INFORMATIONSWISSENSCHAFT
Theorie, Methode und Praxis
SCIENCES DE L'INFORMATION
Théorie, méthode et pratique

2018 - 1

Beiträge der 21. Jahrestagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen» Basel, 28. Februar und 1. März 2017

Georg Büchler (Hg./éd.)

Weiterbildungsprogramm in Archiv-, Bibliotheks- und Informationswissenschaft
Programme de formation continue en archivistique, bibliothéconomie et sciences de l'information

Historisches Institut der Universität Bern
Séction d'histoire de la Faculté des Lettres de l'Université de Lausanne

UNIL | Université de Lausanne

UNIVERSITÄT BERN

Informationswissenschaft:
Theorie, Methode und Praxis
Sciences de l'information:
théorie, méthode et pratique

Beiträge der 21. Jahrestagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen» Basel, 28. Februar und 1. März 2017

Georg Büchler (Hg./éd.)

2018
Bern

Reihe:

Informationswissenschaft:
Theorie, Methode und Praxis
Sciences de l'information:
théorie, méthode et pratique

herausgegeben von / édité par:
Gaby Knoch-Mund, Ulrich Reimer, Barbara Roth-Lochner

Band 5 (2018), Nr. 1

ISBN 978-3-906813-22-6
ISSN 2297-9069
DOI: http://dx.doi.org/10.18755/iw.2018.1
Online-Publikation: Bern Open Publishing, bop.unibe.ch/iw/

Inhalt

Vorwort

Georg Büchler

Am 28. Februar und 1. März 2017 nahmen 147 digitale Archivarinnen und Archivare aus Deutschland, der Schweiz, Österreich, Liechtenstein, Tschechien und Ungarn an der 21. Jahrestagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen» in Basel teil. Für die KOST, die seit der zehnten Ausgabe 2006 an den Arbeitskreistagungen teilnimmt, und für das Staatsarchiv Basel-Stadt war es eine besondere Ehre, die Kolleginnen und Kollegen zum zweiten Mal nach 2009 in der Schweiz begrüssen zu dürfen. Aus bescheidenen Anfängen hat sich die Arbeitskreistagung längst zur wichtigsten deutschsprachigen Veranstaltung im Bereich der digitalen Archivierung entwickelt. Auch dieses Mal nahmen Erfahrungsberichte und Beispiele aus der konkreten Archivpraxis einen wichtigen Raum ein. Daneben fanden grundsätzliche und mehr theoretische Überlegungen ebenfalls ihren Platz.

Die Tagungsbeiträge sind in diesem Band in Artikelform publiziert. Verzichtet wurde dabei auf die Publikation von Beiträgen, die einen bereits wieder überholten Arbeitsstand referierten oder in erster Linie der Information der Fachgemeinde dienten. Alle Präsentationen sind jedoch wie jedes Jahr auf den Webseiten des Arbeitskreises beim Staatsarchiv St. Gallen dokumentiert und zugänglich unter http://www.staatsarchiv.sg.ch/home/auds/.[1]

Die herausragende Stellung praxisnaher Beiträge an den Tagungen des Arbeitskreises impliziert nicht einen Verzicht auf ein theoretisches Fundament. Die erste Session der Basler Tagung war deshalb Modellen und Grundsatzüberlegungen gewidmet. Lambert Kansy (Staatsarchiv Basel-Stadt) und Martin Lüthi (Staatsarchiv St. Gallen) präsentierten das Resultat eines gemeinsamen KOST-Projekts, das *KOST-Diskussionspapier AIS-Modell*. Dieses definiert in möglichst generischer Weise die archivischen Kernprozesse, die dazugehörigen Informationsobjekte und notwendige Schnittstellen zu Umsystemen und soll so als Ausgangspunkt für eine vertiefte Diskussion archivischer Anforderungen an das AIS, die zentrale Fachanwendung eines Archivs dienen. Auf ein so bekanntes wie umstrittenes Konzept fokussierte Prof. Dr. Christian Keitel vom Landesarchiv Baden-Württemberg in seinem Beitrag *Der einzige Kompass, den wir haben. Zur Kritik der Designated Community*. Gegen vergangene und aktuelle Kritik am OAIS-Konzept der Designated Communities unterstrich er die Notwendigkeit, archivische Entscheide explizit

1 Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.2

auf Annahmen über die künftigen Nutzer des Archivguts und deren Nutzungsziele zu gründen – implizit tun wir dies ohnehin.

Die zweite Session hatte den Anspruch, Licht in eine selten öffentlich beleuchtete Ecke der digitalen Archivierung zu bringen, nämlich in Fragen zu den *Kosten der digitalen Archivierung*. Detaillierte Zahlen präsentierte Dr. Krystyna Ohnesorge vom Schweizerischen Bundesarchiv BAR, die in ihrem Referat mit dem Titel *Digitale Archivierung im BAR – Ein Blick hinter die Kulissen* das Angebot des BAR zur digitalen Archivierung für Dritte vorstellt und die Überlegungen erläuterte, die zur Preisgestaltung geführt haben. Dr. Julia Krämer-Riedel, Historisches Archiv der Stadt Köln, und Tobias Schröter-Karin, LWL-Archivamt für Westfalen, referierten über *Archivierung im Verbund: Kosten der digitalen Archivierung am Beispiel von DiPS.kommunal*. DIPS.kommunal bietet als Teil des Digitalen Archivs NRW den kommunalen Archiven digitale Archivierung als Service an. In einer solchen Konstellation ist es absolut zentral, die Kosten für diese Dienstleistung nachvollziehbar zu machen und im Detail zu rechtfertigen. Die Hostinglösung docuteam cosmos stand im Zentrum des Referats *Kostenmodelle für digitale Archivierung – Vergleich von Theorie und Praxiserfahrung* von Dr. Tobias Wildi, docuteam. Er diskutierte verschiedene Modelle zur Kostenanalyse und -berechnung in der digitalen Langzeitarchivierung, die in den letzten Jahren publiziert worden waren, und zeigte ihre Limiten in der praktischen Anwendung auf. Das Kostenmodell von docuteam beruht auf vier Kostenkategorien und umfasst verschiedene Berechnungen von Kostenfaktoren. Da das Angebot noch neu und in Entwicklung ist, wird auf eine Publikation in Artikelform zum gegenwärtigen Zeitpunkt verzichtet. Die Session wurde beschlossen von Maurice Heinrich vom Deutschen Archäologischen Institut mit dem Referat *Archivierung von digitalen Forschungsdaten der Altertumswissenschaften – Kosten- und Finanzierungskonzepte*. Das vorgestellte Modell ist auf die Situation und die spezifischen Anforderungen einer Forschungsrichtung ausgerichtet und versucht, Projektfinanzierung und institutionelle Verankerung miteinander zu verbinden. Auch dieses Referat ist im vorliegenden Band nicht publiziert.

Dass viele Archive in den letzten Jahren eine gewisse Übung im präzisen Umgang mit Dateiformaten erworben haben, zeigte sich in der dritten, sehr praktisch ausgerichteten Session. Dr. Kai Naumann vom Landesarchiv Baden-Württemberg und Dr. Christoph Schmidt vom Landesarchiv Nordrhein-Westfalen analysierten unter dem Titel *Chancen und Risiken des Einsatzes verlustbehafteter Bildkompression in der digitalen Archivierung* den Informationsverlust beim Scannen sowie bei der Umwandlung von TIFF in JPEG2000. Die Ergebnisse ihrer Untersuchung stellen den noch weitgehend geltenden archivischen Konsens über die Superiorität von TIFF als Bildarchivformat stark in Frage. Martin Kaiser (KOST) berichtete über eine gross angelegte *TIFF-Korpus-Analyse* der KOST zusammen mit drei ihrer

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.2

Trägerarchive, in deren Rahmen vier Millionen TIFF-Dateien aus den drei Archiven mit verschiedenen Validierungs- und Charakterisierungswerkzeugen analysiert worden waren. Damit liegen nun erstmals Daten über ein sehr grosses Korpus realer TIFF-Dateien vor, welche fundierte Aussagen und gezielte Planungen ermöglichen. Der erste Kongresstag wurde beschlossen von Stephanie Kortyla und Christian Treu (Sächsisches Staatsarchiv), die anhand einiger prägnanter Beispiele die *Nutzen und Grenzen der Formatidentifizierung* beim Preservation Planning auszuloten versuchten. Die Limiten der Formatdatenbank PRONOM und des Identifikationstools DROID erfordern einen bewussten Einsatz und grosse Fachkenntnis der beteiligten Archivarinnen und Archivare.

Noch tiefer in die konkrete Arbeit tauchte die Arbeitskreistagung in der vierten Session ein, *Praxisberichte: Behördenberatung, Fachverfahren, Webarchivierung*. Vor der Archivierung steht die Behördenberatung. Dazu präsentierten eingangs Dr. Christine Friederich und Dr. Martin Schlemmer (Landesarchiv Nordrhein-Westfalen) *Das E-Government-Gesetz NRW und die Praxis der Behördenberatung – ein Werkstattbericht aus dem Landesarchiv NRW*. Das E-Government-Gesetz hat die Rahmenbedingungen für die Behördenberatung in Nordrhein-Westfalen verändert und beim Landesarchiv zu einer Neuorganisation dieser besonders für digitale Unterlagen zentralen Aufgabe geführt. Eine ähnliche Fragestellung untersuchte der Beitrag *Anwendung «Fachverfahren in der Bundesverwaltung» – Behörden-Fachdatenanwendungen im archivischen Überblick* von Joachim Rausch und Marion Teichmann (Deutsches Bundesarchiv). Mit der vorgestellten Anwendung hat das Bundesarchiv die frühere Praxis der Erhebungen mittels Fragebogen abgelöst und zugleich die Möglichkeiten der Benutzbarkeit und Auswertung der erhobenen Informationen stark verbessert. Dieses Referat war als Produktpräsentation angelegt und ist deshalb im vorliegenden Band nicht enthalten. Zur eigentlichen Archivierung aus Fachverfahren (bzw. in der schweizerischen Terminologie Fachanwendungen) referierten danach Ursina Rodenkirch-Brändli und Bernhard Stüssi (Staatsarchiv Graubünden) unter dem Titel *Archivierung aus Fachanwendungen im Staatsarchiv Graubünden: ein Werkstattbericht*. Sie präsentierten zwei aktuelle Projekte mit Anwendungen aus dem Strassenverkehrsamt und der Sozialversicherungsanstalt und fragten insbesondere nach der Definition einer das Verwaltungshandeln korrekt abbildenden Akteneinheit in beiden Fällen. Ein weiterer archivischer Kernprozess, der bei der digitalen Archivierung, und dort vor allem bei der Archivierung aus Fachverfahren, grosse Änderungen erfordert, ist die Erschliessung. Dr. Sigrid Schieber (Hessisches Hauptstaatsarchiv) fragte dazu: *Die Erschließung strukturierter Massendaten aus Datenbanken – was ist nötig, um solche Daten interpretierbar und benutzbar zu machen?* Über die klassisch archivische Erschliessung hinaus ist nämlich von der Binnenerschliessung der Datenbankinhalte zu sprechen. Diese muss der

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.2

Archivnutzerin und dem Archivnutzer ermöglichen, die Archivalien zu interpretieren und zu verstehen, und muss dazu unter anderem auch den Bewertungs- und Übernahmeprozess in geeigneter Form dokumentieren. Da Frau Dr. Schieber diese Thematik am Deutschen Archivtag weiterentwickelt hat, wird sie in diesem Rahmen publiziert werden. Ein eher marginales, aber an den Arbeitskreistagungen schon fast klassisches Thema rundete diese Session ab: das Referat *Webarchivierung in der Praxis: Erfahrungen des Staatsarchivs Basel-Stadt* von Kerstin Brunner und Olivier Debenath (Staatsarchiv Basel-Stadt). Es bot einen Rückblick auf vier Jahre systematische Bewertung und Archivierung von Webseiten der kantonalen Verwaltung, und zwar sowohl auf die archivfachlichen als auch auf die technischen Aspekte. Die gemachten Erfahrungen und die daraus gezogenen Lehren werden die zukünftige Webarchivierung in Basel-Stadt gestalten.

Erst in den letzten Jahren sind Zugang und digitaler Lesesaal langsam ein Thema geworden. Die fünfte Session beleuchtete dazu zwei Aspekte: Zbyšek Stodůlka vom Nationalarchiv Prag berichtete über die *E-Identität als Schlüssel zu den Dienstleistungen des digitalen Archivs*. Im Rahmen von E-Government nehmen elektronische Identifizierung und Vertrauensdienste für elektronische Transaktionen eine zentrale Rolle ein. Für ein Archivportal eröffnet diese neue Infrastruktur neue Möglichkeiten, aber auch vielfältige Herausforderungen. Dr. Beat Gnädinger, Staatsarchiv Zürich referierte abschliessend zum Thema *Digital Access - Ja! Aber wie? Die Online-Werkzeuge des Staatsarchivs Zürich*. Mit diesem Rückblick auf die Entwicklung im Online-Zugang zu Archivmetadaten steckte er den Rahmen ab, in welchem der zukünftige Zugriff auf digitale Primärdaten erfolgen kann. Das Referat hatte einen Überblickscharakter und wurde für die Publikation nicht ausgearbeitet.

Die 21. Jahrestagung des Arbeitskreises AUdS war geprägt von einer gewachsenen, vielfältigen Community. Ich hoffe, dass dieser Eindruck auch bei der Lektüre des Tagungsbandes durchscheint. Nicht abbilden kann dieser Band die vielfältigen und reichen Diskussionen in den Pausen und beim Abendessen, welche immer einen besonderen Reichtum der Arbeitskreistagung darstellen. Die Diskussion geht weiter, 2018 in Marburg, 2019 in Prag. Die KOST freut sich, auch weiterhin dazu beitragen zu können.

Bern, im Februar 2018
Georg Büchler,
Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen
KOST

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.2

Ein konzeptionelles Modell für Archivinformationssysteme. Das KOST-Diskussionspapier AIS-Modell[1]

Lambert Kansy, Martin Lüthi

Einleitung

Der vorliegende Beitrag stellt die Ergebnisse des Projekts «Referenzmodell AIS» der Koordinationsstelle für die dauerhafte Archivierung von elektronischen Unterlagen (KOST) zur Entwicklung eines konzeptionellen Modells für Archivinformationssysteme vor.[2]

Ausgehend von Erfahrungen der beiden Staatsarchive Basel-Stadt und St.Gallen bei der Einführung ihres Archivinformationssystems zwischen 1997 und 2003 sowie weiteren Projekten im Umfeld der Archivierung von digitalen Unterlagen und zuletzt bei der Arbeit an der Realisierung eines digitalen Zugangs wurde deutlich, dass es an grundlegenden konzeptionellen Überlegungen mangelt, wie Archivinformationssysteme standardisiert werden können und gestaltet sein können, um den Bedürfnissen der Anwender nachhaltig zu entsprechen.

Gemeinsam konzipierten die KOST und die Staatsarchive der Kantone Basel-Stadt, Bern und St.Gallen ein Projekt, mit dem die Grundlagen geschaffen werden sollten, um die Anforderungen an archivische Informationssysteme möglichst generisch zu definieren und eine Basis zu schaffen, auf der aufbauend eine Standardisierung dieser Anforderungen erfolgen kann.

Ausgehend von den archivischen Geschäftsprozessen wurden Anforderungen an Archivinformationssysteme definiert und Hilfsmittel für Hersteller wie Archive entwickelt, um die konkrete Realisierung beziehungsweise Evaluation solcher Systeme zu unterstützen.

---

1 Die zum vorliegenden Beitrag zugehörige Präsentation ist abrufbar auf der Website des Arbeitskreises für die Archivierung von Unterlagen aus digitalen Systemen: http://www.staatsarchiv.sg.ch/home/auds/21/_jcr_content/Par/downloadlist_0/DownloadListPar/download.ocFile/1_Kansy_L%C3%BCthi.pdf. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Informationen zu dem Projekt KOST 14-026 finden sich auf der Website der KOST: https://kost-ceco.ch/cms/index.php?14-026_de. Der Projektverlauf im ersten Jahr sowie die Darstellung der Gründe, die zu dem Projekt geführt haben, wurden auf der Tagung des Arbeitskreises für die Archivierung von Unterlagen aus digitalen Systemen 2016 in Potsdam vorgestellt. Siehe dazu die Präsentation des Vortrags auf der Website des Arbeitskreises: http://www.staatsarchiv.sg.ch/home/auds/20/_jcr_content/Par/downloadlist_2/DownloadListPar/download.ocFile/Pr%C3%A4sentation%20AUdS%20M%C3%A4rz%202016%20Potsdam%20V%201%200.pdf und Kansy, Lambert;Lüthi, Martin: Ein Referenzmodell für Archivinformationssysteme, Potsdam 2017 (Publikation bevorstehend).

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.3

Hintergrund und Ziele

Der Einsatzbereich von Archivinformationssystemen, mit denen Archivgut verwaltet und in denen hierzu Arbeitsprozesse abgebildet und gesteuert werden, ist aufgrund der geringen Anzahl von Archivinstitutionen per se klein und der Softwaremarkt hierfür entsprechend limitiert. Als eigenständige Systeme entstanden sie national wie international seit den 1990er Jahren. Diese Entwicklung wurde stark von der geringen Standardisierung sowohl der Prozesse wie auch der Nutzung von Metadaten im Archivbereich beeinflusst. Zwar hätten mit den Standards des Internationalen Archivrats, die in dem gleichen Zeitraum entwickelt wurden, entsprechende Standards für die Verwaltung von Archivgut zur Verfügung gestanden. Die Gleichzeitigkeit der Entwicklung von Standards und Systemen hat hier aber interessanterweise nicht zu einer frühen Standardisierung der Systeme geführt. Die Standardisierung erfolgte dabei nicht gleichermassen für alle Arbeitsprozesse. Sie konzentrierte sich auf die Verwaltung von Archivgut und dessen Beschreibung. Vor- und nachgelagerte Prozesse wurden weniger bis gar nicht einbezogen. So fehlen Standards zur Beschreibung von Prozessen und deren Arbeitsobjekten im Bereich der Bewertung, aber auch im Bereich der Ablieferung und Übernahme

Es resultierten heterogene Anforderungen an die Features von Archivinformationssystemen mit Datenmodellen, die nicht auf Interoperabilität angelegt sind, und uneinheitliche Schnittstellen für den Import und für den Export. Vorhandene Schnittstellen sind in ihrem Funktionsumfang oft stark eingeschränkt und nicht dokumentiert respektive offengelegt. Es handelt sich somit um eigentliche Silo-Anwendungen. Ein weiteres Merkmal heutiger Archivinformationssysteme ist die häufig gegebene integrierte Architektur, in der eine Vielzahl von Funktionalitäten zur Abbildung diverser Aufgaben und Prozesse innerhalb eines IT-Systems zusammengefasst werden. Der Funktionsumfang der daraus resultierenden Systeme ist jedoch von System zu System verschieden. Hinzu kommt, dass diese Funktionen fest aneinander gekoppelt sind. Diese Charakteristika, heterogene Anforderungen, Silo-Anwendungen und fest integrierte Systeme, führen zu einem hochgradigen Vendor-Lock-In. Archive müssen hohe Aufwände und Kosten einsetzen, um Archivinformationssysteme auszutauschen oder in Systemlandschaften zu integrieren. Auch ist der Datenaustausch mit erhöhten Kosten aufgrund der geringen Standardisierung der Schnittstellen verbunden. Die Weiterentwicklung der archivischen Standards kann damit ebenso wenig aufgenommen werden wie Entwicklungen in der Arbeit der öffentlichen Verwaltung wie von Firmen oder branchenübergreifende Informationsportale für Kulturgut und die Entwicklung von *linked data*.

Auf diesem Hintergrund entstand bei den Projektbeteiligten das Bedürfnis, diese gewachsene Situation aufzubrechen und eine generische Prozess- und Informationsarchitektur zu entwickeln, um die Weiterentwicklung der bestehenden Systeme

systematisch gestalten und die Veränderungen in der Arbeit der Archive in den letzten 20 Jahren einbeziehen zu können. Die angestrebte Informationsarchitektur basiert auf dem Konzept der losen Koppelung, die eine modulare Systemarchitektur erlaubt. Damit können einzelne Komponenten ohne weiteres ausgetauscht und die Abhängigkeiten zwischen diesen Komponenten auf ein Minimum reduziert werden durch die Festlegung von Schnittstellen zwischen diesen und nicht nur zwischen dem Archivinformationssystem und weiteren IT-Systemen, mit denen es interagiert. Auf diese Weise soll eine Architektur geschaffen werden, die Interoperabilität und Datenaustausch befördert. Auch sollen Entwicklungen der archivischen Standards aufgenommen werden können, wie sie sich seit kurzem abzeichnen.

In dem KOST-Diskussionspapier AIS-Modell wird ein auf diesen Grundlagen aufgebautes Archivinformationssystem entwickelt. Das Resultat des Projekts ist jedoch kein ausgearbeitetes Lösungskonzept und auch keine Detailspezifikation eines Informationssystems. Es bietet auch keine Referenzimplementierung an, sondern ein Konzept, welche archivischen Geschäftsprozesse in einem Archivinformationssystem abgebildet werden sollen, und eine Umsetzung der Prozesse in eine Informationsarchitektur.

Das KOST-Diskussionspapier bildet die Basis für eine archivfachliche Diskussion, in der vertieft über die Anforderungen an die Abbildung und Steuerung von Arbeitsprozessen in Archivinformationssystemen sowie die Verwaltung von Informationen über Archivgut in derartigen Informationssystemen diskutiert wird. Auf den Ergebnissen der Diskussion und einer folgenden Überarbeitung kann dieses Modell die Grundlage einer Standardisierung von Anforderungen an Archivinformationssysteme sein.[3]

Projektmethodik und -verlauf

Für die Erarbeitung eines generischen Ansatzes erwies sich die Zusammensetzung des Projektteams aus Vertretern der drei Staatsarchive Basel-Stadt, Bern und St.Gallen als äusserst förderlich, da auf diese Weise unterschiedliche lokale Ausprägungen und Nutzungsformen des Archivinformationssystem eingebracht und somit der Gefahr entgegengewirkt werden konnte, spezifische Aspekte zu verallgemeinern. Die Mitwirkung der KOST brachte eine weitere Erweiterung der konzeptionel-

3 Diese Zielsetzung resp. der Verzicht auf die Schaffung eines Standards war anfänglich nicht enthalten. In der ersten Projektdefinition wurde die Schaffung eines eCH-Standards angestrebt. Es zeigte sich jedoch im ersten Jahr der Projektarbeit, dass dies zum einen mit den verfügbaren Ressourcen und innerhalb des gesetzten Zeitrahmens nicht möglich war und zum anderen auch den Verzicht auf eine breitere Fachdiskussion bedeutet hätte. Gerade diese aber, so wurde deutlich, ist unabdingbar für eine Standardisierung.

len Überlegungen mit sich. Ohne diesen breitgefächerten Erfahrungshintergrund wäre die so rasche Erarbeitung des AIS-Modells nicht möglich gewesen.

Zu Beginn des Projekts befasste sich das Projektteam mit der Definition des Begriffs Archivinformationssystem. Zudem wurde eine Reihe von Prämissen für die Modellierung des AIS-Modells entwickelt. Im weiteren Verlauf wurden zuerst die archivischen Geschäftsprozesse modelliert und beschrieben, da ein prozessbasiertes Verständnis der Systemspezifikation zugrunde gelegt wurde. Auf dieser Basis wurden anschliessend die Informationsobjekte des Archivinformationssystems erarbeitet mit ihren Eigenschaften und Funktionen. Dabei wurde der Umfang des Archivinformationssystems in mehreren Iterationen immer präziser definiert. Auch wurden hierbei die Schnittstellen zwischen dem Archivinformationssystem und Fremdsystemen definiert. Ein wichtiger und (zeit-)intensiver Arbeitsschritt bestand in der Festlegung des Funktionsumfangs des Archivinformationssystems. Diesbezüglich zeichnete sich das Projekt durch eine zunehmende Verschlankung aus.

Mit einer ersten Fassung des AIS-Modells wurde im August und September 2016 eine Reihe von Experten im Bereich der Archivinformatik um eine Review des Entwurfs angefragt. Dies war erforderlich, um zu verhindern, dass Vorannahmen unerkannt bleiben und trotz aller Bemühungen auf lokale Besonderheiten abgestellt wird. Die Reviewteilnehmer haben dem Projekt eine grosse Anzahl wertvoller und weiterführender Rückmeldungen gegeben, die im September/Oktober 2016 in den Entwurf eingearbeitet wurden. Der Grundtenor der Rückmeldungen war dabei positiv. Es zeigte sich, dass die Grundausrichtung, das Archivinformationssystem ausgehend von den Geschäftsprozessen zu modellieren, richtig war und dass auch der Modellierung der Prozesse weitgehend zugestimmt werden konnte.

Die Projektergebnisse wurde schliesslich in Form eines KOST-Diskussionspapiers mit einer Reihe von Anhängen publiziert.[4] Letztere erlauben eine Nachnutzung der Prozessmodellierungen ebenso wie der Spezifikation der Informationsobjekte und der Schnittstellen.

Grundsätze der Modellierung und Spezifikation

Methodisch wurde im Projekt eine an TOGAF[5] angelehnte Sicht auf die Architektur von Informationssystemen verwendet. Diese verwendet die Metapher eines «Architekturhauses». Die Gesamtarchitektur entspricht dabei einem Gebäude, die einzelnen Stockwerke repräsentieren jeweils die verschiedenen Architekturschichten. Die Modellierung eines solchen Architekturhauses erfolgt von oben nach unten. Im

4 Siehe http://kost-ceco.ch/cms/index.php?ais_conceptual_model_de.

5 TOGAF ist ein Standard von The Open Group zur Darstellung von Informatiksystemarchitekturen. Siehe http://togaf.org/.

obersten Stockwerk wird die Geschäftsarchitektur anhand der Geschäftsprozesse definiert. Davon abgeleitet wird die Anwendungsarchitektur beschrieben, bestehend aus Informationsobjekten und Anwendungskomponenten. Schliesslich folgt die Definition der Technologiearchitektur. Das Bild des Architekturhauses erlaubt es, die zusammengehörenden Geschäftsprozesse, Informationsobjekte und Anwendungen sowie Technologiekomponenten aufeinander bezogen darzustellen und zugleich den Primat der Geschäftsprozesse vor der Objekt- und Anwendungsmodellierung und der Technologieschicht durchzusetzen. Auf diese Weise werden die unterschiedlichen Perspektiven[6] auf die Gesamtarchitektur adäquat gewichtet.

Die Modellierung der Geschäftsprozesse erfolgte mit BPMN 2.0[7] respektive mit dem auf BPMN 2.0 basierenden eCH-Standard eCH-0158.[8] Zusätzlich zu diesen Standards werden Prozesslandkarten eingesetzt, um Prozessgruppen zu gliedern und übersichtlich zu gestalten.

Bei der Erarbeitung der Objektmodellierung wurde ArchiMate® 2.1 eingesetzt.[9] Dieser Modellierungsstandard erlaubt eine präzise Modellierung der Anwendungsschicht und der Informationsobjekte. Er bietet zudem die Möglichkeit, die beiden Architekturebenen *Geschäftsprozesse* und *Anwendungen* respektive *Informationsobjekte* in der Darstellung miteinander zu verbinden und so die Verbindungen beider Ebenen aufzuzeigen.

Für die detaillierte Beschreibung der Informationsobjekte und der Schnittstellen wurden Klassendiagramme aus UML verwendet.[10] Das Ergebnis wird sowohl in einer grafischen Ausgabe als auch in einer XML-Struktur verfügbar gemacht.

Während das AIS-Modell die Architekturebenen der Geschäftsprozesse und der Informationsobjekte und Anwendungen, letztere als Schnittstellen, detailliert beschreibt, beschränken sich die Aussagen zur Technologieschicht auf eine Reihe von grundlegenden Anforderungen, die jedoch keine Implementierungsvorgaben darstellen.

6 Business, Application, Technology. Siehe ArchiMate 2.1 Specification: http://pubs.opengroup.org/architecture/archimate2-doc/.

7 BPMN 2.0 ist ein offener Standard der Object Management Group (OMG) zur Modellierung von Geschäftsprozessen. Neben einer grafischen Darstellung der Prozessmodelle können diese auch als XML-Strukturen ausgegeben und in dieser Form softwareübergreifend verwendet werden. Siehe http://www.omg.org/spec/BPMN/.

8 eCH-0158 ist ein Standard von eCH und definiert ein Subset von BPMN 2.0. Er ist ausgerichtet auf die Darstellung der Prozesssicht aus Geschäftsoptik. Er schliesst die weitergehenden Möglichkeiten von BPMN 2.0 bewusst aus. Siehe https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0158.

9 Archimate ist ein Standard von The Open Group und ist frei verfügbar. Siehe http://www.opengroup.org/subjectareas/enterprise/archimate-overview.

10 UML steht für Unified Modelling Language und ist eine etablierte Methode für die Modellierung von Informationsobjekten mit ihren Eigenschaften und Methoden. Sie wird von der Object Management Group OMG betreut und wurde als ISO 19505-1 und ISO 19505-2 standardisiert. Siehe http://www.omg.org/spec/UML/.

Grundlegend für die Erarbeitung des Referenzmodells war die Entscheidung, nicht alle archivischen Geschäftsprozesse im Archivinformationssystem abzubilden, sondern nur eine Teilmenge davon. Das Archivinformationssystem soll diejenigen Prozesse abbilden und die dazugehörigen Informationsobjekte verwalten, die mit dem Archivgut als Objekt archivischer Tätigkeit zentral verbunden sind. Es handelt sich um Prozesse, die spezifisch für die Arbeit der Archive sind respektive eine spezifisch archivische Ausprägung aufweisen. Die übrigen archivischen Prozesse, die an andere Systeme delegiert werden können, sollen über definierte Schnittstellen eingebunden werden. Bei der Definition der Geschäftsprozessarchitektur wurde der Grundsatz verfolgt, die Bearbeitung analogen und digitalen Archivguts nach Möglichkeit in einem Prozess gemeinsam zu beschreiben und lediglich auf Ebene der Aufgaben und Aktivitäten zu trennen.

Die Modellierung der Anwendungsarchitektur basiert auf dem Grundsatz der losen Koppelung zwischen dem Archivinformationssystem und Fremdsystemen. Daraus resultiert eine Reihe von Schnittstellen, die einerseits die Integrationsfähigkeit von Archivinformationssystemen in komplexe Systemlandschaften, anderseits die Kompatibilität zwischen verschiedenen Archivinformationssystemen gewährleisten. Grundsätzlich kommt dieses Architekturprinzip auch bei der internen Organisation des Archivinformationssystems zu Anwendung, um eine modulare Weiterentwicklung zu ermöglichen.

Das Archivinformationssystem wird auf diese Weise konzeptionell nicht als eine monolithische Anwendung verstanden, die alle archivischen Geschäftsprozesse abbilden muss, sondern als zentrales Element einer Systemlandschaft, die als ganzes die archivische Arbeit abbildet. Im Zentrum derselben steht das Archivinformationssystem, da es die Verwaltung des Archivguts zur Aufgabe hat. Es bildet die Basis für Systeme, die weitere Geschäftsprozesse abbilden, in denen Archivgut verwendet wird, aber weitere Informationsobjekte massgeblich sind oder hinzukommen.

Ergebnisse

Definition und Abrenzung

Als Archivinformationssystem (AIS) wird ein Informationssystem mit folgenden Eigenschaften verstanden:

- Es bildet zentrale archivische Geschäftsprozesse ab.
- Es stellt die Verwaltung von Archivgut unabhängig von seinem Informationsträger sicher.
- Es implementiert die notwendigen Informationsobjekte und gliedert sie zu funktionalen Einheiten.

— Es besitzt dokumentierte Schnittstellen zwischen den funktionalen Einheiten und zu Fremdsystemen.[11]
— Es ermöglicht eine über viele Jahre hinweg konsistente Datenhaltung.

Archivische Prozesse und Informationsobjekte

Die Gesamtheit der Geschäftsprozesse eines Archivs bildet den Kontext, in dem das Archivinformationssystem realisiert wird. Nicht alle dieser Prozesse werden durch das Archivinformationssystem gesteuert oder bilden sich darin ab. Die nicht enthaltenen Prozesse werden in der Regel durch andere Systeme abgebildet und diese können über Schnittstellen mit dem Archivinformationssystem verbunden sein. Um den Kontext des Archivinformationssystems zu verdeutlichen, wurden daher in einem ersten Schritt alle Geschäftsprozesse modelliert, sofern sie archivspezifisch sind.

Ebenso wurden die wesentlichen Informationsobjekte aller Geschäftsprozesse festgelegt. Auch wenn nicht alle dieser Objekte Teil des Archivinformationssystems sind, war dies erforderlich, um wesentliche Schnittstellen von diesem zu Fremdsystemen definieren zu können.

Die archivischen *Prozesse* werden in neun Prozessgruppen gegliedert, die in ihrer Gesamtheit die archivischen Aufgaben umfassen. In Abhängigkeit ihrer Komplexität werden sie weiter in Teilprozesse untergliedert.[12]

11 Fremdsysteme sind andere in der Organisation genutzte und eventuell über Schnittstellen mit dem AIS verbundene Systeme.

12 Vgl. für weitere Details das KOST-Diskussionspapier AIS-Modell, S. 4f. sowie den Anhang A, Prozessmodellierung.

Abbildung 1: Prozesslandkarte archivische Prozessgruppen

Die Prozessgruppe 1, Steuerung, umfasst Querschnittsprozesse, die in allen anderen Prozessgruppen verwendet werden, welche Fachprozesse abbilden.

Die Prozessgruppe 2, Bewertung, umfasst die Aktivitäten des Archivs im Zusammenhang mit der Überlieferungsbildung und -sicherung.

Die Prozessgruppe 3, Übernahme, schliesst sachlich unmittelbar an die Bewertung an, in zeitlicher Hinsicht jedoch keinesfalls zwingend. Sie umfasst die Aktivitäten im Zusammenhang mit der Ablieferung von Unterlagen an das Archiv respektive die damit verbundenen Abläufe im Archiv.

Die Prozessgruppe 4, Erschliessung, beschreibt alle Aktivitäten, die mit der Ordnung, Strukturierung und Verzeichnung von übernommenem Archivgut zu tun haben.

Die Prozessgruppe 5, Benutzung, umfasst die folgenden Aktivitäten: 5.1. Benutzer bearbeiten, 5.2. Benutzer beraten, 5.3 Anfragen bearbeiten, 5.4 Archivgut bereitstellen, 5.5 Ausleihfristen verwalten, 5.6. Archivgut zurücknehmen und 5.7 Ausleihe abschliessen. Teil der Bearbeitung von Benutzern ist die eigentliche Benutzerverwaltung wie auch die Verwaltung und Zuweisung von Zugangsberechtigungen.

Die Prozessgruppe 6, Magazinverwaltung, gliedert sich aufgrund des Einflusses der Materialität des Archivguts in zwei unabhängige Prozessbereiche für die Verwaltung des analogen und des digitalen Magazins. Sie werden dennoch als eine

Prozessgruppe betrachtet, da die Zielsetzung identisch ist: die sichere Aufbewahrung von Archivgut und die Verwaltung der dafür benötigten (Hilfs-)Mittel.

Die Prozessgruppe 7, Bestandserhaltung, umfasst alle Tätigkeiten, die sich der Erhaltung des Archivguts widmen. Sie gliedert sich ebenfalls aufgrund der Materialität des Archivguts in zwei unabhängige Prozessbereiche für die analoge und digitale Bestandserhaltung.

Aus Prozesssicht unmittelbar an die Prozessgruppe 5, Benutzung, schliesst die Prozessgruppe 8, Reproduktion an. Diese Gruppe beschreibt die Bearbeitung von Reproduktionsaufträgen von Benutzern. Reproduktionsaufträge entstehen immer aus einem Benutzungsfall und stossen einen Digitalisierungsprozess an. Sie können mit Bezahlvorgängen verbunden sein.

Die Prozessgruppe 9, Digitalisierung, beschreibt Abläufe zur Herstellung von Digitalisaten analogen Archivguts. Hierzu gehören Arbeitsschritte in der Planung, der internen oder externen Durchführung, der Auslieferung und der eventuellen Magazinierung der Digitalisate im digitalen Magazin zum Zweck der Nachnutzung.

In den vorgängig definierten Geschäftsprozessen werden zahlreiche *Informationsobjekte* verwendet, die nachstehend aufgeführt sind.

Prozessgruppe 1, Steuerung:

- Auswertung
- Stammdaten
- Grunddaten des Aktenbildners
- Schutzfristkategorie
- Bestätigung
- Report

Prozessgruppe 2, Bewertung:

- Bewertung
- Aktenbildner-Bewertung
- Unterlagen-Bewertung
- Ablieferungsvereinbarung
- GEVER-Dossier

Prozessgruppe 3, Übernahme:

- Angebot
- Ablieferung
- Übernahmeinformationspaket (Submission Information Package SIP)
- Ingest

Prozessgruppe 4, Erschliessung:

- Archivplan
- Verzeichnungseinheit

- Archivinformationspaket (Archival Information Package AIP)
- Findmittel

Prozessgruppe 5, Benutzung:

- Benutzer
- Rolle
- Berechtigung
- Ausleihe
- Auslieferungsinformationspaket (Dissemination Information Package DIP)

Prozessgruppe 6, Magazinverwaltung:

- Standort
- Fläche
- Behältnis

Prozessgruppe 7, Bestandserhaltung:

- Erhaltung
- Monitoring Archivgut
- Monitoring Technologie
- Erhaltungsplanung
- Restaurierungsbericht
- Notfallplan

Prozessgruppe 8, Reproduktion:

- Reproduktionsauftrag
- Rechnung
- Lieferung

Prozessgruppe 9, Digitalisierung:

- Digitalisat

AIS-unterstützte Prozesse und Objekte

Für das AIS-Modell wurden 9 *Prozessgruppen* definiert. Damit wird das archivische Arbeitsfeld abgesteckt, innerhalb dessen das AIS sich befindet. Das Archivinformationssystem deckt jedoch nur diejenigen Prozesse ab, in denen das Archivgut selber im Zentrum steht. Es sind dies die Geschäftsprozesse Bewertung, Übernahme und Erschliessung. Nur für diese erfolgt in diesem Kapitel eine Vertiefung der Prozessmodellierung auf Ebene Teilprozess und teilweise Prozessschritt sowie die Spezifikation der erforderlichen Informationsobjekte inklusive ihrer Eigenschaften und Methode. Es handelt sich also hier um eine starke Reduktion gegenüber der Darstellung aller archivischen Prozesse und Informationsobjekte. Der Funktionsumfang des Archivinformationssystems entspricht weitgehend der Funktionalen Einheit Daten-

verwaltung (*Data Management*) des OAIS-Modells, vertieft und erweitert diese jedoch.[13]

Aus Sicht des konzeptionellen Modells besteht die Prozessgruppe *Bewertung* aus den drei Prozessen *Aktenbildner bewerten* (2.1), *Unterlagen bewerten* (2.2) und *Ablieferungsvereinbarung abschliessen* (2.3) Die Bewertung deckt die Aufgabe der Ermittlung des archivischen Wertgehalts von Registraturgut ab. Es ist auch möglich, dass Bewertung erst nach der Übernahme und Erschliessung als Nachbewertung stattfindet.

Die Prozessgruppe *Übernahme* besteht aus den Prozessen *Angebot prüfen* (3.1), und wahlweise *analoges Archivgut übernehmen* (3.2) oder *digitales Archivgut übernehmen* (3.3). Während bei der Bewertung die Materialität der Unterlagen keinen Einfluss auf den Ablauf hat, weicht die Übernahme analoger Unterlagen von derjenigen digitaler Unterlagen ab. Daher müssen beide Fälle getrennt behandelt werden.

Die Prozessgruppe *Erschliessung* setzt sich aus den Prozessen *Erschliessung planen* (4.1), *analoges Archivgut erschliessen* (4.2) respektive *digitales Archivgut erschliessen* (4.3), *Abschlussarbeiten durchführen* (4.4) sowie *Bestand freigeben* (4.5) zusammen. Erschliessung kann auch zu einem späteren Zeitpunkt als Nacherschliessung erfolgen. Auch bei der Erschliessung ist ein Unterschied zwischen analogen und digitalen Materialien zu machen. Die Abweichungen betreffen die Ordnung von Archivgut, die bei digitalem Archivgut in der Regel nicht respektive deutlich anders als bei analogem Archivgut erfolgt, und die Magazinierung respektive Speicherung im digitalen Magazin.

Das AIS-Modell konzentriert sich, analog zum Vorgehen bei den Prozessen, auf die Spezifikation der genuinen *AIS-Objektklassen*. Das bedeutet wiederum eine Reduktion der Gesamtheit aller Objekte auf diejenigen, die gemäss einer Minimalspezifikation nicht an Drittsysteme delegiert werden können. Das AIS-Referenzmodell erlaubt es, weitere Objektklassen in einem AIS abzubilden, solange diese Minimalspezifikation erfüllt wird. In der Folge wird der Einfachheit halber auf die Darstellung der Fremdsysteme verzichtet und nur die AIS-Innensicht gezeigt.

13 Zum OAIS-Referenzmodell respektive ISO 14721:2012 siehe https://public.ccsds.org/Pubs/650x0m2.pdf oder https://www.iso.org/standard/57284.html.

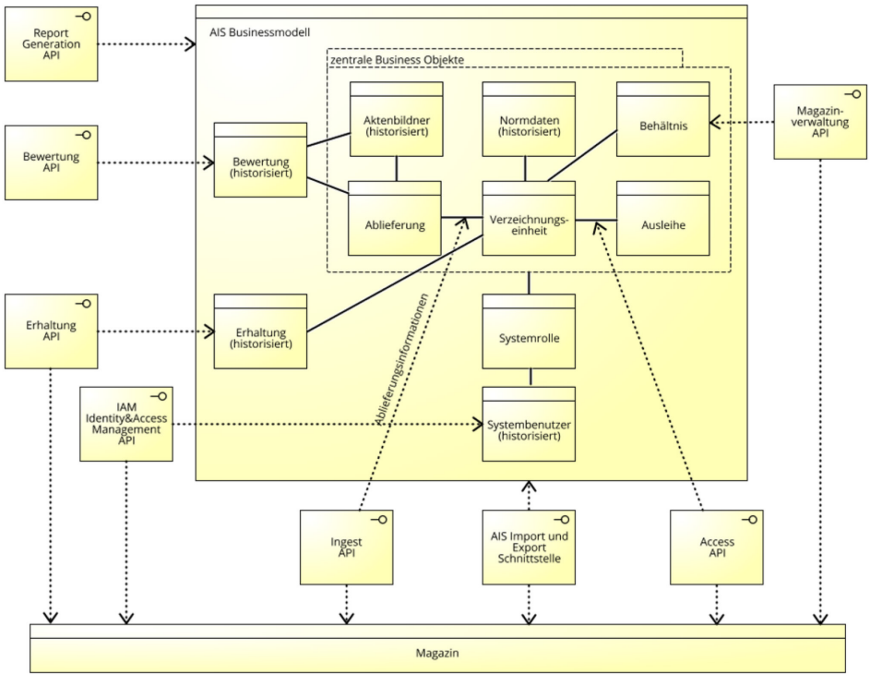

Abbildung 2: AIS-Innensicht

Folgende Objekte werden für das Archivinformationssystem aus den Geschäftsprozessen heraus abgeleitet:

- Ablieferung
- Aktenbildner
- Normdaten
- Verzeichnungseinheit

Einige der nicht durch das AIS abgedeckten Prozesse werden jedoch in diesem durch Informationsobjekte referenziert, die von entsprechenden Schnittstellen zu Fremdsystemen angesprochen werden. Es sind dies Behältnis, Ausleihe, Bewertung und Erhaltung. Diese Objekte werden als Nachweisobjekte bezeichnet, da mit ihnen kein Prozess gesteuert wird, sondern weil sie dokumentierte Information beinhalten, die die Ergebnisse von Prozessabwicklungen in anderen Systemen auf Dauer im AIS verfügbar machen.

