Ordnungsstrukturen von der Floppy zur Festplatte. Zur Vereinnahmung komplexer digitaler Datensammlungen im Archivkontext

Jürgen Enge
Zentrum für Information, Medien und Technologie
Hochschule für angewandte Wissenschaft und Kunst Hildesheim/Holzminden/Göttingen
Goschentor 1
31134 Hildesheim
juergen.enge@hawk-hhg.de

Heinz Werner Kramski
Wissenschaftliche Datenverarbeitung
Deutsches Literaturarchiv Marbach
Schillerhöhe 8–10
71672 Marbach
heinz.werner.kramski@dla-marbach.de

Tabea Lurk
Konservierung & Restaurierung
Hochschule der Künste Bern
Fellerstrasse 11
3027 Bern
tabea.lurk@bfh.ch

Abstract: Der vorliegende Beitrag geht auf die wachsenden Herausforderungen ein, mit denen Archive bei der Übernahme komplexer digitaler Datensammlungen, wie z. B. Nachlässen, konfrontiert sind. Nach einer kurzen Einleitung in die Problematik wird im Rückblick auf digitale Datenzugänge des DLA der letzten 10 Jahre die Übernahme und Vereinnahmung Floppy-basierter Datensammlungen vorgestellt. Während die Handhabung dieser Daten in der Archivvorstufe aufgrund der relativ überschaubaren Datenmenge und Speicherstruktur noch teilweise »händisch« erfolgen kann, wächst die Herausforderung bei Datensammlungen, die ganze Festplatten oder Computersysteme umfassen. Auf ihnen sind neben inhaltlich relevanten Daten der Autorinnen oder Autoren auch Fremddaten abgelegt, die aus Korrespondenzen, der Zusammenarbeit mit anderen Nutzern oder Recherchezwecken resultieren. Hinzu kommen Programm- und Systemdateien, die nicht notwendig mit der Arbeit der Autorinnen oder des Autoren zusammen hängen. Vor allem in Fällen, in denen die Dateneigner mitunter selbst programmiert haben oder an spezifischen Software(-konfigurationen) oder der Rechnerperipherie Hand angelegt haben, wird die Suche der »archivrelevanten« Daten zur Herausforderung. All dies ist beim Bestand »Friedrich Kittler« exemplarisch der Fall. Der zweite Teil des Aufsatzes stellt das softwarebasierte Werkzeug »Indexer« vor, das die Datenanalyse automatisiert und die Inhalte über einen technologisch breit abgestützten Volltext-Index durchsuchbar macht. Auch wenn das Werkzeug klassische

Archivprozesse wie die Selektion und die inhaltliche Beurteilung der Daten keinesfalls ersetzt, kann es die Arbeit in der Archivvorstufe doch grundsätzlich erleichtern.

1 Problemaufriss

Peter Lyman von der UC Berkeley School of Information hat in einer Studie schon für das Jahr 2002 errechnet, dass weltweit neu entstehende Information zu einem ganz überwiegenden Teil auf magnetischen Datenträgern gespeichert wird, und insbesondere Papier praktisch keine Rolle mehr spielt: »Ninety-two percent of new information is stored on magnetic media, primarily hard disks. Film represents 7% of the total, paper 0.01%, and optical media 0.002%.« [Ly03, 1f].

Mit einer gewissen Verzögerung erreicht dieser Trend die Gedächtnisorganisationen, die ihre traditionellen Aufgaben der Bewahrung, Erschließung und Bereitstellung nun auf digitale Objekte ausdehnen. Auch das Deutsche Literaturarchiv Marbach (DLA) schließt digitale Objekte in seinen Sammelauftrag explizit ein.[1] Nachdem in diesem Bereich in den letzten Jahren eine erste Welle an digitalen Zugängen stattgefunden hat und bewältigt wurde, zeigt sich nun, mit fortschreitender Kapazität der überlieferten Datenträger, auch innerhalb der digitalen Sammlungen eine neue qualitative Stufe.

Hier soll es im Folgenden um die »born-digitals« gehen, also um trägergebundene digitale Unikate, die mit Nachlässen, Vorlässen usw. erworben werden, und die aus verschiedenen Gründen besonders problematisch sind [KB11, 142]. Volltexte, die durch Transkription gewonnen werden, Digitalisate analoger Quellen, digitale Dokumente, die online oder offline publiziert werden und auch reine AV-Medien bleiben in diesem Beitrag unberücksichtigt.

1.1 Leistung und Grenzen des bestehenden DLA-Workflows für digitale Nachlassobjekte

Der erste Marbacher Nachlass mit digitalen Bestandteilen war im Jahr 2000 der des Schriftstellers Thomas Strittmatter (1961–1995), der als Dramatiker (Viehjud Levi) und Romanautor (Raabe Baikal) bekannt wurde. Neben 19 Kästen mit konventionellem Papier-Material fanden sich ein Atari Mega ST2 (betriebsfähig), eine externe Festplatte Atari Megafile 30 (defekt) und 43 Disketten (Atari 3,5“ einseitig, 360 KB; Atari 3,5“ doppelseitig, 720 KB; Mac 3,5“ 400 KB Zone Bit Recording; Mac, 3,5“, 1,4 MB).

[1] »Die Sammlungen überliefern Zeugnisse der Entstehung, Verbreitung, Wirkung, Deutung und Erforschung literarischer und geistesgeschichtlich bedeutsamer Werke und des Lebens und Denkens ihrer Autorinnen und Autoren in handschriftlicher und gedruckter, bildlicher und gegenständlicher, audiovisueller und digitaler Form.« [DL13]

