Schweizerisches Sozialarchiv

Christian Koller
(Hrsg.)

AKTEN DER 27. TAGUNG DES ARBEITSKREISES ARCHIVIERUNG VON UNTERLAGEN AUS DIGITALEN SYSTEMEN

(5./6. März 2024 in Zürich)

Impressum:

Herausgeber: Christian Koller
2024, Schweizerisches Sozialarchiv, Zürich
www.sozialarchiv.ch

ISBN 978-3-033-10930-8

Inhalt

VORWORT

Christian Koller

Die Archivierung digitaler Dokumente ist in der Archivwelt seit geraumer Zeit ein zentrales Thema. Bereits seit 1997 führt der in Deutschland, Österreich, der Tschechischen Republik, der Schweiz, Ungarn, Belgien, Luxemburg, Frankreich und Dänemark aktive Arbeitskreis „Archivierung von Unterlagen aus digitalen Systemen“ (AUdS) jährlich eine Fachtagung durch. Die Konferenz 2024 fand am 5. und 6. März in Zürich im Kirchgemeindehaus Paulus statt.[1] Organisatoren waren eine Gruppe Zürcher Archive sowie die schweizerische Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST). Dem engeren Organisationskomitee gehörten Martin Akeret (Archiv der Universität Zürich), Marlis Betschart (Stadtarchiv Winterthur), Christian Koller (Schweizerisches Sozialarchiv), Isabelle Mehte (KOST), Sander Ouwendijk (Stadtarchiv Zürich), Bernhard Stüssi (Staatsarchiv des Kantons Zürich) und Sonja Vogelsang (Archiv für Zeitgeschichte an der ETH Zürich) an.

Die Bedeutung des Themas widerspiegelte sich in der sehr großen Zahl der Teilnehmenden. Neben den vor Ort anwesenden 120 Archivarinnen und Archivare machten über 300 Personen von der Möglichkeit der remote-Teilnahme Gebrauch. In verschiedenen Panels und parallelen Barcamps diskutierte die Tagung insgesamt 29 Präsentationen von Vertreterinnen und Vertretern staatlicher, privater und kirchlicher Archive sowie Archivdienstleistungsunternehmen, Bibliotheken und Museen, die die ganze Breite der Thematik ausloteten.[2] Angesprochen wurden dabei technische, archivfachliche, konzeptuelle, organisatorische und juristische Belange und Projekte. Einmal mehr zeigte sich in beeindruckender Weise, dass Archive entgegen ihrem Image als Staubfängerinstitutionen an der Spitze des technischen Fortschritts marschieren und den digitalen Wandel, mit dem sie durch die Übernahme, Erhaltung und Zugänglichmachung von „digital born“ Akten sich digitalisierender Organisationen und Behörden sowie die Digitalisierung der eigenen Betriebsabläufe, Metadaten, Benutzungsdienstleistungen und Bestandeserhaltungsmaßnahmen gleich in zweifacher Hinsicht befasst sind, wesentlich mitprägen. Damit

[1] Tagungsberichte: Bollen, Timo/Neffgen, Ines (2024), „27. Tagung des Arbeitskreises Archivierung von Unterlagen aus digitalen Systemen (AUdS)“, in: *Archiv Theorie und Praxis* 77/3, S. 262-264; Koller, Christian (2024), „Digitale Archivierung: Die AUdS-Tagung in Zürich“, in: *SozialarchivInfo* 2, S. 6-8, https://www.zora.uzh.ch/id/eprint/260883; Friederich, Christine (2004), „Elektronische Archivierung: Das war die AUdS-Konferenz 2024“, in: *Sax Archiv Blog: Neues aus dem Sächsischen Staatsarchiv*, 2.4.2024, https://saxarchiv.hypotheses.org/30291.

[2] Die Präsentationen sind auf der Webseite des Staatsarchivs des Kantons St. Gallen abrufbar: https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2024.html

tragen sie im digitalen Zeitalter zur Stärkung demokratischer Legitimität, rechtsstaatlicher Mechanismen und zivilgesellschaftlicher Aktivitäten bei.

I.

GRUSSBOTSCHAFTEN:
OHNE ARCHIVE KEIN DEMOKRATISCHER RECHTSSTAAT

Demokratien sind keine Selbstläuferinnen

Jacqueline Fehr

Liebe Archivarinnen und Archivare
Liebe Digitalisierungsfachleute
Geschätzte Organisatorinnen und Organisatoren der Tagung
Sehr geehrte Damen und Herren

Ich freue mich, Sie hier in Zürich zu begrüßen und hoffe, dass Sie alle eine gute Anreise hatten. Bereits zum 27. Mal tagt der Arbeitskreis „Archivierung von Unterlagen aus digitalen Systemen“. Sie können also eine stolze Bilanz vorweisen. Und Sie beschäftigen sich mit einem Thema, das gut zum Kanton Zürich passt. Denn eines kann ich Ihnen sagen: digitale Systeme haben wir hier viele, ja, sehr viele. Der digitale Werkzeugkasten, den sich öffentliche Verwaltungen in der Schweiz vor vielleicht 40 oder 50 Jahren zugelegt haben, ist seither ständig gewachsen. Immer wieder wurden neue Werkzeuge beschafft. Nicht selten passten sie zu den bereits bestehenden. Aber nicht immer.
Gleichzeitig wurden alte Werkzeuge abgelöst. Oft gelangen so Verbesserungen. Aber nicht immer. Und manchmal wurde der Werkzeugkasten selbst ausgetauscht. Manche alten Werkzeuge fanden auch dort wieder Platz, andere passten nicht mehr hinein, noch andere passten nicht mehr zueinander. Sie sehen: Unser digitaler Werkzeugkasten ist heute zwar insgesamt leistungsfähig, aber auch groß – und vor allem bunt.
Die Regierung ist sich schon länger bewusst, dass die Vielfalt mit der Zeit ein wenig zu vielfältig wurde. Deshalb geben wir inzwischen Gegensteuer. Gerade jetzt löst der Kanton Zürich seine alte Hardware- und Software-Umgebung komplett ab und wechselt auf mobile Geräte, auf die jüngste Windows-Umgebung und auf Microsoft 365. Damit erzielen wir viele Verbesserungen. Der Wechsel wirft aber auch Fragen auf, vor allem zur Datensicherheit und zu den Abhängigkeiten, in die wir uns mit dem Einsatz eines so umfassenden Systems begeben. Diese Fragen müssen wir formulieren, die Unsicherheiten und Abhängigkeiten erkennen, die damit verbundenen Risiken abwägen und sie wo möglich minimieren. Dieser Prozess ist bei uns zurzeit im Gang, und er ist anspruchsvoll. Aber ich verstehe ihn auch als Chance.
Es wäre naiv, zu glauben, dass staatliche Organisationen in ihrer ganzen Größe und Komplexität heute funktionieren können, ohne von zahlreichen Systemen, Anbietern und anderen Stakeholdern abhängig zu sein. Mehr noch: Die Welt ist vernetzter denn je und digitaler denn je, Tendenz

zunehmend. Es geht deshalb nicht darum, Abhängigkeiten grundsätzlich zu meiden oder zu leugnen. Sondern es geht darum, die Welt in ihrer Komplexität zu akzeptieren und zu versuchen, darin eine Rolle zu spielen. Sie zu verstehen und sie mitzugestalten, und zwar auf eine verantwortungsvolle Weise. Das gilt auch in Bezug auf das digitale Equipment öffentlicher Organe und in Bezug auf die digitalen Daten, die wir bewirtschaften. Lassen Sie mich einen kurzen Blick darauf werfen.

Was unser Basis-Equipment betrifft, scheint mir die Sache relativ einfach. Unsere Abhängigkeit ist da ebenso groß wie diejenige von Unternehmen und Privaten: Wir haben kaum Einfluss auf die Chip-Herstellung, ebenso wenig wie auf die Produktion der restlichen Hardware, die wir nutzen. Und auch die Gestaltung der führenden Betriebssysteme und Office-Anwendungen richtet sich nicht oder kaum nach unseren Vorgaben. Unser Einfluss ist allenfalls indirekt: Unser Konsum- und Nutzungsverhalten wird registriert und beeinflusst die Produktestrategien von morgen. Es hilft nicht, unsere kleine Rolle in diesem Markt zu beklagen. Aber es wichtig, sich hier der eigenen Kleinheit bewusst zu sein.

Grösser ist unser Einfluss bei den zahlreichen Fachanwendungen, die öffentliche Verwaltungen einsetzen. Funktionalität und Leistungsfähigkeit der verfügbaren Systeme hängen weitgehend von unserer Nachfrage ab. Auch die Weiterentwicklung können wir beeinflussen. In vielen Fällen gibt es gemischte Gremien, Teams von Anbietern und Kunden, die gemeinsam planen, wie die Fachanwendung von morgen aussieht. Das ist gut und sinnvoll, vorausgesetzt, dass wir über die nötige Bestellerkompetenz verfügen.

Gleichzeitig ist die Marktrealität in der Schweiz auch schwierig: Oft wird Fachsoftware von kleinen Firmen entwickelt und angeboten. Nicht selten kommen solche Lösungen in die Jahre, sie veralten technisch und sind mit der Zeit nicht mehr genügend sicher. Und sie wirken irgendwann am Bildschirm einfach erschreckend veraltet. Da können wir besser werden, können unsere Marktmacht besser nutzen. Das ist zwar verbunden mit Arbeit und mit der Übernahme von Verantwortung, aber wir stehen hier in der Pflicht. Denn oft sind Fachapplikationen die Gefäße, in denen wir sensible Daten über Menschen pflegen.

Damit komme ich zum Kernpunkt dessen, was ich Ihnen zur Begrüßung hier in Zürich mitgeben will. Und damit wird es nicht nur persönlich, sondern auch ernst.

Meine Damen und Herren, keine private Firma ist auch nur annähernd für so viele schützenswerte Personendaten verantwortlich wie die öffentlichen Organe. Polizei- und Strafverfolgungsbehörden, Justizvollzug, Gesundheitsinstitutionen, Zivilstands- und Grundbuchämter, Steuerbehörden, Gerichte – sie alle nutzen Fachapplikationen für ihre Arbeit, und sie alle pfle-

gen sensible und sehr sensible Daten – und zwar über sämtliche Menschen, die in unseren Ländern leben.

Wir sind verantwortlich dafür, dass diese Daten nur zu Zwecken erfasst werden, die das Gesetz vorsieht. Wir sind verantwortlich dafür, dass sie richtig sind. Und wir sind verantwortlich dafür, dass sie nur so lange in produktiven Systemen gehalten werden wie nötig. Dann müssen sie daraus verschwinden. Und spätestens hier kommen die Archive, kommen Sie ins Spiel: Ihr Auftrag ist es, die Tätigkeit des Staats, für den Sie arbeiten, anhand von Originaldaten zu überliefern. Das heißt, Sie müssen sämtliche Daten, die ihre staatlichen Organe produzieren, so strikte wie möglich bewerten und so deren Kerngehalt freilegen. Dann müssen Sie die Daten in Ihre eigenen Systeme übernehmen, und zwar so, dass sie auch morgen und übermorgen noch lesbar und verständlich sind. Sie müssen also nicht nur den Entstehungskontext mit überliefern, sondern sich auch noch für das richtige Datenformat entscheiden und die nötigen Metadaten mitnehmen in die Zukunft. Und damit es Ihnen nicht langweilig wird: Alle Daten, die Sie in ihre Hoheit übernehmen, müssen sie so lange schützen wie nötig, sie dann aber so schnell wie möglich öffentlich zugänglich machen. So erfüllen Sie Ihren Auftrag.

Ich weiß: Das ist eine anspruchsvolle Aufgabe. Und ich weiß auch, dass diese Aufgabe erfüllbar ist. Vor allem dann, wenn die Archive gut miteinander zusammenarbeiten, so wie Sie. Und was ich bei all dem auch weiß. Sie haben eine anspruchsvolle, wichtige, spannende, sinnstiftende und damit auch dankbare Aufgabe. Seien Sie stolz auf das, was Sie tun.

Sie tragen mit Ihrer Arbeit dazu bei, dass demokratische Rechtsstaaten stabil bleiben. Dass sich mündige Bürgerinnen und Bürger jederzeit ein eigenes Bild historischer Begebenheiten machen können, und zwar beruhend auf primären Fakten, auf Originaldaten. Sie halten diese stabil und verfügbar. Damit bieten Sie den Menschen Gelegenheit, sich aus ihrer Perspektive heraus immer wieder neu mit der Vergangenheit auseinanderzusetzen. Denjenigen Fragen nachzugehen, die aktuell wichtig sind. Abzuklären, wie eine bestimmte Entwicklung verlief. Festzustellen, wie eine Behörde seinerzeit entschieden hat. Rechte einzufordern, manchmal auch erst nach Jahrzehnten. Kritisch und gleichzeitig verantwortungsvoll umzugehen mit der Vergangenheit, um heute die Zukunft kritisch und verantwortungsvoll mitzugestalten.

Meine Damen und Herren, Demokratien sind keine Selbstläuferinnen, das erfahren wir in jüngster Zeit wieder viel deutlicher als auch schon. Demokratien sind ausgerichtet auf faire Verteilung von Macht und von Gütern, auf Ausgleich, auf Partizipation, auf Mitsprache, auf Widerspruch. Demokratien sind Systeme, die darauf angelegt sind, dass einzelne Bäume nicht zulasten ihrer Umgebung in den Himmel wachsen, diese in den Schatten stellen und auch dann noch riesigen Schaden anrichten, wenn sie stürzen.

Denjenigen Menschen, die den Hals nicht vollkriegen, die nicht genug Macht kriegen können, nicht genug Geld, wirken Demokratien also direkt entgegen. Entsprechend stark geraten demokratische Systeme ins Visier von Despoten, Narzissten und Autokraten. Demokratien müssen sich gegen solche Angriffe aktiv wehren. Wenn sie das nicht tun, wenn sie sich nicht gegen totalitäre Ansprüche stellen, sondern selbstgefällig zurücklehnen, schwächen sie sich und gefährden sich letztlich selbst.

Die öffentlichen Archive sind ein wichtiges Element im demokratischen Abwehrdispositiv. Sie stellen der Allgemeinheit ein Fundament zur Verfügung, auf dem sachliche Auseinandersetzungen über verschiedenste Themen möglich sind – und zwar immer wieder und immer wieder neu. Faktenbasiert und kritisch. Dieser Auftrag, den Sie seit der Französischen Revolution haben, hat sich in den letzten 200 Jahren nicht geändert. Und er ändert sich auch mit der Digitalisierung der Gesellschaft, der öffentlichen Verwaltungen und der Archive nicht. Im Gegenteil. Er wird noch anspruchsvoller und noch wichtiger. Denn Sie wissen es selbst am besten: Es ist anspruchsvoller, Originaldaten aufzubewahren als Originalakten.

Dafür, dass Sie sich bemühen, Ihren Auftrag wahrzunehmen und zu erfüllen, dafür, dass Sie sich zum Austausch hier in Zürich treffen und sich gegenseitig bei der Erfüllung ihrer Aufgabe unterstützen, danke ich Ihnen herzlich. Mit Staatsarchivar Beat Gnädinger diskutiere ich oft über die Fragen, die ich soeben angetippt habe. Von ihm habe ich viel über genau das gelernt, was ich ihnen soeben als Nachdenk-Angebot mitgegeben habe.

Ich freue mich, ihm nun das Wort zu geben und wünsche Ihnen eine ertragreiche Tagung.

Vielen Dank für Ihre Aufmerksamkeit.

Archivieren heißt Demokratie ernst nehmen

Beat Gnädinger

Sehr geehrte Frau Regierungsrätin (oder verständlicher für die ausländischen Gäste: Sehr geehrte Frau Innenministerin)
Liebe Organisatorinnen und Organisatoren der Tagung
Liebe Kolleginnen und Kollegen
Sehr geehrte Damen und Herren

Ich freue mich meinerseits, Sie herzlich hier in Zürich zu begrüßen, sowohl persönlich als auch im Namen der KOST und der sechs Zürcher Archive, die diese Tagung zusammen organisiert haben.
Sie haben es gehört: Meine Chefin meint es ernst. Sie meint es ernst mit der Demokratie, mit den Archiven – und mit uns Archivfachleuten. Darüber bin ich froh. Denn es gibt noch immer viele politisch Verantwortliche, die die Rolle öffentlicher Archive im demokratischen Rechtsstaat nicht wirklich verstanden haben oder sie zu wenig ernst nehmen. Meist meinen es diese Leute nicht bös. Vielmehr haben sie Archiven gegenüber einfach eine mehr oder weniger gleichgültige Haltung. Sie halten die Archive nicht für wichtig genug, denn diese befassen sich ja ausschließlich mit der Vergangenheit – währenddem der eigene Blick selbstverständlich konsequent und entschlossen in die einzig relevante Richtung gerichtet ist, nämlich in die Zukunft. Ich beklage mich nicht über diese, nun ja, nur bedingt reflektierte Haltung. Denn in unserem beruflichen Alltag leiden wir ja meist nicht direkt darunter, dass wir uns mehrheitlich im Windschatten der großen Politik bewegen. Und wir halten auch aus, dass Archivarinnen und Archivare in den Status-Rankings der Berufe nicht nur nicht so weit oben liegen wie die Ärzte oder die Astronautinnen – sondern dass sie in diesen Rankings nicht einmal vorkommen. Aber wir dürfen uns von unserem eigenen Großmut nicht davon abhalten lassen, unsere Aufgabe ernst zu nehmen, mit ihr Wirkung zu erzielen, sie möglichst vielen Menschen nahe zu bringen, bei Bedarf entschieden dafür einzustehen – und uns dauernd aktiv zu bemühen, unsere Aufgabe zeitgemäß zu erfüllen. Ich verstehe unter diesen eher programmatischen Ansprüchen Verschiedenes und versuche, Ihnen das an drei Beispielen zu erläutern.
Beispiel 1 bezieht sich auf den Zugang zu den Akten über die eigene Person. Heute kommen täglich Leute zu uns, die auf der Suche sind nach Akten über fürsorgerische Zwangsmaßnahmen, über eine Adoption, über eine Fremdplatzierung, über einen medizinischen Eingriff, über

eine Strafmaßnahme – also eben: nach Akten über die eigene Person. Bis vor wenigen Jahren war das nicht so. Was ist seither geschehen? Vor vielleicht 10, 15 Jahren setzte sich die Erkenntnis durch, dass die Schweiz zahllose Menschen aus verschiedenen gesellschaftlichen Randgruppen während langen Jahrzehnten grob und achtlos, ja, in vielerlei Hinsicht unrechtmäßig behandelt hat.

Der Staat, private Institutionen und die Gesellschaft schufen Mitte des 19. Jahrhunderts ein System, das es erlaubte, der Allgemeinheit nicht genehme Lebensbahnen sozusagen mit einer Bewegung aus dem Handgelenk für immer zu ändern. Nur zu oft führten diese Eingriffe in Sackgassen, ins Elend oder sogar in den Tod. Erst im letzten Drittel des 20. Jahrhunderts geriet das System stark unter Druck und musste geändert werden. Und erst vor wenigen Jahren schafften es die Betroffenen, die Öffentlichkeit in genügendem Maß auf die Missstände aufmerksam zu machen, symbolische Entschädigungen zu erwirken, wissenschaftliche Untersuchungen anzustoßen – und ihr Recht auf Akteneinsicht durchzusetzen.

Dieses Recht bestand *de iure* zwar schon länger. Aber bei dessen Wahrnehmung stießen viele Betroffene in der Praxis lange auf Widerstände, auch in den zuständigen Kanzleien und Archiven. Oft hörten sie: „Wir haben nichts mehr." „Aus Datenschutzgründen können wir Ihnen die Unterlagen nicht geben." „Die Beantwortung Ihrer Anfrage ist sehr aufwändig. Leisten Sie einen Kostenvorschuss." Oder ganz banal: „Ihre Ansprüche sind verjährt."

Erst als sich die Staatsarchive vor rund 10 Jahren entschieden, ihre alte, passive Haltung aufzugeben und den Betroffenen ihr Fachwissen aktiv zur Verfügung zu stellen, setzten Veränderungen ein: Durch Merkblätter erhielten Gemeinden mehr Klarheit über ihre Rechte und Pflichten, ebenso wie Betroffene. Die Archive selbst sammelten Erfahrungen bei der Sicherung auch von kleinen Datenspuren in ihren Unterlagen, bei der kantonsübergreifenden Aktensuche, bei der Zusammenarbeit mit Opferhilfen, beim Umgang mit Betroffenen. Mussten wir uns bis dahin oft sagen lassen „Archive geben nicht alles heraus, was sie haben", „Archive unterstützen die Behörden, nicht die Kundschaft", hat sich das Bild inzwischen gewandelt. Es hat sich gewandelt, weil wir uns ernsthaft gefragt haben, worin unser Job im Kern besteht – und weil wir anschließend entsprechend gehandelt haben. Dadurch werden die Schweizer Archive heute anders wahrgenommen als gestern. Sie werden anders wahrgenommen, weil sie Verantwortung übernommen haben.

In meinem zweiten Beispiel geht es um Nacherschließung, Digitalisierung und Beständeerhaltung. Archive sind für Unterlagen und Informationen aus vielen Jahrhunderten verantwortlich. Einige Dokumente in unseren Häusern haben mehr als ein Jahrtausend auf dem Buckel, andere sind erst wenige Jahre alt, der Rest liegt irgendwo dazwischen. Das macht unsere Aufgabe viel-

fältig und spannend. Und seit nicht nur audiovisuelle Dokumente zu Pergament- und Papierbeständen hinzukommen, sondern auch digitale Daten in schnell wachsenden Mengen, ist unser Job noch spannender geworden. Aber damit wird nicht nur unsere Zeitachse immer länger, sondern es wird auch die Breite unserer Aufgaben immer grösser.

Wir wissen, wie wir Pergament erhalten. Wir wissen, wie wir Tintenfraß stoppen. Wir bewahren industriell gefertigtes Papier aus dem frühen 20. Jahrhundert vor dem Zerfall. Wir schauen, dass VHS-Videobänder auch morgen noch lesbar sind. Und inzwischen befassen wir uns mit einer Vielzahl von Dateiformaten, die wir nicht nur lesbar halten, sondern so pflegen, dass die originalen Informationen, die sie repräsentieren, erhalten bleiben. Gleichzeitig steigen die Ansprüche der Öffentlichkeit an die Verfügbarkeit unserer Daten und an unsere Reaktionszeiten. Archive, die ihre Verzeichnungsdaten nicht auf dem Netz haben, rutschen immer mehr in eine Wahrnehmungslücke. Archive, die ihre zentralen Serien nicht als digital lesbare Texte aufs Netz stellen und die Digitalisate auf Nachfrage hin nicht innert nützlicher Frist zur Verfügung stellen können, geraten unter Druck.

Und damit wir es nicht zu einfach haben, kommt dazu: Die Verzeichnung unserer Bestände hat eine Halbwertszeit. Es genügt nicht, Findmittel aus dem 18., 19. oder 20. Jahrhundert einfach zu digitalisieren und dabei zu hoffen, dass sie den Ansprüchen weiterhin genügen. Wenn wir das, wozu wir verpflichtet sind, auch wirklich wollen – nämlich, dass alle unsere Bestände möglichst gleich gut genutzt werden können, müssen wir auch die Findmittel dazu auf dem Stand der Zeit halten. Das heißt: Neben all unseren alten und neuen Beständen haben wir auch unsere Findmittel aktuell zu halten, sonst geben wir sie einer schleichenden Entwertung preis.

Das Staatsarchiv Zürich hat deshalb die Erhaltung der Bestände sowie die Nacherschließung und Digitalisierung zu Hauptprozessen „befördert“ und in den letzten 15 Jahren zwei eigene entsprechende Abteilungen aufgebaut. Sie beanspruchen inzwischen zusammen gegen 30 Prozent unserer personellen Ressourcen. Aktuell scheint uns, dass die Proportionen ungefähr stimmen.

Damit komme ich zu meinem dritten und letzten Beispiel: Digitales Know-how, verteilt auf alle archivischen Hauptprozesse. Schon länger bauen wir in allen Abteilungen sukzessive digitale Kompetenzen auf. Viele Mitarbeitende des Staatsarchivs haben in ihrem Curriculum inzwischen einen Informatik-Hintergrund in der einen oder anderen Form. Eine wichtige Verstärkung in diesem Bereich sind für uns auch die Studierenden und die Zivildienstleistenden. Aber wir sind noch nicht da, wo wir hinkommen müssen in den nächsten Jahren. Wir brauchen noch mehr digitale Kompetenzen und personelle Ressourcen, und zwar in allen Prozessen, für die wir verantwortlich sind.

Es ist völlig klar, warum: Die Überlieferungsbildung beschäftigt sich immer mehr mit digitalen Ablieferungen. Die Erschließung muss diese verarbeiten und für die spätere Nutzung bereitstellen. Die Kundendienste brauchen eine Plattform, auf der wir unsere Informationen in allen Ausprägungen öffentlich verfügbar machen können. Denn Benutzung ist heute zunehmend konnotiert mit Begriffen wie Open Government Data oder Semantische Suche. Und, wie bereits gesagt, Nacherschließung und Beständeerhaltung sind ohnehin Prozesse, die ohne digitales Know-how nicht auskommen.

Aber gleichzeitig sind alle unsere analogen Kenntnisse und Fähigkeiten nach wie vor gefragt. Wir alle sind weiterhin – und wohl noch für manche Jahre – mit analogen Aktenangeboten und mit unerschlossenen Beständen konfrontiert. Und ohnehin, so hoffe ich zumindest, denkt kein Archiv daran, analoge Bestände zu digitalisieren und die analogen Bestände daraufhin zu vernichten. Das heißt: Was wir heute können, müssen wir auch morgen noch können. Aber zusätzlich haben wir alles abzudecken, was im digitalen Zeitalter auf uns zukommt, und zwar auf der Angebots- und auf der Nachfrageseite – und natürlich dazwischen, also im Archiv selbst.

Ich sage deshalb, was ich schon zu verschiedenen Gelegenheiten gesagt habe, gern auch hier noch einmal: Unser Beruf war noch nie so vielfältig und so anspruchsvoll wie heute. Und wir hatten auch noch nie so gute Werkzeuge und Voraussetzungen wie heute, um unseren Auftrag zu erfüllen. Die Digitalisierung erweitert unseren Job um vielerlei Aspekte und Dimensionen, um viele Aufgaben – und um viele Möglichkeiten. Aber unser Grundauftrag bleibt unverändert, und er bleibt, wenn Sie an die Worte von Frau Regierungsrätin Fehr zurückdenken, ausgesprochen wichtig.

In diesem Sinn danke ich Ihnen meinerseits dafür, dass Sie sich heute und morgen austauschen über den Stand der Arbeiten in Ihrem Spezialgebiet, dass Sie über ihren Zaun hinausschauen, dass Sie gemeinsam mit anderen weiterkommen wollen, dass wir als Archive zusammen zur Qualität und Widerstandsfähigkeit des demokratischen Rechtsstaats beitragen.

Vielen Dank.

II.

EIN JAHRZEHNT DIGITALES ARCHIV: RÜCKBLICK UND AUSBLICK

Das Nationale Digitale Archiv der Tschechischen Republik nach 10 Jahren: Was die neue Generation anbieten kann

Zbyšek Stodůlka

Die ersten Schritte auf dem Weg zum digitalen Archiv

Die Anfänge der Überlegungen zur Nutzung von EDV-Technik für archivische Tätigkeiten im damaligen Staatlichen Zentralarchiv, dem Vorläufer des heutigen Nationalarchivs, wurden erstmals Ende der 1960er-Jahre konzipiert. In den 1970er-Jahren wurde unter der Schirmherrschaft der Tschechoslowakischen Wissenschaftlich-Technischen Gesellschaft eine Gruppe von Experten gegründet, die sich mit der Einführung neuer Technologien im Archivwesen beschäftigte. Ziel war der Austausch von Erfahrungen und die Übernahme bewährter Verfahren aus dem Ausland. Zu den Hauptaktivitäten dieser Gruppe zählten: 1. die Erstellung kontrollierter Vokabulare (Thesauri) zur Beschreibung der Inhalte von Archivalien und Archivbeständen; 2. die Entwicklung von datenbankgestützten Findmitteln, insbesondere fondsübergreifender Art, die sekundär aus bestehenden Hilfsmitteln zusammengesetzt wurden; 3. die Anfertigung neuer Hilfsmittel, die eine Umwandlung in ein digitales Format erlaubten; 4. die praktische Auseinandersetzung mit der Nutzung von Computern bei Exkursionen zu Rechenzentren; und 5. die Archivierung von mit Computern erzeugten Aufzeichnungen. Prägende Persönlichkeiten und maßgebliche Treiber dieser Entwicklungen waren unter anderem PhDr. Jaroslav Honc, PhDr. Václav Babička und PhDr. Tomáš Kalina.

Trotz einzelner Teilergebnisse – wie der Erstellung von Indexen mit der KWOC-Methode (Keyword Out of Context) und der Analyse zur Entwicklung eines Archiv-Informationssystems (z. B. ARCHAIS) – war ein substanzieller Fortschritt erst nach der Anschaffung eigener EDV-Technik in der zweiten Hälfte der 1980er-Jahre möglich. Es wurde ein Referat für Automatisierung geschaffen, die später als eigenständige Abteilung geführt wurde. Inspiriert durch das französische System PRIAM entstanden verschiedene Datenbanken zu Aktivitäten der Provenienzstellen sowie thematische Hilfsmittel und computergestützte analytische Ausgaben. Das erste Programm, das dem universellen Beschreiben von Archivalien diente, war das 1991 von Mitarbeitern entwickelte Programm SPIS, basierend auf der von der UNESCO bereitgestellten Datenbankanwendung CDS/ISIS (Kalina, 2004, S. 202-214).

Entwicklung des Nationalen Digitalarchivs

Dieses Arbeitsfeld war somit sowohl theoretisch als auch praktisch dazu befähigt, ein umfassendes Projekt zur Einrichtung einer zentralisierten Einheit für staatliche Archive zu entwickeln, das sich mit der langfristigen Aufbewahrung und Zugänglichmachung von Dokumenten in digitaler Form befasst. Dieses Vorhaben wurde durch den Beschluss der Regierung der Tschechischen Republik Nr. 11 vom 7. Januar 2004 bestätigt. Im Rahmen der Erarbeitung eines technologischen Projekts, das 2008 abgeschlossen wurde, wurden sowohl zentralisierte (nationale Einrichtung) als auch dezentralisierte Lösungen (Einrichtungen bei staatlichen Regionalarchiven) untersucht. Aufgrund des Bedarfs an qualifizierten Kapazitäten und der erforderlichen Infrastruktur wurde entschieden, die zentrale Einrichtung beim Nationalarchiv in Prag zu errichten, welche unter der Bezeichnung Nationales Digitalarchiv (NDA) firmiert.

Die Konzeption der Digitalisierung der Verwaltungsaktivitäten unter Nutzung elektronischer Systeme für das Dokumentenmanagement gemäß dem umgesetzten MoReq2-Standard (in Form des Nationalen Standards für elektronische DMS/VBS) wurde in die Novellierung des Archivgesetzes und seiner Durchführungsverordnungen im Jahr 2009 eingeschlossen.

Dies bedeutete einen erheblichen qualitativen Fortschritt und setzte klare Anforderungen sowohl an die abgebenden Stellen als auch an digitale Archivlösungen, um die Bewertung, Übernahme, Speicherung und Bereitstellung dieses digital geborenen Archivguts sicherzustellen. In der Praxis nutzten zahlreiche Provenienzstellen bereits damals verschiedene Informationssysteme zur Schriftgutverwaltung, der Übergang zur elektronischen Schriftgutverwaltung erfolgte jedoch weitaus langsamer, als ursprünglich erwartet.

Die langfristige Finanzierung des Nationalen Digitalarchivs wurde 2013 durch Regierungsbeschluss Nr. 611 gesichert. In den Jahren 2011 bis 2013 wurde ein digitales LZA-System konzipiert. Angesichts der hohen Anforderungen öffentlicher Ausschreibungen entschied sich das Nationalarchiv Ende 2013, ein eigenes System zu entwickeln, das auf den spezifischen Bedürfnissen der Institution basierte.

Als Kernlösung wurde das kanadische Open-Source-System Archivematica ausgewählt. Dieses System wurde von der kanadischen Firma Artefactual Systems Inc. seit 2009 entwickelt, initiiert durch das UNESCO-Programm Memory of the World. Heute wird das System in zahlreichen Archiven und Bibliotheken, insbesondere in den USA und Kanada, eingesetzt, und seine Popularität wächst auch in Europa und Asien. Zu den Vorteilen gehören die Verfügbarkeit, Anpassungsfähigkeit, aktive Benutzercommunity, definierte AIPs, Standardkonformität (METS, PREMIS, BagIt, usw.), definierte Prozesse (Microservices) und die Möglichkeit zur individuellen Anpassung der Migrationsstrategien (Format Policy Rules).

Das LZA-System wurde so weit wie möglich für die automatisierte Verarbeitung von SIPs angepasst. Hierzu wurde eine vorgegebene Verzeichnisstruktur implementiert, die einzelnen Schritte des Übernahmeprozesses sowie genutzte Werkzeuge zur Formatidentifikation, Metadatenauslese und Migrationsprozesse definiert. Darüber hinaus wurden eine automatisierte Antwortschnittstelle für die Angaben (Größe, Formate, AIP-IDs, usw.) integriert und ein eigener Steuermechanismus namens DISTRIBUTOR zur Verwaltung und Kontrolle der Speicherung eingeführt.

Das Nationalarchiv entwickelte mit internen Ressourcen eine Anwendungsschnittstelle zur Bewertung und Übernahme von Unterlagen aus DMS/VBS (ERMS), die den Export von SIP-Paketen (XML-Schema NSESSS im METS-Format) unterstützt. Im Mai 2014 erfolgte die erste Bewertung von Archivalien und im April 2015 wurde das erste Archivgut mittels dieser neuen Infrastruktur erfolgreich digital archiviert (Stodůlka 2016a).

