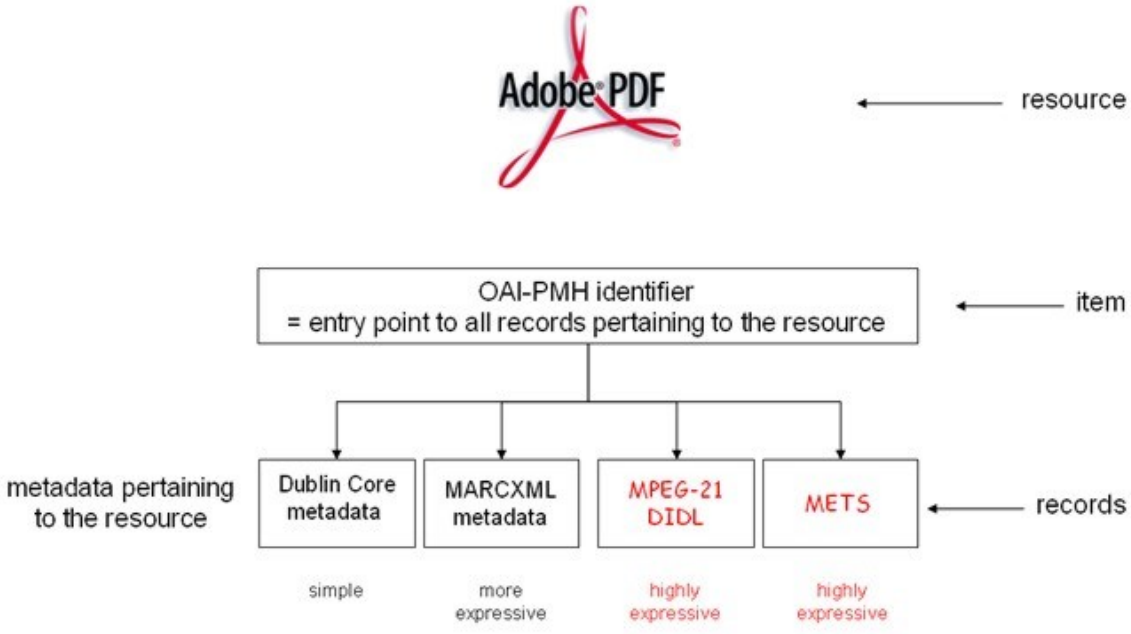

Abbildung 2: OAI-PMH Datenmodell[67]

Vom Item wiederum kann es verschiedene Abzüge geben. Diese werden Records genannt und beziehen sich anhand des oben genannten Identifiers auf das richtige Item. Records können in unterschiedlich komplexen und ausdrucksstarken Formaten erzeugt werden, basieren aber alle auf XML. Kleinster gemeinsamer Nenner dafür ist das Dublin Core Metadata Set, welches von jedem Data Provider bereitgestellt bzw. unterstützt werden muss.[68] Für eine wissenschaftliche Arbeit würde sich beispielsweise auch MARCXML anbieten. Die verschiedenen Metadatenformate unterscheiden sich in ihrer Tiefe bzw. Granularität und in ihrer fachlichen Ausrichtung. Die ausgegebenen Records werden entweder im Item gespeichert oder dynamisch bei der Anfrage generiert. Ein Record kann durch die Kombination von Item-Identifier, Metadatenpräfix (also dem Metadatenformat) und dem Zeitstempel des Versendens an den Harvester eindeutig identifiziert werden.

Records sind die Antworten, die ein Harvester auf eine GetRecord- oder ListRecord-Anfrage (s.u.) bekommt, und folgen einem bestimmten XML-Schema. Sie beginnen mit einem Kopf, dem Header, der neben dem Identifier und dem Zeitstempel der letzten Änderung am jeweiligen Record auch Angaben zur Zugehörigkeit des Items zu einem oder mehreren Sets (s.u.) enthält. Optional kann im Header auch angegeben werden, dass das gewünschte Metadatenformat mittlerweile nicht mehr unterstützt wird, sollte dies der Fall sein. Darauf folgen die eigentlichen Metadaten. Zunächst wird mit einem Kürzel, dem Metadatenpräfix, angegeben, um welches

[67] Vgl. Van de Sompel, Herbert; Nelson, M. L.; Lagoze, C.; Warner, S. (2004): Resource Harvesting within the OAI-PMH Framework

[68] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2008): The Open Archives Initiative Protocol for Metadata Harvesting

Format es sich handelt, beispielsweise „oai_dc“ für Dublin Core. Fernerhin werden der Namensraum des Formats und das zu Grunde liegende XML-Schema des Records durch URLs angegeben. Je nach Metadatenformat sehen die darauf folgenden Tags mit den Feldern für Autor, Titel, etc. anders aus. Zuletzt können ein oder mehrere About-Container folgen. Auch diese müssen einem XML-Schema entsprechen und enthalten Informationen über die Metadaten wie rechtliche Hinweise oder Angaben zur Herkunft.

Die bereits erwähnten Sets sind eine Möglichkeit des Data Providers, die Items in Gruppen zusammenzufassen. Dies ist besonders für das selektive Harvesting nützlich. Die Gestaltung der Sets bleibt dem Data Provider überlassen. Möglich sind einfache Listen von z.B. Dokumenttypen (Preprint, Abschlussarbeit, Zeitschriftenaufsatz, etc.) bis hin zu hierarchischen Strukturen von Sets und Subsets zur inhaltlichen Klassifizierung und Erschließung. Dazu gibt es den Parameter „setSpec“ für jeden Knotenpunkt in der hierarchischen Struktur. Dieser Parameter gibt den Pfad von der obersten Ebene des Sets bis zur aktuellen an. Dies wird mittels für das Repository eindeutigen Setidentifiern gehandhabt. Fernerhin hat jedes Set oder Subset einen menschenlesbaren Namen und kann eine Beschreibung erhalten.

3.2.1.2 Anfragen

OAI-PMH bietet sechs verschiedene Typen von Anfragen.[69] Auf diese wird auch als Verb oder Befehl Bezug genommen. Fast alle dieser Anfragen können mit bestimmten Parametern (Arguments) modifiziert werden. Im Folgenden ein Überblick über die sechs Anfragetypen. Jeweils drei dienen dazu, Informationen über das Repository zu erlangen bzw. Metadaten zu extrahieren. Beispielanfragen sind jeweils in den Fußnoten hinterlegt. [70]

Über die Identify-Anfrage kann der Harvester Informationen über das Repository erhalten. Zwingend anzugebende Informationen seitens des Repositoriums sind sein Name, seine Base-URL, die von ihm unterstützte Version von OAI-PMH, der älteste Zeitpunkt von Änderungen im Repositorium, eine Angabe darüber ob und wie die Löschung von Records dokumentiert sind, die zeitliche Granularität, mit der das Repositorium durchsucht werden kann (z.B. auf die Stunde genau oder auf den Tag genau) und mindestens eine Administrator-E-Mail-Adresse für den Fall, dass größere Probleme oder Fragen auftreten. Weitere Informationen sind optional.[71]

Anhand der Anfrage ListMetadataFormats lässt sich herausfinden, welche Metadatenformate das Repository, inklusive des obligatorischen Dublin Core, beherrscht. Die Anfrage kann mit einem Identifier-Parameter verknüpft und damit nur auf ein bestimmtes Item bezogen werden.[72]

Der Befehl ListSets gibt die Set-Hierarchie des Repositorys aus, sofern eine solche vorhanden ist. Andernfalls wird als Antwort der Fehler „noSetHierarchy“ ausgegeben.[73]

---

[69] Vgl. ebd.

[70] Aus Gründen der Übersichtlichkeit wird auf die XML-Antworten bei den jeweiligen Beispielen verzichtet. Da die Anfragen aber auf HTTP basieren, fungieren sie als Hyperlink zu den jeweiligen Antworten, die dann im Browser angesehen werden können.

[71] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=Identify

[72] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=ListMetadataFormats

[73] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=ListSets

Mit dem Befehl GetRecord kann ein einzelner Abzug eines Items angefordert werden. Dazu sind der Identifier des Items, sowie das gewünschte Datenformat als Metadatenpräfix zwingend als Parameter anzugeben.[74]

ListRecords fordert gleich mehrere Records an. Auch hier ist das Metadatenformat ein erforderlicher Parameter. Mit diesem Befehl kann das gesamte Repository „abgeerntet“ werden. Optional können aber auch einengende Parameter eingefügt werden. So lässt sich selektives Harvesting anhand von Zeitstempel (von / bis) oder Zugehörigkeit zu bestimmten Sets betreiben.[75]

Im Gegensatz zu ListRecords liefert die Anfrage ListIdentifiers nur die Header der Records und damit lediglich die Informationen von Identifier, Zeitstempel und Setzugehörigkeit eines Items. Auch bei diesem Befehl ist das Metadatenformat anzugeben. Die Verwendung von Zeitstempel und Setzugehörigkeit sind ebenfalls möglich.[76]

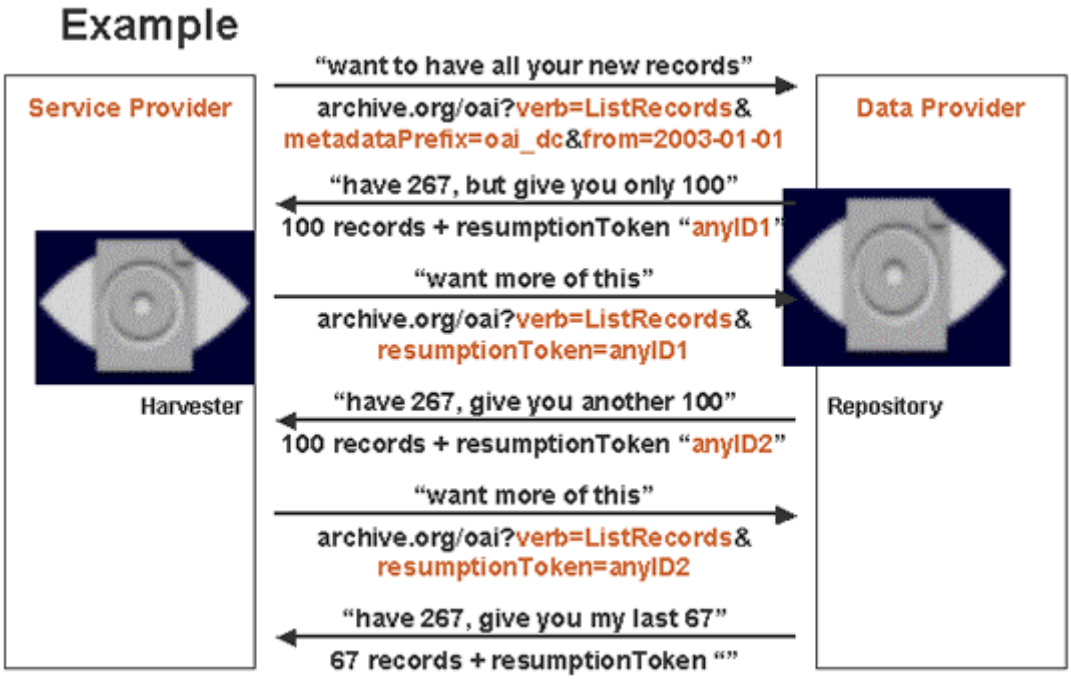

Abbildung 3: Flow Control in OAI-PMH[77]

Bei den Anfragen listSets, listIdentifiers und listRecords kann es sein, dass das Repository mit sehr großen Listen antworten muss und diese aus Gründen des Rechenaufwands zunächst nur unvollständig ausgibt. Der Harvester hat dann die Möglichkeit den Rest der Antwort anzufragen. Dieses Verfahren wird auch als Flow Control bezeichnet.

[74] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=GetRecord&identifier=oai:elib.uni-stuttgart.de-opus:1773&metadataPrefix=oai_dc

[75] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=ListRecords&metadataPrefix=oai_dc&set=ddc:020

[76] Beispiel: http://elib.uni-stuttgart.de/opus/oai2/oai2.php?verb=ListRecords&metadataPrefix=oai_dc

[77] Vgl. Carpenter, L.; Heery, R. (2003): 3. Main Technical Ideas of OAI-PMH

In OAI-PMH läuft die Flow Control anhand eines „Resumption Token“ ab. Dieses wird der ausgegebenen, unvollständigen Antwort als Identifier angefügt. Der Harvester schickt bei der Anfrage nach dem Rest der Liste das Resumption Token mit, sodass das Repositorium die Anfrage wiedererkennt und die richtigen zusätzlichen Daten übertragen kann. Liefert das Repository den letzten Teil der Liste, so ist der Wert des Resumption Token leer und der die Übertragung damit beendet. Die Angabe der Gesamtgröße der Liste ist Pflicht.

3.2.2 Verwendete Software und Verfahrensweise

Zu den grundlegenden Aspekten der Implementierung und des Betriebs von Harvestingsoftware gibt es einige frei verfügbare Erläuterungen und Anleitungen, wie etwa im Forum der OAI[78] oder in den Guidelines.[79] Die Programme für das Harvesting mit OAI-PMH sind u.A. auf der Homepage der Open Archives Initiative[80] aufgelistet:

Für diese Arbeit wurde der Net::OAI::Harvester[81] verwendet. Dieser ist eine Erweiterung der Programmiersprache Perl[82] und bietet für die Befehle listRecords und listIdentifiers jeweils eine Variante: listAllRecords und getAllIdentifiers verarbeiten automatisch die Resumption Tokens und liefern alle angeforderten Ergebnisse.[83] [84]

Net::OAI::Harvester wurde auf einem Amazon EC2-Server mit dem Betriebssystem Debian 7 installiert und betrieben, um die anfallenden Datenmengen abspeichern und verrechnen zu können.[85] Das Harvesting wurde über selbst verfasste Perl-Skripte abgewickelt und die gewonnenen Metadaten pro Repositorium über die Perlerweiterung „Excel::Writer::XLSX“[86] in einer Exceltabelle gespeichert.

Das Herunterladen der PDF-Dateien (Resource Harvesting) stellt diverse Anforderungen. So sind etwa die in den Metadaten bereitgestellten Identifier in ihrer Form nicht übergreifend spezifiziert, sodass es sich dabei um verschiedene Arten handeln kann, wie etwa eine DOI-Nummer, eine URL oder eine Zitation. Das DINI-Zertifikat verpflichtet zur Verwendung eines beständigen Identifikators (Persistent Identifier), welcher „in Form einer operablen URL“[87] anzugeben ist. Eine solche URL muss auch in den exportierten Metadaten eines Titels im Dublin Core Feld <dc:identifier> angegeben werden.

---

[78] Vgl. Carpenter, L.; Heery, R. (2003): OAI-PMH Online-Tutorial

[79] Vgl. Lagoze, C.; Van de Sompel, Herbert; Nelson, M.; Warner, S. (2005): Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting

[80] Vgl. Open Archives Initiative (o.J.): PMH-Tools

[81] Vgl. Summers, E.; Emmerich, M. (o.J.): Net::OAI::Harvester

[82] Vgl. Comprehensive Perl Archive Network (2013): Perl Source

[83] Vgl. Summers, E. (2004): Building OAI-PMH Harvesters With Net::OAI::Harvester

[84] Vgl. Summers, E.; Emmerich, M. (o.J.): Net::OAI::Harvester

[85] Prof. Magnus Pfeffer hat sich freundlicherweise bereiterklärt, für diese Arbeit einen bereits erworbenen EC2-Server der Hochschule der Medien bereitzustellen.

[86] Vgl. McNamara, J. (2014): Excel::Writer::XLSX

[87] Vgl. Deutsche Initiative für Netzwerkinformation e. V. (2013): DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste 2013, S. 24

Eine vorhandene URL bedeutet damit aber nicht automatisch einen direkten Link zur PDF-Datei des Dokumentes.[88] Der Link kann auch lediglich zu einer Website mit den Titelinformationen des Dokuments führen, welche wiederum den Link zur Datei enthält. Man nennt solche Seiten auch Landing Page, Splash Page oder Frontdoor. Die Bereitstellung eines direkten Links zum Volltext wird von der DINI empfohlen, ist aber nicht verpflichtend.[89]

Gibt es keine Konventionen über die Bereitstellung der Metadaten kann es auch vorkommen, dass die URL in den DC Feldern <dc:relation> oder <dc:format> (Beschreibung des Formats gefolgt von der URL) vorkommen. Expressivere Metadatenformate als Dublin Core beinhalten oft von vorn herein festgelegte Felder für die URL der Ressource, um diesem Problem zu entgehen.[90] Allerdings sind diese Formate nicht auf allen Repositorien präsent. Selbst wenn ein Repositorium ein oder mehrere expressive Metadatenformate (z.B. MARC-XML) unterstützt, ist damit nicht sichergestellt, dass auch von allen Items ein entsprechender Abzug darin verfügbar ist. Daher garantiert noch immer nur das Anfordern von Dublin Core Metadaten eine vollständige Liste von Datensätzen eines Repositoriums.

Mit Crawlern wie HTTrack[91] kann man diesen Problemen ausweichen, indem man ein Abbild der gesamten Webseite des Repositoriums macht. Allerdings wird dabei weitaus mehr als die benötigten PDF-Dokumente heruntergeladen. Diese müssen anschließend erst aus der Ordnerstruktur extrahiert werden. Deshalb und auf Grund der größeren heruntergeladenen Datenmenge wird auf diese Herangehensweise verzichtet.

Stattdessen wird mit Hilfe verschiedener Perl-Erweiterungen[92] ein eigener Crawler konstruiert und in das Skript integriert. Da durch das DINI-Zertifikat ein Identifikator in Form einer URL obligatorisch ist, setzt der Crawler an diesen an, statt etwa an einer URN (Uniform Resource Name). Sofern eine URN vorhanden ist enthält sie zwar auf jeden Fall einen Link zum Volltextdokument, müsste aber erst noch über einen Resolver aufgelöst werden.

Das Harvesting-Skript ist so konstruiert, dass es Dublin Core Metadatensätze aller Dokumente eines Repositoriums anfordert und aus dem Feld für Identifikatoren die Inhalte, die mit http bzw. https beginnen, also die operablen URLs extrahiert und an den Crawler weitergibt. Dieser überprüft, ob die URL auf „.pdf“ endet. Trifft dies zu, wird die PDF-Datei, die sich hinter der URL verbirgt heruntergeladen. Ist die URL kein Link zum Volltext, sondern zu einer Landing Page, wird diese Seite nach weiteren Links durchsucht. Diese Links werden gespeichert und wiederum daraufhin überprüft, ob sie auf „.pdf“ enden und gegebenenfalls heruntergeladen. Die Dateien werden anhand der URL aus dem Identifikatorenfeld benannt. Dabei müssen allerdings alle Sonderzeichen durch Unterstriche ersetzt werden, da viele andere Sonderzeichen zur Benennung von Dateien nicht zulässig sind. An den Namen wird eine Ziffer angehängt, falls sich auf einer Seite mit dem gleichen Identifier mehrere PDF-Dateien befinden. So wird verhindert, dass

[88] Vgl. ebd., S. 47

[89] Vgl. ebd., S. 50

[90] Vgl. Van de Sompel, Herbert; Nelson, M. L.; Lagoze, C.; Warner, S. (2004): Resource Harvesting within the OAI-PMH Framework

[91] Vgl. Roche, X. (2014): HTTrack Website Copier

[92] Vgl. Eblen, J. (2000): Web crawling in Perl

nur eine Datei heruntergeladen wird, welche dann die vorherige ersetzt. Die Ziffer beginnt für jeden Identifikator wieder bei eins.

Probleme ergaben sich allerdings bei Repository-Systemen, die die Identifier-URL per HTTP Redirect zu einer URL zum PDF-Volltext auflösen. Das bedeutet, dass ein Identifier, der nicht auf „.pdf“ endet beim Abruf sofort zu einer URL aufgelöst wird, die die PDF-Datei liefert, ergo auf „.pdf“ endet. Dies führte zu Fehlern beim Crawling, da das Perl-Skript nicht erkennen konnte, dass es bereits eine PDF-Datei vor sich hat, und daher versucht hat, diese nach Links zu PDF-Dateien abzusuchen.

3.3 Validierung

Die Validierung eines Formates führt einen Abgleich zwischen einer Datei und den formatspezifischen Eigenschaften durch. Sie ist damit von der Formaterkennung und der Metadatengewinnung abzugrenzen.[93] Die geernteten Dublin Core Metadaten liefern im Feld <dc:format> lediglich die Information „application/pdf“ entsprechend der MIME-Type Klassifizierung[94]. Das bedeutet, dass es sich bei der entsprechenden Ressource um eine nicht näher spezifizierte PDF-Datei handelt. Die Validierung der heruntergeladenen Dateien soll die genaue PDF- bzw. PDF/A-Version aufzeigen.

3.3.1 Hintergrund

Zunächst wird eine Metadatengewinnung durchgeführt, d.h. in die Metadatenfeldern der Datei geschaut und festgestellt, welche Version die PDF-Datei „behauptet“ zu sein. Die Metadaten von PDF-Dateien sind in den Dokumenteinstellungen und im Document Information Dictionary (DocInfo) zu finden.[95] Bei ISO-konformem PDF/A werden die Metadaten im XMP-Format (Extensible Metadata Platform) abgelegt. XMP basiert auf dem Resource Description Framework (RDF), welches vom World Wide Web Consortium als Standard für XML-basierte Metadaten anerkannt ist, und wurde im Jahre 2001 mit der PDF Version 1.4 eingeführt.

Das XMP-Framework ist in der Lage, Metadaten aus verschiedenen Schemas wiederzugeben, darunter neben dem XMP-Schema auch das Adobe-PDF-Schema und Dublin Core. Da XMP erweiterbar (extensible) ist, lassen sich auch individuelle Schemas erzeugen und darstellen. Einträge in der DocInfo können dank XMP aus verschiedenen Metadatenschemas stammen und dennoch gültig sein. Beispielsweise kann das Programm, dass die entsprechende PDF-Datei erzeugt hat, entweder im Feld „xmp:CreatorTool“ oder im Feld „pdf:Producer“ nachgewiesen sein. Die Information über die angebliche PDF-Version findet sich im Feld „pdf:PDFVersion“

[93] Vgl. Funk, S. E.; Neubauer, M. (2010): Formatcharakterisierung. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7:16

[94] Vgl. Freed, N.; Kucherawy, M.; Baker, M.; Hoehrmann, B. (2015): Media Types

[95] Vgl. Rosenthol, L. (2011): PDF/A Metadaten XMP, RDF & Dublin Core

bzw. in den XMP-Tags für PDF/A.[96] Das Jahr der Erstellung bzw. der letzten Änderung am Dokument (Last Modification) ist ebenfalls in der DocInfo zu finden.