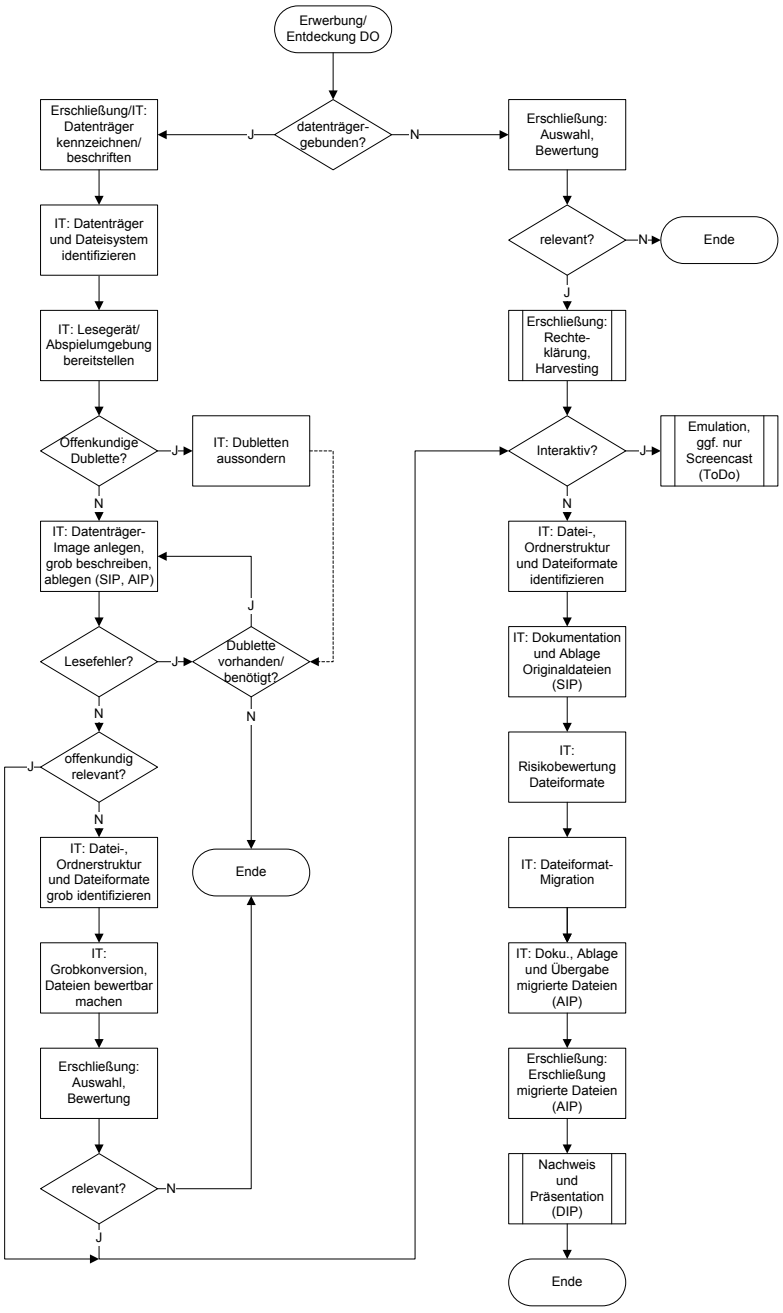


Abbildung 1: Workflow zur Bearbeitung digitaler Nachlassobjekte im DLA Marbach

Das DLA hat an diesem Beispiel einen Workflow zur Erhaltung und Erschließung von digitalen Nachlassobjekten entwickelt und in den Jahren darauf verfeinert, der gut auf statisches, textuelles Material mit überschaubarem Umfang anwendbar ist.[2] Bis 2011 wurden rund 300 Datenträger (überwiegend Disketten) aus 30 Beständen gesichert und ca. 28.000 Dateien in stabile Formate migriert. Abbildung 1 gibt einen Überblick der Abläufe, die im Folgenden dann kurz erläutert werden.

Die Erwerbung erfolgt ähnlich wie bei konventionellem Material durch das Archiv,[3] das heißt, es geht ein physisches Objekt (Hardware, Datenträger) in den Besitz des DLA über.[4] Im nächsten Schritt erfolgen die Bereitstellung einer geeigneten Abspielumgebung und eine erste Sichtung. Hier wird zunächst versucht, offenkundige physischer Dubletten zu identifizieren und auszuscheiden (alle Datenträger bleiben jedoch als potentielle Ausstellungsstücke und als Reserve im Fall von Lesefehlern erhalten).

Im nächsten Schritt wird eine Sektor-Image-Kopie des gesamten Datenträgers angelegt. Hier kommen selbstgeschriebene Scripte und im Wesentlichen das Tool »ddrescue« unter Cygwin zum Einsatz, gelegentlich, bei wichtigen, fehlerhaften Medien oder besonderen Diskettenformaten auch die Hardware-Software-Kombination »Kryoflux« [Kr13]. In diesem Schritt werden auch elementare deskriptive und technische Metadaten, eine MD5-Prüfsumme und ein rekursives Dateilisting nach einer 2002 selbst entworfenen (XML-)Konvention festgehalten. Die Ablage erfolgt im Dateisystem in einem Ordner »0_Original-Disk«, wobei Unterordner »disk01« usw. die einzelnen Datenträger als Gliederungsprinzip erhalten. Das DLA praktiziert also die Erhaltung der Informationsobjekte durch Trennen von ihrem ursprünglichen Träger.

In einem weiteren Ordner »1_Original« werden anschließend Kopien der lesbaren Originaldateien abgelegt.[5] Er bildet die Grundlage des Ordners »2_Konvertiert«, der formatmigrierte, langzeitstabile Entsprechungen der Originaldateien aufnimmt. Je nach Ausgangsmaterial kommt hier PDF/A oder CSV zum Einsatz (frühe Konversionen liegen nur als RTF vor). Fotos im JPG-Format werden nicht konvertiert. Dies ist ein arbeitsintensiver Schritt, der viele manuelle Einstellungen der verschiedensten Konvertierprogramme erfordert. (Der Bearbeitungsaufwand bis zu dieser Stufe liegt erfahrungsgemäß durchschnittlich bei ca. zwei Stunden pro Diskette.)

[2] Für Details siehe [KB11].
[3] Hier ist die Abteilung »Archiv« des DLA gemeint.
[4] Ben Goldman macht darauf aufmerksam, dass damit noch keinesfalls eine Akzessionierung im Sinne einer intellektuellen Aneignung und Bewertung stattfindet: »As far as our internal administration was concerned, these disks [floppy disks, zip disks, CDs and DVDs] were already accessioned, usually as part of much larger, mostly paper-based collections and following protocols established for analog collections. But this only makes sense logically if you consider disks – or digital media of any sort – to be items in collections, deserving of the same consideration we might give to individual documents. It is more appropriate, I submit, to think of digital media as containers of items which require the kind of archival administration we might normally reserve for boxes in a collection. In this sense, the data (files and folders) found in these containers had not been accessioned at all.« [Go11]
[5] Gelöschte Dateien, für die sich die Editionsphilologie teilweise auch interessiert (vgl. [Ri10]) sind nicht Gegenstand des Standard-Workflows. Sie können aber bei Bedarf aus den Volume-Images gewonnen werden. Forensische Information unterhalb der Ebene der erstellten Sektor-Images (z.B. magnetische Flusswechsel) werden nicht erhalten. Hier musste eine pragmatische Entscheidung getroffen werden, was als »signifikante Eigenschaft« gelten kann.

