Beiträge des Workshops "Digitale Langzeitarchivierung" auf der Informatik 2013

am 20.09.2013 in Koblenz

nestor edition Sonderheft 1

Beiträge des Workshops „Digitale Langzeitarchivierung“ auf der Informatik 2013

am 20.09.2013
in Koblenz

nestor edition - Sonderheft 1

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

http://www.langzeitarchivierung.de

nestor Kooperationspartner:

- Bayerische Staatsbibliothek
- Bibliotheksservice-Zentrum Baden-Württemberg
- Bundesarchiv
- Computerspiele Museum Berlin
- Deutsche Nationalbibliothek
- FernUniversität Hagen
- Georg-August-Universität Göttingen / Niedersächsische Staats- und Universitätsbibliothek Göttingen
- GESIS - Leibniz-Institut für Sozialwissenschaften
- Goportis - Leibniz-Bibliotheksverbund Forschungsinformation
- Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen
- Humboldt-Universität zu Berlin
- Institut für Deutsche Sprache
- Konrad-Zuse-Zentrum für Informationstechnik Berlin
- Landesarchiv Baden-Württemberg
- Landesarchiv Nordrhein-Westfalen
- PDF/A Competence Center
- Stiftung Preußischer Kulturbesitz / SMB - Institut für Museumsforschung
- Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden

URN: urn:nbn:de:0008-2014012419

[http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2014012419]

Die Schriftenreihe nestor edition präsentiert ausgewählte wissenschaftliche Arbeiten mit dem Schwerpunkt Langzeitarchivierung. Sie wird in loser Folge von nestor – Kompetenznetzwerk Langzeitarchivierung herausgegeben. Damit entsteht ein Forum, in dem Beiträge zu verschiedenen Aspekten der digitalen Langzeitarchivierung einer breiten Öffentlichkeit zugänglich gemacht werden.
Die Arbeiten werden von ausgewiesenen Experten aus den jeweiligen Fachgebieten für die nestor edition gezielt ausgewählt, wenn sie einen besonderen Beitrag zu wichtigen Themenfeldern oder zu neuen wissenschaftlichen Forschungen auf dem Gebiet leisten.

Bemerkungen zu dieser Publikation, aber auch Vorschläge für die Aufnahme weiterer Beiträge in der Edition gerne an: VL-nestor@dnb.de

Für die Partner von nestor – Kompetenznetzwerk Langzeitarchivierung
Reinhard Altenhöner und Armin Straube
Deutsche Nationalbibliothek

Vorwort

Am 20.9.2013 fand auf der Jahrestagung der Gesellschaft für Informatik ein Workshop zur digitalen Langzeitarchivierung statt, der durch sehr interessante Vorträge und angeregte Diskussionen geprägt war. Die Beiträge des Workshops sind im Rahmen der Proceedings der Informatik 2013 als Band der Lecture Notes in Informatik (LNI) erschienen.

Auf dem Workshop entstand die Idee, mit einem Sonderheft der nestor-edition allen an der digitalen Langzeitarchivierung Interessierten einen einfachen Zugriff auf die Beiträge zu ermöglichen.

Wir möchten allen Autoren für die freundliche Zustimmung zur Zweitveröffentlichung danken. Ein besonderer Dank geht an die Mitglieder des Programmkomitees des Workshops: Thomas Bähr, Dr. Katharina Ernst, Yvonne Friese, Dr. Claus-Peter Klas, Prof. Dr. Marc Wilhelm Küster und Prof. Dr. Andreas Rauber.

Wir hoffen, dass die Beiträge mit der Veröffentlichung in diesem Sonderheft der nestor-edition ein breites Publikum finden.

Die Organisatoren des Workshops
Dr. Steffen Schilke und Armin Straube

Inhaltsverzeichnis

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

Jürgen Enge[1] Heinz Werner Kramski[2] Tabea Lurk[3]

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

[2] Wissenschaftliche Datenverarbeitung, Deutsches Literaturarchiv Marbach, Schillerhöhe 8–10, 71672 Marbach, heinz.werner.kramski@dla-marbach.de

[3] 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] »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]

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).

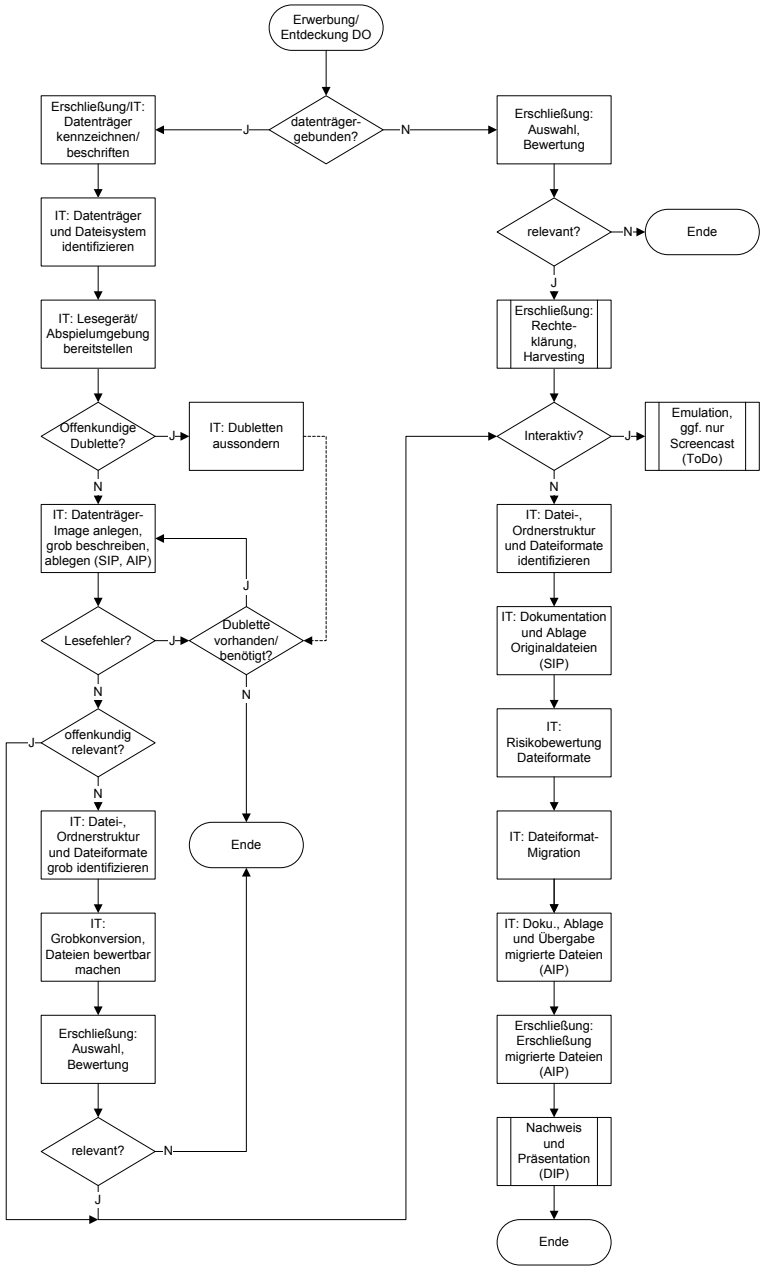


Abbildung 1: Workflow zur Bearbeitung digitaler Nachlassobjekte im DLA Marbach

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

[2] Für Details siehe [KB11].

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.)

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

---

[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.

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 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.

[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.

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.

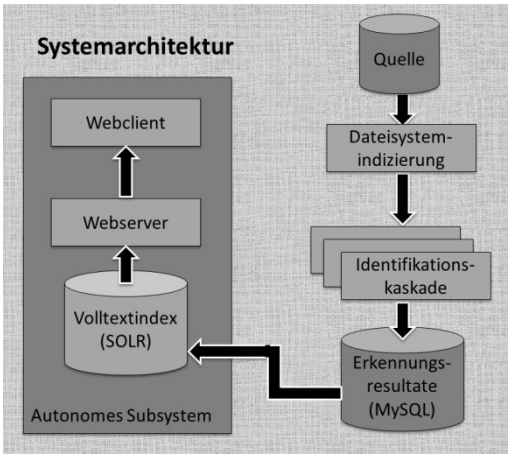


Abbildung 2: Systemarchitektur »Indexer«

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.

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.

[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.

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 Abel,...
12	2027724	NULL	NULL	NULL	
12	2027725	application/msword	NULL	Application-Name: Microsoft Word 8.0 Author: pvh C...	 Friedrich Kittler DRACULAS VERM Som...
12	2027726	NULL	NULL	NULL	
12	2027727	application/msword	NULL	Application-Name: Microsoft Word 8.0 Author: pvh C...	Friedrich Kittler ECHO Narziß erkennt sic

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.

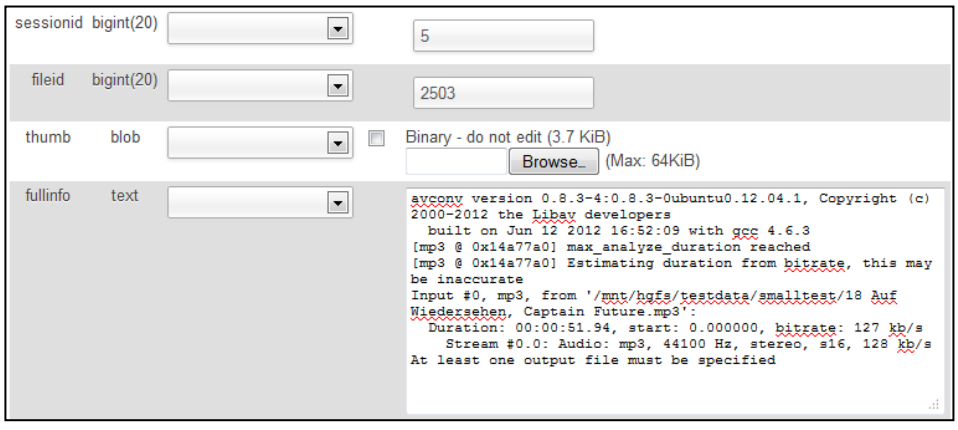


Abbildung 6: Datenbankeintrag von AVCONV

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.

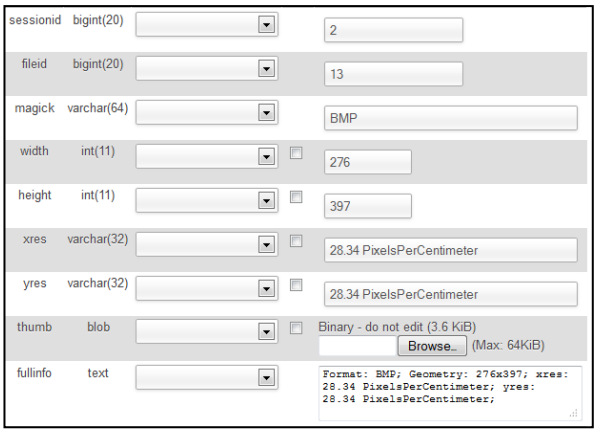


Abbildung 7: Datenbankeintrag ImageMagick

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.

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

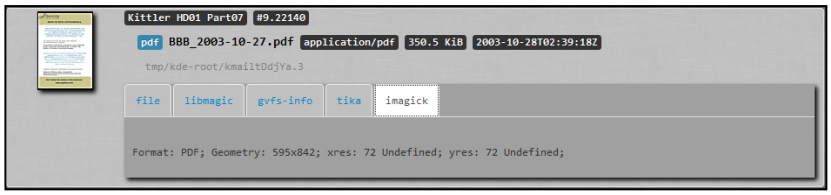


Abbildung 9: Erkennungskaskade – imagick

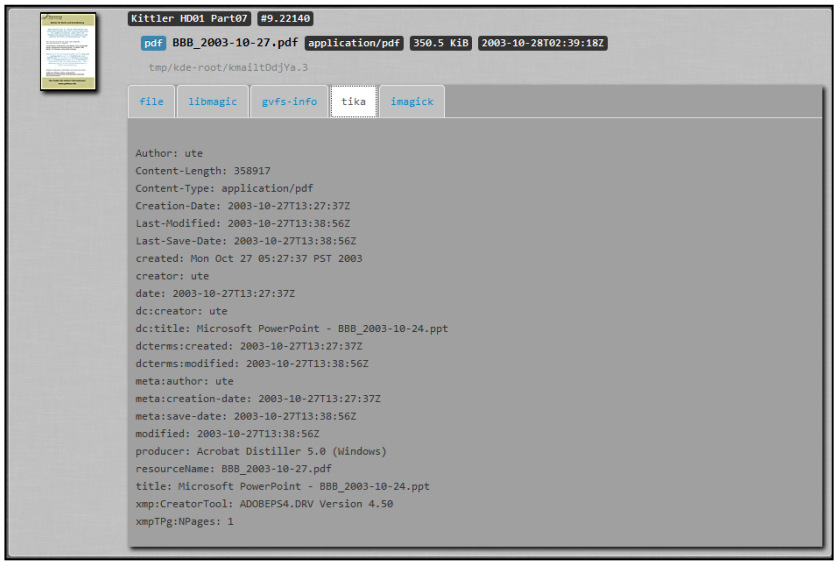


Abbildung 10: Erkennungskaskade – tika

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].

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.

[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].

datorium – ein neuer Service für Archivierung und Zugang zu sozialwissenschaftlichen Forschungsdaten

Wolfgang Zenk-Möltgen Monika Linne
Datenarchiv für Sozialwissenschaften, GESIS - Leibniz Institut für Sozialwissenschaften,
Unter Sachsenhausen 6-8, 50667 Köln
wolfgang.zenk-moeltgen@gesis.org monika.linne@gesis.org

Abstract: Forschungsdaten in den Sozialwissenschaften werden trotz zahlreicher Anstrengungen noch nicht ausreichend archiviert und zugänglich gemacht. Das GESIS Datenarchiv führt daher mit datorium einen neuen Service für Wissenschaftlerinnen und Wissenschaftler ein, ihre Forschungsdaten auf einfache Weise zu archivieren und Anderen zugänglich zu machen. Die Konzeption und Umsetzung des Serviceangebots von datorium wird in zwei Phasen vollzogen: Zunächst wird der Zugang zu den Standard-Langzeitarchivierungsprozessen des Datenarchivs über datorium vereinfacht. In einem zweiten Schritt wird datorium als einfache Möglichkeit des data sharings ausgebaut, die neben einer bitstream preservation auch ein Angebot für eine optionale Langzeitarchivierung der Daten durch GESIS enthält. Im Beitrag werden die Funktionen von datorium, die Software für die Implementierung und die Einbettung in die Arbeitsprozesse des Datenarchivs zur Langzeitarchivierung und die Distribution dargestellt. Ziel des GESIS Datenarchivs ist es, mehr Forschungsdaten mit einer größeren thematischen Breite für die wissenschaftliche Verwendung bereit zu stellen und zu archivieren.

1 Einleitung

Das bisherige Dienstleistungsangebot des GESIS Datenarchivs wird im Jahr 2013 um das digitale Daten-Repositorium datorium erweitert. Dieses erlaubt es, dass Forschungsdaten selbst hochgeladen, beschrieben und weitergegeben werden können, enthält als Basissicherung eine bitstream preservation und bietet eine Anbindung an die Langzeitsicherungsangebote des GESIS Datenarchivs. datorium stellt so eine komplementäre Ergänzung zu den bisherigen Sicherungs- und Distributionsangeboten des Archivs dar. Auf diesem Wege soll die Kultur des data sharings, die das Archiv seit über 50 Jahren unterstützt und fördert, weiter vorangetrieben werden und die thematische Breite und Anzahl der durch das Archiv geförderten Nutzungen von Forschungsdaten erhöht werden. Damit folgt das Datenarchiv auch den vielen internationalen und nationalen Empfehlungen zur Verbesserung der Archivierung und des Zugangs zu öffentlich finanzierten Forschungsdaten [Wi2012, Eu2010, Oe2013]. GESIS reagiert damit auf eine sich verändernde Datenlandschaft, in der Wissenschaftlerinnen und Wissenschaftler schnelle und flexible Werkzeuge fordern, die es ihnen insbesondere erlauben, ihre Forschungsergebnisse zu publizieren und mit Anderen zu teilen.

Ein wesentlicher Bestandteil der Arbeit des Datenarchivs besteht in der Generierung von Metadaten [ZH2012]. Sie sind zentral für die Auffindbarkeit, die Nutzbarkeit und die Langzeitarchivierung der archivierten Daten. Mit datorium sollen Wissenschaftlerinnen und Wissenschaftler die Möglichkeit erhalten, ihre Daten, Metadaten und relevanten Dokumente eigenständig in das Repositorium einzupflegen sowie dazugehörige Publikationen zu verlinken. Darüber hinaus können sie selbst definieren, welchen Nutzergruppen sie den Zugriff auf ihre Daten erlauben. Den Forscherinnen und Forschern wird dadurch einerseits ermöglicht, ihre Forschungsergebnisse kostenfrei zu archivieren und andererseits zeitnah Anderen zur Verfügung zu stellen. Darüber hinaus erhalten sie die Möglichkeit, dass die Forschungsdaten auf einfache Weise in die bestehende Langzeitarchivierung des GESIS Datenarchivs übernommen werden können. Eine solches Angebot kann daher dazu dienen, der Profession Zugang zu bisher unerschlossenen Forschungsprojekten zu ermöglichen.

2 Datenarchivierung

Der Prozess der Archivierung sozialwissenschaftlicher Daten im Verständnis der GESIS-Abteilung Datenarchiv für Sozialwissenschaften umfasst die Schritte Akquisition, Dateneingang, Datenaufbereitung und Dokumentation, Datenbereitstellung und Langzeitarchivierung. Die derzeitige Praxis der Datenaufbereitung und Dokumentation lässt sich dabei grob in eine Standardarchivierung und eine Added-value-Archivierung unterteilen [Ma2012]. Für die Langzeitarchivierung versteht sich das Archiv nach dem OAIS-Referenzmodell als Organisation, die die

Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit der digitalen Forschungsdaten für ihre Nutzbarkeit durch die Zielgruppe der Sozialwissenschaften übernommen hat [vgl. DS2010].

Im Rahmen der gegenwärtigen Standardarchivierung durch das Datenarchiv (siehe Abb. 1) durchläuft jede Studie nach Dateneingang zunächst eine ausführliche Dateneingangs-kontrolle [Ma2012]. Dieser Prozess beinhaltet u.a. eine technische Kontrolle bzgl. der verwendeten Formate, Lesbarkeit oder Virenfreiheit der Daten. Darüber hinaus findet eine Überprüfung auf Vollständigkeit und Nutzbarkeit der Daten, Erhebungsinstrumente sowie der Dokumente statt. Außerdem wird kontrolliert, ob die Daten mit dem beschriebenen Projekt übereinstimmen und ob die Daten konsistent sind (Wild Codes, fehlende Werte, Gewichtung korrekt, question routing, Vercodungsfehler, Dopplungsfehler etc.). Ein weiterer wichtiger Arbeitsprozess der Standardarchivierung ist außerdem die Datenschutzkontrolle, in der z.B. die Quelle der Erhebung oder Datenschutzverletzungen überprüft werden.

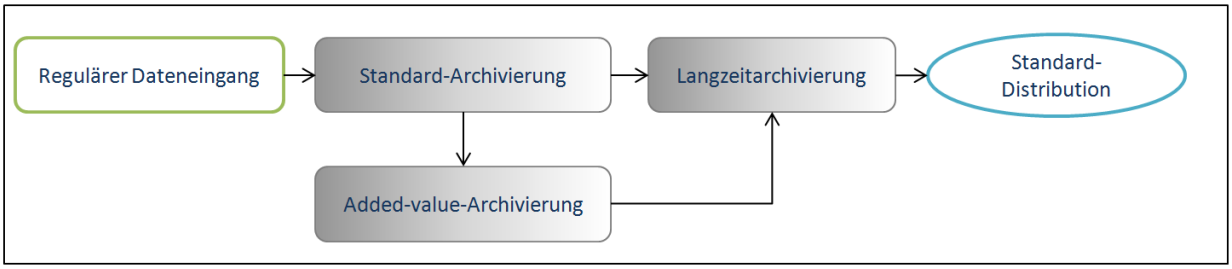


Abbildung 1: Workflow der gegenwärtigen Datenarchivierung

Im Anschluss an die Dateneingangskontrolle und nach eventuellen Fehlerkorrekturen in Absprache mit den Datengeberinnen und Datengebern findet eine Überführung der zu archivierenden Daten und Dokumente in geeignete Langzeitsicherungs- und Distributionsformate statt. Im weiteren Verlauf wird eine Studienbeschreibung erstellt, welche die inhaltlichen, methodischen und technischen Charakteristika der Studie enthält. In diesem Zuge werden außerdem eine Studiennummer und eine DOI für die aktuellste Version der Studiendokumentation vergeben [Ze2012]. Nach einer Bearbeitung der wichtigsten Metadatenfelder erfolgt die Sicherung im Archivspeicher und die Publikation der Studie über den Datenbestandskatalog. Die Publikation von Metadatenfeldern, deren Bearbeitung aufgrund einer höheren Komplexität länger dauert, kann zu einem späteren Zeitpunkt erfolgen. Zur Langzeitarchivierung gehört für das Archiv neben dem Erhalt der physischen und logischen Informationen der Studie auch der Prozess der Beobachtung aktueller Speichertechnik und -formate, sowie die Anpassung der digitalen Objekte durch Migration oder Emulation [Ma2012].

Für ausgewählte Studien oder besondere Studienkollektionen bietet das Datenarchiv den Service einer Added-value-Archivierung an (vgl. Abb. 1). Die Studien durchlaufen zunächst alle Arbeitsprozesse der Standardarchivierung und erhalten im Anschluss eine besondere Datenaufbereitung (Datenbereinigung; Standardisierung; zeit- und/oder ländervergleichende Integration/Kumulation; Harmonisierung; Ergänzung mit Kontextdaten/Aggregatdaten). Darüber hinaus findet eine umfassende Produktion von strukturierten Metadaten auf Variablenebene statt (z.B. vollständige Frage- und Antworttexte – teilweise multilingual; Intervieweranweisungen; besondere Anmerkungen zur Dokumentation; abweichende Länder- oder Wellenspezifika), deren Publikation über Online-Portale wie z.B. ZACAT oder Variable Overviews erfolgt [BZ2011]. Außerdem werden im Rahmen der Added-value-Archivierung weitere Kontextinformationen (z.B. themenspezifische oder vergleichbare Fragen; Trendvariablen) hinzugefügt sowie Codebücher, Datenhandbücher, Variablenreports oder Methodenberichte generiert.

In beiden Fällen, also sowohl innerhalb der Standardarchivierung als auch der Added-value-Archivierung erfolgt die Studiendokumentation und Langzeitsicherung durch das Archiv, welches die Verantwortung für den Erhalt der langfristigen Nutzbarkeit und Interpretierbarkeit der ihm anvertrauten digitalen Objekte übernimmt. Das GESIS Datenarchiv hat dafür eine Preservation Policy verabschiedet und bereitet sich auf eine Zertifizierung als vertrauenswürdiges digitales Archiv im Rahmen des „Data Seal of Approval“ [DS2010] vor. Die Überprüfung und Aufbereitung der Daten nach den neuesten Standards sowie Dokumentation und Sicherung von dazugehörigen Studienmaterialien sind der Garant für spätere Sekundäranalysen und für die Vergleichbarkeit mit anderen Studien.

3 Bestand

Zahlreiche Forscherinnen und Forscher aus Institutionen der akademischen sowie der kommerziellen Markt- und Meinungsforschung nutzen das GESIS Datenarchiv für Sozialwissenschaften zur Archivierung und Distribution ihrer Studien. Zurzeit sind ca. 5.100 Studien von insgesamt etwa 2.200 Primärforschenden von fast 2.400 Institutionen im Bestand des Archivs. Davon sind die Forschungsdaten von knapp 2.900 Studien nach einer Registrierung direkt im Download für Sekundärnutzende verfügbar. Alle weiteren Studien sind über ein Warenkorbsystem bestellbar oder unter speziellen Bedingungen zu nutzen. Dies eröffnet Wissenschaftlerinnen und Wissenschaftlern die Möglichkeit, ohne eigenen Erhebungsaufwand empirisch zu arbeiten und auf Bestehendem aufzubauen. Vorhandenes Wissen wird so optimal und kostengünstig genutzt.

Zurzeit archiviert das Datenarchiv für Sozialwissenschaften digitale Datensätze aus allen Bereichen der Sozialwissenschaften, wenn die Studie Aussagen über die deutsche Bevölkerung oder über Teile von ihr erlaubt, die Untersuchung von deutschen Forschenden durchgeführt wurde, unabhängig davon, ob sich die Untersuchung auf Deutschland bezieht oder nicht, oder wenn die Studie ganz allgemein für die sozialwissenschaftliche Gemeinde von Interesse sein könnte [Ma2012]. Für eine vollständige Archivierung benötigt das Datenarchiv von den Primärforschenden alle Materialien, die für eine Sekundäranalyse notwendig sind. Dies umfasst mindestens die Daten, wenn möglich aufbereitet für die direkte Verwendung in einer Statistiksoftware, außerdem das Erhebungsinstrument und eine methodische Beschreibung.

Durchschnittlich wurden mit diesen Aufnahmekriterien innerhalb der letzten Jahre etwa 170 neue Studien pro Jahr vom Datenarchiv akquiriert, was zum Stand von jetzt ca. 5.100 veröffentlichten Studien geführt hat. Im Vergleich dazu kann man sehen, dass die DFG seit 1999 insgesamt 87.898 Projekte gefördert hat, von denen alleine 15.303 aus den Geistes- und Sozialwissenschaften stammen [Df2013]. Ebenso kann man von etwa 6.000 bis 7.000 neuen oder aktualisierten Forschungsprojekten alleine in Deutschland, Österreich und der Schweiz ausgehen, die jährlich in das sozialwissenschaftliche Forschungsinformationssystem SOFIS aufgenommen werden [Ko2012: 40] und aktuell zu einer Zahl von 51.585 Projekten geführt haben [So2013]. Auch wenn diese Zahlen nicht nur empirische Datenerhebungsprojekte umfassen, wird hier deutlich, dass das Akquisitionsvolumen für Forschungsdaten im Datenarchiv klare Steigerungs-möglichkeiten aufweist. Hier kann datorium einen Beitrag leisten, indem die Aufnahme eines Forschungsprojektes wesentlich einfacher gestaltet wird. Darüber hinaus können mit datorium auch andere Projektarten, die über die bisher von GESIS hauptsächlich archivierten Studien hinaus gehen, archiviert werden, wie etwa Dissertationsprojekte oder kleinere eigenfinanzierte Projekte.

Mit dem Angebot von datorium verfolgt GESIS demnach das Ziel, das Akquisitionsvolumen zu erhöhen und auch solche Forschungsprojekte zu erfassen, welche über die o.g. Archivierungskriterien des Datenarchivs hinausgehen. Sowohl die Anzahl der erfassten Studien, der Datengebenden, als auch der Daten- und Dokumentendownloads durch Nutzerinnen und Nutzer sollte durch ein solches Repositorium erheblich gesteigert werden können. Dies bedeutet, dass sich auf beiden Seiten, sowohl bei den Datengebenden als auch bei den Datennutzenden, möglicherweise neue Nutzungsprofile erschließen. Dies soll insbesondere durch die geplanten sozialen Vernetzungsmöglichkeiten innerhalb von datorium unterstützt werden, so dass sich neue produktive Kontakte mit synergetischen Effekten zwischen Wissenschaftlerinnen und Wissenschaftlern ergeben können. Darüber hinaus wird – durch die niedrigschwellige Möglichkeit für eine zunächst einfache Sicherung der Forschungsdaten und die Option zur Übernahme in die Langzeitsicherung des Datenarchivs – bei Forschenden ein größeres Bewusstsein für die Notwendigkeit der digitalen Langzeitarchivierung geschaffen.

Wissensproduktion ist ein Prozess, der darauf basiert, dass neue Erkenntnisse schnell verbreitet werden, damit sie von anderen Forscherinnen und Forschern weiterverwendet werden können. Wesentlich ist deshalb die Weitergabe von Forschungsergebnissen in der Forschungsgemeinschaft zum Zweck des Erkenntnisgewinns. Studien konnten zeigen, dass die Bereitstellung von Forschungsdaten auch die Zitationen der Publikationen über die Forschungsergebnisse von Forschenden erhöht [PDF2007]. Dieser Reputationsgewinn ist ein wesentlicher Anreiz für Forschende zur Archivierung und Bereitstellung ihrer Daten.

Darüber hinaus erleichtert die Publikation von Forschungs¬erkenntnissen der wissenschaftlichen Gemeinde das Auffinden und die Zitation von relevanten wissenschaftlichen Forschungsergebnissen. In diesem Sinne eignet sich ein niedrigschwelliges Webangebot in Form des datorium-Portals besonders gut, um die Verbreitung und Nutzung von wissenschaftlichen Arbeiten in einem größeren Spektrum als bisher zu unterstützen und zu erleichtern. Mit dem Angebot von datorium durch das GESIS Datenarchiv werden die Forscherinnen und Forscher daher auch bei der Beachtung von Empfehlungen zu Open Access unterstützt [Re2013].

4 Vorgehen

Zunächst wurden für das Projekt datorium ein Konzept und die Spezifikationen für den Betrieb des Repositoriums erarbeitet. Dabei wurde die Implementierung in zwei Phasen geplant: Phase I (s. Abb. 2) ermöglicht es Wissenschaftlerinnen und Wissenschaftlern, ihre Forschungsdaten über datorium selbständig einzupflegen und dazugehörige Dokumente hochzuladen. Die technische Implementierung dieses Datenzugangs von Forschungsprojekten in das Datenarchiv steht seit Juni 2013 zur Verfügung. Die Daten durchlaufen daraufhin die auch bisher bei der Standardarchivierung durchgeführten Prozesse durch das Datenarchiv und werden in Abhängigkeit des Studieninhalts und der Datenaufbereitung in entsprechenden Online-Portalen publiziert (Datenbestandskatalog, ZACAT, CESSDA Data Portal, Variable Overview, HISTAT).

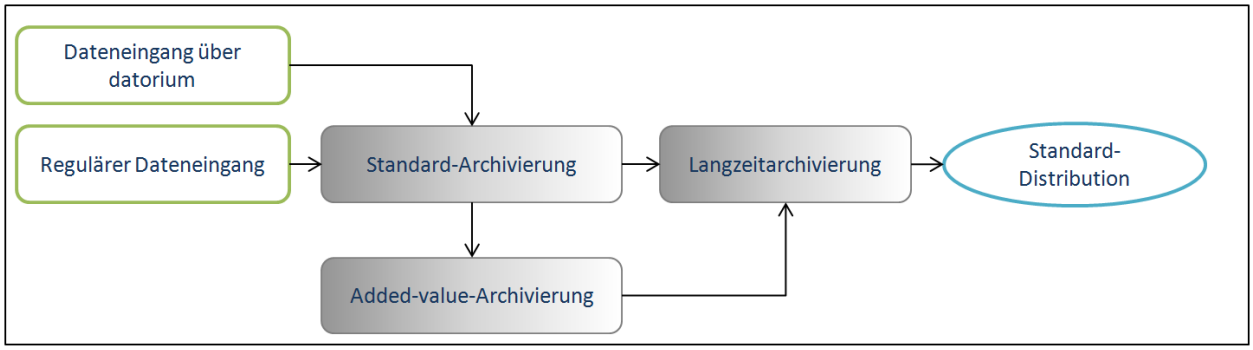


Abbildung 2: Workflow datorium und regulärer Dateneingang – Phase I

In dieser Phase I der Implementierung von datorium wird der Dateneingang interessierten Datengeberinnen und Datengebern zur Verfügung gestellt. Diese können den neuen Zugang nutzen, um auf einfache Weise ihre Daten für die Archivierung und Distribution an das Datenarchiv zu übergeben. Zusätzlich können sie die Nutzungsfreundlichkeit testen und tragen damit auch zu einer Weiterentwicklung des Angebots bei.

Im weiteren Projektverlauf wird in einer Phase II (s. Abb. 3) das Angebot von datorium einen eigenen Datenbestand enthalten, der nicht durch die Standardprozesse des Archivs behandelt wird. Nach Implementierung dieser Phase II durchlaufen alle weiteren und neu hochgeladenen Forschungsdaten im datorium-Portal einen neuen Review-Prozess durch das Datenarchiv. In diesem wird entschieden, ob die Forschungsdaten – falls durch die Datengebenden gewünscht – im Standardarchiv archiviert werden oder ob sie ohne eine weitere Bearbeitung durch das Datenarchiv in datorium gehalten und dort publiziert werden. Für Studien in datorium ist eine Basissicherung als bitstream preservation vorgesehen, denn diese deckt die unmittelbaren Bedürfnisse nach Sicherung und Bereitstellung ab. Wenn Wissenschaftlerinnen und Wissenschaftler eine echte Langzeitarchivierung durch das Datenarchiv wünschen, können sie dies bei der Einstellung ihrer Daten und Metadaten in datorium angeben. Eine Entscheidung über diesen Wunsch nach Aufnahme in das Standardarchiv kann allerdings nur im Review-Prozess durch das Datenarchiv getroffen werden, da eine Übernahme der Daten in die Standardarchivierung mit Ressourcenaufwand verbunden ist.

Darüber hinaus können die Nutzenden in datorium auch bestimmen, ob ihre Daten publiziert oder ohne Sichtbarkeit nach außen lediglich archiviert und langfristig gesichert werden sollen. Im Fall der Archivierung ohne Publikation findet keine Datenaufbereitung durch das Datenarchiv statt.