Zudem muss die Validationssoftware aber noch die Datei selbst überprüfen und damit feststellen, ob die angegebene und die tatsächliche PDF-Version übereinstimmen. Dazu wird die Datei auf die bereits erwähnten Faktoren wie beispielsweise Kryptofunktionen untersucht. Da bei der Validierung einer großen Anzahl Dokumente erhebliche Probleme in der Performanz auftreten können, wird in der Praxis eher dazu geraten, bereits im Vorfeld die Arbeitsschritte bei der Erstellung von PDF/A zu überprüfen.[97]

3.3.2 Verwendete Software und Verfahrensweise

Für einzelne Dokumente sind Online-Validatoren[98] eine sinnvolle Anwendung. Für die Überprüfung größerer Dateimengen sind sie aber ungeeignet. Für diese Massenvalidierung existieren verschiedene kommerzielle und nicht-kommerzielle Lösungen. Die PDF-Association listet verschiedene Softwarelösungen zur Validierung auf ihrer Webseite auf[99] und stellt zudem eine Möglichkeit, PDF-Validatoren zu testen, in Form absichtlich nicht konformer PDF/A-Dateien bereit.[100]

Entgegen einiger Empfehlungen ist das frei verfügbare Programm JHOVE[101] nicht zur PDF/A-Validation geeignet.[102] JHOVE beherrscht zwar Validierung, allerdings nicht für alle Dateiformate.[103] Der Acrobat Preflight Validator von Adobe gilt als sehr zuverlässig[104] und ist innerhalb einer kostenlosen Testversion von Acrobat Pro verfügbar; allerding läuft er nicht auf Linux/Unix-Systemen, weshalb er für diese Arbeit ebenso ausscheidet wie der Apache Preflight, welcher zwar frei verfügbar ist, jedoch eine größere Fehlerquote aufweist.[105] Der pdfaPilot von Callas Software kann Dateien lediglich auf ihre Konformität zum PDF/A-1-Standard überprüfen.[106]

Für diese Arbeit wurde der 3 Heights PDF Validator[107] von PDF-Tools in einer Testversion genutzt. Er ist für Linux erhältlich und kann alle PDF/A-Varianten validieren. Zum auslesen des letzten Änderungsdatums aus der DocInfo wurde das kostenfreie „pdfinfo“ [108] verwendet.

---

[96] Vgl. ebd.

[97] Vgl. Merz, T. (2011): Validierung von PDF/A

[98] Vgl. Solid Documents (2015): Free PDF/A Validator

[99] Vgl. PDF Association (2015): Products

[100] Vgl. PDF/A Competence Center (2011): Isartor Test Suite

[101] Vgl. JSTOR (2009): JHOVE

[102] Vgl. Friese, Y. (2014): Langzeitverfügbarkeit sichern: PDF-Validierung durch JHOVE?

[103] Vgl. Funk, S. E.; Neubauer, M. (2010): Formatcharakterisierung. In: Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K.: nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7:17

[104] Vgl. Oettler, A. (2013): PDF/A kompakt 2.0, S. 12

[105] Vgl. van der Knijff, Johan (o.J.): Identification of preservation risks in PDF with Apache Preflight

[106] Vgl. Callas Software GmbH (o.J.): Wichtigste Funktionen

[107] Vgl. PDF Tools AG (2015): PDF Validator

[108] Vgl. SPI Inc. (2015): Package: poppler-utils in wheezy

Die Dateien können nicht automatisch gegenüber allen möglichen PDF-Versionen abgeprüft werden. Daher werden sie über das Argument „-cl ccl“ (claimed conformance and level) gegenüber der angegebenen Version validiert. Über die Berichtsfunktion („-rs“) kann in Erfahrung gebracht werden, ob das tatsächliche Format dem tatsächlichen entspricht und welche Vorgaben eine Datei nicht erfüllt. Die Ansteuerung der Programme zur Validierung wurde in das Perl-Skript integriert.

4 Auswertung

Die Base-URLS der ausgewählten Repositorien sind beim Directory of Open Access Repositories (DOAR)[109] oder bei der OAI[110] verzeichnet. In Anbetracht der oben genannten Kriterien und Anforderungen an Repositorien und des Grades der Automatisierung der Harvesting- und Validierungsprozesse wurden die unten aufgeführten Repositorien ausgewählt. Darunter befinden sich auch Repositorien ohne DINI-Zertifikat, die einen Vergleich zwischen zertifizierten und nicht zertifizierten Repositorien ermöglichen sollen.

4.1 Aussortierung von Repositorien

Leider mussten einige Repositorien wegen Fehlern bei Zugriff und Download aus dem Umfang dieser Arbeit herausgenommen werden. Einige gaben gar keine Antwort auf die OAI-PMH-Anfrage, andere lieferten leere Rückmeldungen oder keine PDF-Dateien zurück. Der häufigste Fehler wurde wahrscheinlich dadurch verursacht, dass die bereitgestellten operablen URLs als Identifier per Redirect auf die eigentliche PDF-Datei verwiesen und nicht auf eine Landing Page. Solche und ähnliche Fehler tauchten bei einigen Repository-Systemen auf, weshalb diese entweder aus der Untersuchung ausgenommen werden mussten oder nicht alle PDF-Dateien heruntergeladen und analysiert werden konnten. Dabei waren einige Repository-Systeme wie etwa OPUS weniger anfällig, aber dennoch nicht fehlerfrei.

Wegen instabiler Internetverbindungen konnten zudem keine Repositorien ab einer Größe von ca. 20.000 Items (laut Open DOAR) bearbeitet werden. Die Dauer des Harvesting- und Validierungsvorgangs dauerte schlicht zu lange. Daher musste auf die Auswertung größerer institutioneller und fachlicher Repositorien, wie etwa die Universitäten Bielefeld und Frankfurt, die Zentralbibliotheken für Medizin und Wirtschaft oder das Social Science Open Access Repository, verzichtet werden.

Die meisten Repositorien lieferten unproblematische Ergebnisse. Bei einigen gab es aber mehr Einträge in der Ergebnistabelle als Dublin-Core-Metadatensätze. Dies lässt sich zum einen darauf zurückführen, dass zu einer Ressource (und damit zu einem Item) schlichtweg mehrere Dateien gehören können; dazu zählen beispielsweise Anhänge, einzelne Kapitel oder einzelne Dokumente für Titelblatt oder Abstracts. Zum anderen ist auf Grund der Diskrepanz zwischen Einträgen und Metadatensätzen anzunehmen, dass das Perl-Skript versucht hat, weitere PDF-Dateien herunterzuladen, die sich auf der entsprechenden Landing Page befinden, aber nicht zur Ressource gehören.

Daneben gab es oft auch mehr Einträge in der Ergebnistabelle als heruntergeladene PDF-Dateien. In diesen Fällen gab der Validator im Report die Fehlermeldung aus, dass die entsprechende Datei nicht vorhanden ist. Eine eindeutige Ursache für diesen Fehler konnte nicht be-

[109] Vgl. University of Nottingham (2014): OpenDOAR - Summaries - Germany
[110] Vgl. Open Archives Initiative (2014): Registered Data Providers

stimmt werden. Mögliche Ursachen sind veraltete Links, Download-Sperrungen für die Dateien seitens der Website oder Fehlfunktionen bei den Anweisungen zum Anfordern oder Crawlen der Webseite im Perl-Skript. Der Anteil an fehlerhaften Einträgen war in manchen Fällen unverhältnismäßig hoch und die Anzahl der übrigen Einträge reichte bei Weitem nicht an die bei Open DOAR angegebenen Dokumentanzahlen heran. Daher wurden diese Repositorien nicht untersucht. Der Fehler trat auch in anderen Fällen auf, allerdings so schwach ausgeprägt, dass die Werte in die Analyse der PDF-Dateien miteinbezogen wurden.

Die TU Freiberg musste von der Analyse ausgeschlossen werden, da die per Harvesting heruntergeladenen PDF-Dateien ausschließlich aus gescannten Bildern bestanden, obwohl beim normalen Zugriff über den Browser normale Textdateien zu finden sind. Die Ursache für diesen Fehler ist nicht bekannt.

Das Repositorium der RWTH Aachen wurde während der Bearbeitungszeit auf ein neues Repository-System umgestellt. Wegen einer fehlenden OAI-PMH-Base-URL konnten daher nach dem 19.1.2015 keine Dateien heruntergeladen und analysiert werden. Die TU Hamburg-Harburg lieferte zu jedem Eintrag mindestens zwei Dateien, von denen die zweite nicht geöffnet werden konnte. Wegen der Dateigröße kann es sich dabei aber weder um eine Dublette, noch um ein Titelblatt oder Abstract handeln. Um die Ergebnisse nicht zu verfälschen wurde auf eine Auswertung hierbei verzichtet.

Der Dokumentenserver der Universität Hohenheim lieferte nur etwa halb so viele Dateien, wie erwartet. Auch hören die Modifikationsdaten nach dem Jahr 2010 praktisch auf. Ein Blick auf die Wachstumsstatistik des Repositoriums bei Open DOAR zeigt erst im Jahr 2012 wieder Werte für das Repositorium. Es ist also möglich, dass das Dokumentensever zwei Jahre lang außer Betrieb war, etwa für eine Umstellung des Repository-Systems, und alle seitdem veröffentlichten Dateien nicht über das Perl-Skript erreicht wurden, da ihre Nachweise nicht per OAI-PMH zugänglich sind. Ähnliches zeigte sich auch bei der Universität Passau; dort wurden nur etwa 400 von 2700 Dokumenten heruntergeladen.

Insgesamt konnte das Skript auf 24 Repositorien erfolgreich angewandt werden.

4.2 Auswertung der einzelnen Repositorien

Als Grundlage dienen hier die Fehlermeldungen des 3-Heights-PDF-Validators[111]. Sie werden in untenstehender Tabelle übersetzt und ggf. erläutert. Aus Gründen des Umfangs wird meist nur auf die häufigsten Fehler eingegangen. Offensichtlich falsche Informationen, wie etwa unrealistische Jahreszahlen, wurden zum Teil in den Grafiken behalten, um einen ganzheitlicheren Überblick zu gewähren. Auch wenn es der Anschaulichkeit nicht sehr dienlich ist, wird wegen des Umfangs der Arbeit auf einige Grafiken verzichtet und die relevanten Informationen schriftlich dargelegt, wenn die Ergebnisse bereits genannten Ergebnisse ähneln. Auf kleine Repositorien wird wegen Geringfügigkeit nur oberflächlich eingegangen.

[111] Vgl. PDF Tools AG (2014): 3-Heights PDF Validator Shell, User Manual, S. 19

Englisch	Deutsch
The file format (header, trailer, objects, xref, streams) is corrupted.	Die Datei ist beschädigt.
The document doesn't conform to the PDF reference (missing required entries, wrong value types, etc.).	Die Datei entspricht nicht dem Format demgegenüber es geprüft wurde.
The file is encrypted.	Die Datei ist verschlüsselt / beinhaltet Kryptofunktionen.
The document contains device-specific color spaces.	Die Datei enthält Farbräume, die nicht auf allen Geräten dargestellt werden können.
The document contains illegal rendering hints (unknown intents, interpolation, transfer and halftone functions).	Die Datei enthält ungültige Angaben zum Rendering / zur Darstellung.
The document contains alternate information (images).	Die Datei enthält alternative Beschreibungen für Bilder.
The document contains embedded PostScript code.	Die Datei enthält PostScript Code.
The document contains references to external content (reference XObjects, file attachments, OPI).	Die Datei verlinkt auf Inhalte, die sich außerhalb der Datei befinden.
The document contains fonts without embedded font programs or encoding information (CMAPs)	Die Datei enthält eine oder mehrere Schriftarten, die nicht eingebettet sind oder keine Angaben zur Kodierung haben.
The document contains fonts without appropriate character to Unicode mapping information (ToUnicode maps)	Die Datei enthält eine oder mehrere Schriftarten ohne angemessene Informationen zum Mapping in Unicode.
The document contains transparency.	Die Datei enthält transparente Ebenen.
The document contains unknown annotation types.	Die Datei enthält unbekannte Arten von Anmerkungen.
The document contains multimedia annotations (sound, movies).	Die Datei enthält multimediale Anmerkungen.
The document contains hidden, invisible, non-viewable or non-printable annotations.	Die Datei enthält versteckte, unsichtbare, nicht anschaubare oder nicht druckbare Anmerkungen.

The document contains annotations or form fields with ambiguous or without appropriate appearances.	Die Datei enthält Anmerkungen oder Formularfelder mit mehrdeutigen oder gänzlich ohne angemessene Erscheinung.
The document contains actions types other than for navigation (launch, JavaScript, ResetForm, etc.)	Die Datei enthält Aktionen / Aktionstypen, die nicht für die Navigation eingesetzt werden.
The document's metadata is either missing or inconsistent or corrupt.	Die Metadaten der Datei sind nicht vorhanden, unstimmig oder beschädigt.
The document doesn't provide appropriate logical structure information.	Die Datei liefert nicht ausreichende Information zu seiner logischen Struktur.
The document contains optional content (layers).	Die Datei enthält Ebenen.

Tabelle 1: Fehlermeldungen des 3-Heights-PDF-Validators

Bei jedem Repositorium wird zunächst ein stichpunktartiger Überblick über die Richtlinien bezüglich der Langzeitarchivierung und den angebotenen Hilfestellungen gegeben. Aus Gründen der Relevanz und der Übersichtlichkeit wird bei den Anforderungen der Repositorien auf nicht textuelle Formate verzichtet und i.d.R. nur das Präsentationsformat angegeben. Sowohl bei den Hilfestellungen, als auch bei den LZA-Richtlinien verweisen einige Repositorien auf die entsprechenden Webseiten anderer Einrichtungen.

4.2.1 Albert-Ludwigs-Universität Freiburg

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
    - Format: PDF Pflichtformat, PDF/A empfohlen, Ablehnung von Dateien mit Kryptofunktionen, HTML bedingt möglich
    - Dauer: keine zeitliche Einschränkung bezüglich der Verfügbarkeit[112]
- Hilfestellungen: Bereitstellung detaillierter Tutorials[113] und Arbeitsplätze zur Erstellung von PDF-Dateien

Der Publikationsserver der Universität Freiburg lieferte insgesamt 8666 PDF-Dateien (Stand: 19.1.2015), von denen weniger als 0,7 % PDF/A-Dateien sind. Die größten Anteile haben die PDF-Versionen 1.4 und 1.6.

[112] Vgl. Universitätsbibliothek Freiburg (2015): Häufig gestellte Fragen (FAQ) - Universitätsbibliothek Freiburg

[113] Vgl. Universitätsbibliothek Freiburg (2015): Publizieren im PDF-Format (ein Online-Tutorial) - Universitätsbibliothek Freiburg

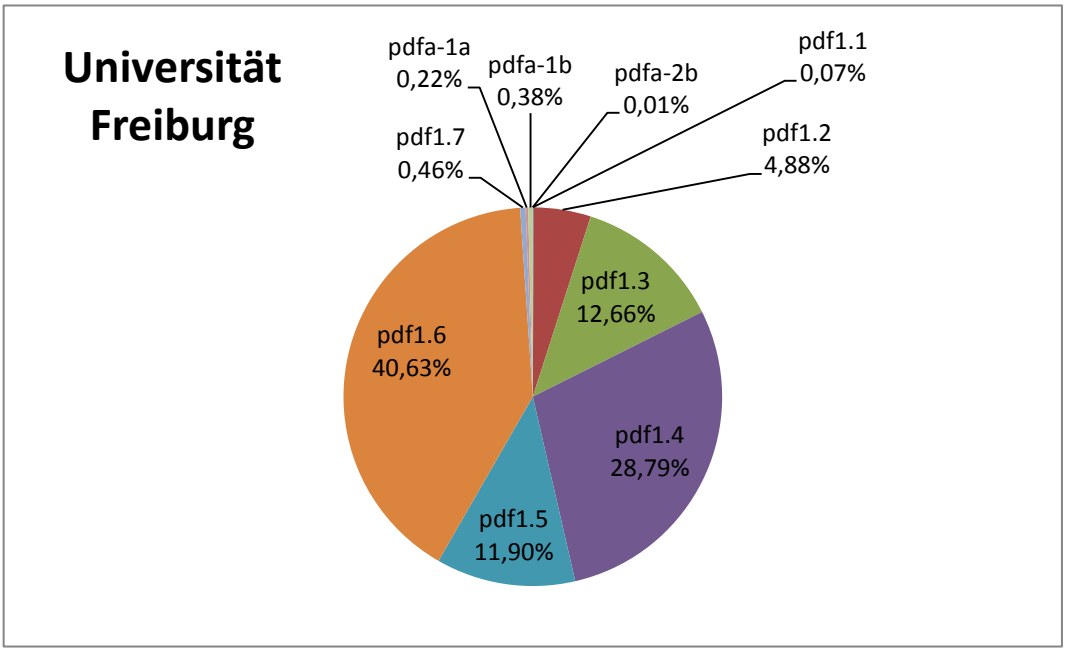

Abbildung 4: Anteil der PDF-Formate an der Universität Freiburg

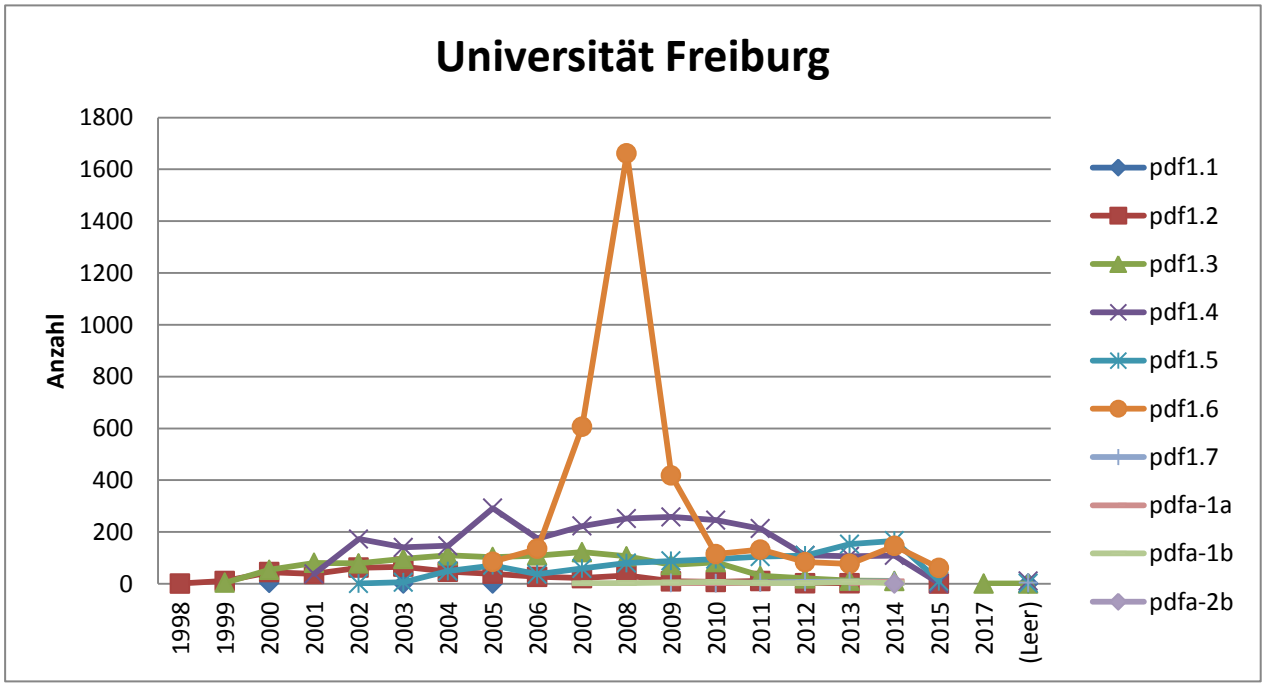

Abbildung 5: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Freiburg

Bei Version 1.6 ist die schlagartig erhöhte Anzahl von erstellten Dateien in den Jahren 2007 bis 2009 besonders auffällig. Eine denkbare Erklärung dafür ist, dass zu diesem Zeitpunkt ein größeres Paket an Dateien aus einem anderen Repository-System oder einem Digitalisierungsprojekt angefallen ist. Zudem zeigt sich, dass die Validierungssoftware nicht alle Fehler aufdecken kann. Das Änderungsdatum einer Datei ist auf das Jahr 2017 datiert; der Validator hat den entsprechenden Eintrag aber nicht mit einer Fehlermeldung über widersprüchliche Metadaten versehen. Insgesamt weisen 106 Dateien Fehlerhafte Metadaten auf, 312 Dateien enthalten nicht eingebettete Schriftarten und 3133 Dokumente entspre-

chen nicht dem angegebenen Dateiformat. Insgesamt sind trotz des detaillierten Tutorials mehr als ein Drittel der Dateien nicht fehlerfrei.

4.2.2 Deutsches Institut für Internationale Pädagogische Forschung (DIPF)

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF bzw. PDF/A, ohne Kryptofunktionen, durchsuchbarer Volltext[114]
  - Dauer: Langzeitverfügbarkeit, Garantie für fünf Jahre
- Hilfestellungen: Links zu Anleitungen[115] [116] an der Humboldt-Universität zu Berlin

Das Repositorium des DIPF lieferte insgesamt 22426 Einträge; davon waren aber 7557 Einträge mit den eingangs erwähnten Dateifehlern (Stand: 19.1.2015). Diese Einträge konnten naturgemäß nicht validiert werden und fallen daher aus der Beobachtung heraus. Damit bleiben noch 14869 Dateien übrig. Von diesen machen ca. 26 % PDF/A-Dateien aus, zumeist PDF/A-1b. Den größten Anteil haben wieder die Formate 1.4 und 1.6.

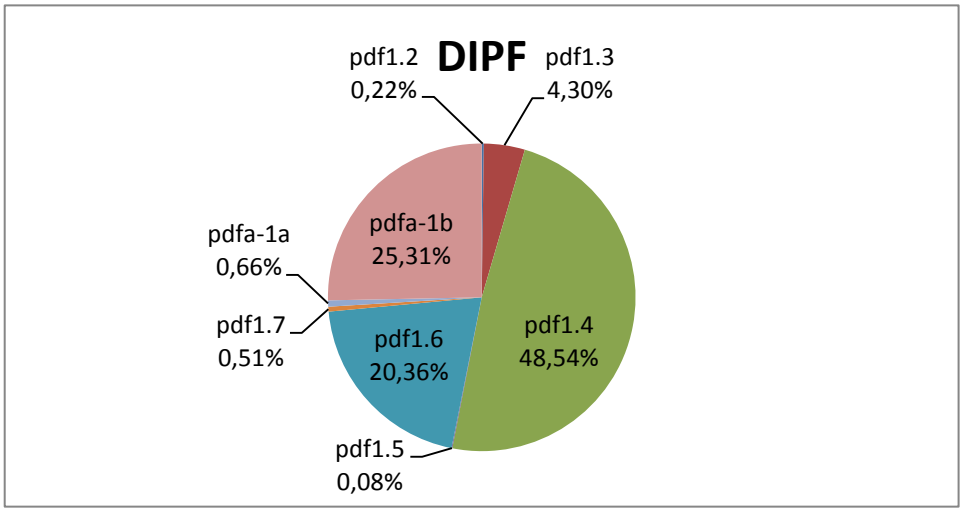

Abbildung 6: Anteil der PDF-Formate am Dt. Institut für Internationale Pädagogische Forschung

Besonders ins Auge springt der Verlauf der neuen Dateien im Format 1.4. Nachdem eher wenige und im Jahr 2012 sogar gar keine Dateien in diesem Format verändert wurden steigt die Quote in den Jahren 2013 und 2014 schlagartig an. Zeitgleich fallen die Werte für PDF 1.6 und PDF/A-1b, die beinah parallel verlaufen, rapide bis auf Null ab. Im Jahr 2014 wurden sogar ausschließlich PDF 1.4 Dateien modifiziert. Aus den Leitlinien und den Hilfestellungen lässt sich

[114] Vgl. Deutsches Institut für Internationale Pädagogische Forschung (2013): Fachportal Pädagogik - pedocs - Leitlinien des Dokumentenservers pedocs

[115] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Elektronisches Publizieren

[116] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Technische Hinweise zur Erstellung und Korrektur von PDF-Dokumenten

dafür keine direkte Erklärung herleiten. Möglich ist eine interne, bisher noch nicht kommunizierte Richtlinie, die die veröffentlichten Dateien auf das Format 1.4 beschränkt. Dies erscheint vor dem Hintergrund, dass die Anleitung für die Erstellung von PDF- und PDF/A-Dokumenten zumindest in Version 0.8 seit 2013 vorhanden ist, schwer nachzuvollziehen. Ein Fehler der Validationssoftware ist hierbei nicht auszuschließen. Wiederum ist möglich, dass es sich um ein Datenpaket aus einem Digitalisierungs- oder Konvertierungsprojekt handelt, oder dass in den Jahren 2009 bis 2011 Anstrengungen zur Erlangung des DINI-Zertifikats unternommen und später aus Gründen des Aufwands wieder fallen gelassen wurden.