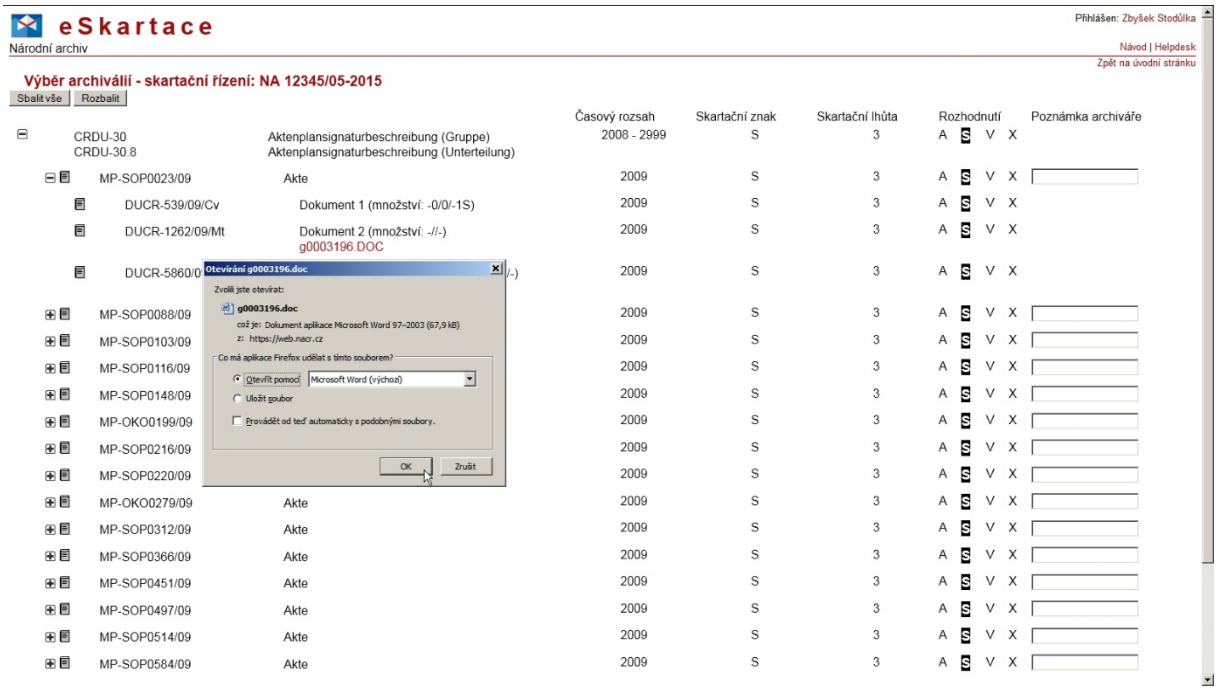


Abbildung 1: Umgebung zur Bewertung von Unterlagen aus DMS/VBS (eAkte) aus dem Jahr 2015

Das Nationale Archivportal: Schnittstelle der digitalen Dienstleistungen für Archive, Provenienzstellen und Forschende

Parallel zu ersten Erfahrungen mit Betrieb des digitalen Archivs liefen seit 2015, insbesondere jedoch im Jahr 2016, Arbeiten an einer Portallösung, die bestehende Anwendungen integriert und eine einheitliche Benutzeroberfläche für die Kommunikation zwischen Archiven, Aktenbildnern und Forschern bietet: das Nationale Archivportal. Die Anwendung erweitert die Möglichkeiten zur Bewertung und Übernahme von Archivalien auf unstrukturierte Daten. Es

wurden Module zur Bewertung und Sortierung unstrukturierter Dateien (später auch für Datenbanken) entwickelt sowie Werkzeuge zur Verwaltung und Präsentation von Archivbeständen. Ebenso entstand ein Benutzerverwaltungssystem mit lokalen Administratoren in den Gebietsarchiven und zentralen Administratoren im Nationalen Digitalarchiv (NDA) (Stodůlka 2016). Zur Koordination der Aktivitäten wurde auf der Open-Source-Plattform Redmine das Umfeld MoPED (Methodische Plattform für elektronische Dokumente) eingerichtet, das sich schrittweise auch für weitere interessierte Archivare öffnete und heute eine Kommunikationsplattform für weitere Akteure in der Branche darstellt, wie z. B. die Arbeitsgruppe für Hochschularchive. Mit dem neuen Projekt NDA II war es möglich, die bestehenden Module des Portals grundlegend zu modernisieren, insbesondere vor dem Hintergrund, dass die Anwendung für die Bewertung und Übernahme aus DMS/VBS grundlegend überarbeitet werden musste. Dies war notwendig aufgrund der neuen Version des Nationalstandards für elektronische Systeme der Schriftgutverwaltung, der schließlich im Jahr 2017 veröffentlicht wurde und das bestehende Datenpaket SIP unter anderem um ein Audit-Log erweiterte. In der Praxis zeigte sich auch die Notwendigkeit eines Offline-Werkzeugs. Seitdem wurde eine Möglichkeit geschaffen, eine Excel-Tabelle zu erstellen, in der Entscheidungen markiert und in das Portal zurückimportiert werden können. Die Möglichkeiten zur Anreicherung von Metadaten wurden grundlegend erweitert, einschließlich der Neuzuordnung von Sachgruppen anstelle der zukünftigen archivischen Bearbeitung, durch die Nutzung einer einheitlichen Erfassung von Metadaten aus verschiedenen Datenpaketentypen im Format EAD 3.

Zugleich war klar, dass fortschrittlichere Methoden des Benutzerzugriffs entwickelt werden mussten, die sich aus der Einführung der elektronischen Identität sowie aus den wachsenden Anforderungen im Bereich der Cybersicherheit ergaben (Stodůlka 2018).

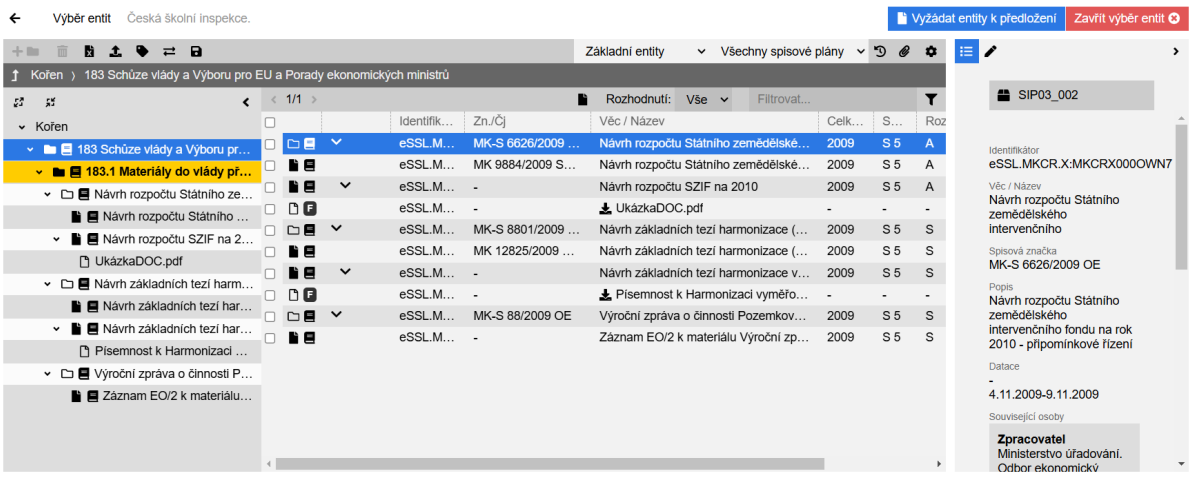
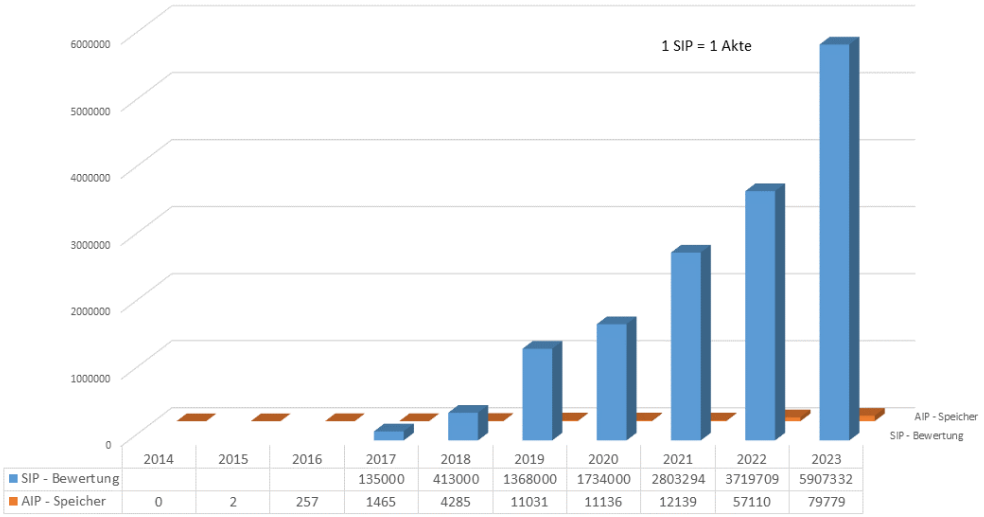


Abbildung 2: Umgebung von NArP 2.0 zur Bewertung von Unterlagen aus DMS/VBS (eAkte)

Aufgrund des schrittweisen Übergangs zu externen Lieferanten für bestimmte Softwareentwicklungsbereiche wurde die Zahl der Programmierer reduziert und das Team schrittweise um zwei Spezialisten für digitale Archive erweitert. Um die Qualität der digitalen Archivalien sicherzustellen, wurde ab 2018 der SIP-Validator als Webanwendung eingeführt, die bei der Validierung und Einreichung von SIPs im Portal hilft. SIPs, die als nicht valid erachtet werden, gelangen nicht in die Bewertung und verbleiben bis zur nächsten Bewertung im DMS/VBS der abgebenden Stelle. Derzeit sind 147 Prüfpunkte definiert, darunter die Struktur von METS, Aktenplan, Aufbewahrungsfristen und Formate u. a. (Validátor SIP, 2024).



	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
SIP - Bewertung				135000	413000	1368000	1734000	2803294	3719709	5907332
AIP - Speicher	0	2	257	1465	4285	11031	11136	12139	57110	79779

Abbildung 3: Statistik der bearbeiteten SIPs zur Bewertung und der gespeicherten AIPs

Eine wichtige Rolle spielt die seit 2016 organisierte Fortbildung, die als Kombination aus Präsenz- und Fernunterricht über das LMS Moodle angeboten wird. Sie gliedert sich in folgende Module: Modul A: NArP 2.0 – Bewertung und Übernahme; Modul B: NArP 2.0 – Erschließung und Zugang zum Archivgut; Modul C: NArP 2.0 – Verwaltung; Modul D: Nachweis des nationalen Archivguts in IS PEvA II; Modul E: Zentrales Archivmodul für Normdateien – CAM. Ab 2016 wurden zunächst die Koordinatoren mit Fokus auf Behörde- und Archivberatung in Gebietsarchiven ausgebildet, die wiederum andere Archivare schulten. Diese Koordinatorenstellen wurden neu geschaffen, um gezielt die Qualität der Beratung und die Zusammenarbeit zwischen Archiven und NDA zu verbessern. Während der COVID-19-Pandemie wurde die Schulung zentral online durchgeführt, unterstützt durch Moodle-Materialien, Zoom-Demonstrationen, YouTube-Videos und praktische Aufgaben, die die Teilnehmer:innen online abschließen mussten.

Insgesamt wurden auf diese Weise 220 Archivar:innen zertifiziert, und das Konzept erhielt positives Feedback.

Es ist hervorzuheben, dass die Beratung zur korrekten Erstellung von SIPs für IT-Lieferanten und Provenienzstellen erhebliche personelle Ressourcen der Mitarbeitenden des digitalen Archivs in Anspruch nahm. Ebenso beanspruchte deren Mitwirkung an verschiedenen Gesetzesnovellierungen zur Digitalisierung der Schriftgutverwaltung erhebliche Kapazitäten.

In der Praxis wurde deutlich, dass abgebende Stellen, insbesondere aus der zentralen Staatsverwaltung, nicht in der Lage waren, die technologischen Fortschritte nachzuvollziehen und die gestellten Anforderungen zu erfüllen. Vor dem Hintergrund anhaltender Defizite in der Funktionalität von elektronischer Schriftgutverwaltung leitete die Regierung im Jahr 2019 eine umfassende Überprüfung ein. Diese Evaluation erstreckte sich auf die DMS/VBS von 72 zentralen staatlichen Behörden. Die Überprüfung ergab systemische Defizite und wies auf zentralen Verbesserungsbedarf hin. Die Provenienzstellen wurden verpflichtet, einen Maßnahmenplan zu erstellen, um die festgestellten Defizite zu beheben. Diese verstärkte Aufmerksamkeit auf Leitungsebene führte zu positiven Entwicklungen, etwa der Ablösung bestehender Anbieter durch leistungsfähigere (zur Umwandlung der Tätigkeiten bei der Aussonderung und Bewertung vgl. Kunt/Lechner/Pokorný, 2020).

Die COVID-19-Pandemie stellte einen weiteren Einflussfaktor dar. Während die Behörden verstärkt versuchten, ihre Arbeit in den digitalen Raum zu verlagern, nahm die Zahl der Institutionen, die Unterlagen zur Bewertung und Übernahme mittels NArP angeboten haben, signifikant zu.

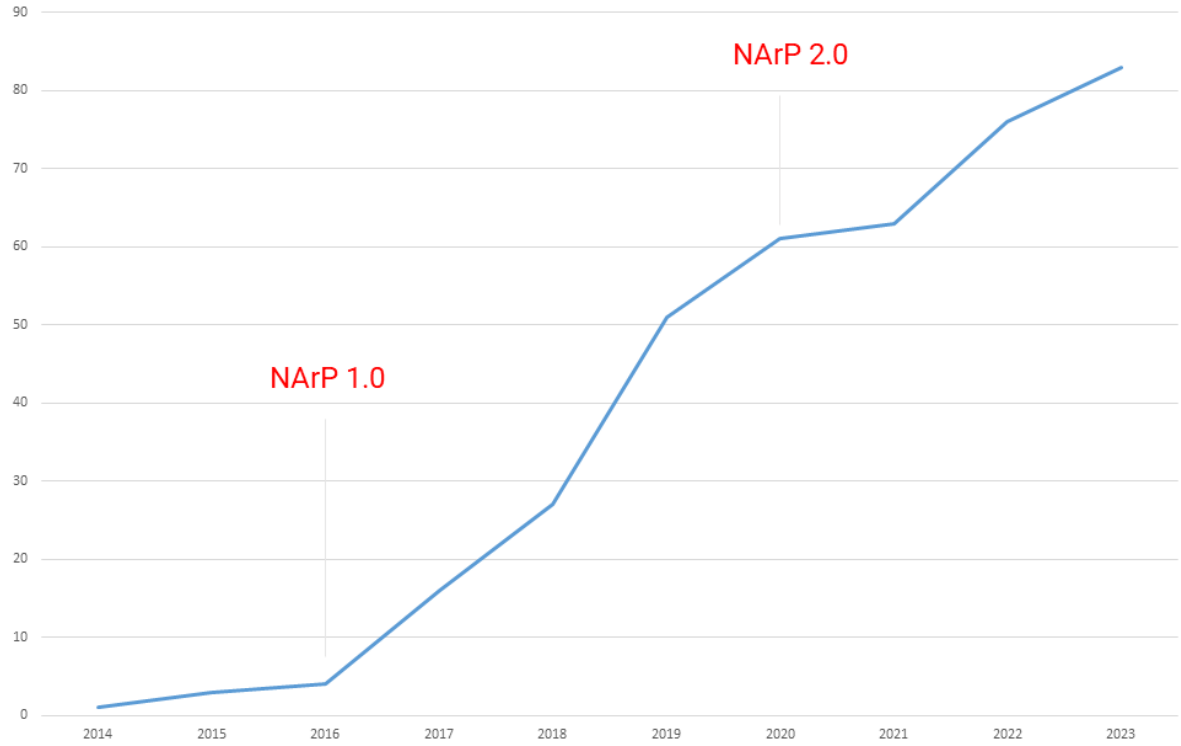


Abbildung 4: Anzahl der Archive, die die Dienste des NArP nutzen, nach Jahren

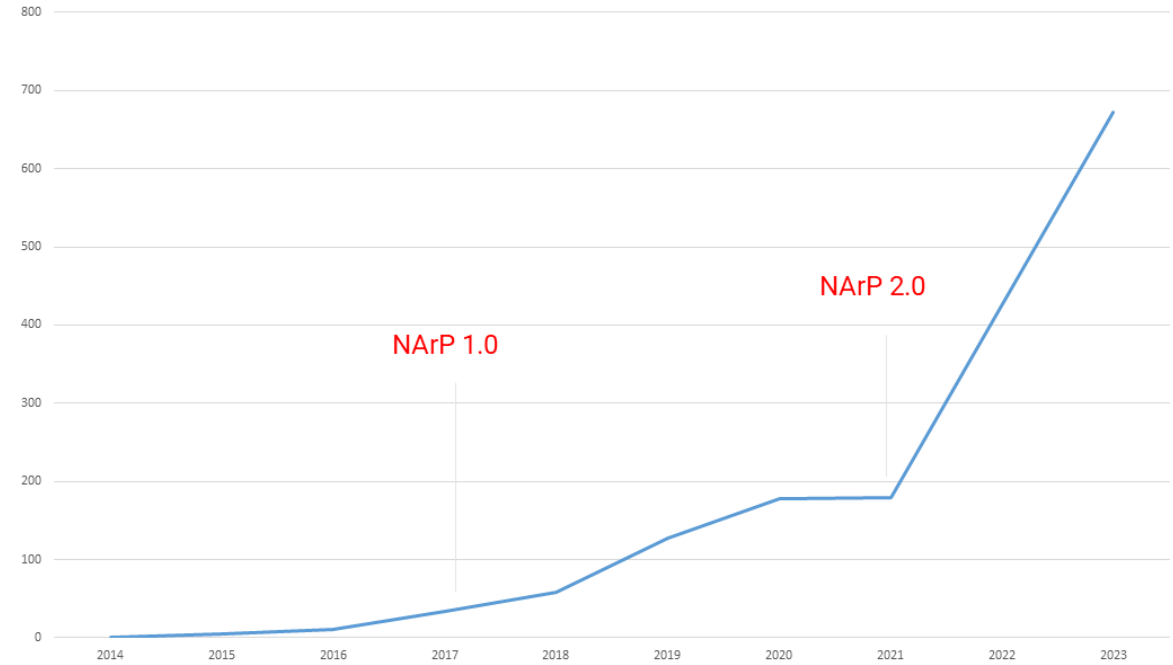


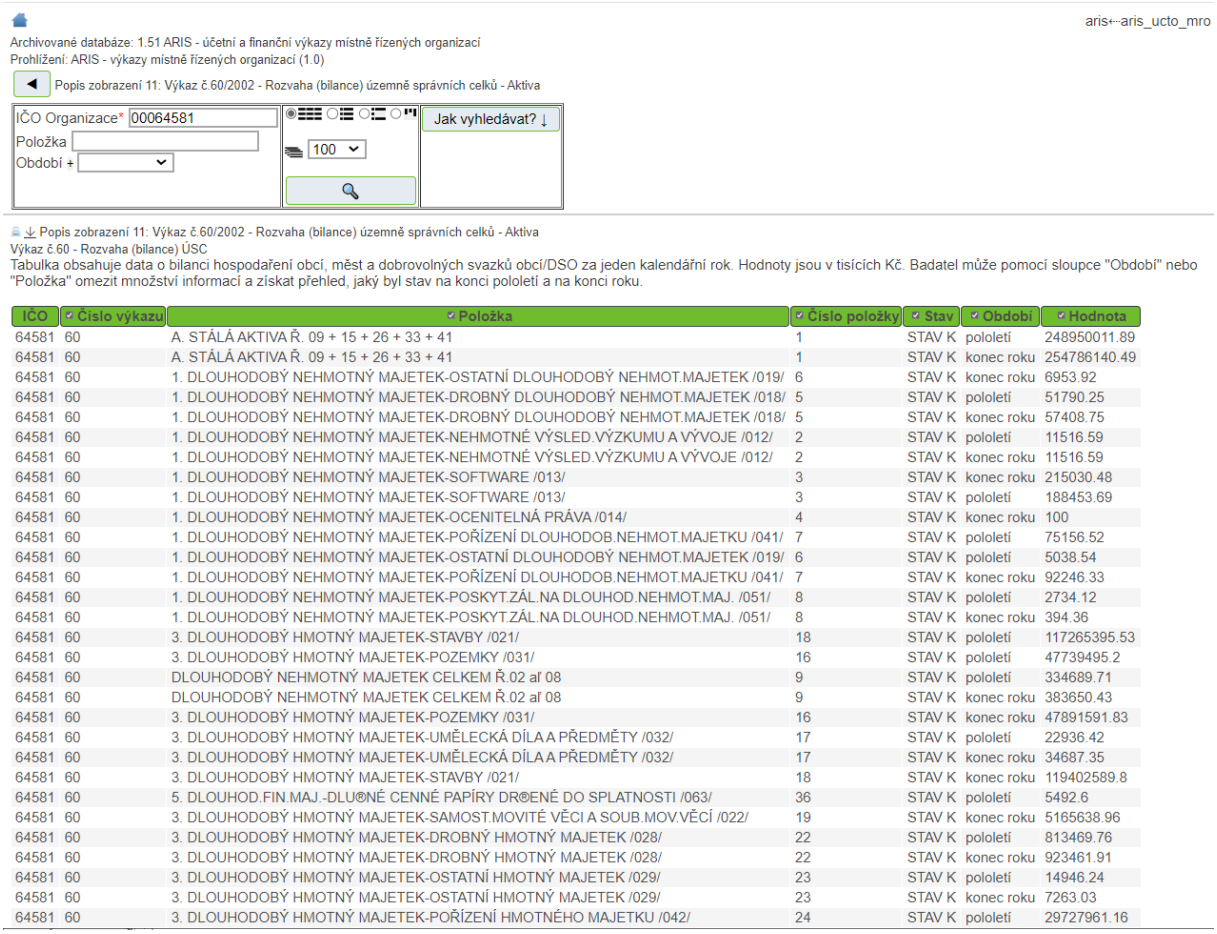
Abbildung 5: Anzahl der Provenienzstellen, die die Dienste des NArP nutzen, nach Jahren

Internationale Inspiration und Erfahrungsaustausch

Im Bereich der digitalen Archivierung treten die Mitarbeitenden nicht nur auf zahlreichen nationalen Kongressen und Konferenzen in den Bereichen Archivwesen, Schriftgutverwaltung, e-Government-Entwicklung und Digitalisierung der öffentlichen Verwaltung auf. Äußerst wertvoll für die Entwicklung des digitalen Archivs war die jährliche Teilnahme an den Sitzungen

der AUdS, bei denen der Entwicklungsstand verglichen und Inspiration für die nächsten Schritte gewonnen werden konnten. Für den Erfahrungsaustausch in spezifischen Bereichen der Entwicklung digitaler Archive wurde seit 2015 eine Reihe von Workshops unter dem Titel „Archives in Digital Age“ (AiDA) organisiert, an denen Kollegen aus Deutschland, Polen, der Slowakei und Ungarn teilnahmen (Naumann/Stodůlka/Vojáček 2016). Im Jahr 2019 wurde das Nationalarchiv Mitglied des DLM-Forums und engagierte sich in mehreren Initiativen, darunter die Standardisierung von Informationspaketen (E-ARK/eArchiving/DILCIS-Board) und die Arbeitsgruppe EAG „Archiving by Design“, die sich mit der nachhaltigen Archivierung von Informationen bereits in der Entstehungsphase beschäftigt.

Im Zuge der Übernahme von Datenbanken, die bisher nur über ad-hoc Exporte in mehr oder weniger geeigneten Formaten zugänglich waren (CSV, XML, u. a.), wurden seit 2017 positive Ergebnisse mit dem SIARD-Format und der Implementierung des slowenischen Werkzeugs dbDIPview erzielt (Rechtorik 2022). Es wird angestrebt, die Prinzipien des „Archiving by Design“ bereits bei der Systemgestaltung oder beim Systemumbau zu verankern.



IČO	Číslo výkazu	Položka	Číslo položky	Stav	Období	Hodnota
64581	60	A. STÁLÁ AKTIVA Ř. 09 + 15 + 26 + 33 + 41	1	STAV K	pololetí	248950011.89
64581	60	A. STÁLÁ AKTIVA Ř. 09 + 15 + 26 + 33 + 41	1	STAV K	konec roku	254786140.49
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-OSTATNÍ DLOUHODOBÝ NEHMOT.MAJETEK /019/	6	STAV K	konec roku	6953.92
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-DROBNÝ DLOUHODOBÝ NEHMOT.MAJETEK /018/	5	STAV K	pololetí	51790.25
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-DROBNÝ DLOUHODOBÝ NEHMOT.MAJETEK /018/	5	STAV K	konec roku	57408.75
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-NEHMOTNÉ VÝSLED.VÝZKUMU A VÝVOJE /012/	2	STAV K	pololetí	11516.59
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-NEHMOTNÉ VÝSLED.VÝZKUMU A VÝVOJE /012/	2	STAV K	konec roku	11516.59
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-SOFTWARE /013/	3	STAV K	konec roku	215030.48
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-SOFTWARE /013/	3	STAV K	pololetí	188453.69
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-OCENITELNÁ PRÁVA /014/	4	STAV K	konec roku	100
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-POŘÍZENÍ DLOUHODOB.NEHMOT.MAJETKU /041/	7	STAV K	pololetí	75156.52
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-OSTATNÍ DLOUHODOBÝ NEHMOT.MAJETEK /019/	6	STAV K	pololetí	5038.54
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-POŘÍZENÍ DLOUHODOB.NEHMOT.MAJETKU /041/	7	STAV K	konec roku	92246.33
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-POSKYT.ZÁL.NA DLOUHOD.NEHMOT.MAJ. /051/	8	STAV K	pololetí	2734.12
64581	60	1. DLOUHODOBÝ NEHMOTNÝ MAJETEK-POSKYT.ZÁL.NA DLOUHOD.NEHMOT.MAJ. /051/	8	STAV K	konec roku	394.36
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-STAVBY /021/	18	STAV K	pololetí	117265395.53
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-POZEMKY /031/	16	STAV K	pololetí	47739495.2
64581	60	DLOUHODOBÝ NEHMOTNÝ MAJETEK CELKEM Ř.02 ať 08	9	STAV K	pololetí	334689.71
64581	60	DLOUHODOBÝ NEHMOTNÝ MAJETEK CELKEM Ř.02 ať 08	9	STAV K	konec roku	383650.43
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-POZEMKY /031/	16	STAV K	konec roku	47891591.83
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-UMĚLECKÁ DÍLA A PŘEDMĚTY /032/	17	STAV K	pololetí	22936.42
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-UMĚLECKÁ DÍLA A PŘEDMĚTY /032/	17	STAV K	konec roku	34687.35
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-STAVBY /021/	18	STAV K	konec roku	119402589.8
64581	60	5. DLOUHOD.FIN.MAJ.-DLU®NÉ CENNÉ PAPÍRY DR®ENÉ DO SPLATNOSTI /063/	36	STAV K	pololetí	5492.6
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-SAMOST.MOVITÉ VĚCI A SOUB.MOV.VĚCÍ /022/	19	STAV K	konec roku	5165638.96
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-DROBNÝ HMOTNÝ MAJETEK /028/	22	STAV K	pololetí	813469.76
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-DROBNÝ HMOTNÝ MAJETEK /028/	22	STAV K	konec roku	923461.91
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-OSTATNÍ HMOTNÝ MAJETEK /029/	23	STAV K	pololetí	14946.24
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-OSTATNÍ HMOTNÝ MAJETEK /029/	23	STAV K	konec roku	7263.03
64581	60	3. DLOUHODOBÝ HMOTNÝ MAJETEK-POŘÍZENÍ HMOTNÉHO MAJETKU /042/	24	STAV K	pololetí	29727961.16

Abbildung 6: Zugang über dbDIPview zur archivierten Datenbank ARIS (Automatisiertes Budgetinformationssystem, das bis zum Jahr 2009 vom Finanzministerium betrieben wurde)

Für die Archivierung von Social-Media-Daten, wie beispielsweise von X (ehemals Twitter)-Accounts, wurde ein finnisches Skript weiterentwickelt, um die Sammlung des tschechischen Twitter-Inhalts zu gewährleisten. Da dieses Netzwerk eine wichtige Plattform für die Kommunikation von Politikern und Institutionen mit der Öffentlichkeit darstellt, umfasst es derzeit mehr als 500 Accounts von Institutionen sowie prominenten Persönlichkeiten des öffentlichen Lebens.

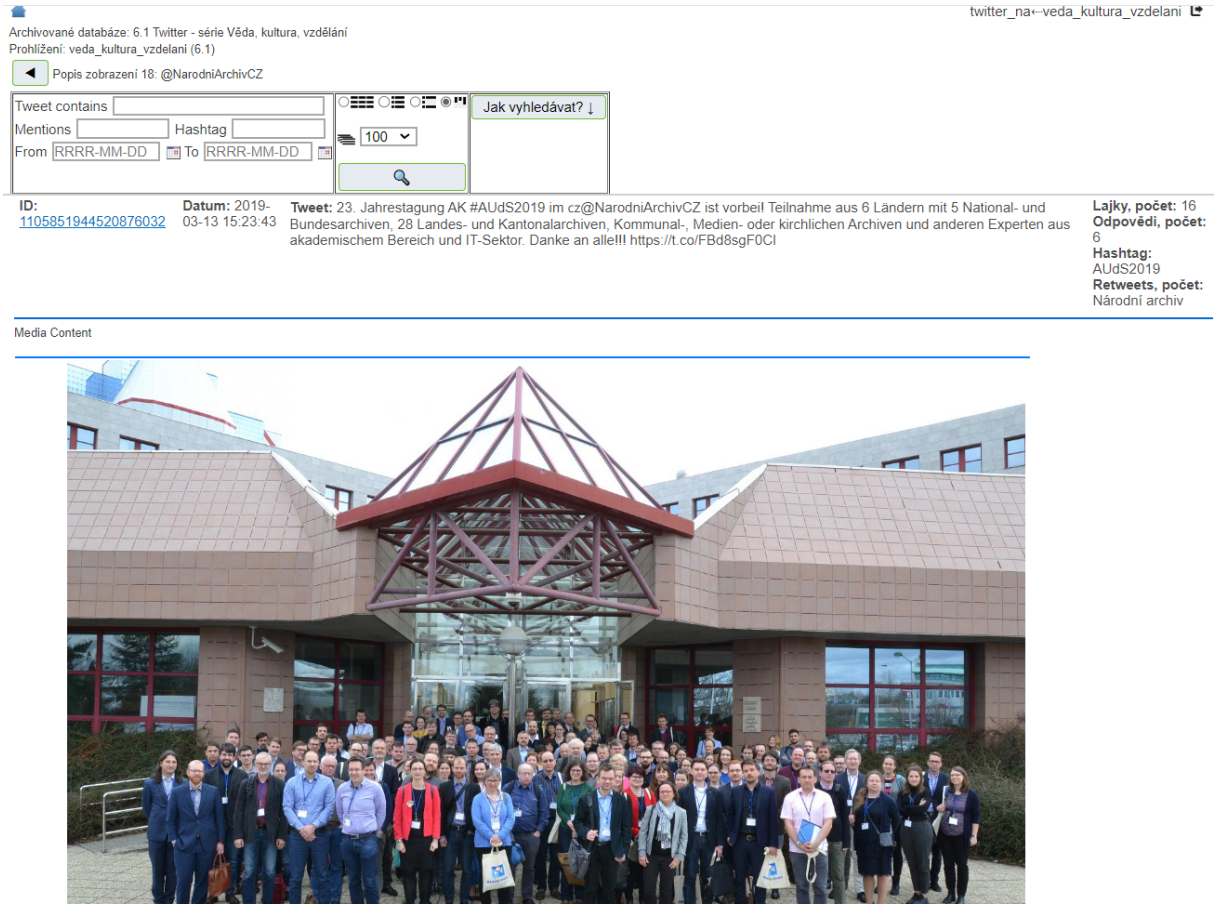


Abbildung 7: Zugang zu archivierten Tweets (von Nationalarchiv zur 23. Jahrestagung AUdS 2019)

Ein weiteres Projekt widmete sich der Archivierung von Geodaten, bei dem eine spezifische Methode zur Archivierung für die tschechischen Anforderungen entwickelt und ein Werkzeug namens ArchiGIS geschaffen wurde, das die Erstellung von SIP-Paketen gemäß dem E-ARK-Standard ermöglicht (Nationalarchiv 2022).