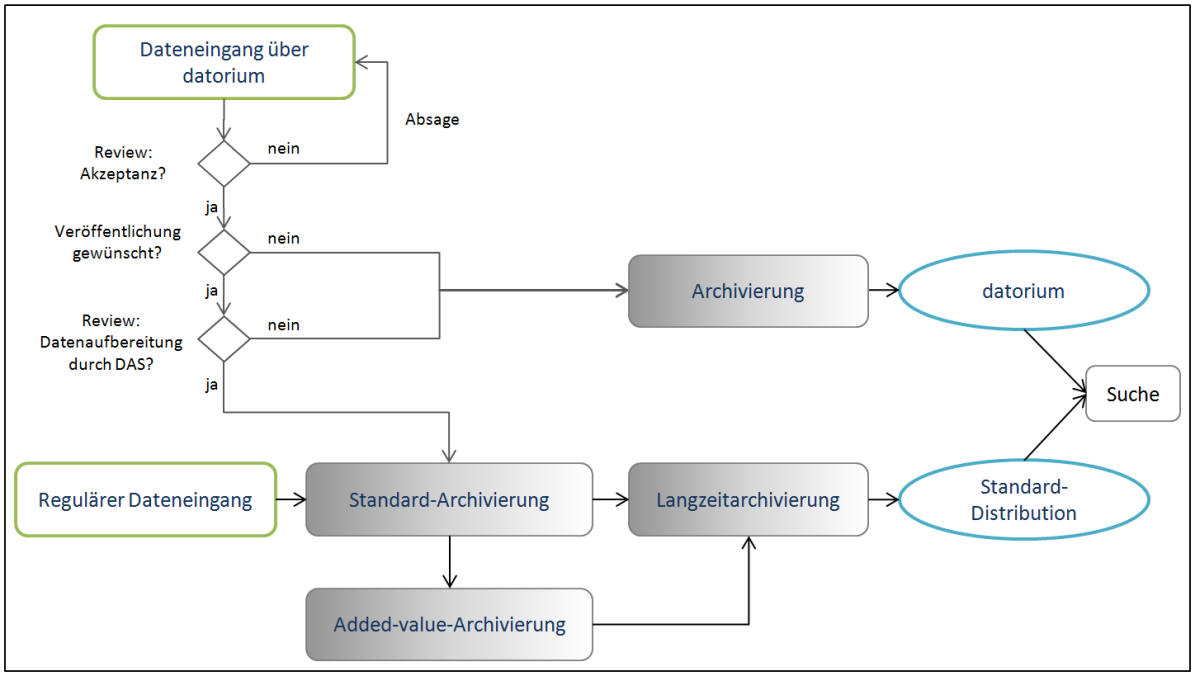


Abbildung 3: Workflow datorium und regulärer Dateneingang – Phase II

Für Datennutzende sollen in Phase II alle Bestände des Datenarchivs gemeinsam durchsuchbar gemacht werden, so dass nur ein Zugangspunkt angelaufen werden muss. Dazu wird die Suche im Datenbestandskatalog so ausgebaut werden, dass sie die bisherigen Bestände des Datenarchivs und die Bestände in datorium umfasst (s. Kap. 9). Für den Zugang zu den Daten durch Dritte werden im Standardarchiv weiterhin die bisherigen Regeln gelten. Für datorium können die Datengebenden selbst die Zugänglichkeit ihrer Forschungsdaten bestimmen.

5 Zielgruppen und Inhalte

GESIS versteht sich als Teil der globalen sozialwissenschaftlichen Forschungsgemeinschaft. Allgemein richtet sich das Angebot des Instituts primär an Forscherinnen und Forscher der empirischen Sozialforschung mit einem Schwerpunkt auf den Fachgebieten Soziologie und Politikwissenschaft sowie an solche der Sozialwissenschaften insgesamt. Weitere Zielgruppen liegen im politischen, sozialen und kommerziellen Umfeld der Sozialwissenschaften.

Ein Großteil der Datengeberinnen und Datengeber im Datenarchiv kommt zurzeit aus größeren sozialwissenschaftlichen Forschungseinrichtungen mit dem Schwerpunkt auf Langzeitstudien, wie z.B. ALLBUS, European Values Study, International Social Survey Programme, Eurobarometer, Politbarometer oder German Longitudinal Election Study. Mit dem Service des datorium-Portals wird deshalb u.a. eine Erhöhung der archivierten Einzelstudien angestrebt, indem beispielsweise Forschungsdaten von Datengeberinnen und Datengebern aus kleineren Forschungsprojekten akquiriert werden. Die thematische Breite sozialwissenschaftlicher Forschungsdaten, etwa aus der Medienforschung, der Gesundheitsforschung, Wirtschaftsforschung oder Umweltforschung soll verbessert werden. Denkbar ist außerdem eine Ausweitung der Zielgruppe auf andere wissenschaftliche Disziplinen (wie z.B. Geistes- oder Erziehungswissenschaften), um interdisziplinäre Forschungssynergien zu unterstützen.

Das geplante Daten-Repositorium soll wissenschaftliche Forschungsbeiträge der genannten Zielgruppen enthalten, die bisher noch keinen Eingang in das Standardarchiv von GESIS gefunden haben. Somit kann ein solches Repositorium eine sinnvolle Ergänzung zu existierenden Archivbeständen darstellen. Kleinere Forschungsprojekte z.B. aus DFG-Förderung oder vergleichbaren Drittmittelprojekten sind aktuell im Bestand des Archivs tendenziell eher unterrepräsentiert. Die geförderten Projekte der DFG oder die nachgewiesenen Projekte in SOFIS bieten ein breites mögliches Erschließungspotenzial für datorium (s. Kap. 3).

Darüber hinaus können Forscherinnen und Forscher z.B. ihre Ergänzungen zu einem Archivdatensatz von GESIS – z.B. Kontextdaten zu einem Eurobarometer – ins Repositorium laden. Oder sie können ihre noch unveröffentlichten Forschungsprojekte publizieren und diesbezügliche Aufsätze, Berichte sowie Vorträge hervorheben oder aber auf die aktuelle Forschungsentwicklung ihrer Studie hinweisen, wodurch andere Forscherinnen und Forscher den

Fortschritt des Projekts verfolgen können. Außerdem eignet sich datorium für die Aufnahme von Daten, die im Rahmen von Qualifizierungsarbeiten, z.B. zu Dissertationen oder Habilitationen, erstellt worden sind. Ebenfalls geplant ist die Aufnahme von Analyse-Syntaxen zur Generierung von Replikations¬datensätzen, die der scientific community für Reanalysen zur Verfügung gestellt werden sollen. Weitere Repositoriumsinhalte könnten Daten zu wissenschaftlichen Aufsätzen sein, die in Zeitschriften publiziert werden sollen, wenn diese Regeln für die Archivierung und Zugänglichmachung verabschiedet haben [VS2012].

Wissenschaftlerinnen und Wissenschaftler können durch diese Form der Web-Publikation Anerkennung für ihre Forschungsarbeiten mit Daten erhalten, auch wenn ihre inhaltlichen wissenschaftlichen Publikationen bisher nicht in einem etablierten Journal veröffentlicht wurden. Vorteile gegenüber einer Publikation von Forschungsdaten auf einer eigenen Homepage sind, dass die Sichtbarkeit in einem disziplinären Repositorium deutlich höher ist, und dass zunächst die mittelfristige, später auch die langfristige Sicherung, Verfügbarkeit und Identifizierung der Forschungsdaten gewährleistet sind.

6 Daten-Review

Die in datorium eingestellten Daten und Metadaten durchlaufen nach Fertigstellung von Phase II nicht mehr den Dokumentations- und Archivierungsprozess der Standardarchivierung (vgl. Abb. 3). Eine Prüfung dieser Daten auf Vollständigkeit, Sinnhaftigkeit, Format, Einhaltung des Datenschutzes, etc. erfolgt ab Phase II mittels eines Reviews durch GESIS innerhalb von datorium. Im Reviewprozess wird u.a. eine Studiennummer zugewiesen, die keine inhaltliche Bedeutung trägt und ähnlich wie die im Standardarchiv verwendeten ZA-Nummern zur schnellen Identifizierung einer Studie dient. Dieser Reviewprozess soll in stark standardisierter Form nur eine grundlegende Prüfung leisten, damit er für eine große Anzahl an Studien durchführbar ist. Nach der Einführungsphase von datorium wird eine Evaluation erfolgen, ob die Einzelheiten des Reviewprozesses neu festzulegen sind, um dieses Ziel zu erreichen. Eine hochgeladene Studie wird erst dann nach außen publiziert, wenn sie im Rahmen der Dateneingangskontrolle die Kriterien des Reviews erfüllt und eine Freischaltung durch das Datenarchiv erhält.

In Ausnahmefällen wird bei Erfüllung spezieller Qualitätskriterien im Reviewprozess von GESIS entschieden, ob eine Studie aus dem datorium-Portal in das Standardarchiv übernommen wird und hier den Aufbereitungsprozess und die Dokumentation der Standardarchivierung durchläuft. Dies können inhaltliche Kriterien sein, z.B. eine hohe Relevanz des Untersuchungsgegenstandes, oder auch methodische Kriterien, wie die Repräsentativität der Stichprobe oder ein innovatives Untersuchungsdesign. Für solche Studien ist dann auch die Langzeitarchivierung gegeben, also eine langfristige Sicherung und Interpretierbarkeit der Daten z.B. durch Formatmigration, die zu einer höheren Sicherheit der zukünftigen Nutzbarkeit für diese Studie führt. Da die Ressourcen des GESIS-Datenarchivs begrenzt sind, muss im Review-Prozess eine Auswahl getroffen werden. Denkbar ist bei verfügbaren Ressourcen auch, dass eine Added-value-Dokumentation einer solchen Studie durchgeführt wird. Dieses wird in solchen Ausnahmefällen mit den Datengeberinnen und Datengebern einzeln vereinbart.

7 Registrierung von DOI-Namen

Zur sicheren und dauerhaften Zitierung von Forschungsdaten bietet GESIS in Kooperation mit DataCite, der internationalen Initiative zur Unterstützung des Zugangs zu digitalen Forschungsdaten, den DOI Registrierungsservice für sozial- und wirtschaftswissenschaftliche Daten in Deutschland da|ra an [HZ2011]. Ebenso wie andere persistente Identifikatoren, etwa Handle oder URN, ist der DOI-Name unveränderlich und identifiziert ein Objekt unmittelbar, also nicht lediglich eine Eigenschaft des Objekts, wie beispielsweise die Adresse, an der es platziert ist. Darüber hinaus wird ein Objekt durch den DOI-Namen mit aktuellen und strukturierten Metadaten verknüpft, wozu dann auch die Adresse oder URL gehört, an der das Objekt zu finden ist. Die Benutzung von DOIs zur Zitation von Zeitschriftenaufsätzen ist bereits seit langem etabliert. Durch die Vergabe von DOIs für Forschungsdatensätze kann daher die Zitation dieser Daten erleichtert werden. Dies ist auch eine zentrale Voraussetzung für eine dauerhafte Verknüpfung von Forschungsergebnissen in Form von Berichten und Publikationen und den ihnen zugrunde liegenden Primärdaten. Die Metadaten der Forschungsdaten in datorium werden in da|ra zur Verfügung gestellt und werden dann dort auch recherchierbar sein [Ha2013].

Dieser Service der Datenregistrierung zur Zitierung und Verlinkung von elektronischen Ressourcen soll ebenfalls in datorium angeboten werden. In Phase I erfolgt die Vergabe einer DOI noch über die Standardarchivierung im Datenbestandskatalog mit dem Präfix „10.4232“. In Phase II wird eine direkte Registrierung der DOI-Namen aus datorium heraus mit dem Präfix „10.7802“ erfolgen, da datorium dann einen zusätzlichen Bestand an Daten enthalten wird. Objekte dürfen aus datorium nicht gelöscht werden, sondern lediglich der Zugang zum Objekt kann

beschränkt werden. Die dazugehörigen Metadaten zum Objekt bleiben also auf jeden Fall erhalten, so dass eine Zitation dieses Objekts auch später noch möglich ist.

8 Regeln für die Nutzung

Die Benutzung des Services von datorium erfolgt nach Zustimmung zu den allgemeinen Geschäftsbedingungen. Darüber hinaus können die Datengeberinnen und Datengeber in Phase II durch die Auswahl einer vorgegebenen Zugangskategorie flexibel definieren, welchen Personen oder Gruppen sie ihre Daten und hinterlegten Dokumente zugänglich machen wollen. Folgende Zugangskategorien sind vorgesehen:

a) Freier Zugang: Die Forschungsdaten sind ohne Einschränkung für alle registrierten Nutzer und Nutzerinnen zugänglich, ohne dass eine Einwilligung der Datengebenden eingeholt werden muss.

b) Eingeschränkter Zugang: Für den Zugang zu den Forschungsdaten muss eine Einwilligung der Datengebenden über das System eingeholt werden. Diese bestimmen eigenständig, welchen registrierten Nutzerinnen und Nutzern sie den Zugang erlauben.

c) Kein Zugang: Datengeberinnen und Datengeber möchten Forschungsdaten ausschließlich archivieren und nicht für Andere zugänglich machen. Diese Daten werden nicht veröffentlicht.

Die Zugangsberechtigung zu eingeschränkt zugänglichen Daten für Sekundäranalysen muss bei den Datengeberinnen und Datengebern beantragt werden. Diese können eigenständig entscheiden, welchen Nutzerinnen und Nutzern sie ihre Daten zugänglich machen möchten. Für diesen Zweck wird es im System eine Möglichkeit geben, den Zugang zu diesen Daten zu beantragen. Daraufhin wird vom System eine E-Mail an die in den Anmeldedaten hinterlegte E-Mail-Adresse der Datengeberin oder des Datengebers gesendet. Diese wird auch in Kopie an den Service von datorium zugestellt, so dass eine Dokumentation der Kontaktaufnahme möglich ist. Die Angefragten können sich in datorium einloggen und die Zugangsberechtigung erteilen oder verweigern. Reagieren Datengeberinnen oder Datengeber nicht innerhalb einer definierten Zeit, nimmt GESIS Kontakt zu ihnen auf, um den Vorgang zu bearbeiten. Für den Fall, dass Datengeberinnen und Datengeber nicht mehr erreichbar sind (z.B. im Todesfall), müssen diese bereits vorher in der Vertragsvereinbarung bestimmen, welche Zugangskategorien und -bedingungen für ihre Daten gelten sollen.

9 Technische Implementierung

Für die Implementierung des datorium-Portals wurde entschieden, eine bereits existierende Repository-Software zu nutzen, bei der bereits auf Erfahrungen in der Anwendung zurückgegriffen werden konnte. Dazu wurde eine ausführliche Evaluation verschiedener Repository-Softwarelösungen durchgeführt. Verbreitete Lösungen für ähnliche Anwendungen sind DSpace, EPrints, Fedora, MyCoRe, und OPUS, CKAN, Git, Mendeley oder Dataverse (vgl. [Op2004], [Ku2012], [Lo2006], [Za2011], [Cr2011]). Auf dieser Basis fiel die Wahl auf DSpace, eine nicht-kommerzielle Open Source Plattform zum Betrieb eines Dokumenten¬servers. Ausschlaggebende Faktoren für die Wahl von DSpace waren u.a. die flexible Erweiterbarkeit von Metadatenfeldern, ein auf GESIS anpassbares Rechtemanagement, die freie Definition von Workflows, die benutzungsfreundliche Programmier¬¬möglichkeit einer multilingualen Oberfläche, sowie die positiven Erfahrungen durch den Betrieb von SSOAR. Für den Einsatz im GESIS-Datenarchiv lässt sich DSpace in die vorhandenen Mechanismen gut integrieren. Dies sind Eigenschaften, auf die bei der Verwendung von anderen Systemen hätte verzichtet werden müssen, und die die unabhängige Gestaltung des Portals und die Anpassung an neue mögliche Nutzungsanforderungen ermöglichen.

Der Schwerpunkt von DSpace liegt auf der Erfassung, Speicherung und Distribution von digitalen Ressourcen [DHD2011]. Die Software wurde vom Massachusetts Institute of Technology und der Forschungsabteilung von Hewlett Packard in Anlehnung an das OAIS-Referenzmodell entwickelt und ist architektonisch in der Lage, Strategien zur Langzeitverfügbarkeit von digitalen Ressourcen zu unterstützen. Die Verbreitung findet unter der BSD-Lizenz statt. Der Einsatz dieser Software findet primär in Universitäten, Bibliotheken und Forschungseinrichtungen statt.

Prinzipiell wird im Rahmen der technischen Implementierung berücksichtigt, dass möglichst keine neuen Suchschnittstellen erzeugt werden müssen, sondern vorhandene, z.B. des Datenbestandskatalogs, verwendet werden können. Für die Implementierung in Phase II ist eine gemeinsame Bestands-Durchsuchung des datorium-Portals und der Distributionswege des Standardarchivs geplant. Dazu wird ein Suchindex auf der Basis von Solr verwendet werden, wie er bereits bei da|ra in Verwendung ist. Da das Metadatenschema von datorium auf dem Datenbestandskatalog DBK aufbaut und auch gemeinsame kontrollierte Vokabulare enthält, ist die semantische

Integration der Suche möglich. Dieses stellt eine enorme Erleichterung für die Nutzenden dar, da sie mittels einer integrierten Suche in datorium und im DBK alle bei GESIS verfügbaren Forschungsdaten finden können. Darüber hinaus werden die Metadaten auch für Recherchen in da|ra verfügbar sein, wo Nutzende breit gefächerte Suchanfragen durchführen und auch Bestände von anderen Institutionen finden können.

10 Ausblick

Das Angebot der Phase I von datorium zur Aufnahme von Forschungsdaten ins Archiv für die Standardarchivierung steht bereits seit Juni 2013 zur Verfügung. Dieses Angebot wird mit einzelnen Datengeberinnen und Datengebern getestet und verbessert. Gegenwärtig arbeitet GESIS an der Umsetzung der Phase II von datorium. Dabei müssen Datengeberinnen und Datengeber die alternative Möglichkeit erhalten, die Archivierung und Distribution ihrer Forschungsdaten über das datorium-Portal eigenständig und selbstbestimmt durchzu¬führen. Die bisherige Planung sieht vor, dass datorium zur Aufnahme von Forschungsdaten mit Phase II gegen Ende des Jahres 2013 zur Verfügung stehen soll. Zusätzlich dazu ist bereits eine Weiterentwicklung des Angebots in Zusammenarbeit mit den Einrichtungen DIW (Deutsches Institut für Wirtschaftsforschung Berlin), WZB (Wissenschaftszentrum Berlin für Sozialforschung) und ZBW (Deutsche Zentralbibliothek für Wirtschaftswissenschaften Leibniz-Informationszentrum Wirtschaft) im Rahmen eines Drittmittelantrags für die Jahre 2014-2016 geplant. Hiermit soll sowohl eine technisch als auch konzeptionell anspruchsvollere Entwicklung zu einer Realisierung der Anforderungen aus der Profession führen. Wenn mit diesen Arbeiten eine größere thematische Breite und eine höhere Anzahl an gesicherten und langzeitarchivierten, sowie zur Verfügung gestellten Forschungsdaten in den Sozialwissenschaften erreicht werden kann, werden damit die Empfehlungen zahlreicher Organisationen zur Verbesserung der Langzeitarchivierung und des Zugangs zu öffentlich finanzierten Forschungsdaten in den Sozialwissenschaften umgesetzt.

11 Literatur

[BZ2011] Brislinger, E.; Zenk-Möltgen, W.: Findings of the original language documentation for European Values Study (EVS) 2008. Präsentation auf der IASSIST 2011 “Data Science Professionals: A Global Community of Sharing”. Vancouver, 31.05. 2011. Online http://www.iassistdata.org/downloads/2011/2011_b3_brislinger_etal.pdf [Zugriff am 18.04.2013], 2011.

[Cr2011] Crosas, M.: The Dataverse Network. An Open-Source Application for Sharing, Discovering and Preserving Data. In: D-Lib Magazine, Vol. 17, Nr. 1/2 doi:10.1045/january2011-crosas, 2011.

[Df2013] DFG: GEPRIS Datenmonitor. Online http://gepris.dfg.de/gepris/OCTOPUS/ ?module=gepris&task =showMonitor [Zugriff am 18.04.2013], 2013.

[DHD2011] Droogmans, L.; Hollister, V.; Donohue, T.: DSpace institutional repository platform. Online: https://atmire.com/labs17/bitstream/handle/123456789/7641/ 0000082812f0000000267.pdf?sequence=1 [Zugriff am 19.04.2013], 2011.

[DS2010] Dobratz, S.; Schoger, A.: Grundkonzepte der Vertrauenswürdigkeit und Sicherheit. In (Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K., Hrsg.): nestor-Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. Kap. 5.2. urn:nbn:de:0008-2010071949, 2010.

[Eu2010] European Union: Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. A submission to the European Commission. Online http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [Zugriff am 20.2.2013], 2010.

[Ha2013] Hausstein, B.; Quitzsch, N.; Jeude, K.; Schleinstein, N.; Zenk-Möltgen, W.: da|ra Metadata Schema. Version 2.2.1. GESIS Technical Reports, 2013/03, 2013.

[HZ2011] Hausstein, B.; Zenk-Möltgen, W.: da|ra – Ein Service der GESIS für die Zitation sozialwissen-schaftlicher Daten. In (Schomburg, S.; Leggewie, C.; Lobin, H.; Puschmann, C., Hrsg.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010, Köln, 2., ergänzte Fassung, 2011; S. 139–147.

[Ko2012] Kolle, Ch.: Wissenschaftliche Literaturrecherche. In (Berninger, I.; Botzen, K.; Kolle, Ch.; Vogl, D.; Watteler, O.): Grundlagen sozialwissenschaftlichen Arbeitens. Verlag Barbara Budrich, Opladen & Toronto, 2012; S. 33-61.

[Ku2012] Kunze, S. R.: Ein Personalisierungskonzept für Dataset-Repositorys am Beispiel von CKAN. In (Technische Universität Chemnitz – Fakultät für Informatik, Hrsg.): Studentensymposium Informatik Chemnitz 2012. Tagungsband zum 1. Studentensymposium Chemnitz vom 4. Juli 2012, Chemnitz, 2012; S. 27-37.

[Lo2006] Loeliger, J.: Collaborating With Git. A variety of clever tools allow you to share your software. In: Linux Magazine, June 2006.

[Ma2012] Mauer, R.: Das GESIS Datenarchiv für Sozialwissenschaften. In (Altenhöner, R.; Oellers, C., Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Scivero, Berlin, 2012; S. 197-215.

[Oe2013] OECD: New Data for Understanding the Human Condition: International Perspectives. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. Online http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf [Zugriff am 18.04.2013], 2013.

[Op2004] Open Society Institute: A Guide To Institutional Repository Software. New York, 3rd edition. Online http://www.soros.org/openaccess/pdf/ OSI_Guide_to_IR_Software_v3.pdf [Zugriff am 20.02.2013], 2004.

[PDF2007] Piwowar, H. A.; Day, R. S., Fridsma, D. B.: Sharing Detailed Research Data Is Associated with Increased Citation Rate. In: PLoS ONE, Vol. 2, Nr. 3, S. e308. doi:10.1371/journal.pone.0000308, 2007.

[Re2013] RECODE: Policy RECommendations for Open Access to Research Data in Europe (RECODE) project. Online: http://recodeproject.eu/ [Zugriff am 18.04.2013], 2013.

[So2013] SOFIS: Projekte. Online http://www.gesis.org/sofiswiki/Kategorie:Projekte [Zugriff am 18.04.2013], 2013.

[VS2012] Vlaeminck, S; Siegert, O.: Welche Rolle spielen Forschungsdaten eigentlich für Fachzeitschriften? Eine Analyse mit Fokus auf die Wirtschaftswissenschaften. In (RatSWD Working Paper Series) No. 210, November 2012.

[Wi2012] Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informations-infrastrukturen in Deutschland bis 2020. Drs. 2359-12 vom 13.7.2012. Berlin. Online http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf [Zugriff am 20.02.2013], 2012.

[Za2011] Zaugg, H.; West, R. E.; Tateishi, I.; Randall, D. L.: Mendeley: creating communities of scholarly inquiry through research collaboration. In: TechTrends, Vol. 55, Nr. 1, S. 32–36. doi: 10.1007/s11528-011-0467-y, 2011.

[Ze2012] Zenk-Möltgen, W.: The metadata in the data catalogue DBK at the GESIS data archive. Präsentation auf dem RatSWD Workshop "Metadata and Persistent Identifiers for Social and Economic Data". Berlin, 7.-8.05.2012, Online http://www.ratswd.de/ver/docs_PID_2012/Zenk-Moeltgen_PID2012.pdf [Zugriff am 18.04.2013], 2012.

[ZH2012] Zenk-Möltgen, W.; Habbel, N.: Der GESIS Datenbestandskatalog und sein Metadatenschema, Version 1.8. GESIS Technical Reports, 2012/01, Online http://www.gesis.org/fileadmin /upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-01.pdf [Zugriff am 18.04.2013], 2012.

Digitale Langzeitarchivierung von Videokunst

Dipl.-Rest. Andreas Weisser

restaumedia, Reischstr. 6, D-79102 Freiburg, weisser@restaumedia.de

Abstract: Die digitale Langzeitarchivierung von audiovisuellen Daten stellt insbesondere für kleine oder mittlere Archive sowie Sammlungen und Museen eine große Herausforderung dar. Abgesehen von der Fragestellung, in welchen Formaten (Codecs) die audiovisuellen Inhalte gespeichert werden sollen, bleibt oft auch die Frage nach einem sicheren Speicherkonzept unbeantwortet. Vielfach werden externe Festplatten oder optische Datenträger zur Langzeitarchivierung verwendet – ohne die enormen Risiken zu kennen, die diese Datenträger beinhalten.

Der Vortrag stellt die Langzeitarchivierungsstrategie der Julia Stoschek Collection vor, einer der anerkanntesten Privatsammlungen im Bereich Videokunst (http://www.julia-stoschek-collection.net/). Exemplarisch werden die Strategie zur Langzeitarchivierung der Videokunstwerke sowie die Lösung zur Speicherung von born-digital content vorgestellt. Die mehrstufige Strategie zur Sicherung der Videokunstwerke basiert einerseits auf der physikalischen Lagerung von Medien in einem speziell designten Mediendepot und andererseits der redundanten Speicherung der Digitalisate. Diese werden z.T. vor Ort sowie in Zusammenarbeit mit einem Dienstleister zur Speicherung von audiovisuellen Inhalten in einem „deep archive“ (robotergestützte Tape-Library) mehrere Hundert Kilometer entfernt gesichert. Zusätzlich wird eine exakte Erfassung der Metadaten bei der Akquisition der Kunstwerke sowie ein permanentes Obsoleszenz-Monitoring der in der Sammlung befindlichen Medienformate und Codecs durchgeführt.

1 Die Sammlung

Die Julia Stoschek Collection ist eine internationale private Sammlung zeitgenössischer Kunst mit dem Fokus auf zeitbasierten Medien. Sie wurde 2007 in einer alten Rahmenfabrik in Düsseldorf für die Öffentlichkeit zugänglich gemacht. Neben dem Erweitern des Sammlungsbestandes sind die Restaurierung und der Erhalt der Kunstwerke Schwerpunkte der Sammlungstätigkeit. Derzeit umfasst die wachsende Sammlung mehrere Hundert Medienkunstwerke auf Filmen, Dias, Videobändern, Optischen Datenträgern, Festplatten, USB-Sticks sowie weiteren Medien.[1]

2 Risiken für audiovisuelle Archive/Sammlungen

Gerade bei heterogenen Sammlungen, die eine große Vielfalt an unterschiedlichen Medientypen beinhalten, können Erhalt und Pflege eine zeitintensive Herausforderung darstellen. Denn einerseits müssen die Inhalte dauerhaft gesichert und bei Kopier- und Abspielvorgängen möglichst unverändert transferiert werden. Andererseits gilt es, die Datenträger, bei denen es sich um optische Datenträger, Bänder, Flash-Speicher oder Festplatten handelt, physisch unversehrt zu bewahren.

Wie andere Kunstwerke auch, benötigen Medienkunstwerke konservatorische Betreuung und Aufmerksamkeit, um nicht durch vorzeitige Alterung Schaden zu nehmen. Nicht nur die Art und Weise der Lagerung entscheidet über den Bestand des Kunstwerks, auch technische Faktoren spielen eine wichtige Rolle. Durch den technischen Fortschritt werden Medienformate aber auch Codecs in immer kürzeren Abständen obsolet („Obsoleszenz von Formaten“). Bereits nach wenigen Jahren ist es häufig sehr komplex, passende Abspielgeräte für die jeweiligen Datenträger zu finden.

Selbst wenn diese vorhanden sind, muss deren einwandfreie Funktion gewährleistet sein, um einen Datenträger nicht durch technische Fehlfunktion zu beschädigen. Häufig bereitet auch die Alterung der Speichermedien Probleme. So sind beispielsweise Videobänder des Formats U-matic häufig von einer chemischen Abbaureaktion betroffen, die dazu führt, dass die Magnetschicht klebrigen Abrieb aufweist – betroffene Bänder können ohne vorherige Behandlung und Reinigung nicht problemfrei abgespielt werden. Auch VHS, das Hauptdistributionsformat der 1980er und 1990er Jahre, bleibt nicht von Alterungsreaktionen verschont.

Neben Alterung und Obsoleszenz beeinflussen weitere Faktoren den Langzeiterhalt von Medienkunstwerken. Diese sind zum Teil eng miteinander verwoben, sodass gewisse Abhängigkeiten bestehen und ein einzelner Faktor nicht für sich alleine betrachtet werden darf. Für den Erhalt wichtig sind,

[1] Quelle: www.julia-stoschek-collection.net

- Das Klima und die Lagerung der Medien,
- der Zustand der Medien (und Abspielgeräte),
- Lebenserwartung von Formaten und Medien,
- eine Strategie zur Langzeitarchivierung – sowie deren Komplexität.

3 Vorgehensweise

Diese Rahmenbedingungen geben zwangsläufig eine Reihenfolge für die Entwicklung einer Strategie zur Langzeitarchivierung vor. Während die Abfolge der einzelnen Schritte auch auf die meisten anderen Sammlungen übertragbar ist, sind es die Ergebnisse und Schlussfolgerungen nicht unbedingt. Denn jede Sammlung ist einzigartig, hat unterschiedliche Schwerpunkte und oft auch unterschiedliche Anforderungen. Deshalb sind die hier vorgestellten Lösungen und Handlungsempfehlungen auch nicht für jedes Archiv oder jede Sammlung gleich sinnvoll oder nützlich.

4 Bestandsanalyse

Vor der Entwicklung einer Strategie zur Langzeitarchivierung sollte eine Bestandsanalyse durchgeführt werden, bei der auch darüber nachgedacht werden sollte, wie sich der Medienbestand in der Zukunft verändern wird. Diese Maßnahme bildet die Grundlage aller folgenden Planungen: denn nur wenn bekannt ist, welche Formate sich in welchem Zustand in der Sammlung befinden, können weitere Schritte geplant werden. Insbesondere bei älteren Medientypen, die über einen längeren Zeitraum nicht mehr abgespielt wurden, ist Vorsicht geboten. Um Komplikationen auszuschließen, müssen die Bänder vor dem ersten Abspielen auf Alterungsschäden und ihren Zustand untersucht werden.

5 Strategie zur Langzeitarchivierung

Audiovisuelle Datenträger bestehen immer aus zwei Komponenten: der gespeicherten Information und dem eigentlichen Trägermedium. Deshalb muss eine Strategie zur Langzeitarchivierung immer beide Faktoren berücksichtigen. Oberste Priorität hat dabei jedoch der unverfälschte Erhalt des Inhalts.

Das Trägermedium sollte trotzdem nicht unbeachtet bleiben, denn es gehört zum Gesamtkunstwerk – insbesondere dann, wenn es künstlerisch gestaltet ist. Einige Medien in der Stoschek Collection tragen beispielsweise die Künstler/innen-Signatur oder wurden in aufwändigen Verpackungen erworben. Doch selbst wenn dies nicht der Fall ist, so handelt es sich immerhin um das Ankaufsmedium, welches das vom Künstler autorisierte Kunstwerk trägt.

Bereits 2005 wurde für die Julia Stoschek Collection eine Archivierungsstrategie entwickelt, die einerseits dem originalen Träger ein optimiertes Umfeld garantiert und gleichzeitig den Inhalt dauerhaft bewahrt. Diese Strategie musste allerdings – bedingt durch den schnellen technischen Fortschritt – immer wieder den aktuellen Bedürfnissen angepasst werden. Erstmals geschah dies beim Wechsel von Standard-Definition (SD) zu High-Definition (HD) und auch das allmähliche Verschwinden von proprietären Videomedien wurde bereits implementiert.

Während der Entwicklungsphase der Strategie wurden unterschiedliche Möglichkeiten zur Langzeitarchivierung skizziert und erörtert. Sie basierten jedoch alle auf der Grundannahme, dass der Erhalt der audiovisuellen Medienkunstwerke langfristig nur auf der digitalen Ebene zu gewährleisten sein wird, da analoge Lösungen zukünftig nicht mehr zur Verfügung stehen. Um einer unüberschaubaren Format-Vielfalt zu begegnen wird bereits vor dem Ankauf einer Arbeit mit der Galerie oder der Künstlerin bzw. dem Künstler festgelegt, in welchem Format die Master geliefert werden sollen.

Dadurch gelingt es, die heterogene Sammlung auf der digitalen Ebene auf wenige, etablierte und einheitliche Zielformate zusammenzufassen. Hierdurch wird der Aufwand zur Pflege reduziert, da lediglich eine überschaubare Anzahl von Formaten regelmäßig überprüft und durch Obsoleszenz-Monitoring abgesichert werden muss. Gleichzeitig wird die gesamte Betreuung vereinfacht. Denn häufig führen komplexe Pläne dazu, dass sie aus Zeit- oder Ressourcenmangel nicht umgesetzt werden (können).