Hinzu kommen die Objektklassen Systemrolle und Systembenutzer, die für den Zugriff auf das System benötigt werden.

Die Objektklassen kapseln alle notwendigen Eigenschaften der Informationsobjekte und besitzen Methoden,[14] um lesend und schreibend darauf zuzugreifen. Die

14 Die hier bezeichneten Methoden beziehen sich auf die Innensicht des AIS.

Eigenschaften der Objekte orientieren sich an den einschlägigen Standards zur archivischen Erschliessung des ICA: ISAD(G), ISAAR(CPF) und ISDF.[15] Diese werden jedoch erweitert und um andere Standards ergänzt.[16]

Schnittstellen

Neben der vertieften Prozessmodellierung und der Spezifikation der Informationsobjekte ist die Definition der Schnittstellen des Archivinformationssystems zentral, wenn dieses im Zentrum einer archivischen Systemlandschaft stehen und zugleich als offenes Informationssystem realisiert werden soll.

Die Schnittstellen definieren, wie und auf welche Art ein Archivinformationssystem mit Fremdsystemen agieren kann.[17] Dabei sind zwei Arten zu unterscheiden. Im ersten Fall nutzt das AIS die Funktionalität eines anderen Systems für seine eigenen Zwecke, stellt diese Funktionalität also nicht selber zur Verfügung. Das ist beispielsweise bei der Benutzerverwaltung der Fall. Hier muss eine standardisierte Fremdschnittstelle vom Archivinformationssystem implementiert werden (im Beispiel eine IAM Application wie LDAP oder AD). Im zweiten Fall stellt das AIS einen Teil seiner inneren Funktionalität und Daten (das heisst, Methoden zum Lesen und Schreiben der Datenstruktur eines Teils seiner Objektklassen) einem Fremdsystem zur Verfügung. Diese zweite Art von Schnittstellen kann wiederum in zwei Fälle unterteilt werden. Im ersten Fall funktioniert das Archivinformationssystem als Informationsquelle für ein Fremdsystem, z.B. für einen Digitalen Lesesaal oder eine webbasierte Metasuche. Der Zugriff erfolgt hier nur lesend. Im zweiten Fall beliefert ein Fremdsystem das Archivinformationssystem mit neuen Informationen; der Zugriff erfolgt hier lesend und schreibend. Die Interaktion mit dem AIS kann dabei Hauptzweck des Fremdsystems sein oder auch nur Teil eines grösseren Prozesses. Beispielsweise liefert ein Ingestprozess neben dem Umwandeln eines SIP in ein AIP auch wichtige Informationen an das Archivinformationssystem.

15 ISAD(G) siehe http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition. ISAAR(CPF) siehe http://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd. ISDF siehe http://www.ica.org/en/isdf-international-standard-describing-functions.

16 Der vom ICA im September 2016 zur Kommentierung freigegebene Entwurf des neuen Erschliessungsstandards «Records in Context» wird in dem Referenzmodell nicht berücksichtigt, da es sich noch nicht um einen definitiven Standard handelt. Seine Verwendung erscheint jedoch als attraktiv, da damit die Inkonsistenzen zwischen den bestehenden ICA-Standards beseitigt werden und neuere Entwicklungen wie Linked Data und die Nutzung von Ontologien von «Records in Context» unterstützt werden. Zu «Records in Context» siehe http://www.ica.org/en/egad-ric-conceptual-model.

17 Ein Fremdsystem ist z.B. ein Ingest- oder ein PreIngest-Tool oder eine Applikation «Digitaler Lesesaal».

Abbildung 3 zeigt das Archivinformationssystem eingebettet in eine Menge von Fremdsystemen.[18]

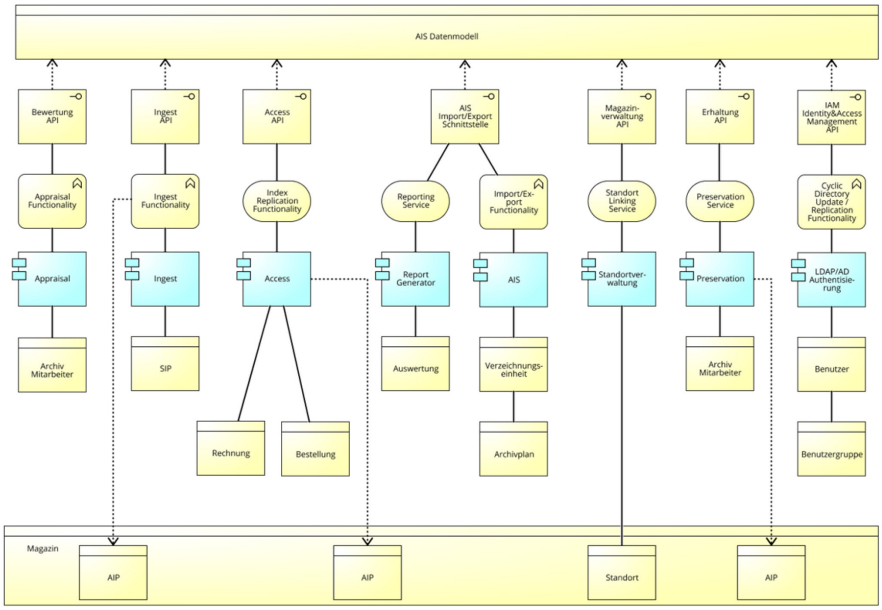

Abbildung. 3: AIS-Aussensicht

In dieser Darstellung lassen sich die notwendigen Schnittstellen erkennen, die nach der zugrundeliegenden Funktionalität benannt sind: *Ingest*, *Access*, *Bewertung*, *Erhaltung*, *AIS Import/Export*, *IAM*, *Magazinverwaltung* und *Report*.

An den Schnittstellen wird auf die Objekte des Archivinformationssystems durch Fremdsysteme mit entsprechenden Methoden *lesend*, *suchend* (in einer Erweiterung des lesenden Zugriffs) oder *schreibend* zugegriffen. Diese Unterscheidung der Zugriffsart erlaubt eine Systematisierung der Schnittstellen in einer CRUD-Tabelle (*Create, Retrieve or Read, Update and Delete*).

	Aktenbildner				Ablieferung				Normdaten				VE				Ausleihe				Behältnis				Rolle				AIS-Benutzer			
	provenance				accession				authority record				unit of description				loan				container				systemRole				systemUser			
	C	R	U	D	C	R	U	D	C	R	U	D	C	R	U	D	C	R	U	D	C	R	U	D	C	R	U	D	C	R	U	D
Ingest	-	-	-	-	-	X	X	-	-	-	-	-	X	X	(x)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Access	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-	X	X	X	-	-	X	-	-	-	-	-	-	-	-	-	-
Magazinverwaltung	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	X	-	-	-	-	-	-	-	-	-
IAM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	-
Report Generation	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-	-	X	-	-
Bewertung	X	X	X	X	X	X	X	[X]	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Erhaltung	-	-	-	-	-	-	-	-	-	-	-	-	X	X	X	[X]	-	-	-	-	X	X	X	[X]	-	-	-	-	-	-	-	-
AIS [Import]/Export	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-	[X]	X	-	-

Abbildung 4: CRUD-Tabelle

18 Die beiden Darstellungen der Aussensicht und der Innensicht sind als Archimate-Diagramme gezeichnet.

Eine detaillierte Beschreibung der Schnittstellen erfolgt in der UML-Notation. Sie legt die Attributnamen und Attributtypen der von aussen sichtbaren AIS-Objekte fest. Desgleichen werden die Methoden und ihre Parameter definiert.

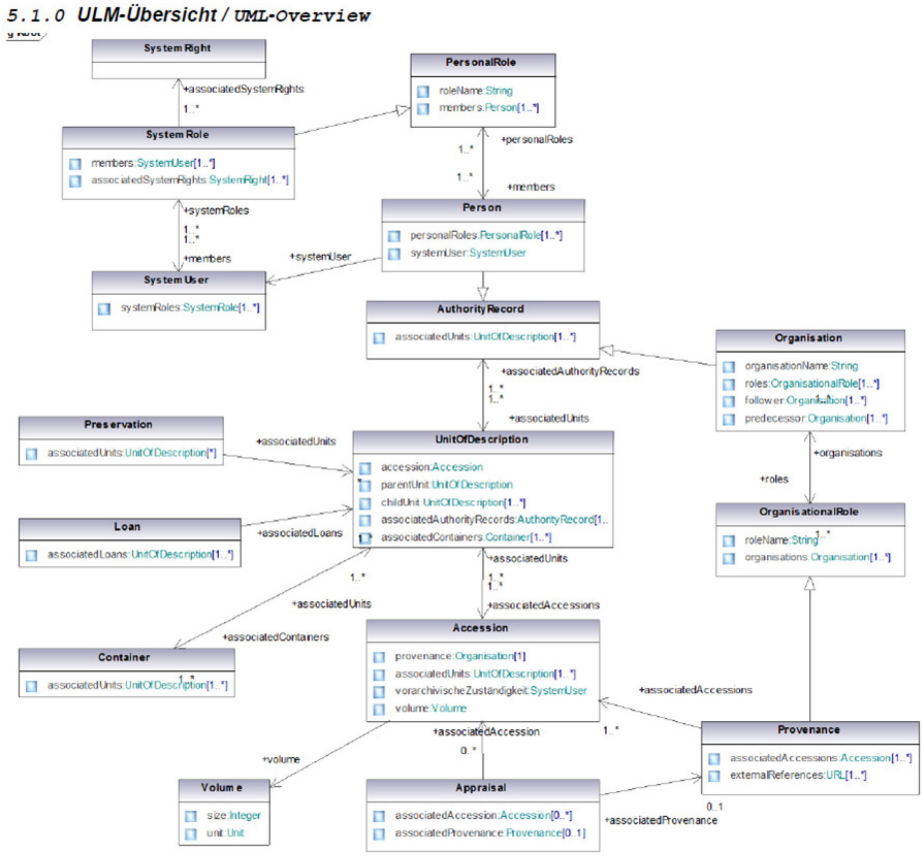

Abbildung 5: UML-Diagramme im Überblick

Anforderungskatalog

Die aus Sicht des konzeptionellen Modells erforderlichen Anforderungen an eine Umsetzung werden in einem Anforderungskatalog abgebildet. Er gliedert sich in folgende Bereiche:

- funktionale Anforderungen,
- Anforderungen an die Lösungsarchitektur,
- technische Anforderungen und
- Anforderungen an Datenschutz und Informationssicherheit.

Im Anforderungskatalog sind die von einer Anforderung betroffenen Prozesse und Informationsobjekte ersichtlich. Zudem werden wesentliche Rahmenbedingungen

gesetzt, die bei einer Lösung zu berücksichtigen sind. Er dient überdies als Hilfsmittel bei der Einführung neuer Archivinformationssysteme oder der einheitlichen Beurteilung bestehender Systeme.

Fazit und Ausblick

Mit dem vorliegenden Ergebnis des KOST-Projekts *AIS-Referenzmodell* konnte gezeigt werden, dass die Entwicklung eines generischen Archivinformationsmodells ausgehend von den archivischen Geschäftsprozessen und einer Reihe von Prämissen für die Architekturentwicklung des Archivinformationssystems möglich ist.

Somit liegen Prozessmodellierungen und Objekt(klassen)-Definitionen ebenso vor wie Spezifikation der Schnittstellen zwischen dem Archivinformationssystem und relevanten Fremdsystemen. Diese Ergebnisse sollten in einem nächsten Schritt einer kritischen Betrachtung unterzogen werden.

Bereits während der Projektarbeit und insbesondere als Resultat der Review konnten folgende Themenbereiche identifiziert werden, die vertieft betrachtet werden sollten:

- Ist die Entscheidung, eine Reihe von archivischen Prozessen nicht in das AIS aufzunehmen, sinnvoll, oder welche Änderungen der Abdeckung der archivischen Prozesse durch das AIS müssten erfolgen?
- Sind die archivischen Prozesse grundsätzlich vollständig und korrekt modelliert worden? Welche Ergänzungen sind erforderlich?
- Welche (Metadaten-)Standards werden bei der Definition der Objekte und ihrer Eigenschaften sowie der Schnittstellen verwendet?
- Wie kann die Weiterentwicklung der ICA-Standards mit dem neuen Standard *Records in Context* aufgenommen werden?
- Wird mit der getroffenen Architekturentscheidung das Ziel erhöhter Interoperabilität und architektonischer Offenheit nachhaltig erreicht?

Der einzige Kompass, den wir haben. Zur Kritik der Designated Community

Christian Keitel

Welche Rolle spielen die künftigen Nutzer im Denken der Archivarinnen und Archivare? Sind sie der einzige Zielpunkt aller archivischen Bemühungen oder ist diese Debatte vielmehr eine Scheindebatte? Deutet der vom OAIS-Standard in die Welt gesetzte Begriff der Designated Community vielleicht sogar auf einige blinde Flecken im Umfeld der bisherigen archivwissenschaftlichen Theoriebildung? Oder vernebelt er vielmehr das, was bisher klar und eindeutig war?

Designated Community

Sehen wir zunächst auf die Entstehung des Begriffs. Der Begriff der «Designated community» wurde im Rahmen des OAIS-Standards entwickelt.[1] In der allerersten, von Don Sawyer und Lou Reich verfassten und im September 1994 vorgestellten Version des späteren Standards wurde der Begriff zwar noch nicht explizit genannt, aber doch schon deutlich erkennbar umrissen:

> *«Archived Information: Information, represented by digital data, that is being preserved for* public access *over the long (indefinite) term. The information is deemed to be understandable* to one or more segments of the public.*»*[2]

Bereits auf dem ersten zum Entwurf abgehaltenen US-Workshop im Oktober 1995 wurden diese Annahmen diskutiert:

> *«There were questions of how much access is required - should be for a designated community».* Und: *«Perhaps change our archive definition to: a repository that intends to preserve information for use by a designated community.»*[3]

In der 2. Rohfassung von OAIS wurde die allgemeine Öffentlichkeit daher durch eine oder mehrere Designated Communities ersetzt.[4] Von nun an sollte sich das übrigens schon in der ersten Version angelegte Spannungsverhältnis zwischen der Definition des Archivs und derjenigen der Designated Community durch die ganze

1 Lee, Christopher A.: Defining Digital Preservation Work: A Case Study of the Development of the Reference Model for an Open Archival Information System, 2005, S. 158.
2 Digital-Archiving Information Services Reference Model, 14.9.1995, Auszeichnungen vom Autor.
3 ISO Archiving Standards - First US Workshop – Minutes vom 11. und 12.10.1995.
4 Reference Model for Digital Archiving Standards vom 19.12.1995.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.4

Versionsgeschichte von OAIS hindurch erhalten. Auf der einen Seite wird das Archiv ganz allgemein und ohne weitere Konkretisierung durch seinen Bezug auf die Designated Community definiert. Auf der anderen Seite wird die Designated Community dadurch bestimmt, dass sie auf die Verstehbarkeit der digitalen Objekte abzielt. Der Terminus stellt daher sowohl einen allgemeinen Fluchtpunkt aller Bestrebungen als auch das konkrete Ziel einer spezifischen Teilaufgabe dar.

Der Standardisierungsprozess setzte sich über insgesamt 20 nationale US-Workshops und 13 internationale Workshops bis Ende 2001 fort.[5] Das Protokoll zum 4. US-Workshop verrät uns, dass der Begriff ganz bewusst gewählt wurde, um eine noch breiter zu fassende Nutzergruppe einzugrenzen.[6]

Während des weiteren Standardisierungsprozesses erscheint manchmal der Terminus «Designated User Community»,[7] manchmal können wir auch von «Designated Consumer Community»[8] lesen, letzten Ende wurden diese Varianten aber verworfen. Ein anderer Ansatz war das erstmals im dritten internationalen Workshop im November 1996 erschienene «Designated Information Object". Auch dieser Begriff hielt sich nicht lange und mutierte schließlich zu *Content Information*.

Ein längeres Leben war dem Begriff der «Knowledge Base» vergönnt. Erstmals genannt wird dieser Begriff in der zweiten Version des White Book im Oktober 1997.[9] Seitdem ist sie ein fester Bestandteil des Standards. Ausgangspunkt war die Representation Information, die die eigentlich zu archivierende Content Information erst aufrufbar und verstehbar macht. Darüber hinaus wird manchmal noch weitere Representation Information benötigt, um die zunächst benannte Representation Information zu erklären. Auch für diese Representation Information könnten weitere Angaben notwendig werden. In der Theorie kann sich deshalb ein endloser Zirkel von Metadaten aufbauen, die benötigt werden, um andere Metadaten zu beschreiben, die wiederum andere Metadaten beschreiben und so fort. Die Kenntnis künftiger

5 Entsprechende Übersichten finden sich im Internetarchiv (https://web.archive.org/web/20061009230759/http://nssdc.gsfc.nasa.gov/nost/isoas/overview.html und https://web.archive.org/web/20061009231724/http://nssdc.gsfc.nasa.gov/nost/isoas/us/past_workshops.html) sowie bei Lee, OAIS (s. Anm. 1). (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

6 4th US Workshop, 10-11.7.1996: «Lou regarding 2.2.6 add terms regarding making material available to designated community. We should be careful not to make all this applicable to a broader community.»

7 Erstmalige Nennung im Reference Model for Digital Archiving Standards Version 2 vom 19.12.1995, der Begriff verbleibt in den Entwürfen bis zu CCSDS 650.0-W-1.2 White Book.

8 Der Begriff wird im Draft red book genannt. Ein Teil blieb bis zur ersten Version des ISO-Standards als Kapitelüberschnitt «3.2.3 Determines Designated Consumer Community» erhalten.

9 CCSDS 650.0-W-1.1.

Nutzer und damit deren Knowledge Base begrenzen diesen potentiell endlosen Zirkel.[10]

Diese wenigen Bemerkungen können das weite Feld der Debatten nur andeuten, die während der Erarbeitung des Standards geführt wurden. Besonders eindrücklich werden diese Debatten im Oktober 2000 in Review-Kommentaren zusammengefasst.[11] Der International Council for Scientific and Technical Information (ICSTI) fragte an, ob anstelle von Designated Community nicht besser von Consumer Community oder Primary Consumer Community die Rede sein sollte. Die Antwort der Autoren: «The reason for using 'designated' is to make clear that an active identification of the relevant Consumer Community is to be made.»[12] In anderen Worten sollte der Archivar nicht nur vage über die Designated Community nachdenken, sondern sich explizit auf sie beziehen.

Ausgehend von der Definition der Designated Community[13] kritisierten auch verschiedene Reviewer, dass die Verstehbarkeit nicht mehr in der Zuständigkeit des Archivs liege und der bloße Zugang alles sei, was vom Archiv erwartet werden könne. Überhaupt, merkten die Reviewer an, was bedeute «unabhängig verstehbar» («independently understandable»)? Die US-Arbeitsgruppe antwortete, dass mit «unabhängig» der Bezug zum Urheber der Archivalien gemeint sei und fährt dann fort: «This is a traditional function of an archive.»[14]

Archivwissenschaftliche Positionen

Tatsächlich dürfte es schwer fallen, ein der Designated Community vergleichbares Konzept bei den klassischen Autoren der Archivwissenschaft zu finden. Die 1898 erschienene niederländische *Handleiding voor het ordenen en beschrijven van archieven* beginnt mit folgender Archiv-Definition: «Ein Archiv ist die Gesamtheit der geschriebenen, gezeichneten und gedruckten Dokumente, in dienstlicher Eigenschaft von irgend einer Behörde oder einem ihrer Beamten empfangen oder ausgefertigt, wofern diese Dokumente bei der Behörde oder deren Beamten bestimmungsgemäß

10 Der Begriff erscheint erstmals in CCSDS 650.0-W-2.0 WHITE BOOK, October 15, 1997. In der vorangehenden Version wird er noch nicht genannt (CCSDS 650.0-W-1 WHITE BOOK, April 10, 1997).

11 OAIS Critiques/RIDS 25 October 2000.

12 OAIS Critiques (wie Anm. 11). Allerdings stimmte die Arbeitsgruppe zu, dass die Designated Community bei Zeitschriften am ehesten mit dem für die Erschließung von Zeitschriften verwendeten Begriff der «primary audience» gleichgesetzt werden könne.

13 «Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities.» OAIS Critiques (wie Anm. 11).

14 OAIS Critiques (wie Anm. 11).

verbleiben sollen.»[15] Die Definition war sowohl von der Versammlung der niederländischen Archivare als auch der Reichsarchivare 1890 bzw. 1891 einstimmig angenommen worden. Ihr Schwerpunkt liegt auf der Herkunft und Verwahrung der Unterlagen. Nutzer kommen in ihr – im Gegensatz zur OAIS-Definition eines Archivs – nicht vor.

Eine in vielen Punkten vergleichbare Archiv-Definition stellte Sir Hilary Jenkinson an den Anfang seines *Manual of Archival Administration* von 1922, um dann zu ergänzen: «To this Definition we may add a corollary. Archives were not drawn up in the interest or for the information of Posterity."[16] Überhaupt stehe die Erhaltung der Archivalien unter den Aufgaben des Archivars an erster Stelle, während die Erfüllung der Nutzer-Bedürfnisse nur sekundär zu berücksichtigen sei. Diese Prioritätensetzung sei auf gar keinen Fall zu verändern.[17] Die Begründer des deutschsprachigen Fachs argumentierten ähnlich.

Aus der Rückschau erscheint es kaum als Zufall, dass der Begriff der Designated Community eben nicht von klassischen Archivaren und Archivwissenschaftlern erfunden wurde. Zu sehr waren die Diskurse fixiert auf Fragen der Abgrenzung (z.B. von den Historikern, den Bibliothekaren etc.), zu wichtig erschien das Ziel der eigenen Neutralität. Es kann daher nicht verwundern, dass von Seiten der klassisch geprägten Archivarinnen und Archivare Kritik am Konzept der Designated Community geübt wurde. Die Kritik lässt sich in drei Punkten zusammenfassen:

1.) Traditionelle Archive stehen allen Teilen der Bevölkerung offen. Das Konzept der Designated community scheint diesen Ansatz zu sehr zu verengen.

Der Terminus der Designated Community war wie oben beschrieben von den OAIS-Autoren bewusst eingesetzt worden, um die allgemeine Öffentlichkeit näher einzugrenzen. Zugleich verweist der Standard aber auch darauf, dass eine Designated Community aus zahlreichen User Communities bestehen kann. Damit eröffnet sich eine standardkonforme Möglichkeit, über die Definition verschiedener Unter-Gruppen auch die allgemeine Öffentlichkeit näher zu bestimmen.

15 Zitiert nach der deutschen Übersetzung, Samuel Muller, Johan Adriaan Feith, Robert Fruin, Anleitung zum ordnen und beschreiben von Archiven, Leipzig 1905, hier S. 1.

16 Hilary Jenkinson : A Manual Of Archive Administrations. - Oxford, 1922, S. 11.

17 Jenkinson (wie Anm. 16), S. 15: «It is not his business to deal with questions of policy to decide whether twenty thousand pounds, or one thousand or nothing should be spent on printing transcripts of his Archives; whether the student would be best served by having the Archives in a Metropolis, or in the Provinces; at what date modern 'confidential' Archives should be thrown open to the public. He will doubtless take an intelligent interest in such subjects, but as an Archivist he is not concerned with them: they are questions for Historians, Politicians, Administrators; whom, at most, he may advise."

2.) Designated community scheint nur ein abstraktes Konzept ohne Relevanz für die konkret im Archiv anstehenden Aufgaben zu sein.

Auf den ersten Blick scheinen verschiedene Punkte im Standard und in den verwandten Standards für diese Annahme zu sprechen scheinen. Selbst OAIS konkretisiert den Begriff kaum. Stattdessen lernen wir, dass der Aspekt der künftigen Verstehbarkeit sowohl bei den SIPs (Übergabepaketen) als auch bei den AIPs (Archivierungspaketen) geprüft werden sollte.[18] Da sich die Knowledge Base der Designated Community ändern könne, solle sie überwacht werden. In diesen Punkten wird die allgemeine Idee auf verschiedene Punkte übertragen, ohne dass das anfängliche Abstraktionsniveau konkreter werden würde.

Etwas anders sieht es dann im Teil 6.1. aus. Hier unterscheidet der Standard verschiedene Archivtypen und im Wesentlichen zwei Typen von Designated Communities. Während ein sogenanntes unabhängiges Archiv nur eine lokale Designated Community besitze, habe ein Verbundsarchiv (federated archive) sowohl eine lokale als auch eine globale Designated Community. In anderen Worten werden hier die verschiedenen Zielgruppen durch ihre Zugehörigkeit zu geographischen Regionen unterschieden. Diese vom Standard selbst vorgenommene Interpretation des Begriffs überzeugt nicht, da sich die denkbaren Nutzerinteressen kaum nach geographischen Kriterien unterscheiden lassen.

Aber auch bei den anderen Standards der OAIS-Familie wird das Bild nicht wirklich klarer. PAIMAS schlägt eine Validierung der Content Information durch Repräsentanten der Designated Community vor.[19] Designated Community und heutige Nutzer werden so weitgehend miteinander gleichgesetzt. An anderer Stelle wird die Designated Community in den Bereich der sogenannten Producer[20] gerückt. Diese seien in der Lage, die Erwartungen der Designated Community zu bestimmen.[21] Einige Abschnitte später wird eben diese Möglichkeit im selben Standard in Frage gestellt.[22] Die Einschätzung der Nutzerinteressen durch die Producer wird schwierig, wenn wir an Schellenbergs Unterscheidung von Primär- und Sekundärzweck denken.[23]

18 Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012, S. 4-13.

19 Producer-Archive Interface Methodology Abstract Standard (PAIMAS), CCSDS 651x0m1, May 2004, S. 3-26.

20 Producer erstellen die später zu archivierende Information. In der Welt der klassischen Archive können sie mit den abgebenden Stellen gleichgesetzt werden.

21 PAIMAS (wie Anm. 19), S. 3-3 (P2).

22 PAIMAS (wie Anm. 19), S. 3-6: «However, it should be noted that for some institutional and/or governmental Archives neither the Producer nor the Archive has a precise idea of how the information to be preserved will be used. Even with scientific observation Archives, 10 years after data production, scientific data is used in ways that the Producers could not even imagine.»

23 Schellenberg, Theodore R.: Modern Archives: Principles and Techniques. Chicago 1956.

Auch die ISO 16363 verbindet an verschiedenen Punkten den archivischen Entscheidungsprozess mit dem Konzept der Designated Community. Zwar sind die Angaben zumeist sehr allgemein formuliert, was der Funktion der Norm als Zertifizierungsstandard geschuldet sein dürfte. Zwei Formulierungen sind allerdings bemerkenswert: «The preservation policy might then include information about the expected level of comprehensiveness by the repository's Designated Community for each Archival Information Package.»[24] Und: «The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest.»[25] Offenbar können die Anforderungen der Designated Community unterschiedlich und in verschiedenem Umfang erfüllt werden. Hier deutet sich eine erste Perspektive für die Nutzbarmachung des Konzepts der Designated Community an.

Bereits bestehende Archive haben das Konzept vor allem auf zweierlei Wegen rezipiert. Die eine Gruppe nennt zwar den abstrakten Term, ignoriert ihn aber bei der Ausgestaltung der praktischen Entscheidungsprozesse. Die Archive dieser Gruppe sehen zumeist die interessierte Öffentlichkeit als potentielle Nutzer an. Die andere Gruppe besitzt dagegen eine sehr kleine und überschaubare Designated Community. Diese Archive beschreiben das Konzept und die Knowledge Base häufig auf eine eher konkrete Art und Weise.[26] Auch in der unmittelbar auf OAIS bezogenen Fachdiskussion ergeben sich keine wirklich überzeugenden Fortentwicklungen. Entweder werden heutige Nutzer und Designated Community weitgehend in eins gesetzt[27] oder sollen anstelle der Designated Community die Urheber der Unterlagen befragt werden.[28]

Letzen Endes schweigen sich die Standards der OAIS-Familie ebenso wie ihre Rezeption weitgehend darüber aus, wie denn das Konzept der Designated Community in die Praxis umgesetzt werden könnte. Nur in der Rezeption durch kleinere Archive mit einer klar umrissenen Gruppe künftiger Nutzer können wir Hinweise zur praktischen Umsetzung finden. Es gibt daher gute Gründe, dieser Frage weiter unten im Rahmen konkreter Anwendungsbeispiele noch einmal intensiver nachzugehen.

24 Audit and Certification of Trustworthy Digital Repositories, CCSDS 652.0-M-1, Magenta Book 2011, S. 3-7.

25 Audit and Certification (wie Fußnote 24), S. 4-23.

26 Z.B. Shaon, Arif u.a.: Long-Term Sustainability of Spatial Data Infrastructures: A Metadata Framework and Principles of Geo-Archiving, in: iPRES 2011 – 8th International Conference on Preservation of Digital Objects, Singapore 2011, S. 120-129.

27 Vgl. Kärberg, Tarvo: Digital Preservation and knowledge in the public archives: for whom?, in: Archives and Records 35 Nr. 2 (2014), S. 126-143. Der Autor versucht, über automatisierte Verfahren das Verhalten heutiger Nutzer zu ermitteln und davon Erkenntnisse für die Ermittlung der Designated Community zu ziehen.

28 Vgl. Bischoff, Frank M.: Bewertung elektronischer Unterlagen und die Auswirkungen archivarischer Eingriffe auf die Typologie zukünftiger Quellen, in: Archivar 67 (2014), S. 40-52.

3.) Designated Community kann nicht eindeutig bestimmt werden.

Wenn wir über die mutmaßliche Designated Community unseres Archivs reden, reden wir über die Zukunft. Leider ist alles Reden über die Zukunft spekulativ. Spekulationen sollten nun in den Augen vieler Archivarinnen und Archivare so gut es geht vermieden werden. Hinzu kommt, dass diese Spekulationen jeweils von einzelnen Menschen vorgenommen werden, sie also notwendigerweise subjektiv eingefärbt sind. Benjamin Bussmann kommt daher in seiner Masterarbeit zum Ergebnis, dass Designated Communities aufgrund ihrer spekulativen und subjektiven Bestimmung nicht eindeutig bestimmt werden können und deshalb aus den Konzeptionen digitaler Archive und Archivierung entfernt werden sollten.[29]

OAIS selbst widerspricht dieser Ansicht nur teilweise: «The degree to which Content Information and its associated PDI conveys information to a Designated Community is, in general, quite subjective. Nevertheless, it is essential that an Archive make this determination in order to maximize information preservation.»[30]

Weshalb insistieren die Autoren von OAIS so beharrlich auf dem Konzept der Designated Community? Vielleicht sollten wir zur Beantwortung dieser Frage zwischen unseren Möglichkeiten zur Bestimmung der Designated Community und deren Funktion im archivischen Entscheidungsfindungsprozess unterscheiden. Beschreibungen von Designated Communities müssen notwendigerweise von Menschen vorgenommen werden; auf dieselbe Art werden Entscheidungen stets von Menschen vorgenommen. Sowohl die Beschreibung der künftigen Nutzer als auch die auf dieser Basis getroffenen Entscheidungen tragen daher unvermeidlich subjektive Anteile. Sobald wir aber solche Entscheidungen akzeptieren, sollten wir das Motiv der Subjektivität auch nicht gegen die Designated Community wenden. Wir können es schlicht nicht vermeiden, subjektiv gefärbte Entscheidungen zu treffen.[31]

Allerdings gibt es unterschiedliche Wege, trotz aller Rahmenbedingungen zu den angestrebten Zielen zu kommen. Dabei erscheint es sinnvoll, sich über die Art der Spekulationen zu unterhalten. Es macht einen Unterschied, ob wir Aussagen darüber treffen, was wir heute Abend essen oder darüber, was wir in drei Jahren am Tag X zu Abend essen werden. Heute Abend dürften es zwei Käsebrote sein, aber ob uns das auch in drei Jahren zusagen wird? Es ist daher möglich, ungeachtet aller Subjektivität und Spekulationen Aussagen über die Zukunft zu treffen, die uns wahrscheinlicher erscheinen als andere Aussagen über die Zukunft.

29 Bussmann, Benjamin: Die Bestandserhaltung digitaler Informationen mittels der Definition von signifikanten Eigenschaften. Masterarbeit im berufsbegleitenden Fernstudiengang Archivwissenschaft an der Fachhochschule Potsdam. Düsseldorf 2015. S. 99.

30 OAIS (wie Anm. 18), S. 3-4.

31 Vgl. Keitel, Christian: Prozessgeborene Unterlagen. Anmerkungen zur Bildung, Wahrnehmung, Bewertung und Nutzung digitaler Überlieferung, in: Archivar 67 (2014), H. 3, S. 278-285.

Die am wenigsten spekulative Aussage, die wir über unsere Archive treffen können, ist die Annahme, dass sie in Zukunft Nutzer haben werden. Wenn wir diese Annahme ernsthaft in Zweifel ziehen würden, wären wir eigentlich nicht in der Lage, unsere Archive mit der bisher gekannten Selbstverständlichkeit weiter zu betreiben. Erst das Konzept der Designated Community macht diese Annahme aber explizit. Erst jetzt ist es möglich, diese bislang implizite Annahme auch zu kritisieren und praktische Lösungskonzepte zu entwickeln.

Heutige und künftige Nutzer

Die Autoren von OAIS grenzten die Designated Commmunity deutlich von den Wünschen der heutigen Nutzer ab. Die heutigen Benutzerwünsche können nicht 1:1 zum Ausgangspunkt für die archivische Bewertung herangezogen werden. Auf der anderen Seite sehen die heutigen Nutzer mit einem vielleicht unverstellten Blick auf die Archive, den die Archivarinnen und Archivare selbst gar nicht haben können. Außerdem verfügen sie über Nutzungserfahrungen, die dem archivischen Stammpersonal in aller Regel ebenfalls abgehen. In gesellschaftspolitischer Hinsicht können wir über derartige pragmatische Überlegungen hinaus auch fragen, ob es nicht im Interesse der einzelnen gesellschaftlichen Gruppen liegen müsste, dass ihr Handeln und ihre Erfahrungen auch angemessen an die zukünftigen Nutzer überliefert werden. Öffentlich-rechtliche Archive, die ihre Legitimation über die Parlamente oder von Stadt- oder Kreisversammlungen erhalten, könnten sich daher in der Pflicht sehen, diesen Wünschen auch nachzukommen. Hans Booms hat bereits 1972 die Frage gestellt, ob Bewertung nicht eine gesamtgesellschaftliche Aufgabe sei und entsprechend dazu die Bewertungsentscheidungen auch von den einzelnen gesellschaftlichen Gruppen vorbereitet, wenn nicht sogar getroffen werden müsste.[32] Diese zunächst nur im deutschen Sprachraum sich entwickelnde Diskussion lebte Ende der 1980er Jahre noch einmal auf, nachdem der Artikel in der kanadischen Fachzeitschrift Archivaria auf Englisch erschienen war.[33] Es gibt daher sowohl archivpolitische wie auch ganz pragmatische Gründe, weshalb wir beim Sinnieren über die Designated Community auch die Interessen der heutigen Nutzer im Blick haben sollten. Zugleich ermöglicht es uns die von OAIS vorgenommene Unterscheidung der heutigen und der angenommenen künftigen Nutzer, die heutigen Nutzungssituationen zu transzendieren.

32 Booms, Hans: Gesellschaftsordnung und Überlieferungsbildung, in: Archivalische Zeitschrift 68 (1972), S. 3-40.

33 Booms, Hans: Society and the Formation of a Documentary Heritage: Issues in the Appraisal of Archival Sources, in: Archivaria 24 (1987), S. 69-107.

Dabei erweist es sich freilich immer wieder als sehr schwierig, den heutigen Nutzern Antworten auf diese archivischen Fragen zu entlocken. 2008 hat der Verfasser dieser Zeilen beispielsweise zusammen mit Peter Haber die Historiker, also die Gruppe der professionellen Archivnutzer gefragt, wie denn aus ihrer Sicht genuin digitale Archivalien beschaffen sein müssten.[34] Nachdem auf diese Anfrage keine Antwort einging, wurde die Frage vor dem größeren Forum von HSozCult wiederholt. Gefragt wurde:

— «Wo arbeiten Historiker bereits heute mit genuin digitalen Quellen (Quellen, die digital entstanden und geblieben sind)?
— Ist es für die Forschungen erheblich oder unerheblich, dass diese Quellen in digitaler Form vorliegen? Warum?
— Welche Eigenschaften sollten digitale Quellen für die Forschungen besitzen (z.B. Durchsuchbarkeit, statistische Auswertbarkeit etc.)?
— Welche Typen digitaler Quellen (z.B. Webseiten, Blogs, elektronische Akten) erscheinen heute in besonderem Maß interessant für künftige Historiker?
— Welche Bereiche der heutigen Informationsgesellschaft (z.B. bestimmte Vereine oder Gerichte) sollten für künftige Generationen archiviert werden?
— Oft sind die rechtsverbindlichen Quellen noch auf Papier, während die digitalen Formen zugleich leichter zu benutzen sind. Wie lassen sich diese beiden Aspekte (Rechtsverbindlichkeit / digitale Benutzbarkeit) für eigene Forschungsprojekte gewichten?»[35]

Leider ging auf die gestellten Fragen keine einzige konkrete Antwort ein.[36] Anfang 2018 scheinen sich diese Verhältnisse nicht grundlegend geändert zu haben. Es ist immer noch nicht erkennbar, wie diese Fragen aus Sicht der Geschichtswissenschaft zu beantworten sind.

Auch die Archive und die ihnen zugeschriebene Wissenschaft sind noch weit von einem in sich kohärenten und für die archivische Praxis hilfreichen Bild der Designated Community entfernt. Dennoch können wir einige Beispiele benennen, in denen mit dem Konzept schon gearbeitet wird – Beispiele, die zeigen, weshalb es sinnvoll sein dürfte, weiterhin am Konzept der Designated Community festzuhalten.

34 http://weblog.histnet.ch/archives/tag/digitale-quellen.

35 Anfrage 'Forschen mit Digitalen Quellen' vom 4.12.2008 auf http://hsozkult.geschichte.hu-berlin.de/forum/id=1055&type=anfragen.

36 Keitel, Christian: Über den Zusammenhang zwischen Quellenkritik und Informationserhalt. Ergebnisse der Anfrage «Forschen mit ‚digitalen Quellen'», http://hsozkult.geschichte.hu-berlin.de/forum/type=anfragen&id=1055.

Anwendungsbeispiele

Zertifizierung

Die heutigen Ansätze zur Zertifizierung digitaler Archive können auf den Bericht der Commission on Preservation and Access von 1996 zurückgeführt werden.[37] Nach über 20 Jahren liegen nun drei Zertifizierungsverfahren vor, die alle keine konkrete technische Implementierung erwarten. Sie benötigen daher einen anderen Ausgangspunkt und finden diesen nicht überraschend in der Designated Community. Es ist daher nur konsequent, dass dieser Ausgangspunkt auch stets am Anfang der jeweiligen Standards steht:

- Die «Core Trustworthy Data Repositories Requirements» beginnen mit der Definition des Archivs («repository») und seiner Designated Community.[38]
- Der ISO Standard 16363 und seine inhaltliche Entsprechung, der CCSDS Standard «Audit and Certification of Trustworthy Digital Repositories», nennen die Designated Community als viertes Haupt-Kriterium (viele Kriterien sind untergliedert): «The repository shall have defined its Designated Community and associated knowledge base(s) and shall have these definitions appropriately accessible.»[39]
- In der DIN 31644 erscheint die Benennung der Zielgruppen als drittes Kriterium.[40]

Jeder dieser Ansätze geht davon aus, dass die Vertrauenswürdigkeit digitaler Archive nur in Bezug auf die anzunehmenden künftigen Nutzer näher bestimmt werden kann. Nach der Vergabe des nestor-Siegels an mittlerweile vier Archive können diese Annahmen auch ganz praktisch bestätigt werden.[41] Die Angemessenheit der Prozesse und Maßnahmen eines digitalen Archivs kann nur vor dem Hintergrund der vom Archiv selbst gesetzten Zielsetzungen beurteilt werden. An erster Linie ist dabei die Designated Community zu berücksichtigen.

37 Preserving Digital Information / Task Force on Archiving Digital Information, Commission on Preservation and Access. - Washington D.C., 1996.

38 Core Trustworthy Data Repositories Requirements, v.01.00, https://drive.google.com/file/d/0B4qnUFYMgSc-eDRSTE53bDUwd28/view. Die Requirements wurden von den Vertetern von Data Seal of Approval und dem World Data System vereinbart und lösen das bekanntere Data Seal of Approval ab.

39 Audit and Certification of Trustworthy (wie Anm. 23), S. 3-5. ISO-Norm 16363:2012, Space data and information transfer systems -- Audit and certification of trustworthy digital repositories.

40 DIN 31644: Information und Dokumentation - Kriterien für vertrauenswürdige digitale Langzeitarchive, 2012-04.

41 Der Autor leitet zusammen mit Dr. Astrid Schoger die nestor-AG Zertifizierung, die den nestor-Siegel vergibt, http://www.langzeitarchivierung.de/Subsites/nestor/DE/Siegel/siegel_node.html.

Bewertung

2007 beschloss das Landesarchiv Baden-Württemberg, die Bewertung personenbezogener Unterlagen auf neue Beine zu stellen. Einerseits erbrachte das sogenannte DOT-Modell teilweise zu große Übernahmemengen, andererseits sollten auch die elektronischen Fachverfahren berücksichtigt werden. Eine Arbeitsgruppe fand insgesamt fünf verschiedene Nutzungsziele:[42]

- Ein Familienforscher könnte daran interessiert sein, zu seinen Vorfahren wenigstens einige Grundinformationen zu finden. Da das Archiv nicht alle Nutzer im Voraus kennen kann, sollten in diesem Fall von allen Mitarbeitern wenigstens einige grundlegende Informationen archiviert werden.
- Ein Nutzer könnte an einer sozialwissenschaftlichen quantitativen Auswertung interessiert sein.
- Ein Nutzer könnte am Leben eines zeittypischen Mitarbeiters interessiert sein.
- Ein anderer Nutzer könnte dasselbe Interesse auf eine berühmte Persönlichkeit beziehen.
- Schließlich könnte ein Nutzer noch ein Interesse daran haben zu untersuchen, wie die Einrichtung gearbeitet und funktioniert hat.

In den ersten beiden Fällen dürften künftige Nutzer vor allem an der Übernahme von Daten aus Fachverfahren interessiert sein, in den beiden darauf folgenden Fällen an der Übernahme einzelner Akten. Im zuletzt genannten Fall könnten sowohl Fachverfahren als auch Akten von Interesse sein. Die hier über künftige Nutzer getroffenen Aussagen sind einerseits auf einer hohen Abstraktionsebene. Andererseits ermöglichen erst diese Annahmen die bei der Bewertung notwendige abwägende Unterscheidung zwischen verschiedenen Unterlagengruppen. Die Annahmen sind daher praktikabel und mit Gewinn umsetzbar. Seit der Publikation des Artikels fragt der Autor dieser Zeilen in fast allen Fortbildungsveranstaltungen, ob die Teilnehmer nicht noch eine weitere Nutzungsmöglichkeit benennen könnten. Bislang wurde auf diese Frage noch keine Ergänzung genannt, was indirekt als Beleg dafür gewertet werden kann, dass die Überlegungen der Arbeitsgruppe doch eine gewisse Dauerhaftigkeit beanspruchen können.