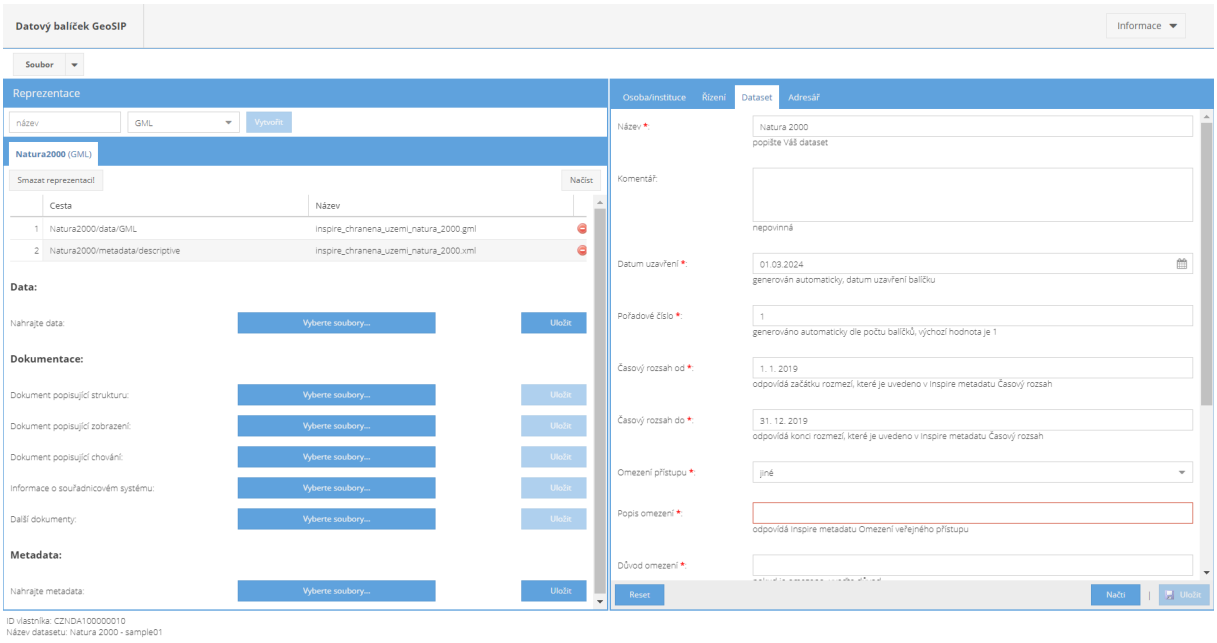


Abbildung 8: Das Programm ArchiGIS zur Erstellung von SIP-Paketen mit Geodaten nach dem E-ARK-Standard

Aus dem laufenden Betrieb ergab sich die Notwendigkeit, die Identifikation von Formaten zu verbessern. Durch die Zusammenarbeit mit dem britischen Nationalarchiv konnte das Formatregister PRONOM um weitere Formate ergänzt werden. Neben den heimischen Formaten wie der Office-Software-Familie 602 aus den 1980er- und 1990er-Jahren oder den Formaten FO/ZFO aus dem Bereich der tschechischen Datenspeicher und Formulare wurde auch die Identifikation international weit verbreiteter Formate erweitert. Zu diesen Formaten zählen insbesondere alle Varianten des PDF/A-Formats, weiter das OGC GeoPackage, das für geodatenbezogene Informationen genutzt wird, und ASiC (Associated Signature Containers), ein Format für die sichere Aufbewahrung digitaler Signaturen u. a.

In den Jahren 2022 und 2023 wurde in Zusammenarbeit mit Medienarchiven (Nationales Film-, Tschechisches Rundfunk-, Tschechisches Fernseharchiv) und der Tschechischen Technischen Universität eine moderne Version des Nationalen Standards der Formate für die Archivierung vorbereitet. Dieses Dokument wurde am 9. Juni 2023 vom Regierungsrat für die Informationsgesellschaft offiziell genehmigt und als Teil der Wissensbasis der E-Government-Architektur integriert, was die Interoperabilität und den Austausch von Daten zwischen verschiedenen Systemen erleichtert. Der Standard wurde so gestaltet, dass er eine hohe Flexibilität bei digitalen Archiven ermöglicht, insbesondere im Bereich des Risikomanagements, und hilft dabei, sich an unterschiedliche Herausforderungen und technische Anforderungen anzupassen. Darüber hinaus beschränkt sich der Standard nicht nur auf die Definition von Formaten, sondern enthält auch spezifische Empfehlungen für den Export in spezialisierten SIP-Paketen, um den vielfältigen Anforderungen der verschiedenen Datentypen gerecht zu werden. Im Gegensatz zur

derzeitigen Verordnung Nr. 259/2012 deckt diese Methodik eine Vielzahl von Datentypen ab, darunter: 1. Audio; 2. Binärdateien (ausführbare Dateien); 3. Datenbanken und strukturierte Daten; 4. E-Mails; 5. Textdokumente; 6. Internet und Intranet (Web); 7. Komprimierte Dateien; 8. Container; 9. Disk-Images; 10. Bilder (Raster); 11. Bilder (Vektor); 12. Geodaten (GIS); 13. Soziale Netzwerke (Medien); 14. Technische Zeichnungen und Modelle (CAD, BIM etc.); 15. Buchhaltungsunterlagen; 16. Audiovisuelle und kinematografische Dokumente (Národní standard formátů pro archivaci 2023).

Perspektiven und nächste Schritte

Mehrere Projekte zur Modernisierung des aktuellen digitalen Archivs stehen in Aussicht. Es wird an der Entwicklung einer Methodik und eines Verfahrens zur Archivierung von E-Mail-Konten gearbeitet, einschließlich der Implementierung des Capstone-Ansatzes (Stodůlka 2023). Im Laufe dieses Jahres soll auch ein von der Technologischen Agentur der Tschechischen Republik geleitetes Projekt abgeschlossen werden, das die Entwicklung eines Austauschformats für AIPs zwischen digitalen Archiven sowie die Definition von Kommunikationsdiensten zwischen digitalen Archiven und Erschließungssystemen zum Ziel hat, da es in der Archivlandschaft der Tschechischen Republik mehrere solcher Systeme gibt. Die Methodik zur Erteilung der Genehmigung für digitale Archive wird kontinuierlich weiterentwickelt, wobei das Nationalarchiv als fachlicher Gutachter beteiligt ist. Zur Erleichterung der Evaluierung hat es im Jahr 2018 die Methodik gemäß dem nestor-Siegel übersetzt, die von der Arbeitsgruppe des nestor für Zertifizierung genehmigt wurde (Pracovní skupina nestoru pro certifikaci, 2018).

Durch das Modul der Nationalen Digitalen Forschungsstelle (NDB), die am 1. Januar 2026 in Betrieb genommen wird, wird es möglich sein, die Verwaltung von Forschenden und deren Zugang zu elektronischem Archivgut zentral zu gewährleisten.

Seit 2023 ist das neue Projekt NDA III in Umsetzung, das die Implementierung eines neuen LZA-Systems mit neuen technologischen Möglichkeiten und erhöhten Anforderungen vorsieht. Die maximale verarbeitbare Größe eines Datenpakets (SIP) soll bis zu 8 TB betragen.

Geplant sind unter anderem:

- die Implementierung der E-ARK Informationspaket-Architektur (vor allem Inhaltstypen wie GeoSIP, SIARD, u. a.)
- die Integration von KI zur Metadatenanreicherung, OCR, maschinellen Übersetzung, Spracherkennung und Speech-to-Text
- die Unterstützung der Verarbeitung von Geodaten und Datenbanken
- die systematische Unterstützung von Emulationstechnologien

- die Erweiterbarkeit und Skalierbarkeit des gesamten Systems.

Das Projekt zielt darauf ab, moderne Technologien zu integrieren, um die langfristige Aufbewahrung und Nutzung digitaler Archive zu gewährleisten und dabei die Anforderungen verschiedener Datenformate und -typen zu erfüllen.

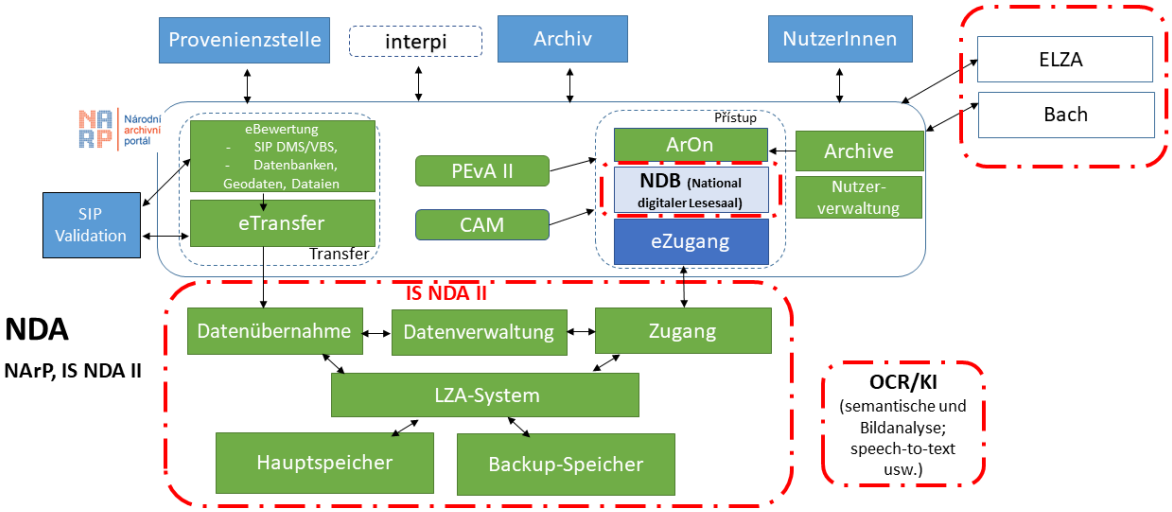


Abbildung 9: Architektur des NDA mit rot markierten Komponenten, die in den kommenden Jahren modernisiert werden.

Bibliografie

Kalina, T. (2004), 'Informatika a výpočetní technika', in: Benešová, E. et al. (Hg.), *Aby na nic a na nikoho nebylo zapomenuto: K jubileu ústředního archivu českého státu 1954–2004*. Praha: Státní ústřední archiv v Praze, S. 199-217.

Kunt, M., Lechner, T., Pokorný, R. (2020), 'Elektronické skartační řízení'. *Archivní časopis* 70/1, S. 52-73.

Národní standard formátů pro archivaci, 2023, https://archi.gov.cz/znalostni_baze:archivni_formaty (3.10.2024).

Naumann, K., Stodůlka, Z., Vojáček M., (2016), 'Mitteleuropäischer Perspektivvergleich in Prag', *Der Archivar* 69/1, S. 36-37.

Nationalarchiv (2022), *Transformace digitálních prostorových dat pro účely trvalého uložení v digitálním archivu*, Praha: Národní archiv, https://www.nacr.cz/wp-content/uploads/2024/02/Geodata_DEF.pdf (3.10.2024).

Pracovní skupina nestoru pro certifikaci (2018), 'Vysvětlující informace k Pečeti nestoru pro důvěryhodné digitální archivy', *Nestor Materialien* 17, http://nbn-resolving.de/urn:nbn:de:0008-2018071005 (3.10.2024).

Rechtorik, M. (2022), 'DB archiving and use at Czech authorities', in: Naumann, K. (Hg.), *Databases for 2080 workshop proceedings*. Stuttgart: Landesarchiv Baden-Württemberg, urn:nbn:de:101:1-2022071903 (3.10.2024).

Stodůlka, Z. (2016a), 'Digitales Archiv mit eigenen Kräften? Erfahrungen und Herausforderungen', in: *Mitteilungen des österreichischen Staatsarchivs: Digitale Archivierung – Innovationen, Strategien, Netzwerke*. Bd. 59. Innsbruck: Studien Verlag, S. 33-38.

Stodůlka, Z. (2016b), 'Archivportal: Neue Wege der Überlieferungsbildung und Nutzung in der Tschechischen Republik', in: *20. Jahrestagung des Arbeitskreises Archivierung von Unterlagen aus digitalen Systemen*, Potsdam, https://www.sg.ch/content/dam/sgch/kultur/staatsarchiv/auds-2016/daten---%C3%BCberneh-men-und-verarbeiten/03_STODULKA_AUDS_2016_Stodulka_praesentation_V01.pdf (3.10.2024).

Stodůlka, Z. (2018), 'E-Identität als Schlüssel zu den Dienstleistungen des digitalen Archivs', *Informationswissenschaft: Theorie, Methode und Praxis* 5/1, S. 110-124, https://doi.org/10.18755/iw.2018.13 (3.10.2024).

Stodůlka, Z. (2023), 'Archivierung der elektronischen Kommunikation: Ein Europäisches Dilemma zwischen Transparenz und Datenschutz', in: *26. Jahrestagung des Arbeitskreises Archivierung von Unterlagen aus*

digitalen Systemen, Mannheim, https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2023/_jcr_content/Par/sgch_downloadlist_1776496378/DownloadListPar/sgch_download_1303965982.ocFile/12_AUdS_2023_Stodulka_Elektronische_Kommunikation_V09.pdf

Validátor SIP. 2024, https://validatorsip.nacr.cz/ (3.10.2024).

Archivische Bewertung im digitalen Zeitalter

Maria von Loewenich

Einleitung

Es ist inzwischen eine Binsenweisheit, dass sich im Zuge der digitalen Transformation die Formen, in denen relevante Informationen bearbeitet, verwaltet und gesichert werden, stark verändert haben und mit hoher Dynamik weiter verändern. Neben die klassische Aktenführung sind seit den 1950er-Jahren zunächst Fachverfahren und seit den 1990ern E-Mail-Postfächer und Dateiablagen getreten. Und in den vergangenen Jahren ist dann eine Vielzahl weiterer Systeme hinzugekommen wie etwa Kollaborationstools, Messenger-Dienste und Business-Intelligence-Systeme. Die Archivwelt beschäftigt sich mit dieser Entwicklung seit nunmehr fast 30 Jahren. Der Schwerpunkt lag dabei vor allem auf den Fragen, wie genuin digitale Unterlagen ins Archiv übernommen und wie sie dort aufbewahrt und dauerhaft erhalten werden können. Des Weiteren wurde lange Zeit erwartet, dass bald flächendeckend E-Akte-Systeme eingeführt würden und dort, ebenso wie im Analogen, die maßgebliche Überlieferung entstehen werde (Keitel 2009). Vor etwa 15 Jahren setzte sich dann allmählich die Erkenntnis durch, dass die Einführung von E-Akte-Systemen vielerorts auf sich warten lässt und gleichzeitig in vielen anderen Systemen längst überlieferungsrelevante Inhalte entstanden sind, um die sich Archive kümmern sollten. Der Fokus verschob sich daraufhin verstärkt auf die Übernahme von Fachverfahren, Dateiablagen und E-Mail-Postfächern (vgl. exemplarisch Jacobs 2023, Axer 2019, Benauer 2020). Weniger in den Blick genommen wurde bisher, welche Auswirkungen die Veränderungen, die mit der digitalen Transformation einhergehen, auf die archivische Methodik und insbesondere die Methoden der archivischen Bewertung haben.[1] Dieser Thematik möchte sich der vorliegende Beitrag daher weiter annähern.

Die archivische Bewertung hat bekanntlich das Ziel, nach möglichst objektivierbaren Kriterien die Archivwürdigkeit von Unterlagen festzustellen. Des Weiteren soll im Zuge der archivischen Bewertung eine möglichst gleichmäßige, inhaltlich aussagekräftige und schlanke Überlieferung entstehen. Archivarinnen und Archivare versuchen dazu, strukturelle Merkmale zu identifizieren, mit denen die zu bewertenden Unterlagen gruppiert und kategorisiert werden können (vgl. Menne-Haritz 2001). Im deutschen Kontext sind das u. a. die Provenienz und der Trägerstoff der Unterlagen sowie der logische Sinnzusammenhang meist einer Akte oder eines Vorgangs.

[1] Umfassend hat sich mit dieser Frage bisher vor allem Verena Türck in ihrer Transferarbeit im Rahmen der Laufbahnprüfung für den höheren Archivdienst an der Archivschule Marburg auseinandergesetzt (vgl. Türck 2014).

Sie liegen in unterschiedlicher Gewichtung vielen methodischen Ansätzen zugrunde, und zwar häufig, ohne dass sich der Archivar oder die Archivarin dessen bewusst ist. Bei allen drei genannten Bezugspunkten haben sich im Zuge des digitalen Transformationsprozesses Änderungen ergeben. Diese Änderungen haben entscheidenden Einfluss darauf, so die These dieses Beitrags, wie gut sich die Ziele der archivischen Bewertung erreichen lassen. Die Bedeutung der drei Kriterien sollte daher überprüft und gegebenenfalls angepasst werden.

Die Provenienz der Unterlagen

Die Provenienz ist seit langer Zeit einer der wichtigsten Bezugspunkte der archivischen Arbeit. Sie ist nicht nur ein wesentlicher Faktor bei der Bewertung, sondern gibt in der Regel auch den Ordnungsrahmen vor, in dem Archivgut den Nutzenden bereitgestellt wird. Sie ist also tief im archivischen Denkprozess verankert. Eine besondere Rolle spielt die Provenienz beim sogenannten Federführungsprinzip, das vor allem das deutsche Bundesarchiv als wesentliches Instrument der Bewertung nutzt (vgl. Bundesarchiv 2011). Es sieht als ersten Bezugspunkt vor, wer der Produzent der jeweiligen Unterlagen ist. Diese Information wird im Anschluss damit abgeglichen, in welchem Maße der Produzent der Unterlagen in der Bundesverwaltung für die entsprechenden Aufgaben zuständig ist und ob diese Aufgaben als grundsätzlich überlieferungsrelevant eingestuft werden. Ist der Produzent der Unterlagen federführend für die Bewältigung einer relevanten Aufgabe verantwortlich, sind die entsprechenden Unterlagen als archivwürdig zu bewerten. Ist er dagegen nur an der Wahrnehmung einer Aufgabe beteiligt, so werden seine Unterlagen in der Regel kassiert. Hinter diesem Vorgehen steht die Grundannahme, dass bei derjenigen Organisationeinheit, die federführend für eine Aufgabe zuständig ist, alle wesentlichen Informationen zusammenlaufen und so dort eine vollständige und zugleich verdichtete Überlieferung entsteht. Bei anderen Modellen wie der horizontalen und vertikalen Bewertung (Kretzschmar 1996) oder den Dokumentationsprofilen, wie sie vor allem im kommunalen Bereich verwendet werden (Becker 2009), steht zwar zunächst ein Aufgaben- oder gesellschaftlicher Bereich im Fokus. Im nächsten Schritt wird aber auch hier auf die Akteure und ihre Unterlagen und damit auf die Provenienz abgestellt.

Eine der Entwicklungen, die sich aus der digitalen Transformation der Verwaltung ergibt, ist, dass in zunehmendem Maße behördenübergreifend gearbeitet wird. Das heißt, dass nicht jede Stelle ihre eigenen Unterlagen führt, sondern mehrere Behörden oder Institutionen im selben System relevante Informationen festhalten, austauschen und nachhalten. Schon lange im Blick ist dieses Vorgehen bei Fachverfahren, wie sie beispielsweise die Polizei oder die Finanzverwaltung nutzen. Relativ neu ist hingegen der zunehmende Einsatz von Kollaborationstools wie

zum Beispiel Jira, Teams oder Share Point. Sie bieten Werkzeuge wie Kanban-Boards oder Wikis, mit denen die institutionenübergreifende Zusammenarbeit organisiert und auch dokumentiert werden kann. Bei der Weiterentwicklung des zentralen E-Akte-Systems in der Bundesverwaltung wird so beispielsweise das Kollaborationstool Jira eingesetzt. Dort können Change Requests als Tickets eingestellt werden, die im Anschluss auf einem Kanban-Board verschiedene Prozessphasen durchlaufen. Alle Behörden, die an der Weiterentwicklung des Systems beteiligt sind, haben auf Jira Zugriff. Sie kommentieren dort die jeweiligen Change Requests, hinterlegen wesentliche Informationen zum Arbeitsprozess und nutzen Jira als Medium, um sich über den Fortgang der Diskussion und der Umsetzung eines Change Requests zu informieren.

In einer Konstellation wie der geschilderten befindet sich die vollständige und verdichtete Überlieferung häufig nicht mehr in den Unterlagen der federführend verantwortlichen Stelle, sondern im gemeinsam genutzten System, in diesem Fall im Kollaborationstool Jira. Denn die Beteiligten senden ihre Zuarbeiten, Einschätzungen usw. nicht mehr an die federführende Stelle, sondern hinterlegen sie selbst direkt im System. Zwar führen die beteiligten Organisationseinheiten meist parallel dazu auch noch eigene Akten und Vorgänge, sie enthalten aber häufig nicht oder nur sehr unvollständig die Inhalte, die im gemeinsam genutzten System nachgehalten sind. Es ist daher in solchen Fällen problematisch, unhinterfragt provenienzorientierten Bewertungsansätzen zu folgen. Vielmehr sollte in diesen Fällen angestrebt werden, Unterlagen aus dem gemeinsam genutzten System zu übernehmen. Je nach Konstellation kann es natürlich sinnvoll sein, diese Unterlagen um Unterlagen insbesondere der federführend verantwortlichen Stelle zu ergänzen.

Der Bedeutungsverlust der Provenienz als archivischer Bezugspunkt ist auch noch in anderer Hinsicht zu beachten. Wie eingangs schon angerissen, spielt sie nicht nur bei der archivischen Bewertung eine Rolle, sondern auch bei allen sich daran anschließenden Ordnungsprozessen. Denn vor allem die Tektonik bildet standardmäßig Provenienzen ab. Die Fachverfahren, Kollaborationstools und anderen Systeme, die institutionsübergreifend genutzt werden, passen jedoch nicht in dieses Schema. Es besteht deshalb die Gefahr, dass das Objekt dem archivischen Ordnungsrahmen angepasst wird und nicht umgekehrt. Oder anders formuliert: Es besteht die Gefahr, dass Informationen, die im Entstehungszusammenhang zusammengehörten, gemäß den beteiligten Provenienzen getrennt werden. Damit wird der archivische Grundsatz verletzt, den vorarchivischen Ordnungsrahmen zu respektieren. Das nachträgliche Aufteilen von Unterlagen nach Provenienzen ist also in diesem Sinne als unzulässig zu bewerten. Zumindest in der internationalen archivischen Diskussion ist diese Problematik schon seit geraumer Zeit präsent. Seit

über zehn Jahren arbeitet der International Council on Archives (ICA) am Erschließungsstandard Records in Context, der ermöglicht, einem Objekt mehrere Institutionen zuzuordnen und den starren hierarchischen Ordnungsrahmen der Tektonik zu überwinden (vgl. Wildi 2023). Dieser Standard muss nun allerdings erst noch Eingang in die im deutschsprachigen Raum eingesetzten Archivinformationssysteme finden.[2]

Der Trägerstoff der Unterlagen

Der zweite Bezugspunkt der archivischen Bewertung, der hier besprochen werden soll, ist der Trägerstoff, also das Medium, auf dem die jeweiligen Informationen aufgezeichnet sind. Das erscheint gegebenenfalls zunächst überraschend. Denn bereits seit mindestens 15 Jahren ist es eine anerkannte Tatsache, dass bei digitalen Unterlagen der Trägerstoff kein Bezugspunkt mehr sein kann. Allerdings wurde in diesem Kontext der Trägerstoff stets als die konkrete Festplatte, eine Diskette oder dergleichen definiert. Weniger reflektiert wurde, dass auch das System, in dem Informationen nachgehalten werden, eine Art spezifischen Trägerstoff darstellt.

Auch im analogen Bereich kamen Karteien, Kartensammlungen usw. zum Einsatz, das vorherrschende Arbeits- und Dokumentationsmittel war jedoch die Akte. Im Zuge der digitalen Transformation sind neben die klassische Aktenführung nun, wie eingangs schon ausgeführt, zahlreiche andere Systeme getreten, in denen relevante Informationen bearbeitet und nachgehalten werden. Die Aktenführung und die daneben eingesetzten Systeme stehen nicht als Solitäre nebeneinander, vielmehr bestehen zwischen ihnen zahlreiche Verbindungen und Überscheidungen. Zum Teil werden Daten zwischen den Systemen über Schnittstellen ausgetauscht, zum Teil besteht der Zusammenhang der Informationen nur auf der inhaltlichen Ebene. Als Beispiel dafür, wie eine solche Systemlandschaft aussehen kann, kann das deutsche Bundesarchiv dienen (vgl. Abb. 1): Im Zentrum befindet sich die analoge und digitale Aktenführung. Mit den Akten in Verbindung stehen zahlreiche Fachverfahren, über die einzelne Aufgaben oder Arbeitsschritte erledigt werden. Das reicht von der Zeiterfassung (FAZIT) und der Beantragung und Abrechnung von Dienstreisen (TMS) bis hin zum Archivinformationssystem BASYS, mit dem nicht nur das Archivgut verwaltet wird, sondern auch wesentliche Prozesse der Benutzung gesteuert werden. Neben diesen Systemen werden zahlreiche weitere eingesetzt, die einen eher inoffiziellen Charakter haben wie etwa Dateiablagen, APEX-Datenbanken und Outlook-Postfächer.

[2] Zu dieser Frage hat der Ausschuss „Archivische Fachinformationssysteme“ der Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA) am 23. April 2024 im Landesarchiv Berlin eine Tagung veranstaltet. Vgl. die Präsentationen unter: https://www.bundesarchiv.de/das-bundesarchiv/kooperationen-und-partner/kla/archivische-fachinformationssysteme/ (21.11.2024).

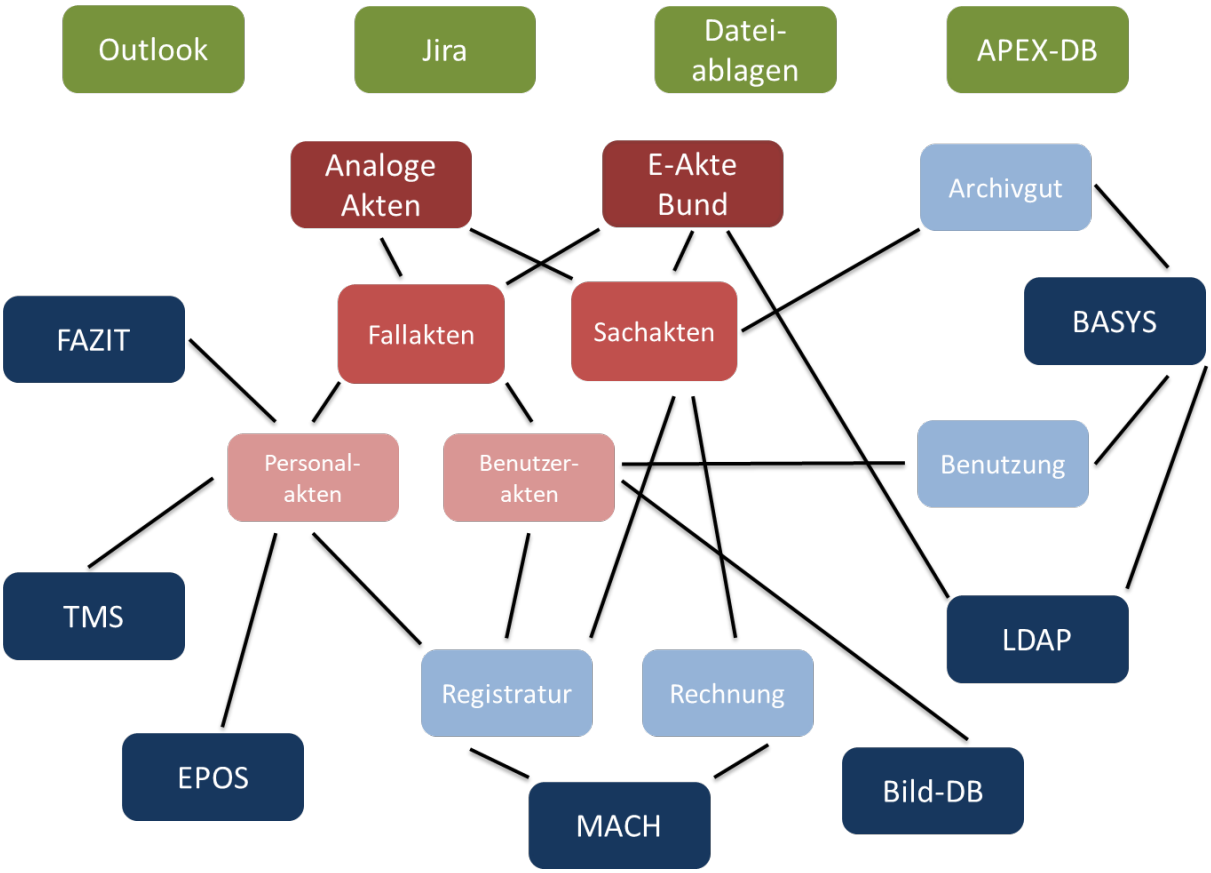


Abbildung 1

Die Verbindungen und Überschneidungen der Systeme führen dazu, dass Informationen zum Teil redundant vorliegen. Ein häufig anzutreffendes Beispiel dafür ist, dass ein Sachverhalt primär in einem Fachverfahren bearbeitet wird, das Ergebnis der Bearbeitung aber am Ende des Arbeitsprozesses beispielsweise in Form einer PDF-Datei in ein E-Akte-System importiert wird. Eines der klassischen Ziele der Bewertung ist, eine möglichst dichte Überlieferung ohne Redundanzen zu bilden. Dieses Ziel liegt auch den Ansätzen Federführungsprinzip und horizontale und vertikale Bewertung zugrunde. Die konsequente Umsetzung dieser Prinzipien kann in dem Fall, dass Informationen sowohl in einem Fachverfahren als auch in einer Akte vorliegen, zu dem Schluss führen, die Daten aus dem Fachverfahren als kassabel zu bewerten. Denn in der Akte sind ja bereits alle wesentlichen Informationen in verdichteter Form abgebildet. Diese Vorgehensweise blendet aber aus, dass die Form der Unterlagen im digitalen Bereich wesentlich von dem jeweiligen System geprägt wird, in dem die Unterlagen verwahrt werden. Dieselben Informationen können einen gänzlich anderen Charakter haben je nachdem, ob sie etwa als Daten in einem Fachverfahren gespeichert sind oder in einer Datei in einem E-Akte-System liegen. Und mit dieser anderen Form geht auch einher, dass sich die Informationen auf andere Weise auswerten und nutzen lassen.

Informationen, die in Form von Daten vorliegen, eignen sich in besonderem Maße für quantifizierende Ansätze, wie sie vor allem in den Sozial- und Wirtschaftswissenschaften angewendet werden. Bei diesen Ansätzen liegt der Fokus nicht auf Grundsatzentscheidungen oder dem besonderen oder exemplarischen Einzelfall, sondern auf der Summe aller Fälle. Beim Nachhalten in Akten sind diese Informationen zwar ebenfalls vorhanden, sie sind dort aber in der Regel dem jeweiligen Einzelfall zugeordnet. Sie eignen sich damit nicht oder nur sehr eingeschränkt dazu, Informationen quantifizierend auszuwerten. Es ist also bei der Bewertung digitaler Unterlagen nicht, wie in der Vergangenheit postuliert, vom Trägerstoff zu abstrahieren. Vielmehr sind die verschiedenen Systeme und ihre spezifischen Strukturen in die Bewertungsentscheidung mit einzubeziehen. Je nach Konstellation kann das dazu führen, dass nicht entweder die Akten oder die Daten im Fachverfahren als archivwürdig bewertet werden, sondern beides.
Ein Beispiel, bei dem beide Überlieferungsformen archivwürdig sein können, sind Informationen zu Beschäftigten einer Behörde. In der Verwaltung werden inzwischen fast flächendeckend Fachverfahren eingesetzt, in denen Personalstammdaten nachgehalten werden, also Basisdaten zu allen Beschäftigten wie zum Beispiel ihre organisatorische Zuordnung, ihre Besoldungsstufe/Entgeltgruppe und ihre bisherigen dienstlichen Verwendungen. Solche Daten können von Interesse sein, um beispielsweise die Geschäftsverteilung in einer Behörde zu einem bestimmten Zeitpunkt nachzuvollziehen oder quantifizierende Aussagen über die Personalstruktur einer Stelle oder eines ganzen Verwaltungszweigs zu treffen. Neben diesen Systemen stehen die Personalakten, die diese Informationen ebenfalls enthalten, aber in der Regel deutlich darüber hinausgehen. Sie sind vor allem dann relevant, wenn Informationen zu einer konkreten Person gesucht werden. Die Wahrscheinlichkeit, dass das eintritt, steigt, je prominenter die jeweilige Person war. Daher werden in der Regel insbesondere Personalakten aus Leitungsbereichen als archivwürdig bewertet. Werden nun, um beim Beispiel zu bleiben, Personalstammdaten zu allen Beschäftigten einer Behörde sowie Personalakten zu herausragenden Persönlichkeiten übernommen, so ergeben sich Redundanzen in der Überlieferung. Sie sollten natürlich nicht unkritisch und unreflektiert entstehen. Überwiegt jedoch der spezifische Wert der jeweiligen Unterlagen, so sollten sie hingenommen werden.

Akte und Vorgang als logische Sinneinheiten

Der letzte Bezugspunkt, der hier betrachtet werden soll, ist die Akte bzw. der Vorgang als logische Sinneinheit. Es ist archivwissenschaftlicher Konsens, dass nur im Ausnahmefall in die Ordnung einer Akte oder eines Vorgangs eingegriffen wird. Hintergrund ist, dass auch die konkrete Ordnung eines Schriftgutobjekts eine wesentliche Kontextinformation darstellt. Denn die

Veränderung des Kontextes kann den Sinn der Informationen beeinflussen und im schlimmsten Fall sogar verfälschen (vgl. International Council on Archives 1996). Doch nicht nur im archivischen Rahmen wird der Akte eine besondere Stellung zugeschrieben. Auch in der deutschen Rechtstradition ist die Akte eine wichtige, Informationen strukturierende Einheit. Dabei wird zwischen zwei Aktenbegriffen unterschieden, und zwar zwischen der Akte im materiellen Sinne und der Akte im formellen Sinne (Schneider 2024, Rn. 26). Die Aktenführung im formellen Sinne meint das gemeinsame Aufbewahren von Dokumenten in einem Schriftgutobjekt. Im analogen Bereich ist das zum Beispiel ein Leitz-Ordner, im digitalen Bereich ein Datencontainer in einem E-Akte-System. Unter einer Akte im materiellen Sinne werden dagegen alle aktenrelevanten Informationen verstanden, die denselben Sachverhalt betreffen. Dabei ist nicht relevant, wo und in welcher Form sie gespeichert bzw. gelagert werden. Auch Informationen, die an unterschiedlichen Orten zu finden sind, werden als Teile derselben Akte begriffen, sofern sie in einem logischen Sinnzusammenhang stehen. Im analogen Bereich sind die Akte im materiellen und formellen Sinne in der Regel identisch. Inzwischen sind neben die Akte, wie bereits ausgeführt, jedoch zahlreiche andere Systeme getreten, in denen Informationen bearbeitet und nachgehalten werden. Das hat zur Folge, dass Informationen, die gemäß dem materiellen Aktenbegriff einer Akte zuzuordnen sind, nun nicht mehr zwangsläufig auch in einer Akte oder einem Vorgang im formellen Sinne zu finden sind. Die Akte bzw. der Vorgang stellen also gegebenenfalls keine abschließende logische Sinneinheit mehr dar.