Schließlich wird eine Kopie des gesamten Ordners »2_Konvertiert«, als »3_Geordnet« an die Abteilung Archiv übergeben, die nur dort Schreibrechte auf die Dateien besitzt. Sie ordnet die Dateien nach inhaltlichen (gattungsbezogenen) Kategorien des Hausstandards »Memo«, beschreibt sie in dem zentralen Nachweisinstrument »Kallías« und stellt Verknüpfungen zu sogenannten Multimedia-Sätzen her, über die sich die digitalen Dokumente von berechtigten Nutzern in Kallías öffnen und anzeigen lassen. Diese Stufe ist jedoch erst für einen kleinen Teil des digitalen Bestandes umgesetzt.

Der bisherige Workflow stellt vor allen die Bitstream-Erhaltung der gefährdeten Datenträger sicher und gewährleistet die Formatmigration der enthaltenen statischen Dateien. Da die erstellten Volume-Images auch in virtuellen Maschinen gemountet werden können, ist gleichzeitig die Grundlage für Emulationsansätze gelegt, die jedoch noch am Anfang stehen.

Es gibt einige systematische Mängel, die in einem geplanten DFG-Projekt ausgeräumt werden sollen, etwa die fehlende Orientierung an Standards oder die Tatsache, dass zwar Prüfsummen und technische Metadaten zu Datenträgern, nicht aber zu einzelnen Dateien systematisch festgehalten werden. Auch stand bisher die reine Sicherung im Vordergrund; eine Präsentation digitaler Objekte für die lokale Benutzung, die auch die urheber- und persönlichkeitsrechtlichen Einschränkungen individuell berücksichtigt, ist noch ein Desiderat.

Ein grundsätzliches Problem besteht aber darin, dass das an wenigen Disketten entwickelte Verfahren nicht für große Datenmengen skaliert. An mehreren Punkten des Workflows ist eine Entscheidung notwendig, welches Material als relevant anzusehen ist und den weiteren Aufwand rechtfertigt: diese ist bisher eher implizit gefallen, etwa schon bei der Übergabe einiger eindeutig beschrifteter Disketten an das Referat »Wissenschaftliche Datenverarbeitung«. Bei größeren, unübersichtlichen Datenmengen wird das Dilemma besonders deutlich, dass die Relevanz von vielen Dateien nicht ohne aufwändige Analyse- und Konvertierarbeiten beurteilt werden kann, die man sich für irrelevantes Material eigentlich sparen muss.

1.2 Der Nachlass Friedrich Kittlers als Paradigma neuer Herausforderungen

Mit dem digitalen Nachlass des Medienwissenschaftlers Friedrich Kittler (1943–2011) stellen sich nun ganz konkret quantitativ und qualitativ neue Fragen. Der Nachlass umfasst nach heutigem Stand mindestens fünf PCs unterschiedlichen Alters aus Wohnung und Büro. Diese sind teils mit ihren Festplatten bereits als Hardware in Marbach, teils nur als Festplatten-Image. Der Hauptrechner ist noch in Berlin, weil er für die geplante Edition der selbstgeschriebenen Software noch als Hardware-Referenz benötigt wird.[6] Dabei handelt es sich nicht um »einfache« DOS- oder Windows-PCs, sondern überwiegend um von Kittler und seinen Mitarbeitern selbst angepasste, einander ablösende

[6] Dass Kittler auch selbst (Grafik-)Programme geschrieben hat, die sich einer einfachen Formatmigration entziehen und die per Emulation erhalten werden müssen, wird in diesem Beitrag nur insofern berücksichtigt, als Kittler-Quelltexte als solche z.B. von mitgelieferten Musterlösungen der Entwicklungsumgebungen unterschieden werden müssen.

Linux-Installationen, die aber auch ältere MS-DOS-Partitionen mit früheren Versionsständen seiner Quelltexte und wissenschaftlichen Beiträge etc. mitführen. Bis auf zwei ältere SCSI-Platten mit SGI-Disklabeln, die aus einer Workstation stammen, konnten die meisten Partitionen inzwischen erfolgreich unter VMware gemountet und einer ersten Sichtung unterzogen werden. Somit kann jede weitere (auch maschinelle) Analyse zumindest unabhängig von der Original-Hardware stattfinden, zumal diese teilweise nur noch mit langen Timeouts und besorgniserregenden Geräuschen startet.

Der Festplattenbestand wird begleitet von 330 3,5“- und 6 5,25“-Disketten mit FAT-, ext2- und Minix-Dateisystemen in recht gutem Zustand sowie von 94 überwiegend selbst gebrannten optischen Medien, die sehr viele Lesefehler aufweisen. Nur ein kleiner Teil der Datenträger konnte bisher eindeutig als Massenware (z. B. c’t-Beilagen) oder als vorkonfektionierte Installations- und Treibermedien identifiziert werden. Ein großer Teil scheint wiederum Datensicherungen von DOS- und Linux-PCs zu verschiedenen Zeitpunkten zu enthalten, wobei es sich sowohl um installierte Anwendungen und Entwicklungswerkszeuge, als auch um Dateien »von Kittlers Hand« handeln kann. Auch Zusendungen von anderen Personen sind darunter. Während von fast allen magnetischen und optischen Medien rekursive Dateilistings möglich waren, steht die Image-Kopie der Disketten noch aus. Das bewährte DLA-Tool »FloppImg« wird dabei wegen des hohen Anteils an ext2-Dateisystemen nicht zum Einsatz kommen können.