So bildeten die beiden Sätze „Eine Kopie ist keine Kopie“ und „Keep it simple“ die Grundlage für die Entwicklung für das „Drei-Säulen-Modell“. Denn das Vorhandensein einer einzigen Kopie setzt das Kunstwerk einem zu großen Verlustrisiko durch Alterung, technischen Defekten, Diebstahl, Bedienungsfehler oder Format-Obsoleszenz aus.

Um diesen Bedrohungen zu entgehen, wird die Fortdauer des Werkes auf drei unabhängige Säulen verteilt, die auch beim Ausfall einer einzelnen Säule noch eine tragfähige Basis bilden. Deshalb werden neben dem Original noch zwei Kopien auf verschiedenen Trägern und in unterschiedlichen digitalen Formaten erstellt.

Abbildung 1: Die schematische Darstellung des „Drei-Säulen-Modells“ veranschaulicht die mehrgliedrige Sicherung der audiovisuellen Inhalte. Während die beiden roten Säulen für die Nutzung gesperrt sind, dient die grüne Säule der Benutzung.

Die erste Säule besteht aus den von der Julia Stoschek Collection angekauften Datenträgern, den „Originalen“. Diese werden in einem klimatisierten und gesondert gesicherten Mediendepot gelagert. Als zweite Säule fungiert eine digitale 1:1 Kopie (Datenträger und/oder File), das als Backup (Sicherheitskopie) dient. Mit der dritten Säule werden alle Benutzungen abgedeckt. In Form eines gebräuchlichen Medientyps (z.B. DVD, Blu-ray, File) dient die dritte Säule zum Beispiel als Ausstellungskopie. Dadurch werden die Verwendung des Originalträgers und des Backups verhindert.

6 Mediendepot

Basis und „Schatzkammer“ der Sammlung ist jedoch das Medienkunstdepot. Weil schwankende sowie hohe Temperatur- und Luftfeuchtigkeitswerte für Bänder und Filme schädlich sind, war dies bei der Planung einer der wichtigsten Aspekte. Insbesondere auf niedrige und stabile Feuchtigkeitswerte wurden angestrebt, da dies einer der wichtigsten Faktoren bei der Langzeitarchivierung ist [Sc12]. Als Optimalwerte gelten 25-30% relative Feuchte (±5%/24H) bei 8-10°C (±1°C/24H) [Sc12]. Da diese Werte jedoch nur unter großem technischem Aufwand realisierbar sind, wurden als Kompromiss 35% relative Luftfeuchtigkeit sowie Temperaturen um 15°C realisiert. Auch für andere Medien wie Filme und Dias sind diese Bedingungen sinnvoll.

Um die Klimaschwankungen auch beim Betreten oder Bestücken des Depots so gering wie möglich zu halten, verfügt das individuell für die Julia Stoschek Collection konzipierte und geplante Medienkunstdepot über zwei Schleusen: eine Personenschleuse verhindert abrupte Klimawechsel wenn Mitarbeiterinnen und Mitarbeiter das Depot betreten. Die zweite Schleuse ist für Bänder und Medien konzipiert, die ins Mediendepot eingelagert werden. Sie werden von der Klimaanlage in der Schleuse langsam akklimatisiert bevor sie dann in einer Rollregalanlage aufbewahrt werden. Die hängenden Rollregale sind kugelgelagert und ermöglichen eine optimale Raumnutzung. Um alle Risiken für die gelagerten Videobänder auszuschließen, wurden die einbrennlackierten Regalböden vor der Bestückung auf Restmagnetisierung untersucht. Zudem verhindert die vollständige Erdung der Regale eine statische Aufladung.

Weil Staub und Luftschadstoffe eine ernsthafte Gefahr für Medienkunstwerke darstellen, wird die Luft vor und nach der Konditionierung mehrfach gefiltert. Gleichzeitig sorgen Rauch- und Wassermelder sowie eine mehrstufige Alarmsicherung und Videoüberwachung für einen umfassenden Schutz vor Gefahren und Diebstahl.

7 Speicherstrategie: Formate und Codecs

Obwohl es sich bei audiovisuellen Datenträgern um ein vergleichsweise junges Medium handelt, sind etliche analoge und digitale Formate heute vom Markt verschwunden oder bereits ausgestorben. Über die Langlebigkeit eines Medienformats entscheiden neben der physischen Alterung auch die Verfügbarkeit von Abspielmöglichkeiten, sowie die Marktpolitik einzelner Hersteller und Patentinhaber. Hinzugekommen ist, dass proprietäre physische Medien in den letzten Jahren weitgehend von proprietären Software Codecs abgelöst wurden. Somit muss die prognostizierte „Haltbarkeit" von Software und Schnittsystemen in die Bewertung von zukunftsträchtigen Zielformaten mit einfließen.

Aus konservatorischer Sicht ist es zudem problematisch, dass nahezu alle digitalen Videoformate über daten-reduzierende Mechanismen verfügen, um mit dem beträchtlichen Informationsgehalt der bewegten Bilder fertig zu werden. Der dafür gebräuchliche Begriff „compression" verschleiert dabei jedoch seine tatsächliche Bedeutung: es handelt sich meist um eine verlustbehaftete Datenreduktion („lossy compression"). Im Klartext bedeutet dies, dass bestimmte Anteile der Bild- oder Toninformationen verworfen werden. Dies geschieht einerseits, um Speicherplatz zu sparen, andererseits lässt sich eine geringere Datenmenge meistens einfacher verarbeiten und deshalb auch besser darstellen, da weniger Rechenleistung benötigt wird.

Deshalb muss bei der digitalen Langzeitarchivierung (dLZA) sehr genau geprüft werden, welche Zielformate ausgewählt werden. Bei der Wahl eines ungeeigneten Zielformats besteht die Gefahr, dass es zu sichtbaren Veränderungen des Erscheinungsbildes kommt. So kann z.B. die verlustbehaftete Kompression den Charakter eines Videokunstwerkes nachhaltig beeinflussen – und das Werk somit substantiell verändern. Zudem besteht die Gefahr, dass in der Zukunft eventuell notwendige weitere Kopier- oder Bearbeitungsvorgänge nochmals Verluste verursachen können und damit die Qualität weiter sinkt, wenn bereits das Ausgangsformat stark datenreduziert ist.

Im Idealfall sollte für die digitale Langzeitarchivierung deshalb ein Format gewählt werden, das auf eine verlust-behaftete Datenreduzierung verzichtet. Ist dies nicht möglich, sollten zumindest Formate ausgewählt werden, die nur über eine geringe Datenreduzierung verfügen und weit verbreitet sind.

7.1 Bandbasierte Medienkunst

Für Erwerb und Langzeitarchivierung im Mediendepot werden in der Julia Stoschek Collection je nach Quellmaterial unterschiedliche Zielmedien oder Formate verwendet. SD-Arbeiten werden auf DigitalBetacam Kassetten erworben und gelagert – bei Arbeiten, die in HD entstanden, kommen HDCAM Kassetten zum Einsatz. DigitalBetacam hat sich seit seiner Einführung 1993 durch Sony als ein weltweiter Quasi-Standard im Fernsehen für SD-Video etabliert. Für die Verwendung als Backup sprechen zudem die relativ milde Kompression der Daten (ca. 2:1) und die in den zurückliegenden Jahren erwiesene Qualität und Zuverlässigkeit. Gleiches gilt für die HDCAM-Familie, die sich ebenfalls als langlebiges und zuverlässiges Format für HD Arbeiten etablieren konnte – bei hohen Datenraten und geringer Datenreduktion.

Auch wenn die Langzeitarchivierung von proprietären Medien wie DigitalBetacam oder HDCAM anachronistisch erscheint, so macht sie doch (noch) Sinn – und dies aus ganz praktischen Gründen, die auf dem eingangs zitierten Satz „keep it simple" basieren. Denn Videokassetten funktionieren immer nach dem Schlüssel-Schloss-Prinzip: jedes Kassettenformat „passt" nur in das dafür vorgesehene Abspielgerät (mit Ausnahmen). Und dies ist bereits mit bloßem Auge erkennbar. Niemand würde versuchen, eine MiniDV Kassette in einem DigitalBetacam Abspielgerät abzuspielen – und umgekehrt. Somit wird schon eines der größten Probleme der dateibasierten Langzeitarchivierung eliminiert: die oftmals nicht ohne Hilfsmittel sichtbare Kompatibilität von Codec und Abspielgerät bzw. Rechner. Gleichzeitig handelt es sich um ein physisches Format, das relativ einfach archiviert werden kann. Denn die Kassetten werden im Mediendepot im Regal gelagert.

7.2 Filebasierte Medienkunst

Doch auch bei dieser Strategie gab es zunächst wenige Ausnahmen, die mittlerweile zur Regel geworden sind. Künstlerinnen und Künstler arbeiten kaum mehr in bandbasierten Workflows. Klassische Videomedien sind praktisch ausgestorben. Stattdessen wird dateibasiert gedreht und gearbeitet. Und das schlägt sich dann in den gelieferten Arbeiten nieder. Mittlerweile werden die meisten Arbeiten auf Flash-Cards, USB-Sticks, mobilen Festplatten oder Miniaturrechnern (z.B. Mac Mini) geliefert. Insbesondere bei HD Inhalten ist letzteres eine häufige

Variante. Aus diesem Grund wurde die Strategie zur Langzeitarchivierung um eine Säule für „born-digital content“ erweitert.

Angelehnt an das OAIS Referenzmodell werden für die langfristige Archivierung des audiovisuellen Kunstwerks (d.h. das Informationspaket, bestehend aus Inhaltsdaten und Erhaltungsmetadaten)[2] nicht nur die eigentlichen Files gespeichert, sondern auch alle ermittelbaren technischen und deskriptiven Metadaten in der Datenbank erfasst und vorgehalten.

Beim Erwerb und Eingang (=Übernahme nach dem OAIS Referenzmodell[3]) in die Sammlung wird zunächst eine Sichtkontrolle und einfache Qualitätskontrolle des audiovisuellen Kunstwerks durchgeführt. Anschließend werden die technischen Metadaten aus dem File extrahiert und gemeinsam mit den von der Künstler/in oder Galerie übermittelten Informationen über Inhalt, Aufführungsbestandteile und -besonderheiten in der Datenbank gespeichert. Die technischen Metadaten werden mit der frei verfügbaren Software Mediainfo[4] ermittelt und als PDF File in die Dokumentation eingepflegt. Zukünftig wird für jedes in die Sammlung übernommene File eine MD5 Checksumme gebildet und gespeichert werden, um die Datenintegrität nachverfolgen zu können.

Während für die dLZA von Audiodateien klare Richtlinien und Empfehlungen existieren [Br09], fehlen diese für Videoinhalte.[5] Dort sind ähnliche Empfehlungen angekündigt - jedoch noch nicht publiziert worden. Vielmehr wurde von verschiedenen Seiten darauf hingewiesen, dass für die Langzeitarchivierung möglichst verlustfrei arbeitende Formate ausgewählt werden sollten.[6]

Die Wahl eines Formats für die filebasierte Langzeitarchivierung erfolgt immer im Kontext der aufbewahrenden Institution, seiner “Lieferanten” und “Kunden” sowie der vorhandenen personellen und ökonomischen Möglichkeiten. Homogene Sammlungen, wie z.B. audiovisuelle Archive, werden hierbei sicherlich andere Wege gehen (können) als z.B. Museen und kleinere Sammlungen, die mit heterogenem Sammlungsbestand arbeiten. Zeitgenössische Videokunst entsteht auf vielfältigste Weise und stammt aus allen denkbaren Regionen und Kulturkreisen. Aus technischer Sicht handelt es sich um eine riesige Bandbreite an möglichen Formaten die vornehmlich aus dem Consumer-Bereich stammen. Teilweise werden Formate gemischt oder auch in kreativer Form kombiniert. In vielen Fällen handelt es sich um professionell arbeitende Künstlerinnen und Künstler, die über großen technischen Sachverstand verfügen - in einigen Fällen jedoch auch nicht.

Ziel der Sammlung ist es, ein Videokunstwerk bereits beim Erwerb vom Künstler/in oder der Galerie in den Formaten zu erhalten, in denen es a) produziert wurde (=Original/Master), b) langzeitarchiviert werden soll (= Sicherheitskopie/Submaster) sowie c) vorgeführt werden soll (=Ausstellungskopie). Nur so kann zweifelsfrei gewährleistet werden, dass jede Fassung auch der künstlerischen Intention entspricht.

Bei der Festlegung, in welchem Format eine dLZA bei der Julia Stoschek Collection erfolgen soll, wurde auf diesen Umstand Rücksicht genommen. Es musste sich daher um ein Format handeln, das von Künstlerinnen und Künstlern im Produktionsalltag auch erstellt werden kann.

Vor der Akquisition eines Werkes werden der Künstlerin/dem Künstler oder der Galerie “Production Guidelines” übermittelt, in denen die jeweiligen Formate definiert sind. Gleichzeitig soll ein umfassender Fragenkatalog zur Entstehung und Aufführung des Werks dabei helfen, auch in Zukunft genügend Informationen für notwendige Rückschlüsse zu sammeln. So werden neben den technischen Metadaten zur Auflösung, Codec, Aspect Ratio, Datenrate, Spurbelegung und vielen weiteren Punkten, auch Informationen zur verwendeten Soft- und Hardware bei der Entstehung wie auch bei der Präsentation abgefragt.

Für die Langzeitarchivierung werden von den Künstlern deshalb 10 bit uncompressed (4:2:2) Files im Quicktime Container angefragt, da diese mit sehr vielen Schnittsystemen kompatibel sind. Dieses File Format ist weit verbreitet und gut dokumentiert. Dadurch ist eine breite Unterstützung – sowohl von professionellen wie auch semiprofessionellen Schnittsystemen – vorhanden. Dies gilt auch für Software-Applikationen und Hardware.

---

[2] nestor materialien 16, Frankfurt 2012, S. 21

[3] nestor materialien 16, Frankfurt 2012, S. 32

[4] http://mediainfo.sourceforge.net/

[5] Empfehlungen für die Langzeitarchivierung von Video wurden von der IASA für Ende 2012 angekündigt jedoch noch nicht veröffentlicht. Somit existieren lediglich Projekt- bzw. Erfahrungsberichte wie z.B. von der Library of Congress, PrestoSpace und einzelnen Universitäten sowie Museen.

[6] u.a. wurden im Rahmen der PrestoSpace, PrestoPrime und Prestocentre verschiedene Publikationen erstellt, die sich mit digitalen Formaten zur Langzeitarchivierung befassten. Im Kontext des PrestoSpace Projekts wurde die erste praktikable Adaption von MXF/JPEG2000 in einer verlustfreien Variante entwickelt. Gleichzeitig hat die Library of Congress Leitlinien entwickelt, wie eine digitale Langzeitarchivierung im Archivkontext aussehen sollte.

Andere Formate, wie z.B. MXF/JPEG2000, die sich für die Langzeitarchivierung von Videokunst sehr gut eignen[7] [We11] wurden zwar in Erwägung gezogen, aufgrund mangelnder Unterstützung im künstlerischen Kontext und der wachsenden Sammlung jedoch nicht weiter verfolgt.

8 Speicherstrategie: Medien

Diese Datenmengen ökonomisch, sicher und langfristig zu speichern war eine der nächsten Herausforderungen. Externe Festplatten schieden nach kurzer Prüfung aus: neben ungeklärten Kompatibilitätsfragen, die vermutlich in wenigen Jahren zu Problemen geführt hätten, war auch die Haltbarkeit von Festplatten in Frage gestellt. Untersuchungen von Google [Pi07] und dem Computer Science Department der Carnegie Mellon University [Sc07] zeigten, dass bereits nach kurzer Nutzungsdauer erhebliche Ausfälle bei den in Servern verbauten Festplatten zu verzeichnen waren. Somit kamen nur redundant arbeitende Speicherlösungen in Frage.

Deshalb wurden die Videodaten zunächst auf dem Server gelagert, der auch die IT Infrastruktur der Sammlung bedient. Dies erfolgt im RAID-Level 1, bei der alle Daten redundant im System vorgehalten werden. Doch auch hier besteht das Risiko, dass im Havariefall alle Daten – und damit auch die Kunstwerke – verloren gehen. Sollte es zu einem Brand, Wasserrohrbruch oder Vandalismus kommen, ist die Speicherung aller Daten an einem Ort äußerst riskant. Dem Leitsatz folgend, der besagt „eine Kopie ist keine Kopie“ wurde deshalb nach einer zusätzlichen externen Speicherlösung gesucht. Dieses Offsite Storage sollte redundant, kostengünstig und vor allem sicher vor externen Einflüssen sein. Ein schneller Zugriff bzw. ein schnelles Rückspielen der Daten war nicht das wichtigste Kriterium.

Zunächst wurden Cloud-Modelle größerer Speicheranbieter untersucht. Hierbei wurde schnell klar, dass einige zwar relativ günstig sind, aber auch Risiken bergen können. Viele Dienstleister haben nicht nur ihren Firmensitz im Ausland sondern meistens auch die Rechenzentren, in denen die Daten gelagert werden. Sie unterliegen somit im Falle von Komplikationen ausländischer Rechtsprechung, was zumindest finanziell unkalkulierbare Folgen haben kann. Gleichzeitig kann nie sicher gesagt werden, wo sich und in welcher „Nachbarschaft“ sich die eigenen Daten tatsächlich befinden. Denn die Daten werden so abgelegt, wie sie eingespielt werden – eine logische Zuordnung erfolgt nur über die Verknüpfung einer Datenbank. Konkret kann das bedeuten, dass die Videokunstwerke auf Laufwerken gemeinsam mit Daten aus Versicherungen, Industrie oder aber sonstigen Kunden gespeichert werden. Im Jahr 2011 wurde ein Fall in den USA publik[8], bei dem das FBI aus Ermittlungsgründen einen Server einer Schweizer Firma beschlagnahmte, weil sie illegales Datenmaterial dort vermutete. Sämtliche anderen Nutzer hatten ohne Vorwarnung keinerlei Zugriff mehr auf die gespeicherten Inhalte. Die virtuelle „Nachbarschaft“ bei der Datenlagerung kann so zu ungeahnten Komplikationen führen. Und auch der Fall einer Insolvenz oder Pleite eines Speicheranbieters sollte berücksichtigt werden. Was geschieht dann mit den Daten, die sich im Ausland auf Laufwerken des Dienstleisters befinden? Wandern sie in die Konkursmasse oder können sie problemlos aus dem Gesamtspeicher herausgelöst werden. Wie sehen die Besitzverhältnisse eines Speicheranbieters aus? Gehört die Hardware einer Leasing-Gesellschaft, oder befinden sie sich tatsächlich im Besitz des Anbieters?

Aufgrund der komplexen Fragestellungen wurde beschlossen, eine Lösung mit einem lokalen Speicheranbieter zu erarbeiten, um diese Fragestellungen eingehen zu können. Hierbei wurden unterschiedliche Lösungsansätze untersucht und verglichen. Da schnelle Zugriffszeiten keine Priorität hatten und die Kosten gering gehalten werden sollten, kamen Lösungen, die auf Speicherbändern basieren, schnell in die engere Auswahl. Eine robotergestützte Tape-Library bietet den Vorteil, dass sie aufgrund niedrigerer Betriebskosten deutlich günstiger betrieben werden kann als vergleichbar große Festplatten-Server. Diese sind permanent „on“, verbrauchen viel Energie und erzeugen Wärme, die teuer abgeführt werden muss. Ein Datenband hingegen wird nur dann bewegt, wenn die dort gespeicherten Informationen abgerufen werden sollen.

Die entwickelte Lösung basiert auf einer Tape-Library mit IBM Jaguar-Tapes, die von einem Speicheranbieter in Norddeutschland betrieben wird. Das redundante System ist auf die Speicherung von audiovisuellen Inhalten ausgelegt und verfügt über eine leicht zu administrierende Oberfläche, die es ermöglicht, klare Schreib- und Leserechte zu erteilen. Im System werden beim Ingest automatisch Vorschau-Videos erzeugt, die es ermöglichen, die gespeicherten Inhalte in einfacherer Qualität zu sichten ohne die im „Deep-Archive“ gespeicherten Master-Files

[7] siehe hierzu den Projektbericht zur Langzeitarchivierung des Essl Museums. Bei der Langzeitarchivierung des analogen Videobestandes wurde eine Digitalisierung der SD Inhalte in Kooperation mit dem Phonogrammarchiv in MXF/JPEG200 vorgenommen. Die Archivierung neu akquirierter Videokunstwerke erfolgt wie bei der Julia Stoschek Collection.

[8] Reuters: Web hosting firm says FBI took servers in raid. 22.06.2011 (www.reuters.com/article/2011/06/22/us-cybersecurity-raid-idUSTRE75L4S820110622)

zu bewegen. Diese Vorschau-Videos werden auf einem dem System vorgeschalteten Server gehostet. Insbesondere bei Ausstellungsvorbereitungen ist dies ein großer Vorteil, da sie ohne Wartezeiten bereitstehen und von allen Mitarbeiterinnen und Mitarbeitern angesehen werden können. Um das beschriebene Risiko der ungewollten virtuellen Nachbarschaft auszuschließen, wurde das System so weiterentwickelt, dass die Inhalte der Sammlung physisch getrennt gespeichert werden. In diesem „Deep Archive" werden die Daten ausschließlich auf gesondert gekennzeichneten Bändern gespeichert, die sich im Besitz der Julia Stoschek Collection befinden. Auf diesen Privat-Bändern werden nur sammlungseigene Daten gespeichert. Damit wird einerseits ausgeschlossen, dass eine Beschlagnahmung wie im geschilderten Fall der staatsanwaltlichen Untersuchungen stattfinden kann und andererseits ist für den Fall einer Insolvenz des Dienstleisters vorgesorgt. Die Bänder befinden sich im Besitz der Stoschek Collection und können nicht in der Konkursmasse aufgehen sondern zweifelsfrei als Sammlungseigentum identifiziert und herausgelöst werden. Durch diese zusätzliche Sicherheitsstufe stiegen zwar die Initialkosten leicht, da die eigenen Bänder immer komplett (und nicht anteilig des tatsächlichen Speicherbedarfs) berechnet werden – doch die laufenden Kosten blieben hierdurch unangetastet.

Eine systeminterne automatisierte Kontrolle gewährleistet die Qualität und Datenintegrität der gespeicherten Daten durch Checksummenvergleiche. Trotz der Individualisierung durch die eigenen Bänder bleiben die Kosten bei diesem System überschaubar und können mit den Aufwendungen für eine selbst gemanagte Lösung konkurrieren.

9 Fazit

Aus den beiden Leitsätzen „Eine Kopie ist keine Kopie" und „keep it simple" wurde die „Drei-Säulen-Strategie" zur Langzeitarchivierung der Julia Stoschek Collection entwickelt. Mit dieser Strategie gelingt es, die heterogene Sammlung von Video- und Filmkunst langfristig zu sichern und zu erhalten. Grundlagen sind einerseits das Mediendepot, das besonders klimatisiert und gesichert, alle physischen Datenträger beherbergt. Daneben steht das digitale Archiv, das alle Files, die sich im Sammlungsbestand befinden, redundant speichert. Dieses digitale Archiv wird von einem externen Dienstleister kostengünstig betrieben und bildet das Backup der Sammlung. Durch die Ermittlung und Speicherung aller relevanten Metadaten zu den einzelnen Medienkunstwerken ist eine genaue Beschreibung der Files gewährleistet und ermöglicht auch in näherer Zukunft eine exakte Identifizierung der Files.

Ein häufig unterschätztes Restrisiko bleibt jedoch: aufgelöst in Nullen und Einsen werden die Medienkunstwerke in einer „Black-Box" verstaut. Wenn aber die Kommunikation zwischen Rechner und Speichermedium irgendwann einmal versagen sollte, bedarf es erheblichen Aufwand, die Daten wieder herzustellen. Unter konservatorischen Gesichtspunkten spielt daher die Dokumentation der Daten eine noch wichtigere Rolle als bisher. Die Pflege und Erfassung der Metadaten sowie das Obsoleszenz Monitoring werden in der Digitalen Ära immer wichtiger. Denn Codecs und Formate verändern sich fortlaufend. Inkompatibilitäten sind daher vorprogrammiert. Archive, Sammlungen und Museen sind deshalb in der Pflicht, sich mit ihrem Bestand intensiv zu befassen.

Aus konservatorischer Sicht ist es sinnvoll, die Sammlung auf ein breites und belastbares Fundament zu stellen. Nur so kann die Gefahr eines Verlusts der Kunstwerke minimiert werden. Gleichzeitig sollte aber bedacht werden, dass die Pflege einer digitalen Medienkunstsammlung Ressourcen und Mittel bindet:
In der digitalen Zukunft werden Umkopieren und Transkodieren zum Alltag gehören. Dass dies nicht immer kostenneutral und mit eigenem Personal zu stemmen sein wird, sollte schon beim Ankauf der Kunstwerke bedacht werden.

10 Literaturverzeichnis

[Br09] IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, ed. by Kevin Bradley. Second edition 2009. (= Standards, Recommended Practices and Strategies, IASA-TC 04). International Association of Sound and Audiovisual Archives.

[Ne12] nestor materialien 16, Frankfurt 2012

[Pi07] Pinheiro, Eduardo, Wolf-Dietrich Weber and Luiz André Barroso. "Failure Trends in a Large Disk Drive Population." Paper Präsentation bei: 5th USENIX Conference on File and Storage Technologies - Paper, 12-13.02.2007: 1-13 San Jose, California. 08.08.2007 http://209.85.163.132/papers/disk_failures.pdf

[Sc07] Schroeder, Bianca and Garth A. Gibson. “Disk Failures in the Real World: What does an MTTF of 1,000,000 Hours Mean to You?” Paper Präsentation bei: 5th USENIX Conference on File and Storage Technologies – Paper, 12-13.02.2007: San Jose, California. 26.03.2007
http://www.usenix.org/events/fast07/tech/schroeder/schroeder_html/index.html

[Sc12] Schüller, Dietrich; IASA TC-05 - A Preview; Prepared for TELDAP; Taipei, 23 February 2012

[We11] Andreas Weisser, Ute Kannengiesser, Langzeitarchivierung neuer Medien im Essl-Museum; in: IIC-Restauratorenblätter, Band 30, Klosterneuburg 2011, S. 171-176

DA-NRW: Eine verteilte Architektur für die digitale Langzeitarchivierung

Sebastian Cuy Martin Fischer Daniel de Oliveira Jens Peters
Johanna Puhl Lisa Rau Manfred Thaller

Historisch-Kulturwissenschaftliche Informationsverarbeitung Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln
{sebastian.cuy, martin.fischer, d.de-oliveira, jens.peters, johanna.puhl, l.rau, manfred.thaller}@uni-koeln.de

Abstract: Die Regierung von Nordrhein-Westfalen plant den Aufbau eines landes-weiten Langzeitarchivs für digitale Inhalte aus dem Bereich des kulturellen Erbes. Dieses soll zugleich als ein Pre-Aggregator für die Präsentation der Daten in der Deutschen Digitale Bibliothek und Europeana dienen.
Als technische Umsetzung wurde ein Softwaresystem entworfen, das ausschließlich auf Open Source Komponenten aufsetzt um solch ein verteiltes, selbst-validierendes digitales Archiv zu realisieren. Die genannte Lösung bietet den Einlieferern mittels sogenannter Contracts (XML-basierte Verträge) einen hohen Grad an Kontrolle über die auf die eingelieferten Objekte angewandten Methoden für Archivierung und Publikation.

1 Das DA-NRW Projekt und seine Zielstellung

Das Digitale Archiv NRW [DA13c] wurde 2010 mit dem Ziel ins Leben gerufen, eine Struktur zu errichten, in der Kultureinrichtungen aller Sparten - Archive, Bibliotheken und Museen, aber ggf. auch andere, wie Bodendenkmal-ämter - des Landes Nordrhein-Westfalen ihre digitalen Daten langfristig kostengünstig verwahren können.

Zunächst wurde ein Prototyp mit einer mehrfach redundanten Speicherarchitektur und Schnittstellen zu anderen Archivierungsinitiativen sowie deren Metadatenstandards entwickelt. Mittlerweile, nach umfangreichen Tests, befindet sich das Projekt in einer Verstetigungsphase, in der die Softwarelösung für den Produktionsbetrieb angepasst wird und gleichzeitig ein organisatorischer Rahmen errichtet wird, der den Dauerbetrieb erlaubt. Gegenstand dieses Papiers ist die Softwarelösung – DA-NRW Suite [Th13] – die einen technischen Kern zur Realisierung von Langzeitarchivierungslösungen beliebigen organisatorisch administrativen Zuschnitts anbietet, nicht die angestrebte organisatorisch-administrative Umsetzung.

Das DA-NRW soll Digitalisate von körperlichen Dokumenten (Retrodigitalisate), aber auch Born-Digital- Materialien archivieren. Wichtige Einlieferer sind Bibliotheken, Archive und Museen. Es ist eine besondere Herausforderung im Projekt, den mitunter recht unterschiedlichen Anforderungen der jeweiligen Sektoren an ein digitales Archiv gerecht zu werden. Die genannten Einrichtungen verfolgen häufig sehr unterschiedliche Aufträge und Ziele: Die Aufgabe von Bibliotheken ist es, Publikationen der Öffentlichkeit zugänglich machen; dagegen verwahren Archive auch private, personengebundene Daten wie Nachlässe oder kommunale Dokumente, die strengen Datenschutzrichtlinien unterliegen. Bibliotheken sammeln weitestgehend gedrucktes Material, welches eine relativ einheitliche Struktur aufweist, während beispielsweise Museen vielfältige dreidimensionale Werke oder audiovisuelles Material beherbergen.

Das DA-NRW muss sich also sowohl auf unterschiedliches Material verschiedener technischer und struktureller Komplexität einstellen als auch die unterschiedlichen Einschränkungen für die Veröffentlichung einzelner Bestände berücksichtigen.

Das im Folgenden beschriebene System und insbesondere die entwickelte Softwarelösung wurden in einer nach den Prinzipien des „Open Archival Information Systems“ (OAIS) [Co12] aufgebauten Architektur realisiert. Die Begrifflichkeiten für Pakete aus dem OAIS-Modell: SIP (Submission Information Package), AIP (Archival Information Package), DIP (Dissemination Package), sowie die Bezeichnungen für die Workflowkomponenten Ingest, Data Management, Access und Presentation werden also stets im Sinne dieses Standards verwendet.

Das DA-NRW gewährleistet die langfristige Speicherung des eingelieferten Materials an verteilten Lokationen mithilfe redundanter Speicherung. Ein besonderes Merkmal der entwickelten Technik ist die regelmäßige automatische Überprüfung der Integrität der Pakete an den einzelnen Speicherorten auf der Basis hinterlegter Checksummen.

2 Abläufe im DA-NRW

Die folgende Abbildung stellt den organisatorischen Aufbau des Projekts dar:

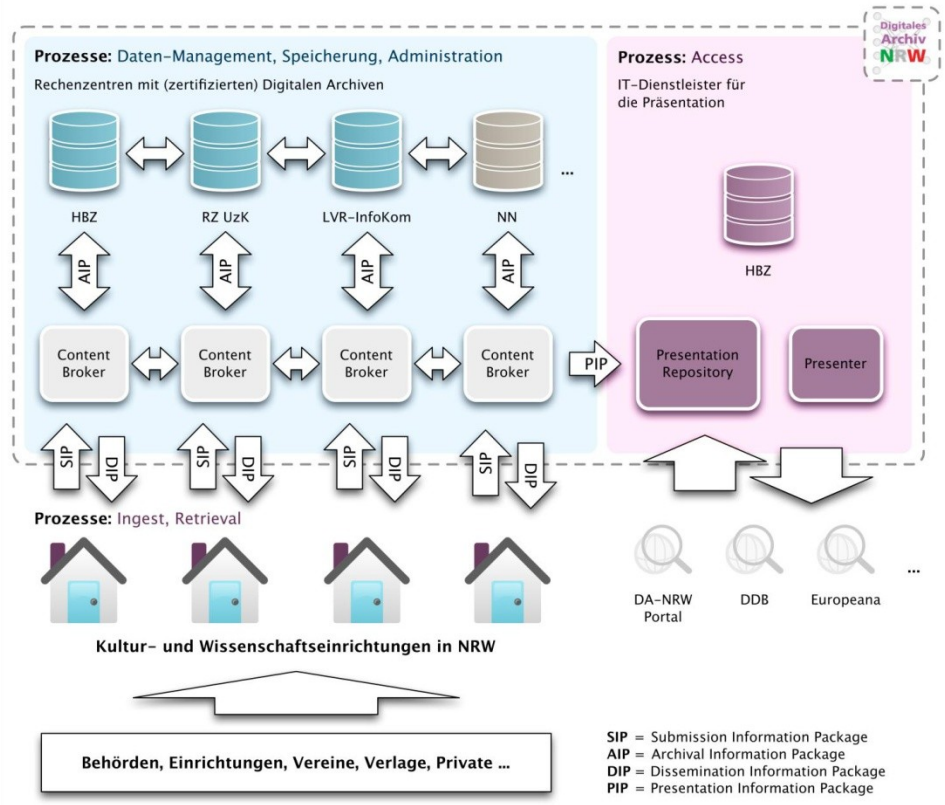


Abbildung 1: Architektur des DA-NRW

In der dargestellten Skizze stellen sowohl ContentBroker als auch das Presentation Repository eigene Software-Entwicklungen im Rahmen des Projekts dar.