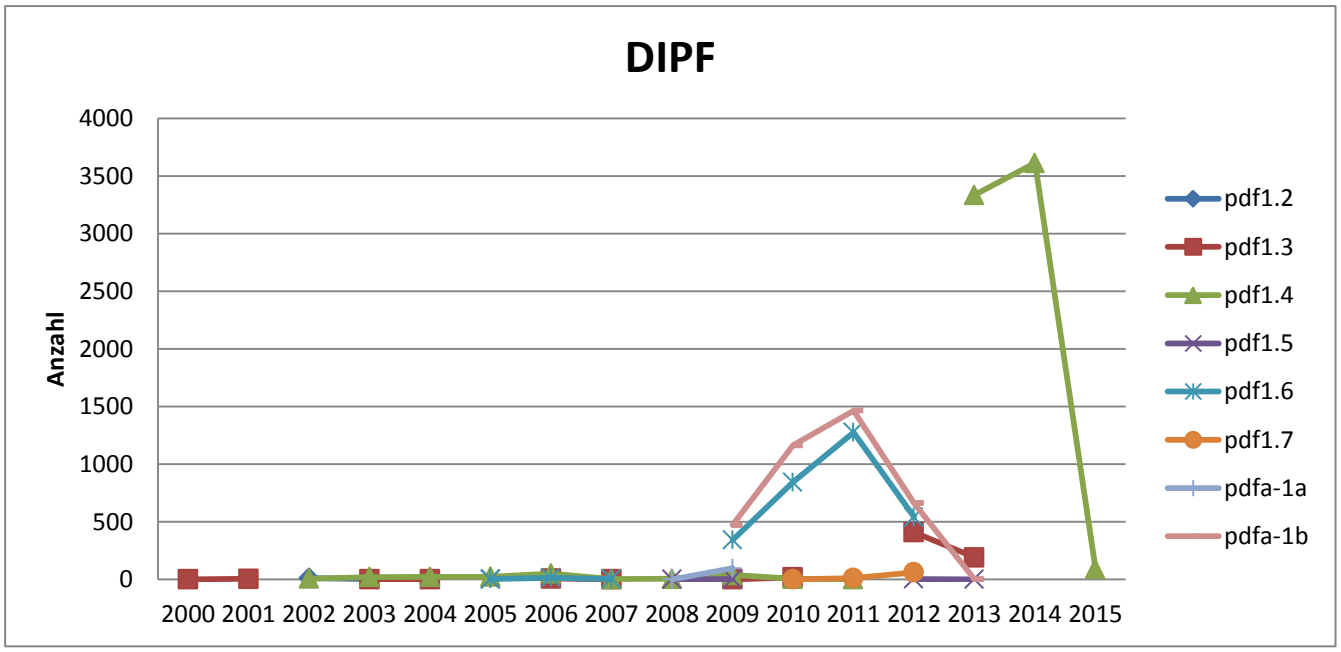

Abbildung 7: Art und letztes Änderungsdatum von PDF-Dateien am Dt. Institut für Internationale Pädagogische Forschung

Mit 6331 Vorkommen ist der häufigste Fehler, dass die Dateien nicht der angegebenen PDF-Referenz entsprechen. Bis auf 18 Stück trifft dieser Fehler auf alle angeblichen PDF/A-1b-Dateien zu, sowie auf ungefähr 80 % der PDF 1.6 Dateien. Mit 86 Mal trat der Fehler relativ selten bei PDF 1.4 Dateien auf. Dies lässt vermuten, dass sowohl die Werte aus der obigen Abbildung, als auch die PDF-Referenz-Fehler für 1.6 und PDF/A-1b eine gemeinsame Ursache haben könnten, wie beispielsweise fehlerhafte Erstellungs- oder Validierungsprozesse, die schließlich im Jahr 2013 komplett behoben wurde. Weitere nennenswerte Fehler sind ungültige Farbräume (730), nicht eingebettete Schriftarten (190) und fehlerhafte Metadaten (505).

4.2.3 Fachhochschule Köln

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF
  - Dauer: mindestens fünf Jahre[117]
- Hilfestellungen: Link zum PDF-Tutorial der Universität Freiburg[118]

Der Dokumentenserver der FH Köln lieferte 53 Dateien (Stand: 19.1.2015). Davon waren mit 45 Stück die meisten im Format 1.6. Sechs Dateien waren in 1.5. und zwei in 1.4. Da das älteste Modifikationsdatum der Dateien bei 2012 liegt und das Repositorium sehr wenige Dateien enthält kann der leicht absteigenden Tendenz der Anzahl eingereichter Dateien keine fundierte Deutung zu Grunde gelegt werden. Zwei Dateien entsprechen nicht der PDF-Referenz; eine beinhaltete eine nicht eingebettete Schriftart.

Dass das Repositorium 2010 das DINI-Zertifikat erhielt, die analysierten Dateien aber frühestens 2012 verändert wurden, deutet darauf hin, dass alle PDF-Dateien auf irgendeine Weise verändert, möglicherweise sogar nach 1.6 konvertiert wurden.

4.2.4 Hochschule Konstanz Technik, Wirtschaft und Gestaltung (HTWG)

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF für Online-Einreichung, beliebiges Textformat bei Übermittlung an die Bibliothek[119]
  - Dauer: Keine Angabe
- Hilfestellungen: Kurze Erklärung zu Metadaten[120], Abgabe in beliebigem Format und Konvertierung durch die Bibliothek

Abzüglich Dubletten und Dateifehler konnten insgesamt 123 Dateien erfolgreich überprüft werden (Stand: 19.1.2015). Den Löwenanteil machen wieder die normalen PDF-Formate, besonders 1.3 und 1.4 aus. Nur zwei der Dateien sind in einem PDF/A-Format. Die geringe Gesamtanzahl an Dateien erklärt auch die sporadische Verteilung der PDF-Formate: Version 1.7 beispielsweise wurde erst im Jahr 2014 verwendet, wobei zeitgleich der Anteil anderer Formate stark abnimmt. Dies könnte von der Wahrnehmung des Angebots der Bibliothek rühren, beliebige Formate einreichen zu können und die Bibliothek die Konvertierung vornehmen zu lassen. 1.3 und 1.4 wurden nach Spitzen in den Jahren 2003 und 2005 nur noch vereinzelt verwendet. Auch bei diesem Repositorium zeigt sich die genannte Schwäche des Validators: Eine Datei,

[117] Vgl. Fachhochschule Köln Hochschulbibliothek (2013): Cologne Open Science - Leitlinien von Cologne Open Science

[118] Vgl. Universitätsbibliothek Freiburg (2015): Publizieren im PDF-Format (ein Online-Tutorial) - Universitätsbibliothek Freiburg

[119] Vgl. Hannemann, B. (o.J.): OPUS

[120] Vgl. Bibliothek der HTWG Konstanz (o.J.): OPUS 4 | Hilfe: Was sind Metadaten?

die als letztes Änderungsdatum 1970 angibt, wurde nicht mit einer Fehlermeldung bezüglich Metadaten versehen.

4.2.5 Forschungszentrum Jülich

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF erwünscht[121] [122], theoretisch aber HTML, XML, Word, TXT und weitere möglich[123]
  - Dauer: „langfristige Verfügbarkeit“[124]
- Hilfestellungen: Erläuterungen zu den Metadaten[125]

Von den 5241 Einträgen beim FZ Jülich waren 151 fehlerhaft (Stand: 19.1.2015). Über die übrigen 5090 Dateien sind die normalen PDF-Formate mit Ausnahme von Version 1.1 und 1.7 beinahe gleichmäßig verteilt. Mit ca. 0,12 % sind PDF/A-Formate kaum vertreten.

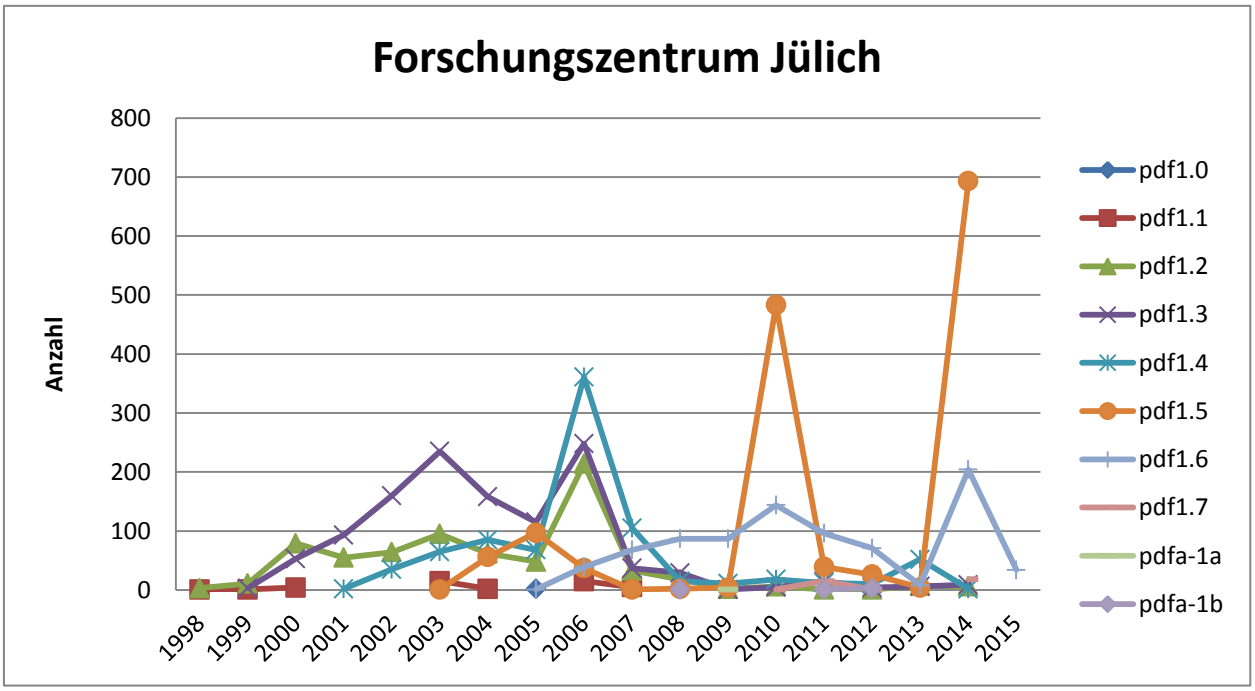

Abbildung 8: Art und letztes Änderungsdatum von PDF-Dateien am Forschungszentrum Jülich

Besonders präsent sind die im Abstand von drei bis vier Jahren auftretenden Spitzen an modifizierten bzw. veröffentlichten Dateien. Sie markieren womöglich die Durchführung entsprechend

[121] Vgl. Zentralbibliothek des Forschungszentrum Jülich (o.J.): JUWEL - Volltextserver: FAQ

[122] Auf der Seite ist zu lesen: „Unabhängig vom Ausgangsformat sollte jedes Dokument als pdf-Datei auf dem Server abgelehnt [sic] werden.“ Es wird vermutet, dass es sich hierbei um einen Fehler handelt und eigentlich „abgelegt“ heißen soll.

[123] Vgl. Zentralbibliothek des Forschungszentrum Jülich (o.J.): Von JUWEL unterstützte Formate

[124] Vgl. Zentralbibliothek des Forschungszentrum Jülich (o.J.): JUWEL - Volltextserver: FAQ

[125] Vgl. Zentralbibliothek des Forschungszentrum Jülich (o.J.): JUWEL - Volltextserver: Hilfe

lange dauernder (Digitalisierungs-) Projekte oder Migrationen. Ab 2006 sind die Unterschiede zwischen den Hoch- und Tiefphasen zunehmend krasser. Zeitgleich lässt sich eine Ausdünnung der verwendeten Formate feststellen. Nachdem 2003 und besonders 2006 noch mehrere Formate – besonders 1.2, 1.3 und 1.4 – verwendet wurden zeigt sich ab 2010 ein eindeutiger Trend zur Verwendung von PDF 1.5 und in weitaus geringerem Maße zu 1.6. Da die Leitlinien, Hilfestellungen und Anforderungen diese Formate nicht explizit fordern liegt der Grund für diese Entwicklung möglicherweise in der Konvertierung älterer Formate in ein einheitliches oder schlichtweg in der Verbreitung der Formate durch entsprechende Erstellungs- bzw. Exporttools.

249 Dateien wiesen fehlhafte Metadaten auf, 150 entsprachen nicht der angegebenen PDF-Version und 85 enthielten nicht eingebettete Schriftarten. Weitere Fehler traten nur vereinzelt auf.

4.2.6 Johannes Gutenberg-Universität Mainz

- DINI-Zertifikat: 2007
- LZA-Richtlinien:
  - Format: PDF, PDF/A empfohlen, ohne Kryptofunktionen
  - Dauer: mindestens fünf Jahre[126]
- Hilfestellungen: Kurze Anleitung und Checklisten[127]

3075 Dateien wurden von der Universität Mainz heruntergeladen und überprüft (Stand: 19.1.2015). PDF/A-Dateien machen dabei weniger als drei Prozent aus. Allerdings ist das PDF/A-1b-Format mit 2,5 % stärker vertreten als an anderen Repositorien.

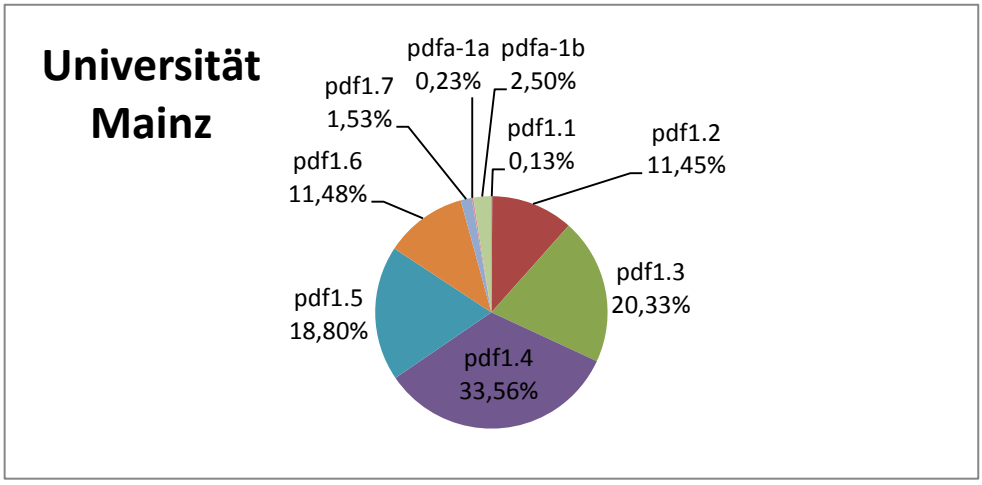

Abbildung 9: Anteil der PDF-Formate an der Universität Main

[126] Vgl. Universitätsbibliothek Mainz (o.J.): Leistungsspektrum und Nutzungsrichtlinien

[127] Vgl. Universitätsbibliothek Mainz (o.J.): Anleitung zum Eintragen und Veröffentlichen eines Dokuments

Auch zeitlich sind die verschiedenen Formate insgesamt gleichmäßiger verteilt als die anderer Repositorien. Version 1.4 ist über längere Zeit hinweg das dominierende Format und wird erst 2013 durch Version 1.5 aus dieser Position verdrängt. Seit 2007, dem Jahr des Erwerbs des DINI-Zertifikats, sind auch PDF/A-Dateien vertreten. 2011 ist eine deutliche Spitze der PDF/A-1b-Dateien zu erkennen und zum gleichen Zeitpunkt fällt der Wert für PDF 1.3 deutlich. Ob dies mit der Veröffentlichung der Kurzanleitung zur PDF-Erstellung zusammenhängt kann nicht gesagt werden, da für diese kein Jahr bekannt ist.

Wieder zeigen sich offensichtlich falsche Metadaten zum letzten Modifikationsdatum, die keine entsprechende Fehlermeldung lieferten und wieder sind die häufigsten Fehler die der Nichtübereinstimmung mit dem angeblichen Format (387), der Verwendung von nicht eingebetteten Schriftarten und der fehlerhaften Metadaten.

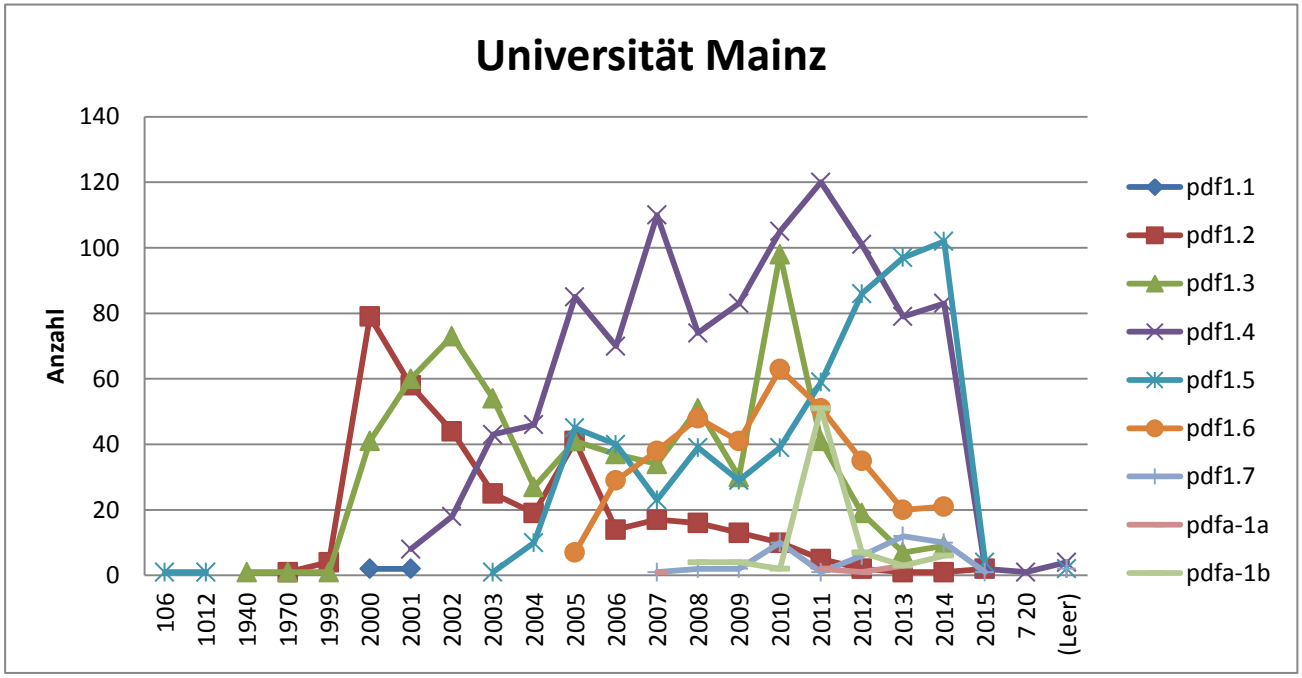

Abbildung 10: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Mainz

4.2.7 Justus-Liebig-Universität Gießen

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF ohne Kryptofunktionen, PDF/A-1b bevorzugt[128], PDF/A empfohlen[129]
  - Dauer: Garantie für mindestens zehn Jahre[130]
- Hilfestellungen: Keine

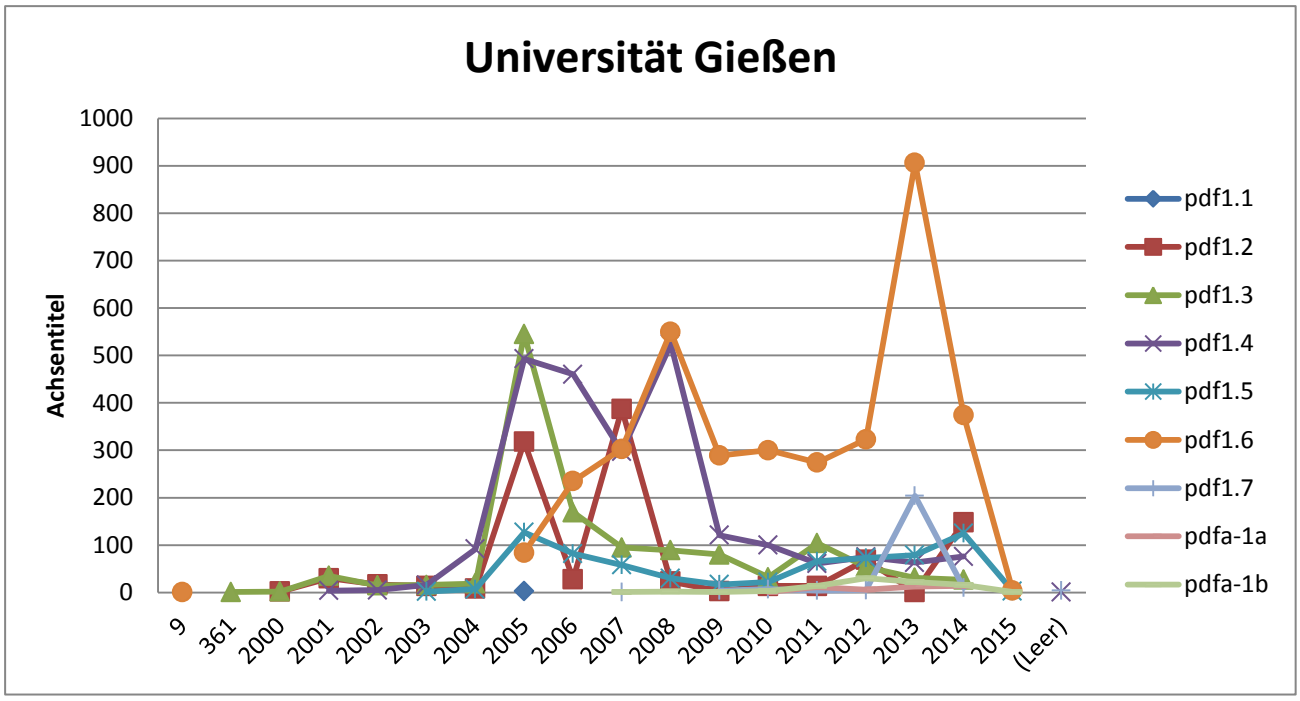

Abbildung 11: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Gießen

Das Repositorium der Universität Gießen ergab 9490 überprüfbare Dateien (Stand: 21.1.2015) und zeichnet ein den meisten Repositorien ähnliches Bild der Formatanteile, in dem besonders die Versionen 1.4 und 1.6 dominieren und die PDF/A-Formate zusammen knapp anderthalb Prozent ausmachen. Die Empfehlung von PDF/A bzw. die Bevorzugung von PDF/A-1b schlägt sich nicht in den erhobenen Daten nieder. Zwar ist PDF/A-1b zu größerem Teil vorhanden als etwa PDF/A-1a, allerdings weisen auch Repositorien ohne die entsprechende Vorgabe dieses Verhältnis auf, teilweise sogar mit deutlicherem Unterschied.

