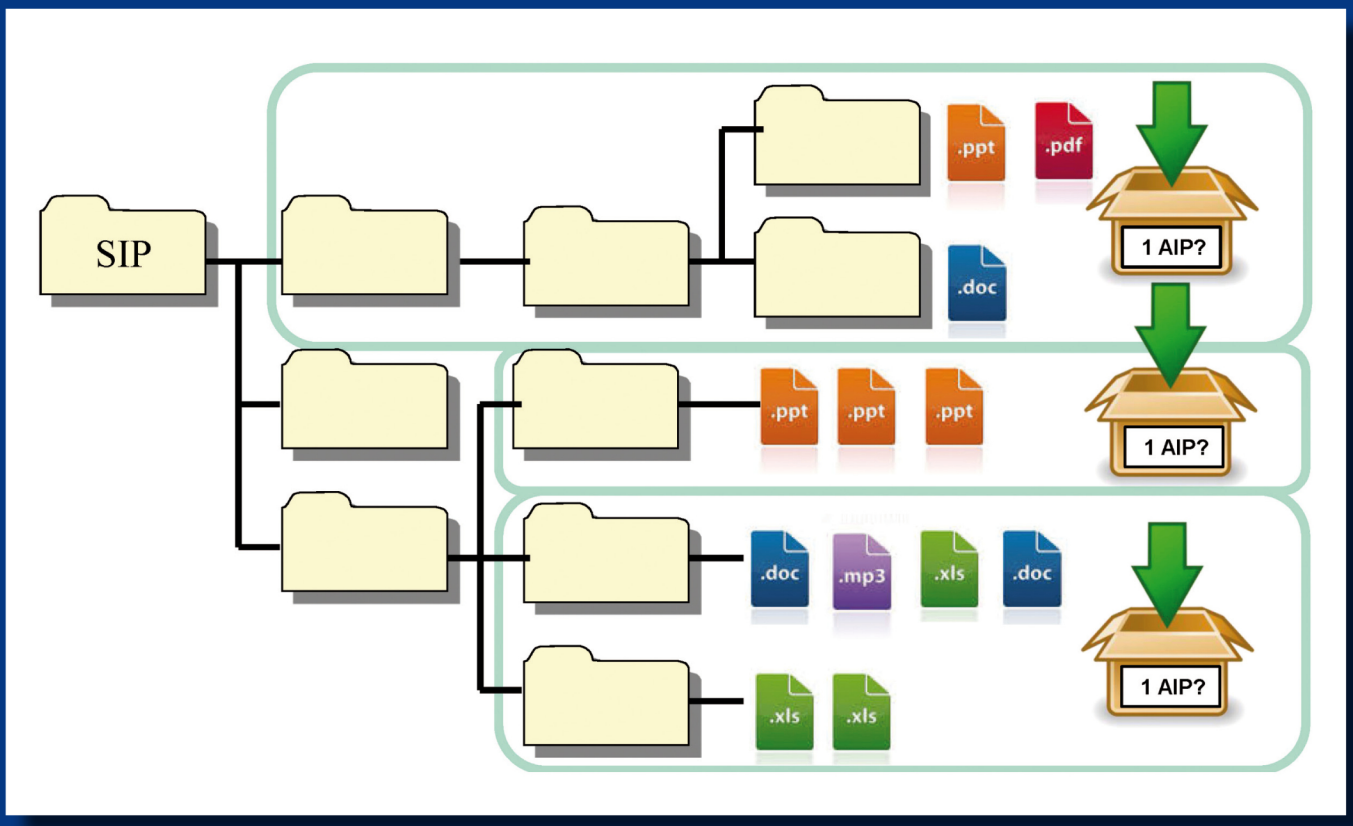
Kreative digitale Ablagen und die Archive



Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns

Sonderveröffentlichungen der Staatlichen Archive Bayerns
Nr. 13

Kreative digitale Ablagen und die Archive

Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns

Herausgegeben von
Kai Naumann und Michael Puchta

Staatliche Archive Bayerns

KLA Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder

München 2017

Sonderveröffentlichungen der Staatlichen Archive Bayerns

herausgegeben von den Staatlichen Archiven Bayerns

Schriftleitung: Julian Holzapfl

Nr. 13: Kreative digitale Ablagen und die Archive. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns, herausgegeben von Kai Naumann und Michael Puchta

Satz und Gestaltung: Karin Hagendorn

Umschlaggestaltung: Karin Hagendorn

Druck: Grafik und Druck GmbH Peter Pöllinger, Landsberger Str. 318a, 80687 München

ISSN 1618-0739

ISBN 978-3-938831-81-6

Inhalt

Vorwort

Die Archivierung digitaler Unterlagen ist seit Jahren ein publikumsträchtiges Tagungsthema. Von den darunter diskutierten Fragen scheint der Umgang mit Dateiablagen, E-Mails und Intranetseiten einen besonders hohen Stellenwert zu haben. Das zeigt die Zahl der Workshopteilnehmer, von denen die Generaldirektion der Staatlichen Archive Bayerns am 22./23. November 2016 geradezu überrannt wurde. Die Veranstalter hielten freudigen Herzens stand und boten ein abwechslungsreiches Programm, in dem die Besucherinnen und Besucher durch Publikumsdiskussionen beteiligt wurden und selbst die Arbeit mit entsprechenden Software-Werkzeugen kennenlernen konnten.

Was Dateisammlungen oder Fileablagen sind, ist noch nicht im Konsens definiert – zwei Vorschläge zur Definition befinden sich auf S. 7 sowie auf S. 44 dieses Bandes. Sie beschäftigen aber die Archive seit langem im Windschatten von anderen Themen. Der 2009 verstorbene Rechtsprofessor und ehemalige Präsident des Bundesverfassungsgerichts Ernst Benda arbeitete mit einem E-Akten-System, das eigentlich für mittelständische Firmen konzipiert war, und 2005 vermachte er dem Bundesarchiv große Teile seines schriftlichen Lebenswerks in dieser Form.[1] Von einer definierten, normbasierten Ablieferungsschnittstelle war keine Rede. Das Bundesarchiv nahm die Daten trotzdem an.

Dies war die erste Begegnung der Herausgeber mit dem Phänomen. Während im fachinternen Diskurs seitdem DMS-Einführungen, Konzeption digitaler Archivsoftware, Bestandserhaltungsfragen und andere Archivaliengattungen wie Fachanwendungen, Geodaten, Statistikdaten und Internetseiten große Aufmerksamkeit fanden, blieb dagegen die Bewertung und Aufbereitung von Dateisammlungen, von E-Mail-Konten und Intranetseiten, von unstrukturiert genutzten DMS etc. in den Publikationen etwas zurück.

Es gab verwaltungsinterne Bestrebungen, eine unkontrollierbar anmutende Form der Schriftgutverwaltung zu verhindern – also den „Formatzoo“, wie ein Kollege sich ausdrückte, außen vor zu halten. Dieser Versuch ist in den meisten Verwaltungsbereichen als gescheitert anzusehen.

Die abgebenden Stellen entwickeln sich in zwei Richtungen: Die einen besitzen und benutzen ein E-Akten-System, das systematische Aussonderungsschnittstellen hat sowie eine strukturierte Dokumentenablage ermöglicht und betreiben parallel kreative digitale Ablagen. In diesen Fällen besteht für die zuständigen Archive in der Regel kein Anlass, eine Dateisammlung oder E-Mail-Konten zu übernehmen, sofern deren Inhalte in die elektronischen Akten eingeflossen sind.

Doch es gibt auch eine Vielzahl von Behörden in den unterschiedlichsten Verwaltungsbereichen von Arbeitsbeschaffung bis Zivilverteidigung, deren Handlungsgegenstände und Handlungsweise beinahe ausschließlich aus Netzlaufwerken und E-Mail-Konten hervorgehen. Es muss Aufgabe der Archive sein, diese Stellen durch gute Beratung von der Notwendigkeit einer geordneten Schriftgutverwaltung zu überzeugen. Doch auch bei diesen Stellen muss Überlieferungsbildung betrieben werden können. Dieser Band soll deshalb der Fachgemeinde Hinweise dafür geben, mit solchen Formen der Schriftgutverwaltung auf Augenhöhe umzugehen.

[1] Nachlass Ernst Benda, Bundesarchiv Koblenz N 1564-MD.

Hierzu wurde von Annekathrin Miegel, Sigrid Schieber und Christoph Schmidt aus ihren praktischen Erfahrungen heraus ein Handlungsvorschlag formuliert. Susanne Belovari beschreibt ihren Umgang mit einer vorwiegend audiovisuellen Dateisammlung und führt Bewertungskriterien auf. Niklas Konzen und Kristina Starkloff beleuchten zwei verschiedene Sonderformen von kreativen digitalen Ablagen, nämlich unstrukturiert genutzte E-Akten-Systeme und E-Mail-Konten. Vor, während und nach der Tagung hat Kai Naumann unter Beteiligung vieler Praktiker die nötigen Arbeitsschritte und die verwendeten Werkzeuge mitgeschrieben und in einen Text zusammengefasst. Den größten Teil des Buchs nehmen Beschreibungen von einzelnen Werkzeugen ein, wobei unterschieden wird zwischen Querschnittswerkzeugen, die Bewertung und Übernahme insgesamt abdecken, und Spezialwerkzeugen, die einzelne Schritte in diesem Prozess erledigen.

Die Herausgeber möchten folgenden Beteiligen verbindlichst danken:

- den Referentinnen und Referenten für ihre Vorträge und Manuskripte
- den Softwareherstellern für ihre teils kostenlosen, teils kostenpflichtigen Produkte und die Informationen dazu
- den Referenten, Herstellern und weiteren Praktikern im deutschsprachigen Raum für die vielen Hinweise für den Text „Arbeitsschritte zur Aufbereitung“
- den Kolleginnen und Kollegen vom Ausschuss Digitale Archive der KLA für ihre vielfältige Beteiligung
- der Generaldirektorin der Staatlichen Archive Margit Ksoll-Marcon und allen ihren damals beteiligten Mitarbeiterinnen und Mitarbeitern
- den Gästen der Tagung für ihre gute Laune und Disziplin trotz teils beengter Verhältnisse
- und nicht zuletzt Karin Hagendorn von der Generaldirektion der Staatlichen Archive Bayerns für die Erstellung der Druckvorlage, und Julian Holzapfl für die Betreuung dieser Publikation.

Der Zweck der Veranstaltung war unter anderem, einen Überblick über Bedarf und Angebot zur Bewältigung entsprechender archivischer Überlieferung zu schaffen. Der Vergleich verschiedener Softwareprodukte war hierfür notwendig, kann aber fachlich durchdachte Entscheidungen für oder gegen bestimmte Software nicht ersetzen.

Die Beteiligten freuen sich auf Rückmeldungen zu einschlägigen Entwicklungen aus der Informatik, den Gedächtnisinstitutionen und den übrigen Disziplinen im Informations- und Dokumentationsbereich. Insbesondere die Fortentwicklung übergreifender Querschnittswerkzeuge ist ein dringendes Anliegen, um einem Verlust der frühen Überlieferung des digitalen Zeitalters zuvorzukommen.

Kai Naumann, Michael Puchta

Vom richtigen Umgang mit kreativen digitalen Ablagen

Annekathrin Miegel, Sigrid Schieber und Christoph Schmidt

Einleitung

Der richtige Umgang mit Dateisammlungen[1] stand lange Jahre nicht im Fokus der archivfachlichen Beschäftigung mit elektronischen Unterlagen. Der reinen Lehre nach arbeitet die Verwaltung ja auch im digitalen Zeitalter aktenmäßig, und wo es keine Akten gibt, da gibt es doch wenigstens Fachverfahren. Daten in Dateisammlungen, mithin: Einzeldateien in Filesystemen oder E-Mail-Postfächer, die ohne maschinell nachvollziehbare Struktur abgelegt werden und ohne das Wissen des Anlegenden oft kryptisch bleiben, bilden in dieser angenommenen besten aller Welten nur Zwischenstadien der Informationsverarbeitung, Übergangsformen ohne dauerhafte Relevanz. Das ganze Interesse der Archive galt daher den strukturierten Daten.

Sobald aus der Theorie jedoch die archivische Praxis erwuchs, stellte sich auf fast allen Ebenen der Verwaltung heraus, dass die Bürokratie viel weniger strukturiert und aktenmäßig arbeitet, als sie es eigentlich sollte. Da ersetzt die Dateisammlung die Akte („Akte? Was ist denn eine Akte?“), da werden besonders wichtige Dokumente aus E-Mailpostfächern ganz bewusst nicht mehr ausgedruckt, Protokolle und Niederschriften nur noch elektronisch vorgehalten. Die Archive stellt dies vor eine zweifache Herausforderung: Zum einen im Rahmen der Behördenberatung, in der sie dringlicher denn je auf die Grundsätze einer rechtskonformen Schriftgutverwaltung hinweisen müssen, und zum anderen im praktischen Umgang bei der Archivierung dieses Materials.

Der vorliegende Text widmet sich der Frage, wie diese zweite Herausforderung angegangen werden kann, welche Probleme regelmäßig auftreten und welche Lösungsansätze denkbar sind. Er versteht sich dabei weniger als ein archivwissenschaftlicher Diskussionsbeitrag denn als archivpraktische Handlungsanregung, die vor allem aus den konkreten Erfahrungen ihrer AutorInnen erwachsen ist. Seine Gliederung folgt, soweit dies möglich ist, der klassischen Abfolge der Arbeitsschritte im Archivierungsprozess.

Prüfung der archivierungsrelevanten Form

Bei einer Anbietung von Dateisammlungen ist zunächst zu fragen, ob die angebotene Form tatsächlich die archivierungsrelevante Form ist. Die archivierungsrelevante Form ist stets die rechtsrelevante, vollständigste Dokumentation des behördlichen Handelns. Sofern eine regelkonforme analoge oder digitale Aktenführung existiert und die Informationen aus der Dateisammlung in diese eingeflossen sind, sind die jeweiligen Akten rechtsrelevant und werden archiviert. Auf die Übernahme der Dateisammlung kann in diesem Fall verzichtet werden. Umgekehrt kommen für eine Archivierung stets nur diejenigen Dateisammlungen in Frage, die nicht oder nicht vollständig in die Aktenführung oder in ein anderes strukturiertes Informationssystem eingeflossen sind. Bei Bestandsbildnern ohne geregelte Aktenführung, beispielsweise privaten Nachlassgebern, kommt dieser Fall regelmäßig vor; in der regulären

[1] Menge von Einzeldateien, die von einem oder mehreren Bearbeitern zur Erledigung einer oder mehrerer Aufgaben über einen bestimmten Zeitraum erstellt und nach individuellen Ordnungskriterien zusammengestellt wurden. Die Dateien liegen auf einer Ebene und/oder hierarchisch in einer Verzeichnisstruktur vor. Es können in einer Dateisammlung unterschiedlichste Dateiformate enthalten sein.

Bürokratie tritt er leider häufiger auf als es wünschenswert wäre. Als besonders prominentes Beispiel aus der jüngeren Zeit sind hier die dienstlichen E-Mails des ehemaligen baden-württembergischen Ministerpräsidenten Mappus zu nennen, die zwar wichtige Dokumente staatlichen Handelns darstellen, die aber entweder gar nicht oder aber nicht vollständig in die Aktenführung der Landesregierung eingeflossen sind.[2]

Archivwürdigkeit vs. Archivfähigkeit

Da die nicht-aktenmäßige Informationshaltung rechtlich wie formal-inhaltlich einen Ausnahmefall im bürokratischen System darstellt (oder zumindest darstellen sollte), ist das Verhältnis von Archivwürdigkeit zur Archivfähigkeit bei diesem Material besonders kritisch zu prüfen – zumal die Archivfähigkeit von Dateisammlungen oft besonders problematisch ist. So präsentieren sich dem Archiv gerne intransparente Strukturen und schlecht bis gar nicht dokumentierte Sachbetreffe, die große Arbeitsaufwände im Archivierungsprozess nach sich ziehen. Sinnzusammenhänge, die sich erst im Kontext mehrerer Dokumente erschließen, lassen sich in manchen Fällen nur mit hohem Aufwand erschließen. Hinzu kommen oftmals technische Schwierigkeiten mit nicht archivfähigen Dateien, die im Einzelfall eine sehr intensive Eingangsbearbeitung erfordern. Ökonomisch problematisch sind auch Dateien, die sich nach dem heutigen Stand der Technik überhaupt nicht in ein archivfähiges Format wandeln lassen, wie z.B. ausführbare Dateien. Sollen diese Dateien bereits bei der Bildung der ersten Repräsentation als nicht-archivfähig kassiert werden oder mit viel Vertrauen in die Entwicklungen der Zukunft mit gespeichert werden?

In jedem Fall sollte vor der Übernahme einer Dateisammlung eine sorgfältige Analyse des Verhältnisses zwischen fachlich begründeter Archivwürdigkeit und fachlich und wirtschaftlich begründeter Archivfähigkeit stehen.

Zeitschnitte

Führt die Behörde Dateisammlungen anstelle von Akten, werden meist auch die klassischen Regeln der Aktenführung nicht eingehalten. Verzeichnisse werden in den seltensten Fällen geschlossen, Aufbewahrungsfristen nicht vermerkt und angebotene Daten nach der Übergabe ans Archiv nicht gelöscht und im schlimmsten Fall weiterbearbeitet. Zeitnahe Übernahmen lassen sich auch aufgrund fehlender Erhaltungsstrategien in den Behörden nicht vermeiden. Das Archiv ist daher oft gezwungen, Zeitschnitte einer Dateisammlung zu übernehmen, die jeweils einen Anteil an Dateien doppelt enthalten. Daraus ergeben sich zwangsläufig weitere Fragen: Lassen sich diese Zeitschnitte sauber voneinander abgrenzen? Wie soll mit Dateien umgegangen werden, die in einem Zeitschnitt bereits enthalten waren, im neuen Zeitschnitt in einer neuen Version (=weiterbearbeitet) vorliegen? Was ist mit den Dateien, die zwischen zwei Zeitschnitten angelegt und wieder gelöscht wurden?

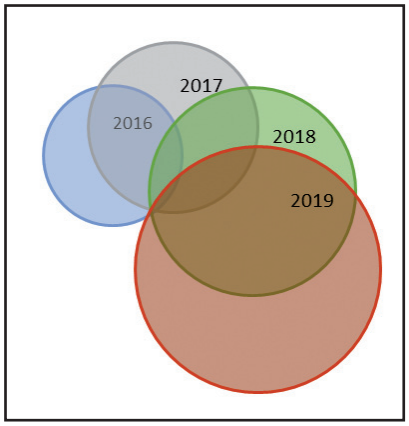


Abb 1: Schematische Darstellung der Entwicklung einer Dateisammlung im Laufe der Zeit

[2] Urteil des Verwaltungsgerichtshofs Baden-Württemberg vom 30. Juli 2014 (1 S 1352/13), Pressemeldung vom 4.8.2014, *http://vghmannheim.de/pb/,Lde_DE/2271892/?LISTPAGE=2271610*, Urteil des Verwaltungsgerichts Karlsruhe vom 27. Mai 2013 (2K 3249/12), Pressemeldung vom 31.5.2013, *http://vgkarlsruhe.de/pb/,Lde_DE/1740647/?LISTPAGE=1740537*.

Rekonstruktion von Strukturen

Obwohl Dateisammlungen oft ein für Außenstehende wenig verständliches Erscheinungsbild haben, liegt ihnen meistens doch eine rationale Struktur zu Grunde. Diese ist jedoch nicht ohne Weiteres zugänglich. Daher ist es wichtig, zu Beginn des Archivierungsprozesses alle verfügbaren Informationen zum Inhalt, zur Komposition, zum Zweck und Entstehungskontext der Sammlung zusammenzutragen. Wer hat was erstellt? Welche Aufgaben wurden hier von wem erledigt? Gibt es eventuell eine externe Dokumentation zur Dateisammlung? Können bei der Übernahme Informationen über die Struktur von den SachbearbeiterInnen eingeholt werden? Je mehr Informationen gesammelt werden, desto einfacher gestalten sich die sich anschließenden einzelnen Arbeitsschritte der Archivierung, insbesondere die Bewertung, die Paketierung und die Erschließung.

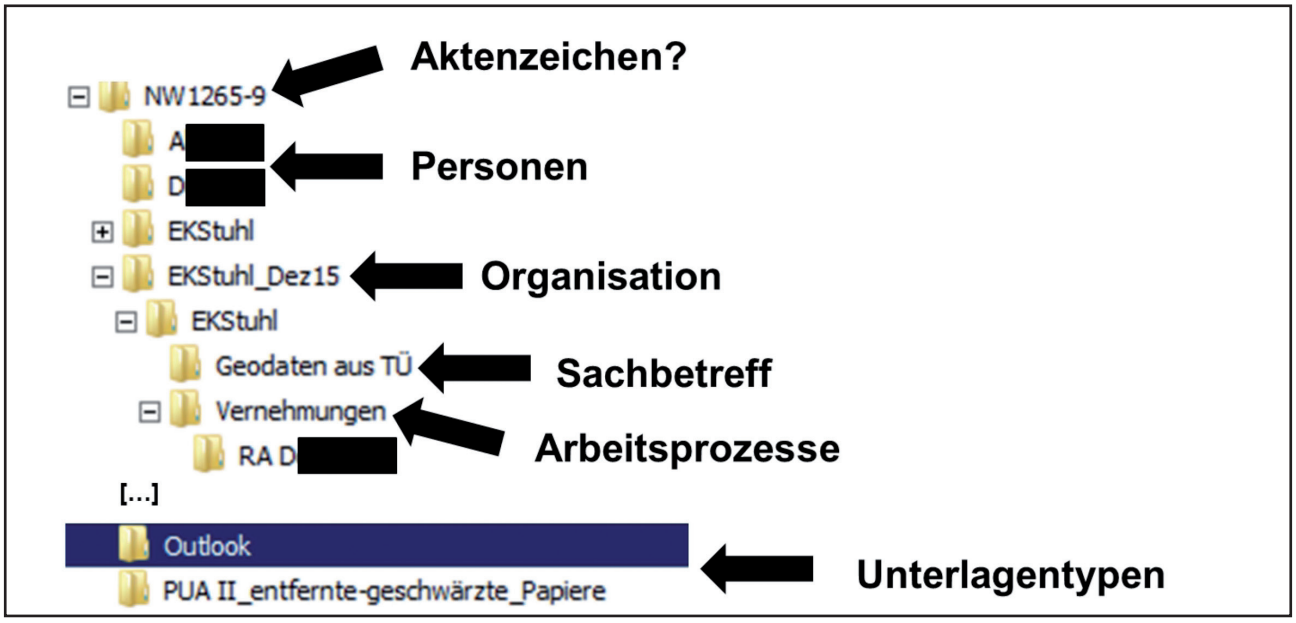


Abb. 2: Beispiel für eine schwer verständliche Struktur (Screenshot NRW Landeskriminalamt)

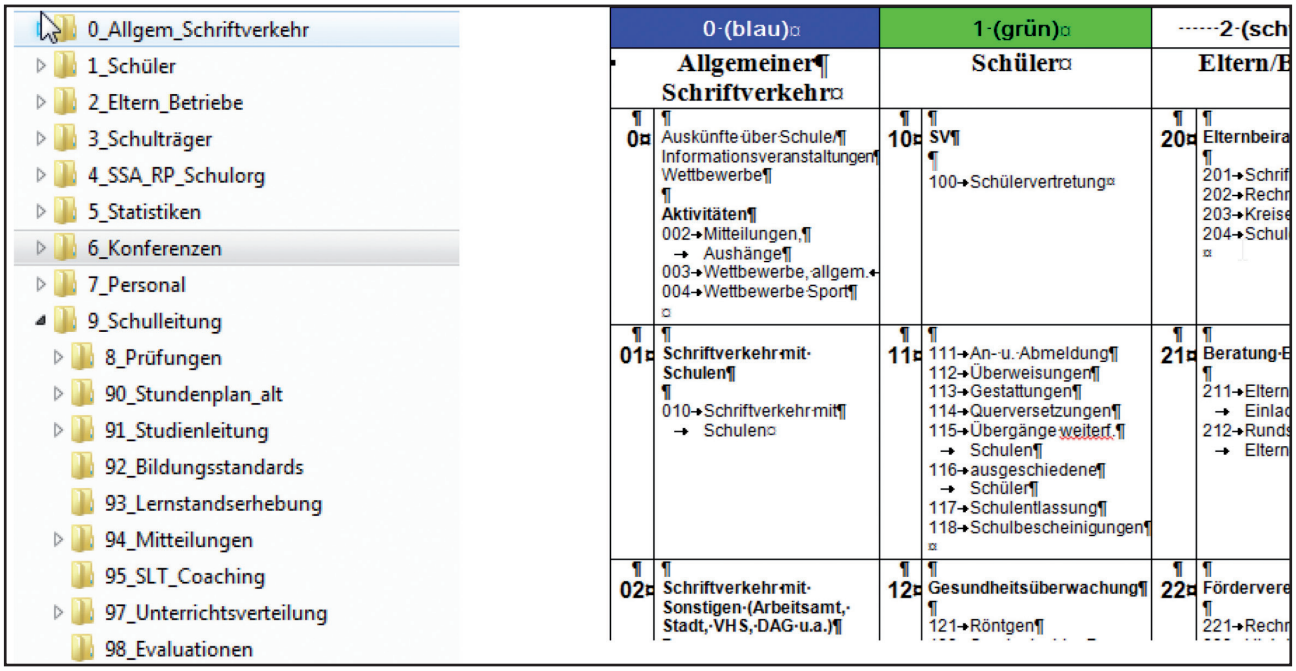


Abb. 3: Beispiel für eine dokumentierte und gut verständliche Struktur (Screenshot der Dateisammlung eines hessischen Gymnasiums und der darin enthaltenen Dokumentation zur Ablagestruktur)

Bewertung

Auch die Bewertung von Dateisammlungen richtet sich nach den klassischen Grundsätzen der Überlieferungsbildung. Soll die Dateisammlung in ihrer Gesamtheit archiviert oder kassiert werden oder kann eine Auswahlarchivierung erfolgen? Diese Frage ist gerade bei großen Dateisammlungen mit intransparenten Strukturen meist nicht ohne weiteres zu beantworten. Die Kenntnis über die Entstehungszusammenhänge ist – wie bereits erwähnt – für das Erkennen und Erhalten von vorarchivischen Strukturen und Sinnzusammenhängen von zentraler Bedeutung.

Dateiendung	Dateien	Beschreibung
Office Dateien	144773	Dokumente und Dateien von Office Prograr
.xlt	16	Microsoft Excel-Vorlage
.pot	343	Microsoft PowerPoint 97-2003-Vorlage
.pdf	6940	Adobe Acrobat Document
.mde	29	Microsoft Access MDE Database
.odt	25	OpenDocument Text
.dot	4902	Microsoft Word 97-2003-Vorlage
.odp	6	OpenDocument Präsentation
.pptx	90	Microsoft PowerPoint-Präsentation
.doc	103339	Microsoft Word 97-2003-Dokument
.ppt	83	Microsoft PowerPoint 97-2003-Präsentation
.xlsm	3	Microsoft Excel-Arbeitsblatt mit Makros
.docx	1377	Microsoft Word-Dokument

Abb. 4: Formatanalyse einer Dateisammlung

Die elektronische Datenverarbeitung eröffnet aber auch neue Möglichkeiten für die Bewertung. Der Einsatz von Analysetools zur Auswertung der Metadaten schafft einen schnellen Überblick über die Dateisammlung. Auf diese Weise kann zum Beispiel mit wenig Aufwand der Verzeichnisbaum mit den darin enthaltenen Dateiobjekten ausgeworfen werden. Die Analyse von Dateimengen, -größen und -formaten, Entstehungszeiträumen, Bearbeitern, Versions- und Dublettenprüfung etc. liefert wichtige Hinweise zur Zusammensetzung der Dateisammlung. Mittels der Dokumentenvorschau und Volltextsuchen in den Dokumenten kann zudem eine kursorische inhaltliche Sichtung erfolgen.

Im einfachsten Fall ist die gesamte Dateisammlung archivwürdig. Doch meist stellt sich schon bei der ersten Sichtung heraus, dass einige Dateien nicht archivwürdig oder auch gar nicht archivfähig sind. Dabei stellt sich grundsätzlich die Frage, auf welcher Ebene die Bewertung erfolgen soll und kann. Die meisten Dateisammlungen entziehen sich den klassischen Regeln der Schriftgutverwaltung. Sinnzusammenhänge und Strukturen sind angesichts komplexer Verzeichnisstrukturen und rudimentärer Verzeichnis- und Dokumententitel kaum oder nur mit hohem Aufwand nachvollziehbar und erschweren eine Bewertung auf der Ebene der vorarchivischen Formierung. Auf den ersten Blick erscheint in diesem Fall eine

Komplettarchivierung als pragmatische Lösung, um Sinnzusammenhänge und Strukturen zu erhalten und den Aufwand für die Bewertung im Rahmen zu halten. Unter Umständen werden auf diese Weise jedoch im großen Stil redundante Dateien (wie z.B. Dubletten, Versionen) oder nicht archivfähige Dateien (wie z.B. Programm- und Systemdateien, Viren) archiviert.

Stellt sich bei einer ersten Sichtung dagegen heraus, dass die Kassation einzelner Teile der Lieferung fachlich sinnvoll und ökonomisch möglich ist, muss im nächsten Schritt geklärt werden, auf welcher Ebene die Bewertung erfolgen soll. Gut strukturierte Ablagen können auf Verzeichnisebene bewertet werden. Im Idealfall orientiert sich die Verzeichnisstruktur an einem klassischen Aktenplan oder weist vergleichbare Ordnungskriterien auf. Wird dabei das Verzeichnis als formierte Einheit betrachtet, deren Entstehungskontext erhalten bleiben soll, sind Kassationen einzelner Dateien innerhalb dieser Strukturen ausgeschlossen.

Abb. 5: Gut strukturierte Dateisammlung, die eine Bewertung auf Verzeichnisebene erlaubt

Den stärksten Eingriff stellt die Bewertung auf Dateiebene dar. Allerdings ist die inhaltliche Sichtung von Einzeldateien je nach Umfang der Dateisammlung mit einem hohen Arbeitsaufwand verbunden. Eine mitunter sehr effiziente Alternative oder auch Ergänzung zur klassischen Bewertung kann die softwareunterstützte Auswahl nach technischen Kriterien wie z.B. nach Dateiformaten, sowie die Kassation von Dubletten und Versionen darstellen. Besonders die Ermittlung von Dubletten und Versionen bietet mitunter ein erhebliches Kassationspotential. Die Kassation einzelner Dateien bedeutet jedoch in jedem Fall die Zerstörung des unmittelbaren Entstehungskontextes und besonders bei der technischen Auswahl einen meist nur schwer abschätzbaren Informationsverlust.

Der Umgang mit Redundanzen ist dabei grundsätzlich zu klären. So enthalten besonders Dateisammlungen mit E-Mail-Korrespondenz einen hohen Anteil redundanter Elemente. Ähnlich verhält es sich mit Backup-Dateien. Wie werden Dateien aus Backups von den übrigen Dateien abgegrenzt? Ist es sinnvoll, durch die Wiederherstellung von Backups die Zahl der Dubletten um ein Vielfaches zu erhöhen, nur um dann noch mehr Zeitschnitte der gleichen Dateisammlung vorliegen zu haben?

Dennoch können beide Verfahren, die technische Auswahl und die Bewertung nach fachlichen Kriterien, durchaus gewinnbringend kombiniert werden. So wurde zum Beispiel bei der Übernahme einer Serversicherung einer Schule, die in großer Zahl Programm- und Systemdateien enthielt, durch eine technische Vorbewertung die Anzahl der Dateien um 75 % reduziert. Da die Auswahl auf Verzeichnisebene stattfand, blieb der Entstehungskontext innerhalb der vorarchivisch formierten Einheiten erhalten.

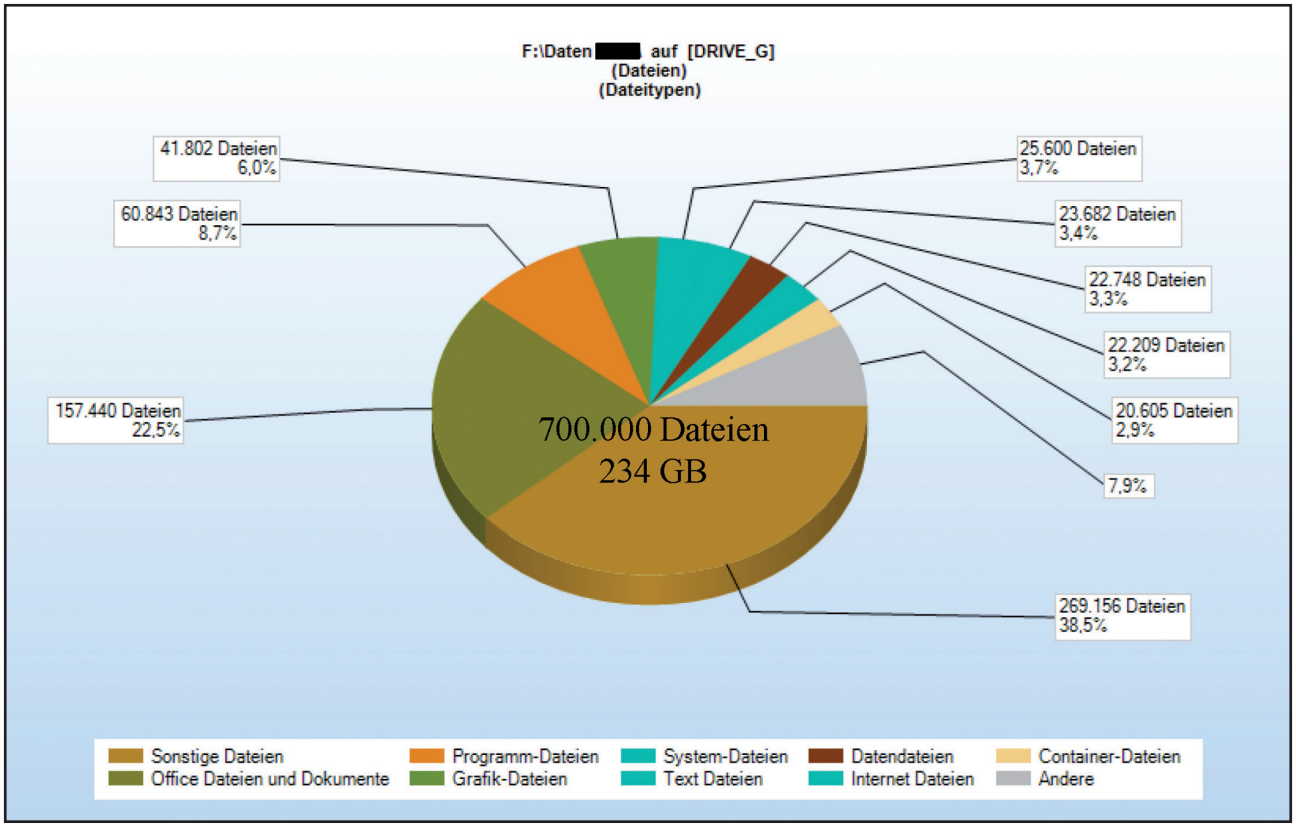


Abb. 6: Dateiformatanalyse über ca. 700.000 Dateien der Serversicherung der hessischen Odenwaldschule

Einen Ausweg aus dem Dilemma, dass durch die Kassationen auf Dateiebene der Entstehungszusammenhang zerstört werden könnte, ist die Dokumentation der vorarchivischen Verzeichnisstruktur und aller Dateien des Zugangs. In einer Bestandsaufnahme können vor der Bearbeitung durch das Archiv alle Dateien mit Verzeichnispfad, Dateinamen, Hashwerten etc. nachgewiesen werden. Darüber hinaus sollte die Vollständigkeit und Integrität der Daten während des gesamten Bearbeitungsprozesses regelmäßig geprüft und alle Eingriffe seitens des Archivs, insbesondere Kassationen, protokolliert werden.

Dokumente und Einstellungen\Administrator\Anwendungsdaten\Adobe\Acrobat\7.0\AdobeCMapFnt07.lst	4014a2ad4a0b71b8487adfcfc63a5b31
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Adobe\Acrobat\7.0\AdobeSysFnt07.lst	be697889689b6727390162fc8fe994de
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Adobe\Acrobat\7.0\JavaScripts\glob.settings.js	57f3d8f5bcc781fb4a36b750bdd0aeaa
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Adobe\Acrobat\7.0\UserCache.bin	94993866b43cfe22bdf0aca8e6d36173
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fax\FritzFax.dbf	2763a71f21db7f61a9db8ba3f050c17e
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fax\FritzFax.dbk	82843ba29bd014ca06fc06d159f99f9b
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fax\FritzFax.mdk	98857ff30322f281f98e539961ff6080
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fax\FritzFax.mdx	abab51bf4a3773513dfbef483288574e
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fon\FRITZFON.DBF	53a89df81143824688346449df23ab89
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fon\FRITZFON.MDX	6e3fab7c4d59e6cb817e9b68a1c3c638
Dokumente und Einstellungen\Administrator\Anwendungsdaten\FRITZ!\Fon\FritzFon.Wav	ee570d060af6ff4e6f02bd220d268fb1
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Macromedia\Flash Player\macromedia.com\support\fl	373b89e5f1d9b2cf502d7a04999b1bcf
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\Address Book\Administrator.wab	983384fbb9cda612716e3639177c6913
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\Address Book\Administrator.wa~	983384fbb9cda612716e3639177c6913
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v1.1.4322\security.confi	c8be206b08bf692cc539f7f528b028ac
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v1.1.4322\security.confi	18f1ae83368510b76a6d696ddbefe230
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v1.1.4322\security.confi	c8be206b08bf692cc539f7f528b028ac
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v2.0.50727.42\security.c	766828fe64e86592391a9ce1b9956e39
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v2.0.50727.42\security.c	8f7eac32d2bc5a42a7899252cfcde44f
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v2.0.50727.832\security.	766828fe64e86592391a9ce1b9956e39
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CLR Security Config\v2.0.50727.832\security.	1b839851a70e1733019ec8ea2c92976b
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CryptnetUrlCache\Content\7B2238AACCEDC	e0520a90f3236cf00bcdee6d516c450b
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CryptnetUrlCache\Content\A8FABA189DB7D	bf618cabb5afe6813eec78f1ea1f63b5
Dokumente und Einstellungen\Administrator\Anwendungsdaten\Microsoft\CryptnetUrlCache\Content\F482C95F83F1B5	8f7db8a3b898929890e236a84906ed59

Abb. 7: Bestandsaufnahme einer Dateisammlung (SIP) mit Verzeichnispfad, Dateiname und Hashwert

Technische Eingangsbearbeitung

Neben der Bewertung stellt vor allem die technische Aufbereitung der Daten für den Archivierungsprozess eine Herausforderung dar. Erste Hürden tun sich meist schon bei der Zugangssicherung von Dateisammlungen mit tiefen Verzeichnispfaden auf, da überlange Dateipfade eine einfache Sicherung auf Windowssystemen verhindern können.

Unstrukturierte Dateisammlungen enthalten zudem oft viele unterschiedliche Dateiformate, passwortgeschützte oder verschlüsselte Dateien, Containerformate, invalide oder korrupte Dateien und andere ‚Problemfälle', die sich zum Teil nur mit erheblichem Aufwand in archivfähige Formate konvertieren lassen.

Dateisammlungen sind zudem nicht selten auch ein Sicherheitsrisiko für die IT-Infrastruktur. Schadsoftware wird in gepackten Archiven von den üblichen Virenscannern nicht immer erkannt. Für die Bewertung und AIP-Bildung müssen diese jedoch entpackt werden.

Vor einer Archivierung müssen all diese Probleme möglichst effizient identifiziert und gelöst werden – durch Entpacken, Kassieren, Konvertieren usw. Dazu gehört übrigens auch die Frage, ob die notwendige technische Infrastruktur überhaupt zur Verfügung steht und die Bearbeitungsverfahren den anfallenden Datenmengen gewachsen sind.

Bildung von Archivinformationspaketen

Bei der Archivierung von analogem Schriftgut spielt die Frage der Formierung von Archivalieneinheiten eher eine nachgeordnete Rolle. Die von der abgebenden Stelle formierte (Akten)einheit bleibt in der Regel als Archivalieneinheit erhalten. Dieser Grundsatz lässt sich bedingt noch auf elektronische Akten übertragen, insofern diese in einem den Regeln der Schriftgutverwaltung genügenden Dokumentenmanagementsystem entstanden sind. Bei der Übernahme von Dateisammlungen stellt sich dagegen grundsätzlich die Frage der Paketierung: Welche Dateien eines Übergabeinformationspakets (SIP) sollen zusammengefasst werden und bilden später ein Archivinformationspaket (AIP)?

So können zum Beispiel gleichförmige Reports aus einer Datenbank auf der Ebene der Einzeldatei (eine Datei bildet ein AIP), nach dem Ordnungskriterium des Entstehungsjahrgangs (alle Dateien eines Jahrgangs bilden ein AIP) oder des gesamten Übernahmeinformationspakets (alle Dateien eines SIPs bilden ein AIP) gruppiert werden. Je nachdem, ob ein enger oder weiter Fokus für die Paketierung gewählt wird, bewegt sich die Spannbreite der zu bildenden Einheiten zwischen einem und mehreren hunderten oder tausenden Archivinformationspaketen. Dateisammlungen mit komplexen Verzeichnisstrukturen lassen sich dagegen meist nur schwer nach einheitlichen Ordnungskriterien formieren. Meist bleibt als einzige Alternative, individuell je AIP zu entscheiden, welche Verzeichnisse zusammengefasst werden sollen. Dies setzt jedoch eine vertiefte inhaltliche Sichtung des Materials voraus.

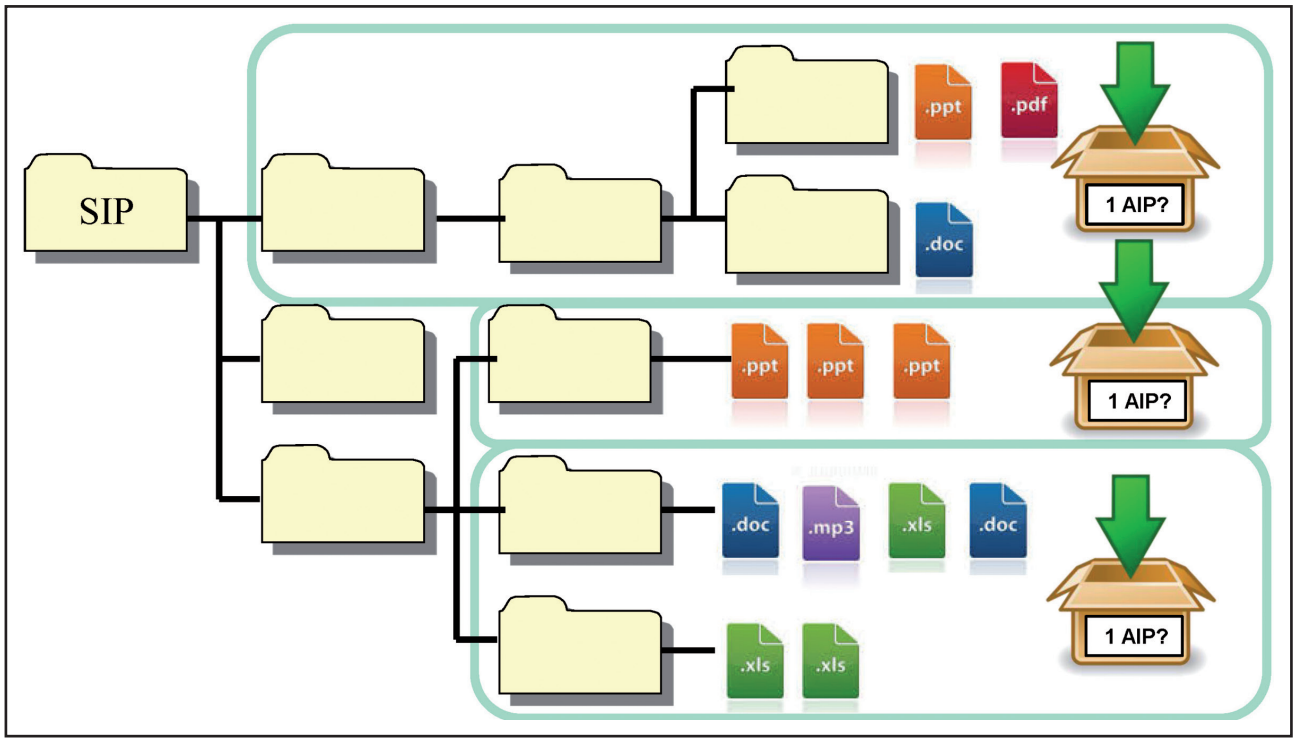


Abb. 8: AIP-Bildung bei einer beliebig strukturierten Dateisammlung

Die Entscheidung, die in diesen Fällen getroffen werden muss, ist abhängig vom Datenmodell des betreffenden Übernahmearchivs, von den angestrebten Nutzungsszenarien sowie von technischen Rahmenbedingungen. Wichtig ist außerdem eine Abwägung zwischen dem Aufwand bei der Bewertung und der AIP-Bildung und dem Aufwand der langfristigen Erhaltung. Eine oberflächliche Bewertung führt in der Regel zur Bildung sehr großer oder minimal kleiner AIPs, die eine Herausforderung für die Nutzung und die Erhaltung darstellen können. Umgekehrt führt eine intensive inhaltliche Durchdringung des Materials oft zur Bildung genau angemessener AIPs, die in Bestandserhaltung und Nutzung weniger Probleme verursachen.

Erschließung

Da in aller Regel die Anzahl der gebildeten AIPs auch der Anzahl der Verzeichnungseinheiten entspricht, resultiert der grundsätzliche Erschließungsaufwand für unstrukturierte Dateisammlungen aus dem Ingestprozess und den dort getroffenen Setzungen. Je nachdem, welche Entscheidung vorab getroffen wurde, kann der Erschließungsprozess hier sehr arbeitsintensiv sein – z.B. wenn sehr viele AIPs oder sehr große AIPs mit sehr vielen Einzeldateien gebildet wurden oder wenn die Strukturen des Materials so diffus sind, dass eine inhaltliche Durchdringung seiner Inhalte vorab nicht möglich war. Auch an dieser Stelle wird deutlich, dass es im Umgang mit unstrukturiertem Material von grundlegender Bedeutung ist, den gesamten Prozess der Archivierung von Anfang an mitzudenken und insbesondere die Folgen der unterschiedlichen Paketierungsoptionen so genau wie möglich vorab zu prüfen.

Nutzung

Enthalten die aus den Daten einer unstrukturierten Dateisammlung gebildeten AIPs mehr als eine Datei und wurden diese nicht durch eine nachträglich hinzugefügte Strukturdatei (z.B. METS) ergänzt, so kann sich die Nutzung als problematisch erweisen. Je nachdem, wie anwenderfreundlich die Nutzung sein soll, kann es notwendig werden, eine eigene Nutzungsinfrastruktur vorzuhalten, die das Auslieferungsinformationspaket (DIP) in seiner Struktur interpretieren und die benötigten Einzelviewerkomponenten koordinieren kann. Zudem kann es für das Verständnis des Materials notwendig sein, die ursprünglichen Strukturen des archivierten Materials nachvollziehen zu können. Dies kann über die Bereitstellung der entsprechenden Metadaten, zum Beispiel der Bestandsaufnahme der Struktur des SIP vor Bewertung, Kassation und AIP-Bildung, oder die Berücksichtigung der Ausgangsstrukturen bei der Bildung der Archiveinheiten erreicht werden.

Anpassung des Archivierungsworkflows?

Bei der Übernahme von analogen oder digitalen Akten folgen üblicherweise verschiedene Arbeitsschritte aufeinander: Bewertung (vor Ort), Übernahme, Magazinierung der bereits in der Behörde formierten Einheiten und schließlich – teilweise erst Jahre später – die Erschließung.

Betrachtet man nun die zuvor beschriebenen Schritte der Übernahme von Dateisammlungen, so fällt auf, dass es hier größere Abweichungen geben kann – vor allem, wenn man sich für eine Bewertung mit Teilkassation und eine Bildung von mehreren AIPs pro Dateisammlung entscheidet. In diesem Fall lassen sich die Arbeitsschritte der Bewertung, Formierung (AIP-Bildung/Magazinierung) und Erschließung bei der Eingangsbearbeitung von Dateisammlungen oft nicht von einander trennen. Aus pragmatischen Gründen werden Dateisammlungen selten vor Ort in der Behörde bewertet: Die Bewertung ist aufwändig und die komplette Übernahme auf einem Datenträger verhältnismäßig einfach. Daher werden Dateisammlungen oft als Ganzes übernommen und erst im Nachhinein im Detail bewertet. Die Sichtung des Materials ist gerade bei schwach strukturierten Ablagen zeitaufwändig, so dass die Bewertung, Formierung und Erschließung aus Gründen der Zeitökonomie besser in einem Arbeitsschritt oder in einer flexiblen Abfolge durchgeführt werden sollten. Dies kann sich in der Praxis jedoch als problematisch erweisen, da viele Archivierungssysteme technisch-organisatorisch festgelegte Workflows voraussetzen.

Fazit

Wir haben in diesem Beitrag versucht, die ersten praktischen Erfahrungen mit der Übernahme von Dateisammlungen im Hessischen Landesarchiv und im Landesarchiv Nordrhein-Westfalen auszuwerten und dabei herauszuarbeiten, welchen grundsätzlichen Fragen sich ein Archiv vor der Übernahme von Dateisammlungen stellen und was es bei der Übernahme bedenken sollte. Naturgemäß liegt der Hauptaugenmerk dabei vor allem auf den mannigfaltigen Fallstricken, die eine solche Übernahme mit sich bringen kann. Es wurden viele mögliche Probleme genannt, für die es aktuell noch keine Patentlösungen gibt, sondern höchstens Lösungsansätze.

Um jedoch auf die eingangs getroffene Feststellung über die Verbreitung von Dateisammlungen zurückzukommen – Dateisammlungen sind im Behördenalltag zwar nicht regelkonform, aber auch nicht selten. Häufiger als uns ArchivarInnen lieb ist, sind sie die einzige Form, in der archivwürdige Informationen vorliegen. Es ist also keine Option, Dateisammlungen grundsätzlich als „zu schwierig“ zu deklarieren und nicht zu übernehmen. Wenn Archive „Fehler“ bei der Übernahme von Dateisammlungen machen, weil die technischen oder personellen Ressourcen noch nicht ausreichend oder die konzeptionellen Überlegungen noch nicht weit genug gediehen sind, so ist das zwar bedauerlich, sehr viel bedauerlicher wäre es jedoch für die spätere Forschung, wenn archivwürdige Dateisammlungen gar nicht erst übernommen würden.

Rasche und einfache Bearbeitung von Dateisammlungen: ein MPLP-Ansatz

Susanne Belovari

In einem dreiwöchigen Kooperationsprojekt entwickelten das *Staatsarchiv Ludwigsburg*[1] und Susanne Belovari vom *Universitätsarchiv* der *University of Illinois* einen ersten einfachen Zugang, um Dateisammlungen rasch und mit wenig und ein facher Software bewerten und bearbeiten zu können. Möglich wurde dieses Projekt erst durch die fachliche Betreuung von Corinna Knobloch vor Ort.