Die Einlieferer – Einrichtungen des kulturellen Erbes in NRW – reichen ihre Daten über eine entsprechende Schnittstelle in das DA-NRW ein (Ingest), wo sie über den ContentBroker und die Datenhaltungsschicht verteilt auf den „Knoten“ der Betreiber gespeichert werden (Preservation Management, Storage). Von dort können die Einlieferer die Daten auch wieder anfordern (Retrieval). Mit „Knoten“ wird dabei die Hardware eines technischen Dienstleisters (Rechenzentrums) bezeichnet. Diese kann alle Softwarekomponenten der DA-NRW Suite unterstützen, aber auch auf eine Teilmenge beschränkt werden, wodurch sie dann als „intelligenter Mirror“ erscheint.

Neben diesen Komponenten zur Langzeitspeicherung gibt es den Präsentations-Bereich. Der ContentBroker sendet dafür entsprechende Daten an das so genannte Presentation Repository, welches als Aggregator für Portale fungiert und somit den Zugriff auf dafür frei gegebene Pakete ermöglicht.

OAIS konform durchläuft ein zur Archivierung in das DA-NRW-System eingespieltes Informationspaket mehrere Stadien. Für die Einlieferung wird zunächst ein *SIP* (Submission Information Package) gebildet. Sein Inhalt ist frei vom Einlieferer zu bestimmen. Empfohlen wird, Einheiten inhaltlich zusammenhängender Daten (intellectual entities), als Einzelpakete abzulegen. Ein so erstelltes SIP enthält die zu archivierenden Originaldaten und Metadaten, aber explizit die Archivierung betreffende Zusatzinformationen, beispielsweise darüber, ob die Daten konvertiert und veröffentlicht werden dürfen.

Nach erfolgreicher Übertragung auf einen Knoten des DA-NRW, wird das SIP dort zunächst auf Konsistenz geprüft: D.h., es werden einerseits alle mitgelieferten Checksummen verifiziert, andererseits wird festgestellt, ob jede

mitgelieferte Datei in den Metadaten enthalten ist bzw. ob alle in den Metadaten referenzierten Dateien auch tatsächlich angeliefert wurden.

Für die Archivierung wird aus dem SIP ein *AIP* (Archival Information Package) gebildet, das zusätzlich konvertierte Formate der ursprünglich eingelieferten Dateien in einer weiteren Repräsentation enthält. In der neuen Repräsentation werden Formate verwendet, die als besser geeignet für eine langfristige Speicherung identifiziert wurden (Ein Beispiel für die Einschätzung von Dateiformaten bietet das Florida Digital Archive [FI12]).

Da der Erfolg derartiger Konversionsmaßnahmen nach dem Stand der Technik jedoch nicht automatisch verifizierbar ist, wird das ursprüngliche Format in jedem Fall auch langzeitarchiviert. Ein AIP bekommt während der Verarbeitung einen URN (Uniform Resource Name) zugewiesen und konstituiert damit ein Objekt im Sinne des DA-NRW. Objekte können durch Einlieferung weiterer SIPs, zusätzliche oder neue Fassungen der gleichen Dateien enthalten, und so ergänzt oder ersetzt werden. Das System fasst solche, namentlich durch Angabe einer bereits bestehenden URN gekennzeichneten Objekte, als Deltas auf.[1] Deltas werden genauso wie SIPs im DA-NRW verarbeitet und als weitere AIPs abgespeichert, gehören aber logisch zu einem bereits bestehenden Objekt. Diese „Ergänzung oder Ersetzung" bestehender Komponenten erfolgt jedoch strikt im Sinne einer Versionierung: Auf den vorigen Zustand des Objekts kann ggf. zurückgegangen werden.

Soll das Objekt wieder aus dem Archiv bereitgestellt werden, so wird ein *DIP* (Dissemination Information Package) gebildet. Dieses enthält in Form einer sogenannten „Oberflächenansicht" die neuesten Versionen aller Daten des Objekts, berücksichtigt also sowohl durch Konvertierung oder eine spätere Migration entstandene neue Dateiversionen sowie durch Deltas hinzugekommene Dateien, welche wiederum auch in konvertierter Form vorliegen können.

Eine weitere Form der Dissemination von Paketen stellen *PIP*s (Presentation Information Packages[2]) dar, die Derivate der Dateien des Objektes enthalten, die in der Web-Präsentation zum Einsatz kommen.

Die praktische Umsetzung der genannten organisatorischen und informationsverarbeitenden Konzepte bildet die DA-NRW Software Suite. Die folgenden Bestandteile bilden die Kernkomponenten:

- Die **Datenmanagementschicht** (Realisiert mit der Software „iRODS" [IR13]) sorgt für die redundante Speicherung und das Ressourcen-Management.
- Der **„ContentBroker"** verarbeitet und beauftragt die Verteilung der Pakete.
- Das **„Presentation Repository"** stellt die Schnittstelle zu externen Portalen dar.

Weiterhin zählen zur DA-NRW Software Suite zwei Komponenten, die als Schnittstellen zum System fungieren und damit besonders für die Einlieferer relevant sind:

- Der **„SIP-Builder"** kommt im Pre-Ingest zur Anwendung und unterstützt die Einlieferer dabei, ihre Objekte in SIPs zu wandeln.
- Die **„DA-Web"-Oberfläche** ist die zentrale Schnittstelle zwischen den Einlieferern und dem DA-NRW. Hierüber können Pakete hochgeladen, in der Verarbeitung überprüft und auch wieder angefordert werden.

[1] SIPs, die nicht als Delta eingeliefert werden, werden in Abgrenzung dazu auch als Ersteinlieferung bezeichnet.

[2] Auf abstrakter Ebene handelt es sich bei den PIPs um DIPs im Sinne des OAIS Modells. Die Unterscheidung wurde im Rahmen des DA-NRW getroffen um zwischen zwei unterschiedlichen und häufig verwechselten Situationen zu unterscheiden. Die Bezeichnung DIP (Dissemination Information Package) verwenden wir für eine an die Technologie zum Zeitpunkt des Retrieval angepasste Form des Objekts, die prinzipiell alle im AIP enthaltene Information bereitstellt. Als PIP (Presentation Information Package bezeichnen wir eine Teilmenge dieser Information, die, unbeschadet der langfristigen Sicherung der gesamten Informationsmenge, für eine transiente Nutzung extrahiert wird.

3 Datenaufbereitung / (Pre-) Ingest

Um die im DA-NRW eingehenden Daten und Metadaten korrekt interpretieren zu können, müssen diese bereits bei der Anlieferung in einer geeigneten Form vorliegen.

Der Pre-Ingest als das Verfahren, Daten in geeignete, also vom DA-NRW verarbeitbare Form zu bringen, findet in der Regel bei den Besitzern der digitalen Objekte beziehungs-weise bei von den einliefernden Institutionen beauftragten Unternehmen statt. Um sowohl Einlieferern mit bereits automatisierten SIP-Generierungsprozessen als auch kleineren Institutionen ohne große IT-Abteilung gerecht zu werden, bietet das DA-NRW zwei Möglichkeiten zur Erstellung von SIPs:

Zum einen ist es den Nutzern freigestellt, selbstständig SIPs zu produzieren. Dabei müssen die Vorgaben der SIP-Spezifikation des DA-NRW [DA13a] eingehalten werden.

Zum anderen stellt das Digitale Archiv NRW ein Softwarewerkzeug – den SIP-Builder – zur Verfügung, der die Erstellung DA-NRW-konformer SIPs unterstützt. Dies geschieht, indem vorhandenes Datenmaterial in eine entsprechende Struktur überführt und mit Informationen über die rechtlichen Bedingungen für die Verwaltung der Daten im DANRW versehen wird.

Diese Angabe von rechtlichen Einschränkungen ist ein wichtiger Bonus für die Weiterverarbeitung der SIPs im DA-NRW: Mithilfe dieser Angaben wird festgelegt, ob die zu archivierenden Daten auch von Webportalen über das Presentation Repository abgefragt werden dürfen und, falls ja, in welcher Qualität. Darüber hinaus kann auch die Migration bzw. die Konvertierung in langzeitfähige Dateiformate unterbunden werden.

Das online verfügbare XML-Schema des DA-NRW [DA13b] dient als Grundlage zur Vergabe dieser rechtlichen Einstellungen. Bei der aus dem SIP-Builder resultierenden XML-Datei handelt es sich um den so genannten Contract, der eine Erweiterung des PREMIS-Schemas darstellt [Li13].

Der SIP-Builder steht sowohl als Command-Line-Interface-Version (CLI-Version) als auch als Version mit grafischer Oberfläche (GUI: Graphical User Interface) zur Verfügung. Die CLI-Variante eignet sich für den Einsatz im Rahmen per Skript automatisierter Lösungen, während die GUI-Variante naturgemäß Nutzerinteraktion erfordert, aber auch durch die grafische Oberfläche intuitiver zu bedienen ist.

Unabhängig davon, auf welche Weise die SIPs im Pre-Ingest erstellt werden, obliegt die Definition der Paketgranularität dem Einlieferer.

Die Entscheidung darüber, auf welchem technischen Weg die im Pre-Ingest erstellten Submission Information Packages in die DA-NRW Software Suite eingespeist werden, legt der Einlieferer in Kooperation mit dem gewünschten Abgabeknoten fest. Bei großen Abgabemengen bieten sich hier sicherlich Verfahren unter Einbeziehung der internen, technischen Abgabeschnittstelle im ContentBroker an. Für kleinere Mengen besteht die Möglichkeit SIPs über die Benutzerschnittstelle DA-Web in Kombination mit einer generischen WebDAV-Schnittstelle [We09] abzugeben.

4 Benutzerschnittstellen im DA-NRW

Das DA-NRW-System verfügt neben einer technischen Schnittstelle zur Abgabe großer Datenmengen über eine Benutzerschnittstelle, welche als Web-GUI realisiert ist.

Mittels dieser Web-GUI, die innerhalb des Projekts zunächst nur zur Visualisierung und Kontrolle der Paketverarbeitung konzipiert war, ist auch eine niedrigschwellige Abgabe von Paketen in das DA-NRW-System möglich, indem die Web-GUI auch Zugriff auf den eingesetzten WebDAV Server bietet. Damit bündelt die Benutzerschnittstelle die zentralen Anforderungen eines Einlieferers im DA-NRW, bestehend aus einer möglichst einfachen und komfortablen Oberfläche für den Ingest und das Retrieval. Beide Aufgaben werden unterstützt durch die Möglichkeit einer Einsichtnahme in den Verarbeitungsprozess der von der Institution abgegebenen SIPs sowie einer komfortablen Recherche in den aufbewahrten AIPs. Nach Anforderung wird das AIP als DIP aufbereitet und steht nach der Entnahme aus den Langzeitspeichern im institutionseigenen Ausgangsordner für die entsprechende Institution zur Abholung bereit.

Wie oben bereits erwähnt, arbeitet die Benutzerschnittstelle mit einem WebDAV-fähigen Server zusammen, da nur die Kombination mit WebDAV derzeit eine problemlose Abgabe beliebig großer Objekte über die Protokollfamilie HTTP/HTTPS erlaubt. Neben diesen Protokollen können selbstverständlich auch andere technische Lösungen zur Ablieferung am jeweiligen Knoten des DA-NRW verwendet werden (z.B. SFTP, SSH/SCP), um so die interne technische Abgabeschnittstelle des ContentBrokers direkt anzusprechen.

Aus mehreren Gründen wurde der Webschnittstelle bei der Programmierung die Kombi-nation aus WebDAV und HTTP zu Grunde gelegt: Zum einen, weil eine WebDAV-Unterstützung in den meisten Betriebssystemen bereits vorhanden ist, zum anderen ist die Verwendung der Protokolle HTTP/HTTPS zumeist institutions-und damit firewall-übergreifend kein Problem.

Ebenso ermöglicht die genannte Kombination eine relativ niedrigschwellige Abgabe und Entnahme der Pakete für den Benutzer selbst. Niedrigschwellig in diesem Zusammen-hang meint auch, dass beliebig große Objekte mit Mitteln und Werkzeugen abgegeben und entnommen werden können, die nicht nur auf jedem Betriebssystem grundsätzlich verfügbar sind, sondern auch ohne großen benutzerseitigen Installations-und Konfigurationsaufwand auskommen.

Das zukünftige Potenzial der Webschnittstelle besteht insbesondere in einer möglichen Integration von Aufgaben des Pre-Ingests (DA-NRW Software SIP-Builder) in den Funktionsumfang der Webmaske. Auf diese Weise könnte adaptiv auf die vorhandenen Datenobjekte des Benutzers reagiert werden und die Bildung von SIPs im Pre-Ingest noch besser unterstützt werden.

5 Paketverarbeitung

Die interne Verarbeitung von Paketen, beispielsweise die Umwandlung von SIPs zu AIPs, übernimmt im DA-NRW der ContentBroker, eine Software-Eigenentwicklung, die auf allen beteiligten Speicherknoten zum Einsatz kommt und über diverse Schnittstellen an die übrigen Systembestandteile (iRODS, Presentation Repository, Datenbanken) angebunden ist.

Der ContentBroker führt Modifikationen an Paketen in wohldefinierten Schritten, den sogenannten *Actions* durch. Pakete werden hier als technisches Konzept aufgefasst und bezeichnen zusammengehörige Daten, die je nach Bearbeitungsstadium in einer gemeinsamen Ordnerstruktur oder in einem Container zusammengefasst sind. Diese Stadien entsprechen den organisatorischen Einheiten des OAIS-Informationsmodells. In Form einer Queue – also einer Datenbanktabelle – werden die notwendigen und bereits erfolgten Arbeitsschritte für die aktuell in Bearbeitung befindlichen Pakete auf allen Knoten im DA-NRW System aufgelistet.

Dabei verarbeitet der ContentBroker eines Archivknotens *Jobs*, wenn sie in der Queue erscheinen. Da ein Job jeweils einer Action entspricht und somit als wohldefinierter Arbeitsschritt gilt, kann der ContentBroker beliebige Jobs bearbeiten, ohne dass das System seinen wohldefinierten Zustand verlässt. Verschiedene Pakete können daher auch gleichzeitig bearbeitet werden und sich in unterschiedlichen Stadien befinden.

Diese Möglichkeit wird dazu genutzt, die einzelnen Actions priorisieren zu können. Daneben wird dadurch die Möglichkeit eröffnet, verschiedene Actions mitunter gleichen Typs (z.B. Packen von mehreren Paketen) synchron zu verarbeiten. Zusammen gewährleistet die beschriebene Vorgehensweise eine ressourcenschonende Nutzung der Hardware bei möglichst hohem Durchsatz. Die Priorisierung der Actions und die maximale Anzahl der gleichzeitig möglichen Actions pro Actiontyp sind dabei konfigurierbar.

Wesentliche Bestandteile der Paketverarbeitung sind die Identifikation der Formate der in den SIPs enthaltenen Dateien sowie die Konvertierung dieser Dateiformate. Die Konvertierung dient dabei der Sicherstellung der langfristigen Lesbarkeit der Daten einerseits und Zwecken der Publikation im Netz andererseits.

Die tatsächliche Formatidentifikation wird vom ContentBroker an das externe Programm FIDO [Op12] delegiert. Dabei ist FIDO funktional identisch mit DROID [Na13], jedoch mit wesentlich höherer Performanz [Sh10]. FIDO liefert dabei für jede erkannte Datei einen PRONOM-Identifier [Na06] zurück. PRONOM Identifier geben nicht nur Auskunft über das erkannte Dateiformat selbst, sondern enthalten auch Informationen über die Version eines Dateiformates. FIDO erkennt Formate mithilfe regulärer Ausdrücke und hat sich im Betrieb als robust, zuverlässig und schnell erwiesen. Die gefundenen Identifier werden vom ContentBroker mit einer Liste von *Policies*

abgeglichen. Für jedes erkannte Format wird eine Policy aktiviert, wodurch dann die Ausführung von assoziierten Konvertierungsroutinen in Gang gesetzt wird. Konvertierungsroutinen sind für das DA-NRW System zentral festgelegte und technisch auswertbare Beschreibungen von Formatkonvertierungen, die das Zielformat und das zu verwendende Werkzeug spezifizieren. Sie werden auf den entsprechenden Knoten des DA-NRW unterstützt, indem die spezifizierten Werkzeuge (wie z.B. ImageMagick zur Bildkonvertierung [Im13]) in der jeweils geforderten Programmversion installiert werden. Damit das DA-NRW als eine Einheit homogen agiert, sollen Konvertierungsroutinen reproduzierbare Ergebnisse bei Formatkonvertierungen, unabhängig von der sonstigen Hard- und Softwareausstattung des jeweiligen Knotens, liefern.

Ein Beispiel dafür aus einem tatsächlichen Workflow sind SIPs, die Dateien im JPEG-Format enthalten. In diesem Beispiel würde für jede JPEG-Datei eine Policy aktiviert werden, welche wiederum eine Konvertierungsroutine anstößt, die die Datei unter Verwendung von ImageMagick in das JPEG 2000 Format umwandelt. Abschließend würden die generierten Dateien zusätzlich zu den Originaldateien in das im Verlauf der Paketverarbeitung generierte AIP eingefügt.

6 Datenhaltung im DA-NRW

Eine Kernaufgabe des Projekts ist die Entwicklung einer informationstechnischen Lösung zur Unterstützung von Datenmanagementprozessen und der Ansteuerung der Speicherressourcen an den Knoten. Den einliefernden Einrichtungen gegenüber soll „das" DA-NRW gemäß der bereits gezeigten Architekturdarstellung möglichst als *ein* einziges, gesamtheitliches System erscheinen. So ist es gewünscht, dass die Einlieferung von SIPs weitestgehend an allen Knoten in gleicher Weise möglich ist – das System soll sich also überall in gleicher Weise verhalten. So soll sowohl die Paketverarbeitung der SIPs verteilt auf allen Knoten möglich sein, wie auch der spätere Zugriff (Retrieval) auf die Objekte von jedem Knoten aus.

Im Rahmen der Anbindung von beteiligten Rechenzentren ist es eine besondere Herausforderung, die bereits existierenden, unterschiedlich gewachsenen Hardwarelandschaften in das architektonische Gesamtkonzept möglichst nahtlos zu integrieren. In der aktuellen Systemarchitektur werden neben herkömmlichen SAN-Speichern auch Bandarchive eingesetzt, welche den Vorteil besonders niedrigen Strombedarfs und damit eines besonders nachhaltigen Speicherkonzepts bieten.

Hier besteht ein eindeutiger Zielkonflikt zwischen den abstrakten Anforderungen der Langzeitarchivierung und der Notwendigkeit, ein System in die Vorgaben einer bestehenden Landschaft von Dienstleistern kurzfristig einzubinden. Aus Sicht der Langzeitarchivierung wird die Vorgabe angestrebt, dass Speichermedien mindestens ein Jahr lang ohne Stromzufuhr überlebensfähig sein sollen. Dies läuft, wenn wir uns gleichzeitig auf weit verbreitete Medien konzentrieren, also Technologien ausschließen, die so experimentell sind, dass die benötigten Geräte ohne weiteres wieder vom Markt verschwinden können, derzeit zwangsläufig auf Bänder hinaus; das ökologische und ökonomische Argument, das besonders stichhaltig wird, wenn wir davon ausgehen, dass auf „langzeitarchivierte" Daten u.U. mehrere Jahrzehnte nicht zugegriffen wird, wurde schon erwähnt. Da eine Reihe von IT Dienstleistern, die eingebunden werden sollen, in den letzten Jahren jedoch die Entscheidung getroffen haben aus anderen Gründen auf die Bandtechnologie zu verzichten, müssen derzeit andere Speichertechnologien in Kauf genommen werden.

Eine weitere, nicht zu vernachlässigende Anforderung liegt Im Aufbau schneller und automatischer Abgleichmechanismen, die zwischen den Knoten benötigt werden. Ferner sollten einzelne Knoten theoretisch in der Lage sein, spezielle Technologien, etwa Konvertierungsroutinen für spezielle Formate, dem gesamten System zur Verfügung stellen zu können.

Insbesondere die Anforderung, das System nach außen als ein Verbundsystem betreiben und benutzen zu können, machte den Einsatz einer Rechnerstruktur nötig, die gemeinhin als Grid-Architektur bezeichnet wird. Das DA-NRW folgt hier dem Definitionsvorschlag von Ian Foster, der ein Grid als ein System beschreibt, in dem die Nutzung gemeinsamer Ressourcen dezentral koordiniert wird und standardisierte, offene Protokolle und Schnittstellen verwendet werden, um in der Konsequenz nicht triviale Dienste anzubieten. [Fo02]

Zum Betrieb des DA-NRW wurde als klassische „Middleware"-Komponente die Soft-ware iRODS des RENCI-Forschungslabors [Re13] der University of North Carolina (UNC) in Chapel Hill, USA verwendet. Diese zeichnet sich durch ihre langjährige Genese als Grid-Software im Open-Source-Bereich aus.

iRODS (Akronym für *Integrated Rule-Oriented Datamanagement System*) basiert auf der Software SRB [Sa12], die bereits einen vom tatsächlichen Aufbewahrungsort getrennten, logischen Namensraum liefert [Ra10]. Ein logischer Namensraum bedeutet: Das System iRODS trennt den tatsächlichen physikalischen Speicherort einer Datei von seiner hierarchischen Ordnerstruktur, der fortan als logischer Namensraum fungiert. Ein Dateiobjekt des logischen Namensraumes kann so gleichzeitig mehrere physikalische Aufbewahrungsorte aufweisen, besitzt aber nur einen Bezeichner. Dieser Bezeichner fungiert als logische Adresse des Objekts. Der genannte Mechanismus erleichtert den einheitlichen Zugriff auf die Daten durch den ContentBroker und die Integritätssicherung über iRODS.

Im Gegensatz zu seinem Vorgänger SRB bietet iRODS zusätzlich eine Metadatenverwaltung und eine Komponente zum Datenmanagement, die so genannte *RuleEngine*. Mittels dieser können Zugriffs- und Verwaltungsfunktionen auf den Daten ausgeführt werden [Ra10]. Datenobjekte können in der Folge mit Metadaten versehen und gleichartige Verwaltungs- und Bearbeitungsfunktionen auf das Objekt angewendet werden. Die Verwaltungsfunktionen regeln Zugriffsrechte und Operationen auf den Daten, wie z.B. das automatische Erstellen und Ablegen von Kopien an mehreren Lokationen.

Die Software iRODS wird bereits in vielen internationalen Projekten für die Aufbewahrung digitaler Objekte eingesetzt und besitzt gerade eine hohe Reputation für das Datenmanagement von großen Einzelobjekten[3]. Da die Software außerdem bereits in einer Vielzahl von häufig akademisch getriebenen Projekten in der Verwaltung sehr großer Datenmengen verwendet wird, ist es absehbar, dass sich der Funktionsumfang von iRODS auch auf die erwarteten Datenmengen[4] des DA-NRW gut anwenden lässt. iRODS ist auf einer Vielzahl von UNIX-Derivaten lauffähig und somit unproblematisch in der Kompatibilität.

Die Datenhaltung im DA-NRW wird in einem Verbund von iRODS-Servern erledigt, die in der Lage sind, die unterschiedlichen Hardwaresysteme zur Langzeitspeicherung an den Knoten anzusteuern. Auf diese heterogenen Speichersysteme wird das vollständig gebildete AIP übertragen; in der iRODS-Nomenklatur wird es dorthin „repliziert". Durch die Verwendung unterschiedlicher Speichersysteme in der Datenhaltung – zum Einsatz kommen block- und hierarchisch orientierte Speichersysteme verschiedener Hersteller – wird eine sinnvolle Diversität der Aufbewahrungsmedien hergestellt [UI09]. Die Datenhaltung der AIPs selbst erfolgt in unkomprimierten TAR Containern, die eine dem BagIt-Standard (Derzeit als IETF draft auf dem Wege zur verbindlichen Standardisierung [Bo11]) folgende Ordner-Binnenstruktur aufweisen und mit einer von iRODS erstellten Prüfsumme versehen werden.

Die Verwendung von AIPs, die sowohl auf der logischen, als auch auf der physikalischen Ebene ausschließlich auf öffentlich dokumentierte, breit eingesetzte und nicht proprietäre Standards aufsetzen, halten wir für eine der zentralen Anforderungen an den technischen Rahmen einer Langzeitarchivierungslösung. Hier stimmen wir der von TITAN erhobenen Forderung aus dem Multimediabereich voll zu [Ma08]. Da der am angegebenen Ort referenzierte AXIS Standard nicht weiter entwickelt zu worden scheint, und damit jedenfalls das Prinzip des breiten Einsatzes verletzt, wurde er nicht weiter berücksichtigt.

Nach der Replikation der AIPs, also der Kopie an die anderen Speicherknoten, wird die Prüfsumme des zuvor übertragenen Objekts dort erneut berechnet und mit derjenigen am Ausgangsknoten verglichen. Daraufhin wird eine valide Kopie des AIPs in der iRODS-internen Verwaltungsdatenbank iCAT angemeldet. In der Folge obliegt es im Rahmen des Projekts angepassten Systemprozessen, das Objekt über iRODS zyklisch zu prüfen.

Diese Prüfprozesse wurden hauptsächlich als umfangreiche iRODS Rules implementiert und enthalten eine Vielzahl von so genannten MicroServices, die zusammenhängend einen Workflow definieren, der die Objekte auf ihre Integrität hin überprüft. Für diese Integritätsprüfung werden die Objekte zyklisch tatsächlich von den Langzeitmedien angefordert, gelesen und ihre aktuelle Prüfsumme, in Form eines MD5 Ausdrucks, mit derjenigen verglichen, die zum Zeitpunkt ihrer Erstellung auf dem jeweiligen Medium generiert wurde. Bereits während der initialen Replikation eines neuen Objekts auf ein Langzeitmedium findet im Nachgang des Kopierprozesses solch eine Überprüfung erstmals statt. Diese Überprüfung dient neben der Integritätssicherung des Pakets auf dem Medium ebenfalls der generellen Synchronizitätsprüfung, da die Überprüfung nur dann als valide gilt, wenn alle Replikationen die gleiche Prüfsumme aufweisen und wenn die geforderte Anzahl an minimal erforderlichen Replikationen eines Objekts tatsächlich vorhanden ist.

---

[3] Viele Projekte aus dem High-Performance-Computing-Bereich, wie das CCIN2P3. Unter anderem auch die Bibliothèque nationale de France (BnF) und das European Union Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN).

[4] Nach Bedarfserhebungen zu Beginn des Projekts gehen wir von einer Startkapazität von etwa 200 TB aus.

Allen voran die Erfordernis eines gesonderten Logging-Prozesses im Fehlerfall, als auch die Verwendung von speziellen Medientypen, wie u.a. des Tivoli Storage Managers (TSM) von IBM[5]5 und der generellen Nutzung von WORM Medien [Av09] stellten hier besondere Herausforderungen an die Datenhaltungsschicht dar, die eine aufwändigere Anpassung der Software erforderlich machten.

Die Migration von nicht mehr langzeitarchivierungsfähigen Formaten in geeignetere Dateiformate ist nicht Aufgabe dieser Schicht. Konzeptuell wird die Notwendigkeit der Migration von Datenobjekten mit bestimmten Formateigenschaften extern (durch eine kooperative Technology Watch) festgestellt. Der ContentBroker stellt auf Grund der zu den AIPs gespeicherten technischen Informationen dann fest, welche AIPs betroffen sind, entnimmt in einem solchen Fall die betroffenen Pakete aus dem Langzeitspeicher und migriert die entsprechenden Dateien. Aus der Perspektive der iRODS-Schicht handelt es sich hier um einen gewöhnlichen Ingest & Retrieval-Prozess aus dem iRODS Repository (Client / Serverarchitektur). Gleiches gilt für das geplante Repackaging gemäß dem OAIS-Modell.

Da es sich als zweckmäßig herausgestellt hat, den ContentBroker in JAVA zu implementieren, kommuniziert diese Anwendung über eine eigene Schnittstelle, basierend auf der Jargon API, mit dem iRODS Server. Der ContentBroker ist so auch in der Lage, während der Paketverarbeitung eine verteilte Verarbeitung zu initiieren, wobei der iRODS Server den notwendigen physikalischen Kopierprozess an einen anderen Knoten übernimmt. Auf diese Weise können bestimmte Knoten eine dedizierte technische Verarbeitungskompetenz im Rahmen der Paketverarbeitung übernehmen.

7 Präsentation

Die Veröffentlichung von digitalisiertem und digital geborenem Kulturgut im WWW ist in den letzten Jahren zu einer Selbstverständlichkeit für viele Institutionen geworden. Besonders die Entwicklung großer, spartenübergreifender Portale (Europeana, DDB) hat diesen Trend noch verstärkt und die Zugänglichkeit zu Kulturobjekten im Internet erheblich verbessert. Dennoch ist zur Teilnahme an solchen Initiativen ein gewisser technischer Aufwand für die Bereitstellung entsprechender Schnittstellen notwendig, der insbesondere von kleineren Institutionen oft nicht geleistet werden kann. Aus diesem Grund beinhaltet das DA-NRW eine zentrale Komponente zur Speicherung, Verwaltung und Vermittlung von Präsentationsobjekten, das Presentation Repository. Diese Präsentationsobjekte werden in Anlehnung an das OAIS-Modell im DA-NRW *PIP*s (Presentation Information Packages) genannt und bezeichnen speziell für das Web aufbereitete Derivate der eingelieferten Dateien.

Die Aufgabe des Presentation Repository ist im Wesentlichen die Aggregation der unterschiedlichen, digitalen Objekte. Dabei werden nicht nur Metadaten gespeichert und verfügbar gemacht, sondern auch die dazugehörigen Primärdaten. Dieses Feature entbindet Datenlieferanten von der Notwendigkeit, eigene Webserver betreiben zu müssen und garantiert die zuverlässige Verfügbarkeit konsistenter Objekte auch über längere Zeiträume hinweg.

Zur Kommunikation und Datenübermittlung an Dritte implementiert das Presentation Repository standardisierte, auf Webtechnologien aufbauende Schnittstellen für die automatisierte Abfrage kompletter (Teil-) Bestände zur lokalen Weiterverarbeitung (Harvesting) und die direkte Live-Suche in den vorhandenen Präsentations-und Metadaten. Hier zeigt sich ein weiterer Vorteil der zentralen Aggregation von Präsentationsobjekten: Komponenten zur Bereitstellung von Schnittstellen müssen nicht redundant bei den Institutionen betrieben werden und auf technologische Veränderungen kann an zentraler Stelle reagiert werden.

Das Presentation Repository ist direkt in den Workflow des DA-NRW eingebunden und kann damit von den Funktionen zur Formatverwaltung und -konvertierung des ContentBrokers profitieren. Die Wandlung in für die Präsentation im Web notwendige Derivate kann so parallel und analog zu den Konvertierungen in langzeit-archivfähige Formate stattfinden. Auch die Konvertierung der Metadaten nach standardisierten Schemata wird dabei vorgenommen. Die dabei entstehenden PIPs können dann an das Presentation Repository übertragen werden, wo über die angebotenen Schnittstellen auf sie zugegriffen werden kann, so dass sie von verschiedenen Diensten abgerufen werden und in unterschiedlichen Kontexten im Web präsentiert werden können.

[5] Die Anbindung ist mittels des universellen Treibers MSS realisiert [IR12].

Um den Aufwand bei der Entwicklung des Presentation Repository möglichst gering zu halten und im Rahmen der verfügbaren Mittel eine stabile Anwendung zur Präsentation der veröffentlichten Objekte anbieten zu können, wurde bereits zu Anfang des Projekts die Entscheidung gefällt, eine Open-Source-Repository-Software einzusetzen. Aufgrund der Flexibilität und der modularen Struktur fiel die Wahl auf Fedora Commons [Fe13].

Leider traten im Laufe des Projekts und bei steigender Datenmenge auch die Schwächen von Fedora zu Tage. Insbesondere der integrierte Mulgara-Triple-Store zur Verwaltung des Resource Index, aber auch einfache IO-Operationen sorgten mit steigender Anzahl der von Fedora verwalteten Objekte für erhebliche Performanzprobleme. Aus diesen Gründen wird Fedora in der kommenden Projektphase durch eine andere Repository- Lösung ersetzt werden. Die Entscheidung, ob es sich dabei um ein anderes Open-Source-Produkt oder um eine Eigenentwicklung handeln wird, steht zum aktuellen Zeitpunkt noch aus.

Bezüglich der Schnittstellen des Repository hat sich gezeigt, dass Dritte im Wesentlichen über zwei Arten von Schnittstellen auf die Präsentationsdaten zugreifen. Zum einen hat sich OAI-PMH [La02] als Format für den Austausch von Metadaten zwischen Repositorien etabliert. Zum anderen ist es aber notwendig, eine zusätzliche Schnittstelle anzubieten, die eine Live-Suche in den vorhandenen Daten ermöglicht, damit auch Anwendungen ohne eigenen Suchindex gezielte Abfragen im Datenbestand ausführen können. OAI-PMH konnte mit Hilfe des Fedora-Plugins *oaiprovider*, welches auf PrOAI basiert [Pr09], realisiert werden. Die Implementation der Suchschnittstelle erfolgte auf Basis von Elasticsearch [El13].

8 Ausblick/Fazit

Die dargestellte Implementation eines digitales Archivs macht deutlich, dass eine Langzeitarchivierungslösung, die sich durch Bitstream Preservation an mehreren Lokationen, automatische Konvertierung und optionale Präsentation auszeichnet, nicht nur prinzipiell möglich, sondern auch ressourcensparend praktisch umsetzbar ist. Auch wenn im Laufe des Projekts bei manchen Komponenten und Konzepten Schwächen zu Tage traten, so zeigte sich doch, wie eine Vielzahl von bekannten Standards und Tools zusammen mit schlanken, diese integrierenden Eigenentwicklungen miteinander kombiniert werden kann, damit eine möglichst robuste, praxistaugliche Verbundlösung für eine möglichst breite Zielgruppe angeboten werden kann.