Ein Beispiel für das Auseinanderfallen von der Akte im materiellen und formellen Sinne sind die Benutzungsakten des deutschen Bundesarchivs. Das Bundesarchiv verwaltet, wie bereits erwähnt, in seinem Archivinformationssystem BASYS nicht nur Archivgut, sondern steuert darüber auch wesentliche Benutzungsprozesse. So wird zum Beispiel im System nachgehalten, welche Akten ein Benutzer bzw. eine Benutzerin eingesehen hat. Früher wurden diese Informationen nach Ende einer Benutzung als PDF-Datei exportiert und in die Benutzungsakte geheftet. Da aber nicht immer eindeutig war, wann eine Benutzung abgeschlossen war, ist das häufig nicht geschehen. Die Akten blieben in der Folge unvollständig. Inzwischen verzichtet das Bundesarchiv aktiv darauf und definiert damit BASYS, was diese Informationen betrifft, als Teil der Benutzungsakte.

Werden nun ausschließlich, um beim Beispiel zu bleiben, Benutzungsakten ohne die Daten aus BASYS als archivwürdig bewertet, so sind die Akten im materiellen Sinne nicht mehr vollständig. Es ist natürlich möglich, in einem solchen Fall im Zuge der Bewertung zu entscheiden, Informationen, die sich in einem anderen System befinden, abzuschneiden, sofern damit der archivische Auftrag, Rechte zu sichern, nicht verletzt wird. Diese Entscheidung sollte aber

bewusst getroffen werden. Keine Option ist dagegen, weiterhin anzunehmen, dass Akten und Vorgänge grundsätzlich in sich abgeschlossene logische Sinneinheiten sind, die bei der Bewertung unhinterfragt als Bezugspunkt genutzt werden können.

Ein weiterer, in diesem Kontext relevanter Aspekt ist, dass in den Systemen, die neben die klassische Aktenführung getreten sind, Informationen nicht mehr zwangsläufig in logischen Sinnzusammenhängen im Sinne einer Akte oder eines Vorgangs strukturiert werden. Besonders augenfällig ist das bei Fachverfahren. Dort werden Daten, die in einer Beziehung zueinanderstehen, an unterschiedlichen Orten im System verwahrt, bei relationalen Datenbanken beispielsweise in verschiedenen Tabellen. Die jeweilige logische Sinneinheit entsteht erst in der Ansicht, also zum Beispiel, wenn die Daten zu einer Person aufgerufen werden. Schon bei wenig komplexen Fachverfahren gibt es jedoch nicht nur eine mögliche Ansicht der Daten, sondern viele verschiedene, je nachdem welche Abfrage gestellt wird.

Um auch das mit einem kurzen Beispiel zu illustrieren: Das deutsche Bundesarchiv hat die zentrale Datenbank der Treuhandanstalt (ISUD) übernommen.[3] Die Treuhandanstalt war eine Behörde, die die volkseigenen Unternehmen der DDR nach der deutschen Wiedervereinigung sanieren, privatisieren oder, falls nötig, abwickeln sollte (vgl. u. a. Pötzl 2019). In der Datenbank finden sich zu jedem Unternehmen Kerninformationen, wie zum Beispiel der Name, der Standort, die Branche, der Käufer oder Details zum Vertragsabschluss. Nach allen diesen Informationen lassen sich Treffermengen generieren. Dabei kann es sich zum Beispiel um alle Unternehmen in einem Ort, alle Unternehmen einer Branche, alle Unternehmen, die dieselbe Person gekauft hat, usw. handeln. Bei jeder dieser Abfragen entsteht eine eigene logische Sinneinheit. In einem Fachverfahren gibt es also nicht mehr nur eine logische Sinneinheit, wie sie eben eine Akte oder ein Vorgang bilden, sondern unzählige mögliche logische Sinneinheiten.

Unter anderem das Positionspapier des Arbeitskreises Bewertung des Verbands deutscher Archivarinnen und Archivare (VdA) zu Fachverfahren aus dem Jahr 2014 geht davon aus, dass bei vielen Fachverfahren nur eine Auswahl der darin enthaltenen Daten als archivwürdig bewertet werden sollte (VdA-Arbeitskreis „Archivische Bewertung“ 2014). Vor allem bei komplexen, möglicherweise noch mit anderen Systemen vernetzten Fachverfahren ist es jedoch mitunter schwierig, alle Konstellationen im Blick zu behalten. Es besteht daher die Gefahr, dass ein Fachverfahren auf eine Fragestellung oder Verwendungszweck hin bewertet wird. Das kann in der Folge zu einer Verzerrung der ursprünglich vorhandenen Daten führen, wie sie mit dem Paradigma verhindert werden soll, dass logische Sinneinheiten bei der Bewertung nicht verändert werden dürfen.

[3] Bundesarchiv, B 412/135408.

Das deutsche Bundesarchiv bewertet unter anderem aus diesem Grund Fachverfahren inzwischen in der Regel entweder vollständig als archivwürdig oder kassabel. Nur im Ausnahmefall wird eine Auswahlentscheidung getroffen. Das bislang einzige Beispiel für eine solche Ausnahme sind die Daten der Bundesagentur für Arbeit (BA). Sie sind so umfangreich, dass sie selbst unter den heutigen Speicherbedingungen für elektronische Daten nicht vollständig überliefert werden können. Nach fachlicher Abstimmung mit der BA und dem ihr angegliederten Institut für Arbeitsmarkt- und Berufsforschung (IAB) werden nun etwa 5 % der Daten ins Bundesarchiv übernommen. Die zu überliefernden Daten werden mit einem Algorithmus ausgewählt, der die tatsächliche Zufälligkeit der Auswahl sicherstellt. Einziger Bezugspunkt ist dabei der jeweilige Kunde bzw. die jeweilige Kundin. Es wird also keine weitere inhaltliche Auswahlentscheidung getroffen, sondern es werden alle Daten zu einer Person übernommen.

Fachverfahren sind nur eine Art von Systemen, die neben die klassische Aktenführung getreten sind. Es stellt sich also die Frage, wie es sich in Bezug auf logische Sinneinheiten etwa bei Dateiablagen, E-Mail-Postfächern und Kollaborationstools verhält. Dateiablagen sehen einer klassischen Aktenführung auf den ersten Blick sehr ähnlich. Es gibt verschiedene Hierarchiestufen, und diese Hierarchiestufen sind in der Regel vom Allgemeinen zum Speziellen hin aufgebaut. Trotzdem unterscheiden sich Dateiablagen an einem entscheidenden Punkt von einer ordnungsgemäßen Aktenführung. Denn in Dateiablagen existiert keine klare Trennung von ordnungsgebendem Rahmen, den bei der Aktenführung der Aktenplan vorgibt, und konkreten Schriftgutobjekten, also den Akten und Vorgängen. Es ist also häufig nur schwer zu entscheiden, wo eine logische Sinneinheit beginnt und wo sie endet. Hinzu kommt noch, dass die Antwort darauf in jedem Teilabschnitt der Dateiablage eine gänzlich andere sein kann.

Noch unübersichtlicher ist die Situation bei E-Mail-Postfächern und Kollaborationstools. Denn diese Systemtypen kennen noch weniger verlässliche Strukturmerkmale. In E-Mail-Postfächern lassen sich zwar in der Regel Ordner zu Sachzusammenhängen anlegen. Sie sind jedoch nicht zwingend erforderlich und können außerdem parallel zu den systemseitig vorgegebenen Ordnern Posteingang und Postausgang bestehen, ohne von diesen klar abgegrenzt zu sein. Und unter dem Begriff Kollaborationstool werden ganz unterschiedliche Systeme zusammengefasst, die diverse ordnungsgebende Mechanismen haben. Es können Ordnerstrukturen vorhanden sein, die klassischen Dateiablagen ähneln. Informationen können aber auch in Wikis, in Forumsbeiträgen oder als Tickets hinterlegt sein. Das Prinzip klar zu definierender logischer Sinneinheiten im Sinne einer ordnungsgemäßen Aktenführung ist daher auch hier eher im Ausnahmefall zu erwarten.

Bei allen drei Arten von Systemen ist also ebenso wie bei Fachverfahren Vorsicht geboten, die archivische Bewertung auf zu tiefer Ebene anzusetzen. Denn die Gefahr ist relativ hoch, logische Zusammenhänge nicht zu erkennen und in der Folge zu zerstören. Im deutschen Bundesarchiv werden daher Dateiablagen, die zum Beispiel eine bestimmte Organisationseinheit geführt hat, in der Regel vollständig übernommen. Mit E-Mail-Postfächern und Kollaborationstools hat das Bundesarchiv bisher noch keine praktische Erfahrung, wird in Zukunft aber auch dort mit großer Umsicht vorgehen.

Resümee

Die digitale Transformation führt dazu, dass die Methoden der archivischen Bewertung kritisch hinterfragt werden müssen. Das Fortführen analoger Prinzipien kann der Qualität der Überlieferung schaden und sogar dazu führen, dass es zu inhaltlichen Verzerrungen kommt. Besondere Bedeutung kommt dabei den Bezugspunkten zu, die Bewertungsmodellen und -konzepten zugrunde liegen, häufig, ohne dass sich Archivarinnen und Archivare dessen bewusst sind. Der vorliegende Beitrag hat versucht, das anhand der Provenienz und des Trägerstoffs der Unterlagen sowie der Akte bzw. des Vorgangs als logischen Sinneinheiten zu zeigen. Sie stehen aber nur exemplarisch für weitere Bezugspunkte, die der archivischen Bewertung zugrunde liegen. Als Beispiel sei hier nur das Lebenszyklusmodell genannt, dessen Bedeutung es ebenfalls zu überdenken gilt. Es ist daher dringend geboten, die Diskussion über die archivische Bewertung neu zu entfachen und sie konsequent unter digitale Vorzeichen zu stellen.

Bibliografie

Axer, Christine (2019), ‘Überlieferungsbildung in Zeiten flüchtiger Strukturen’, in: *Verlässlich, richtig, echt – Demokratie braucht Archive! 88. Deutscher Archivtag in Rostock*, hg. v. Verband deutscher Archivarinnen und Archivare (Tagungsdokumentation zum Deutschen Archivtag, 23), Fulda, S. 99-107.

Becker, Irmgard Christa (2009), ‘Arbeitshilfe zur Erstellung eines Dokumentationsprofils für Kommunalarchive. Einführung in das Konzept der BKK zur Überlieferungsbildung und Textabdruck’, *Der Archivar* 62, S. 122-131.

Benauer, Maria (2020), ‘E-Mails, ihr Wert und ihre Bewertung’, *Scrinium* 74, S. 87-115.

Bundesarchiv (2011), *Strategiepapier Bewertungsgrundsätze (Dokumentationsprofil) des Bundesarchivs für Unterlagen der Bundesrepublik Deutschland*, Stand: 17. Mai 2011, https://www.bundesarchiv.de/assets/bundesarchiv/de/Downloads/Erklaerungen/beratungsangebote-strategiepapier-bewertungsgrundsaetze-dokumentationsprofil-des-bundesarchivs-fuer-unterlagen-der-bundesrepublik.pdf (21.11.2024).

International Council on Archives (1996), *Code of Ethics*, Peking, https://www.ica.org/app/uploads/2023/12/ICA_1996-09-06_code-of-ethics_EN.pdf (21.11.2024).

Jacobs, Rainer (2023), ‘Effiziente Verfahren, Echte Daten. Die Übernahme von Informationen aus Fachverfahren in das Bundesarchiv’, *Archiv. Theorie und Praxis* 76, S. 25-27.

Keitel, Christian (2009), ‘Elektronische Archivierung in Deutschland. Eine Bestandsaufnahme’, in: *Für die Zukunft sichern! Bestandserhaltung analoger und digitaler Unterlagen. 78. Deutscher Archivtag in Erfurt*, hg. v. Verband deutscher Archivarinnen und Archivare (Tagungsdokumentation zum Deutschen Archivtag, 13), Fulda, S. 115-128.

Kretzschmar, Robert (1996), ‘Vertikale und horizontale Bewertung. Ein Projekt der staatlichen Archivverwaltung Baden-Württemberg’, *Der Archivar* 49, S. 257-260.

Menne-Haritz, Angelika (2001), ‘Archivische Bewertung: Der Prozess der Umwidmung von geschlossenem Schriftgut zu auswertungsbereitem Archivgut’, *Schweizerische Zeitschrift für Geschichte* 51, S. 448-460.

Pötzl, Norbert F. (2019), *Der Treuhand-Komplex. Legenden, Fakten, Emotionen*, Hamburg.

Schneider, Jens-Peter (2024), ‘Verwaltungsverfahrensgesetz. § 29 Akteneinsicht durch Beteiligte’, in: Friedrich Schoch und Jens-Peter Schneider, *Verwaltungsrecht*, 5. Aufl. München 2024, S. 25-29.

Türck, Verena (2014), *Veränderung von Bewertungsgrundsätzen bei der Übernahme digitaler Unterlagen? Untersuchung von Bewertungsentscheidungen anhand baden-württembergischer Beispiele*. Transferarbeit im Rahmen der Laufbahnprüfung für den Höheren Archivdienst an der Archivschule Marburg (47. Wissenschaftlicher Lehrgang), o. O., https://www.landesarchiv-bw.de/sixcms/media.php/120/Transferarbeit_VerenaTuerck_02.pdf (21.11.2024).

VdA-Arbeitskreis „Archivische Bewertung“ (2014), *Bewertung elektronischer Fachverfahren*, Stand: Dezember 2014, https://www.vda.archiv.net/fileadmin/user_upload/pdf/Arbeitskreise/Archivische_Bewertung/Bewertung_Fachverfahren_Positionen_StandDez2014.pdf (21.11.2024).

Wildi, Tobias (2023), ‘Die Erweiterung des Provenienzprinzips: Der neue Records in Context-Standard’, *Archiv. Theorie und Praxis* 76, S. 166-173.

Prozesshandbuch Digitale Übernahme und Erschließung

Lambert Kansy und Kerstin Brunner

Ausgangslage und Zielsetzung

Vor rund fünfzehn Jahren begann im Staatsarchiv Basel-Stadt (StABS) die Entwicklung von Prozessen und Instrumenten, um digitale Dokumente aus der kantonalen Verwaltung in die Langzeitarchivierungslösung des StABS übernehmen zu können. Von 2010 bis 2012 wurden im Rahmen des Projekts „Informatisierung III“ die grundlegenden Werkzeuge hierfür realisiert; vorausgegangen war ein Konzipierungsprozess. Die Firma scope solutions entwickelte die Softwarekomponente scopeIngest mit dem Modul „Übernahmen“ in scopeArchiv und dem Ingest-Applikationsserver. Als digitales Repository wurde das open source Fedora Commons Repository gewählt und mit den Ingestkomponenten integriert.

Parallel zur Entwicklung der Werkzeuge unternahm das StABS zwischen 2010 und 2013 mit dem *Organisationshandbuch Digitale Archivierung StABS* einen ersten Anlauf, die Prozesse, Rollen, Zuständigkeiten und Akteure bei der digitalen Übernahme und Erschließung zu definieren. Die Aufnahme des Regelbetriebs begann jedoch erst 2014. Für rund drei Jahre bestand ein Pilotbetrieb. In dieser Phase entwickelte sich die digitale Archivierung mit ihren unterschiedlichen Geschäftsfällen (Übernahmen, Bewertung sowie Begleitung von Systemeinführungen) langsam weiter. Dies bot die Möglichkeit, die neuen Werkzeuge auszutesten, Mängel zu beheben und Erfahrungen mit der definierten Organisationsstruktur zu sammeln.

Bereits 2007/2008 waren erste genuin digitale Unterlagen und Daten als Archivgut übernommen worden. Dazu gehörten etwa Daten von Großrechnerapplikationen, Webseiten der kantonalen Verwaltung, die im Rahmen der Verwaltungsreform RV09 teilweise tiefgreifend neu strukturiert werden sollten, und Unterlagen aus dem elektronischen Journal der Kantonspolizei. Bis 2017 blieb die Anzahl der Fälle noch sehr überschaubar. Erst danach nahmen die Geschäftsfälle pro Jahr deutlich zu. Von den ersten Schritten bis anfangs 2024 hat das StABS rund 1111 Ingests durchgeführt und abgeschlossen.

2016 genehmigte der Große Rat des Kantons Basel-Stadt das Vorhaben „Ausbau und Weiterentwicklung des Archivinformationssystems des Staatsarchivs (Digitales Archiv 2.0)“. Dieses Investitionsprojekt soll auch den Bereich der digitalen Archivierung weiterentwickeln. Das 2022 gestartete Teilprojekt „p-transfer“ dient dabei der Neugestaltung von Prozessen und Kommunikation in der Überlieferungsbildung zwischen Archiv und Provenienzstellen. Beim Start des Teilprojekts „p-transfer“ 2020 zeigte sich deutlich, dass die Übernahme und Erschließung

genuin digitaler Unterlagen noch immer nicht konsolidiert erfolgte. Die bisher entwickelten organisatorischen Lösungen mit Prozessen und Rollen entsprachen nicht der Realität digitaler Archivierung. Die Übernahme und Erschließung digitaler Ablieferungen musste immer noch projektmäßig bearbeitet werden.

Aufgrund dieser Entwicklung wurde eine Überarbeitung organisatorischer Grundlagen der digitalen Archivierung respektive des wichtigen Teilbereichs Übernahme und Erschließung ins Auge gefasst. Die Erfahrungen aus mehr als fünfzehn Jahren Übernahme, Sicherung und Erschließung von genuin digitalen Unterlagen als digitales Archivgut sollten ausgewertet, zusammengefasst und standardisiert werden. Dabei sollten die Abläufe, Rollen, Zuständigkeiten und Akteure im StABS auf der Basis umfangreicher Erfahrungen neu definiert werden. Dies stellte zugleich eine Gelegenheit dar, das bislang nur einem kleinen Kreis von Beteiligten zugängliche Wissen auch weiteren Archivarinnen und Archivaren im StABS zugänglich zu machen und einen Prozess des Wissenstransfers in Gang zu setzen. Die voranschreitende digitale Transformation der Kantonsverwaltung führt dazu, dass die Übernahme und Erschließung von mehr Mitarbeitenden als bisher beherrscht werden muss. Die erfahrungsbasierte Erstellung eines Prozesshandbuchs für die Übernahme und Erschließung bot daher eine sehr willkommene Möglichkeit, den Wissenstransfer mit einem Standardisierungsschub zu verbinden.

Mit der Erstellung des Prozesshandbuchs waren folgende weitere Zielsetzungen verbunden: Durch die Neuerarbeitung und Modellierung der Prozesse, durch die Definition von Rollen, Zuständigkeiten und Beteiligten sollte eine stabile und einheitliche Praxis ermöglicht werden, die auf konkreten Erfahrungen beruht. Auch sollten die Schnittstellen der Ablauf- mit der Aufbauorganisation geklärt werden, mit dem Einbezug der Abteilungen Vorarchiv, Erschließung und Archivinformatik. Ferner galt es, die bisher bekannten Anwendungsfälle auszudifferenzieren und zu strukturieren. So sollte die digitale Übernahme und Erschließung künftig auch transparent und nachvollziehbar dokumentiert werden.

Um einen Beitrag an die archivübergreifende Diskussion zur digitalen Archivierung zu leisten, wurde beschlossen, das StABS-Prozesshandbuch *Digitale Übernahme und Erschließung* zu publizieren. Die Beschreibung der StABS-Praxis soll es ermöglichen, in einen fachlichen Austausch mit anderen Archiven und deren Praktiken sowie mit Aktenbildnern und deren Sicht auf die Praxis treten zu können.

Überblick, Grundlagen und zentrale Prozesse

Inhalt des Prozesshandbuchs „Digitale Übernahme und Erschließung“

Das Prozesshandbuch gliedert sich in acht Kapitel. Nach einer Einleitung, welche eine Übersicht sowie gesteckte Ziele umfasst, befasst sich das zweite Kapitel mit archivfachlichen, rechtlichen und methodischen Grundlagen.

Die Kapitel 3 „Lebenszyklus digitaler Unterlagen“ und 4 „Prozessgruppen“ bilden das Herzstück des Prozesshandbuchs. Kapitel 3 stellt den Lebenszyklus digitaler Unterlagen von der Produktion beim Aktenbildner bis hin zur Benutzung im Archiv vor; es positioniert die Übernahme und Erschließung im Gesamtkontext der digitalen Archivierung. Kapitel 4 geht im Detail auf Arbeitsschritte und Abläufe sowie Zuständigkeiten und Verantwortlichkeiten der beteiligten Akteure im Staatsarchiv Basel-Stadt ein; es hat diesbezüglich regelgebenden Charakter. Mit Hilfe von Prozessmodellierungen werden die einzelnen Bestandteile des Gesamtprozesses schematisch dargestellt.

Kapitel 5 „Anwendungsfälle“ umfasst bisher bearbeitete Ablieferungen nach Verarbeitungsprofil. Jeweils ausführliche Beschreibungen des gesamten Vorgehens inklusive „lessons learnt“ erläutern Übernahme und Verzeichnung von SIP nach eCH-0160, Dateiablagen und Dateisammlungen, Datenbanken und Fachapplikationen sowie Gever-Systeme nach jeweiligen Ablieferungen und Ablieferungsgruppen. So wird nachvollziehbar, wie die einzelnen Anwendungsfälle aus konkreten Ablieferungen entstanden sind. Dieses Kapitel dient primär dem StABS-internen Gebrauch, als fundierte Eigendokumentation und Nachschlagewerk. Darüber hinaus liefert es aber auch ein Repertoire an Referenzfällen, das laufend erweitert werden soll. Die Dokumentation dieser Referenzfälle soll als Basis für Weiterentwicklungen im Umgang mit verschiedenen künftigen Ablieferungen dienen. So wird es möglich, zu neuen Erkenntnissen zu gelangen und übergeordnete Prozesse und spezifische Vorgehensweisen technischen Veränderungen gewinnbringend anzupassen.

Die Kapitel 6 „Glossar“, 7 „Anhang“ und 8 „Bibliographie“ komplettieren das Prozesshandbuch. Es werden unter anderem Handreichungen zu Anwendungen, Werkzeugen und bestimmten Abläufen zur Verfügung gestellt.

Kapitel 2 „Grundlagen“

Als zentraler Orientierungspunkt erwies sich für das StABS unter anderem der Standard PAIMAS. In einem ersten Schritt erfolgte eine Übersetzung des Standards und der verschiedenen Kapitel mit zugehörigen Einzelpunkten in die deutsche Sprache. Danach wurden sämtliche Punkte auf StABS-eigene Begrifflichkeiten und Bedürfnisse hin überprüft und – wo nötig – angepasst. Als Fazit ließ sich daraus ein Korsett für die verschiedenen Prozesse ableiten, wie

sie in Kapitel 4 des Prozesshandbuches abgebildet sind. Die Prozessschemata wurden nach dem Modellierungsframework BPMN (Business Process Model and Notation) und nach dem einschlägigen schweizerischen eCH-Standard 0158 gestaltet. Darüber hinaus sind archivfachliche und rechtliche Grundlagen mit in das Handbuch aufgenommen worden. Dabei wurden immer wieder Verweise auf einschlägige Textstellen im bereits bestehenden und ebenfalls online verfügbaren *Handbuch Erschließung* des StABS angebracht.

Kapitel 3 „Lebenszyklus digitaler Unterlagen"

Kapitel 3 beschreibt den kompletten Lebenszyklus von digitalen Unterlagen, beginnend mit der Produktion und Aufbewahrung der Dokumente beim Aktenbildner bis hin zur Nutzung der archivierten Geschäfte und Bestände im StABS. Diese Beschreibung hat Eingang in das Prozesshandbuch gefunden, um eine Einbettung der spezifischen Prozesse im Bereich Übernahme und Erschließung anhand des gesamten Ablaufes zu ermöglichen. Konkret werden die folgenden Schritte in einem jeweils eigenen Unterkapitel beschrieben: Produktion und Aufbewahrung, Angebot/Anbieten, Bewertung, Ablieferungsvereinbarung, Pre-Ingest und Ablieferung/Transfer, Ingest, Verzeichnung, Post-Processing, Preservation und Benutzung.

Nebst der erwähnten Beschreibung wurde pro Lebenszyklus-Abschnitt jeweils eine gleichbleibende Matrix erstellt. Diese enthält Angaben zu Akteuren, Auslösern, Voraussetzungen, involvierten Systemen, Ergebnissen, Folgeschritten und Verzweigungen. Als Beispiel ist hier die Matrix für den Schritt „Ablieferungsvereinbarung" (Kapitel 3.5 im Prozesshandbuch) abgebildet.

Akteure	- Aktenbildende Stelle - Abteilung Vorarchiv - Archivleitung - Ggf. Abteilung Erschliessung - Ggf. Abteilung Informatik
Auslöser	- Abgeschlossener Bewertungsvorgang
Voraussetzungen	- Konsolidierter Bewertungsentscheid - eCH-Standard 0160[15] und 0165[16] - Liste aktueller archivwürdiger Dateiformate
Involvierte Systeme	- Geschäftsverwaltungssystem des StABS
Ergebnis / Output	- Von der Archivleitung unterzeichnete Ablieferungsvereinbarung
Folgeschritt	- Die aktenbildende Stelle initiiert die Generierung des SIP - Pre-Ingest und Ablieferung / Transfer
Verzweigungen	

Abbildung 1: Matrix Prozessschritt Ablieferungsvereinbarung

Kapitel 4 „Prozessgruppen“

Kapitel 4 befasst sich vertieft mit den Prozessgruppen Übernahme und Erschließung. Diverse Prozesse und Prozessgruppen werden nach Vorgaben von BPMN respektive eCH-0158 modelliert und darüber hinaus ergänzend strukturiert beschrieben. Dieser Teil des Prozesshandbuches hat für die Arbeit im StABS Vorgabecharakter und definiert Zusammenarbeit, Abhängigkeiten, Zuständigkeiten sowie Arbeitsschritte.

Bei den Prozessen wird zwischen vorgelagerten Prozessen und Kernprozessen unterschieden. Vorgelagerte Prozesse sind im Prozesshandbuch nicht detailliert behandelt, da der Fokus auf Übernahme und Erschließung liegt. Es handelt sich um folgende Prozesse:

- Kontaktaufnahme zwischen Aktenbildner und Archiv
- Anbietungen
- Beratungen (zum Beispiel im Bereich Records Management oder auch bei der Einführung einer neuen Software, deren Inhalte es später zu archivieren gilt)
- Bewertungsvorgänge (das Staatsarchiv unterscheidet zwischen Mikro- und Makrobewertungen)

Zu den Kernprozessen, welche im Detail erläutert werden, gehören:

- Ablieferungsvereinbarung
- Übernahme und Pre-Ingest (SIP-Generierung und Metadatenstandards, Transfer, SIP-Prüfung und -Validierung, Ingest-Vorbereitungen, gegebenenfalls Konvertierungen)
- Ingest (in seinen verschiedenen Ausformungen)
- Post-Processing (dem Ingest nachgelagerte Arbeiten wie die Vervollständigung von Verzeichnungsarbeiten, Qualitätssicherung, Abschlussarbeiten und Löschprozess)

In der Erläuterung des Kernprozesses „Ablieferungsvereinbarung“ hält das StABS die folgenden Punkte fest: Art der Strukturierung des Paketes, auszuwählende Dateiformate, Datenstruktur, Erfassungsregeln und Art der Metadaten, Übergabekanal, Ablauf zu Generierung und Prüfung eines Testpaketes, Kosten und diesbezügliche Zuständigkeiten, Löschprozedere, rechtliche Aspekte, Periodizität von Ablieferungen, Konditionen zu Ablehnungen von Datenpaketen. Die in der Vereinbarung enthaltenen Punkte sind nach technischen und inhaltlichen Gesichtspunkten gruppiert.

Mit der „Ablieferungsvereinbarung“ hat sich das StABS im Zuge des Projektes „p-transfer“ und bei der Gestaltung des Prozesshandbuches ein Werkzeug geschaffen, das in dieser Form im Regelfall noch nicht angewandt, obwohl es im Archivgesetz erwähnt ist. Wie bei einem Bewertungsentscheid sollen mit Hilfe der Ablieferungsvereinbarung Verbindlichkeiten geschaffen werden. Mit der Konsolidierung der Ablieferungsvereinbarung durch die Mitglieder des Teams

Digitale Archivierung (bestehend aus Mitgliedern der Abteilungen Informatik, Vorarchiv und Erschließung) soll nebst dem Mehr-Augen-Prinzip die Expertise aller Involvierten gewinnbringend mit einbezogen werden. Verbindlichen Charakter erhält das Papier mit der abschließenden Prüfung durch die Archivleitung sowie der Unterschrift seitens Aktenbildner und Archivleitung.
Unter „Pre-Ingest“ werden Transfer und Validierung eines Paketes zusammengefasst. Einzelne Aspekte können auch schon bei der Ausarbeitung der Ablieferungsvereinbarung zum Bereich Pre-Ingest gehören: nämlich dann, wenn in der Ablieferungsvereinbarung Parameter festgehalten werden, die Strukturierung oder erforderliche Metadaten sowie deren Form für ein Ablieferungspaket festlegen. In diesem Falle erfolgt ein erster und entscheidender Erschließungsschritt also bereits im Zuge des Prozesses, welcher zeitlich mehrheitlich in den vorarchivischen Bereich gehört. Klar nicht zum Bereich Pre-Ingest gehört aber der vorarchivische Schritt der Bewertung.
Zum Pre-Ingest kann des Weiteren die Prüfung von Testpaketen gehören, wie auch Arbeiten in Zusammenhang mit Dateikonvertierungen oder der StABS-internen Zusammenstellung eines Ingestpaketes bei Anlieferung von einer Ordnerstruktur.
„Ingest und Verzeichnung“ können automatisiert in einem Schritt vor sich gehen. Dies ist zum Beispiel dann der Fall, wenn ein Ingest eines eCH-0610-Pakets durchgeführt wird. Bei einem solchen Ingest werden die Verzeichnungseinheiten zusammen mit den Metadaten aus dem SIP (Submission Information Package) direkt ins AIS (Archivinformationssystem) geschrieben. Es können aber auch Verzeichnungseinheiten mit allen notwendigen und verfügbaren Metadaten vor dem Ingest manuell angelegt werden. Die zur Verzeichnungseinheit gehörenden Dateien werden danach verknüpft und der Ingest wird erst dann pro bereits existierende Verzeichnungseinheit angestoßen.

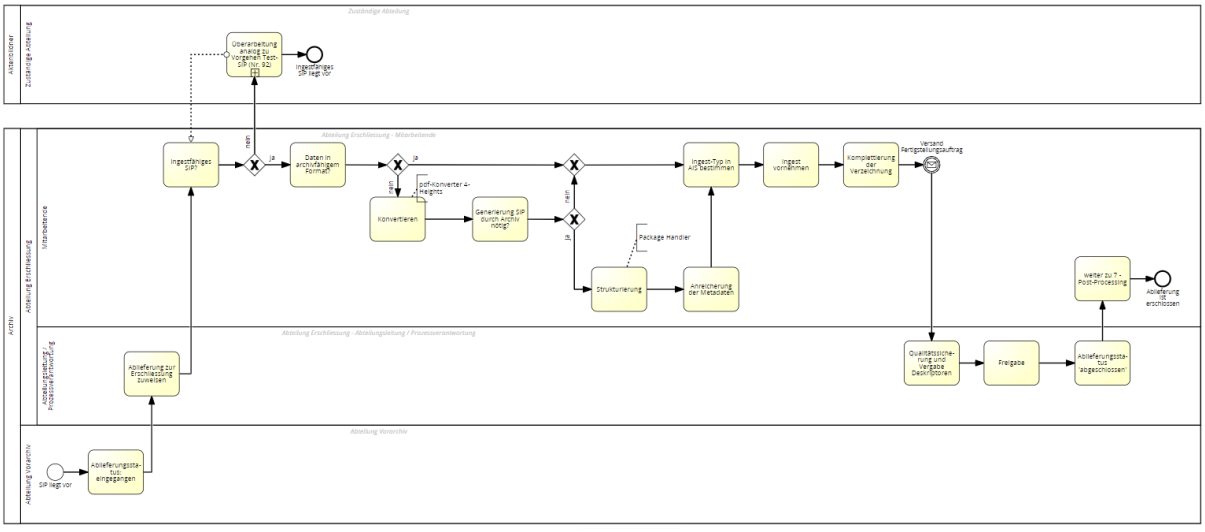


Abbildung 2: Prozessdiagramm „Ingest und Verzeichnung“

Unter „Post-Processing“ ist die Komplettierung der Verzeichnungsinformation im AIS zu verstehen. Hierzu gehören unter anderem das Setzen der korrekten Schutzfristen, die Einstellungen zur Sichtbarkeit der Metadaten, ob das Archivgut über den Digitalen Lesesaal publiziert werden darf und ob urheberrechtliche Parameter zu beachten sind. Des Weiteren gehören zum Post-Processing eine abschließende Qualitätssicherung gemäß 4-Augen-Prinzip, die Freigabe der verzeichneten Einheiten im AIS und das Anstoßen des Löschprozesses, welches wiederum den Aktenbildner mit einbezieht. Qualitätssicherung und Löschprozess werden jeweils im StABS-internen Geschäftsverwaltungssystem dokumentiert und dauerhaft aufbewahrt.