Dateisammlungen sind typischerweise riesig, unstrukturiert, haben tiefe Ordnerhierarchien und sehr viele (fast identische) Duplikate, Nichtessentielles, schwierige Dateiformate und unpräzise Metadaten. Bewertung und „Verdichtung" sind daher meist unabdingbar, um die Größe und Komplexität solcher Sammlungen zu reduzieren und zukünftige Benutzer nicht nach der berühmten Nadel im Heuhaufen suchen zu lassen. Obwohl Archive auch für digitale Sammlungen verantwortlich sind, fehlen uns oft Ressourcen (Zeit, Geld, Hardware, Arbeitskraft, Fachwissen usw.) sowie erprobte einfache Programme und Workflows, um Dateisammlungen zu bearbeiten.

Viele von uns suchen daher schon länger nach einfachen Bewertungs- und Bearbeitungszugängen ähnlich der Herangehensweise „More Product Less Process" (MPLP) für analoge Unterlagen.[2] MPLP-Ansätze versuchen, ein akzeptables Gleichgewicht zu schaffen zwischen dem Risiko, essentielle Dateien zu übersehen, und Anforderungen, die aus dem Mangel an Personal und Speicherplatz hervorgehen und eine tiefgehende Bearbeitung schwierig oder unmöglich machen. Wie bei analogen Sammlungen werden wir auch bei Dateisammlungen hybride Bearbeitungszugänge entwickeln müssen, d.h. sammlungsspezifische Kriterien festlegen, nach denen wir bewerten, bearbeiten und Entscheidungen und Eingriffe dokumentieren. Außerdem werden wir genauso bereit sein müssen, digitale Unterlagen möglicherweise „verrotten" zu lassen, d.h. sie zu lagern, obwohl sie weder perfekt bearbeitet noch in Standardformate migriert sind.

Um digitale Ressourcen zu bearbeiten, braucht es allerdings angemessenes digitales Werkzeug, genau so wie es früher Schaufeln und später Bagger brauchte, um Steinkohle zu fördern. Erwünschte *einfache Zugänge* zur digitalen Bewertung erfordern daher *einfache oder einfach erlernbare Programme*, die uns helfen, Bewertungen zu automatisieren, zu beschleunigen und zu vereinfachen. Bearbeitungsschritte erfolgen dann auch mittels Programmen, die ja nichts weiter als Bitketten sind, welche von IT und Firmen mit ihren eigenen Bedürfnissen, Zielsetzungen und Profitdenken zusammengefügt wurden. Daher sollte es selbstverständlich sein, dass wir oder unsere Fachorganisationen anfangen, Softwarekapazitäten zu testen und zu bewerten. Denn nur so können wir sicherstellen, dass Programme Bearbeitungsschritte zuverlässig und dokumentierbar durchführen. Wo aber fängt man an?

[1] Das Staatsarchiv Ludwigsburg ist Abteilung 5 des Landesarchivs Baden-Württemberg.

[2] Mark Greene und Dennis Meissner, More Product, Less Process: Revamping Traditional Archival Processing. In: The American Archivist: Fall/Winter, Vol. 68 (2005), Nr. 2, S. 208–263. Für erste digitale MPLP Überlegungen: More Product, Less Process for Born-Digital Collections: Reflections on CurateCamp Processing, Gastposting von Meg Phillips, Electronic Records Lifecycle Coordinator for the National Archives and Records Administration. *https://blogs.loc.gov/thesignal/2012/08/more-product-less-process-for-born-digital-collections-reflections-on-curatecamp-processing/* (aufgerufen am 20.9.2016).

Software-Tests bezüglich der Entfernung von Doppelungen und leeren Dateien/Ordnern

Angesichts mangelnder Ressourcen (Zeit, Geld, Hardware, Arbeitskraft, Fachwissen usw.) empfehle ich dort anzusetzen, wo wir uns im weitesten Sinn den größten Nutzen und den geringsten Aufwand versprechen. Hier stellt die Anzahl an digitalen Duplikaten sicherlich alles in den Schatten, was Archive an analogen Duplikaten gewohnt sind. Das war auch bei unserer Projektsammlung der Fall. Wie bekannt, wird seit längerem ein bestimmter Grad an analoger Redundanz aus Speicher-, Zeit-, Kontext- und anderen Gründen akzeptiert. Im Sinne von MPLP erarbeiteten etwa amerikanische ArchivarInnen individuelle Lösungen, die entweder gar nichts, nur leicht ersichtliche Kopien (z.B. große Berichte) oder innerhalb von Ordnern z.B. ein Bündel leicht wahrnehmbarer, weil farbiger Informationsblätter, entfernten. Ähnlich differenzierte Zugänge werden gewiss auch für digitale Duplikate notwendig sein; es wurden auch schon welche vorgeschlagen, deren Zugangsweisen oder Programme aber noch nicht getestet wurden.[3] Zweifellos können solche doppelten (und leeren) Dateien und Ordner aber nur mit Software effizient entfernt werden.

Ich testete daher zehn Programme zur Deduplizierung mehrfach anhand einer Stichprobe, die 10 % der Dateien (inklusive langer Dateinamen), 20 % der Ordner, ein Drittel der Duplikate und eine Vielfalt an Formaten der Projektsammlung beinhaltete. Acht der zehn Softwareprodukte versagten entweder schon beim einfachen Identifizieren des Inhaltes (Dateien, Ordner, Größe), beim Identifizieren der Anzahl doppelter Dateien (welches natürlich alle weiteren Schritte beeinflusste) oder bei der korrekten Deduplizierung. Außerdem waren sie bezüglich Softwareeinstellungen, graphischen Darstellungen und Berichts- und Dokumentationsfunktionen unzulänglich. Mein Richtwert für korrekte Identifizierung war die „Eigenschaften"-Funktionalität des Windows Explorers sowie die „Get Info"-Funktionalität von Mac und schlussendlich die Resultate von *Tree Size Pro* (genutzt wurde die 30-Tage-Testversion).[4]

Bei zusätzlichen Tests der zwei übriggebliebenen Programme, *Duplicate File Detective* (v. 6.0.76) und *TSP*, konnte nur *TSP* alle Duplikate regelmäßig identifizieren und entfernen und bestach durch einfache Handhabung, Einstellungsmöglichkeiten, Geschwindigkeit (< 1 Min.), großartige Grafikeinstellungen und Berichtsfunktionen, welche die Verifizierung und Dokumentation ermöglichen.[5]

In einer zweiten Serie von Tests überzeugte auch hier *TSP* im Vergleich zu *Remove Empty Directories* bezüglich besserer und korrekter Identifizierung und Löschung leerer Dateien/Ordner und Berichtsfunktionen.[6]

[3] Die *Bentley Historical Library* der University of Michigan, USA, fand sich mit „einem bestimmten Grad an Redundanz in Sammlungen ab und verwendet Duplikaterkennung hauptsächlich für die Identifizierung ganzer Ordner oder Verzeichnisse." In: Mike Shallcross, The Work of Appraisal in the Age of Digital Reproduction, Teil des Mellon *ArchivesSpace-Archivematica-DSpace Workflow Integration* Projektes. *http://archival-integration.blogspot.com/2015/06/the-work-of-appraisal-in-age-of-digital.html* (aufgerufen am 20.9.2016).

[4] CloneSpy v. 3.24., Directory List and Print v. 3.21 (freeware), Duplicate Cleaner v. 3.2.7., Duplicate File Detective v. 6.0.76, Duplicate File Eraser v. 2.0.1.0., Duplicate File Finder v. 6.0. (Freeware), Fast Duplicate Eraser, Fast Duplicate File Finder v. 4.7.0.1., WinMerge 2.14.0., Tree Size Pro 6.3 (genutzt wurde die 30-Tage-Testversion).

[5] Wir sollten hier unrealistische Ansprüche vermeiden: ArchivarInnen, die mit analogen Unterlagen arbeiten, dokumentieren ja auch weder ihre Bewertungskriterien noch Bewertungsschritte in allen Einzelheiten.

[6] Wahrscheinlich basiert TSPs gute Performance auf der Entwicklungspartnerschaft mit Microsoft und anderen, auf Grund dessen die Firma neue Spezifikationen im Voraus erhält.

Auf Grund der Testresultate und weil es TreeSize (in der Basis-Version) als Freeware und zugleich (als TreeSize Pro) mit wirklich erschwinglicher Lizenz gibt, beschloss ich, TSP zur groben Bewertung zu verwenden. Und ich wollte ausprobieren, ob es zur qualitativen Bewertung taugt. Ein einfacher Zugang bedeutet nämlich auch, dass ich nur ein oder sehr wenige Bearbeitungsprogramme einsetzen muss.

Bearbeitungsschritte: Ein Beispiel anhand der Johannes-Wagner-Schule Nürtingen für Hörgeschädigte und Sprachbehinderte

1. Rasche Durchsicht: erste Bewertungsfragen und Kriterien, potentielle Risiken

Nachdem ich mich über die abgebende Stelle und die Abgabe informiert hatte, sichtete ich die Projektdateisammlung (1960–2015) der Schule visuell und danach mit Hilfe von *TSP.* Ursprünglich beinhaltete die Sammlung 677 GB, 65057 Dateien, 3638 Ordner und eine komplexe (mehr als 10stufige) Ordnerhierarchie. An Formaten lagen hauptsächlich Fotos und AV-Dateien, digitalisierte Schmalfilme und Zeitungsartikel sowie Office-Dateien vor. Mit dabei war auch eine, leider ungenügende, Inventarliste. Visuell schätzte ich, dass ca. 70 % der Ordner und Dateien doppelt existierten.[7]

Im Gegensatz zur analogen Bewertung konnte ich nun *zuerst* mit Hilfe guter Software und Segmentanalyse mit der Löschung der Duplikate beginnen, ohne ein Inventar anfertigen zu müssen. Wenn vom Archiv gewollt, werden so bestimmte Datei- und Ordnerarten gelöscht, riesige Sammlungen reduziert und so die weitere qualitative Bewertung vereinfacht.[8] In unserem Fall versprach uns eine solche Löschung die größte Rendite mit kleinstem Aufwand und machte die rasche qualitative Bewertung erst praktikabel, da die Sammlung um über zwei Drittel reduziert wurde.

2. Grobe Bewertung

In meinem Fall umfasste die Entfernung von Dopplungen zwei Schritte: das manuelle Löschen doppelter Hauptordner und das Löschen doppelter Dateien mit Hilfe von Software.

2.1 Entfernung von Doppelungen

- Manuell (60 Min.): Löschung der visuell ersichtlichen, doppelten Hauptordner und Hauptverzeichnisse (minus zwei Drittel Speicherplatz).
- Software (< 1 Min.): mit *TSP dokumentierte* Löschung der doppelten Dateien (minus 8 % Dateien)

[7] Ob eine beschleunigte Bewertung ein unakzeptables Risiko darstellt, muss im Einzelfall geklärt werden. Unsere Sammlung enthielt keine sensiblen Unterlagen wie Schülerdaten. Aus Datenschutz- und Urheberrechtsgründen wird sie in absehbarerer Zeit nicht im Internet zugänglich sein.

[8] Gemäß der Fachliteratur beinhalten abgegebene Dateisammlungen 20 % bis 50 % an (Fast-)Duplikaten. Während in Englisch-sprechenden Ländern analoge Duplikate ohne Weiteres entfernt werden, ohne deren ursprüngliches Lokalität zu dokumentieren, wird in Europa der Verlust des „Kontextes“ bei Löschung digitaler Doppelungen als Argument gegen solche Löschungen verwendet. Aus meiner Erfahrung geht bei den wenigsten jener Löschungen immanent bedeutungsvoller Kontext verloren oder stellt angesichts mangelnder Archivressourcen einen „akzeptablen“ Verlust dar. Sollte das nicht der Fall oder das Risiko inakzeptabel sein, schlage ich *hard links* vor, die dokumentieren, wo doppelte Dateien vor deren Löschung existierten.

2.2 Entfernung leerer Ordner und leerer, temporärer und rein technischer Dateien

- Software (wenige Sek.): mit *TSP* Filter- und graphischen Ansichtsoptionen gelang eine *rasche* Identifikation und *dokumentierte* Löschung leerer Ordner und leerer, temporärer oder rein technischer Dateien.

3. Qualitative Bewertung

Da ich an einer beschleunigten qualitativen Bearbeitung interessiert war, wollte ich Dateien natürlich weder umbenennen noch umstrukturieren noch verschieben, was ja auch dem archivarischen Respekt für die Provenienz, die vorgefundene Ordnung und, wie ich meine, die vorgefundenen Bezeichnungen widerspräche. Um die qualitative Bewertung zu priorisieren und zu beschleunigen, musste das verwendete Programm rasche, visuelle Übersichten sowie Segmentanalysen anbieten und einfache Workflows ermöglichen.

3.1 Entscheidung bezüglich Bewerten oder Nichtbewerten

Unsere Projektsammlung hatte nun zirka 30 % der ursprünglichen Größe und Dateienanzahl sowie 40 % der ursprünglichen Ordner. An diesem Punkt kann ich dann entscheiden, ob sich eine tiefere qualitative Bewertung mit Hinblick auf z.B. Verdichtung der Akten, Speicherplatz, Arbeitszeit, Erhaltungsformat, Risiken, Kosten, Programme usw. auszahlt.

Mit *TSP*, besonders dessen graphischen Ansichten, konnte ich innerhalb weniger Sekunden nach Format, Alter, Größe, Ordner usw. filtern, Resultate verifizieren und Material oder Ordner leicht vorreihen, bei denen sich gemäß meiner selbst gewählten Kriterien die Bewertung auszahlt (Abb. 1). Nachdem Verzeichnisduplikate gelöscht worden waren, hatten wir jetzt drei Hauptverzeichnisse: Foto- und Video-Archiv (abgekürzt FVA, 98 %), Schulportfolio (1,6 %) und Themen (0,5 % Speicherplatz). Hier entschied ich mich, z.B. Video- und Bilddateien zu bewerten (7,3 % an Videodateien belegten drei Viertel des Speicherplatzes), Office-Dateien aber wegen der geringen Speicher- und Migrationsprobleme zurückzustellen

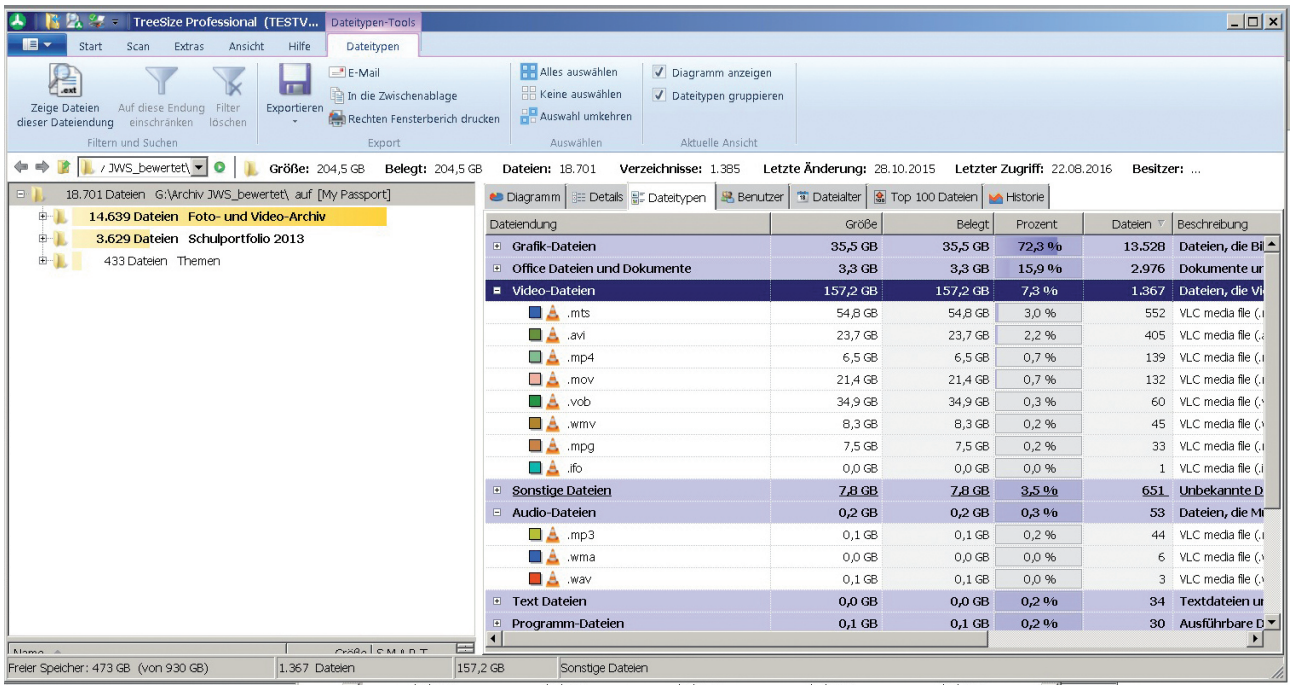


Abb. 1: Links: drei Hauptordner. Rechts: Dateitypen für Hauptordner, Foto- und Video-Archiv (FVA)

(Abb. 2). Ich übernahm ältere Video- und Fotoordner bis 2004 (< 2 % der Dateien) komplett, da jeweils wenige Dateien das historische Geschehen dokumentierten. Jüngere Ordner von 2009–2014 (> 85 % der Dateien) unterzog ich einer qualitativen Bewertung (Abb.3).

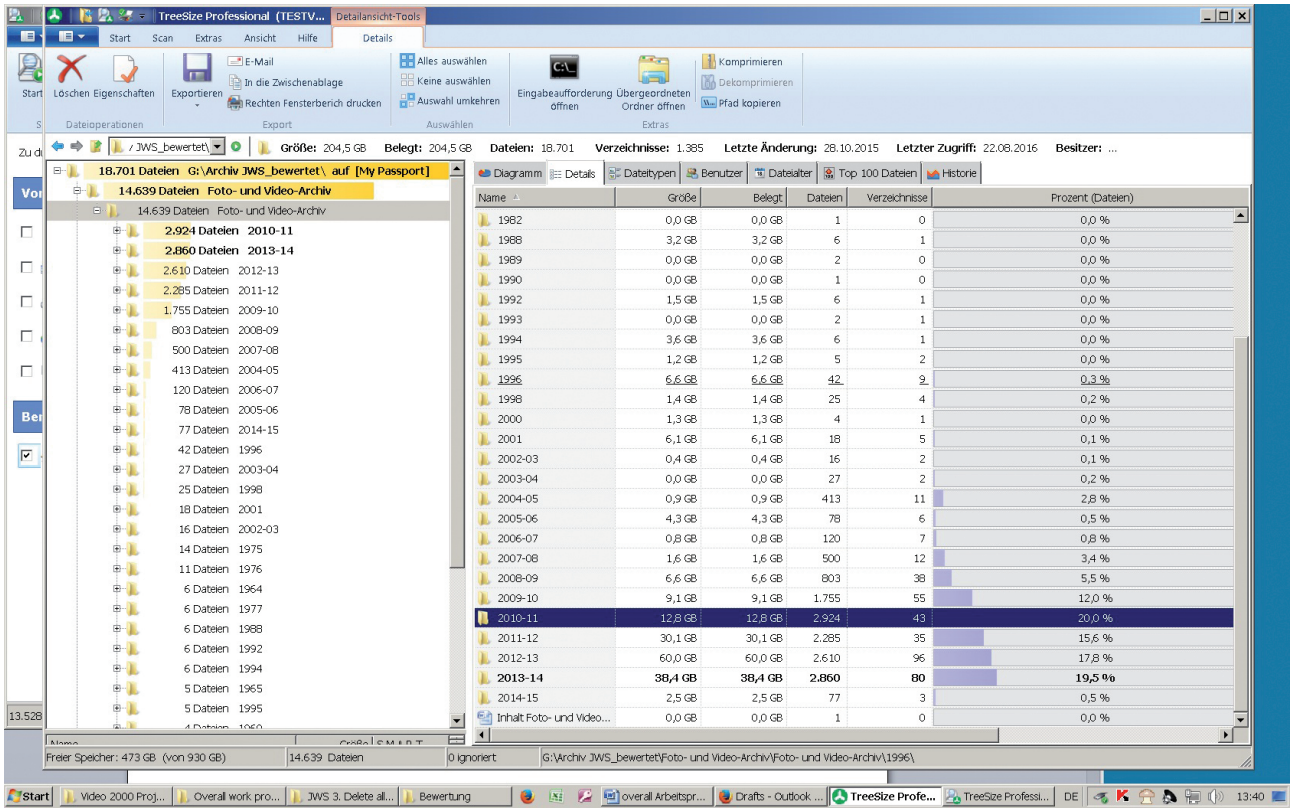


Abb. 2: Links: FVA Schuljahrordner. Rechts: Schuljahre nach Größe + Dateienanzahl geordnet

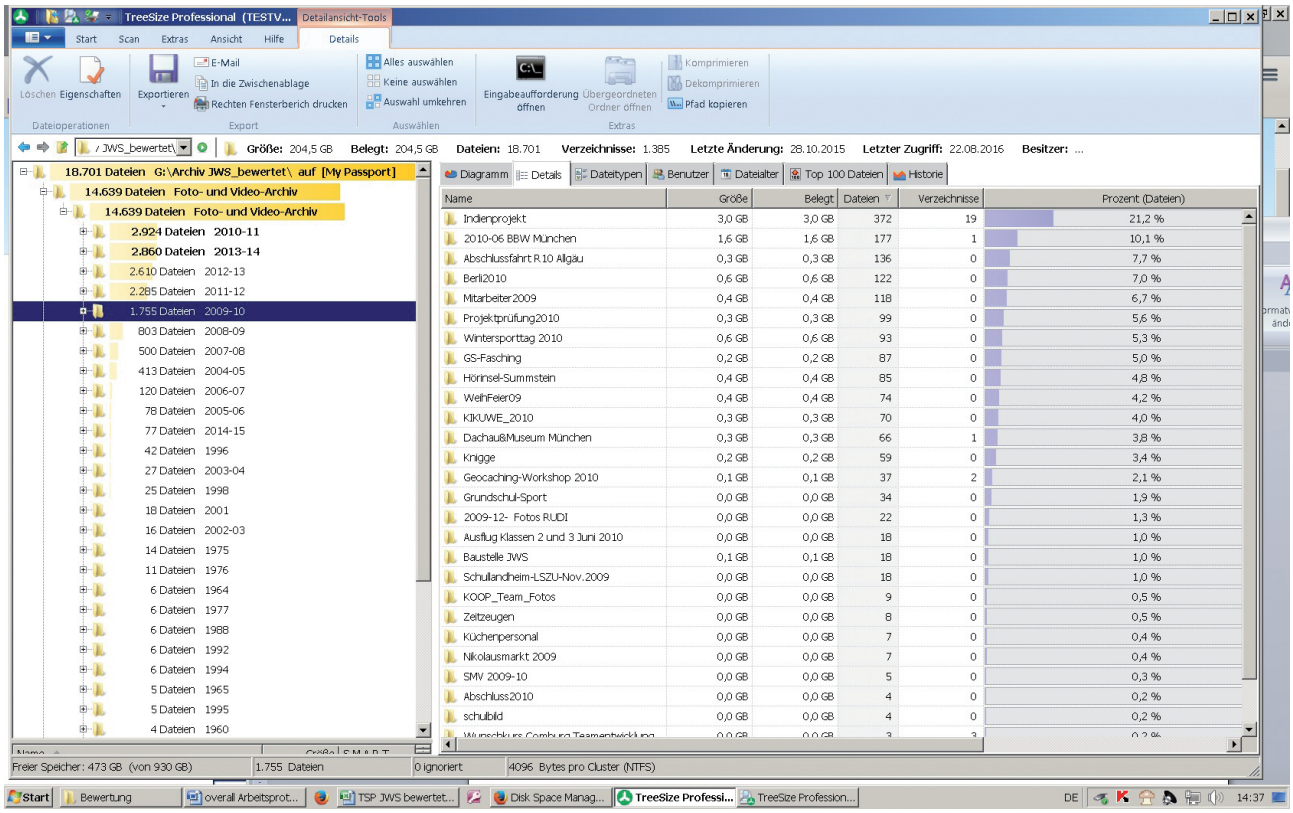


Abb. 3: Links: FVA Schuljahr 2009–10. Rechts: Unterordner nach Größe + Dateienanzahl geordnet

3.2 Qualitative Bewertungskriterien (die auch während der Bewertung entwickelt werden)

Qualitative analoge Bewertung ist eine wahrlich komplexe Aktivität, die wir meist so geschwind durchführen, dass wir den Prozess kaum erklären oder reproduzieren können. Die qualitative digitale Bewertung wird wahrscheinlich nicht anders sein. Leider ist es in unserem Fachbereich noch immer nicht Usus, einzugestehen, dass Besonderheiten einer Sammlung, eines Archivs, der ArchivarInnen, unseres Hintergrundwissens und andere kontextbezogene Faktoren immer (auch) die Auswahl qualitativer Bewertungskriterien bedingen. Hier habe ich über die Jahre gelernt, zwischen den folgenden Kriterien, die (m) eine qualitative analoge Bewertung beeinflussen könnten, zu unterscheiden: persönliche Faktoren und professionelle Bewertungskriterien, jene, die ich schon vor, und jene, die ich während und auf Grund der Bewertung entwickle und hinzunehme.

Für meine erste digitale qualitative Bewertung waren folgende persönliche, inhaltliche und visuell-spezifische Kriterien ausschlaggebend:

Persönlicher Hintergrund (ein Beispiel)

- Holocaust-Restitutionshistorikerin und Archivarin
- interdisziplinäre akademische Ausbildung
- einige schwerstbehinderte Freunde
- jahrelanges Mitglied des ursprünglichen Wiener Blindenchors
- viele Familienmitglieder im Lehrberuf
- Interesse/Hintergrund bezüglich praktischer Fähigkeiten und Handwerk
- reguläre (nicht digitale) Archivarin
- MPLP Zugang zu analogen Sammlungen

Inhaltliche Kriterien (ein Beispiel)

- geschichtliche Dokumentation (auch langfristige Konsequenzen des Nationalsozialismus)
- Gruppen: z.B. Studenten, Eltern, Lehrer, Köche
- Infrastrukturen: z.B. Internat, Labors, Klassenzimmer, Garten
- Schulprogramm und dessen Vielfältigkeit: z.B. Gebärdensprache, handwerkliches Lernen, Exkursionen
- Teilabschnitte einer Aktivität: z.B. E-Bike und Wildbienenhotel bauen
- zeitgenössische Themen: z.B. Theater gegen Mobbing, Kochen und Diät
- Schuljahr und Höhepunkte
- Vielfalt innerhalb einer Veranstaltung: z.B. Spiele auf der Landschulwoche, Weihnachtsmarkthandwerk
- Aufführung, nicht aber die Proben: z.B. Theateraufführung, Jahresabschlussfeier
- bemerkenswerte Gendergleichheit bei Aktivitäten
- Neuheiten: z.B. Hörgerätemoden, Interessen/Stärken von Hörgeschädigten, frühe Integrationsbeispiele (z.B. Skifahren)
- pädagogisches Training und Lehrhilfsmittel

Visuell-spezifische Kriterien (ein Beispiel)

(1) Art/Inhalt der Abbildung: z.B. Profil/Hinterkopffotografien statt Frontalfotografien, um getragene Hörgeräte zu dokumentieren

(2) Übersicht: z.B. zusammenfassendes Weihnachtsfestvideo mit Text, nicht hingegen kurze Ausschnittclips

(3) Qualität der Aufnahme/Format, ästhetische Qualität aber auch bezüglich zukünftiger Reproduzierbarkeit: z.B. Teamfoto, nicht aber Fußballvideoclip

3.3 Workflow

Ich fand sogleich heraus, dass es zur Bearbeitung der Dateien am effizientesten war, *TSP und* die Icon- und Preview-Funktionen des Windows Explorer zu verwenden. Mit TSP Segmentanalyse konnte ich z.B. rasch nach Größe oder Format priorisieren. TSP war auch sehr stark im Öffnen und Abspielen der Video- und Audiodateien. Hingegen erlaubte die Icon- und Preview-Funktion (wie sie jedes Betriebssystem hat) die rasche Durch- und Übersicht sowie Bewertung der Fotos innerhalb eines Ordners. Im Gegensatz zur analogen Bewertung bekam ich so innerhalb weniger Minuten eine Übersicht sowie detaillierte Analysen und konnte fundiert entscheiden, ob und was bewertet werden sollte.

3.4. Bewertungsbeispiele

Die folgenden Beispiele zeigen, wie ein ganzes Schuljahr und dann Einzelfälle bewertet oder nicht bewertet wurden. Innerhalb der priorisierten Schuljahre, 2009 bis 2015, ging ich chronologisch vor, damit mir durch die ursprüngliche Bilderchronologie sammlungsspezifische Bewertungskriterien bewusst wurden. Nur so konnte ich z.B. herausfinden, ob die bemerkenswerte Gendergleichheit oder Videos über Mobbing oder Pantomimentanz einzigartig und neu waren oder schon Routine darstellten und daher vielleicht über die Jahre hinweg nicht mehr dokumentiert werden mussten.

3.4.1. Bewertung eines FVA Schuljahres, 2012–13

Sehen wir uns den größten Ordner (60 GB) an, „Schuljahr (SJ) 2012–13“, in dem Videos 80 % des Speicherplatzes und Fotos 82 % der Dateien ausmachten (Abb. 4). Der Unterordner „Weihnachten“, einer von mehreren über die Weihnachtsfeier, ist hier beispielhaft. Von seinen 50 Dateien waren 18 Video Clips ohne Text oder Kontext, von denen ich alle löschte außer einem, der etwas Neues zeigte: einen phosphoreszierten Pantomimentanz im Dunkeln, für den Gehörgeschädigte natürlich genauso geeignet sind wie Nicht-Gehörgeschädigte (Abb. 5).

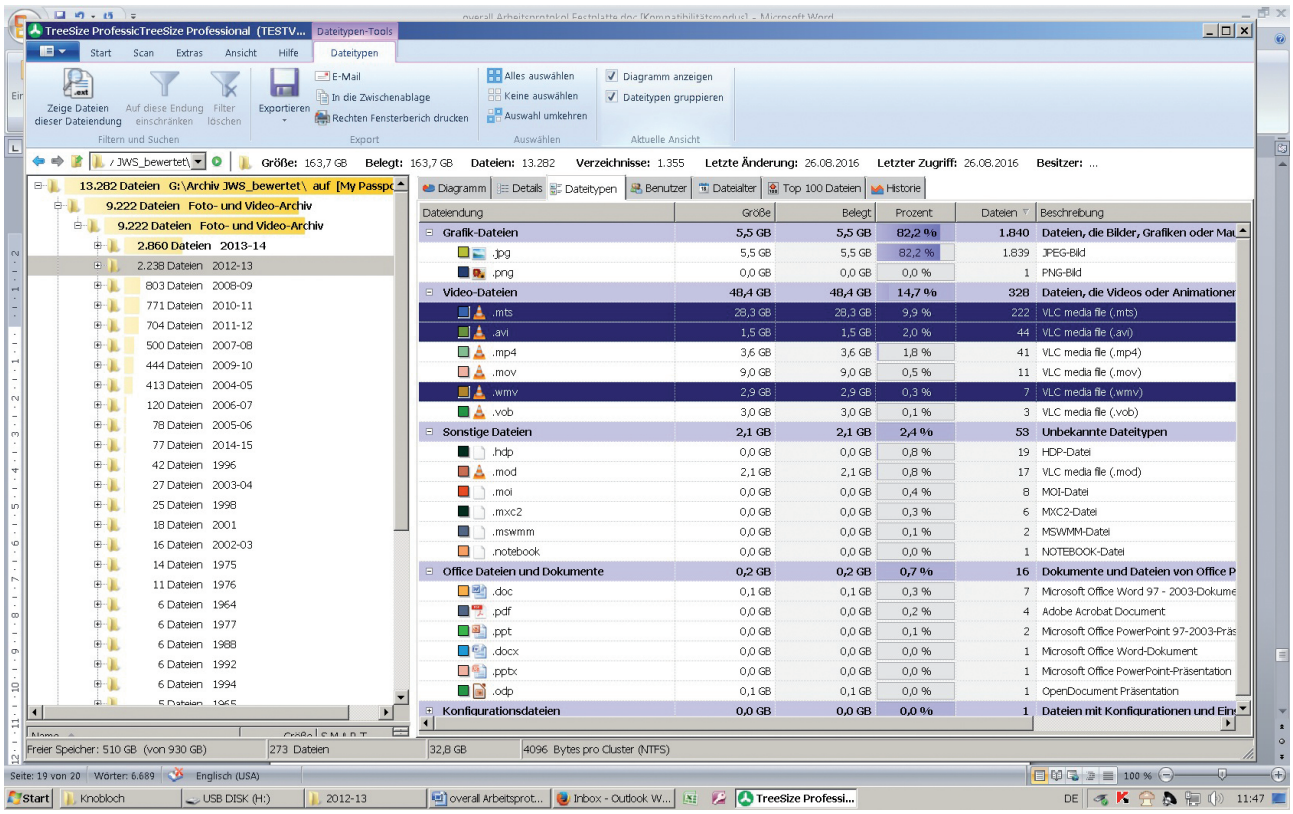


Abb. 4: Vor der Bewertung: Dateiformate für Schuljahr 2012–13

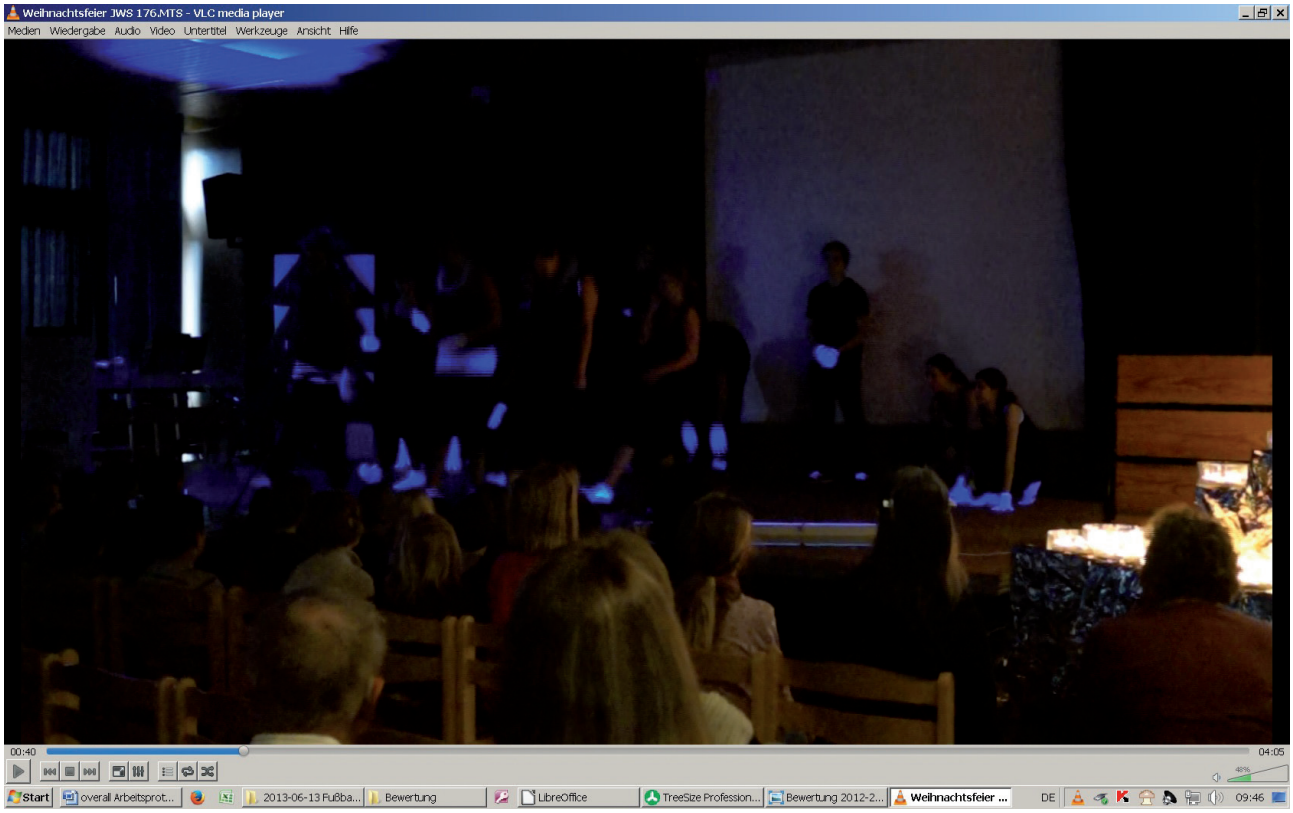


Abb. 5: Weihnachtsfeier, Schuljahr 2012–13: Pantomimentanz

Wie in den Vorjahren enthielt der Unterordner „Schuljahr“ ein zusammenfassendes Video der genannten Weihnachtsfeier inklusive Schilderung und daher Kontext, das ich aufbewahrte. Ich übernahm *alle* Videos, in denen Kinder die Gebärdensprache lernten, selbst die mit schlechter Videoqualität, da Gebärdensprache zentral für die Schule und bis dahin nicht durch Fotos dokumentiert worden war. Von den 22 Videoclips der Kochgruppe behielt ich eines, welches Kochen und Schülerinterviews enthielt. Insgesamt übernahm ich ein Viertel der ursprünglich 135 Ordnerdateien.

Der Unterordner „Fußballturnier“ beinhaltete 159 Dateien, von denen 158 Fotos und Videoclips von zumeist geringer Qualität waren (Abb. 6). Letztere löschte ich – warum unzusammenhängende kurze Fußballvideos archivieren? – und übernahm nur einige Fotos mit hoher Auflösung, die das Fußballspielen dokumentieren und in der Zukunft gut reproduziert werden können. Bewertungskriterien hier waren z.B. die Dokumentation von (a) Mädchen und Buben, die gemeinsam spielten, (b) getragene Hörgeräte, (c) Trophäen, (d) Teams, (e) Spielstatistiken, (f) Zuschauer und Trainer (Abb. 7).

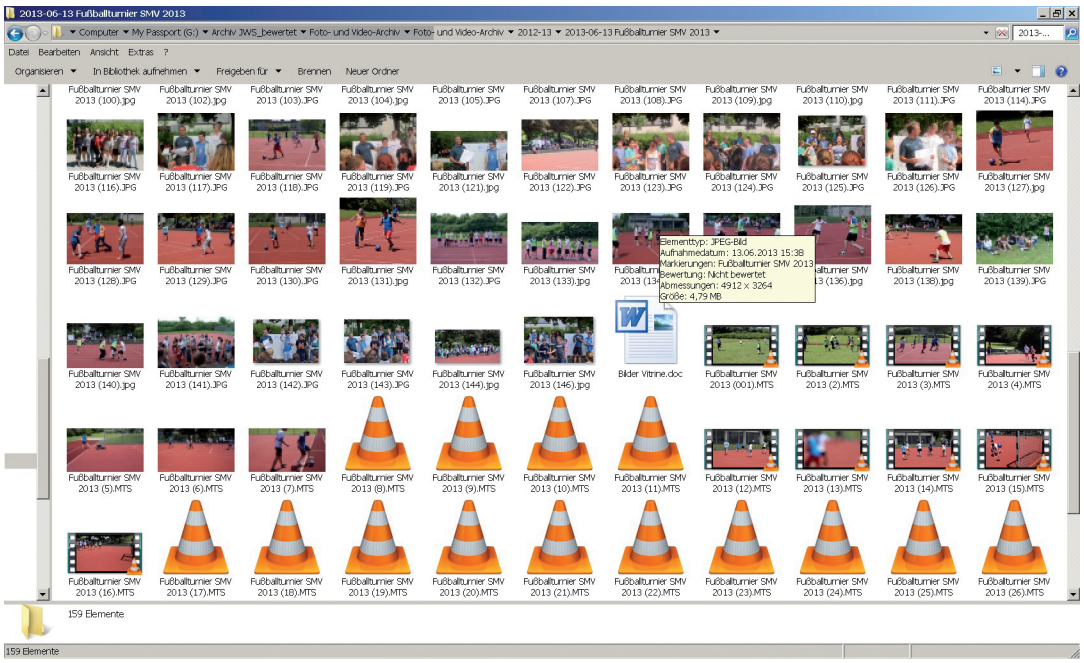


Abb. 6: Vor der Bewertung: Ordner Fußballturnier Schülermitverwaltung (SMV) 2013 (Icon View)

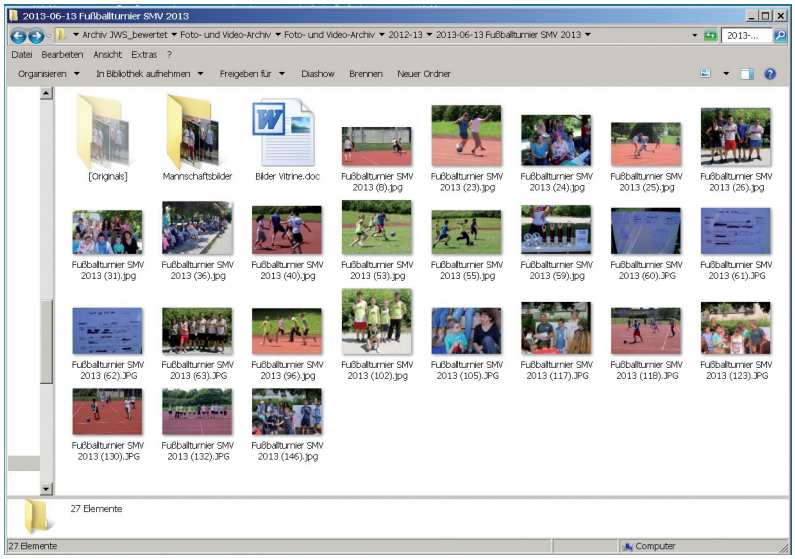


Abb. 7: Nach der Bewertung: Ordner Fußballturnier SMV 2013 (Icon View), siehe auch abgebildete Hörgeräte

Es gab keinen Grund, die jährlichen, gleichartigen und langen Ostermessenvideoclips aufzuheben. Ein paar Fotos waren genug, um zu dokumentieren, dass die Messe stattfand. Selbstverständlich übernahm ich mehr Dateien, wenn sie Schulgeschichte dokumentierten, z.B. Lehrerabschiedsfeiern, ein Treffen ehemaliger Schüler oder einzelne Klassenfotos. Innerhalb von zweieinhalb Stunden konnte ich so zirka 90 % der Videos und 60 % der Fotos löschen und hatte schlussendlich ein Sechstel der ursprünglichen Speichergröße des Schuljahres.

3.4.2. FVA Einzelbewertungsbeispiele – Schuljahr 2010–11

Der Ordner für das „Schuljahr 2010–11“ (12.8 GB) hatte die meisten Dateien, 98 % Fotos, von denen fast die Hälfte in einem Unterordner die Schülerreise nach London dokumentierten. Mehrere SchülerInnen hatten die Fotos aufgenommen (fast 90 % von einer Schülerin), die Touristenattraktionen und SchülerInnen zeigten (Abb. 8). *Unsere* Aufgabe ist es aber, die Schule zu dokumentieren, die die Reise organisierte und nicht London oder persönliche Erlebnisse der Jugendlichen. Deswegen übernahm ich jeweils nur ein Foto, das, soweit ich es beurteilen konnte, alle Jugendlichen und Lehrer zeigte, schaute mir kurz ein paar Videos an, um meinen ersten auf deren Titel beruhenden Eindruck zu überprüfen (z.B. *Changing of the Guard* 1–6 oder *Big Ben*), und löschte dann alle Videos. Innerhalb von zehn Minuten hatte der Unterordner nur noch 0,2 % seiner ursprünglichen Größe (5 GB).

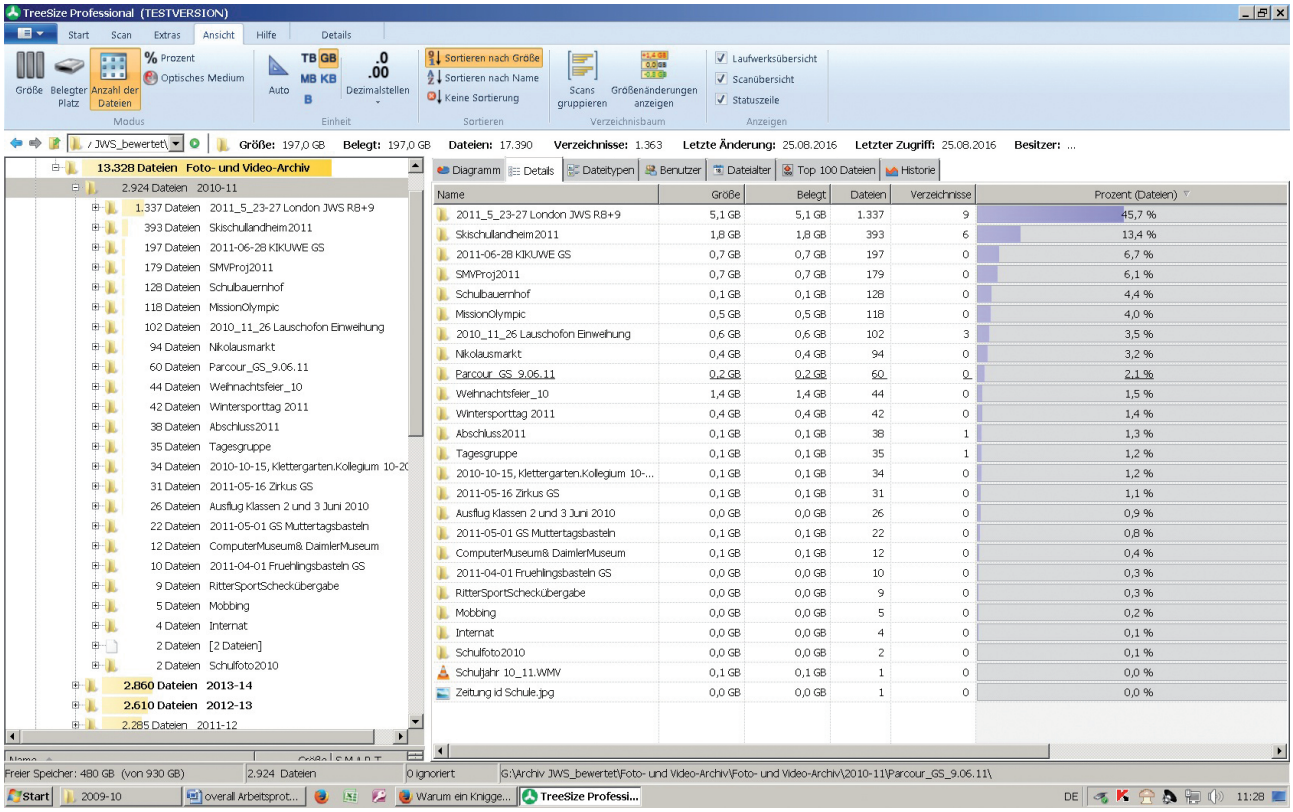


Abb. 8: Vor der Bewertung: 2011 Londonreise JWS R8+9

Ebenso ging ich beim zweitgrößten Unterordner „Skischullandheim“ vor, für den ich nur eine zusammenfassende PowerPoint-Präsentation und einige wenige Fotos der Aktivitäten behielt. Videos, in denen SchülerInnen herumblödelten oder eine Katze streichelten, wurden gelöscht. Für den drittgrößten Unterordner „Weihnachtsfeier“ behielt ich das erste derartige Zusammenschnittvideo der jährlichen Weihnachtsfeier, da es auch Text beinhaltete. Wegen der schlechten Videoqualität übernahm ich zusätzlich mehrere Fotos, auch weil deren Bewertung größen- und formatsmäßig (0,1 GB) keine Priorität war.

Im Gegensatz dazu hob ich alle Videos des Projektordners „Mobbing“ auf. Mobbing hatte damals eine neue hohe Priorität in Schulen und Kinder zeigten schauspielerisch in den Videos, was bei Mobbing passiert und wie sie reagieren sollten. Eine Bewertung für kleinere Unterordner zahlte sich nach unserer Kosten-Nutzenrechnung (z.B. Zeit versus ersparter Speicherplatz) nicht aus.

3.4.3. FVA Einzelbewertungsbeispiele – Schuljahr 2009 und 2013–14

Anbei noch ein paar Beispiele, um die inhaltsbezogene Bewertung zu verdeutlichen. Ich übernahm z.B. alle Mitarbeiterfotos des Ordners „Mitarbeiter 2009“ als essentielle Schuldokumentation (Abb. 9).

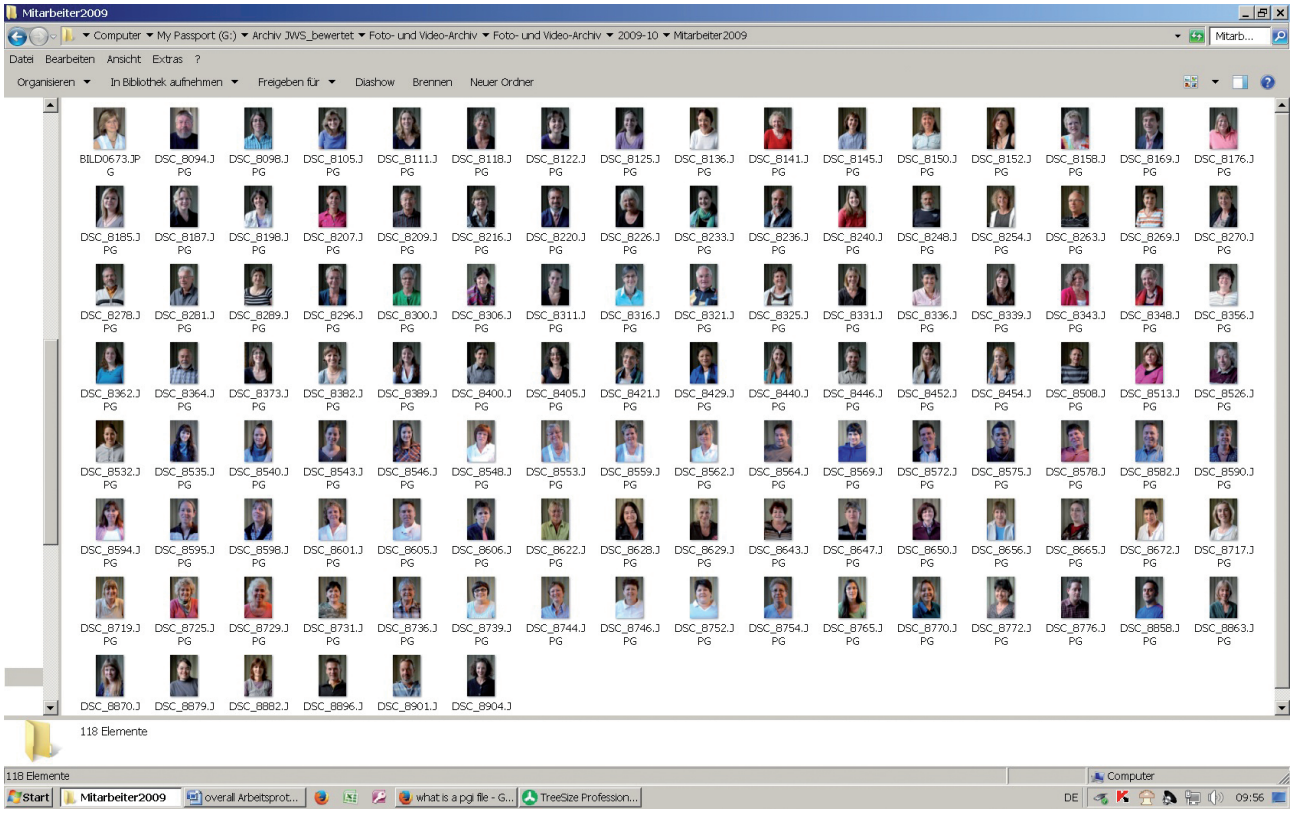


Abb. 9: Bewertungsbeispiel, in dem nichts gelöscht wurde: Mitarbeiter 2009 (Icon View)

Wie ich durch meine Bewertung lernte, waren praxisorientierte Projekte eine weitere wirkliche Stärke der Schule, z.B. das Bauen eines elektrischen Miniatur-Motorrads oder eines Wildbienenhotels 2013–14 und die Vielfalt an Handwerk, Sport und Spielen bei Schulfesten und Reisen, bei denen z.B. Spiele für Gehörgeschädigte auch verschiedenartig sein können. Für solche praxisorientierten Programme versuchte ich, alle Schritte, Aktivitäten und deren vielfältige Produkte zu dokumentieren – und diese Art von Bewertung brauchte Zeit, war mir aber wichtig und den Aufwand wert. Weiters übernahm ich für das Schuljahr 2013–14 z.B. alle Bildungsvideos für Lehrer, da sie Lehrmitteländerungen dokumentierten, löschte jedoch fast alle der 359 Videoclips des „5km-Laufes," die Teilausschnitte laufender Kinder zeigten, und behielt stattdessen einige Lauffotos.