Mittelfristig ist die Erweiterung der bestehenden Architektur besonders in den Bereichen Obsoleszenzüberwachung und Präsentation geplant. So soll es möglich sein Dateien, die in zukünftig als obsoleszent markierten Formaten vorliegen, zu diesem Zeitpunkt automatisiert in andere Formate zu migrieren. Um die Zugänglichkeit des Archivmaterials für einen breiten Nutzerkreis mit Hilfe von Webportalen zu verbessern, wurde begonnen ein komplexeres, domänenübergreifendes Metadatenmodell auf Basis von EDM [Eu13] zu implementieren.

Grundsätzliche Herausforderungen für die Zukunft bestehen allerdings in einer besseren Unterstützung des gesamten Pre-Ingest Prozesses. Zwar haben sich inzwischen, sowohl für Primär-als auch für Metadaten, Standards etabliert, jedoch zeigt sich, dass die An-wendung und Interpretation dieser Standards in der Praxis sehr unterschiedlich ausfällt, wodurch eine automatisierte Verarbeitung erheblich erschwert wird. Außerdem variiert die Qualität der vorhandenen Daten je nach technischer, personeller und finanzieller Ausstattung der jeweiligen Institution stark. Die Entwicklung technischer Lösungen zur Unterstützung der Abgabe vor Ort kann hier helfen, Hürden abzubauen.

9 Literaturverzeichnis

[Av09] Avisaro: Digitale, wechselbare Speichermedien. Hannover, 2009.
URL: http://www. avisaro.com/tl/datentraeger.html (12.06.2013).

[Bo11] Boyko, A. et al: The BagIt File Packaging Format (V0.97). Washington D.C., 2011.
URL: http://tools.ietf.org/html/draft-kunze-bagit-09 (29.04.2013).

[Co12] Consultative Committee for Space Data Systems (CCSDS) – Recommendation for Space Data System Practices: Reference Model for an Open Archival Information System (OAIS). Magenta Book, Washington D.C., USA, 2012.
URL: http://public.ccsds.org/ publications/archive/650x0m2.pdf (16.04.2013).

[DA13a] DA-NRW: Wiki – Technische Spezifikationen. Köln, 2013.
URL: http://da-nrw.hki.uni-koeln.de/projects/danrwpublic/wiki/SIP-Spezifikation (17.04.2013).

[DA13b] DA-NRW: Contract Schema. Köln, 2013.
URL: http://www.da-nrw.de/schemas/ contract/v1/DA-NRW-contract-1.xsd (17.04.2013).

[DA13c] DA-NRW: Projektwebseite. Köln, 2013.
URL: http://www.danrw.de/ (17.04.2013).

[El13] Elasticsearch: Elasticsearch. Flexible and powerful open source, distributed real-time. Amsterdam, 2013.
URL: http://www.elasticsearch.org/ (17.04.2013).

[Eu13] Europeana: Europeana Data Model (EDM) Documentation. Den Haag, 2013.
URL: http://pro.europeana.eu/edm-documentation (29.04.2013).

[Fe13] Fedora Commons: Fedora Commons Repository Software. Winchester (USA), 2013.
URL: http://fedora-commons.org/ (17.04.2013).

[Fl12] Florida Digital Archive: Recommended Data Formats for Preservation Purposes. Florida, 2012.
URL: http://fclaweb.fcla.edu/uploads/recFormats.pdf (17.04.2013).

[Fo02] Foster, I.: What is the Grid? A Three Point Checklist. Chicago, 2002.
URL: http://www. mcs.anl.gov/~itf/Articles/WhatIsTheGrid.pdf (16.04.2013).

[Im13] ImageMagick: Convert, Edit, and Compose Images. Pennsylvania, 2013.
URL: http://www.imagemagick.org/script/index.php (17.04.2013).

[IR13] IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. North-Carolina, 2013.
URL: http://www.irods.org (16.04.2013).

[IR12] IRODS: MSS Driver. North-Carolina, 2012.
URL: https://www.irods.org/index.php/ Universal_Mass_Storage_System_driver (12.06.2013).

[La02] Lagoze, C. et. al. (Hg.): The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). USA, 2002.
URL: http://www.openarchives.org/OAI/openarchives protocol.html (17.04.2013).

[Li13] Library of Congress: PREMIS – Preservation Metadata Maintenance Activity. Washington D.C., 2013.
URL: http://www.loc.gov/standards/premis/ (17.04.2013).

[Ma08] Maréchal, G. (TITAN): The AIP is the cornerstone of the implementations of OAIS. San Francisco, 2008.
URL: http://lib.stanford.edu/files/pasig-spring08/GuyMarechal_ FedArchivesAIP.pdf (29.04.2013).

[Na06] National Archives: The technical registry PRONOM. Surrey (UK), 2006.
URL: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx (17.04.2013).

[Na13] National Archives: DROID: file format identification tool. Surrey (UK), 2013.
URL: http://www.nationalarchives.gov.uk/information-management/projects-and-work/ droid.htm (17.04.2013).

[Op12] Open Planets Foundation: FIDO. Wetherby (UK), 2012.
URL: http://www. openplanetsfoundation.org/software/fido (17.04.2013).

[Pr09] Proai: OAI Provider Library. USA, 2009.
URL: http://proai.sourceforge.net/ (17.04.2013).

[Ra10] Rajasekar, A. et. al.: iRODS Primer: Integrated Rule-Oriented Data System. Kalifornien, 2010.

[Re13] Renaissance Computing Institute (RENCI): About. North-Carolina, 2013.
URL: http://www.renci.org/about (16.04.2013).

[Sa12] San Diego Supercomputer Center (SDSC): SRB – The DICE Storage Resource Broker. North-Carolina / Kalifornien, 2012.
URL: http://www.sdsc.edu/srb/index.php/ Main_Page (17.04.2013).

[Sh10] Sharpe, R.: FIDO / DROID comparison. Beitrag auf: FIDO – a high performance format identifier for digital objects. Open Planets Foundation. Wetherby (UK), 17.11.2010.
URL: http://www.openplanetsfoundation.org/blogs/2010-11-03-fido-%E2%80%93-highperformance-format-identifier-digital-objects (12.06.2013).

[Th13] Thaller, M. (Hg.): Das Digitale Archiv NRW in der Praxis. Eine Softwarelösung zur digitalen Langzeitarchivierung. Hamburg, 2013.

[Ul09] Ulrich, D: Bitstream Preservation. In Neuroth, H. et. al.: nestor Handbuch – Eine kleine Enzyklopädie der Langzeitarchivierung. Version 2.0, Boizenburg, 2009.
URL: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel _346.pdf (17.04.2013).

[We09] WebDAV: Specifications. USA, 2009.
URL: http://www.webdav.org/specs/ (12.06.2013).

Emulation as an Alternative Preservation Strategy - Use-Cases, Tools and Lessons Learned

Dirk von Suchodoletz Klaus Rechert Isgandar Valizada Annette Strauch
Albert-Ludwigs University Freiburg, Professorship in Communication Systems
Ulm University, Communication- and Information Center

Abstract: Emulation has evolved into a mature digital preservation strategy providing, inter alia, functional access to a wide range of digital objects using their original creation environments. In contrast to format migration strategies a functional, emulation-based approach requires a number of additional components. These can be provided by Emulation-as-a-Service, implemented and developed as a distributed framework for various emulation-based services and technologies for long-term preservation and access. This paper presents three distinct applications of the emulation-strategy to preserve complex scientific processing systems, to render complex interactive and dynamic digital objects, and to implement a universal migration-workflows utilizing the original environments in which objects were created.

1 Introduction

It can be stated that today's memory institutions are faced with an extensive range of increasingly complex digital objects (DOs) from a wide range of domains. Business processes but also (commercial) research and development are being carried out by using computer systems producing solely digital results. The material to be made accessible over longer periods includes not only the traditional digital objects such as PDFs or images of digitized items. These objects are complemented by primary research data, all types of software artifacts, ranging from educational material to computer games, digital art as well as machines of famous people. In terms of DOs, even 10 years are a huge timespan when trying to keep those objects accessible and usable. Access to individual artifacts or complete software environments as well as reconstructing and re-enacting related processes may be required at some point.

At the moment, migration is the method most often deployed and trusted by memory institutions for long-term archiving and access of digital objects. This strategy takes digital objects through the constantly changing digital environment, made up of changing hardware and software configurations. It usually requires translation of the objects inner structures to an up-to-date schema. Although these translations make it possible to use and render digital objects in actual computer environments, such an approach unnecessarily limits the number of object types that can be archived. Moreover, suitable migration tools, are usually not available for dynamic and interactive digital objects. A further problem that mostly concern businesses and research is the data-centric view of a migration strategy. Modern research involves not only data but also heavily relies on dedicated workflows and tool-chains. Thus, a pure data-centric strategy misses important pieces that are necessary for an authentic re-enactment, for instance to replicate research results or to reproduce a complex digital environment with its complete internal context to be preserved.

Emulation as a long-term digital preservation strategy shifts migration to the hardware level. It complements migration, as it is able to provide a valuable service: by using emulators, a more accurate reproduction of current digital objects and their processing environment can be verified and, to a certain extent, long-term access is guaranteed. While emulation or format-migration strategies may also be applied in the future (i.e. to save costs, react only if required) authenticity may be difficult to verify.
This paper deals with three applications for emulation strategies. Each application is underpinned by a practical use-case that has been examined during the bwFLA project. [1]
One of the project aims is, to provide emulation services in an easily understandable and easily usable fashion. The preservation of a complex scientific processing system is discussed as a first concrete use-case. The workflow and supporting software framework is explained in order to to safeguard ingest and access to these environments. A second use-case shows tools and workflows for rendering complex digital objects, presenting an example for digital art. Finally, a format migration-workflow is demonstrated, utilizing emulation as a scalable and universal migration tool. The respective past desktop environment is wrapped in such a way, that every human interaction is abstracted

[1] 1bwFLA - Functional Long-Term Access, http://bw-fla.uni-freiburg.de, (20. June 2013).

by an automated procedure. Each achieved result and the various problems encountered during the involved process are discussed and evaluated.

2 Emulation-based Preservation Strategies

For more than 15 years, there has been a vital debate on using emulation as a strategy to ensure long-term access to digital records. Although emulation has always been an essential complement of migration, especially for dynamic and interactive artifacts, it is often considered to be too expensive and to require a good deal of technical skills to be a viable solution. As emulation does not require changes to the object or its structure, but migrates the underlying hardware, the original state of the DO and its authenticity is preserved. While technical issues have been mostly solved at present, tackling the costs and scalability is now a major challenge. [BT13]

Emulation does not operate on the object directly but rather addresses the environment in which it was used to create the object. This means, for example, the replication of software and/or hardware through other software. In the best case, it will not make any difference whether the object is handled through an emulated or original environment. Emulators, i.e. specialized software applications running in digital environments, preserve or alternatively replicate original environments. Research on emulation as a long-term archiving strategy has matured since the first reports on archiving of digital information in 1996 [oPAG96], fundamental experiments with emulation executed by Rothenberg [Rot95] and the theoretical and practical work within the long-term preservation studies of IBM and the Netherlands National Library [vDS02]. The next phase was reached by the EU projects PLANETS and KEEP. The former looked into the inclusion of emulation into preservation workflows [vdHvS09, BKK + 09], while the latter was focused primarily on media transfer, emulators and emulation frameworks [PAD + 09, LKMvdH11].

Digital objects cannot be used on their own. They require a suitable digital working environment in order to be accessed. This context must combine suitable hardware and software components so that its creation environment or a suitable equivalent is generated, depending on the type of the DO. No matter which emulator is being chosen, the contextual information of the original environment of the digital artifact in which it has been created is always required. Questions such as ”for which operating systems is WordPerfect 6.0 compatible with?” or ”which tool generated this specific statistical analysis presented in that paper” are less obvious today than years ago. To overcome this knowledge gap, a formalization process is needed to compute the actual needs for an authentic rendering environment. Several concepts like view paths [vDS02, vdHvS09] or specific technical metadata schemata [DA12, DPDC12] were proposed.

While the technical challenges developing emulators are not considered in this paper, usability and accessibility of emulators for non-technical users are crucial. Emulation technology usually resembles a specific computer system. Since the number of different ancient and current computer systems is limited, the number of required emulator-setups is also limited. In order to allow non-technical individuals to access deprecated original environments, the tasks of setting up and configuring an emulator, injecting and retrieving digital objects in and from the emulated environment have to be provided as easy-to-use services. Making these services web-based allows for a large and virtually global user base to access and work with emulated systems. As a result of the bwFLA project, ”Emulation-as-a-Service”[vSRV13] was developed which includes a framework and workflows to build novel cost-effective services for digital objects and associated processes so that long-term access is guaranteed with a predictable quality [RVvL12].

3 Preservation of Complex Machines

In certain situations, preserving the full original machine is inevitable. Also, there are several good reasons for making images of entire computer environments and for maintaining the ability to render them over a period of time. For instance, researchers can be provided with the ability to experience individual users’ or representative users’ old information environments such as politicians’, artists’ and other famous persons information environments. Another example is to get an average/representative user’s desktop from a particular time period for accessing old web pages or certain documents and spreadsheets. Scientific research environments produce additional examples as computer workstations installations have grown and have been modified over time. During perennial research projects, fluctuation of staff is common and consequently system knowledge is volatile. Complex software workflows have been created, usually involving tailored software components and highly specialized tool-chains paired with non-standard system tweaks. Rebuilding the system from scratch is an often

complicated and time-consuming task involving manual labor and requiring technically-skilled personnel. [MRN + 12]

As another example, information stored in Electronic Document and Records Management Systems (EDRMS) can be ruined from an archival perspective if they are taken out of their EDRMS as their context can be lost. Here, it would theoretically be possible to define a metadata standard and preserve sufficient metadata to capture the context that the file came from; however in practice this is extremely difficult. Thus, preserving the complete workstation seems to be more economical and, full functionality of the system can be retained with minimal effort if carried out properly. [CvSC13]

In a very direct application of emulation strategy physical machines are migrated into virtual or emulated ones by making a direct image of a computers hard disk. This image file, representing a virtual disk, can then be run again attached to emulated hardware. Full system preservation can be described as a migration on the hardware layer. This transformation optimally preserves the original system and contained objects with all context in such a way that everything "behaves" in the emulated environment as it did on the original system [Lof10, vSC11]. Limitations in emulators may affect the result, but in many preservation scenarios few technical limitations like a limited screen resolution or color depth do not necessarily decrease the quality and usability of the emulated environment.

Practical aspects of a system imaging workflow are being described in [WRCv12]. During the bwFLA project a first prototype of a semi-automated workflow got implemented which tries to simplify the several involved preservation steps [RVvL12]. The full system-preservation ingest workflow is split into three stages:

Preparation & Characterization In a first step the user characterizes the computer system to be preserved (in the following denoted as "preservation target") by determining the original operating system and computer architecture to select a purpose-made imaging mini environment to boot on the original hardware. Using the original hardware ensures disk adaptor compatibility and allows to gather additional technical metadata such as information on the CPU, amount of memory and original peripherals deployed. To actually perform the imaging process on the original hardware, the preservation target requires certain technical capabilities, e.g. USB-port or optical disk drive (CD-/DVD) and the ability to boot removable media. Furthermore, a (standard) network adapter is required to transfer the image data into the image archive. To ensure the necessary conditions, the user is interactively questioned if the preservation target meets these requirements. Depending on the choices made, the imaging process is prepared either to be carried out on the preservation target, or on a different (modern) computer system. The latter option requires a dismantling of the preservation target and the removal of the hard disk.

A knowledge base on different operating systems regarding their compatibility with emulators and hardware dependencies is part of the framework. The user is presented with known issues based on his previous selections and step-by-step guides describing user-actions to be carried out on the preservation target. Such tasks may include reconfiguration of the system to default graphic drivers or disabling of external dependencies during the boot-process (e.g. mounting network shares, connections to license servers, etc.). External dependencies may be restored in the emulated environment in a post-processing step. Finally, a specially tailored bootable image is generated and made available for download. The bootable image is either to be written on a USB pen-drive or CD/DVD. The medium is able to boot the preservation target using a preconfigured system that contains necessary configuration and credentials to connect to the repository backend to upload at a later point the generated image.

System Imaging By using the software provided, the target machine is activated and the preservation workflow is executed. An automated process launches the imaging process, the gathering of relevant hardware information about the target machine and the uploading of this data to the frameworks data backend. At the moment, the only interactive choice allows the user to select the drive to be preserved, if multiple options are available. By default, the standard boot-device is chosen. Currently, only complete block device preservation is supported. Single partitions and multiple disk configurations including special setups like RAID are planned for inclusion in future work.

Verification & Submission In a last step, the generated disk image is post-processed to be used in an emulator. This may include steps to pre-load required peripheral drivers for the new emulated hardware or the disabling of the original driver configuration. Finally, an emulation component, part of the bwFLA framework, is invoked with the preserved system image. The result is presented to the user for approval. If approval is granted, the image is submitted together with generated technical metadata to a dedicated repository for further processing. The workflow ends with the opportunity for users to update or amend the frameworks software knowledge base.

With integration of full system-preservation workflows into a distributed digital preservation framework, a common knowledge base on preparing, imaging and re-enacting ancient computer systems can be built, providing step-by-step instruction even to non-technical users. Due to the integrated feedback loop, the owner of a machine, subject to preservation is able to describe and test desired properties of the system for later access. Furthermore, external dependencies, either functional or data services are identified. Both, interfaces to the services as well as their availability should be documented and monitored as part of preservation planning.

3.1 Use-Case OS/2-DB2-based Scientific Environment

At the Linguistics Department at the University of Freiburg, a long-running research project was finally shut down. It had started in the 1970ies with documentations into local dialects of the south-west region of Germany.[2] Later on, it became one of the first projects in this particular field of the humanities to move from laborious paper based evaluation to a computerized environment. The collected data, entered into a relational database, was then used to create customized dialect maps depending on various input parameters which were published in numerous publications and theses. Several specific workflows and even custom font and symbol shapes were created to produce PostScript output from a TeX file source. Many researchers and PhD students contributed to the project and refined the workflows over time. The system can still be run and even now produces up-to-date language maps from the data source. Unfortunately, no actual user of the system has a full understanding of how the system was setup and configured.

The system was put together in 1993 and consists of one x86 server machine running OS/2 version 2.1, running a IBM DB2 database and six x86 clients, offering access to the database over a now deprecated LAN infrastructure. Various workflows for map generation could be executed on the clients from network shares. By the end of the project, the server was still fully functional and at least three clients were working completely, two of them more partially.

It was not possible to boot the imaging mini-environment on the original hardware for several reasons such as missing proper boot devices or incompatible network adaptors. Thus, the hard disks had to be removed and connected to a suitable imaging machine. Any pre-processing like resetting the disk driver to general IDE or the hardware specific SVGA driver to compatible VGA was not considered for the lack of deeper OS/2 configuration knowledge.

The SCSI disk got easily imaged to a container file. Additionally, this file back was written back onto a newer SCSI disk as a backup for the (by then) still in use database machine. This test served at the same time as proof of an identical copy of the original. The replacement worked as expected in the original server. The installed system started exactly as it was previously shut down. Different to expectations as most people would assume, IDE to be more widespread and compatible, the dumping of parallel port IDE disks was more challenging compared to the SCSI counterparts. A working IDE adaptor was required which not only correctly recognizes the disk but also produces reliable disk images.

3.2 Discussion

Both the server and the client environments were finally re-run successfully after the relevant hardware drivers for disk, network and VGA had been modified for the new environment. Each machine setup regarding the relevant components like the DB2 database were exactly in the same state in which the original machine was shut down before imaging. Nevertheless, the involved tasks to revive OS/2 in its new environment were rather complex. Not only that the disk imaging process ran into unexpected problems, additionally a couple of post-processing steps were required to boot the original operating system completely. There are two virtual machines and emulators available (QEMU and VirtualBox) supporting OS/2. A couple of changes to the original configuration were to be made to successfully complete the system migration and start the environment. The original desktop screen resolution was demoted to standard VGA as no compatible driver is available to support the original screen resolution. Another issue was the network connection, which is required to access the database and shared folders from the clients. The original Token Ring network was migrated to Ethernet as no equivalent was available in any emulator.

[2] The project was relevant enough to get added to the permanent exhibition of the Uniseum, the Freiburg University Museum, http://www.uniseum.de

Since the object considered for system imaging is much more complex, the measurement of success and completeness of the workflows is not completely clear yet and it remains a topic of ongoing research. The process can be seen as a kind of migration affecting certain parts and aspects of the object. Usually, lower layer components of the hardware-software-stack like the CPU or drivers are affected. These should not influence the object of interest but can definitely do so. The presented method is a suitable and efficient preservation strategy for highly complex and deprecated systems where detailed knowledge of the system is not available anymore. With a full system-preservation, a device is preserved as a "black box" with somewhat limited utility for future use since details on inner mechanics and constructions are not covered by this process and hence potentially lost. If the system setup is completely known and comparably easy to reproduce, a slightly different method can achieve better results with a smaller footprint regarding the object size.

4 Preserving Environments and Processes

The electronic collections of libraries, museums and archives are growing and have an increasingly relevant role in their holdings. These objects are increasingly complex and may require certain software environments to run or render properly. Standard digital preservation methods can lose important parts of DOs and can not address properties like non-linearity or interaction as required, e.g. for electronic teaching material, encyclopedias, multimedia objects, computer games or digital art.

Often, the formats of those DOs are outdated and can no longer be run or rendered on today's systems. Emulation can provide the required digital environments suitable for a given object type. In order to deal with the different classes of objects, and also to cope with their special requirements, emulation can be applied on original system-environments to arbitrarily render DOs. To re-enact a digital object in its original system-environment, a number of additional components and configurations are required.

A traditional method to discover a digital object's runtime dependencies is querying a filetype database like PRONOM [BCHB07]. Several tools have been proposed to resolve software dependencies such as DROID[3] makes use of "file-magic" fingerprints, while other tools utilize system library resolving mechanisms [Jac11]. While these tools and techniques provide valueable information to users, they do not guarantee generation of a suitable rendering environment regarding, for instance, completeness, quality and conflicting dependencies. Identification of file-type and linking applications is only the first step. This information needs to be extended to a viewpath description [vdHvS09] with required additional software, a suitable operating system and hardware emulator. In order to preserve a re-enactable rendering environment, any dependencies from interactive applications to operating system and hardware components need to be identified.

Having a complete viewpath description for a digital object is not sufficient for providing access to it. The system-environment described by the viewpath has to be recreated first. Instantiation of a viewpath implies that all software components from the operating system to the object's rendering application are installed, configured and operational. In most cases, a viewpath instantiation will not be possible without manual user interaction with the emulated environment (e.g. software installation and configuration). However, a significant challenge when dealing with outdated software packages is the diminishing knowledge on how to handle the installation and configuration processes properly. One potential solution is to automate the different installation steps for each relevant package. Another possible approach is to minimize dependency on this knowledge by providing automated configuration and execution within virtualized environments [WB10].

In contrast to a full system-preservation the bwFLA system-environment preservation workflow makes use of the user's knowledge to identify all necessary components of the object's rendering environment such that the rendering environment is complete and there are no dependency conflicts (Fig. 1).
Furthermore, preserving the knowledge on installation, configuration and deployment of software components ensures the recreation process of past system-environments. By providing a preview of the emulated and recreated environment during ingest, the user is able to test if the chosen setup meets the desired rendering quality and functionality. If ingest workflows allow the shifting of quality control to the object producer, memory institutions can ensure the availability and completeness of rendering environments. The technical workflows are split into two different ingest procedures, one handling digital objects to be prepared for long-term access and the other for ingesting missing software dependencies and creating rendering environments (Fig. 1).

[3] The DROID Project, http://droid.sourceforge.net/, (20. June 2013).

As a precondition for this workflow, it is assumed that the digital object is already part of an archive and is available as an Archival Information Package (AIP), containing institution-specific metadata. Furthermore, it is assumed that the contributor has knowledge of the object's rendering environment and is aware of the objects significant properties and expected behavior. Finally, a dedicated software archive infrastructure is necessary.

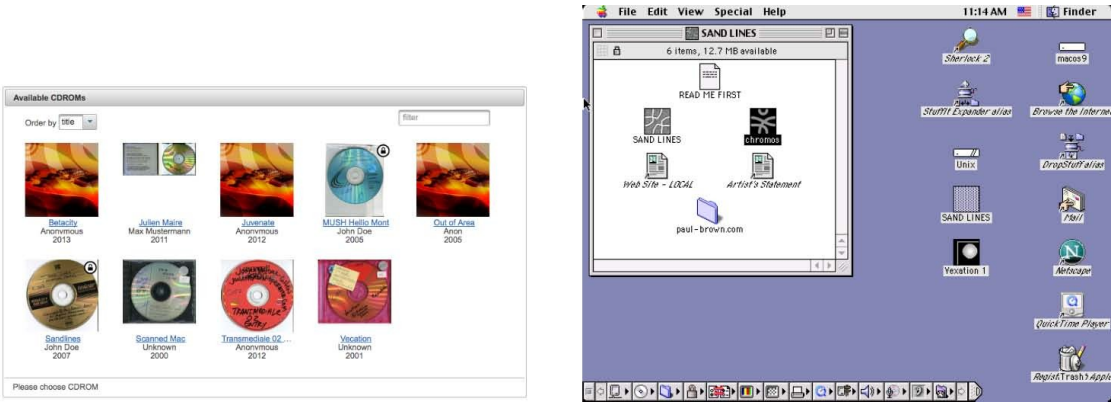
In a first step, the digital object is imported from an archive, metadata and object manifestation is normalized so that it is useable within the bwFLA framework (WF-I.0 in Fig. 1). In the next step, a query for a suitable rendering environment is executed. The software archive suggests the known rendering environments, which the contributor is able to choose from (WF-I.1). If no suitable rendering environment is available, the contributor is redirected to the software archive ingest procedure (WF-I-SW). If a suitable rendering environment has been identified, it has to be instantiated, i.e. all software components defined by the descriptive viewpath have to be installed and configured. This task has to carried out manually if no automated installation routine or pre-configured (i.e. cached) image is available (I-WF.2). The final steps of the workflow allow the user to fine-tune and configure the object's environment (I-WF.4) and to assess the rendering quality (WF-I.3) [GR12]. If the rendering quality is approved, metadata is generated and made available for further processing or storing.
The software archive ingest workflow starts with importing a single software component, e.g. a rendering application (WF-I-SW.0). During this process the user is able to provide detailed descriptive information about the object. This description is used as archival metadata for indexing and search purposes. Furthermore, this description contains information on the license model and intellectual property rights, as well as pointers to additional information like documentation and manuals (WF-I-SW.1). In a second stage of the workflow, the software component's dependencies are determined. If a required dependency is not known or not available in the software archive, it must first be ingested into the software archive by using a recursive invocation of the ingest workflow for this missing dependency. Through an installation and test procedure (WF-I-SW.3), the software component's functionality and completeness of the identified viewpath is verified. For each successfully ingested dependency object the viewpath is extended accordingly. The resulting viewpath then represents a suitable, manually tested and confirmed software environment. The generated metadata information might also include user feedback about the quality and/or costs of the produced technical metadata.

4.1 Use-Case Dynamic and Interactive Objects

Memory institutions like the "transmediale"[4] archive require versatile strategies to preserve, curate and display complex digital objects like digital art or similar types of objects which cannot be directly migrated. For this task, additional software components need to be preserved and enriched with additional information (metadata) like operation manuals, license keys, setup How-Tos and usage knowledge. Furthermore, each software component defines its own soft- and hardware dependencies. To ensure long-term access to digital objects through emulation, not only availability of technical metadata (e.g. TOTEM schema viewpath descriptions [ADP10]) is required, but these viewpaths also need to be tested and evaluated by users who are aware of the digital object's environment properties and performance. This can be achieved by the bwFLA framework with which one can perform a structured installation process of a reference workstation.

The bwFLA access workflow is able to provide a base technology for this task in order to keep digital artifacts alive. In bwFLA, different basic system-environments for manifold purposes were created. The base images include several Apple Macintosh systems offering Mac OS 7.5, 8.5 and 9.0 and various DOS/Windows environments, including the full range of the relevant early Windows versions from 3.11 till 98. On top of the base images, several access scenarios can be run. At the moment, the most advanced is a curation and access workflow for digital art. The use-case offers access to CD-ROM art which was produced, presented and distributed to end-users at the "transmediale" art fair (Fig. 2).

[4] It is an art festival based in Berlin running for over 25 years which forms new connections between art, culture and technology.



(a) Available CD-ROMs of digital art. (b) Digital object rendered in emulated environment.

Figure 2: bwFLA: Functional access to digital art collection.

4.2 Discussion

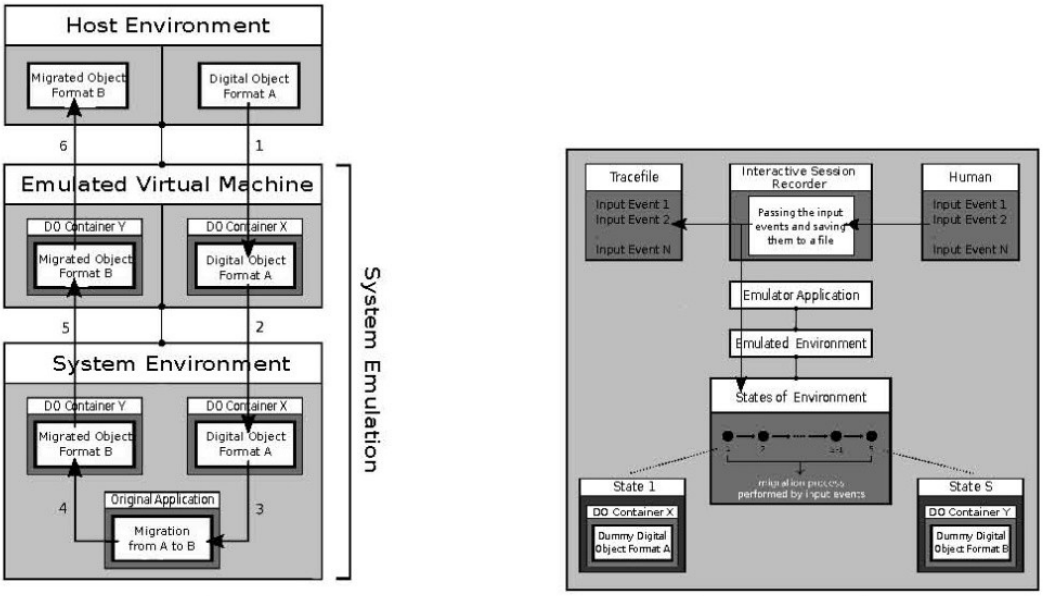
The direct access to many DOs and interaction still offers the most complete and authentic experience. Especially dynamic and/or interactive objects like digital art and many other multimedia material cannot be (easily) preserved using format migration. The proposed workflow requires significant manual user interaction and seems costly and time-consuming at first sight. However, regarding preservation of current digital objects, the basic rendering environment is quite stable, concerning software and hardware dependencies. Usually, the main differences can be found on the top layer of the viewpath description, i.e. only a few additional steps are required if the software archive already contains suitable viewpath descriptions of today's common digital objects. The ingest workflow could be further accelerated by employing caching strategies on created software images and by automation of installation tasks. However, the extra costs in terms of manual interaction during object ingest may reduce the long-term preservation planning costs since only the bottom layer (i.e. emulator) of the viewpath needs to be taken into account.

5 Migration-through-Emulation

Accessing objects directly in their original environment is desirable for authentic reproduction, but can be too laborious and costly if used for manual format migration like required for object normalization in archive ingest. If no suitable migration tools are available for a certain object type, the original application developed by the software producer is the best candidate for handling a specific artifact. Migration-through-Emulation (MtE) describes the concept of using the original or a compatible environment of a designated DO running in a virtual machine and thus replacing the original hardware and/or software stack. This approach avoids the often impossible alteration and adaptation of outdated software to present-day environments. A virtual machine runs within the host environment which contains the selected original system-environment suitable for handling a certain type of digital objects. The original system-environment is either reproduced from original software stored in the software archive or cloned from a prototypical original system.

To make MtE deployable in large-scale preservation scenarios without relying on user interaction, the user's function is replaced by a workflow execution engine [RvSW + 09]. This requires appropriate interfaces in order to use emulators [vS10]. In contrast to simple command-line input-output migration tools, a MtE service builds on aforementioned technologies, e.g. using system emulation, a controlled rendering environment for certain types of digital objects but also an abstract description of all interactive commands to be carried out in order to perform a certain migration. Such a description consists of an ordered list of interactive input actions (e.g. key strokes, mouse movements) and expected observable output from the environment (e.g. screen- or system-state) for synchronization purposes (Fig. 3).

The *migration component* (MC) is the main module visible to the end-user, by exposing a simple *migrate* interface for (possibly) complex DO migration from format fmt A to format fmt B . The user requests a migration by providing a (set of) digital object(s) to be migrated, the requested final format, and a set of parameters. These may restrict the migration path length set quality or cost criteria for the migration process. The individual migration steps are identified. Figure 3 illustrates the general mode of operation of a MC. Based on the resulting identified migration path, the MC instantiates each node as a single migration unit. Beside this, the MC takes care of intermediate results and, if necessary, error reporting and recovery.



(a) MtE functional components. (b) Migration workflow

Figure 3: Abstract and generalized MtE workflow in the bwFLA framework.