Bestandserhaltung

Auch die nestor-Arbeitsgruppe «Digitale Bestandserhaltung» begreift die Designated Community als zentralen Referenzpunkt für alle anstehenden archivischen

42 Ernst, Albrecht et al.: Überlieferungsbildung bei personenbezogenen Unterlagen, in: Archivar 61 (2008), S. 275-278.

Entscheidungen.[43] Die Wahl zwischen den beiden von OAIS selbst vorgeschlagenen Bestimmungen des Begriffs fiel daher eher auf die Definition des Archivs als auf die Definition der Designated Community selbst. Der Verstehbarkeit kommt dabei zwar eine wichtige, aber keine herausgehobene Rolle zu. Stattdessen wurde die Designated Community mit einem Kompass gleichgesetzt. Dieser mag zwar subjektiv und spekulativ sein, dennoch bietet er die einzige vertretbare Orientierung bei zahlreichen anstehenden Entscheidungen. Es gibt schlicht keinen anderen Kompass, der bei archivischen Entscheidungsprozessen eine vergleichbare Funktion übernehmen könnte.

Bei Wanderungen ist es nun eine Sache, dass der Kompass zuverlässlich nach Norden zeigt und eine völlig andere, welche sonstigen Bedingungen noch bei der Wegfindung berücksichtigt werden müssen. Berge, Flüsse und andere Umstände müssen dabei berücksichtigt werden. Vergleichbar hierzu nennt auch die nestor-Arbeitsgruppe vier allgemeine Faktoren, die in die Entscheidungsfindung einbezogen werden sollten: Finanzierbarkeit, Authentizität, Angemessenheit und Automatisierbarkeit.[44] In dieser Landschaft kann der Kompass namens Designated Community zwar eine gute Orientierung ermöglichen. Zugleich kann es aber auch sein, dass das Archiv nicht in der Lage ist, die angenommenen Bedürfnisse künftiger Nutzer vollständig zu erfüllen.

Soweit der große Rahmen. Die konkreten Entscheidungen sind dann zumeist auf die einzelnen Archivalien bezogen. Es ist sicherlich nicht übertrieben, diese objektbezogenen Entscheidungen zum Kern der archivischen Arbeit zu erklären. Der Leitfaden zur digitalen Bestandserhaltung beschreibt drei konkrete Faktoren, die bei solchen Entscheidungen berücksichtigt werden müssen:

- Die Objekte: Die meisten Archive beherbergen eine große Menge unterschiedlicher Objekte. Schon aus finanziellen Gründen dürften die wenigsten in der Lage sein, jedes Objekt eigenständig zu bearbeiten. Es führt also kein Weg an einer künftigen Automatisierung vorbei, und um diese zu ermöglichen, schlägt der Leidfaden vor, die Objekte in gleichartige Gruppen aufzuteilen. Diese Informationstypen sollen dem Archiv auf einem möglichst einfachen Weg erlauben, einen Überblick über seine Objekte zu gewinnen.
- Die Designated Community: Jedes Archiv sollte in einer Policy seine Designated Communities beschreiben und sich dabei auch auf den Detaillierungsgrad dieser Beschreibung festlegen. So kann die Beschreibung «Historiker», «Sozialhistoriker» oder «Sozialhistoriker des späten 20. Jahrhunderts» nen-

43 Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung, Version 1.0, verfasst und herausgegeben von der nestor-Arbeitsgruppe Digitale Bestandserhaltung, Version 2.0, Frankfurt/Main 2012.

44 Leitfaden (wie Fußnote 43), S.6 f.

nen. Dennoch können auch solche Beschreibungen leicht zu abstrakt sein. Die nestor-Arbeitsgruppe schlug daher einen dritten Begriff vor:

— Nutzungszweck: Unter diesem Begriff sollen die funktionalen Möglichkeiten beschrieben werden, mit denen das Objekt in Zukunft genutzt wird. Der Nutzungszweck ist von inhaltlichen Bestimmungen deutlich zu unterscheiden. Es ist daher besser, von «statistischer Auswertbarkeit» zu sprechen als von einer «Suche nach der Verteilung von Männern und Frauen». Auch beim Nutzungszweck liegt es an jedem einzelnen Archiv, den Detaillierungsgrad seiner Beschreibung festzulegen. Diese konkreten Ausformulierungen können auf vier allgemeine Nutzungszwecke zurückgeführt werden: Wahrnehmung des Gesamtobjekts (z.B. die Lektüre eines Romans); Suche nach einer bestimmten Information (z.B. nach der ersten Nennung einer bestimmten Romanfigur); Weiterverarbeitung des Objekts oder von Teilen davon (z.B. die Verwendung einer statistischen Datenreihe in einer von Nutzer selbst angelegten größeren Datenbank); oder Ausführung des Objekts (z. B. Spielen eines Computerspiels).

Nutzungsziele sind in dem Leitfaden eine unabhängige Einheit. Ein Nutzungsziel kann daher von verschiedenen Designated Communities verfolgt werden. Ebenso kann ein digitales Objekt verschiedene Nutzungsziele ermöglichen.

Objektbezogene Entscheidungen im Archiv sind daher Entscheidungen, bei denen eine auf den Objekten selbst basierende Kategorie mit zwei externen Kategorien (Designated Communities und Nutzungsziele) abgeglichen werden muss. Es muss sowohl das Gegebene (also das Objekt) als auch die Zielsetzung (was getan werden sollte) berücksichtigt werden.

Erst wenn wir Annahmen über die künftigen Nutzer und Nutzungsziele machen, können wir die Archivalien bewerten und erhalten. Diese Annahmen können manchmal notgedrungen sehr vage sein und nach weiteren Konkretisierungen verlangen. Dennoch können wir auf sie nicht verzichten. Erst über die Einbeziehung von Nutzern und Nutzungsformen wird unser eigenes Handeln im Archiv transparent, erst dann kann es weiterentwickelt werden.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.4

Digitale Archivierung im Schweizerischen Bundesarchiv – Ein Blick hinter die Kulissen

Krystyna W. Ohnesorge

Entwicklung von Lösungen für die digitale Archivierung

Die sich ständig ändernden IKT-Werkzeuge haben den Effekt, dass eine Interoperabilität der Daten während einer langen Zeitperiode von 30 oder 50 Jahren selten erreicht wird. Die Archivierung von digitalen Informationsobjekten stellt daher eine herausfordernde Aufgabe dar. Für die Erschliessung, Sicherung und Vermittlung von digitalem Archivgut sind spezielle Hilfsmittel zur Verfügung zu stellen. Zudem bedarf die langfristige Erhaltung der Informationsobjekte über den technologischen Wandel hinweg aufwändige Migrationen auf neue Formate und Systemplattformen. Angesichts der Komplexität, Heterogenität und der erforderlichen Sicherheit der Daten sind diese Aufgaben keineswegs trivial. Sie beanspruchen hohe Investitionen und setzen eine Kontinuität beim Aufbau von Methoden und Lösungen sowie beim archivisch-technischen Wissen voraus. Um die Archivierung von digitalem Archivgut erfolgreich anzugehen, müssen wir uns aber nicht nur mit den technischen Möglichkeiten auseinandersetzen, sondern auch mit den organisatorischen Rahmenbedingungen. Dazu gehört beispielsweise die Entwicklung von organisatorischen Szenarien zur Übernahme, Sicherung und Wiederverwendung von digitalem Archivgut.

Digitale Archivierung im Betrieb

Das Schweizerische Bundesarchiv (BAR) hat sich in den vergangenen achtzehn Jahren umfassend mit der digitalen Archivierung im nationalen wie internationalen Kontext auseinandergesetzt. Das BAR hat Strategien, Konzepte und Prozesse für die digitale Archivierung erarbeitet und diese mit Hilfe von umfangreichen Informatiklösungen im Betrieb umgesetzt. Die Archivierungsstrategie des BAR beruht auf dem Migrationsprinzip.[1] Die Archivierung von digitalen Unterlagen erfolgt in einer begrenzten Anzahl von genau spezifizierten, standardisierten und vom BAR

1 Policy digitale Archivierung, 2009, https://www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/policy_digitale_archivierung.pdf.download.pdf/policy_digitale_archivierung.pdf. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.5

publizierten Dateiformaten.[2] Diese Formate sind für die Archivierung von digitalen Unterlagen besonders geeignet und wurden vom BAR ausgewählt und ausdrücklich gutgeheissen. Das BAR bestimmt jährlich die archivtauglichen Formate im Rahmen eines festgelegten Prozesses und publiziert diese als verbindliche Vorgabe für die Ablieferung von digitalen Unterlagen für die Bundesverwaltung (BV).

Die Lösung Digital Information Repository (DIR), welche die Kernprozesse Ablieferung, Sicherung, Erhaltung und Vermittlung von digitalen Daten aus Fachapplikationen (Datenbanken) sowie digitalen Geschäftsunterlagen aus GEVER-Systemen und aus Fileablagen über standardisierte Schnittstellen unterstützt, wurde im Einklang mit internationalen Standards entwickelt und bereits 2009 eingeführt. Heutzutage findet man im digitalen Magazin des BAR 18 TB an digitalen Unterlagen, wobei es sich hier ausschliesslich um die sogenannten «digitally born data» aus der Bundesverwaltung handelt.

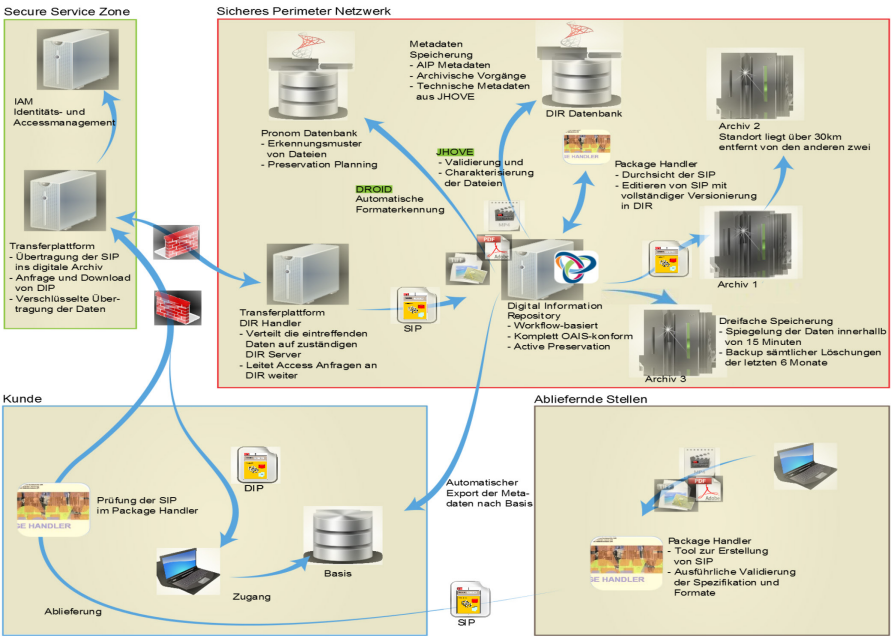

Abbildung 1: Die Applikationslandschaft des BAR für die digitale Archivierung

2 Archivtaugliche Dateiformate, 2014, https://www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/archivtaugliche_dateiformate.1.pdf.download.pdf/archivtaugliche_dateiformate.pdf.

Ausbau: Weiterentwickeln und verbessern

Das BAR hat seine Lösungen nicht nur auf der funktionalen Applikationsebene weiterentwickelt und verbessert, sondern in dieser Zeit auch die Hardware regelmässig erneuert und den neuesten IKT- sowie Sicherheitsstandards der Bundesverwaltung angepasst. Wir haben beispielsweise die Anwendung *SIARD Suite* und das SIARD-Format für die Archivierung von relationalen Datenbanken entwickelt sowie den *Package Handler*, mit dem man SIP (Submission Information Packages) gemäss dem Standard eCH-0160 erstellen und validieren kann. Beide Anwendungen stellt das BAR der archivischen Gemeinschaft gratis zur Verfügung. Der Betrieb der Infrastruktur für die digitale Archivierung (Soft- und Hardware) wird in Zusammenarbeit mit dem zentralen IT-Leistungsanbieter der Bundesverwaltung, dem Bundesamt für Informatik und Telekommunikation (BIT), sichergestellt. Entsprechend den Sicherheitsanforderungen des BAR wird das digitale Archivgut an drei geografisch getrennten Standorten gespeichert; davon ist einer über 30 km von den beiden anderen entfernt. Die Sicherheitskopien von SIPs werden zudem zweimal täglich mithilfe eines Backups gesichert, das das BAR bis auf ein Jahr zurückverfolgen kann. Die Infrastruktur läuft in einem dedizierten Perimeter-Netzwerk innerhalb der sicheren BV-Netz-Zone, welches extra für das BAR abgetrennt und abgesichert ist.

2014 hat das BAR eine weitere Anwendung mit dem Namen Transferplattform in Betrieb genommen, die eine automatische und sichere Übernahme von SIP von der abliefernden Stelle ins BAR garantiert und die Identifikation sowie Authentifikation der Benutzer mittels einer sogenannten Zwei-Faktoren-Technologie ermöglicht. Die Transferplattform gestattet es dem BAR auch, die digitale Archivierung als Dienstleistung für Dritte anzubieten.[3]

In der neuen Strategie 2016–2020[4] setzt das BAR die 2006 begonnene Transformation zu einem digitalen Archiv konsequent fort, um allen Interessierten den Zugriff auf Informationen aus dem Archiv und deren Verarbeitung orts- und zeitunabhängig zu ermöglichen. Das BAR baut dafür beispielsweise seine öffentliche digitale Informationsinfrastruktur aus, die den Kunden erlauben wird, selbstständig zu recherchieren und direkt auf die gefundenen Informationen beziehungsweise Unterlagen zuzugreifen, um sie auszuwerten und weiterzuverarbeiten. Zudem entwickelt das BAR Hilfsmittel zur digitalen Auswertung von Quellen und erforscht wichtige Themen, insbesondere im Bereich der Verwaltungswissenschaften.

3 Digitale Archivierung im BAR, 2017, https://www.bar.admin.ch/dam/bar/de/dokumente/kundeninformation/dokumentation_produktdigitalearchivierung.pdf.download.pdf/dokumentation_produktdigitalearchivierung.pdf.

4 Strategie Bundesarchiv 2016-2020, 2015, https://www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/Strategie%20Bundesarchiv%202016-2020.pdf.download.pdf/Strategie_Bundesarchiv_2016-2020.pdf.

Dadurch erweitert es die Kenntnisse sowohl über die Verwaltungsgeschichte als auch über die im BAR archivierten Bestände. Diese neue Ausrichtung wird dem BAR zukünftig erlauben, seine Dienstleistungspalette im digitalen Bereich gezielt zu erweitern.

Angebot «Digitale Archivierung für Dritte»

Die Ausgaben für den Betrieb und den Ausbau von digitalen Archivierung nehmen aber immer weiter zu. Im Alleingang werden die Bereitstellung und der Unterhalt der benötigten Infrastrukturen und Dienstleistungen immer schwieriger. Das BAR stellt deshalb seine digitale Archivierungsinfrastruktur und die dazu gehörenden Softwarelösungen auch bundesverwaltungsexternen Kunden zur Nutzung zur Verfügung. Kantone, Gemeinden und Institutionen mit einem öffentlich-rechtlichen Auftrag können ihre Daten beim Bundesarchiv digital sichern und sich bei der langfristigen Archivierung beraten lassen. Durch die mehrfache Nutzung können die Kosten auf mehrere Partner verteilt werden, was einen wirtschaftlichen Betrieb der digitalen Archive garantiert.

Die Kernprozesse wie die Beratung der abliefernden Stellen in der eigenen Verwaltung, die Bewertung, welche Unterlagen archiviert werden sollen, das Führen eines Metadatenkatalogs und somit die Erschliessung sowie die Vermittlung des Archivguts bleiben weiterhin in der Kompetenz des Drittkunden des BAR. Somit verbleibt die gesamte Kommunikation mit den eigenen internen und externen Kunden in der Kompetenz des jeweiligen Archivs. Das BAR kommuniziert ausschliesslich mit den Archivaren des Kunden und übernimmt die Archivierung von digitalen Unterlagen als sogenannter «Full Service Provider».

Kosten bei 1 TB Initialmenge	
Investitionskosten System	44‘750 CHF
Investitionskosten 1 TB Speicher	6‘200 CHF
Total Investitionskosten (einmalig)	**50‘950 CHF**
Direkte Betriebskosten System	20‘648 CHF
Indirekte Betriebskosten System	53‘729 CHF
Betriebskosten 1 TB Speicher	6‘100 CHF
Total Betriebskosten (jährlich)	**80‘477 CHF**
Kosten bei einem jährlichen Wachstum um 1 TB	
Total Investitionskosten (einmalig)	**6‘200 CHF**
Total Betriebskosten (jährlich)	**6‘100 CHF**

Abbildung 2: Kostenmodell des BAR

Die Dienstleistung des BAR «digitale Archivierung für Dritte» wird zu Selbstkosten angeboten, d. h. sie basiert auf einer Vollkosten-Rechnung. Die Kosten setzen sich dabei aus Investitions- und Betriebskosten zusammen. Die einmaligen Investitionskosten umfassen dabei die sogenannten Systemkosten und die Investitionskosten für den Speicher. Zu Systemkosten zählen beispielsweise die Lizenzkosten für DIR, die Anbindung durch die Transferplattform beim Kunden und die Projektkosten (etwa 21 Tage). Die Betriebskosten umfassen die direkten und indirekten Kosten. In den direkten Betriebskosten sind der Support für das DIR und der Betrieb der Transferplattform sowie der Support durch BAR-Mitarbeitende (etwa 12-13 Tage pro Jahr) berechnet. In den indirekten Betriebskosten, d. h. den Kosten, die nicht nach direktem Verursacherprinzip, sondern mithilfe von Schlüsselgrössen berechnet werden, sind der Anteil an der Abschreibung der Infrastrukturen und der Zuschlag für die Betriebskosten enthalten.

Das Beispiel in Abbildung 2 zeigt auf, wie die Rechnung mit der Initialmenge von 1 TB berechnet wird. Die einmaligen Investitionskosten bei 1 TB betragen dabei rund CHF 51’000. Die jährlichen Betriebskosten bei 1 TB betragen ca. CHF 80’500. Die Speicherkosten sind mengenabhängig. Die Mindestgrösse einer verfügbaren Speicherplatzeinheit beträgt im BAR-Angebot 0,5 TB.

Kosten teilen – Synergien nutzen

Als institutioneller Betrieb des Bundes garantiert das BAR höchste Standards an Verfügbarkeit, Datensicherheit und Beständigkeit. Die Dienstleistung des Bundesarchivs «digitale Archivierung für Dritte» bietet einen Lösungsansatz für die Optimierung der Ausgaben für die digitale Archivierung durch die Nutzung von Synergien zwischen Archiven. Mit dem transparenten Kostenmodell ist für die Kunden klar ersichtlich, welche Kosten initial und für den Betrieb anfallen. Der Nutzen dieses Betriebsmodells liegt darin, dass alle Kosten in Zusammenhang mit dem Aufbau, dem Betrieb und der Weiterentwicklung der IT-Infrastruktur für ein eigenes digitales Magazin entfallen. Auch alle Massnahmen für die Erhaltung des Archivguts werden durch das BAR durchgeführt.

Die Vorteile der Nutzung der Dienstleistungen des BAR durch die Archive der öffentlichen Verwaltung liegen bei der Einsparung der teilweise erheblichen Investitionen für den Aufbau eines digitalen Archivs und der Mitnutzung von innovativen Infrastrukturen für digitale Archivierung. Zudem profitieren die Kunden vom technischen und spezialisierten Knowhow des BAR und müssen, was vor allem für kleinere Archive interessant ist, dieses nicht mit eigenem Personal sicherstellen. Das BAR garantiert zudem die kontinuierliche Weiterentwicklung der Lösungen für die digitale Archivierung.

Archivierung im Verbund

Kosten der digitalen Langzeitarchivierung am Beispiel von DiPS.kommunal

Julia Krämer-Riedel, Tobias Schröter-Karin

Einleitung

Elektronische Langzeitarchivierung ist kein günstiges Unterfangen. Davon wissen diejenigen Einrichtungen ein Lied zu singen, die in den vergangenen Jahren schon mit dem Aufbau eines Langzeitarchivs beschäftigt waren. Am Beispiel von DiPS.kommunal, der Lösung, die der Landschaftsverband Westfalen-Lippe und das Historische Archiv der Stadt Köln nutzen, und den praktischen Erfahrungen, die beide Einrichtungen bislang auf dem Gebiet der Langzeitarchivierung gesammelt haben, soll ein Einblick in Kostenfaktoren gegeben werden, die beim Aufbau und beim Unterhalt eines elektronischen Langzeitarchivs berücksichtigt werden müssen. Die Betrachtungen stützen sich auf Überlegungen, die im Zuge der Kostenkalkulation für das Digitale Archiv Nordrhein-Westfalen (DA NRW)[1] bzw. DiPS.kommunal angestellt wurden. Letztlich geht es um die Frage, wie ein Kostenmodell für die digitale Langzeitarchivierung aussehen muss. Bei DiPS.kommunal wurde geprüft: Wo fallen Aufwände an? Welche Kostenfaktoren gibt es überhaupt? Was muss in die Kalkulation mit einfließen? Diejenigen Archive, die sich in naher Zukunft auf den Weg machen, um ein elektronisches Archiv aufzubauen, werden sich mit eben diesen Fragen beschäftigen müssen.

Allerdings – dies sei den Ausführungen vorangestellt – ist es sehr schwierig, hieb- und stichfeste Zahlen für «die» Kosten der Langzeitarchivierung zu liefern. Die Aufwände, die seit Jahren in die Entwicklung der Digital Preservation Solution (DiPS)[2] geflossen sind, lassen sich rückwirkend kaum mehr aufschlüsseln und genau beziffern, haben doch eine Vielzahl von Institutionen und Personen am Aufbau dieser Lösung mitgewirkt.[3] Und für einen Vergleich der Kosten der analogen und digitalen Archivierung fehlt sowohl in Köln also auch in Münster schlicht eine seriöse Datengrundlage. Die Frage «Was kostet mich die digitale Langzeitarchivierung?» wird an dieser Stelle also nicht abschließend beantwortet werden kön-

1 Informationen zum Projekt DA NRW unter: https://www.danrw.de/. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Informationen zur Digital Preservation Solution (DiPS) unter: http://www.stadt-koeln.de/leben-in-koeln/kultur/historisches-archiv/dips-digital-preservation-solution.

3 Hoppenheit, Martin; Schmidt, Christoph; Worm, Peter: Die Digital Preservation Solution (DiPS). Entstehung, Grundlagen und Einsatzmöglichkeiten eines Systems zur elektronischen Langzeitarchivierung. In: Archivar 69 (2016), S. 375-382. http://www.archive.nrw.de/archivar/hefte/2016/Ausgabe_4/Ausgabe_4-16.pdf.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.6

nen. Die Archivlandschaft und die einzelnen Standortbedingungen sind viel zu heterogen, als dass eine konkrete Summe genannt werden könnte, die auf alle übertragbar wäre. Der Aufbau eines Langzeitarchivs ist in jedem Fall ein sehr individuelles Projekt. Nichtsdestotrotz gibt es einige Kostenfaktoren, die bei jedem Projekt bedacht werden müssen und die je nach Ausgangslage kostenmäßig unterschiedlich ins Gewicht fallen können.

Ziel dieser Ausführungen kann daher nur sein, die zweifellos hohen Kosten der digitalen Archivierung nachvollziehbar zu machen – und damit vielleicht auch entsprechend nachvollziehbare Argumente denjenigen an die Hand zu geben, die ein Langzeitarchiv aufbauen müssen.[4] Am Beispiel von DiPS.kommunal, das unter dem Dach des DA NRW im Verbund genutzt werden kann, sollen schließlich die Möglichkeiten, die eine solche Verbundlösung aus organisatorischer Sicht bietet, beleuchtet werden.

Kostenfaktoren bei der Langzeitarchivierung

Beim Aufbau und Betrieb eines elektronischen Langzeitarchivs müssen einige Kostenfaktoren berücksichtigt werden. Die Frage «Was kostet die digitale Langzeitarchivierung?» wurde schon mehrfach, auch auf dieser Tagung, gestellt. Susanne Fröhlich hat am Beispiel des «Digitalen Archivs Österreich» 2012 und 2015 die Kosten für den Aufbau eines Langzeitarchivs aufgezeigt.[5] Es gibt hierzu auch veröffentlichte allgemeine Modelle, die Kostenfaktoren benennen und auf dieser Grundlage die Kosten für ein Langzeitarchiv unter bestimmten Bedingungen errechnen bzw. Formeln entwickeln, nach denen die Kosten unter Berücksichtigung der variablen Parameter (z.B. Größe der zu archivierenden Daten) berechnet werden können.[6]

4 Hierzu: Sandner, Peter: 10 FAQs. Argumente zu Bedarf und Notwendigkeiten der digitalen Archivierung, in: Keitel, Christian; Naumann, Kai (Hg.): Digitale Archivierung in der Praxis: 16. Tagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen». Stuttgart 2013, S. 57-70. http://www.staatsarchiv.sg.ch/home/auds/16/_jcr_content/Par/downloadlist/DownloadListPar/download_0.ocFile/Sandner_10_FAQs.pdf.

5 Fröhlich, Susanne: Kostenfaktoren in digitalen Archiven. Erfahrungen des Digitalen Archivs Österreich, in: Keitel, Christian; Naumann, Kai (Hg.): Digitale Archivierung in der Praxis: 16. Tagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen». Stuttgart 2013, S. 31-49. http://www.staatsarchiv.sg.ch/home/auds/16/_jcr_content/Par/downloadlist/DownloadListPar/download_1.ocFile/Froehlich_Kostenfragen_in_digitalen_Archiven.pdf. Zum Vortrag von Susanne Fröhlich auf der 19. Tagung des Arbeitskreises am 10./11. März 2015 in Wien siehe: dies., Ein Showcase [Präsentation]. http://www.staatsarchiv.sg.ch/home/auds/19/_jcr_content/Par/downloadlist_3/DownloadListPar/download.ocFile/_Fr%C3%B6hlich,%20Susanne_%20Digitales%20Archiv%20%C3%96sterreich%20-%20Ein%20Showcase%20%5BPr%C3%A4sentation%5D.pdf. Zu den Kosten der digitalen Archivierung vergleiche außerdem die Beiträge von Gabriele Stüber und Peter Sandner, die ebenfalls auf der 16. AUdS-Tagung gehalten wurden: http://www.staatsarchiv.sg.ch/home/auds/16.html.

6 Schmitt, Karlheinz: Kosten der digitalen Archivierung. Ein mögliches Vorgehensmodell und erste Erfahrungen, in: Keitel, Christian; Naumann, Kai (Hg.): Digitale Archivierung in der Praxis: 16. Ta-

Nicht zuletzt das nestor-Handbuch verweist auf solche Modelle.[7] Ein prominentes Beispiel ist das LIFE-Projekt, eine im Jahr 2005 begonnene, mehrjährige Initiative der British Library und des University College London – also eine nicht von Archiven ausgehende Initiative –, die es sich zum Ziel gesetzt hatte, ein Modell zu erarbeiten, mit dessen Hilfe Einrichtungen die Kosten für ihre Langzeitarchivierung besser kalkulieren können.[8] Ein anderes Beispiel wäre das Projekt DP4lib (Digital Preservation for Libraries), an dem u.a. die Deutsche Nationalbibliothek und die Niedersächsische Staats- und Universitätsbibliothek Göttingen beteiligt waren.[9]

Im LIFE-Projekt ging man bei der Betrachtung von den einzelnen, aufeinanderfolgenden Arbeitsschritten bei der Langzeitarchivierung aus und beschrieb, an welcher Stelle welche Kostenfaktoren einzukalkulieren sind. So beginnt der «Lebenszyklus» mit der Entstehung, mit Erwerb, Auswahl/Bewertung und Übernahme der Unterlagen, es folgen Erschließung/Katalogisierung, Signaturenvergabe, Erhaltung, Konservierung, Speicherung, Abruf/Wiederauffindung, Benutzung/Ausgabe. Dazu fallen auch Kosten außerhalb des Lebenszyklus an, z.B. Verwaltung/Administration oder Systeminfrastruktur.[10]

Man sieht an den Begrifflichkeiten, dass es sich nicht um ein rein archivisches Projekt handelte.[11] Man erkennt aber auch dahinter das OAIS-Modell und seine Module Datenübernahme/Ingest, Datenaufbewahrung/Archival Storage, Datenmanagement, Systemverwaltung, Preservation Planning und Zugriff/Access;

---

gung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen». Stuttgart 2013, S. 19-29. http://www.staatsarchiv.sg.ch/home/auds/16.html.

7 Vgl. hierzu Kapitel 14 «Geschäftsmodelle» des nestor-Handbuches mit Beiträgen von Achim Oßwald (14.1 Einführung), Thomas Wollschläger und Frank Dieckmann (14.2 Kosten, 14.3 Service- und Lizenzmodelle), in: nestor-Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Hrsg. von Neuroth, Heike et al. Version 2.3, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php. Siehe im Einzelnen die Artikel von A. Oßwald (14.1 Einführung), F. Dickmann und Th. Wollschläger (14.2 Kosten, 14.3 Service- und Lizenzmodelle).

8 Zum LIFE-Projekt: http://www.life.ac.uk/.

9 Zum Projekt DP4Lib: http://www.dnb.de/EN/Wir/Projekte/Archiv/dp4lib.html. Zur Übertragbarkeit des Kostenmodells auf Archive: Ucharim, Michael: DP4Lib als Kostenmodell für die digitale Langzeitarchivierung im Archivwesen? (Transferarbeit 2013). https://www.landesarchiv-bw.de/sixcms/media.php/120/55275/Transferarbeit2013_Ucharim.pdf.

10 Abschlussbericht des LIFE 3-Projekts: http://www.life.ac.uk/3/documentation.shtml und http://www.life.ac.uk/3/docs/life3_report.pdf, insbesondere Grafik auf S. 7 (27.08.2017). Siehe außerdem: Strodl, Stefan; Rauber, Andreas: A cost model for small scale automated digital preservation archives. https://fedora.phaidra.univie.ac.at/fedora/get/o:294219/bdef:Content/get; Weatley, Paul: Costing the Digital Preservation Lifecycle More Effectively. https://fedora.phaidra.univie.ac.at/fedora/get/o:294138/bdef:Content/get; Hagel, Harald; Minkus, Michael et. al., Entwicklung von Organisations- und Geschäftsmodellen zur Langzeitarchivierung digitaler Objekte aus DFG-geförderten Digitalisierungsprojekten. Studie im Auftrag der Deutschen Forschungsgemeinschaft. April 2009. https://www.digitale-sammlungen.de/content/dokumente/2009_04_Studie_Organisations_und-Geschaeftsmodelle.pdf.

11 Starkloff, Kristina: Übertragbarkeit des Kostenmodells zur Langzeitarchivierung LIFE auf den archivischen Bereich (Transferarbeit 2013). https://www.landesarchiv-bw.de/sixcms/media.php/120/55273/Transferarbeit2013_Starkloff.pdf.

Bereiche, die sich im Übrigen auch auf die analoge Archivierung übertragen lassen. Unter Archival Storage wären hier z.B. dann neben Magazinfläche auch Klimatisierung und Verpackungsmaterial zu berücksichtigen.[12]

In jedem dieser Bereiche fallen für bestimmte Aufwände Kosten an, die sich auch nach Kostenarten trennen lassen: Personalkosten, Betriebskosten, Lizenzkosten, Entwicklungskosten, Kosten für Dienstleistungen und Anschaffungen, eventuell Fortbildungskosten oder Kosten für Dienstreisen etc. Und viele Kosten fallen bei der analogen Archivierung genauso an wie bei der elektronischen Archivierung: Ein Beispiel wären Bau- und Unterhaltskosten für das Gebäude oder Personalkosten, zeitliche Aufwände für Behörden-/Dienststellenbetreuung, Bewertung, Überlieferungsbildung, Aktenübernahme und Magazinierung und Kosten für Bestandserhaltung (Verpackung, Klimatisierung, ggf. Restaurierung, Reproduktion/Digitalisierung).[13]

An dieser Stelle wäre es natürlich schön gewesen, ein Beispiel zu haben, um die Kosten beider Archivierungsarten gegeneinanderzustellen und zu prüfen, was günstiger ist: elektronische oder analoge Archivierung. Allerdings fehlte hierzu eine seriöse Datengrundlage. Eine solche Rechnung bedürfte Vergleichszahlen mehrerer Archive verschiedener Größe und Ausstattung, andernfalls bliebe das Ergebnis zwangsläufig eng auf ein Archiv bezogen und ließe sich nicht verallgemeinern bzw. übertragen (vgl. z.B. die von Standort zu Standort unterschiedlichen Kosten für Miete/Lagerfläche). Bau- und Betriebskosten für das Archivgebäude müssten in die Kosten der digitalen Archivierung anteilig miteinberechnet werden, denn auch die

12 Vgl. Kapitel 4 des nestor-Handbuches (Version 2.3) zum Referenzmodell OAIS: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2010062438.

13 Leider fehlen an dieser Stelle belastbare Zahlen für einen Kostenvergleich. Faktoren für eine Wirtschaftlichkeitsbetrachtung verschiedener Aufbewahrungsformen (papierne Verwahrung, Mikrofilm, Digitalisierung und digitale Archivierung, Mikrofilm) liefert ein Beitrag von Steffen Schwalm: Schwalm, Steffen: Speicherung. Ermittlung der Wirtschaftlichkeit unterschiedlicher Aufbewahrungsformen, in: Ernst, Katharina (Hg.): Erfahrungen mit der Übernahme digitaler Daten. Bewertung, Übernahme, Aufbereitung, Speicherung, Datenmanagement: 11. Tagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen». Stuttgart 2007, S. 30-35. http://www.staatsarchiv.sg.ch/home/auds/11.html. Die Kommunale Gemeinschaftsstelle für Verwaltungsvereinfachung (KGSt) erstellte 1985 ein Organisationsgutachten, das einen Überblick über die einzelnen Faktoren gibt, die bei der Einrichtung eines Archivs zu berücksichtigen sind, darunter die notwendige Personalausstattung, Tätigkeiten/Arbeitsverteilung, Bauten/Räume, Kooperationen sowie Prozesse/Arbeitsabläufe. Banner, G. et. al., Kommunales Archiv. KGSt-Gutachten 1985. Die digitale Archivierung wird als kosten- und personalintensiver Faktor bei künftigen Organisationsmodellen mit zu berücksichtigen sein. Praktisches Hilfsmittel zur Errechnung von Aufwänden ist in diesem Zusammenhang die Arbeitshilfe der Bundeskonferenz der Kommunalarchive beim Deutschen Städtetag «Grundlagen kommunalarchivischer Arbeit» (2012). Hier werden für bestimmte Arbeitsbereiche (z.B. Vorfeldberatung oder Bewertung) Zeitaufwände (in Minuten) genannt. In Kombination mit den Personalkosten (die z.B. über die von der KGSt herausgegebenen Tabelle zu verschiedenen Personalstufen ermittelt werden können) ließen sich im Einzelfall Aufwände relativ genau beziffern. Die BKK-Arbeitshilfe ist abrufbar unter: http://www.bundeskonferenz-kommunalarchive.de/empfehlungen/Arbeitshilfe_Grundlagen_kommunalarchivischer_Arbeit_2014-06-14.pdf.

elektronische Archivierung findet überwiegend im Büro statt. Zudem entfällt mit Aufbau eines Langzeitarchivs nicht die Verpflichtung, auch weiterhin das analoge Aktenmaterial zu erhalten und zu betreuen. Fazit und unterm Strich wichtiges Argument gegenüber einer Verwaltung: Mit der elektronischen Langzeitarchivierung kommen neue zusätzliche Aufgaben dazu, d.h. es wird teurer, und das Archiv benötigt mehr Personal.[14]

Erst die Erfahrungen der nächsten Jahre werden zeigen, an welchen Stellen gegenüber der analogen Archivierung erhöhte Aufwände bei der elektronischen Archivierung entstehen. Eine vorsichtige Prognose wäre, dass v.a. im Bereich der «Vorfeldarbeit», bei der Betreuung der Aktenproduzenten/abgebenden Stellen und der Bewertung elektronischer Unterlagen höhere Aufwände zu erwarten sein werden, da die elektronischen Daten, Fachanwendungen usw. in ihrem technischen Umfeld verstanden und auf ihre Archivierbarkeit und die Möglichkeiten der späteren Nutzbarkeit geprüft werden müssen.[15]

Nicht vergessen werden darf, dass es mit Fortschreiten der Technik und der Digitalisierung der öffentlichen Verwaltungen immer wieder Daten geben wird, die erst aussonderungsfähig gemacht werden müssen. Ein elektronisches Langzeitarchiv kann noch so gut durchdacht sein, es wird niemals über alle technischen Erfordernisse für den Ingest künftiger Daten verfügen, sondern muss stets erweitert und angepasst werden. Nicht jede Fachanwendung, die in der Verwaltung zum Einsatz kommt, verfügt von vornherein über eine Exportschnittstelle. Solange das zuständige Archiv diese Daten als nicht archivwürdig einstuft, ist dies letztlich auch nicht erforderlich. Die Bewertungspraxis einzelner Archive ist jedoch unterschiedlich, sodass auch im Falle einer Verbundlösung der Verbund nur dort effizient agieren und sich für den Bau einer Aussonderungsschnittstelle oder eines Importkanals einsetzen kann, wo es eine ausreichend große Anzahl von Archiven gibt, die die Daten aus diesem speziellen Fachverfahren auch tatsächlich archivieren möchten.

Man kann durchaus die Frage stellen, ob dies Auswirkungen auf die künftige Überlieferungsbildung haben wird. Wenn nicht ausreichend Geld und Personal für ein Übernahmeprojekt vorhanden sind, die elektronischen Daten zu übernehmen,

14 «Digitale Archivierung erfolgt zusätzlich zur bisherigen Archivierung.» Zitat von P. Sandner (siehe Anm. 4), S. 62.

15 Zu den erforderlichen Arbeitsschritten im Vorfeld der Übernahme von Daten z.B. aus Fachverfahren oder Dokumentenmanagementsystemen wurden bereits verschiedene Erfahrungsberichte verfasst, als Beispiele: Worm, Peter: Standardisierung der Aussonderung aus den elektronischen Personenstandsregistern, in: Archivar 70 (2017), S. 9-15. http://www.archive.nrw.de/archivar/hefte/2017/Ausgabe_1/Archivar_1_2017.pdf. Konzen, Niklas: Übernahme von E-Akten aus kommunalen Dokumentenmanagementsystemen in das Langzeitarchiv DIMAG. Ein Vorschlag zur praktischen Umsetzung anhand von Fallbeispielen aus den DMS der Stadt Kirchheim unter Teck und des Landratsamts Karlsruhe (Transferarbeit 2016). https://www.landesarchiv-bw.de/sixcms/media.php/120/60857/Transferarbeit2016_Konzen.pdf.

d.h. archivfähig zu machen, dann werden sie möglicherweise nicht übernommen werden oder das Übernahmeprojekt wird solange vertagt, bis der technische Aufwand und die Kosten für die Archivierung in keinem vernünftigen Verhältnis mehr zueinander stehen. Die Folge: eine Überlieferungslücke. Betroffen können Daten aus einem Fachverfahren sein, aber auch Foto- oder Filmsammlungen, die in einem besonderen Bild- oder Filmformat vorliegen, für die das eigene Langzeitarchiv zu diesem Zeitpunkt noch keine Ingestmöglichkeit und Erhaltungsstrategie für die dauerhafte Lesbarkeit und künftige Nutzbarkeit der Daten bietet. Auch der Erhalt des Bitstreams könnte, sollten sich die Daten später als nicht mehr interpretierbar erweisen, auf diese Weise zum Datenverlust führen. Doch auch in den Fällen, in denen sowohl das jeweilige Langzeitarchiv über einen Eingangskanal als auch das exportierende System über eine Aussonderungsschnittstelle verfügt (z.B. eine XDomea2.2-Schnittstelle zur Übernahme strukturierter Daten aus Dokumentenmanagementsystemen), ist oftmals eine Anpassung des Mappings zwischen Export- und Importschnittstelle erforderlich. Diese Aufwände werden bei allen Archiven, unabhängig davon, ob sie ein eigenes Langzeitarchiv betreiben oder an einem Verbund partizipieren, künftig in höherem oder geringerem Maße anfallen. Im Gegensatz zur archivischen «Vorfeldarbeit» im analogen Bereich bedarf es bei diesen elektronischen Übernahmen des technischen Know-hows – Kenntnisse, die künftig bei der Archivausbildung noch deutlich stärker berücksichtigt werden müssen. Zugespitzt formuliert: Die digitale Langzeitarchivierung verzeiht keine Fehler oder Nachlässigkeiten in der Vorfeldarbeit. Einmal übernommene Daten sind dauerhaft archiviert und können nicht nachkassiert werden. Aus diesem Grund ist es unerlässlich, dass die Archivarinnen und Archivare künftig auch über das notwendige Verständnis zur Vorbereitung der elektronischen Daten für die Übernahme in das digitale Magazin verfügen und ausreichend Zeit in die Vorbereitung von Akten- bzw. Datenübernahmen investiert wird.

An anderer Stelle wird die Arbeit bei der elektronischen Archivierung möglicherweise aber auch effektiver: Vielleicht lassen sich Aufwände z.B. für Erschließung stellenweise bei der elektronischen Archivierung im Vergleich zum Umgang mit analogem Material reduzieren, indem Metadaten automatisiert in ein Archivinformationssystem übernommen werden können.[16] Und für die Archivierung von Fachverfahrensdaten oder bestimmte Datentypen gilt: Der anfängliche Aufwand, diese Daten archivfähig zu machen und entsprechende Einlieferungskanäle zu definieren, ist sehr groß. Sind diese Arbeiten jedoch einmal getätigt worden, reduziert sich der Aufwand künftiger Übernahmen aus diesen Fachverfahren bzw. besonderer

16 Schröter-Karin, Tobias: Vereinfachte Erschließung mit DiPS.kommunal. Artikel im Blog des LWL-Archivamtes vom 20.04.2017. https://archivamt.hypotheses.org/5008.

Datentypen auf ein Minimum. Lediglich Anpassungen sind vorzunehmen, wird an der Datenstruktur im Ursprungssystem etwas verändert. Diese Überlegungen müssten dann auch in künftige Kostenmodelle zur elektronischen Archivierung einfließen.

Im Folgenden wurde, ohne Anspruch auf Vollständigkeit, eine grobe Auflistung von Kostenfaktoren für den Aufbau und den Unterhalt eines elektronischen Langzeitarchivs erstellt, um zu zeigen, an welchen Stellen Aufwände entstehen und Kosten anfallen. Als Orientierung dienten dabei Überlegungen, die im Projekt DA NRW bzw. der Lösung DiPS.kommunal angestellt wurden. Auf dieser Grundlage wurde weiter kategorisiert: Wo entstehen die Aufwände/Kosten? Entstehen diese beim Betreiber, beim Rechenzentrum, beim Archiv/Kunden oder beim Softwareentwickler/Hersteller? Und es wurde nach Kostenarten unterschieden: Personalkosten, Kosten für Software-Lizenzen, Anschaffungskosten für Hardware sowie Wartungs- und Betriebskosten (wie Strom, Miete), Kosten für Dienstleistungen usw. Auf dieser Grundlage wurde ein Kosten- und Geschäftsmodell entwickelt, das die Basis der Konditionen bildet, zu denen eine Teilnahme am Verbund möglich ist.