3.4.4. Bewerten oder Nichtbewerten der weiteren Hauptverzeichnisse

Ich sah mir dann die zwei weiteren Hauptverzeichnisse, „Schulportfolio" und „Themen," an und stoppte meine Bewertung nach einer Stunde. Ich hatte z.B. 100 der größten redundanten Schulportfolio-Textdateien gelöscht, es brachte mir aber fast keine Speicherplatzverringerung, und die 410 (!) Fotos in „Themen" waren nach kurzer Durchsicht essentiell.

4. Resultat der Bewertung

Innerhalb von vier Tagen reduzierte ich die Sammlung auf 74,2 GB (-89 %), 1314 Ordner (-64 %), und 8467 Dateien (-87 %) und führte eine qualitative Bewertung von zirka 15.000 Foto- und Videodateien durch (Abb. 10 und 11). Ich entwickelte einen einfachen Workflow, der ein unkompliziertes praktisches Programm und die Icon- und Preview-Funktion des Betriebssystems verwendet. Die Verdichtung der AV- und Foto-Dateien und die kleine Sammlung an Office-Dateien schaffen nun verringerten Erhaltungsaufwand für das Archiv (z.B. geringere Software- und Migrationsprobleme gelöschter komplexer File Formate und geringere Kosten) sowie geringe Suchanforderungen für zukünftige Benutzer, die Office-Dateien ja mit Hilfe normaler Suchfunktionen leicht durchforsten können.

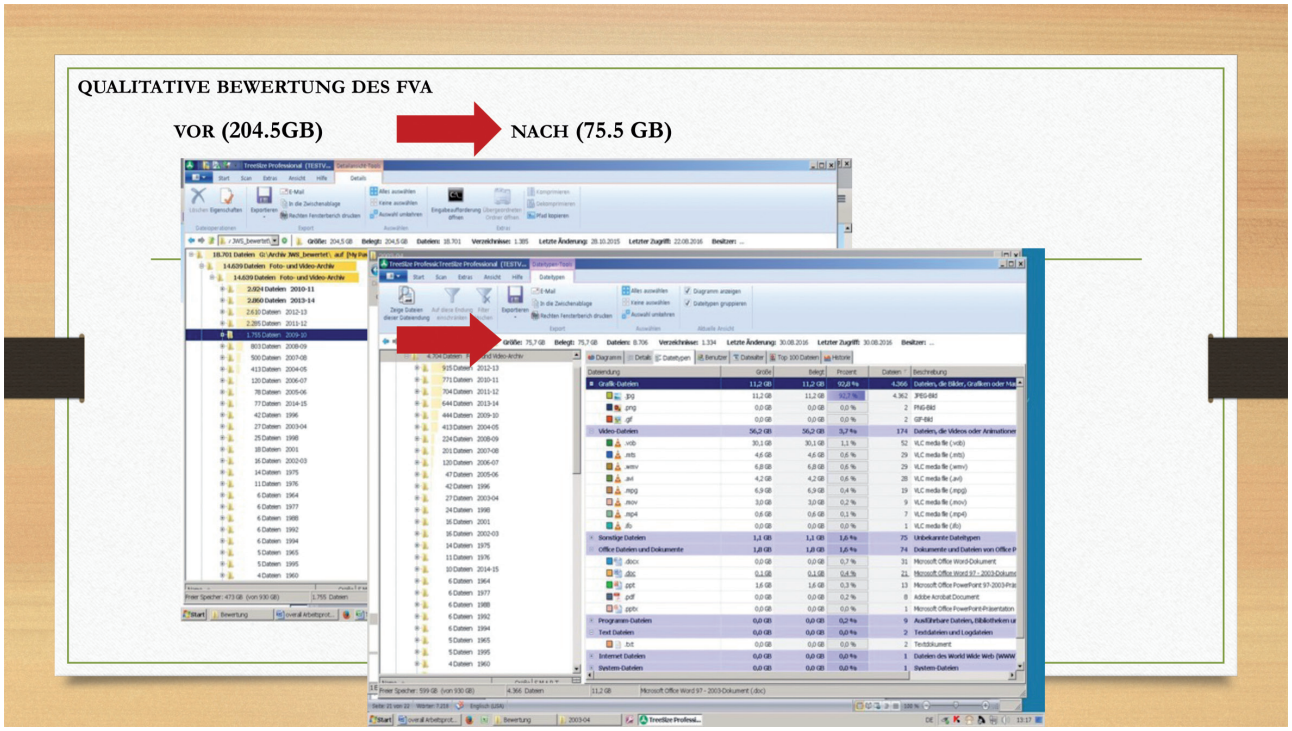


Abb. 10: Resultat der Qualitativen Bewertung des FVA

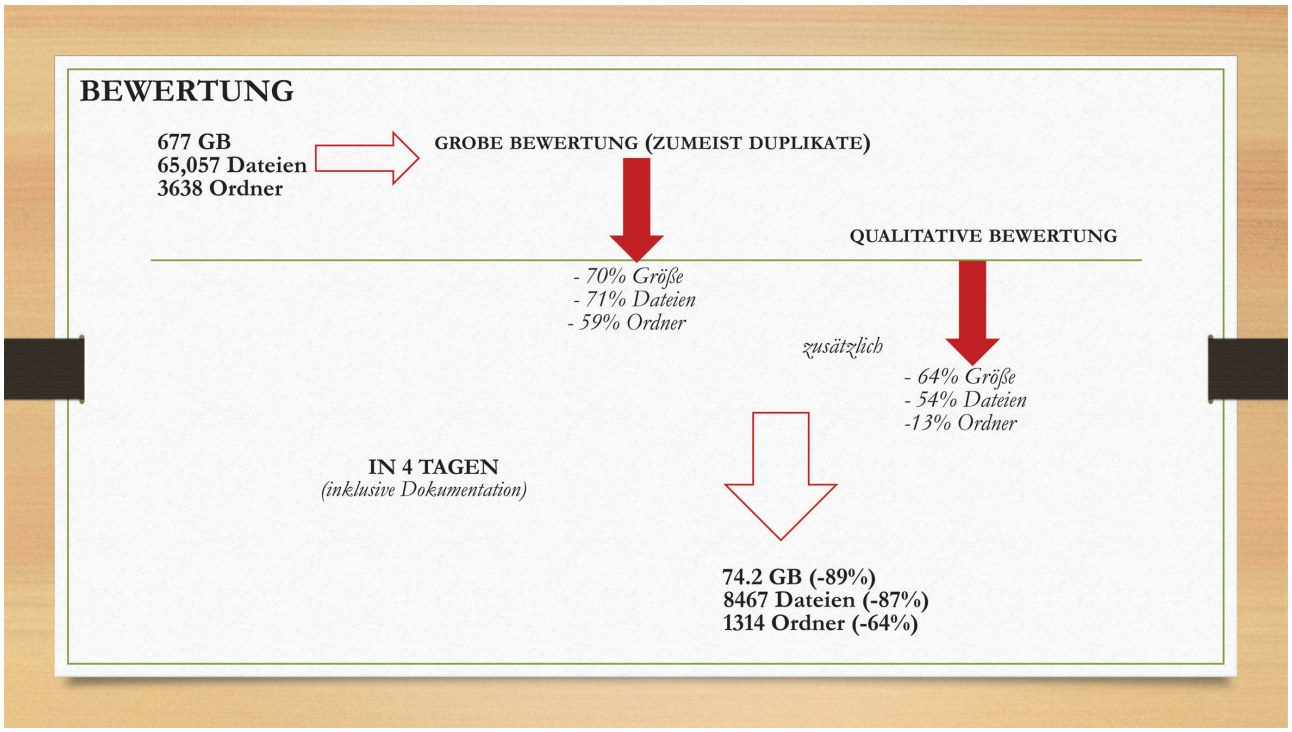


Abb. 11: Gesamtresultat der Groben und der Qualitativen Bewertung

Wie es mit der Projektsammlung von Susanne Belovari weiterging

Kai Naumann

Dieser Text ist kein eigenständiger Beitrag, sondern nur ein kurzer Bericht, wie die von Susanne Belovari bearbeitete Sammlung (vgl. S. 17–29) letztlich zu einem ordentlich verzeichneten und magazinierten Unterlagenbestand wurde.

Die bearbeiteten Dateien der Johannes-Wagner-Schule sind nun als Bestand FL 240/4 I beim Staatsarchiv Ludwigsburg katalogisiert.[1] Das Staatsarchiv legt Wert auf eine parallele Findbarkeit digitaler und analoger Unterlagen und hat daher beide Unterlagenarten in einem Findmittel zusammengeführt. Der digitale Teil wurde mit dem DIMAG IngestTool in Archivinformationspakete (AIP) umgewandelt und in das Digitale Archivsystem DIMAG überführt. Einzelne Dateitypen[2] wurden wegen des allzu großen Aufwands in der Bestandserhaltung für den Ingest ausgefiltert. Der Dateityp MS Access (MDB) wurde ebenfalls nicht übernommen, da Datenbankdateien auffielen, die Patientendaten der Schüler enthielten. Solche Dateien wurden datenschutzgerecht vernichtet. Der Dateityp ZIP wurde ebenfalls nicht übernommen, da ein Entpacken aufwändig geworden wäre und die Dateisammlung bereits ohne solches Entpacken den gewünschten Informationswert besitzt.

Die Dateisammlung gehorchte zwei unterschiedlichen Strukturprinzipien:

a) Videos und Fotos waren in chronologischen Ordnern (Kalender- oder Schuljahre) organisiert.

b) Das Schulportfolio war in Sachordnern strukturiert, die ihrerseits in beliebiger Tiefe Unterordner aufweisen.

Im Folgenden werden die Arbeitsgänge, die sich anschlossen, teils getrennt nach a) und b) geschildert.

1. AIP-Fokus setzen

Für den Bereich a) ließ sich der Fokus für die AIP-Erzeugung automatisch aufgrund der Struktur setzen, und zwar wurde mit der 2. Ebene diejenige fokussiert, die die einzelnen Ereignisse wiederspiegelte. Für den Bereich b) wurde die 3. Ebene fokussiert, aber es waren Abweichungen erforderlich, die durch manuelles Umstellen einiger AIPs in der Ebenenstruktur von Ebene 2 auf die einheitliche Ebene 3 erfolgten.

2. AIPs für den Ingest in das Digitale Archivsystem erzeugen

Die AIPs wurden mit dem DIMAG-Ingesttool erzeugt. Als Eingangsmetadaten dienten die Ordnernamen. Durch Lookup-Tabellen wurde jedem AIP-Ordnernamen eine eindeutige Bestellsignatur[3] im Bestand FL 240/4 I zugewiesen. Die inneren Strukturinformationen der

[1] Link zum Bereich Foto- und Videosammlung: *http://www.landesarchiv-bw.de/plink/?f=2-5529298.* Link zum Bereich Schulportfolio: *http://www.landesarchiv-bw.de/plink/?f=2-5529787*.

[2] BAT, DAT, DLL, HDP, INF, PGI, CDR, DB, EMM, ERG, GID, ICO, INI, JS, LNG, MMAP, MXC2, NOTEBOOK, PLV, PUB, SAM, SE, SHS, SYS, TTR, UNT, URL, URL, WHI, WMLP, WRI.

[3] Dies könnte mit dem neuen DIMAG IngestTool 2.0 (Mai 2017) auch durch eine eingebaute Funktion erledigt werden.

Ordner werden, falls vorhanden, in DIMAG in das Feld „relativer Dateipfad“ transferiert. Sie können auf diesem Weg später wieder betrachtbar gemacht werden. Wie die Pfadangabe die verschiedenen Erschließungsfelder bediente, ist aus Abb. 1 und 2 ersichtlich. Für die Datierung von Teil b) bestanden keine Möglichkeiten, das Datum der jüngsten und der ältesten Datei im AIP automatisiert zu ermitteln. Daher wurde die Laufzeit der gesamten Sammlung (2002–2014) für jedes einzelne AIP gesetzt. Die Datierung laut Metadaten der abgebeben Dateien ist aber in DIMAG enthalten (Feld „Letzte Änderung (abg. Stelle)“) und kann zu gegebener Zeit auf der Ebene des AIP erhoben und in Scope angeführt werden, wenn Bedarf besteht.

Das Feld „Referenzen extern“ wurde nicht in DIMAG importiert, sondern floss in ein Scope-Sortierfeld ein, um die Zuordnung von Bestelleinheiten auf entsprechende Strukturobjekte zu erleichtern.

3. Erschließungsmetadaten für den Ingest in das Archivinformationssystem erzeugen

Die Erschließungsmetadaten für die AIP-Ebene wurden mit ScopeArchiv Übernahmeassistent in ScopeArchiv überführt. Die Metadaten für die Ordungsebene darüber wurden später manuell erzeugt. Hierfür wurden in Scope-Archiv Sortierfelder angelegt, die eine Sortierung der Einheiten in der Reihenfolge der übergeordneten Strukturobjekte erlaubten. Nachdem alles erledigt war, wurde auch in der Publikumspresse des Landesarchivs eine Meldung lanciert.[4]

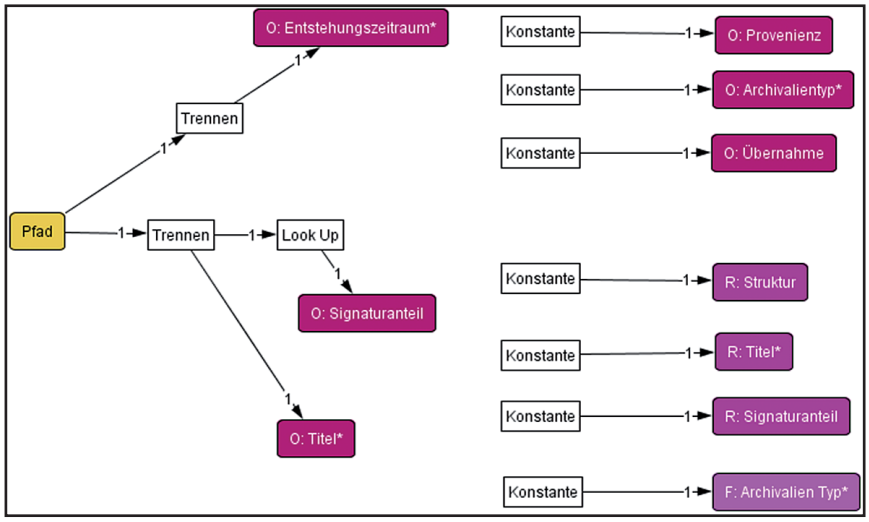


Abb. 1: Mapping für den Ingest von Teil a) Foto- und Videosammlung

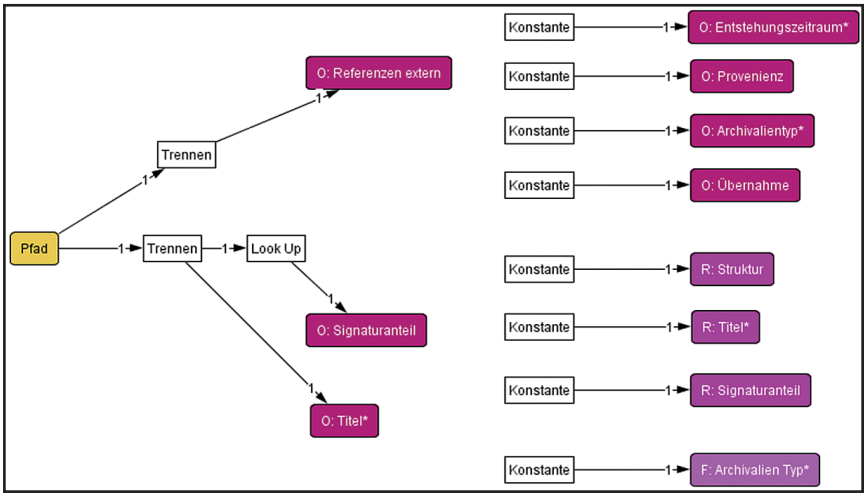


Abb. 2: Mapping für den Ingest von Teil b) Schulportfolio

[4] Archivnachrichten Nr. 54 (März 2017), *https://www.landesarchiv-bw.de/web/61800,* im PDF auf S. 44.

Fileablagen im Gewand von E-Akten: Was ein DMS mit einer Dateisammlung gemeinsam hat

Niklas Konzen

1. Einleitung

Der vorliegende Beitrag stellt einige Beobachtungen vor, die aus der Erstellung meiner Transferarbeit für die archivarische Staatsprüfung Anfang 2016 erwachsen sind.[1] Die Arbeit verfolgte die Fragestellung, in wie weit sich Aussonderung, Bewertung und Archivierung von E-Akten umsetzen lassen, wenn die elektronische Schriftgutverwaltung (SVG) der abgebenden Stelle von DOMEA/EVA-Empfehlungen[2] abweicht – was nach wie vor in den meisten kommunalen und staatlichen Verwaltungen der Fall sein dürfte. Diese Fragestellung wurde anhand von elektronischen Akten der Stadtverwaltung Kirchheim unter Teck und des Landratsamts Karlsruhe untersucht. Mit ausschlaggebend für die Auswahl dieser beiden Verwaltungen war, dass beide zum Zeitpunkt der Untersuchung das DMS Regisafe public verwendeten, also eine im kommunalen Bereich sehr verbreitete Software. Außerdem strebten beide eine Nutzung der Fachanwendung DIMAG für die digitale Archivierung ihrer elektronischen Unterlagen an, eines vom Landesarchiv Baden-Württemberg entwickelten Systems, das im Rahmen einer Kooperationslösung auch von baden-württembergischen Kommunalarchiven genutzt werden kann.[3] Es stand also zu erwarten, dass die Erfahrungen der Untersuchung auch für andere Kommunen relevant sein würden.

Im Rahmen der Untersuchung wurde die jeweilige Praxis der elektronischen Schriftgutverwaltung in den Beispielfällen mit den Vorgaben von DOMEA / EVA verglichen und auf Abweichungen untersucht, dann die möglichen Vorgehensweisen für den Export je einer Test-E-Akte aus Regisafe, ihre Aufbereitung, Bewertung und Übernahme in DIMAG praktisch erprobt und dokumentiert. Bei den ausgewählten Test-E-Akten handelte es sich nicht um Dummy-Akten, sondern um Unterlagen, die im laufenden Geschäftsbetrieb der jeweiligen Verwaltung entstanden waren.

Ein Ergebnis des Vergleichs zwischen Empfehlungen und Praxis der elektronischen SGV in den Beispielfällen bestand in der Feststellung, dass in beiden DMS Ablagestrukturen existierten, die in mehrfacher Hinsicht eher Dateisammlungen bzw. Fileablagen ähnelten als einer EVA-konformen E-Akte. Dieser Aspekt ist Gegenstand der folgenden Darstellung. Zunächst werden kurz die Ausgangsbedingungen der elektronischen Schriftgutverwaltung in den Fallbeispielen skizziert, dann ein paar Beispiele für kreative Fileablagen im DMS vorgestellt.

[1] Niklas Konzen, Übernahme von E-Akten aus kommunalen Dokumentenmanagementsystemen in das Langzeitarchiv DIMAG: Ein Vorschlag zur praktischen Umsetzung anhand von Fallbeispielen aus den DMS der Stadt Kirchheim unter Teck und des Landratsamts Karlsruhe, (Transferarbeit) Stuttgart 2016. https://www.landesarchiv-bw.de/sixcms/media.php/120/60857/Transferarbeit2016_Konzen.pdf (aufgerufen am 7.4.2017).

[2] Das Organisationskonzept Elektronische Verwaltungsarbeit (EVA) des Bundesinnenministeriums unterstützt die deutsche öffentliche Verwaltung bei der Auswahl und beim Betrieb von Anwendungen zur digitalen Schriftgutverwaltung und Prozessunterstützung, http://www.verwaltung-innovativ.de/DE/E_Government/orgkonzept_everwaltung/orgkonzept_everwaltung_artikel.html. Das Vorgängerkonzept DOMEA wird vom Ministerium für eine Übergangszeit weiterhin zur Verfügung gestellt.

[3] Miriam Eberlein – Christian Keitel – Manfred Waßner, „DIMAG“ wird kommunal: Ein digitales Langzeitarchiv für Städte und Gemeinden in Baden-Württemberg, in: Digitale Archivierung. Innovationen – Strategien – Netzwerke. Tagungsband zur 19. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ (Mitteilungen des Österreichischen Staatsarchivs 59/2016), Wien 2016, S. 21–31. Folien unter *http://www.staatsarchiv.sg.ch/home/auds/19.html.*

2. Ausgangsbedingungen für die Entstehung der Ablagestrukturen in den beiden Fallbeispielen

Beide untersuchten Verwaltungen nutzten zur Zeit der Untersuchung das DMS Regisafe public, ein Produkt der Hans Held GmbH. Im Landratsamt Karlsruhe war Regisafe seit 2004 im Einsatz, in Kirchheim unter Teck seit 2008. Kirchheim hat allerdings inzwischen das Vertragsverhältnis beendet.[4]

In beiden Verwaltungen gab es bis zum Zeitpunkt der Untersuchung kaum normative Vorgaben für die elektronische Schriftgutverwaltung. Die Aktenordnung des Landratsamts Karlsruhe stammte aus dem Jahr 1994 und war noch nicht für die elektronische Aktenführung angepasst worden. In Kirchheim wurde die Aktenordnung zwar bei DMS-Einführung aktualisiert und durch eine Dienstvereinbarung EDV weiter ausgeführt und erläutert, viele Prozesse der elektronischen Schriftgutverwaltung, insbesondere Aussonderung und Übernahme, jedoch nur unzureichend oder gar nicht reguliert. Festgeschrieben wurde jedoch in beiden Fällen die Verwendung des Boorberg-Aktenplans als Grundlage für die Gliederung des Aktenbestandes, der auch jeweils im DMS hinterlegt ist. Aktenzeichen werden − abweichend von EVA − in beiden Verwaltungen mit Sachbetreffen des Aktenplans gleichgesetzt, eine Praxis, die in vielen Kommunalverwaltungen verbreitet ist.[5]

So wenig Regulierung es seitens der Verwaltungsleitung gab, so viel Freiheit ermöglichte auch das verwendete DMS, jedenfalls in der Systemkonfiguration, die im Rahmen dieser Untersuchung vorgefunden wurde. In Regisafe entspricht eine Akte bzw. ein Aktenzeichen der kleinsten Aktenplaneinheit. Die Bearbeitung von Unterlagen wird daher in der Regel nicht auf Akten-, sondern allenfalls auf Vorgangs- oder Dokumentebene abgeschlossen. Unterhalb des Aktenzeichens werden „Teilakten / Vorgänge“ (im folgenden: „Vorgänge“)[6] gebildet, die entweder andere Vorgänge oder „Schriftstücke“ (im folgenden: „Dokumente“) enthalten können. Jedem Dokument sind mindestens eine, häufig auch mehrere Dateien zugeordnet, z.B. eine E-Mail mit Anhängen. Es wird keine dreistufige Objekthierarchie Akte-Vorgang-Dokument vorgegeben, sondern das System erlaubt unterhalb des Aktenzeichens bis zu vier Vorgangsebenen.[7]

Einerseits gab es also in beiden Fallbeispielen keine normativen Vorgaben darüber, wie eine elektronische Akte strukturiert sein soll und im DMS kaum funktionale Beschränkungen für die Gestaltung der elektronischen Ablage, andererseits waren Aktenbildung, Aktenzeichenvergabe und Registrierung in beiden Verwaltungen nicht zentralisiert, sondern in der Regel Aufgaben des einzelnen Sachbearbeiters. Diese Kombination von Faktoren begünstigte die Entstehung eines wuchernden Dschungels von Ablagestrukturen, die eben mehr mit individuellen, kreativen Fileablagen gemein haben, als mit dem DOMEA-Ideal einer E-Akte.

3. „Fileablagen“ im DMS

Auf den ersten Blick unterscheidet sich die im DMS vorgefundene Ablagestruktur insofern von einer Fileablage, als der Bestand an elektronischen Unterlagen durch einen Aktenplan gegliedert wird. Außerdem erfüllen die DMS in den Fallbeispielen ansatzweise die EVA-Anforderung, dass Akte, Vorgang und Dokument in der elektronischen Aktenablage jeweils

[4] Vgl. Konzen (wie Anm. 1) S. 6.
[5] Konzen (wie Anm. 1) S. 6–7.
6 Im Interesse von eindeutiger und ökonomischer Ausdrucksweise wurden hier Regisafe-spezifische Begriffe an die in DOMEA benutzte Terminologie angeglichen.
[7] Konzen (wie Anm. 1) S. 10.

Containerobjekte mit eigenen Metadatensätzen sein sollen, mit der Einschränkung, dass diese Metadatensätze nach EVA-Kriterien unvollständig sind.[8] Hinsichtlich der EVA-Anforderung, dass jedes Dokument einem Vorgang und jeder Vorgang einer Akte zugeordnet sein muss,[9] zeigen sich bei genauerem Hinschauen gewisse Schwachstellen: In beiden Fallbeispielen hatten einzelne Sachbearbeiter Vorgänge oder Dokumente mitten im Aktenplan angelegt.[10] Allerdings handelt es sich dabei um Einzelfälle.

Schwerer dagegen wiegen einige Auffälligkeiten der Ablagestrukturen unterhalb der Aktenebene. Unter manchen Aktenzeichen verbergen sich „Bearbeiternester", die nicht wesentlich anders aussehen als eine Explorer-Fileablage. Ein idealtypisches Beispiel:

[**Aktenzeichen**]

└**Vorgang**: „Arbeitsbereich N. N."

└**Dokumente**: Urlaubs- und Dienstreiseanträge, Abrechnungen, Posteingänge, -ausgänge, weitere Unterlagen zum Aufgabenbereich von N. N.

└**Vorgang**: „Arbeitsbereich N. N./2012"

└**Dokumente**: Wie oben, jedoch ausschließlich aus 2012

└**Vorgang**: „Arbeitsbereich N. N./2012/Aufgabenbereich XY"

└**Dokumente**: Arbeitsunterlagen zum Aufgabenbereich XY

└**Vorgang**: „Arbeitsbereich N. N./2013"

(...)

Zwar ist diese Struktur einer Aktenplanposition zugeordnet, hinter dieser Position verbirgt sich allerdings faktisch der persönliche Ablagebereich eines einzelnen Mitarbeiters − ein Ordnungsprinzip, nach dem auch viele Fileablagen strukturiert sind. Die weitere Untergliederung dieses persönlichen Ablagebereichs folgt keinen klaren und eindeutigen Regeln, sondern es kommen parallel unterschiedliche Ordnungsmerkmale zum Einsatz (Themen, Chronologie o.ä.), die keiner systematischen Hierarchie folgen. Die Ähnlichkeit solcher „E-Akten" zu kreativen Fileablagen wird noch ausgeprägter, wenn es sich um besonders umfangreiche Strukturen handelt. Ein Beispiel dafür sind die Testdaten aus Kirchheim unter Teck: Dabei handelt sich um sämtliche Vorgänge, die im DMS der Stadt Kirchheim unter dem Aktenzeichen (bzw. der Aktenplanposition) „048.751 Bürokommunikation, Vorgangsbearbeitung, DMS" abgelegt waren. Diese Ablagestruktur wurde hier nach dem Export aus Regisafe mit Hilfe eines Kommandozeilenbefehls in Form von Explorer-Ordnern visualisiert (vgl. Abb. 1), da es in Regisafe nicht möglich war, sämtliche Gliederungsebenen eines Aktenzeichens gleichzeitig zu öffnen.

[8] Vgl. Konzen (wie Anm. 1) S. 10–13.

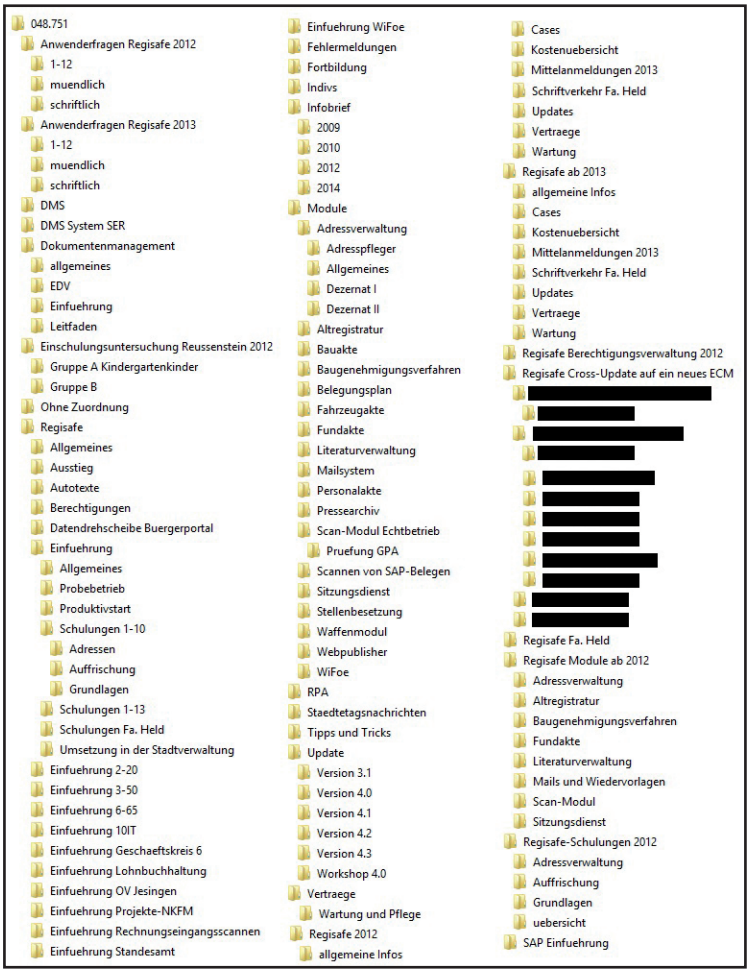
[9] Bundesministerium des Innern, Organisationskonzept Elektronische Verwaltungsarbeit – Baustein E-Akte, o.O. 2012, *http://www.verwaltung-innovativ.de/SharedDocs/Publikationen/Organisation/e_akte.pdf?__blob=publicationFile&v=2* (aufgerufen am 1.3.2016), S. 7 f., 9, 14.

[10] Konzen (wie Anm. 1) S. 10.

Tabelle 1: Umfang der Testdaten aus dem DMS der Stadt Kirchheim unter Teck (Aktenzeichen 04 8.751: Bürokommunikation, Vorgangsbearbeitung, Dokumentenmanagement DMS)

Anzahl der Dateien	2825
Anzahl der Dokumente („Schriftstücke“)	2318 (davon 46 ohne Zuordnung zu einem Vorgang)
Anzahl der Vorgänge („Teilakte/Vorgang“)	144 (davon 25, die anderen Vorgängen übergeordnet sind und selbst keine Dokumente enthalten)
Datenumfang	959 MB

Wie die Abbildung verdeutlicht, haben die Sachbearbeiter hier die Möglichkeit, bis zu vier Vorgangsebenen zu erstellen, voll ausgenutzt, was die Gesamtstruktur sehr unübersichtlich macht. Angesichts der Vorgangstitel stellt sich die Frage, ob es sich bei diesen Vorgängen wirklich um Vorgänge im Sinne einer Dokumentation von Prozessen handelt oder um sachthematische Dateisammlungen. Die Bilanz ist gemischt: Bei der näheren Untersuchung der jeweils einem Vorgang zugeordneten Dokumente stellte sich teilweise heraus, dass es sich nur um thematische Materialsammlungen handelte oder dass zwar ein Prozess abgebildet wurde, der jeweilige Prozess aber nur unvollständig dokumentiert war. Diese Unvollständigkeit war unter anderem bedingt durch die Fortexistenz von Papierunterlagen parallel zum DMS.[11] Zugleich wird angesichts der Vorgangstitel deutlich, dass hier wie beim Beispiel des „Bearbeiternestes“ unterschiedliche Ordnungsprinzipien auf unsystematische Weise verwendet wurden und dass es dabei zu inhaltlichen Überschneidungen kommen konnte.



Ordnerstruktur des SIP aus Kirchheim (Az. 048.751 Bürokommunikation, Vorgangsbearbeitung, DMS)

[11] Zum hybriden Charakter der Überlieferung in den beiden Fallbeispielen vgl. Konzen (wie Anm. 1) S. 8.

Tabelle 2: Inkonsistenzen der Ablagestruktur im Kirchheimer Beispiel (Az. 048.751)

Vorgang auf Ebene 1	Zahl der enthaltenen Dokumente (mit allen untergeordneten Verzeichnissen)
Anwenderfragen Regisafe 2012	45
DMS	11
Dokumentenmanagement	10
Einfuehrung REGISAFE	1
Ohne Zuordnung	45
Regisafe	1874
Regisafe 2012	46
Regisafe ab 2013	62
Regisafe Berechtigungsverwaltung 2012	12
Regisafe Cross-Update auf ein neues ECM	112
Regisafe Fa. Held	3
Regisafe Module ab 2012	66
Regisafe-Schulungen 2012	25

Tabelle 2 verdeutlicht, welche der 144 Vorgänge auf der ersten Ebene unterhalb der Aktenebene angelegt sind und wieviele Dokumente sie enthalten, wobei alle, d.h. bis zu drei, tiefer liegenden Vorgangsebenen mitgerechnet wurden. Bereits auf der ersten Gliederungsebene sind Vorgänge angelegt, von denen einige Teilmengen von Schriftstücken enthalten, die eigentlich von der Logik her anderen Einheiten zuzuordnen wären. Zum Beispiel müsste „Regisafe 2012“ dem Vorgang „Regisafe“ als Teilmenge untergeordnet sein, statt auf gleicher Ebene zu erscheinen; der Vorgang „Regisafe Berechtigungsverwaltung 2012“ müsste eigentlich Teilmenge von „Regisafe 2012“ sein etc.

Tabelle 3: Vorgangsdubletten im Kirchheimer Beispiel (Az. 048.751)

Kleinere Vorgänge auf Ebene 1–2	Untergeordnete Vorgänge in „Regisafe“
048.751/Regisafe Module ab 2012	048.751/Regisafe/Module
048.751/Regisafe Berechtigungsverwaltung 2012	048.751/Regisafe/Berechtigungsverwaltung
048.751/Einführung REGISAFE	048.751/Regisafe/Einführung
048.751/Regisafe Cross-Update auf ein neues ECM	048.751/Regisafe/Update
048.751/Regisafe Fa. Held	048.751/Regisafe/Allgemeines/Anfragen Fa. Held
048.751/Anwenderfragen Regisafe 2012	048.751/Regisafe/Allgemeines/Anwenderfragen
048.751/Regisafe-Schulungen 2012	048.751/Regisafe/Einführung/Schulungen (...)

Beim Vergleich von Vorgängen auf unterschiedlichen Ebenen werden schnell zahlreiche inhaltliche Doppelungen deutlich: die weniger umfangreichen Vorgänge auf Ebene 1 überschneiden sich inhaltlich mit Vorgängen, die dem Vorgang „Regisafe“ auf Ebene 1 zugeordnet sind. Diese Doppelung setzt sich auf Dateiebene fort. Eine Dublettenprüfung mit dem Freeware-Tool CloneSpy ergab, dass die Testdaten aus Kirchheim insgesamt 226 Prüfsummendubletten in 105 Dateikategorien enthielten (also 8 % der Gesamtzahl der Dateien).[12]

[12] Konzen (wie Anm. 1) S. 16–18.

Damit weisen diese Ablagestrukturen viele Merkmale auf, die Ulrich Schludi als typische Merkmale von Fileablagen nennt:[13]

1. Bedingt dadurch, dass der Bearbeiter eine große Zahl von Ordnerebenen und beliebig viele Ordner pro Ebene erstellen kann, entstehen sehr umfangreiche und unübersichtliche Ablagestrukturen.
2. Es besteht keine klare Hierarchie zwischen unterschiedlichen Ordnungsmerkmalen, sondern diese werden in unsystematischer Weise und oft parallel auf gleicher Ebene angewandt.
3. Die Ablage wird nicht nach Prozessen, sondern nach Themen, Sachen, Objekten strukturiert.
4. Es existieren z.T. mehrere Verzeichnisse für den gleichen Sachverhalt, es bestehen also inhaltliche Überschneidungen zwischen den Verzeichnissen.
5. Einzeldokumente sind nicht nur innerhalb der Vorgänge auf der jeweils untersten Hierarchieebene zu finden, sondern auf jeder Ebene.
6. Dateien werden zum Teil mehrfach an unterschiedlichen Orten abgelegt.
7. Es wird nicht bewusst eine Akte für die Ablage formiert, d.h. der Sachbearbeiter entlastet die Ablage nicht von Dokumenten, die nicht aktenrelevant sind.

Diese Merkmale treffen alle auf einen Teil der Testdaten aus dem DMS der Stadt Kirchheim mit dem Aktenzeichen 048.751 sowie die darüber hinaus erwähnten „Bearbeiternester“ an anderen Stellen des Aktenplans zu, einige davon − insbesondere die Punkte 3, 5 und 6 − auch auf die Testdaten aus Karlsruhe. In wie weit dieser Befund repräsentativ für die Ablagestruktur in den untersuchten DMS ist, konnte aufgrund der beschränkten Einsichtsrechte im Rahmen der Untersuchung nicht abschließend beurteilt werden − es ist aber sicherlich kein seltenes Phänomen.

Schludi nennt als weitere Merkmale von Fileablagen eine fehlende Trennung zwischen Aktenplan und Aktenbereich.[14] Dies ist in den beiden Untersuchungsfällen ebenfalls gegeben, da einerseits Aktenzeichen und Akte in den untersuchten DMS gleichgesetzt wurden, zum anderen einzelne Vorgänge und Dokumente ohne Aktenzeichenzuordnung mitten im Aktenplan angelegt wurden (letzteres betraf allerdings, wie erwähnt, nur wenige Verzeichnisse). Schließlich hält Schludi fest, dass bei Fileablagen die Metadaten zur Steuerung des Aussonderungsprozesses − d.h. das Datum der letzten Änderung als Anhaltspunkt für den Abschluss der Bearbeitung − nicht auf Ordnerebene, sondern auf Dateiebene zu finden sind.[15] Selbst hier waren die untersuchten Ablagen im DMS der Fileablage nicht überlegen: Zwar bietet Regisafe die Möglichkeit, Dokumente und Vorgänge abzuschließen und Aufbewahrungsfristen zu setzen; da diese Funktion von den Sachbearbeitern nicht genutzt wurde, war jedoch keine Auswahl aussonderungsreifer Unterlagen auf Vorgangsebene möglich.

Die Herausforderungen, die aus der Nutzung von Fileablagen in öffentlichen Verwaltungen aus archivarischer Sicht im Hinblick auf Aussonderung, Bewertung und Übernahme entstehen, können also auch bei elektronischen Unterlagen aus DMS auftreten: Die Verwen-

[13] Ulrich Schludi, Zwischen Records Management und digitaler Archivierung. Das Dateisystem als Basis von Schriftgutverwaltung und Überlieferungsbildung. In: Kai Naumann – Peter Müller (Hrsg.), Das neue Handwerk. Digitales Arbeiten in kleinen und mittleren Archiven, Stuttgart 2013, S. 20–38. URL: *http://www.landesarchiv-bw.de/sixcms/media.php/120/59735/Das%20Handwerk%20Inh.96S.pdf* (aufgerufen am 18.3.2016), v.a. S. 23–25, 31–33.
[14] Schludi (wie Anm. 12) S. 24.
[15] Ebd., S. 29.

dung eines DMS führt nicht zwangsläufig dazu, dass die aussonderungsreifen Unterlagen eine nachvollziehbare, aktenmäßige Struktur aufweisen und folglich mit vertretbarem Zeitaufwand einer inhaltlichen Bewertung unterzogen werden können. Ein DMS schützt nicht zwangsläufig vor inhaltlichen Redundanzen in der Ablage und ermöglicht nicht zwangsläufig eine Steuerung des Aussonderungsprozesses über Metadaten auf Vorgangsebene. Um zu verhindern, dass Ablagestrukturen in DMS faktisch die Gestalt von Fileablagen annehmen, ist es notwendig, dass die betreffende Verwaltung entsprechende normative Vorgaben konzipiert, deren Anwendung durch die Sachbearbeiter durchsetzt und Systeme nutzt, die durch funktionale Einschränkungen die Entstehung allzu kreativer Strukturen unterbinden. Allerdings ist davon auszugehen, dass Archive noch lange vor der Herausforderung stehen werden, elektronische Unterlagen aus DMS zu übernehmen, die typische Merkmale von Fileablagen aufweisen.

Überlieferung von E-Mail-Konten als genuin digitale Unterlagen. Archivwürdigkeit, Übernahmemethodik und Einblicke in die Entwicklung eines Werkzeugs

Kristina Starkloff

Ein Archiv, das nur für Personen und Ämter zuständig ist, die nach den Regeln der Schriftgutverwaltung DIN ISO 15489-1 arbeiten, könnte das Thema E-Mail-Archivierung eigentlich mit gutem Gewissen ausklammern. Schließlich würde dort jeder vorgangsrelevante Schritt in das bereits eingeführte und reibungslos funktionierende Vorgangsbearbeitungssystem, Dokumentenmanagementsystem oder auch in die Stehordner und Hängeregister abgelegt. Darunter fiele zweifellos auch die E-Mail, die vielfach die analogen Schreiben ersetzt hat. Hier mag lediglich das Problem auftreten, dass die elektronische Post einige wichtige Parameter nicht mehr enthält, die im analogen Bereich selbstverständlich schienen. So ist gegenüber der Papierwelt das Phänomen des anonymen Briefs viel häufiger, weil kein vorgedrucktes Briefpapier mehr die Adressangaben des Absenders festhält und die E-Mail-Adresse nicht selten ein Akronym oder Pseudonym ist. Diese Lücken können inzwischen z.B. durch Kontaktformulare ausgeglichen werden, die bestimmte Angaben unumgänglich einfordern. Ganz wie in der analogen Welt sind die Angaben trotz allem nicht.

Doch was, wenn es an einer bestimmten Stelle keine geordnete Schriftgutverwaltung mehr gibt? Dann werden E-Mail-Konten der leitenden Personen zur einzigen Möglichkeit, die Handlungsgegenstände und die Handlungsweise dieser Einrichtung nachzuvollziehen. Und nach und nach erkennen Archive, dass E-Mail-Konten (Accounts) in manchen Fällen auch eine sinnvolle Ergänzung zur traditionellen Überlieferungsbildung sein können, da sie aufgrund ihrer im Alltag so vielfältig einsetzbaren Eigenschaft so manche zusätzliche Information enthalten.

Dieser Text beschreibt zum ersten, wie E-Mail-Archivierung mit dem Hauptthema dieses Buches zusammenhängt. Danach folgt eine kurze Darstellung der Eigenschaften und Strukturierungsformen von E-Mails und der Ansatzpunkte für die Archivierung. Zum dritten wird ein Projekt zur Auswahl und Übernahme von E-Mails aus E-Mail-Konten (Accounts) vorgestellt.

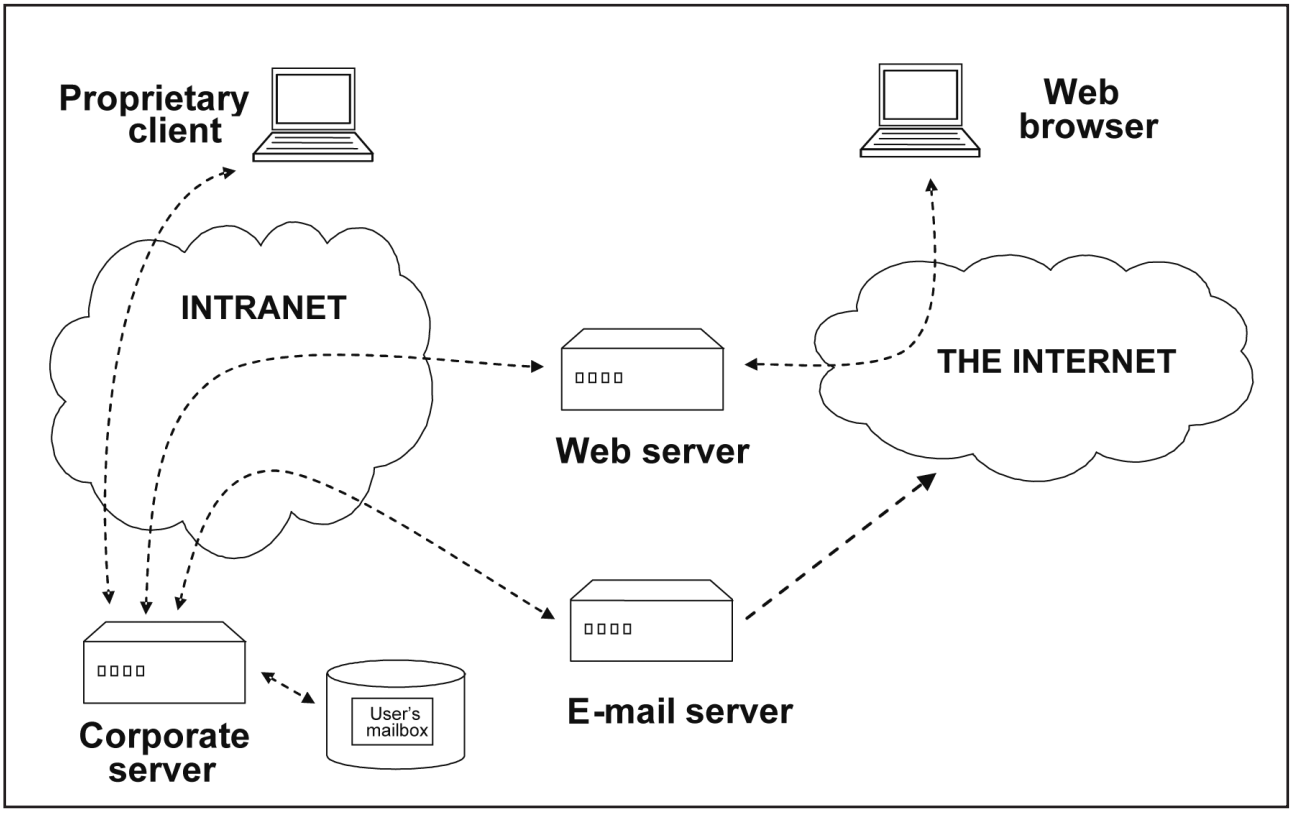
1. E-Mail-Sammlungen und Dateisammlungen

Zweifellos fallen E-Mail-Accounts unter die in diesem Buch gewählte Definition von „kreativen digitalen Ablagen“: Sie enthalten Unterlagen aller Art, was insbesondere in der Formatvielfalt von Anhängen deutlich wird. Ihre Inhalte gehen vielfach über einen rein entwurfsmäßigen Charakter hinaus. Sie werden von ihren Inhabern mitunter in Verzeichnisbäumen angelegt, wobei die Struktur je nach individueller Ordnungsliebe stark variieren kann. Nur sehr selten sind intellektuelle Einheiten so klar voneinander abgegrenzt, wie es bei klassischen Papierakten der Fall ist. E-Mail-Clients bieten aufgrund von Suchmöglichkeiten eine weitere Form der Strukturierung nach Bedarf. Bei Angabe eines Stichworts werden alle Nachrichten mit dieser Textfolge angezeigt. Auch nach E-Mail-Adressen oder besser Namen von natürlichen Personen können E-Mails gruppiert werden. Eine endgültige Ablage nach solchen Ordnungskriterien ist denkbar und mit wenig Arbeitsaufwand verbunden. Entsprechend ist davon auszugehen, dass sich die Felder der archivischen Aufbereitung von Dateisammlungen oder Intranetseiten, ja selbst von manchen inkonsequent eingesetzten E-Akten-Systemen, gegenseitig befruchten können.

2. Der Aufbau einer E-Mail und die Ansatzpunkte für die Übertragung ans Archiv

Wer sich mit E-Mails als genuin digitalen Unterlagen beschäftigt, muss sich mit ihrem Aufbau und Formaten beschäftigen. Verschiedene Publikationen widmen sich der Beschreibung der E-Mails, E-Mail-Accounts und dem technischen Umfeld, was an dieser Stelle nicht wiederholt werden soll.[1] Im Grunde sind vor allem zwei Beobachtungen relevant, die im Folgenden näher erläutert werden.

1. Beim Betrachten der Infrastruktur für E-Mails (siehe Abbildung) wird deutlich, dass es verschiedene Stadien im Lebenszyklus einer E-Mail gibt. Je nachdem in welchem sich die E-Mail befindet, liegen die Nachrichten in einheitlichen oder verschiedenen Container-Formaten vor.
2. E-Mails bestehen – verkürzt dargestellt – aus drei Teilen:
   a) (Erweiterter) Header
   b) Body (der Text der E-Mail)
   c) Anlagen



Infrastruktur der Übermittlung von E-Mails in Institutionen.
Aus: InterPARES 3 Project, General Study 05 – Keeping and Preserving E-Mail, S. 11

[1] Siehe exemplarisch: The InterPARES 3 Project, TEAM Italy, General Study 05 – Keeping and Preserving E-mail, September 2008, *http://www.interpares.org/ip3/display_file.cfm?doc=ip3_italy_gs05a_final_report.pdf* (aufgerufen am 02.05.2017).

Zu Punkt 1):

Nahezu jeder Nutzer von E-Mails bedient sich inzwischen des Komforts eines E-Mail-Clients. Neben den oben genannten Such- und Sortierfunktionen bieten diese inzwischen eine übersichtliche Darstellung, Textformatierungsmöglichkeiten und teilweise sogar sogenannte „Archivierungsfunktionen“. So praktisch diese Clients für den aktiven Gebrauch auch sind, der Umgang mit den dadurch je nach Client konvertierten Container-Formaten erweist sich im Bereich der Langzeitarchivierung als problematisch oder zumindest umständlich. Um dies zu umgehen, lohnt ein Blick auf die Infrastruktur von E-Mails. Allen Clients gemeinsam ist, dass sie ausgehende E-Mails in das Standardformat MIME (Multipurpose Internet Mail Extensions) überführen müssen, um einen Datenaustausch zu ermöglichen. In der anderen Richtung kommen eingehende E-Mails im MIME-Format an und müssen dann konvertiert werden. Somit liegt zu einem gewissen Zeitpunkt das einheitliche Container-Format MIME vor. Dies gilt es zu nutzen. Neben der Tatsache, dass es sich um einen stabilen Standard handelt, zählt zu den weiteren Vorteilen, dass alle Teile der E-Mail – außer manchen Anhängen – in Standardformaten codiert sind und mit geringem Aufwand in archivfähige Formate überführt werden können. Für die Übergabe- oder Übernahmeprozesse ist es – je nach der individuellen Voraussetzung – sinnvoll, soweit möglich auf der „Serverebene“ anzusetzen.