Beziehung zwischen Ablauf- und Aufbauorganisation

Der klassische vorarchivische Prozess der Ablieferungsvorbereitung und Übernahme (die Bewertung ausgenommen) wird durch den Einbezug von spezifischem Fachwissen aus den Abteilungen Informatik (spezifisches Fachwissen zu technischen Anwendungen zum Beispiel) und Erschließung (Definition von Strukturen, Ordnung und erforderlichen Metadaten) aufgebrochen. Es wird ein höheres Maß an Austausch, Kommunikation und Einbezug verschiedener Akteure aus verschiedenen Abteilungen zum richtigen Zeitpunkt erforderlich. Ebenso notwendig ist der Einbezug aller notwendigen Arbeitsschritte und Abklärungen vor einer Ablieferung. Denn nach erfolgter Ablieferung gestalten sich Umstrukturierungen und Anreicherungen von Metadaten sowie konservatorische Maßnahmen im Zuge der Erschließung komplexer oder gar unmöglich, im Unterschied zur Erschließung klassisch-analogen Archivguts.

Der Erschließungsprozess digitalen Archivguts unterscheidet sich auch sonst von demjenigen für analoges Archivgut. Die Erschließung digitaler Unterlagen verläuft wie bereits punktuell dargelegt in zeitlich gestaffelten Clustern, wobei aber die Erschließungsgrundsätze wo immer möglich denen von analogem Archivgut entsprechen sollen (zum Beispiel erforderliche Metadaten und deren Form). Zeitlich gestaffelte Cluster können im Zuge der Erschließung von digitalem Archivgut bis zu drei vorkommen:

1. Im Zuge des vorarchivischen Prozesses, zum Beispiel bei der Erstellung der Ablieferungsvereinbarung, oder immer dann, wenn Instruktionen zur Gruppierung oder Vorbereitung von Datenpaketen herausgegeben werden.
2. Mit dem Ingestprozess, in welchem nebst den Verzeichnungseinheiten die Metadaten aus dem SIP ins AIS übernommen werden (dieser Schritt kann bei einem bestimmten Ingest-Typ wegfallen). In der Regel handelt es sich um die folgenden zentralen Datenelemente, die entweder vom AIS generiert oder aus dem SIP ins AIS geschrieben werden: Signatur, Titel, Entstehungszeitraum, Verzeichnungsstufe und Erschließungsgrad.

3. Im Zuge der Komplettierung von Erschließungsinformationen im AIS. Im StABS ist es (noch) nicht der Fall, dass sämtliche Verzeichnungsinformationen einem SIP beigegeben werden. Es handelt sich hierbei um die folgenden zentralen Elemente, welche sich vornehmlich auf Sichtbarkeit von Metadaten, Archivgut und damit auch auf die Benutzung beziehen: Urheberrechtsstatus, Schutzfrist, Sichtbarkeit der Metadaten im Archivportal, Informationen zur physischen Beschaffenheit des Archivguts oder die Zugänglichkeit. Mit dem Wert „Zugänglichkeit“ wird im AIS des StABS die Publikation digitalen Archivguts über den Digitalen Lesesaal gesteuert.

Anwendungsfälle

Kapitel 5 „Aufbau und Inhalt“

Die in Kapitel 5 enthaltenen Anwendungsfälle zeigen auf, wie einzelne Übernahmen entstanden sind und welches die jeweiligen Schritte dabei sind. Die nachstehende Tabelle gibt einen Überblick der unterschiedlichen Anwendungsfälle.

#	Bezeichnung	Beispiel
UC01.01	Proprietäres SIP + Konvertierung nach eCH-0160	Polizeijournal ARAP
UC01.02	Natives SIP eCH-0160	Kantonsblatt
UC02.01	Dateiablage: Verzeichnisstrukturen	Ablieferung Sozialdemokratische Partei, Klientendossiers IV-Stelle und Kantonspolizei Administrativmassnahmen, Hostdaten Zentrale Informatikdienste
UC02.02	Dateiablage: WARC-Container	Webseiten kantonaler Dienststellen
UC03.01	Datenbank: SIARD-Container	Einwohnerinformationssystem

Abbildung 3: Übersicht Anwendungsfälle

Für jeden Anwendungsfall werden folgende Informationen zusammengetragen:

- Beschreibung des Falls
    - Art der Unterlagen
    - Involvierte Akteure
    - Besonderheiten
    - Dokumentation
    - Archivsignatur
    - Ablieferungsvereinbarung (technische Verarbeitungsregel scopeIngest)

    - Konvertierung von Dateien
    - Erschließungsarbeiten
- Lessons learnt
    - Besonderheiten
    - Do's & dont's
    - Abweichung von Vorgaben
    - Maßnahmen zur Einhaltung der Vorgaben

Beispielfall „Arap"

Da das Kapitel „Anwendungsfälle" in der Online-Version des Handbuchs nicht verfügbar ist, soll hier ein Anwendungsfall näher betrachtet werden.

Das StABS übernimmt das elektronische Polizeijournal „Arap" jährlich, vorhanden sind Unterlagen ab 2009, die Übernahmen setzten 2010 ein. Eine integrale Übernahme erfolgt für das Journal, das Kernmetadaten aller erfassten Fälle beinhaltet. Die eigentlichen Dossiers werden in Auswahl übernommen – mit Sample-Bildung nach Zufallszahlen durch das StABS – und bilden eine zweite Serie an Unterlagen. Da im zugrundeliegenden IT-System der Kantonspolizei keine Ablieferungsschnittstelle nach eCH-0160 implementiert werden konnte, werden die Unterlagen in einem spezifischen SIP-Format an das StABS geliefert und von diesem über eine Migration in ein eCH-0160-SIP ingestfähig gemacht.

Die Ablieferungsvorbereitung und das Sampling laufen folgendermaßen ab:

- Der Aktenbildner legt zum vereinbarten Zeitpunkt jährlich die Daten „Geschäftsdossier" auf den SFTP-Server des StABS.
- Die zuständige Person aus dem Team Vorarchiv erfasst die Ablieferung mittels scopeAblieferungen, Status angekündigt.
- Die zuständige Person aus der Abteilung Archivinformatik erfasst ein Ablieferungsprotokoll, zählt die Fälle aus, stellt mit einem Konfidenzintervall von 99% mit 3% Fehlertoleranz das Sample zusammen. Dabei werden allfällige Doppeleinträge herausgesucht und manuell von der zuständigen Person der Abteilung Archivinformatik entfernt.
- Eine Liste der ausgewählten Geschäfte wird an die zuständige Person bei der Kantonspolizei verschickt.

Die Bildung des SIP besteht aus den folgenden Schritten:

- Die zuständige Person bei der Kantonspolizei bildet das SIP und stellt es via SFTP-Server dem StABS zur Verfügung. Die weitere Verarbeitung des SIP wird durch die verantwortliche Person der Abteilung Archivinformatik im StABS vorgenommen.

- Die durch die Kantonspolizei abgelieferten Daten im SIP werden auf Korrektheit und Vollständigkeit überprüft.
- Das SIP wird in das eCH-0160 Format transformiert.
- Danach wird das SIP neu gepackt und mittels KOST-Val validiert. Eine Information über diesen Arbeitsschritt geht an die zuständige Person der Abteilung Vorarchiv

Der nun folgende Prozess der Erschließung wird angestoßen, indem die zuständige Person aus der Abteilung Vorarchiv im Tool scopeAblieferungen den Status auf „eingegangen“ anpasst. Damit wird die Abteilungsleitung Erschließung darüber in Kenntnis gesetzt, dass die Ablieferung zur weiteren Bearbeitung bereit ist. Die Abteilungsleitung weist die Ablieferung der zuständigen Person aus der Abteilung Erschließung zwecks Bearbeitung zu.
Die Verzeichnung besteht aus den folgenden Schritten:

- Ingest des SIP inklusive Verknüpfung auf die entsprechende Ablieferung und Generierung und Transfer ins AIS der erforderlichen Metadaten der Verzeichnungseinheiten
- Ergänzung der Metadaten auf Stufe Serie
- Ergänzung der Metadaten für den Eintrag zum Geschäftsjournal
- Ergänzung der Metadaten für die Einträge zu den Arap-Geschäften in Auswahl

Beurteilung der Arbeit (Fazit und Ausblick)

Rückblickend auf die von 2022 bis 2024 dauernde Erarbeitung des Prozesshandbuchs „Digitale Übernahme und Erschließung“ lässt sich festhalten, dass erst aufgrund konkreter praktischer Erfahrungen Prozesse definiert und ausgearbeitet werden konnten, die stabil sind und in der Praxis effektiv angewendet werden können. Es waren konkrete Erfahrungen aus Übernahmen notwendig, die als Anschauungsmaterial dienen.
Ebenso zeigte sich in aller Deutlichkeit, dass digitale Archivierung auf alle Abteilungen ausstrahlt und Abläufe verändert werden müssen. So ist eine engere Zusammenarbeit erforderlich und die klassische Arbeitsteilung im Archiv muss überdacht werden. Als Beispiel dient die Verzahnung von Vorarchiv und Erschließung bei der Erarbeitung von Ablieferungsvereinbarungen.
Die digitale Erschließung unterscheidet sich deutlich von der klassischen Erschließung analogen Archivguts. Neu muss die Erschließung die zentralen Aufgaben Sichten, Ordnen und Strukturieren bei standardisierten digitalen Ablieferungen bereits im Zuge des vorarchivischen Prozesses erbringen. Sie fokussiert sich nach dem Ingest auf Kontrolle, Korrektur und Ergänzung von beim Ingest maschinell erstellten Beschreibungsmetadaten. Hinzu kommt die Aktualisierung von Beschreibungen auf Bestands- und Fondsebene sowie die Erstellung von Verweisen zwischen Verzeichnungseinheiten. Unverändert bleibt das Setzen von Schutzfristen und die

Festlegung von Sichtbarkeit. Neu ist die Festlegung der Publikation von digitalem Archivgut im Digitalen Lesesaal – eine aus Sicht der Benutzung zentrale Aufgabe. Darüber hinaus müssen auch urheberrechtliche Aspekte im Zuge der Erschließung bestimmt und die entsprechenden Metadaten verzeichnet werden, damit die Benutzung auf dieser Basis gesteuert werden kann. Ähnliches gilt für die Verzeichnung von digitalisiertem Archivgut. Insgesamt nimmt dadurch die Wichtigkeit einer qualitativ hochstehenden Erschließungsarbeit sogar zu. Fehler in der Erschließung führen in der digitalen Welt schneller und in größerem Ausmaß zur Verletzung von Datenschutz und Persönlichkeitsrechten wie auch des Urheberrechts.

Zugleich darf die Etablierung der Prozesse zur Bearbeitung digitaler Ablieferungen die Prozesse der analogen Archivierung nicht beeinträchtigen. Es braucht bis auf weiteres beide Prozessbereiche – ein kompletter Umbau hin zu einer rein digital arbeitenden Organisation ist bis zum vollständigen Abschluss des digital turns in der Kantonsverwaltung nicht denkbar. Jedoch kann die digitale Praxis Veränderungen in der analogen Archivierung anregen. So könnten Ablieferungsvereinbarungen auch bei analogen Ablieferungen hilfreich sein. Ablauf- und Aufbauorganisation müssen daher gesamthaft neu zusammengebracht werden. Schließlich ist die Befähigung der Mitarbeitenden ein zentraler Aspekt: Mit dem Prozesshandbuch hat das StABS ein Werkzeug für den internen Wissenstransfer respektive Wissensaufbau geschaffen.

Das 2024 veröffentlichte Prozesshandbuch soll eine ganze Reihe von Wirkungen auslösen. Das Vorliegen funktionierender Prozesse und klar beschriebener Anwendungsfälle soll die Praxis normativ anleiten. Der Kompetenzaufbau soll erleichtert werden, etwa bei der Einarbeitung neuer Mitarbeitender oder bei digitalen Erschließungsarbeiten durch Archivarinnen und Archivare, die bislang analoge Ablieferungen erschlossen haben. Insgesamt soll langfristig ein gemeinsames Verständnis für diese Prozesse im Rahmen der digitalen Archivierung geschaffen und so abteilungsübergreifende Zusammenarbeit befördert werden. Für Provenienzstellen sollen genauere Informationen über den Ablauf digitaler Ablieferungen verfügbar sein. Und die Publikation des Prozesshandbuchs auf der Website des StABS soll den archivfachlichen Diskurs zur digitalen Archivierung unterstützen und der interessierten Öffentlichkeit einen Einblick in die StABS-Praxis geben. Schließlich stellt die Prozessmodellierung den Ausgangspunkt für weitere Projekte dar: so etwa im Projekt „p-transfer“ bei der Schaffung einer vorarchivischen Kollaborationsplattform und für die interne Aufgabensteuerung. Künftig dienen die erarbeiteten Prozesse und Rollen überdies als Ausgangspunkt, an den die digitale Bestandeserhaltung im Projekt „p-preserve“ anknüpfen kann, mit dem Prozesse und Instrumentarien für konkrete Praktiken zur Erhaltung digitalen Archivguts erarbeitet und umgesetzt werden können.

Das Prozesshandbuch *Digitale Übernahme und Erschließung* kann auf der Website des StABS als PDF-Datei heruntergeladen werden: https://www.bs.ch/pd/kultur/museen-und-andere-dienststellen/staatsarchiv/strategie-und-konzepte

III.

VERBUNDLÖSUNGEN, EBENENÜBERGREIFENDE KOORDINATION UND SCHNITTSTELLEN ZU FACHANWENDUNGEN

Meldedaten-Archivierung: Die Betrachtung eines vielseitigen Gesamtprozesses aus unterschiedlichen Perspektiven

Antje Scheiding und Henrike Thomas

Einführung

Die Archivierung von Meldedaten ist bundesweit seit vielen Jahren ein Thema in der kommunalen Archivlandschaft. Mit Inkrafttreten des Bundesmeldegesetzes (BMG) wurden 2015 die einzelnen Landesmeldegesetze durch einheitliche Regelungen zu Löschung, Aufbewahrung und Anbietung an die zuständigen Archive abgelöst. In der Folge erhöhten sich für sächsische Kommunen die allgemeinen Aufbewahrungsfristen für Meldedaten nach Wegzug oder Tod von 10 auf 55 Jahre. Bundeseinheitlich wurde die Löschung der Daten nach der Aufgabenerfüllung mit dem Anbietungsvorrang an die zuständigen Archive verfügt.

Auf dem Gebiet der ostdeutschen Bundesländer setzte die Erfassung elektronischer Meldedaten in den neu eingerichteten, kommunalen Meldeämtern zu Beginn der 1990er-Jahre ein. Das elektronische Melderegister ist damit im Regelfall das am längsten eingesetzte Fachverfahren in sächsischen Kommunen. Die Schwerpunkte in der Überlieferungsbildung aus dem elektronischen Melderegister liegen auf den aussonderungsreifen Meldedaten nach Wegzug oder Tod bis zum Stichtag 31.10.2005 und vor allem auf den aufgelösten Personenverkettungen.

Im vorliegenden Bericht werden die gesammelten Erfahrungen und der komplexe Gesamtprozess der Meldedatenarchivierung aus zwei unterschiedlichen Perspektiven beleuchtet: aus der internen Sicht eines Kommunalarchivs (Stadtarchiv Leipzig) und aus der externen Sicht einer koordinierenden Beratungsstelle (Leitstelle elektronisches Kommunalarchiv).[1]

[1] Der Beitrag gibt im Wesentlichen den Sachstand Februar 2024 wieder.

Das elektronische Kommunalarchiv Sachsen (elKA)

Mit dem elektronischen Kommunalarchiv steht den Kommunen in Sachsen seit 2022 eine auf kommunaler Ebene geschaffene Verbundlösung zur Verfügung, um die Herausforderungen in der elektronischen Archivierung gemeinsam zu bewältigen. Hervorgegangen aus einem vierjährigen Aufbauprojekt, ist es der Sächsischen Anstalt für Kommunale Datenverarbeitung (SAKD) gesetzlich zugeordnet. Mit dem elKA wird den Kommunen die technische Infrastruktur eines digitalen Langzeitarchivs (Leistungen eines Rechenzentrums und DIMAG als Archivierungssoftware) zentral bereitgestellt. Darüber hinaus gibt es eine Leitstelle, die koordinierend und beratend tätig ist. Dem Verbund haben sich bisher 42 Kommunen unterschiedlicher Größe angeschlossen – darunter kleine kreisangehörige Gemeinden, sieben Landkreisverwaltungen sowie große Stadtverwaltungen von Mittelstädten und kreisfreien Städten.[2] Das bedeutet, dass die nutzenden Archive vergleichsweise heterogene Voraussetzungen für die Durchführung der Archivierungsprozesse mitbringen. Es ist bekannt, dass die elektronische Archivierung ein sehr komplexes, herausforderndes Themenfeld ist. Aus diesem Grund bietet die Leitstelle neben den rein technischen und administrativen Aufgaben auch unterschiedliche Serviceleistungen an, wie beispielsweise Schulungsangebote (DIMAG, Einstieg in die elektronische Archivierung) und verschiedene Formen des Fachaustausches. Darüber hinaus werden konkrete Unterstützungsleistungen für die Kommunalarchive im Archivierungsprozess angeboten. Das Angebot wird insbesondere für die Verarbeitung und Archivierung von Transferpakten aus Aussonderungsläufen des Meldefachverfahrens MESO Classic oder auch VOIS|MESO in Anspruch genommen.

Abbildung 1: Modell der Kooperationspartner des elektronischen Kommunalarchivs

Prozess der Meldedatenarchivierung

Für die Verarbeitung der Meldedaten hat die Leitstelle einen Standardprozess aufgesetzt, um Massendaten strukturiert, normiert und effizient zu verarbeiten. Eingebaute Kontrollpunkte helfen, Fehlerquellen in der Verarbeitung zu vermeiden. Das standardisierte Vorgehen hat sich ab

[2] Stand August 2024.

Herbst 2023 entwickelt, wobei mit weiteren Optimierungen und dem Ausbau der Automatisierungspotenziale zu rechnen ist.

Der Prozess der Meldedatenarchivierung beginnt mit der Datenbereitstellung der Transferpakete durch die Kommune oder den IT-Dienstleister. Bisher wurden Aussonderungspakete aus dem Fachverfahren MESO Classic oder VOIS|MESO verarbeitet. Die Pakete beinhalten unterschiedliche Nachrichtentypen. Es gibt Hauptdatensätze, welche die Meldedaten nach Ablauf der Frist von zehn Jahren nach Wegzug oder Tod wiedergeben, sowie diverse Teildatensätze, die bei Datenbereinigungsläufen und aufgelösten Personenverkettungen entstehen. Je nach Größe der Gemeinde umfassen die Aussonderungspakete zwischen 10'000 und über 1 Mio. XML-Datensätze, die aus ZIP-Containern zu je 5'000 XML-Dateien bestehen.

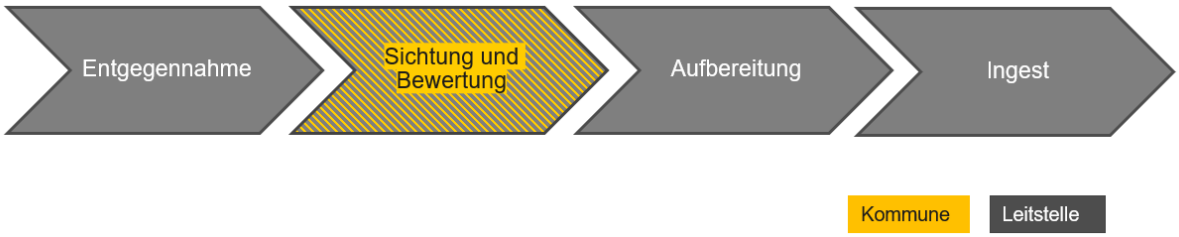


Abbildung 2: Prozess der Meldedatenverarbeitung im elKA

Die erste Phase im Prozess der Meldedatenverarbeitung/Meldedatenarchivierung ist die Entgegennahme, welche die Authentizität und Integrität der Datenlieferung sicherstellen soll und entscheidend für deren Weiterverarbeitung ist. Dazu zählen Aufgaben des Pre-Ingests wie Viren- und Integritätsprüfung, Entpacken der Container, Formaterkennung und perspektivisch die XML-Strukturvalidierung. Diese Prüfschritte sind für die Vertrauenswürdigkeit der Unterlagen und eine spätere Nutzbarmachung sehr wichtig. In der nächsten Phase werden mit Werkzeugen der Massenverarbeitung die Voraussetzungen für eine Bewertungsentscheidung geschaffen, indem das Archiv einen Überblick über den Inhalt des Gesamtpaketes erhält. Da die Transferpakete auch eine Vielzahl von Datensätzen aus Bereinigungsläufen enthalten, werden etwa Nachrichtentypen mit abgelaufenen Ausweisdaten bislang als nicht archivwürdig bewertet.

Für den Gesamtüberblick wird eine Konkordanz in Form einer Excel-Datei über alle XML-Dateien geniert. Diese tabellarische Übersicht ist die Grundlage für die Auswertung hinsichtlich der Archivreife der Datensätze nach Falleintritt und der potenziellen Archivwürdigkeit der Nachrichtentypen. Darüber hinaus werden unterschiedliche Provenienzen identifiziert, etwa aus eingemeindeten Ortschaften oder Datensätzen benachbarter Gemeinden aus der interkommunalen Zusammenarbeit. Die Ergebnisse der Sichtung werden dem Archiv in Form einer

Dokumentation und Datenanalyse zur Verfügung gestellt. Anhand dessen entscheidet das Archiv, welche Datensätze archivwürdig sind und letztlich in das DIMAG überführt werden sollen.

Für den Ingest werden die Datensätze nach Mustern aufbereitet, welche die Leitstelle für die Ablagestruktur in DIMAG und die Bildung der Archivpakete anbietet. Die Übertragung ins DIMAG erfolgt mit IngestTool.[3] Auf diese Weise werden die Archivpakete erzeugt und mit Metadaten aus der Konkordanz (etwa Name, Geburtsdatum, letzte Meldeanschrift und Schutzfristen) ergänzt.

Erkenntnisse aus der Verarbeitung der Meldedaten

Bei der Verarbeitung von Meldedaten handelt sich um Massendaten, die an sich strukturiert und verständlich sind. Jedoch sind die Arbeitsschritte der Entgegennahme und Aufbereitung zeitintensiv und erfordern Werkzeuge, die für die Massenverarbeitung von Daten geeignet sind. Erschwerend kommt hinzu, dass in dem Bereich der Meldedatenarchivierung wenig Erfahrungswerte vorliegen und noch offene Fragen zur Kopplung mit dem Archivfachinformationssystem (AFIS) und zur Nutzbarmachung bestehen. Daher ist in der Prozessgestaltung weiterhin mit Unsicherheiten und Risiken zu rechnen.

Als Optimierungsansätze sieht die Leitstelle weitere Automatisierungsoptionen, etwa durch den Einsatz des IngestProzessModuls.[4] Darüber hinaus wurden bereits die Werkzeuge für die Verarbeitung von Massendaten verbessert und der Arbeitsplatzspeicher der PC-Arbeitsplätze erweitert. Mit einem in Python programmierten Skript wurden etwa eine automatisierte Sortierung nach Nachrichtentypen und die Bildung von Ingestpaketen realisiert. Damit können an jede Kommune angepasste Bewertungskriterien umgesetzt und Ingestlisten mit ausgelesenen Metadaten effizient gebildet werden.

Für die Prozessgestaltung wurde zunächst iterativ-experimentell vorgegangen, um darauf aufbauend die Abläufe laufend zu optimieren. Bewährte Verfahren wurden über Checklisten und eingebaute Qualitätskontrollen normiert, um Fehler und Risiken in der Datenverarbeitung auszuschließen.

Die Erfahrungen haben gezeigt, dass die Leitstelle die Prozesse nicht allein verbessern und optimieren kann. Sie ist auf den engen Austausch mit der Kommune angewiesen. Auch der

[3] DIMAG-IngestTool (DIT) ist ein leistungsfähiges Modul des DIMAG-Verbundes, welches weitestgehend automatisiert eine regelbasierte Paketierung und Bildung von Archivinformationspaketen (AIP) im DIMAG-Kernmodul ermöglicht. Für das Mapping können aus verschiedenen Metadatenquellen (wie Excel oder XML-Dateien) Informationen ausgelesen und verarbeitet werden.

[4] Das IngestProzessModul (IPM) setzt bei der Entgegennahme digitaler Unterlagen im Archiv an. Es steuert verschiedene Bearbeitungsschritte des Pre-Ingests und bindet externe Werkzeuge. Dazu gehören u. a. die Integritäts- und Vollständigkeitsprüfung, Formaterkennung, Formatvalidierung und die Strukturvalidierung.

Erfahrungsaustausch im Entwicklungsverbund des DIMAG ist besonders hervorzuheben, denn durch den Wissenstransfer zwischen den Partnern werden wertvolle Hinweise weitergegeben. Der Fachaustausch im Verbund und das Teilen von Erfahrungswissen tragen letztlich zur optimalen Weiterentwicklung bei.

Überlieferung der Meldedaten in Leipzig

Vorgeschichte

Meldenachweise sind vielgenutzte Quellen bei wissenschaftlichen und privaten Anfragen. Direkter Vorgänger der Meldedaten ist die Meldekartei mit einer Laufzeit von etwa 1955 bis 1993. Sie wurde nach der Wiedervereinigung vom elektronischen Melderegister abgelöst. 2008 erfolgte die Einführung der Meldesoftware MESO Classic und 2023 die Umstellung auf VOIS|MESO (jeweils Software der HSH Soft- und Hardware Vertriebs GmbH).

Die Überlieferung der Meldenachweise im Stadtarchiv Leipzig endete entsprechend Anfang der 1990er-Jahre. Das Sächsische Meldegesetz, welches bis zum 31.10.2015 gültig war, sah eine kurze Aufbewahrungsfrist von zehn Jahren nach Wegzug oder Tod einer Person vor. Hinzu kamen die unmittelbar zu löschenden aufgelösten Personenverkettungen, beispielsweise bei Volljährigkeit (Trennung zwischen Eltern- und Kind-Datensatz). Es galt, den offensichtlichen Missstand der bisher nicht erfolgten Aussonderung aus den Meldedaten zu beheben, der aus Unerfahrenheit und fehlenden technischen Möglichkeiten resultierte.

Den ersten Anlauf unternahm das Stadtarchiv Leipzig 2015. In Zusammenarbeit mit der Meldebehörde wurden testweise Aussonderungen durchgeführt. Grundlage der Exporte aus MESO Classic waren XML-Dateien, die nach dem herstellereigenen Standard xarchivo erstellt wurden.[5] Die Beschaffung der Software archivo für die ausgesonderten Meldedaten wurde fachlich ausgeschlossen, da eine fachverfahrensunabhängige Archivierung und Nutzbarmachung zu priorisieren ist.[6]

Das Stadtarchiv Leipzig analysierte die Daten umfassend hinsichtlich Struktur und Inhalt. Anschließend wurden Unstimmigkeiten und Verständnisfragen über die Meldebehörde mit dem Verfahrenshersteller geklärt. Eine Unstimmigkeit bestand in der Dokumentation der seinerzeit gültigen XML-Schema-Definition 1.2 (XSD), in der die Werte „true“ und „false“ für die

[5] XML-Schema-Definition (XSD) veröffentlicht unter https://www.hsh-berlin.com/modules.php?name=HSH_Content&cid=92&download=2460 (letzter Aufruf am 22.08.2024).

[6] Software archivo als „Zwischenarchiv“ für Meldedaten, entwickelt von Archiven zusammen mit dem MESO-Hersteller HSH; s. a. Worm (2016).

Auskunftssperre fälschlicherweise umgekehrt beschrieben waren. Dieser Fehler wurde in der Version 1.3 behoben.[7]

```
<?xml version="1.0" encoding="UTF-8"?><!--HSH Soft- und Hardware Vertriebs
GmbH--><!--Version XMeldBibliothek: 17.0.2.13521--><xarchivo:archivuebergabe.ARC001
xmlns:xarchivo="http://www.hshsoft.de/xarchivo"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.hshsoft.de/xarchivo
http://www.hshsoft.de/xarchivo/archivschnittstelle_xarchivo_v1.3.1.xsd" version="1.3.1"
fassung="2023-01-16" produkt="MESO" produkthersteller="HSH Soft- und Hardware Vertriebs
GmbH" produktversion="2.35.0">
<xarchivo:nachrichtenkopf>
<xarchivo:ereignisnummer>1</xarchivo:ereignisnummer>
<xarchivo:ereignis>Tod/Sterbefall</xarchivo:ereignis>
<xarchivo:speziellesereignis>Weitere Löschung von Daten bei Archivierung der
Person</xarchivo:speziellesereignis>
<xarchivo:erstellungszeitpunkt>2023-09-12T22:16:59</xarchivo:erstellungszeitpunkt>
<xarchivo:tagesvorgangszaehler>127874</xarchivo:tagesvorgangszaehler>
<xarchivo:absender>
<xarchivo:behoerdenkennung>ags:14713000,hsh:1436500000</xarchivo:behoerdenkennung>
<xarchivo:organisationseinheit>
<xarchivo:bezeichnung>Stadt Leipzig</xarchivo:bezeichnung>
```

Abbildung 3: Kopf einer XML-Datei nach Ereignis "Tod/Sterbefall"

Der Schwerpunkt des Stadtarchivs Leipzig lag von Beginn an auf der Nutzbarkeit: Wie können die Meldedaten recherchierbar gemacht und Auskünfte daraus erteilt werden? Über das Einlesen der massenhaften XML-Dateien in Microsoft Excel wurde mit Microsoft Access eine Datenbank erstellt. Mit entsprechendem Layout, vorgefertigten Anfragen und eingerichteten Druckfunktionen entstand somit in Eigenleistung eine einfache und zweckdienliche Rechercheмöglichkeit. Als Einstieg konnte eine Personen- oder Adressrecherche gewählt werden. Jederzeit konnte dabei die originäre XML-Datei ermittelt werden – das eigentliche Archivgut. Auf diese Weise konnten auch die Personenverkettungen in einer separaten Access-Datenbank zugänglich gemacht werden.

Das Projekt verzögerte sich bis 2018 und wurde ohne Durchführung der finalen Aussonderung pausiert. Entscheidender Antrieb für die Weiterarbeit war die Entwicklung des elektronischen Kommunalarchivs. Aus heutiger Perspektive wird deutlich, dass wichtige Bestandteile der Archivierung von Meldedaten damals nicht betrachtet wurden. Dazu zählen insbesondere die Bildung von Archivpaketen sowie der Nachweis im AFIS.

Die Erstübernahme 2023

Mit Inbetriebnahme des DIMAG im Stadtarchiv Leipzig im April 2022 waren die Voraussetzungen der Archivierung gegeben. Die Dringlichkeit der Übernahme, die schon lange aufgrund der vorliegenden archivreifen Daten bestand, verstärkte sich abrupt durch den bevorstehenden

[7] Beide Versionen der XSD sind veröffentlicht auf der Internetseite des Verfahrensherstellers: https://www.hsh-berlin.com/modules.php?name=HSH_Content&cid=92. Betroffen ist der Eintrag unter <xs:element name="auskunftssperre" type="xs:boolean"> (22.08.2024).

Softwarewechsel von MESO Classic auf VOIS|MESO im Herbst 2023. Die Aussonderung sollte noch mit dem alten System stattfinden, denn zum einen waren zu Testzwecken bereits Exporte durchgeführt worden und zum anderen sind mögliche Risiken einer Datenmigration in das neue VOIS|MESO abgewandt.

Seit 2008 löste die Meldebehörde Leipzig sogenannte Archivierungsläufe in MESO Classic aus. Dabei prüfte die Meldesoftware, ob es Daten gibt, die aus dem aktiven Datenbestand entfernt werden müssen. Auf Grundlage der Archivierungsläufe konnten Exporte angestoßen werden. Das Stadtarchiv Leipzig hatte sich mit der Meldebehörde darauf verständigt, die Exporte in Jahresscheiben zu bündeln (2008 bis 2023). In der Summe übergab die Meldebehörde über 1 Mio. XML-Dateien mit einer Gesamtspeichergröße von 5,7 GB.