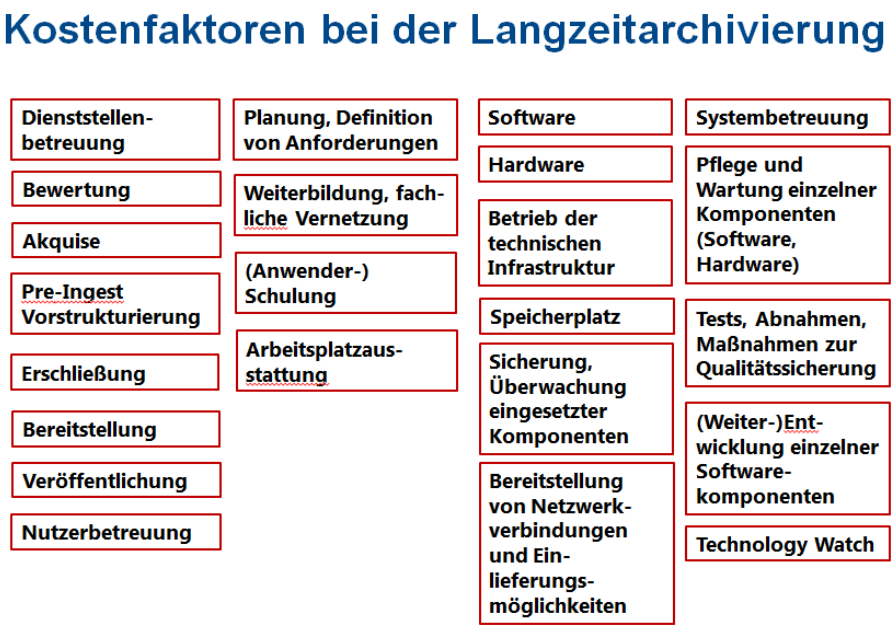

Abbildung 1: Kostenfaktoren bei der Langzeitarchivierung

Einige dieser Faktoren fallen genauso im analogen Bereich an. Auch ohne Langzeitarchiv müssen in den meisten Fällen Kosten z.B. für Anschaffung, Wartung und Sicherung von Software, z.B. eines Archivinformationssystems, berücksichtigt wer-

den. Aber man sieht: Es kommen einige Punkte im technischen Bereich hinzu, die bei der Langzeitarchivierung einkalkuliert werden müssen. Wenn eine entsprechende IT-Infrastruktur nicht vorhanden ist, muss diese entweder geschaffen werden oder es muss über die Mitnutzung bestehender Infrastrukturen nachgedacht werden. An dieser Stelle setzte das DA NRW mit seinen Überlegungen zur Schaffung einer Verbundlösung für Nordrhein-Westfalen an.[17]

Idee des Verbundes ist, dass seine Mitglieder viele Punkte aus Bereichen wie Wartung/Pflege oder Systembetreuung zu einem vorher kalkulierten Preis «einkaufen». Man geht davon aus, dass es für die Beteiligten insgesamt günstiger ist, eine bestehende Infrastruktur zu nutzen und anteilig für die Nutzung zu bezahlen, als diese Infrastruktur bei jedem einzelnen Nutzer separat zu entwickeln, aufzubauen und aufrechtzuerhalten. Vor allem aber ist es aus organisatorischer Sicht einfacher, einen Service in Anspruch zu nehmen, anstatt jede einzelne Aufgabe selbstständig und individuell zu lösen.

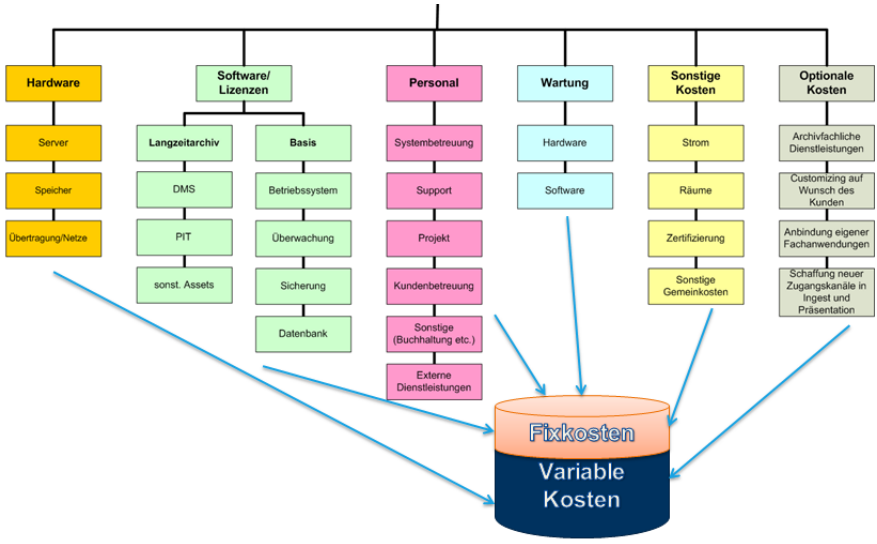

Abbildung 2: Kostenfaktoren für DiPS.kommunal im DA NRW

17 «Weil die notwendigen Sicherungsmaßnahmen für digitale Unterlagen inhaltlich komplex und technisch aufwändig sind, haben Land und Kommunen von Beginn an eine gemeinsame träger-, sparten- und institutionenübergreifende Lösung für eine Langzeitarchivierung angestrebt, die von allen Institutionen in NRW zur dauerhaften Sicherung ihrer digitalen Bestände genutzt werden kann. […] Das DA NRW ist ein informationstechnisches Angebot für alle Einrichtungen, die ihr elektronisches Kulturgut nach dem Archivgesetz und Pflichtexemplargesetz sicher und auf Dauer speichern müssen. […] Zu diesem Zweck arbeiten das Land NRW […] und der Zweckverband KDN – Dachverband kommunaler IT-Dienstleister […] als Arbeitsgemeinschaft zusammen, um eine wirtschaftliche Lösung zur Langzeitarchivierung digitaler und digitalisierter Kulturgüter in Nordrhein-Westfalen anbieten zu können.» Quelle: http://www.danrw.de/.

In Abbildung 2 werden die Kostenfaktoren für DiPS.kommunal, die die Betreiber der Lösung berücksichtigen müssen, nochmals zusammengefasst. Der Preis, den die Verbundteilnehmer zur Nutzung der Lösung bezahlen, errechnet sich aus den Aufwänden für die aufgeführten Bereiche/Arbeiten/Dienste.

Das Projekt Digitales Archiv NRW – eine Verbundlösung für Nordrhein-Westfalen[18]

Aufgrund der eingangs geschilderten Schwierigkeiten beim Aufbau eines eigenen digitalen Langzeitarchivs war bereits seit einigen Jahren abzusehen, dass insbesondere kleinere Kommunen große Probleme haben würden, sich selbstständig um eine gesetzeskonforme, fachgerechte Archivierung ihrer elektronischen Unterlagen zu kümmern.[19]

Daher wurde als gemeinsame Lösung für ein gemeinsames Problem in NRW seit 2009 ein Lösungsverbund entwickelt, der von Land und Kommunen gemeinsam getragen wird. Angestrebt wurde «eine gemeinsame träger-, sparten- und institutionenübergreifende Lösung für eine Langzeitarchivierung […], die von allen Institutionen in NRW zur dauerhaften Sicherung ihrer Bestände genutzt werden kann».[20] Das Land NRW arbeitet dabei mit dem Zweckverband KDN – Dachverband kommunaler IT-Dienstleister[21] zusammen, um insbesondere den nordrhein-westfälischen Kommunen eine unkomplizierte Beteiligung am Lösungsverbund zu ermöglichen. Die Finanzierung des Projekts (ca. 13,6 Mio. Euro für die Projektphase von 2015-2019) wird durch das Land NRW (51 Prozent) sowie die nordrhein-westfälischen Kommunen (49 Prozent) getragen, als leitendes Gremium des Projekts fungiert eine Arbeitsgemeinschaft aus Ministerium[22] und KDN. Die Zusammenarbeit wurde am 7. September 2015 vertraglich festgehalten.[23]

DiPS.kommunal als Spartenlösung für Archive stellt nur einen Teil dieses Lösungsverbundes dar, der mit der Open-Source-Lösung DNS (DA NRW Software

18 Umfangreiche Informationen zum Digitalen Archiv NRW und den darin vereinten Lösungen finden sich unter http://www.danrw.de/.

19 Vgl. das Fazit von Huppertz, Manfred: Besser im Verbund – Kooperationen im Bereich der elektronischen Langzeitarchivierung, in: Archivpflege in Westfalen-Lippe 79 (2013), S. 19-21. Zum Thema der Archivierung im Verbund auch Fischer, Ulrich: Gemeinsame Lösungen für ein gemeinsames Problem. Verbundlösungen für die elektronische Langzeitarchivierung in Deutschland, in: Archivpflege in Westfalen-Lippe 80 (2014), S. 20-25.

20 Vgl. https://www.danrw.de/ueber-das-da-nrw/die-nrw-loesung/, vgl. auch Anm. 17.

21 http://www.kdn.de/cms750/startseite/.

22 2012-2017: Ministerium für Familie, Kinder, Jugend, Kultur und Sport des Landes NRW, seit Juni 2017: Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen.

23 Vgl. Pressemitteilung des Digitalen Archivs NRW vom 07.09.2015, https://www.danrw.de/service/aktuelle-mitteilungen/artikel/news/nordrhein-westfalen-startet-sein-digitales-archiv/.

Suite) eine weitere, spartenübergreifende Langzeitarchivierungslösung im Angebot hat, um dem sparten- und trägerübergreifenden Anspruch gerecht werden zu können.

DiPS.kommunal und DiPS

DiPS.kommunal ist ein Verbundangebot, das sich aus der in mehreren Archiven bereits seit Jahren bewährten Lösung DiPS (Digital Preservation Solution) entwickelt hat. DiPS basiert auf derselben technischen Grundlage wie DiPS.kommunal, ist aber nur als Einzelinstallation lauffähig.

Zwar arbeiten alle DiPS-Nutzer (Bundesarchiv, Landesarchiv NRW, Landesarchiv Rheinland-Pfalz, Stadtarchiv Köln, Stadtarchiv Stuttgart, Archiv des LWL) in einem gemeinsamen Nutzerkreis zusammen, um insbesondere bei der Weiterentwicklung Synergien zu erzielen, jedoch unterhält jeder Nutzer eine eigene Infrastruktur mit der dazugehörenden Betreuung, um DiPS im eigenen Hause einsetzen zu können. Diese Infrastruktur stellt aber eben genau den Kostentreiber dar, der, neben den fachlichen Herausforderungen, die elektronische Langzeitarchivierung so aufwändig und teuer macht, weshalb eine eigene Installation für die meisten potentiellen Nutzer, gerade mittelgroße und kleinere Archive, nicht in Frage kommt. Aus diesem Grund haben sich vor einigen Jahren der LWL (LWL.IT Service Abteilung und LWL-Archivamt [das Archiv des LWL ist Teil des LWL-Archivamtes]) sowie die Stadt Köln (Historisches Archiv und Amt für Informationsverarbeitung) zusammengeschlossen, um die von ihnen bereits genutzte Langzeitarchivierungslösung DiPS für die nordrhein-westfälischen Archive zu einem Verbundangebot im Rahmen des DA NRW weiterzuentwickeln. In der Zwischenzeit ist diese Zusammenarbeit durch eine vertragliche Übereinkunft formalisiert worden.[24]

Von Archiven für Archive: Entwicklungsleitlinien für DiPS.kommunal

Ziel der Verbundlösung war einerseits, für das einzelne Archiv die Kosten durch den Kostentreiber Infrastruktur zu senken, indem große Teile der Infrastruktur zentral in den Betriebsstätten betrieben werden und die Benutzung von DiPS.kommunal weitgehend browserbasiert erfolgt.[25] Die Datenbestände der einzelnen Mandanten werden dabei selbstverständlich streng voneinander getrennt.

Andererseits musste die Ergonomie und Anwendbarkeit von DiPS so weiterentwickelt werden, dass auch die weniger technikaffinen Kolleginnen und Kollegen

24 Zu Geschichte und Gemeinsamkeiten von DiPS und DiPS.kommunal vgl. den Beitrag von Hoppenheit, Schmidt und Worm im Archivar 69 (2016): siehe Anm. 3.

25 Ausnahmen sind nur das Strukturierungswerkzeug PIT.plus sowie der Transferservice zur sicheren Datenübertragung, die lokal bzw. innerhalb der IT-Infrastruktur des Mandanten installiert werden müssen.

(nach einer entsprechenden Schulung) mit dem nötigen Werkzeug umgehen können. Angesichts der weit verbreiteten Befürchtungen, dass das Thema Digitale Langzeitarchivierung für den «Durchschnittsarchivar» zu komplex sein könnte, war die Vereinfachung der Benutzung ein besonderes Anliegen. Ohne einen niedrigschwelligen Einstieg in die Nutzung der nötigen Werkzeuge ist eine breite Akzeptanz einer entsprechenden Lösung nicht zu erreichen.

Obwohl DiPS grundsätzlich eine privatwirtschaftliche Projektentwicklung der Firma SER ist, wurde bei der Weiterentwicklung von DiPS zu DiPS.kommunal Wert darauf gelegt, dass die Entwicklungshoheit für DiPS.kommunal nicht bei der Firma SER, sondern bei der Entwicklergemeinschaft liegt. Zudem kommen, wie bereits in DiPS, primär nicht proprietäre und offen dokumentierte Metadatenschemata und Datenmodelle zum Einsatz, so dass im Fall der Fälle ein Umstieg auf eine andere Langzeitarchivierungslösung ohne Informationsverlust möglich wäre.[26]

Im Bereich der Ingestmöglichkeiten wird ebenfalls auf nachhaltige und kosteneffektive Lösungen gesetzt. Mit Hilfe einer Schnittstelle auf Basis des weitverbreiteten XDomea-Standards (Version 2.2) können aus jeglichen aktenproduzierenden Anwendungen wie Dokumentenmanagementsystemen, Fachanwendungen u.ä. Aussonderungen in das Digitale Magazin übernommen werden. Eine Rückmeldung in Form einer XDomea-Nachricht,[27] mit der die erfolgreiche Übernahme der ausgesonderten Daten in das Digitale Magazin quittiert werden kann, kann auch eine automatisierte Löschung der ausgesonderten Daten im Ursprungssystem anstoßen.

Mit Hilfe eines speziellen Tools, dem PIT.plus (Pre-Ingest-Toolset), können so gut wie alle anderen Daten, die ohne beschreibende Metadaten ausgesondert werden müssen, bearbeitet, beschrieben, strukturiert und übernommen werden. Auch für solche Daten wären Tools denkbar, die aus Daten in einer Ordnerstruktur eine XDomea-Lieferung formen. Eine solche Funktionalität könnte nützlich sein, wenn z.B. die Implementierung einer XDomea-Schnittstelle in einer Fachanwendung nicht möglich oder nicht gewollt ist, aber Daten (und Metadaten) in ein Dateisystem ausgesondert werden können.[28]

Ein individueller Eingangskanal existiert einzig für elektronische Personenstandsregister, die nach dem bundesweiten Standard XPSR 1.8 an die zuständigen

26 Zum verwendeten Datenmodell und den verwendeten Metadatenschemata vgl. den oben genannten Aufsatz von Hoppenheit, Schmidt und Worm (siehe Anm. 3). Das verwendete DiPS-Bundesarchiv-Schema findet sich unter http://www.digitalpreservationsolution.de/.

27 Nachrichtentyp 0506 «Aussonderung.AussonderungImportBestaetigen.0506». Vgl. Spezifikation XDomea v. 2.2.0, S. 439.

28 Ein entsprechender Prototyp für die Übernahme der Gebäudeakten des Bau- und Liegenschaftsbetriebs des LWL wird aktuell im Rahmen einer Bachelorarbeit in der LWL.IT Service Abteilung konzeptioniert und entwickelt.

Archive ausgesondert werden. Eine Aussonderung nach dem XDomea-Standard wäre hier weder möglich noch sinnvoll.[29]

Es ist zwar aus technischer Sicht denkbar, dass zukünftig weitere individuelle Übernahmeschnittstellen für spezielle Datentypen entstehen, aus fachlicher und ökonomischer Sicht soll diese Art der Individualisierung aber nach Möglichkeit vermieden werden, da sie zwangsläufig höhere Wartungs- und Anpassungskosten nach sich zieht, die gegenfinanziert werden müssten. Zudem ist es wahrscheinlich, dass die Gruppe der DiPS.kommunal-Entwickler zusammen mit den DiPS.kommunal-Anwendern mittelfristig eine gewisse Marktmacht entwickeln kann. Mit diesem Hebel könnten Fachverfahrenshersteller dazu bewegt werden, eine passende Aussonderungsschnittstelle in Richtung DiPS.kommunal zu entwickeln und diese im Rahmen der Produktpflege (weitestgehend) kostenneutral für die Anwender bereitzustellen. Eine generische und bereits im Praxisbetrieb (u.U. bei der Konkurrenz) erprobte Schnittstelle wie die XDomea-Schnittstelle stellt hier für Fachverfahrenshersteller eine wesentlich niedrigere Hürde dar als eine vollständige individuelle Neuentwicklung.

Auch bei der Anbindung der Erschließungssoftware wird eine generische Schnittstelle eingesetzt, die Basiserschließungsinformationen in Form einer XML-Datei im offen dokumentierten DiPS-Schema des Bundesarchivs erzeugt. Diese XML-Datei kann anschließend mit Hilfe eines einfachen Datenmappings und einer einfachen Datentransformation in die Erschließungssoftware eingespielt und dort in der tieferen Erschließung (soweit diese noch erfolgen muss) nachgenutzt werden. Das Archiv ist also in der Regel unmittelbar nach dem Ingest der ausgesonderten Daten und dem anschließenden Einspielen der Erschließungsinformationen auskunftsfähig, während größere Erschließungsrückstände zukünftig vermieden werden können.[30] Erfreulicherweise arbeiten die drei einschlägigen Erschließungssoftwarehersteller im kommunalen Bereich bereits an entsprechenden Schnittstellen bzw. haben sie bereits bereitgestellt. Immerhin zwei Hersteller stellen die entsprechende Schnittstelle ohne zusätzliche Lizenz- oder Wartungskosten für ihre Kunden bereit (es können allerdings noch Kosten anfallen, wenn die Softwarefirmen die Kunden bei der Anpassung des Datenmappings unterstützen müssen).

Die sichere Datenübertragung zwischen Mandanten und Betriebsstätte wird mit Hilfe eines so genannten Transferservices sichergestellt, eine Art Synchronisierungsdienst, der die Datenübertragung überwacht und protokolliert. Dieser Übertragungsweg ist auch für Archivgut mit besonderem Schutzbedarf (z.B. Personalakten)

29 Zur Aussonderung aus elektronischen Personenstandsregistern vgl. Worm, Standardisierung der Aussonderung aus den elektronischen Personenstandsregistern, siehe Anm. 15.

30 Vgl. Schröter-Karin, Tobias: Vereinfachte Erschließung mit DiPS.kommunal. Artikel im Blog des LWL-Archivamtes vom 20.04.2017. https://archivamt.hypotheses.org/5008.

geeignet. Die Authentifizierung zwischen Mandant und Betriebsstätte kann zurzeit durch Zwei-Faktor-Authentifizierung oder eine Verbindung über das Verbindungsnetz DOI[31] erfolgen.

Leistungsumfang des Angebots

Aktuell liegt die durch die Mandanten zu erbringende Kostenbeteiligung bei 19.100 € pro Jahr.[32] In der Kostenbeteiligung inbegriffen sind die geschilderten Eingangskanäle und Werkzeuge für strukturierte und unstrukturierte Daten, elektronische Personenstandsregister sowie die sichere Datenübertragung. Über die Bereitstellung der Basiserschließungsinformationen wird die beim Mandanten im Einsatz befindliche Erschließungssoftware in die OAIS-konforme Gesamtarchitektur eingebunden.

Jedem Mandanten wird zu Beginn ein Speicherplatz von 500 Gigabyte bereitgestellt. Weiterer Speicherplatz kann bei Bedarf für einen Preis von 0,26 € pro GB pro Monat (entspricht 1560 € pro Jahr pro 500 GB Speicherplatz) bereitgestellt werden. Wie lange die einzelnen Mandanten mit diesem Speicherplatz auskommen, muss sich in der Praxis zeigen. Bisher geht die Entwicklergemeinschaft davon aus, dass der Speicherplatz bei einer mittelgroßen Kommune drei bis fünf Jahre ausreichen sollte, wenn keine größeren Mengen an Digitalbildern oder audiovisuellem Material eingeliefert werden.

Die durch DiPS.kommunal-Entwicklergemeinschaft und -Anwenderkreis durchgesetzten Datenschnittstellen in Dokumentenmanagementsystemen, Fachanwendungen und Erschließungswerkzeugen sind nicht Teil des Angebots im ökonomischen Sinne, trotzdem bilden sie natürlich einen Teil des «Benefits» der Verbundlösung. Die DiPS.kommunal-Mandanten müssen hier zwar unter Umständen Anpassungen/Konfigurationen an bestehenden Schnittstellen finanzieren, sie müssen aber keine komplette Neuentwicklung finanziell einkalkulieren beziehungsweise bekommen solche Schnittstellen idealerweise sogar kostenneutral angeboten.

Beteiligungsmöglichkeiten

Die Teilnahme an DiPS.kommunal erfolgt über den KDN, aus dessen Leistungsangebot sich die mittelbaren und unmittelbaren Mitglieder ohne Ausschreibung bedie-

31 Das DOI-Netz (DOI steht für Deutschland-Online Infrastruktur) wurde als «verbindende Netzwerkstruktur (Koppelnetz) der Netze der Öffentlichen Verwaltung in Deutschland» errichtet, d.h., es ist ein staatliches Angebot zur Verbindung der öffentlichen Verwaltungseinrichtungen in Deutschland. Vgl.http://www.bva.bund.de/DE/Organisation/Abteilungen/Abteilung_BIT/Leistungen/IT_Produkte/VerbindungsnetzDOI/ProjektDOI/projektdoi_node.html.

32 Die Kostenbeteiligung dient dazu, die entstehenden Infrastrukturkosten zu decken. Mit der Kostenbeteiligung wird kein Gewinn erzielt.

nen können. Die beiden Betriebsstätten bei der Stadt Köln und beim LWL haben jeweils eine eigene regionale Zuständigkeit, die sich an den Zuständigkeiten der nordrhein-westfälischen Landschaftsverbände orientiert. Die Stadt Köln bedient dabei den Raum des Landschaftsverbands Rheinland, der LWL naturgemäß seinen eigenen Zuständigkeitsbereich.

Auch wenn die Verbundlösung DiPS.kommunal im Rahmen des Digitalen Archivs NRW grundsätzlich auf NRW ausgerichtet ist, kann die Dienstleistung DiPS.kommunal aufgrund der Mitgliedschaft des KDN in der Marketing- und Dienstleistungsgesellschaft der öffentlichen IT-Dienstleister in Deutschland ProVitako mittelfristig von allen mittelbaren und unmittelbaren ProVitako-Mitgliedern deutschlandweit abgerufen werden.[33]

Vor- und Nachteile von Verbundlösungen

Auch wenn es, wie oben gezeigt, sowohl in finanzieller als auch in organisatorischer Hinsicht vorteilhaft ist, eine bestehende Lösung zu nutzen und sich an einem Verbund zu beteiligen, kann die Teilnahme an einem Verbund durchaus auch mit Nachteilen verbunden sein, und nicht für jedes Archiv ist dies die optimalste Lösung.[34] Natürlich fallen mit der Beteiligung am Verbund Kostenfaktoren/Aufwandsposten weg, die man als Archiv mit einer «stand-alone-Lösung» mit bedenken muss. Auch ein Teil der Kosten für die Entwicklung weiterer Softwaremodule wird, je nach Organisationsstruktur, über den Verbund abgedeckt.[35] Allerdings sind damit die Möglichkeiten, das System an die eigenen Anforderungen optimal anzupassen, eingeschränkt. Im Interesse einer Einheitlichkeit, mit der Wartungs- und Pflegekosten unter Kontrolle gehalten werden können, nutzen Verbundteilnehmer ein und dieselbe Softwarelösung, ohne dass sich die einzelnen Komponenten voneinander groß unterscheiden. Als Nutzer einer Gemeinschaftslösung bzw. als Teil einer Entwicklergemeinschaft ist das Archiv so an Entscheidungen der Gemeinschaft gebunden und kann keine individuellen Wege gehen, ohne zusätzliche Aufwände und Kosten in Kauf zu nehmen. Bei Sonderkomponenten ist auch deren Pflege und Wartung

33 Entsprechende vertragliche Vereinbarungen zur Dienstleistungsüberlassung werden aktuell (Stand: August 2017) erarbeitet.

34 Fischer, Ulrich: Gemeinsame Lösungen für ein gemeinsames Problem. Verbundlösungen für die elektronische Langzeitarchvierung in Deutschland, in: Archivpflege in Westfalen und Lippe 80 (2014), S. 20-25, hier S. 21f.

35 Wobei an dieser Stelle nochmals an die «archivische Vorfeldarbeit» und die Vorbereitung von elektronischen Aktenaussonderungen erinnert sei. Mit der Planung von Aussonderungsschnittstellen oder auch den Anpassungen beim Import von Daten beispielsweise aus E-Akten-Anwendungen sind mögliche zeitliche, personelle und finanzielle Aufwände verbunden, die ein Archiv einkalkulieren muss, insbesondere, wenn es Daten übernehmen möchte, für die es (noch) keine standardisierten Eingangskanäle oder Übernahmeverfahren gibt.

nicht mehr über den Verbund abgedeckt. Ein Archiv mit besonderen Anforderungen, dessen zu archivierende Daten besonders strukturiert sind oder bei denen es sich um einen besonderen Datentyp handelt, muss prüfen, ob die Lösung, die auf diesem Weg genutzt werden kann, den Anforderungen genügt oder ob so viele Sonderanpassungen erforderlich wären, dass sich eine Beteiligung am Verbund nicht auszahlt.

Bei Verbundlösungen lassen sich verschiedene Konstrukte unterscheiden:[36]

- Archive können sich darauf verständigen, eine einheitliche Speicherlösung und eine einheitliche Software zu nutzen und sich in Form einer Entwicklergemeinschaft zusammentun.
- Archive können sich in einem Verbund für ein einheitliches Softwareprodukt entscheiden, jedoch jeweils eine eigene Speicherlösung wählen.
- Denkbar wäre auch die Nutzung einer eigenen Speicherlösung in Verbindung mit einem einheitlichen Softwareprodukt mit Sonderanpassungen, entsprechend der jeweils besonderen Anforderungen. Die genutzte Software würde sich also in bestimmten Komponenten von der «Basislösung» unterscheiden. Die DiPS-Nutzergruppe, die die von HP und SER entwickelte Lösung verwendet, ist hierfür ein Beispiel.

Im Falle von DiPS.kommunal meint Verbundlösung: Die am Verbund beteiligten Archive nutzen eine Speicherlösung und eine einheitliche Software.[37] Individuelle Anpassungen an der technischen Basis sind dabei nur in sehr geringem Maße möglich. Zur Minimierung der Aufwände bei Wartung, Pflege und Weiterentwicklung, aber auch Aufwänden für Einarbeitung und Schulung, wird eine Einheitlichkeit der verwendeten Systeme angestrebt. Im Normalfall gibt es hier keine abweichenden Teilkomponenten. Nur so können die Betreiber der Lösung gewährleisten, dass der in Anspruch genommene Service jederzeit bei allen Verbundteilnehmern dem vereinbarten Leistungsangebot und dem angestrebten Qualitätsstandard entspricht. Nichtsdestotrotz sind Sonderanpassungen möglich; diese müssen separat bezahlt (und auf eigene Verantwortung gepflegt) werden. Die dauerhafte Funktionsfähigkeit kann nicht mehr über den Verbund sichergestellt werden.

Fazit

Im Beitrag wurden einige wesentliche Kostenfaktoren benannt, um die zweifellos hohen Kosten der elektronischen Langzeitarchivierung nachvollziehbar zu machen. Als Grundlage dienten Überlegungen aus dem Projekt DA NRW, um die entstehenden Kosten zu kalkulieren, sowie Erfahrungen, die im Zuge der Entwicklung der

36 Fischer, Gemeinsame Lösungen für ein gemeinsames Problem, siehe Anm. 34, S. 21f.

37 Verbundlösung meint an der Stelle nicht den Archivverbund, bei dem sich Archive einen Archivar oder ein Gebäude teilen, den/das sie auf der Basis einer Vereinbarung gemeinsam finanzieren.

Lösung DiPS.kommunal in den vergangenen Jahren aufseiten des LWL und der Stadt Köln gemacht wurden. Am Beispiel von DiPS.kommunal wurde gezeigt, was eine Langzeitarchivierung im Verbund kosten kann. Eine Beteiligung an einem Verbund kann je nach Ausgangslage finanziell, aber v.a. aus organisatorischer Sicht lohnenswert sein, da Kosten für die Entwicklung eines Gesamtsystems entfallen und Aufwände für Betrieb, Sicherheit, Wartung oder Weiterentwicklung mit dem finanziellen Beitrag zur Beteiligung am Verbund abgegolten sind. Nicht in jedem Fall muss eine kooperative Lösung für ein Archiv jedoch die geeignete (und günstigere) Lösung sein.

Da die Archive auch im Zeitalter der Digitalisierung weiterhin verpflichtet sein werden, ihre analogen Unterlagen – ob Papier-, Foto- oder Filmdokumente u.a. – zu erhalten, ist die Archivierung elektronischer Daten eine Aufgabe, die neu hinzukommt und als Daueraufgabe auch bestehen bleiben wird. Vielleicht lassen sich Aufwände stellenweise bei der elektronischen Archivierung im Vergleich zum Umgang mit analogem Material reduzieren. Auf der anderen Seite kommen aber auch neue Aufwände hinzu, wie die Einarbeitung in die elektronische Archivierung und Tätigkeiten bei der «Vorfeldarbeit» (z.B. Fachverfahrensbewertung, Schnittstellenplanung). Vermutlich könnte man mit dem in den nächsten Jahren überall anstehenden Einführungs- und Transformationsprozess hin zur digitalen Archivierung in jedem Archiv mindestens eine/n Mitarbeiter/in die nächsten zwanzig Jahre voll auslasten. Dies bedeutet aber zugleich, dass sowohl für die Vorfeldarbeit wie auch für den Vorgang der Archivierung auch im elektronischen Zeitalter qualifiziertes Archivpersonal benötigt wird. Personalkosten und der Bedarf an archivischem Fachwissen bleiben bestehen – auch bei Beteiligung an einer Verbundlösung im Bereich der Langzeitarchivierung, denn dies enthebt den/die Archivar/in nicht der Verpflichtung, den Prozess der elektronischen Archivierung seiner archivwürdigen Daten von Beginn an bis zur Ablage auf dem sicheren Speicher fachlich zu begleiten.[38]

38 Fischer, Gemeinsame Lösungen für ein gemeinsames Problem, siehe Anm. 34, S. 23.

Chancen und Risiken verlustbehafteter Bildkompression in der digitalen Archivierung

Kai Naumann, Christoph Schmidt

Einleitung

Zu den Kernanliegen des archivischen Berufs gehört es, das dem Archiv anvertraute Archivgut vor Beschädigungen und Verlust zu bewahren. Es ist daher nicht verwunderlich, dass die willentliche Herstellung oder Duldung von Verlust in der Archivwelt nur selten ergebnisoffen diskutiert wird. Dies gilt auch und insbesondere für die Nutzung so genannter «verlustbehafteter» Bilddatenformate bei der digitalen Archivierung. Zumindest in der deutschen Archivcommunity besteht heute ein stillschweigendes Übereinkommen, den Einsatz entsprechender Techniken mehr oder minder explizit abzulehnen.[1] Gleichwohl prägen verlustbehaftete Bilddatenformate die außerarchivische digitale Landschaft in hohem Maße. So hat sich etwa JPEG in den vergangenen 20 Jahren zu einem Bilddatenformat entwickelt, das aus vielen Anwendungsgebieten gar nicht mehr fortzudenken scheint. Insbesondere im Kontext des Internet, aber auch in der Digitalfotografie ist JPEG omnipräsent, so dass die Archive in wachsendem Maße gezwungen sind, sich mit Anbietungen auseinanderzusetzen, die verlustbehaftete Bilddatenformate enthalten. Es ist daher nicht verwunderlich, dass das archivische Ablehnungsdogma ebenso langsam wie stillschweigend erodiert. So speichern viele Archive inzwischen Bilder im JPEG-Format, sofern dieses gleichzeitig auch das Produktions- bzw. Anbietungsformat ist. Eine mögliche Neubewertung verlustbehafteter Bilddatenformate ergibt sich jedoch auch in der Digitalisierung aus den wachsenden Bilddatenmengen, die in einigen Bereichen dazu führen, das Kriterium der ökonomischen Effizienz bei der Auswahl geeigneter Speicherformate vorsichtig neu zu gewichten.[2]

1 Diese Ablehnung spiegelt sich auch in verschiedenen Empfehlungen und Vorschriften wider, wie etwa den «Praxisregeln Digitalisierung» der Deutschen Forschungsgemeinschaft (Deutsche Forschungsgemeinschaft: DFG-Praxisregeln «Digitalisierung». DFG-Vordruck 12.151, o.O., [2016], S. 20-21) oder dem «Katalog archivischer Datenformate» der KOST (http://kost-ceco.ch/wiki/whelp/KaD/index.php). (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Als Beispiele wären hier etwa die besonders in den größeren Flächenländern sehr umfangreichen Luftbilddatenbestände der staatlichen Vermessungsverwaltungen zu nennen. In den von der Vermessungsverwaltung und den Archiven gemeinsam erstellten «Leitlinien zur bundesweit einheitlichen Archivierung von Geobasisdaten» wird der Einsatz verlustbehafteter Bildkompressionsverfahren zumindest nicht strikt abgelehnt (Leitlinien zur bundesweit einheitlichen Archivierung von Geo-

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.7

Vor diesem Hintergrund ist es das Ziel des vorliegenden Textes, einen ergebnisoffenen Dialog über verlustbehaftete Speicherformate anzustoßen und zu fördern. Da das gesamte Thema viele unterschiedliche Teilaspekte hat, die jeweils für sich genommen bereits eine intensivere Behandlung verdienten, wäre es im Rahmen des gewählten Publikationsformats unmöglich, diese erschöpfend zu diskutieren. Der Text beschränkt sich daher darauf, Anregungen und Denkanstöße zu bieten, Fragen aufzuwerfen und Vorschläge zu machen.

Am Anfang unserer Überlegungen stehen einige allgemeine Gedanken zum Verlustbegriff, der, wie zu zeigen sein wird, eng mit dem weniger negativ konnotierten Begriff der «Veränderung» verbunden ist. Danach soll anhand einiger Beispiele beleuchtet werden, welche Verluste beim Einsatz verlustbehafteter Kompressionsverfahren unter Laborbedingungen auftreten und wie diese zu bewerten sind. Unsere Darstellungen basieren dabei auf einigen praktischen Versuchen, die wir im Vorfeld der AUdS-Tagung 2017 unternommen haben.[3] Diese haben, darauf sei ausdrücklich hingewiesen, nicht den Anspruch wissenschaftlicher Beweisführung; sie sollen vielmehr zu eigenen Gedanken, Experimenten, Widersprüchen anregen. Nach der Vorstellung einiger praktischer Facetten im Umgang mit verlustbehafteten Kompressionsverfahren sollen dann einige Argumente diskutiert werden, die für den Einsatz verlustbehafteter Formate sprechen. Der Fokus der Überlegungen liegt dabei auf dem am weitesten verbreiteten Format JPEG. Ein knappes Fazit fasst die wesentlichen Ergebnisse des Beitrags zusammen.

Der Verlustbegriff in der digitalen Bestandserhaltung

Die Reprographie gehört zu den Disziplinen, die unter bestimmten Bedingungen Verluste in Kauf nehmen können, denn reprographische Verluste sind zunächst nur Qualitätsveränderungen, ohne dass damit eine Verschlechterung des angestrebten operativen Ergebnisses verbunden sein muss. Veränderungen können nämlich durchaus im Sinne des Erstellers sein, wenn etwa eine Wandlung in JPEG zwar im Detail zu Veränderungen führt, aber das Verschicken eines großen Digitalfotos per E-Mail erlaubt. Auch ein Archiv, das ja gleichsam eine Datei viele 100 Jahre in die Zukunft verschicken soll, kann sich überlegen, welche Qualitätsveränderungen erlaubt und welche unerwünscht sind.[4]

basisdaten. Abschlussbericht der gemeinsamen AdV-KLA-Arbeitsgruppe «Archivierung von Geobasisdaten» 2014-2015, o.O., [2015], S. 15-16).

3 Diese Versuche wurden von einigen Kollegen in den beteiligten Archiven tatkräftig unterstützt. Der Dank der Autoren hierfür gilt vor allem Corinna Knobloch (Landesarchiv Baden-Württemberg) und Martin Hoppenheit und Marcel Werner (Landesarchiv Nordrhein-Westfalen).

4 So auch der Fototechnische Ausschuss der Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA) in seinem Papier von Dezember 2016, allerdings ohne

Wer die Grenzen des Erlaubten definieren will, kann bei der Umwandlung von Rastergrafiken (Pixelgrafiken) in Folgeformate absolute Maßstäbe setzen. Jedes Pixel, das kann die Anforderung sein, soll seine Position im Bild und seine Farbe behalten. Dazu müssen stets auch das Farbmodell (z.B. RGB) und das Farbprofil (z.B. sRGB) erhalten bleiben. Das Ergebnis nennt man eine verlustfreie Formatmigration. Doch es können auch relative Maßstäbe gesetzt werden, die Veränderungen wie die berühmten JPEG-Kompressionsartefakte in gewissen Grenzen erlauben. Gibt es unter den relativen Maßstäben auch für das Archiv hinnehmbare Varianten?

In jedem Fall ist festzustellen, dass in den letzten Jahren für die Archive die Möglichkeiten zugenommen haben, Verluste sichtbar zu machen und sie zu bewerten. Das kostenlose Open-Source-Programmpaket ImageMagick steht bereit, um selbst weniger gebräuchliche Formate wie JPEG2000 zu erstellen und zu analysieren.[5] Die KOST hat für ImageMagick außerdem eine grafische Oberfläche namens KOST-Simy bereitgestellt, die ein Rasterbild Pixel für Pixel mit einem anderen Rasterbild vergleicht, das die gleiche Anzahl Pixel in Höhe und Breite besitzt. Die Differenz zwischen beiden Bildern wird in einer absoluten Zahl an Pixeln, einer prozentualen Quote an Pixeln (in Relation zur Gesamtzahl) und einer Rastergrafik mit Falschfarben festgehalten. Ein Toleranzwert für den Farbunterschied je Pixel und die Differenzquote können eingestellt werden (vgl. unten Versuchsaufbau 1, 3. Absatz).

Diese Differenz als Quote veränderter Pixel ist für Formatmigrationen von Rastergrafiken eine messbare Eigenschaft, die grundsätzlich ins Konzept der digitalen Bestandserhaltung gemäß Nestor-Leitfaden «Digitale Bestandserhaltung»[6] passt. Die Differenzzahl ermöglicht Aussagen zum Erfüllungsgrad der dort genannten Kriterien E1, E4, E5, E6 und E7, aber nur in einem übergreifenden Sinne. Die Kriterien einzeln abzuprüfen, ist mit dieser Zahl nicht möglich. Im Leitfaden wurde für Bilder insgesamt, das heißt sowohl Vektorgrafik als auch Rastergrafik (oder Pixelgrafik), ein Anforderungskatalog formuliert. In den hier nachfolgend dargestellten Experimenten war festzustellen, in welchem Ausmaß die Differenz auftritt, wie sie beeinflussbar ist und ob die im Nestor-Leitfaden genannten Erfüllungsgrade ausreichen, um die Differenz zu beschreiben. Nicht möglich war es, die im Leitfaden geforderten Prüfungen zu vollziehen. Entsprechend würde es sich empfehlen, in der Weiterentwicklung des Leitfadens Rastergrafik und Vektorgrafik zu unterscheiden und auf durch Software prüfbare Eigenschaften noch mehr Wert zu legen.

Überlegungen zur Kompression, http://www.bundesarchiv.de/DE/Content/Downloads/KLA/wirtschaftliche-digitalisierung.pdf?__blob=publicationFile.

5 http://www.imagemagick.org/.

6 Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung, Version 2.0. Verfasst und herausgegeben von der nestor-Arbeitsgruppe Digitale Bestandserhaltung, Frankfurt am Main 2012. http://nbn-resolving.de/urn:nbn:de:0008-2012092400.

Beispiele für den Einsatz verlustbehafteter Bildkompression unter Laborbedingungen

Versuchsaufbau 1

Das erste Beispiel, das experimentell im Staatsarchiv Ludwigsburg bearbeitet wurde, ist ein JPEG-Bilddatenstrom in einer PDF/A-Datei. Der PDF/A-Standard erlaubt seit seiner ersten Ausprägung PDF/A-1 das Einbetten von JPEG-Objekten. Die meisten Kopierer erstellen, wenn man eine farbige Vorlage in PDF/A scannt, eine solche Datei. Es ist also sehr wahrscheinlich, dass sich solche JPEG-Bilddatenströme in Archivbeständen finden und eines Tages (zum Beispiel im Jahr 2056) in ein Folgeformat migriert werden müssen. Da JPEG eine sehr ökonomische Relation zwischen Datenmenge und Pixelanzahl besitzt, würde eine Formatmigration in ein verlustfreies Format (wie heute z.B. PNG oder TIFF) eine erhebliche Vermehrung der Datenmenge um den Faktor 2 bis 3 mit sich bringen. Sollten JPEG-Datenströme in der Größenordnung von 3 TB vorliegen, müssten für die neue Repräsentation nicht weitere 3 TB, sondern weitere 9 TB bereitgestellt werden. Möglich ist, dass Speicherkosten dann keine Rolle mehr spielen werden. Es erscheint aber auch möglich, dass eine finanzielle Erwägung zu der Vorgabe führt, dass eine neue Repräsentation ebenfalls nur 3 TB oder besser weniger verbrauchen darf. Und ebenso könnte es bei einer zweiten Migration im Jahr 2098 sein.

Um dieses Szenario zu simulieren, wurde ein Testbild in Repräsentation R 1 zunächst vom JPEG-Format mit verlustbehafteter Kompression in JPEG2000 (R 2) umgewandelt und anschließend in JPEG (R 3) zurückumgewandelt. Beide Formate JPEG2000 und JPEG stehen hier nur stellvertretend für künftige verlustbehaftete Bildkompressionsformate der Jahre 2056 und 2098.

Es fanden mehrere Versuchsreihen statt, die sich hinsichtlich der verwendeten Qualitätsparameter bei der Kompression unterschieden. Die Differenz wurde jedes Mal mit KOST-Simy festgehalten, wobei der Toleranzparameter «M» (für Medium) zur Anwendung kam. Mit diesem Toleranzparameter werden Fehler ausgeworfen, wenn mehr als 0,001 Prozent der Pixel sich um 5 Prozent oder mehr von ihrem früheren Farbwert unterscheiden. Abweichungen unterhalb dieser Schwellen werden nicht wiedergegeben.