5.1 Use-Case Migration of PPT 4.0 to PDF

Migration tools working in todays environments are not available for all file formats. A good example is the Power Point 4.0 format which was used by the end of the 1990ies. Trying to open it in recent office suites produces a "format not understood"-error. For the evaluation a emulation based migration bwFLA workflow was produced, which transforms a PPT 4.0 input into a PDF file output. The workflow is implemented as a service accessible through the web. Another use case could be (chained) migrations to move the outdated format up to an actual version. Several MtEs could be coupled to read the object in a newer version of the producing application and saving it in that one. Beside keeping the format, a wide range of different output formats supported by the producing applications could be generated. Thus a Microsoft Word 97 file, for example could be migrated to RTF, ASCII text or PDF at the same time for different purposes, such as viewing, indexing or further processing in actual software. Further workflows taking different inputs or producing different outputs can be created. Such procedures can then be used to e.g. normalize objects during archival ingest routines.

5.2 Discussion

To verify reliability of the functional encapsulated workflows, especially interactive workflow recordings a simple large-scale format migration process has been created as a testcase. Input format is a Microsoft DOC file version 8.0. Output format is RTF. The application used to migrate input files was MS Word 97 running on Windows 98. After recreation of the software environment, a single DOC file was injected into the emulation component with the help of a virtual floppy image and the system was started. In the next step, automated user interface interaction was

carried out. In this case, the injected file was opened with the MS Word application and the file was exported as in RTF file format. After shutting down the emulated system, the migrated result was then available to the user.

The workflow has been carried out on 997 different objects of the same file type. In the evaluation 892 (89,47%) completed migrations and 105 failures were observed. Execution of a single instance took about 4 minutes. To investigate the reasons of failed migration workflows further, the experiment got re-run on the 105 objects and analyzed screenshots of 30 randomly chosen failure states. In these failure cases, three types of modal windows appeared on the screen. Note, the interactive workflow recording system is an independent platform and thus relies on graphical screen output and emulator stat. The first error category was due to a notification that there is no free space on the hard drive. Originally, the size was set to 30 MByte but due to the migration of embedded images into specific format MS Word 97 (BMP) more disk-space is required. This type of errors happened only when objects contained graphical content. The second and third error category was due to a MS Word specific warning (file was marked as write-protected and the file contained macros).

6 Conclusion and Outlook

Future users of digital assets significantly benefit from accessible data and user-friendly functional toolsets both in the scientific and cultural heritage as well as in the commercial domain. Emulation services for digital preservation can help to bridge outdated working environments for a wide range of objects and original environments onto todays devices.

After a number of successful national and international initiatives and projects on digital preservation and access it is time to take these results to memory institutions. The bwFLA project on functional long-term archiving, has started implementing and integrating different workflows in ”Emulation-as-a-Service”. These provide a range of different environments to deal with various curational and archival tasks using emulation. For original environments, the bwFLA EaaS supports at the moment 8 different emulators being able to run 15 distinct legacy computer platforms. The platforms range from MacOS 7 running on a Motorola 68K system emulator, PPC based platforms to various x86-based platforms. In a distributed EaaS model the costs of archiving secondary digital objects like operating systems and popular applications can be shared. With mutual specialization niches and specific areas can be covered without losing generality. The shift of the usually non-trivial task of the emulation of obsolete software environments from the end user to specialized providers can help to simplify digital preservation and access strategies. EaaS makes emulation and emulators more easy to handle. In combination with distributed access infrastructure structure, EaaS, preservation planning and preservation costs are fixed, determined only by the number of emulators and emulated systems.

While an emulation approach has technical limitations (e.g. due to external (network) dependencies, digital rights management, license dongles, etc.), the proposed workflows allow to uncover such issues and indicate risks w.r.t. to long-term preservation. By integrating emulation-based services, memory institutions are able to acquire new users and to provide new types of services. In cooperation with today’s object creators, memory institutions can not only gain knowledge on future preservation challenges, but are also able to build a solid knowledge base on current formats.

7 Acknowledgments

The work presented in this publication is part of the bwFLA – Functional Long-Term Access project funded by the federal state of Baden-Württemberg, Germany.

8 References

[ADP10] David Anderson, Janet Delve, and Dan Pinchbeck. Towards a workable, emulation-based preservation strategy: rationale and technical metadata. New review of information networking, (15):110–131, 2010.

[BCHB07] Tim Brody, Leslie Carr, Jessie M.N. Hey, and Adrian Brown. PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services.International Journal of Digital Curation, 2(2), 2007.

[BKK+09] Christoph Becker, Hannes Kulovits, Michael Kraxner, Riccardo Gottardi, Andreas Rauber, and Randolph Welte. Adding quality-awareness to evaluate migration webservices and remote emulation for digital preservation. In Proceedings of the 13th European Conference on Digital Libraries (ECDL09), 2009.

[BT13] Daniel Burda and Frank Teuteberg. Sustaining accessibility of information through digital preservation: A literature review. Journal of Information Science, 2013. Broad literature overview on actual research in the field of DP in business sphere, identification of research questions.

[CvSC13] Euan Cochrane, Dirk von Suchodoletz, and Mick Crouch. Database Preservation Using Emulation – A Case Study. Archifacts, (April):80–95, 2013.

[DA12] Janet Delve and David Anderson. The Trustworthy Online Technical Environment Metadata Database – TOTEM. Number 4 in Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik. Verlag Dr. Kovaˇc, Hamburg, 2012.

[DPDC12] Angela Dappert, Sebastien Peyrard, Janet Delve, and Carol C.H Chou. Describing Digital Object Environments in PREMIS. In 9th International Conference on Preservation of Digital Objects (iPRES2012), pages 69–76. University of Toronto, 2012.

[GR12] Mark Guttenbrunner and Andreas Rauber. A Measurement Framework for Evaluating Emulators for Digital Preservation. ACM Trans. Inf. Syst., 30(2):14:1–14:28, May 2012.

[Jac11] Andrew N. Jackson. Using Automated Dependency Analysis To Generate Representation Information. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011), pages 89–92, 2011.

[LKMvdH11] Bram Lohman, Bart Kiers, David Michel, and Jeffrey van der Hoeven. Emulation as a Business Solution: The Emulation Framework. In 8th International Conference on Preservation of Digital Objects (iPRES2011), pages 425–428. National Library Board Singapore and Nanyang Technology University, 2011.

[Lof10] Mary J. Loftus. The Author's Desktop. Emory Magazine, 85(4):22–27, 2010.

[MRN+12] Rudolf Mayer, Andreas Rauber, Martin Alexander Neumann, John Thomson, and Gonçalo Antunes. Preserving Scientific Processes from Design to Publication. In George Buchanan, Edie Rasmussen, and Fernando Loizides, editors, Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012), Cyprus, September 23–29 2012. Springer.

[oPAG96] Commission on Preservation, Access, and The Research Libraries Group. Report of the Taskforce on Archiving of Digital Information. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf, 1996. [online; last accessed 10.06.2013].

[PAD+09] Dan Pinchbeck, David Anderson, Janet Delve, Getaneh Alemu, Antonio Ciuffreda and Andreas Lange. Emulation as a strategy for the preservation of games: the KEEP project. In DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory, 2009.

[Rot95] Jeff Rothenberg. Ensuring the Longevity of Digital Information. Scientific American, 272(1):42–47, 1995.

[RvSW+09] Klaus Rechert, Dirk von Suchodoletz, Randolph Welte, Maurice van den Dobbelsteen, Bill Roberts, Jeffrey van der Hoeven, and Jasper Schroder. Novel Workflows for Abstract Handling of Complex Interaction Processes in Digital Preservation. In Proceedings of the 6th International Conference on Preservation of Digital Objects (iPRES2009), pages 155–161, 2009.

[RVvL12] Klaus Rechert, Isgandar Valizada, Dirk von Suchodoletz, and Johann Latocha. bwFLA – A Functional Approach to Digital Preservation. PIK – Praxis der Informationsverarbeitung und Kommunikation, 35(4):259–267, 2012.

[vdHvS09] Jeffrey van der Hoeven and Dirk von Suchodoletz. Emulation: From Digital Artefact to Remotely Rendered Environments. International Journal of Digital Curation, 4(3), 2009.

[vDS02] Raymond van Diessen and Johan F. Steenbakkers. The Long-Term Preservation Study of the DNEP project - an overview of the results. IBM Netherlands, Amsterdam, PO Box 90407, 2509 LK The Hague, The Netherlands, 2002.

[vS10] Dirk von Suchodoletz. A Future Emulation and Automation Research Agenda. In Jean-Pierre Chanod, Milena Dobreva, Andreas Rauber, and Seamus Ross, editors, Automation in Digital Preservation, number 10291 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany.

[vSC11] Dirk von Suchodoletz and Euan Cochrane. Replicating Installed Application and Information Environments onto Emulated or Virtualized Hardware. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011), pages 148–157, 2011.

[vSRV13] Dirk von Suchodoletz, Klaus Rechert, and Isgandar Valizada. Towards Emulation as-a-Service – Cloud Services for Versatile Digital Object Access. International Journal of Digital Curation, 8:131–142, 2013.

[WB10] Kam Woods and Geoffrey Brown. Assisted Emulation for Legacy Executables. International Journal of Digital Curation, 5(1), 2010.

[WRCv12] Ian Welch, Niklas Rehfeld, Euan Cochrane, and Dirk von Suchodoletz. A Practical Approach to System Preservation Workflows. PIK – Praxis der Informationsverarbeitung und Kommunikation, 35(4):269–280, 2012

BlogForever: From Web Archiving to Blog Archiving

Hendrik Kalb[1] Paraskevi Lazaridou[2] Vangelis Banos[3]
Nikos Kasioumis[4] Matthias Trier[5]
[1]hendrik.kalb@tu-berlin.de, [2]paraskevi.lazaridou@tu-berlin.de, [3]vbanos@gmail.com, [4]nikos.kasioumis@cern.ch, [5]mt.itm@cbs.dk

Abstract: In this paper, we introduce blog archiving as a special type of web archiving and present the findings and developments of the BlogForever project. Apart from an overview of other related projects and initiatives that constitute and extend the capabilities of web archiving, we focus on empirical work of the project, a presentation of the BlogForever data model, and the architecture of the BlogForever platform.

1 Introduction

The aim of this paper is to introduce blog archiving as a special type of web archiving.

Web archiving is an important aspect in the preservation of cultural heritage [Mas06] and, therefore, several projects from national and international organisations are working on web preservation activities. The most notable web archiving initiative is the Internet Archive[1] which has been operating since 1996. In national level, there are several remarkable activities, mainly from national libraries, to preserve web resources of their national domain. For example, the British Library announced this spring a project to archive the whole .uk domain [Coo13].

Web archiving is always a selective process, and only parts of the existing web are archived [GMC11, AAS+11]. The selection seems often to be driven by human publicity and search engine discoverability [AAS+11]. Furthermore, contrary to traditional media like printed books, web pages can be highly dynamic. Therefore, the selection of archived information comprises not only the decision of what to archive (e.g. topic or regional focus) but also additional parameters such as the archiving frequency per page, and parameters related to the page request (e.g. browser, user account, language etc.) [Mas06]. Thus, web archiving is a complex task that requires a lot of resources.

All active national web archiving efforts, as well as some academic web archives are members of the International Internet Preservation Consortium[2] (IIPC). Therefore, the web archiving tools[3] developed by the IIPC are widely accepted and used by the majority of Internet archive initiatives [GMC11]. However, the approach inherent in these tools has some major limitations. The archiving of large parts of the web is a highly automated process, and the archiving frequency of a webpage is normally determined by a schedule for harvesting the page. Thus, the life of a website is not recorded appropriately if the page is updated more often than it is crawled [HY11]. Next to the harvesting problem of web archiving, the access of the archived information is inadequate for sophisticated retrieval. Archived information can be accessed only on site or page level according to a URI because analysis and management of current web archiving does not distinguish between different kinds of web pages. Thus, a page with a specific structure like blogs is handled as a black box.

The blogosphere, as part of the web, has an increasing societal impact next to traditional media like press or TV. Prominent examples are the influential blogs in political movements in Egypt [Ish08, Rad08] or Iran [Col05]. But there are also other domains that people engage in blogging, e.g. in the fields of arts or science [WJM10], teaching [TZ09] or leisure activities [Chi10]. The blogosphere as an institution has two connotations: On the one hand it is considered as a place where people build relationships – the blogosphere as a social networking phenomenon [AHA07, Tia13]. This view is emphasizing the activity of relating to others. On the other hand, it is also important to recognize that the numerous contributions yield a joint creation - the blogosphere as a common oeuvre, an institution shared by all bloggers and readers [KT12]. However, blogs as other social media are ephemeral and some that described major historical events of the recent past are already lost [Che10, Ent04]. Also the loss of personal diaries in the form of blogs has implications for our cultural memory [O'S05].

[1] http://archive.org
[2] http://netpreserve.org
[3] http://www.netpreserve.org/web-archiving/tools-and-software

The BlogForever[4] project creates a novel software platform capable of aggregating, preserving, managing and disseminating blogs. Through the specialisation in blog archiving, as a subcategory of web archiving, the specific features of the blog as a medium can be exploited in order to overcome limitations of current web archiving.

2 Related work

In the following section, we review related projects and initiatives in the field of web archiving. Therefore, we inspect the existing solutions of the International Internet Preservation Consortium[5] (IIPC) for web archiving and the ArchivePress[6] blog archiving project. Furthermore, we look into several research projects such as Longitudinal Analytics of Web Archive Data[7] (LAWA), Living Web Archives[8] (LiWA), SCalable Preservation Environments[9] (SCAPE), Collect-All ARchives to COmmunity MEMories[10] (ARCOMEM), and the Memento[11] project. Table 1 provides an overview of the related initiatives and projects we examine in this section.

Initiative	Description	Started
ArchivePress	Explore practical issues around the archiving of weblog content, focusing on blogs as records of institutional activity and corporate memory.	2009
ARCOMEM	Leverage the Wisdom of the Crowds for content appraisal, selection and preservation, in order to create and preserve archives that reflect collective memory and social content perception, and are, thus, closer to current and future users.	2011
IIPC projects	Web archiving tools for acquisition, curation, access and search.	1996
LAWA	Development of tools and methods to aggregate, query, and analyse heterogeneous Internet data at large scale.	2010
LiWA	Develop and demonstrate web archiving tools able to capture content from a wide variety of sources, to improve archive fidelity and authenticity and to ensure long-term interpretability of web content.	2009
Memento	Development of a technical framework that integrates current and past Web.	2009
SCAPE	Developing an infrastructure and tools for scalable preservation actions.	2011

Table 1: Overview of related initiatives and projects

[4] http://blogforever.eu
[5] http://netpreserve.org
[6] http://archivepress.ulcc.ac.uk/
[7] http://www.lawa-project.eu/
[8] http://liwa-project.eu/
[9] http://www.scape-project.eu/
[10] http://www.arcomem.eu/
[11] http://www.mementoweb.org/

The IIPC[12] is the leading international organization dedicated to improving the tools, standards and best practices of web archiving. The software they provide as open source comprises tools for

- acquisition (Heritix[13]),
- curation (Web Curator Tool[14] and NetarchiveSuite[15]), and
- access and finding (Wayback[16], NutchWAX[17], and WERA[18]).

They are widely accepted and used by the majority of Internet archive initiatives [GMC11].

The ArchivePress[19] project was an initial effort to attack the problem of blog archiving from a different perspective than traditional web crawlers. To the best of our knowledge, it is the only existing open source blog-specific archiving software. ArchivePress utilises XML feeds produced by blog platforms in order to achieve better archiving [PD09]. The scope of the project explicitly excludes the harvesting of the full browser rendering of blog contents (headers, sidebars, advertising and widgets), focusing solely on collecting the marked-up text of blog posts and blog comments (including embedded media). The approach was suggested by the observation that blog content is frequently consumed through automated syndication and aggregation in newsreader applications, rather than by navigation of blog websites themselves.

The LiWA[20] project aims at the improvement of web archiving technologies. Thereby, it focuses on the areas of archive fidelity [DMSW11, OS10], spam cleansing to filter out fake content [EB11, EGB11], temporal coherence [EB11, BBAW10, MDSW10], semantic evolution of the terminology [TZIR10, TNTR10], archiving of social web material, and archiving of rich media websites [PVM10]. The project aims at the creation of long term web archives, filtering out irrelevant content and trying to facilitate a wide variety of content.

The ARCOMEM project focuses mainly on social web driven content appraisal and selection, and intelligent content acquisition. It aims at the transformation of "archives into collective memories that are more tightly integrated with their community of users and to exploit Social Web and the wisdom of crowds to make Web archiving a more selective and meaning-based process" [RP12]. Therefore, methods and tools are developed and research is undertaken in the areas of social web analysis and web mining [MAC11, MCA11], event detection and consolidation [RDM+11], perspective, opinion and sentiment detection [MF11], concise content purging [PINF11], intelligent adaptive decision support [PKTK12], advanced web crawling [DTK11], and approaches for semantic preservation [TRD11].

The SCAPE project is aiming to create scalable services for planning and execution of preservation strategies [KSBS12]. They address the problem through the development of infrastructure and tools for scalable presservation actions [SLY+12, Sch12], the provision of a framework for automated, quality-assured preservation workflows [JN12, HMS12], and the integration of these components with a policy-based preservation planning and watch system [BDP+12, CLM11].

The LAWA project aims at large-scale data analytics for Internet data. Therefore, it focuses on the development of a sustainable infrastructure, scalable methods, and software tools for aggregating, querying, and analysing heterogeneous data at Internet scale with a particular emphasis on longitudinal data analysis. Research is undertaken in the areas of web scale data provision [SBVW12, WNS+11], web analytics [BB13, PAB13, WDSW12, YBE+12, SW12], distributed access to large-scale data sets [SPNT13, YWX+13, SBVW12], and virtual web observatory [SPNT13, YWX+13, ABBS12].

[12] http://netpreserve.org/
[13] http://crawler.archive.org; an open-source, extensible, Web-scale, archiving quality Web crawler
[14] http://webcurator.sourceforge.net/; a tool for managing the selective Web harvesting process
[15] https://sbforge.org/display/NAS/Releases+and+downloads; a curator tool allowing librarians to define and control harvests of web material
[16] http://archive-access.sourceforge.net/projects/wayback/; a tool that allows users to see archived versions of web pages across time
[17] http://archive-access.sourceforge.net/projects/nutch/; a tool for indexing and searching Web archives
[18] http://archive-access.sourceforge.net/projects/wera/; a Web archive search and navigation application
[19] http://archivepress.ulcc.ac.uk/
[20] http://liwa-project.eu/index.php

The Memento[21] project aims to provide access to the Web of the past in the way that current Web is accessed. Therefore, it proposes a framework that overcomes the lack of temporal capabilities in the HTTP protocol [VdSNS+09]. It is now active Internet-Draft of the Internet Engineering Task Force [VdSNS13].

The aforementioned projects are evidence of various remarkable efforts to improve the harvesting, preservation and archival access of Web content. The BlogForever project, presented in the following, puts the focus on a specific domain of Web, the weblogs.

3 BlogForever project

In the following, we introduce the BlogForever project. In particular, we present three surveys that have been conducted, the BlogForever data model which constitutes a foundation for blog archiving, and the two components of the BlogForever platform.

3.1 Surveys about blogs and blog archiving

Several surveys were conducted in the project to reveal the peculiarities of blogs and the blogosphere, and to identify the specific needs for blog preservation.

Two distinct online questionnaires were disseminated in six languages to blog authors and blog readers. The aim was to examine blogging and blog reading behaviour, the perceived importance of blog elements, backup behaviour of bloggers, perceptions and intentions for blog archiving and blog preservation. Complete responses were gathered from 512 blog authors and 428 blog readers. One finding was that the majority of blog authors rarely consider archiving of their blogs. This increases the probability of irretrievable loss of blogs and their data, and, therefore, justifies efforts towards development of independent archiving and preservation solutions. Additionally, the results indicated a considerable interest of readers towards a central source of blog discovery and searching services that could be provided by blog archives [ADSK+11].

A large-scale evaluation of active blogs has been conducted to reveal the adoption of standards and the trends in the blogosphere. Therefore, 259,390 blogs have been accessed and 209,830 retrieved and further analysed. The evaluation revealed the existence of around 470 blogging platforms in addition to the dominating WordPress and Blogger. There are also a large number of established and widely used technologies and standards, e.g. RSS, Atom feeds, CSS, and JavaScript. However, the adoption of metadata standards like Dublin Core[22], Open Graph[23], Friend of a Friend[24] (FOAF), and Semantically Interlinked Online Communities[25] (SIOC) varies significantly [BSJ+12, ADSK+11].

Another survey, aiming on the identification of specific requirements for a blog archive, comprised 26 semi-structured interviews with representatives of different stakeholder groups. The stakeholder groups included blog authors, blog readers, libraries, businesses, blog provider, and researchers. Through a qualitative analysis of the interviews, 114 requirements were identified in the categories functional, data, interoperability, user interface, performance, legal, security, and operational requirements, and modelled with the unified modelling language (UML). While several of the requirements were specifically for blogs (e.g. comments to a blog may be archived even if they appear outside the blog, for example in Facebook), various requirements can be applied on web archives in general [KKL+11].

3.2 The BlogForever data model

While it seems that it is almost impossible to give an exclusive definition for the nature of blogs [Gar11, Lom09], it is necessary for preservation activities to identify blogs' properties [SGK+12]. This is even more crucial for the BlogForever platform which aims on sophisticated access capabilities for the archived blogosphere. Therefore, the different appearances of blogs were examined, and a comprehensive data model was created.

[21] http://www.mementoweb.org/
[22] http://dublincore.org/
[23] http://ogp.me/
[24] http://www.foaf-project.org/
[25] http://sioc-project.org/

The development of the data model was based on existing conceptual models of blogs, data models of open source blogging systems, an empirical study of web feeds, and the online survey with blogger and blog reader perceptions. Thus, it was possible to identify various entities like [SJC+11]:

- Core blog elements, e.g. blog, post, comments,
- Embedded content, e.g. images, audio, video,
- Links, e.g. embedded links, blogroll, pingback,
- Layout, e.g. css, images,
- Feeds, e.g. RSS, Atom, and
- User profiles and affiliations.

The full model comprises over forty single entities and each entity is subsequently described by several properties, e.g. title, URI, aliases, etc. Figure 1 shows, therefore, the high level view of the blog core. The directions of the relationships between the primary identified entities of a weblog are indicated by small triangles [SJC+11].

Beside the inherent blog properties, additional metadata about archiving and preservation activities are captured, stored, and managed. For example, information regarding the time of harvesting of a blog or the legal rights of the content have to be documented as well. Furthermore, additional data may emerge as well as annotations from the archive users, like tags or comments.

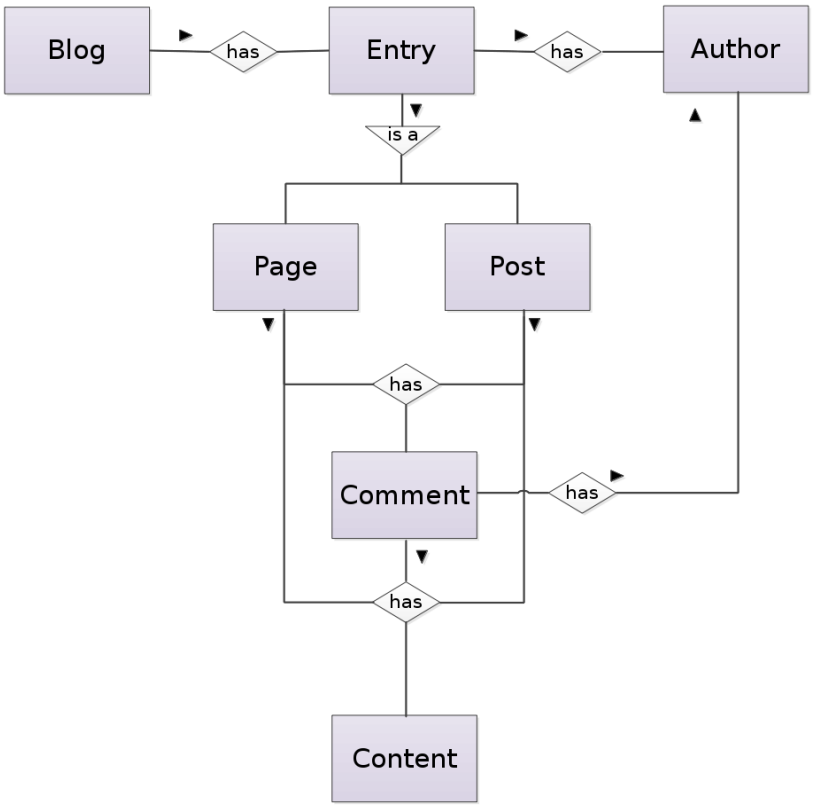


Figure 1: Core of the generic blog data model [SJC+11, p. 45]

3.3 BlogForever platform components

The BlogForever platform consists of the spider component and the repository component. The segmentation into two distinct parts with a well-defined communication interface between them makes the platform more flexible because the components can be developed separately or even replaced if necessary.

The spider component is responsible for harvesting the blogs. It comprises of several subcomponents as shown in figure 2. The Inputter is the starting point, where the list of blogs that should be monitored is maintained. The list should be manually defined instead of using ping servers in order to enable the harvesting of qualified blogs and avoid spam blogs (also known as splogs). All blog URLs collected by the Inputter have to pass through the Host Analyzer, which approves them or blacklists them as incorrect or inappropriate for harvesting. Therefore, it parses each blog URL, collects information about the blog host and discovers the feeds that the blog may provide. The System Manager consists of the source database and the scheduler. While the source database stores all monitored blogs, including various metadata like filtering rules and extraction patterns, the scheduler determines when the blogs are checked for updates. The Worker is responsible for the actual harvesting and analysing of the blog content. Therefore, it fetches the feeds of the blogs as well as HTML content. Both are analysed in order to identify distinct blog elements. Further parsing enables the creation of an XML representation of the identified information and entities, and the identification and harvesting of embedded materials.

Finally, the Exporter delivers the extracted information together with the original content and embedded objects to the repository component [RBS+11].

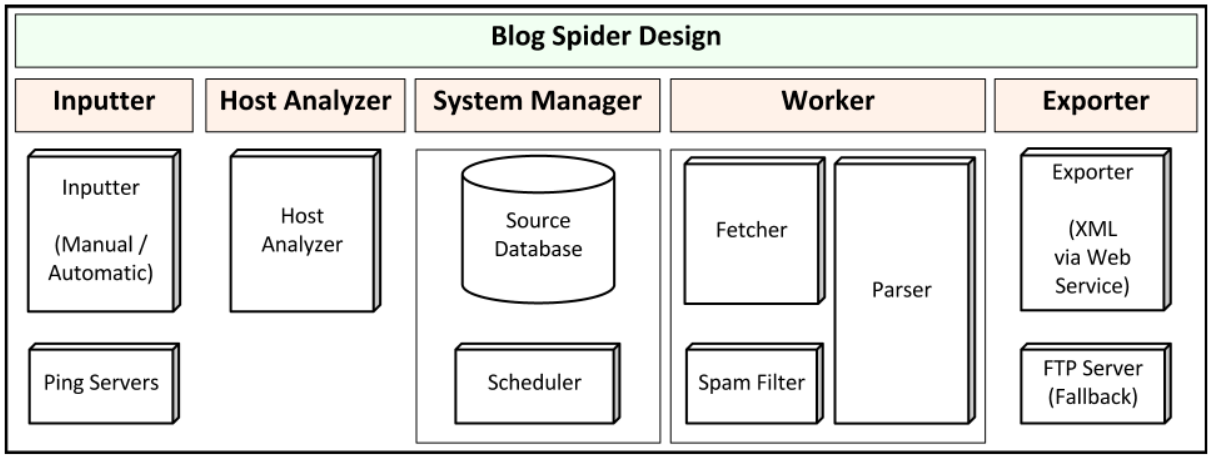


Figure 2: BlogForever spider component design [RBS+11]

The repository component represents the actual preservation platform. It facilitates the ingest, management, and dissemination of the harvested blog content and the extracted information. The repository component is based on the open source software suite Invenio[26], and the subcomponents are shown in figure 3.

New blogs for archiving are announced through the Submission as single blogs or bulk submissions. Thereby, a topic and a license can be indicated. The repository component informs in turn the spider about changes in the list of blogs to monitor. The submission of new blogs in the repository component enables the management of the archived blog selection through one point. The Ingest receives and processes the packages that spider component delivers. It conducts validity checks before the information is transferred to the internal storage. The Storage consists of databases and a file system. It manages the archived data and is responsible for the replication, incremental backup, and versioning. The latter is necessary to keep every version of an entity, e.g. a post, even if the entity has been updated. The Core Services comprise indexing, ranking, digital rights management (DRM), and interoperability. Indexing is performed to enable high speed searching on the archived content. Additionally, the search results can be sorted or ranked, e.g. according to their similarity. The DRM facilitates the access control on the repository's resources. Interoperability is a crucial aspect to facilitate a broader dissemination and integration

[26] http://invenio-software.org

into other services. Therefore, the repository component supports beside others the protocols of the Open Archive Initiative[27] (OAI), the OpenURL format, the Search/Retrieval via URL[28] (SRU), and Digital Object Identifiers (DOI). Finally, the User Services provide the functionalities of searching, exporting, personalising, and collaborating to the archive users. Searching can be performed through a search phrase in a single text field but also more enhanced search strategies are possible through the focussing on specific metadata (e.g. title, author) and the use of regular expressions. The retrieved metadata can be exported in several formats (e.g. Dublin Core[29], MODS[30]) for further processing. Additionally, users can create personal collections and configure notifications that keep them informed about changes in their collection. Collections can also be shared with other users. The possibility to comment and rate any repository content facilitates further collaboration.

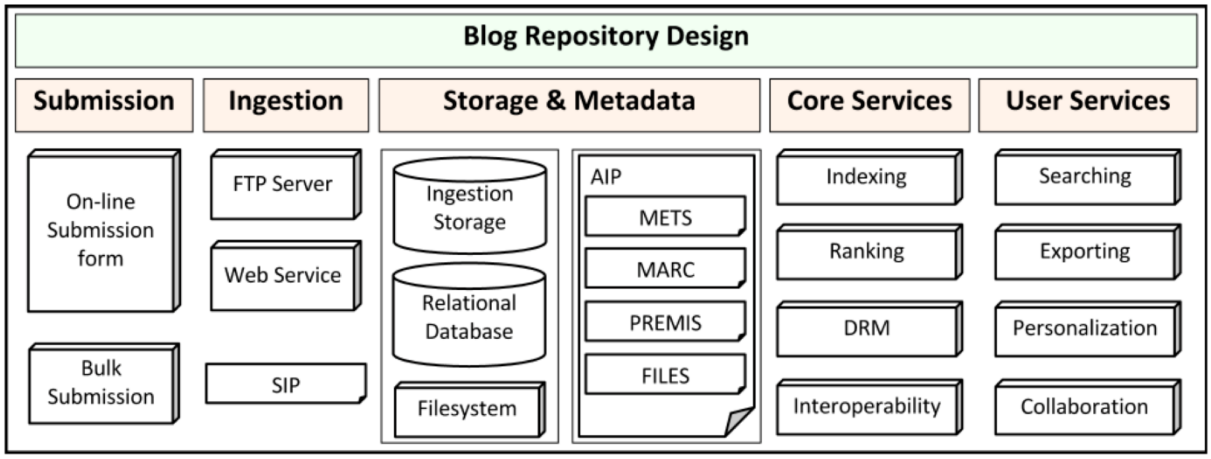


Figure 3: BlogForever repository component design

4 Conclusion

In this paper, we introduced the BlogForever project and blog archiving as a special kind of web archiving. Additionally, we gave an overview about related projects that constitute and extend the capabilities of web archiving. While there are certainly several other aspects to present about the BlogForever project and its findings, we focused on an overview of the empirical work, the presentation of the foundational data model, and the architecture of the BlogForever platform. The software will be available as open source at the end of the project and can be adopted especially by memory institutions (libraries, archives, museums, clearinghouses, electronic databases and data archives), researchers and universities, as well as communities of bloggers. Furthermore, guidelines and recommendations for blog preservation will be provided but could not be introduced in this paper. Two institutions plan already to adopt the BlogForever platform. The European Organization for Nuclear Research (CERN) is going to create a physics blogs archive to maintain blogs related to their research. The Aristotle University of Thessaloniki is going to create an institutional blog archive to preserve university blogs.

The approach of the BlogForever platform is dedicated but not limited to blog archiving. News sites or event calendars have often the same structure and characteristics as blogs (e.g. The Huffington Post). Thus, they could be also archived with BlogForever. However, it should be also emphasized that blogs are just one type of Web content and social media. Other types may cause different challenges but create also additional opportunities for exploitation. Therefore, additional research should be conducted in the future to further improve, specialise and support the current status of web archiving.

[27] http://www.openarchives.org/
[28] http://www.loc.gov/standards/sru/
[29] http://dublincore.org/
[30] http://www.loc.gov/standards/mods/

5 Acknowledgments

This work was conducted as part of the BlogForever[31] project co-funded by the European Commission Framework Programme 7 (FP7), grant agreement No.269963.

6 References

[AAS+ 11] Scott G Ainsworth, Ahmed Alsum, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. How much of the web is archived? In Proceeding of the 11th annual international ACM/IEEE joint conference, page 133, New York, New York, USA, 2011. ACM Press.

[ABBS12] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index maintenance for time-travel text search. In the 35th international ACM SIGIR conference, pages 235–243, New York, New York, USA, 2012. ACM Press.

[ADSK+ 11] Silvia Arango-Docio, Patricia Sleeman, Hendrik Kalb, Karen Stepanyan, Mike Joy, and Vangelis Banos. BlogForever: D2.1 Survey Implementation Report. Technical report, BlogForever Grant agreement no.: 269963, 2011.