Zu Punkt 2):

Abgesehen von den Anhängen besteht jede E-Mail zwangsläufig aus einem erweiterten Header und dem Body, der meist die wesentliche Information beinhaltet. Wie das InterPARES 3 Project[2] in seiner Publikation zu E-Mails aufzeigt, enthält der Header eine Vielzahl an Angaben, die technisch unkompliziert als Metadaten genutzt werden können. Entsprechend muss das Archiv lediglich entscheiden, welche Informationen für Erschließung und Zugriff relevant sind, und diese in ein archivisches Fachinformationssystem (AFIS, OAIS-Funktionsbereich Data Management) oder ein digitales Archivsystem (OAIS-Funktionsbereich Storage) übernehmen. Ein manueller Eingriff in die Erschließung ist obsolet, was angesichts des Umfangs vieler E-Mail-Accounts beruhigt.

3. Ein Werkzeug zur E-Mail-Archivierung?

Die Anforderungen an ein Werkzeug zur E-Mail-Archivierung sind schnell formuliert: Sämtliche archivwürdige Teile eines E-Mail-Accounts sollen kontextuell verbunden bleiben und in archivierungsfähige Formate überführt werden. Um die verschiedenen Bereiche des Lebenszyklus einer digitalen Langzeitarchivierung unkompliziert zu gewährleisten, sollte eine ähnliche Praktikabilität wie die in der Clientumgebung erhalten bleiben (Ordnerbaum, Suche, Liste der E-Mails).

Aufgrund dieser Kriterien scheiden schon die meisten kommerziellen Programme aus, welche meist lediglich auf gesetzliche Aufbewahrungsfristen (meist bis 20 Jahre) in Unternehmen ausgerichtet sind. Auch die Idee, ein einheitliches PDF zu erstellen, das alle drei Teile einer E-Mail enthält, mag die meisten Informationswerte erhalten, den Zugriff jedoch erschweren. Die Schwierigkeit im Zugriff liegt vor allem daran, dass eine Vielzahl an Formaten als E-Mail-Anhänge versendet werden, die bei weitem nicht alle in das PDF-Format konvertiert werden können.

[2] Ebd. S. 36.

Die Entwicklung eines E-Mail-Werkzeugs als Scrum-Projekt

Im Archiv der Max-Planck-Gesellschaft (AMPG) bilden Nachlässe und darin die Korrespondenz der Wissenschaftler eine wichtige und stark genutzte Überlieferung. Vielfach ersetzten E-Mails seit einigen Jahren analoge Schreiben. Somit wurde schon vor längerer Zeit deutlich, dass die Übernahme der E-Mail-Überlieferung eine notwendige Maßnahme ist.

Neben der Klärung organisatorischer Modalitäten wie der Frage, ob E-Mails privat genutzt werden dürfen, anbietungspflichtig oder nicht sind, benötigen Archive ein Werkzeug, das die nachhaltige und weitgehend automatisierte Übernahme in ein Langzeitarchiv technisch unterstützt und auch für einen Archivar mit geringem IT-Hintergrund anwendbar ist.

Übergangslösung

Zunächst galt es, Überlieferungsverluste möglichst gering zu halten, weshalb eine Übergangslösung in Form einer Übernahme auf Serverebene eingerichtet wurde. Zu diesem Zweck wurde ein E-Mail-Server aufgesetzt. Abgebende Stellen können diesen mit ihrem gewohnten E-Mail-Client verbinden und die E-Mails in selbstbestimmter Auswahl auf ein extra eingerichtetes und mit Passwort geschütztes Konto kopieren. Der Vorteil für den Abgebenden ist, dass er keine Software installieren und erlernen muss und die E-Mails bei der abgebenden Stelle erhalten bleiben. Für das Archiv besteht der Vorteil darin, dass die Angaben, wie die Absenderadresse, das ursprüngliche Datum und ggf. die Ordnerstruktur unverändert vorliegen, und so auf eine spätere Weiterverarbeitung „warten“ können. Eine relative Sicherheit kann durch den Einsatz einer Firewall und die Passwortvergabe garantiert werden. Im Falle einer Abgabepflicht könnten die Nachrichten auch komplett z.B. durch die zuständige IT übergeben werden.

Nun ist das Archiv fähig, E-Mails im MIME-Standardformat zu bekommen, zur OAIS-konformen Archivierung wird aber das nächste passende Werkzeug benötigt.

Scrum-Projekt und Kooperationen

Da das Projekt „E-Mail-Archivierung“ zu komplex erschien, um in ein allumfassendes Pflichten heft gefasst zu werden, und einige Anforderungen oder Lösungsansätze anfangs noch unklar waren, entschied sich das Archiv der MPG zu einer agilen Softwareentwicklung in Form eines Scrum-Projekts.

Mit dem Anfang der 1990er Jahre entwickelten Scrum-Konzept soll verhindert werden, dass ein „fertiges“, aber im ungünstigsten Fall nicht passendes Produkt entsteht. Durch den engen Kontakt zwischen Archivar und Programmierer kann der laufende Prozess gesteuert werden. In regelmäßigen Treffen werden Missverständnisse als solche identifiziert und aufgeklärt, nicht funktionale Bereiche angesprochen und verbessert, aber auch rechtzeitig erkannt und behoben. Bereichernd erweisen sich die Vorschläge und Ideen beider Seiten, die in dem geschützten Rahmen besprochen, verworfen oder umgesetzt werden.

Zudem sollte ein weiteres Vorgehen probiert werden: Neben dem Ziel, ein passendes Produkt zu entwickeln, spielte die Frage nach Umsetzbarkeit mit geringem Budget eine entscheidende Rolle. Wenn Entwicklungen im Bereich der digitalen Langzeitarchivierung grundsätzlich von größeren finanziellen Möglichkeiten abhängig wären, würde dies die meisten kleineren Archive in eine passive Rolle drängen. Ob das der Fall ist oder Handlungsspielräume bestehen, galt es zu hinterfragen.

Eine wichtige Voraussetzung war es, Kooperationspartner im Bereich der Entwicklung zu finden. Darüber hinaus galt es eigene „Erfahrungslücken“, vor allem die fehlenden praktischen Erfahrungen in einem Produktivsystem der digitalen Langzeitarchivierung auszugleichen. Erstes Ziel war die Erstellung eines Pflichtenhefts, was das Landesarchiv Baden-Württemberg konzeptionell maßgeblich unterstützte. Das als dynamisches Konzept verstandene Ergebnis wurde dem Max-Planck-Institut für Informatik vorgelegt. Weitere sehr hilfreiche und vor allem praxisbezogene Hinweise flossen ein.

Mit dem Ergebnis begannen Mitarbeiter des Max-Planck-Archivs die Suche nach Kooperationspartnern in Berliner Hochschulen. Ein Seminar in Zusammenarbeit mit der Freien Universität Berlin, Institut für Informatik, AG Technische Informatik entstand. Drei Studentengruppen befassten sich in Folge mit der E-Mail-Archivierung, unterteilten die Bereiche in logische Arbeitseinheiten und überprüften vorhandene Produkte (z.B. das Softwarepaket der Stanford Universität ePADD[3]) auf ihren Inhalt und potentielle Anwendbarkeit. Mit dem Ende des Seminars waren einige Zwischenergebnisse erzielt. Fraglich ist nun, ob es ein weiteres Format, z.B. eine Masterarbeitvergabe geben könnte oder wie der Fortgang organisiert wird. Da nun grundlegende Kenntnisse und erste Ergebnisse vorliegen, ist die benötigte Kompetenz und Zeit leichter zu kalkulieren. Entsprechend mag es leichter sein, eine Abschlussarbeit zu vergeben, was aufgrund der sehr speziellen Thematik bisher nicht gelang.

Sicher ist in jedem Fall, dass sowohl die Archivare als auch die Studenten bislang viel lernten. Neben dem gegenseitigen Verständnis des jeweils anderen Kontextes sind gute Teilergebnisse entstanden, einige Archivarswünsche mussten der Realität weichen und dafür flossen kreative Ideen ein. Wie und wann genau mit einem fertigen Produkt zu rechnen ist, kann derzeit kaum präzisiert werden. Die vorhandenen Komponenten befinden sich in der Testphase, eine „Gebrauchsanweisung für Archivare“ ist begleitend in Arbeit und muss anschließend ausprobiert werden. Sicher scheint jedoch, dass das Produkt mit der Fertigstellung ein in sich schlüssiges und funktionales Werkzeug sein wird, das anderen Archiven zur Verfügung gestellt wird.

[3] *https://library.stanford.edu/projects/epadd* (aufgerufen am 2.5.2017).

Welche Schritte erfordert die Aufbereitung von Dateisammlungen und welche Querschnitts- und Spezialwerkzeuge werden gebraucht?

Kai Naumann

Die folgende Darstellung richtet sich an Archivare[1] und Informatiker, die Dateisammlungen übernehmen, bewerten und für die archivische Lagerung aufbereiten wollen, sowie an Informatiker, die Werkzeuge zu diesem Zweck schreiben.

Dateisammlungen (auch bekannt als Fileablagen) sind Ablagen für digitale Unterlagen aller Art, die eine Organisation oder eine Einzelperson auf Dateisystemen in Verzeichnisbäumen anlegt und die über einen rein entwurfsmäßigen Charakter hinausgehen. Dateisammlungen entstehen seit den 1990er Jahren, verstärkt ca. seit dem Jahr 2000. Eine klare Abgrenzung intellektueller Einheiten ist nur vorhanden, wenn in Anlehnung an Aktenpläne, Fall- oder Ortskennzeichen oder andere Taxonomien gearbeitet wird. Wo dies wie oft nicht der Fall ist, entsteht eine individuelle Ordnung, die Dateisammlungen unübersichtlich und – zumindest ohne Volltextrecherche – schwer benutzbar macht.

Im Folgenden sind Bearbeitungsschritte beschrieben, die derzeit bei den meisten Projekten nacheinander manuell durchgeführt werden. Dies erfolgt mit drei Zielen:

1. Um Interessierten, die ebenfalls Dateisammlungen bearbeiten wollen, die Arbeit zu erleichtern, werden die Schritte im Einzelnen erklärt.
2. Auch sollen einzelne Werkzeuge genannt werden, die zum Automatisieren des Schritts in Frage kommen.
3. Da Dateisammlungen aber eigentlich zu zahlreich sind, um Einheit für Einheit Schritt für Schritt einzeln manuell abzuarbeiten, ist zu zeigen, welche Werkzeuge (als sog. Querschnittswerkzeuge) viele Schritte begleiten. Dies erfolgt mit dem Ziel, den Entwicklern solcher Werkzeuge die Optimierung zu erleichtern und potenziellen Nutzern die Entscheidung für ein Querschnittswerkzeug zu erleichtern.

Nicht nur der Verfasser hat an diesem Text mitgeschrieben, sondern eine ganze Reihe weiterer Praktiker, nämlich Marco Birn, Marguérite Bos, Birgit Hartenstein, Martin Hoppenheit, Ulrike Jachemich, Maria Kobold, Nicole Martini, Annekathrin Miegel, Fabian Näser, Anne Kathrin Pfeuffer, Joachim Rausch, Christoph Schmidt, Heike Simon und Andreas Steigmeier. Ihnen sei hiermit herzlichst gedankt!

Abgrenzung dieses Dokuments

Da die Arbeit mit Dateisammlungen noch am Anfang steht, soll das Dokument im Zweifel eher als Diskussionsgrundlage und nicht als Normvorschlag verstanden werden. Einige Fragen der Archivierung, die dem Ingest nachgelagert sind, klammert diese Beschreibung aus. Diese müssen im größeren Kontext betrachtet werden, denn sie betreffen alle Arten von digitalem Archivgut gleichermaßen. Dieses Dokument ist auch keine Marktanalyse, die Kaufempfehlungen begründen kann. Bewusst wurde darauf verzichtet, die einzelnen Arbeitsschritte zu gewichten. Welcher Arbeitsschritt welche Relevanz hat, ist nur im konkreten Projektkontext festzustellen.

[1] Zur besseren Lesbarkeit werden in diesem Text Funktionsbezeichnungen für Menschen in der männlichen Form genutzt. Selbstverständlich gelten diese für Frauen und Männer gleichermaßen.

Anwendbarkeit auf verwandte Ablageformen

Nicht nur reine Fileablagen, sondern auch Kollaborationsplattformen (z.B. Confluence, SharePoint) oder E-Mail-Konten haben Eigenschaften, die in großen Teilen ebenfalls mit den geschilderten Arbeitsschritten bewältigt werden können. Hierfür ist es oftmals nur erforderlich, bestimmte Begriffe wie Datei oder Verzeichnisordner in andere wie E-Mail oder Seitenbereich umzumünzen.

Zur Terminologie

So jung und gut vernetzt die Disziplin der digitalen Archivierung auch ist, haben sich dennoch bereits unterschiedliche Begrifflichkeiten in den verschiedenen Arbeitsgruppen ergeben. Auch haben Begriffe von Institution zu Institution leicht unterschiedliche Bedeutungen. Vor und während der Lektüre dieses Texts ist es daher bei einigen Begriffen (z.B. SIP, AIP) sinnvoll, im Glossar am Ende des Texts die Bedeutung in diesem Dokument nachzuschlagen.

Charakteristika der Bearbeitung von Dateisammlungen

Bei der Eingangsbearbeitung von Dateisammlungen treten regelmäßig einige Probleme auf, die seitens des Archivs bewältigt werden müssen.

Motivationsproblem. Dateisammlungen sind vielen Mitarbeitern im Archivwesen fremd. Daraus folgt, dass solches Archivgut manchmal nur ungern bearbeitet wird. Die Anfangswiderstände beim Umgang mit fremdartigem Archivgut müssen überwunden werden. Immerhin gleichen Dateisammlungen in mancher Hinsicht einer Sammlung von Papierdokumenten, sodass diese Sorte digitaler Unterlagen für Neulinge in digitaler Archivierung naheliegt, wenn die Größe und die Vielfalt nicht überhand nehmen.

Automatisierungsproblem. Dateisammlungen können erheblich Arbeitskraft binden, weil viele Schritte erforderlich sind, die bei klassischem Archivgut fehlen. Zudem können diese Schritte mit einer normalen Büro-PC-Ausstattung nicht automatisiert werden. In der Folge kann ein Projekt sich totlaufen.

Überschaubarkeitsproblem. Dateisammlungen werden, weil Speicherplatz selten eine Rolle spielt und die Gliederungsmöglichkeiten grenzenlos sind, nach wenigen Jahren unüberschaubar. Daher müssen in der Aufbereitung Einheiten mit thematischer oder zeitlicher Eingrenzung und mit einer überschaubaren Menge an enthaltenen Unterlagen unterschieden werden können.

Paketierungsproblem. Sobald die Dateisammlung überschaubar geworden ist, müssen die späteren Bestelleinheiten konkret gebildet werden. Selten wird aus einer Sammlung eine, sondern meist eine Vielzahl von Bestelleinheiten erzeugt. Diese Aufgabe muss fachlich durchdrungen („Welche Einheiten sind zu bilden?“) und technisch („Wie formiere ich die Einheiten?“) umgesetzt werden. Die technische Umsetzung ist Teil des Mappingproblems.

Mappingproblem. Anders als echte Dokumenten-Management-Systeme haben Dateisammlungen individuelle und selten dokumentierte Konzepte für Strukturen und Metadaten. Deshalb müssen Methoden und Hilfsmittel vorhanden sein, um Ordnungsstrukturen und Metadaten der Bestelleinheit und des Einzelschriftstücks zu verstehen, zu extrahieren und in ein einheitliches, archivfähiges Modell zu überführen. Die Ordnungsstrukturen sind letztlich auch Metadaten, aber mit einem besonderen Charakter.

***Vielfaltsproblem*.** Kreative digitale Ablagen enthalten oftmals eine bunte Auswahl verschiedenster Dateitypen, weshalb die archivische Bestandserhaltung teilweise schon an diesem Punkt einsetzen sollte. Dateitypen, die von Obsoleszenz bedroht sind, sollten nach Möglichkeit bereits im Vorfeld in langfristig brauchbare Formate umgewandelt werden. Dateitypen, die mit Hilfe der gewählten Bestandserhaltungsstrategie gar nicht erhalten werden können, sollten früher oder später aus dem Prozess ausgefiltert oder gar nicht übernommen werden. Handelt es sich um eher zufällige Beigaben, sind Ausnahmen denkbar. Wenn zum Beispiel in einer Sammlung von Vogelstimmen unter 934 WAV-Tondateien ein oder zwei BMP-Bilddateien dabei sind, könnten manche Archive in Abwägung von Kosten und Nutzen auch ohne ein Bestandserhaltungskonzept für das BMP-Format eine Archivierung der Vogelstimmen mitsamt den BMP-Dateien in Erwägung ziehen.

Phasen der Bearbeitung von Dateisammlungen

1. Analyse

In dieser Phase gewinnt der Archivar über die Dateisammlung einen Überblick im Hinblick auf die Unterlagen, ihre Struktur und die Eigenschaften. Hier werden auch Viren oder unerwünschte Dateiformate erkannt.

In dieser Phase sollte festgehalten werden, was genau dem Archiv angeboten wurde.

2. Nachbewertung und SIP-Formierung

In dieser Phase wird festgelegt, welche Unterlagen archiviert werden sollen und in welcher Struktur sie übernommen werden. Die Strukturen können automatisiert anhand von vorhandenen Kriterien oder manuell festgelegt werden. Auch können die Strukturen mit Metadaten angereichert (erschlossen) werden.

Hier ergeben sich erwünschte Veränderungen, deren Ursachen transparent sein sollten. Unerwünschte Veränderungen sollten ausgeschlossen sein.

3. SIP-Erzeugung für Digitales Archiv und AFIS

Schließlich werden die im Schritt 2 festgelegten Parameter angewendet. Es entstehen Metadaten und Primärdaten in einer Form, die an ein Digitales Archivsystem (z.B. Archivematica, DIMAG, DIPS, DSpace) und ein Archivisches Fachinformationssystem (z.B. ArchivesSpace, Arcinsys, Augias, scopeArchiv) übergeben werden können.

Nach dem Ingest der Unterlagen sollte nachweisbar sein, welche Veränderungen sich in Schritt 1 bis 3 an den Unterlagen ergaben.

1.a–j) Die Analysephase im Einzelnen

1.a) Dateisammlungstyp auswählen

Ein Assistent öffnet sich, der zunächst fragt, ob es sich bei der Dateisammlung z.B. um ein Netzlaufwerk, eine USB-Festplatte, ein E-Mail-Konto oder eine Intranet-Umgebung handelt.

Dieser Schritt ist derzeit mehr eine Vision als eine konkrete Anforderung. Er setzt voraus, dass eine Software-Umgebung existiert, die sich mit allen im Beispiel genannten Quellablagen verbinden und auf die im Folgenden benötigten Objekte zugreifen kann.

1.b) Containerformate (gesamte Ablieferung oder einzelne Teile) auspacken

Gelegentlich sind die Dateisammlung insgesamt oder Teile von ihr in Containerformaten gepackt. Diese sollten vor der Bearbeitung aufgelöst werden. Es kann für das Backup (vgl. 1.j) erforderlich und sinnvoll sein, dass die Container neben den entpackten Dateien belassen werden.

Mit Containerformat sind Dateiformate gemeint, die ausschließlich eine oder mehrere Dateien umhüllen. Nicht nur verlustfreie Kompressionsformate (z.B. ZIP oder WARC), sondern auch Verschlüsselungsformate (z.B. VeraCrypt) gehören dazu. Nicht gemeint sind Container, die einen Dokumentenkontext herstellen, wie PDF/A, Office-Formate oder Videoformate.

1.c) Virenscan

Solch ein Scan ist vor der Übernahme in ein Digitales Archiv unabdingbar, wird aber vielfach als Querschnittsaufgabe der allgemeinen IT nebenbei erledigt, sofern die Aufbereitung auf Arbeitsplatzrechnern erfolgt. Dieser Schritt sollte nach dem Entpacken der Containerformate erfolgen, da diese vorher nicht geprüft werden können.

Im Idealfall sollte der Virenscan sich auch an späteren Punkten wiederholen, z.B. für den Fall, dass einzelne Dateien von der abgebenden Stelle nachgeliefert werden.

Wird eine Datei aufgrund eines Virenscans desinfiziert oder gelöscht, sollte dieser Schritt vermerkt werden bzw. in 3.d) nachvollziehbar sein.

1.d) Prüfsummen und IDs für jede Datei vergeben

Aus mehreren Gründen sollte jede Datei in der Sammlung eine eindeutige Kennung, besser eine Prüfsumme bekommen, um nachweisen zu können, dass in der Obhut des Archivs, vor allem in den folgenden Arbeitsschritten, keine Datei versehentlich oder böswillig gelöscht oder verändert wurde. Die Ergebnisse sollten in einer übergreifenden Bestandsaufnahme festgehalten werden.

1.e) IDs für jeden Verzeichnisordner vergeben

Das gleiche wie für die Dateien gilt auch für Ordner. Ordnernamen sind Metadaten, die ebenfalls verloren gehen oder verfälscht werden können. Die Ergebnisse sollten in einer übergreifenden Bestandsaufnahme festgehalten werden.

1.f) Normalisieren von Datei- und Ordnernamen (z.B. Eliminieren von Umlauten u.a. Sonderzeichen, Vergabe von IDs)

Dieser Schritt erfolgt, um den Umgang mit den Dateien für die beteiligten Softwareprodukte zu erleichtern, denn Dateinamen können schwierige Sonderzeichen enthalten, die manche Software zum Absturz bringt.

Vereinzelt können Dateinamen irreführend sein und die Benutzung erschweren, z.B. eine DOC-Datei, die die Endung DOCX hat. Hier ist ein Umbenennen der Datei nötig, der Vorgang sollte aber festgehalten werden.

Der Schritt kann mit 1.d und 1.e in Verbindung stehen.

1.g) Dateien analysieren

Die Analyse stellt Informationen bereit, die für die Folgeschritte besonders in Phase 2 relevant werden, z.B. Dateiformate, Datum/Zeit, eingebettete Metadaten (z.B. IPTC, EXIF).

Unter Umständen können hier Ungereimtheiten auftauchen, die die Lesbarkeit der Dateien beeinträchtigen. Beispielsweise können Dateien eine falsche Dateiendung haben, weshalb sie später umbenannt werden müssen.

1.h) Verzeichnisordner analysieren, z.B. nach Ebene in der Baumstruktur, Datum/Zeit

Bei diesem Schritt wird auch deutlich, welche Informationsmengen sich hinter welchem Teil des Verzeichnisbaums verbergen.

1.i) Verweisobjekte (Hyperlinks, Dateiverknüpfungen) analysieren

Dieser Schritt ist meist nicht erforderlich, da Verknüpfungen selten in Dateisammlungen verwendet werden. Er sei – auch mit Blick auf Intranet-Lösungen – der Vollständigkeit halber erwähnt.

1.j) Backup des Zugangs erstellen

Dieser Schritt bietet sich an, wenn in der Folge an den Daten weitergearbeitet wird, denn durch das Backup kann nachvollzogen werden, welche Dateien gelöscht oder sonst verändert wurden. Irrtümliche Veränderungen können so bis zum Ende des Gesamtablaufs rückgängig gemacht werden.

Fallstricke gibt es beim Kopieren folgende:

- Pfadangaben, die mehr als 200 Zeichen haben, werden ignoriert und nicht mitkopiert
- Dateien mit bestimmten Attributen im Dateisystem werden ignoriert und nicht mitkopiert

2.a–g) Die Phase der Nachbewertung und SIP-Formierung im Einzelnen

2.a) Bei Bedarf Duplikate finden und löschen

Das Herausfiltern bzw. der Abgleich vermeintlich redundanter Ordner oder Dateien kann automatisiert erfolgen. Hierfür gibt es bereits ein paar Werkzeuge, die Unterschiede und Gemeinsamkeiten feststellen und hierzu einen Report erzeugen. Bei Bedarf können diese exportiert oder gelöscht werden.

Grundsätzlich sollten aber nicht alle Duplikate ausgemerzt werden müssen. Auch in Papierakten finden sich seit Einführung des Fotokopierers massenhaft Duplikate. Duplikate entsprechen durchaus einer geordneten Schriftgutverwaltung, wenn gute Gründe dafür bestehen, ein Dokument an zwei Ordnungspositionen abzulegen.

Sollen Duplikate aus Speicherplatzgründen vermieden werden, so ist es am sinnvollsten, die Duplikatkontrolle auf der Speicherebene (also in einem Teilmodul des Digitalen Archivsystems) anzusiedeln.

Das Ergebnis der Löschung von Duplikaten sollte vermerkt werden bzw. in 3.d) nachvollziehbar sein.

2.b) Suche nach Hinweisen auf sensible personenbezogene Daten

Dokumente, die besonderen Geheimhaltungsvorschriften unterliegen könnten, enthalten bestimmte Textmuster, die von einer Volltextsuche erkannt werden können. Hierfür könnten hinterlegte Vokabulare (z.B. Bezeichnungen von Krankheiten beim Arztgeheimnis) oder reguläre Ausdrücke (z.B. Ansetzungsregeln für Kreditkartennummern beim Bankgeheimnis) verwendet werden.

Entsprechende Dokumente werden dann markiert. Es besteht die Möglichkeit, in der Erschließung entsprechende Sperrvermerke zu setzen oder die Dokumente mangels Relevanz für die Überlieferung auszulassen.

2.c) Nachbewerten, auch automatisiert (z.B. nach Datenmengen, Formattypen, Entstehungszeit)

Die Analyse der Dateisammlungen nach den verschiedenen Attributen der Dateien kann für die Bewertung hilfreich sein. Zu beachten ist, dass durch Kopieren bzw. Verschieben der Dateien deren Metadaten nicht geändert werden dürfen.

Mit speziellen Prüfsummen-Datenbanken können z.B. alle Systemdateien eines Betriebssystems oder anderer Software ausgefiltert werden, wenn gewünscht. Auch können Dateien mit bestimmten Namen oder Dateiendungen (z.B. thumbs.db) ausgefiltert werden.

Das Ergebnis der Nachbewertung sollte vermerkt werden bzw. in 3.d) nachvollziehbar sein.

2.d) Formatkonversionen zur Vereinheitlichung

Dieser Punkt hat eher deklaratorischen Charakter, da diese wichtige Aufgabe nicht nur bei Dateisammlungen, sondern auch bei E-Akten oder archivierten Webseiten in großem Stil ansteht. Insofern sei hier nur gesagt, dass der Schritt an dieser Stelle sinnvoll ist und Ergebnisse des Schritts festgehalten werden sollten. Hier die Werkzeuge zu seiner Bewältigung zu nennen, würde aber den Rahmen dieses Papiers leider sprengen.

Wichtig ist die Erkenntnis, dass bestimmte Metadaten aus Primärdatendateien bei Formatmigrationen verloren gehen können (z.B. IPTC-Metadaten in einer Bilddatei). Daher sollten solche Metadaten zuvor in Bestandsaufnahmen oder der Datenhaltung des Querschnittswerkzeugs gesichert werden.

2.e) Die einzelnen SIPs automatisiert formieren

Eine automatisierte Formierung von SIPs ist oftmals erforderlich, da die Grenzen intellektueller Einheiten in Dateisammlungen unklar formuliert sind. Hierfür gibt es zwei Varianten.

2.e.1) Variante: Merkmale für die SIP-Formierung automatisch aufgrund Ordnernamen setzen

Das Paketieren der Verzeichnispfade und Dateien kann, wenn die Analyse der Strukturen abgeschlossen ist, anhand von verschiedenen Metadaten automatisiert erfolgen. In dieser Variante ist es das klassische Metadatum des Ordnernamens.

2.e.2) Variante: SIP abgrenzen durch Gruppieren nach bestimmten Metadaten => d.h. Formierung „künstlicher“ Abgrenzungen

Eine andere Variante ist, sich andere Metadaten zunutze zu machen. Zum Beispiel könnte es sinnvoll sein, alle Office-Dateien älteren Typs (DOC, MDB, PPT, XLS) zusammenzuziehen, oder eine Fotosammlung chronologisch in Monatspakete (ermittelt nach Erstelldatum der Dateien) aufzuspalten.

Diese Option kann auch sinnvoll sein, wenn die eigentlichen Ingestprozesse für den Transfer in das Digitale Archivsystem Mengenbeschränkungen aufweisen, z.B. pro Ingest nur 2 GB eingespielt werden können.

2.f) Die SIPs händisch formieren bzw. die in 2.c erzeugten SIP-Abgrenzungen händisch anpassen

Nachdem durch die automatische Bildung der SIPs eine Struktur vorliegt, kann es sinnvoll sein, diese durch manuelle Eingriffe weiter zu verbessern.

2.g) Vorgezogene Erschließung, d.h. Beschreibungen zu Dateien, Strukturobjekten oder Verweisobjekten (z.B. Aktenzeichen, Ablieferungsnummern) hinzufügen

Auch bei der vorgezogenen Erschließung gibt es zwei Varianten, die im Folgenden erläutert werden.

2.g.1) Variante: Händischer Editor

Es ist möglich, ein Objekt aufzurufen und in einem Formular bestimmte Metadaten zu ergänzen, idealerweise bereits mit archivischem Mapping. Hier erhobene Metadaten (z.B. IPTC-Tags in einem JPEG) können aber auch ersatzweise per Schritt 2.g.3) in ein archivisches Mapping mitgenommen werden.

2.g.2) Variante: Mehrfachaktualisierung

Dieser Schritt kann sinnvoll sein, wenn z.B. alle Fotos eines Ingests von einem Fotografen stammen und dessen Urheberrechtsvermerk übergreifend gesetzt werden soll.

2.g.3) Variante: Lookup-Prozesse (Nachschlagen aus anderen Datenquellen, z.B. XML, CSV)

Nicht selten werden der Dateisammlung Metadaten mitgegeben, zum Beispiel bei fallbezogener Ablage aus einem Fachverfahren oder aus einer archivischen Bestandsaufnahme bei der Sichtung. Diese Metadaten können durch maschinelles Nachschlagen mit einbezogen werden, wenn z.B. der Dateiname auf die Metadaten referenziert.

3.a–e) Die Phase der SIP-Erzeugung für Digitales Archiv und Archivinformationssystem im Einzelnen

3.a) Metadaten und Primärdaten für den Ingest in das Digitale Archivsystem bereitstellen. Soweit Dateipfade als Metadatum vorliegen, werden sie mitgegeben

Dieser Schritt bedeutet, dass die in Schritt 2 definierten Umwandlungen ausgeführt werden.

Erst jetzt sollten die angebotenen digitalen Unterlagen, sofern dies nötig ist, umgruppiert und verändert werden. Es ist aber auch denkbar, dass bereits in Phase 2 Löschungen und Veränderungen stattfanden. Dann kommt dem Backup 1.j) besondere Bedeutung zu.

3.b) Erschließungsmetadaten für den Ingest in das AFIS erzeugen

Archivische Fachinformationssysteme (AFIS) sind die Online-Kataloge der Archive. Die Metadaten, die zur Katalogisierung (Erschließung) gebraucht werden, müssen entweder aus den Metadaten zu 3.a) erzeugt werden oder separat bereitgestellt werden. Wichtig ist stets, dass eine Kennung oder ID eine – zumindest im Rahmen des Zugangs – eineindeutige Verbindung zwischen Erschließungsobjekt und zugehörigem SIP herstellt.

Hier gibt es zwei Varianten:

3.b.1) Variante: Erschließungsmetadaten enthalten nur Objekte auf SIP-Ebene und darunter

In dieser Variante wird dem AFIS lediglich eine lange Liste von Erschließungseinheiten in einem Zugang/SIC übergeben, die einer Ablieferungsliste entspricht. Diese kann viele 1000 Positionen haben, was die Übersicht verschlechtert.

3.b.2) Variante: Erschließungsmetadaten enthalten auch Strukturobjekte oberhalb der SIP-Ebene (z.B. Teilbestand, Findbuchkapitel)

In dieser Variante können dem AFIS auch Strukturobjekte übergeben werden, die Gliederungsmöglichkeiten schaffen, um große Zugänge zu untergliedern. Eine Lieferung von digitalisierten Karten kann so beispielsweise in Maßstabsbereiche, darunter in Entstehungsjahrzehnte untergliedert werden.

*3.c) F***ür die SIP-Ebene** *werden menschenlesbare IDs erzeugt (z.B. fortlaufende Nummern, Bestellsignaturen)*

Im Lesesaal brauchen Nutzer von Archivgut die Möglichkeit, sich Notizen zu bestimmten Quellen zu machen. Hierfür sind kurze Signaturen mit Buchstaben und Ganzzahlen erforderlich (z.B. AB Nr. 1234), die im Zuge der Aufbereitung erzeugt werden sollten.

3.d) Validierung, ob die gewünschten Schritte durchlaufen wurden

Validierung ist als Gültigkeits- und Tauglichkeitsprüfung der durchlaufenen Schritte zu verstehen. Diese Prüfung kann mehr oder weniger gründlich ausfallen, ausgehend von den zwei nachfolgenden Varianten.

3.d.1) Variante: Vergleichsmöglichkeit zwischen Anfangs- und Endzustand

Hier kann vom Anfang der Analysephase eine Bestandsaufnahme mit Prüfsummen und Ordner-IDs aus 1.d) und 1.e) hinzugezogen werden, die nun am Prozessende mit dem Bestand abgeglichen wird. Komfortable Vergleichsmöglichkeiten mit grafischer Darstellung sind bislang aber nicht bekannt.

3.d.2) Variante: Detaillierte Prüfung

Denkbar ist, dass jeder Schritt in den hier genannten Prozessen in einem Protokoll festgehalten und auf Dauer nachgewiesen wird. Auch können wenigstens folgenreiche und schwerwiegende Schritte (vgl. 3.e) gesondert protokolliert und nachgewiesen werden. Noch eine Variante wäre, dass der Bereitstellungsprozess 3.a) genauer mitprotokolliert wird. Solche Protokolleinträge könnten zum Schluss nochmals überprüft werden.

3.e) Metadaten erstellen, die die Schritte des Prozesses und die verwendeten Parameter beschreiben

In Protokollmetadaten zu einzelnen Objekten oder Berichten in Dokumentform kann nachvollzogen werden, wie der Übernahmeprozess genau ablief.

Die Werkzeuge

Spezialwerkzeuge

Als Spezialwerkzeug wird hier ein Tool bezeichnet, das nur einen oder mehrere isolierte Arbeitsschritte erfüllt, aber nicht durch den Gesamtprozess führt. Da die Spezialwerkzeuge im Internet einen gewissen Bekanntheitsgrad erlangt haben, genügt zum Finden eine Suchmaschine, sodass hier auf Links verzichtet wird. Die meisten Werkzeuge sind im Anhang beschrieben.

Querschnittswerkzeuge

Als Querschnittswerkzeuge werden hier Software-Tools verstanden, die mehrere oder alle der hier dargestellen Arbeitsschritte in einen Bearbeitungsprozess integrieren. Manche dieser Werkzeuge sind nicht im Internet frei verfügbar, sondern dort nur beschrieben. Es gibt weitere Querschnittswerkzeuge, die in dieser Aufstellung nicht erwähnt sind, z.B. DILA Import Preparation Tool (vgl. Aufsatz von Anne Kathrin Pfeuffer in diesem Band) oder den Archivematica Appraisal Tab[2].

ByteBarn – Das vom Sächsischen Staatsarchiv entwickelte Werkzeug wird in diesem Band beschrieben. Präsentationsfolien: *http://www.staatsarchiv.sg.ch/home/ auds/20/_jcr_content/Par/downloadlist_0/DownloadListPar/download_2.ocFile/HUTH_AUDS2016_StA_Huth_final.pdf*

DIMAG IngestTool – Das IngestTool wurde vom Hessischen Landesarchiv entwickelt und steht Partnern des DIMAG-Entwicklungsverbunds zur Verfügung. Vgl. das Informationsblatt unter *http://dimag-wiki.la-bw.de/xwiki/bin/view/%C3%96ffentli che+Software+und+Informati onen/*

docuteam packer – Die schweizerische Firma docuteam stellt diese Software kostenlos zur Verfügung. *https://wiki.docuteam.ch/doku.php?id=docuteam:packer*. Erläuterungen unter *http://www.docuteam.ch/angebot/digitales-archiv-docuteam-cos mos/software-as-a-service/*

IngestList – IngestList ist ein Modul der DIMAG-Software, das nicht nur DIMAG-Partnern, sondern aller Welt kostenlos zur Verfügung steht. *http://ingestlist.sf.net*

Package Handler – Package Handler wurde vom Schweizerischen Bundesarchiv entwickelt. Gegen Registrierung kann jedermann Package Handler herunterladen und benutzen. *https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/ package-handler.html*

PreIngest Toolset – Das Tool wurde von Hewlett-Packard im Auftrag des DiPS-Verbunds entwickelt und steht Partnern des DiPS-Verbunds zur Verfügung.

[2] Der Archivematica Appraisal Tab ist ein 2016 fertiggestelltes, mit Geldern der Andrew G. Mellon Stiftung gefördertes Zusatzmodul für Archivematica. Vgl. den Abschluss-Blogpost der Projektbetreuer: *http://archival-integration.blogspot.com/2016/11/the-end-is-just-new-beginning.html*.

Zur Entwicklung Folien von Nils Hoppe: *http://www.staatsarchiv.sg.ch/home/auds/17/ _jcr_content/Par/downloadlist_0/DownloadListPar/download.ocFile/Praesentation%20Hoppe.ppsx*

Zum Einsatz Folien von Heike Simon: *http://www.staatsarchiv.sg.ch/home/auds/20/_ jcr_content/Par/downloadlist_0/DownloadListPar/download.ocFile/SIMON_AUdS_praesentation_PIT_simon.pdf*

startext COMO – Das Produkt befand sich zum Redaktionsschluss noch in der Entwicklung, wurde aber bereits im Entwicklungsstadium öffentlich vorgeführt. *http://www.startext.de/*

Rechtlich-organisatorische Bedingungen der Werkzeuge

Das PreIngest Toolkit PIT und das DIMAG IngestTool sind nur für Partner der Verbünde DIPS bzw. DIMAG erhältlich.

Der Package Handler wird interessierten Benutzern nach Akzeptanz des Lizenzvertrages gebührenfrei zur Verfügung gestellt.

Die Produkte AccessData FTK Imager, Altova XMLSpy, Ingestamatic, MS Excel, oXygen, SleuthKit Autopsy, startext COMO, TreeSize Pro sind gebührenpflichtige Software.

Die Produkte ClamAV, docuteam packer, IngestList, LibreOffice Calc sind Open Source Software.

Technische Grenzen der Werkzeuge

Welche Datenmenge, welche Objektanzahl verträgt welches Werkzeug?

Assistenzwerkzeug	Grenzangabe
ByteBarn	unbekannt
DIMAG IngestTool	einzelne Prozesse mit >20.000 Dateien schwerfällig
docuteam packer	Dateigröße: Grenze ab 2 GB bekannt, Ausräumen der Grenze in Arbeit; Dateienanzahl: keine Grenze bekannt
IngestList	Anzahl der verarbeitbaren Dateien hängt vom Arbeitsspeicher ab. Minimum 6000 Dateien.
Package Handler	Keine obere Grenze für Dateigröße bekannt; bei Paketen, die größer als 2 GB sind, wird die Validierung aber deutlich langsamer (von fünf Minuten bis zu einer Stunde).
PreIngest Toolkit	Gesamtcontainer > 3 GB können nicht immer zuverlässig bearbeitet werden.
startext COMO	unbekannt

Umgang der Werkzeuge mit Metadatenstandards

Package Handler erzeugt und validiert SIP[3] (Aufbau und Metadaten) gemäß dem eCH-0160-Standard sowie gemäß darüber hinausgehenden Anforderungen des Schweizerischen Bundesarchivs (z.B. Anforderungen zu Schutzfristen, Ablieferungsnummer).

[3] SIP hier im Sinne von Submission Information Collection.

Der docuteam packer erzeugt Paketmetadaten gemäß METS (Profil Matterhorn METS), PREMIS und EAD. Er kann eCH-0160-Metadaten entgegennehmen.

Das DIMAG IngestTool erzeugt Paketmetadaten für das DIMAG-Kernmodul, die dem allgemeinen Standard SOAP entsprechen und in allen DIMAG-Installationen verarbeitet werden, aber sonst bislang nicht formell standardisiert sind.

Das DIMAG IngestTool kann Metadaten beliebiger Standards entgegennehmen, da es Metadatenfelder aus CSV oder XML gezielt auf die Ebene AIP, Repräsentation oder Datei mappen kann.

Raster über Bearbeitungsschritte und Erfüllungsgrade

Im Idealfall sollen die zu dem betreffenden Schritt genannten Werkzeuge in der Lage sein, den betreffenden Schritt jeweils über eine ganze Serie von Dateien automatisch abzuarbeiten. Dies ist aber noch lange nicht überall der Fall.

Die Fähigkeiten der genannten Software gelten zum Stand April 2017.

Die Wichtigkeit der einzelnen Schritte ist hier nicht thematisiert und muss im konkreten Fall ermessen werden. Das „beste“ Querschnittswerkzeug lässt sich nicht aus der Anzahl der erfüllten Punkte ableiten.

Die Nummerierung der einzelnen Schritte und Phasen dient nur zur Orientierung und geben keine zeitliche Reihenfolge vor. Die zeitliche Abfolge sollte aus den konkreten Gegebenheiten entwickelt werden, sofern technisch und konzeptuell möglich.

- ● Punkt wird erfüllt
- ◓ Punkt wird teilweise erfüllt
- ○ Hersteller hält spätere Erfüllung für denkbar

Bearbeitungsschritt	ByteBarn	docuteam packer	IngestList	DIMAG IngestTool	Startext COMO	Package Handler	PreIngest Toolset	Spezialwerkzeug (vorzugsweise kostenloses)
Phase 1: Analyse								
1.a) Dateisammlungstyp auswählen							◓[1]	
1.b) Containerformate (gesamte Ablieferung oder einzelne Teile) auspacken		◓[2]			○		◓[3]	7zip, WinZip, bei forensischen Formaten AccessData FTK Imager
1.c) Virenscan								Avira, ClamAV, Kaspersky
1.d) Prüfsummen und IDs für jede Datei vergeben		●	●		●	●	●	TreeSize Pro, Manifest Maker
1.e) IDs für jeden Verzeichnisordner vergeben		●		◓[4]	●	●	●	
1.f) Normalisierung von Datei- und Ordnernamen (z.B. Eliminieren von Umlauten u.a. Sonderzeichen, Vergabe von IDs)	●	◓[2]		●	○	●	●	IrfanView (Batchfunktion), Total Commander
1.g) Dateien analysieren, z.B. nach ungeeigneten Dateiformaten, Datum/Zeit, eingebettete Metadaten	●	◓[2]	●		●	◓	●	Sleuthkit Autopsy
1.h) Verzeichnisordner analysieren, z.B. nach Ebene in der Baumstruktur, Datum/Zeit	◓	◓		◓	◓	◓	◓	Total Commander, TreeSize Pro[5]
1.i) Verweisobjekte (Hyperlinks, Dateiverknüpfungen) analysieren							◓[6]	

Bearbeitungsschritt	ByteBarn	docuteam packer	IngestList	DIMAG IngestTool	Startext COMO	Package Handler	PreIngest Toolset	Spezialwerkzeug (vorzugsweise kostenloses)
1.j) Backup der Dateisammlung erstellen		●[7]				●[7]		Data Accessioner, Total Commander, Windows Explorer
Phase 2: Nachbewertung und SIP-Formierung								
2.a) Bei Bedarf Duplikate finden und löschen					○	●	●	CloneSpy, TreeSize Pro, WinMerge
2.b) Suche nach Hinweisen auf sensible personenbezogene Daten								Sleuthkit Autopsy
2.c) Nachbewerten, auch automatisiert (z.B. nach Datenmengen, Formattypen, Entstehungszeit)				◓[8]	●		●	Data Accessioner, Total Commander, TreeSize Pro
2.d) Formatkonversionen zur Vereinheitlichung	Dieser Punkt ist hier nur deklaratorisch erfasst. Da z.B. bei E-Akten dieser Schritt ebenfalls erforderlich ist, sollte er übergreifend anderweitig ausgeführt werden.							
2.e) Die einzelnen SIPs automatisiert formieren								
2.e.1) Variante: Merkmale für die SIP-Formierung automatisch aufgrund von Ordnernamen setzen				●	○	●	●	Stapelverarbeitung per Scriptsprache (manuelle Vorbereitung bzw. Auswahl durch weiteres Tool, wie z.B. TreeSizePro)
2.e.2) Variante: SIP abgrenzen durch Gruppieren nach bestimmten Metadaten => d.h. Formierung „künstlicher“ Abgrenzungen				◓[9]	●[9]		◓[9]	Total Commander, TreeSize Pro (beide nicht automatisiert), Stapelverarbeitung per Scriptsprache

Bearbeitungsschritt	ByteBarn	docuteam packer	IngestList	DIMAG IngestTool	Startext COMO	Package Handler	PreIngest Toolset	Spezialwerkzeug (vorzugsweise kostenloses)
2.f) Die SIPs händisch formieren bzw. die in 2.c erzeugten SIP-Abgrenzungen händisch anpassen		●		●	●	●	●	Total Commander, TreeSize Pro
2.g) Vorgezogene Erschließung, d.h. Beschreibungen zu Dateien, Strukturobjekten oder Verweisobjekten (z.B. Aktenzeichen, Ablieferungsnummern) hinzufügen.		●			●	●	●	
2.g.1) Variante: Händischer Editor		●	◓[10]		●	●	●	Data Accessioner, Ingestamatic (bei Fotos), Notepad++
2.g.2) Variante: Mehrfachaktualisierung					●	●		Ingestamatic (bei Fotos), MS Excel, LO Calc, oXygen, XMLSpy
2.g.3) Variante: Lookup-Prozesse (Nachschlagen aus anderen Datenquellen, z.B. XML, CSV)		◓[2]		●		◓[11]		LWL METS Generator
Phase 3 SIP-Erzeugung für Digitales Archiv und für Archivinformationssystem								
3.a) Metadaten und Primärdaten für den Ingest in das Digitale Archivsystem bereitstellen. Soweit Dateipfade als Metadatum vorliegen, werden sie mitgegeben.		●	◓[12]	●	●	●	●	Data Accessioner
3.b) Erschließungsmetadaten für den Metadatentransfer in das AFIS erzeugen								MS Excel, LO Calc, oXygen, XMLSpy

Bearbeitungsschritt	ByteBarn	docuteam packer	IngestList	DIMAG IngestTool	Startext COMO	Package Handler	PreIngest Toolset	Spezialwerkzeug (vorzugsweise kostenloses)
3.b.1) Variante: Erschließungsmetadaten enthalten nur Objekte auf SIP-Ebene und darunter		●		●	●	●	●	
3.b.2) Variante: Erschließungsmetadaten enthalten auch Strukturobjekte oberhalb der SIP-Ebene (z.B. Ablieferung, Findbuchkapitel, Teilbestand)					●	●		
3.c) Das Tool vergibt für die SIP-Ebene menschenlesbare IDs (z.B. fortlaufende Nummern, Bestellsignaturen).				◓[13]	○	◓[14]	●	
3.d) Validierung, ob die gewünschten Schritte durchlaufen wurden								
3.d.1) Variante: Vergleichsmöglichkeit zwischen Anfangs- und Endzustand			◓[15]		○		●[16]	TreeSize Pro (Vergleich mit gespeichertem Scan)
3.d.2) Variante: Detailliertere Prüfung	◓[17]	[1]			○	●	●	
3.e) Metadaten erstellen, die die Schritte des Prozesses und die verwendeten Parameter beschreiben				●	◓[18]		●	

Fußnoten zur Tabelle:

[1] PreIngest Toolset unterscheidet zwischen SIARD-Containern und Dateisystemen.

[2] Wird von docuteam feeder im Nachgang ausgeführt.

[3] In einigen Programmversionen enthaltenes Feature. Entpackt werden ZIP-Formate, TAR-Formate sowie E-Mail-Container (EML, MSG).

[4] DIMAG IngestTool vergibt keine IDs, arbeitet aber nach einem festen Sortieralgorithmus und kann per Lookup-Tabellen solche IDs vergeben.

[5] Während die Querschnittswerkzeuge hier nur einen Überblick über die Baumstruktur geben, kann TreeSize Pro einzelne Ordnerbereiche auch nach Zeit oder nach Dateityp untergliedern.

[6] Wenn E-Mail-Container ausgepackt wurden, sind Verweise auf den Namen des Objekts möglich.

[7] Das Programm holt alle Dateien einer Sammlung zunächst in ein Arbeitsverzeichnis, ehe es weitere Schritte vollzieht. Die Ausgangsdaten bleiben unangetastet. Bei docuteam packer können die Ausgangsdaten optional gelöscht werden, wenn der Datenträger überzulaufen droht.

[8] DIMAG IngestTool lässt das Filtern nach Dateiendungen oder bestimmten Teilen des Dateinamens zu.

[9] Formierung bei IngestTool und PIT nur nach Merkmalen in Dateinamen und Dateinamensendung. Bei startext COMO kann auch nach eingebetteten Metadaten und Primärdaten der Dateien gruppiert werden.

[10] IngestList erlaubt es, auf Dateiebene beschreibende Metadaten anzubringen. Auf SIP-Ebene können nur Protokollmetadaten angelegt werden, die DIMAG später übernimmt.

[11] Package Handler kann bei der Erstellung des SIP per CSV Metadaten und auch die Struktur von Ordnungssystempositionen (Strukturobjekten) und Dossiers (AIPs) importieren. Dadurch kann ein Teil der archivischen Erschließung direkt ins SIP übernommen werden.

[12] IngestList kann eine Bestandsaufnahme nur in genau ein AIP überführen.

[13] IngestTool kann per Lookup-Prozess eine Serie von IDs zuweisen. Der Einbau einer Zählerfunktion, die automatisch fortlaufende Nummern erzeugt, ist für spätere Versionen angedacht.

[14] Package Handler vergibt IDs nur für die Nummerierung der Strukturobjekte (dort als „Ordnungssystemposition“ bezeichnet).

[15] IngestList erstellt schon bei der abliefernden Stelle eine Bestandsaufnahme, die später begleitend zu den erstellten AIPs im digitalen Archivsystem abgelegt werden kann. Der Vergleich zwischen Ur-Bestandsaufnahme und Bestand im Archiv ist aber Nutzersache und nicht automatisiert.

[16] PIT ermöglicht am Schluss des Übernahmeprozesses einen automatisierten Rückblick auf die Aktionen im Rahmen des Ingests und prüft, ob die zur Übernahme vorgesehenen Dateien noch unverändert vorhanden sind.

[17] ByteBarn protokolliert Veränderungen an den Dateien mit.

[18] Bislang werden nur Bewertungsentscheidungen (Löschungen) automatisch protokolliert. Händische Einträge sind später denkbar.

Glossar

Dieses Glossar ist keine Einführung in alle Begriffe der Archivinformatik. Da viele Begriffe in der Community nicht immer identisch verstanden werden, gelten diese Definitionen nur im Kontext dieses Dokuments, um die Leser auf die Bedeutung wichtiger Begriffe vorzubereiten.