Nr.	R0	P1	D1	R1	P2	D2	R2	S
1	3,7 MB	90	< 0,001 %	2,9 MB	100	< 0,001 %	5,4 MB	12,0 MB
2	3,7 MB	80	< 0,001 %	1,4 MB	90	< 0,001 %	1,6 MB	6,7 MB
3	3,7 MB	90	< 0,001 %	2,9 MB	90	< 0,001 %	1,5 MB	8,1 MB
4	3,7 MB	90	< 0,001 %	2,9 MB	80	0,04 %	0,8 MB	7,4 MB
5	3,7 MB	80	< 0,001 %	1,4 MB	80	0,05 %	0,8 MB	5,9 MB
6	3,7 MB	70	< 0,001 %	0,9 MB	80	0,07 %	0,8 MB	5,4 MB
7	3,7 MB	60	0,009 %	0,6 MB	100	0,01 %	4,3 MB	8,6 MB
8	3,7 MB	50	0,06 %	0,4 MB	50	1,13 %	0,4 MB	4,5 MB

Tabelle 1: Versuchsreihe über Veränderungen an Pixeln, die über einen farblichen Unterschied von 5% pro Pixel hinausgehen, nach Formatmigrationsprozessen unter verschiedenen Parametern. (R1: Ausgangsgröße, P1: Parameter JPEG2000, D1: Differenz zu R0, R1: Neue Dateigröße, P2: Parameter JPEG, D2: Differenz zu R0, R3: Neue Dateigröße, S: Summe Dateigrößen R0-R2)

Es lassen sich folgende Beobachtungen machen:

Wie die Versuchsreihen 1 bis 3 zeigen, ist es im Lauf zweier verlustbehafteter Formatmigrationen nicht nur möglich, die Farbwerte im vordefinierten Toleranzbereich zu halten, sondern gleichzeitig den Speicherbedarf der neuen Repräsentationen zu verringern.

Versuchsreihen 4 bis 6 zeigen, dass mit dem JPEG2000-Algorithmus bei der ersten Wandlung noch wesentlich geringere Dateigrößen erreichbar sind, ohne dass die von KOST-Simy gesteckte Grenze überschritten wird. Der zweite Migrationsschritt nach JPEG verfehlt dann aber die Schwelle des Erlaubten.

Versuchsreihen 3 und 7 zeigen, dass eine zu intensive Anwendung des einen Algorithmus unerwünschte Folgen hat, wenn das Einhalten der Grenze beim Folgealgorithmus das Ziel sein muss. Während in Versuchsreihe 7 der Speicherbedarf der R 2 mit 0,6 MB sehr niedrig wird, zieht der Speicherbedarf der R 3 mit 4,3 MB über den Ausgangsbedarf der R 1 hinaus an, weil ein hoher Qualitätsparameter erforderlich ist, um die Differenzgrenze einzuhalten. Es liegt nahe, dies auf die für jeden Algorithmus spezifische Form der Kompressionsartefakte zurückzuführen. Dass der Folgealgorithmus diese Muster mit seinen eigenen spezifischen Mustern in Einklang bringen muss, führt zu größeren Datenmengen. Deshalb erscheint es sinnvoll, das Maß an Artefakten in allen Formatmigrationen auf das Nötigste zu beschränken.

Insgesamt ist die Versuchsreihe 2 diejenige, die unter Einhaltung der gewünschten Qualitätsstandards eine Sicherung mit dem geringstmöglichen Speicherplatz (für die drei Repräsentationen zusammen) erlaubt.

Versuchsaufbau 2

Ein weiterer Versuch, der im Technischen Zentrum des Landesarchivs NRW in Münster durchgeführt wurde, bestand darin, die Grenzen des Schadens einer wiederholten Anwendung verlustbehafteter Kompressionsverfahren zu ermessen. Im Experiment wurde eine JPEG-Datei ohne zusätzliche inhaltliche Bearbeitung immer wieder mit dem gleichen höchsten Qualitätsfaktor neu verlustbehaftet komprimiert.

Anzahl Neukompressionen	Veränderung gegenüber R 1
10	1,419 %
100	3,865 %
1000	3,874 %

Tabelle 2: Versuchsreihe über Veränderungen an Pixeln, die über einen farblichen Unterschied von 5 % pro Pixel hinausgehen, nach mehrfacher JPEG-Kompression. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Hierbei zeigte sich, dass die Veränderungsrate, die sich aus häufig wiederholten JPEG-Kompressionen ergibt, nach vielen Wiederholungen immer geringer ausfällt. Das heißt, die ersten Wiederholungen führen zu stärkeren Veränderungen als die späteren, bis schließlich nur noch minimale Veränderungen feststellbar sind. Bemerkenswert an dem Ergebnis dieses Versuches ist vor allem, dass sich die bildlich nachweisbaren Veränderungen bei reinen Neukompressionen zumindest im gewählten Beispiel in recht engen Grenzen halten und nicht exponentiell «aufschaukeln».

Versuchsaufbau 3

Ebenfalls aus Münster stammt das Beispiel einer Rastergrafik, die sich von ihrem Pixelmuster kaum für eine verlustbehaftete Formatmigration eignet. Das Experiment zeigte, dass sich diese mangelnde Eignung auch in Kennzahlen nachweisen lässt. Es handelt sich um eine standardisierte Landkarte im Maßstab 1:25000 (DTK 25), die aus nur 18 verschiedenen, flächig angelegten Farbtönen besteht. Die Farbtiefe betrug 8 Bit, die Grafik war als TIFF-Datei mit einer proprietären RLE-Kompression von Apple verlustfrei komprimiert. Die JPEG-Konversion führte zu Veränderungen, die in Tabelle 3 dokumentiert sind.

Datei	TIFF (R 1)	JPEG (R 2)
Größe/Pixel	3200x3200	3200x3200
Kompression	RLE (Mac-Variante)	JPEG
Verwendete Farbwerte	18	36504
Größe	1,1 MB	5,7 MB
Veränderung zu R 1	-	9,282 %
Farbtiefe	8 Bit	24 Bit

Tabelle 3: Versuch der Formatmigration von TIFF nach JPEG, der sich anhand von automatisiert erhobenen Kennzahlen als Fehlschlag identifizieren lässt. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Dass eine JPEG-Umwandlung in diesem Fall kein erfolgversprechender Weg ist, zeigt sich an der erheblichen Veränderungsrate gegenüber dem Ausgangsbild, an der Vergrößerung der Datenmenge und an der Vervielfachung der Farbwerte. Interessant ist, dass bei dem hier vorliegenden standardisierten Bildtyp DTK 25 jede Vermehrung der effektiven Farbwerte über 18 hinaus einen Informationsverlust bedeuten würde, denn die Farbwerte haben im kartografischen Kontext jeweils eine exakte Bedeutung (z.B. Hellblau für Gewässer).

Versuchsaufbau 4

Der vierte Versuch, der im Staatsarchiv Ludwigsburg stattfand, beschäftigt sich mit dem Verhältnis zwischen rein digitalen Prozessen und den analog-digitalen Übergangsprozessen beim Scannen. Im Kollegenkreis waren die Verfasser darauf hingewiesen worden, dass die meisten Scanner nicht in der Lage sind, eine Vorlage zweimal in eine identische Bitfolge zu verwandeln.

Für den Versuch wurde eine aquarellierte Zeichnung im Folioformat zweimal unmittelbar hintereinander ohne Verrücken der Vorlage gescannt. Der Unterschied der beiden Scans betrug nach den oben (Versuch 1) genannten Maßstäben 1,388 % der Pixel. Bei anderen Scannermodellen, an denen der gleiche Versuch geplant war, war ein Vergleich der zwei erstellten Scans gar nicht möglich, weil diese Scanner jedes Mal eine leicht unterschiedliche Anzahl von Pixeln auswarfen.

Aus diesem Versuch ergibt sich, dass bei der Migration von einem Muster auf physischen Trägern in eine digitale Rastergrafik oft eine Variationsbreite vorliegt, die die Veränderungsrate bei verantwortungsvoll durchgeführten verlustbehafteten Formatmigrationen erheblich übersteigt. Mit anderen Worten: ein Scan hinterlässt unter Umständen viel mehr Qualitätsunsicherheiten als eine verlustbehaftete Formatmigration – nur weil dies bislang weniger bekannt ist, kann man es nicht ignorieren.

Abbildung 1: Ein daumennagelgroßer Ausschnitt aus dem Scan eines Aquarells.

Abbildung 2: Differenzbild zwischen Scan und JPEG-komprimierter Darstellungsform. Schwellenwerte wie in Versuchsaufbau 1, 3. Absatz.

Abbildung 3: Differenzbild zwischen Scan A und Scan B, ohne Verrücken der Vorlage. Schwellenwerte wie in Versuchsaufbau 1, 3. Absatz.

Versuchsaufbau 5

Ein oft vorgebrachtes Argument gegen komprimierte Dateiformate insgesamt zielt auf deren fehlende «Robustheit» ab. Gemeint ist damit die Toleranz eines Dateiformats gegen Bit-Rot, also: die ungeplante Veränderung von Bits durch technische Defekte. Tatsächlich lassen sich Unterschiede in der Robustheit empirisch nachweisen.

Im Technischen Zentrum des Landesarchivs NRW in Münster wurden hierzu jeweils 50 identische TIFF- und JPEG-Dateien einem künstlichen Bit-Rot unterzogen, indem einzelne Bits bzw. Bitfolgen verändert wurden. Danach wurden die nachweisbaren Bildinformationsverluste mit den bereits bekannten Mitteln gemessen.

Bit-Rot in %	Verlustquote in Pixeln TIFF (unkomprimiert)	Verlustquote in Pixeln JPEG
0,01 %	55,76 %	99,86%
0,001 %	10,46 %	97,43%
0,0002 %	2,81 %	92,79%

Tabelle 4: Auswirkungen von Bit-Rot auf die Bildqualität unkomprimierter und komprimierter Dateiformate. Veränderungsraten ermittelt mit KOST-Simy, Parameter vgl. Versuchsaufbau 1, 3. Absatz.

Das Ergebnis des Versuchs ist eindeutig: Je dichter das Pixelmuster im Datenstrom komprimiert ist, desto schneller können auch kleine Veränderungen zu großen Problemen führen. So vertragen bei den hier in Frage stehenden Bildformaten unkomprimierte TIFF-Dateien im Schnitt eine deutlich höhere Anzahl fehlerbedingter Veränderungen als stark komprimierte JPEG-Dateien.[7]

Daraus zu folgern, dass unkomprimierte Dateiformate für den Langzeiterhalt per se besser seien als komprimierte, ist jedoch problematisch. Denn mit etwas Pech kann bei jedem Dateiformat die Veränderung bereits sehr weniger einzelner Bits zu einem kompletten Informationsverlust führen. Zudem basiert die Bewertung eines Dateiformats nach Robustheit auf der Annahme, Bit-Rot sei in einem Langzeitarchiv ein akzeptables oder zumindest ein unvermeidliches Phänomen. Dies entspricht weder dem fachlichen Diskussionsstand darüber, wie ein vertrauenswürdiges Langzeitarchiv einzurichten sei, noch der tatsächlichen Praxis. Vertrauenswürdige digitale Archive begegnen der Gefahr des Bit-Rot mit ausgereiften und bewährten Mechanismen der technischen Bitstream Preservation. Weswegen also sollte Robustheit als Qualitätskriterium für Langzeittauglichkeit aufrechterhalten werden? Der Archivinformatiker Gary McGath meinte dazu kürzlich in seinem Blog:

> *«Banning compression from archives in the hope of minimizing the damage from bit rot is a foolish preservation strategy.»*[8]

Gute Gründe für JPEG?

Die geschilderten Laborversuche, die sich problemlos um einige zusätzliche Szenarien erweitern ließen, haben den Fokus unserer bisherigen Überlegungen auf die Diskussion möglicher Risiken bestimmter Bildkompressionsverfahren gelegt, und zwar immer mit Blick auf die Gefahr des Informationsverlusts. Das Kriterium des

7 Vgl. hierzu auch: Heydegger, Volker: Just One Bit in a Million: On the Effects of Data Corruption in Files. In: Agosti, Maristella (u.a.) (Hg.): Research and Advanced Technology for Digital Libraries. ECDL 2009 Berlin / Heidelberg 2009, S. 315-326.

8 McGath, Gary: Bit-rot tolerance doesn't work, in: Mad File Format Science [Blog]: https://madfileformatscience.garymcgath.com/2016/10/18/bit-rot-tolerance/

(begrenzten) Informationsverlusts kann jedoch bei einer an den signifikanten Eigenschaften eines Objekts und der ökonomischen Leistungsfähigkeit eines Archivs ausgerichteten Bestandserhaltungsstrategie nur ein Entscheidungskriterium unter mehreren sein. In diesem abschließenden Kapitel sollen daher einige Aspekte angesprochen werden, die möglicherweise für die Archivierung verlustbehaftet komprimierter Dateien in einem digitalen Langzeitarchiv sprechen.

Konkret geschieht dies am Beispiel JPEG. Denn abgesehen davon, dass JPEG bereits ohnehin in den Archiven angekommen ist, gibt es eine ganze Reihe guter Gründe, JPEG als Langzeitformat in Betracht zu ziehen. So ist JPEG offen standardisiert[9] und äußerst weit verbreitet, was eine sehr hohe Restlebenszeit des Formats und eine geringe Verdrängungsgefahr erwarten lässt. Viewer, Konverter und Werkzeuge zur Qualitätssicherung sind in großer Zahl frei verfügbar – ein Vorteil, der JPEG insbesondere gegenüber JPEG2000 auszeichnet. Durch seine hohe Verbreitung wird die irgendwann anstehende Aufgabe der Formatmigration zudem große Teile der IT-Welt betreffen, zumindest, sofern diese ein Interesse am Erhalt älterer Datenbestände hat. Die Archive werden diese Aufgabe also technisch nicht alleine bewältigen müssen.

Last but not least bieten hochkomprimierte Dateiformate ökonomische Vorteile in allen Funktionsbereichen des digitalen Archivs. In der Archivwelt, die gerne (vermeintliche) fachliche Optimalstandards ohne eine kritische Überprüfung der Angemessenheit im Einzelfall einfordert, wird über den Einfluss ökonomischer Faktoren auf Entscheidungsprozesse nur ungern gesprochen. Dieses Ethos wird durch die deutsche Archivgesetzgebung zwar gestützt[10] – gleichwohl sind öffentliche Archive einer begrenzten Budgetierung unterworfen. Gerade in größeren Archiven werden diese Grenzen in der Begeisterung für einzelne Arbeitsvorhaben gerne ausgeblendet. Da sie aber trotzdem existieren und das Gesamtmaß der Gestaltungsmöglichkeiten bestimmen, besteht ohne eine reflektierte und verzahnte Steuerung des gesamten Arbeitsbereichs die Gefahr der «kalten Kassation». Anders ausgedrückt: In einer Welt begrenzter Ressourcen verhindert jede Entscheidung für eine Maßnahme die Durchführung einer anderen Maßnahme. Dieses Phänomen ist naturgemäß nicht zu verhindern, ließe sich jedoch durch eine flexible Prüfung fachlicher Angemessenheit unter Berücksichtigung der vorhandenen Ressourcen kontrollieren und strategisch steuern.

9 Die JPEG-Komprimierung ist normiert in ISO/IEC 10918.

10 So legt z.B. das Archivgesetz des Landes Nordrhein-Westfalen fest, dass über die Archivwürdigkeit angebotener Unterlagen «das zuständige Archiv unter Zugrundelegung fachlicher Kriterien» entscheidet (§ 2 Abs. 6 Satz 2 ArchivG NRW). Entsprechende Regelungen finden sich in den meisten anderen deutschen Archivgesetzen.

Bei der digitalen Archivierung könnte die Wahl eines Rasterbildformats, das den signifikanten Objekteigenschaften angemessen ist, eine wichtige Stellschraube gegen die «kalte Kassation» sein.

Hierzu ein konkretes Rechenbeispiel. In Nordrhein-Westfalen wird seit kurzem unter dem Dach des Lösungsverbundes «Digitales Archiv NRW» eine mandantenfähige digitale Archivierungslösung für Kommunen angeboten. Die Kosten für die Nutzung dieses Systems «DIPS.kommunal» belaufen sich für ein Archiv als Kunde voraussichtlich auf etwa € 20.000,- pro Jahr. In dieser Pauschalsumme enthalten sind 500 GB Netto-Datenspeicher; jedes weitere GB kostet € 3,12 pro Jahr.

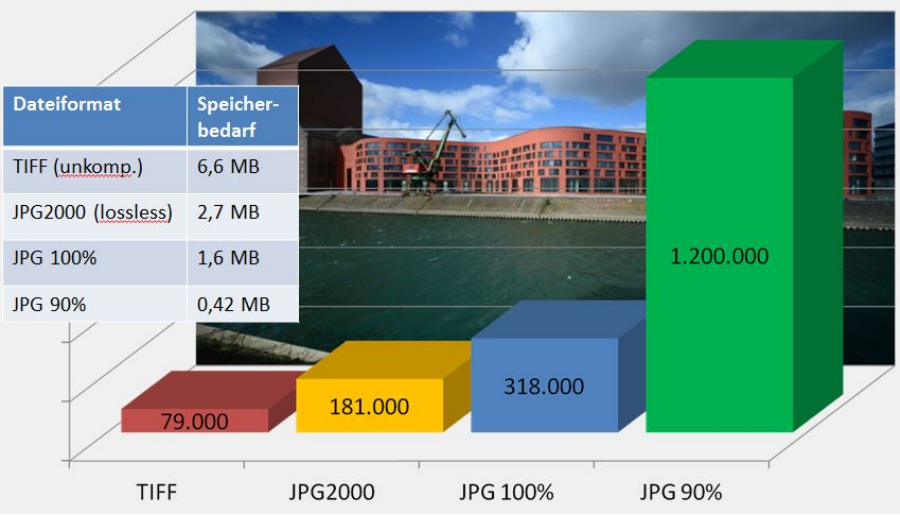

Abbildung 4: Beispielbild für unterschiedliche Speichervolumen

Das gezeigte Beispielbild ließe sich bei gleicher Auflösung als unkomprimierte TIFF-Datei mit einer Größe von 6,6 MB ca. 79.000 mal speichern, als verlustfreie JPEG2000-Datei mit einer Größe von 2,7 MB ca. 181.000 mal, als JPEG mit geringster Komprimierung und einer Größe von 1,6 MB rund 318.000 mal und als JPEG mit 90% Qualität und einer Größe von 0,42 MB etwa 1,2 Millionen mal. Die jährlichen Speicherkosten pro Bild betrügen (gerundet) € 0,25 (TIFF), € 0,11 (JPEG2000), € 0,06 (JPEG 100%), € 0,02 (JPEG 90%).

Der Einsatz verlustbehafteter Bildkompression kann somit in Archiven ohne unbegrenzte Ressourcen vor dem Totalverlust durch «kalte Kassation» schützen.

Fazit

Welche Erkenntnisse und Anregungen lassen sich nun aus den vorgestellten Überlegungen zum Umgang mit verlustbehafteten Bilddatenformaten festhalten?

Zunächst die Erkenntnis, dass ein geeignetes Instrumentarium für Vergleiche verschiedener Rasterbildformate zur Verfügung steht. ImageMagick und KOST-Simy ermöglichen sowohl eine vollautomatisierbare, in den Konvertierungsworkflow integrierbare Qualitätskontrolle für größere Datenmengen als auch händische Prüfungen mit Hilfe einer grafischen Benutzeroberfläche. Mit automatisiert erhobenen Kennzahlen lassen sich der Erfolgsgrad und der Umfang eingetretener Bildveränderungen einer verlustbehafteten Formatmigration bestimmen und bewerten. Auch lässt sich, basierend auf den Eigenschaften des Bilds, ein Optimum zwischen der Kennzahl Speicherbedarf und den Kennzahlen für Qualität ermitteln. Falsche Parameter können in diesem Prozess erkannt und korrigiert werden. Als allgemeine Regel könnte z.B. gelten: «Wenn ein Bild von Format A in Format B überführt wird, und sich mehr als 0,001 Prozent der Pixel verändern, sind Parameter oder Algorithmus ungeeignet.» Die Qualitätsparameter entsprechender Regeln sind bedarfsorientiert und nach fachlich wie ökonomischen Maßstäben zu bestimmen. Zum zweiten haben die Laborversuche verdeutlicht, dass die Entscheidung über den Einsatz verlustbehafteter Bilddatenformate am besten einzelfallabhängig zu treffen ist und von den angenommenen signifikanten Eigenschaften des digitalen Archivguts abhängig sein sollte.

Für die Weiterentwicklung des nestor-Leitfadens Bestandserhaltung wirft dies die Frage auf, ob die dort genannten, recht abstrakten Kriterien für die Auswahl von Bilddatenformaten mit den Erfüllungsgraden «Ja» / «Nein» /«für das menschliche Auge nicht erkennbar» in dieser Form praktisch ausreichend sein können. Als belastbarer hat sich in den Versuchen der Einsatz von Schwellenwerten erwiesen, die sich aus Kennzahlen zu Veränderungsgrad, Anzahl der Farbwerte und Datenmenge speisen. Es könnte daher sinnvoll sein, den im Leitfaden genannten Informationstyp «Bild» in zwei oder mehr Informationstypen aufzuspalten, zu denen dann passende Schwellenwerte definiert werden können. Nachdem schon Veronika Krauß 2016 in Potsdam wertvolle Anstöße in diese Richtung gegeben hat,[11] konnten auch die Verfasser dieses Beitrags nur einige weitere Schritte gehen.

Drittens hat sich in den Laborversuchen zumindest JPEG als ein Format erwiesen, das auch bei mehrfacher Kodierung und Dekodierung relativ widerständig

11 Krauß, Veronika; Bahrami, Arefeh: Ist das Bild noch das Bild? Authentizität digitaler Objekte unter Formattransformationen in Kooperation mit dem Thüringischen Hauptstaatsarchiv, Vortragsfolien von der 20. Tagung des Arbeitskreises Archivierung von Unterlagen aus digitalen Systemen am 1./2.3.2016 in Potsdam, online unter http://www.staatsarchiv.sg.ch/home/auds/20.html. Tagungsband erscheint demnächst.

gegen größere Veränderungen des Pixelbestandes ist. Insbesondere die mehrfache Neukodierung ohne willentliche Manipulation des Bildes führte zumindest im gewählten Beispiel nicht zu katastrophalen Pixelveränderungen. Umgekehrt formuliert: Die unabsichtliche Erzeugung von Artefakten, die mit dem menschlichen Auge erkennbar sind, ist schwieriger als erwartet. Ob die eingetretenen Veränderungen akzeptabel sind oder nicht, hängt freilich von den zuvor bestimmten signifikanten Eigenschaften des Informationsobjekts ab.

Viertens ist deutlich geworden, dass insbesondere im Umgang mit Digitalisaten die Bildqualität nur teilweise vom gewählten Bilddatenformat abhängig ist: So greift ein Scanprozess unter Umständen viel stärker und vor allem schwerer kalkulierbar in die Abbildungstreue einer Grafik ein als eine kontrolliert durchgeführte verlustbehaftete Formatmigration.

Last but not least hat sich gezeigt, dass der Einsatz eines verlustbehafteten Bilddatenformats wie JPEG auch einige gravierende Vorteile in der digitalen Langzeitarchivierung mit sich bringen kann. Wir glauben verstanden zu haben, dass die verlustbehaftete Kompression weniger Risiken birgt als bislang vermutet und einige Vorteile, insbesondere hinsichtlich der Langzeitverfügbarkeit des Formats und seiner ökonomischen Perspektiven, mit sich bringt. Gleichwohl bedarf es im Umgang mit verlustbehafteten Bilddatenformaten eines sehr sorgfältigen Risikomanagements, welches unbedingt auf genauen Kenntnissen der dem Format zu Grunde liegenden Technik basieren muss. Eine verlässliche Bestandserhaltung setzt voraus, dass man das Material, mit dem man es zu tun hat, gut kennt!

Wer genau hinschaut, stellt fest, dass verlustbehaftete Kompression nicht nur in den Reprowerkstätten der Archive schon längst zum Arbeitsalltag gehört. Langsam bildet sich ein intuitives Verständnis für ihre Möglichkeiten heraus, das uns zwar nicht vor kleinen Fehlern, aber vor großen Katastrophen bewahrt und uns vor dem Hintergrund hoher Speicherkosten ein Stück weit mehr Handlungsfähigkeit verschafft.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.7

TIFF-Korpus-Analyse

Martin Kaiser, Claire Röthlisberger-Jourdan, Georg Büchler

Ausgangslage

TIFF ist gegenwärtig und seit langem das meistgebrauchte Format zur Archivierung von unkomprimierten Bilddaten.[1] Es handelt sich um ein flexibles, anpassungsfähiges Dateiformat, das über die Jahre eine Vielzahl von Erweiterungen und Ergänzungen erfahren hat. Daneben bietet es die Möglichkeit, Metadaten in anderen Standards (wie IPTC, EXIF oder ICC) einzubetten. Diese Flexibilität und Ausprägungen machen TIFF jedoch zu einem komplexen Dateiformat und bergen Risiken für die digitale Archivierung. Die Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen KOST hat deshalb bereits 2014 eine Empfehlung zum Preservation Planning für TIFF-Dateien publiziert, welche basierend auf der Baseline-TIFF-Spezifikation ein archivtaugliches TIFF zu definieren versucht.[2] 2015 bis 2017 strebte ein gemeinsames Projekt[3] des *Digital Humanities Lab* an der Universität Basel (DHLab), der *Universität Girona* und der Firma *Easy Innova* an, eine erweiterte Baseline-Spezifikation in eine *ISO Recommendation* zu überführen, um so dem Umstand abzuhelfen, dass TIFF eine offene Spezifikation der Firma Adobe, jedoch kein ISO-Standard ist.[4]

Damit eine solche Empfehlung nicht nur auf theoretischen Überlegungen beruht, sondern sich auf eine fundierte Analyse echter archivischer Daten stützen kann, haben es die KOST und das DHLab im Rahmen dieses Projekts unternommen, mehrere Millionen Dateien aus drei Archiven systematisch zu untersuchen. Parallel dazu wurden an diesem Korpus auch etliche bekannte und in der Archivwelt verbreitete Analysetools getestet.

---

1 Dieser Beitrag ist eine überarbeitete Version des Artikels «TIFF-Korpus-Analyse» auf der KOST-Website, https://kost-ceco.ch/cms/index.php?tiff-data-analysis_de. Wir danken den Kollegen in den beteiligten KOST-Trägerarchiven für ihre Mitarbeit an der Korpus-Analyse: Marcel Büchler (Schweizerisches Bundesarchiv); Martin Lüthi und Vedat Akgül (Staatsarchiv St. Gallen); Markus Loch und Lambert Kansy (Staatsarchiv Basel-Stadt). Unser Dank geht ebenfalls an das Projektteam des TI-A-Projekts, im Besonderen an Peter Fornaro und Erwin Zbinden vom Digital Humanities Lab der Universität Basel. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 https://kost-ceco.ch/cms/index.php?preservation_tiff_de.

3 http://ti-a.org/.

4 Ein entscheidender Nachteil dieser Konstellation zeigt sich seit Ende 2016: Die aktuelle TIFF-Spezifikation (Version 6.0), früher publiziert unter http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf, ist kommentarlos von der Adobe-Website verschwunden. Sie bleibt an anderen Quellen greifbar, zum Beispiel beim Internet Archive unter https://web.archive.org/web/20091223030231/http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.8

Korpus und Fragestellungen

TIFF-Korpus

Die Staatsarchive Basel-Stadt und St. Gallen und das Schweizerische Bundesarchiv stellten für die Untersuchung ihre TIFF-Sammlungen von je etwa 12 TB zur Verfügung. Die Bestände sind, was Alter, Grösse und Ursprung betrifft, in allen Archiven sehr heterogen. Eine summarische Aufstellung zeigt Tabelle 1.

Staatsarchiv Basel-Stadt		
Typ	Anzahl	Grösse
2000	1‘950	
2001	2‘300	
2002	200	
2003	100	
2004	10‘000	
2005-2015	750‘000	
Total	**764‘550**	**12.4 TB**
Staatsarchiv St. Gallen		
	Anzahl	Grösse
Total	**870‘000**	**12.83 TB**
Schweizerisches Bundesarchiv		
Typ	Anzahl	Grösse
Archiv-TIFFs	1‘700‘000	5-6 TB
Digitalisate	7300	0.46 TB
Digitalisate von Dritten	14‘000‘000	6 TB
Total	**15‘707‘300**	**11-12 TB**

Tabelle 1: Anzahl und Grösse der analysierten TIFF-Dateien

Fragestellungen und Analyseprogramme

Für die Analyse der TIFF-Dateien wurden die folgenden Programme ausgewählt:

MD5-Berechnung	MD5 Das Berechnen des MD5-Schlüssels gehört nicht zu den Analysemodulen, garantiert aber die Lesbarkeit der Datei.
Formaterkennung	file http://gnuwin32.sourceforge.net/packages/file.htm Mit der Formaterkennung durch file werden falsch gekennzeichnete Dateien erkannt.
Formatvalidierung	JHOVE http://jhove.openpreservation.org/

	Die JHOVE-Validierung ermittelt die grundlegende Struktur der TIFF-Datei. Wichtig sind hier Status und InfoMessage.
Formatvalidierung	DPF-Manager http://www.preforma-project.eu/dpf-manager.html Der DPF-Manager ist eine Alternative zu JHOVE aus dem PREFORMA-Projekt.
Validierung und TIFF-Tag-Extraktion	checkit_tiff: a conformance checker for baseline TIFFs https://github.com/SLUB-digitalpreservation/checkit_tiff checkit_tiff wurde von der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden entwickelt.
TIFF-Tag-Extraktion	tiffhist http://dhlab.unibas.ch/ Das vom DHLab entwickelte C++-Programm extrahiert alle TIFF-Tags in eine CSV-Tabelle (TIFF-Tag, Datentyp und Wert)
EXIF-Extraktion	ExifTool (EXIF-Extraktion) http://owl.phy.queensu.ca/~phil/exiftool/ Eingebettete EXIF- und XMP-Metadaten werden extrahiert.
Thumbnail-Generierung	ImageMagick http://www.imagemagick.org/ Für jede Datei wird mit ImageMagick ein sehr kleines Thumbnail generiert. In diesem Schritt wird somit die Payload oder Bitmap der TIFF-Datei untersucht. Eine erfolgreiche Konvertierung belegt die korrekte Implementierung von Komprimierung und Farbraum.

Tabelle 2: Verwendete Analyseprogramme

Vorgehen im Detail

Weil davon auszugehen war, dass viele Dateien noch einer Schutzfrist unterstehen oder urheberrechtlich geschützt sind, sollten die zu untersuchenden Korpora die beteiligten Archive nicht verlassen. Auch sollten über Dateinamen oder Pfadnamen keine Rückschlüsse auf die Archivbestände gezogen werden können. Deswegen wurden in einem ersten Schritt die TIFF-Dateien aus dem jeweiligen Archivsystem auf USB- oder NAS-Platten kopiert und diese anschliessend für die weitere Untersuchung vom Netzwerk des jeweiligen Archivs abgehängt. Das Kopieren nahm wegen organisatorischer Herausforderungen und wegen der Datenmenge etwa 3 Monate in Anspruch.

Um die Anforderungen an den Umgang mit grossen Datenmengen, langen Programmlaufzeiten und sicherer Anonymisierung erfüllen zu können, wurde beschlossen, Programmausführung und Logverwaltung mit einer Datenbank und einem speziellen Analyse-Loop-Programm zu realisieren. Dieses musste sowohl im Linux- als auch im Windows-Umfeld eingesetzt werden können. Ein Abbruch in einem Analyseschritt durfte die nächsten Tools nicht beeinflussen. Als Datenbank wurde SQlite und als Programmiersprache für das Überwachungsprogramm Golang

gewählt.[5] SQlite hat den Vorteil, dass weder Server noch Administration notwendig sind. Golang ist eine kompilierte Sprache und auf allen Plattformen verfügbar, hat ein API zu SQLite und ist einfacher als C/C++. Alle Programme, Datenbankmodelle, Scripts und SQL-Abfragen sind auf GitHub verfügbar; eine detaillierte Installationsanleitung findet sich auf der KOST-Website.[6]

Diese Konstellation erlaubte es, die Ausführung der Analysemodule von der Auswertung der Log- oder Systemausgabe vollständig zu trennen. Die Log- oder Systemausgabe zu jedem Analyseschritt wurde für die spätere Auswertung festgehalten. Um der Anforderung der vollständigen Anonymisierung gerecht zu werden, wurden die Logdateien beim Schreiben gefiltert und Pfad und Dateinamen entfernt. Die eigentliche Auswertung erfolgt anschliessend vollständig offline, entweder im Archiv oder ausgelagert. Die Vorteile dieses Vorgehens sind, dass verschiedene Auswertungen auch zeitlich versetzt möglich sind und dass die Fragestellungen und Auswertungsmethoden während der Arbeit noch verändert werden können. Für die Auswertung stehen nach der Analyse insgesamt etwa 35 GB Log-Informationen zur Verfügung.

Analyse-Loop-Programm

Das Analyse-Loop-Programm liest alle TIFF-Dateien des Korpus vom NAS und führt mit der jeweils gelesenen Datei durch Aufrufen von externen Programmen mehrere Analyseschritte aus. Der Loop-Prozess ist zweiteilig und besteht aus der Initialisierung der Prozessdatenbank und der eigentlichen Analyse.

Der Initialisierungsschritt (Abbildung 1) erstellt die Datenbank und schreibt für jede TIFF-Datei einen Eintrag mit dem Pfad und Dateinamen als Schlüssel. Die Initialisierung kann mehrfach aufgerufen werden und fügt so neue Verzeichnispfade zur Datenbank hinzu. Um eine spätere anonymisierte Auswertung ausserhalb der Archive zu ermöglichen, werden Dateinamen und Dateipfad, welche allenfalls Rückschlüsse auf den Inhalt der Dateien erlauben würden, in einer separaten Tabelle (*namefile*) gehalten.

5 Siehe https://www.sqlite.org/ und https://golang.org/.

6 Siehe https://github.com/KOST-CECO/TiffAnalyseProject beziehungsweise https://kost-ceco.ch/cms/index.php?tiff-data-analysis_de.

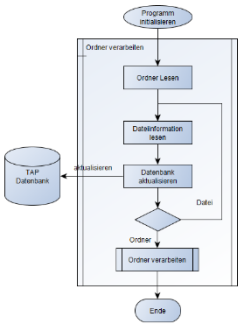

Abbildung 1: Initial Loop liest sämtliche TIFF-Dateien

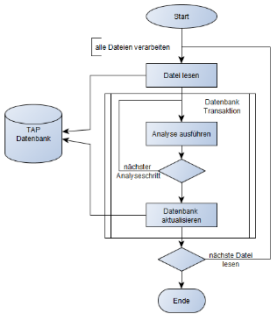

Abbildung 2: Process Loop führt Analyseschritte aus

Im Analysefall (Abbildung 2) werden die Dateieinträge in der Datenbank abgearbeitet und die Analysetools im Kommandozeilenmodus aufgerufen. Durch die Verwendung einer Datenbank ist jederzeit ein Abbrechen und Neustarten der Analyse möglich. Die Analyse umfasst folgende Schritte:

- Der Dateipfad und der Pfad zur Logdatei werden dem Analyse-Loop-Programm im Kommandozeilenmodus übergeben.
- Falls das Analyse-Loop-Programm kein Logfile im Append-Modus öffnen kann, hängt es den Log-Output entweder an die Logdatei an oder schreibt ihn in die Datenbank.
- Der aktuelle Offset der Logdatei wird in der Datenbank gespeichert.
- Der Exitstatus des Analysetools wird in der Datenbank festgehalten.
- Es kann festgelegt werden, ob der Systemoutput des Analysetools in eine spezielle Ausgabedatei geschrieben oder in der Datenbank gespeichert werden soll.
- Eine Logrotation verhindert allzu grosse Logdateien.

Datenmodell der Analysedaten

Die Tabellen *keyfile* und *namefile* enthalten den primären Verzeichnisscan, also die Namen aller Dateien mit Dateigrösse und Erstellungszeit, soweit diese aus dem Lesen der Verzeichnisstrukturen erstellt werden können. Zum Ausführen der Analysemodule werden die notwendigen Informationen aus der Tabelle *analysetool* ausgelesen: Programmname und Pfad, Logdatei, Datei bzw. BLOB für den Systemoutput. Die Tabelle *status* hält den Exitstatus des Analyseprogramms fest. Die Tabellen *logindex* und *sysindex* speichern entweder den Dateinamen und den Offset in die jeweilige Logdatei oder den gesamten Output der eben analysierten Datei in ein entsprechendes LOB.

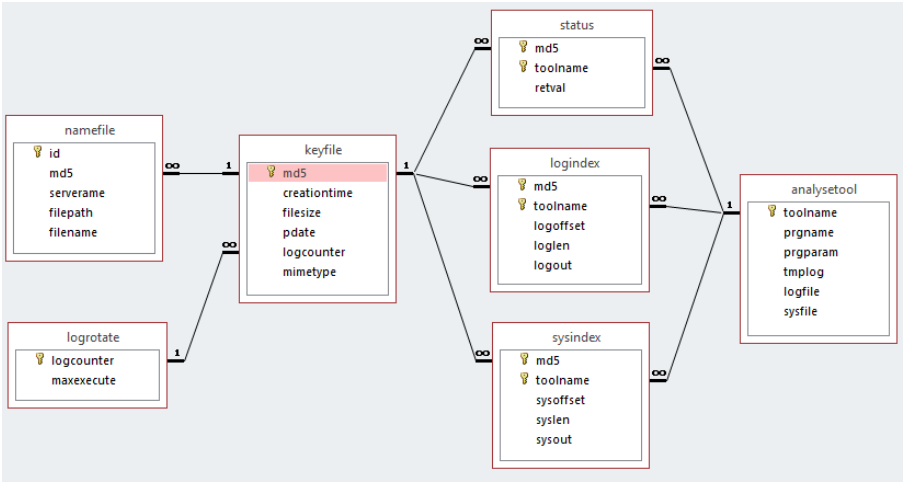

Abbildung 3: Das Datenmodell in grafischer Darstellung

tablename	name	description
analysetool	toolname	Name des registrierten Analyseprograms in Kurzform
	prgname	Pfad und Dateiname zum Analyseprograms
	prgparam	Parameter des Analyseprogramms mit Wildcards %file% und %log%
	tmplog	Temporäre Logdatei: ersetzt Wildcards %log% beim Ausführen des Analyseprogramms, Fehlen meint keine Logdatei schreiben
	logfile	Pfad und Dateiname der mit diesem Analyseprogramms verbunden Logdatei: Ist kein Logfile definiert, wird in LOB «logout» gespeichert
	sysfile	Pfad und Dateiname der mit diesem Analyseprogramms verbunden Ausgabedatei: Ist kein Sysfile definiert, wird in LOB «sysout» gespeichert
keyfile	md5	MD5-Hashwert und Referenz zum namefile
	creationtime	Entstehungszeitpunkt der Datei laut Dateisystem
	filesize	Dateigrösse in Byte

	pdate	Zeitpunkt und Flag für den Abschluss der gesamten Analyse
	logcounter	Zähler für «logfile» bzw. «sysfile» beginnend mit Eins
	mimetype	MIME Type (Internet Media Type) der Datei gemäss der Magic Number
logindex	md5	MD5-Schlüssel der TIFF-Datei
	toolname	Kurzname des Tools
	logoffset	Offset in die Ausgabedatei analysetool.logfile
	loglen	Länge des Logausgabe
	logout	Vollständige Logausgabe des Analysetools oder Name der Logdatei
logrotate	logcounter	Zähler für «logfile» bzw. «sysfile» beginnend mit eins
	maxexecute	Maximale Verarbeitungsschritte pro «logfile» bzw. «sysfile»
namefile	id	Referenz zu «keyfile»
	md5	MD5-Hashwert
	servername	Name des NAS-Servers oder des zugeordneten Laufwerkbuchstabens
	filepath	Dateipfad
	filename	Dateiname mit Dateiextension
status	md5	MD5-Schlüssel der TIFF-Datei
	toolname	Kurzname des Tools
	retval	Rückgabewert des Tools[7]
sysindex	md5	MD5-Schlüssel der TIFF-Datei
	toolname	Kurzname des Tools
	sysoffset	Offset in die Ausgabedatei analysetool.sysfile
	sylen	Länge der Konsolenausgabe
	sysout	Vollständige SystemOut-Ausgabe des Analysetools (stderr und stdout) oder Name der Logdatei

Tabelle 3: Das Datenmodell mit Tabellen

Einschränkungen bei der Analyse

Nach den ersten Tests hat sich schnell gezeigt, dass die verwendeten Tools sehr unterschiedliche Rechenzeiten pro TIFF-Datei erfordern. Tabelle 4 dokumentiert die Rechenzeiten pro Tool über 1000 Dateien unterschiedlicher Grösse (mittlere Grösse ~5.5 MB), indexiert gegenüber *tiffhist*:

7 Siehe dazu http://www.hiteksoftware.com/knowledge/articles/049.htm.

Tool	Zeit (s)	Faktor
tiffhist	257	1.0
dpf-manager	257	1.0
file	266	1.0
exiv2	267	1.0
exif	335	1.3
checkit_tiff	503	2.0
jhove	697	2.7
ImageMagick	3424	13.3
Total	6006	23.4

Tabelle 4: Rechenzeiten pro Tool über 1000 Dateien unterschiedlicher Grösse

Um die Analysezeit nicht ausufern zu lassen, wurde beschlossen, ImageMagick nur über einem sehr kleinen Teilbestand und JHOVE nur etwa über der Hälfte der Dateien auszuführen. Das aus unserer Sicht wichtigste Tool zur Extraktion von TIFF-Tags, das vom DHLab entwickelte *tiffhist*, wurde hingegen auf allen Dateien ausgeführt.

Auswertung

Die Analyseresultate stehen der gesamten Fachgemeinschaft zur Verfügung. Sie sind zu diesem Zweck auf der Website der KOST publiziert und erläutert[8]. Alle Interessierten sind eingeladen, diese Daten für ihre eigenen Forschungen zu benützen; besonders angesprochen sind dabei die Hochschulen und Fachhochschulen.

Abschliessend folgen hier zwei ausgewählte Beispiele für eine mögliche Auswertung. Weitere Beispiele sind auf der KOST-Website dokumentiert.

8 Siehe https://kost-ceco.ch/ftp_space/TIFF-Analyse/. Der Downloadspace enthält die Analysedatenbank als SQL Loader Script tap.sql.gz (md5: 33b406a083472fb3853c1d07169bd640) und die Logdateien in einem TAR-File log.tgz (d44b169f1d8048637c5502be646f8a85). In separaten Dateien ist die später fertiggestellte Analyse der 10 Millionen Dateien des Schweizerischen Bundesarchivs dokumentiert: 10-tap.sql.gz (fcb11b650d9c64a2f908a561934e6fd) und 10-log.tgz (d317855606603bca8033ab0ae21ea03e). Alle Dateien sind gzip-komprimiert.

Verteilung TIFF-Komprimierung

In diesem Beispiel werden auf Grund der von *tiffhist* in die Logdatei geschriebenen Tag-Informationen die Verteilung und Werte des Compression Tags 259 untersucht. Das Tag 259 ist in der TIFF-Spezifikation folgendermassen erläutert:

Compression

Data can be stored either compressed or uncompressed.