[AHA07] Noor Ali-Hasan and Lada A Adamic. Expressing Social Relationships on the Blog through Links and Comments. Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2007.

[BB13] Klaus Berberich and Srikanta Bedathur. Computing n-Gram Statistics in MapReduce. In 16th International Conference on Extending Database Technology (EDBT '13), Genoa, Italy, 2013.

[BBAW10] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. A Language Modeling Approach for Temporal Information Needs. In 32nd European Conference on IR Research (ECIR 2010), pages 13–25, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[BDP+ 12] Christoph Becker, Kresimir Duretec, Petar Petrov, Luis Faria, Miguel Ferreira, and Jose Carlos Ramalho. Preservation Watch: What to monitor and how. In Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012), pages 215–222, Toronto, 2012.

[BSJ+ 12] Vangelis Banos, Karen Stepanyan, Mike Joy, Alexandra I. Cristea, and Yannis Manolopoulos. Technological foundations of the current Blogosphere. In International Conference on Web Intelligence, Mining and Semantics (WIMS) 2012, Craiova, Romania, 2012.

[Che10] Xiaotian Chen. Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs. Internet Reference Services Quarterly, 15(1):21–33, February 2010.

[Chi10] Tara Chittenden. Digital dressing up: modelling female teen identity in the discursive spaces of the fashion blogosphere. Journal of Youth Studies, 13(4):505–520, August 2010.

[CLM11] Esther Conway, Simon Lambert, and Brian Matthews. Managing Preservation Networks. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011), Singapore, 2011.

[Col05] Stephen Coleman. Blogs and the New Politics of Listening. The Political Quarterly, 76(2):272–280, 4/2005.

[Coo13] Robert Cookson. British Library set to harvest the web, 2013.

[DMSW11] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. The SHARC framework for data quality in Web archiving. The VLDB Journal, 20(2):183–207, March 2011.

[DTK11] Katerina Doka, Dimitrios Tsoumakos, and Nectarios Koziris. KANIS: Preserving k- Anonymity Over Distributed Data. In Proceedings of the 5th International Workshop on Personalized Access, Profile Management and Context Awareness in Databases (PersDB 2011), Seattle, 2011.

[EB11] Miklós Erdélyi and András A Benczúr. Temporal Analysis for Web Spam Detection: An Overview. In TWAW 2011, Hyderabad, India, 2011.

[EGB11] Miklás Erdélyi, András Garzó, and András A Benczúr. Web spam classification. In the 2011 Joint WICOW/AIRWeb Workshop, pages 27–34, New York, New York, USA, 2011. ACM Press.

[Ent04] Richard Entlich. Blog Today, Gone Tomorrow? Preservation of Weblogs. RLG DigiNews, 8(4), 2004.

[Gar11] M Garden. Defining blog: A fool's errand or a necessary undertaking. Journalism, September 2011.

[GMC11] Daniel Gomes, Joao Miranda, and Miguel Costa. A Survey on Web Archiving Initiatives. In Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt, editors, Research and Advanced Technology for Digital Libraries, volume 6966 of Lecture Notes in Computer Science, pages 408–420. Springer Berlin / Heidelberg, 2011.

[31] http://blogforever.eu/

[HMS12] Reinhold Huber-Mörk and Alexander Schindler. Quality Assurance for Document Image Collections in Digital Preservation. In Proceedings of the 14th International Conference on Advanced Concepts for Intelligent Vision Systems, pages 108–119, Brno, Czech Republic, 2012. Springer.

[HY11] Helen Hockx-Yu. The Past Issue of the Web. In Proceedings of the ACM WebSci'11, Koblenz, Germany, 2011.

[Ish08] Tom Isherwood. A new direction or more of the same? Political blogging in Egypt. Arab Media & Society, September 2008.

[JN12] Bolette Ammitzbøll Jurik and Jesper Sindahl Nielsen. Audio Quality Assurance: An Application of Cross Correlation. In Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012), pages 144–149, Toronto, 2012.

[KKL+ 11] Hendrik Kalb, Nikolaos Kasioumis, Jaime Garcia Llopis, Senan Postaci, and Silvia Arango-Docio. BlogForever: D4.1 User Requirements and Platform Specifications. Technical report, BlogForever Grant agreement no.: 269963, 2011.

[KSBS12] Ross King, Rainer Schmidt, Christoph Becker, and Sven Schlarb. SCAPE: Big Data Meets Digital Preservation. ERCIM NEWS, 89:30–31, 2012.

[KT12] Hendrik Kalb and Matthias Trier. THE BLOGOSPHERE AS ŒUVRE: INDIVIDUAL AND COLLECTIVE INFLUENCES ON BLOGGERS. In ECIS 2012 Proceedings, page Paper 110, 2012.

[Lom09] Stine Lomborg. Navigating the blogosphere: Towards a genre-based typology of weblogs. First Monday, 14(5), May 2009.

[MAC11] Silviu Maniu, Talel Abdessalem, and Bogdan Cautis. Casting a Web of Trust over Wikipedia: an Interaction-based Approach. In Proceedings of the 20th International Conference on World Wide Web (WWW 2011, Hyderabad, India, 2011.

[Mas06] Julien Masanes. Web Archiving. Springer-Verlag, Berlin, Heidelberg, 2006.

[MCA11] Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. Building a signed network from interactions in Wikipedia. In Databases and Social Networks (DBSocial '11), pages 19–24, Athens, Greece, 2011. ACM Press.

[MDSW10] Arturas Mazeika, Dimitar Denev, Marc Spaniol, and Gerhard Weikum. The SOLAR System for Sharp Web Archiving. In 10th International Web Archiving Workshop, pages 24–30, Vienna, Austria, 2010.

[MF11] Diana Maynard and Adam Funk. Automatic Detection of Political Opinions in Tweets. In Proceedings of MSM 2011: Making Sense of Microposts. Workshop at 8th Extended Semantic Web Conference (ESWC 2011), pages 88–99, Heraklion, Greece, 2011. Springer Berlin Heidelberg.

[O'S05] Catherine O'Sullivan. Diaries, Online Diaries, and the Future Loss to Archives; or, Blogs and the Blogging Bloggers Who Blog Them. The American Archivist, 68(1):53–73, 2005.

[OS10] Marilena Oita and Pierre Senellart. Archiving Data Objects using Web Feeds. In 10th International Web Archiving Workshop, pages 31–41, Vienna, Austria, 2010.

[PAB13] Bibek Paudel, Avishek Anand, and Klaus Berberich. User-Defined Redundancy in Web Archives. In Large-Scale and Distributed Systems for Information Retrieval (LSDS-IR '13), Rome, Italy, 2013.

[PD09] Maureen Pennock and Richard M. Davis. ArchivePress: A Really Simple Solution to Archiving Blog Content. In Sixth International Conference on Preservation of Digital Objects (iPRES 2009), pages 148–154, San Francisco, USA, 2009. California Digital Library.

[PINF11] George Papadakis, Ekaterini Ioannou, Claudia Niederee, and Peter Fankhauser. Efficient entity resolution for large heterogeneous information spaces. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), pages 535–544, New York, New York, USA, 2011. ACM Press.

[PKTK12] Nikolaos Papailiou, Ioannis Konstantinou, Dimitrios Tsoumakos, and Nectarios Koziris. H2RDF: Adaptive Query Processing on RDF Data in the Cloud. In Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion), pages 397–400, New York, New York, USA, 2012. ACM Press.

[PVM10] Radu Pop, Gabriel Vasile, and Julien Masanes. Archiving Web Video. In 10th International Web Archiving Workshop, pages 42–47, Vienna, Austria, 2010.

[Rad08] Courtney Radsch. Core to Commonplace: The evolution of Egypt's blogosphere. Arab Media & Society, September 2008.

[RBS+ 11] M. Rynning, V. Banos, K. Stepanyan, M. Joy, and M. Gulliksen. BlogForever: D2. 4 Weblog spider prototype and associated methodology. Technical report, 2011.

[RDM+ 11] Thomas Risse, Stefan Dietze, Diana Maynard, Nina Tahmasebi, and Wim Peters. Using Events for Content Appraisal and Selection in Web Archives. In Proc. of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), Bonn, Germany, 2011.

[RP12] Thomas Risse and Wim Peters. ARCOMEM: from collect-all ARchives to COmmunity MEMories. In 21st international conference companion on World Wide Web (WWW '12 Companion, pages 275–278, New York, New York, USA, 2012. ACM Press.

[SBVW12] Marc Spaniol, Andras A Benczur, Zsolt Viharos, and Gerhard Weikum. Big Web Analytics: Toward a Virtual Web Observatory. ERCIM NEWS, 89:23–24, 2012.

[Sch12] Rainer Schmidt. SCAPE — An Architectural Overview of the SCAPE Preservation Platform. In Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES 2012), pages 85–88, Toronto, 2012.

[SGK+ 12] Karen Stepanyan, George Gkotsis, Hendrik Kalb, Yunhyong Kim, Alexandra I. Cristea, Mike Joy, Matthias Trier, and Seamus Ross. Blogs as Objects of Preservation: Advancing the Discussion on Significant Properties. In iPres 2012, pages 218–224, Toronto, Canada, 2012.

[SJC+ 11] Karen Stepanyan, Mike Joy, Alexandra Cristea, Yunhyong Kim, Ed Pinsent, and Stella Kopidaki. Blogforever: Weblog Data Model. Technical report, 2011.

[SLY+ 12] Arif Shaon, Simon Lambert, Erica Yang, Catherine Jones, Brian Matthews, and Tom Griffin. Towards a Scalable Long-term Preservation Repository for Scientific Research Datasets. In The 7th International Conference on Open Repositories (OR2012), Edinburgh, UK, 2012.

[SPNT13] George Sfakianakis, Ioannis Patlakas, Nikos Ntarmos, and Peter Triantafillou. Interval Indexing and Querying on Key-Value Cloud Stores. In 29th IEEE International Conference on Data Engineering, Brisbane, Australia, 2013.

[SW12] Marc Spaniol and Gerhard Weikum. Tracking entities in web archives. In 21st international conference companion on World Wide Web (WWW '12 Companion, pages 287–290, New York, USA, 2012. ACM Press.

[Tia13] Q Tian. Social Anxiety, Motivation, Self-Disclosure, and Computer-Mediated Friendship: A Path Analysis of the Social Interaction in the Blogosphere. Communication Research, 40(2):237–260, February 2013.

[TNTR10] Nina Tahmasebi, Kai Niklas, Thomas Theuerkauf, and Thomas Risse. Using word sense discrimination on historic document collections. In Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10), pages 89–98, New York, New York, USA, 2010. ACM Press.

[TRD11] Nina Tahmasebi, Thomas Risse, and Stefan Dietze. Towards automatic language evolution tracking A study on word sense tracking. In Proc. of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn 2011), Bonn, Germany, 2011.

[TZ09] Vicente Torres-Zuniga. Blogs as an effective tool to teach and popularize physics: a case study. Latin-American Journal of Physics Education, 3(2):4, 2009.

[TZIR10] Nina Tahmasebi, Gideon Zenz, Tereza Iofciu, and Thomas Risse. Terminology Evolution Module for Web Archives in the LiWA Context. In 10th International Web Archiving Workshop, pages 55–62, Vienna, Austria, 2010.

[VdSNS+09] Herbert Van de Sompel, Michael L Nelson, Robert Sanderson, Lyudmila L Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time Travel for the Web. Technical report, November 2009.

[VdSNS13] Herbert Van de Sompel, Michael L Nelson, and Robert Sanderson. HTTP framework for time-based access to resource states, 2013.

[WDSW12] Yafang Wang, Maximilian Dylla, Marc Spaniol, and Gerhard Weikum. Coupling label propagation and constraints for temporal fact extraction. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL '12), pages 233-–237. Association for Computational Linguistics, July 2012.

[WJM10] Xiaoguang Wang, Tingting Jiang, and Feicheng Ma. Blog-supported scientific communication: An exploratory analysis based on social hyperlinks in a Chinese blog community. Journal of Information Science, 36(6):690–704, December 2010.

[WNS+ 11] Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, András A Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. Longitudinal Analytics on Web Archive Data: It's About Time! In 5th Biennial Conference on Innovative Data Systems Research (CIDR '11), pages 199–202, Asilomar, California, USA, 2011.

[YBE+ 12] Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. Natural language questions for the web of data. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12), pages 379–390, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[YWX+ 13] Qiang Yan, Xingxing Wang, Qiang Xu, Dongying Kong, Danny Bickson, Quan Yuan, and Qing Yang. Predicting Search Engine Switching in WSCD 2013 Challenge. In Workshop on Web Search Click Data 2013 (WSCD2013), 2013.

Vertrauenswürdige und beweiswerterhaltende elektronische Langzeitspeicherung auf Basis von DIN 31647 und BSI-TR-03125

Steffen Schwalm[1] · Ulrike Korte[2] · Detlef Hühnlein[3]

[1] BearingPoint GmbH, Kurfürstendamm 207-208, 10719 Berlin, steffen.schwalm@bearingpoint.com
[2] Bundesamt für Sicherheit in der Informationstechnik, Godesberger Allee 185-189, 53175 Bonn, ulrike.korte@bsi.bund.de
[3] ecsec GmbH, Sudetenstraße 16, 96247 Michelau, detlef.huehnlein@ecsec.de

Abstract: Es besteht eine hohe Notwendigkeit, nicht nur in der öffentlichen Verwaltung, sondern auch in Unternehmen, Geschäftsprozesse zu digitalisieren und für die elektronischen Dokumente und Daten auch in ferner Zukunft die Lesbarkeit, Verfügbarkeit sowie die Integrität, Authentizität und Verkehrsfähigkeit gewährleisten zu müssen. Besondere Herausforderungen existieren in diesem Umfeld beim dauerhaften Erhalt der Beweiskraft der elektronisch signierten Dokumente. Vor diesem Hintergrund entwickelt der DIN-Arbeitskreis NA 009-00-15-06 AK „Arbeitskreis Beweiswerterhaltung kryptographisch signierter Dokumente“ den DIN-Standard 31647, der auf der Technischen Richtlinie TR 03215 „Beweiswerterhaltung kryptographisch signierter Dokumente“ des Bundesamtes für Sicherheit in der Informationstechnik (BSI) aufsetzt. Dieser Beitrag stellt die wesentlichen Inhalte und das mögliche Zusammenspiel des DIN-Standards und der BSI-TR-03125 (TR-ESOR) vor.

1 Einleitung

Die Nutzung der Informationstechnologie für Abwicklungen von Geschäftsprozessen ist allgemein etabliert. Geschäftsrelevante Unterlagen liegen zunehmend ausschließlich elektronisch vor. Elektronische Dokumente können jedoch aus sich heraus weder wahrgenommen noch gelesen werden. Sie liefern aus sich heraus auch keine Hinweise für ihre Integrität und Authentizität sowie die Ordnungsmäßigkeit im elektronischen Rechts- und Geschäftsverkehr. Gleichzeitig bestehen jedoch umfassende Dokumentations- und Aufbewahrungspflichten, deren Dauer zwischen zwei und 110 Jahre oder dauernd umfasst, die einen langfristigen Nachweis von Authentizität, Integrität und Nachvollziehbarkeit elektronischer Unterlagen erfordern. Während dieser Fristen muss es zudem möglich sein, die Dokumente Prüfbehörden oder Gerichten vorzulegen und anhand der Daten die genannten Nachweise zu führen. Dies erfordert eine langfristige Verkehrsfähigkeit der Unterlagen. Die Nutzung kryptographischer Mittel, wie fortgeschrittene oder qualifizierte elektronischer Signaturen und qualifizierte Zeitstempel, ermöglicht nach geltendem Recht die Erhaltung des für die Nachweisführung notwendigen Beweiswerts, ohne die Verkehrsfähigkeit einzuschränken (siehe [F 06], [Ro07], [BMWi 07]).

Besondere Herausforderungen existieren in diesem Umfeld bei der Beweiswerterhaltung der elektronisch signierten Dokumente, da die Sicherheitseignungen der eingesetzten kryptographischen Algorithmen mit der Zeit abnehmen können, so dass im Rahmen der langfristigen Aufbewahrungsdauer signierter Dokumente zusätzliche Maßnahmen für den Erhalt der Beweiskraft notwendig sind. Die BSI-TR-03125, die auch Eingang in § 6 EGovG (Elektronische Aktenführung) gefunden hat, wurde auf der Grundlage bestehender rechtlicher Normen sowie nationaler und internationaler technischer Standards entwickelt und liefert eine modular aufgebaute, logische Gesamtkonzeption und technische Spezifikation für die beweiswerterhaltende Langzeitspeicherung kryptographisch signierter Daten und Dokumente im Rahmen der gesetzlichen Aufbewahrungsfristen. Die DIN-Norm 31647 beschreibt, basierend u.a. auf der Technischen Richtlinie BSI-TR-03215 „Beweiswerterhaltung kryptographisch signierter Dokumente“ des Bundesamts für Sicherheit in der Informationstechnik (BSI), dem OAIS-Modell sowie der DIN 31644, grundsätzliche fachliche und funktionale Anforderungen an ein generisches System zur Beweiswerterhaltung kryptographisch signierter Dokumente unter Wahrung der Authentizität, Integrität, Verlässlichkeit, Verkehrs- und Migrationsfähigkeit der Dokumente mindestens für die Dauer der geltenden Aufbewahrungsfristen.

Der vorliegende Beitrag stellt den aktuellen Arbeitsstand hinsichtlich der Entwicklung der DIN 31647 und das Zusammenspiel mit der existierenden BSI-TR-03125 vor und ist folgendermaßen gegliedert: Abschnitt 2 erläutert die grundsätzlichen Anforderungen an die Aufbewahrung elektronischer Unterlagen. Der Abschnitt 3 enthält einen Überblick über die DIN-Norm 31647. Abschnitt 4 greift ausgewählte Aspekte der BSI-TR-03125 (TR-ESOR) auf. In Abschnitt 5 werden die wesentlichen Ergebnisse des Beitrags kurz zusammengefasst und um einen Ausblick auf zukünftige Entwicklungen ergänzt.

2 Grundsätzliche Anforderungen an die Aufbewahrung elektronischer Unterlagen

Um insbesondere im Bereich der öffentlichen Behörden eine Abgrenzung zum normativ definierten Begriff der Archivierung, die die dauerhafte Aufbewahrung archivwürdiger Unterlagen im zuständigen öffentlichen Archiv umfasst, hat sich für die Aufbewahrung im Rahmen geltender Aufbewahrungsfristen der Terminus „Langzeitspeicherung“ gemäß ([BMI 12], [BSI-TR-03125], [BarchG]) bzw. gemäß äquivalenten Gesetzen der Länder und Archivsatzungen der Kommunen etabliert und wird dementsprechend auch im Text verwendet.

Während dieser Aufbewahrungsfristen muss die Authentizität, Integrität und Nachvollziehbarkeit (siehe auch [BSI-TR-RESISCAN], Tabelle 6) gegenüber Prüfbehörden, Gerichten, etc. nachgewiesen werden können, was neben der Erhaltung der Unterlagen selbst vor allem die Erhaltung des Beweiswerts dieser Unterlagen erfordert (siehe [F 06], [Ro07]). Durch die Verwendung geeigneter kryptographischer Mittel, wie sie bei der qualifizierten elektronischen Signatur und bei qualifizierten Zeitstempeln zum Einsatz kommen, kann ein hoher Beweiswert erzielt und langfristig erhalten werden.

So ermöglichen fortgeschrittene oder qualifizierte elektronische Signaturen und qualifizierte Zeitstempel nach geltendem Recht die zur eindeutigen Nachweisführung notwendige Beweiswerterhaltung direkt am eigentlichen Dokument, da Signaturen und Zeitstempel direkt am Dokument oder der digitalen Akte bzw. dem Vorgang, in der sich das Dokument befindet, angebracht werden.

Der Beweiswert ist also eine inhärente Eigenschaft der jeweiligen elektronischen Unterlagen. Dementsprechend müssen Maßnahmen zur Beweiswerterhaltung auch direkt an den elektronischen Unterlagen ansetzen. Dies bedingt quasi die Langzeitspeicherung selbsttragender Archivpakete im Sinne geltender Standards und Normen (z.B. OAIS-Modell, [BSI-TR-03125] Kap.3.1).Dabei ergibt sich der folgende Lebenszyklus elektronischer Unterlagen.

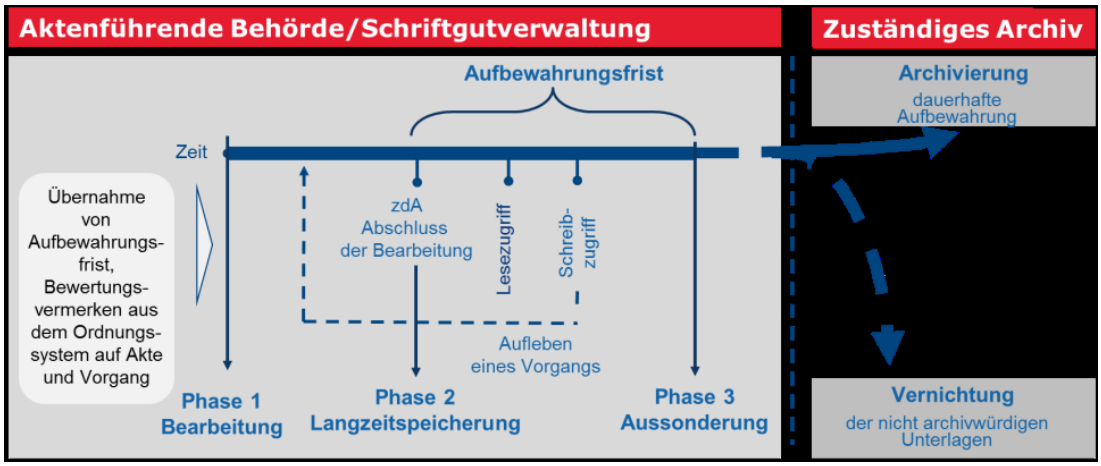


Abbildung 1: Lebenszyklus elektronischer Unterlagen

Da die Sicherheitseignung der den kryptographischen Mitteln zugrundeliegenden Algorithmen im Kontext der technischen Entwicklung abnehmen kann, ist es insbesondere bei qualifizierten elektronischen Signaturen notwendig, diese vor Ablauf der Sicherheitseignung zu erneuern. Dies erfolgt durch eine Nachsignatur (§ 17 SigV), also der Anbringung einer neuen qualifizierten elektronischen Signatur sowie eines qualifizierten Zeitstempels. Hierbei genügt die Erstellung von qualifizierten Zeitstempeln, sofern diese mittels einer qualifizierten elektronischen Signatur erzeugt wurden. Außerdem kann ein solcher Zeitstempel mehrere Dokumente, ihre Hashwerte oder einen aus solchen Hashwerten gebildeten Merkle-Hashbaum gemäß RFC 4998 bzw. RFC 6283 umfassen, was eine sehr wirtschaftliche Nachsignatur einer Vielzahl von Dokumenten ermöglicht[1]. Die Nachsignatur muss dabei jeweils alle vorhergehenden Signaturen und Zeitstempel einschließen. Sofern auch die Sicherheitseignung der der Signatur zugrundeliegenden Hashalgorithmen ausläuft, sind zunächst neue Hashwerte mit einem geeigneten Algorithmus zu berechnen, bevor die Nachsignatur unter Verwendung qualifiziert signierter qualifizierter Zeitstempel erfolgt.

Hinzu kommt eine sichere Speicherung und Datenhaltung, um den Anforderungen hinsichtlich Datensicherheit und Datenschutz gerecht zu werden. Dabei ist es z.B. für die öffentliche Verwaltung nicht ausreichend, einzelne

[1] Umgekehrt würde bei Einsatz der von ETSI für die langfristige Archivierung von Signaturen standardisierten {C,X,P}AdES-A – Formate für jede zu konservierende Signatur ein eigener Zeitstempel benötigt werden.

Dokumente oder Daten aufzubewahren. Vielmehr muss die Langzeitspeicherung den Entstehungskontext bzw. den Aktenzusammenhang wahren. Es gilt, Verwaltungsentscheidungen für die gesamte Dauer der Aufbewahrungsfristen nachvollziehbar und beweissicher zu halten. Nur so kann der bestehende Beweiswert erhalten und Kosten für die aufwändige Rekonstruktion der Unterlagen vermieden werden. Der Beweiswert und damit die Behandlung elektronischer Unterlagen vor Gericht wird in §§ 371a ff. ZPO geregelt. Diese Regelungen gelten gem. § 98 VwGO auch für die öffentliche Verwaltung. Für die öffentliche Verwaltung ist darüber hinaus zu beachten, dass der Beweis anhand von Akten und in der Folge den Dokumenten geführt wird (§ 99 VwGO). Hierfür ist es also notwendig, erst einmal Akten zu bilden und im Aktenzusammenhang aufzubewahren.

Diese Anforderungen und Rahmenbedingungen gelten für alle elektronischen Unterlagen, unabhängig davon, in welchen Verfahren oder Ablage diese gehalten werden. Um die geltenden Anforderungen an die Aufbewahrung elektronischer Unterlagen zu erfüllen, gilt es, aktuelle nationale sowie internationale Standards und Normen zu berücksichtigen. Hierzu zählt insbesondere das in ISO 14721 definierte Open Archival Information System (OAIS) - Modell als zentrale Norm, die auch eine inhaltliche Basis der DIN 31647 bildet, sowie die in RFC 4998 und RFC 6283 standardisierte Evidence Record Syntax, die im Regelfall Zeitstempeln gemäß RFC 3161 umfasst.

3 Die DIN 31647 (Entwurf)

3.1 Einführung

Das in ISO-14721:2012 genormte OAIS gilt als zentrale Norm zur Langzeitspeicherung und Archivierung elektronischer Unterlagen. Es beschreibt die notwendigen Komponenten (Prozesse) und Informationspakete zur langfristigen oder dauerhaften Aufbewahrung digitaler Daten. Ursprünglich entwickelt wurde es für die Archivierung von Forschungsdaten in der Raumfahrt, aber inzwischen hat es sich weltweit zur Langzeitspeicherung und Archivierung durchgesetzt.

Die nachstehende Grafik zeigt das OAIS-Modell im Überblick:

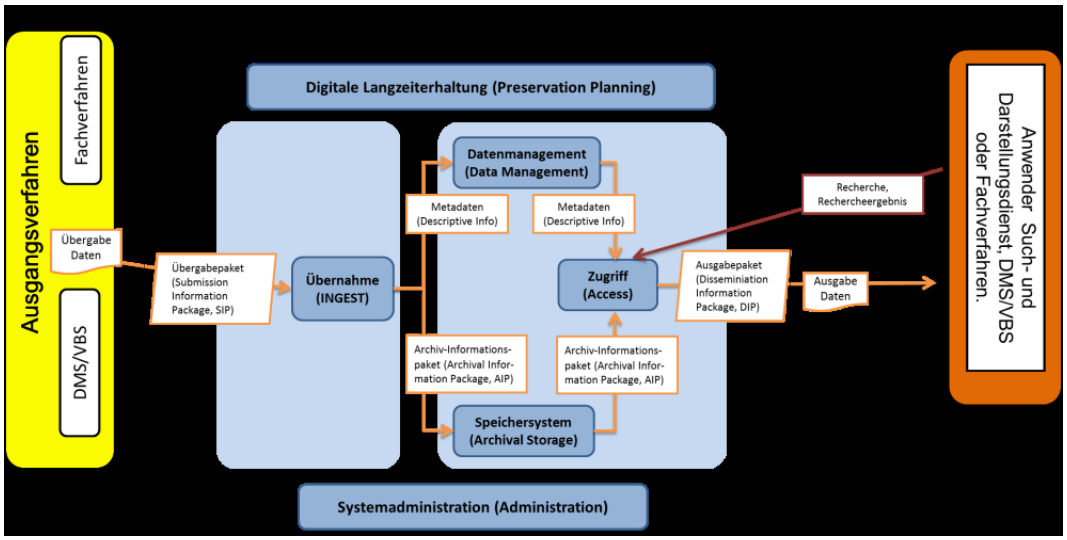


Abbildung 2: Open Archival Information System

Eine OAIS-konforme Langzeitspeicherung ist vollständig *hersteller- und systemneutral.* Das OAIS bezieht sich ausschließlich auf die Aufbewahrung und langfristige Erhaltung der elektronischen Unterlagen selbst, denn Hard- und Software unterliegen regelmäßigen Veränderungen aufgrund der IT- und wirtschaftlichen Entwicklung. Die Aufbewahrungspflichten beziehen sich zudem ausschließlich auf die Unterlagen selbst, nicht auf die Verfahren und die Hardware. Insofern gilt es, die Unterlagen selbst aufzubewahren in Form selbsttragender Archivinformationspakete. Eine solche system- und datenträgerneutrale Langzeitspeicherung umfasst für alle aufzubewahrenden Unterlagen die folgenden Prozesse mit den jeweils angegebenen grundlegenden Eigenschaften.

- Prozesse

- o Ingest (Übernahme),
- o Archival Storage (Speicher/elektronisches Magazin),
- o Daten Management (Metadatenverwaltung),
- o Access (Nutzung),
- o Preservation Planning (digitale Bestandserhaltung),
- o Systemadministration,

- Informationspakete
  - o Submission Information Package – SIP (Übergabepaket): aufzubewahrende Daten aus dem laufenden Verfahren,
  - o Archival Information Package – AIP (Archivinformationspaket): Form, in der die Daten aufbewahrt werden,
  - o Dissemination Information Package – DIP (Ausgabepaket): Form, in der die Daten beim Zugriff abhängig von den Zugriffsrechten ausgegeben werden.

Das OAIS geht dabei grundsätzlich davon aus, dass selbsttragende Informationspakete, die sog. Archivinformationspakete (AIP), im Archivspeicher abgelegt werden. Dies bedeutet, dass die Datenpakete selbst alle Informationen zur Interpretation, Lesbarkeit, Nutzbarkeit, Verständlichkeit, Recherche und zu den beweisrelevanten Nachweisen der Integrität und Authentizität der aufzubewahrenden Unterlagen in standardisierter und herstellerunabhängiger Form (i.d.R. in einem XML-basierten Paket)enthalten. Dies umfasst also:

- Metadaten,
- Inhalts-/Primärdaten sowie die
- zur Sicherung von Authentizität und Integrität notwendigen Daten.

In der Umsetzung wird das OAIS-Modell in Deutschland ergänzt durch folgende Normen:

- DIN 31644 zur Vertrauenswürdigkeit digitaler Langzeitarchive,
- DIN 31645 zur Spezifizierung der Datenübernahme,
- DIN 31646 hinsichtlich Persistent Identifier.

3.2 Normative Einordnung der DIN 31647

Die DIN 31647 formuliert fachliche und funktionale Anforderungen an ein generisches System zur Beweiswerterhaltung kryptographisch signierter Dokumente unter Wahrung der Authentizität, Integrität, Nachvollziehbarkeit, Verfügbarkeit, Verkehrs- und Migrationsfähigkeit der Dokumente mindestens für die Dauer der geltenden Aufbewahrungsfristen.

Die in der Norm 31647 beschriebenen Funktionen zur Beweiswerterhaltung kryptographisch signierter Dokumente stellen kein eigenes System dar. Sie ergänzen vielmehr einen OAIS-konformen Langzeitspeicher, z.B. ein vertrauenswürdiges digitales Langzeitarchiv (dLZA) im Sinne der DIN 31644, um die notwendigen Maßnahmen zur Beweiswerterhaltung kryptographisch signierter Dokumente. Die Norm 31647 folgt dem Prozess- und Informationsmodell des Referenzmodells OAIS und der DIN 31645. und ergänzt die DIN 31645 auf Basis der TR-03125 (TR-ESOR) des Bundesamtes für Sicherheit in der Informationstechnik (BSI) um die zur Beweiswerterhaltung notwendigen Informationen und Funktionen. Das bedeutet, dass ein dLZA, in dem die Beweiswerterhaltung kryptographisch signierter Dokumente erfolgen soll, neben den Anforderungen der DIN 31644 und der DIN 31645 auch denjenigen der DIN 31647 genügen muss. Entsprechend den in Kap. 2 beschriebenen Anforderungen an eine beweissichere Langzeitspeicherung umfasst dies

- Beweiswerterhaltung (DIN 31647, TR-03125)[2],

[2] Beweiswerterhaltung im Sinne der DIN 31647 bedeutet, den Beweiswert der in einem digitalen Langzeitarchiv aufbewahrten elektronischen Informationen über die Dauer des Aufbewahrungszeitraumes zu erhalten und so die mit der Aufbewahrung bezweckten Rechtsfolgen elektronischer Unterlagen mindestens für die Dauer der

- Informationserhaltung (OAIS, DIN 31644, DIN 31645, DIN 31646, etc.)

und erfordert somit die Verbindung der etablierten Standards und Normen zu einer Gesamtlösung, die beide Teile adressiert. Das bedeutet, dass Fragen nach der Informationserhaltung kryptographisch signierter Dokumente keine Bestandteile der DIN 31647 darstellen, sondern für diese Aspekte vielmehr im Rahmen Preservation Planning eigenständig entsprechende Maßnahmen, z.B. unter Aufgreifen des Migrations- oder Emulationsverfahrens, zu treffen sind[3].