Die Zahl der Kittler-Medien übersteigt also das gesamte Archiv digitaler Nachlassobjekte der letzten 10 Jahre. Die Anzahl von Dateien, die gesichtet und klassifiziert werden müssen, liegt schon jetzt schätzungsweise über 1,6 Millionen, obwohl noch nicht alle Volumes zugänglich sind. Die Sichtung und Klassifikation – die ja der eigentlichen Relevanzbeurteilung und Formatmigration vorausgehen muss – wird auch dadurch erschwert, dass Kittler nicht mit Standardverzeichnissen wie »/home« gearbeitet hat, sondern seine Dateien (immer als »root«) z. B. in »/usr/ich« abgelegt hat. Es ist daher nicht auszuschließen, dass sich auch sonst in der Dateisystemhierarchie individuelle Spuren finden, die erhalten werden müssen. Auch bei der Dateibenennung geht Kittler eigene Wege: ».doc« ist oft nicht das, was heutige Anwender erwarten, und Textdateien treten auch mit den Extensions ».utf« oder ».lat« auf, was wohl den Zeichensatz wiedergibt.

Es ist daher klar, dass sich die weiteren Bearbeitungsschritte auf Software-Werkzeuge stützen müssen und auch die eigentliche Erschließung nicht mehr im klassischen Verfahren stattfinden kann, sondern wahrscheinlich im Dialog von Forschern und Archiv z. B. in speziellen, projekthaften Erschließungsgruppen.

Besonders folgende Software-Funktionen wären hilfreich:

- IDs und Prüfsummen für Einzeldateien vergeben und erzeugen
- echte Datei-Dubletten erkennen und ausscheiden
- MIME-Typen trotz ungewöhnlicher Extensions erkennen
- Dateien kennzeichnen, die mit hoher Wahrscheinlichkeit von Kittler stammen (Hinweise liefern z. B. Speicherorte, MIME-Typen usw.)

- insbesondere Musterprogramme der Entwicklungsumgebungen und Libraries von Quelltexten Kittlers unterscheiden (Hinweise liefern z. B. im Quelltext enthaltene Kommentare)
- im Gegenzug Systemdateien und Standard-Software kennzeichnen, um sie ausscheiden/ausblenden zu können (Hinweise liefern z. B. sekundengenaue Häufungen von Änderungsdaten)
- insbesondere MS-DOS-, MS-Windows-, Linux-Konsol- und Linux-X11-Executables erkennen, um z. B. für Ausstellungen gezielt Emulationen aufbauen zu können
- mit (alten) Viren infizierte Executables erkennen und kennzeichnen
- die Änderungshistorie einzelner Dateien innerhalb der komplexen Überlieferung von Platten und Datensicherungen erkennen und darstellen
- den Werkzusammenhang innerhalb der komplexen Überlieferung erkennbar machen bzw. Erschließungserkenntnisse festhalten und Annotationen ermöglichen
- vertrauliche Dateien als solche kennzeichnen und ausblenden.

2 Lösungsansatz

Der Prototyp des Software-Werkzeugs »Indexer« bewältigt einige der zuvor erwähnten Anforderungen, indem er eine Reihe an technischen Analyseverfahren bereitstellt, die nacheinander abgearbeitet werden. Die ineinandergreifenden Arbeitsroutinen operieren teilweise redundant, um die Qualität der Ergebnisse zu verbessern. Sie lassen sich generell in drei Bereiche unterteilen: Zunächst wird ein initiales Verzeichnisses über alle Daten erstellt, es folgt die Ausführung einer sogenannten »Identifikations-Kaskade« und schließlich wird eine Such- und Nutzeroberfläche mit Volltextindex erzeugt, auf welche das webbasierte Such- und Nutzerinterface zugreift.

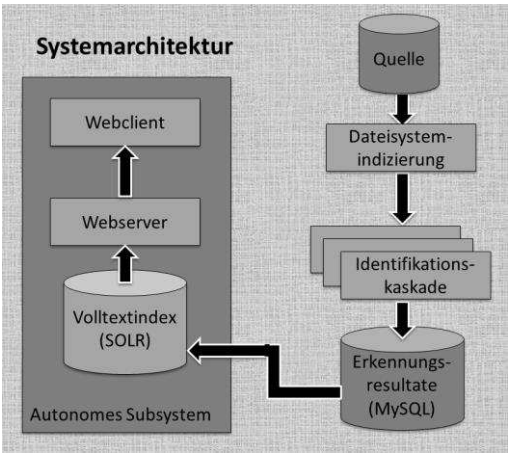


Abbildung 2: Systemarchitektur »Indexer«

Alle Verfahren operieren insofern »archivkonform«, als die Authentizität und Integrität der Daten nicht tangiert wird. Das System hinterlässt also keinerlei eigene Daten innerhalb der komplexen digitalen Archivalie(n). Sämtliche Metadaten, die im Rahmen des Analyseprozesses entstehen, werden gemeinsam mit der Information des Zugriffspfads in parallelen Datenhaltungssystemen abgelegt.

2.1 Erfassung und Indizierung

Prinzipiell benötigt der Indexer »lesenden« Zugriff auf die Dateisysteme der Archivalie.[7] Häufig wird das zu untersuchende Disk-Image daher in Dateiform beispielsweise in einer virtuellen Maschine gemounted. Im ersten Schritt wird dann das Dateisystem der Disk eingelesen. Dazu erlaubt der Indexer die Angabe einer sogenannten »SessionID «. Die SessionID ermöglicht es, verschiedene Dateisysteme mit eigenen IDs in die Indexer-Datenbank zu übernehmen, so dass beispielsweise mehrere unterschiedliche Nachlässe oder Objektgruppen später auch separat betrachtet werden können. Die Tabelle mit den grundlegenden Dateiinformationen enthält folgende Angaben:

- sessionid: die ID des Archivierungsdurchgangs
- fileid: Eindeutige Datei-Identifikationsnummer innerhalb einer Session
- parentid: ID des Dateiordners, in der der Verzeichniseintrag zu finden ist
- name: Datei- oder Ordnername
- path: Pfad des Verzeichniseintrags
- filetype: Typ, wobei unter der Typ der Datei, Verzeichnis, Verweis angegeben werden
- filesize: Dateigröße
- sha256: Prüfsumme (kann auch zur Authentizitätsprüfung weiterverwertet werden)
- filectime: Erstellungsdatum
- filemtime: Änderungsdatum
- fileatime: Datum des letzten Zugriffs (Achtung, diese Angaben sind häufig falsch. Fehler entstehen, wenn Dateisysteme in beschreibbarem Modus gemounted wurden!)
- stat: sämtliche Informationen des Unix-Aufrufs stat()[8]
- archivetime: Zeitpunkt der Indizierung.