Auch hier findet sich eine generelle Tendenz zur Reduktion der Diversität der Formate und zur Verwendung eines einheitlichen Formates. Die deutliche Durchsetzung von PDF 1.6 als Quasi-Standard in den letzten Jahren ist vor dem Hintergrund, dass es keinerlei Hilfestellungen für Publizierende gibt, durchaus erstaunlich. Möglich ist, dass ältere Dateien in einem Projekt in

[128] Vgl. Justus-Liebig-Universität Gießen (2014): Anleitung zur Veröffentlichung von Dissertationen in GEB

[129] Vgl. Justus-Liebig-Universität Gießen (2011): Häufig gestellte Fragen – Faqs

[130] Vgl. Justus-Liebig-Universität Gießen (o.J.): Dokumenten- und Publikationsserver GEB der JLU Gießen - Leitlinien

das Format konvertiert wurden. Das würde auch die Spitze von über 900 modifizierten Dateien von PDF 1.6 im Jahr 2013 erklären, wobei natürlich nicht auszuschließen ist, dass diese durch eine Migration oder aus einem Digitalisierungsprojekt resultiert.

Wieder sind Fehler bei den Metadaten (364), nicht eingebetteten Schriftarten (224) und der PDF-Referenz (609) häufig anzutreffen. Dazu kommen neben anderen Mängeln auch 357 korrumpierte, also beschädigte Dateien. Auch hier wurden die Dateien mit hochgradig unrealistischen Metadaten (z.B. das Modifikationsjahr 9) nicht mit einem Metadatenfehler bedacht.

4.2.8 Philipps Universität Marburg

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF/A, PDF/A-1a empfohlen[131]
  - Dauer: „Dauerhafte Bereitstellung“[132]
- Hilfestellungen: Persönliche Unterstützung bei der Konvertierung (bei erheblichem Aufwand kostenpflichtig)[133], kurze Erläuterung zu PDF/A, sowie Hinweise auf die Erstellung[134]

Obwohl die Universität Marburg kein DINI-Zertifikat besitzt bestehen die 561 Dateien bestehen zu ca. 28 % aus PDF/A-Dateien und zu etwa ähnlich großen Anteilen aus PDF 1.4 und 1.5 (Stand: 19.1.2015). Dies lässt sich durchaus auf die Leitlinien zurückführen, in denen PDF/A-Dateien gefordert werden. Das empfohlene PDF/A-1a ist allerdings nicht so häufig vorhanden wie die Basic-Variante, was wahrscheinlich daran liegt, dass Textverarbeitungsprogramme wegen der geringeren Komplexität eher den Export zu PDF/A-1b unterstützen. Allerdings ist der Anteil von PDF/A-1a-Dateien im Vergleich zu anderen Dokumentenservern außergewöhnlich hoch.

Die Universitätsbibliothek bietet neben Hinweisen seit 2011 auch eine persönliche Beratung bei der Konvertierung in das gewünschte Format. Im Folgejahr traten PDF/A-Formate erstmals vermehrt auf. Zur Relativierung muss aber hinzugefügt werden, dass in diesem Jahr sämtliche Formate, ergo insgesamt einfach mehr Dateien eingereicht bzw. zumindest modifiziert wurden. Da das Repositorium laut Open DOAR auch Dokumente ab dem Jahr 1997 enthält[135], scheint ein Digitalisierungs- oder Konvertierungsprojekt durchaus plausibel. Der starke Abfall an modifizierten Dateien ab dem Jahr 2013 korreliert mit der schlagartig verringerten Gesamtanzahl an Dateien auf dem Server. Eine denkbare Erklärung wäre, dass zuvor kapitelweise veröffentlichte Dokumente wieder zu einer Datei zusammengeführt wurden, wodurch die Gesamtzahl der Dateien sinkt und zugleich die Zahl veränderter Dateien im Vorfeld ansteigt.

---

[131] Vgl. Glaser, T. (2013): FAQs

[132] Vgl. Universitätsbibliothek Marburg (2011): Benutzungsordnung für den Publikationsserver der Universitätsbibliothek an der Philipps-Universität Marburg

[133] Vgl. ebd.

[134] Vgl. Glaser, T. (2013): FAQs

[135] Vgl. University of Nottingham (2015): OpenDOAR - Summaries - Worldwide, Query: Publikations- und Dokumentenserver der Universitätsbibliothek Marburg

Der häufigste Mangel ist wieder der PDF-Referenz-Fehler (90), gefolgt von fehlerhaften Metadaten (33) und nicht eingebetteten Schriftarten (22). Wie bei den meisten anderen Repositorien kamen diverse andere Fehler vereinzelt vor.

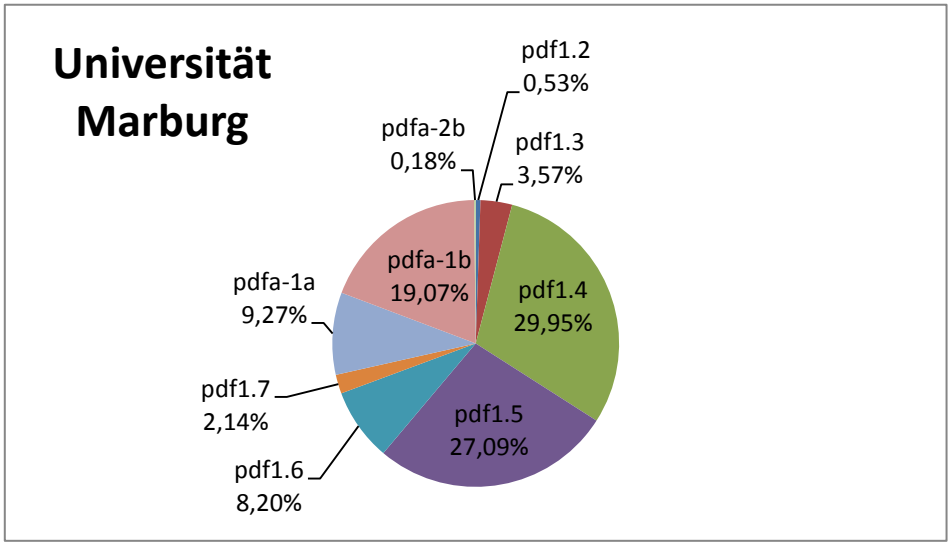

Abbildung 12: Anteil der PDF-Formate an der Universität Marburg

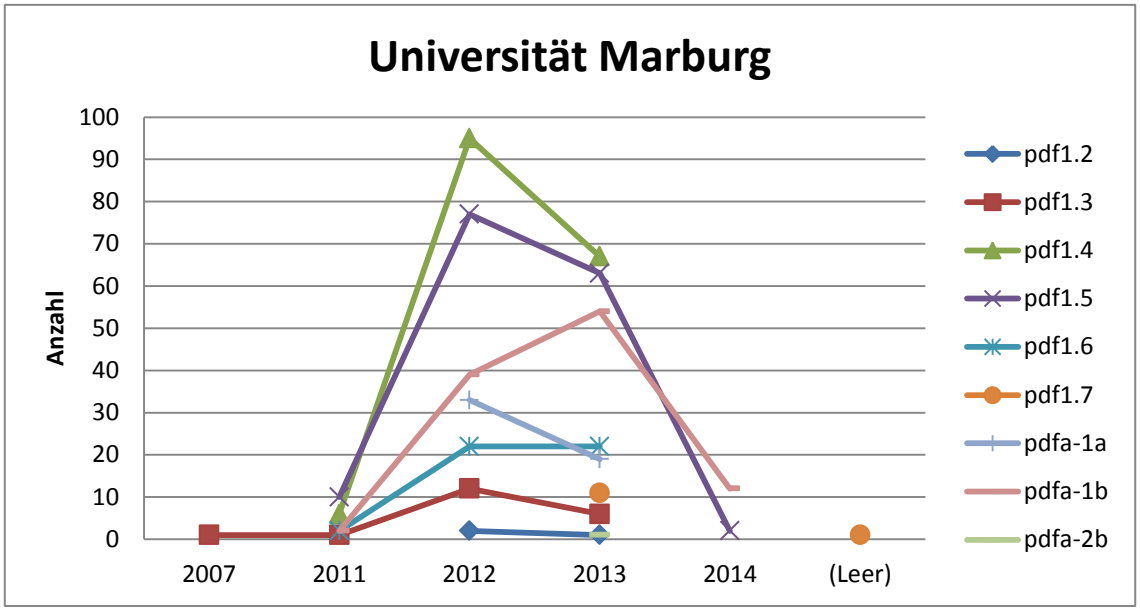

Abbildung 13: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Marburg

4.2.9 Ruhr-Universität Bochum

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF ohne Kryptofunktionen[136], HTML
  - Dauer: Keine Angabe
- Hilfestellungen: Links zu Tools zur Konvertierung[137], persönliche Beratung, Konvertierung[138]

Abzüglich der 19 Dateifehler ergab das Repositorium der Universität Bochum 12508 Einträge (Stand: 20.1.2015). Ganz offensichtlich ist das Format 1.6 vorherrschend. PDF/A-Dateien sind zu weniger als einem halben Prozent vorhanden. In den Jahren von 1999 bis 2008 ist gut zu erkennen, dass stets das damals aktuellste PDF-Format am meisten verwendet wurde, jedenfalls mit einer Verzögerung von einem Jahr. Auf PDF 1.7 trifft dies aber nicht mehr zu, da ab dem Jahr 2010 schlagartig die Anzahl an Dokumenten im Format 1.6 aufschnellt. Dieser Umstand lässt sich mit dem von der Bibliothek angebotenen Service der Konvertierung nach PDF erklären, der womöglich zu dieser Zeit angetreten wurde, erklären. Dies ist jedenfalls eine plausible Erklärung für das nachfolgende hohe Niveau der Anzahl neuer Dateien in 1.6. und der steten, wenn auch deutlich geringeren Anzahl neuer Dateien in PDF/A-1b.

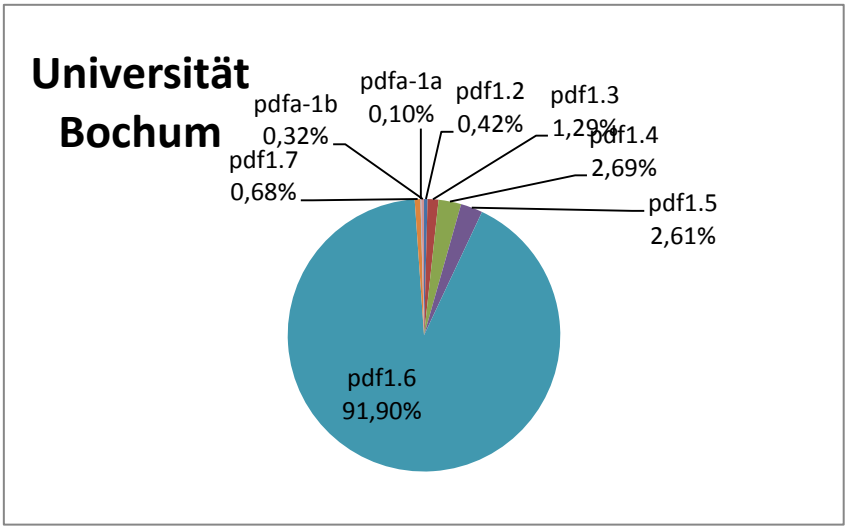

Abbildung 14: Anteil der PDF-Formate an der Universität Bochum

Über 900 Dateien entsprechen nicht der PDF-Referenz, über 200 verwenden nicht eingebettete Schriftarten und knapp 80 weisen den Metadatenfehler auf. Zudem sind 37 Dateien auf irgendeine Weise verschlüsselt, was konträr zu den Leitlinien und Anforderungen der Universität steht. 25 Dokumente beinhalten gerätespezifische Farbräume.

[136] Vgl. Universitätsbibliothek Bochum (2015): Zugelassene Datenformate und Layout für Elektronische Dissertationen

[137] Vgl. ebd.

[138] Vgl. Universitätsbibliothek Bochum (2014): Dissertationen elektronisch publizieren

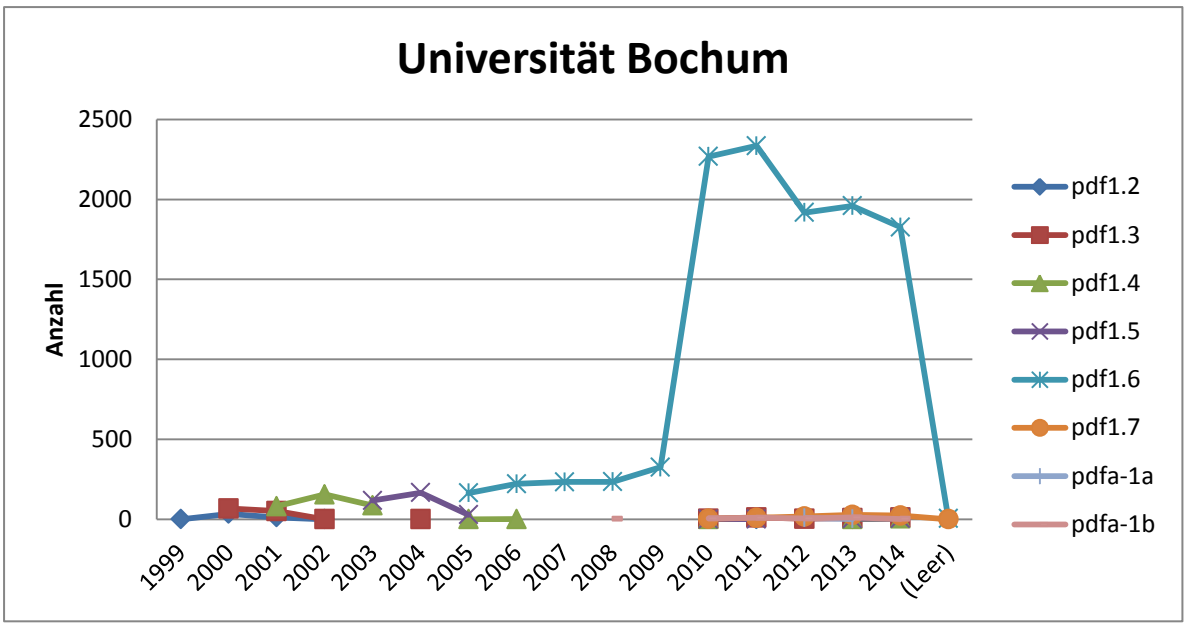

Abbildung 15: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Bochum

4.2.10 Schloss Dagstuhl Leibniz-Zentrum für Informatik

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF[139]
  - Dauer: Langzeitverfügbarkeit[140]
- Hilfestellungen: Keine

4392 Dateien lieferte der „Dagstuhl Research Online Publication Server“ (DROPS), von denen fast 90% im Format 1.4 oder 1.5 sind (Stand: 19.1.2015). Der Anteil an PDF/A-1b-Dateien ist mit 0,07 % (entspricht drei Dateien) verschwindend gering, was in Anbetracht der Tatsache, dass die Einrichtung keine bestimmte PDF-Version verlangt, nicht verwunderlich ist. Dagegen ist trotz der fehlenden Spezifikationen und Hilfestellungen kein „Wildwuchs“ wie in einigen anderen Repositorien festzustellen.

Dies zeigt sich auch in den Fehlermeldungen. Zwar gibt es 668 Referenzfehler, 49 Font-Einbettungsfehler und 12 Metadatenfehler, aber es fehlt die bei anderen Repositorien zu beobachtende Zerstreuung einer recht geringen Anzahl an Fehlermeldungen auf viele verschiedene Arten von Fehlern. Lediglich drei Meldungen über beschädigte Dateien sind erfasst.

[139] Vgl. Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH (2015): Schloss Dagstuhl : About DROPS

[140] Vgl. ebd.

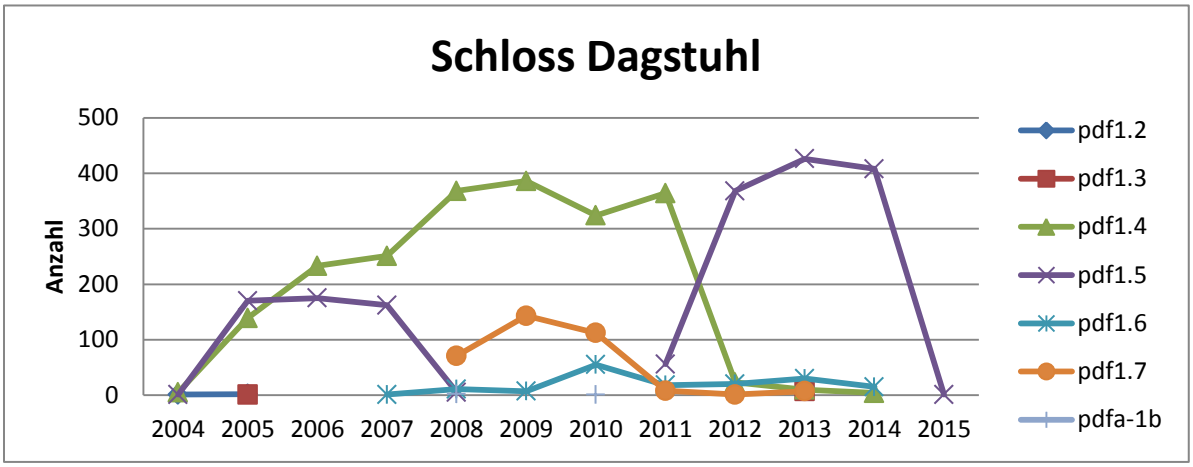

Abbildung 16: Art und letztes Änderungsdatum von PDF-Dateien am Schloss Dagstuhl Leibniz-Zentrum für Informatik

4.2.11 Technische Universität Chemnitz

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF Pflichtformat, PDF/A bevorzugt, zusätzliche Dateien in anderen Formaten möglich[141]
  - Dauer: Garantie für fünf Jahre
- Kurze Anleitungen und Links zu Tools zur Erstellung und Konvertierung von PDF-Dateien, Dokumentvorlagen[142], Link zum Tutorial der Universität Freiburg

Das Repositorium der TU Chemnitz lieferte zunächst weitaus mehr Einträge, als es laut Open DOAR Items hat. Allerdings waren von diesen fast 6963 Einträgen 3186 Stück mit Dateifehlern (Stand: 19.1.2015). Die 3777 übrigen sind zwar auch deutlich mehr als die Anzahl an Items, allerdings ist diese Zahl durchaus plausibel, da zu einer gecrawlten Seite oft mehrere Dateien gehören, teilweise sogar bis über 40 Stück. Dazu zählen u.A. Schaubilder, Diagramme, einzelne Kapitel, Abstracts und Titelseiten.

Mit 1,01% 1a, 3,68% 1b und 0,05% 2b nehmen PDF/A-Dateien einen größeren Anteil der Dokumente ein, als auf anderen Repositorien. Dies mag an den bereitgestellten Hilfen, Anleitungen und Dokumentvorlagen zum Thema PDF-Erstellung liegen. Insgesamt ist die Verteilung der Formate wenig auffällig. Auch die Verläufe der Arten und letzten Änderungsdaten zeigen bis auf eine etwas erhöhte Spitze von PDF 1.2 im Jahr 2010 keine Besonderheiten auf. Ebenso reihen sich die ausgegebenen Fehlermeldungen mit ein: PDF-Referenz (595), Metadaten (128) und nicht eingebettete Schriftarten (105) machen das Gros der Fehler aus. Lediglich die 41 Fehler

[141] Vgl. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - Chemnitz: Hilfe

[142] Vgl. Technische Universität Chemnitz (2014): Universitätsbibliothek: Publizieren

bezüglich der Anmerkungen oder Formularfelder mit mehrdeutiger oder ohne angemessene Erscheinung scheinen erwähnenswert.

4.2.12 Technische Universität Kaiserslautern

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF ohne Kryptofunktionen
  - Dauer: Garantie für fünf Jahre[143]
- Hilfestellungen: Kontakt, Link zum Tutorial der Universität Freiburg, Link zu den Formatvorlagen der Humboldt-Universität zu Berlin, kurze Erklärung zu Metadaten[144]

Ähnlich wie die TU Chemnitz lieferte auch die TU Kaiserslautern zunächst deutlich mehr Einträge. Nach Abzug der fehlerhaften Einträge sowie der Dubletten blieben noch 2787 Dateien zu überprüfen (Stand: 19.1.2015). Davon sind weniger als ein Prozent in PDF/A-Formaten; den Großteil machen die Versionen 1.2 (ca. 43%), 1.3 (ca. 15%) und 1.4 (ca. 20%) aus. Besonders 1.2 und 1.3 hatten in den Jahren von 1996 bis 2001 den größten Zuwachs. Danach glichen sich die Anzahlen modifizierter Dateien weitestgehend an. Ausnahme hierbei bildet die Spitze von PDF 1.2 im Jahr 2012.

Die verlinkten Tutorials und Formatvorlagen haben sich nicht merklich auf die Zahlen ausgewirkt. Zumindest aber gibt es keine verschlüsselten Dateien, wie es die Leitlinien und das DINI-Zertifikat vorsehen. Dafür zeigen sich wieder Fehler bei der PDF-Referenz (228), eingebetteten Schriftarten (80), beschädigten Dateien (42) und Metadaten (41). Weitere Fehler wurden nicht ermittelt.

4.2.13 Universität des Saarlandes

- DINI-Zertifikat: 2004
- LZA-Richtlinien:
  - Format: PDF ohne Kryptofunktionen, PDF/A empfohlen[145]
  - Dauer: Garantie für fünf Jahre
- Hilfestellungen: Erläuterungen zu Metadaten[146], Links zu Tools zur PDF-Erstellung[147], Link zum Tutorial der Universität Freiburg

Die Universität des Saarlandes lieferte 5293 Einträge (Stand: 19.1.2015). Auch hier bleibt der Anteil an PDF/A-Formaten unter einem Prozent. Im Jahr 2014 wurden offenbar keine Dateien in einem der Archiv-Formate eingereicht. Bis auf eine Spitze von 18 Dokumenten im Jahr 2011 bleiben die Zahlen mit maximal vier Dokumenten sehr niedrig. Der erhöhte Wert von 2011 erklärt sich durch eine insgesamt erhöhte Anzahl von Dokumenten, auch bei anderen Formaten.

[143] Vgl. Universitätsbibliothek Kaiserslautern (2014): KLUEDO | Leitlinien
[144] Vgl. Universitätsbibliothek Kaiserslautern (o.J.): KLUEDO | Hilfe
[145] Vgl. Universität des Saarlandes (o.J.): SciDok faq
[146] Vgl. Universität des Saarlandes (o.J.): SciDok-Hilfe
[147] Vgl. Universität des Saarlandes (o.J.): Vorbereitung elektronischer Dokumente

Der Großteil der Dateien sind in den Formaten 1.4 (33,52%), 1.5 (16,78%), 1.6 (16,69%), 1.3 (15,48%) oder 1.2 (9,71%). Der Verlauf der Modifikationsdaten lässt erkennen, dass seit 2004, dem Jahr der DINI-Zertifizierung, insgesamt mehr Dateien erstellt / bearbeitet und eben auf das Repositorium geladen wurden, mit einer Hochphase von 2008 bis 2011.

Von den 165 vom Validator ausgegebenen Metadatenfehlern gehörte wieder keiner zu den Dateien mit merkwürdigen Jahreszahlen als Modifikationsdatum (z.B.1942 oder 2711). 139 Dateien enthalten nicht eingebettete Schriftarten; 388 entsprechen nicht dem angegeben Format. Keine der Dateien ist verschlüsselt, was konform der Richtlinien des Repositoriums ist.