Die nachstehende Grafik verdeutlicht die normative Einordnung der DIN 31647 (Entwurf):

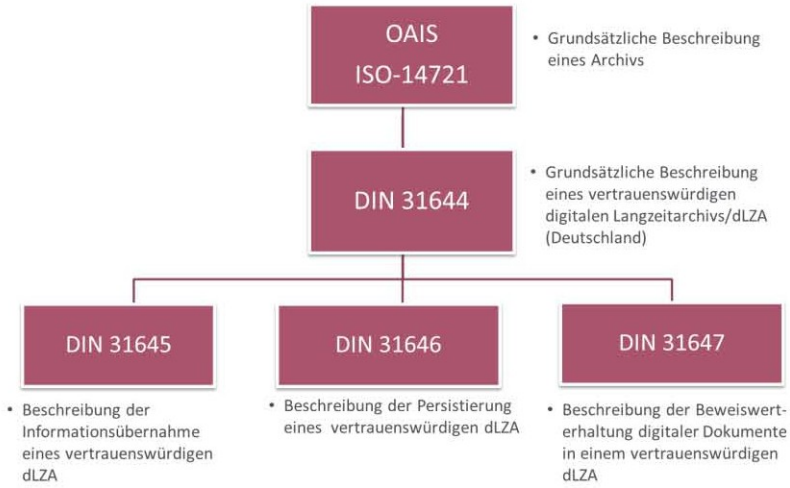


Abbildung 3: Normative Einordnung der DIN 31647 (Entwurf)

Unter kryptographisch signierten Dokumenten werden alle elektronischen Datenobjekte verstanden, deren Integrität und ggf. Authentizität durch die Verwendung kryptographischer Sicherungsmittel (z.B. Signaturen, Zeitstempel) nachgewiesen werden und damit deren Beweiswert erhalten werden soll. Hintergrund ist, dass diese Sicherungsmittel einen langfristigen mathematisch eindeutigen Integritäts- und ggf. Authentizitätsnachweis und die vollständige Beweiswerterhaltung ermöglichen.

Im Ergebnis entsteht so die branchen- und anwendungsfallübergreifende Darstellung eines generischen Systems zur Beweiswerterhaltung kryptographisch signierter Unterlagen anhand standardisierter kryptographischer Funktionen insbesondere auf Basis von Hashwerten und qualifizierten Zeitstempel.

3.3 Konkretisierung des OAIS-Modells zur Beweiswerterhaltung nach DIN 31647

Aufbauend auf einem allgemeingültigen Vokabular konkretisiert die DIN 31647 das Datenmodell des OAIS sowie die Prozesse und Informationspakete.

Im Kontext der Beweiswerterhaltung sind dabei grundsätzlich alle Informationspakete, SIP, AIP und DIP, relevant, da die Beweiswerterhaltung eine Funktion innerhalb eines OAIS-konformen dLZA darstellt.

Innerhalb der Provenance Information (Herkunftsinformation) beschreiben die Information Property Description, welche Eigenschaften der Inhaltsdatenobjekte zu den signifikanten Eigenschaften gehören. Zum Zwecke des Beweiswerterhalts muss der Anwender / Produzent kryptographisch zu schützende Datenobjekte kenntlich machen und ihre zu erhaltenen Eigenschaften durch spezielle Erhaltungsmetadaten, genannt Beweisdatenbeschreibung,

---

vorgeschriebenen Aufbewahrungszeiträume zu gewährleisten. Technisch wird dies durch Beweisdaten und beweisrelevante Daten, also kryptographische Sicherungsmittel und deren Erhaltung realisiert.

[3] Das Projekt TransiDoc hat z.B. eine Lösung zur Nutzung des Migrationsverfahrens als Methode zur Informationserhaltung kryptographisch signierter Dokumente erarbeitet (Vgl. http://www.transidoc.de).

innerhalb der Information Property Description aufführen. Diese Beschreibung muss mindestens die folgenden Informationen enthalten:

- auf welche Datenobjekte sich kryptographische Sicherungsmittel beziehen,
- welche Eigenschaften (Integrität und/oder Authentizität) der Datenobjekte geschützt werden sollen und
- welche Anforderungen an die Prüfung der Sicherungsmittel und ihre Ergänzung um Validierungsdaten gestellt werden.

Die Beweisdatenbeschreibung dokumentiert faktisch, wie Integrität und Authentizität nachgewiesen werden und stellt damit die Basis zur Beweiswerterhaltung kryptographisch signierter Dokumente dar.

Für die Beweiswerterhaltung sind Beweisdaten und beweisrelevante Daten notwendig. Beweisdaten dienen dem Nachweis der Integrität und Authentizität. Ein Beweisdatensatz (sog. Technische Beweisdaten) auf Basis von RFC4998 bzw. RFC6283 enthält u.a. Archivzeitstempel über die gespeicherten Archivdatenobjekte sowie weitere Informationen, die die Richtigkeit und die Gültigkeit elektronischer Signaturen zum Signaturzeitpunkt sowie die rechtzeitige Signaturerneuerung nachweisen. Beweisrelevante Daten sind Signaturen bzw. Zeitstempel zu genau einem Datenobjekt inklusive der für die Prüfung der Signatur bzw.- Zeitstempel notwendigen Prüfdaten. Diese ermöglichen es, die Eigenschaften Integrität und / oder Authentizität von digitalen Objekten nachzuweisen und damit die Basis zur Beweiswerterhaltung kryptographisch signierter Dokumente zu bilden. Insofern ist es zur Beweiswerterhaltung kryptographisch signierter Dokumente notwendig, Beweisdaten und beweisrelevante Daten in den Fixity Information der Erhaltungsmetadaten eines Informationspakets nach OAIS abzulegen. Das Archivinformationspaket muss damit einen vollständigen Satz der Beweisdaten und beweisrelevanten Daten innerhalb der Fixity Information der Erhaltungsmetadaten umfassen.

Sofern kryptographisch signierte Dokument ins dLZA übernommen werden, muss bereits das Übernahmeinformationspaket (SIP) die Signaturen als Beweisdaten in den Erhaltungsmetadaten umfassen. Der Schutz des Beweiswertes kann sich dabei sowohl auf in Transferpaketen (SIP) enthaltene Beweisdaten als auch auf dort ungeschützte Inhaltsdaten beziehen. Für ursprünglich ungeschützte, also nicht kryptographisch signierte Datenobjekte können zum Zeitpunkt der Einlagerung Integritätsnachweise erzeugt und nachweisbar erhalten werden. In Transferpaketen enthaltene kryptographische Authentizitätsnachweise (elektronische Signaturen) sind explizit auszuweisen.

Das dLZA muss darüber hinaus, entsprechend den geltenden Zugriffsrechten, es ermöglichen, im Ausgabeinformationspaket (DIP) nutzerbezogen Beweisdaten und beweisrelevante Daten als Teil der Erhaltungsmetadaten auszugeben.

Innerhalb von SIP und AIP können Beweisdaten und beweisrelevante Daten sowohl als Teil eines digitalen Datenobjekts oder in Form separater Datenobjekte vorliegen. In jedem Fall sind Beweisdaten und beweisrelevante Daten einem digitalen Objekt zugeordnet. Beweisdaten, beweisrelevante Daten und AIP bilden eine logische Einheit und müssen für den gesamten Zeitraum der Aufbewahrung in einem untrennbaren Zusammenhang stehen. Credential Data bilden im Sinne von OAIS einen „protection shield“ [OAIS] für digitale Objekte.

Im Ergebnis beschreiben die Provenance Information den Aussteller eines Dokuments sowie die in den Fixity Information enthaltenen beweisrelevanten Daten und Beweisdaten und verweisen auf diese. Die eigentlichen kryptographischen Daten für den Authentizitäts- und Integritätsnachweis im Kontext der Beweiswerterhaltung werden schlussendlich in den Fixity Information nachgewiesen.

Die nachstehende Grafik verdeutlicht dieses Prinzip:

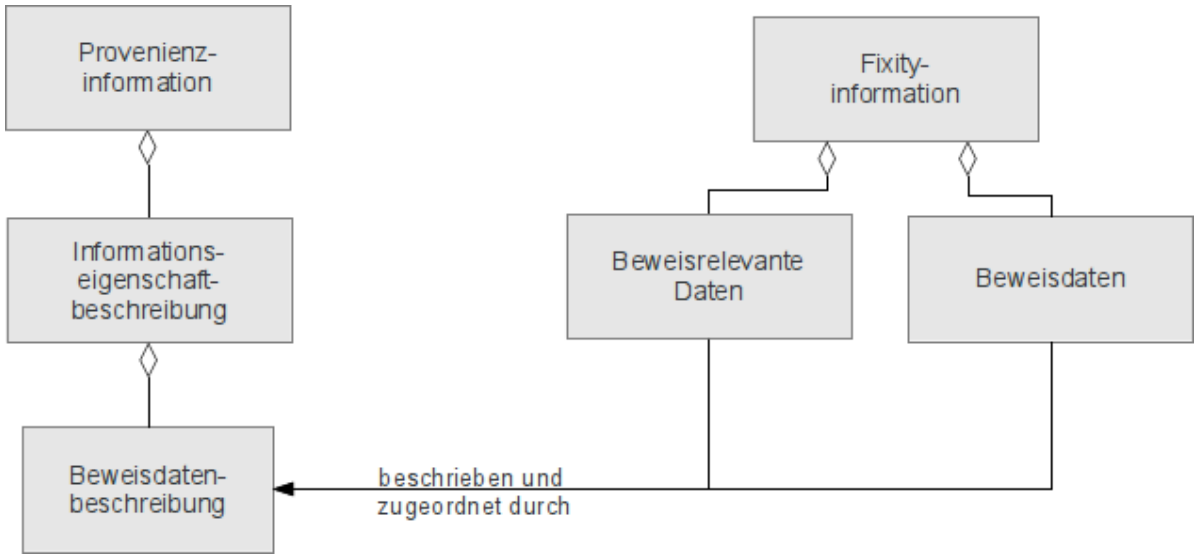


Abbildung 4: Einordnung der Metadaten für den Beweiswerterhalt

Konkretisierungen hinsichtlich der Prozesse gem. OAIS-Modell ergeben sich aus aktueller Sicht vor allem bei:

- INGEST,
  - o Aufnahme beweisrelevanter Daten sowie der technischen Beweisdaten im AIP,

- Data Management,
  - o Vergabe einer eindeutigen AOID je AIP,

- Preservation Planning (Beweiswerterhaltung),
  - o Ergänzung der Informationserhaltung um die Maßnahmen zur Beweiswerterhaltung, also die periodische Nachsignatur sowie die Hashwerterneuerung der kryptographischen Sicherungsmittel, bevor deren Sicherheitseignung abläuft,

- Access,
  - o Abruf der AIP einschl. der beweisrelevanten Daten sowie der technischen Beweisdaten.

Die Konkretisierung bildet eine inhaltliche Basis zur Darstellung der notwendigen Funktionen an ein generisches System zur Beweiswerterhaltung kryptographisch signierter Dokumente. Die Norm soll dabei die grundlegenden Funktionen beschreiben. Die Darstellung basiert auf der TR-03125, ermöglicht in der Umsetzung jedoch explizit auch eine die TR-03125 ergänzende oder über diese hinausgehende Realisierung. Die DIN-Norm ist hier technik- und anwendungsneutral.

3.4 Anforderungen und Funktionen eines generischen Systems zur Beweiswerterhaltung

Ein generisches System zur Beweiswerterhaltung erfüllt z.B. die folgenden Anforderungen:

- Erhaltung und Erweiterung des fachlich-logischen Zusammenhangs von AIP:
  - o Sofern eine Versionierung von AIP möglich ist, darf der Beweiswert der bereits archivierten Dokumente im Ursprungs-AIP nicht verletzt werden, was im Grundsatz durch eine Versionierung des AIP, mit eigener Beweissicherung und eindeutiger Versionsnummer, für die Versionen innerhalb des AIP erreicht wird,

- Hashen von AIP:
  - o Erzeugung kryptographischer, sicherheitsgeeigneter Hashwerte[4],
  - o Vermeidung von Mehrdeutigkeiten bei signierten XML-Daten,
  - o Möglichst frühzeitiges Hashen der AIP zur Integritäts-/Authentizitätssicherung, Aufbau eines Hashbaums (RFC 4998 bzw. RFC 6283) und Versiegelung des Baums mit einem Archivzeitstempel (RFC 3161 und RFC 5652),
  - o Sortierung, Konkatenation und Kanonisierung von Daten,
  - o Nachvollziehbarkeit der Beweiswerterhaltung einschl. Wahrung von Hard- und Softwareneutralität sowie Interoperabilität,
  - o Migrationsfähigkeit durch selbsttragende AIP.

Ein generisches System zur Beweiswerterhaltung beinhaltet insbesondere die folgenden Funktionen:

- Einholung und Prüfung der beweisrelevanten Daten des SIP/Version:
  - o Signaturprüfung, Einholung beweisrelevanter Daten (z.B. Zertifikatsinformationen, Sperrdaten) und Einlagerung im AIP/Version,

- Erzeugen von technischen Beweisdaten des AIP/Version:
  - o Evidence Record gem. RFC 4998 bzw. RFC 6283 einschl. Archivzeitstempel und Nachweis über Gültigkeit elektronischer Signaturen zum Signaturzeitpunkt sowie die rechtzeitige Signaturerneuerung / Hasherneuerung,
- Abruf der technischen Beweisdaten und beweisrelevanten Daten des AIP/Version,
- Prüfung der technischen Beweisdaten,
- Erhaltung durch Erneuerung der technischen Beweisdaten des AIP,
- Nachsignatur und Hasherneuerung.

4 Technische Richtlinie TR-ESOR (TR 03125)

Das BSI hat die Technische Richtlinie 03125 „Beweiswerterhaltung kryptographisch signierter Dokumente“ (TR-ESOR) ebenfalls auf Basis der Standards RFC 4998 und RFC6283 und der Ergebnisse der vorausgegangenen Projekte ArchiSig und ArchiSafe mit dem Ziel bereitgestellt, die Integrität und Authentizität archivierter Daten und Dokumente bis zum Ende der gesetzlich vorgeschriebenen Aufbewahrungspflicht unter Wahrung des rechtswirksamen Beweiswertes zu erhalten.

Thematisch behandelt die Technische Richtlinie dabei:

- Daten- und Dokumentenformate,
- Austauschformate für Archivdatenobjekte und Beweisdaten,
- Empfehlungen zu einer Referenzarchitektur, zu ihren Prozessen, Modulen und Schnittstellen als Konzept einer Middleware,
- zusätzliche Anforderungen für Bundesbehörden sowie
- Konformitätsregeln für die Konformitätsstufe 1 „logisch-funktional“ und die Konformitätsstufe 2 „technisch-interoperabel“ .

[4] Die Sicherheitseignung wird aktuell durch das Bundesamt für Sicherheit in der Informationstechnik und die Bundesnetzagentur festgestellt und veröffentlicht.

Auf der Basis des vorliegenden Anforderungskatalogs können Anbieter und Produkthersteller zu dieser Richtlinie 03125 konforme Lösungsangebote entwickeln, die auf Basis der Konformitätsstufe 1 „logisch-funktional“ bzw. der Konformitätsstufe 2 „technisch-interoperabel“ zertifiziert werden können.

4.1 TR 03125 Referenzarchitektur

Aus den funktionalen Anforderungen für den Erhalt des Beweiswerts wurde in der TR 03125 eine modulare Referenzarchitektur abgeleitet, die nachfolgend erklärt wird.

Die in der TR-ESOR für Zwecke des Beweiswerterhalts kryptographisch signierter Daten entwickelte Referenzarchitektur (siehe Abb. 5) besteht aus den folgenden funktionalen und logischen Einheiten:

- Das „ArchiSafe-Interface“ (TR-S. 4) bildet die Eingangs-Schnittstelle zur TR-ESOR-Middleware und bettet diese in die bestehende IT- und Infrastrukturlandschaft ein.

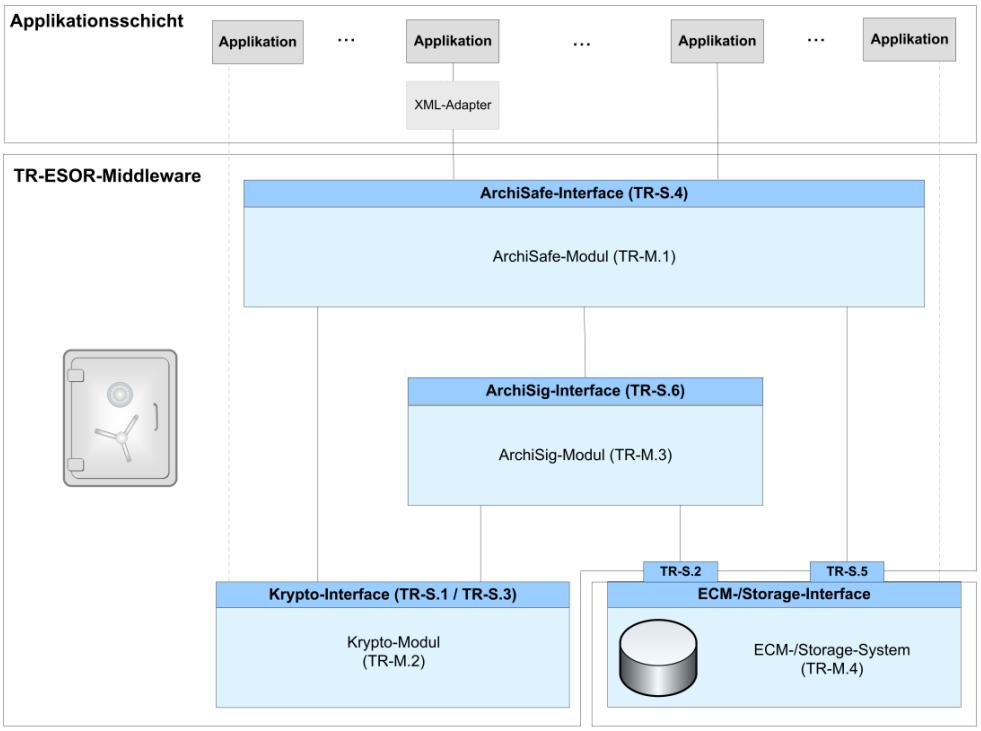


Abbildung 5: TR 03125 Referenzarchitektur

- Das „ArchiSafe-Modul“ (TR-M.1) regelt den Informationsfluss in der Middleware, sorgt dafür, dass die Sicherheitsanforderungen an die Schnittstellen zu den IT-Anwendungen umgesetzt werden und gewährleistet eine Entkopplung von Anwendungssystemen und Enterprise Content Management (ECM)/Langzeitspeicher. Die Sicherheitsanforderungen dieses Moduls sind im „Common Criteria Protection Profile for an ArchiSafe Compliant Middleware for Enabling the Legally compliant Long-Term Preservation of Electronic Documents (ACM_PP)” [BSI-PP-0049] definiert.
- Das „Krypto-Modul“ (TR-M.2) mit den Eingangsschnittstellen TR-S.1 und TR-S.3 stellt die kryptographischen Funktionen bereit, welche für den Beweiswerterhalt kryptographisch signierter Dokumente wesentlich sind. Das Krypto-Modul stellt Funktionen zur Erstellung (optional) und Prüfung elektronischer Signaturen, zur Nachprüfung elektronischer Zertifikate und zum Einholen qualifizierter Zeitstempel sowie weiterer beweisrelevanter Daten für die Middleware zur Verfügung. Das Krypto-Modul muss die Anforderungen des Gesetzes über Rahmenbedingungen für elektronische Signaturen (SigG) und der Verordnung zur elektronischen Signatur (SigV) erfüllen. Die Aufrufschnittstellen des Krypto-Moduls sollen nach dem eCard-API-Framework (vgl. [BSI-TR-03112], [OASIS-DSS] und [BSI-TR-03125-E]) gestaltet sein, um die Integration und Austauschbarkeit kryptographischer Funktionen zu erleichtern.

- Das „ArchiSig-Modul“ (TR-M.3) mit der Schnittstelle TR-S. 6 stellt die erforderlichen Funktionen für die Beweiswerterhaltung der digital signierten Unterlagen gemäß [RoSc06] zur Verfügung. Auf diese Weise wird gewährleistet, dass die in § 17 SigV geforderte Signaturerneuerung einerseits gesetzeskonform und anderseits performant und wirtschaftlich durchgeführt werden kann und somit dauerhafte Beweissicherheit gegeben ist. Das ArchiSig-Modul bildet in der Middleware faktisch das zentrale Modul zur technischen Beweiswerterhaltung und wird daher im Folgenden näher beschrieben:
  - Um den Nachweis der Integrität und damit auch der Authentizität eines Archivdatenobjekts (AIP) auch noch nach langer Zeit führen zu können, werden Hashwerte der jeweiligen Archivdatenobjekte zusätzlich in einem Merkle-Hashbaum gespeichert. Die Hashwerte werden mit einem qualifizierten Zeitstempel mit qualifizierter elektronischer Signatur geschützt. Der Zeitstempel zusammen mit der Liste der Hashwerte wird als Archivzeitstempel (engl. Archive-Time Stamp) bezeichnet. Archivbetreiber sind gehalten, die Sicherheitseignung der eingesetzten kryptographischen Algorithmen regelmäßig zu überprüfen.
  - Wenn nur der eingesetzte Signaturalgorithmus absehbar seine Sicherheitseignung verliert, aber der eingesetzte Hashalgorithmus beibehalten werden kann, ist es ausreichend, eine Zeitstempelerneuerung durchzuführen. Für diesen Zweck wird vor Eintritt dieses Zustandes ein neuer Zeitstempel über dem zuletzt erzeugten Zeitstempel erzeugt. Auf Basis dieses Prozesses entsteht mit der Zeit eine Folge von Archivzeitstempel, die in einer „ArchiveTimeStampChain“ enthalten sind ([DIN 31647], Kap. 3.3.2.).
  - Der kryptographisch geschützte Hashbaum ermöglicht so ein wirtschaftliches Verfahren zur Erneuerung elektronischer Signaturen, da nur ein zusätzlicher Zeitstempel pro Hashbaum benötigt wird.
  - Falls der eingesetzte Hash-Algorithmus in absehbarer Zeit seine Sicherheitseigenschaften verliert, muss eine Hashbaum-Erneuerung durchgeführt werden. Hierzu werden für alle Archivdatenobjekte neue Hashwerte berechnet und mit einem neu erzeugten Archivzeitstempel versehen. Das Ergebnis wird in einen neuen „ArchiveTimeStamp Chain“ eingefügt, so dass eine Sequenz von „ArchivTimeStampChains“ , eine sogenannte „ArchiveTimeStampSequence“ , entsteht.
  - Darüber hinaus unterstützt das ArchiSig-Modul u.a. auch den Abruf und Prüfung technischer Beweisdaten für den Nachweis der Integrität und Authentizität eines gespeicherten elektronischen Dokumentes und dessen Existenz zu einer bestimmten Zeit mittels der im RFC 4998 und im RFC 6283 standardisierten Beweisdaten (engl. Evidence Record). Bei der Erzeugung eines Evidence Records wird aus dem gesamten Hashbaum der reduzierter Hashbaum (siehe [RFC4998], [RFC6283]) für das entsprechende Archivdatenobjekt oder die entsprechende Archivdatenobjekt-Gruppe gewonnen und in eine ArchivTimeStampChain eingefügt, die wiederum in einer ArchivTimeStampSequence eingebettet wird (siehe auch [DIN 31647], Kap. 3.3.2).
- Das ECM- bzw. das Langzeitspeicher-System mit den Schnittstellen TR-S. 2 und TR-S. 5, das nicht mehr Teil der Technischen Richtlinie 03125 TR-ESOR ist, sorgt für die physische Archivierung/Aufbewahrung.

Die in Abb. 5 dargestellte IT-Referenzarchitektur soll die logische (funktionale) Interoperabilität künftiger Produkte mit den Zielen und Anforderungen der Technischen Richtlinie ermöglichen und unterstützen (siehe auch [BSI-TR-03125-C.1] und [BSI-TR-03125-C.2]).

Diese strikt plattform-, produkt-, und herstellerunabhängige Technische Richtlinie [BSI-TR-03125] hat einen modularen Aufbau und besteht aus einem Hauptdokument und Anlagen, die die funktionalen und sicherheitstechnischen Anforderungen an die einzelnen Module, Schnittstellen und Formate der TR-ESOR-Middleware beschreiben.

5 Zusammenspiel der DIN 31647 sowie der TR-03125

Das OAIS-Modell definiert grundsätzliche Anforderungen an Informationspakete und Prozesse zur Aufbewahrung elektronischer Unterlagen. Die TR-03125 des BSI dagegen beschreibt die spezifischen Anforderungen an die Beweiswerterhaltung digitaler Unterlagen unter Verwendung kryptographischer Methoden. Die Technische Richtlinie lässt dabei die Anwendung der definierten Komponenten und Prozesse sowohl für vor der Langzeitspeicherung, signierte als auch unsignierte Unterlagen zu. Dies ermöglicht den Aufbau ganzheitlicher elektronischer Langzeitspeicherlösungen, für alle elektronischen Unterlagen, wie dies z.B. bei der Bundesagentur für Arbeit, dem Bundesministerium für Gesundheit einschl. Geschäftsbereich oder dem DVZ Mecklenburg-Vorpommern bereits umgesetzt wird.

Die TR-03125 definiert grundlegende Prozesse und Anforderungen an Informationspakete aus Sicht der Beweiswerterhaltung. Damit wird das OAIS-Modell gezielt um die notwendigen Abläufe und Anforderungen für den Anwendungsfall der beweissicheren Aufbewahrung ergänzt. Die Technische Richtlinie enthält hierfür sehr detaillierte Anforderungen und Prozessabläufe sowohl für die Langzeitspeicherung im Allgemeinen (Definition Standardformate und -schnittstellen) als auch die Beweiswerterhaltung im Besonderen. Daneben werden die notwendigen Komponenten und Schnittstellen definiert sowie eine Referenzarchitektur empfohlen. Die TR-03125 fokussiert jedoch einzig und allein auf die Beweiswerterhaltung und Aspekte der technischen Interoperabilität. Außer dem Hinweis auf mögliche Standardformate werden Prozesse, Funktionen und Anforderungen an Informationspakete zur Erhaltung der elektronischen Unterlagen selbst (Datenerhaltung) und damit insbesondere deren Lesbarkeit ausgeblendet. Hier unterstützt das OAIS-Modell im Verbund mit der DIN 31644 und DIN 31645, die wiederum die spezifischen Fragen der Beweiswerterhaltung außer Acht lassen.

Zur Umsetzung der beweissicheren Langzeitspeicherung bedarf es damit einer Verknüpfung aus TR-03125 und OAIS-Modell sowie der diese untersetzenden Normen die DIN 31644 und DIN 31645. Hierbei sind Fragen der Beweiswerterhaltung als auch der Erhaltung der elektronischen Unterlagen selbst zu berücksichtigen. Im Fokus steht dabei jeweils die Erzeugung und Ablage selbsttragender AIP in herstellerneutraler und damit zur langfristigen Aufbewahrung geeigneter Form. Die hierfür notwendigen grundsätzlichen Anforderungen und generischen Funktionen beschreibt die DIN 31647 (Entwurf), die insbesondere unter Verwendung der TR 03125 umgesetzt werden kann.

6 Zusammenfassung

Es besteht eine hohe Notwendigkeit, nicht nur in der öffentlichen Verwaltung (E-Government) sondern auch in Unternehmen, Geschäftsprozesse zu digitalisieren und für die elektronischen Dokumente und Daten auch in ferner Zukunft die Lesbarkeit, Verfügbarkeit sowie die Integrität, Authentizität und Verkehrsfähigkeit gewährleisten zu können. Besondere Herausforderungen existieren in diesem Umfeld beim dauerhaften Erhalt des Beweiswerts kryptographisch signierter Dokumente. Vor diesem Hintergrund entwickelt der DIN-Arbeitskreis NA 009-00-15-06 „Beweiswerterhaltung kryptographisch signierter Dokumente“ mit der DIN 31647 eine verbindliche DIN-Norm. Diese setzt auf der Technischen Richtlinie TR 03215 „Beweiswerterhaltung kryptographisch signierter Dokumente“ des Bundesamtes für Sicherheit in der Informationstechnik (BSI) auf. Darüber hinaus bezieht der Normungsentwurf weitere maßgebliche Standards und Normen explizit

ein. Hierzu zählen: DIN 31644, DIN 31645 sowie das in ISO-14721:2012 genormte OAIS-Modell. Die DIN 31647 soll diese etablierten Standards zur digitalen Langzeitspeicherung um die notwendigen Anforderungen an die Beweiswerterhaltung kryptographisch signierter Dokumente ergänzen.

Im Zusammenspiel von OAIS-Modell, DIN 31644, DIN 31645 und TR-03125 wird eine ganzheitliche beweissichere Aufbewahrung elektronischer Unterlagen auf Basis geltender Standards und Normen ermöglicht. Diese Symbiose aus Anforderungen und Regeln aus dem Umfeld der Gedächtnisorganisationen sowie dem Records Management ermöglicht so den Aufbau wirtschaftlicher und langfristiger Lösungen. Mit der DIN 31647 wird eine normative Grundlage zur zielgerichteten Konzeption und Entwicklung einer Langzeitspeicherung erarbeitet, die eine Beweiswerterhaltung in standardisierter Form unter Berücksichtigung der Maßgaben des OAIS-Modells mittels insbesondere selbsttragender AIPs und damit Herstellerneutralität und so eine langfristig sichere Aufbewahrung elektronischer Unterlagen ermöglicht.

7 Literaturverzeichnis

[BArchG] Gesetz über die Sicherung und Nutzung von Archivgut des Bundes (Bundesarchivgesetz - BArchG) vom 6. Januar 1988 (BGBl. I S. 62), zuletzt geändert durch § 13 Abs. 2 des Informationsfreiheitsgesetzes vom 5. September 2005 (BGBl. I S. 2722).

[BMI 12] Organisationskonzept elektronische Verwaltungsarbeit. Baustein E-Akte. Bundesministerium des Innern (Hrsg), Berlin 2012.

[BMWi 07] Handlungsleitfaden zur Aufbewahrung elektronischer und elektronisch signierter Dokumente. Bundesministerium für Wirtschaft und Technologie (Hrsg.), Berlin 2007.

[BSI-PP-0049] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Common Criteria Protection Profile for an ArchiSafe Compliant Middleware for Enabling the Long-Term Preservation of Electronic Documents (ACM_PP)*, Version 1.0, 2008.

[BSI-TR-03125] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Beweiswerterhaltung kryptographisch signierter Dokumente (TR-ESOR)*, TR 03125, Version 1.1,, 2011. https://www.bsi.bund.de/ContentBSI/Publikationen/TechnischeRichtlinien/tr03125/index_htm.html.

[BSI-TR-03125-B] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Anlage B zu [BSI-TR-03125], Profilierung für Bundesbehörden*, Ver. 1.1, 2011.

[BSI-TR-03125-C.1] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Anlage C.1 zu [BSI-TR-03125], Conformity Test Specification (Level 1 – Functional Conformity)*, Ver. 1.1, 2012.

[BSI-TR-03125-C.2] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Anlage C.2 zu [BSI-TR-03125], Conformity Test Specification (Level 2 – Technical Conformity)*, geplant für 2013.

[BSI-TR-03125-E] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Anlage E zu [BSI-TR-03125]: Konkretisierung der Schnittstellen auf Basis des eCard-API-Frameworks*, TR 03125 Version 1.1, 2011.

[BSI-TR-03125-F] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Anlage F zu [BSI-TR-03125], Formate und Protokolle*, Version 1.1, 2011.

[BSI-TR-RESISCAN] Bundesamt für Sicherheit in der Informationstechnik (BSI): *Rechtssicheres ersetzendes Scannen (TR-RESISCAN)*, Version 1.0 2013.

[DIN 31644:2012-04] Information und Dokumentation - Kriterien für vertrauenswürdige digitale Langzeitarchive.

[DIN 31645:2011-11] Information und Dokumentation - Leitfaden zur Informationsübernahme in digitale Langzeitarchive.

[DIN 31646:2013-01] Information und Dokumentation - Anforderungen an die langfristige Handhabung persistenter Identifikatoren (Persistent Identifier).

[DIN-31647] DIN 31645:2013-Entwurf: Beweiswerterhalt kryptografisch signierter Dokumente.

[EGovG-RE] Referentenentwurf der Bundesregierung: *Gesetz zur Förderung der elektronischen Verwaltung sowie zur Änderung weiterer Vorschriften*, Bearbeitungsstand 16.03.2012, über: http://www.bmi.bund.de/SharedDocs/Downloads/DE/Gesetzestexte/Entwuerfe/Entwurf_EGov.html.

[F 06] S. Fischer-Dieskau: Das elektronisch signierte Dokument als Mittel zur Beweissicherung. Baden-Baden 2006.

[OASIS-DSS] OASIS: *Digital Signature Service Core, Protocols, Elements, and Bindings*, Version 1.0, http://docs.oasis-open.org/dss/v1.0/oasis-dss-core-spec-v1.0-os.html, 2007.

[RFC3161] C. Adams, P. Cain, D. Pinkas, R. Zuccherato: *Internet X.509 Public Key Infrastructure – Time-Stamp Protocol (TSP)*, IETF RFC 3161, http://www.ietf.org/rfc/rfc3161.txt, 2001.

[RFC4998] T. Gondrom, R. Brandner, U. Pordesch: *Evidence Record Syntax (ERS)*, IETF RFC 4998, http://www.ietf.org/rfc/rfc4998.txt, August 2007.

[RFC6283] A. J. Blazic, S. Saljic, T. Gondrom: *Extensible Markup Language Evidence Record Syntax (XMLERS)*, IETF RFC 6283, http://www.ietf.org/rfc/rfc6283.txt, Juli 2011.

[RoSc06] A. Rossnagel, P. Schmücker (Hrsg.): *Beweiskräftige elektronische Archivierung. Ergebnisse des Forschungsprojektes „ArchiSig – Beweiskräftige und sichere Langzeitarchivierung digital signierter Dokumente“*, Economica Verlag, 2006.

[Ro07] A. Rossnagel: Langfristige Aufbewahrung elektronischer Dokumente, Anforderungen und Trends, Baden-Baden, 2007.