[7] Die Dateisysteme werden hierzu unter Linux mit Hilfe des »mount«-Befehls read-only verfügbar gemacht. Das Dateisystem der Quelldaten ist dabei irrelevant, solange es vom lesenden Linux System unterstützt wird.
[8] Dass durch den stat()-Aufruf teilweise redundante Daten entstehen, die in früheren Informationen bereits enthalten sind, soll hier nicht stören.

Als eineindeutiger Identifikator (Signatur) für die einzelnen Dateien wird die Kombination aus Session-ID und FileID verwendet.

2.2 Identifikations-Kaskade

Nach der Erfassung und Vergabe der eineindeutigen Identifikatoren folgt eine »Identifikations-Kaskade« zur Erstellung des Volltextindexes, bei welcher ausgewählte, nacheinander geschaltete Analysewerkzeuge schrittweise angewandt werden. Um die Erkennungsqualitäten zu verbessern, werden gezielt partiell redundante Werkzeuge (Tools) eingesetzt. Die unterschiedlichen Werkzeuge sind auf spezifische Aspekte der Formaterkennung sowie unterschiedliche Formate spezialisiert und können damit auch Formate identifizieren, die auf den ersten Blick unklar scheinen.

Bereits bei der initialen Indizierung der Dateien wird als Identifikationsbibliothek libmagic angewendet, welches versucht, den MIME-Type und das Encoding festzustellen.

sessionid	fileid	mimetype	mimeencoding	description
13	2034350	text/rtf	us-ascii	Rich Text Format data, version 1, ANSI
13	2034351	application/msword	binary	Microsoft Word Document
13	2034352	application/msword	application/mswordbinary	CDF V2 Document, Little Endian, Os: Windows, Versi...
13	2034353	application/msword	binary	Microsoft Word Document
13	2034354	text/rtf	us-ascii	Rich Text Format data, version 1, ANSI
13	2034355	application/msword	application/mswordbinary	CDF V2 Document, Little Endian, Os: Windows, Versi...
13	2034356	text/rtf	us-ascii	Rich Text Format data, version 1, ANSI
13	2034357	application/msword	binary	Microsoft Word Document
13	2034358	application/octet-stream	binary	data
13	2034359	text/x-tex	unknown-8bit	LaTeX document text
13	2034360	application/msword	application/mswordbinary	CDF V2 Document, Little Endian, Os: Windows, Versi...
13	2034361	text/rtf	us-ascii	Rich Text Format data, version 1, ANSI
13	2034362	application/msword	binary	Microsoft Word Document
13	2034363	application/msword	application/mswordbinary	CDF V2 Document, Little Endian, Os: Windows, Versi...
13	2034364	text/rtf	us-ascii	Rich Text Format data, version 1, ANSI
13	2034365	application/msword	binary	Microsoft Word Document

Abbildung 3: Datenbankauszug der libmagic Erkennung

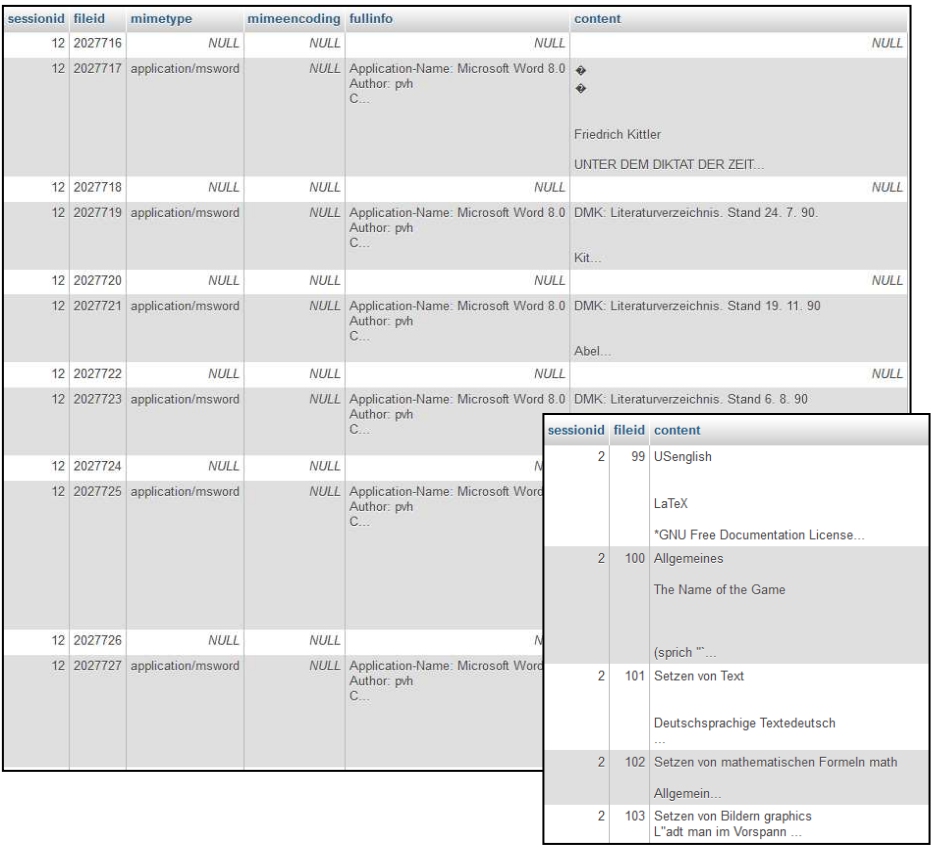
Im nächsten Schritt wird die MIME-Type-Erkennung des gvfs-info-Tools eingesetzt, um eine »zweite Meinung« einzuholen.

sessionid	fileid	mimetype	fullinfo
13	2034355	application/msword	display name: t_realti.doc edit name: t_realti.doc...
13	2034356	application/rtf	display name: t_realti.rtf edit name: t_realti.rtf...
13	2034357	text/plain	display name: t_realti.txt edit name: t_realti.txt...
13	2034358	application/octet-stream	display name: t_schrif.dfv edit name: t_schrif.dfv...
13	2034359	text/x-tex	display name: t_sonder.tex edit name: t_sonder.tex...
13	2034360	application/msword	display name: t_sra.doc edit name: t_sra.doc name:...
13	2034361	application/rtf	display name: t_sra.rtf edit name: t_sra.rtf name:...
13	2034362	text/plain	display name: t_sra.txt edit name: t_sra.txt name:...