4.2.14 Universität Duisburg-Essen

- DINI-Zertifikat: 2007
- LZA-Richtlinien:
    - Format: „verbreitet[es]“ Format, z.B. PDF oder HTML, ohne Kryptofunktionen, Empfehlung von PDF/A und weiteren, optional Originalformat[148]
    - Dauer: Dauerhafte Archivierung und Verfügbarkeit, Garantie für fünf Jahre
- Hilfestellungen: Feste Ansprechpartner, individuelle Autorenschulungen möglich[149]

Abzüglich der 3291 Dateifehler zählt das Repositorium der Universität Duisburg-Essen 10633 Einträge (Stand: 19.1.2015). Das ist bedeutend mehr, als die 2355 angegebenen Items bei Open DOAR. Der Grund hierfür liegt darin, dass viele Werke sowohl im Ganzen, als auch kapitelweise veröffentlicht wurden.

Zusammen sind die PDF/A-Varianten 1a (1,01%), 1b (2,34%) und 2b (0,06%) zu etwa 3,4% vertreten. Damit ist der Anteil größer, als bei den meisten anderen Repositorien. Dies lässt sich darauf zurückführen, dass die Universität die Verwendung von PDF/A empfiehlt, die zusätzliche Einreichung der Arbeiten im Original und damit die kontrollierte Konvertierung nach PDF/A ermöglicht und individuelle Autorenschulungen anbietet. Der erhöhte Anteil von 1b rührt wahrscheinlich von der geringen Komplexität und den Exportfunktionen der meisten Textverarbeitungsprogramme. Die größten Anteile haben aber die normalen PDF-Formate 1.4 (34,41%), 1.3 (23,14%) und 1.2 (18,35%).

Betrachtet man die Jahre der letzten Modifikation stellt man fest, dass bis zum Jahr 2008 weniger Formate, dafür in größerer Anzahl bearbeitet wurden. Dominant sind dabei die oben genannten, normalen PDF-Formate. Diese weisen Spitzen zu jeweils anderen Zeiten auf, was auf die Versionsgeschichte und die damalige Verbreitung der Formate zurückzuführen ist. Nach einem gemeinsamen Hoch im Jahr 2003 und einem gesamten Tief im Jahr 2006 folgt im Jahr der DINI-Zertifizierung 2007 eine Spitze bei den Formaten 1.3 und 1.4; ab 2008 fallen die Werte zurück auf das eher niedrige Niveau der anderen Formate, wobei das Format 1.5 einen dezenten Aufwärtstrend verzeichnet.

[148] Vgl. Universität Duisburg-Essen (o.J.): DuEPublico: Leitlinien

[149] Vgl. Universität Duisburg-Essen (o.J.): DuEPublico Ansprechpartner und Infos für Autoren

Gemäß den Richtlinien ist keine Datei verschlüsselt. Die Fehler zeichnen ein ähnliches Gesamtbild wie die meisten anderen Repositorien. Beinah eintausend Dateien entsprechen nicht der PDF-Referenz (davon sind 114 PDF/A-Dateien), knapp 300 haben Schriftarten nicht eingebettet, 132 haben falsche Metadaten. Wiederum sind nicht die Dateien mit offensichtlich falschen Angaben im Feld „LastModification“ betroffen. Hinzu kommen diverse beschädigte Dateien, spezifische Farbräume, Schriftarten ohne Unicode-Mapping und vereinzelt weitere.

4.2.15 Universität Potsdam

- DINI-Zertifikat: 2007
- LZA-Richtlinien:
    - Format: PDF/A-1 oder PDF/X-3 / PDF/X-4[150], HTML, zusätzlich alle Originalformate
    - Dauer: Garantie für fünf Jahre, Langzeitverfügbarkeit für XML[151]
- Hilfestellungen: Erläuterung verschiedener Präsentations- und Originalformate[152], ausführliches Tutorial zur Erstellung und Konvertierung von PDF- und PDF/A-Dateien mit verschiedenen Textverarbeitungsprogrammen mit Links zu Tools und Plugins[153], Erläuterung zu Metadaten[154], Kontakt

Der Publikationsserver der Universität Potsdam lieferte 6489 Einträge abzüglich Dateifehler (Stand: 20.1.2015). Hier zeigt sich ein echter Ausnahmewert: beinahe 48% der Dateien geben an, PDF/A-Dateien zu sein. Dies hängt eindeutig damit zusammen, dass die Universität ausdrücklich PDF/A-1 oder PDF/X fordert. Zusätzlich werden die Originalformate verlangt, sodass daraus ggf. PDF-Dateien erzeugt werden können. Zusätzlich werden ausführliche Erläuterungen und Anleitungen sowie Links zu benötigten Tools bereitgestellt, die es den Autoren erleichtern, PDF/A-Dateien zu erstellen.

Relativiert wird die beeindruckend hohe Zahl dadurch, dass die Metadaten- und PDF-Referenzfehler auf sehr viele der PDF/A-Dateien zutreffen. Das muss nicht bedeuten, dass all diese Dateien tatsächlich nicht im jeweiligen Format vorliegen, es zeugt aber von einer erhöhten Wahrscheinlichkeit, dass sie nicht vollkommen valide sind. Ergänzend muss angemerkt werden, dass es keine Hinweise darauf gibt, dass der verwendete 3-Heights-PDF-Validator PDF/X-Dateien erkennen kann, wodurch diese Fehlermeldungen entstehen könnten. Auffällig ist zudem, dass PDF/A-Dateien erst seit 2007 – dem Jahr des Erwerbs des DINI-Zertifikats – modifiziert bzw. erstellt wurden. Seitdem ist PDF/A-1b fast durchgehend das dominierende Format.

Dass sich trotz der strikten Richtlinien noch immer normale PDF-Dateien auf dem Server befinden, die erst in der jüngeren Vergangenheit modifiziert oder erstellt wurden, kann zum einen

[150] Vgl. Universität Potsdam (o.J.): Formate
[151] Vgl. Universität Potsdam (o.J.): Leitlinien
[152] Vgl. Universität Potsdam (o.J.): Formate
[153] Vgl. Universität Potsdam (o.J.): Tutorial zur PDF-Erstellung
[154] Vgl. Universität Potsdam (o.J.): OPUS-Hilfe: Was sind Metadaten?

daran liegen, dass die Vorgaben etwa für Interne nicht so strikt gelten (oder einfach nicht eingehalten werden) oder die Überprüfung der Dateien schlichtweg nicht alle unpassenden Formate erkennt.

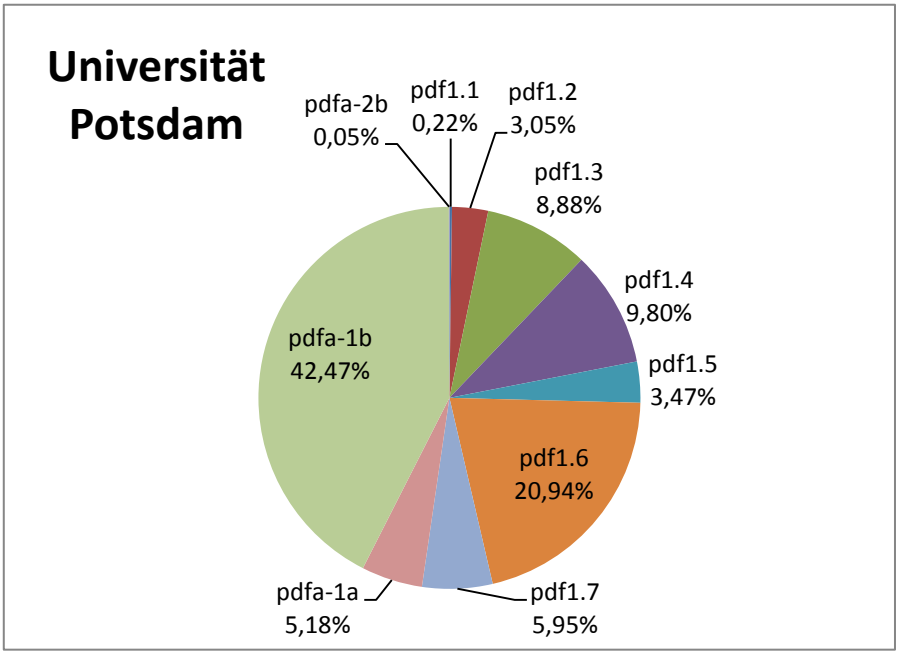

Abbildung 17: Anteil der PDF-Formate an der Universität Potsdam

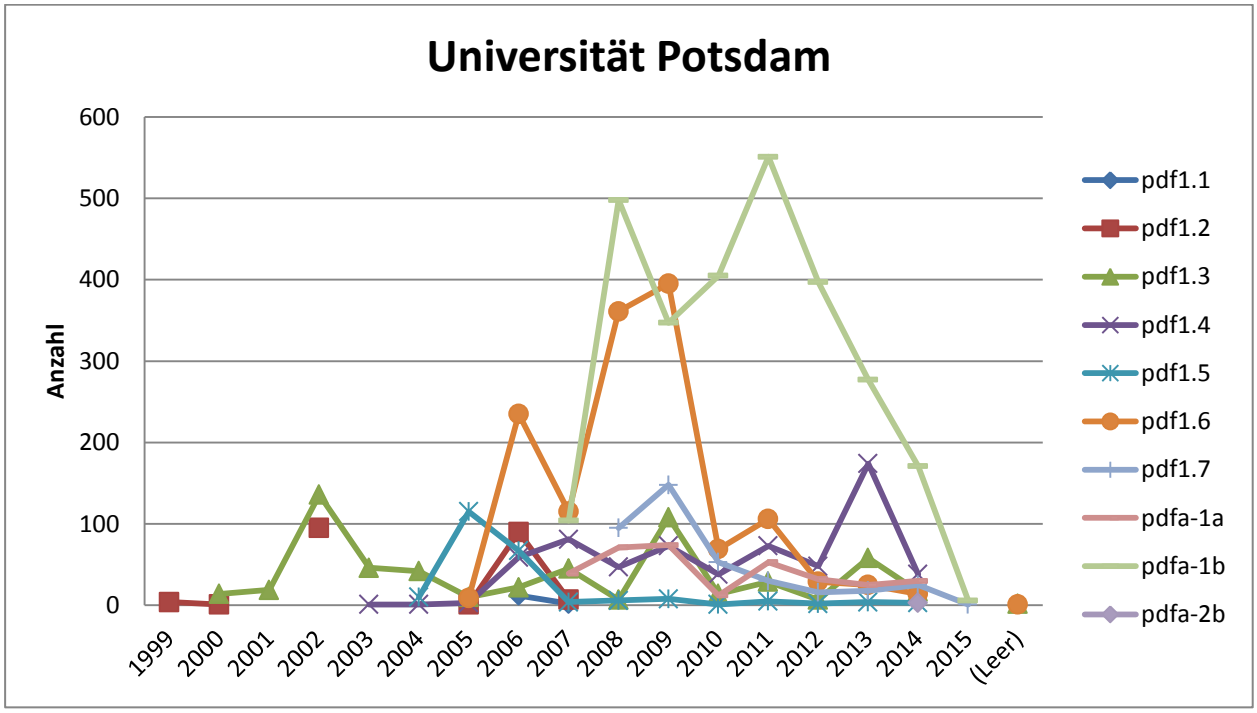

Abbildung 18: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Potsdam

4.2.16 Universität Siegen

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: PDF, ohne Kryptofunktionen, zusätzlich Originalformat[155]
  - Dauer: mindestens zehn Jahre[156]
- Hilfestellungen: Anmerkungen[157] und Links[158] zur Erstellung von PDF-Dateien

Abzüglich Dubletten ergab der Publikationsserver der Universität Siegen 784 Einträge (Stand: 19.1.2015). Mit ca. 45% sind die meisten davon im Format 1.4. Darauf folgen 1.6 mit etwa 20% und 1.3 und PDF/A-1b mit jeweils etwa 10%. Die PDF/A-Formate sind wie folgt vertreten: 0,51% 1a, 10,2% 1b, 0,13% 2a und 2,55% 2b. Der etwas erhöhte Anteil könnte darauf basieren, dass die Universitätsbibliothek zusätzlich zur PDF-Datei das Originalformat annimmt, um daraus ggf. eine PDF-Datei zu erzeugen.

Die gesammelten Dateien lassen bei den eingereichten Dokumenten einen Trend von der Dominanz eines Formats hin zu einer Pluralität der Formate, besonders in den letzten zwei Jahren erkennen. Dabei blieb die Anzahl der eingereichten Dokumente aber ähnlich.

Wieder sind PDF-Referenz, eingebettete Schriftarten und mangelhafte Metadaten die häufigsten Fehler. Hinzu kommen neben vereinzelten anderen Fehlern aber noch die Verwendung gerätespezifischer Farbräume (16 Meldungen) und Transparenz (19 Meldungen).

4.2.17 Universität Stuttgart

- DINI-Zertifikat: 2004
- LZA-Richtlinien:
  - Format: PDF ohne Kryptofunktionen[159]
  - Dauer: Mindestens fünf Jahre
- Hilfestellungen: Links zu Tools zur Erstellung von PDF-Dateien[160], Link zum Tutorial der Universität Freiburg

9692 Einträge ergab der Hochschulschriftenserver der Universität Stuttgart (Stand: 21.1.2015). Die Anteile der normalen PDF-Formate (ausgenommen 1.1) sind relativ gleichmäßig verteilt; Version 1.4 und 1.6 nehmen aber mit jeweils ca. 24% den größten Teil ein. Die drei Archivformate 1a, 1b und 2b bleiben zusammen unter einem halben Prozent.

Der Verlauf der modifizierten Dateien pro Format lässt einen starken Aufschwung von 1.6 im Jahr 2009 erkennen, welcher ab 2011 wieder abflacht; dabei wird das Format von 1.5 als füh-

[155] Vgl. Universitätsbibliothek Siegen (o.J.): Informationen für Autoren
[156] Vgl. Universitätsbibliothek Siegen (o.J.): Leitlinien des Publikationsservers OPUS Siegen
[157] Vgl. Universitätsbibliothek Siegen (o.J.): PDF-Dokumente erstellen
[158] Vgl. Universitätsbibliothek Siegen (o.J.): Technische Anforderungen
[159] Vgl. Universitätsbibliothek Stuttgart (2011): OPUS-FAQ
[160] Vgl. Universitätsbibliothek Stuttgart (2011): Vorbereitung elektronischer Dokumente

rendes allmählich abgelöst. Dessen Werte stiegen seit 2008 kontinuierlich, während die aller anderen seit 2011 sinken. Was dieses „Comeback“ von 1.5 verursacht ist nicht ersichtlich.

Bei den Fehlermeldungen gleicht der Dokumentenserver den meisten anderen Repositorien. Es gibt keine verschlüsselten Dokumente, was mit den Vorgaben d’accord geht. Eine Auswirkung der DINI-Zertifizierung oder der Hilfestellungen ist nicht zu erkennen.

4.2.18 Universität Ulm

- DINI-Zertifikat: 2004
- LZA-Richtlinien:
  - Format: PDF[161], möglichst 1.4[162]
  - Dauer: prinzipiell unbegrenzt, Garantie für mindestens zehn Jahre (Archivierungsdauer) und fünf Jahre „Online-Verbreitung“[163]
- Hilfestellungen: Helpdesk, Überblick über Metadaten, Konvertierung in PDF kann von der Bibliothek vorgenommen werden[164], Erläuterungen und Links zu Tools zur PDF-Erstellung[165], Link zum Tutorial der Universität Freiburg

Das Repositorium der Universität Ulm ergab 3422 Einträge, von denen nur etwa ein halbes Prozent PDF/A-Dateien, nämlich 1a und 1b sind (Stand: 20.1.2015). Die übrigen verteilen sich hauptsächlich auf die Versionen 1.2 bis 1.6; mit ca. 34% ist 1.4 am häufigsten vertreten.

Seitdem das Repositorium 2004 das DINI-Zertifikat erhalten hat ist die Anzahl der modifizierten bzw. erstellten Dateien insgesamt deutlich gestiegen. Im letzten Jahr fielen aber die Werte für alle Formate außer 1.6 deutlich ab; dieses erreichte einen neuen Höchstwert. Möglich ist, dass zu diesem Zeitpunkt das Angebot der Bibliothek, die Originaldateien in ein passendes PDF-Format zu konvertieren, begonnen und gut angenommen wurde. Dies widerspricht allerdings der Vorgabe seitens der Universität, möglichst Version 1.4 einzureichen, welches bis 2013 das führende Format war. Die vom Validator ausgegebenen Fehlermeldungen ähneln in Art und Häufigkeit denen der meisten anderen Repositorien.

[161] Vgl. Universität Ulm (o.J.): VTS | FAQ
[162] Vgl. Universität Ulm (o.J.): VTS | PDF-Erstellung
[163] Vgl. Universität Ulm (o.J.): VTS | Leitlinien
[164] Vgl. Universität Ulm (o.J.): VTS | Akzeptierte Dateiformate
[165] Vgl. Universität Ulm (o.J.): VTS | PDF-Erstellung

4.2.19 Virtuelle Fachbibliothek Fachinformation für Politikwissenschaft, Verwaltungswissenschaft und Kommunalwissenschaften (ViFaPol)

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF, PDF/A empfohlen[166]
  - Dauer: „mittel- und langfristige Verfügbarkeit“[167]
- Hilfestellungen: Keine

Es wurden 5219 Dateien aus der ViFaPol, einem disziplinären Repositorium, heruntergeladen (Stand: 21.1.2015). Nur neun Stück davon sind PDF/A-Dateien; die restliche Verteilung erstreckt sich hauptsächlich über die PDF-Versionen 1.2 bis 1.6. Dies verwundert nicht, da das Repositorium keinerlei Hilfestellungen gibt und PDF/A nur empfiehlt, aber nur PDF verlangt.

Die Modifikationsdaten spiegeln weitestgehend die Verbreitung der Formate 1.4 und 1.6 wider; im letzten Jahr fielen aber die Werte aller Formate deutlich ab. Die Fehlermeldungen weisen insofern eine kleine Überraschung auf, dass mit 499 Erscheinungen der Fehler mangelhafter Metadaten am häufigsten ist; dahinter rangieren die Nichtübereinstimmung mit der angegebenen PDF-Version (392) und nicht eingebettete Schriftarten. Weitere Fehler traten vereinzelt auf.

4.2.20 Virtuelle Fachbibliothek Psychologie (PsyDok)

- DINI-Zertifikat: 2004
- LZA-Richtlinien:
  - Format: PDF vorgesehen, PDF/A empfohlen, Ablehnung von Dateien mit Kryptofunktionen [168]
  - Dauer: Minimum fünf Jahre[169]
- Hilfestellungen: Informationen zu Metadaten[170], Leitfaden zur technischen Vorbereitung der Dokumente[171]: Link zum Tutorial der Universität Freiburg, Link zu Software zur PDF-Erzeugung / -Konvertierung

Das disziplinäre Repositorium PsyDok lieferte 3499 Dateien, von denen 53,02% PDF 1.7 entsprechen (Stand: 19.1.2015). Darauf folgen die Versionen 1.4 (21,03%) und 1.3 (14,12%). Trotz der Empfehlung von PDF/A sind nur vier Dokumente in einem der Archivformate.

In den Jahren 2011 bis 2013 ist eine sehr große Spitze bei der Anzahl an Modifikationen beim 1.7-Format zu erkennen, deren Ursache nicht eindeutig festzumachen ist. Wiederum sind Digitalisierungs- und Konvertierungsprojekte denkbar. Ein Zusammenhang mit den veröffentlichten

[166] Vgl. Staats- und Universitätsbibliothek Hamburg (2013): Unsere Leitlinien : eDoc.ViFaPol
[167] Vgl. Staats- und Universitätsbibliothek Hamburg (2013): Veröffentlichen : eDoc.ViFaPol
[168] Vgl. Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok FAQ
[169] Vgl. Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok - Policy
[170] Vgl. Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok-Hilfe
[171] Vgl. Saarländische Universitäts- und Landesbibliothek (o.J.): Vorbereitung elektronischer Dokumente

Hilfestellungen ist in Anbetracht der Schwere und Plötzlichkeit des Ausschlages eher unwahrscheinlich.

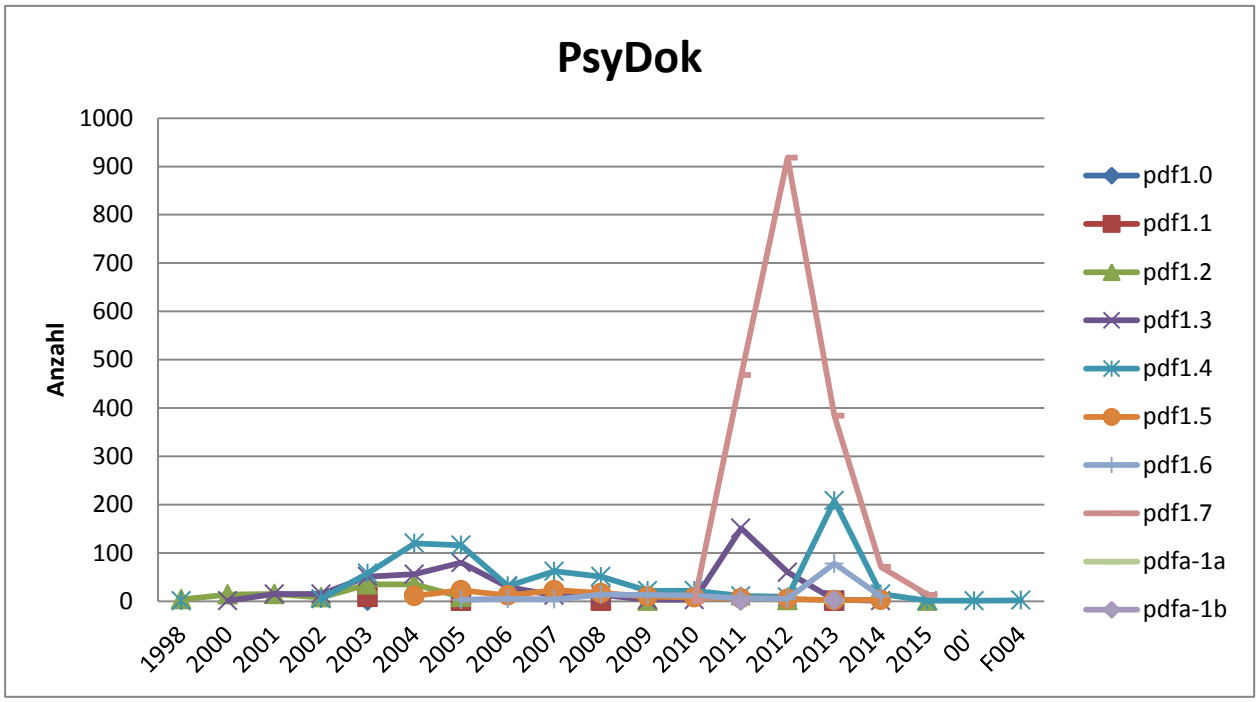

Abbildung 19: Art und letztes Änderungsdatum von PDF-Dateien bei der Virtuellen Fachbibliothek Psychologie (PsyDok)

Auch bei PsyDok wurden Dateien mit fehlerhaften Einträgen im Feld LastModified (z.B: 00‘ oder F004) nicht mit einem Metadatenfehler versehen. Mit 174 mal ist der PDF-Referenz-Fehler wieder der häufigste.