Jahr	belegter Speicher in MB	Anzahl XML-Dateien
2008	4,55	1.129
2009	33,38	7.940
2010	614,46	102.864
2011	121,34	22.523
2012	139,22	24.337
2013	222,77	39.521
2014	204,51	34.569
2015	56,54	13.028
2016	1.630,00	302.302
2017	236,58	39.061
2018	440,31	104.422
2019	768,37	208.649
2020	335,66	68.234
2021	298,43	53.001
2022	301,46	50.394
2023	278,48	44.850
Summe	**5.686,06**	**1.116.824**

Tabelle 2: Übernommene Datenmengen nach Jahresscheiben

Bewertung:

In den Exporten sind unterschiedliche Nachrichtentypen enthalten (vgl. Tabelle 2). Pro Person und Nachrichtentyp entsteht eine XML-Datei. Es kann somit zu einer Person 1-n verschiedene XML-Dateien über die Gesamtdauer der Meldezeit in Leipzig entstehen. In Leipzig wurden 15 verschiedene Nachrichtentypen ausgegeben, die in Anzahl, zeitlichem Auftreten und letztlich in ihrer Aussagekraft stark variieren. Die bedeutendste Überlieferung ist der Nachrichtentyp „Personarchivierung". Hier sind – ähnlich zur analogen Meldekartei – alle Meldeadressen einer Person mit Ein- und Auszugsdatum aufgeführt. Daneben wurden die Nachrichtentypen zur Auflösung eines Familienverbands sowie eines Beziehungsendes als archivwürdig bewertet (Personenverkettungen). Sie liefern die wichtigen Verkettungen zwischen Personen (v. a. Eltern-Kind und Ehepartner/-innen). Allen Nachrichtentypen ist ein XML-Abschnitt „Betroffener" gemein, in dem Name(n), Geburtsdatum und die zum Löschzeitpunkt aus dem aktiven Meldedaten gültige Meldeadresse enthalten sind. An dieser Stelle befindet sich zudem die eineindeutige Kennung der Person (ID). Über diese ID können verschiedene Nachrichtentypen eineindeutig einer Person zugewiesen werden, was auch die Problematik der Wiederherstellung der Personenverkettungen löst. Die Personenverkettungen, die woanders als unvollständige Datensätze erst nach Zusammenführung mit dem Hauptdatensatz archiviert werden sollen (vgl. Worm, 2016), sind in Leipzig und den anderen sächsischen Kommunen Teil der zu archivierenden Daten. Die Aufbewahrungsfrist ist abgelaufen und die darin enthaltenen Informationen sind archivwürdig. Richtig ist, dass die große Masse der Daten aufgrund personenbezogener Schutzfristen über Jahre nicht beauskunftet werden können, doch lange Schutzfristen gibt es in Archiven auch in anderen Überlieferungen.

Die häufig vorkommenden Nachrichttypen „Datenbereinigung" und „Löschen Aktive" wurden vor der finalen Bewertungsentscheidung umfangreich analysiert. Im Nachrichtentyp „Datenbereinigung" sind Daten zum Nachweis der Richtigkeit bei der Ausstellung von Pässen und Ausweisen sowie waffen- und sprengstoffrechtlichen Verfahren enthalten. Sie entstehen unter Angabe des Datums der ausführenden Behörde, wenn etwa eine waffenrechtliche Erlaubnis erlosch oder entzogen wurde. Bei „Löschen Aktive" geht es ausschließlich um die Entnahme der entsprechenden Passdaten aus den aktiven Meldedaten. Bei den waffen- und sprengstoffrechtlichen Daten waren gewisse Informationswerte feststellbar, doch wurde gemäß dem Federführungsprinzip auf eine Archivierung über die Meldebehörde verzichtet. Weshalb diese Nachrichtentypen in unterschiedlichen Jahren unterschiedlich häufig auftraten, wurde vorerst nicht weiter untersucht.

Die übrigen Nachrichtentypen traten nur in geringer Anzahl auf und enthielten mit Ausnahme des Abschnitts „Betroffener“ entweder gar keine Daten oder nur den Eintrag „Ausschluss von der Wählbarkeit“ unter Wahlrechtsausschluss, etwa durch Wegzug ins Ausland.

Lfd. Nr.	Nachrichtentyp	Anzahl	Anmerkung
1	Personarchivierung	532.080	
2	Datenbereinigung	289.319	ab 2016 (BMG)
3	Auflösung Familienverband Eltern	101.204	
4	Auflösung Familienverband Kind	82.964	
5	Löschen Aktive	79.360	ab 2019
6	Beziehungsende	30.027	
7	Auflösung Familienverband Kind Tod	1.505	Bedeutung: gesetzliche/-r Vertreter/-in verstorben
8	Tod/Sterbefall	130	ab 2013
9	Abmeldung von Amts wegen unbekannt	116	nur 2013-2016
10	Wegzug unbekannt	93	nur 2013-2019
11	Wegzug Ausland	19	nur 2013-2018
12	Abmeldung von Amts wegen ins Ausland	4	nur 2014-2016
13	Wegzug Inland	1	nur 2011
14	Wegzug von NEW [Nebenwohnsitz]	1	nur 2017
15	Wegzug HAW [Hauptwohnsitz] zu NEW	1	nur 2014

Tabelle 3: Auftreten verschiedener "Nachrichtentypen" in der Übernahme Leipzig

Ablagestruktur in DIMAG und Bildung der Archivpakete

Für die Bildung der Archivpakete (oder der Verzeichnungseinheiten aus Perspektive des AFIS) gibt es vereinfacht nur zwei Möglichkeiten: gebündelt oder einzeln. Die Bündelung mehrerer XML-Dateien zu einem Archivpaket erscheint aus technischer Sicht von Vorteil, denn es reduziert massenhafte Daten auf eine leichter handhabbare Menge an Intellektuellen Entitäten. Die archivinterne fachliche Diskussion und anschließende Befragung anderer Fachkolleg:innen

führte schnell in eine andere Richtung: Die fassbare Einheit ist die Person. Jede Person wird einzeln betrachtet und damit einzeln erfasst. Ein Vergleich zur analogen Meldekartei wirkt nur auf den ersten Blick konträr. Die Meldekartei ist als ein einziger Datensatz im AFIS nachgewiesen. Bei den Meldedaten soll jede Person einen eigenen Datensatz erhalten. Doch angenommen, die Meldekartei würde digitalisiert werden, so entspräche jede Karteikarte einem Digitalisat und wäre im Anschluss eine eigenständige Verzeichnungseinheit im AFIS. Der Vorzug der Meldedaten ist, dass eine Erfassung pro Person überhaupt möglich ist. Ein Archivpaket ist eine Person und bezieht sich auf eine XML-Datei. Es wird dabei nicht das Ziel verfolgt, mehrere XML-Dateien zusammenzuführen, die zu einer Person im Laufe der Meldezeit in Leipzig entstehen können (Auflösung von Personenverkettungen, Personarchivierung). Gemäß § 13 Abs. 1 und 2 BMG laufen die Aufbewahrungsfristen gespeicherter Daten bei einer betroffenen Person zu verschiedenen Zeitpunkten aus. Je nach Fristablauf führt das zu unterschiedlichen Nachrichtentypen und Aussonderungszeitpunkten. Die Zusammenführung wird durch die Recherche ermöglicht.

Strukturvarianten:
Die hohe Zahl der Archivpakete muss durch eine geeignete Struktur (oder auch Klassifikation) gegliedert werden. In Anlehnung an die Erfahrungen aus der Meldekartei wurden im Stadtarchiv Leipzig drei Strukturvarianten entwickelt und exemplarisch durch die Leitstelle elKA an zwei Aussonderungsjahren erprobt (2013 und 2016, vgl. Tabelle 1).

1. „Der Klassiker“ nach Nachnamen A-Z:
Die naheliegende Art ist die Strukturierung nach Nachnamen A-Z. Viele Archivaliengruppen mit Personenbezug haben eine alphabetische Grundordnung. Die Meldekartei ist ebenfalls alphabetisch geordnet. Für die Aussonderungsjahre 2013 und 2016 ergab der Strukturierungstest unter der Annahme von 1'000 Archivpaketen, die maximal pro Strukturknoten angelegt werden sollten, eine Anzahl von 157 beziehungsweise von 323 durchzuführenden Ingestläufen.[8]

2. „Die Straßenkartei“ nach letzter Anschrift:
Attraktiv erschien die Struktur nach der letzten gemeldeten Wohnanschrift. Auf einen Blick ließen sich Bewohner:innen einer Meldeadresse überblicken, wenn auch durch stadtinterne Wegzüge niemals vollständig. Vorbild waren die Straßenkarteien aus dem Stadtgebiet Leipzig und einigen eingemeindeten Orten. Bei der Erprobung wurden die Strukturpunkte

[8] Die Annahme von 1.000 Archivpakten beruhte auf einer Empfehlung des AFIS-Herstellers Startext zur Darstellung der Verzeichnungseinheiten im AFIS. Diese Anzahl wurde mittlerweile revidiert, sodass größere und damit weniger Ingestpakete anzunehmen sind. Gleichzeitig besteht auch eine Empfehlung für das IngestTool von 1.000 Objekten je Ingestlauf, wobei es keine feste technische Grenze gibt.

Hausnummerngenau gebildet, was zu einer hohen Anzahl von errechneten Ingestläufen aufgrund der notwendigen Strukturpunkte von 24'689 (2013) und 54'802 (2016) führte. Eine Reduzierung auf Straßennamen hätte die gesetzte 1'000er-Grenze in einigen Fällen überschritten.

3. „Die Schutzfristenhilfe“ nach Geburtsjahr:
Wichtigste Aufgabe bei der Beauskunftung von Meldedaten nach Sächsischem Archivgesetz ist die Prüfung der archivischen Schutzfristen.[9] Eine Strukturierung nach dem Geburtsjahr der Personen ermöglich auf den ersten Blick die Einteilung in „kann beauskunftet werden“ und „kann nicht beauskunftet werden“. Nach und nach könnten bei dieser Strukturvariante ganze Strukturpunkte nach akribischer Prüfung zur selbstständigen Recherche bereitgestellt werden. Die Anzahl der errechneten Ingestläufe liegt mit 357 (2013) und 2'649 (2016) in der Mitte der drei Varianten.

Das Stadtarchiv Leipzig reihte sich am Ende in die Entscheidungen anderer sächsischer Archive ein und entschied sich für den Klassiker nach Nachnamen A-Z. Begründet wurde es fachlich durch die etablierte Form für personenbezogene Archivalien und technisch durch die jeweils zu erwartenden Aufwände der Ingestläufe. Die Durchführung des Ingest bindet schließlich Personalkapazitäten, die wichtigste Ressource im elKA-Verbund.

Recherche, Auskunft und Ausblick
Von Beginn an sollte bei der Meldedatenarchivierung eine ergiebige Recherche im Archiv ermöglicht werden. Mit dem dargestellten Prozess kann der Zugang zu den Meldedaten bereits durch die generierte Konkordanz und den Ingest in DIMAG gewährleistet werden. Die Anschaffung weiterer Software oder Module durch die Archive ist daher nicht notwendig.
Um die elektronische Archivierung zu komplettieren, liegt die Zielsetzung im Stadtarchiv Leipzig auf der Kopplung von AFIS und DIMAG einschließlich des DIMAG-Access-Tools. Die Recherche der archivierten Meldedaten soll im AFIS beginnen und die Anzeige der Primärdatei (XML-Datei) im Access-Tool die Meldeauskunft komplettieren. Seit August 2024 steht dem Stadtarchiv Leipzig eine Teststellung der AFIS-DIMAG-Schnittstelle und der Anbindung des DIMAG-Access-Tool an das AFIS zur Verfügung. Von deren Konfiguration und Weiterentwicklung wird der abschließende Ingest der Meldedaten abhängig gemacht. Dabei wird eingehend die Leistungsfähigkeit im Verhalten mit Massendaten geprüft.

[9] Für die vorliegenden Meldedatensätze ist die Schutzfrist von 100 Jahren nach Geburt der betroffenen oder angehörigen Person gemäß § 10 Abs. 1 Nr. 3b SächsArchivG bei Wegzug oder aufgelösten Personenverkettungen relevant. Die zehnjährige Frist nach Tod einer Person ist für die Sterbefälle sowohl nach der kurzen zehnjährigen Aufbewahrungsfrist nach § 26 Abs. 4 SächsMG als auch nach § 13 Abs. 1 und 2 BMG abgedeckt.

Was wird beauskunftet? Das Archivgesetz schützt große Teile der Meldedaten für viele Jahre. Unproblematisch sind die übernommenen Meldedaten nach Tod einer Person. Auch bei den weggezogenen Personen werden Auskünfte anhand der erfassten Geburtsdaten möglich sein. Die aufgelösten Personenverkettungen bei Volljährigkeit eines Kindes sind kaum nutzbar, müssen aber mindestens für Ermittlungs- und Strafverfolgungsbehörden recherchiert werden können.

Bereits bei Übernahme der Meldedaten war offensichtlich, dass beim Nachrichtentyp Personarchivierung die Aufbewahrungsfrist nach BMG nicht eingehalten wurde. In der Analyse wurde festgestellt, dass lediglich die fünfjährige Transferfrist zur Separierung aus dem aktiven Meldedatenbestand berücksichtigt wird. Das entsprach nicht der Erwartungshaltung der Meldebehörde und des Stadtarchivs. Eine vorfristige Übernahme ist zwar laut BMG möglich, bindet das Archiv jedoch an das BMG und die dort aufgeführten Aufgaben der Meldebehörde.[10] Die angestrebte Trennung zwischen Meldedaten nach BMG und Meldedaten nach Archivgesetz bringt dem Stadtarchiv Leipzig rechtliche Sicherheit. In der Erstübernahme 2023 waren 532'080 XML-Dateien mit dem Nachrichtentyp Personarchivierung enthalten. Alle Wegzüge und Sterbefälle, die auf dem Stadtgebiet bis zum 31.10.2005 erfolgten, waren mit Inkrafttreten des BMG am 1.11.2015 aussonderungsreif. Ab diesem Stichtag gilt für Wegzüge und Sterbefälle die Gesamtaufbewahrungsfrist von 55 Jahren.[11] Eine erste Schätzung reduziert die archivreifen XML-Dateien zur Personarchivierung um über die Hälfte auf etwa 220'000.[12] Die Übrigen werden ab dem Jahr 2060 nach Ablauf der Gesamtaufbewahrungsfrist regulär über VOIS|MESO ausgesondert.

Die Meldebehörde und das Stadtarchiv Leipzig unterstützen zusammen mit der Leitstelle elKA und weiteren sächsischen Kommunen die Verbesserung der Aussonderung aus VOIS|MESO, um einen resilienten, transparenten, protokollierten und fristgerechten Aussonderungsprozess im Meldeverfahren zu erreichen. Ein Schwerpunkt ist dabei die Einführung eines technischen Vollständigkeitsabgleichs zwischen den auszusondernden und tatsächlich übergebenen Daten.

Fazit zur Meldedatenarchivierung im Verbund

Es ist ein Trugschluss zu glauben, dass die Herausforderungen bei der Meldedatenarchivierung kleiner werden. Im Gegenteil, denn desto mehr Erfahrungen zusammengetragen werden, umso

[10] Vgl. § 16 Abs. 2 BMG.
[11] Schreiben des Sächsischen Städte- und Gemeindetags vom 8.12.2014 zur Klärung der Aufbewahrungsfristen mit Inkrafttreten des Bundesmeldegesetzes.
[12] Eine abschließende Zahl konnte noch nicht ermittelt werden, da alle Aussonderungsjahre (vgl. Tabelle 1) Altfälle enthalten. Aus dem umfangreichen Aussonderungsjahr 2016 sind z. B. rund 50.000 Personarchivierungsdatensätze archivreif.

mehr Besonderheiten und neue Fragen treten auf. Meldedatenarchivierung ist und bleibt komplex. Bisher konnten im Verbund 16 sächsische Kommunen bei der Archivierung ihrer Meldedaten unterstützt werden, wobei sieben davon zum Abschluss gebracht wurden. Bei den noch offenen Vorgängen sind entweder inhaltliche, organisatorische, strukturelle, rechtliche oder technische Probleme aufgetreten. Der vorgestellte Bearbeitungsprozess löst nur einen Teil der Arbeitsschritte zur Archivierung von Meldedaten, wie die Datenanalyse und die Aufbereitung der Ingestpakete.

Trotzdem zeigt die Zusammenarbeit im Verbund an dieser Stelle ihre Stärken: Das Teilen von Wissen und Erfahrungswerten führt zu kooperativer Kompetenz, von der alle Kooperationspartner profitieren. Wechselseitige Informationsflüsse und die Steuerung und Normierung der Abläufe reduzieren die Risiken in der Datenverarbeitung und tragen letztlich zur Qualitätssicherung und Weiterentwicklung der Archivierungsprozesse bei.

Aus Perspektive des Archivs wird durch den Verbund eine Gemeinschaft erzeugt, die im lernenden Arbeitsfeld der elektronischen Archivierung unterstützend wirkt. Selbstredend wird dem Archiv vor Ort Arbeit abgenommen und der gesamte Prozess der Archivierung beschleunigt, insofern Entscheidungen zeitnah getroffen und alle Rahmenbedingungen geklärt sind. Die Normierung der Überlieferungsbildung bei Meldedaten sorgt für Nachhaltigkeit. Der vielseitige Gesamtprozess der Meldedatenarchivierung ist richtungsweisend für das zukünftige Wirken des elektronischen Kommunalarchivs Sachsen.

Bibliografie

Worm, Peter, 2016, „Was ändert sich mit dem Bundesmeldegesetz für die Archive in NRW?“, in: *archivamtblog – Neues aus dem Archivwesen in Westfalen-Lippe*, https://archivamt.hypotheses.org/4265 (22.08.2024).

10 Fachverfahren, 3 E-Akten-Systeme und 1 Aussonderungslösung: Zur bevorstehenden bundesweiten Überlieferung der E-Akten der Justiz

Bastian Gillner

In den meisten Archiven stellen mittelalterliche Urkunden die ältesten Stücke und die früheste Überlieferungsschicht dar. Urkunden sind Schriftstücke, die Recht schaffen, Recht beweisen und Recht verkörpern (vgl. Hochedlinger, 2009, S. 25). Sie dokumentieren Entscheidungen über Privilegien und Besitz, auch und gerade im Fall von verletzten Rechten, von widerstreitenden Interessen und von kleinen wie großen Konflikten. Heute, tausend Jahre später, haben die Archive ein ungebrochenes Interesse an der Überlieferung von normierenden und konfligierenden Rechten, zeigt sich hier doch das Aushandeln von soziopolitischen wie privaten Machtverhältnissen und das Funktionieren bzw. Nicht-Funktionieren von gesellschaftlichen Mikro- und Makrostrukturen aller Art.

In einem historischen Fundamentalprozess der Rechtsstaatsbildung sind neben zivilrechtliche Entscheidungen über Recht und Besitz auch strafrechtliche Sanktionen für deviantes Verhalten und vielfältige fachgerichtliche Entscheidungen für Unmengen von fachlichen Einzelfragen getreten. Sie alle liegen in massenhafter Form bei Gerichten und Justizbehörden vor und es wird nicht mehr lange dauern, bis sie alle nur noch in rein digitaler Form existieren. Nicht mehr mittelalterliches Pergament oder neuzeitliches Papier wird das Medium der Justizüberlieferung sein, sondern die elektronische Akte auf einem digitalen Speicher. Denn die digitale Transformation, die nunmehr auch schon seit einem Vierteljahrhundert ihre volle Wirkmacht entfaltet, hat natürlich auch vor der Justiz nicht haltgemacht.

Maßgeblich sind für die gesamte bundesdeutsche Justiz hier das Gesetz zur Förderung des elektronischen Rechtsverkehrs von 2013 sowie das Gesetz zur Einführung der elektronischen Akte in der Justiz und zur weiteren Förderung des elektronischen Rechtsverkehrs von 2017. E-Justice-Gesetz wird ersteres Gesetz gemeinhin genannt und E-Justice ist das Schlagwort, unter dem die digitale Transformation hier subsumiert wird: Alle justiziellen Verfahrensabläufe und jeglicher Rechtsverkehr sollen elektronisch ablaufen. E-Justice ist damit das Äquivalent der dritten Staatsgewalt zum E-Government, also kurz gesagt die Überführung staatlichen Handelns in das digitale Zeitalter. Und E-Justice ist dabei keineswegs mehr Zukunftsmusik: Äußerst nüchtern formuliert das zweite der genannten Gesetze: „Weitere Änderung [...] zum 1. Januar

2026: [...] Die Akten werden elektronisch geführt.“ Somit muss die gesamte Justiz mit dem beginnenden Jahr 2026 elektronische Akten führen. Diese Frist ist als spätester Zeitpunkt zu verstehen, tatsächlich arbeiten gegenwärtig bereits viele Gerichte oder auch schon ganze Gerichtsbarkeiten mancher Länder elektronisch. Auch wenn die letzten Papierakten mit längerer Aufbewahrungsfrist noch über Jahre hinweg in die Archive tröpfeln werden, so ist vollkommen klar, dass die Justizüberlieferung zum Ende der aktuellen Dekade eine rein digitale Form haben wird. Nicht nur für die Justiz, auch für die Archive bedeutet diese Transformation einen gewaltigen Epochenbruch. Wie die Archive diese Herausforderung zu meistern gedenken und wie die zukünftige Überlieferungsbildung mittels E-Akten der Justiz aussehen soll, das möchte folgender Beitrag skizzieren. Verpflichtet fühlt er sich dabei der frühen Pionierarbeit, die Bernhard Grau mit seinen Überlegungen zur Nachnutzung des Kommunikationsstandards XJustiz für die Aktenaussonderung auf der Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“ (AUdS) 2013 geleistet hat – bemerkenswerte zehn Jahre vor dem tatsächlichen Start in die praktische Projektarbeit von Archiven und Justiz zur Aussonderung von E-Akten (vgl. Grau, 2014).

„Die Akten werden elektronisch geführt“, so verlangt es der Gesetzgeber im genannten gleichnamigen Gesetz von der Justiz. Dieser schlichte Satz hat eine enorme Tragweite, denn die Akte, genauer gesagt die Verfahrensakte, ist fest im normativen Regelwerk der Justiz verankert: „Jeder Geschäftsvorgang erhält ein Aktenzeichen, unter dem alle dazugehörigen Dokumente [...] zu führen sind. [...] Zu einem Geschäftsvorgang gehören alle Anträge, Erklärungen, Handlungen und Entscheidungen, die [...] eine Angelegenheit betreffen, mit der das Gericht oder die Staatsanwaltschaft befasst ist [...]“, so formuliert etwa die gemeinsame Aktenordnung der Landesjustizverwaltungen (Aktenordnung, 2024, § 2). Die Akte ist damit zentrales Instrument im Prozess der Rechtsprechung, sie trägt alle Informationen zu Prozessführung und Urteilsfindung. Erinnert sei an den bekannten vormodernen Rechtssatz „Quod non est in actis, non est in mundo“ – was nicht in den Akten ist, existiert nicht. Für die Justiz ist das keineswegs eine unverbindliche historische Referenz, es ist nach wie vor ein gültiges Prinzip von gewichtiger Grundsätzlichkeit. Allein Informationen in der Akte haben Relevanz, andernorts abgelegte Informationen sind nicht Teil der Rechtsprechung.

Somit ist und bleibt das zentrale Informationsobjekt der Justiz auch im digitalen Zeitalter die (E-)Akte. Diese fundamentale Tatsache gilt es insbesondere vor der Entwicklung zu betonen, die die (E-)Akte der Verwaltung erfahren hat. Als mit der beginnenden digitalen Transformation auf einmal Dateisysteme, Mailsysteme, Fachverfahren und viele andere IT-Systeme in der Verwaltung Einzug hielten, erfuhr die dortige Aktenführung einen massiven Substanzverlust: Das

einstmals einheitliche Informationsobjekt Akte zerfaserte, ja zerfledderte, in unterschiedliche nachgeordnete Ordnungssysteme (vgl. Ernst, 2017; Gillner, 2019 und 2024; Unger/Schmalzl, 2020). Da eine solche Fragmentierung der Akte aber weder rechtskonform noch praktisch ist, muss die Verwaltung aktuell hart um eine ordentliche Schriftgutverwaltung ringen (vgl. Bischoff, 2018; Gillner, 2019; Schlemmer, 2025). Die Justiz hingegen hat glücklicherweise das sinnvolle Prinzip der Aktenführung niemals verlernt. Nicht zuletzt für die Archive und ihre Konzeption einer Aussonderungslösung ist das von hohem Wert.

Im Vorlauf zur Aussonderungsthematik verdienen zwei Spezifika der E-Akte der Justiz noch eine einleitende Erwähnung: Erstens verdeckt die Rede von „der“ E-Akte der Justiz die Tatsache, dass ebendiese E-Akte der Justiz kein einheitliches System ist. Vielmehr sind es drei unterschiedliche E-Akten-Lösungen, die in den Justizverwaltungen von Bund und Ländern im Einsatz sind. Sie tragen die Namen e2A (elektronischer ergonomischer Arbeitsplatz), eAS (E-Akte als Service) und eIP (elektronisches Integrationsportal). Die Verteilung dieser drei Systeme über die Länder hinweg folgt keiner erkennbaren Logik; was warum wo eingesetzt wird, liegt irgendwo in den Tiefen des deutschen Föderalismus begründet (vgl. EDV-Länderbericht, 2023). Allerdings sind mit diesen drei E-Akten-Lösungen dann auch alle Gerichte in Deutschland versorgt (bzw. werden nach vollständigem Rollout versorgt sein).

Zweites kennt die Justiz, bei aller Aktenförmigkeit ihrer Arbeit, doch auch Fachverfahren (vgl. Dässler/Schwarz, 2010; Naumann, 2010; Keitel, 2011; Raselli, 2014; Pilger, 2015; Habersack, 2015; Steffenhagen, 2019; Gillner, 2023; Holzapfl u. a., 2023; Jacobs, 2023). Mehr noch, sie kennt sie nicht nur, sie hat eine Vielzahl von ihnen mit sehr zentralen Funktionalitäten im alltäglichen Masseneinsatz. Diese Fachverfahren waren deutlich eher im Einsatz als die E-Akte und sie haben bereits mit der analogen Verfahrensakte zusammengewirkt, denn sie dienen der Verfahrensorganisation in der Justiz: Überblick über alle laufenden Verfahren, Haltung von Stammdaten, Zuweisung von Personen, Terminen und Räumen etc. Entstanden sind diese Fachverfahren in den 1990/2000er Jahren, als alle Bereiche der Justiz merkten, dass ihre Arbeit durch digitale Instrumente deutlich erleichtert und effektiviert werden konnte. Typisch für diese Zeit war dann aber auch das Schaffen von Einzellösungen statt dem Anstreben einer übergreifenden Gesamtkonzeption; jeder Bereich schuf sein eigenes Fachverfahren. So nutzt die Zivilgerichtsbarkeit EUREKA, forumSTAR oder JUDIKA, die Strafgerichtsbarkeit MESTA oder web.sta, die Fachgerichtsbarkeiten EUREKA-Fach, FOKUS, JUSTUS oder VG/FG. Die Justiz arbeitet zwar gegenwärtig an einer einheitlichen integrierten Gesamtlösung (dem gemeinsamen Fachverfahren GeFa), aber deren flächendeckender Produktivbetrieb wird wohl nicht mehr in dieser Dekade erreicht werden. Realität ist also gegenwärtig eine heterogene

Fachverfahrenslandschaft aus ca. 10 unterschiedlichen Systemen mit bunter Streuung über Gerichtsbarkeiten und Landesgrenzen hinweg.

Zentrale Frage ist nun, wie es die Archive schaffen können, aus dieser heterogenen Systemlandschaft eine vernünftige Überlieferung zu bilden. Als wäre eine integrierte Lösung für drei E-Akten-Verbünde und zehn Fachverfahrens-Verbünde noch nicht kompliziert genug, so müssen sie auch noch ihre eigenen unterschiedlichen Archiv-Systeme in eine funktionsfähige Aussonderungs- und Archivierungslösung einbinden. Die Komplexität der Aufgabe war den Archiven früh bewusst. Etwa seit dem Start der ersten E-Akten-Piloten 2016 hatten die staatlichen Archive Deutschlands das Thema Justiz-E-Akte im Blick (vgl. Schweizer, 2021). Da es sich bei den Verbünden um länderübergreifende IT-Systeme handelt, waren es weniger Bundes- und Landesarchive, die als Einzelakteure tätig wurden, sondern vielmehr der Ausschuss Records Management der Konferenz der Leiterinnen und Leiter der Archivverwaltungen von Bund und Ländern (KLA). Dieser Ausschuss hat die Aufgabe, elektronische Systeme im länderübergreifenden Einsatz zu identifizieren, archivrelevante Inhalte und aussonderungsfähige Datenobjekte zu bestimmen sowie die Schaffung von Aussonderungsschnittstellen und -workflows zu begleiten. Auch übernimmt er den Austausch mit länderübergreifenden Gremien auf Behördenseite. Entsprechend suchte der Ausschuss nach Ansprechpartnern in der Justiz und fand sie in der AG Einheitlicher Strukturdatensatz, einer Unter-AG der AG IT-Standards der Bund-Länder-Kommission für Informationstechnik in der Justiz. Die föderale Gremienstruktur mutet hier unübersichtlich an, inhaltlich hatten beide Seiten hier aber genau den richtigen Ansprechpartner gefunden und konnten beginnen, ein Verständnis für die jeweiligen Anliegen zu finden (vgl. Grau, 2021).

Eine Aussonderungslösung, wie auch immer sie im Detail aussehen sollte, so machte die Justiz früh deutlich, muss auf einem Kernelement aufsetzen; und das ist der Austauschstandard XJustiz. XJustiz gehört zum XÖV-Rahmenwerk, einer Sammlung von Standards für den elektronischen Datenaustausch der öffentlichen Verwaltung auf der Basis von Nachrichten in XML-Syntax und zugehörigen Prozessen (vgl. Ernst, 2021; Hoppenheit/Schmidt, 2021). XJustiz ist der Standard, mit dem die Justiz *sämtliche* Kommunikationsprozesse abwickelt, sei es die Einreichung eines Anwaltsschreibens an ein Gericht, die Übermittlung eines Bußgeldverfahrens von einer Polizeibehörde an eine Staatsanwaltschaft, die Eintragung in ein Handels- oder sonstiges Register, die Übermittlung einer Fahndungsmitteilung oder die Abgabe einer ganzen Verfahrensakte von einem Gericht an ein anderes. Diese und noch viele, viele Kommunikationsszenarien mehr bringt der XJustiz-Standard in eine maschinenlesbare XML-Form, d.h. er definiert, welche Metadaten für welchen Zweck übermittelt werden müssen. Entsprechend ist dieser

Standard auch nicht ganz klein, die Spezifikation der aktuell gültigen Version 3.4.1 umfasst 1070 Textseiten. In diese Kommunikationsszenarien gemäß XJustiz-Standard muss und wird sich auch die Aussonderung von Akten an die Archive einfügen.

Somit bestehen also drei wesentliche Komponenten, die die Aussonderungslösung für E-Akten der Justiz maßgeblich bestimmen:

1. die E-Akte (die die Dokumente und deren Metadaten enthält)
2. das Fachverfahren (das die fachlichen Metadaten führt)
3. der XJustiz-Standard (der die Datenstruktur vorgibt).

Aufbauend auf diesen Komponenten haben Archive und Justiz nun mit der Konzeption einer Aussonderungslösung begonnen: In einem ersten Schritt schufen Archive und Justiz eine Reihe von Aussonderungsnachrichten, die sich die Nachrichtenfolge des xdomea-Standards zum Vorbild nehmen. Die Justiz wird also eine Anbietungsnachricht mit den Metadaten aller auszusondernden Akten an das zuständige Archiv schicken. Das Archiv wird jede Akte mit „A[rchivieren]“ oder „V[ernichten]“ bewerten und diese Information mit einer Bewertungsnachricht an die Justiz zurücksenden. Dann übersendet die Justiz alle als archivwürdig bewerteten Akten an das Archiv, d.h. alle Primärdaten und alle Metadaten. Das Archiv bestätigt schließlich die erfolgreiche Übernahme in das digitale Magazin mit einer Importbestätigung. Die archivwürdigen Akten liegen damit im Archiv vor und werden dann bei der Justiz gelöscht.

Interessanter als dieser von der Verwaltungs-E-Akte bekannte Workflow ist nun aber der Aufbau der XJustiz-Nachrichten. Im Mittelpunkt steht hier die Nachricht Justiz-zu-Archiv, die als Anbietungs- oder als Aussonderungsnachricht profiliert sein kann und sich im Wesentlichen darin unterscheidet, ob Primärdaten mitgeliefert werden. Diese Nachricht Justiz-zu-Archiv trägt eine erhebliche Menge von Metadaten, die wichtige Informationen für das Data Management im Archiv darstellen. Auf oberster Ebene setzt sich die Nachricht aus dem Nachrichtenkopf und den Fachdaten zusammen. Der Nachrichtenkopf ist im Wesentlichen für technische Attribute wie Version oder Nachrichten-ID wichtig, archivisch interessanter sind die Fachdaten. Unter den Fachdaten liegt das Aussonderungsobjekt, d.h. die einzelne Verfahrensakte. Zusammengesetzt wird das Aussonderungsobjekt wiederum aus den Schriftgutobjekten und den Fachdaten Aussonderung. Unter den Schriftgutobjekten finden sich alle Informationen über die Akte und die enthaltenen Teilakten und/oder Dokumente. Hier sind Aktenzeichen, Aktentyp, Laufzeit u. ä. Metadaten ablesbar. Unter den Fachdaten Aussonderung hingegen finden sich alle Informationen zum vorliegenden Verfahren, also die Beteiligten, der Verfahrensgegenstand, das Sachgebiet u. ä. Auch wenn die Archive vielleicht zehn oder zwanzig zentrale Metadaten für

ihre Zwecke benötigen, so stecken doch erheblich mehr Informationen in jeder Nachricht, insgesamt eine knapp dreistellige Zahl an Metadaten.