Tag = 259 (103.H)
Type = SHORT
Value = 1 -10, values > 32766 (proprietary values)

1	No compression
2	CCITT 1D
3	Group 3 Fax
4	Group 4 Fax
5	LZW
6	JPEG TIFF/6-.0 marked as deprecated
7	JPEG TIFF TechNote2 1995
8	Adobe Deflate
9	JBIG bw
10	JBIG color
32773	PackBits

Abbildung 4: Definition des Compression Tag in der TIFF-Spezifikation[9]

Ein einfaches Windows- oder Linux-Shellscript erzeugt aus den Logdateien die Übersicht in Tabelle 5.

9 TIFF Revision 6.0 (1992), siehe oben Anm. 4, S. 17.

Compression	Basel-Stadt	St. Gallen	Bundesarchiv	Bundesarchiv extern	Total	Prozent
none	732'696	550'675	560'956	217'843	2'062'170	52%
CCIT 1D	0	0	564	0	564	0%
Fax Group 3	0	0	20'041	0	20'041	1%
Fax Group 4	22'745	317	602'571	1'207'376	1'833'009	46%
LZW	31	15'651	19'534	0	35'216	1%
old JPEG	0	0	0	0	0	0%
JPEG	0	15'427	12'095	0	27'522	1%
Adobe Deflate	0	0	1'593	0	1'593	0%
JBIG bw	0	0	0	0	0	0%
JBIG color	0	0	0	0	0	0%
Pack Bits	0	0	0	0	0	0%
other	0	0	0	0	0	0%
Total	755'472	582'070	1'217'354	1'425'219	3'980'115	100%

Tabelle 5: Verteilung der Kompressionsarten im Korpus

Vergleich zweier Tools (exiftool und exiv2)

Die Auswertung der Ausgaben verschiedener Tools ist bei der Speicherung der Logausgabe in der Datenbank relativ einfach. Das SQL-Script vergleicht die Ausgabe von exiftool und exiv2 zur jeweils gleichen Datei und gibt das Resultat in einer HTML-Datei aus:

```
.output exiftool&exiv2.html
.mode asci
SELECT "<!DOCTYPE html><HTML><head><style> table { font-family: arial, sans-serif;
border-collapse: collapse; width: 100%; } td, th { border: 1px solid #dddddd; text-
align: left; padding: 8px; } tr:nth-child(even) { background-color: #dddddd; }
</style></head>
<BODY><PRE><TABLE>";

.mode html
SELECT
--  sys1.md5,
    sys1.toolname,
    sys1.sysout,
--  sys2.md5,
    sys2.toolname,
    sys2.sysout
FROM
    (SELECT md5, toolname, sysout from sysindex WHERE toolname = "exif") sys1
INNER JOIN
    (SELECT md5, toolname, sysout from sysindex WHERE toolname = "exiv2") sys2
ON
    sys1.md5 = sys2.md5;

.mode ascii
SELECT "</TABLE></PRE></BODY></HTML>";
.exit
```

Abbildung 5: SQL-Script zum Toolvergleich

Das Resultat, die Darstellung der Ausgabedatei in einem Browser, zeigt Abbildung 6.

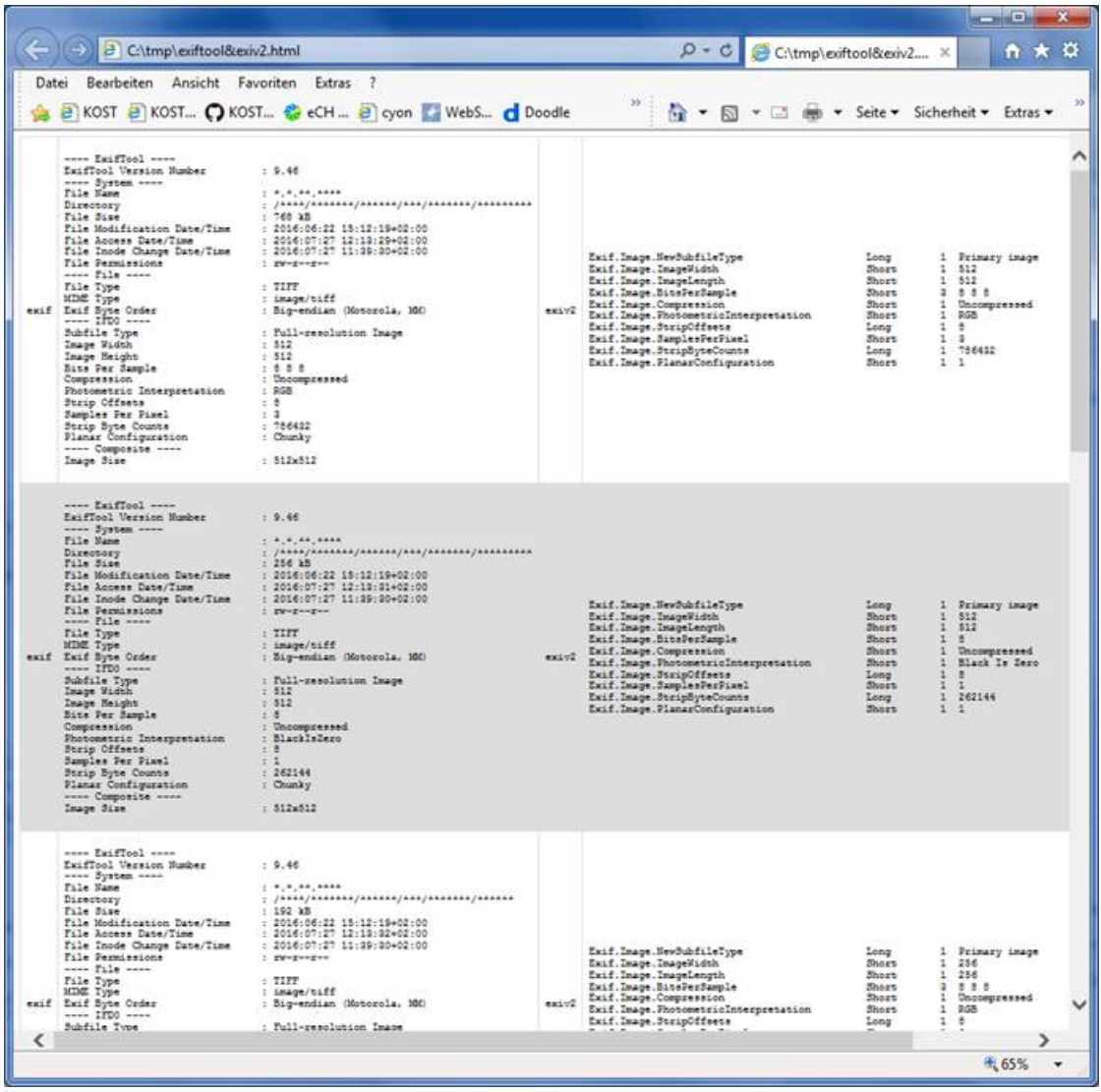

Abbildung 6: Resultate des Toolvergleichs

Fazit

Die TIFF-Korpus-Analyse der KOST ermöglicht es der Fachgemeinschaft, Empfehlungen und Strategien zum Umgang mit TIFF-Dateien im Archiv auf konkrete Eigenschaften real existierender Dateien aus verschiedenen Archiven und Entstehungskontexten abzustützen. Die Publikation der Analysewerkzeuge und des Vorgehens erlaubt es zudem, weitere Korpora in vergleichbarer Weise zu analysieren und die Datenbasis damit zu vergrössern.

Nutzen und Grenzen der Formaterkennung

(Zu-)Fälle bei PRONOM und DROID

Stephanie Kortyla, Christian Treu

Ziel jeder Archivierung ist Nutzbarmachung. Im digitalen Bereich sind Objekte nur mit Hilfsmitteln, genauer mit einer technischen Darstellungsumgebung wahrnehmbar. Digitale Objekte werden in Repräsentationen abgelegt, von denen jede aus mindestens einer Datei besteht.[1] Folglich liegt pro Objekt mindestens ein Dateiformat vor. Doch um welches handelt es sich explizit? Die Dateinamenserweiterung, auch Extension genannt, kann trügen, unbekannt sein oder auch fehlen. Eine Identifizierung des Formats, als Erstbestimmung oder Verifizierung einer Vermutung, ist unabdingbar. Mit diesem Anhaltspunkt können Daten lesbar und nutzbar gemacht werden. Nicht nur für den aktuellen, sondern auch für einen zukünftigen Gebrauch von Objekten, z.B. im Zuge von Bestandserhaltung wie Formatmigration, sind passende Werkzeuge auszuwählen. Formaterkennung bildet nur einen ersten Schritt im Lebenszyklus von Unterlagen, die für eine Langzeitspeicherung und/ oder dauernde Archivierung vorgesehen sind.[2]

Das Sächsische Staatsarchiv betreibt seit 2013 sein Elektronisches Staatsarchiv (el_sta) und setzt sich fortlaufend mit Formaterkennung auseinander. Im Folgenden wird auf die in der Praxis eingesetzten Werkzeuge mit ihrem Nutzen und ihren Grenzen näher eingegangen.[3] Beim el_sta werden das technische Register

1 Vgl. PREMIS Editorial Committee (Hg.): PREMIS Data Dictonary for Preservation Metadata.Version 3.0., 2015, S. 8. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Im hiesigen Beitrag soll es dagegen nicht um Bewertung von «archivtauglichen» Formaten gehen.

3 Mit den aufgekommenen Fragen wandte sich das Sächsische Staatsarchiv im Sommer 2015 in der Community zunächst an die nestor-AG Formaterkennung und daraufhin an Jay Gattuso von der Neuseeländischen Nationalbibliothek, siehe unten. Die nestor-AG hat auf ihrer Wiki-Seite einige der hier geschilderten Gegebenheiten ebenfalls aufgenommen. Vgl. Tunnat, Yvonne: PRONOM. Persistenz von PUIDs. Wiki-Unterseite der nestor AG-Formaterkennung, 2016. https://wiki.dnb.de/display/NESTOR/PRONOM%3A+Persistenz+von+PUIDs. Nutzererfahrungen mit PRONOM und DROID sind bisher marginal veröffentlicht worden: http://coptr.digipres.org/DROID_(Digital_Record_Object_Identification). Vgl. auch Gattuso, Jay: Throughput efficiencies and misidentification risks in DROID, 2012. http://ndha-wiki.natlib.govt.nz/assets/NDHA/Reading/MSB+DROID+v1-05.pdf. Ders.: Evaluating the historical persistence of DROID asserted PUIDs, 2012. Auf dieser Website sind weitere Testberichte zu finden. Vgl. auch Jackson, Andy: Formats over Time: Exploring UK Web History. In: iPres2012. Proceedings of the 9th International Conference on Preservation of Digital Objects, 2012, S. 155-158. https://ipres-conference.org/ipres12/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf. Vgl. auch Tarrant, David; Carr, Les: LDS³: Applying Digital Preservation Principals to Linked Data Systems. In: iPres2012. Proceedings of the 9th International Conference on Preservation of Digital Objects, 2012, S. 77-84, hier S. 83. https://ipres-conference.org/ipres12/sites/ipres.ischool.utoronto.ca/files/

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.9

bzw. die Formatdatenbank PRONOM, die Schnittstelle zum eigentlichen Werkzeug bzw. die sogenannte Signature-File sowie das Erkennungswerkzeug DROID eingesetzt. Hierbei handelt es sich um eines der gängigsten Verfahren zur Formatidentifizierung.[4]

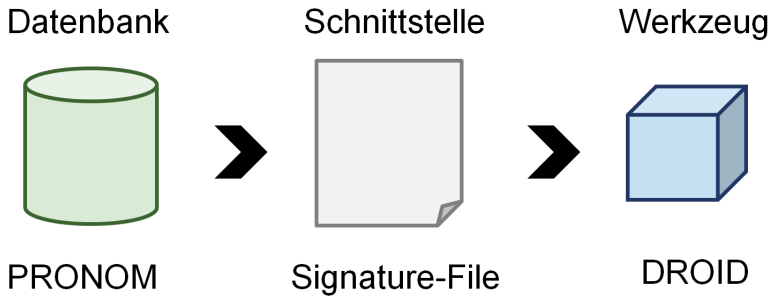

Abbildung 1: Zur Formaterkennung eingesetzte Komponenten, hier PRONOM, PRONOM-Signature-File, DROID. Eigene Darstellung

Formaterkennung – wofür?

Durch Erkennung und Analyse der vorliegenden Formate lässt sich eine Übersicht erstellen, die dem Preservation Planning als Grundlage dienen kann. Zudem lassen sich durch die Formaterkennung bereits vor dem Ingest etwaige Unstimmigkeiten und Fehler, wie falsche Endungen und unter Umständen beschädigte Dateien, erkennen und beheben. Unerwünschte System- und Vorschaudateien (z. B. Thumbs.db) lassen sich ebenso automatisiert erkennen und filtern. Die bei der Ana-

iPres%202012%20Conference%20Proceedings%20Final.pdf. Vgl. auch Töwe, Matthias; Geisser, Franziska; Suri, Roland E.: To Act or Not to Act – Handling File Format Identification Issues in Practice. Poster in: iPRES2016. 13th International Conference on Digital Preservation. Proceedings, 2016, S. 288f. https://ipr16.organizers-congress.org/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf.

4 Einen Überblick über weitere Formaterkennungswerkzeuge bietet z.B. folgende Website: http://coptr.digipres.org/Category:File_Format_Identification. «The 'big 3' file format identification tools, DROID, Tika and File (…)», in: Wheatley, Paul; Pennock, Maureen: Supporting practical preservation work and making it sustainable with SPRUCE. In: iPres2013. Proceedings of the 10th International Conference on Preservation of Digital Objects, 2013, S. 73-77, hier S. 74. Bewertungen von Erkennungswerkzeugen sind zu finden bei: Knijff, Johan van der; Wilson, Carl: Evaluation of Characterisation Tools. Part 1: Identification. SCAPE Project, 2011. http://scape-project.eu/wp-content/uploads/2014/08/SCAPE_PC_WP1_identification21092011.pdf sowie bei Röthlisberger-Jourdan, Claire; Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST): Formaterkennung und Formatvalidierung: Theorie und Praxis, 2012. https://kost-ceco.ch/cms/index.php?format_validation_de.

lyse generierten technischen Metadaten können in die Metadaten des AIP aufgenommen werden.

Nicht zuletzt bietet die Formaterkennung auch die Möglichkeit des Risikomanagements. Um spätere Kosten für spezielle Viewer und lizensierte Fachanwendungen einzudämmen oder wenigstens deren Bedarf kritisch beurteilen zu können, bietet sich die Begrenzung und Kontrolle der eingehenden Formate an. Probleme, die sich zum Beispiel aus obsoleten oder proprietären Formaten ergeben, können besser eingeschätzt werden. In aggregierter Form können die technischen Metadaten der bereits erfolgten Ingests in einer Statistik zusammengefasst werden.

Erkennung vs. Validierung

Hauptziel der Formaterkennung ist die Identifizierung eines bestimmten Dateiformates. Die Validierung überprüft hingegen die Regelhaftigkeit bzw. die Normkonformität einer Datei[5] und grenzt dabei gegebenenfalls Formatvarianten voneinander ab. Dies geht weit über die bloße Erkennung hinaus, erfordert aber spezielle Tools. Ein Datei kann nutzbar und dennoch nicht valide sein, wenn sich z. B. eine Datei als PDF/A zu erkennen gibt, aber nur dem PDF-1.4-Standard entspricht.[6] In der Praxis existieren bisher Validatoren für nur wenige Formate.

Formaterkennung – wie ?

Um Dateiformate zu identifizieren, gibt es verschiedene Ansätze. Drei Methoden sollen hier beispielhaft dargestellt werden, wobei die zuletzt genannte den größten Nutzwert bietet. Für viele selbstverständlich, aber bei weitem nicht ausreichend, ist die Formaterkennung anhand der Dateiendung oder Extension. Die Vergabe von Dateiendungen auf verschiedenen Betriebssystemen ist weder obligatorisch noch einheitlich. Ein Bild im JPEG Interchange Format (JIF) kann die Endungen .jpeg, .jpe, .jpg oder schlicht keine Endung tragen. Endet eine Datei auf .pdf, kann dies repräsentativ für eine von 37 PDF-Varianten stehen (nach aktuellen Stand der Signature-File V90). Vergleichbares ergibt sich auch bei der Bestimmung des MIME Types. Nicht immer lässt sich der MIME Type ermitteln und selbst wenn, ist dieser nicht sehr trennscharf. Die Standards HTML 4.01, XHTML 1.0 und HTML5 können sich alle den gleichen MIME Type, hier «text/html», teilen. Für eine genauere Unterscheidung kann eine Validierung vorgenommen werden.

5 Vgl. Röthlisberger-Jourdan, Formaterkennung und Formatvalidierung (siehe Anm. 4), S. 4.
6 PDF/A-1 basiert auf PDF 1.4.

Die zuverlässigste und eindeutigste Formaterkennung erfolgt durch die Analyse und Identifizierung charakteristischer Bytesequenzen,[7] so genannter *Signatures*.[8] Dabei wird versucht, bestimmte Bitfolgen, auch *Magic Numbers* genannt, zu erkennen und einem Format zuzuordnen. Diese Muster können am Anfang, am Ende oder einer anderen Stelle im Bitstrom auftauchen. Bisweilen enthalten die Signatures versteckte Botschaften in ASCII oder in hexadezimaler Kodierung, welche zum Beispiel in einem Hex-Editor lesbar gemacht werden können. So sind alle zip-basierten Dateiformate mit der Magic Number «pk», für den Entwickler Phil Katz, versehen. Die Signature kann aber auch vergleichsweise offensichtlich, wie «%PDF-1.4» für PDF 1.4 oder «DOCTYPE HTML[…]» für HTML sein.

Nutzen der Formaterkennung

PRONOM

Um die Signaturen und weiterführende Informationen zu verschiedenen Formaten zusammenzutragen, bedarf es einer Datenbank. Die bisher verbreitetste Datenbank dieser Art ist die 2002 durch das britische Nationalarchiv (TNA) begonnene PRONOM-Datenbank.[9] Kernaspekt von PRONOM ist die Bereitstellung und Pflege von eindeutigen Identifiern für Formate (Persistent Unified IDentifier) in einer kostenfreien, webbasierten Datenbank. Ergänzungs- und Verbesserungsvorschläge durch die Community werden regelmäßig durch TNA eingepflegt.[10] Aus der Datenbank wird wiederum mehrmals im Jahr eine DROID Signature-File generiert, deren Hauptzweck es ist, die Grundlage zur automatisierten Formaterkennung zu stellen. Als XML-Datei bildet sie die Schnittstelle zwischen Datenbank und Erkennungstool.[11]

DROID

Das Tool DROID (Digital Record Object IDentification) wurde ebenfalls von TNA entwickelt und der Community im Jahre 2005 zum ersten Mal zur Verfügung gestellt. 2017 ist die Version v6.3 veröffentlicht worden.[12] Die Java-Anwendung kann

7 Vgl. Röthlisberger-Jourdan, Formaterkennung und Formatvalidierung (siehe Anm. 4).

8 Unter unixoiden Betriebssystemen wird ein solcher Ansatz mit dem Tool «file» schon länger verfolgt.

9 http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm. Andere «Projekte» wie GDFR bzw. UDFR konnten sich nicht etablieren. Vgl. http://www.udfr.org/.

10 Inwieweit PRONOM zukunftssicher, etwa finanziell gesehen, bzw. zweckmäßig ausreichend ist, soll im Rahmen dieses Beitrags nicht erörtert werden.

11 Eine Übersicht über bisherige Signature-Files ist zu finden unter http://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm.

12 Eine Bedienungsanleitung findet sich hier: The National Archives (TNA): Droid User Guide, 2017. http://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf.

plattformunabhängig über Kommandozeilenanweisung oder per GUI (Desktop-Version) eingesetzt werden. Sie kann in Prozess-/ Systemlandschaften integriert werden. DROID ist weltweit verbreitet und frei verfügbar.

DROID differenziert Typen, das heißt, es unterscheidet zwischen Dateien, Verzeichnissen und «Archiv»-Dateien (z.B. zip, tar). Das Werkzeug ermöglicht Stapelverarbeitung und führt die Analyse verzeichnisübergreifend durch. Die Struktur der zu analysierenden Einheit ist für das Werkzeug somit irrelevant. Die dabei angewandten Methoden zur Erkennung sind Signature-, Extension- und Containeridentifizierung, und nur wenn keine dateiinterne Signature gefunden wird, folgt eine Erkennung über die Extension. Voraussetzung ist hier, dass überhaupt eine existiert, andernfalls bleibt das Format unerkannt. Für eine Erkennung über Signatures ist die Existenz einer Extension unerheblich. Über ein Scanprofil können Einstellungen für eine Analyse vorgenommen werden. Ergebnis ist eine Informationssammlung über die gescannten Objekte, die z.B. im GUI tabellarisch ausgegeben wird. Optional können die Ergebnisse in verschiedenen Varianten exportiert werden, so dass beispielsweise die zu Objekten erstellten Metadaten in nachfolgenden Prozessschritten weiterverarbeitet werden können. DROID ermöglicht so Angaben zu Extension (auch Anzeige, wenn nicht vorhanden), PUID (PRONOM), Formatname und -version, Mime-Type, Erkennungsmethode und -status (erkannt | nicht erkannt | Ambiguität), optional Hashwert. Gelegentlich treten Performanceprobleme auf (Scanprofileinstellung, Absturz). Aufgrund dessen sind in der Community vereinzelt bereits Überlegungen angestellt, die komplexen Signature-Files zu reduzieren.[13]

Grenzen der Formaterkennung

DROID

Nachfolgend wird auf Gegebenheiten und sich daraus ergebende Konsequenzen im Umgang mit DROID und PRONOM eingegangen. Gemäß dem Prinzip, dass zuerst eine Erkennung über Signatures angestoßen wird und im negativen Fall anschließend eine Erkennung über Extensions, muss berücksichtigt werden, dass Formate zufällig über Muster verfügen können, welche als Signature eines anderen Formates

13 Reduktion des Umfangs der von der einsetzenden Institution akzeptierten Formate bei Hoppenheit, Martin: Minimizing the DROID signature file, 2017. http://hoppenheit.info/blog/2017/minimizing-the-droid-signature-file/. Auf diese Weise konkret umgesetzt für die Praxis ist die Signature-File bei der KOST für den KOST-Val, s.u.; Syntaxreduktion durch Auslassen von sog. Shift-Bytes bei Spencer, Ross: Hacking the DROID Signature File: Keep It Simple Stupid! 2012. http://exponentialdecay.co.uk/blog/hacking-the-droid-signature-file-keep-it-simple-stupid/.

erkannt werden,[14] so dass die Formaterkennung zwar technisch korrekt verläuft, das Ergebnis praktisch aber eine falsche Formatangabe liefert. So erging es dem Staatsarchiv beispielsweise mit einer Übernahme von Daten in Plain Text mit spezieller Kodierung, hier EBCDIC (fmt/159). Zu erwarten war eine Erkennung über Extension (hier .ebcdic), was in den meisten Fällen zutraf. Im Scanergebnis traten aber auch unerwartete Formate auf, so z.B. das Microsoft Owner File Format (fmt/473). Bei Überprüfung des Scanergebnisses für jene Datei wurde deutlich, dass DROID im ersten Prüfvorgang eine Signature fand, so dass die Formaterkennung für diese Datei nach diesem Prozessschritt beendet war. Einzig der Warnhinweis eines Mismatches, genauer dass die tatsächliche Extension (hier .ebcdic) nicht zum vermeintlich erkannten Format (MS Owner File) gehört, deutete daraufhin, dass hier evtl. ein Erkennungsproblem vorliegen könnte. Wie aber bereits erwähnt, ist die Extension, sofern sie überhaupt vorhanden ist, keine Garantie für Korrektheit. Im hiesigen Fall war die Extension (.ebcdic) allerdings zutreffend, so dass die durch die Formaterkennung automatisch erhobenen Metadaten für den Archivierungsprozess (DROID ist in die Prozesslandschaft im Background des el_sta integriert) nach AIP-Generierung und noch vor Ingest in das Repository manuell anzupassen waren.[15] Bei einem Ingest von knapp 90 AIP mit insgesamt über 1000 Dateien, wobei EBCDIC nur eines von mehreren Formaten war, und aufgrund geringer Erfahrung in Bezug auf das EBCDIC-Format wurde eine Qualitätskontrolle stichprobenhaft durchgeführt.

Um diesen Fehler zu reproduzieren, kann ein simpler Test mit einer Datei eines vermeintlichen Formats durchgeführt werden. Dieser ist auch bei Röthlisberger-Jourdan[16] beschrieben. Eine einfache Plain Text-Datei optional mit .pdf-Extension beinhaltet ausnahmslos die zur Signature-Erkennung notwendigen Angaben (hier: «%PDF-1.4» sowie «%%EOF»). DROID findet im ersten Schritt zur Formatidentifizierung bereits eine Signature, hier für das Format PDF 1.4 (fmt/18) und gibt demzufolge als Ergebnis fmt/18 aus. Dabei ist offensichtlich, dass die PDF-Datei nicht funktionsfähig ist respektive sich nicht über entsprechende Software wie einen PDF-Reader darstellen lässt. Hier wird deutlich, dass die Erkennungssoftware erwartungsgemäß und zuverlässig arbeitet, einem Nutzer jedoch

14 Hierauf weisen z.B. Dunckley und Rankin hin: Dunckley, Matthew; Rankin, Stephen: The Use of File Description Languages for File Format Identification and Validation, 2007, S. 1. https://epubs.stfc.ac.uk/work/50089.

15 Mitcham wirft eine Frage zum «over-ride» auf: «to over-ride file identifications - eg – 'I know this isn't really xxxx format so I'm going to record this fact' (and record this manual intervention in the metadata)». Mitcham, Jenny: File identification… let's talk about the workflow, 2015. http://digital-archiving.blogspot.de/2015/11/file-identification-lets-talk-about.html.

16 Vgl. Röthlisberger-Jourdan, Formaterkennung und Formatvalidierung (siehe Anm. 4), S. 4f.

auch bewusst sein muss, dass Ergebnisse stets stichprobenhaft kritisch geprüft werden sollten.

Ein weiterer Aspekt, der seit Inbetriebnahme des el_sta aufgetreten ist, betrifft unvorhergesehene Scandifferenzen. Diese treten bei Einsatz verschiedener Erkennungswerkzeuge[17] bzw. Datengrundlagen (hier verschiedene Signature-File-Versionen) auf. Bedingt durch die Fortschreibung bzw. Aktualisierung der Signature-Files werden derlei Scanergebnisdifferenzen kontinuierlich zu erwarten sein.[18] Dahingehend stellt sich die Frage, in welchem Ausmaß Differenzen zu akzeptieren wären und ob bei wissentlicher Datengrundlagenänderung (z.B. Veröffentlichung einer neuen Signatur-File mit relevanten Änderungen in Bezug auf eigene Echt-Ingests)[19] eine Re-Identifizierung[20] durchzuführen wäre. Diskutabel wäre hier der Grad der Granularität, beispielsweise im Hinblick auf Dateiformatversionen. So werden PDF-Dokumente zunächst in Hauptgruppen wie PDF, PDF/A, PDF/E, PDF/UA und PDF/X unterschieden. In den Gruppen wiederum gibt es weitere Unterversionen wie PDF 1.0-1.7, PDF/A-1a sowie PDF/A-1b etc. In PRONOM sind aktuell (Stand Signatur-File v90 vom 30.03.2017) 37 verschiedene PDF-Versionen mit PUID aufgenommen. Eine falsche Zuordnung eines Formats (wie oben beschrieben anhand von EBCDIC) wäre gegenüber einer fehlerhaften Version (PDF/A-1a oder PDF 1.4) stärker zu gewichten. Abschließend sei auf die KOST verwiesen, die eine eigene Signature-File (KaD-Signature-File[21]) aus den PRONOM-Signature-Files generiert, in der einige Formatversionen und demzufolge auch mehrere PUID zusammengefasst sind. Diese Signature-File wird im eigens erstellten Validierungswerkzeug KOST-Val eingesetzt.[22]

Neben der sich ändernden Datengrundlage (Signature-Files) können die Scandifferenzen auch mit der Scanprofileinstellung begründet werden. Seit der Version v6.3 bietet DROID die Option, anstelle der gesamten Datei (full scan) lediglich einen abgesteckten, selbstgewählten Abschnitt zu scannen. Bei diesem Scanmodus, dem sog. «Max-Byte-Scan» (MBS), wird sowohl Dateianfang als auch

17 Neben DROID kamen bspw. KOST-Val und Fits zum Einsatz.

18 Gattuso, Jay als Kommentar zu Blogeintrag von Wheatley, Paul: Don't panic!: What we might need format registries for, 2012. http://openpreservation.org/blog/2012/07/05/dont-panic-what-we-might-need-format-registries/. Auch Mitcham greift einige Fragen zu Ergebnisdifferenzen und (Nicht-)Identifikation auf: Mitcham: File identification… (siehe Anm. 15).

19 Zumindest sollte die jeweils eingesetzte Signature-File-Version dokumentiert sein. Im weiteren Sinne auch bei Tarrant und Carr: «[…] facts might include the format identification information (at the time)». Tarrant, Carr, LDS³ (siehe Anm. 3), S. 82.

20 Töwe, Geisser, Suri, To Act or Not to Act (siehe Anm. 3), S. 288f. Spencer ist pro Rerun bei neuer Signature: «Pronom is a continuum», Spencer, Ross: Tweet von @beet-keeper, 2015. https://twitter.com/beet_keeper/status/626890544631812097.

21 Vgl. https://github.com/KOST-CECO/KaD_SignatureFile.

22 https://kost-ceco.ch/cms/index.php?kost_val_de. Hierbei ist zu erwähnen, dass das Werkzeug KOST-Val in erster Linie zur Validierung einiger Formate eingesetzt wird.

-ende auf Signatures hin abgetastet. Signatures finden sich häufig in diesen Abschnitten. Es wird deutlich, dass je kleiner der Wert respektive Abschnitt ist, desto schneller die Erkennung verläuft. Allerdings ist auch eine höhere Fehlerquote zu erwarten. Hierauf wird später anhand eines Beispiels näher eingegangen. Je größer der Scanabschnitt (an Dateianfang und -ende) ist, desto länger dauert der Scanprozess, umso wahrscheinlicher ist jedoch die Erkennungsquote und umso kleiner die zu erwartende Fehlerquote. Zudem muss auch bedacht werden, dass bei kleineren Dateien der MBS nicht vorteilhaft ist, wenn der doppelte MBS-Wert die jeweilige Dateigröße überschreitet, da sich die Scanabschnitte von Dateianfang und -ende mittig überlappen und somit ein Teil doppelt gescannt würde.[23] Der empfohlene MBS-Wert liegt bei 65536 Bytes. Dieser basiert auf umfangreichen Tests mit verschieden großen Dateien und Formaten durch die Neuseeländische Nationalbibliothek und ist auch der Default-Wert für die MBS-Einstellung in Droid.[24]

Wie angedeutet, kann immer eine gewisse Fehlerquote innerhalb der Erkennungsquote bestehen.[25] Als Beispiel in Bezug auf MBS mögen Formatversionen dienen, die auf einer anderen Version basieren, wie bei einigen Fällen in der PDF-Gruppe.[26] So basieren auf PDF 1.4 und PDF 1.7 jeweils weitere PDF-Versionen. Konkret basiert beispielsweise PDF/A-1a auf PDF 1.4. Dies spiegelt sich in der Signature wider So beginnen beide Dateien mit dem Signature-Bestandteil %PDF-1.4 und enden mit %%EOF.[27] Diese Signature-Teile treten dabei an fest definierten Stellen auf. DROID gibt so zuverlässig das Format fmt/18 für PDF 1.4 aus. Bei PDF/A-1a kommt zusätzlich als Distinktionsmerkmal ein Signature-Teil vor, der variabel im Bitstrom auftritt. Das Problem wird hierbei bereits sichtbar: Es kann nicht vorhergesehen werden, an welcher Stelle dieser Teil auftritt, ergo kann kein zuverlässiger MBS-Wert vorab definiert werden. Dementsprechend ist das Ergebnis nicht vorhersehbar. Liegt der variable Signature-Teil der PDF/A-1a-Datei außerhalb des Scanbereichs, wird der Signature-Teil am Dateiende (%%EOF) von DROID aufgegriffen, PDF 1.4 zugeordnet und als Ergebnis ausgegeben.

23 Gattuso, Throughput efficiencies and misidentification risks in DROID (siehe Anm. 3), S. 12. Schaubild zum MBS auf S. 4

24 Vgl. ebd., S. 7. Für umfangreiche Tests s. weitere Quellen von Gattuso.

25 Auch Bachmann u.a. weisen darauf hin. Vgl. Bachmann, Steffen; Ernst, Katharina: Formaterkennung – Ziele, Herausforderungen, Lösungsansätze. In: Manke, Matthias (Hg.): Auf dem Weg zum digitalen Archiv. Stand und Perspektiven von Projekten zur Archivierung digitaler Unterlagen. 12. Tagung des Arbeitskreises «Archivierung von Unterlagen aus digitalen Systemen» am 2. und 3. März 2011 in Schwerin, 2012, S. 69-73, hier S. 72.

26 Auf die Problematik bei der Erkennung von bestimmten PDF-Versionen weisen auch Knijff und McGath hin. Vgl. Knijff, Johan van der: PDF version numbers based on deprecated mechanism #114. Mit Kommentar von McGath, Gary, 2016. https://github.com/digital-preservation/droid/issues/114.

27 %%EOF ist für die Signature-Erkennung von PDF/A-1a irrelevant. Bei einem Full Scan würden sowohl die festen als auch das variable Signature-Teil erfasst. Um hier einem Konflikt vorzubeugen, ist in PRONOM vermerkt, dass PDF/A-1a eine höhere Priorität gegenüber PDF 1.4 genießt.

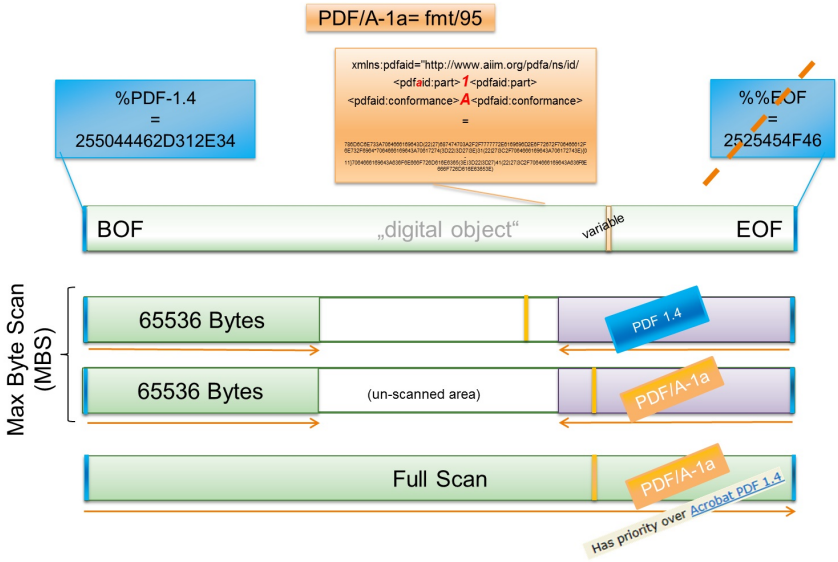

Abbildung 2: Je nach Auftreten des variablen Signature-Bestandteils in der PDF/A-1a-Datei wird jener vom Max-Byte-Scan erfasst oder nicht. Bei Nichterfassung findet DROID den Signature-Bestandteil von PDF 1.4 und gibt dies als Ergebnis aus. Darstellung in Anlehnung an Gattuso, Throughput efficiencies and misidentification risks in DROID, (siehe Anm. XX), S. 4. http://ndha-wiki.natlib.govt.nz/assets/NDHA/Reading/MSB+DROID+v1-05.pdf.

Wie schwerwiegend diese Scandifferenz ist, die sich auf eine Version einer Formatgruppe bezieht, muss jede Institution selbst werten.

PRONOM – Weiterentwicklung der Datengrundlage

Wie bereits angedeutet, wird PRONOM kontinuierlich durch TNA und die Community weiterentwickelt. Angestrebt wird dabei, ca. 100 neue PUIDs pro Jahr einzupflegen. Es kommen also regelmäßig neue Einträge hinzu, jedoch werden ältere gegebenenfalls modifiziert, zusammengelegt oder als überholt («Deprecated») gekennzeichnet. Die Zusammenhänge zwischen PUID, Signature und Extension sind bisweilen relativ komplex. Beispielsweise wird das PNG 1.1 Format unter der PUID «fmt/12» geführt, es existieren dafür in der Signature File die internen IDs 183, 184 und 185.[28] Andere Formate wie Plain Text («x-fmt/111») besitzen hinge-

28 Die Signature File vergibt für die Signatures interne IDs, welche bestimmten Formaten zugeordnet werden. Zum Aufbau der Signature-File, vgl. z.B. Gattuso, Jay (2012): How to write a new signature

gen gar keine Signature. Zudem kann die «generic» Signature mit der ID 78 nicht spezifisch einem Format zugeordnet werden, sondern verweist auf zwei verschiedene Versionen von Excel-Dateien.

Seit 2010 gibt es einen sprunghaften Anstieg bei den Signatures zu verzeichnen, was deren gewachsene Bedeutung für die Formaterkennung unterstreicht. Insgesamt gibt es jedoch mehr Dateiendungen (Extensions) als Formate und mehr Formate als Signatures.[29] Die dynamische Weiterentwicklung von PRONOM ist für die Nutzer nicht nur mit verbesserter Erkennung von Formaten verbunden. Unweigerlich stellt sich auch die Frage nach der Persistenz, denn die Daten, die vor Jahren bei einem Ingest generiert wurden, können nun veraltet sein. Neue Erkenntnisse zu Formaten sind für die Persistenz der Datenbank relativ unproblematisch, da der alte Informationsstand nicht falsch, sondern aktualisiert ist. Zunächst konnten beispielsweise die verschiedenen Microsoft Office Formate nur unter dem Sammelformat OLE2 Compound Document (fmt/111) erkannt werden. Heutzutage ist eine Unterscheidung möglich.[30]

Eine größere Hürde für die Persistenz[31] der erkannten Formate ergibt sich bei den ursprünglich vorläufigen, zusammengelegten oder überholten PRONOM-Einträgen. Da die Community schneller weitere PUID benötigte,[32] als TNA diese in PRONOM einpflegen konnte, bestand die Möglichkeit, vorläufige Einträge zu generieren. Diese sollten durch ein vorangestelltes «x» markiert werden (x-fmt). Da auf etliche x-fmt bereits verwiesen war, stellte sich eine spätere Löschung als problematisch heraus. Aus Stabilitätsgründen werden die vorhandenen x-fmt-Einträge daher weiterhin gepflegt. Zukünftig sollen jedoch keine weiteren vorläufigen PUIDs vergeben und bestehende zum Teil migriert werden.[33] Da sowohl die endgültigen als auch die vorläufigen PUIDs aus demselben Zahlenpool geschöpft haben, ohne in einem inhaltlichen Zusammenhang zu stehen, ergibt sich daraus eine gewisse Ambiguität: Das PDF 1.4 Format hat die PUID fmt/18, das Format Comma Separated Values besitzt die PUID x-fmt/18. Insgesamt haben aktuell ca. 450 PUIDs uneindeutige Nummern, die nur durch das vorangestellte x zu unterscheiden sind.

---

file for DROID. A guide by NLNZ [National Library of New Zealand]. http://openpreservation.org/system/files/how%20to%20write%20a%20sig%20file%20v1.1.pdf.

29 Vgl. Diagramm in Young, Paul: Identifiying digital file formats – a collaborative effort (2016). http://blog.nationalarchives.gov.uk/blog/identifying-digital-file-formats-collaborative-effort/.

30 Vgl. Tunnat, PRONOM (siehe Anm. 3).

31 «File format registries are expected to be persistent, trustworthy, and publicly discoverable.» Barve, Sunita (2007): File Formats in Digital Preservation. In: Proceedings of the International Conference on Digital Libraries, S. 239-248. http://dlissu.pbworks.com/f/File+format1.pdf.

32 Es wurde dadurch möglich, die Metadateien für ein AIP mit einem PUID zu bestücken, bevor diese als endgültige Formate in die PRONOM-Datenbank eingeflossen waren.

33 Vgl. http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm.

Aufgrund von Problemen in der Praxis der Formaterkennung wurden PUIDs zum Teil umstrukturiert. Die PUIDs fmt/7 bis fmt/10 waren für das Format TIFF bestimmt. Da aus technischen Erwägungen diese TIFF Varianten nun unter fmt/353 zusammengefasst werden, sind die vorgenannten PUIDs als veraltet gekennzeichnet und verweisen nun aus Stabilitätsgründen auf fmt/353.[34] Mit Stand Anfang 2017 sind 64 PUIDs aus vergleichbaren Gründen zurückgezogen wurden.[35]

Nicht abschließend geklärt ist indes die Frage, wie mit einem anstehenden Ingest verfahren werden soll, wenn noch kein PUID vorhanden ist.[36]

— Abwarten/ Ingest aufschieben.
— Signature bei PRONOM einreichen, damit PUID zugewiesen werden kann.[37]
— Ingest ohne PUID durchführen, nicht zuletzt auch deswegen, da, wie erwähnt, die Datenbasis der Formaterkennung und somit das Ergebnis (z.B. Scandifferenzen) aus verschiedenen Gründen dynamisch ist.

Das el_sta hat bereits eine Handvoll Ingests ohne PUID vorgenommen, so zum Beispiel mit SQLite-Dateien. Zeitgleich wurde der Vorschlag für eine neue Signature von verschiedenen Seiten bei TNA/PRONOM eingereicht. Inzwischen existiert für derartige Dateien ein PUID (fmt/729). Bisher hat das Staatsarchiv aber generell noch keinen Rerun der Formaterkennung durchgeführt.

Fazit

Sowohl die Datengrundlage (= PRONOM, seit 2002) als auch die Schnittstelle (= Signature-File,[38] seit 2005) als auch das Werkzeug (= DROID, seit 2005) sind im Laufe der letzten Jahre weiterentwickelt worden, umfangreicher und mächtiger geworden. Trotz aller Verbesserungen kann zu einem bestimmten Zeitpunkt jedoch immer nur eine gewisse Erkennungsquote erreicht werden. Innerhalb dieser Quote existiert auch immer eine Fehlerquote, wie oben beschrieben wurde.

34 Vgl. Tunnat: PRONOM (siehe Anm. 3).
35 Suche nach «deprecated» in der PRONOM Datenbank. http://www.nationalarchives.gov.uk/PRONOM/BasicSearch/proBasicSearch.aspx?status=new.
36 Frage auch bei Töwe, Geisser, Suri, To Act or Not To Act (siehe Anm. 3) sowie bei Mitcham, File identification… (siehe Anm. 15).
37 Signature-Vorschläge können unter https://www.nationalarchives.gov.uk/contact-us/submit-information-for-pronom/ eingereicht werden.
38 Ebenso werden parallel Container-Signatures gepflegt. Für weitere Informationen siehe http://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm.