4.2.21 Westfälische Wilhelms-Universität Münster

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF/A-1 als Mindestanforderung[172], PDF/A-1b ausreichend, PDF/A-1a erwünscht[173]
  - Dauer: „Langzeitarchivierung“[174]
- Hilfestellungen: Informationen zu PDF/A und eigene Anleitungen zur Erstellung sowie Link zur Humboldt-Universität zu Berlin[175]

Das Repositorium der Universität Münster lieferte 7731 Dateien (Stand: 19.1.2015). Auch hier rührt die erhöhte Zahl daher, dass Werke teils kapitelweise veröffentlicht wurden. Wenn auch nur mit zwei Dateien, so ist mit PDF/A-3b doch erstmals ein Vertreter des PDF/A-3-Formats

[172] Vgl. Universitäts- und Landesbibliothek Münster (2013): Dateiformate - Publizieren an der WWU
[173] Vgl. ebd.
[174] Vgl. Universitäts- und Landesbibliothek Münster (2013): WWU-Publikationsserver - Leitlinien
[175] Vgl. Universitäts- und Landesbibliothek Münster (2013): Datei-Erstellung - PDF/A

vorhanden. Von 2b gibt es immerhin zehn Dateien (0,13%), von 1a 94 Stück (1,22%) und von 1b sogar 973 Dateien, die einen Anteil von 12,59% an allen Dateien ausmachen. Den Löwenanteil liegt aber bei den Versionen 1.3 und 1.4 mit etwa 30% bzw. 26%. Dahinter rangieren 1.2, 1.6, 1.5 und 1.7.

Besonders auffällig ist, dass im Jahr 2014 über 750 PDF/A-1b-Dateien geändert bzw. erstellt wurden. Dies ist wahrscheinlich eine direkte Folge der im Jahr 2013 verabschiedeten Vorgabe, PDF/A-1 als mindesten Standard zu fordern: 1a ist gewünscht, 1b ist aber ausreichend. Die Beliebtheit des letzteren wird auf die Exporttools einschlägiger Textverarbeitungsprogramme zurückgeführt. Zudem bietet die Universität entsprechende Anleitungen an. Dass dennoch Nicht-PDF/A-Dateien verändert und wohl veröffentlicht wurden lässt sich zumindest teilweise durch Fehler bei der Validation erklären.

Die aufgetretenen Fehler (558 falsche PDF-Referenzen, 167 nicht eingebettete Fonts und134 mangelhafte Metadaten) zeichnen ein ähnliches Bild wie bei anderen Repositorien. Einzig die 59 beschädigten Dateien können als kleine Ausnahme gelten.

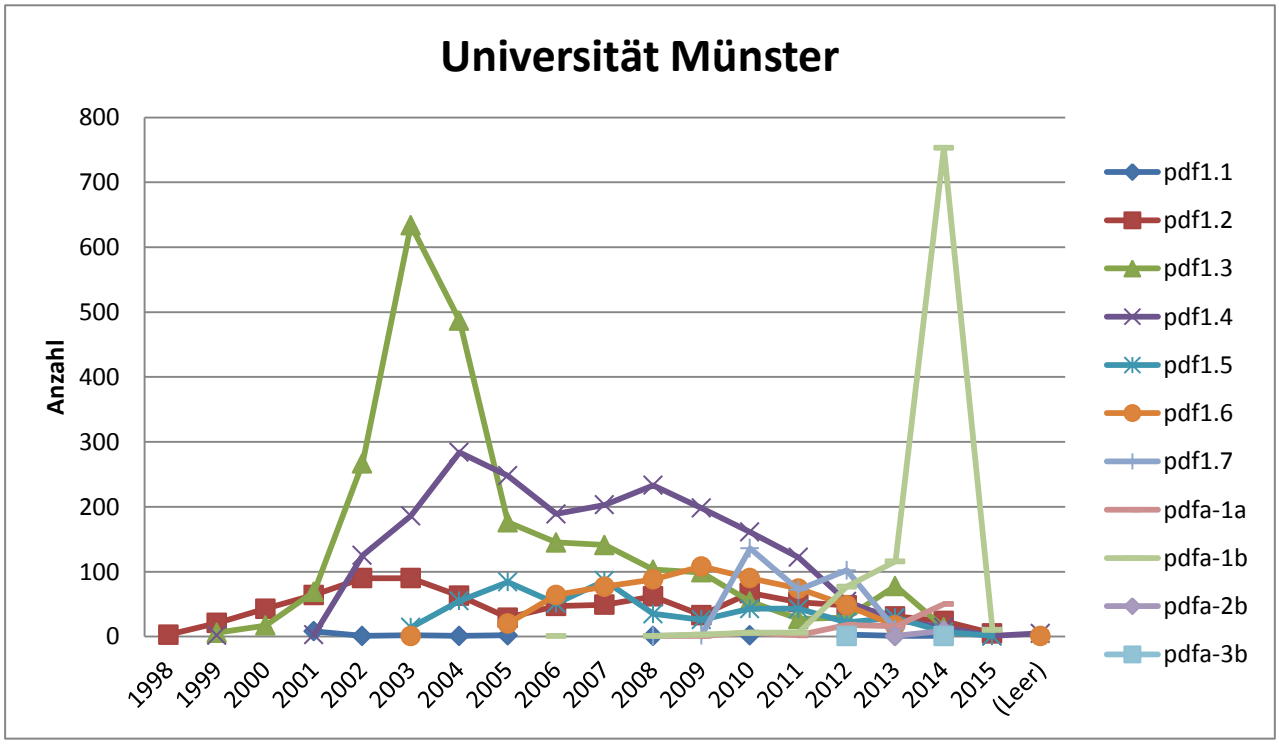

Abbildung 20: Art und letztes Änderungsdatum von PDF-Dateien an der Universität Münster

4.2.22 Hochschule Heilbronn

- DINI-Zertifikat: Keins
- LZA-Richtlinien:
  - Format: PDF
  - Dauer: Keine Angabe
- Hilfestellungen: Überblick über Metadaten, Kontakt[176]

Abzüglich der Dubletten und fehlerhaften Einträge in der Ergebnistabelle lieferte der OPUS-Server der Hochschule Heilbronn 70 Dateien, von denen über die Hälfte im Format 1.5 ist; der Rest besteht aus 1.4 und einzelne Dateien in 1.3 (Stand:19.1.2015). Wegen der geringen Anzahl Dateien sind die meisten Informationen wenig aussagekräftig. Allerdings zeigen sich bei den ausgegebenen Fehlermeldungen ähnliche Tendenzen wie bei anderen Repositorien. Dies könnte ein Indiz dafür sein, dass sich gerade beim PDF-Referenz-Fehler eine weitere Schwäche des 3-Heights-PDF-Validators offenbart.

4.2.23 Humboldt-Universität zu Berlin

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
  - Format: Dissertationen in PDF[177], Habilitationsschriften[178] sowie Magister-, Bachelor- und Diplomarbeiten[179] in PDF/A
  - Dauer: Garantie von fünf Jahren für PDF/A-Dateien[180]
- Hilfestellungen: Arbeitsgruppe „Elektronisches Publizieren“, Links zu Tools zur Konvertierung[181], Konvertierung kann ggf. Computer- und Medienservice / Bibliothek (bei hohem Aufwand gebührenpflichtig) übernehmen, Anleitungen zur Erstellung von PDF-Dateien[182], Erläuterungen zu Anforderungen und Empfehlungen an PDF-Dateien[183]

Abzüglich der 14446 Einträge mit Dateifehlern ergab der Publikationsserver der Humboldt-Universität 16785 überprüfbare Einträge, die auch der bei Open DOAR angegebenen Anzahl an Items entsprechen (Stand: 22.1.2015). Mit etwas mehr als elf Prozent liegt der Anteil an PDF/A-Dateien höher als bei vielen anderen Dokumentenservern (2,17% 1a, 9,1% 1b). Die weiteren Dateien sind hauptsächlich in den Formaten 1.2 (ca. 22%), 1.3 (ca. 15%) und 1.4 (ca. 35%).

---

[176] Vgl. Hochschule Heilbronn (o.J.): OPUS 4 | Hilfe

[177] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Dissertationen: Was muss eingereicht werden?

[178] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Habilitationsschriften: Was muss eingereicht werden?

[179] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Magister- und Diplomarbeiten: Was muss eingereicht werden?

[180] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Leitlinien

[181] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): FAQ: Elektronisches Publizieren

[182] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Technische Hinweise zur Erstellung und Korrektur von PDF-Dokumenten

[183] Vgl. Arbeitsgruppe 'Elektronisches Publizieren' (2014): Anforderungen an die technische Qualität des PDF-Dokumentes einer Dissertation

Der erhöhte Anteil von PDF/A-Dateien ist eine Auswirkung der Vorgabe, dass Habilitationsschriften sowie Magister-, Bachelor- und Diplomarbeiten in entsprechenden Formaten eingereicht werden müssen, im Zusammenspiel mit dem Konvertierungsservice und den bereitgestellten Anleitungen und Hintergrundinformationen.

Besonders auffällig sind die extrem hohen und schlagartig angestiegenen Zahlen an bearbeiteten bzw. erstellten Dateien im Format 1.4 im Jahr 2003 (über 2300 Stück) und für die Formate 1.2 und 1.3 im Jahr 2009 (über 2600 bzw. über 1000 Stück). Da die Wachstumsstatistik bei Open DOAR eine Berichtslücke in den Jahren 2009 bis 2011 aufweist, kann nicht gesagt werden, ob die gesamte Anzahl an Dokumenten im Jahr 2009 unverhältnismäßig stark angestiegen ist oder nicht. Wäre sie angestiegen, würde dies neu eingestellte Dokumente bedeuten, was auf ein Migrations- oder Digitalisierungsprojekt schließen ließe. Wäre die Zahl der normalen Wachstumsrate gefolgt, hätte dies auf eine Konvertierung alter Dateien von verschiedenen Formaten in die genannten schließen lassen können, was ggf. eine Maßnahme im Zuge des Erwerbs des DINI-Zertifikat im Folgejahr 2010 gewesen sein könnte.

Bei den Fehlermeldungen stellt die Humboldt-Universität keine wirkliche Ausnahme dar. Zusätzlich zu den generell weit verbreiteten Fehlern bei angegebenem Format, fehlerhaften Metadaten und eingebetteten Schriftarten ist nur die etwas erhöhte Anzahl an beschädigten Dateien nennenswert.

4.2.24 Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden

- DINI-Zertifikat: 2010
- LZA-Richtlinien:
    - Format: PDF ohne Kryptofunktionen, zusätzlich Originalformat möglich[184]
    - Dauer: keine zeitliche Beschränkung, langfristige Archivierung, mindestens fünf Jahre
- Hilfestellungen: Kurze Anleitungen und Links zu Tools zur Erstellung und Konvertierung von PDF-Dateien[185]

Das Repositorium der SLUB Dresden lieferte 16364 verwertbare Einträge; von den ursprünglichen ca. 32000 Einträgen wies also etwa die Hälfte den bekannten Fehler auf, dass die entsprechende Datei nicht vorhanden sei (Stand: 23.1.2015). Die Anzahl der ausgewerteten Dateien entspricht aber ungefähr der bei Open DOAR angegebenen Zahl an Items.

Die Verteilung der Dokumente auf die verschiedenen PDF-Versionen hält keine Überraschungen bereit. Am stärksten sind 1.4 (ca. 29%), 1.6 (ca.20%) sowie 1.3 und 1.5 (jeweils ca. 18%). PDF/A-Formate sind mit etwas über vier Prozent vertreten, hauptsächlich 1b (3,25%) und 1a (0,91%). Auch der Verlauf der Modifikationsdaten zeigt kein ungewöhnliches Bild.

---

[184] Vgl. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - SLUB: Publizieren

[185] Vgl. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - SLUB: Technische Hinweise zur PDF-Erstellung

Dafür ist der Metadatenfehler mit 1465 mal fast genauso oft aufgetreten wie der PDF-Referenz-Fehler. Dahinter folgt wie gewöhnlich der Fehler nicht eingebetteter Schriftarten vor vereinzelten anderen Mängeln. Der Validator versah Dateien mit Jahresangaben wie etwa „6+01“ wieder nicht mit einer Fehlermeldung. Da nur eine verschlüsselte Datei festgestellt werden konnte entspricht das Repositorium dahingehend seinen Leitlinien und Vorgaben.

4.3 Gesamtauswertung

Insgesamt wurden 154.964 Dateien von 24 Repositorien ausgewertet. Davon stammen etwa 110.000 von den 15 DINI-zertifizierten Repositorien und etwa 40.000 von den neun nicht-zertifizierten.[186] Der Anteil von PDF/A-Formaten liegt insgesamt bei ungefähr 7,7%. Bei Repositorien mit Zertifikat liegt dieser Wert bei etwa 9,6%, bei denen ohne bei etwa 3%. Das häufigste Archiv-Format ist Version 1b, welches die geringsten Anforderungen stellt und von allen gängigen Export-, Erstellungs- und Konvertierungstools unterstützt wird. Die Version 3b kommt nur bei den Repositorien ohne Zertifikat vor, wohingegen Version 2a nur bei denen mit Zertifikat vorkommt. Die Formate PDF/A-2u PDF/A-3a und PDF/A-3u sind überhaupt nicht vorzufinden.

Generell sind die PDF-Formate 1.4 und 1.6 am stärksten vertreten. Bei zertifizierten Repositorien liegt die ältere Version 1.4 mit ca. 31% vorn, bei den anderen ist es 1.6 mit etwa 40%. Daher weisen diese Repositorien auch anteilig weniger alte Formate auf. Allerdings zählt zu diesen Repositorien auch das der Universität Bochum, welches zu über 90% Version 1.6 beherbergt.

Da für viele Repositorien oft die absoluten Bestandszahlen fehlen, lässt sich nur schwer eine Aussage darüber treffen, ob alte Dateien in neue Formate konvertiert werden. Die oft beobachteten Spitzen in der Veränderung neuerer Formate in den betreffenden Jahren würden bei einer normalen Wachstumsrate des Gesamtbestands auf Konvertierungsprojekte schließen lassen. In Anbetracht der Tatsache, dass die meisten Repositorien aber nur einige Jahre der Verfügbarkeit garantieren, erscheint die Vermutung plausibel, dass die hinzugefügten Dateien zu nicht unerheblichen Teilen aus Migrationen oder Digitalisierungen stammen. Dies schließt Konvertierungsprojekte allerdings nicht gänzlich aus.

Es zeigt sich, dass bis zur ersten Verwendung von neuen PDF/A-Formaten ungefähr so viel Zeit vergeht wie bei normalen PDF-Formaten, nämlich etwa zwei bis drei Jahre. Auch wenn die absoluten Zahlen für das letzte Jahr ein wenig gesunken sind, kann man sagen, dass die Archivformate recht gut angenommen werden. Allerdings ist aus Perspektive der Langzeitarchivierung doch zu wünschen, dass die Anteile der PDF/A-Formate an den veröffentlichten Dokumenten steigen.

Anhand der Repositorien HU Berlin und Uni Münster lässt sich ablesen, dass bei Verpflichtung zum PDF/A-Format und entsprechenden Informations- und Hilfsangeboten die Rate an LZA-tauglichen Dateien sehr deutlich steigerbar ist. Bei der bloßen Empfehlung der Archivformate bleibt die Hemmschwelle der Autoren noch zu hoch, um sich mit dem Thema zu beschäftigen,

[186] Bei der Gesamtauswertung wurden Einträge mit offensichtlich unrealistischen oder fehlenden Jahreszahlen der Übersicht wegen entfernt.

selbst wenn ausreichend Informationsmaterial angeboten wird. Daher sind nur bei ausdrücklicher Verpflichtung zu PDF/A auch deutliche Unterschiede in der Quantität entsprechender Formate zu erkennen. Ein Unterschied zwischen naturwissenschaftlich-technisch und geisteswissenschaftlich geprägten Institutionen ist nicht zu erkennen.

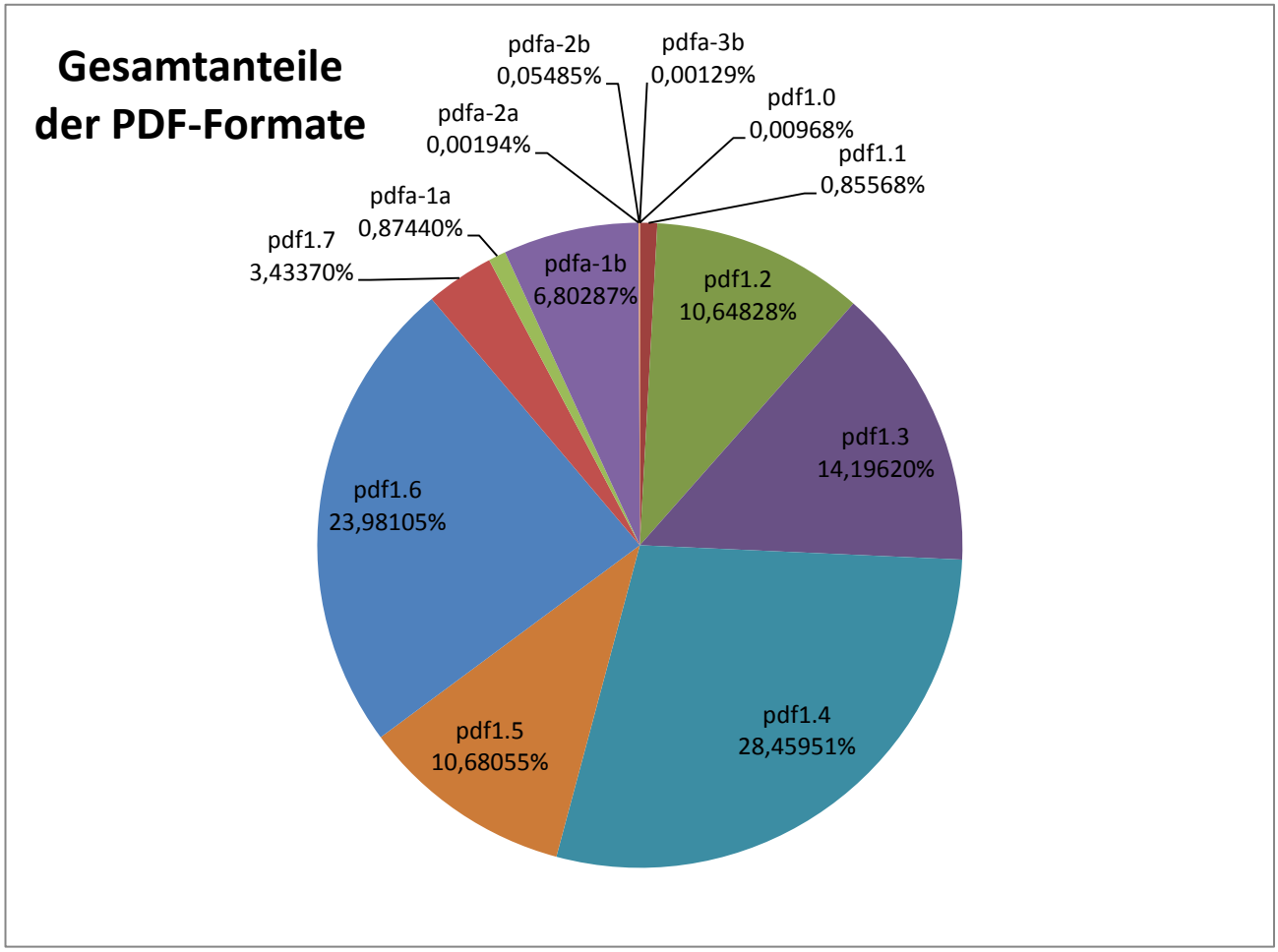

Abbildung 21: Anteil der PDF-Formate insgesamt

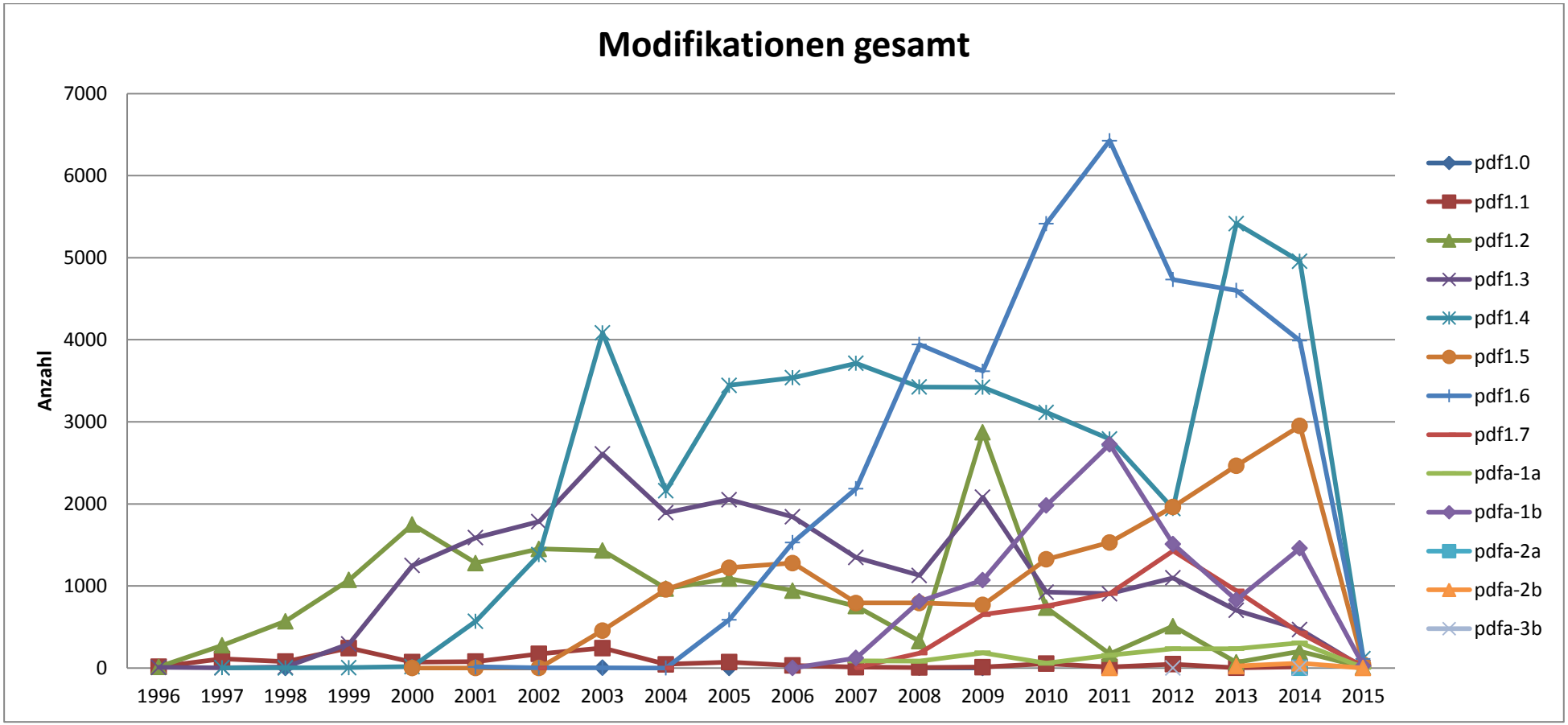

Abbildung 22: Art und letztes Änderungsdatum von PDF-Dateien insgesamt

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Gesamt
pdf1.0			1		2			3		2			3	1		3					15
pdf1.1	16	111	79	244	72	78	173	244	46	72	32	11	6	11	53	12	47	4	15		1326
pdf1.2	14	277	570	1073	1748	1279	1452	1432	968	1090	944	754	327	2874	735	173	511	69	202	9	16501
pdf1.3	8	4	11	294	1250	1588	1785	2608	1893	2052	1843	1347	1128	2081	924	905	1099	704	469	6	21999
pdf1.4		4	2	5	18	570	1382	4086	2159	3445	3537	3713	3424	3422	3115	2790	1943	5416	4956	115	44102
pdf1.5					1	1	1	457	957	1223	1279	792	792	769	1324	1531	1964	2465	2950	45	16551
pdf1.6						15	3	2	1	589	1531	2186	3943	3617	5414	6427	4733	4601	3990	110	37162
pdf1.7							2		2			5	182	652	756	907	1428	941	427	19	5321
pdfa-1a												87	84	186	58	155	236	236	307	6	1355
pdfa-1b											1	123	813	1072	1980	2727	1511	829	1459	27	10542
pdfa-2a																			3		3
pdfa-2b																1		24	59	1	85
pdfa-3b																	1		1		2
Gesamt	**38**	**396**	**663**	**1616**	**3091**	**3531**	**4798**	**8832**	**6026**	**8473**	**9167**	**9018**	**10702**	**14685**	**14359**	**15631**	**13473**	**15289**	**14838**	**338**	**154964**

Tabelle 2: Art und letztes Änderungsdatum von PDF-Dateien insgesamt

Zur Auswertung der Fehlermeldungen ist zu sagen, dass ein Fehlverhalten des PDF-Validators keineswegs ausgeschlossen werden kann. Die vielen Meldungen, dass eine Datei nicht der angeblichen Version entspräche, sind dergestalt erstaunlich, dass sie mit sehr großem Abstand am häufigsten auftreten. Es kann angenommen werden, dass der Validator hierbei besonders empfindlich reagiert und Verstöße gegen die PDF-Referenz meldet, wo eigentlich keine vorliegen. Beispielsweise könnten PDF/A-1b-Dateien wegen kleinerer Unstimmigkeiten für 1.4-Dateien gehalten werden, dem Format auf dem PDF/A-1 basiert. Stimmen die Reporte des Validators aber, so ist ein deutlicher Mangel in der Qualität der Metadaten der Dokumente zu konstatieren, weil damit knapp 14% aller Dateien nicht der PDF-Version entsprächen, die in ihren Metadaten angegeben wird. Dieser Umstand könnte an den Export- und Erstellungstools liegen, die der PDF-Datei eine falsche oder teils auch gar keine Information über die Formatversion mitgeben.