Abbildung 4: Datenbankauszug der gvfs-info Erkennung

Ein etwas komplexeres Werkzeug wird im dritten Schritt mit Apache Tika eingesetzt. In diesem Durchgang wird neben der MIME-Type-Erkennung und der Analyse des Encodings bei Texten auch gleich der Volltext extrahiert und in die zugehörige Datenbanktabelle geschrieben. Hier ist zwar die Rate der Fehl-Erkennungen geringer als bei den vorherigen Werkzeugen und auch die Erkennungsrate ist insgesamt etwas schlechter, allerdings kommt der Volltext-Extraktion im Weiteren eine zentrale Rolle zu.

Da der von Apache Tika extrahierte Volltext häufig keine direkt nutzbare Basis für das »Mining« im Archiv darstellt, können weitere Volltext-Extrahierer eingesetzt werden. Im Prototyp des Indexers ist zum Beispiel detex im Einsatz, der Texte aus Dateien des MIME-Types »text/x-tex« Inhalte extrahiert. Bei dieser Extraktion werden alle TeX-Kommandos entfernt, um den für die Volltextsuche semantisch relevanten Textanteil herauszufiltern. Das bedeutet, dass der rohe Text ohne Formatanweisungen zur Recherche verwendet werden kann.



sessionid	fileid	mimetype	mimeencoding	fullinfo	content
12	2027716	*NULL*	*NULL*	*NULL*	*NULL*
12	2027717	application/msword	*NULL*	Application-Name: Microsoft Word 8.0 Author: pvh C…	 Friedrich Kittler UNTER DEM DIKTAT DER ZEIT…
12	2027718	*NULL*	*NULL*	*NULL*	*NULL*
12	2027719	application/msword	*NULL*	Application-Name: Microsoft Word 8.0 Author: pvh C…	DMK: Literaturverzeichnis. Stand 24. 7. 90. Kit…
12	2027720	*NULL*	*NULL*	*NULL*	*NULL*
12	2027721	application/msword	*NULL*	Application-Name: Microsoft Word 8.0 Author: pvh C…	DMK: Literaturverzeichnis. Stand 19. 11. 90 Abel…
12	2027722	*NULL*	*NULL*	*NULL*	*NULL*
12	2027723	application/msword	*NULL*	Application-Name: Microsoft Word 8.0 Author: pvh C…	DMK: Literaturverzeichnis. Stand 6. 8. 90
12	2027724	*NULL*	*NULL*	*N*	
12	2027725	application/msword	*NULL*	Application-Name: Microsoft Word Author: pvh C…	
12	2027726	*NULL*	*NULL*	*N*	
12	2027727	application/msword	*NULL*	Application-Name: Microsoft Word Author: pvh C…	

sessionid	fileid	content
2	99	USenglish LaTeX *GNU Free Documentation License…
2	100	Allgemeines The Name of the Game (sprich "`…
2	101	Setzen von Text Deutschsprachige Textedeutsch …
2	102	Setzen von mathematischen Formeln math Allgemein…
2	103	Setzen von Bildern graphics L"adt man im Vorspann …

Abbildung 5: Datenbankauszüge der Tika- und detex-Erkennung

Die unterschiedlichen Resultate der »generischen« Erkennungswerkzeuge lassen sich auf die verschiedenen Erkennungsalgorithmen und -datenbanken zurückführen. In widersprüchlichen Fällen ist häufig eine Einzelentscheidung durch den Nutzer/das Archiv nötig.

Die nächsten Erkennungsschritte setzen auf die »Erkenntnisse« der vorherigen Durchläufe auf und verfeinern die Resultate durch weitere Informationen. So werden nun auch die technischen Metadaten jener zeitbasierten Medien erfasst, deren MIME-Type von gvfs-info als »video/*« oder »audio/*« erkannt wurde. Sie werden weiter mit Hilfe des Programms avconv (früher ffmpeg) untersucht, wobei detaillierte technische AV-Metadaten extrahiert werden. Zudem werden nun auch Thumbnails für die Indexer-Oberfläche generiert. Bei Videos wird automatisch ein Screenshot erzeugt und für Audiodateien ein Sonogramm (Eigenentwicklung) generiert.

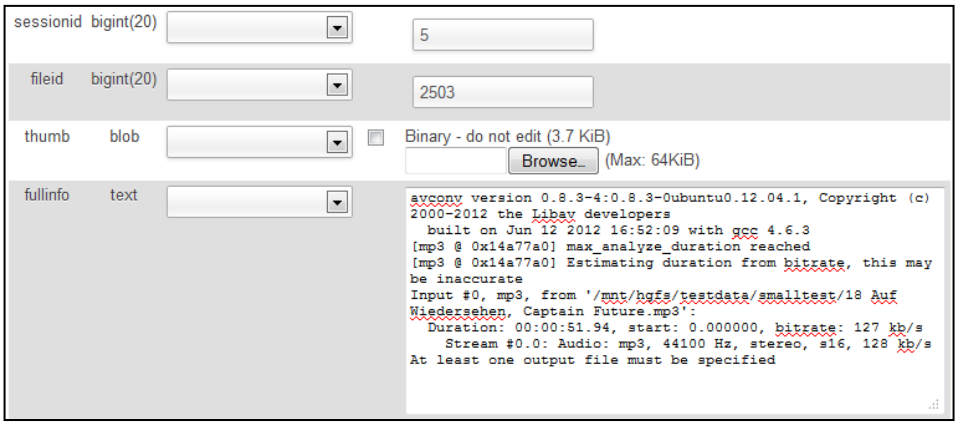


Abbildung 6: Datenbankeintrag von AVCONV

Bild- und PDF-Daten werden mit Hilfe von ImageMagick analysiert. Das Tool bindet alle Dateien mit dem MIME-Type »image/*« und »application/pdf« ein. Ähnlich wie bei AV-Daten wird auch hier ein Thumbnail für die Such-Oberfläche erzeugt.