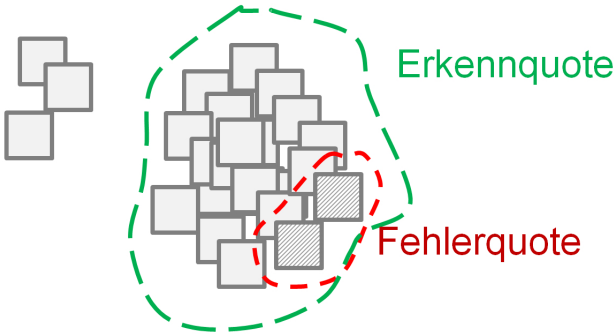

Abbildung 3: Bei der Formaterkennung besteht immer nur eine gewisse Erkennungsquote, innerhalb derer wiederum eine Fehlerquote zu erwarten ist. Eigene Darstellung.

Formaterkennung an sich ist «a work in progress, and therefore could not be considered complete at this time»,[39] sagt Jay Gattuso von der NZLZ im Jahre 2012. Er fügt hinzu: «Perhaps it is only possible to have a relative 'truth'".[40] Oben geschilderte An- und Auffälligkeiten im Umgang mit PRONOM und DROID sollen andeuten, dass bei Formaterkennungswerkzeugen Nutzen, aber auch Grenzen existieren. Formaterkennung bildet im Umfeld von Langzeitspeicherung und digitaler Archivierung nur einen ersten Schritt, jedoch gleichzeitig auch eine wichtige Voraussetzung.[41] Es sollte ein verlässlicher Überblick darüber zu schaffen sein, was im eigenen Repository vorliegt.[42] Dennoch soll auch darauf hingewiesen sein, dass Werkzeuge und Datengrundlagen daher immer nur als temporär und pragmatisch anzusehen sind.

Wie kann eine Institution auf dieser Grundlage handeln? Kurzfristig können andere Werkezuge hinzugezogen bzw. DROID komplett ersetzen werden. Als Alternative böte sich das Programm Siegfried von Richard LeHane an.[43] Dieses Werkzeug arbeitet u.a. auf der Basis von PRONOM-Signature-Files. Langfristig

39 Gattuso, Jay als Kommentar zu Blogeintrag von Wheatley: Don't panic! (siehe Anm. 18).

40 Gattuso, Throughput efficiencies and misidentification risks in DROID (siehe Anm. 3), S. 14.

41 Bachmann; Ernst, Formaterkennung (siehe Anm. 25), S. 69; Spencer, Ross: Generation of a Skeleton Corpus of Digital Objects for the Validation and Evaluation of Format Identification Tools and Signatures. In: The International Journal of Digital Curation 8 (1), 2013, S. 120-130, hier S. 120. http://www.ijdc.net/index.php/ijdc/article/view/8.1.120.

42 Weitergehende Informationen z.B. über Content Profiling bei Petrov, Petar; Becker, Christoph: Large-scale content profiling for preservation analysis, 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.303.6330&rep=rep1&type=pdf.

43 Für weitere Informationen vgl. LeHane, Richard: Siegfried – a PRONOM-based, file format identification tool, 2014. http://openpreservation.org/blog/2014/09/27/siegfried-pronom-based-file-format-identification-tool/.

wäre eine noch stärkere Vernetzung auf Arbeitsebene anzustreben bzw. die schon vorhandenen Plattformen auszubauen. Das nestor Netzwerk bietet mit seinen AGs, Praktikertagen, Workshops und Publikationskanälen vielfältige Optionen.[44] Weitere Einrichtungen wie die KOST oder der LWL (hier Archivberatung zur elektronischen Archivierung) wären wünschenswert. Nicht zuletzt sollten Anwendertreffen und Fachtagungen näher in den Fokus gerückt werden.

44 Siehe Website von nestor. http://www.langzeitarchivierung.de/.

Das E-Government-Gesetz NRW und die Praxis der Behördenberatung

Ein Werkstattbericht aus dem Landesarchiv NRW

Christine Friederich, Martin Schlemmer

Einleitung

Mit Behörden haben Archive jeden Tag zu tun. Behörden produzieren bei ihrer Tätigkeit Unterlagen, die von Archivarinnen und Archivaren bewertet, übernommen, erschlossen und der Nutzung zugänglich gemacht werden. Manchmal werden Behörden auch zu Nutzern, wenn sie bereits an das Archiv abgegebene Unterlagen noch einmal für ihre Arbeit benötigen. Archive und Behörden begegnen sich also in unterschiedlichen Szenarien. Bei der Behördenberatung durch das Archiv steht die Behörde mit ihrer Funktion als Schriftgutproduzentin im Mittelpunkt. Da es allein von der Behörde abhängt, ob vollständige und aussagekräftige Akten entstehen, ist die erfolgreiche Beratung zur Schriftgutverwaltung im Vorfeld eine entscheidende Voraussetzung dafür, dass aussagekräftiges Archivgut entstehen kann. Dieser Zusammenhang ist nicht neu, hat durch die in den letzten Jahren verabschiedeten E-Government-Gesetze in Deutschland und das «Organisationskonzept elektronische Verwaltungsarbeit»[1] – welches das bisherige DOMEA-Konzept abgelöst hat – als fachliche Grundlage jedoch an Aktualität gewonnen. Für Archive ergeben sich angesichts dieser geänderten Rahmenbedingungen neue Möglichkeiten und Herausforderungen in ihrer Behördenberatung: Ist eine Aktualisierung der eigenen Beraterrolle erforderlich? Unterscheidet sich die Behördenberatung für herkömmliche Unterlagen so sehr von derjenigen für digitale Unterlagen, dass dafür eigene Überlegungen nötig sind? Um dem nachzugehen, werden wir im Folgenden in einem Werkstattbericht aus dem Landesarchiv Nordrhein-Westfalen vorstellen, wie das E-Government-Gesetz NRW die Praxis der Behördenberatung verändert hat und wie wir mit den daraus entstandenen Herausforderungen umgehen. Zunächst werden wir die rechtlichen und organisatorischen Rahmenbedingungen für Behördenberatung im Kontext des E-Governments in NRW darlegen. In einem zweiten Schritt werden wir die damit verbundenen Herausforderungen knapp skizzieren und dann unser Beratungsangebot näher erläutern. Schließlich werden wir ein kurzes Fazit

1 Vgl. http://www.cio.bund.de/Web/DE/Architekturen-und-Standards/Organisationskonzept-E-Verwaltung/organisationskonzept_e-verwaltung_node.html. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.10

ziehen. Wir möchten so einen Beitrag leisten zur Debatte über Digitalisierung in Verwaltung und Archiven und den fachlichen Erfahrungsaustausch befördern.[2]

Rahmenbedingungen: Das E-Government-Gesetz NRW

Verlieren wir zunächst ein paar Worte zu den (rechtlichen) Rahmenbedingungen unserer Beratungstätigkeit für die Landesverwaltung von Nordrhein-Westfalen. Grundstein ist hier das im Juli 2016 in Kraft getretene «Gesetz zur Förderung der elektronischen Verwaltung in Nordrhein-Westfalen» (E-Government-Gesetz NRW / EGovG NRW), das umfangreiche Maßnahmen zur Digitalisierung und Modernisierung der Landesverwaltung auf den Weg gebracht hat. Darin ist auch der für fast alle Landesbehörden verpflichtende Umstieg auf die elektronische Aktenführung bis zum 1. Januar 2022 festgelegt.[3] Da hier aktuell auch der Schwerpunkt der Beratungstätigkeit des Landesarchivs NRW liegt, werden wir uns im Folgenden auf diesen Bereich konzentrieren.

Generell ist mit der Umsetzung des E-Government-Gesetzes NRW eine gezielte Steuerung der dafür notwendigen Maßnahmen verbunden, mit dem Ziel der Standardisierung und Vereinheitlichung des E-Governments.[4] Geleitet und koordiniert werden diese Maßnahmen vom «Beauftragten der Landesregierung Nordrhein-Westfalen für Informationstechnik», dem CIO[5]. Diese Konzeption wirkt sich auch auf den Bereich der elektronischen Aktenführung aus. So wird es voraussichtlich ab 2018 ein landeseinheitliches Dokumentenmanagementsystem (DMS) geben sowie Vorgaben für die elektronische Schriftgutverwaltung in Form von Regelungswerken (z.B. Muster-Aktenordnung), Handreichungen oder Best-Practice-Empfehlungen. Für die Umsetzung des E-Government-Gesetzes NRW sind bestimmte organisatorische Formen vorgesehen, etwa ein Projektmanagement nach festgelegten Standards, das wiederum in ein beim CIO gebündeltes, standardisiertes Programmmanagement zusammenläuft. Auch die Notation für die Beschreibung von Geschäftsprozessen ist landeseinheitlich festgelegt.[6] Diese Beispiele sollen nur

2 Vgl. hierzu Sulzbacher, Cornelia; Marckhgott, Gerhart: Tempora mutantur – nos et mutemur in illis!. In: Scrinium 69 (2015), S. 146-163.

3 § 9 EGovG NRW.

4 Maßgeblich hierfür sind v.a. der «Masterplan zur Umsetzung des E-Government-Gesetzes und zur weiteren Modernisierung in der Landesverwaltung Nordrhein-Westfalen» der Landesregierung Nordrhein-Westfalen sowie das «Konzept Programm- und Projektmanagement» des CIO. Beide Texte befinden sich aktuell noch im Entwurfsstadium.

5 Seit dem 1.11.2013 ist Hartmut Beuß Chief Information Officer (CIO). Die Stabsstelle des CIO ist beim Ministerium für Inneres und Kommunales des Landes Nordrhein-Westfalen (MIK) angesiedelt.

6 Bekanntgabe der landeseinheitlichen Notation für die Durchführung von Geschäftsprozessanalysen gemäß § 12 des E-Government-Gesetzes Nordrhein-Westfalen, Runderlass des Ministeriums für Inneres und Kommunales vom 21. Dezember 2016, MBl. NRW. 2017 S. 16.

einen kleinen Einblick ermöglichen, ohne jedoch einen Anspruch auf Vollständigkeit zu erheben.

Das E-Government-Gesetz NRW legt in § 11 den Beratungsauftrag des Landesarchivs NRW bei der Umsetzung des Gesetzes fest. Dieser Auftrag resultiert aus der im Archivgesetz des Landes NRW verankerten Beratungsaufgabe, die Landesverwaltung «bei der Verwaltung, Aufbewahrung und Sicherung ihrer Unterlagen» zu unterstützen.[7] Das Landesarchiv NRW versteht die im EGovG formulierte Beratungsaufgabe als Querschnittsaufgabe und hat diese dem im August 2016 neu eingerichteten Dezernat F 4: Elektronische Unterlagen als Arbeitsschwerpunkt übertragen. Die archivische Behördenberatung zur Umsetzung des E-Government-Gesetzes NRW wird folglich zentral durchgeführt. Aktuell sind in diesem Bereich vier Kolleginnen und Kollegen des höheren Dienstes tätig, die derzeit vier Ministerien intensiv betreuen.

Einen zweiten Schwerpunkt bildet die Mitarbeit in ressortübergreifenden Arbeitsgruppen zur Umsetzung des E-Government-Gesetzes NRW. Diese Arbeitsgruppen sind damit beauftragt, Standards und Best Practices für die elektronische Verwaltungsarbeit in Nordrhein-Westfalen zu erstellen. Das Landesarchiv NRW ist vorrangig in die Arbeitsgruppe E-Akte eingebunden und dort federführend bei den Arbeitspakten «Aktenplan und Aufbewahrung», «Verwaltungsvorschrift E-Akte», « (Muster-)Aktenordnung» sowie «Handreichungen zur Schriftgutverwaltung». Eine Mitwirkung besteht beim Arbeitspaket «Definition Landesstandard 1.0», der die Standard-Konfiguration für das geplant landesweit einzusetzende DMS beschreiben wird. Die aktive Mitgestaltung aller wichtigen landesweiten Standards der Schriftgutverwaltung – zum Teil sogar unter Federführung des Landesarchivs NRW – ist in der Bundesrepublik leider nicht selbstverständlich und für uns das Ergebnis langjähriger Vorfeldarbeit. Für das Landesarchiv NRW ist die Arbeit in den entsprechenden Gremien allerdings nicht nur eine gute Gelegenheit, sich selbst innerhalb der Landesverwaltung zu präsentieren. Vielmehr geht es darum, die Grundlagen für eine möglichst reibungslose, qualitativ hochwertige spätere Archivierung elektronischer Unterlagen zu schaffen. Zudem steht die Formulierung von Landesstandards auch in unmittelbarem Zusammenhang mit unserer Beratungstätigkeit, da hier Fragen beantwortet werden, die sich im Rahmen der Beratung häufig ergeben.[8]

Mit der Aufgabe Beratung ist im Kontext des E-Government-Gesetzes NRW jedoch nicht nur das Landesarchiv NRW betraut. Dazu kommt mit dem Competence Center Digitalisierung (CCD), das beim zentralen IT-Dienstleister IT.NRW

7 § 3 Abs. 6 ArchivG NRW.

8 Vgl. auch Gillner, Bastian; Pilger, Kathrin: Von Erinnerungspolitik und E-Government. Archive in der Rolle als politische Berater. In: Storm, Monika (Red.): Archive in der Wissensgesellschaft (Tagungsdokumentation zum deutschen Archivtag, 21) [in Vorbereitung].

angesiedelt ist, ein Akteur, der – mit beträchtlichem Personaleinsatz – unter anderem für die Beratung im technischen Bereich zuständig ist. Auch die Inanspruchnahme externer Beratungsunternehmen durch Behörden ist im Rahmen des Umsetzungsprozesses vorgesehen.

Vor diesem Hintergrund geht es also auch um eine Standortbestimmung für die archivische Behördenberatung. Denn wenn wir nur ein Akteur unter vielen sind, worin besteht dann unsere spezifische Beratungsleistung? Welche Kompetenzen und Wissensinhalte können wir Behörden zur Verfügung stellen, damit diese besser gerüstet sind auf ihrem Weg zum E-Government?

Herausforderungen: Wer sind wir und was tun wir da eigentlich?

Selbstverständlich ist Behördenberatung keine neue Aufgabe für Archive, sondern gehört zu den Kernaufgaben. Auch entspricht in der digitalen Welt das Beratungsziel grundsätzlich dem in der analogen Behördenberatung: Es geht darum, die Behörden in die Lage zu versetzen, eine Schriftgutverwaltung zu betreiben, die den gesamten Lebenszyklus im Blick hat und die aussagekräftige, rechtskonforme und letztlich auch archivfähige Akten und Vorgänge hervorbringt.

Geändert haben sich die Rahmenbedingungen für die Behördenberatung und damit auch die Möglichkeiten für das Landesarchiv NRW, in diesem Bereich zu agieren. Doch auch die zu beratenden Behörden befinden sich durch das E-Government-Gesetz NRW in einer neuen Situation. Mit der Verpflichtung zum Umstieg auf die E-Akte rückt das in den meisten Behörden bislang eher vernachlässigte Thema Schriftgutverwaltung auf einmal in den Mittelpunkt. Funktionale Defizite in der Schriftgutverwaltung, die bisher vielleicht etwas störend, aber vernachlässigbar schienen, treten nun offen zu Tage und werden zum Hindernis im Prozess beim Umstieg auf die E-Akte. Dabei zeigt sich häufig, dass ausreichendes Wissen über Schriftgutverwaltung in der Behörde fehlt, sodass diese «Kompetenzlücke» von außen geschlossen werden muss. Das Landesarchiv NRW ist deshalb ein gefragter Ansprechpartner für Know-how über Schriftgutverwaltung. Dabei unterscheiden wir uns von den anderen Beratungsakteuren vor allem dadurch, dass wir Wissen aus drei unterschiedlichen Bereichen mitbringen:

- Wir sind Experten für digitale Schriftgutverwaltung
- Wir sind als Betreiber des digitalen Archivs für die Landesverwaltung Experten für die Aussonderung und digitale Archivierung von elektronischen Unterlagen und können damit den kompletten Lebenszyklus von Akten und Vorgängen fachlich begleiten; und
- Wir sind aufgrund unseres Aufgabenzuschnitts und unserer Erfahrung mit zahlreichen Behörden vernetzt und haben ressortübergreifend Einblicke in

unterschiedliche Verwaltungszweige, sodass wir eine breite Perspektive in unsere Beratung einbringen können.

Üblicherweise werden wir ganz zu Beginn des Umstellungsprozesses auf die E-Aktenführung von den Behörden mit einbezogen, weil sich bereits an dieser Stelle zeigt, dass grundlegende Kenntnisse und Instrumentarien der Schriftgutverwaltung geklärt sein müssen, bevor mit der eigentlichen Umsetzung begonnen werden kann. Das bietet den Vorteil, dass wir bereits bei der Neuausrichtung der Grundlagen der behördlichen Schriftgutverwaltung im Zuge der E-Akten-Einführung mitwirken können, also insbesondere bei der Überarbeitung des Aktenplans, der Aktenordnung oder bei der Festlegung der Aufbewahrungsfristen.

Für die Behörden stellt sich die eigene Schriftgutverwaltung oft als «black box» dar. Das hängt damit zusammen, dass in vielen Behörden die Schriftgutverwaltung in den letzten Jahrzehnten geradezu atomisiert wurde und eine zentrale Steuerung fehlte. So ist es mitunter den einzelnen Organisationseinheiten überlassen, den Aktenplan ab einer bestimmten Gliederungsebene selbst auszugestalten. Das führt zu einer sehr heterogenen Strukturierung, zu einer Wiederholung von Querschnittsaufgaben unterhalb verschiedener Hauptgruppen und zu einer nicht angemessenen Gliederungstiefe. Teilweise vermischen sich dann auch noch Aktenplan und Aktenverzeichnis, indem für jede neu angelegte Akte auch eine neue Aktenplanposition im Aktenplan geschaffen wird. Im Extremfall entstehen dann ausufernde «Aktenpläne», die mehrere 10.000 Zeilen in Excel umfassen und so praktisch nicht mehr nutzbar sind. Zugleich fehlt es an Überblick über die so gewachsenen Strukturen, wenn Aktenpläne nicht zentral zusammengeführt werden. Fileablagen auf Organisationsebene, E-Mail-«Archive» der Bearbeitenden, Fachverfahren und rudimentäre Papier-Restakten, deren Kontext und Entstehungszusammenhang sich kaum mehr herstellen lassen, erschweren einen Einblick, geschweige denn einen Überblick über den Stand der Dinge. Die Verpflichtung zur Aktenmäßigkeit, der das Verwaltungshandeln unterliegt, lässt sich so nicht mehr erfüllen.

Die E-Akte erscheint hier manchmal als eine regelrechte Verheißung. Dies gilt nicht nur für die Archiv-, sondern auch für die Behördenseite. Doch der Wunschtraum, dass mit der Nutzung der E-Akte auch alles gleich besser wird, ist eben genau dies: ein Wunschtraum. Ein DMS kann aus einer nicht funktionierenden Schriftgutverwaltung keine funktionierende machen. Rheinisch salopp formuliert: «Murks bleibt Murks». Oder noch drastischer, mit dem CEO von Telefónica Deutschland, Thorsten Dirks, auf dem Wirtschaftsgipfel der «Süddeutschen Zeitung» 2015: «Wenn Sie einen Scheißprozess digitalisieren, dann haben Sie einen scheiß digitalen Prozess.»[9] Diese Erkenntnis wird keineswegs nur vom Archiv in

9 Zitiert nach http://www.computerwoche.de/g/die-besten-it-sprueche-2015,106507,3.

die Behörden hineingetragen. Vielmehr macht der geplante Umstieg auf die elektronische Aktenführung vielen Behörden ihre Defizite in der Schriftgutverwaltung erst bewusst. «Kenne dich selbst» ist deshalb eine der ersten Voraussetzungen, die eine Behörde mitbringen muss, damit wir sinnvolle Beratungsarbeit leisten können.

Bei unserer praktischen Beratungstätigkeit hat sich bislang vor allem die Heterogenität der Behörden und ihres Vorgehens beim Umstieg auf die E-Akte als Herausforderung dargestellt. So stoßen wir auf ganz unterschiedliche Behördenkulturen, mit unterschiedlichen Belegschaften, unterschiedlichen Erfahrungen mit der E-Akte nach dem alten DOMEA-Konzept und mit unterschiedlichen Veränderungsgeschwindigkeiten. Es gibt Behörden, die den Sprung ins kalte Wasser wagen, und diejenigen, die über mehrere Zwischenschritte sich ganz langsam dem Thema E-Akte annähern. Auch die jeweilige Organisationsstruktur zur Umsetzung des E-Government-Gesetzes spielt eine Rolle, da es hier sehr unterschiedliche Varianten gibt. Teilweise sind Projektgruppen dafür verantwortlich, teilweise findet die Umsetzung des E-Government-Gesetzes ganz oder überwiegend durch die Linienorganisation statt.

Vor diesem Hintergrund mussten wir unsere Rolle als Beratende erst finden. Die Erwartungen der Behörden sind hoch, gleichzeitig mussten wir uns aber darüber bewusst werden, dass wir in dem Gesamt-Prozess «Umstieg auf die E-Akte» eben nur ein ganz bestimmtes Segment beratend unterstützen können. Für uns heißt das, dass wir Know-how und fachliche Standards für die digitale Schriftgutverwaltung in den Gesamtprozess einbringen und den Behörden bislang fehlende Kenntnisse in der Schriftgutverwaltung vermitteln. Ziel ist, dass die Behörden auf der Grundlage eines ausreichenden fachlichen Wissensstands die für sie jeweils passenden Entscheidungen beim Umstieg auf die E-Akte treffen können. Keinesfalls können wir Behörden jedoch solche Entscheidungen abnehmen oder Aufgaben übernehmen, die in den Zuständigkeitsbereich der Behörde fallen. So wird beispielsweise immer wieder irrtümlich angenommen, dass wir für die Festsetzung der behördlichen Aufbewahrungsfristen zuständig seien. Wir weisen dann darauf hin, dass nur die Behörde selbst wissen und entscheiden kann, wie lange bestimmte Unterlagen aufbewahrt werden müssen. Ähnlich verhält es sich mit der Erstellung eines Aktenplans. Auch hier können wir nicht für die anfragende Behörde deren kompletten Aktenplan erstellen. Wir können jedoch dabei behilflich sein, indem wir grundlegende Kenntnisse vermitteln, die zur Erstellung eines guten Aktenplans nötig sind. Mittlerweile haben wir gelernt, diese Grenzen unseres Beratungs-Portfolios aktiver zu kommunizieren, um nicht Erwartungen zu wecken, die wir letztlich nicht erfüllen können. Wir möchten also als Themenexperten wahrgenommen werden, was wir mittlerweile auch erfolgreich vermitteln konnten.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.10

Doch auch unsere fachlichen Inhalte stellen uns immer wieder vor Herausforderungen. Vor dem Hintergrund einer Digitalisierung der Aktenführung müssen auch die Grundlagen der Schriftgutverwaltung neu überdacht und an machen Stellen entsprechend nachjustiert werden. So bringt etwa die im «Organisationskonzept elektronische Verwaltungsarbeit»[10] vorgeschriebene Objekthierarchie «Akte – Vorgang – Dokument»[11] allerhand Auswirkungen beispielsweise auf die notwendige Gliederungstiefe des Aktenplans mit sich. Oft stehen wir mit einem Bein tief im «alten Preußen» und mit dem anderen in der neuen Welt der E-Akte. Deshalb haben wir in den letzten Monaten auch intensiv beispielsweise an den Grundbegriffen der Schriftgutverwaltung gearbeitet. Diese Erkenntnisse fließen in unser Beratungsangebot ein.

Beratungsangebot

Unser Beratungsangebot umfasst unterschiedliche Formate. Zum gegenseitigen Kennenlernen und zum Abklären des Beratungsbedarfs der Behörde sowie der Beratungsangebote des Landesarchivs NRW findet zunächst ein gemeinsames Gespräch statt. Im Anschluss wird in einem vom Landesarchiv NRW erarbeiteten Fragebogen der Stand der Schriftgutverwaltung abgefragt. Dabei geht es nicht nur darum, dass wir einen genaueren Einblick erhalten, wie Schriftgutverwaltung in der jeweiligen Behörde gehandhabt wird. Vielmehr hat sich gezeigt, dass das Ausfüllen des Fragebogens für die Behörde selbst sehr erhellend ist und oftmals geradezu einen «Aha»-Effekt hervorruft. Vielen wird erst dann deutlich, wie wenig sie eigentlich über die aktuell gepflegte Schriftgutverwaltung in ihrem Haus informiert sind. Deshalb ist der Fragebogen oft der erste Impuls für den Weg hin zu dem bereits beschriebenen «Kenne dich selbst!» als Voraussetzung für den Umstieg auf die E-Akte. In einem zweiten, zeitnahen Gesprächstermin mit der Behörde werden dann der Fragebogen besprochen sowie das weitere Vorgehen festgelegt. Häufig gibt es noch offene Fragen, die erst geklärt werden müssen, bevor weitergearbeitet werden kann, oder es sind noch nicht die passenden Zusammensetzungen der Arbeitsgruppen gefunden etc. Wie genau der weitere «Fahrplan» aussieht, liegt in erster Linie in der Hand der Behörde, die für den zeitlichen und organisatorischen Rahmen der Einführungsphase verantwortlich ist. Generell ist angestrebt, nach dem

10 Das «Organisationskonzept elektronische Verwaltungsarbeit» des Bundesministeriums des Innern besteht aus mehreren Bausteinen, die hier abgerufen werden können: https://www.verwaltung-innovativ.de/SharedDocs/Publikationen/Organisation/e_akte.html.

11 Die terminologischen Probleme bei der Übersetzung der Objekthierarchie in andere Verwaltungssysteme oder Sprachen wurden bereits mehrfach u. a. im Kontext der Normierungsbemühungen in diesem Bereich formuliert. In der Schweiz entspricht die Akte in etwa einem Dossier, der Vorgang einem Subdossier, in Österreich einem Akt bzw. einem Akt-Einzelstück

zweiten Gesprächstermin in die inhaltliche Arbeitsphase in Form von Workshops einzusteigen. Die Workshops sind halb- oder ganztägig angelegt und sollen den Behörden praxisnah das notwendige Know-how vermitteln, um die anstehenden Aufgaben in der digitalen Schriftgutverwaltung fachgerecht zu erfüllen. Aktuell arbeiten wir an einem Workshop-Portfolio zu folgenden Themen:

- Grundbegriffe und Grundsätze der digitalen Schriftgutverwaltung
- Aktenplan und Aktenordnung
- Scanprozesse / Ersetzendes Scannen
- Aufbewahrung, Aussonderung, Archivierung
- Herausforderungen bei der Einführung von Dokumentenmanagementsystemen (Do‘s & Don‘ts)
- Aktenfreie Verwaltung? Fachverfahren und Dokumentenmanagementsysteme

Im Moment sind vor allem die Angebote zu den Grundlagen der digitalen Schriftgutverwaltung und zu Aktenplan und Aktenordnung gefragt. Das hängt damit zusammen, dass dies meist die ersten Schritte sind, die auf dem Weg zur E-Akte gemacht werden müssen. Drei der vier von uns betreuten Ministerien stehen derzeit kurz vor dem Eintritt in die «Workshop-Phase», eines hat bereits einen Workshop absolviert. Die Resonanz darauf war positiv, insbesondere auf den praktischen Teil in Form eines Planspiels zur Erstellung eines Aktenplans. Dennoch sehen wir auch hier Potenzial zur Weiterentwicklung und haben begonnen, Inhalt und Ablauf des Workshops einer kritischen Revision zu unterziehen und das Workshop-Konzept auf Grundlage dieser Ergebnisse zu überarbeiten.

Ein weiteres Thema, das häufig bereits bei den ersten Gesprächen genannt wird, ist die Mitarbeiterschulung im Bereich digitale Schriftgutverwaltung. Dies wird meist im Rahmen des Change-Managements geplant und kann ganz unterschiedliche Formate haben, von reinen Informationsveranstaltungen, in denen beispielsweise der überarbeitete Aktenplan vorgestellt und Grundbegriffe der Schriftgutverwaltung erläutert werden, bis hin zu konkreten Praxisschulungen. Auch hier sind wir noch in den Planungen. Allerdings kristallisiert sich bereits heraus, dass unser Part nur der sein kann, Grundlagen der digitalen Schriftgutverwaltung zu vermitteln. Alles, was konkrete organisatorische Umsetzungen, rechtliche Fragen, das Change-Management im engeren Sinne usw. angeht, fällt in den Aufgabenbereich der Behörde.

Fazit

Der Wunsch nach Veränderung und die Bereitschaft, die neuen Entwicklungen und die Umsetzung des E-Government-Gesetzes NRW mitzutragen, sind in den nord-

rhein-westfälischen Landesbehörden deutlich spürbar. Das Landesarchiv NRW wird dabei als kompetenter und verlässlicher Partner wahrgenommen, der die in den Behörden vorhandenen Kompetenzlücken im Bereich der Schriftgutverwaltung schließen helfen kann. Gleichzeitig sind nicht nur die Behörden, sondern auch wir als Landesarchiv NRW Lernende im Sinne einer «lernenden Verwaltung», die neuen Anforderungen mit Offenheit und Bereitschaft zur Veränderung gegenübertritt.[12]

Für das Dezernat F 4 des Landesarchivs NRW waren allein schon die Reflexion über die eigenen Arbeitsschwerpunkte, die Kernkompetenzen und Alleinstellungsmerkmale, die Definition der eigenen Rolle im Beratungsprozess auf dem Weg zur E-Akte sowie die Präzisierung unseres Beratungs-Portfolios äußerst wertvoll. Der Beratungsprozess kann demnach durchaus als beiderseitiger Lernprozess betrachtet werden. Die Behörden und Gerichte sind angehalten, sich auf dem Weg zur E-Akte mit ihrer Schriftgutverwaltung zu befassen; doch auch die beratende Seite sollte die Gelegenheit zur Selbstreflexion nutzen, da dies wiederum der Qualität der Beratung zugute kommt. Wenngleich sich die Schriftgutverwaltung in vielen Behörden weitgehend «verselbstständigt» hat und häufig – wenn überhaupt – lediglich auf der Ebene der untersten Organisationseinheiten geregelt ist beziehungsweise gelebt wird, so bietet doch das EGovG NRW nun die Gelegenheit, die Schriftgutverwaltung für die gesamte Organisation in Augenschein zu nehmen und in ihrer Gänze einer Neuordnung zu unterziehen.

Insgesamt profitiert das Landesarchiv NRW in hohem Maße von seiner Beratungs- und Mitgestaltungsrolle im «E-Government»-Prozess. Es kann sich als Dienstleister für die Verwaltung neu aufstellen und profilieren und kann zugleich alle notwendigen Rahmenbedingungen und Voraussetzungen für die Archivierung elektronischer Unterlagen aktiv und rechtzeitig mitgestalten. Dieser Herausforderung für Archive im digitalen Zeitalter stellen wir uns.

Eine obere Landesbehörde notierte in unseren Fragebogen zum aktuellen Stand der Schriftgutverwaltung, sie sei «in Sachen Schriftgutverwaltung ein weißes Blatt, was grundlegend neu gestaltet werden» könne. Den Aufbruch der Verwaltung im Zuge der Einführung der elektronischen Verwaltung sollten wir in diesem Sinne als Chance begreifen. Wenn wir sie nutzen, könnte diesem Anfang ein gewisser Zauber inne wohnen.

12 Vgl Seibel, Wolfgang: Verwaltung verstehen. Eine theoriegeschichtliche Einführung. Berlin 2016; Land Baden-Württemberg – Stabsstelle für Verwaltungsreform; Mauch, Siegfried (Bearb.): Qualitätsmanagement und lernende Organisation in der Landesverwaltung Baden-Württemberg. Eine Wegbeschreibung zur Förderung der Selbstentwicklungsfähigkeit in der öffentlichen Verwaltung. Stuttgart 1999.

Archivierung aus Fachanwendungen

Ein Werkstattbericht aus dem Staatsarchiv Graubünden

Ursina Rodenkirch-Brändli, Bernhard Stüssi

Seit dem 1. Januar 2016 ist die Archivierung für die Graubündner Behörden verbindlich geregelt. Im «Gesetz über die Aktenführung und Archivierung» werden unter anderem die Nachvollziehbarkeit und die Dokumentierung des Handelns der Behörden als Ziele definiert.[1] Ein Ordnungssystem ist vorgeschrieben, nicht aber das Dossierprinzip. Letzteres schien dem Gesetzgeber offenbar nicht erforderlich für eine funktionierende Aktenführung und Archivierung – obschon seine Einhaltung in der archivischen Praxis auch in Graubünden lange Zeit als Selbstverständlichkeit gesehen wurde. Was in der «Papierwelt» über Jahrhunderte erprobt und einfach vorstellbar ist, bewährt sich in Form von Dateiablagen auch in der elektronischen Welt. Ein grundsätzliches Problem der Archivierung stellt sich allerdings mit der zunehmenden Verbreitung von Fachanwendungen in neuer Schärfe: Zuerst war nicht die Archivierung, sondern die Aufgabe. Das Strassenverkehrsamt vergibt und entzieht Nummernschilder an Autofahrer, die Sozialversicherungsanstalt spricht Invalidenrenten und revidiert diese periodisch. Entsprechend organisieren die Behörden ihre Aktenführung so, dass sie diese Aufgaben wahrnehmen können. Wenn dies ohne Dossiers auch oder gar besser funktioniert: Das Verwaltungshandeln will dennoch dokumentiert sein.

Jean-Yves Rousseau und Carol Couture schlagen eine organische Herangehensweise vor, die das Problem elegant aufzulösen verspricht. Sie unterscheiden zwischen dem «dossier» als Zuständigkeit für ein Projekt, für eine bestimmte Tätigkeit und dem «dossier» als «article» (Akteneinheit), der dokumentarischen Spur, die bei der Erfüllung einer Aufgabe entsteht.[2] «C'est l'analyse de la réalité documentaire à traiter qui fait foi de l'usage approprié de l'unité de travail que constitue l''article'.»[3]

Mit diesem Ansatz gingen wir die Archivierung aus Fachanwendungen zweier Behörden an: des Strassenverkehrsamts und der Sozialversicherungsanstalt.

1 Gesetz über die Aktenführung und Archivierung vom 28. August 2015 (GAA, BR 490.000), Inkrafttreten am 1. Januar 2016.

2 Die KOST-Terminologie schlägt für «article» u. a. «Akteneinheit, Artikel, Verzeichnungseinheit» (nach VSA) oder «Element» (nach ICA) vor.

3 Rousseau, Jean-Yves; Couture, Carol et al.: Les fondements de la discipline archivistique, Québec 1994 (reprint ebd. 2011), S. 122.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.11

Zur Veranschaulichung behandeln wir hier den Entzug von Kontrollschildern und die Invalidenversicherung.

Der Entzug von Kontrollschildern ist eine Routineaufgabe, die stets etwa gleich abläuft und meistens aufgrund nicht vorhandener Haftpflichtversicherung vorgenommen wird. In diesem Falle erhält das Strassenverkehrsamt von der Versicherung via Fachapplikation die Information, dass die Haftpflichtversicherung zu einem Fahrzeug erloschen ist. Das System generiert darauf hin diverse Briefe an den Halter mit der Aufforderung, einen Versicherungsnachweis beizubringen. Wird dem nicht nachgekommen, generiert das System ein Schreiben an die Kantonspolizei zum Einzug der Schilder. Dieser Prozess ist vollautomatisiert und wird durch das Strassenverkehrsamt lediglich überprüft. Das Handeln der Abteilung Fahrzeugzulassung bildet sich als Evidenzwert in den Unterlagen ab: Der Ablauf eines jeden Kontrollschild-Entzugs lässt sich anhand von Meldungseingang und Standardbriefen nachvollziehen. Der Gleichförmigkeit der Aufgabe steht die Gleichförmigkeit der Unterlagen gegenüber; es genügt folglich zur Archivierung eine Bemusterung. Zur Abbildung des Informationswertes können demzufolge die Abläufe nicht hinzugezogen werden. Dazu werden aus dem System Exporte von Stammdaten erstellt und archiviert. Die Überlegungen hinter diesem Vorgang lassen sich im Projekt 14-001 ViaCar/CARI der Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) nachlesen.[4]

Anders verhält es sich mit der Invalidenversicherung: Obwohl formalrechtlich ebenfalls detailliert geregelt, unterscheiden sich die einzelnen IV-Fälle erheblich voneinander. Art und Grad der Invalidität sind von Fall zu Fall verschieden, ebenso der Umgang der betroffenen Person mit ihrem Handicap. Der Zuspruch einer (Teil-)Rente hängt auch von der beruflichen Situation der/des Invaliden ab und nicht zuletzt von ihrer/seiner finanziellen Lage, insbesondere im Zusammenhang mit anderen Versicherungsleistungen. Um einen Überblick über den konkreten Vollzug der Aufgabe «Invalidenversicherung» durch die Sozialversicherungsanstalt (und dessen Wandel im Lauf der Zeit) zu erhalten, ist deshalb nicht nur auf den Evidenzwert, sondern vor allem auf den Informationswert der Unterlagen abzustellen.[5] Eine Bemusterung genügt hier zur Dokumentation nicht; gefragt ist eine Stichprobe von angemessener Grösse.

Die Praxistauglichkeit der Bewertungsentscheide «Bemusterung» (Abteilung Fahrzeugzulassung) und «Stichprobenziehung» (Invalidenversicherung) lässt sich

4 Siehe «14-001 ViaCar/CARI», https://kost-ceco.ch/cms/index.php?14-001_de. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

5 Statistik und zusammenfassende Berichte bieten einen Überblick auf der operativen Ebene der Sozialversicherungsanstalt; die Wandlung des Behördenhandels oder die Auswirkungen desselben auf einzelne Personen lassen sich damit aber kaum nachvollziehen.

gegen die eingesetzten Fachapplikationen prüfen: Letztere sind auf die Arbeitsinhalte und -abläufe der Behörden optimiert. Wenn die gewünschten Muster, respektive Stichproben, ohne grossen Zusatzaufwand (den gilt es bei der Archivierung für die Behörden ohnehin zu vermeiden) aus der Applikation gezogen werden können, so ist von einer angemessenen Abbildung der obgenannten Arbeitsinhalte und -abläufe auszugehen. Sollten dagegen die gewünschten Unterlagen technisch nicht oder nur sehr schwer zu erlangen sein, so deutet dies darauf hin, dass die Archivierung Aufgaben und Tätigkeit der betreffenden Behörde zu wenig berücksichtigt.

Was also sind die Akteneinheiten der Dokumentation des Entzugs von Kontrollschildern oder der Invalidenversicherung? Eine Akteneinheit der Abteilung Fahrzeugzulassung enthält eine kurze Beschreibung des Geschäftsvorgangs, ausgedruckte Dokumente (z. B. Standardbriefe), aus Datensätzen generierte Dokumente (z. B. einen Fahrzeugausweis) sowie Änderungen in Datensätzen (z. B. den Eingang der Meldung, dass ein Kontrollschild entzogen wurde).

Bei der Invalidenversicherung besteht a priori keine Akteneinheit. Aufgrund der Verschiedenheit der einzelnen IV-Fälle gibt es keine generische «réalité documentaire». Gleichwohl ist die Anzahl der möglichen Dokumenttypen beschränkt. Ein IV-Fall ist stets durch eine Anmeldung, Unterlagen zur Beurteilung (z. B. ärztliche Gutachten), einen Vorbescheid und eine Verfügung dokumentiert. Daneben kann eine Vielzahl an weiteren Unterlagen anfallen wie Lohnabrechnungen, Gesprächsnotizen, Handelsregisterauszüge, Einwände, Bewerbungsunterlagen usw.

Inwiefern sind nun die Fachanwendungen der Abteilung Fahrzeugzulassung und der Invalidenversicherung geeignet, solche Akteneinheiten auszugeben? Das System «CARI» enthält keinen Prozessbeschrieb. Der Standardbrief wird aufgrund von geänderten Informationen in der Datenbank automatisch erzeugt; der Versand wird ebenfalls durch einen Datenbankeintrag dokumentiert. Eine «physische» Kopie (beispielsweise als PDF) wird nicht gespeichert. Analog verhält es sich mit den Fahrzeugausweisen. Änderungen in Datensätzen werden ausschliesslich in der Datenbank selber dokumentiert; bei Bedarf können sie über eine Benutzeroberfläche angezeigt werden. Um eine Akteneinheit zu erzeugen, sind manuelle Benutzereingriffe im Einzelfall nötig: Die Beschreibung des Geschäftsvorgangs wird von einer Vertretung des Strassenverkehrsamts verfasst, Standardbriefe und Fahrzeugausweis werden als Muster erstellt und von der Benutzeroberfläche werden Screenshots gemacht. Der Aufwand ist beträchtlich, lässt sich aber für eine Bemusterung rechtfertigen.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.11

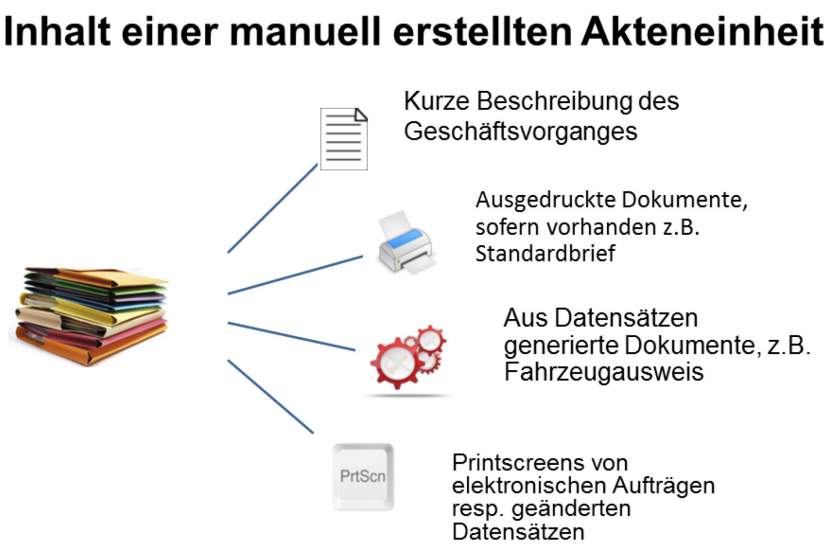

Abbildung 1: Manuell erstellte Akteneinheit der Abteilung Fahrzeugzulassung

Das auf einer Content-addressed-storage-Datenbank beruhende System der Sozialversicherungsanstalt enthält Dokumente in den Formaten TIFF und DOC. Geschäftsfälle werden als Prozessschritte behandelt und manuell eröffnet und geschlossen: Zu derselben Person können beispielsweise sowohl eine IV-Anmeldung als auch eine IV-Taggeld-Anmeldung oder eine IV-Taggeld-Revision als separate Geschäftsfälle vorhanden sein. Bei allen drei Vorgängen wird ein so genanntes Auslösedokument erstellt, zu dem allenfalls weitere, als «Begleitdokumente» bezeichnete Unterlagen abgelegt werden. Allein im Bereich Invalidenversicherung gibt es 126 verschiedene Auslösedokumente und damit Geschäftsfall-Arten. Es ist möglich, eine Akteneinheit (als so genanntes «virtuelles Dossier») auszugeben, indem nach Sozialversicherungsnummer und Dokumenttypen gefiltert wird. Für Abklärungen bei Fachstellen und Gerichtsfälle wurden solche Filter bereits im System implementiert. Für die Invalidenversicherung werden 432 Dokumenttypen berücksichtigt. Die vorhandenen Unterlagen werden zu einem PDF zusammengefasst und automatisch durch eine File History und ein Inhaltsverzeichnis ergänzt. Der Aufwand ist mässig und erlaubt auch eine breitere Übernahme in Form einer Stichprobenziehung.