Archival Information Package (AIP)	Im Sinne von OAIS ein Archivpaket im Archivsystem, das separat bestellbar ist.
Archivisches Fachinformationssystem (AFIS)	Ein AFIS ist das Modul eines Archivs, das sämtliche Informationseinheiten (physisch und digital) recherchierbar und bestellbar macht.
Querschnittswerkzeug	Ein Software-Werkzeug, das mehrere oder alle der hier dargestellen Arbeitsschritte unterstützt.
Spezialwerkzeug	Ein Software-Werkzeug, das nur einzelne der hier dargestellten Arbeitsschritte unterstützt.
Strukturobjekt	Ein übergreifendes Metadatum, das AIPs ordnend zusammenfasst. Ein Strukturobjekt kann im DA einem AIP untergeordnet sein oder im AFIS zu einem ordnenden Merkmal (z.B. Ablieferung, Findbuchkapitel, Teilbestand) werden. z.B. bei Fileablagen: Pfadangabe z.B. bei Intranetseiten: Eintrag in der Navigationsleiste
Submission Information Collection (SIC)	s.u.
Submission Information Package (SIP)	Im Sinne von OAIS ein Informationspaket, das noch nicht archiviert wurde, das aber den Eingangsspezifikationen des DA entspricht. Aus Sicht dieses Papiers sind SIPs und AIPs in ihrer intellektuellen Abgrenzung jeweils identisch. Den Autoren ist bewusst, dass SIP vielfach für die Gesamtheit der späteren AIPs (Zugang) steht. Laut OAIS wäre dies als SIP in der Ausprägung SIC (Submission Information Collection) zu bezeichnen.
Verweisobjekt	Ein Metadatum, das Dateien oder Verzeichnisordner miteinander verbindet. z.B. bei E-Mails: Hyperlinks z.B. bei Aktenplänen: Verweise auf andere Aktenzeichen

Analyse und Datenaufbereitung von digitalen Ablagen mit TreeSize Professional und Total Commander

Marco Birn

1. Verwendung

Bei TreeSize Professional und Total Commander handelt es sich um lizenzpflichtige Software aus den Bereichen Speicherplatz- und Datei-Management. Sie kommen insbesondere bei der Phase 1: Analyse von Dateisammlungen und Phase 2: Nachbewertung und SIP-Formierung zum Einsatz.[1]

Im Gegensatz zu anderen in dieser Publikation vorgestellten Tools wurden die beiden Programme nicht speziell für den Preingest oder Ingest digitaler Objekte entwickelt. Deshalb müssen bei ihrer Verwendung einige Aspekte beachtet werden:

1. Zunächst sollte immer eine Kopie der digitalen Ablage erstellt werden. Andernfalls würden die Originaldaten verändert. Eine Bearbeitung sollte nur innerhalb der Kopie erfolgen.
2. Eingriffe werden nicht protokolliert. Alle Bearbeitungsschritte müssen manuell dokumentiert werden.

2. TreeSize Professional (TSP)

TSP ist ein Speicherplatz-Manager, der die Festplattenstruktur analysieren und grafisch darstellen kann. Die Software ermöglicht die Anzeige einzelner Verzeichnisse eines Laufwerks inklusive aller Unterverzeichnisse und kann zunächst als Analysetool für den ersten Überblick eingesetzt werden. Die Scanergebnisse des ausgewählten Datenträgers bzw. der Ordnerstruktur können dann manuell gefiltert und nach bestimmten Metadaten formiert werden. Zudem bietet das Programm eine automatisierte Dublettenbereinigung an.

2.1 Ansicht und Navigation

Nach dem Einscannen einer Verzeichnisstruktur über das Navigationsfeld erscheint im rechten Fenster die „Details“-Ansicht im Listenformat (Abb. 1).

2.2 Nachbewertung: Bereinigung obsoleter Dateien

Unter „Extras“ auf der Menüleiste lassen sich mit einem Klick auf die Schaltfläche „TreeSize Dateisuche öffnen“ vordefinierte Suchmuster auswählen (Abb. 2). Interessante Funktionen sind u. a. die Anzeige von Dubletten und obsoleter Daten wie z.B. temporärer Dateien. Nach Auswahl einer Dateisuche öffnet sich die Dateisuche (Abb. 4). Hier lassen sich die voreingestellten Suchmuster bearbeiten und beliebig erweitern. In der Ergebnisliste können die angezeigten Dateien dann gelöscht werden.

[1] Vgl. Beitrag von Kai Naumann in diesem Heft, Welche Schritte erfordert die Aufbereitung von Dateisammlungen und welche Querschnitts- und Spezialwerkzeuge werden gebraucht? S. 44–60, hier S. 46..

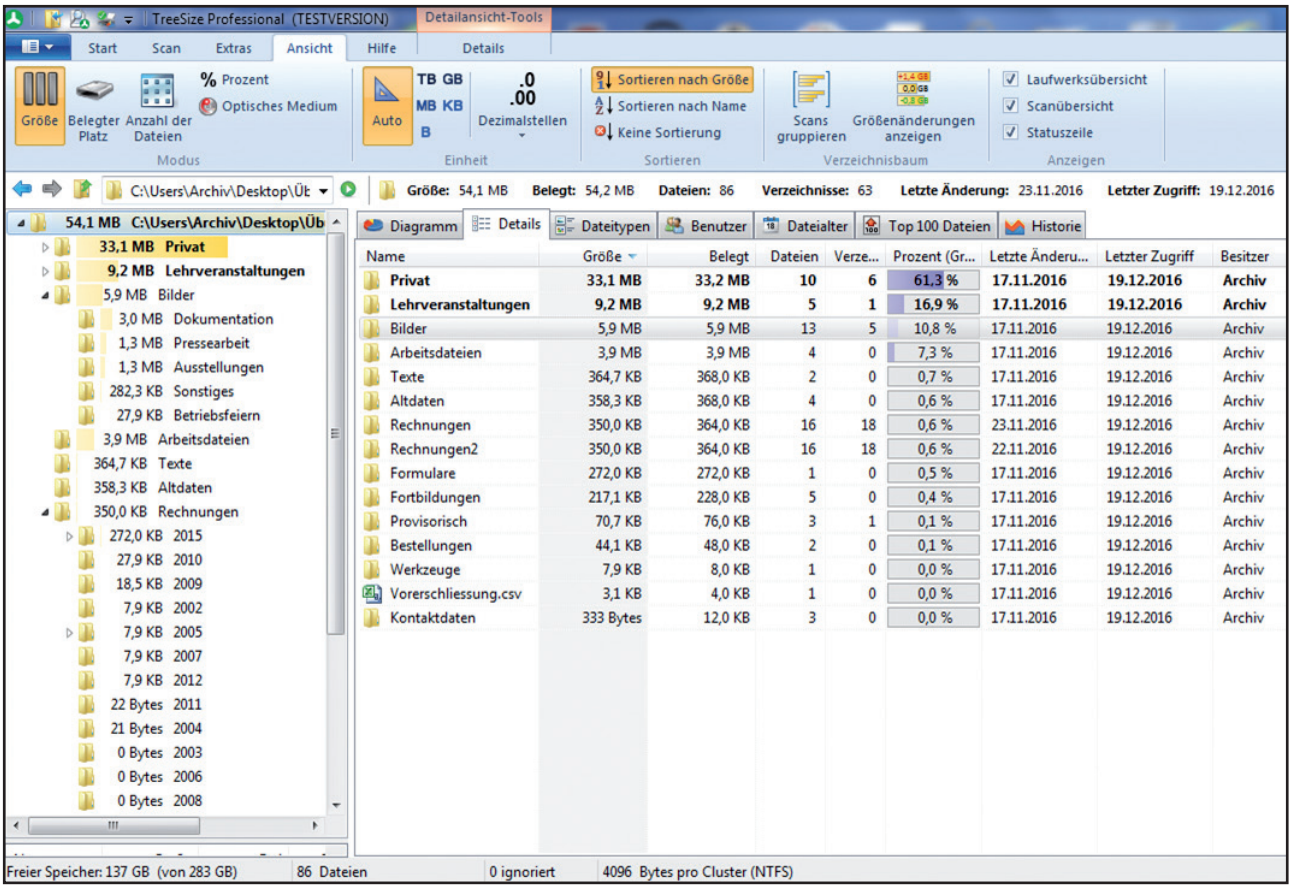


Abb. 1: Links: Navigationsfeld mit Verzeichnis; rechts: Ansicht der Scanergebnisse

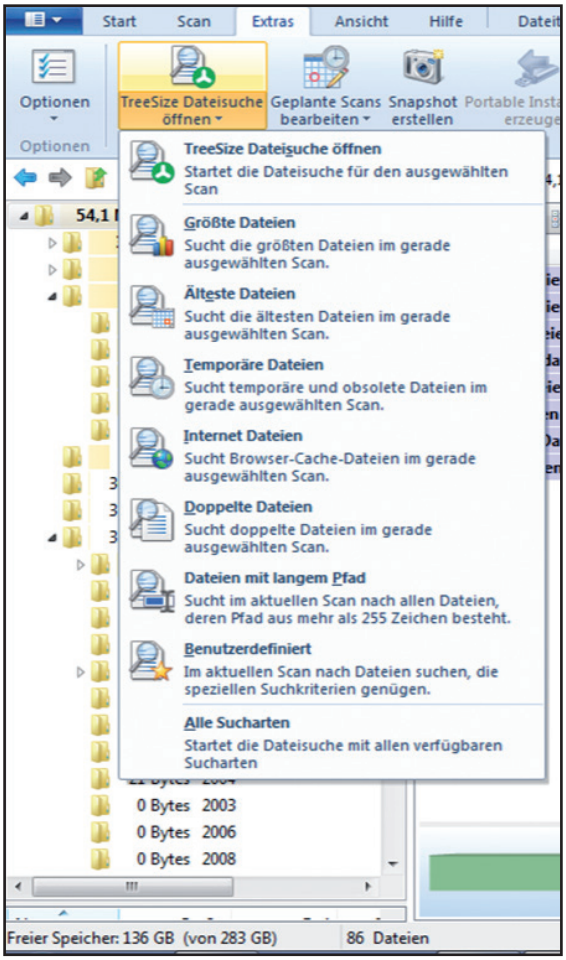


Abb. 2: Das Kontextmenü der Dateisuche

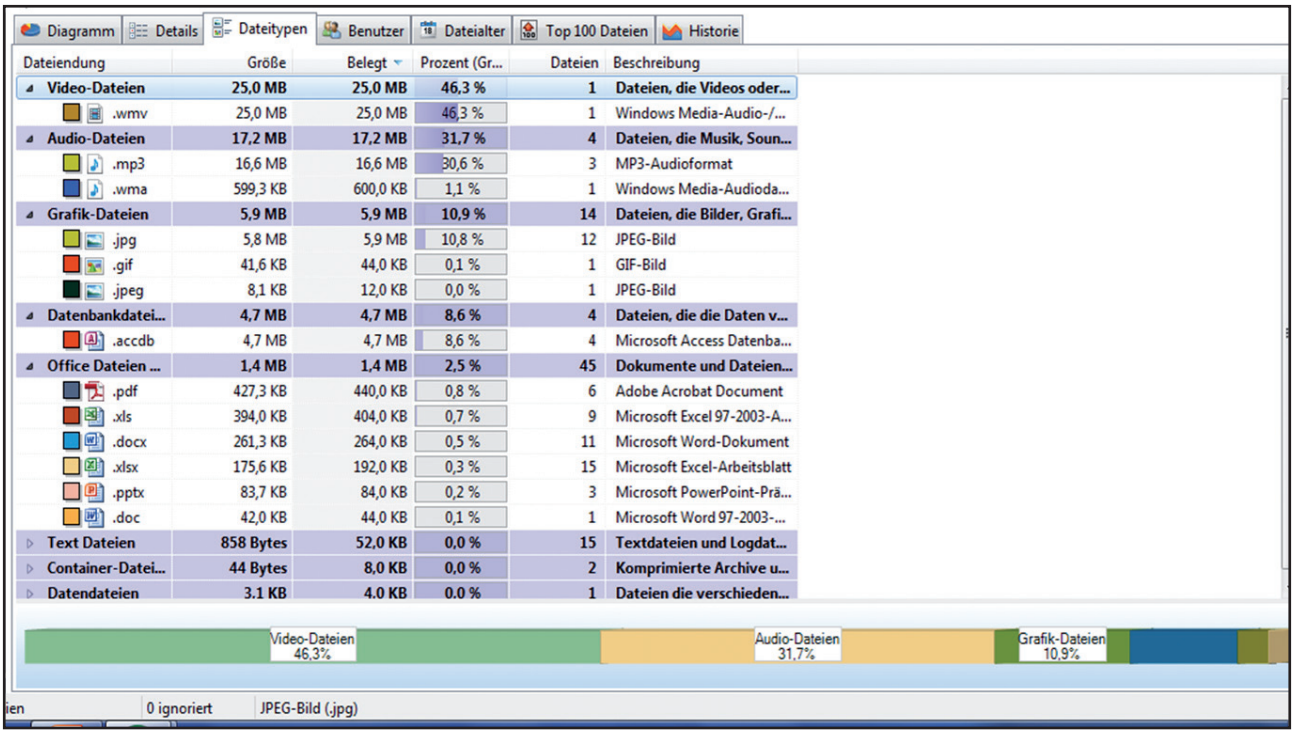


Abb. 3: Ansicht der vorhandenen Objekttypen und Dateiformate

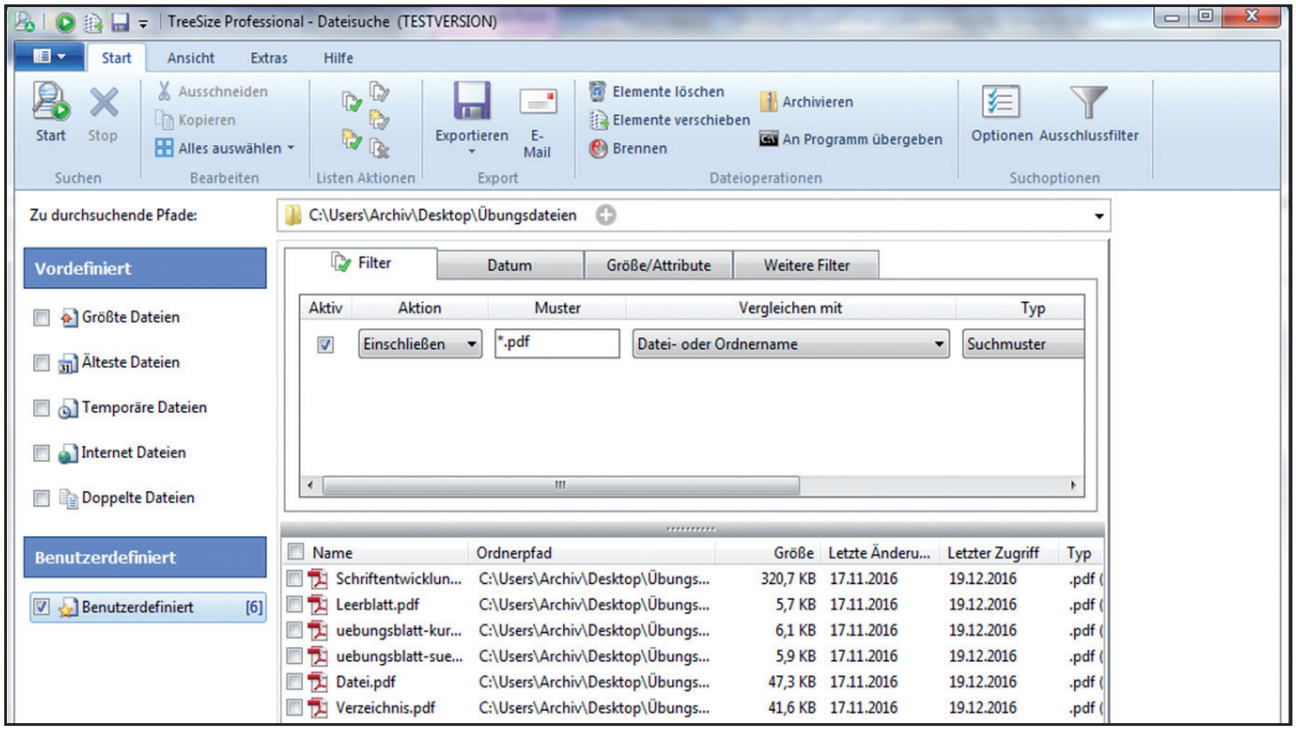


Abb. 4: Dateisuche und Filteransicht

2.3 SIPs formieren: Filtern und Gruppieren nach bestimmten Metadaten

In der „Details“-Ansicht (Abb. 3) über den gleichnamigen Reiter lassen sich alle Dateiformate nach Objektart anzeigen. Mit einem Rechtsklick auf das jeweilige Feld lassen sich automatisch entweder alle Objektarten oder nur eine bestimmte Dateiendung filtern.

Alternative 1: Über den Reiter „Dateitypen“ in der Menüleiste und das Feld „Zeige Dateien dieser Endung“ gelangt man über ein neues Fenster zur Filteransicht (Abb. 4). Hier lassen sich alle beliebigen Dateiendungen im gewählten Verzeichnis filtern.

Alternative 2: Über den Reiter „Extras“ in der Menüleiste und das Feld „TreeSize Dateisuche öffnen“ gelangt man ebenfalls zur Filteransicht (Siehe auch Kap. 2.2).

In der Filteransicht können die Dateien nach bestimmten Metadaten selektiert, anschließend in der unten befindlichen Ergebnisliste kopiert und schließlich in einem neuen Ordner als künftiges SIP eingefügt werden.

2.4 Scans exportieren und vergleichen

Um eine Bestandsaufnahme der Fileablage langfristig zu sichern, können die Scanergebnisse über die Schaltfläche „Exportieren“ in unterschiedlichen Formaten gespeichert werden (Abb. 5 und 6). Zu einem späteren Zeitpunkt können so mögliche Veränderungen festgestellt werden.

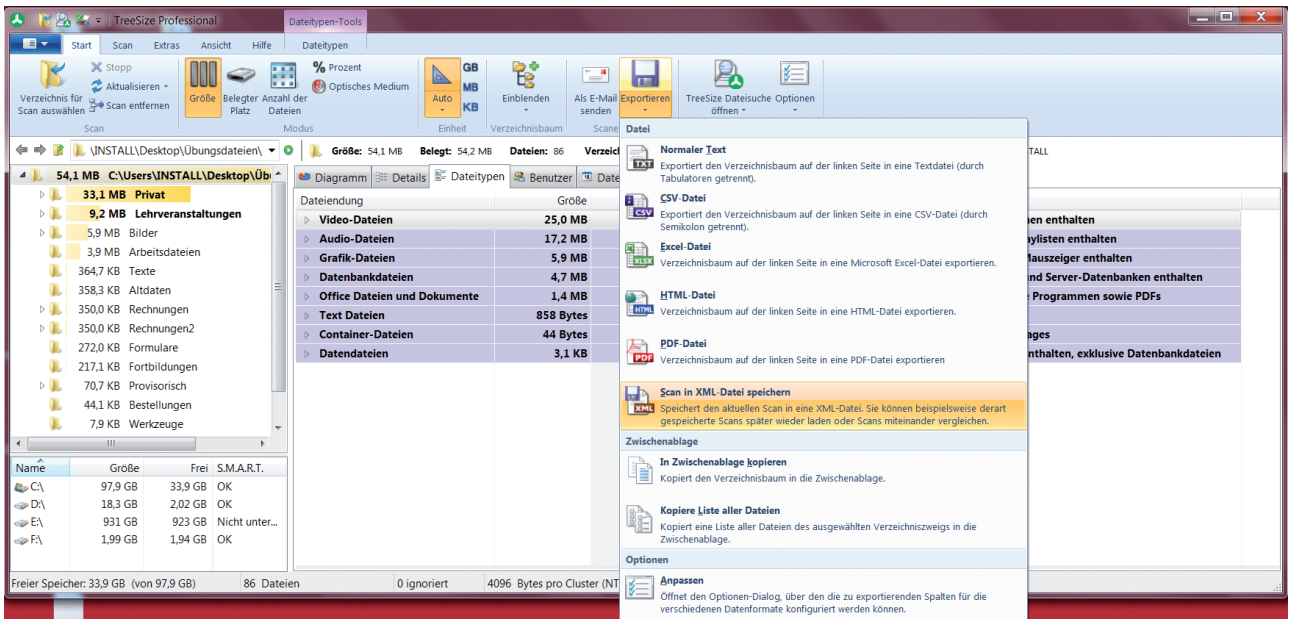


Abb. 5: Scan exportieren

A1 TreeSize Professional Bericht, 28.04.2017 11:14

	A	B	C	D	E	F	G	H	I
1	**TreeSize Professional Bericht, 28.04.2017 11:14**								
2	**C:\Users\INSTALL\Desktop\Übungsdateien\ auf [Windows]**								
3	**Laufwerk: C:\ Größe: 97,9 GB Belegt: 64,0 GB Frei: 33,9 GB 4096 Bytes pro Cluster (NTFS)**								
4									
5	**Absoluter Pfad**	**Größe**	**Belegt**	**Dateien**	**Verzeichnisse**	**Prozent**	**Letzte Änderung**	**Letzter Zugriff**	**Besitzer**
6	C:\Users\INSTALL\Desktop\Übungsdateien\	58,3 MB	58,4 MB	87	63	100,0%	28.04.2017	28.04.2017	INSTALL
7	C:\Users\INSTALL\Desktop\Übungsdateien\Privat\	33,1 MB	33,2 MB	10	6	56,9%	17.11.2016	16.02.2017	INSTALL
8	C:\Users\INSTALL\Desktop\Übungsdateien\Bilder\	10,0 MB	10,1 MB	14	5	17,2%	28.04.2017	28.04.2017	INSTALL
9	C:\Users\INSTALL\Desktop\Übungsdateien\Lehrveranstaltungen\	9,2 MB	9,2 MB	5	1	15,7%	17.11.2016	16.02.2017	INSTALL
10	C:\Users\INSTALL\Desktop\Übungsdateien\Arbeitsdateien\	3,9 MB	3,9 MB	4	0	6,7%	17.11.2016	16.02.2017	INSTALL
11	C:\Users\INSTALL\Desktop\Übungsdateien\Texte\	364,7 KB	368,0 KB	2	0	0,6%	17.11.2016	16.02.2017	INSTALL
12	C:\Users\INSTALL\Desktop\Übungsdateien\Altdaten\	358,3 KB	368,0 KB	4	0	0,6%	17.11.2016	16.02.2017	INSTALL
13	C:\Users\INSTALL\Desktop\Übungsdateien\Rechnungen\	350,0 KB	364,0 KB	16	18	0,6%	16.02.2017	16.02.2017	INSTALL
14	C:\Users\INSTALL\Desktop\Übungsdateien\Rechnungen2\	350,0 KB	364,0 KB	16	18	0,6%	16.02.2017	16.02.2017	INSTALL
15	C:\Users\INSTALL\Desktop\Übungsdateien\Formulare\	272,0 KB	272,0 KB	1	0	0,5%	17.11.2016	16.02.2017	INSTALL
16	C:\Users\INSTALL\Desktop\Übungsdateien\Fortbildungen\	217,1 KB	228,0 KB	5	0	0,4%	17.11.2016	16.02.2017	INSTALL
17	C:\Users\INSTALL\Desktop\Übungsdateien\Provisorisch\	70,7 KB	76,0 KB	3	1	0,1%	17.11.2016	16.02.2017	INSTALL
18	C:\Users\INSTALL\Desktop\Übungsdateien\Bestellungen\	44,1 KB	48,0 KB	2	0	0,1%	17.11.2016	16.02.2017	INSTALL
19	C:\Users\INSTALL\Desktop\Übungsdateien\Werkzeuge\	7,9 KB	8,0 KB	1	0	0,0%	17.11.2016	16.02.2017	INSTALL
20	C:\Users\INSTALL\Desktop\Übungsdateien*.*	3,1 KB	4,0 KB	1	0	0,0%	17.11.2016	16.02.2017	
21	C:\Users\INSTALL\Desktop\Übungsdateien\Kontaktdaten\	333,0 Bytes	12,0 KB	3	0	0,0%	17.11.2016	16.02.2017	INSTALL
22									

Abb. 6: Exportergebnis eines Scans als Excel-Datei

Zudem ermöglicht die Funktion „Mit gespeichertem Scan vergleichen“ (Abb. 7) einen automatischen Abgleich zweier zu unterschiedlichen Zeiten ausgeführten Bestandsaufnahmen. Voraussetzung hierfür ist, dass der Vergleichsscan im XML-Format gespeichert wurde. Im vorliegenden Beispiel (Abb. 8) wurde eine zusätzliche Bild-Datei in den Ordner „Bilder“ geschoben. Nach dem Abgleich beider Scans zeigt TSP im Verzeichnis alle Veränderungen rot eingefärbt an. In diesem Fall enthält die Fileablage im entsprechenden Ordner eine zusätzliche Datei und zusätzliches Speichervolumen in Höhe von 4,2 MB.

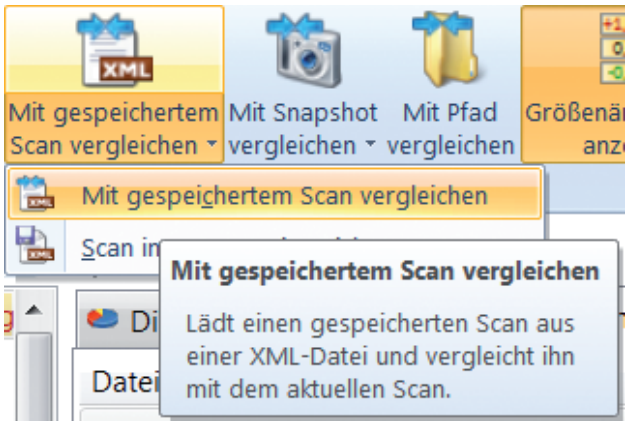


Abb. 7: Scanvergleich starten

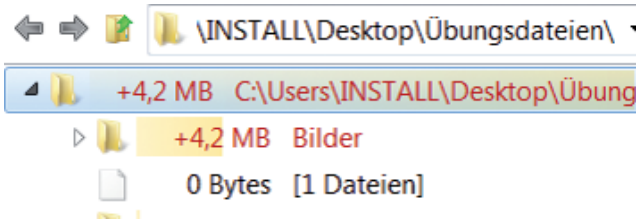


Abb. 8: Unterschiede des Scanvergleichs im Verzeichnis

3. Total Commander (TC)

TC ist ein Tool zur Dateiverwaltung und stellt eine Alternative zum Windows Explorer dar. Augenscheinlichster Unterschied zu Letzterem ist das zweite Ansichtsfenster. Dateien können somit per Drag & Drop schnell verschoben und mit zwischen den Fenstern befindlichen Funktionstasten bearbeitet werden. Dies kann bei der manuellen Formierung von SIPs hilfreich sein. Weitere wichtige Funktionen im Rahmen des Preingests sind die virtuelle Entfernung von Ordnerstrukturen und das Mehrfach-Umbenenn-Tool.

3.1 Ansicht und Navigation

Im linken Fenster wird die Verzeichnisstruktur eingelesen, während das rechte Fenster ermöglicht, hier „künstliche“ Strukturobjekte zu bilden (Abb. 9). Per Drag & Drop können Dateien kopiert oder verschoben und neue Ordner angelegt werden.

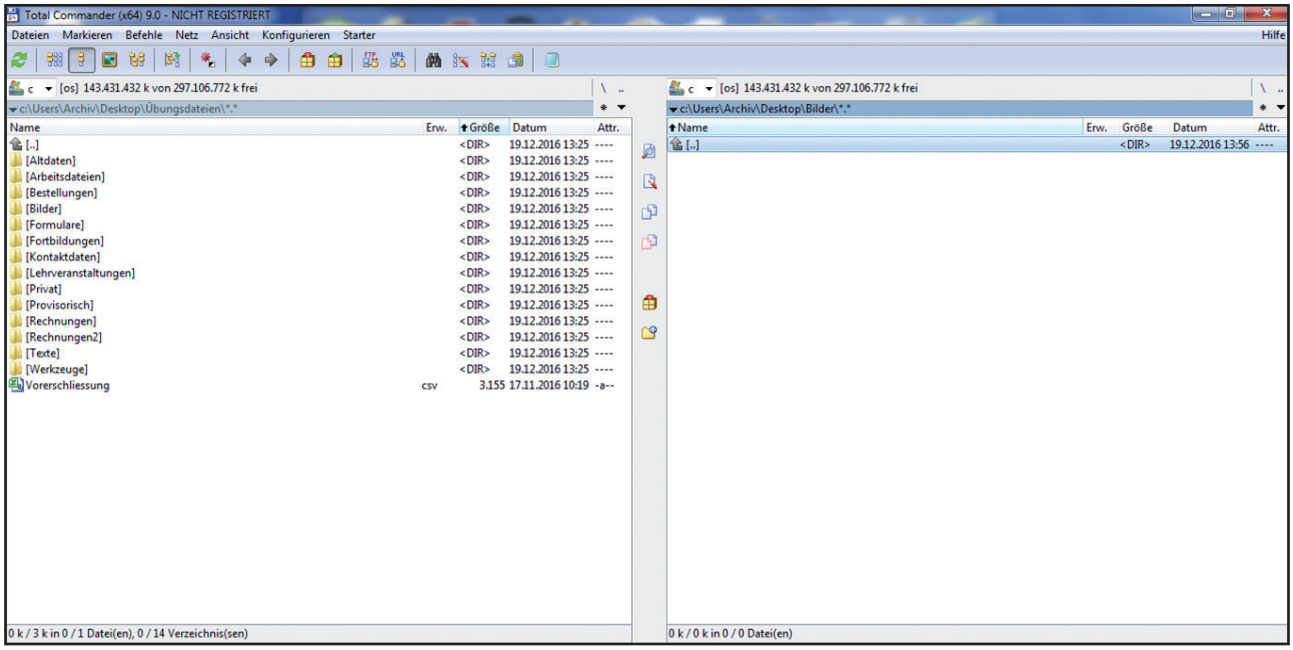


Abb. 9: Links: Originalverzeichnis; rechts: Fenster zum Sortieren/Formieren von SIP

3.2 Strukturobjekte analysieren: Alle Dateien einer digitalen Ablage anzeigen

Über die Schaltfläche „Zweigansicht: Alle Dateien in allen Unterverzeichnissen“ (Abb. 10) wird die Ordnerstruktur virtuell abgebaut und alle in der digitalen Ablage befindlichen Dateien werden als Liste angezeigt (Abb. 11). Diese Liste kann im Folgenden nach bestimmten Metadaten, z.B. nach Dateitypen oder Größe, sortiert und gefiltert werden.

Bild 10: Schaltfläche: Verzeichnisse virtuell abbauen

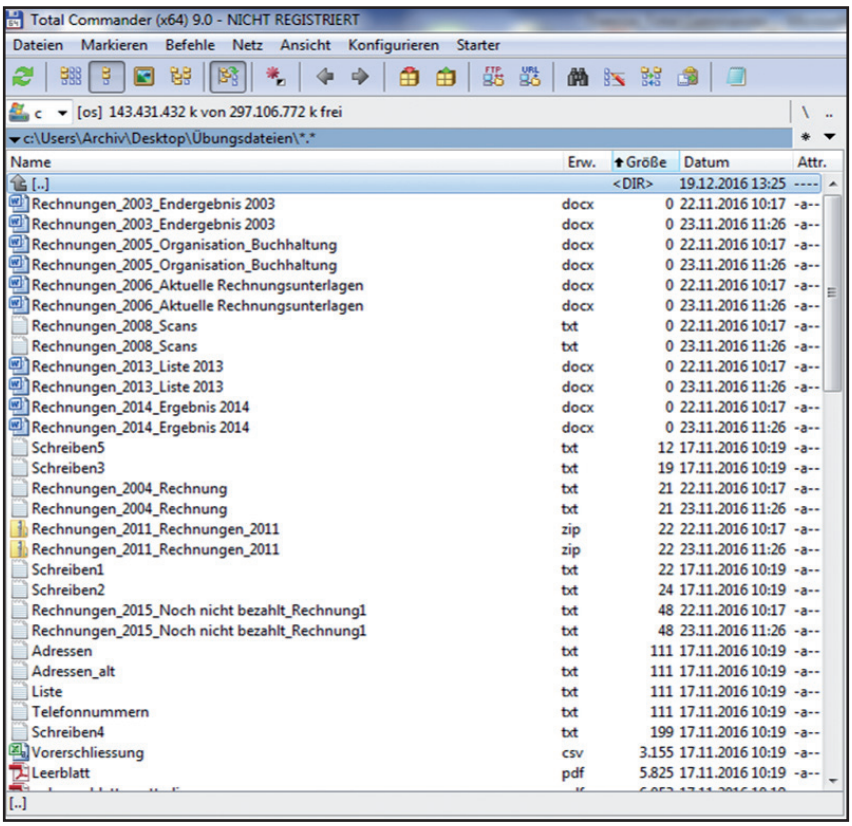


Abb. 11: Alle Dateien in der digitalen Ablage werden angezeigt

3.3 Normalisierung von Objektnamen: Mehrfach-Umbenenn-Tool

Mit diesem Tool ermöglicht TC z. B. die Eliminierung von Umlauten oder Sonderzeichen. Gleichzeitig können aber auch weitere Informationen aus den Metadaten mit in den Dateinamen aufgenommen werden.

Beispiel Dateipfad: In einer digitalen Ablage befindet sich an einer Stelle der Ordner „Rechnungen“. Alle Dateien aus der darunterliegenden Verzeichnisstruktur sollen künftig ein AIP sein (z. B. IO 1: Rechnungen). Mit vorangegangener Aktion aus 3.2 erhält man eine Liste aller vorhandenen Dateien des Verzeichnisses. Alle Dateien sollen den Pfad ab dem Ordner „Rechnungen“ im Dateinamen tragen.

Im ersten Schritt (Abb. 13) wird definiert, dass der neue Dateiname aus dem Pfad [=tc Pfad], dem Dateinamen [N] und der Erweiterung [E] bestehen soll. Zugleich werden alle Sonderzeichen „\“ durch „_“ ersetzt. Im unteren Feld wird direkt der künftige Dateiname angezeigt. Mit „Start!“ werden diese Dateinamen verändert.

In einem zweiten Schritt (Abb. 14) wird der neue Name [N] verändert, indem im Feld „Suchen nach“ der vordere Teil des Pfades, d.h. die dem Ordner „Rechnungen“ übergeordnete Verzeichnisstruktur, eingegeben wird. Da das Feld „Ersetzen durch“ freigelassen wird, wird dieser Teil des Pfades gelöscht.

Die ursprünglich zum Teil wenig aussagekräftigen Dateinamen beinhalten im vorliegenden Beispiel (Abb. 15) nun die Informationen, dass die Dateien aus dem Ordner „Rechnungen“, den nach Rechnungsjahren benannten Ordnern und den weiteren darunter befindlichen Verzeichnissen stammen.

Auf der Ebene der Repräsentation 1 (ursprüngliches Abgabeformat) wird der Dateiname wohl in den meisten Fällen aus Gründen der Authentizität unverändert bleiben. Es kann

jedoch aus verschiedensten Gründen sinnvoll sein, Dateien, die im Rahmen von Preingest-Workflows aufbereitet und konvertiert werden, mit dem vorgestellten Tool umzubenennen und zusätzliche Informationen aus den Metadaten in den Dateinamen aufzunehmen.

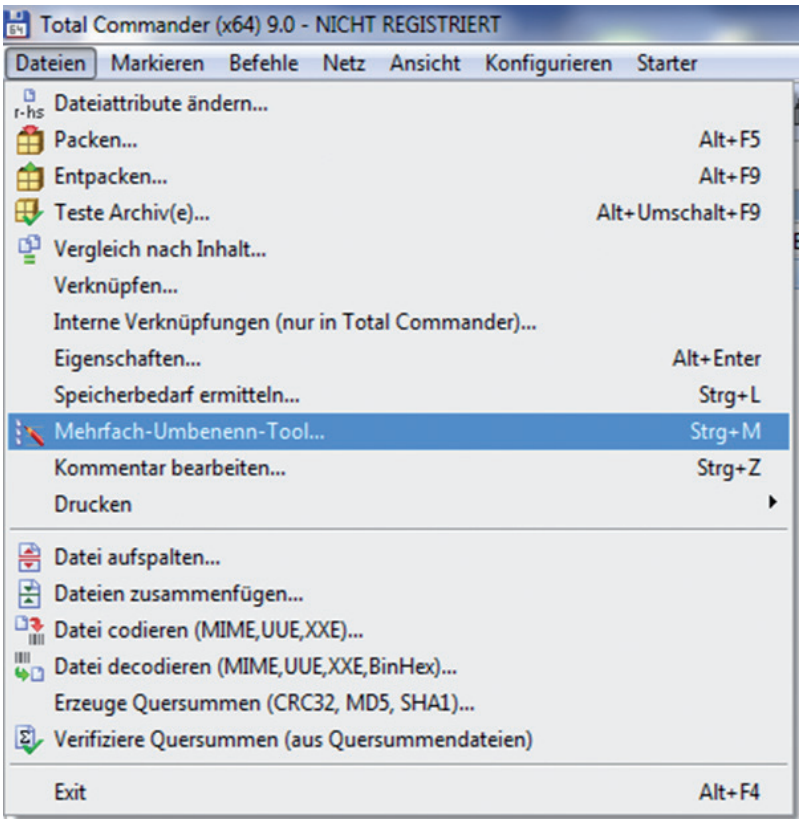


Abb. 12: Über die Menüleiste „Dateien“ lässt sich das Mehrfach-Umbenenn-Tool starten

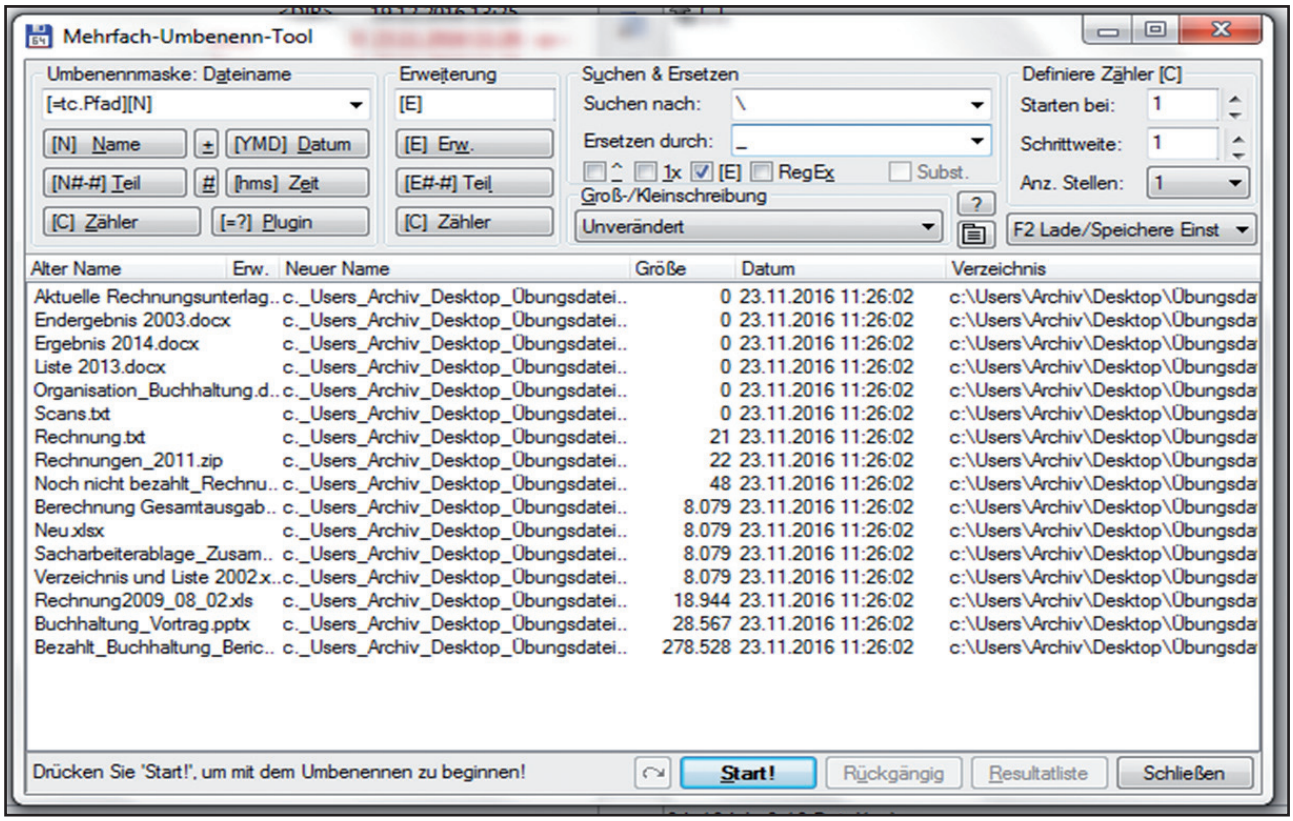


Abb. 13: Mehrfach-Umbenenn-Tool: Schritt 1 – Übernahme des Pfades in den Dateinamen

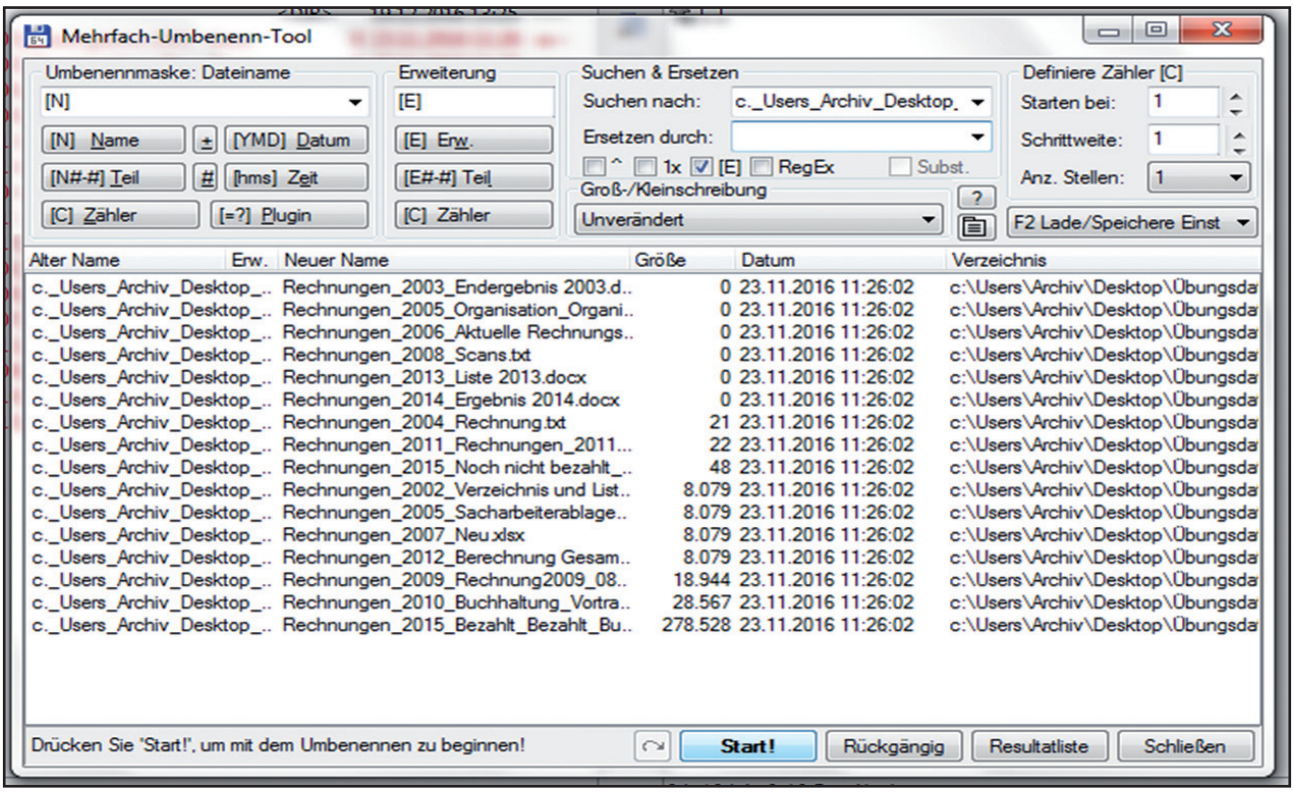


Abb. 14: Mehrfach-Umbenenn-Tool: Schritt 2 – Eliminierung des übergeordneten Pfad-Teils

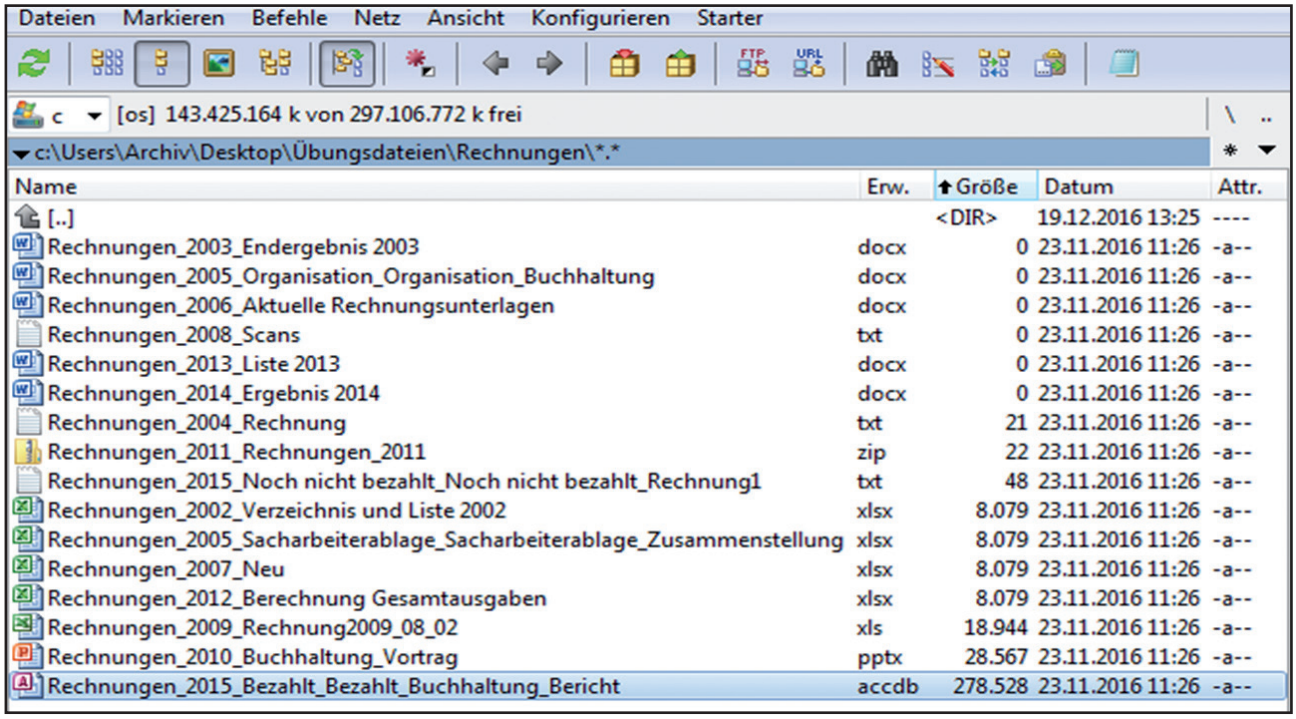


Abb. 15: Ansicht aller umbenannten Dateien des künftigen AIP

3.4 SIPs formieren: Drag & Drop

Unter 3.2. wurde gezeigt, wie mit einem Mausklick Verzeichnisstrukturen virtuell abgebaut werden können. Über das zweite Fenster können nun alle Dateien eines Verzeichnisses oder nur Dateien mit bestimmten Eigenschaften zu SIPs gruppiert werden.

Beispiel Fotodokumentation: Es sollen SIPs einer umfangreichen Fotodokumentation zu einem stillgelegten Asylbewerberheim formiert werden. Dabei werden mehrere Bilder eines Gebäudeteils zu einem SIP (z.B. IO 1 = Außenanlage, IO 2 = Küche etc.). Nachdem im Zielverzeichnis die den AIPs entsprechenden Ordner angelegt wurden, können alle zusammengehörigen Bilddateien per Drag & Drop in den entsprechenden Ordner gezogen werden. Jeder Ordner ist nun ein SIP und kann als ein IO in das digitale Archiv übertragen werden.

4. Musterworkflow

Da die Tools direkt in die Verzeichnisstruktur eingreifen, sollte, wie eingangs erwähnt, immer eine Kopie der Fileablage erstellt werden. TSP ermöglicht zunächst eine grobe Analyse der Verzeichnisstruktur und kann anschließend im Rahmen einer Nachbewertung Dubletten und obsolete Daten eliminieren. TC kann durch den einfachen Abbau der Ordnerstruktur zur weiteren Analyse der Strukturobjekte herangezogen werden. Mit dem Mehrfach-Umbenenn-Tool können anschließend Objektnamen normalisiert und weitere Informationen in den Dateinamen aufgenommen werden. Schließlich ist über die Zwei-Fenster-Ansicht eine einfache SIP-Formierung per Drag & Drop möglich. Die Tools unterstützen keine Konvertierung oder Validierung von Formaten. Diese Schritte wären bei Bedarf durch andere Tools zu gewährleisten.

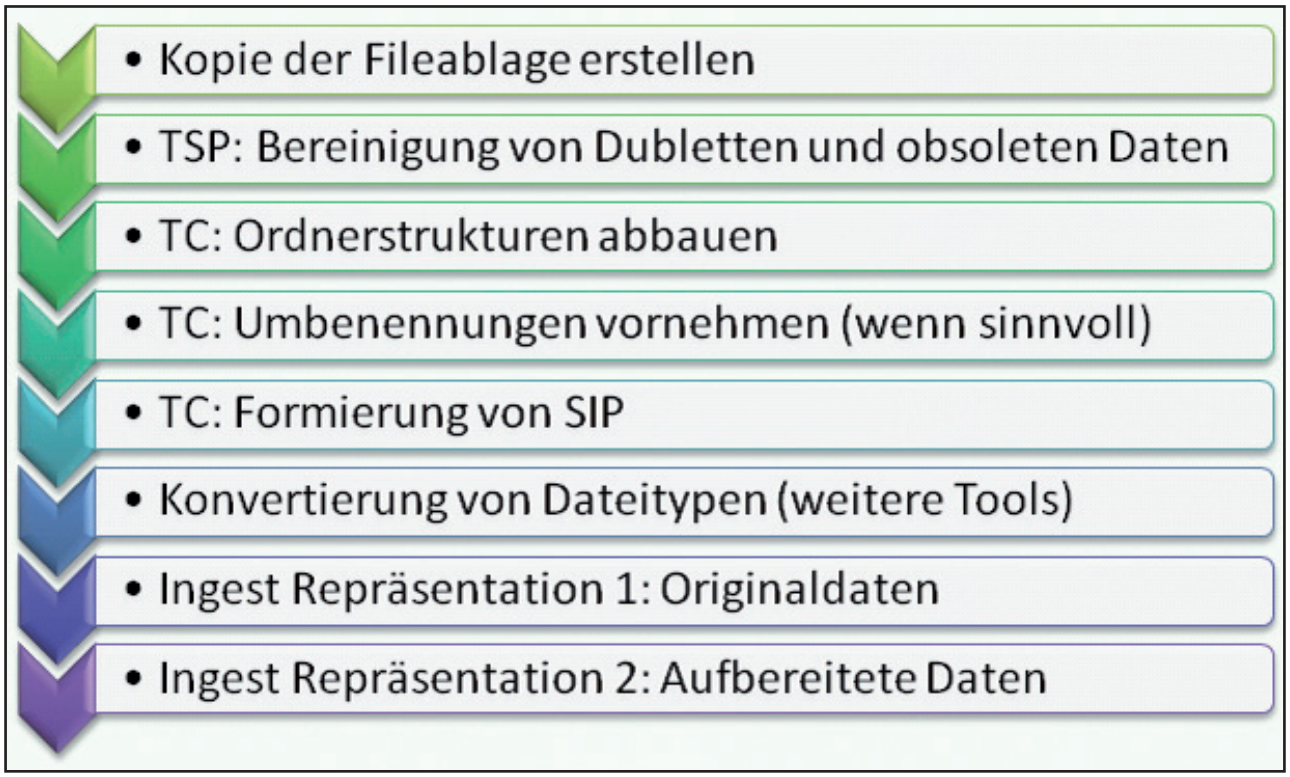


Abb. 16: Ein möglicher Preingestworkflow unter Einbeziehung von TSP und TC

Bytebarn – Datenbanklösung des Sächsischen Staatsarchivs zur Archivierung von Dateiverzeichnissen

Karsten Huth, Peter Bayer

Ausgangslage: Probleme mit kreativen digitalen Ablagen

Digitale Archivierung versucht, die unendlich vielen möglichen Formen, die der digitale Alltag produziert, auf ein überschaubares Maß zu reduzieren, um sie für unbegrenzte Zeit in einem nutzbaren Zustand zu erhalten. Die im Alltag am häufigsten auftretende Form einer kreativen digitalen Ablage ist das Dateienverzeichnis. Jeder, der mit Computern arbeitet, nutzt Verzeichnisse zur spontanen Ablage seiner Dateien. Sie sind intuitiv verständlich und unkompliziert zu handhaben. Fast alle gebräuchlichen Betriebssysteme bieten hierarchische Verzeichnissysteme an. Für die Digitale Archivierung ist der Umstand der immens weiten Verbreitung zunächst von Vorteil. Das Archiv weiß aus eigener praktischer Erfahrung, womit man es zu tun hat. Dennoch klagen Archive oft, wenn ihnen Datenträger mit Dateiverzeichnissen angeboten werden. Wo liegen die Probleme?