Auf der anderen Seite wäre es aber auch seitens der veröffentlichenden Institutionen ratsam, die eingehenden Dateien besser zu überprüfen und ggf. zu korrigieren. Ob für diesen bislang nicht verpflichtenden Arbeitsschritt Mittel für Zeit- und Personalaufwand sowie Software in die Hand genommen werden, muss jedoch stark angezweifelt werden. Vor dem Hintergrund, dass die Validierung ohnehin nicht vollkommen korrekt funktioniert und – wie im Falle des für diese Arbeit verwendeten 3-Heights-PDF-Validators – ganz offensichtliche Fehler w.z.B. Textstrings in Feldern für Jahreszahlen keine Fehlermeldung erzeugen, ist diese Entscheidung auch sehr gut nachzuvollziehen. Die Registrierung widersprüchlicher Angaben, wie etwa eine PDF-Version, die 2001 auf den Markt kam, aber ein letztes Änderungsdatum von 1999, wäre hierbei auch ein Schritt nach vorn.

Insgesamt gibt es 123682 Dateien, die keinen Fehler aufweisen; davon sind 3233 in PDF/A, was verglichen mit der Anzahl PDF/A-Dateien insgesamt wiederum heißt, dass über 8700 PDF/A-Dateien fehlerhaft und damit nicht einwandfrei LZA-geeignet sind. Auf der positiven Seite steht immerhin, dass mit insgesamt nur 44 Stück kaum verschlüsselte Dateien zu finden sind und dass einige Fehler überhaupt nicht auftauchen, wobei auch hier angemerkt werden muss, dass eine Fehlfunktion des Validators nicht auszuschließen ist.

Fehlermeldung	Anzahl Auftritte
Die Datei ist beschädigt.	1253
Die Datei entspricht nicht dem Format demgegenüber es geprüft wurde.	21106
Die Datei ist verschlüsselt / beinhaltet Kryptofunktionen.	44
Die Datei enthält Farbräume, die nicht auf allen Geräten dargestellt werden können.	1505
Die Datei enthält ungültige Angaben zum Rendering / zur Darstellung.	66
Die Datei enthält alternative Beschreibungen für Bilder.	0
Die Datei enthält PostScript Code.	0
Die Datei verlinkt auf Inhalte, die sich außerhalb der Datei befinden.	12
Die Datei enthält eine oder mehrere Schriftarten, die nicht eingebettet sind oder keine Angaben zur Kodierung haben.	4298
Die Datei enthält eine oder mehrere Schriftarten ohne angemessene Informationen zum Mapping in Unicode.	154
Die Datei enthält transparente Ebenen.	493
Die Datei enthält unbekannte Arten von Anmerkungen	3
Die Datei enthält multimediale Anmerkungen.	0
Die Datei enthält versteckte, unsichtbare, nicht anschaubare oder nicht druckbare Anmerkungen.	210
Die Datei enthält Anmerkungen oder Formularfelder mit mehrdeutigen oder gänzlich ohne angemessene Erscheinung.	89
Die Datei enthält Aktionen / Aktionstypen, die nicht für die Navigation eingesetzt werden.	5
Die Metadaten der Datei sind nicht vorhanden, unstimmig oder beschädigt.	7223
Die Datei liefert nicht ausreichende Information zu seiner logischen Struktur.	356
Die Datei enthält Ebenen.	44

Tabelle 3: Häufigkeit der Fehlermeldungen insgesamt

5 Fazit & Ausblick

Diese Arbeit bietet nur einen sehr begrenzten Einblick in die vorgestellte Thematik, da nur eine Auswahl an Repositorien untersucht werden konnte. Offen bleibt, wie eine Validierung für alle Repositorien in Deutschland oder vielleicht sogar weltweit zu realisieren sein könnte. Um die in dieser Arbeit angefallenen Fehler beim Parsen, Crawlen und Herunterladen zu vermeiden, könnte mit Hilfe von Software wie HTTrack mit lokalen Kopien der Webseiten gearbeitet werden. Voraussetzungen sind dabei natürlich die notwendige Rechenleistung und Speichergröße. Ideal wäre zudem eine Validationssoftware, die nicht nur Dateien gegenüber einem bestimmten Schema abprüfen und sagen kann, ob sie dem Schema entspricht, sondern aus der Analyse der Datei die richtige PDF-Version ableiten kann. Generell ist es auf Grund der genannten Schwächen ratsam, mehrere Validatoren zu benutzen, um fehlerhafte Analysen möglichst gut abfangen zu können. Aus den im Kapitel „Validierung“ genannten Gründen war dies bei dieser Arbeit nicht möglich.

Es lässt sich durchaus festhalten, dass einige Repositorien die Empfehlungen der DINI zur Langzeitarchivierung sehr gut annehmen. Die meisten erledigen jedoch nur die Pflicht und nicht die Kür. LZA ist durchaus ein relevantes Thema, auch und gerade für Open Access Repositorien, denn: Was nutzt der freie Zugang zu Dateien, die nicht mehr gelesen werden können? Open Access besitzt damit auch eine zeitliche Dimension, die erst langsam wahrgenommen und bearbeitet wird. Insgesamt gesehen ist die Verbreitung der PDF/A-Formate aber noch nicht sehr weit fortgeschritten. Um dies zu ändern müssen bereits veröffentlichte Dateien in neue, LZA-taugliche Formate konvertiert werden. Dazu benötigt es Konvertierungssoftware, die möglichst zuverlässig große Mengen verarbeiten kann. Firmen wie Callas, Adobe und PDFTools, aber auch Gruppen aus der Open Source Bewegung entwickeln ihre Produkte kontinuierlich weiter. Das Anfordern der Originalformate wie etwa „.docx“ oder „.odt“ macht dabei nur bedingt Sinn. Zwar können aus ihnen aktuelle PDF-Dateien erzeugt werden, allerdings nur so lange, wie die Originalformate selbst noch ausgelesen werden können. Mittel- bis langfristig stehen die Einrichtungen wieder vor dem Problem, eine große Menge Dateien in diversen Formaten migrieren zu müssen, sofern sie die Originalformate nicht nach dem Übergabeprozess in ein archivtaugliches Format umwandeln und in einem Paket mit den Metadaten vorhalten, vergleichbar einem Archival Information Package beim OAIS Referenz Modell.

Um die Verbreitung von PDF/A weiter zu fördern müssten auch die Exporttools der Textverarbeitungsprogramme dahingehend angepasst werden, dass sie standardmäßig die Archivformate verwenden. Dies ist natürlich nur aus archivischer Sicht praktikabel. Viel eher sollten die an PDF/A interessierten Einrichtungen ihre Anforderungen anpassen. Die Erstellung von PDF/A bedeutet für die meisten Publizierenden einen Mehraufwand ohne persönlichen Nutzen. Daher sollten Einrichtungen, denen Langzeitarchivierung ein wichtiges Anliegen ist, unbedingt PDF/A zur Pflicht machen und nicht nur eine Empfehlung aussprechen. Dass die Anteile der PDF/A-Dateien dadurch signifikant gesteigert werden kann ist exemplarisch an den Universitäten Marburg, Potsdam und Münster abzulesen.

Literaturverzeichnis

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Anforderungen an die technische Qualität des PDF-Dokumentes einer Dissertation. URL: http://edoc.hu-berlin.de/e_autoren/anforderungen-diss.php (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Elektronisches Publizieren. URL: http://edoc.hu-berlin.de/e_autoren/faq/ep0.php (11.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): FAQ: Elektronisches Publizieren. URL: http://edoc.hu-berlin.de/e_autoren/faq/ep0.php?arbeit=Habilitationsschriften&index=habil.php&nav=habil (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Leitlinien. URL: http://edoc.hu-berlin.de/e_info/leitlinien.php (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Dissertationen: Was muss eingereicht werden? URL: http://edoc.hu-berlin.de/e_autoren/was-diss.php?arbeit=Dissertationen&nav=diss&nav_this=Was%20muss%20eingereicht%20werden?&index=index.php (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Habilitationsschriften: Was muss eingereicht werden? URL: http://edoc.hu-berlin.de/e_autoren/was-ha-bil.php?arbeit=Habilitationsschriften&nav=habil&nav_this=Was%20muss%20eingereicht%20werden?&index=habil.php (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Publizieren von Magister- und Diplomarbeiten: Was muss eingereicht werden? URL: http://edoc.hu-berlin.de/e_autoren/was-diplom.php?arbeit=Magister-%20und%20Diplomarbeiten&nav=diplom&nav_this=Was%20muss%20eingereicht%20werden?&index=diplom.php (12.01.2015).

Arbeitsgruppe 'Elektronisches Publizieren' (2014): Technische Hinweise zur Erstellung und Korrektur von PDF-Dokumenten. URL: http://edoc.hu-berlin.de/e_autoren/vorlage-pdf.php (11.01.2015).

Aschenbrenner, Andreas (2010): Repository Systeme. Archivsoftware zum Herunterladen. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 11.3-11.6. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_411.pdf (10.11.2014).

Aschenbrenner, Andreas; Wollschläger, Thomas (2010): File Format Registries. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-

Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.19-7.22. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_448.pdf (15.11.2014).

Association for Digital Document Standards e.V. (o.J.): PDF Association. The Future of PDF. URL: http://www.pdfa.org/ (08.09.2014).

Association for Digital Document Standards e.V. (o.J.): The PDF/A Competence Center. URL: http://www.pdfa.org/competence-centers/pdfa-competence-center/ (08.09.2014).

Berthold, Henrike (2014): Die Eignung des Dateiformates PDF für die Langzeitarchivierung. Probleme und Lösungen, Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden. Dresden, 02.06.2014. URL: http://www.opus-bayern.de/bib-info/volltexte//2014/1627/pdf/EignungPdfSlub.pdf (22.09.2014).

Bibliothek der HTWG Konstanz (o.J.): OPUS 4 | Hilfe: Was sind Metadaten? URL: http://opus.htwg-konstanz.de/home/index/help/content/metadata (15.01.2015).

Borghoff, Uwe M. (2003): Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente. 1. Aufl. Heidelberg, Dpunkt.

Brübach, Nils (2010): Das Referenzmodell OAIS. Unter Mitarbeit von Manuela Quetisch, Hans Liegmann und Achim Oßwald. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 4.3-4.14. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_438.pdf (07.11.2014).

Brübach, Nils (2010): Die Überarbeitung und Ergänzung des OAIS. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 4.15-4.16. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_397.pdf.

Callas Software GmbH (o.J.): Wichtigste Funktionen. URL: http://www.callassoftware.com/callas/doku.php/de:products:pdfapilotcli:keyfeatures (18.01.2015).

Carpenter, Leona; Heery, Rachel (2003): 2. History and development of OAI-PMH. URL: https://www.oaforum.org/tutorial/english/page2.htm (28.11.2014).

Carpenter, Leona; Heery, Rachel (2003): 3. Main Technical Ideas of OAI-PMH. URL: https://www.oaforum.org/tutorial/english/page3.htm (28.11.2014).

Carpenter, Leona; Heery, Rachel (2003): OAI-PMH Online-Tutorial. URL: https://www.oaforum.org/tutorial/ (13.09.2014).

Comprehensive Perl Archive Network (2013): Perl Source. URL: http://www.cpan.org/src/README.html (27.11.2014).

Deutsche Initiative für Netzwerkinformation e. V. (2013): DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste 2013. DINI-Arbeitsgruppe „Elektronisches Publizieren“.

Version 4.0. Göttingen (DINI Schriften, 3). URL: http://edoc.hu-berlin.de/series/dini-schriften/2013-3/PDF/3.pdf (10.11.2014).

Deutsche Initiative für Netzwerkinformation e. V. (2014): DINI Check - OAI Validator. URL: http://oanet.cms.hu-berlin.de/validator/pages/validation_dini.xhtml (08.12.2014).

Deutsche Initiative für Netzwerkinformation e. V. (2014): DINI-Zertifikat. URL: http://www.dini.de/dini-zertifikat/ (08.09.2014).

Deutsche Initiative für Netzwerkinformation e. V. (2014): Liste der Repositorien. URL: http://dini.de/dini-zertifikat/liste-der-repositorien/ (11.12.2014).

Deutsches Institut für Internationale Pädagogische Forschung (2013): Fachportal Pädagogik - pedocs - Leitlinien des Dokumentenservers pedocs. URL: http://www.pedocs.de/leitlinien.php?la=de (11.01.2015).

Dublin Core Metadata Initiative (2015): Dublin Core Metadata Element Set, Version 1.1. URL: http://dublincore.org/documents/dces/ (16.01.2015).

Eblen, John (2000): Web crawling in Perl. A quick tutorial. URL: http://martinainscow.pwp.blueyonder.co.uk/L1-Perl/crawling_in_perl_page.html (22.12.2014).

Ergül, Arif; Böhm, Alexander; Schmidt, Elena; Hissen, Simon; Sariklis, Theodosios (2012): Erfolgsfaktoren für die Durchsetzung von PDF/A als weltweiter Standard für elektronische Langzeitarchivierung. In: Information - Wissenschaft & Praxis 63, H. 6, S. 361–366. URL: http://www.degruyter.com/view/j/iwp.2012.63.issue-6/iwp-2012-0074/iwp-2012-0074.xml?format=INT (10.11.2014).

Fachhochschule Köln Hochschulbibliothek (2013): Cologne Open Science - Leitlinien von Cologne Open Science. URL: http://opus.bsz-bw.de/fhk/doku/leitlinien.php (11.01.2015).

Freed, Ned; Kucherawy, Murray; Baker, Mark; Hoehrmann, Bjoern (2015): Media Types. URL: http://www.iana.org/assignments/media-types/media-types.xhtml (16.01.2015).

Friese, Yvonne (2014): Langzeitverfügbarkeit sichern: PDF-Validierung durch JHOVE? URL: http://www.pdfa.org/2014/12/langzeitverfugbarkeit-sichern-pdf-validierung-durch-jhove/?lang=de (03.01.2015).

Funk, Stefan E. (2010): Digitale Objekte und Formate. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.3-7.8. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_445.pdf (15.11.2014).

Funk, Stefan E.; Neubauer, Matthias (2010): Formatcharakterisierung. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.13-7.18. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_447.pdf (02.01.2015).

Glaser, Timo (2013): FAQs. URL: http://www.uni-marburg.de/bis/digitale_bibliothek/archivserver/pubservfaq (15.01.2015).

Hannemann, Bernd (o.J.): OPUS. URL: http://www.htwg-konstanz.de/OPUS.940.0.html (15.01.2015).

Hirtle, Peter B. (o.J.): The History and Current State of Digital Preservation in the United States. Metadata And Digital Collections, A Festschrift in Honor of Thomas P. Turner. URL: http://cip.cornell.edu/DPubS?service=UI&version=1.0&verb=Display&handle=cul.pub/1238609304 (17.11.2014).

Hochschule Heilbronn (o.J.): OPUS 4 | Hilfe. URL: http://opus-hshn.bsz-bw.de/home/index/help#contact (11.01.2015).

Huth, Karsten (2010): Textdokumente. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 17.3-17.7. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_421.pdf (15.11.2014).

Informationsplattform Open Access (2013): Repositorien. URL: http://open-access.net/de/allgemeines/was_bedeutet_open_access/repositorien/ (08.09.2014).

International Organization for Standardization (o.J.): ISO 14721:2012 - Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model. URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284 (06.11.2014).

International Organization for Standardization (2005): ISO 19005-1:2005 - Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1). URL: http://www.iso.org/iso/catalogue_detail?csnumber=38920 (24.11.2014).

International Organization for Standardization (2011): ISO 19005-2:2011 - Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2). URL: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50655 (24.11.2014).

International Organization for Standardization (2012): ISO 19005-3:2012 - Document management -- Electronic document file format for long-term preservation -- Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3). URL: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229 (24.11.2014).

International Organization for Standardization (2014): ISO 32000-1:2008 - Document management. Portable document format -- Part 1: PDF 1.7. URL: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=51502 (19.11.2014).

International Organization for Standardization (2014): ISO/DIS 32000-2 - Document management. Portable document format -- Part 2: PDF 2.0. URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=63534 (19.11.2014).

JSTOR (2009): JHOVE. JSTOR/Harvard Object Validation Environment. URL: http://jhove.sourceforge.net/ (02.01.2015).

Justus-Liebig-Universität Gießen (o.J.): Dokumenten- und Publikationsserver GEB der JLU Gießen - Leitlinien. URL: http://geb.uni-giessen.de/geb/GEB_Leitlinien.pdf (12.01.2015).

Justus-Liebig-Universität Gießen (2011): Häufig gestellte Fragen – Faqs. URL: http://geb.uni-giessen.de/geb/faqs.php?la=de (12.01.2015).

Justus-Liebig-Universität Gießen (2014): Anleitung zur Veröffentlichung von Dissertationen in GEB. URL: http://geb.uni-giessen.de/geb/doku/veroeff_diss.php?la=de (12.01.2015).

Lagoze, Carl; Van de Sompel, Herbert; Nelson, Michael; Warner, Simeon (2005): Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting. URL: http://www.openarchives.org/OAI/2.0/guidelines.htm (21.11.2014).

Lagoze, Carl; Van de Sompel, Herbert; Nelson, Michael; Warner, Simeon (2008): The Open Archives Initiative Protocol for Metadata Harvesting. URL: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm (21.11.2014).

Ludwig, Jens (2010): Auswahlkriterien. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 7.9-7.12. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_446.pdf (15.11.2014).

McNamara, John (2014): Excel::Writer::XLSX. URL: http://search.cpan.org/~jmcnamara/Excel-Writer-XLSX-0.81/lib/Excel/Writer/XLSX.pm#DOWNLOADING (05.01.2015).

Merz, Thomas (2011): Validierung von PDF/A. URL: http://www.pdfa.org/2011/09/validierung-von-pdfa/?lang=de (02.01.2015).

nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (2012): nestor - Publikationen. URL: http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/publikationen_node.html;jsessionid=B402855AA2967E395D9E6180AA934F97.prod-worker3 (03.11.2014).

nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland (2013): nestor - Über uns. URL: http://www.langzeitarchivierung.de/Subsites/nestor/DE/Header/Ueberuns/ueberuns_node.html (03.11.2014).

nestor-Arbeitsgruppe OAIS-Übersetzung / Terminologie (2013): Referenzmodell für ein Offenes Archiv-Informationssystem. Deutsche Übersetzung 2.0 (nestor-materialien, 16). URL: http://files.d-nb.de/nestor/materialien/nestor_mat_16-2.pdf (09.11.2014).

Neuroth, Heike; Oßwald, Achim; Scheffel, Regine; Strathmann, Stefan; Huth, Karsten (2010): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. URL: http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf (03.11.2014).

Oettler, Alexandra (2013): PDF/A kompakt 2.0. PDF für die Langzeitarchivierung. Berlin, Association for Digital Document Standards e. V. URL: http://www.pdfa.org/wp-content/uploads/2013/03/PDFA-kompakt-2_0_screen.pdf (09.11.2014).

Open Archives Initiative (o.J.): PMH-Tools. URL: http://www.openarchives.org/pmh/tools/tools.php (13.09.2014).

Open Archives Initiative (2008): OAI-PMH XML Schema. URL: http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd (24.11.2014).

Open Archives Initiative (2014): Registered Data Providers. URL: http://www.openarchives.org/Register/BrowseSites (11.12.2014).

PDF Association (2015): Products. URL: http://www.pdfa.org/products/?cn=645 (02.01.2015).

PDF Tools AG (2014): 3-Heights PDF Validator Shell, User Manual. Version 4.4. URL: http://www.pdf-tools.com/public/downloads/manuals/vals.pdf (16.01.2015).

PDF Tools AG (2015): PDF Validator. URL: http://www.pdf-tools.com/pdf/pdf-validator-pdfa-validieren-iso.aspx (03.01.2015).

PDF/A Competence Center (2011): Isartor Test Suite. URL: http://www.pdfa.org/2011/08/isartor-test-suite/ (02.01.2015).

Roche, Xavier (2014): HTTrack Website Copier. Free Software Offline Browser. URL: http://www.httrack.com/ (08.12.2014).

Rosenthol, Leonard (2011): PDF/A Metadaten XMP, RDF & Dublin Core. URL: http://www.pdfa.org/2011/09/pdfa-metadaten-xmp-rdf-dublin-core/?lang=de (27.11.2014).

Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok - Policy. URL: http://psydok.sulb.uni-saarland.de/sulb/policy.php (11.01.2015).

Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok FAQ. URL: http://psydok.sulb.uni-saarland.de/doku/faq.php (11.01.2015).

Saarländische Universitäts- und Landesbibliothek (o.J.): PsyDok-Hilfe. URL: http://psydok.sulb.uni-saarland.de/doku/hilfe_formular.php (11.01.2015).

Saarländische Universitäts- und Landesbibliothek (o.J.): Vorbereitung elektronischer Dokumente. URL: http://psydok.sulb.uni-saarland.de/doku/vorbereitung_dok.php (11.01.2015).

Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - Chemnitz: Hilfe. URL: http://monarch.qucosa.de/faq/ (11.01.2015).

Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - SLUB: Publizieren. URL: http://slub.qucosa.de/?id=3318 (12.01.2015).

Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden (o.J.): Qucosa - SLUB: Technische Hinweise zur PDF-Erstellung. URL: http://slub.qucosa.de/faq/technische-hinweise-zur-pdf-erstellung/ (12.01.2015).

Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH (2015): Schloss Dagstuhl : About DROPS. URL: http://www.dagstuhl.de/publikationen/publikationsserver-drops/about-drops/ (15.01.2015).

Schweizerische Nationalbibliothek (2011): Das OAIS-Modell. URL: http://www.nb.admin.ch/nb_professionnel/01693/01696/01876/01878/index.html?lang=de (09.11.2014).

Seminar für Sprachwissenschaft (2012): Glossar. URL: http://www.sfs.uni-tuebingen.de/nalida/de/doku/glossar.html#H (14.01.2015).