Die XJustiz-Nachricht mit ihren Informationen zu Akte *und* Verfahren verschleiert nun ein wenig ein Spezifikum der E-Akten-Führung in der Justiz. Die E-Akte und das Fachverfahren sind nämlich zwei relativ unabhängig nebeneinander bestehende Systeme. Im Fachverfahren werden die Gerichtsverfahren verwaltet, in der E-Akte hingegen die Dokumente abgelegt. In der alltäglichen Justizpraxis gibt es keine unbedingte Verbindung zwischen beiden Komponenten. Die Geschäftsstelle, die Zuständigkeiten, Personaleinsätze und Verhandlungstermine organisiert, benötigt dafür keine Inhalte aus der E-Akte. Der Richter wiederum, der seine Fälle bearbeitet, ist dafür kaum an dem Fachverfahren interessiert. Erst, wenn ein Verfahren und die dazugehörige Akte abgegeben werden, etwa an ein anderes Gericht oder eben an ein Archiv, wird es notwendig, die inhaltlichen Daten aus der Akte und die beschreibenden Daten aus dem Fachverfahren miteinander zu verbinden. Um eine Aussonderung zu ermöglichen, muss eine technische Funktionalität also Daten aus dem E-Akten-System und dem Fachverfahren zusammenbringen. Mittlerweile hat die Justiz die Schaffung einer solchen Funktionalität unter dem Namen XJustiz-Merger zugesagt, für die Archive ist das eine absolut notwendige Komponente, um das archivisch benötigte Gesamtpaket an Metadaten zu erhalten. Bis dieser XJustiz-Merger produktiv geht, wird noch ein wenig Zeit vergehen und den ersten Übernahmen werden in manchen Ländern somit noch die Fachverfahrensdaten fehlen. Bewertung und Erschließung nur mit Aktenzeichen und Laufzeit, aber ohne Beteiligte und Delikt wird eine temporäre Herausforderung darstellen, die der Übergangszeit geschuldet ist.

Diese Eigenheiten der Justiz-E-Akte sind Gegenstand der aktuellen Zusammenarbeit von Archiven und Justiz. Gespräche finden seit 2018 regelmäßig statt und haben beide Seiten von der Notwendigkeit einer gemeinsamen Aussonderungslösung überzeugt, die schließlich auch in einem Konzept festgeschrieben werden konnte. Angesichts der Vielzahl der Akteure ist das durchaus kein geringer Erfolg. Intensiviert wurde die Zusammenarbeit 2023 mit der Initiierung eines gemeinsamen Umsetzungsprojekts von KLA und BLK, das die Realisierung der Aussonderungslösung vorantreibt. Unter der Leitung des Projektbüros der BLK-AG IT-Standards finden gegenwärtig monatliche Treffen von Vertretern der Archive, der E-Akten-Verbünde und der Fachverfahrens-Verbünde statt. In diesem Rahmen sind zwei wesentliche Teilprojekte definiert worden, nämlich das Teilprojekt Datenaustauschformat mit allen Fragen zum XJustiz-Standard und den dortigen Kommunikationsszenarien, und das Teilprojekt Übertragungsweg mit allen Fragen zum konkreten Datenfluss von den Gerichten in die Archive.

Das Teilprojekt Datenaustauschformat befasst sich primär mit den XJustiz-Nachrichten zur Anbietung und Aussonderung. Diese Nachrichten existieren bereits seit der Version 3.0 aus dem Jahr 2019 im XJustiz-Standard und sind (zu einem größeren Teil) aus generischen Komponenten einerseits und (zu einem kleineren Teil) aus spezifischen Komponenten andererseits zusammengesetzt. Die generischen Komponenten finden auch Anwendung in ganz anderen Kommunikationsszenarien, die die Justiz mit XJustiz abwickelt. Beispielsweise sind die sogenannten Instanzdaten mit dem Aktenzeichen, dem Sachgebiet, dem Verfahrensgegenstand u. ä. Bestandteil der Anbietungs- und Aussonderungsnachricht, aber ebenso natürlich bei einer Vielzahl weiterer Anwendungsfälle. Aufgrund dieser Vielgestaltigkeit der Anwendungsfälle sind die einzelnen Elemente der Nachrichten aber zumeist als mögliche Felder definiert, nicht als Pflichtfelder. Da der Aussonderungsprozess bzw. die archivische Weiterverarbeitung aber bestimmte Elemente unbedingt benötigt, wurden im Teilprojekt Datenaustauschformat die Nachrichten für die beiden Anwendungsfälle Anbietung und Aussonderung profiliert. Aus der großen Menge aller Elemente wurden bestimmte Elemente als Pflichtfelder definiert, die in der Nachricht vorkommen müssen. Damit ist gewährleistet, dass alle benötigten Informationen für Bewertung, Erschließung und Recherche an die Archive gelangen (und auch automatisiert weiterverarbeitet werden können).

Zur Profilierung zählte aber nicht nur die Bestimmung von Pflichtfeldern, sondern auch die Aufnahme von spezifischen Elementen in die Nachrichten. Gerade im Bereich spezieller Rechtsgebiete sind die Archive auf Metadaten angewiesen, die nicht in den generischen Komponenten des Standards enthalten sind. Beispielsweise werden über die Nachrichten auch Handelsregisterverfahren ausgesondert werden. Für Archive sind hier Metadaten wie Gründungsdatum, Geschäftszweck oder Kapital der jeweiligen Unternehmen interessant, diese sind aber nicht in den generischen Komponenten der Nachrichten enthalten. Hier wurden in den Anbietungs- und Aussonderungsnachrichten eigene Elemente geschaffen, die diese Metadaten transportieren. Damit sollten für alle Anwendungsfälle – die angesichts der Größe der Justiz nicht wenige sind – die archivischerseits benötigten Metadaten definiert sein.

Im Detail hat die Profilierung zu zahlreichen Folgethemen geführt, die ebenfalls im Teilprojekt behandelt wurden. Beispielhaft sei an dieser Stelle der Umgang mit Codes in XJustiz genannt. XJustiz transportiert an verschiedenen Stellen Kodierungen statt Klartext, so dass die Metadaten bspw. den Wert „60“ aufführen statt „Verstoß gegen das Betäubungsmittelgesetz“ oder „115“ statt „Nebenkläger(in)“. Die Informationen in XJustiz sind also nicht unmittelbar verständlich, was für die Archive in der Übernahme, insbesondere aber mit Blick auf die langfristige Verständlichkeit in der Nutzung ein erhebliches Problem darstellt. Glücklicherweise hat

hier die Justiz die archivischen Bedenken ernst genommen und eine ergänzende Übermittlung von Klartext zugesagt.

Die Brücke zwischen dem Teilprojekt Datenaustauschformat und dem Teilprojekt Übertragungsweg bilden die Aussonderungsprozesse, also die konkreten Workflows beim Ablauf einer Aussonderung. Die Kernprozesse der Aussonderung – Anbietung, Bewertung, Aussonderung, Importbestätigung – sind über die genannten XJustiz-Nachrichten definiert, doch weitere Prozesse harren noch ihrer Ausgestaltung. Insbesondere ist hier die Aktenautopsie im Rahmen der Bewertung zu nennen. Die Anbietungsnachricht transportiert einen umfangreichen Metadatensatz, der eine weitgehende Bewertung ermöglicht. Allerdings ist eine archivische Bewertungspraxis auf alleiniger Basis von Metadaten noch keine Realität, selbst wenn diese – wie in manchen Landesarchiven geplant – mit ergänzenden Informationsressourcen wie der Gemeinsamen Normdatei GND oder Rechtsprechungsdatenbanken abgeglichen werden (vgl. Naumann, 2020). In letzter Konsequenz kann immer der fachliche Bedarf der bewertenden Archivarinnen und Archivaren bestehen, in eine Akte hineinschauen zu müssen.

Zwischen den XJustiz-Szenarien „Anbietungsverzeichnis übergeben“ und „Bewertungsverzeichnis übergeben“ muss also ein Prozess zur Aktenautopsie stehen. Hier könnte das bundeseinheitliche Akteneinsichtsportal der Justiz eine Rolle spielen, aber wie dieser Prozess ausgestaltet werden kann, ist zum gegenwärtigen Zeitpunkt noch unklar. Die andere große Unbekannte bei der Aussonderung ist daneben eine Löschroutine in den Fachverfahren. Anscheinend löschen die Fachverfahren bestimmte Inhalte nach Fristen, die nicht deckungsgleich mit den Aufbewahrungsfristen der Akten sind. Entsprechend stehen diese Metadaten zum Zeitpunkt der Aussonderung nicht mehr zur Verfügung und können vom XJustiz-Merger nicht in eine Anbietungsnachricht geschrieben werden. Sollten die betroffenen Metadaten eine Relevanz für die Archive haben, müsste eine Aussonderung vorfristig vor der ersten Löschfrist erfolgen. Übliche Aussonderungspraktiken würden damit eine spürbare Modifikation erfahren. Die Projektgruppe muss hier also (Rahmen-)Prozesse definieren, die über einen Nachrichtenaustausch und eine Datenübertragung hinausgehen.

Ähnliche Grundsatzarbeit verlangt schließlich auch das Teilprojekt Übertragungsweg. Um die Daten – die Primärdaten aus der E-Akte wie die Metadaten aus den Fachverfahren – von der Justiz zu den Archiven zu bekommen, hat die Justiz die Nutzung ihrer eigenen Datenaustausch-Infrastruktur vorgeschlagen, nämlich das elektronische Gerichts- und Verwaltungspostfach (EGVP). In dieser Infrastruktur gibt es auch eine archivischerseits nutzbare Behördenkomponente, nämlich das besondere Behördenpostfach (beBPo). Da die Justiz sämtliche

Kommunikation über die EGVP-Infrastruktur abwickelt, liegt es nahe, hierüber auch die Aussonderungen laufen zu lassen. Details sollen dann im nächsten Projektschritt geklärt werden.
Als zentrale Frage wird hier sicherlich das Massenproblem in den Mittelpunkt treten. Das gilt zum einen für die technische Leitungsfähigkeit des EGVP. Die Justiz betreibt das EGVP im bundesweit flächendeckenden Echteinsatz, entsprechende Erfahrungswerte liegen also vor, auch wurden bereits Lasttests gefahren. Allerdings sind die archivischen Bedarfe doch eine eigene Dimension für sich. Aktenabgaben mögen im Justizalltag eine Normalität sein, die das EGVP problemlos schultert, bei einer Anbietung jedoch werden *alle* Akten eines (Erledigungs-)Jahrgangs in Metadatenform übermittelt werden müssen. Fünfstellige Aktenmengen werden hier Normalität sein, bei großen Gerichten und Staatsanwaltschaften gar sechsstellige Zahlen. Die Übernahmen nach der Bewertung werden demgegenüber zwar deutlich geringer sein, aber auch Übernahmen von vielleicht hundert archivwürdigen Akten dürften quantitativ alle Aktenabgaben des Justizalltags in den Schatten stellen. Die Aussonderung wird also ganz eigene Ansprüche an die Leistungsfähigkeit des Gesamtsystems stellen.
Das Massenproblem gilt zum anderen aber auch für die Bedienbarkeit des beBPo. Ähnlich einem Maileingang werden dort massenhaft Anbietungs- und Aussonderungsnachrichten auflaufen. Diese Nachrichten müssen ihrer Herkunft und ihrer Aufgabe nach gefiltert und sortiert werden. Händisch dürfte diese Aufgabe im Echtbetrieb kaum möglich sein. An das beBPo müssen also Funktionalitäten gekoppelt werden, die Nachrichten automatisiert weiterverarbeiten, idealerweise bereits im Zusammenspiel mit den Ingestfunktionalitäten der digitalen Archivsysteme. Die Projektgruppe muss hier den Wohlfühlraum der archivischen Fachlichkeit verlassen und sich harten IT-Themen widmen; ohne entsprechende Unterstützung der archivischen IT-Einrichtungen wird das wohl nicht funktionieren.
Was bleibt als Fazit? Im Bereich der Justiz werden die staatlichen Archive tatsächlich eine einheitliche Aussonderungslösung über alle Bundes-, Länder-, Verbund- und Fachlichkeitsgrenzen hinweg schaffen. Der Druck der Justiz, einheitliche und standardkonforme Prozesse zu realisieren, hat die Archive zu diesem Unikum bewogen. Mit dem gemeinsamen Umsetzungsprojekt von BLK und KLA verfügen die Beteiligten über das passende Instrument, um dieses Ziel zu erreichen. Die Arbeiten sind fortgeschritten und haben erste gute Lösungen wie etwa die Profilierung der Anbietungs- und Aussonderungsnachrichten erbracht. Antworten auf weitere Fragen, insbesondere zu Aussonderungs- und Übertragungsprozessen, müssen noch gefunden werden. Viel Zeit bleibt hierfür nicht mehr, denn das Jahr 2026 zur flächendeckenden Realisierung der elektronischen Aktenführung in der Justiz rückt näher. In ersten Bundesländern haben die ersten Pilotgerichte bereits aussonderungsfähige Akten vorliegen. Von Jahr zu Jahr wird deren

Anzahl nun deutlich steigen. Die praktische Alltagsarbeit mit den E-Akten der Justiz liegt also auch für die Archive nicht mehr in weiter Ferne. Für Fachlichkeit und Technik wird dieser Umstieg eine große Herausforderung. Doch die skizzierte Aussonderungslösung wird für alle staatlichen Archive in Deutschland den Weg bieten, diese praktische Arbeit erfolgreich zu gestalten.

Bibliografie

Aktenordnung für die Gerichte der ordentlichen Gerichtsbarkeit und Staatsanwaltschaften (AktO) (2024).

Bischoff, Frank M. (2018), 'E-Government und Records Management als Kernkompetenz und Beratungsaufgabe öffentlicher Archive: Zur Beteiligung des Landesarchivs Nordrhein-Westfalen bei der Einführung der elektronischen Verwaltung in Landesbehörden', in: Maier, Gerald (Hrsg.), *Archive heute – Vergangenheit für die Zukunft. Archivgut – Kulturerbe – Wissenschaft.: Zum 65. Geburtstag von Robert Kretzschmar*, Stuttgart: Kohlhammer, S. 123-139.

EDV-Länderbericht (2023), https://justiz.de/laender-bund-europa/BLK/laenderberichte/index.php (30.9.2024).

Dässler, Rolf / Schwarz, Karin (2010), 'Archivierung und dauerhafte Nutzung von Datenbankinhalten aus Fachverfahren: Eine neue Herausforderung für die digitale Archivierung', *Archivar* 63, S. 6-18.

Ernst, Katharina (2017), 'Welche Zukunft hat die Akte?', in: Storm, Monika (Red.), *Transformation ins Digitale* (Tagungsdokumentationen zum Deutschen Archivtag 20), Fulda: Selbstverlag des VdA, S. 67-75.

Ernst, Katharina (2021). 'Standards und Normen im Bereich der Langzeitarchivierung', *Archivar* 74, S. 62-70.

Gesetz zur Förderung des elektronischen Rechtsverkehrs (2013).

Gesetz zur Einführung der elektronischen Akte in der Justiz und zur weiteren Förderung des elektronischen Rechtsverkehrs (2017).

Gillner, Bastian (2019), 'Good Governance als Kollateralnutzen oder: Wie Archive mit der E-Akte Verwaltungshandeln und Überlieferung verbessern können', in: Herrmann, Tobias (Hrsg.): *Verlässlich, richtig, echt – Demokratie braucht Archive!* (Tagungsdokumentationen zum Deutschen Archivtag 23), Fulda: Selbstverlag des VdA, S. 39-50.

Gillner, Bastian (2023), 'Überlieferungsbildung aus Fachverfahren: Herausforderungen im archivischen Vorfeld', *Archiv theorie & praxis* 76, S. 6-14.

Gillner, Bastian (2024), 'Aktenführung, Vorgangsbearbeitung, Datenhaltung: Digitale Verwaltungspraxis als Ausgangspunkt für eine digitale Quellenkritik', in: Becker, Irmgard Christa u. a. (Hrsg.), *Archivists meet Historians: Transferring source criticism to the digital age* (Veröffentlichungen der Archivschule Marburg 71), Marburg: Archivschule Marburg, S. 15-33.

Grau, Bernhard (2014), 'XJustiz: Überlegungen zur Nachnutzung des Kommunikationsstandards der Justiz für die Aktenaussonderung', in: Nolte, Burkhard (Red.), *Standards, Neuentwicklungen und Erfahrungen aus der Praxis zur digitalen Archivierung: 17. Tagung des Arbeitskreises 'Archivierung von Unterlagen aus digitalen Systemen' am 13. und 14. März 2013 in Dresden*, Halle (Saale): Mitteldeutscher Verlag, S. 89-97.

Grau, Bernhard (2021), 'Die Aussonderung elektronischer Unterlagen der Justiz: Herausforderungen der Zusammenarbeit im Dschungel Bund-Länder-übergreifender Verfahrenspflegestellen, Gremien und Verbünde', in: Becker, Irmgard Christa u. a. (Hrsg.), *E-Government und digitale Archivierung* (Veröffentlichungen der Archivschule Marburg 67), Marburg: Archivschule Marburg, S. 233-253.

Habersack, Michael u. a. (2015), 'Erste Schritte bei der Bewertung elektronischer Fachverfahren: Eine Handreichung für kommunale Archive', in: *Kooperation ohne Konkurrenz. Perspektiven archivischer Kooperationsmodelle: 48. Rheinischer Archivtag 2014* (Archivhefte 45), Bonn: Habelt, S. 220-229.

Hochedlinger, Michael (2009), *Aktenkunde: Urkunden- und Aktenlehre der Neuzeit*, Wien: Böhlau.

Holzapfl, Julian u. a. (2023), 'Quick Wins und dicke Bretter: Übernahme und Archivierung von Fachverfahren', *Archiv theorie & praxis* 76, S. 15-24.

Hoppenheit, Martin / Schmidt, Christoph (2021), 'xdomea und die Archive: Fragen und Antworten', *Archivar* 74, S. 71-75.

Jacobs, Rainer (2023), 'Effiziente Verfahren, echte Daten: Die Übernahme von Informationen aus Fachverfahren in das Bundesarchiv?, *Archiv theorie & praxis* 76, S. 25-27.

Keitel, Christian (2011), *Eine andere Art der Dokumentation. Anmerkungen zur Bewertung umfassender Informationssysteme*, https://www.landesarchiv-bw.de/sixcms/media.php/120/52529/Workshop_Keitel_andere_Art.pdf (30.9.2024).

Naumann, Kai (2010), 'Übernahme von Daten aus Fachanwendungen: Schnittstellen, Erhaltungsformen, Nutzung', in: Wolf, Susanne (Hrsg.), *Neue Entwicklungen und Erfahrungen im Bereich der digitalen Archivierung: Von der Behördenberatung zum Digitalen Archiv: 14. Tagung des Arbeitskreises 'Archivierung von Unterlagen aus digitalen Systemen' vom 1. und 2. März 2010 in München* (Sonderveröffentlichungen

der Staatlichen Archive Bayerns 7), München: Generaldirektion der Staatlichen Archive Bayerns, S. 26-36.
Naumann, Kai (2020), ‘Neues vom Bewertungsautomaten: Workshop über Selesta in Stuttgart und Ludwigsburg’, *Archivar* 73, S. 63-64.
Pilger, Andreas (2015), ‘Bewertung elektronischer Fachverfahren: Diskussionspapier des VdA-Arbeitskreises Archivische Bewertung’, *Archivar* 68, S. 90-92.
Raselli, Donato (2014), *Verfahren zur Langzeitarchivierung von Datenbankinhalten aus Fachanwendungen und die Dokumentation dazugehöriger Prozessvorgänge* (nestor edition 7), Frankfurt am Main: nestor – Kompetenznetzwerk Langzeitarchivierung c/o Deutsche Nationalbibliothek.
Schlemmer, Martin (2025), ‘„E-Akte! Wir arbeiten doch schon digital!“ Wozu dient unter wem nutzt eine gute elektronische Aktenführung? Erfahrungen aus der Behördenberatungspraxis in der Landesverwaltung NRW’, *Scrinium* 79 (i. Dr.).
Schweizer, Verena (2021), ‘E-Akte – XJustiz – Fachverfahren: Entwicklung eines Aussonderungsworkflows für die E-Akte Justiz in Baden-Württemberg’, in: Irmgard Christa Becker u. a. (Hrsg.), *E-Government und digitale Archivierung* (Veröffentlichungen der Archivschule Marburg 67), Marburg: Archivschule Marburg, S. 119-128.
Steffenhagen, Björn (2019), ‘Praktische Möglichkeiten und Grenzen der Übernahme von Fachverfahren’, *Archive in Sachsen-Anhalt* (2019), S. 8-9.
Unger, Michael / Schmalzl, Markus (2020), ‘Digitales Verwaltungshandeln nachvollziehbar archivieren oder: Was ist die (E)Akte?’, *Archivar* 73, S. 371-378.

Konzeption einer Archivschnittstelle zum künftigen Personalmanagementsystem des Freistaates Sachsen

Christine Friederich und Karsten Huth

Einleitung

Der Freistaat Sachsen plant im Rahmen seiner Digitalstrategie „sachsen digital 2030“ (sachsen digital 2030 (2022)) flächendeckend den sukzessiven Umstieg auf ein einheitliches elektronisches Personalmanagementsystem (ePM.SAX) ab 2025. Dazu gehört auch die Einführung der elektronischen Personalakte (ePA). Die Umsetzung erfolgt in mehreren Schritten. Gestartet wurde 2023 mit dem Pilotprojekt ePM.SMI im Geschäftsbereich des Sächsischen Staatsministeriums des Innern (SMI). Das Sächsische Staatsarchiv ist seitdem am Pilotprojekt beteiligt.

Das Ende des Lebenszyklus von elektronischen Personalakten wird konsequent von Anfang an mitgedacht und die Voraussetzungen dafür geschaffen, elektronische Personalakten künftig regelkonform auszusondern. Aktuell erarbeitet das Staatsarchiv dafür gemeinsam mit den Projektpartnern die Aussonderungsschnittstelle und den Aussonderungsprozess. Es handelt sich im Folgenden also um einen Werkstattbericht, der den Ist-Stand zum gegenwärtigen Zeitpunkt widerspiegelt. Der Beitrag stellt deshalb kein fertiges Ergebnis vor, sondern zeigt an einem konkreten Beispiel auf, wie archivfachliche, organisatorische und technische Anforderungen in den Aussonderungsprozess in das Pilotprojekt eingebracht und – soweit sich das bereits sagen lässt – umgesetzt werden.

Im Folgenden wird zuerst kurz das Pilotprojekt ePM.SMI vorgestellt. In einem zweiten Schritt werden die Anforderungen und die Grundkonzeption des Aussonderungsprozesses erläutert. Drittens werden die technischen Überlegungen und praktischen Möglichkeiten einer Archivschnittstellte zur Aussonderung der elektronischen Personalakte vorgestellt.

Das Pilotprojekt ePM.SMI

Das Pilotprojekt ePM.SMI erarbeitet zunächst für den Geschäftsbereich des SMI die wesentlichen Grundlagen für das künftige einheitliche Personalmanagementsystem im Freistaat. Zum Geschäftsbereich des SMI zählen u. a. Polizeibehörden, die Landesdirektion Sachsen (LDS) – eine große Mittelbehörde mit vielfältigen Aufgaben –, aber auch kleinere Behörden wie das Statistische Landesamt oder das Staatsarchiv. Die Leitung des Pilotprojekts liegt beim SMI, die Gesamtprojektleitung des Projekts ePM.SAX bei der Sächsischen Staatskanzlei. Organisatorisch ist das Pilotprojekt in mehrere Scrum-Teams sowie das Teilprojekt 2: Rollout gegliedert,

die jeweils unterschiedliche Themenschwerpunkte verfolgen, etwa Personalmanagement oder Schnittstellen. Die Scrum-Teams und das Teilprojekt 2 setzen sich zusammen aus Expertinnen und Experten aus verschiedenen Behörden sowie aus dem Staatsbetrieb Sächsische Informatik Dienste (SID), dem zentralen IT-Dienstleister des Freistaats, und externen Dienstleistern, die vor allem Aufgaben im Projektmanagement sowie in der technischen Umsetzung wahrnehmen. Das Staatsarchiv ist Mitglied im Teilprojekt 2, das sich im Wesentlichen um folgende Aufgaben kümmert:

1. (Ersetzendes) Scannen

Bestehende Papier-Personalakten, die noch nicht abgeschlossen sind, werden im Rahmen des Umstiegs ersetzend gescannt („Bestandsakten-Scan"). Allein für das Pilotprojekt wird geschätzt, dass etwa 14,5 Millionen Seiten gescannt werden müssen. Zudem wird im Produktivbetrieb das Scannen von Papier-Posteingängen erforderlich sein. Die dafür benötigten Prozesse, die technische Ausstattung und die Infrastruktur werden im Teilprojekt 2 erarbeitet.

2. Datenmigration der Personal(stamm)daten

Bereits jetzt werden Personal(stamm)daten im Fachverfahren PVS elektronisch vorgehalten. Diese Daten müssen in das neue Personalverwaltungssystem übertragen werden. Alle notwendigen Voraussetzungen und die Umsetzung werden im Teilprojekt 2 erarbeitet.

3. Aussonderung

Die Aussonderung betrifft sowohl die elektronische Personalakte als auch die Personaldaten. Für beide Fälle werden Prozesse und Schnittstellen konzipiert und umgesetzt.

Zusätzlich zur Projektstruktur gibt es verschiedene Arbeitsgruppen, die aus Vertreterinnen und Vertretern der betroffenen Behörden und Dienststellen bestehen. Dazu gehören etwa regelmäßige Treffen der sogenannten „Einführungsbeauftragten", die den Umstieg auf das elektronische Personalmanagementsystem in ihren jeweiligen Häusern betreuen, oder die AG Scannen, die sich praxisorientiert mit den Anforderungen und der Umsetzung des (ersetzenden) Scannens befasst und damit die Arbeit des Teilprojekts 2 unterstützt.

Das Staatsarchiv ist im Teilprojekt 2 mit insgesamt drei Personen vertreten: Eine Archivarin aus dem Grundsatzreferat für den Bereich der Überlieferungsbildung sowie zwei Vertreter des elektronischen Staatsarchivs el_sta, die die technische Umsetzung betreuen. Bei Bedarf wird Staatsarchiv-intern zusätzlich auf die praktische Expertise der Fachabteilungen zurückgegriffen. Die Mitarbeit im Pilotprojekt ist zeitaufwändig: Die Mitglieder des Teilprojekts 2 treffen sich wöchentlich per Videokonferenz zu Besprechungen, um Sachstände, offene Punkte und nächste Schritte abzustimmen. Daran nimmt das Staatsarchiv üblicherweise mit einem Vertreter

teil. Dazu kommen gesonderte, meist wöchentliche Termine mit Dienstleistern, an denen alle Projektbeteiligten aus dem Staatsarchiv teilnehmen, um fachlich an der Konzeption des Aussonderungsprozesses und der Aussonderungsschnittstelle zu arbeiten. Aktuell wird auf Grundlage eines vom Staatsarchiv bereits im Herbst 2023 erstellten Grobkonzepts „Aussonderung" das Feinkonzept finalisiert. Im Feinkonzept sind die in ePM.SMI für die Aussonderung benötigten Anpassungen enthalten, die programmiert, getestet und schließlich implementiert werden sollen. Dafür sind perspektivisch weitere Ressourcen einzuplanen.

Archivfachliche Anforderungen an die Aussonderung der elektronischen Personalakte

Bei der elektronischen Personalakte (ePA) handelt es sich um das Fachverfahren SAP HCM, das an das E-Akte-System VIS.SAX über eine Schnittstelle angebunden ist. Es ist eine dynamische Akte, deren Metadaten und Strukturinformationen im Fachverfahren vorgehalten werden. Die in der Personalakte bzw. deren Teilakten enthaltenen Dokumente werden bei Bedarf über das Fachverfahren aufgerufen, aus VIS.SAX bereitgestellt und im Fachverfahren angezeigt. Die Bearbeiterinnen und Bearbeiter arbeiten ausschließlich über das Fachverfahren.
Das bedeutet, dass der gesamte Aussonderungsprozess über das Fachverfahren abgebildet werden wird. Dabei ist zu beachten, dass bereits bei der Anlage und der Bearbeitung der elektronischen Personalakte wesentliche (technische) Voraussetzungen erfüllt werden müssen, um der personalaktenführenden Stelle einen rechtskonformen Ablauf des Lebenszyklus, einschließlich der Aussonderung, zu ermöglichen. Diese Anforderungen müssen in SAP HCM zum Teil gesondert durch einen Dienstleister programmiert und umgesetzt werden. Im vorliegenden Fall erfolgt dies im Wesentlichen über einen eigenen sogenannten SAP-Infotyp „Aussonderung".
Voraussetzung für einen korrekten Ablauf der Aussonderung sind:

- *Akte schließen (z.d.A.-Verfügung) und Start der Aufbewahrungsfrist*

Es muss möglich sein, nach dem Ende der Bearbeitung die Personalakte abzuschließen. Der Abschluss erfolgt durch den Bearbeiter oder die Bearbeiterin. Mit dem Abschluss der Bearbeitung wird gleichzeitig der Start der Aufbewahrungsfrist initiiert. D. h. ab dem Zeitpunkt des Aktenschlusses – bzw. mit Ablauf des Kalenderjahrs des Aktenschlusses – beginnt die Aufbewahrungsfrist zu laufen. Die Personalakte kann dann noch aufgefunden und eingesehen, aber nicht mehr bearbeitet werden. Zudem ist hier festzulegen, ob und wie es möglich sein soll, im Rahmen einer festgelegten Transferfrist die Reaktivierung der Personalakte für die Bearbeitung wieder zuzulassen. Das ist allein schon deshalb empfehlenswert, um versehentlich oder fälschlicherweise abgeschlossene Personalakten wieder für die Bearbeitung freigeben zu können.

- *Korrekter Ablauf der Aufbewahrungsfrist und von vorfristigen Löschungsgeboten*

Für jede Personalakte muss die Dauer der Aufbewahrungsfrist festgelegt sein, mit deren Ablauf die Aussonderung angestoßen wird: Die Personalakte ist dann aussonderungsreif und wird dem Staatsarchiv je nach Bewertungsentscheidung angeboten, übergeben oder ohne vorherige Anbietung vernichtet. Um zu verhindern, dass alle Personalakten aller anbietungspflichtigen Stellen zum gleichen Zeitpunkt ausgesondert werden, ist vorgesehen, im System einen Zeitplan zu hinterlegen, nach dem sukzessive zu abgestimmten Zeitpunkten für jede personalaktenführende Stelle die Aussonderung erfolgt.

Zu beachten ist, dass es in der Personalakte Dokumente geben kann, die aufgrund von Rechtsvorschriften Löschungsgeboten unterliegen, die unabhängig vom Zeitpunkt des Aktenschlusses und der Dauer der Aufbewahrungsfrist gelten. Das ist z. B. bei Disziplinarunterlagen in Personalakten von Beamtinnen und Beamten der Fall (§ 16 Abs. 3 Sächsisches Disziplinargesetz). Es muss also gewährleistet sein, dass diese zum korrekten Zeitpunkt dem Staatsarchiv angeboten und übergeben oder vernichtet werden (Abschn. III., Ziff. 3 Verwaltungsvorschrift über die Aussonderung von Personalakten (VwV AusPersAkten)). Dafür wird unabhängig von der für die gesamte elektronische Personalakte hinterlegten Aufbewahrungsfrist über den Infotyp Aussonderung eine gesonderte „Verweildauer“ auf Dokumentenebene hinterlegt.

- *Festlegung der Aussonderungsart und des daran geknüpften Aussonderungswegs*

Die Aussonderungsart legt je nach hinterlegter Bewertungsfestlegung den Prozessablauf für die Aussonderung fest. Die Bewertungsfestlegungen A (Archivieren), B (Bewerten) und V (Vernichten) setzen also unterschiedliche Prozesse in Gang. Ziel ist, dass der Aussonderungsprozess möglichst automatisiert und mit möglichst geringer Intervention durch den Archivar bzw. die Archivarin ablaufen kann. Voraussetzung dafür ist, dass für möglichst viele Personalakten finale Bewertungsentscheidungen bestehen, also A oder V.

Das Staatsarchiv hat Bewertungsfestlegungen für Personalakten getroffen, die in einer eigenen Verwaltungsvorschrift niedergelegt sind, der VwV über die Aussonderung von Personalakten (VwV AusPersAkten). Wesentlich sind dort die Bewertungsfestlegungen A („übergeben“) und B („anbieten“), für den Rest gilt implizit oder explizit V.

A sind im Wesentlichen die Grundakten und bestimmte Teilakten der Personalakten z. B. von Mitgliedern der Staatsregierung, von bestimmten Besoldungs- und Entgeltgruppen, von Leitungen von Behörden, Gerichten und sonstigen Stellen sowie von allen Beschäftigten, die am 1. März, 1. Juni, 1. September und 1. Dezember eines Jahres geboren sind. B sind z. B. die Grundakten und bestimmte Teilakten von herausragenden Persönlichkeiten. In SAP sind

umfangreiche Metadaten zu den einzelnen Personen vorhanden, sodass sich darüber auch die Bewertungskriterien abbilden lassen.
Bei der Konzeption des Aussonderungsprozesses ist zudem zu beachten, dass voraussichtlich fast alle elektronischen Personalakten zunächst hybrid sein werden, d.h. dass sie neben dem elektronischen auch über einen Papier-Anteil verfügen werden. Grund dafür sind bestehende Rechtsvorschriften, die ein ersetzendes Scannen von bestimmten Dokumentarten nicht erlauben, sodass diese weiterhin in ihrer Ursprungsform aufbewahrt werden müssen.
Die verschiedenen Aussonderungsprozesse für die Bewertungsfestlegungen A, B und V sind folgende:

- *Bewertungsfestlegung V:*

Standardeinstellung bei Anlage einer Personalakte wird V sein, wenn nicht bereits bei Anlage der Akte ein Kriterium für eine andere Bewertungsfestlegung vorhanden ist (z. B. Geburtsdatum). Der weit überwiegende Teil aller Personalakten fällt in diese Kategorie. Diese elektronischen Personalakten werden nach Ablauf der Aufbewahrungsfrist ohne Anbietung gelöscht.

- *Bewertungsfestlegung A:*

Diese Personalakten werden direkt an das elektronische Staatsarchiv übergeben und archiviert. Es ist vorgesehen, dass Personalakten, sobald ein Metadatum ein Archivwürdigkeitskriterium erfüllt (bspw. Besoldungsstufe), automatisiert von der Standard-Einstellung V auf die Einstellung A gesetzt wird.