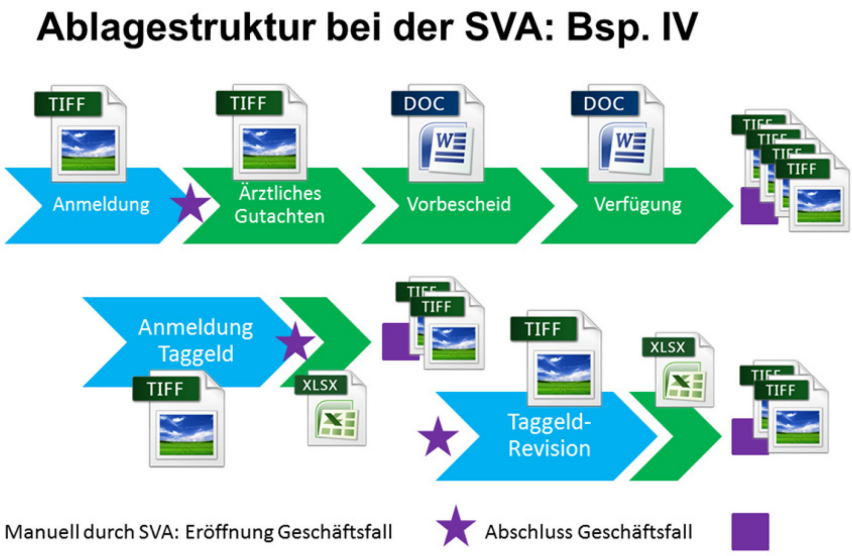

Abbildung 2: Ablagestruktur bei der Sozialversicherungsanstalt

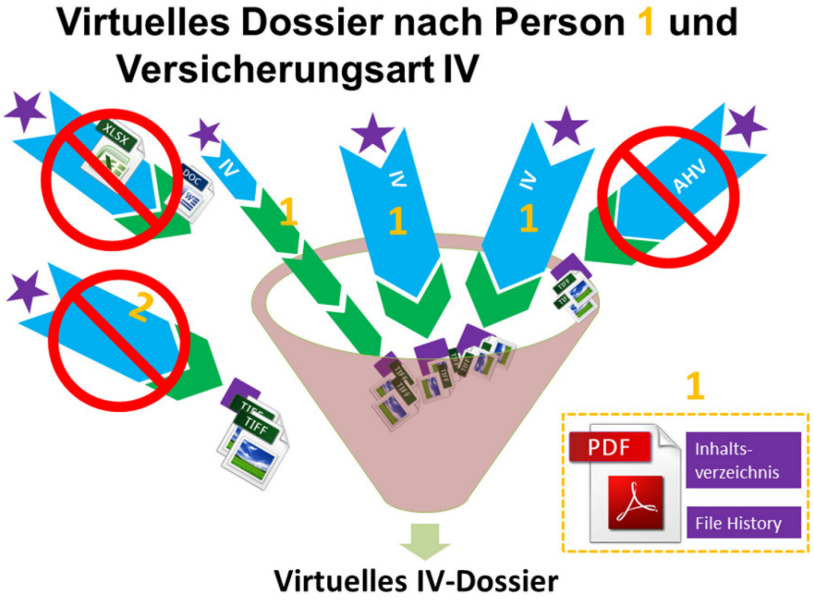

Abbildung 3: Erstellen eines «virtuellen Dossiers» bei der Sozialversicherungsanstalt

Es stellte sich heraus, dass die Analyse der Aufgaben und Tätigkeiten von Strassenverkehrsamt und Sozialversicherungsanstalt und die auf dieser Grundlage vorgenommenen Bewertungsentscheide mit den bei den Behörden vorhandenen Fachapplikationen umsetzbar sind. Das von Rousseau und Couture skizzierte Vorgehen hat sich bewährt: Die Aufgabenanalyse und die Suche nach der dokumentarischen Spur dieser Aufgaben sind auch bei der Archivierung aus Fachanwendungen ein zielführender Ausgangspunkt.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.11

Arbeitsbericht zur Archivierung von Netzressourcen im Staatsarchiv des Kantons Basel-Stadt

Kerstin Brunner, Olivier Debenath

Der vorliegende Bericht dokumentiert die Vorgehensweise zur und die Erkenntnisse aus der aktuellen Praxis der Archivierung von Netzressourcen im Staatsarchiv des Kantons Basel-Stadt. Der Text ist in zwei Teile gegliedert: Der erste Teil beschreibt die archivischen Überlegungen und Entscheidungen hinter dem gewählten Vorgehen. Der zweite Teil des Arbeitsberichts befasst sich mit der Unternehmensarchitektur der technischen Lösung zur Sicherung von Webseiten.

Archivische Aspekte zur Webarchivierung

Anfänge

Mit dem Thema der Archivierung von Netzressourcen beschäftigt sich das Staatsarchiv Basel-Stadt seit 2006. Erste Sicherungen fanden bereits Ende 2008 statt. Dies geschah im Hinblick auf eine bevorstehende Verwaltungsreorganisation, welche tiefgreifende Umstrukturierungen im Verwaltungsapparat mit sich brachte, was sich wiederum in der grundlegenden Überarbeitung oder gar Abschaltung diverser Webseiten niederschlug.

Aufgrund der geringen verfügbaren Ressourcen für die Archivierung von Webinhalten erfuhr das Thema erst ab Ende 2011 einen neuen Arbeitsschub. Über mehrere Monate hinweg wurden verschiedene Ansätze innerhalb der schweizerischen Bibliotheks- und Archivlandschaft studiert. Die Analyse von Intranet- und Internetseiten des Kantons führte zur Identifikation geschäftsrelevanter Inhalte. Eine hausinterne Sicherung ausgewählter Seiten erwies sich als notwendig, was eine Evaluation bestehender Arbeitsweisen und Konzepte zur Folge hatte.

Es wurde entschieden, zwischen 2014 und 2016 eine Pilotphase durchzuführen, anhand derer einerseits Erfahrungen für eine regelmässige, beständige und gut durchdachte Sicherungsstrategie für Netzressourcen gemacht werden sollten, andererseits Prozesse implementiert wurden. Kosten und Personalaufwand sollten bis zur Evaluation der Pilotphase möglichst gering bleiben.

Aufgrund der hohen Anzahl an verfügbaren Seiten und Inhalten wurde ein Bewertungsvorschlag für die Pilotphase in Arbeit genommen. Ausserdem wurden

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.12

verschiedene Open-Source-Tools für die Sicherung der Seiten getestet; der Entscheid fiel schliesslich auf das Web Curator Tool (WCT).[1]

Bewertung

Zum Kernbereich der Überlieferungsbildung gehören Netzpublikationen der kantonalen Verwaltung ohne den Einbezug weiterer ‚Anbieter' aus dem halbstaatlichen oder privaten Umfeld.[2] Eine Übernahme in Auswahl wird im Bereich der Schulen verfolgt (Vorgabe: alle drei Bezirke sollten vertreten sein sowie nach Möglichkeit laufende Bestrebungen zur Schulharmonisierungs-Aktion HARMOS mit abgebildet werden). Nach einer Prüfung der Intranet-Angebote wurde auf die Sicherung derselben verzichtet, da die Inhalte überwiegend organisatorischen und administrativen Zwecken dienen. In Bezug auf partnerschaftliche Organisationen der beiden Halbkantone Basel-Stadt und Basel-Landschaft wurde auf den Bewertungsentscheid bezüglich der Zuständigkeit für physische Unterlagen zurückgegriffen. Temporäre Internetangebote, Abschaltungen, Relaunches oder ausschliesslich im Internet vorhandene Angebote sichert das Staatsarchiv Basel-Stadt nach Möglichkeit, jedoch kann kein periodisches und umfassendes Monitoring solcher Seiten erfolgen.

Gesamthaft kamen so rund 150 Webseiten zusammen, welche periodisch einmal jährlich geharvestet werden. Im Zentrum steht primär die Sicherung von Inhalten; Fragen zu Urheber- und Persönlichkeitsrechten werden im Nachgang der Pilotphase diskutiert.

Verzeichnung

Unter der Abteilung ‚Sammlungen' haben wir einen Fonds ‚Webarchiv' angelegt. Unter diesem Zweig existiert pro Dienststelle ein Bestand, je nach URL werden Serien vergeben.

Abbildung 1: Verzeichnungsstruktur Webarchiv, Ausschnitt

1 Siehe http://webcurator.sourceforge.net/. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Nicht berücksichtigt werden demnach öffentlich-rechtliche Körperschaften, welche nur bestimmte Aufgaben im Auftrag des Kantons erledigen; dasselbe gilt für Personen / Organisationen / Firmen etc., die den Kanton als solchen zwar ‚prägen', aber nicht der Verwaltung angehören oder dieser in irgendeiner Form angebunden sind.

Den Snapshot[3] verzeichnen wir in einem Dossier. Das zugehörige Formular wurde spezifisch für Netzressourcen generiert und enthält Metadaten-Felder für die WCT Target Instance ID, die Anzahl Dateien und das Dateivolumen.

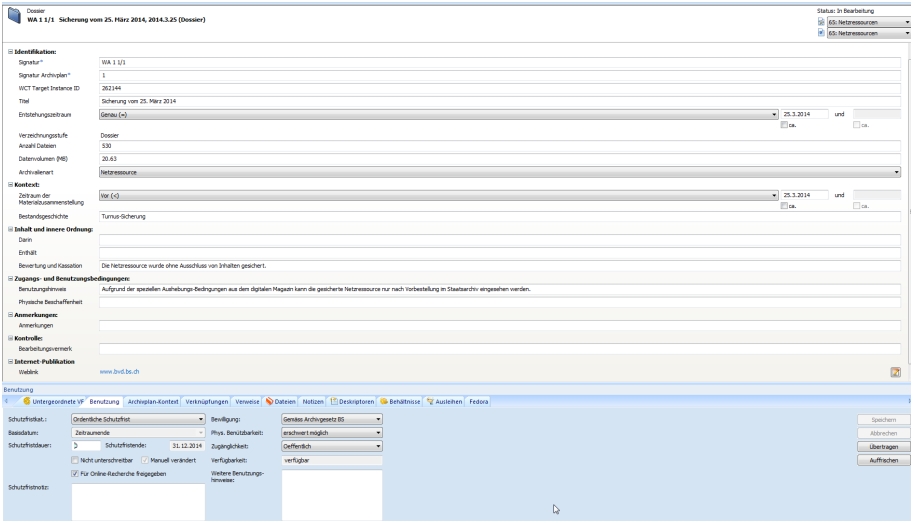

Abbildung 2: Web-Snapshot als Dossier verzeichnet

Die gesicherten Daten werden via Ingest in die Verzeichnungseinheit übernommen. Der Zugriffspfad auf die WARC-Daten in unserem Digitalen Magazin wird angelegt.

Benutzung

Die vom WCT zur Verfügung gestellte Oberfläche dient lediglich zur internen Qualitätssicherung. In Zukunft wird die Benutzung via den Digitalen Lesesaal stattfinden, der sich momentan noch im Aufbau befindet. Um die gesicherten Netzressourcen jedoch bereits jetzt nutzbar machen zu können, steht eine Übergangslösung zur Verfügung: In der Verzeichnungseinheit ist ein Weblink auf eine URL eingetragen, welche die Sichtung der Seite mittels Wayback-Server ermöglicht.

Aufwand in Zahlen

Den Harvest der 150 Seiten wickelt das System in einer Zeitspanne von ungefähr dreissig Stunden im Hintergrund ab. Zeitintensiv gestalten sich jedoch das Verzeichnen der gesicherten Webseiten und das Ermitteln der zugehörigen Metadaten. Hinzu kommen das Verlinken der WARC-Dateien, die daran anschliessende Durch-

3 Momentaufnahme einer Webseite, bzw. ein Screenshot einer kompletten Webseite.

führung des Ingests, das Setzen des Links auf den Wayback-Server sowie eine stichprobenmässige Qualitätskontrolle. Pro URL werden so grob geschätzt fünfzehn Minuten aufgewendet, für die gesamte jährliche Sicherung folglich rund vierzig bis fünfzig Stunden. Dazu kommen zusätzliche Aufwände für Problemfälle.

Aspekte zur Unternehmensarchitektur der Webarchivierung

Zusammenfassung

Die Archivierung von Webseiten im Staatsarchiv Basel-Stadt erstreckt sich über vier Bereiche. Zunächst geht es um das Auslesen der zu speichernden Webseite (Harvesting): Netzressourcen, die über eine URL erreichbar sind, werden in eine spezielle Archivdatei im Web Archive File Format WARC[4] geschrieben. Die gesicherten Inhalte liegen darin in serialisierter Form als XML-Struktur vor. Damit sie mit einem Browser betrachtet werden können, muss von der WARC-Datei ein Index erstellt werden (Indexing); dieser wird in eine CDX[5]-Datei geschrieben. Damit liegt die gesicherte Website vollständig vor. Jetzt erst wird sie archiviert. Dazu werden die WARC- und CDX-Dateien in ein OAIS-konformes AIP gepackt, verzeichnet und in ein Repository abgelegt (Verzeichnung/Ingest). Die gesicherte und archivierte Webseite kann jetzt benutzt werden (Benutzung).

Prozessdarstellung

Die im Staatsarchiv Basel-Stadt eingeführte und betriebene Archivierung von Webseiten lässt sich durch vier lose gekoppelte Prozesse darstellen. Die Bereiche Harvesting, Indexing, Ingest und Benutzung werden jeweils als separate Spur[6] abgebildet, innerhalb derer die zugehörigen Prozesse ablaufen. Diese Spuren sind als organisatorische Verantwortlichkeiten für die beinhalteten Prozesse zu verstehen.

4 Web ARChive file format, WARC: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.
5 CDX: https://archive.org/web/researcher/cdx_file_format.php.
6 In BPMN als swimlane dargestellt.

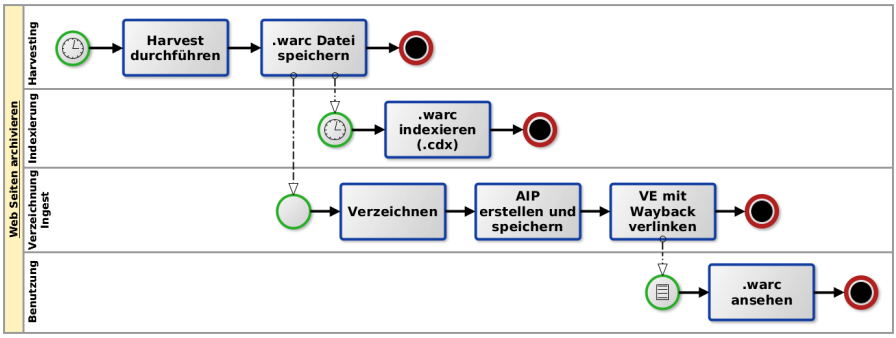

Abbildung 3: Prozessdarstellung

Die lose Koppelung zwischen den Prozessen legt zwar die Reihenfolge, nicht aber den exakten Zeitpunkt der Ausführungen fest. So wird beispielsweise der Ingest-/Verzeichnungsprozess je nach organisatorischer Ressourcenplanung irgendwann, auf jeden Fall aber immer erst nach dem zugehörigen Harvestprozess ausgeführt. Die Entkoppelung der Prozesse ermöglicht eine effiziente Organisation der verfügbaren Personalressourcen. Sie ist die Voraussetzung für die Skalierung hin zu einer breiten Serienproduktion.

Informationsarchitektur

Die beschriebenen Prozesse der Webarchivierung – dargestellt durch die abgerundeten Vierecke – stehen im Verhältnis zu verschiedenen Geschäftsobjekten (gelbe Vierecke) und Datenobjekten (blaue Vierecke). Während Geschäftsobjekte aus Sicht der Archivleitung eine strategische Bedeutung haben, sind die Datenobjekte für die technisch-operative Implementation wichtig.

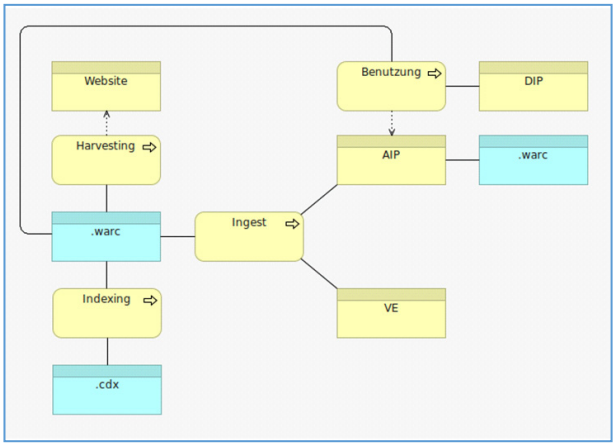

Abbildung 4: Informationsarchitektur

Die Objektdarstellung zeigt auf, dass eine als WARC-Datei gesicherte Netzressource zweimal vorgehalten wird, einmal durch den Benutzungsprozess und einmal durch den Ingestprozess. Diese Redundanz ist ein Kompromiss, der den Benutzungsansprüchen geschuldet ist: Die in einem AIP verpackte WARC-Datei muss im Falle einer Benutzung entpackt, indexiert und an einen geeigneten Ort kopiert werden. Diese Schritte können je nach Grösse und Komplexität der aufgerufenen Website sehr ressourcenintensiv sein und entsprechend lange dauern. Aus diesem Grund liegen bereits entpackte und indexierte Kopien der WARC- und CDX-Dateien für die Benutzung bereit. Die Verfügbarkeit dieser Arbeitskopien ist deutlich geringer gewichtet als jene der archivierten AIPs. Diese Benutzungskopien können jederzeit mit endlichem Aufwand aus den AIPs wieder hergestellt werden.

Anwendungsarchitektur

Um die charakterisierten Prozesse und Informationsobjekte zu implementieren, verwendet das Staatsarchiv Basel-Stadt folgende Anwendungskomponenten:

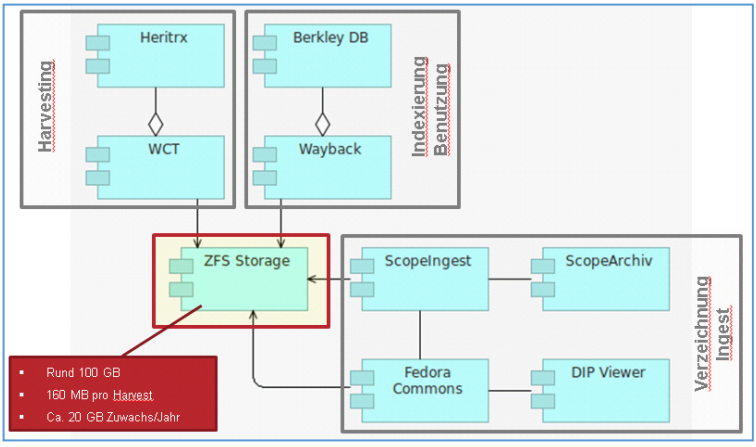

Abbildung 5: Anwendungsarchitektur

Der Harvesting-Prozess wird mit Heritrix[7] als Webcrawler und WCT[8] (Web Curator Tool) als Workflow Management Tool umgesetzt. Pro zu speichernde Webseiteninstanz wird ein Heritrix-Prozess ausgeführt. Die verschiedenen Heritrix-Prozesse werden mit Hilfe von WCT verwaltet und gesteuert. Für die Indexierung und Benutzung werden Berkley DB als Indexer und OpenWayback[9] als WARC-Viewer eingesetzt. Der WARC-Viewer ermöglicht es, eine gesicherte Webseite

7 Siehe: http://crawler.archive.org/index.html.
8 Siehe: http://webcurator.sourceforge.net/.
9 Siehe: https://github.com/iipc/openwayback/wiki.

unter Aufruf ihrer ursprünglichen URL in einem Webbrowser zu betrachten. Für die Verzeichnung und für den Ingest der zu archivierenden Webseite verwenden wir ScopeArchiv, ScopeIngest und Fedora Commons als Repository. Für die Speicherung der WARC- und CDX-Dateien sowie der AIPs wird eine redundante, ZFS[10]-basierte Speicherlösung verwendet.

Technologiearchitektur

Die Anwendungskomponenten werden vom Staatsarchiv Basel-Stadt in einer eigenen UNIX/Sparc-Umgebung betrieben. Gründe für die Wahl dieser Technologiearchitektur sind ihre grosse Ausfallsicherheit und ihre robuste Skalierbarkeit. Heritrix, WCT, OpenWayback, und Fedora Commons sind jeweils Java-Servlet[11]-Anwendungen. Sie werden in einem separaten Tomcat Servlet Container in einer separaten Solaris-Zone betrieben. Heritrix und WCT werden dabei aus Sicherheitsgründen in doppelter Ausführung vorgehalten: Einerseits für Webseiten ausserhalb der kantonalen Domäne – geschützt durch einen Forward Proxy – und andererseits für Webseiten innerhalb der bs.ch-Domäne. Die ZFS Storage Appliance ist ein dediziertes System. ScopeIngest muss innerhalb einer X86-Umgebung betrieben werden. Die Virtualisierung erfolgt dort über Oracle Virtual Box respektive Oracle VM (OVM).

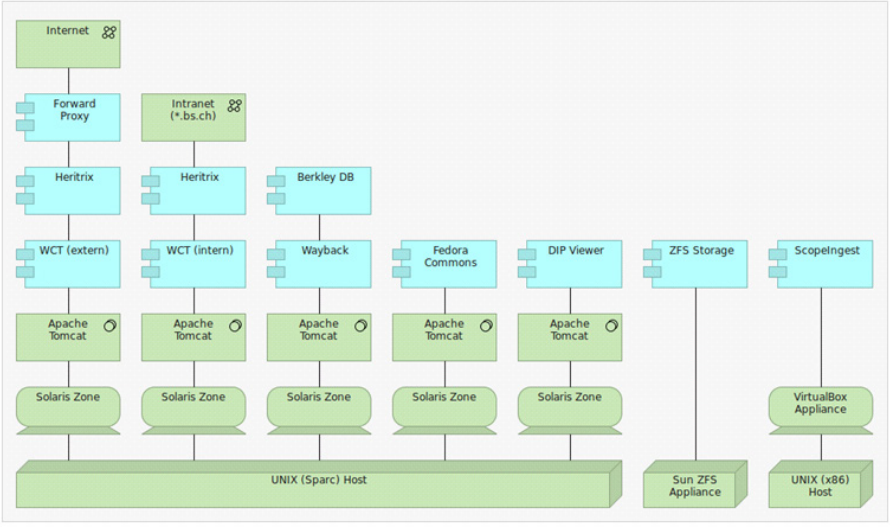

Abbildung 6: Technologiearchitektur

10 Siehe: https://de.wikipedia.org/wiki/ZFS_(Dateisystem).
11 Siehe Java Community Process JSR 315, JSR 340 resp. JSR 369.

Die abgebildete Technologiearchitektur wird tatsächlich parallel, in zwei unabhängigen und räumlich entfernten Standorten betrieben. Dadurch werden eine höhere Ausfallsicherheit[12] und die Trennung von Test und Produktion erreicht

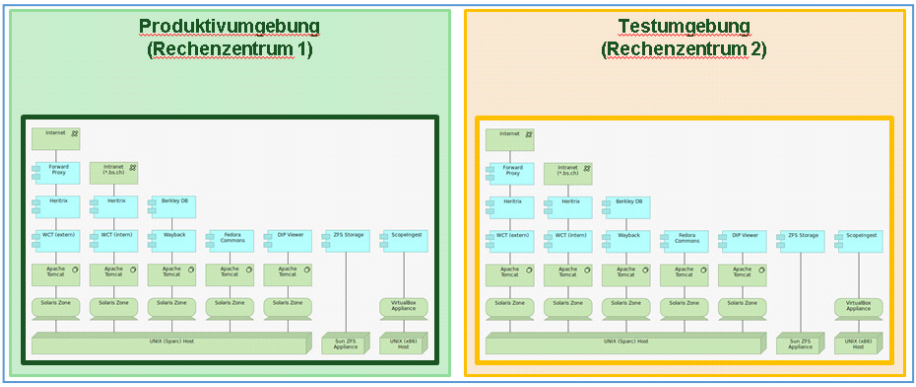

Abbildung 7: Produktiv- und Testumgebung

Fazit

Werden in einem Archiv bereits OAIS-konform digitale Inhalte archiviert, erfordert die Archivierung von Webseiten beziehungsweise Netzressourcen technologisch kein grosses zusätzliches Innovationspotential. Die bereits verwendeten Komponenten AIS und Ingestserver/Repository können gleichermassen eingesetzt werden. Lediglich das Harvesting, die Indexierung und die Benutzung erfordern zusätzliche Werkzeuge. Diese sind aber quelloffen verfügbar. Fachlich und organisatorisch erfordert die Archivierung von Netzressourcen dagegen erhebliche Planungs- und Koordinationsaufwände. Ihre Struktur, ihr Lebenszyklus und ihre Verfügbarkeit unterscheiden sich stark von dokumentartigem Archivgut.

12 Konkret: Reduktion der allfällig zu erwartenden Downtime und Data Loss Time.

E-Identität als Schlüssel zu den Dienstleistungen des digitalen Archivs

Zbyšek Stodůlka

In der heutigen Informationsgesellschaft müssen die Archive neuen Herausforderungen standhalten. Die Veröffentlichung der Findmittel und meistbenutzten digitalisierten Quellen ist angesichts der Menge der gespeicherten Archivalien nicht mehr als die Spitze des Eisberges. In der digitalen Welt, die heute mehr denn je gesellschaftliche Interaktion beinhaltet, stehen die Archive wegen der Bedingung des Präsenzstudiums der Archivalien in der Konkurrenz mit anderen Gedächtnisinstitutionen in einer sehr ungünstigen Stellung.

Die deutsche Bibliothekarin Ute Schwens fasste das Benutzerverhalten bei der Deutschen Digitalen Bibliothek prägnant als Trennung der Spreu von Weizen zusammen: Die Benutzer (besonders diejenigen der jüngeren Generation) ziehen die Daten mit digitalen Objekten bei ihrer Forschung vor.[1] Mit vielen Ausnahmen und Spezifika ist die Nutzung der Archivalien in europäischen Archiven in der Regel möglich, wenn sie älter als 30 Jahre sind, mit spezifischen Regelungen für Archivalien mit personenbezogenen Daten, Urheberrechtsbedingungen u.a. Es gibt also eine Menge rechtliche und technische Hindernisse, die den Fernzugriff bei der Benutzung ausschließen. Auf der anderen Seite ist es die Aufgabe der Archive, die sich verändernden Kommunikationstechniken für den Zugang zu den Archivalien proaktiv auszunutzen.[2]

Mit der weiteren Entwicklung des digitalen Archivs im Prager Nationalarchiv entwickelte sich 2016 die Debatte, wie wir die angenommene e-IDAS-Verordnung für die archivischen Zwecke anpassen können. In der tschechischen öffentlichen Verwaltung ist für die Implementierung dieser Verordnung das Ministerium des Inneren zuständig.[3]

Das Ziel dieser Verordnung ist die Überwindung verschiedener Hindernisse bei der elektronischen Kommunikation und dadurch die Unterstützung der digitalen

1 Ute Schwens: Die Bedeutung von Kulturportalen für Archive und Forscher, in: Irmgard Christa Becker, Gerald Maier, Karsten Uhde und Christina Wolf (Hrsg.), Netz werken. Das Archivportal-D und andere Portale als Chance für Archive und Nutzung. Beiträge zum 19. Archivwissenschaftlichen Kolloquium der Archivschule Marburg, Marburg 2015, S. 19-45, hier S. 29.

2 Grundsätze des Zugangs zu Archiven. Angenommen von der Jahresgeneralversammlung des ICA am 24. August 2012, S. 10.

3 Verordnung (EU) Nr. 910/2014 des Europäischen Parlaments und des Rates vom 23. Juli 2014 über elektronische Identifizierung und Vertrauensdienste für elektronische Transaktionen im Binnenmarkt und zur Aufhebung der Richtlinie 1999/93/EG.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.13

Wirtschaft. Artikel 1 dieser Verordnung betont: «Die wirtschaftliche und soziale Entwicklung setzt Vertrauen in das Online-Umfeld voraus. Mangelndes Vertrauen führt dazu, dass Verbraucher, Unternehmen und öffentliche Verwaltungen nur zögerlich elektronische Transaktionen durchführen oder neue Dienste einführen bzw. nutzen, vor allem, wenn sie die Befürchtung hegen, dass es an Rechtssicherheit mangelt.»

Diese Verordnung setzt die Entstehung neuer Dienstleistungen nicht nur im kommerziellen Sektor, sondern auch beim Angebot von der Seite der öffentlichen Verwaltung, voraus. Aus diesem Grund legt sie das Fundament für eine europäisch verbindende Infrastruktur mit garantierten Vertrauensdiensten, wie aus Artikel 9 sichtbar ist: «In der Regel können Bürger ihre elektronischen Identifizierungsmittel nicht verwenden, um sich in einem anderen Mitgliedstaat zu authentifizieren, weil die nationalen elektronischen Identifizierungssysteme ihres Landes in anderen Mitgliedstaaten nicht anerkannt werden. Aufgrund dieses elektronischen Hindernisses können Diensteanbieter die Vorteile des Binnenmarktes nicht vollständig ausschöpfen. Gegenseitig anerkannte elektronische Identifizierungsmittel werden die grenzüberschreitende Erbringung zahlreicher Dienstleistungen im Binnenmarkt erleichtern, und Unternehmen können grenzüberschreitend tätig werden, ohne beim Zusammenwirken mit öffentlichen Verwaltungen auf viele Hindernisse zu stoßen.» Neben der Zertifizierung verschiedener Vertrauensdienste ist dabei die Schaffung einer anerkannten elektronischen Identität durch eine Nationale Identitätsautorität (NIA) zentral.

Zur Benutzung einer Dienstleistung mit Authentisierung (z.B. Behörde, später Banken, Versicherungsanstalten, Netzanbieter usw.) stellt der Benutzer zunächst ein Gesuch um Authentisierung an die NIA, worauf ihm diese die Liste der Identitätsbetreiber gemäß dem Sicherheitsniveau der Dienstleistung übermittelt.[4] Nach der Authentisierung übergibt der Identitätsbetreiber der NIA das Ergebnis, und die NIA kann dem Dienstleistungsbetreiber die Identität bestimmen, auch mit verifizierten Angeben über den Benutzer. Für die grenzüberschreitende Authentifizierung wird das Sicherheitsniveau «substantiell» oder «hoch» verlangt.[5]

Strategisch erwartet man bei der öffentlichen Verwaltung im maximalen Maß die Digitalisierung der Dienstleistungen, die der Öffentlichkeit angeboten werden, weiter ihre wesentliche Vereinfachung, Beschleunigung, höhere Effizienz und einen Qualitätsanstieg.

4 Durchführungsverordnung (EU) 2015/1502 der Kommission vom 8. September 2015 zur Festlegung von Mindestanforderungen an technische Spezifikationen und Verfahren für Sicherheitsniveaus elektronischer Identifizierungsmittel gemäß Artikel 8 Absatz 3 der Verordnung (EU) Nr. 910/2014 des Europäischen Parlaments und des Rates über elektronische Identifizierung und Vertrauensdienste für elektronische Transaktionen im Binnenmarkt.

5 Verordnung (EU) Nr. 910/2014, Artikel 6 Abs. 1.

Die Lösung verlangt Anpassungen nicht nur auf der zentralen, sondern auch auf niedrigeren Verwaltungsebenen. Mitte 2017 werden in der Tschechischen Republik die Novellierung des Gesetzes über Personalausweise und die Annahme des Gesetzes über die elektronische Identifizierung erwartet. Der neue Personalausweis enthält zwingend ein elektronisches Zertifikat. Dieser dient erstens zum elektronischen Signieren und zweitens für den ganzen Bereich des Identitätsmanagements. Die Authentifizierung mit dem neuen Personalausweis in Verbindung mit der NIA bürgt für das höchste Sicherheitsniveau, und die Dienstleistung bekommt aktuelle Angaben über den Benutzer, die sie z.B. zur Hilfe beim Ausfüllen der Formulare ausnutzen kann.

Seit 2016 benötigt jedes IT-Projekt der öffentlichen Verwaltung, dessen Kosten für den Erwerb mehr als 200.000 € oder für den fünfjährigen Betrieb mehr als 1.100.000 € betragen, schon im Stadium der Architekturplanung eine Genehmigung des eGovernment-Hauptarchitekts im Innenmisterium. Der Schwerpunkt liegt dabei auf dem Einsatz von e-Government-Dienstleistungen; gefordert wird in der Regel eine gesicherte Verbindung zu Registern für Personen (physische und juristische) einerseits und zu Registern der Rechte und Aufgaben anderseits (mit einheitlichem Identitätsraum der öffentliche Verwaltung, sog. JIP/KAAS).

Eines der zentralen Themen des digitalen Archivs ist die Identifizierung und Authentifizierung des Benutzers, der oft in mehreren Rollen (z.B. Vertreter der Provenienzstelle und Nutzer) auftreten kann. Schon produktive Lösungen kann man zum Beispiel in Estland oder beim elektronischen Archiv der Slowakei finden. Beim Zutritt zu letzterem verwendet der Benutzer seine eID nicht nur zur Unterscheidung, in welcher Rolle er beim Portal auftritt (Bürger oder Beamte/Archivar mit Mandat-Zertifikat), sondern sie beschränkt auch die Tätigkeit nach der spezifischen Institution und Arbeitsaufgabe (z.B. Bewertung, Nutzen usw.).[6]

6 Používateľská príručka elektronických služieb: Elektronický archiv Slovenska (Benutzerleitfaden der elektronischen Dienstleistungen: Elektronisches Archiv der Slowakei), on-line: https://portal.minv.sk/wps/wcm/connect/sk/site/main/dokumenty-tlaciva/archiv-slovenska.

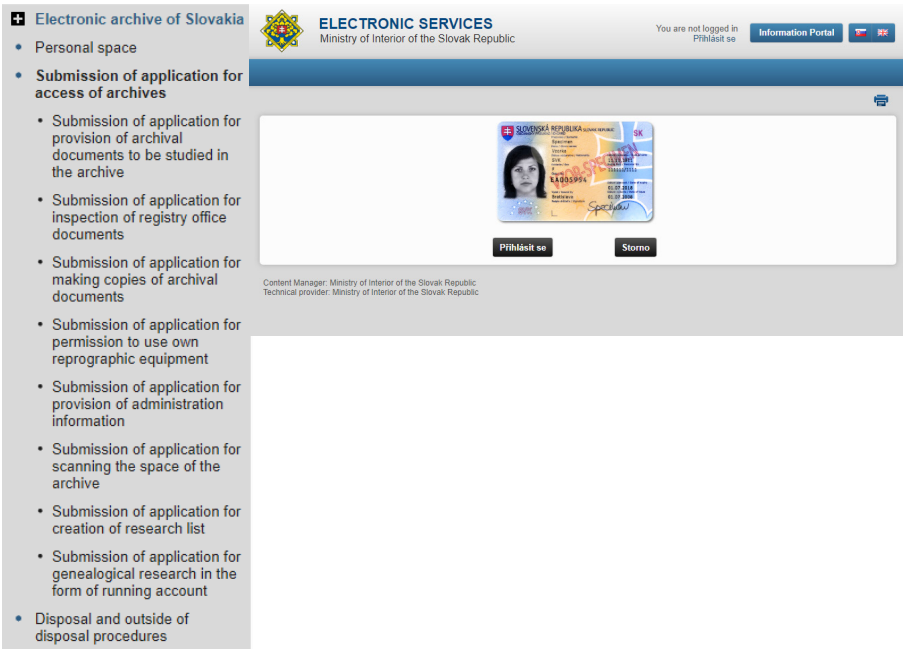

Abbildung 1: Screenshot Elektronisches Archiv der Slowakei

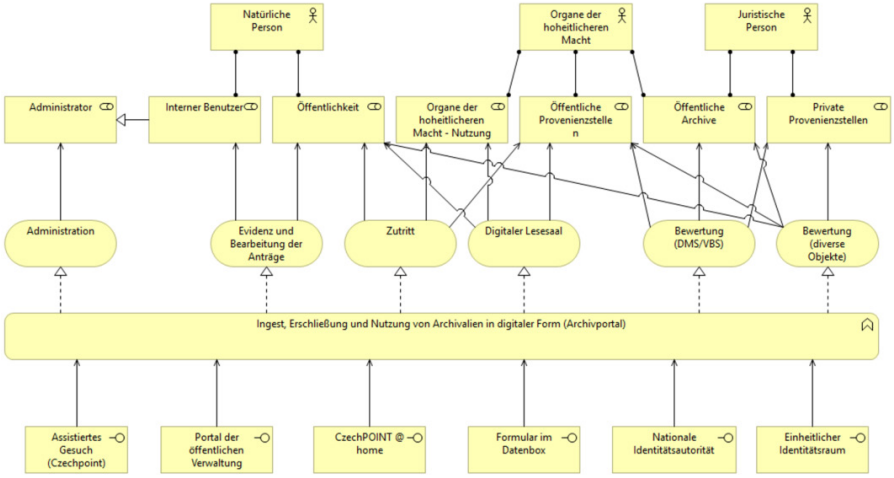

Abbildung 2: Infrastruktur des Einheitlichen Identitätsraums

Beim Antrag für die Refinanzierung des Projektes des digitalen Archivs bis zum Jahr 2020 mussten wir Anforderungen abstimmen und diese Identitätsmanagementinfrastruktur anschließen. Leitmotive sind dabei in erster Linie, Prozesse zu automatisieren, wiederholte Eingabe der Angaben zu verhindern und unsere auch per

zentrales elektronisches Portal (wo der Bürger sich einloggen muss) angebotenen intelligenten Formulare, z.B. Forschungsbogen, auszunutzen. Die Modularität unserer Architektur ermöglicht eine relativ einfache Modifikation der einzelnen Anforderungen.

Unser Digitales Archiv bietet seine Dienstleistungen über das Archivportal an. Zur Authentifizierung von Mitarbeitenden der öffentlichen Verwaltung möchten wir die Infrastruktur des Einheitlichen Identitätsraums, des sogenannten JIP/KAAS ausnutzen, weil dort die Beamten und ihre Tätigkeiten durch die gesetzlichen Rollen definiert sind.

Für den Nutzer planen wir drei Möglichkeiten:

— Identifikation im entsprechenden Archiv oder in einem anderen Archiv. Wenn die Identität überprüft ist, kann der Nutzer die Archivalien in digitaler Form nutzen, auch durch Fernzugriff mit Angaben zum Login zum Archivportal.
— Möglichkeit, ein Gesuch bei einer öffentlichen Kontaktstelle der Verwaltung zu stellen (CzechPoint, z.B. beim Postamt). Nach Ausfüllung des Formulars (z.B. Forschungsbogen) bekommt man Angaben zum Login.
— Ein EU Bürger kann auch seine perssönliche elektronische Identität nutzen. Das Archivportal bekommt dabei die Angaben durch die Anmeldung des Bürgers am sogenannten Portal des Bürgers.

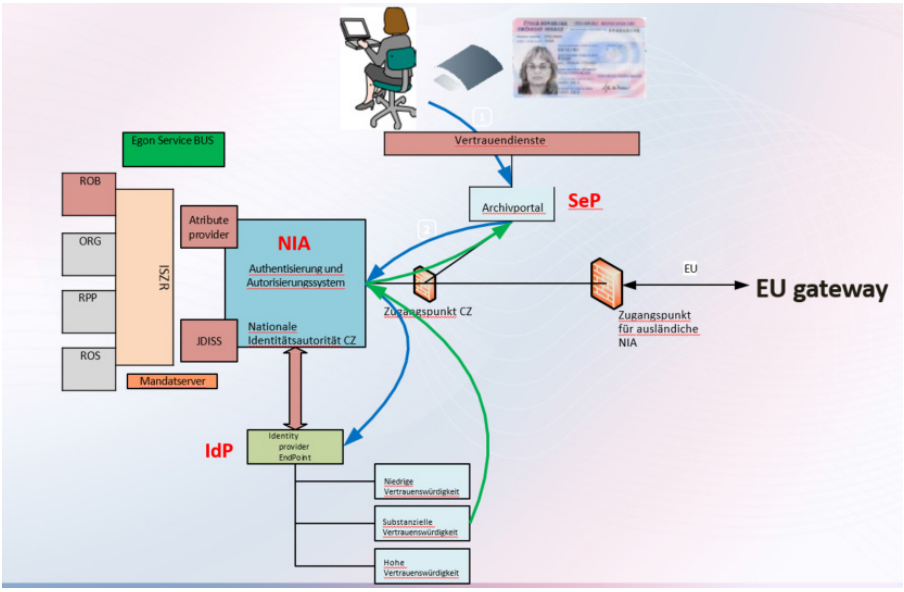

Abbildung 3: Identifikationsmöglichkeiten

Die Datenverknüpfung bringt hoffentlich geringere Kosten, bessere Garantie für die Richtigkeit der Informationen und ein höheres Maß an Sicherheit. Positiv ist ebenfalls, dass sich mit der Weiterentwicklung des Identitätsmanagement auch die Dienstleistungen von DMS/VBS in der öffentlichen Verwaltung weiter verbreiten, z.B. bei der Benutzung durch die Beamten vor Ort in der Behörde. Für diejenigen, die sich nicht in diesen verschiedenen Identitätsräumen befinden (z.B. nicht EU-Bürger), ist weiterhin die traditionelle Lösung möglich, der Besuch im Archiv.

In rechtlicher Hinsicht ermöglicht seit 2012 das Archivgesetz Nr. 499/2004 in § 34 Abs. 6 die Nutzung mit Fernzugriff: Die Benutzung der Archivalien in digitaler Form erfolgt mittels des Nationalportals oder der Portale der digitalen Archive. Entgegen der Befürchtung, dass die Archive beispielsweise bei missbräuchlicher Benutzung von Archivalien, die unter Datenschutz stehen, verantwortlich gemacht würden, kam den Archiven das Verfassungsgericht zu Hilfe. Im Urteil von 20. Dezember 2016 (ÚS 3/14) zur Klage gegen das Archiv der Sicherheitskräfte, die Akte eines Mitarbeiters der tschechoslowakischen Staatssicherheit den Journalisten des Tschechischen Fernsehens vorgelegt zu haben, betonte das Gericht, dass die Verantwortung für Weitergabe und Veröffentlichung bei dem Forscher liegt, der verpflichtet ist, die Genehmigung des Rechteinhabers zu besorgen.[7] Im Falle personenbezogener Daten muss das Archiv selbstverständlich die Zustimmung der Betroffenen zur Nutzung einholen.

Dieser Artikel fasste nur kurz die bisherige Erfahrung des tschechischen Nationalarchivs mit der E-Identität in der archivischen Praxis zusammen. Das Thema hat langfristig das Potenzial, den Fernzugriff zu erleichtern, auch auf Archivalien mit verschiedenen Benutzungseinschränkungen, die nicht veröffentlicht werden können, und zur Debatte über die Informationssysteme der digitalen Archive (besonders der digitalen Lesesäle), über die Kostenfrage, über die Beurkundung in elektronischer und analoger Form und über die Zielgruppen beizutragen.

7 Česká republika, Nález Ústavního soudu Pl. ÚS 3/14 (Tschechische Republik, Urteil des Verfassungsgerichtes Pl-ÚS 3/14), online: https://www.usoud.cz/fileadmin/user_upload/Tiskova_mluvci/Publikovane_nalezy/2017/Pl._US_3_14_vcetne_disentu.pdf. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)