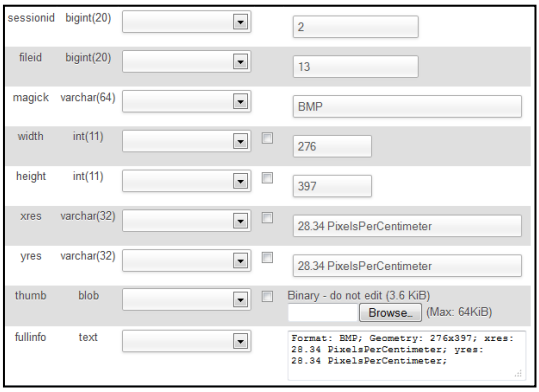


Abbildung 7: Datenbankeintrag ImageMagick

Um das System möglichst flexibel und erweiterbar zu halten, ist die Identifikations-Kaskade des Indexers problemlos erweiterbar. So kann sichergestellt werden, dass sowohl künftige Erkennungswerkzeuge als auch neue Daten- und Formattypen bearbeitet werden können, ohne dass gravierende Veränderungen nötig wären.

2.3 Rechercheinterface

Im Anschluss an die Identifikations-Kaskade wird aus den erkannten und extrahierten Daten ein SOLR[9]-Volltextindex generiert. Er bildet die Voraussetzung für eine Rechercheoberfläche, die leicht handhabbar ist. Die Suchoberfläche lehnt sich an die Erscheinung und Funktionalität gängiger Suchmaschinen an. Über sie erhält der Nutzer Zugriff auf den Volltextindex, wobei je nach Archiv-Policy entweder innerhalb vordefinierter Felder gesucht oder frei recherchiert werden kann. Um das schnelle Erfassen der Inhalte zu erleichtern, werden zu den Treffern neben den extrahierten Metadaten auch die zuvor erzeugten Screenshots, Sonogramme und Textauszüge ausgegeben.

Abbildung 8: Rechercheoberfläche Volltextindex

Der Indexer wird bei der Wiedergabe seiner Inhalte insofern der archivarischen Forderung nach Transparenz, Nachvollziehbarkeit und Reversibilität gerecht, als die Ergebnisse der jeweiligen Analysewerkzeuge und die Skala ihrer Erkennungsrate angezeigt werden können und für den Nutzer somit jederzeit direkt einsehbar sind.

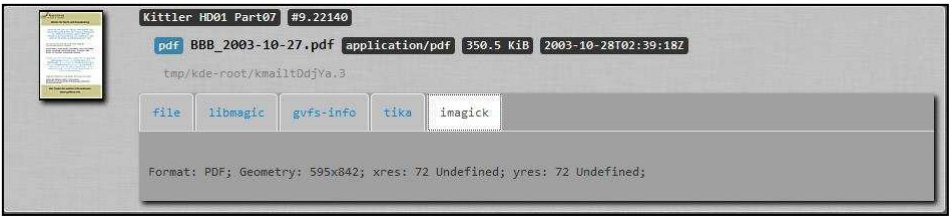


Abbildung 9: Erkennungskaskade – imagick

[9] http://lucene.apache.org/solr/

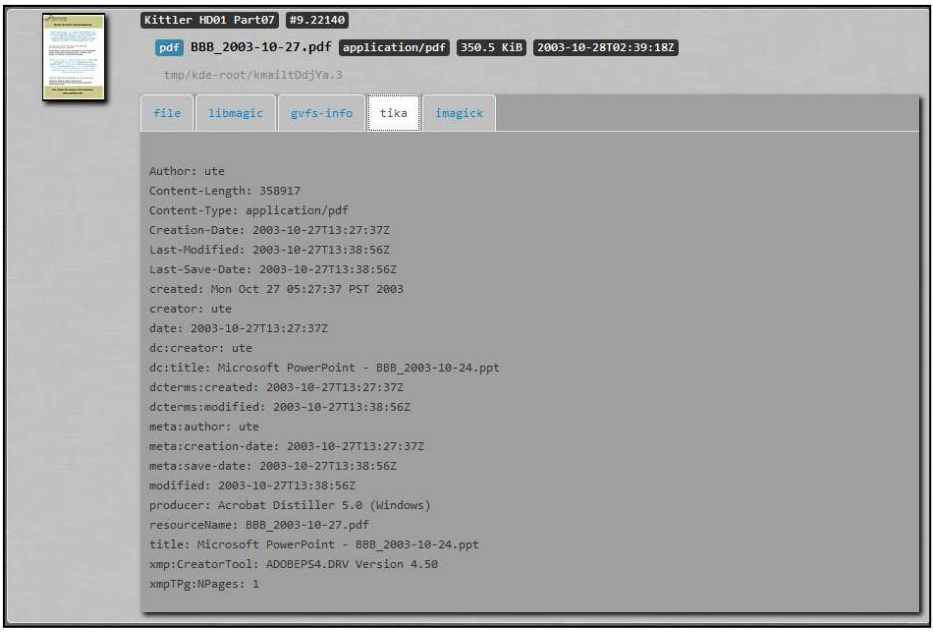


Abbildung 10: Erkennungskaskade – tika

Schließlich wird die zuvor geschilderte Identifikations-Kaskade künftig auch dazu beitragen können, Muster und Ähnlichkeitsstrukturen von Dateien sowie Speicherstrukturen zu erkennen.

Bevor in späteren Arbeitsschritten beispielsweise mittels »pattern matching« Vorschläge über »relevante« Daten(-Objekte) nicht nur erzeugt, sondern auch optimiert werden, die beispielsweise aufgrund von statistischen Wahrscheinlichkeiten gefolgert werden, sind diverse ethisch-semantische Fragen zu prüfen. Im angeführten Kittler-Beispiel wären beispielsweise Hinweise auf die Ablage von Daten denkbar, denn die Speicherstruktur des Autors sah einen eher unüblichen Speicherort vor, der sich von der typischen Ablagekultur eines Standardnutzers unterschied. Hier schließt sich nicht nur technisch sondern auch inhaltlich der Kreislauf des Archivs, insofern ganz grundlegende Fragen und Interessen diskutiert werden müssen. Einige Aspekte sind zu einem guten Teil in den jeweiligen institutionellen Policies geregelt. Darüber hinaus bedarf es aber auch in der Archivvorstufe eines intelligent abgestimmten Wechselspiels zwischen menschlicher und maschineller Intelligenz, deren Zuständigkeiten und Prozesse häufig nur fallspezifisch gelöst werden können.