Problem 1 – Ausufernde Strukturen

Obwohl die meisten Verzeichnissysteme nur zwei Objekttypen kennen (Verzeichnisse und Dateien), können durch fast beliebige Verschachtelungen der Verzeichnisse sehr komplexe und schwer vorhersagbare Strukturen entstehen. Zusätzlich können Dateien auch noch als Container für weitere strukturelle Gebilde dienen (z.B. .zip oder .tar), die das gesamte Gebilde noch komplexer machen. Für Digitale Archive (DA) ergibt sich daraus ein Problem. Viele Ingestprozesse wollen am Anfang die eingehenden Pakete (SIP) auf ihre Gültigkeit hin prüfen. Die Gültigkeit eines SIP beruht meistens auf der Einhaltung einer fest vorgegeben Struktur (s. Abb. 1).

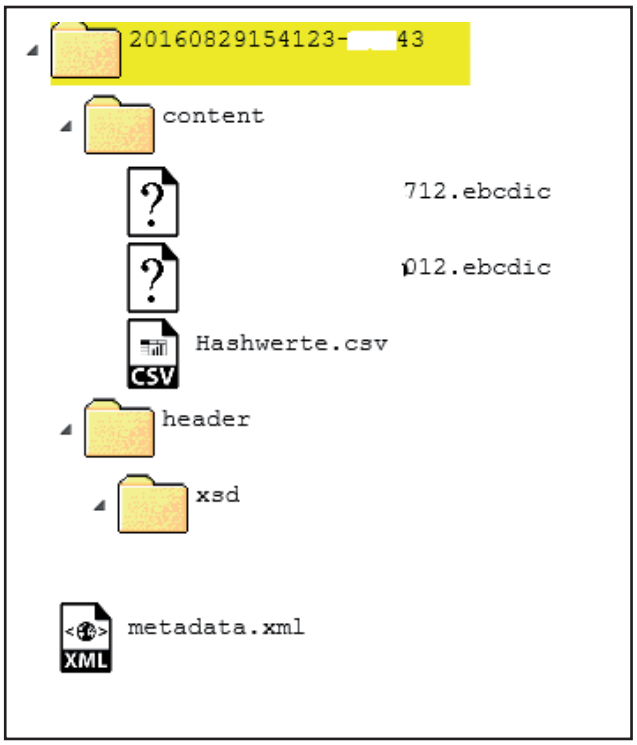


Abb. 1: Gültiges SIP nach Schweizer Standard eCH-0160

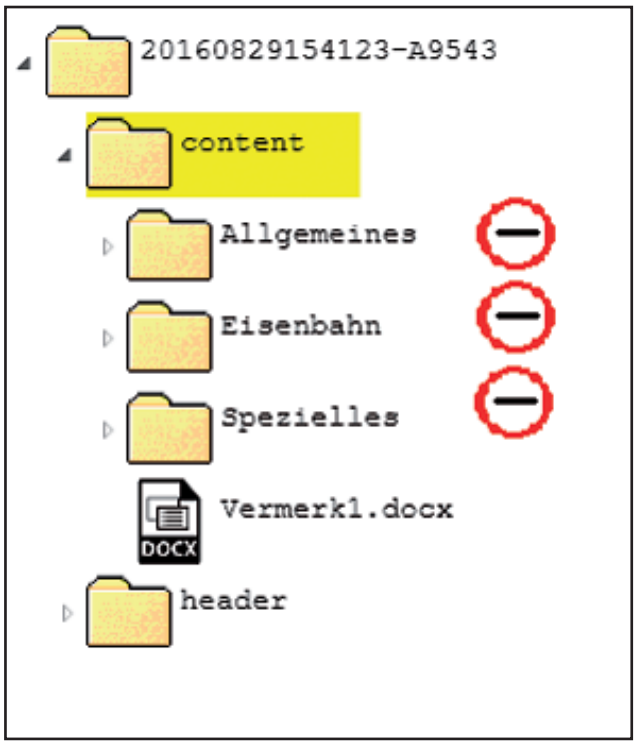


Abb. 2: Ungültiges SIP nach Schweizer Standard eCH-0160

Das Beispiel in Abbildung 1 zeigt die Struktur eines SIP-Standards, die vom Ingestprozess des Elektronischen Staatsarchivs zwingend in exakter Form erwartet wird. Jedes SIP muss auf der zweiten Ebene ein Verzeichnis „content“ und ein Verzeichnis „header“ enthalten. In „content“ liegen die eigentlichen Inhaltsdateien in flacher Reihung, die Gegenstand der Archivierung sind. In „header“ liegen die dazugehörigen Metadaten. Da alle Inhaltsdateien im Verzeichnis „content“ flach abgelegt werden müssen und keine weiteren Verzeichnisse enthalten dürfen, kann ein verschachteltes Dateiverzeichnis nicht über diesen Standard ingestiert werden (s. Abb. 2).

Immer wenn ein DA am Beginn eines Ingestprozesses die Gültigkeit eines SIP beurteilen soll, müssen vorher die Regeln und Strukturen, anhand derer gemessen wird, festgelegt werden. Das muss nicht zwingend der Schweizer Standard eCH-0160 sein, aber feste Regeln müssen sein. Werden diese Regeln und Strukturen vom eingehenden SIP eingehalten, ist es gültig, werden sie auch nur in einem Punkt übertreten, wird es abgelehnt und der Ingest verweigert.

Könnte man nicht einfach auf die Prüfung der Gültigkeit eines SIP verzichten, um dem Problem auszuweichen? Von den Prinzipien her, auf denen ein DA aufgebaut ist, leider nicht. Da das Ziel des Ingests immer möglichst gleichartige AIP im Archivspeicher sind, braucht man verlässliche SIP, auf denen die nächsten Prozessschritte des Ingests aufbauen können.

Ohne weitere technische Hilfsmittel ist man bei der Archivierung von Dateiverzeichnissen zwangsläufig dazu gezwungen, das zu übernehmende Dateiverzeichnis in passende Einzelteile zu zerlegen und auf gültige SIP zu verteilen. Bei dieser Prozedur wird die originale Ordnung der Dinge verändert. Die einzige Chance sie zu erhalten wäre, sie vor dem Ingest zu dokumentieren. Beides (Aufteilung des Verzeichnisses in SIP und Dokumentation der originalen Struktur) erfordert einen hohen Aufwand, den man bei der Archivierung von Dateiverzeichnissen sicher gerne vermeiden würde.

Problem 2 – Die große Format-Wundertüte

Dateiverzeichnisse machen in der Regel keine Vorgaben darüber, welche Dateiformate abgelegt werden dürfen. DA hingegen versuchen, möglichst nur vorher definierte Dateiformate für bestimmte Objekttypen zu übernehmen und dauerhaft zu archivieren. Ähnlich wie bei der ausufernden Struktur wird auch hier die große Freiheit bei der Erstellung des Dateiverzeichnisses bei der Archivierung zu einem Problem. Um die Vorgaben eines DA bei den vorher definierten Dateiformaten zu erreichen, müssten die Dateiverzeichnisse entweder bei der Lieferung nur die vom Archiv erlaubten Formate enthalten, oder sie müssten zumindest in einem Zustand sein, von dem aus eine Konvertierung der Dateien in den erwünschten Zustand möglich wäre. Beides ist fast nie der Fall, so dass Dateiverzeichnisse eigentlich nie von sich aus den Zustand der Archivfähigkeit erreichen. Ohne geeignetes Werkzeug müsste auch hier von Hand gesichtet und gegebenenfalls auch konvertiert werden.

Problem 3 – Keine Metadaten

Metadaten sind unverzichtbar für ein DA. Metadaten werden benötigt, um die Archivalien wieder aufzufinden (inhaltsbeschreibende Metadaten). Metadaten werden benötigt, um eine Übersicht über die Dateiformate und die nötigen technischen Umgebungen bei der Benutzung zu gewährleisten. Sie sind unverzichtbar bei der Aufrechterhaltung der technischen Integrität der Daten (technische Metadaten). Gerade die inhaltsbeschreibenden Metadaten eines Dateiverzeichnisses sind oft spärlich ausgeprägt. Außer den Namen der Verzeichnisse und der Dateien wird oft nichts weiter Verwertbares an das DA übertragen. Bei den

technischen Metadaten sieht es oft nicht besser aus. Das bedeutet, dass auch hier viel von Hand nachgearbeitet werden müsste.

Problem 4 – Alles ist Format, auch das Verzeichnis selbst

Auch wenn alle Dateiverzeichnisse an der Oberfläche zum Nutzer hin ähnlich bis gleich anmuten, sind sie technisch doch bisweilen unterschiedlich realisiert und nur bedingt miteinander kompatibel. Sie sind technisch abhängig von dem Betriebssystem, in dem sie laufen. Damit kann ein Dateiverzeichnis genauso Opfer der technologischen Obsoleszenz werden wie jedes andere Dateiformat auch.

Problem 5 – Das Intranet/Netzwerk des Archivs, der falsche Freund

Manchmal werden schnelle Lösungen gebraucht. Was liegt da bei angebotenen Dateiverzeichnissen näher, als sie in das Verwaltungsnetzwerk des Archivs hinein zu kopieren. Die Struktur kann damit zunächst erhalten werden. Das Problem der Formatvielfalt wird vertagt. Metadaten sind spärlich, zum Wiederauffinden dient letztlich eine Pfadangabe (die oft nicht lange gültig bleibt). Verwaltungsdaten für den täglichen Gebrauch und digitales Archivgut verschmelzen in einem System miteinander. Wird das Netzwerk technisch verändert, müssen alle archivierten Verzeichnisse mit migriert werden. Es gibt bei dieser Lösung kaum Wege, die technische Integrität der Dateien und des Verzeichnisses nachzuweisen. Aus der Not geboren, gibt es manchmal keine andere Möglichkeit. Auf Dauer ist die Sicherung von Archivgut in einem Verwaltungsnetzwerk sicher keine gute Lösung.

Bytebarn

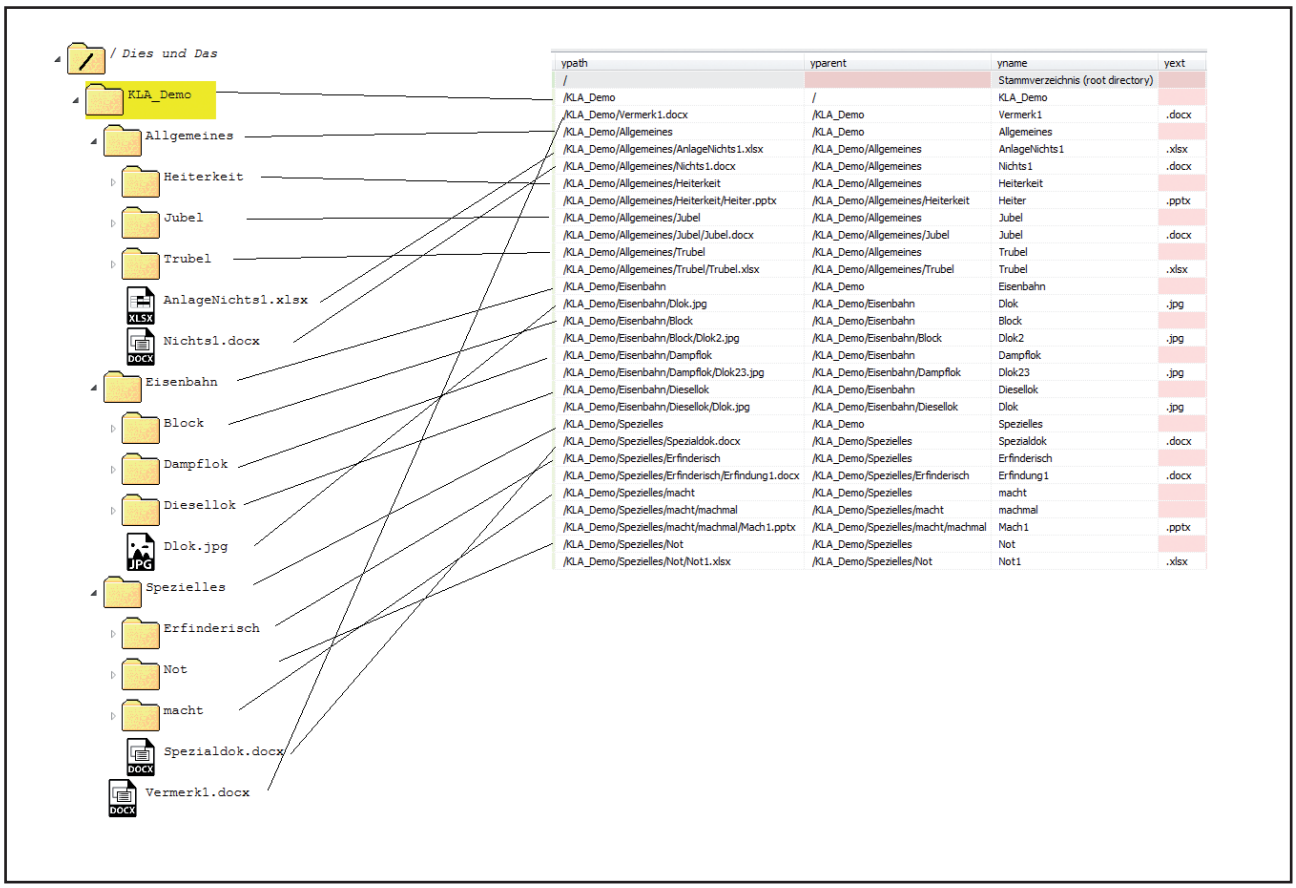
Aufgrund der oben geschilderten Probleme bei der Archivierung von Dateiverzeichnissen wurde im Sächsischen Staatsarchiv die Anwendung Bytebarn entwickelt und im Oktober 2014 mit der ersten Version in Betrieb genommen. Bytebarn basiert auf der Datenbank SQLite, die weltweit im Hintergrund vieler Systeme arbeitet. Bytebarn migriert das Dateiverzeichnis in eine einzelne Datenbanktabelle im SQLite-Format. In dieser Tabelle sind sowohl die originale Verzeichnisstruktur mit allen originalen Dateinamen und den Dateiformaten hinterlegt als auch die Dateien selber (s. Abb. 3).

Mit dieser Migration des kompletten Dateiverzeichnisses in eine Datenbank löst Bytebarn am Sächsischen Staatsarchiv mehrere Probleme, die mit der Archivierung von Dateiverzeichnissen verbunden waren.

Lösung zu 1 – Erhaltung der komplexen Verzeichnisstrukturen bei gleichzeitiger Vereinfachung des SIP

Mit der Verlegung des kompletten Dateiverzeichnisses, welches aus vielen Einzelobjekten besteht, in eine Datenbank, die nur durch eine Datei repräsentiert wird, ist es dem Sächsischen Staatsarchiv möglich, standardisierte überprüfbare SIP zu erzeugen (s. Abb. 4).

Aus einem komplizierten, weitverzweigten Konstrukt wird im SIP eine einzige Datei, die die gesamte Komplexität beinhaltet. Eine vorherige Aufbereitung des Dateiverzeichnisses von Hand ist nicht nötig.



ypath	yparent	yname	yext
/		Stammverzeichnis (root directory)	
/KLA_Demo	/	KLA_Demo	
/KLA_Demo/Vermerk1.docx	/KLA_Demo	Vermerk1	.docx
/KLA_Demo/Allgemeines	/KLA_Demo	Allgemeines	
/KLA_Demo/Allgemeines/AnlageNichts1.xlsx	/KLA_Demo/Allgemeines	AnlageNichts1	.xlsx
/KLA_Demo/Allgemeines/Nichts1.docx	/KLA_Demo/Allgemeines	Nichts1	.docx
/KLA_Demo/Allgemeines/Heiterkeit	/KLA_Demo/Allgemeines	Heiterkeit	
/KLA_Demo/Allgemeines/Heiterkeit/Heiter.pptx	/KLA_Demo/Allgemeines/Heiterkeit	Heiter	.pptx
/KLA_Demo/Allgemeines/Jubel	/KLA_Demo/Allgemeines	Jubel	
/KLA_Demo/Allgemeines/Jubel/Jubel.docx	/KLA_Demo/Allgemeines/Jubel	Jubel	.docx
/KLA_Demo/Allgemeines/Trubel	/KLA_Demo/Allgemeines	Trubel	
/KLA_Demo/Allgemeines/Trubel/Trubel.xlsx	/KLA_Demo/Allgemeines/Trubel	Trubel	.xlsx
/KLA_Demo/Eisenbahn	/KLA_Demo	Eisenbahn	
/KLA_Demo/Eisenbahn/Dlok.jpg	/KLA_Demo/Eisenbahn	Dlok	.jpg
/KLA_Demo/Eisenbahn/Block	/KLA_Demo/Eisenbahn	Block	
/KLA_Demo/Eisenbahn/Block/Dlok2.jpg	/KLA_Demo/Eisenbahn/Block	Dlok2	.jpg
/KLA_Demo/Eisenbahn/Dampflok	/KLA_Demo/Eisenbahn	Dampflok	
/KLA_Demo/Eisenbahn/Dampflok/Dlok23.jpg	/KLA_Demo/Eisenbahn/Dampflok	Dlok23	.jpg
/KLA_Demo/Eisenbahn/Diesellok	/KLA_Demo/Eisenbahn	Diesellok	
/KLA_Demo/Eisenbahn/Diesellok/Dlok.jpg	/KLA_Demo/Eisenbahn/Diesellok	Dlok	.jpg
/KLA_Demo/Spezielles	/KLA_Demo	Spezielles	
/KLA_Demo/Spezielles/Spezialdok.docx	/KLA_Demo/Spezielles	Spezialdok	.docx
/KLA_Demo/Spezielles/Erfinderisch	/KLA_Demo/Spezielles	Erfinderisch	
/KLA_Demo/Spezielles/Erfinderisch/Erfindung1.docx	/KLA_Demo/Spezielles/Erfinderisch	Erfindung1	.docx
/KLA_Demo/Spezielles/macht	/KLA_Demo/Spezielles	macht	
/KLA_Demo/Spezielles/macht/machmal	/KLA_Demo/Spezielles/macht	machmal	
/KLA_Demo/Spezielles/macht/machmal/Mach1.pptx	/KLA_Demo/Spezielles/macht/machmal	Mach1	.pptx
/KLA_Demo/Spezielles/Not	/KLA_Demo/Spezielles	Not	
/KLA_Demo/Spezielles/Not/Not1.xlsx	/KLA_Demo/Spezielles/Not	Not1	.xlsx

Abb. 3: Importieren eines Dateiverzeichnis in eine SQLite Tabelle

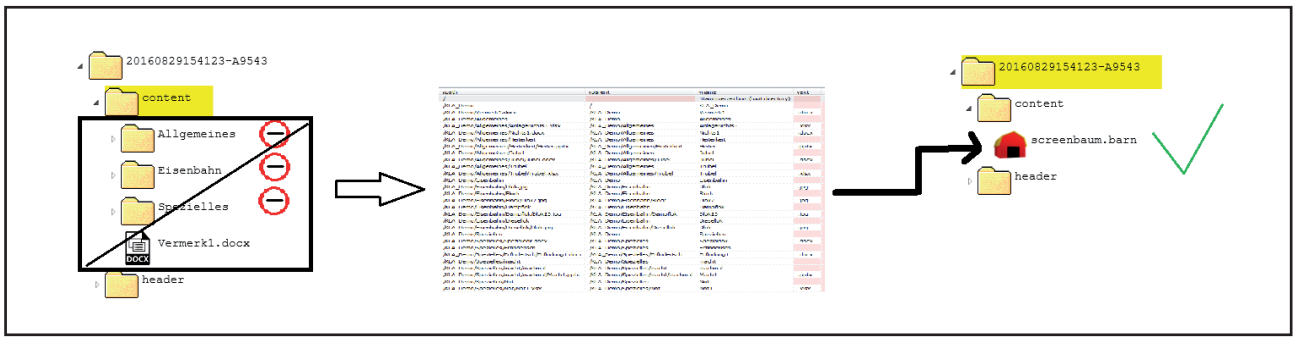


Abb. 4: Integration einer Barn-Datei in ein gültiges SIP nach Standard eCH-0160

Lösung zu 2 – Dokumentation der Formate innerhalb Bytebarn

Beim Import des Dateiverzeichnisses in die Bytebarn Datentabelle wird eine vereinfachte Formaterkennungsprozedur anhand der DROID-Signature File (XML) durchgeführt. Um auch bei großen Dateiverzeichnissen und bei Containerformaten einen zügigen Import zu gewährleisten, werden solche Dateien zunächst nur ausnahmsweise anhand ihrer Dateiendung identifiziert. Sollte weiterer Bedarf oder berechtigter Zweifel an der Korrektheit des erkannten Formats bestehen, bietet Bytebarn die Möglichkeit, die Erkennung über das Werkzeug „Siegfried“[1] durchführen zu lassen (s. Abb. 5).

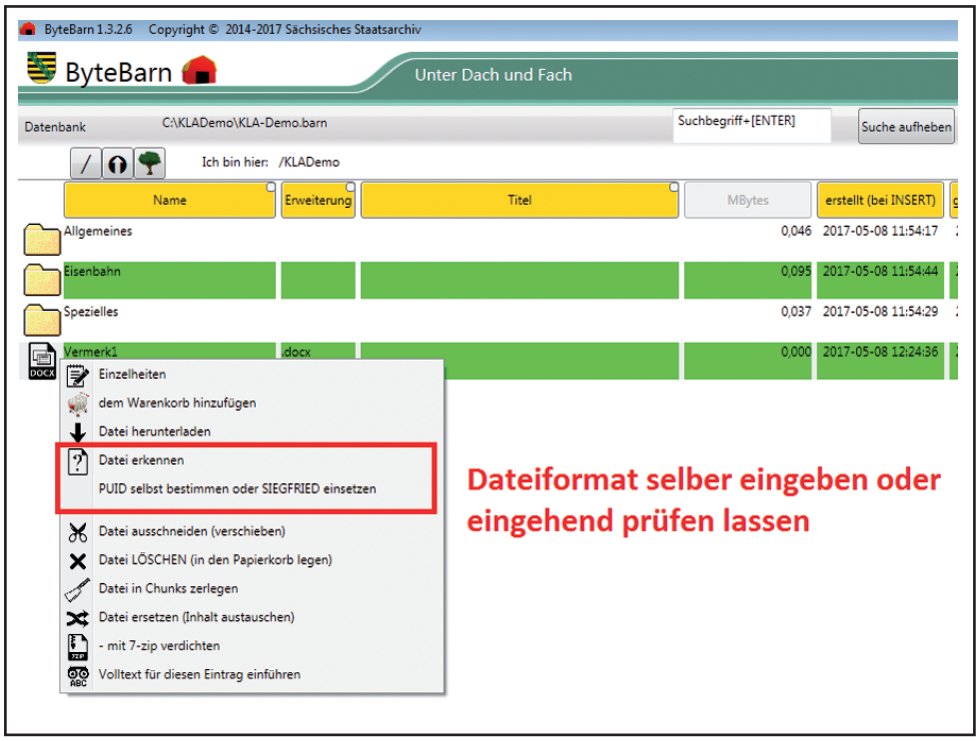


Abb. 5: Menu zur Durchführung einer Formatkontrolle mit dem externen Werkzeug „Siegfried“

Da Bytebarn eine Datenbank ist, können natürlich Abfragen über den Inhalt durchgeführt werden. Eine Statistik über die vorhandenen Dateiformate ist auf Wunsch jederzeit verfügbar. (s. Abb. 6)

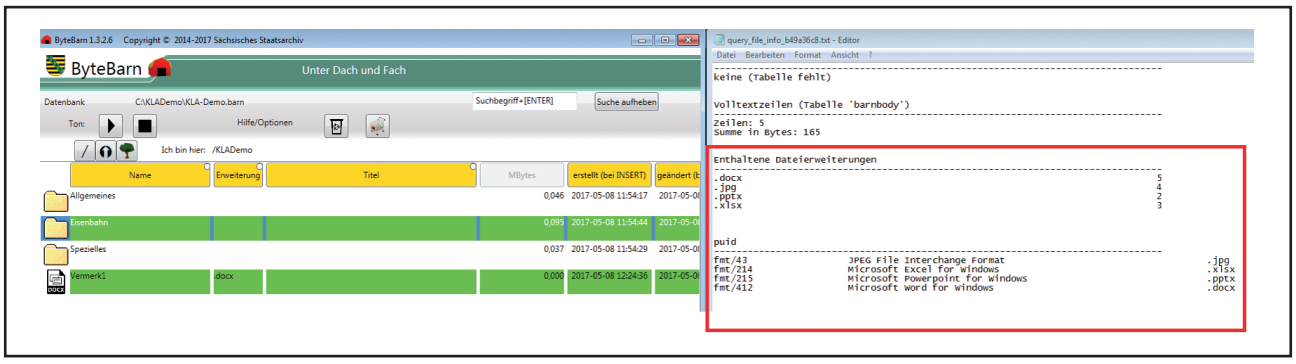


Abb. 6: Übersicht über die in Bytebarn enthaltenen Dateiformate

[1] Richard Lehane, Siegfried is a signature-based file format identification tool *http://www.itforarchivists. com* (aufgerufen am 10.5.2017).

Besteht zusätzlich eine Anbindung von Bytebarn an das Office-Programm LibreOffice können bei Bedarf während des Imports Office-Dateien in PDF/A konvertiert werden. Die originalen Dateien bleiben erhalten und der Austausch der Dateien wird in der Datenbank dokumentiert und ist dauerhaft nachvollziehbar (s. Abb. 7). Bei einem Import sehr exotischer Dateiformate, die nicht im DROID-Signature File aufgeführt sind, kann die automatische Formaterkennung keine Ergebnisse liefern. Falls einem das Format der Dateien bekannt ist, kann die Formatinformation von Hand eingetragen werden.

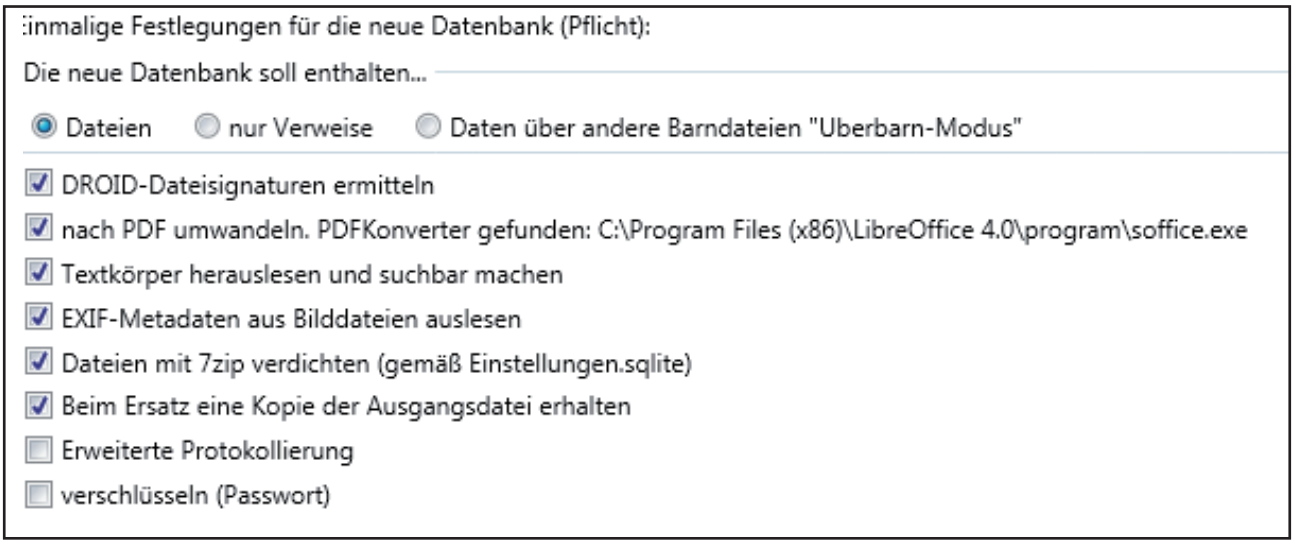


Abb. 7: Optionsmenü für das Neu-Erstellen einer ByteBarn-Datenbank

Lösung zu 3 – Aufwertung durch Metadaten

Ist ein Dateiverzeichnis in Bytebarn importiert, kommen die Vorteile einer Datenbank beim Thema Metadaten voll zur Entfaltung. Während des Imports versucht Bytebarn vorhandene Texte aus den Dateien zu extrahieren, um einen Volltextindex zu erstellen. Bilddateien mit EXIF-Metadaten können ebenfalls indexiert werden. Der Erfolg der Indexierung ist immer abhängig von der Qualität des Ausgangsmaterials. Bei Dateien mit Beschädigungen oder Bilddateien ohne Metadaten fällt das Ergebnis entsprechend dürftig aus. Bei gutem Ausgangsmaterial erhält man hingegen eine voll durchsuchbare Textdatenbank (s. Abb. 8). Nach dem Import des Dateiverzeichnisses können inhaltsbeschreibende Metadaten für jede Datei oder jede Verzeichnisebene hinzugefügt werden (s. Abb. 9). Technische Metadaten werden von Bytebarn automatisiert erfasst (s. Abb. 10).

Wie bereits erwähnt werden Angaben zum Dateiformat automatisch erhoben. Bei einem bekannten Format werden Empfehlungen für ein Darstellungsprogramm gegeben. Zur Prüfbarkeit der Integrität wird ein MD5 Hashwert abgelegt.

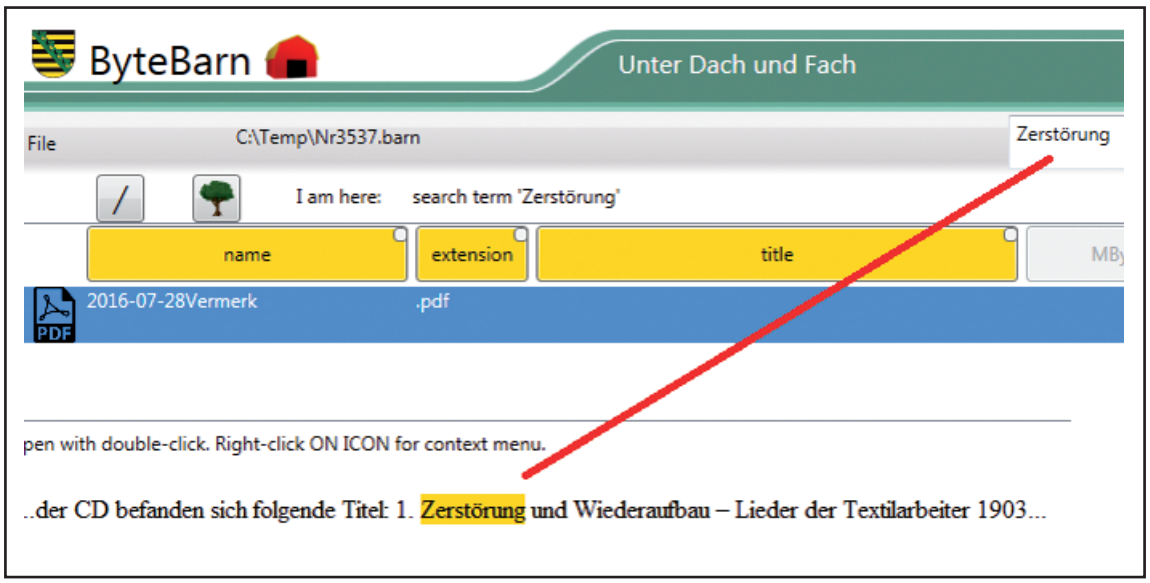


Abb. 8: Gefundene Textstelle bei Volltextrecherche

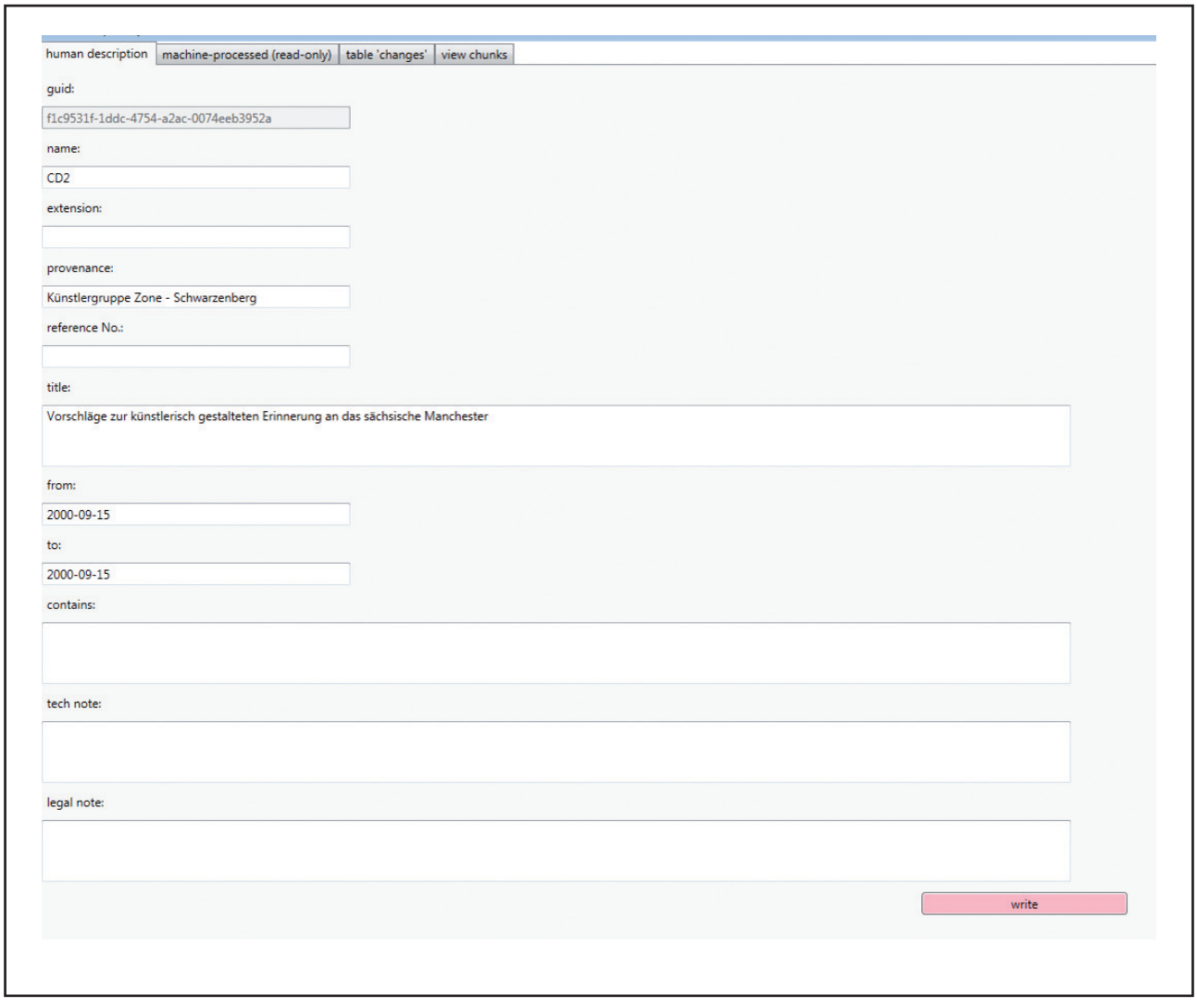


Abb. 9: Frei editierbare Inhaltsmetadaten auf Ebene eines Verzeichnisses

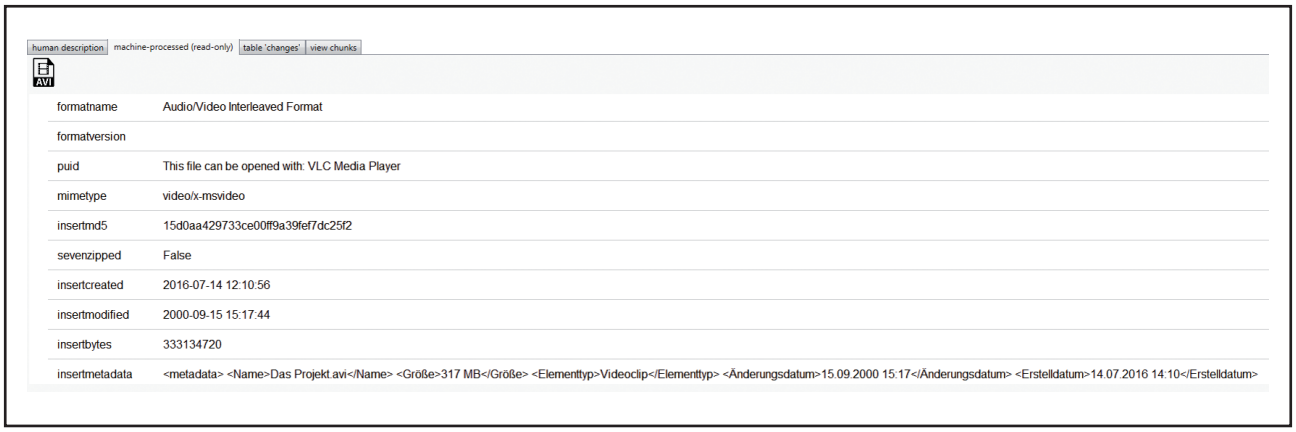


Abb. 10: Identifizierungsergebnis in Bytebarn

Lösung zu 4 – SQLite als Archivformat

Mit der Konvertierung des Dateiverzeichnisses in eine SQLite Datenbank ist die Abhängigkeit vom Betriebssystem aufgelöst. Die Kriterien für ein geeignetes Archivformat sind durch SQLite erfüllt. SQLite ist weit verbreitet und wird sowohl in Windowsumgebungen wie auch auf Linuxplattformen genutzt. Das Format und die nötige Software sind frei verfügbar und gut dokumentiert.

Lösung zu 5 – Als Notlösung bedingt geeignet

Bytebarn ist kein vollumfängliches Digitales Archiv. Sollte jedoch eine sofortige Archivierung von Dateiverzeichnissen für ein Archiv ohne entsprechende Infrastruktur unumgänglich sein, dann wäre die Ablage der archivierten Dateiverzeichnisse in einem Netzwerk als Bytebarn-Dateien sicher eine bessere Alternative als das rohe Ablegen der archivierten Verzeichnisse in einem Verwaltungsnetzwerk. Die Integrität der Daten kann durch Hashwerte geprüft und nachgewiesen werden. Durch die Ummantelung des Archivguts durch die Bytebarn-Datei ist eine Trennung zum Schriftgut der Verwaltung gegeben.

Fazit

Die Entwicklung von Bytebarn hat die Arbeit am Elektronischen Archiv des Sächsischen Staatsarchivs vereinfacht. Haben früher Dateiverzeichnisse einen hohen Aufwand bei der Archivierung verursacht, können sie nun zügig ingestiert werden. Obwohl hinter Bytebarn eine Datenbank steckt, die technisch etwas anderes darstellt als ein Dateiverzeichnis, bemüht sich die Nutzeroberfläche, dem Nutzer die originale Optik eines Dateiverzeichnisses zu bieten. Was noch wichtiger ist: Die originale Ordnungsstruktur bleibt erhalten. Sie kann aber noch mit erklärenden Metadaten angereichert werden, um die Nutzung zu erleichtern. So wird nicht nur der Erhalt einer komplexen Struktur gewährleistet, die Sammlung wird durch zusätzliche Features wie Textsuche und die Extraktion von Metadaten deutlich im Nutzungswert bereichert. Bytebarn protokolliert sich selbst.

Übernahme unstrukturierter Dateisammlungen mit startext COMO

Christian Fabian Näser, Alexander Herschung

Einführung

Die Übernahme von Dateisammlungen stellt die archivische Kernarbeit vor neue und teilweise unbekannte Herausforderungen. Für die Übernahme von strukturiertem digitalen Content gibt es Lösungen, doch wie sieht es für unstrukturierte Dateisammlungen aus? Mit der Digitalisierung des Arbeitsplatzes sowie der privaten Korrespondenz werden unstrukturierte Dateisammlungen die Archive in naher Zukunft immer mehr beschäftigen. Sei es die Datensammlung eines Sachbearbeiters oder die Bildersammlung eines Fotografen; selten liegt das Material geordnet vor. Unstrukturierte Dateisammlungen zu archivieren erfordert gemeinhin einen großen Zeit- und Personalaufwand, da es zunächst gilt, diese Sammlungen zu sichten, zu ordnen und deren Inhalte zu bewerten. Dieser, meist aufwendige, Prozess wird häufig durch die erhebliche Datenmenge und die in vielen Fällen lückenhaften Informationen zu den Dateisammlungen erschwert. Eine Herausforderung stellt hierbei die Bewertung der digitalen Dateisammlung und die Bildung archivischer Übernahmepakete (*AIPs*) aus unstrukturierten Ablieferungen dar. Konnten Papierakten noch händisch schnell durchgeschaut – und so ganze Aktenlieferungen relativ einfach bearbeitet werden – bedarf es im Umgang mit digitalem Content der Hilfe von passenden Werkzeugen.

Mit Blick auf diese Problematik hat die startext GmbH eine einfache und zielführende Lösung entwickelt. Bei der Entwicklung standen eine unkomplizierte und betriebssystemunabhängige Bedienung im Vordergrund.

Bei dem Übernahmeeditor COMO der startext GmbH handelt es sich um eine plattformunabhängige JAVA-Anwendung, die ohne Installation betrieben werden kann. Durch die Verwendung einer JAVA-Komponente ist es so möglich, das Werkzeug auf einem PC, Laptop, oder sogar direkt von einem USB-Stick zu starten. Ein flexibler Einsatz ist so für alle Arbeitsumgebungen gewährleistet.

Das Werkzeug ist eine *standalone* Anwendung und muss nicht mit anderen Programmen der startext GmbH Familie betrieben werden.[1] Das Werkzeug ist intuitiv zu bedienen, es soll sich nicht in erster Linie an Archivare richten, die bereits lange im Thema der digitalen Archivierung arbeiten, sondern soll vor allem den einfachen Einstieg in das Themenfeld ermöglichen.

Der Vorteil des Übernahmeeditors ist, neben der einfachen Installation und der unkomplizierten Bedienung, eine Protokollierung aller vom Benutzer getroffenen Entscheidungen, das bedeutet konkret eine lückenlose Transparenz aller Arbeitsschritte.

[1] Für das startext Archivinformationssystem ACTApro Desk® sind bereits Schnittstellen für die direkte Übernahme der aus dem startext Übernahmeeditor erzeugten Daten geplant.

Von der Übernahme zur Bildung der Übernahmepakete – Die einzelnen Schritte des Übernahmeeditors

1. Einlesen der Primärdaten
2. Extraktion der technischen Metadaten
3. Anlegen von Übernahmepaketen
4. Sortieren und Filtern
5. Bewertungsentscheidung und Zuweisung zu einem Übernahmepaket
6. Erstellen der Übernahmepakete
7. Aussichten

1. Einlesen der Primärdaten

Der Übernahmeeditor arbeitet auf Grundlage einer Ordnerstruktur bzw. von Teilen einer Ordnerstruktur. Beim Einlesen einer Ordnerstruktur in den Übernahmeeditor wird für diese ein Suchindex erstellt. Konkret bedeutet dies, dass die eigentlichen Daten nicht bewegt und nicht in das Werkzeug geladen werden. Somit werden die Originaldaten bei der Bearbeitung innerhalb des Tools nicht verändert, verschoben oder gelöscht. Dies sichert eine schnelle Performance und einen sehr geringen Speicherverbrauch, dadurch gibt es keine Größenbegrenzung der Dateisammlung oder der Anzahl der darin enthaltenen Dateien.

Innerhalb des Werkzeugs kann einfach ein Ordner ausgewählt und die einzelnen Dateien im Werkzeug direkt angezeigt werden.

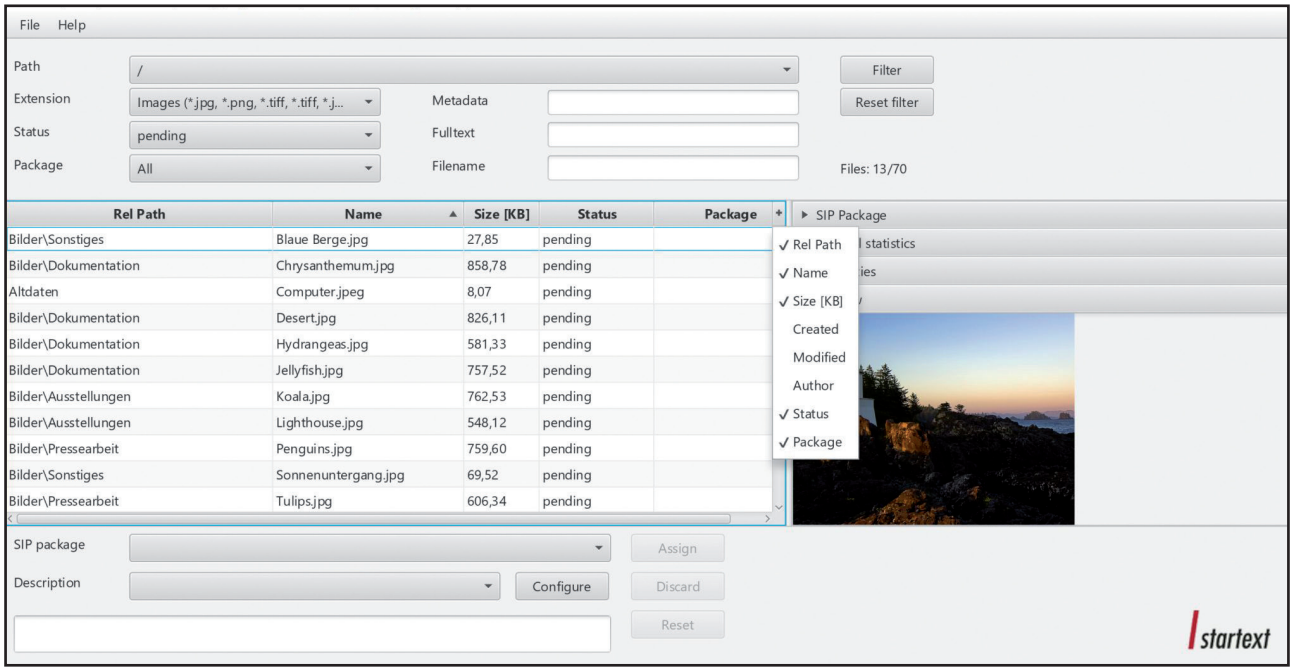


Abb. 1: Darstellung einer Dateiablage im Übernahmeeditor (Dateiname: „Übernahmeeditor_01“)

Die einzelnen Dateien werden in einer Tabellenansicht angezeigt und lassen sich nach mehreren Kriterien sortieren, um eine schnellere Bearbeitung zu gewährleisten. Die Dateien können aus dieser Ansicht heraus direkt aus der Dateiablage geöffnet werden. Beim

Einlesen der Dateien in das Werkzeug wird für jede Datei eine MD5[2] Prüfsumme berechnet. Diese Prüfsumme wird im letzten Bearbeitungsschritt des Werkzeugs, beim Erstellen der einzelnen Pakete, abgeglichen, da diese durch das Öffnen der Dateien verändert werden können. Durch den Einsatz einer MD5 Prüfsumme lässt sich sicherstellen, dass die zu archivierenden Dateien den abgelieferten Dateien entsprechen. Das Öffnen soll an dieser Stelle aber ermöglicht werden, um direkten Zugriff auf den Inhalt der einzelnen Dateien zu ermöglichen. Ein Vorschaufenster soll hierzu eine erste Anlaufstelle bieten und die Auswahl der zu öffnenden Dateien erleichtern. Unter Zuhilfenahme eines Vorschaubildes können so im Zweifelsfall bereits erste Bewertungsentscheidungen getroffen werden.

2. Extraktion der technischen Metadaten

Ein Auslesen der technischen Metadaten ist für die weitere Bearbeitung und Archivierung unabdingbar. Die technischen Metadaten können bei der Bewertung der Archivwürdigkeit helfen und dienen später als Hilfestellung für die tiefere Erschließung. So lassen sich aus verschiedenen technischen Metadaten bereits Laufzeiten oder Ersteller der Dateien ermitteln. Auch Duplikate von Dateien und verschiedene Bearbeitungsversionen derselben Datei werden identifiziert.

Darüber hinaus ist das Speichern der technischen Metadaten für die Einhaltung des OAIS-Modells *(Open Archival Information System)* von zentraler Bedeutung. Technische Metadaten bilden die Grundlagen für die Datenübernahme in ein Langzeitarchiv *(Ingest)* und eine spätere Erhaltungsplanung mit Formatmigration oder Emulation *(Preservation Planning)*.

Für die Indexierung der technischen Metadaten verwendet das Werkzeug die Suchmaschinentechnologie *Apache Lucene*[3].