Shreeves, Sarah L. (2006): Search Interoperability, OAI, and Metadata. An Introduction to the OAI Protocol for Metadata Harvesting, University of Illinois at Urbana-Champaign, 08.12.2006. URL: https://www.ideals.illinois.edu/bitstream/handle/2142/175/METRO_OAIWorkshop.ppt.pdf?sequence=18 (21.11.2014).

Solid Documents (2015): Free PDF/A Validator. URL: http://www.validatepdfa.com/online.htm (02.01.2015).

SPI Inc. (2015): Package: poppler-utils in wheezy. URL: https://packages.debian.org/de/wheezy/poppler-utils (05.01.2015).

Staats- und Universitätsbibliothek Hamburg (2013): Unsere Leitlinien : eDoc.ViFaPol. URL: http://edoc.vifapol.de/opus/leitlinien.php (15.01.2015).

Staats- und Universitätsbibliothek Hamburg (2013): Veröffentlichen : eDoc.ViFaPol. URL: http://edoc.vifapol.de/opus/veroeffentlichen.php (15.01.2015).

Suchodoletz, Dirk von (2009): Emulationen für Archive. Strategien für die Langzeitarchivierung. In: DFN Mitteilungen, H. 76, S. 23–27. URL: https://www.dfn.de/fileadmin/5Presse/DFNMitteilungen/heft76.pdf (07.11.2014).

Summers, Ed (2004): Building OAI-PMH Harvesters With Net::OAI::Harvester. In: Ariadne: Web Magazine for Information Professionals, H. 38. URL: http://www.ariadne.ac.uk/issue38/summers (24.11.2014).

Summers, Ed; Emmerich, Martin (o.J.): Net::OAI::Harvester. URL: http://search.cpan.org/~esummers/OAI-Harvester-1.0/lib/Net/OAI/Harvester.pm (27.11.2014).

Technische Universität Chemnitz (2014): Universitätsbibliothek: Publizieren. URL: https://www.tu-chemnitz.de/ub/publizieren/hss/epublizieren.html (11.01.2015).

The Consultive Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Practices. Washington, CCSDS Secretariat (Recommended Practice, 2). URL: http://public.ccsds.org/publications/archive/650x0m2.pdf (07.11.2014).

Universität des Saarlandes (o.J.): SciDok faq. URL: http://scidok.sulb.uni-saarland.de/doku/faq.php (14.01.2015).

Universität des Saarlandes (o.J.): SciDok-Hilfe. URL: http://scidok.sulb.uni-saarland.de/doku/hilfe_formular.php (14.01.2015).

Universität des Saarlandes (o.J.): Vorbereitung elektronischer Dokumente. URL: http://scidok.sulb.uni-saarland.de/doku/vorbereitung_dok.php (14.01.2015).

Universität Duisburg-Essen (o.J.): DuEPublico Ansprechpartner und Infos für Autoren. URL: http://duepublico.uni-duisburg-essen.de/authoring/ansprechpartner.xml (11.01.2015).

Universität Duisburg-Essen (o.J.): DuEPublico: Leitlinien. URL: http://duepublico.uni-duisburg-essen.de/about/leitlinien.xml (11.01.2015).

Universität Potsdam (o.J.): Formate. URL: http://opus.kobv.de/ubp/doku/formate.php (15.01.2015).

Universität Potsdam (o.J.): Leitlinien. URL: http://opus.kobv.de/ubp/doku/policy.php (15.01.2015).

Universität Potsdam (o.J.): OPUS-Hilfe: Was sind Metadaten? URL: http://opus.kobv.de/ubp/doku/hilfe_formular.php (15.01.2015).

Universität Potsdam (o.J.): Tutorial zur PDF-Erstellung. URL: http://opus.kobv.de/ubp/doku/tutorial.php (15.01.2015).

Universität Ulm (o.J.): VTS | Akzeptierte Dateiformate. URL: http://vts.uni-ulm.de/help/formate.asp (12.01.2015).

Universität Ulm (o.J.): VTS | FAQ. URL: http://vts.uni-ulm.de/help/faq.asp (12.01.2015).

Universität Ulm (o.J.): VTS | Leitlinien. URL: http://vts.uni-ulm.de/policy.asp (12.01.2015).

Universität Ulm (o.J.): VTS | PDF-Erstellung. URL: http://vts.uni-ulm.de/help/pdf.asp (12.01.2015).

Universitäts- und Landesbibliothek Münster (2013): Datei-Erstellung - PDF/A. URL: http://www.uni-muenster.de/Publizieren/veroeffentlichung/dateierstellung/pdfa-erstellung.html (15.01.2015).

Universitäts- und Landesbibliothek Münster (2013): Dateiformate - Publizieren an der WWU. URL: http://www.uni-muenster.de/Publizieren/veroeffentlichung/dateierstellung/dateiformate.html (15.01.2015).

Universitäts- und Landesbibliothek Münster (2013): WWU-Publikationsserver - Leitlinien. URL: http://www.uni-muenster.de/Publizieren/dienstleistungen/repository/leitlinien.html (15.01.2015).

Universitätsbibliothek Bielefeld (2014): Suchmaschine BASE - Bielefeld Academic Search Engine | Die Quellen. URL: http://www.base-search.net/about/de/about_sources_date_dn.php?menu=2 (21.11.2014).

Universitätsbibliothek Bielefeld (2014): Suchmaschine BASE - Bielefeld Academic Search Engine | Über BASE. URL: http://www.base-search.net/about/de/index.php (21.11.2014).

Universitätsbibliothek Bochum (2014): Dissertationen elektronisch publizieren. Ein alternatives Angebot, Leitfaden für Doktoranden. URL: http://www.ub.ruhr-uni-bochum.de/imperia/md/content/benutzung/infomaterial/flyerde.pdf (15.01.2015).

Universitätsbibliothek Bochum (2015): Zugelassene Datenformate und Layout für Elektronische Dissertationen. URL: http://www.ub.ruhr-uni-bochum.de/DigiBib/Tauschseiten/Datenformate.html (15.01.2015).

Universitätsbibliothek Freiburg (2015): Häufig gestellte Fragen (FAQ) - Universitätsbibliothek Freiburg. URL: http://www.ub.uni-freiburg.de/index.php?id=3165 (11.01.2015).

Universitätsbibliothek Freiburg (2015): Publizieren im PDF-Format (ein Online-Tutorial) - Universitätsbibliothek Freiburg. URL: http://www.ub.uni-freiburg.de/index.php?id=3149 (11.01.2015).

Universitätsbibliothek Kaiserslautern (o.J.): KLUEDO | Hilfe. URL: https://kluedo.ub.uni-kl.de/home/index/help (12.01.2015).

Universitätsbibliothek Kaiserslautern (2014): KLUEDO | Leitlinien. URL: https://kluedo.ub.uni-kl.de/home/index/policies (12.01.2015).

Universitätsbibliothek Mainz (o.J.): Anleitung zum Eintragen und Veröffentlichen eines Dokuments. URL: http://archimed.uni-mainz.de/opusubm/informationen-zur-ablieferung.html (11.01.2015).

Universitätsbibliothek Mainz (o.J.): Leistungsspektrum und Nutzungsrichtlinien. URL: http://archimed.uni-mainz.de/opusubm/leistungsspektrum_und_nutzungsrichtlinien.html (11.01.2015).

Universitätsbibliothek Marburg (2011): Benutzungsordnung für den Publikationsserver der Universitätsbibliothek an der Philipps-Universität Marburg. URL: http://www.uni-marburg.de/bis/ueber_uns/vorschr/pubserver.pdf (15.01.2015).

Universitätsbibliothek Siegen (o.J.): Informationen für Autoren. URL: http://www.ub.uni-siegen.de/cms/index.php?id=1173&L=0 (12.01.2015).

Universitätsbibliothek Siegen (o.J.): Leitlinien des Publikationsservers OPUS Siegen. URL: http://www.ub.uni-siegen.de/cms/index.php?id=1316&L=0 (12.01.2015).

Universitätsbibliothek Siegen (o.J.): PDF-Dokumente erstellen. URL: http://www.ub.uni-siegen.de/cms/index.php?id=1188&L=0 (12.01.2015).

Universitätsbibliothek Siegen (o.J.): Technische Anforderungen. URL: http://www.ub.uni-siegen.de/cms/index.php?id=1187&L=0 (12.01.2015).

Universitätsbibliothek Stuttgart (2011): OPUS-FAQ. URL: http://elib.uni-stuttgart.de/opus/doku/faq.php#Wie lange (12.01.2015).

Universitätsbibliothek Stuttgart (2011): Vorbereitung elektronischer Dokumente. URL: http://elib.uni-stuttgart.de/opus/doku/vorbereitung_dok.php (12.01.2015).

University of Nottingham (2014): OpenDOAR - Summaries - Germany. URL: http://www.opendoar.org/find.php?search=&clID=&ctID=&rtID=&cID=81&lID=&rSoftWareName=&submit=Search&format=summary&step=20&sort=r.rName&rID=&ctrl=new&p=1 (11.01.2015).

University of Nottingham (2015): OpenDOAR - Summaries - Worldwide, Query: Publikations- und Dokumentenserver der Universitätsbibliothek Marburg. URL: http://www.opendoar.org/new1find.php (25.01.2015).

Upmeier, Arne (2010): Rechtliche Aspekte. In: Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Karsten Huth (Hg.): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, S. 16.3-16.13. URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_462.pdf (16.11.2014).

Van de Sompel, Herbert; Nelson, Michael L.; Lagoze, Carl; Warner, Simeon (2004): Resource Harvesting within the OAI-PMH Framework. In: D-Lib Magazine 10, H. 12. URL: http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html (29.11.2014).

van der Knijff, Johan (o.J.): Identification of preservation risks in PDF with Apache Preflight. a first impression. URL: http://openpreservation.org/system/files/pdfProfilingJvdK19122012.pdf (03.01.2015).

Zentralbibliothek des Forschungszentrum Jülich (o.J.): JUWEL - Volltextserver: FAQ. URL: http://juwel.fz-juelich.de:8080/dspace/help/faq.jsp#formats (15.01.2015).

Zentralbibliothek des Forschungszentrum Jülich (o.J.): JUWEL - Volltextserver: Hilfe. URL: http://juwel.fz-juelich.de:8080/dspace/help/help.jsp#submit (15.01.2015).

Zentralbibliothek des Forschungszentrum Jülich (o.J.): Von JUWEL unterstützte Formate. URL: http://juwel.fz-juelich.de:8080/dspace/help/formats.jsp (15.01.2015).

Anhang A: Perl-Skript für Harvesting und Validierung am Beispiel der Universität Freiburg

```
#!/usr/bin/perl -w

#Metadaten- und Resource-Harvesting von PDF-Dateien via OAI-PMH sowie Validierung mit pdfinfo und 3-Heights

use Net::OAI::Harvester;
use LWP::Simple;
use LWP::UserAgent;
use HTTP::Request;
use HTTP::Response;
use HTML::LinkExtor;
use Excel::Writer::XLSX;

my $harvester = Net::OAI::Harvester->new(
                        baseURL => 'http://www.freidok.uni-freiburg.de/oai2/oai2.php'
);
my $identify = $harvester->identify();
my $list = $harvester->listAllRecords(
                        metadataPrefix  => 'oai_dc',
```

```
                        );
my @identifiers;

$browser = LWP::UserAgent->new();

$year = '';
$claimed = '';



#Excel Datei
my $exname = $identify->repositoryName();

$row = 0;

my $workbook = Excel::Writer::XLSX->new("$exname.xlsx");
$workbook->compatibility_mode();
                        $worksheet   = $workbook->add_worksheet('neu');
   $worksheet->write($row, 0, 'Datei');
                        $worksheet->write($row, 1, 'Letzte Aenderung');
                        $worksheet->write($row, 2, 'Angebliches Format');
                        $worksheet->write($row, 3, 'Wirkliches Format');
```

```
$worksheet->write($row, 4, 'File Format corrupted');
$worksheet->write($row, 5, 'PDF Reference not conform');
$worksheet->write($row, 6, 'Encrypted');
$worksheet->write($row, 7, 'Color Spaces');
$worksheet->write($row, 8, 'Illegal Rendering Hints');
$worksheet->write($row, 9, 'Alternate Information');
$worksheet->write($row, 10, 'PostScript');
$worksheet->write($row, 11, 'External Content Reference');
$worksheet->write($row, 12, 'Fonts not embedded');
$worksheet->write($row, 13, 'Fonts not Unicode');
$worksheet->write($row, 14, 'Transparenz');
$worksheet->write($row, 15, 'Unknown Annotation Types');
$worksheet->write($row, 16, 'Multimedia Annotation');
$worksheet->write($row, 17, 'Invisivle Annotations');
$worksheet->write($row, 18, 'Form Field Appearance');
$worksheet->write($row, 19, 'Action Types');
$worksheet->write($row, 20, 'Metadata');
$worksheet->write($row, 21, 'Logical Structure');
$worksheet->write($row, 22, 'Ebenen');
$worksheet->write($row, 23, 'Sonstiges');
```

```
#Skriptlaufzeit und Counter
use Time::HiRes qw(gettimeofday tv_interval);
my $t0 = [gettimeofday];
$count = 0;



#Harvesting anhand der Identifier
while ( my $record = $list->next() ) {
                        my $metadata = $record->metadata();
                        @identifiers = $metadata->identifier;
                        foreach $identifier (@identifiers) {
                            if($identifier =~ /^http(.*)\.pdf$/) {
                                #Download
                                $identifier =~ s/\W/_/g;
                                print "Directly downloading ", $identifier, "\n";
                                my $status = getstore($identifier, "$identifier.pdf");
                                        if ( is_success($status) ){
                                            print "Download complete: ", $identifier, "\n";
                                        }
                                        else{
```

```
                print "Fehler beim Download von: ", $identifier, "\n";
            }
#Jahr aus der DocInfo lesen
open(YEAR,"/usr/bin/pdfinfo $identifier.pdf |") || die "Failed: $!\n";
            while ( <YEAR> )
            {
              if (/ModDate/) {
              $year=substr($_,36,4);
              }
            }
#Validierung
print "Validiere: $identifier.pdf\n";
#angebliches PDF-Format auslesen
open(CLAIM,"/home/admin/validator/bin/pdfvalidator -ccl $identifier.pdf |") || die "Failed: $!\n";
            while ( <CLAIM> )
            {
                if (/Conformance/) {
                $claimed=substr($_,13,,)
                }

            chomp($claimed);
```

```
        }
#Datei gegen angebliches Format abprüfen
$row++;

open(VALID,"/home/admin/validator/bin/pdfvalidator -cl ccl -rs -v $identifier.pdf |") || die "Failed: $!\n";
        while ( <VALID> )
        {
            my @report = split(/\n/,$_);

            foreach my $line ( @report ){


                        if ($line =~ /^The file format(.)*/) {
                                $worksheet->write($row, 4, "$line");
                        }

                        elsif ($line =~ /^The document doesn't conform to the PDF reference(.)*/) {
                                $worksheet->write($row, 5, "$line");
                        }
```

```
elsif ($line =~ /^The file is encrypted(.)*/) {
        $worksheet->write($row, 6, "$line");
}

elsif ($line =~ /^The document contains device(.)*/) {
        $worksheet->write($row, 7, "$line");
}

elsif ($line =~ /^The document contains illegal rendering hints(.)*/) {
        $worksheet->write($row, 8, "$line");
}

elsif ($line =~ /^The document contains alternate information(.)*/) {
        $worksheet->write($row, 9, "$line");
}

elsif ($line =~ /^The document contains embedded PostScript code(.)*/) {
        $worksheet->write($row, 10, "$line");
}

elsif ($line =~ /^The document contains references to external content(.)*/) {
```

```
                $worksheet->write($row, 11, "$line");
            }

            elsif ($line =~ /^The document contains fonts without embedded font programs or
encoding information(.)/) {
                $worksheet->write($row, 12, "$line");
            }

            elsif ($line =~ /^The document contains fonts without appropriate character to
unicode mapping information(.)*/) {
                $worksheet->write($row, 13, "$line");
            }

            elsif ($line =~ /^The document contains transparency(.)*/) {
                $worksheet->write($row, 14, "$line");
            }

            elsif ($line =~ /^The document contains unknown annotation types(.)*/) {
                $worksheet->write($row, 15, "$line");
            }
```

```
                elsif ($line =~ /^The document contains multimedia annotations(.)*/) {
                        $worksheet->write($row, 16, "$line");
                }

                elsif ($line =~ /^The document contains hidden(.)*/) {
                        $worksheet->write($row, 17, "$line");
                }

                elsif ($line =~ /^The document contains annotations or form fields with ambigous or without appropriate appearances(.)*/) {
                        $worksheet->write($row, 18, "$line");
                }

                elsif ($line =~ /^The document contains actions types other than for navigation(.)*/) {
                        $worksheet->write($row, 19, "$line");
                }

                elsif ($line =~ /^The document's meta data is either missing or inconsistent or corrupt(.)*/) {
                        $worksheet->write($row, 20, "$line");
                }
```

```
mation(.)*/) {                elsif ($line =~ /^The document doesn't provide appropriate logical structure infor-

                        $worksheet->write($row, 21, "$line");
                }

                elsif ($line =~ /^The document contains optional content(.)*/) {
                        $worksheet->write($row, 22, "$line");
                }



                elsif ($line =~ /^Validating(.)*/) {
                #Tue nichts.
                }
                elsif ($line =~ /^Conformance(.)*/) {
                #Tue nichts.
                }
                elsif ($line =~ /^Done./) {
                #Tue nichts.
                }
```

```
                                        else {
                                        $worksheet->write($row, 23, "$line");
                                        }



                                    }


                                }

                    #Beschreiben der Excel-Datei und Ausgabe zur Kontrolle während der Ausführung


                        $worksheet->write($row, 0, "$identifier");
                        $worksheet->write($row, 1, "$year");
                        $worksheet->write($row, 2, "$claimed");

                }


                if($identifier =~ /^http/){
```

```
#Erst Crawler
my $request = HTTP::Request->new(GET => $identifier);
my $response = $browser->request($request);
    if ($response->is_error()) {
        printf "%s\n", $response->status_line;
    }
$contents = $response->content();

my ($page_parser) = HTML::LinkExtor->new(undef, $identifier);
$page_parser->parse($contents)->eof;
@links = $page_parser->links;
$part = 0; #Variable zur Benennung, falls mehrere PDF-Dateien pro Identifier
#Dann Download
foreach $link (@links) {
    if ($$link[2] =~ /(.*)\.pdf$/) {
        $identifier =~ s/\W/_/g;
        $part++;
        print "Downloading after crawling: $identifier-$part\n";
        my $status = getstore($$link[2], "$identifier-$part.pdf");
            if ( is_success($status) ){
                print "Download complete: $identifier-$part\n";
```

```
        }
        else{
          print "Fehler beim Download von: $identifier-$part\n";
        }


print "Validiere: $identifier-$part.pdf\n";
#Jahr aus der DocInfo lesen
open(YEAR,"/usr/bin/pdfinfo $identifier-$part.pdf |") || die "Failed: $!\n";
            while ( <YEAR> )
            {
              if (/ModDate/) {
              $year=substr($_,36,4);
              }
            }

#Validierung

#angebliches Format auslesen
open(CLAIM,"/home/admin/validator/bin/pdfvalidator -ccl $identifier-$part.pdf |") || die "Failed: $!\n";
            while ( <CLAIM> )
```

```
{
    if (/Conformance/) {
    $claimed=substr($_,13,,)
    }

chomp($claimed);

}

#Datei gegen angebliches Format abprüfen und Fehler in die Tabelle schreiben
open(VALID,"/home/admin/validator/bin/pdfvalidator -cl ccl -rs -v $identifier-$part.pdf |") || die "Failed: $!\n";
$row++;
while ( <VALID> )
{
    my @report = split(/\n/,$_);

    foreach my $line ( @report ){

        if ($line =~ /^The file format(.)*/) {
            $worksheet->write($row, 4, "$line");
```

```
}

elsif ($line =~ /^The document doesn't conform to the PDF reference(.)*/) {
        $worksheet->write($row, 5, "$line");
}

elsif ($line =~ /^The file is encrypted(.)*/) {
        $worksheet->write($row, 6, "$line");
}

elsif ($line =~ /^The document contains device(.)*/) {
        $worksheet->write($row, 7, "$line");
}

elsif ($line =~ /^The document contains illegal rendering hints(.)*/) {
        $worksheet->write($row, 8, "$line");
}

elsif ($line =~ /^The document contains alternate information(.)*/) {
        $worksheet->write($row, 9, "$line");
}
```

```
elsif ($line =~ /^The document contains embedded PostScript code(.)*/) {
        $worksheet->write($row, 10, "$line");
}

elsif ($line =~ /^The document contains references to external content(.)*/) {
        $worksheet->write($row, 11, "$line");
}

elsif ($line =~ /^The document contains fonts without embedded font programs or
encoding information(.)/) {
        $worksheet->write($row, 12, "$line");
}

elsif ($line =~ /^The document contains fonts without appropriate character to
unicode mapping information(.)*/) {
        $worksheet->write($row, 13, "$line");
}

elsif ($line =~ /^The document contains transparency(.)*/) {
        $worksheet->write($row, 14, "$line");
```

```
                }

                elsif ($line =~ /^The document contains unknown annotation types(.)*/) {
                        $worksheet->write($row, 15, "$line");
                }

                elsif ($line =~ /^The document contains multimedia annotations(.)*/) {
                        $worksheet->write($row, 16, "$line");
                }

                elsif ($line =~ /^The document contains hidden(.)*/) {
                        $worksheet->write($row, 17, "$line");
                }

                elsif ($line =~ /^The document contains annotations or form fields with ambigous or
without appropriate appearances(.)*/) {
                        $worksheet->write($row, 18, "$line");
                }

                elsif ($line =~ /^The document contains actions types other than for navigation(.)*/) {
                        $worksheet->write($row, 19, "$line");
```

```
}

elsif ($line =~ /^The document's meta data is either missing or inconsistent or cor-
rupt(.)*/) {
        $worksheet->write($row, 20, "$line");
}

elsif ($line =~ /^The document doesn't provide appropriate logical structure infor-
mation(.)*/) {
        $worksheet->write($row, 21, "$line");
}

elsif ($line =~ /^The document contains optional content(.)*/) {
        $worksheet->write($row, 22, "$line");
}


elsif ($line =~ /^Validating(.)*/) {
#Tue nichts.
}
elsif ($line =~ /^Conformance(.)*/) {
```

```
                    #Tue nichts.
                    }
                    elsif ($line =~ /^Done./) {
                    #Tue nichts.
                    }

                    else {
                    $worksheet->write($row, 23, "$line");
                    }



                }

            }

#Beschreiben der Excel-Datei und Ausgabe zur Kontrolle während der Ausführung

    $worksheet->write($row, 0, "$identifier-$part");
```

```
                    $worksheet->write($row, 1, "$year");
                    $worksheet->write($row, 2, "$claimed");


                }
            }
        }
        }


#Ausgabe der bisherigen Skriptlaufzeit und bisherige Anzahl der bearbeiteten Identifier
$count++;
print "\n";
print "Anzahl: ", $count, "\n";
print "Dauer bislang: ".tv_interval($t0)." Sekunden\n";
print "\n";
print "+"x20, "\n";
print "\n";
}

#Excel-Datei schließen
$workbook->close();
```

```
#Ausgabe der gesamten Skriptlaufzeit und der gesamten Anzahl der bearbeiteten Identifier
print "Die Datei $exname wurde angelegt und beschrieben.\n";
print "Der Vorgang dauerte ".tv_interval($t0)." Sekunden.\n";
print "Es wurden $count Datensätze geerntet.\n";

#Ende
```