3 Zusammenfassung und Ausblick

Wie der vorliegende Beitrag gezeigt hat, gibt es eine ganze Reihe an durchaus praktikablen Ansätzen zum Umgang mit komplexen digitalen Daten an der Schwelle zum Archiv. Dennoch bleibt ein beachtlicher Handlungsbedarf, denn unabhängig von der Tatsache, dass noch keine standardisierten Erfassungsprozesse für derartige Informations-Cluster definiert sind, bleiben grundlegende organisatorische Fragen offen. So es müssen beispielsweise Regeln gefunden werden, die Antworten auf Fragen zur Beurteilung der Inhalte in den weiteren Vereinnahmungsschritten (Appraisal) und der Auswahl (Selection) geben; es sollte geklärt werden, welche (persönlichkeits-, verwertungs-, urheber-, jugendschutz- etc.) rechtlichen Aspekte berücksichtigt werden müssen und/oder ob andere mit Vorsicht zu behandelnden Faktizitäten (Sensitivity) vorhanden sind; Katalogisierungs- und Erfassungsschritte müssen geplant und Zuständigkeiten geklärt werden (Cataloguing/Preparation of records) [NA13]. Neben konservatorischen Aspekten, die im OAIS-Modell unter dem Aspekt des »Preservation Planning« abgehandelt werden, gewinnen bei Planung künftiger Handhabungsroutinen zunehmen kuratorische Fragestellungen an Bedeutung und Aspekte, welche die künftige Vermittlung frühzeitig in den Blick nehmen [DC09]. Das erscheint hier insofern relevant, als durch die Aufbereitung, Zugänglichmachung und (Nach-)Nutzung der Archivalien der Wert der Inhalte im Sinne von sog. »Curation Boundaries« steigt [TH07; SB08]. Zudem hat die Vergangenheit gezeigt, dass sich nicht nur das Verständnis der Inhalte kontinuierlich ändert, sondern dass durch sich ständig verändernde Hardware-Software-Ensembles etc. die einstigen Nutzungskonventionen der Bedienung Änderungen unterworfen sein können. Die Flüchtigkeit semantischer, kultureller und institutioneller Kontexte erfordert eine sorgsame Dokumentation und (historische) Übermittlung. Trotz aller Erfassungs-, Aufbereitungs- und Vermittlungsleistungen muss künftigen Generationen die Möglichkeit gegeben werden, mit ihren Werkzeugen erneut unvoreingenommen recherchieren zu können.

Da all diese Aspekte den künftigen Umgang mit digitalen Archivalien beeinträchtigen können, sollten die zuletzt angedeuteten Fragen möglichst frühzeitig angegangen werden. Gerade im Umgang mit komplexen digitalen Objekten und Datenakkumulationen zeichnen sich derzeit daher zwei scheinbar gegenläufige Tendenzen ab: Einerseits wird – und zwar nicht nur im Archivkontext – der Zeitpunkt der Datenerhebung immer früher angesetzt, also bereits vor der eigentlichen Vereinnahmung (Ingest).[10] Andererseits kommt es im Umfeld der wissenschaftlichen Forschungsarchive vermehrt zur Planung und Umsetzung von Nachnutzungssystemen, welche die Akkumulation des Wissens unterstützten, wobei die schöpferische Arbeit der früheren und späteren Autoren gewahrt wird [WL11].

[10] Exemplarisch für eine solche Vorverlegung der Recherche und Aufarbeitungsvorbereitung, die noch vor der eigentlichen Akquise beginnen, verdeutlicht das Modell zum Ankauf von medienbasierten Gegenwartskunst der Matters-in-Media-Art-Forschung [Ta08].

4 Literaturverzeichnis

Alle URLs sind auf dem Stand vom Juni 2013. Bei reinen Online-Quellen ist als Jahr das der letzten Änderung lt. Seiteneigenschaften zum Zeitpunkt des Abrufs angegeben.

[DC09] Digital Curation Center – DCC (2009): Curation Lifecycle Model, in: http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[DL13] Deutsches Literaturarchiv Marbach [Webseite], in: http://www.dla-marbach.de/dla/index.html.

[Go11] Goldman, Ben: Using What Works: A Practical Approach to Accessioning Born-Digital Archives, in: http://e-records.chrisprom.com/guest-post-ben-goldman/.

[KB11] Kramski, Heinz Werner / von Bülow, Ulrich: »Es füllt sich der Speicher mit köstlicher Habe« – Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach, in: Robertson-von Trotha, Caroline Y./Hauser, Robert (Hg.): Neues Erbe. Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung, Karlsruhe 2011, 141–162. http://uvka.ubka.uni-karlsruhe.de/shop/download/1000024230.

[Kr13] KryoFlux - USB Floppy Controller [Homepage], in: http://www.kryoflux.com/.

[Ly03] Lyman, Peter / Varian, Hal R.: How Much Information 2003, o. O. 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf.

[NA13] The National Archives (2013), Information Management. Records selection and transfer process. in: http://www.nationalarchives.gov.uk/information-management/our-services/selection-and-transfer.htm.

[Ri10] Ries, Thorsten: Die Geräte klüger als Ihre Besitzer. Philologische Durchblicke hinter die Schreibszene des Graphical User Interface, in: Editio 24, 2010, 149-199.

[SB08] Swan, Alma / Brown, Sheridan (2008): The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practive and Future needs. Report to the JISC, in: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf.

[Ta08] Tate (2008): Matters in Media Art. Acquisitions, in: http://www.tate.org.uk/about/projects/matters-media-art/acquisitions.

[TH07] Treloar, Andrew / Harboe-Ree, Cathrine (2008): Data management and the curation continuum: how the Monash experience is informing repository relationships, In: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf.

[WL11] Catharine Ward / Lesley Freiman / Sarah Jones et al. (2011): Making Sense: Talking Data Management with Researchers. In: International Journal of Digital Curation, Vol. 6, No. 2, S. 265-273.