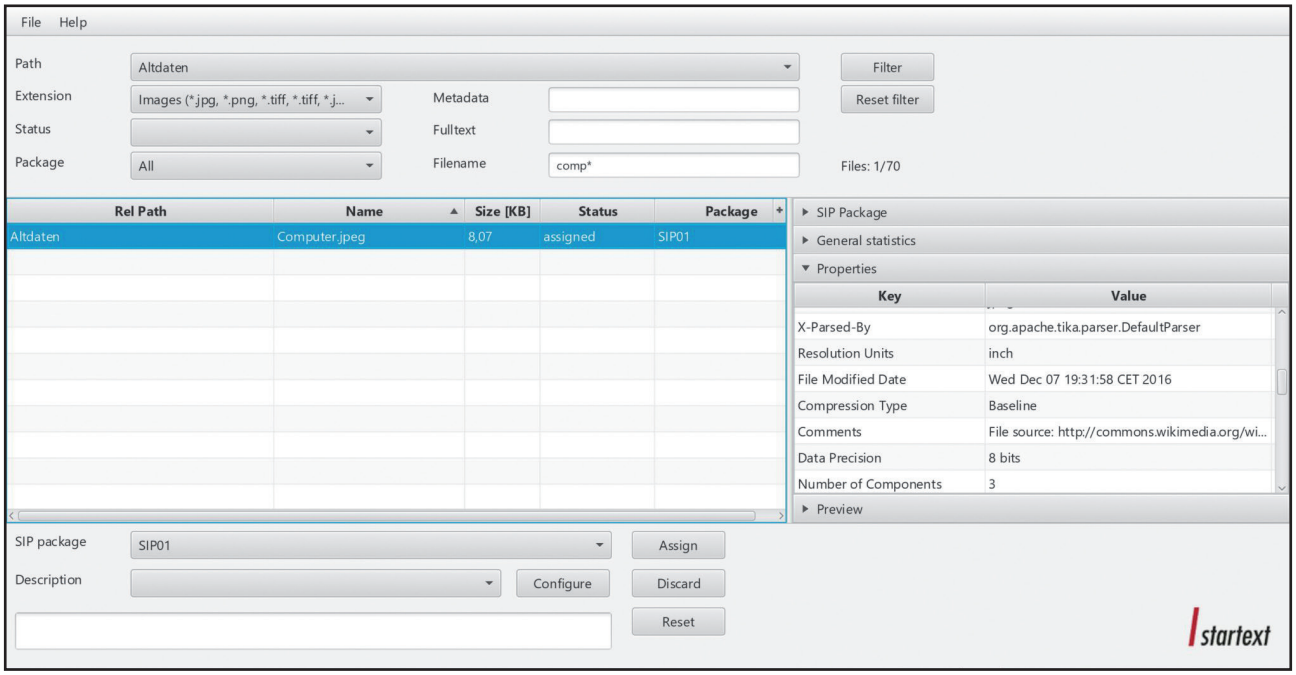


Abb. 2: Die Ansicht der Dateien nach einem bestimmten Kriterium gefiltert (Dateiname: „Übernahmeeditor_02“)

[2] **Message-Digest Algorithm 5** (**MD5**) ist eine gebräuchliche kryptologische Streuwertfunktion, die eine einfache Überprüfung der Datenkorrektheit erlaubt.
[3] Diese wird unter anderen von der Wikipedia und Twitter verwendet, um große Datenmengen schnell zu indexieren und durchsuchbar zu machen.

3. Anlegen von Übernahmepaketen

Im Vorfeld der eigentlichen Bearbeitung müssen im Werkzeug einzelne Übernahmepakete, die SIPs *(Submission Information Package)*, definiert werden. Diese bilden bei der späteren Übernahme der Pakete in ein Archivinformationssystem *(AIS)* die Grundlage für die AIPs *(Archival Information Package)*, die klassischen Verzeichnungseinheiten.

An dieser Stelle können bereits die Titel der einzelnen SIPs festlegt und später in ein Archivinformationssystem übernommen werden.

4. Sortieren und Filtern

Für einen besseren Überblick lassen sich die Dateien nach bestimmten Kriterien, darunter nach technischen Metadaten, filtern. So lassen sich zum Beispiel alle Dateien einer bestimmten Endung – alternativ auch einer Gruppe von Endungen, zum Beispiel der Gruppe Textdokumente – auswählen, um so direkt en bloc ganze Dateiformate zu übernehmen oder zu kassieren. Beispielsweise ließen sich so in einer Dateiablage alle Systemdateien (diese haben meist keinerlei archivischen Wert) lokalisieren und mit einem Klick verwerfen. Dies erleichtert ein schnelles Bearbeiten großer unstrukturierter Dateisammlungen.

Ferner lässt sich die Dateiablage nach ihrem Pfad sortieren. So können systematisch bestimmte Ordnerstrukturen abgearbeitet beziehungsweise in Gänze übernommen oder kassiert werden.

Dank der eingesetzten Suchmaschine *Apache Lucene* wird der Volltext jeder Datei indexiert und ist so über eine Volltextsuche durchsuchbar. Dadurch lassen sich auch Inhalte von Text- oder PDF-Dateien schnell finden, um so Bewertungsentscheidungen effizienter vornehmen zu können. Die Suchfunktion unterstützt sowohl unscharfe und phonetische Suche als auch andere gängige Funktionen, wie z.B. Boolesche Operatoren.

5. Bewertungsentscheidung und Zuweisung zu einem Übernahmepaket

Wurden die zu bearbeitenden Dateien sortiert und gefiltert, kann die eigentliche Bearbeitung – die Vergabe von Bewertungsentscheidungen – beginnen. Dateien können sowohl einzeln als auch als ganze Blöcke bearbeitet werden. An dieser Stelle können Dateien nun entweder übernommen und einem SIP zugeordnet, oder aber verworfen und damit kassiert werden.

Diese Bewertungsentscheidung lässt sich mit einer Begründung hinterlegen. Hierfür ist sowohl ein Freitextfeld als auch eine frei konfigurierbare Dropdown-Liste im System vorhanden. Die verworfenen Dateien werden als solche im Werkzeug markiert, es werden jedoch keine Originaldateien durch das Werkzeug gelöscht. Die Anzahl der zu einem SIP gehörigen Dateien ist nicht begrenzt; auf welcher Grundlage die SIPs gebildet werden, obliegt also dem Bearbeiter. Die Zuordnung lässt sich zu jedem Zeitpunkt nachträglich verändern, Erst durch das Erstellen der SIPs wird die endgültige Zuordnung festgelegt.

Alle Dateien der Dateisammlung lassen sich durch Filtern des Status der angezeigten Dateien – zugeordnet oder noch nicht zugeordnet – bequem bearbeiten, ohne dabei die Übersicht zu verlieren, welche Dateien bereits zugeordnet oder verworfen wurden.

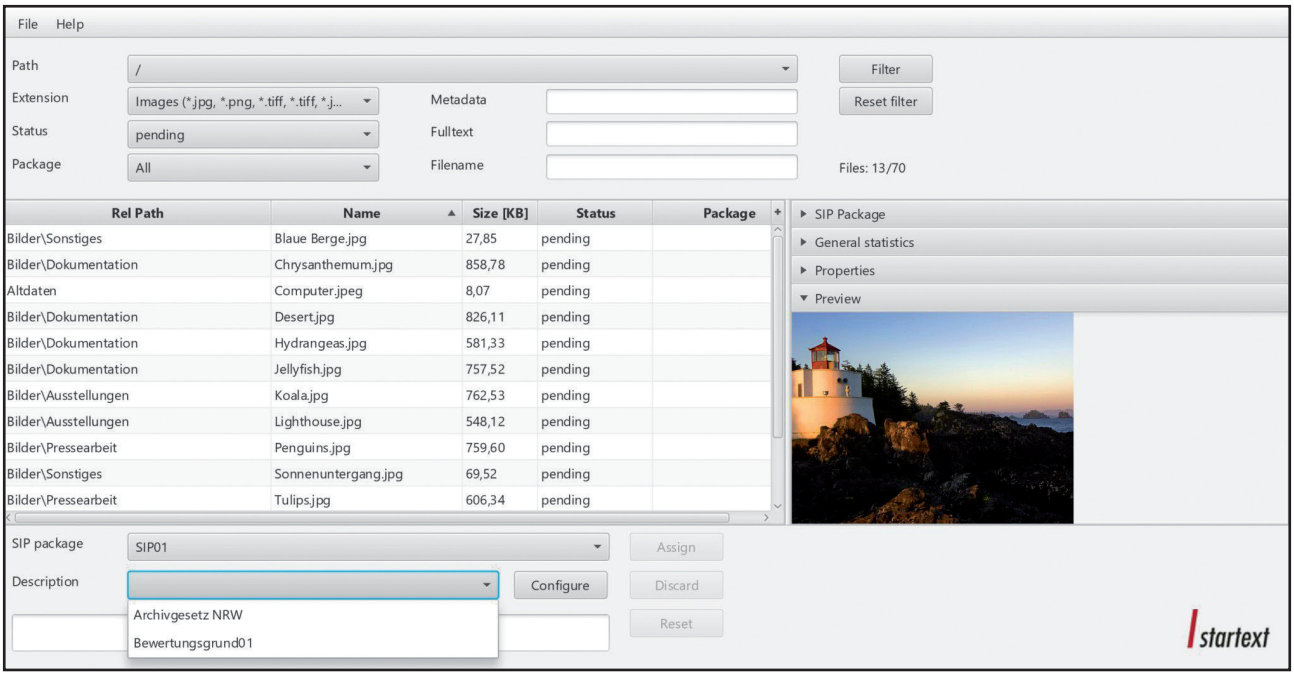


Abb. 3: Hinterlegung einer Begründung für eine getroffene Bewertungsentscheidung (Dateiname: „Übernahmeeditor_03“)

6. Erstellen der Übernahmepakete

Ist die Dateiablage fertig bearbeitet, können die Übernahmepakete als .zip Dateien erstellt werden. Innerhalb jedes Pakets befinden sich nun alle dem Paket zugeordneten Dateien in ihrer ursprünglichen Ordnerstruktur. Strukturpunkte, die keine Dateien innerhalb dieses Pakets beinhalten, werden hierbei nicht mit übernommen.

Darüber hinaus wird für jedes Paket eine METS-XML-Datei *(METS = Metadata Encoding & Transmission Standard)* erstellt. Hierin werden alle Informationen der ursprünglichen Dateiablage gespeichert. Hierzu gehört neben den Angaben der einzelnen Dateien und ihrer technischen Metadaten auch die Bewertungsentscheidung jeder einzelnen Datei. Diese Informationen werden sowohl für übernommene als auch für kassierte Dateien festgehalten. So lassen sich im Nachhinein aus dieser METS-XML-Datei Kassationsprotokolle erstellen. Auch der Volltext wird mit in der METS-XML-Datei gespeichert, sodass dieser für spätere Recherchen mit in ein Archivinformationssystem übernommen werden kann.

Die Verwendung von METS-XML-Dateien ist als archivischer Standard anerkannt, weswegen die Speicherung von Informationen in diesem Format den Vorteil einer systemunabhängigen Weiterverarbeitung bietet: Die automatisierte Weitergabe der Informationen – darunter der Titel des SIPs, die technischen Metadaten sowie die Bewertungsentscheidungen – in beliebige Archivinformationssysteme ist an dieser Stelle sichergestellt.

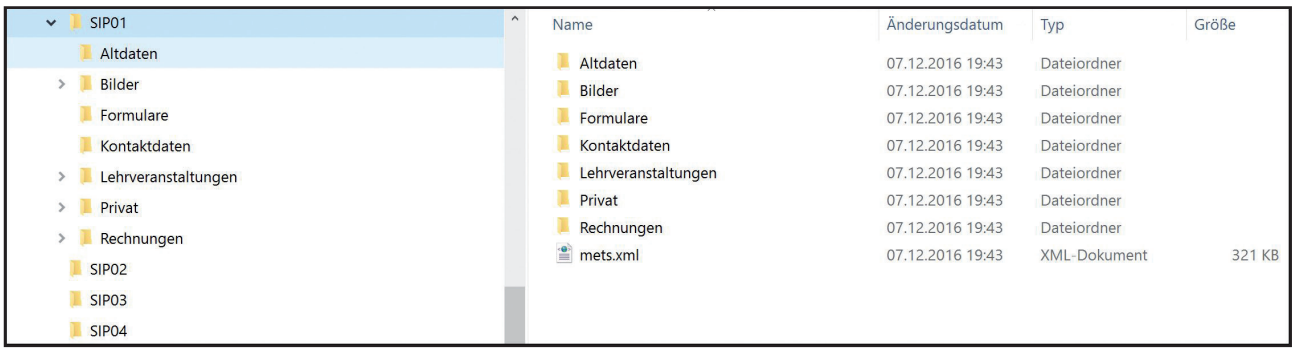


Abb. 4: Die übernommene Ordnerstruktur innerhalb eines SIP (Dateiname: „Übernahmeeditor_04“)

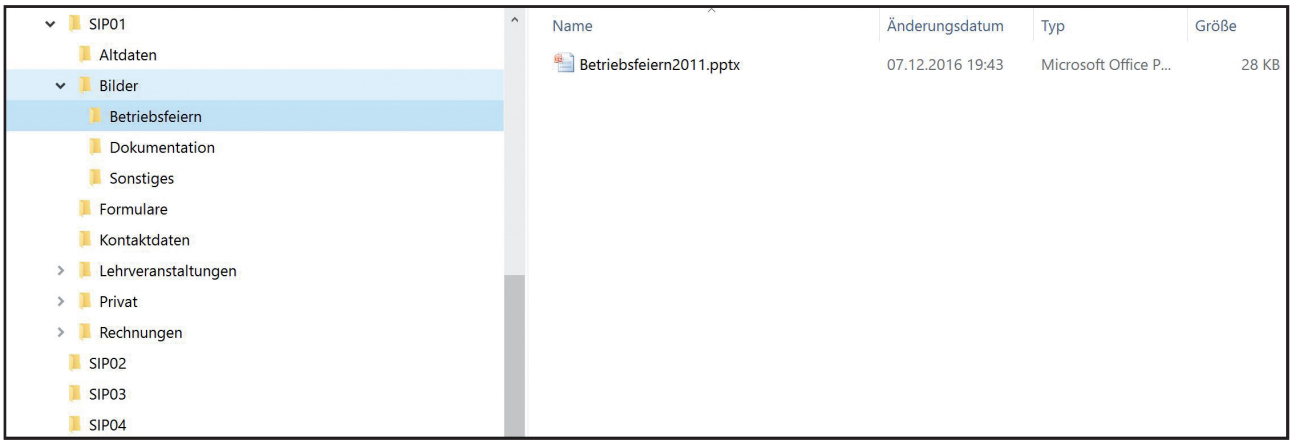


Abb. 5: Die Dateien innerhalb der Ordnerstruktur eines SIP (Dateiname: „Übernahmeeditor_05“)

7. Aussichten

COMO befindet sich aktuell (Stand März 2017) in der finalen Entwicklungsphase. Das Release ist für Mai 2017 geplant.

Weiter geplant ist die Möglichkeit zur Basisbeschreibung der Übernahmepakete durch das Datenformat *Dublin Core*, sowie die komplette Erfassung der Übernahmedaten innerhalb des Werkzeugs. Das Werkzeug wird mit deutscher Benutzeroberfläche erscheinen. Perspektivisch soll so der gesamte Workflow bis hin zur Übergabe von digitalem Content an ein Archivinformationssystem in einem einfach zu bedienenden Werkzeug ermöglicht werden.

Neuigkeiten zum Übernahmeeditor COMO werden auf der Homepage der startext GmbH veröffentlicht: *www.startext.de*.

Der Package Handler des Schweizerischen Bundesarchivs

Kai Naumann

Package Handler[1] wurde mit dem Ziel entwickelt, Ablieferungen (SIP[2]) an das Schweizerische Bundesarchiv (BAR) zu prüfen (Release 2009) und – in einem zweiten Entwicklungsschritt – die Erstellung von SIP durch die Behörden zu ermöglichen (ab Release 2013). Damit sollte auch die Ablieferung von digitalen Unterlagen an das Bundesarchiv erleichtert werden. Auch in der Schweiz ist trotz der Verbreitung von DMS (dort als GEVER bezeichnet) in der Bundesverwaltung eine große Menge historisch wertvoller Unterlagen in Gruppenlaufwerken oder individuellen Ablagen auf Einzel-PCs vorhanden. Zudem gibt es auch archivwürdige Daten in Datenbanken und Fachapplikationen, welche nicht über eine eingebaute Schnittstelle für die SIP-Erstellung für die Ablieferung aufbereitet werden können. Das Bundesarchiv reagierte, indem es mit Package Handler eine Software bereitstellte, die standardkonforme Ablieferungspakete (Submission Information Packages, SIP) erzeugen kann. Ein solches Paket besteht aus einem nach Vorgabe gepackten Container, der behördliche Primärdaten und Metadaten aus Behörde und Archiv enthält (eCH-Standard-160 Archivische Ablieferungsschnittstelle[3]). Package Handler erleichtert den abgebenden Stellen, diese Pakete selbstständig zu erzeugen.

Die Einsatzmöglichkeiten von Package Handler gehen über das Hauptmotiv seiner Entwicklung aber hinaus. Er eignet sich erstens auch dafür, ein anderweitig erzeugtes Ablieferungspaket auf Konformität mit dem Standard eCH-0160 zu validieren. Zweitens kann das Werkzeug im Aufbereitungsprozess dazu verwendet werden, eine Dateisammlung nur zu sichten und zu analysieren, um beispielsweise im Dialog mit der Behörde oder Dienstleistern die Aufbereitung oder Binnenbewertung vorzuplanen. Drittens kann Package Handler dank seiner recht umfassenden Feldausstattung auch zum Vorab-Erschließen digitaler Unterlagen verwendet werden. Viertens kann das entstandene SIP mit seinen Metadaten nicht nur in eCH-0160 kompatible Systeme überführt werden. Sofern nämlich Umwandlungsmöglichkeiten außerhalb des Package Handlers vorhanden sind, welche eine Transformation von eCH-0160 in das gewünschte Paketformat vornehmen können, kann Package Handler auch als Basis für den Ingest in andere Digitale Archivsysteme verwendet werden. Fünftens kann Package Handler auch in der archivischen Nutzung dazu dienen, ein standardkonform gebildetes Archivale zu betrachten, sofern dieses als Paket gemäß eCH-0160 vorliegt.

Da das Schweizerische Bundesarchiv eine Kurzanleitung und eine ausführliche Nutzerdokumentation für jedermann bereithält, wird der Verfasser im Folgenden nur das Grundkonzept schildern und die wesentlichen Funktionsbereiche kurz erwähnen. Anschließend werden die Stärken und die wenigen Schwächen des Werkzeugs beleuchtet.

[1] *https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/package-handler.html* (aufgerufen am 6.4.2017).

[2] Die Definition von SIP (und auch AIP) ist zweideutig. SIP steht an anderen Orten in diesem Band nicht wie hier für die Gesamtheit aller Unterlagen, die in einer Ablieferung im Archiv ankommen, sondern für die späteren bestellbaren Einheiten, die für Package Handler „Dossier“ heißen und meist einer Akteneinheit entsprechend dürften. OAIS bietet mit den Paketvarianten „Collection“ und „Unit“ eine Differenzierungsmöglichkeit zwischen Paketgesamtheiten und Einzelpaketen. In diesem Sinne handelt es sich im Folgenden bei SIP also um Submission Information Collections (SICs).

[3] *https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0160&documentVersion=1.0*

Während in diesem Band die meisten Werkzeuge von Autoren[4] vorgestellt werden, die diese selbst produktiv einsetzen, hat diese Vorstellung eher den Charakter einer Rezension, denn der Autor arbeitet produktiv mit anderen Werkzeugen. Die technischen und inhaltlichen Informationen zum Package Handler im vorliegenden Text wurden mit dem Schweizerischen Bundesarchiv abgestimmt.[5]

Das Grundkonzept von Package Handler

Der Standard eCH-0160 kennt Dossiers (nach bundesdeutscher Terminologie „Akten"), Subdossiers („Vorgänge"), Dokumente und Dateien. Wird Package Handler zur Erstellung eines SIP benutzt, beginnt es zunächst mit einer Analyse der Dateisammlung und schreibt deren Ergebnis in eine Bestandsaufnahme, das Inhaltsverzeichnis des Pakets.

Bei der Entwicklung von Package Handler wurde darauf geachtet, die verschiedenen konzeptuellen Sichten auf die zu bearbeitenden digitalen Unterlagen zu trennen. Dabei wurden Objektsichten von prozessualen Sichten getrennt. Letztere betreffen vor allem Meldungen im Rahmen der Validierung, aber auch Suchergebnisse. Die Objektsichten wurden in einer sehr durchdachten Konzeption dreigeteilt und können mit drei verschiedenen, sogenannten Navigatoren betrachtet und angepasst werden:

- Die Paketsicht (Abb. 1) stellt die digitalen Unterlagen strukturell so dar, wie sie als Dateien und Ordner im Inhaltsverzeichnis des SIP beschrieben werden.[6] Parallel werden einige wenige, von Package Handler erhobene Metadaten zu Datei und Ordner gezeigt.
- Die Ordnungssystem-Sicht (Abb. 2) stellt in einer Baumsicht die digitalen Unterlagen in einer thematischen Struktur dar, die im Lauf der Bearbeitung nach den Maßstäben der Behörde oder des Archivs angepasst werden kann. Ein ordnendes Objekt ohne Primärdaten wird, wenn es oberhalb von Dossiers angelegt ist, „Ordnungssystemposition", kurz OSP genannt. Die Ordnungssystemsicht kann teilautomatisiert aus der Paketsicht entwickelt werden. Hierzu weiter unten mehr.
- Die Dossier-Sicht (Abb. 3) lässt das übergreifende Ordnungssystem weg, und wie in einer Ablieferungsliste für Papierakten sind nur die definierten Dossiers untereinander angeordnet. Die darin enthaltenen Dokumente / Dateien lassen sich darunter ausklappen.

Die drei verschiedenen Sichten stellen nur verschiedene Darstellungsweisen der Inhalte der XML-Datei im SIP dar und werden durch den Package Handler generiert. Auf der Datenhaltungsebene verändert sich die Position der Primärdaten nicht, wenn diese in ein SIP hinein importiert worden sind. Strukturierung und Datenhaltung im SIP bleiben stets voneinander getrennt. Dies gewährleistet ein effizientes Arbeiten, weil es parallele Klärungen auf der inhaltlichen und der technischen Ebene erlaubt. So können in der Paketsicht Dateien von der Behörde noch nachgeliefert werden, während in der Ordnungssystemsicht bereits eine Ordnung der Unterlagen für ein archivisches Findmittel ausgetüftelt wird. Der ursprüngliche Name der Datei (vorarchivisch) bleibt in jedem Fall als Metadatum mit erhalten.

[4] In diesem Text wird im Sinne der sprachlichen Übersichtlichkeit stets die männliche Form verwendet. Bei allen Rollenbezeichnungen sind aber selbstverständlich weibliche Personen ebenfalls gemeint.

[5] An dieser Stelle ganz herzlichen Dank an Marguérite Bos und Nicole Martini, die sich wirklich viel Zeit für Hinweise genommen haben.

[6] Package Handler Release 2015 Nutzerdokumentation, Bern 2016, *https://www.bar.admin.ch/dam/bar/de/dokumente/kundeninformation/benutzerdokumentationpackagehandlerrelease2013.pdf.download.pdf/benutzerdokumentationpackagehandlerrelease2015.pdf* (abgerufen 13.8.2017), S. 31f.

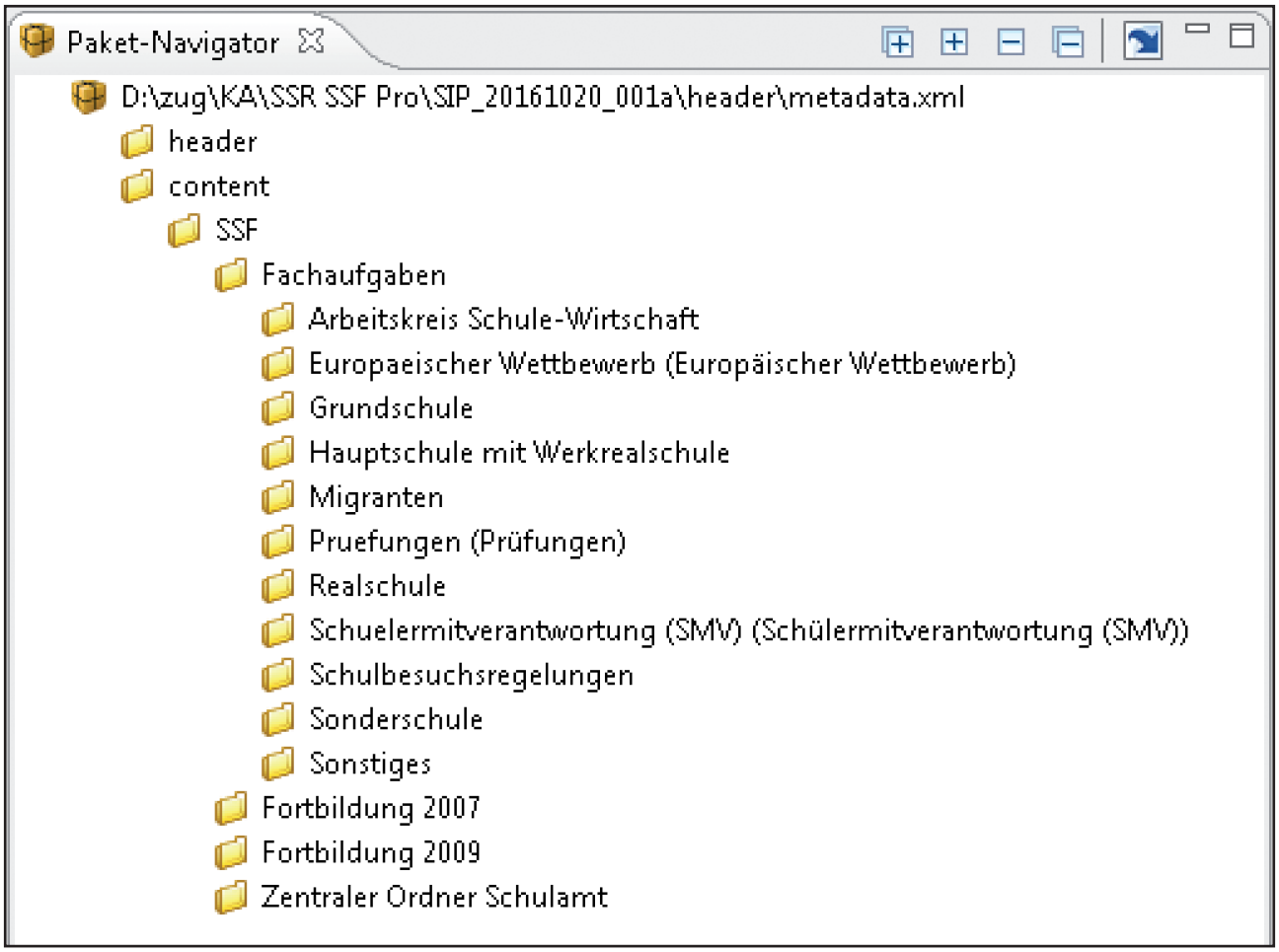


Abb. 1: Paketsicht im Package Handler

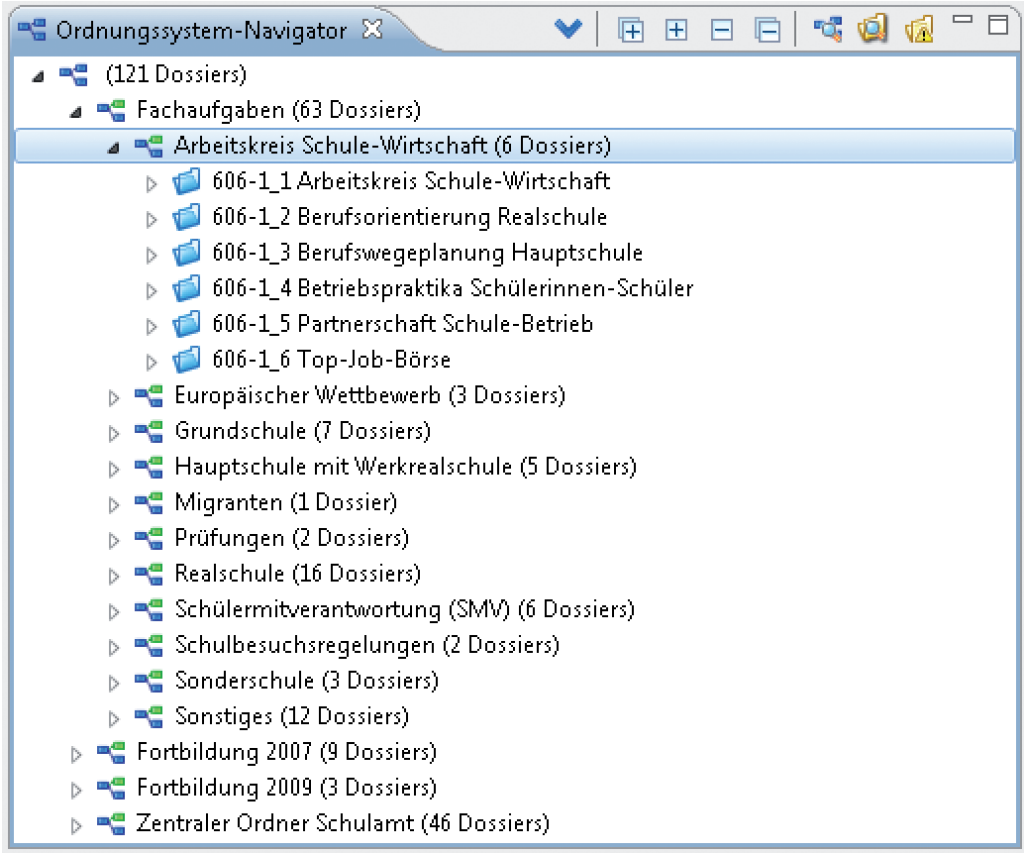


Abb. 2: Ordnungssystem-Sicht im Package Handler

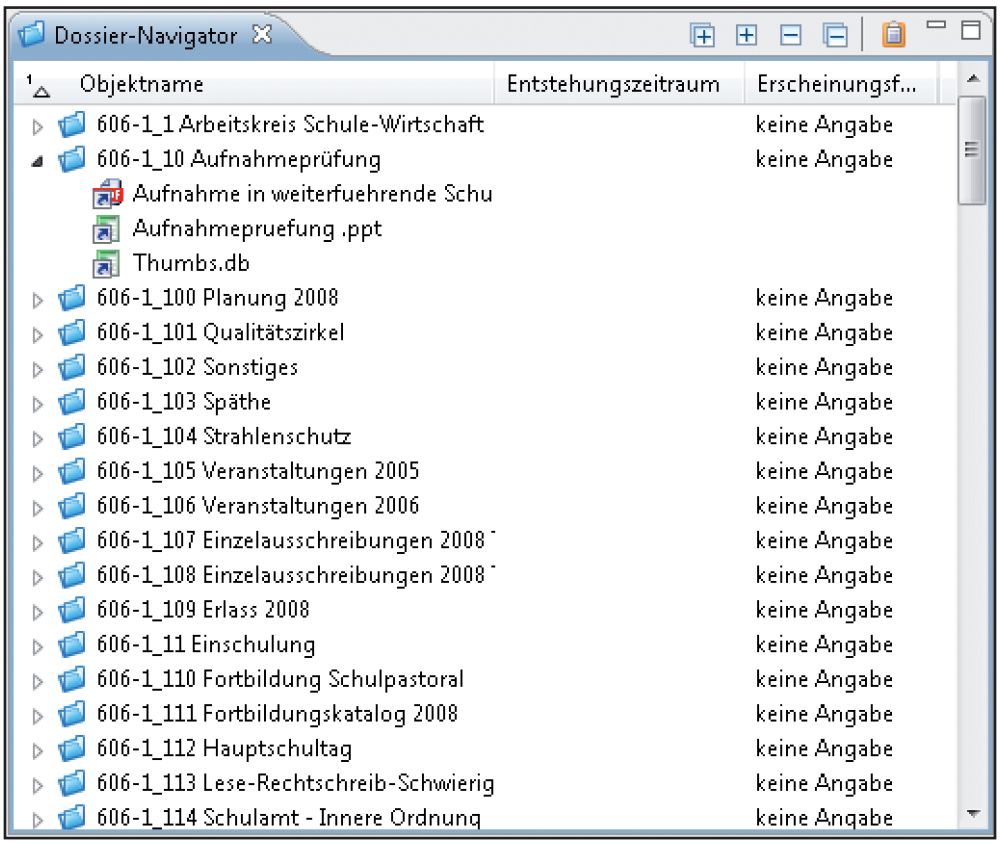


Abb. 3: Dossier-Sicht im Package Handler

Zum Grundkonzept gehört auch, dass Package Handler die Regeln des Standards eCH-0160 in maschinenlesbarer Form verinnerlicht hat. Zu jedem Zeitpunkt kann das SIP daraufhin überprüft werden, ob es alle Regeln bereits einhält (Abb. 4).[7] Schließlich ist noch zu vermerken, dass Package Handler die mit seiner Hilfe vollzogenen Änderungen an den Metadaten einer Ablieferung während der Arbeit mitverfolgt und bis zur Speicherung des SIP ebenfalls protokolliert.

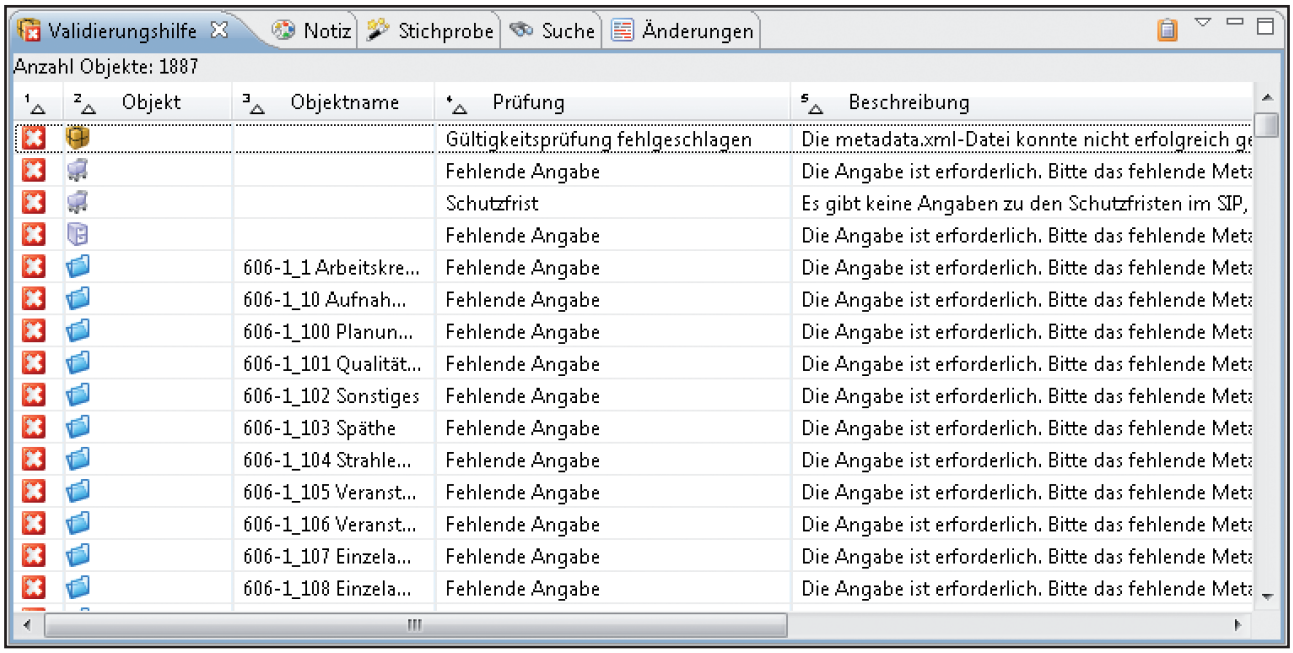


Abb. 4: Validierungshilfe des Package Handlers mit Prüfungsergebnissen

[7] Der umfassende Katalog möglicher Fehlermeldungen findet sich in der Nutzerdokumentation, S. 62–80.

Bedienung von Package Handler

Die Oberfläche ist dreigeteilt in eine Navigatorleiste, ein Hauptfeld für das Anzeigen und Editieren der Metadaten zu einzelnen Objekten und ein weiteres Feld mit der sogenannten Aktionsview, welches in Listenform Resultate der jeweiligen Aktionen (z. B. Suche, Prüfungen etc.) anzeigt. Die Navigatoren stellen Bäume oder Listen der betreffenden Objekte dar. Die drei Fensterbereiche lassen sich gegeneinander flexibel verschieben. Primärdaten können aus den Navigatoren und den Aktionsviews heraus mit der auf dem jeweiligen Rechner installierten Viewersoftware geöffnet werden. Die Editiermasken entsprechen dem Metadatenkatalog gemäss eCH-0160, der auf allgemeinen archivischen Standards für die Erschließung von Archivgut aufbaut (z. B. ISAD(G)).

Um Merkposten beim Bearbeiten zu haben, können Notizen zu einzelnen Objekten angelegt und als Todo-Liste für weitere Aufbereitungsschritte angezeigt werden. Das Werkzeug erleichtert dadurch arbeitsteilige und asynchrone Prozesse (Abb. 5).

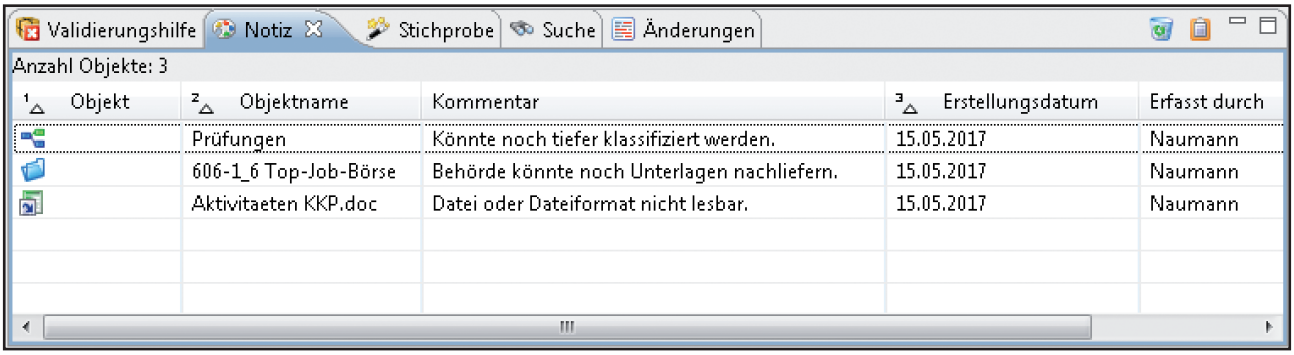


Abb. 5: Todo-Liste aufgrund von Notizen an verschiedenen Stellen des SIP

Beurteilung von Stärken und Schwächen des Werkzeugs

Die massenhafte Verarbeitung von kreativen digitalen Ablagen wird sich nur erreichen lassen, wenn sich der Archivar die behördlichen Metadaten zueigen machen kann. Hierfür hat Package Handler mit einem Zuordnungs-Assistenten („Massenzuweisung“, vgl. Abb. 6) eine praxistaugliche Lösung geschaffen. Im Assistenten kann eingestellt werden, welche Ebene des Verzeichnisbaums (Titel in der Dateiablage) den Titel für künftige Dossiers abgeben soll.[8] So entstehen die Betreffzeilen der Dossiers in der Ordnungssystem-Sicht automatisch aus den Bezeichnungen der entsprechenden Verzeichnisordner in der Paketsicht, können aber auch in der Ordnungssystem-Sicht nachträglich angepasst und korrigiert werden. Bei der SIP-Erstellung kann Package Handler auch vorarchivische Ordnungssysteme wie z.B. Aktenpläne und weitere Metadaten als einfache CSV-Datei importieren und übernehmen.

Die Validierung ist so gelöst, dass alle Regelverstöße übersichtlich dargestellt werden und gruppiert und gefiltert werden können. Package Handler erlaubt es auch, Fehlern nachzugehen, die nicht automatisiert festgestellt werden können, weil es sich hier um qualitative Prüfungen handelt (z. B. Inhalt der Metadatenfelder wie Titel). Um solchen Ungereimtheiten auf die Spur zu kommen, listet eine Stichprobenfunktion dem Bearbeiter einen kleinen, zufällig gewählten Teil der Objekte zur Durchsicht auf.

[8] Nutzerdokumentation, Kapitel 8.3.1. Ordnungssystem aufgrund der Ordnerstruktur im Paket-Navigator erstellen S. 45–48.

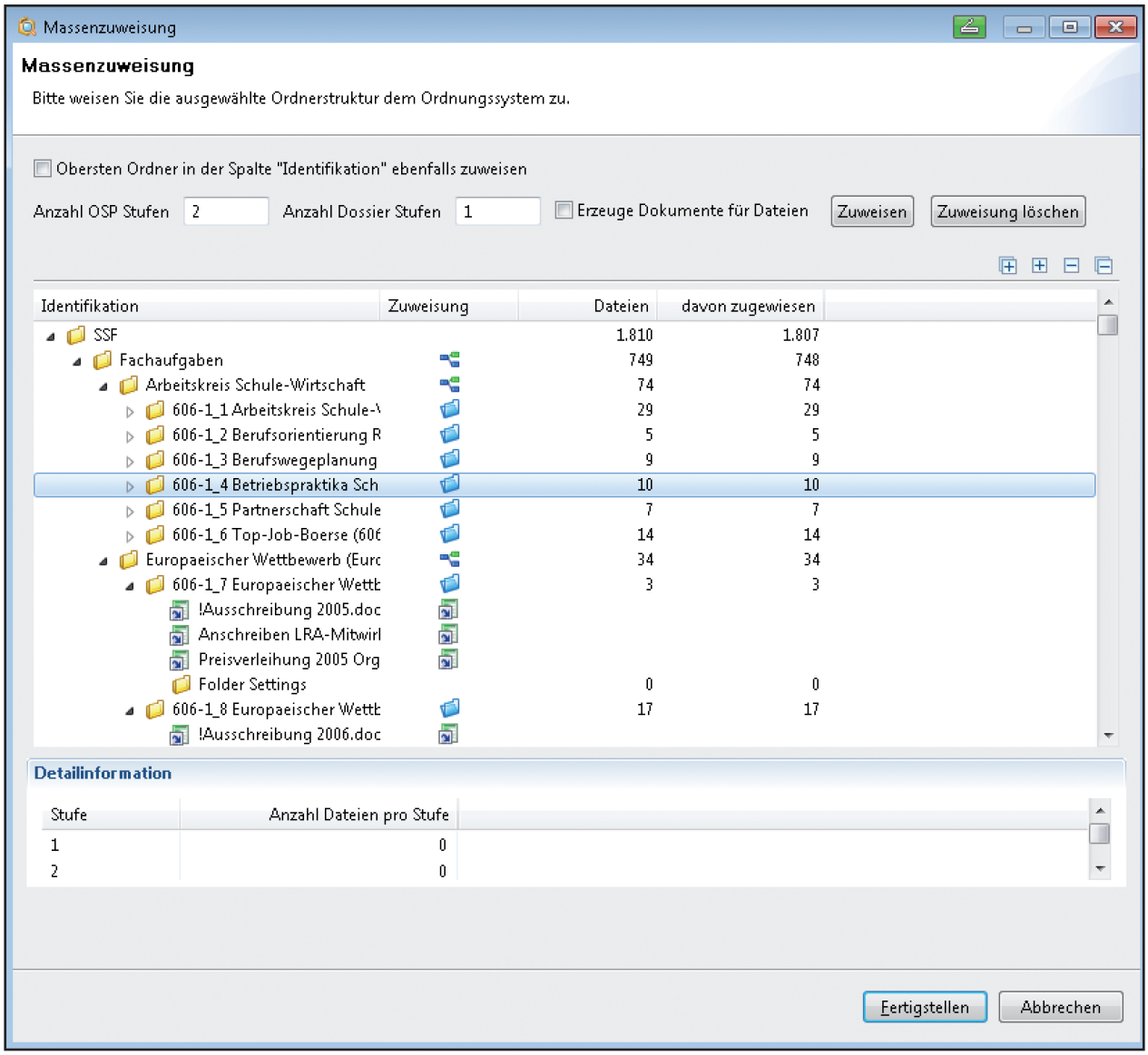


Abb. 6: Halbautomatische Zuordnung von Ebenen eines Dateisystems auf die OSP-Dossier-Dokument Struktur von Package Handler („Massenzuweisung“)

Positiv ist auch, dass die Konzeption die Hybridität heutiger Überlieferung in Rechnung stellt. So ist es möglich, Dossiers mit einem Hinweis zu versehen, dass hier sowohl ein digitaler als auch ein papierener Anteil vorliegt.

Das insgesamt sehr durchdachte Konzept und die solide Verarbeitung von Package Handler lassen es eigentlich nur zu, sich noch mehr Funktionen hinzu zu wünschen. So ist es derzeit möglich, externe Metadaten aus CSV-Daten in die Textfelder von OSPs und Dossiers zu importieren. Hier könnte auch ein Heraussuchen weiterer Metadaten aus anderen Standards ermöglicht werden. Die Protokollierung der getätigten Veränderungen erfasst nur eine Sitzung mit Package Handler bis zum Speichern und Schließen. Hier wäre es bei Ablieferungen mit einem besonders hohen Bedarf an Authentizität vielleicht sinnvoll, wenn Package Handler seine Logs bei Bedarf nicht löscht, sondern eine komplette Bearbeitungsgeschichte des SIP mitschreibt. Auch könnte darüber nachgedacht werden, eine Schnittstelle zu E-Mail-Servern einzurichten, die ein Ordnen und Erschließen dieses Unterlagentyps mit Package Handler ermöglichen würde. Hier ist noch Ausbaupotenzial vorhanden. Lobenswert an Package Handler ist aber nicht nur die technische Seite, sondern auch die Bereitschaft seiner Schöpfer, die Software als Freeware aller Welt bereitzustellen.

DILA Import Preparation Tool

Anne Kathrin Pfeuffer

Strukturiert abgegebene digitale Daten mit eindeutigen, archivfähigen Formaten und im Optimalfall mit Altsignaturen oder sogar Archivsignaturen im Dateinamen lassen sich relativ problemlos in ein Digitales Archiv übernehmen.

Mit den Daten in digitalen Ablagen auf Servern und diversen Datenträgern sieht es oft anders aus. Unstrukturiert, in unterschiedlichsten Formaten (gerne auch veraltet), im schlimmsten Falle namenlos, nicht immer mit stimmigen Datierungen, müssen auch diese Daten bewertet werden. Die Übernahme der als archivwürdig bewerteten Daten in ein digitales Langzeitarchiv gestaltet sich schwierig und aufwendig.

Das Stadtarchiv Braunschweig setzt seit 2011 ein Digitales Langzeitarchiv (DILA) ein. Es setzt sich aus der Archivsoftware, einem Enterprise Content Management System (ECM) und einer für die Langzeitarchivierung geeigneten Speicherlösung zusammen. Analoge wie digitale Archivalien können über die eingesetzte Archivsoftware recherchiert werden.

Die Einspeisung elektronischer Daten zur Speicherung in DILA kann auf zwei Wegen erfolgen: Einerseits durch die direkte Ablage von Daten im ECM-System durch die Archivarin, andererseits über den Import der Daten durch Kollegen in der IT-Abteilung. Egal welche Variante des Imports gewählt wird, die Daten werden mit einigen ihrer Metadaten archiviert. Für diesen Zweck wurde in der IT-Abteilung der Stadt Braunschweig ein kleines Tool

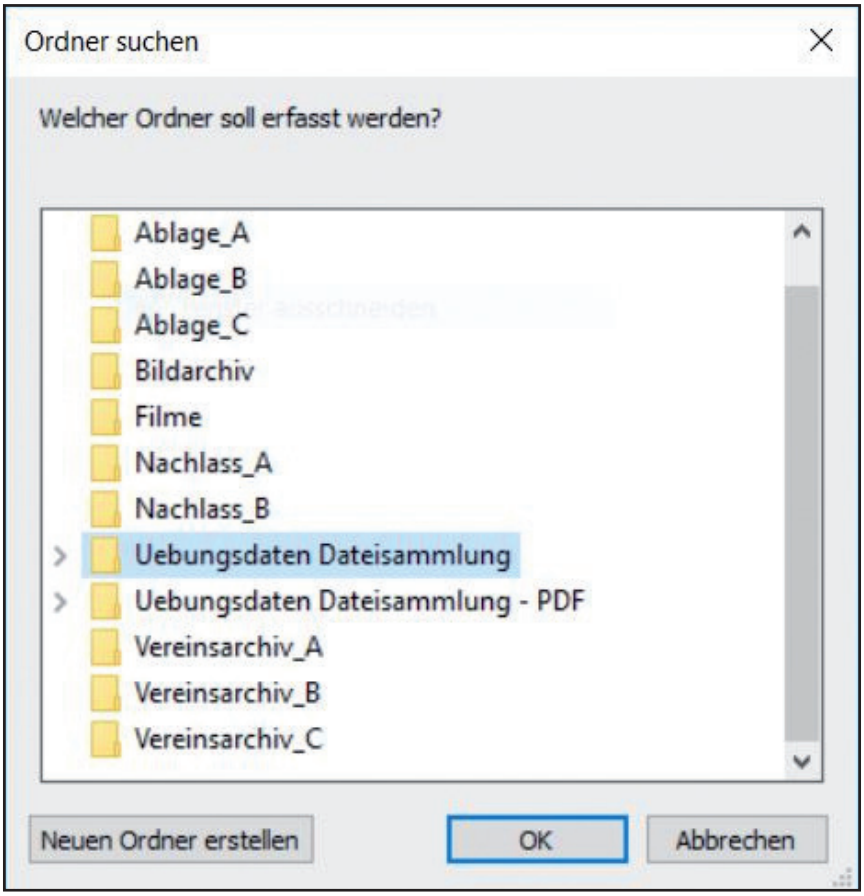


Abb.1: Auswahl des Ordners

entwickelt. Dieses DILA Preparation Tool unterstützt die Vorbereitung der Importe in das Digitale Langzeitarchiv. Nach Start des Tools wird in der Quellverzeichnisstruktur ein Ordner ausgewählt, dessen Dateien archiviert werden sollen. Sollten Unterordner vorhanden sein, so werden diese rekursiv miteinbezogen.

	A	B	C	D	F	G	H	I	J	K	L	M
1	Dateiname	Dateiendung	Erstelldatum	MB	Ordnerpfad	Signatur	Titel	Datierung	Provenienz	Beschreibung	Bestand	Art
2	Vorerschliessung	csv	09.04.2017	0	C:\Ablage_Test\Uebungsdaten Dateisammlung							
3	Computer	jpeg	09.04.2017	0,01	C:\Ablage_Test\Uebungsdaten Dateisammlung\Altdaten							
4	Datei	pdf	09.04.2017	0,05	C:\Ablage_Test\Uebungsdaten Dateisammlung\Altdaten							
5	Kalligraphie-Workshop	docx	09.04.2017	0,26	C:\Ablage_Test\Uebungsdaten Dateisammlung\Altdaten							
6	Verzeichnis	pdf	09.04.2017	0,04	C:\Ablage_Test\Uebungsdaten Dateisammlung\Altdaten							
7	Kalender 08_2016	xls	09.04.2017	0,04	C:\Ablage_Test\Uebungsdaten Dateisammlung\Arbeitsdateien							
8	Nordwind 2007	accdb	09.04.2017	3,88	C:\Ablage_Test\Uebungsdaten Dateisammlung\Arbeitsdateien							
9	uebungsblatt-kurrent	pdf	09.04.2017	0,01	C:\Ablage_Test\Uebungsdaten Dateisammlung\Arbeitsdateien							
10	uebungsblatt-suetterlin	pdf	09.04.2017	0,01	C:\Ablage_Test\Uebungsdaten Dateisammlung\Arbeitsdateien							
11	Kalender Mai 2015	xls	09.04.2017	0,04	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bestellungen							
12	Telefonnummern	txt	09.04.2017	0	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bestellungen							
13	Koala	jpg	09.04.2017	0,74	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Ausstellungen							
14	Lighthouse	jpg	09.04.2017	0,54	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Ausstellungen							
15	Betriebsfeiern2011	pptx	09.04.2017	0,03	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Betriebsfeiern							
16	Chrysanthemum	jpg	09.04.2017	0,84	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Dokumentation							
17	Desert	jpg	09.04.2017	0,81	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Dokumentation							
18	Hydrangeas	jpg	09.04.2017	0,57	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Dokumentation							
19	Jellyfish	jpg	09.04.2017	0,74	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Dokumentation							
20	Penguins	jpg	09.04.2017	0,74	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Pressearbeit							
21	Tulips	jpg	09.04.2017	0,59	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Pressearbeit							
22	Blaue Berge	jpg	09.04.2017	0,03	C:\Ablage_Test\Uebungsdaten Dateisammlung\Bilder\Sonstiges							

Abb. 2: Tabelle mit ausgelesenen Dateieigenschaften

Aus den Dateieigenschaften werden einige ausgewählte ausgelesen. Die gewünschten Dateieigenschaften wurden bei der Konstruktion des Tools durch das Stadtarchiv Braunschweig festgelegt. Nach Beendigung des Auslesens öffnet sich eine mit diesen Daten automatisch generierte Excel-Tabelle.

In dieser Tabelle können Daten dann z.B. sortiert und weitere Daten für die Verzeichnung hinzugefügt werden. Im Stadtarchiv Braunschweig sind dies zwingend die Signatur und ein Feld für die Festlegung des Medientyps. Die Vergabe der Signatur erfolgt in den meisten Fällen bei Born Digitals direkt in der Tabelle. Bei gleichförmigen Massendaten nimmt dies meist nur kurze Zeit in Anspruch. Die Angabe des Medientyps ist für den Import wichtig, da das genutzte ECM-System mehrere für DILA definierte Medientypen unterstützt.

Weitere Verzeichnungsangaben wie beispielsweise Titel und Datierung können ergänzt werden. Liegt bereits eine Verzeichnung in der Archivsoftware vor, können die benötigten Angaben exportiert und in der Tabelle für den Import zusammengefügt werden. Auch bei der Übernahme von digitalen Fotos mit IPTC-Metadaten erfolgt erst ein Import in der Archivsoftware, danach der Ingest der Primärdaten in das ECM. Auch bei anderen Born Digitals wird die erzeugte Tabelle einerseits für den Import der Daten genutzt. Andererseits erfolgt die Verzeichnung dieser Daten durch Import der Verzeichnungsdaten in der Archivsoftware, wo sie später noch angereichert und korrigiert werden können.

Bisher wurden im Stadtarchiv Braunschweig mit dem DILA Preparation Tool seit 2012 über 550 Tabellen für Importe in DILA erstellt.

docuteam packer – Informationspakete bilden und kontrolliert bewirtschaften

Bart Klein, Andreas Steigmeier, Tobias Wildi

Am Anfang stand folgender Wunsch: Die Archivarin eines unserer Kundenarchive suchte ein Werkzeug, mit welchem sie digitale Ablieferungen nicht nur bilden, sondern diese ähnlich einem Aktenbündel oder Ordner auch durchblättern und gegebenenfalls feinbewerten und umstrukturieren könnte. Daraus entstand eine erste Version von „docuteam packer“, einem frei verfügbaren Werkzeug, mit dem Informationspakete gebildet, visualisiert und bearbeitet werden können. Es kann sich sowohl um Ablieferungs- (SIP), Archiv- (AIP) als auch Benutzungspakete (DIP) handeln, weshalb im Folgenden generalisierend von „Paket“ gesprochen wird. docuteam packer wird allerdings hauptsächlich für die Bildung und Bearbeitung von Ablieferungspaketen eingesetzt. Verwendet wird das Werkzeug sowohl in Verwaltungs- und Firmenarchiven wie auch in der Forschungsdatenarchivierung.

Basiert auf offenen Standards

Die Pakete, mit denen docuteam packer umgehen kann, basieren auf dem METS-Standard.[1] Technische und administrative Metadaten werden nach PREMIS abgelegt und die beschreibenden Metadaten nach dem EAD-Standard.[2] Alle drei verwendeten Standards sind offen dokumentiert und weit verbreitet. Zusammen mit dem Staatsarchiv Wallis (CH) haben wir ein METS-Profil ausgearbeitet, das beschreibt, wie die drei Standards aufeinander abgestimmt werden. Das sogenannte „Matterhorn METS“-Profil wurde bei der Library of Congress registriert,[3] auch eine detaillierte Spezifikation des Formats ist verfügbar.[4] Die Matterhorn METS-Pakete können im Ingest-Prozess mit weiteren docuteam-Werkzeugen verarbeitet werden.

docuteam packer eignet sich für die Bildung von Ablieferungspaketen aus Dateiablagen. Von einer vielschichten Ordner- und Dateistruktur wird in packer ein neues Paket angelegt. Das dauert je nach Ablieferungsgröße einen Moment, denn das Werkzeug rechnet von jeder Datei die Checksumme und identifiziert mit Hilfe von DROID[5] die Dateiformate. Wichtige technische Metadaten werden also nicht erst im Ingest (d.h. bei der Überführung ins Archiv), sondern bereits früher beim Anlegen einer neuen Ablieferung protokolliert. Mit Hilfe der Checksummen wird nachgewiesen, dass Dateien beim Transfer ins Archiv nicht mehr verändert wurden. Wenn die neue Ablieferung angelegt ist, dann kann von den gängigsten Dateiformaten eine Vorschau angezeigt werden, so dass die Ablieferung ähnlich einem Papierbündel durchgeblättert werden kann. Damit erhält man auch über große Ablieferungen hinweg rasch eine Übersicht über das gesamte Material.

[1] *http://www.loc.gov/standards/mets/* (aufgerufen am 27.3.2017).
[2] *http://www.loc.gov/standards/premis/v2/index.html* und *http://www.loc.gov/ead/index.html* (aufgerufen am 27.3.2017).
[3] *http://www.loc.gov/standards/mets/profiles/00000041.xml* (aufgerufen am 25.2.2017).
[4] *https://wiki.docuteam.ch/lib/exe/fetch.php?media=oais:spezifikation_matterhorn-mets_20160803_wi.pdf* (aufgerufen am 27.3.2017).
[5] *https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records /droid/* (aufgerufen am 27.3.2017).

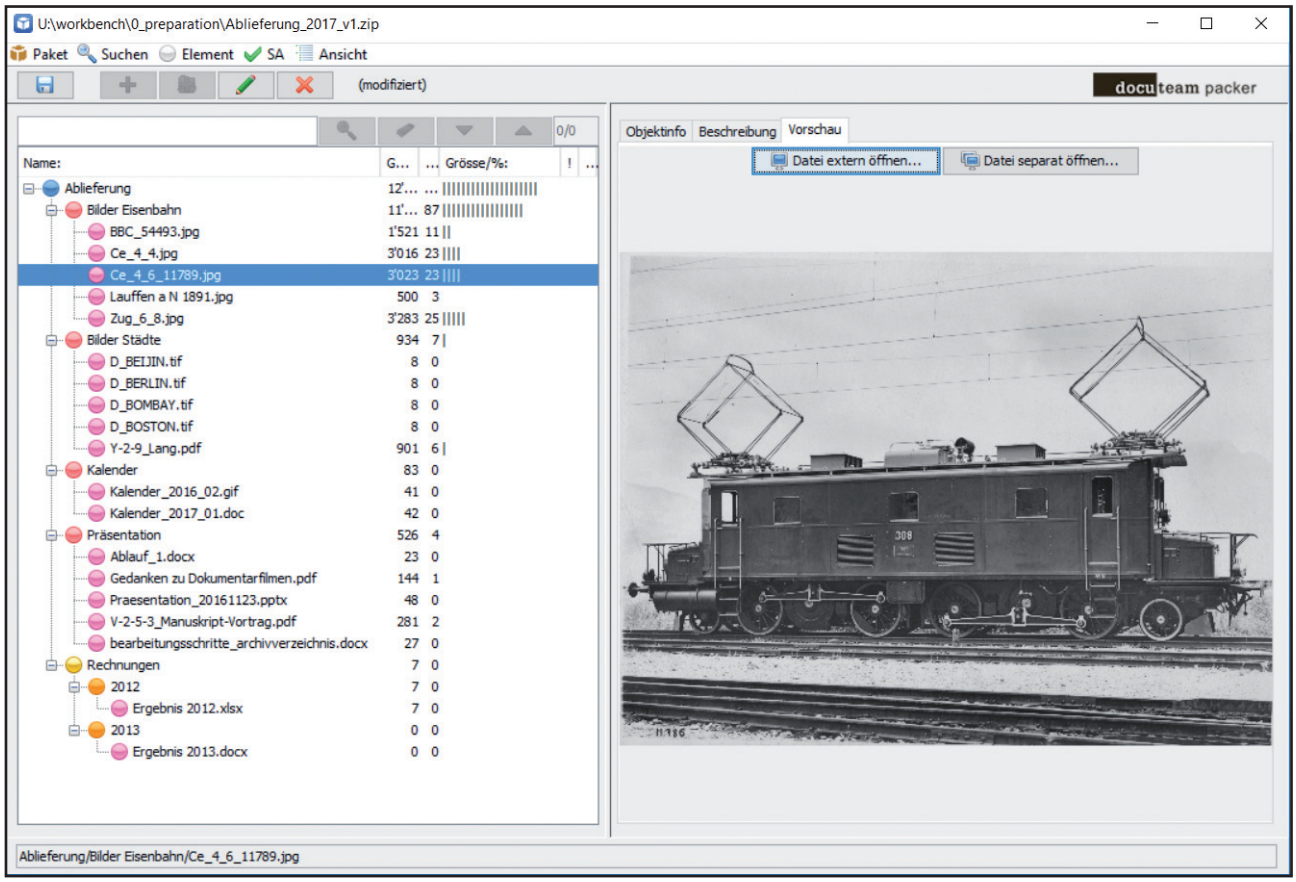


Abb. 1: docuteam packer visualisiert die Struktur einer Ablieferung und zeigt Voransichten der Dateien an

Erlaubt Editierung und Erschließung

docuteam packer erlaubt eine Editierung der Pakete. Dateien und Ordner können umbenannt, gelöscht oder per Drag and Drop an eine andere Stelle im Paket verschoben werden. Wenn weitere Dateien oder Ordner der Ablieferung beigefügt werden sollen, dann geschieht dies am einfachsten ebenfalls über Drag and Drop. Im Hintergrund wird jede dieser Aktionen in den PREMIS Events-Metadaten geloggt. In einem kontrollierten Prozess kann eine Datei durch eine zweite Datei im Sinne einer zusätzlichen Repräsentation ergänzt werden. Diese Funktion kommt in Spezialfällen zur Anwendung, wenn beispielsweise von einer Excel-Datei nicht nur ein PDF/A-Ausdruck archiviert werden soll, sondern auch noch eine CSV-Version.

Zu jedem Ordner und zu jeder Datei im Paket können beschreibende Metadaten erfasst werden. Den verschiedenen Hierarchiestufen in der Ordnerstruktur werden Tektonikstufen (zum Beispiel „Bestand“, „Serie“, „Dossier/Akte“, „Einzelstück“) zugewiesen, wobei sich diese vorgängig konfigurieren lassen. Für die Beschreibung von Ordnern und Dateien wird docuteam packer standardmäßig mit den 26 ISAD(G)-Feldern ausgeliefert. Die Metadatenfelder und die Benennung der Felder können frei konfiguriert werden, solange sie auf dem EAD-Standard basieren. In manchen Kontexten wird docuteam packer mit nur einer Handvoll Metadatenfeldern verwendet, beispielsweise wenn das Werkzeug an einer Hochschule von Forschenden für die Ablieferung von Forschungsdaten an die Hochschulbibliothek genutzt wird. Diese Anpassungsfähigkeit in der Metadatenkonfiguration ist ein wichtiges Merkmal und wird an unserem Werkzeug geschätzt.

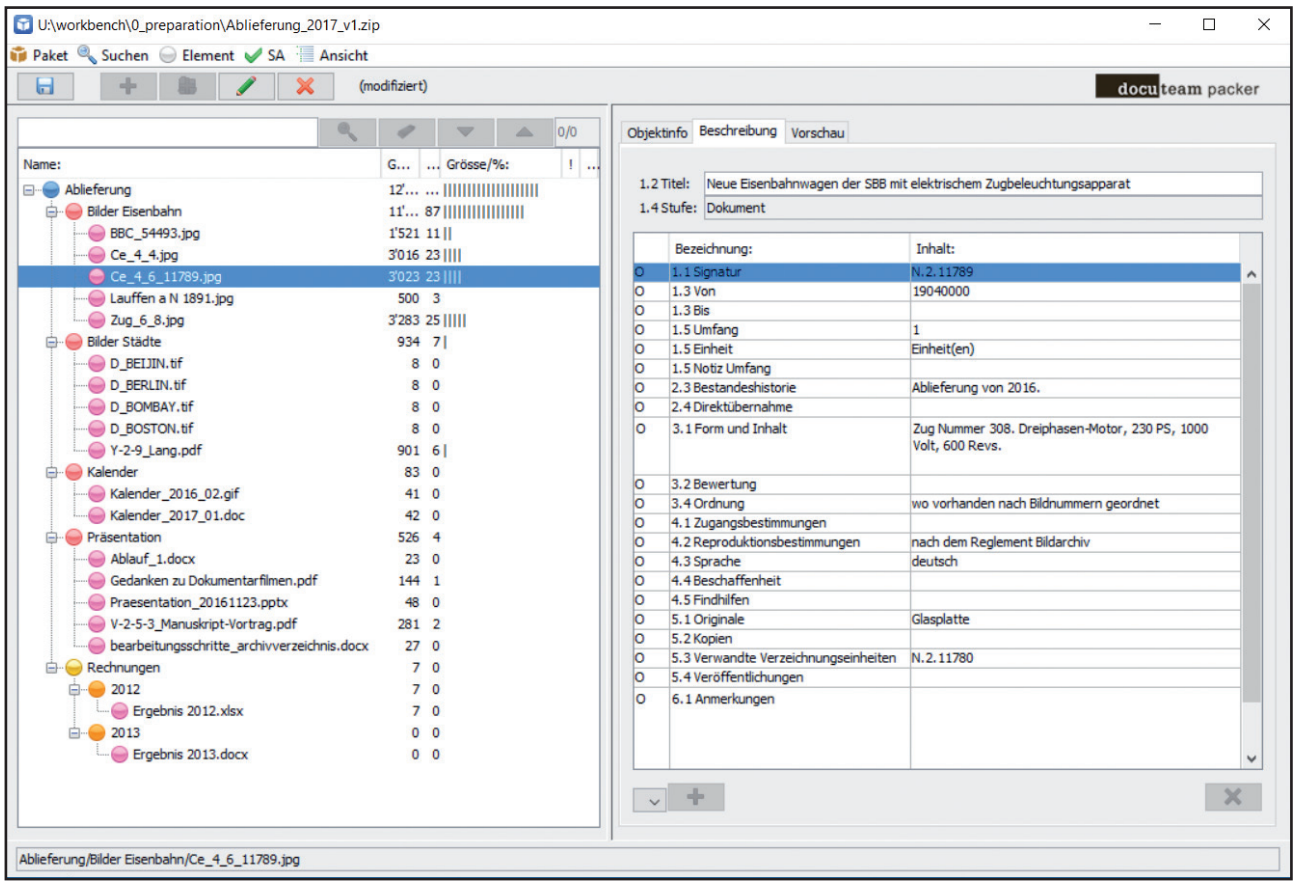


Abb. 2: Auf jeder Verzeichnungsstufe können beschreibende Metadaten erfasst werden, sie werden im EAD-Format gespeichert

Kann SIP direkt dem Ingest übergeben

Die Ablieferung von fertig für die Archivierung vorbereiteten Paketen an den Ingest-Prozess kann über mehrere Wege erfolgen. Mit dem Befehl „Speichern unter“ wird beispielsweise ein Hotfolder als Speicherort ausgewählt, der von unserem Ingest-Dienst „docuteam feeder“ überwacht wird. Wird ein Paket in einen solchen Ordner gespeichert, dann startet automatisch der nachgelagerte Ingest-Prozess. docuteam packer lässt sich aber auch so konfigurieren, dass ein «Submit»-Knopf eingeblendet wird und die Pakete dann über einen SSH-Tunnel verschlüsselt an einen vordefinierten Speicherort verschoben werden. Diese Übermittlungsart wird angewendet, wenn das Werkzeug in einer größeren Organisation zur Anwendung kommt und Pakete ohne große Benutzerschulung ans Archiv abgeliefert werden sollen. Einen Spezialfall stellt die partielle Ablieferung dar. Wenn ein Paket über längere Zeit offen ist, beispielsweise weil es die Daten eines mehrjährigen Forschungsprojekts umfasst, dann sollen in Zwischenschritten zumindest Teile des Pakets ans Archiv abgeliefert werden können. Ordner und Unterordner können für eine Teilablieferung markiert werden, diese wird durchgeführt und die abgelieferten Teile sind darauf für die Bearbeitung gesperrt. So können große Ablieferungen zeitlich gestaffelt in mehreren Etappen erfolgen. Diese Funktion kommt in Verwaltungsarchiven kaum zur Anwendung, ist aber wichtig im Kontext der Forschungsdatenarchivierung.

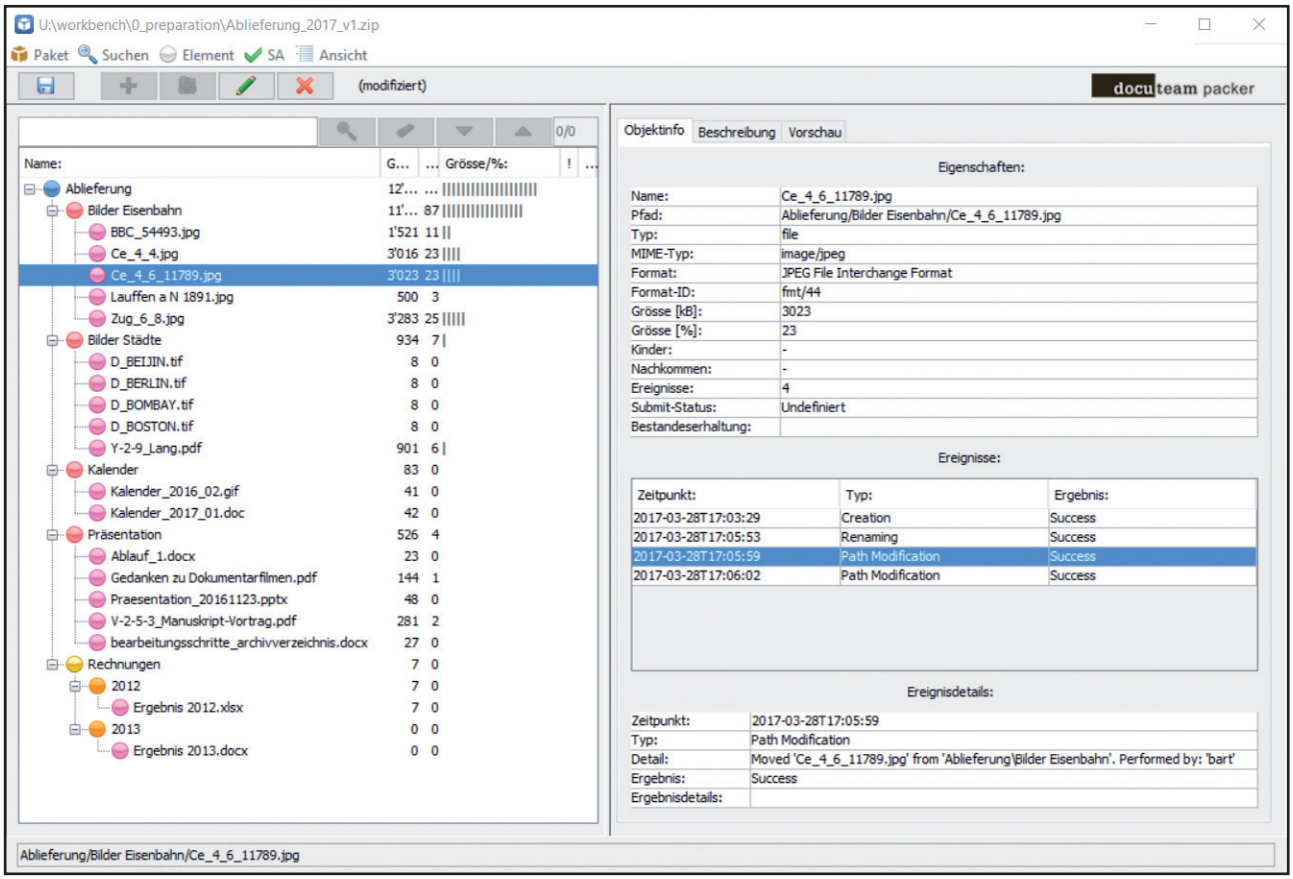


Abb. 3: docuteam packer speichert zu jeder Datei automatisiert eine Anzahl technischer Metadaten und loggt die wichtigsten Aktionen im PREMIS-Format

Ist frei verfügbar und wird laufend weiterentwickelt

docuteam packer steht unter der Open-Source-Lizenz GPL3.[6] Als Bestandteil der docuteam-Werkzeuge für die Umsetzung von OAIS wird er laufend weiterentwickelt. Die aktuelle und die vorangehenden Versionen können frei heruntergeladen werden.[7] Am selben Ort ist auch ein ausführliches Manual für die Benutzung und Konfiguration verfügbar. packer funktioniert „out of the box“, kann per Konfiguration aber auch sehr flexibel an den eigenen Verwaltungs- und Archivkontext angepasst werden. Eine Installation ist nicht notwendig. Nach dem Entpacken der heruntergeladenen Zip-Datei kann docuteam packer auch auf einem USB-Stick betrieben werden. Das ermöglicht seinen Einsatz auch dort, wo die Benutzer nicht die Möglichkeit haben, selbst Software auf ihrem Rechner zu installieren.

docuteam packer entwickeln wir schon seit Jahren weiter. In der Zwischenzeit sind viele neue Funktionalitäten dazugekommen, im Zentrum steht aber nach wie vor der Wunsch nach einem Werkzeug, mit dem man aus Dateisystemablagen möglichst einfach digitale Ablieferungen bilden kann. Insofern leitet uns die Anfrage der eingangs zitierten Archivarin auch heute noch.

[6] *https://www.gnu.org/licenses/gpl-3.0.de.html* (aufgerufen am 27.3.2017).
[7] *https://wiki.docuteam.ch/doku.php?id=docuteam:packer* (aufgerufen am 27.3.2017).

Literatur und Werkzeuge für den Umgang mit kreativen digitalen Ablagen

Kai Naumann

Zur Zeit entstehen ständig weitere Literatur und Tools zum Thema, weshalb ein ergänzender Blick in Online-Bibliographien sinnvoll ist.

Von dieser Aufstellung ausgenommen sind Tools, die in dieser Publikation mit einem eigenen Artikel behandelt werden.

Lehr- und Handbücher allgemein

Adrian Brown, Practical digital preservation. A how-to guide for organisations of any size, London 2013.

Referenzen für Formate und Formaterkennung

Katalog archivischer Datenformate der KOST (Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen)
http://www.kost-ceco.ch/wiki/whelp/KaD/pages/KaD.html

nestor Enzyklopädie 2.3, insbes. Kap. 7.2 *http://nestor.sub.uni-goettingen.de/handbuch*

PRONOM Format Registry der National Archives UK
http://www.nationalarchives.gov.uk/PRONOM

Peter B., Hermann Lewetz, Marion Jaks, Comparing video codecs and containers for archives (Webseite der Österreichischen Mediathek)
http://download.das-werkstatt.com/pb/mthk/info/video/comparison_video_codecs_containers.html#codec_tests

Just Solve the File Format Problem Wiki Das Wiki geht auf eine Initiative von Jason Scott im Jahr 2012 zurück. Es ist recht ausführlich, gerade zu exotischen Formaten.
http://fileformats.archiveteam.org/wiki/Main_Page

Fileformat.info *http://fileformat.info/*. Eine Seite, die man leicht unterschätzt, die aber eine Fülle an Information kanalisiert, auch zu exotischeren Formaten und Codierungen. Der Betreiber sitzt im Staat Pennsylvania (USA).

Tools für Formate und Formaterkennung

FITS verwendet eine Vielzahl erhältlicher Charakterisierungswerkzeuge (derzeit 10), was deren Ergebnisse vergleichbar macht. *http://projects.iq.harvard.edu/fits*

IngestList verwaltet den Übernahmeprozess vom Ausgangssystem bis ins Archiv, es erfasst Dateien und Datenbankauszüge aus Datenbank-Systemen, identifiziert Formate, ermittelt

signifikante Eigenschaften und liefert Dateien mit Metadaten und Protokollinformationen im Archivsystem ab. *http://ingestlist.sf.net*

KOST-Val validiert SIPs inkl. Formatvalidierung (TIFF, JPEG, JPEG2000, SIARD, PDF/A). KOST-Val verwendet dabei unter anderem DROID, PDF/A-Manager und eine beschränkte, aber kostenlose Version des 3-Heights PDF Validator.
http://kost-ceco.ch/cms/index.php?kost_val_de

DROID ermöglicht in der GUI-Version, einen Überblick über die verschiedenen Formate einer Ablieferung zu bekommen.
http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Literatur über Dateisammlungen

AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship (2009–2012), *http://dcs.library.virginia.edu/aims/white-paper/*

ELPAR – die elektronische Parallelregistratur. Kurzdarstellung, 6 Seiten, 2008
http://www.stadtarchiv.mannheim.de/veroeff/schriftgutverwaltung/ELPAR.pdf

Jürgen Enge – Heinz Werner Kramski, „Arme Nachlassverwalter …“ Herausforderungen, Erkenntnisse und Lösungsansätze bei der Aufbereitung komplexer digitaler Dateisammlungen. In: Jörg Filthaut (Hrsg.), Von der Übernahme zur Benutzung. Aktuelle Entwicklungen in der digitalen Archivierung. 18. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“, am 11. und 12. März 2014 in Weimar (Schriften des Thüringischen Hauptstaatsarchivs Weimar 6), Weimar 2015, S. 53–62. Vortragsfolien unter
http://www.staatsarchiv.sg.ch/home/auds/18.html

Niels Hoppe, Das Pre Ingest Toolset (PIT). Ergebnis einer guten Zusammenarbeit der Archive. Entwicklungshistorie und Konzept zur Gesamtlösung. In: Burkhard Nolte – Karsten Huth (Red.), Standards, Neuentwicklungen und Erfahrungen aus der Praxis zur digitalen Archivierung. 17. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ am 13. und 14. März 2013 in Dresden (Veröffentlichungen des Sächsischen Staatsarchivs A 16), Halle/Saale 2014, S. 47–52. Vortragsfolien unter
http://www.staatsarchiv.sg.ch/home/auds/17.html

Karina Jaeger – Maria Kobold, Zwischen Datenwust und arbeitsökonomischer Bewertung. Ein Werkstattbericht zum Umgang mit unstrukturierten Dateisammlungen am Beispiel des Bestandes der Odenwaldschule. In: Archivar 70 (2017) H. 3, S. 307–311.

Corinna Knobloch, Digitale und hybride Quasi-DMS. Befund und Strategiefragen. In: Niels Hoppe, Das Pre Ingest Toolset (PIT) (s. oben), S. 107–118.

Niklas Konzen, Übernahme von E-Akten aus kommunalen Dokumentenmanagementsystemen in das Langzeitarchiv DIMAG: Ein Vorschlag zur praktischen Umsetzung anhand von Fallbeispielen aus den DMS der Stadt Kirchheim unter Teck und des Landratsamts Karlsruhe, (Transferarbeit) Stuttgart 2016.

https://www.landesarchiv-bw.de/sixcms/media.php/120/60857/Transferarbeit2016_Konzen.pdf

Kai Naumann, Dateisammlungen, in: Südwestdeutsche Archivalienkunde (Internetreferenz, erscheint Februar 2018).

Kai Naumann, Digitale und hybride Quasi-DMS. Befund und Strategiefragen. In: Niels Hoppe, Das Pre Ingest Toolset (PIT) (s. oben), S. 99–105.
Vortragsfolien unter *http://www.staatsarchiv.sg.ch/home/auds/17.html*

Ulrich Schludi, Zwischen Records Management und digitaler Archivierung. Das Dateisystem als Basis von Schriftgutverwaltung und Überlieferungsbildung. In: Kai Naumann – Peter Müller (Hrsg.), Das neue Handwerk. Digitales Arbeiten in kleinen und mittleren Archiven, Stuttgart 2013, S. 20–38. *https://www.landesarchiv-bw.de/web/55282*

Heike Simon, Herausforderungen bei der Übernahme von Unterlagen aus Fileablagen. Zum Einsatz des PreIngestToolsets (PIT) im Bundesarchiv, Vortragspräsentation AUdS März 2016, Potsdam, *http://www.staatsarchiv.sg.ch/home/auds/20.html*. Eine Druckfassung erscheint demnächst.

Victoria Sloyan, Born-digital archives at the Wellcome Library: appraisal and sensitivity review of two hard drives. In: Archives and Records 37 (2016) H. 1, S. 20–36, DOI: 10.1080/23257962.2016.1144504, *http://dx.doi.org/10.1080/23257962.2016.1144504*

Isabel Taylor, Eine hydraartige Matrjoschka: Wie wir die Fileablage eines staatlichen Schulamtes bewertet und erschlossen haben, Vortragspräsentation AUdS März 2016, Potsdam, *http://www.staatsarchiv.sg.ch/home/auds/20.html*. Eine Druckfassung erscheint demnächst.

Michael Tobegen, Eine Übernahme von Orthophotos in das Digitale Archiv Nord, Vortragspräsentation AUdS März 2016, Potsdam, *http://www.staatsarchiv.sg.ch/home/auds/20.html*. Eine Druckfassung erscheint demnächst.

Laura Uglean Jackson – Matthew McKinley, It‘s How Many Terabytes? A Case Study on Managing Large Born Digital Audio-Visual Acquisitions. In: International Journal of Digital Curation 11 (2016) S. 64–75. *http://ijdc.net/index.php/ijdc/article/view/11.2.64*

Tools für das Packen und Übertragen von Dateisammlungen

7zip ist eines von vielen Programmen zur verlustfreien Kompression. Es eignet sich, um bei der Übernahme Speicherbedarf und Übertragungsraten überschaubar zu halten. 7Zip hat die Fähigkeit, im Kommandozeilenmodus ein Paket auf Integrität zu prüfen. Hierbei werden die beim Packen gespeicherten Prüfsummen mit dem Inhalt des Pakets verglichen. 7Zip kann auch Pakete mit mehr als 2 GB Inhalt verarbeiten. Darüber hinaus kann 7Zip auch eine (sehr schwache) Verschlüsselung der Inhalte gewährleisten.

Chiasmus ist (für den öffentlichen Dienst kostenlos) beim Bundesamt für Sicherheit in der Informationstechnik erhältlich. Es ermöglicht die Verschlüsselung einzelner Dateien zur Übermittlung per E-Mail oder Datenträger.

Eraser (wie andere vergleichbare Werkzeuge auch) erlaubt das rückstandsfreie Löschen von zu schützenden Daten von Transferdatenträgern. Probleme beim LABW mit Berechtigungen wurden mit dieser Anleitung entschärft:
https://eraser.heidi.ie/forum/threads/eraser-needs-administrative-privileges-to-use.10291/.
Die Software ist verfügbar unter: *http://eraser.sf.net*

Manifest Maker (kostenlos) erzeugt Hashwerte und legt diese nebst den Angaben zum Dateinamen sowie dem Dateipfad in einer strukturierten Textdatei ab.
http://manifestmaker.sourceforge.net/

MD5 File Hasher erstellt von beliebigen Dateizusammenstellungen eine Referenzdatei mit den entsprechenden Prüfsummen. Die Referenzdatei kann später gegen das Original oder Kopien der Dateizusammenstellung abgeglichen werden. Das Ergebnis des Abgleiches weist veränderte, fehlende und neu hinzugekommene Dateien aus. Die Referenzdatei kann bezüglich der neu hinzugekommenen Dateien aktualisiert werden. Anders als in öffentlichen Foren angenommen, überschreitet die Pro-Version keine heute relevante Mengengrenze. Hoher Automatisierungsgrad. *http://www.digital-tronic.de/md5-file-hasher/md5-file-hasher*

md5sum berechnet und kontrolliert MD5-Summen.

VeraCrypt kann große Mengen Dateien sicher verschlüsseln. Falls eingehende Datenträger per Post verschickt werden müssen, ist eine sichere Verschlüsselung angebracht.
https://veracrypt.codeplex.com/

Tools für die Aufbereitung von Dateisammlungen

Adobe Acrobat Professional besitzt seit Version 8 die Fähigkeit, Dokumente auf ihre PDF/A-Konformität zu prüfen. Seit Version 9 ist Acrobat darüber hinaus in vielen Fällen in der Lage, durch Konversion eine PDF/A-Konformität herzustellen.

Autopsy ist ein Open Source Forensiktool, das eine detaillierte Analyse von Festplatten ermöglicht. Z.B. greift es auf die NSRL-Datenbank zu, um Programm- und Systemdateien zu identifizieren, gruppiert Dateien nach Dateiformaten und spürt gelöschte Dateien mittels PhotoRec auf. *http://www.sleuthkit.org./autopsy*

CloneSpy, WinMerge und Duplicate Cleaner schaffen Übersicht in Dateisammlungen mit viel Redundanz, indem sie Dubletten (mehrfach vorkommende Dateien mit gleicher Größe oder gleicher Prüfsumme) auf ein Vorkommen reduzieren. CloneSpy arbeitet mit Prüfsummen und kann Regeln anwenden, an welchem Speicherort eine Datei erhalten bleibt.
http://www.clonespy.com/?Home

CSV2files holt sich Felder aus einer CSV-Datei. Bestimmte Felder können zeilenweise zum Dateinamen deklariert werden, ein anderes zum Inhalt der Datei. Es entstand, um unkonventionell (in Filemaker für Apple MacIntosh) archivierte E-Mails wieder aus der Datenbank herauszuholen. CSV2files ist ein Python-Skript, also muss Python installiert werden. Das Tool ist beim DIMAG-Verbund verfügbar.

Data Accessioner dient zum Bewerten von Dateisammlungen. Es listet den Inhalt einer Dateisammlung auf und macht sich dabei Apache Tika, DROID, ExifTool und JHOVE zunutze. Dublin Core Metadaten können zu jedem Dateiobjekt vergeben werden. Einzelne Dateien oder Ordner können von der Übernahme ausgenommen werden. Vor zu langen Datei-Ordner-Namensketten wird vor dem Kopieren gewarnt. Zuletzt werden die Inhalte in einen Zielbereich kopiert. Eine Bestandsaufnahme im XML-Format kann später zum Weiterverwenden der Metadaten genutzt werden. *http://dataaccessioner.org/*

Dateilistenschreiber und Directory List and Print ermöglichen das Schreiben von Dateiverzeichnissen in eine Textdatei ohne Verwendung der Windows-Kommandozeile.
http://www.sttmedia.de/download=DateilistenSchreiber bzw. *http://www.infonautics.ch/directorylistprint/*

FILEminimizer von balesio erfasst bestimmte Dateitypen in Dateisammlungen und optimiert die Bilddatenströme auf geringeren Speicherplatzbedarf. FILEminimizer Pictures ist kostenlos und erfasst gängige Bilddateitypen. FILEminimizer Suite erfasst auch Office-Formate, PDF und lässt sich in Outlook integrieren.
http://www.balesio.com/fileminimizerpictures/deu/features.php#screenshots

LWL METS-Generator (kostenlos) erzeugt zu regelmäßigen Ordnerstrukturen (v.a. aus Scanwerkstätten) METS-Dateien und ordnet dabei vorhandene Erschließungsdaten anhand Bestellsignatur den Ordnernamen der Digitalisate zu. Der Generator kann auch tief verschachtelte Dateien mit aufnehmen. In semantischer Hinsicht können unregelmäßige baumartige Strukturen aber nicht berücksichtigt werden.
https://www.lwl.org/LWL/Kultur/Archivamt/Archiv_IT/dfg-projekt

Das **National Software Reference Library Reference Data Set (NSRL RDS)** erlaubt das Herausfiltern von Systemdateien aus Dateisammlungen. Es handelt sich um eine Sammlung von Prüfsummen, die die Dateien identifizieren. Es ist aufgeteilt in eine Sammlung bis 1999 und eine Sammlung ab 2000. Eventuell hilfreich bei Nachlässen. Es erfordert einen Server, ist in Autopsy (s. dort) integriert.

Remove Empty Directories (kostenlos) von Jonas John entfernt leere Ordner in einem Verzeichnisbaum. *http://www.jonasjohn.de/red.htm*

Robocopy (kostenlos) Kommandozeilen-Tool von Microsoft zum Kopieren und Synchronisieren von Verzeichnissen. Es bricht beim Kopieren bei Fehlern nicht ab und kann den Erfolg in einer Log-Datei dokumentieren. Seit Windows 7 ist das Tool standardmäßig bei der Windows Installation verfügbar.

Total Commander (Probeversion kostenlos) ist ein Dateimanager mit archivrelevanten Extras (ZIP-Pakete, mehrfaches, auch intelligentes Umbenennen und das Erzeugen von Prüfsummen). *http://www.ghisler.com/deutsch.htm*

TreeSize erlaubt es, eine erste Übersicht über die Datenmengen in Dateisammlungen zu gewinnen. Die TreeSize Free Version ist kostenlos verwendbar, analysiert aber keine Netzlaufwerke, sondern nur den Rechner, auf dem es installiert ist. Die TreeSize Pro-Version schließt eine leistungsfähige und einfach zu bedienende *Dublettensuche* ein.
http://www.jam-software.de

VeraPDF ist ein Open Source Validator für PDF/A in all seinen Ausprägungen. Bisher ist die Verarbeitungsgeschwindigkeit mittelmäßig, weshalb es sich eher zur Zusatzvalidierung und nur für einzelne Dateien eignet. *http://verapdf.org*

Vrenamer (kostenlos) ist ein sehr mächtiges Werkzeug zum kontrollierten Umbenennen von Dateisammlungen. Dank der Möglichkeit, Umbenennungen vorab durchzuspielen und einigen Warnmechanismen komfortabler als Kommandozeilen-Befehle.
http://vrenamer.com/

Unstopable Copier und **ISOBuster** sind Programme, die aus zerkratzten oder teilweise defekten Datenträgern Dateien retten können, indem sie unlesbare Bitfolgen durch Ersatzsequenzen ersetzen. **Achtung**: 1. Diese Herangehensweise empfiehlt sich nur bei Archivalien, deren wesentlicher Gehalt trotz des Fehlens einzelner Bitfolgen erhalten bleibt (z.B. AV-Material). Wo es auf Authentizität und Integrität der Bitfolgen ankommt, ist das Verfahren mit größter Vorsicht anzuwenden! 2. Zu lange Dateipfade auf einem Medium werden u.U. ignoriert, das heißt bestimmte Dateien werden nicht ausgelesen.

Literatur über E-Mail

Corinna Knobloch, Überlegungen zur Übernahme und Archivierung von E-Mail-Konnte. In: Digitale Archivierung. Innovationen – Strategien – Netzwerke. Tagungsband zur 19. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ (Mitteilungen des Österreichischen Staatsarchivs 59/2016), Wien 2016, S. 221–231.
Folien unter *http://www.staatsarchiv.sg.ch/home/auds/19.html*

Corinna Knobloch – Kai Naumann, E-Mails, in: Südwestdeutsche Archivalienkunde (Internetreferenz, erscheint Februar 2018).

Kate Murray, Shaking the Email Format Family Tree. In: *Library of Congress Blog „The Signal“, http://blogs.loc.gov/thesignal/2014/04/shaking-the-email-format-family-tree/*

Gianfranco Pontevolpe – Silvio Salza, General Study 05-Keeping and Preserving E-Mail. In: Inter PARES3 Project, Juni 2009,
http://www.interpares.org/ip3/display_file.cfm?doc=ip3_italy_gs05a_final_report.pdf

Christopher Prom, *Preserving Email. Digital Preservation Coalition (DPC). Technology Watch Report 11-01*, Dezember 2011,
http://dx.doi.org/10.7207/twr11-01 oder *http://www.dpconline.org/publications/technology-watch-reports*

Mike Zuchet, Pilotprojekt zur Langzeitarchivierung digitaler E-Mail-Korrespondenzen des Bundesvorstandes der Vereinten Dienstleistungsgewerkschaft ver.di. In: Christian Keitel – Kai Naumann (Hrsg.), *Digitale Archivierung in der Praxis,* Stuttgart 2013.

Tools für E-Mail

Free Outlook PST File Viewer (kostenlos) ermöglicht das Auslesen und Überprüfen von fremden/übernommenen Outlook-PST-Dateien. *http://www.freeviewer.org/pst/*

MailStore Home (für den privaten Gebrauch kostenlos) überführt E-Mails und Anhänge aus verschiedenen E-Mail-Systemen in eine einheitliche Umgebung und macht diese durchsuchbar. Für eine Konversion in dauerhafte Formate sorgt es nicht.
http://www.mailstore.com/en/mailstore-home-email-archiving.aspx

EPADD ist ein größeres Programmpaket zum Sichern und Nutzen digitaler Archivalien aller Art, das E-Mails unterstützt. *http://library.stanford.edu/projects/epadd/download*

Tools für Bildersammlungen

AsTiffTagViewer zeigt die TIFF-Tags an.

ExifToolGUI kann Fehler in Metadaten (z.B. TIFF) korrigieren, ohne das Bild zu verändern.

FILEminimizer reduziert den Speicherbedarf großer Bildsammlungen. Vgl. unter der Überschrift Dateisammlungen

GIMP ist die Open-Source-Alternative zu Photoshop. Auch GIMP erlaubt die händische Verschneidung von Bildern, um pixelweise Unterschiede festzustellen (Ebenenmodus „Unterschied“).

Ingestamatic (Win) und Shotwell (Linux) sind kostengünstige oder kostenlose Anwendungen zum Ordnen und Erschließen von Digitalbildern.

IrfanView ist ein Viewer, der fast alle Bildformate anzeigen und auch konvertieren kann. Für Archive kostenlos (KOST hat beim Hersteller nachgefragt). Eine Stapelverarbeitung bei der Konvertierung ist möglich. Sogar ganze Verzeichnisbäume können rekursiv in anderen Formaten neu erstellt werden.

KOST-Val validiert TIFF, JPEG2000, JPEG und andere. KOST-Val verwendet dabei unter anderem Jhove, BadPeggy und Jpylyzer.

KOST-Simy erzeugt einen Bildervergleich für die automatische visuelle Kontrolle bei einer Bildkonvertierung (BMP, JPEG, JP2, TIFF, GIF, PNG und teilweise PDF). KOST-Simy verwendet dabei unter anderem ImageMagickCompare und iText.
http://kost-ceco.ch/cms/index.php?kost-simy_de

XnView ist ein Viewer, der fast alle Bildformate anzeigen und auch konvertieren kann. Für gemeinnützige Einrichtungen kostenlos. Eine Stapelverarbeitung bei der Konvertierung ist möglich. *http://www.xnview.com/de/xnview/*

Tools für Intranetseiten

HTTrack ist eine Web-Archivierungssoftware für Windows, die dank einer grafischen Oberfläche erste eigene Versuche beim Archivieren von Internet- oder Intranet-Seiten erlaubt. Das Ergebnis ist eine Dateisammlung, die mit obigen Werkzeugen weiterverarbeitet werden kann. Sind die Seiten im Standardverfahren passwortgeschützt, kann ein Passwort in HTTrack hinterlegt werden. Bei Passwortformularen muss HTTrack aber leider passen.
http://www.httrack.com/

Autorenverzeichnis

Susanne Belovari, University Library, University of Illinois

Dr. Marco Birn, Kreisarchiv Reutlingen

Alexander Herschung, startext GmbH

Karsten Huth, Sächsisches Staatsarchiv

Bart Klein, docuteam GmbH

Dr. Niklas Konzen, Archivschule Marburg

Dr. Annekathrin Miegel, Hessisches Landesarchiv

Christian Fabian Näser, startext GmbH

Dr. Kai Naumann, Landesarchiv Baden-Württemberg

Anne Kathrin Pfeuffer, Stadtarchiv Braunschweig

Dr. Michael Puchta, Generaldirektion der Staatlichen Archive Bayerns

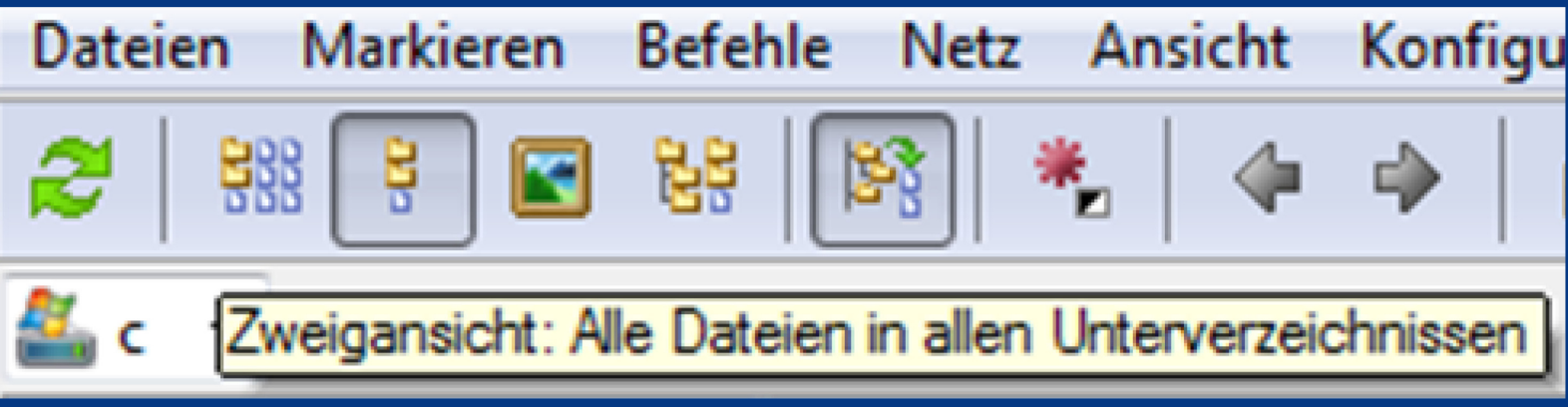
Dr. Sigrid Schieber, Hessisches Landesarchiv

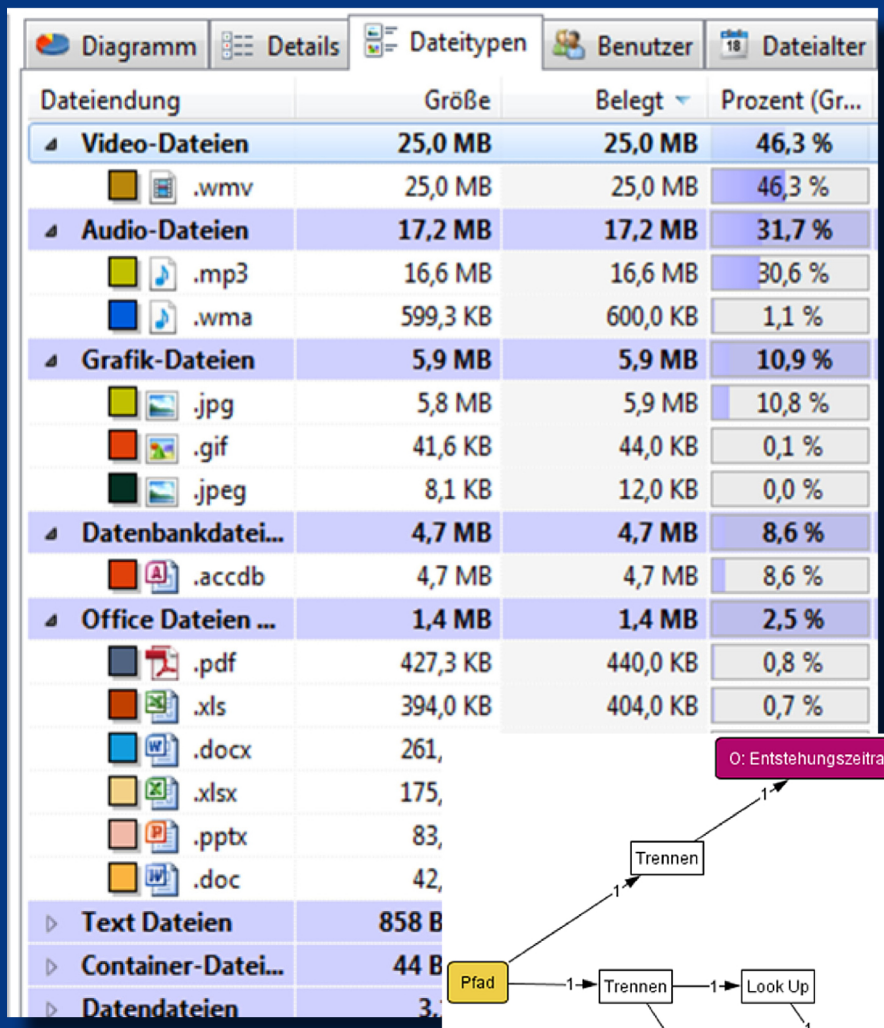
Dr. Christoph Schmidt, Landesarchiv NRW

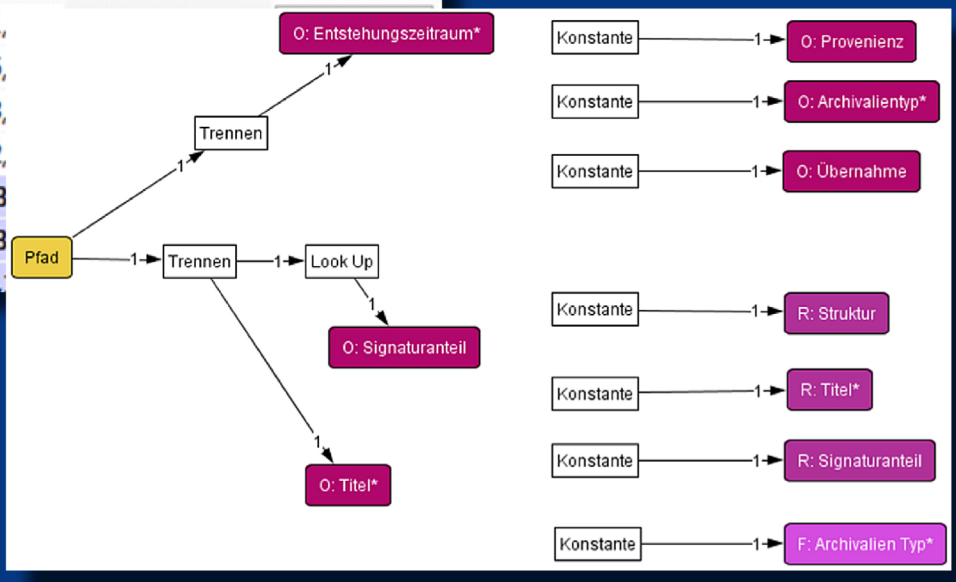
Dr. Kristina Starkloff, Archiv der Max-Planck-Gesellschaft

Andreas Steigmeier, docuteam GmbH

Tobias Wildi, docuteam GmbH







Staatliche Archive Bayerns

KLA Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder

ISSN 1618-0739
ISBN 978-3-938831-81-6