- *Bewertungsfestlegung B:*

Es wird davon ausgegangen, dass Personalakten mit der Bewertungsfestlegung B nur einen sehr geringen Anteil ausmachen. Die Bewertungsfestlegung B kann von den Bearbeiterinnen und Bearbeitern eingetragen werden als „Vorschlag der Behörde“ bzw. auf Vorschlag des zuständigen Archivars bzw. der zuständigen Archivarin. In einem zusätzlichen Freitextfeld soll eine Begründung für den Vorschlag eingegeben werden. Diese Personalakten werden dem Staatsarchiv übermittelt, dort per Aktenautopsie abschließend bewertet und je nach Votum gelöscht oder archiviert.

Signifikante Eigenschaften der elektronischen Personalakte

Die Signifikanten Eigenschaften der Elektronischen Personalakte sind die Struktur der Grundakte sowie der Teilakten zu einer Person. Ebenso signifikant sind die Metadaten zu jeder Personalakte und natürlich die enthaltenen Dokumente.

Die Struktur einer Personalakte bestehend aus der Grundakte und 11 Teilakten wird den Benutzenden in SAP über eine Ordnerhierarchie mit jeweiligen Unterordnern dargeboten. Auf jeder Ebene werden die Metadaten zur Person (z.B. Personalnummer; Geburtsdatum; Familienstand usw.) angezeigt. Nach der Übernahme soll für die Benutzenden im Staatsarchiv eine vergleichbare Optik und Haptik bei der Durchsicht einer elektronischen Personalakte gegeben sein, ohne dass dabei die ursprünglich verwendete SAP-Software zum Einsatz kommt.

Der technische Ablauf der Aussonderung

Zu einem mit dem Archiv festgelegten Termin wird von SAP aus ein Aussonderungslauf gestartet. An das Archiv werden nur Personalakten mit den Bewertungsfestlegungen A oder B übermittelt. Wegen der erwartet geringen Anzahl von Personalakten pro Aussonderungslauf werden auch die noch zu bewertenden Akten bereits vollständig übergeben. Die Bewertung kann dann anhand der vollständigen Personalakte vollzogen werden. Da SAP von Hause aus über eine Schnittstelle verfügt, die die interne Struktur der Personalakte als Ordnerverzeichnis exportiert, kann diese auch für die Aussonderungsschnittstelle verwendet werden. Jeder Aussonderungslauf bekommt eine Nummer, mit welcher der Lauf und die betroffenen Personalakten eindeutig in SAP identifiziert werden können. In SAP wird auch hinterlegt, welcher Prozessschritt des Aussonderungslaufs aktuell anliegt.

Das Aussonderungspaket wird in einer vereinbarten Verzeichnisstruktur auf einem WebDAV-Server innerhalb des Sächsischen Verwaltungsnetzes abgelegt. Der oberste Ordner ist nach der Nummer des Aussonderungslaufs benannt. Danach folgen die Ordner der Personen (benannt nach ihren Personalnummern), die schließlich die Struktur der jeweiligen Akte enthalten. Es werden die Grundakte und alle Teilakten übermittelt. Lediglich die Teilakten „Gesundheit“, „Finanzen“ und „Urlaub“ werden gemäß der VwV AusPersAkten zurückgehalten. Innerhalb der Ordner liegen Dateien, die die eigentlichen Dokumente der Personalakte repräsentieren. Diese werden in einer noch festzulegenden PDF/A-Variante übermittelt. Die Metadaten zu den jeweiligen Personen und den Dokumenten werden als xml-Dateien auf den jeweiligen Ordnerebenen beigelegt.

Das Sachgebiet Elektronisches Staatsarchiv hat Zugriff auf den WebDAV-Server und prüft, ob Aussonderungen vorliegen. Die abgelegten Aussonderungspakete werden vom Server ins Archiv verbracht und dort für den Ingest vorbereitet. Personalakten mit der Bewertungsfestlegung B werden zuerst an die zuständige Fachabteilung zur Bewertung weitergeleitet.

Alle zur Archivierung vorgesehenen Personalakten werden in jeweils eine spezielle SQLite-Datenbank eingelesen, die von einem eigens im Sächsischen Staatsarchiv entwickelten

Programm „Bytebarn“ gelesen werden kann. Bytebarn ermöglicht eine Darstellung der Aktenstruktur, die logisch exakt der Struktur innerhalb des SAP-Systems entspricht und ihr auch optisch nah kommt. Bei einer Benutzung kann man durch die Ordner hindurch navigieren und am Ende das gewünschte Dokument öffnen.
Aus der Personalakte einer Person entsteht so ein AIP im Elektronischen Staatsarchiv, abgebildet und auffindbar durch eine Verzeichniseinheit im Archivinformationssystem des Sächsischen Staatsarchivs.

Ende des Aussonderungsprozesses und Fazit

Nachdem alle Personalakten mit der Bewertungsfestlegung A ingestiert wurden, erstellt das Archiv eine Löschgenehmigung, die über den WebDAV-Server automatisiert in das SAP-System zurückgeschickt wird. Das SAP-System prüft die Löschgenehmigung und entfernt die ausgesonderten Personalakten aus dem System. Dadurch, dass das Sächsische Staatsarchiv schon bei der Erstellung des neuen Verfahrens zum Personalmanagement mit einbezogen wurde, konnte von Anfang an die Aussonderung technisch und organisatorisch implementiert werden. Der im Projekt entwickelte automatisierte Prozess zur Aussonderung ist vom ersten Tag der Inbetriebnahme an aktiv und erleichtert damit die künftigen Arbeitsschritte im Archiv.

Bibliografie

Sächsisches Staatsministerium für Wirtschaft, Arbeit und Verkehr, Referat 41, Grundsatzfragen Digitalisierung (Red.) (2022), *sachsen digital 2030: besser, schneller, sicher* [Online], https://www.digitales.sachsen.de/massnahme-einfuehrung-eines-landeseinheitlichen-elektronischen-personalmanagementsystems-inklusive-elektronischer-personalakte-projekt-epm-sax-6103.html (23.09.2024).

Bayer, Peter (2016), *Dateihaufen unter Dach und Fach: 20. Tagung Arbeitskreis AUdS*, 1.-2. März 2016, Potsdam, Fachhochschule Potsdam [Online], https://www.sg.ch/content/dam/sgch/kultur/staatsarchiv/auds-2016/archivierung-von-unterlagen-mit-besonderen-strukturen/04_BAYER_AUDS2016_StA_Bayer_final.pdf (26.9.2024).

Der Weg der Studierendenakte ins elektronische Langzeitarchiv

Mona Bunse

Studierendenakten sind als Teil der vor Ort gebildeten amtlichen Überlieferung historisch besonders wertvoll, da Hochschulen[1] von den Menschen geprägt werden, die dort arbeiten, studieren und forschen (vgl. Becker et al, 2009, S. 7). Aber nicht nur aufgrund ihres historischen Wertes sollen diese Akten für die Nachwelt erhalten werden. Die Archive gewährleisten den Hochschulen Rechtssicherheit, indem sie sicherstellen, dass die verwaltungsmäßigen Entscheidungsprozesse dauerhaft nachvollziehbar bleiben. Ein direkter Auftrag zur Archivierung ergibt sich aus dem Archivgesetz Nordrhein-Westfalen (NRW). Durch das E-Government-Gesetz NRW sind die Hochschulen zudem angehalten, bis Ende 2025 die elektronische Aktenführung zu realisieren. Dadurch müssen auch Studierendenakten zukünftig digital geführt und anschließend in dieser Form archiviert werden.

Das Projekt *Digitale Langzeitarchivierung an den NRW-Hochschulen* (LZA.NRW)[2] basiert auf den zuvor initiierten eAkte-Projekten der Digitalen Hochschule NRW. Es entwickelt zwei „Golden Master" für die Aussonderung und Archivierung von elektronischen Studierenden- und Personalakten sowie -daten. Zur Umsetzung der digitalen Langzeitarchivierung steht für die Hochschulen die Infrastruktur DiPS.kommunal des Digitalen Archivs NRW (DA NRW) – einer Arbeitsgemeinschaft von Land und Kommunen – zur Verfügung. Auch die Beratung und Unterstützung bei der Einrichtung der DiPS-Mandanten und Schnittstellen an den beteiligten Hochschulen sind Teil des Projektauftrags von LZA.NRW. Dieses Pilotprojekt markiert den ersten Schritt zur digitalen Langzeitarchivierung für alle öffentlichen Hochschulen in Nordrhein-Westfalen. Zudem unterstützt LZA.NRW die Gründung weiterer Hochschularchive – eine wesentliche Voraussetzung für den Einstieg in die digitale Langzeitarchivierung.

Eine wichtige Grundlage für LZA.NRW stellt das Projekt e-Studierendenakte.nrw[3] dar. In dessen Rahmen wurde die Bildung der digitalen Studierendenakte aus den in NRW führenden Campusmanagementsystemen (CaMS) HISinOne und CAMPUSonline unter Verwendung des

[1] Der Begriff Hochschulen wird im Folgenden als Sammelbegriff für Universitäten und Fachhochschulen bzw. Hochschulen für Angewandte Wissenschaften verwendet.

[2] LZA.NRW ist eine gemeinsame Initiative der 30 öffentlichen Fachhochschulen und Universitäten in NRW. Es wird auf Grundlage eines Kooperationsvertrags über eine Umlage finanziert und war ursprünglich auf eine Laufzeit von Oktober 2022 bis September 2024 ausgelegt. Nach einer ersten kostenneutralen Verlängerung um drei Monate sowie einer weiteren Verlängerung um zwei Jahre, wird das Projekt am 31.12.2026 enden. Organisatorisch ist LZA.NRW an der Universität Duisburg-Essen (UDE) angesiedelt.

[3] Das Projekt e-Studierendenakte.nrw wurde am 30.11.2022 abgeschlossen und wird seitdem durch das Kompetenzzentrum E-Akte NRW betreut.

Dokumentmanagementsystems (DMS) d.documents der Firma d.velop (ehemals d.3ecm) fachlich und technisch realisiert. Für die Übernahme dieser Akten hat das Projekt LZA.NRW ein Fachkonzept erarbeitet, nach dem alle archivwürdigen Studierendendaten gemäß dem nestor-Archivstandard über eine XML-Datei aus den CaMS an das DMS übergeben werden können. Aus dem DMS heraus soll anschließend die Aussonderung in das digitale Langzeitarchiv (dLZA) DiPS.kommunal erfolgen.

Der Vorteil der Übergabe aller archivwürdigen Daten an das DMS besteht darin, dass die anschließende Aussonderung allein über das DMS stattfinden kann. Auf diese Weise bleibt die Zahl der Schnittstellen zum dLZA möglichst klein. Nach Ablauf der festgelegten Aufbewahrungsfristen[4] wird die gesamte Akte inklusive der XML-Datei dem zuständigen Archiv angeboten und an DiPS übergeben. Die Nachrichten, mit denen die Systeme kommunizieren, beruhen auf dem xdomea-Standard.[5] Nach Erhalt der Nachricht zur erfolgreichen Aussonderung sind die Ursprungssysteme angehalten, die ausgesonderten Daten endgültig zu löschen. Dadurch soll eine doppelte Datenhaltung verhindert und den Vorgaben der Datenschutz-Grundverordnung (DSGVO) nachgekommen werden.

Gesetzliche Grundlagen

Das Archivgesetz NRW verpflichtet öffentliche Hochschulen des Landes, die Archivierung der bei ihnen entstandenen Unterlagen in eigener Zuständigkeit zu regeln.[6] Dabei gilt: „Archivgut ist auf Dauer sicher zu verwahren. Es ist in seiner Entstehungsform zu erhalten, sofern keine archivfachlichen Belange entgegenstehen.“ Hochschulen müssen außerdem sicherstellen, dass das bei ihnen entstandene Archivgut nicht nur dauerhaft aufbewahrt, sondern auch geschützt wird vor unbefugter Nutzung, Beschädigung oder Vernichtung (ArchivG NRW § 5 Abs. 2).

Unterlagen, die zur Erfüllung der Aufgaben nicht mehr benötigt werden, sind nach Ablauf der Aufbewahrungsfristen den Hochschularchiven anzubieten (vgl. ArchivG NRW § 4). Diese Anbietungspflicht gilt auch für elektronische Akten, die durch das *Gesetz zur Förderung der elektronischen Verwaltung in Nordrhein-Westfalen* (kurz: E-Government-Gesetz NRW) ab dem 31. Dezember 2025 verpflichtend sind (vgl. EGovG NRW § 9 Abs. 3). Dementsprechend wird die im Archivgesetz NRW genannte Entstehungsform für Akten zukünftig (vorwiegend) elektronisch sein. Auf dieser Grundlage werden derzeit zahlreiche Projekte zur eAkten-Einführung an den Hochschulen realisiert.

[4] Im Abschlussbericht des Projekts e-Studierendenakte.nrw werden Aufbewahrungsfristen von maximal 60 Monaten empfohlen.
[5] Mehr dazu in KoSIT, 2021.
[6] Festgehalten ist dies in §§ 1 und 11 ArchivG NRW. Davon ausgenommen sind die staatlichen Kunst- und Musikhochschulen sowie alle Hochschulen in privater bzw. kirchlicher Trägerschaft.

Eine besondere Herausforderung ist dabei die langfristige Interpretierbarkeit elektronischer Daten. § 11 Abs. 1 des E-Government-Gesetzes NRW weist darauf hin, dass zur Erhaltung der Lesbarkeit elektronische Akten möglicherweise in andere Formate überführt werden müssen. Auch das Archivgesetz NRW bleibt in diesem Kontext verbindlich, wie § 11 Abs. 2 des E-Government-Gesetzes betont: „Die Vorschriften des Archivgesetzes Nordrhein-Westfalen […] bleiben unberührt." Damit wird sichergestellt, dass die Anbietungspflicht auch für elektronisches Schriftgut gilt.

Gleichzeitig müssen Hochschulen den Schutz personenbezogener Daten gemäß der Datenschutz-Grundverordnung (DSGVO) und dem Datenschutzgesetz Nordrhein-Westfalen (DSG NRW) gewährleisten. § 10 Abs. 1 des DSG NRW regelt, dass personenbezogene Daten nur gelöscht werden dürfen, nachdem sie dem zuständigen Archiv angeboten und als nicht archivwürdig bewertet wurden. Eine Besonderheit bildet hierbei die Archivierung als „Löschungssurrogat", was bedeutet, dass archivwürdige personenbezogene Daten zwar im Archiv weiterhin vorhanden sind, aber seitens der aktenbildenden Stelle als gelöscht betrachtet werden. Eine Einsichtnahme im Archiv ist daher für die Aktenbildner:innen nicht möglich (vgl. § 6 Abs. 4 ArchivG NRW in Verbindung mit § 10 Abs. 1 DSG NRW).

Der gesetzliche Archivierungsauftrag ist in NRW derzeit noch nicht vollständig umgesetzt, da einige Hochschulen noch keine eigenen Archive eingerichtet haben. Ohne ein solches Hochschularchiv ist es jedoch nicht nur unmöglich, Akten anzubieten und zu übernehmen, sondern es wird auch eine korrekte und langfristige Verwaltung der Daten erschwert. Besonders elektronische Daten sind anfällig für Veränderungen und ihre Lesbarkeit ist stark von der Verfügbarkeit geeigneter Software abhängig. Die digitale Transformation bietet Hochschulen nun die Chance, den Archivierungsauftrag strategisch von Anfang an zu planen. Durch frühzeitige Zusammenarbeit mit Archivexpert:innen können effiziente Prozesse etabliert werden, die sicherstellen, dass Daten revisionssicher und langfristig lesbar bleiben. Entscheidend ist auch die Implementierung einer geeigneten technischen Lösung, die den dauerhaften Schutz archivierter Daten gewährleistet.

Der Lebensweg einer (Studierenden-)Akte

Im Folgenden wird der Lebensweg einer (Studierenden-)Akte nachgezeichnet, um einerseits die Aufbewahrung und die Archivierung klar voneinander abgrenzen zu können und andererseits aufzuzeigen, zu welchem Zeitpunkt eine Akte dem Archiv angeboten werden muss. Abb. 1 visualisiert diesen Lebensweg, wobei einmalige Ereignisse als pinkfarbene Punkte und die einzelnen Phasen als blaue Pfeile dargestellt sind. Zudem werden die Zuständigkeiten

unterschieden, die für eine „aktive“ Akte bei der jeweiligen aktenführenden Stelle und im Anschluss bei dem zuständigen Archiv liegen.

Der Prozess beginnt mit der Anlage einer Akte. Danach folgt die Bearbeitung, die mit dem Vermerk „zu den Akten“ (zdA) abgeschlossen wird. Ab diesem Zeitpunkt beginnt die Aufbewahrungsphase, in der die Akte nicht mehr verändert werden darf. Wie lange eine Akte aufbewahrt wird, hängt von der Art der Akte und den Aufbewahrungsfristen für die darin enthaltenen Dokumente ab. Die Fristen sind so gestaltet, dass die Akte bei Bedarf wieder „aufleben“ kann. Das kann zum Beispiel bei Studierenden zutreffen, die ihr Studium beenden und sich später für ein neues Studium einschreiben. In diesem Fall entspricht die Exmatrikulation der „zu den Akten“-Setzung bzw. Schlussverfügung. Wenn die vorherige Studierendenakte reaktiviert wird, kann die Person ihre alte Matrikelnummer auch im neuen Studiengang nutzen.

Nach Ablauf der Aufbewahrungsfrist wird die Akte ausgesondert und gegebenenfalls archivisch bewertet, soweit das nicht schon vorher geschehen ist. Für Studierendenakten gibt es bereits Empfehlungen zur Archivierung. Je nach Einschätzung und Festlegung des Archivpersonals an den einzelnen Hochschulen kann eine (Nach-)Bewertung vorgenommen werden. Ab diesem Zeitpunkt ist das Hochschularchiv zuständig und nicht mehr die Stelle, die die Akte geführt hat. Die Akte oder Teile davon werden je nach archivischer Bewertungsentscheidung entweder in das Archiv übernommen oder datenschutzgerecht vernichtet. Für digitale Akten ist die Übernahme in ein digitales Langzeitarchiv erforderlich. Die bloße Speicherung auf einem Hochschulserver genügt nicht den Anforderungen an die digitale Langzeitarchivierung, für die eine geeignete Infrastruktur genutzt werden muss (s. Konzeption des Aussonderungs- und Übernahmeprozesses).

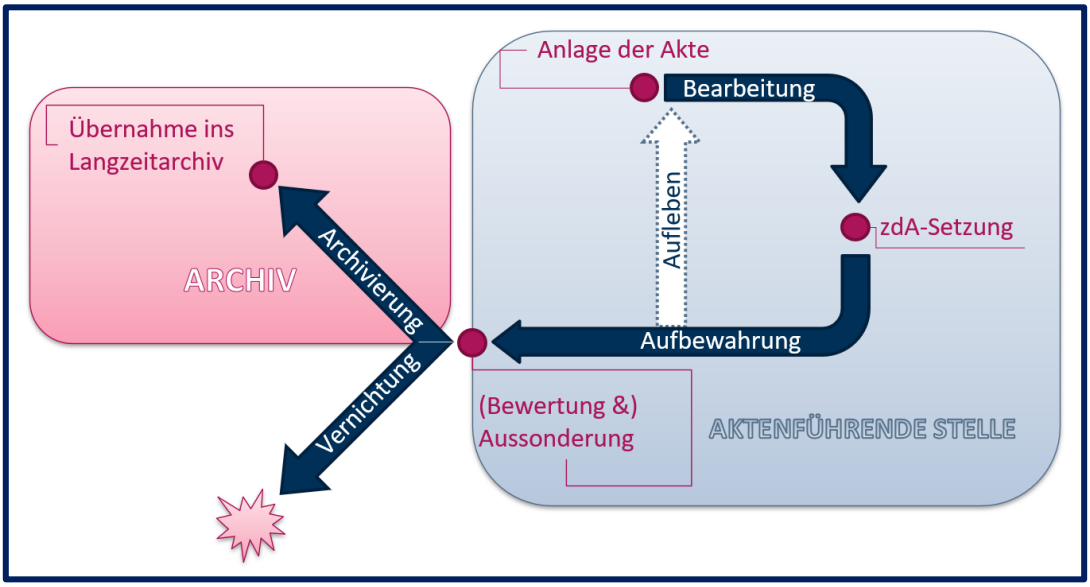


Abbildung 1: Der Lebensweg einer (Studierenden-)Akte

Die eStudierendenakte: Konzeption, Aufbewahrungsempfehlungen & Co.

Um nun speziell für eStudierendenakten die digitale Langzeitarchivierung aller archivwürdigen Inhalte umsetzen zu können, müssen die fachlichen und technischen Parameter dieser Aktenart festgelegt sein. Das Projekt LZA.NRW baut hierzu auf der durch das Projekt e-Studierendenakte.nrw erarbeiteten Masterlösung zur revisionssicheren und datenschutzkonformen Ablage und Verarbeitung sowie zur Aufbewahrung von Daten und Dokumenten der Bewerbenden bzw. Studierenden von NRW-Hochschulen auf. Darin enthalten sind auch Empfehlungen für Aufbewahrungsfristen der einzelnen Bestandteile der eStudierendenakte sowie Empfehlungen für archivische Aussonderungsentscheidungen in Bezug auf Studierendendaten, -akten und -dokumente[7].

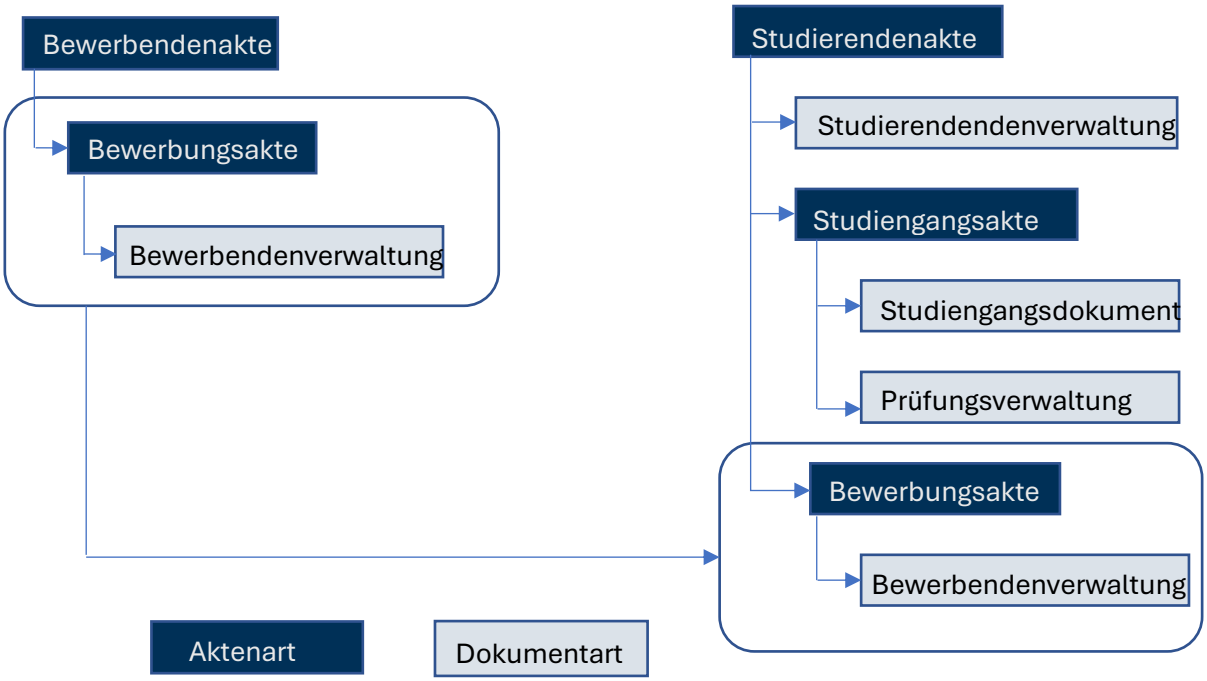


Abbildung 2: Aktenstruktur der eStudierendenakte (Kaltenbach et al, 2022, S. 21, farblich abgeändert)

Zunächst wird im DMS für jede:n Bewerbende:n eine Bewerbendenakte gebildet, unter der einzelne Bewerbungsakten angelegt werden können. Dadurch ist es möglich, einzelne Bewerbungen oder Teilbewerbungen voneinander zu trennen und einzeln auszusondern. Die Bildung der Studierendenakte erfolgt dann bei erfolgreicher Immatrikulation bzw. bei vorhandener Matrikelnummer und die Bewerbungsakte wird in diese integriert. Zudem befinden sich in der Studierendenakte Dokumente zur Studierendenverwaltung sowie die Studiengangsakten. Durch Unterakten werden dabei einzelne Studiengänge voneinander abgegrenzt, wie beispielsweise

[7] Diese Empfehlungen basieren auf Abstimmungen mit Vertreter:innen des Projekts LZA.NRW und der Arbeitsgemeinschaft der Hochschularchive NRW.

Bachelor- und Masterstudiengang. Es sind Anpassungen und Konfigurationen dieser Konzeption möglich, die aber zu Mehraufwand führen können (vgl. Kaltenbach et al, 2022, S. 21 ff.).
Die elektronische Studierendenakte umfasst sämtliche aktenrelevanten Daten und Dokumente einer Person des gesamten Student Life Cycles, jedoch werden im Rahmen dieser Konzeption bislang noch nicht alle archivwürdigen Stammdaten der Studierenden aus den führenden Fachverfahren des Campusmanagements an das DMS übertragen. Um auch diese Daten in das digitale Langzeitarchiv übernehmen zu können, musste das Projektteam von LZA.NRW also einen Weg finden.
Erfolgt nach einer Bewerbung keine Immatrikulation, wird empfohlen, die Bewerbendenakte nach 12 Monaten ab Anlage auszusondern und deren Inhalte mit „V“ (d. h. vernichten) zu bewerten. Bei erfolgreicher Immatrikulation soll die Aussonderung der Bewerbungsakte zusammen mit der jeweiligen Studierendenakte 60 Monate nach Exmatrikulation erfolgen. Deren Inhalte sind archivisch zu bewerten („B“). Auch für Studierenden- und Studiengangsakte ist vorgesehen, die Aussonderung 6 Monate nach dem Exmatrikulationsdatum vorzunehmen. Die einzelnen Dokumente innerhalb dieser Akten erhalten ebenfalls eine Aufbewahrungsfrist von maximal 60 Monaten sowie unterschiedliche Aufbewahrungsarten, die in einer Dokumentenliste des Projekts e-Studierendenakte.nrw festgehalten sind (vgl. Kaltenbach et al, 2022, S. 37).

Archivwürdige Studierendendaten

Welche Einzelinformationen zu Studierenden archivwürdig sind, ist im nestor-Archivstandard zur *Archivierung von Studierendendaten aus Fachverfahren* (s. Tabelle 1) festgelegt. Diese Daten sind nach Ablauf der Aufbewahrungsfrist dem Hochschularchiv anzubieten, sofern sie an der jeweiligen Hochschule geführt werden. Es besteht keine Notwendigkeit, die Daten nur aufgrund ihrer Archivwürdigkeit zu führen. Dies gilt beispielsweise für die Religionszugehörigkeit. Diese kann bei Hochschulen in kirchlicher Trägerschaft eine Rolle spielen, an den meisten Hochschulen wird sie jedoch nicht bei den Studierenden aufgenommen und kann folglich nicht ausgesondert werden.

1. Personenstammdaten	2. Studienverlauf	3. Studienleistungen
1.1 Nachname/einziger Name 1.2 Geburtsname und andere frühere Namen 1.3 Vorname/Vornamen 1.4 Ordensname/ Künstler*innenname/ etc. 1.5 Geschlecht 1.6 Geburtsdatum 1.7 Todesdatum 1.8 Geburtsort 1.9 Staatsbürgerschaft 1.10 Familienstand 1.11 Religionszugehörigkeit 1.12 Matrikelnummer 1.13 Immatrikulationsdatum 1.14 Exmatrikulationsdatum 1.15 Exmatrikulationsgrund 1.16 Hochschulzugangs-berechtigung 1.17 Vorherige Abschlüsse / sonstige Vorbildung 1.18 Heimatanschrift 1.19 Semesteranschrift 1.20 Nachname/einziger Name einer*s gesetzlichen Vertreter*in 1.21 Vorname/Vornamen einer*s gesetzlichen Vertreter*in 1.22 Anschrift einer*s gesetzlichen Vertreter*in 1.23 Herkunftshochschule 1.24 Zielhochschule bei Hochschulwechsel 1.25 Letztes Bearbeitungsdatum	2.1 Hochschulsemester 2.2 Hör- und Rückmeldestatus 2.3 Urlaubssemester, Beurlaubungsgründe 2.4 Angaben zu den belegten Studiengängen 2.5 Bezeichnung des angestrebten Grads 2.6 Angaben zu den belegten Fächern 2.7 Fachsemester	3.1 Bezeichnung der abgeschlossenen Studiengänge 3.2 Bezeichnung der erworbenen Grade 3.3 Geltende Prüfungsordnung 3.4 Abschlussdatum 3.5 Studiengangbezogene Gesamtnote(n) 3.6 Informationen zu einzelnen Prüfungsleistungen 3.6.1 Semester der Prüfungsleistung 3.6.2 Fach 3.6.3 Modulcode 3.6.4 Modulbezeichnung 3.6.5 Veranstaltungsbezeichnung 3.6.6 Typ der Prüfungsleistung 3.6.7 Datum der Prüfungsleistung 3.6.8 Art und Form der Prüfungsleistung 3.6.9 Anerkannte Prüfungsleistung 3.6.10 Externe Hochschule 3.6.11 Note der Prüfung 3.6.12 Credit Points (ECTS-Punkte) 3.6.13 Prüfungsstatus 3.6.14 Prüfer*in 3.6.15 Arbeitstitel 3.6.16 Vermerke zu nicht bestandenen Prüfungen

Tabelle 1: Archivwürdige Einzelinformationen von Studierenden laut nestor-Archivstandard (nestor-AG Archivstandards, 2023, S. 23 ff.)

Konzeption des Aussonderungs- und Übernahmeprozesses

Die wichtigste Grundlage für die Konzeption der Aussonderung und Übernahme von elektronischen Unterlagen bildet das als ISO-Standard 14721 verabschiedete OAIS-Referenzmodell. Das Modell beschreibt den Prozess der Übernahme von Datenpaketen einer produzierenden Stelle (Submission Information Packages bzw. kurz: SIPs) über die Umwandlung zu Archiv-Informationspaketen (AIPs) bis hin zum Zugriff auf das Archivgut in Form von Dissemination Information Packages (DIPs). Die Informationspakete bestehen dabei immer aus dem eigentlichen Inhalt (Primärdaten) und beschreibenden Informationen (Metadaten). Funktional

unterscheidet das OAIS-Modell die sechs zentralen Aufgabenbereiche Datenübernahme (Ingest), Archivspeicher (Storage), Datenverwaltung (Management), Zugang (Access), Erhaltungsplanung (Preservation Planning) und Systemverwaltung (Administration) (vgl. Brühbach, 2010, Kapitel 4.2).

Ein digitales Langzeitarchiv muss folgende Anforderungen, die sich aus dem OAIS-Modell ableiten lassen, erfüllen:

- Vorbereitung auf alle OAIS-Aufgabenbereiche: Planung der Erzeugung von SIPs, AIPs und DIPs, Bereitstellung von Archivspeicher, Einrichtung eines archivischen Fachinformationssystems (AFIS), Regelung von Rollen und Zugriffsrechten sowie Erstellung einer Benutzungsordnung
- Sicherstellung der Erfüllung aller OAIS-Aufgaben
- langfristige Erhaltungsplanung mit Blick auf die Zukunft
- Betreuung durch geschultes Personal, das sich kontinuierlich über aktuelle technische und fachliche Entwicklungen informiert und weiterbildet

Die Langzeitarchivierungslösung DiPS.kommunal erfüllt die Anforderungen gemäß OAIS (vgl. LWL/ Stadt Köln, 2022, S. 5) und bildet eine sichere Infrastruktur für die Langzeitarchivierung der eAkten (und anderer digitaler Unterlagen) von Hochschulen. In der Anbahnungsphase des Projekts LZA.NRW wurde DiPS.kommunal daher ausgewählt und die Konzeption zur digitalen Langzeitarchivierung von eStudierendenakten auf die Nutzung dieser Software ausgerichtet.

Die Campusmanagementsysteme kommunizieren derzeit unidirektional mit dem DMS, indem Daten und Dokumente dorthin übertragen werden. Die Aussonderungs- und Übernahmekonzeption von LZA.NRW berücksichtigt darüber hinaus die archivwürdigen Einzelinformationen der Studierenden gemäß nestor-Archivstandard in Form einer XML-Datei. Das DMS quittiert diese Übergabe. Aus dem DMS werden dann die eStudierendenakten samt XML-Dateien nach DiPS.kommunal übertragen. Für die Übernahme von logisch strukturierten Daten gibt es den xdomea-basierten Eingangskanal eakte. Bei der Aussonderung werden die Primärdaten hier automatisch in inhaltliche Einheiten gegliedert und mit Metadaten im xdomea-XML-Format abgelegt[8] (vgl. LWL/ Stadt Köln, 2022, S. 8). Wenn eine erfolgreiche Übernahme stattgefunden hat, sendet DiPS.kommunal eine Löschrückmeldung an das DMS. Dadurch soll ermöglicht werden, dass die Daten perspektivisch nicht doppelt in verschiedenen Systemen vorgehalten werden, sondern final nur noch im digitalen Langzeitarchiv. Von da aus kann dann über ein

[8] Die Übernahme von unstrukturierten Daten nach DiPS.kommunal kann über das Pre-Ingest-Tool PIT.plus erfolgen.

AFIS auf die archivierten Daten zugegriffen werden. Außerdem kann mithilfe des AFIS die Erschließung erfolgen. Eine Recherche über DiPS.kommunal selbst ist nicht vorgesehen.

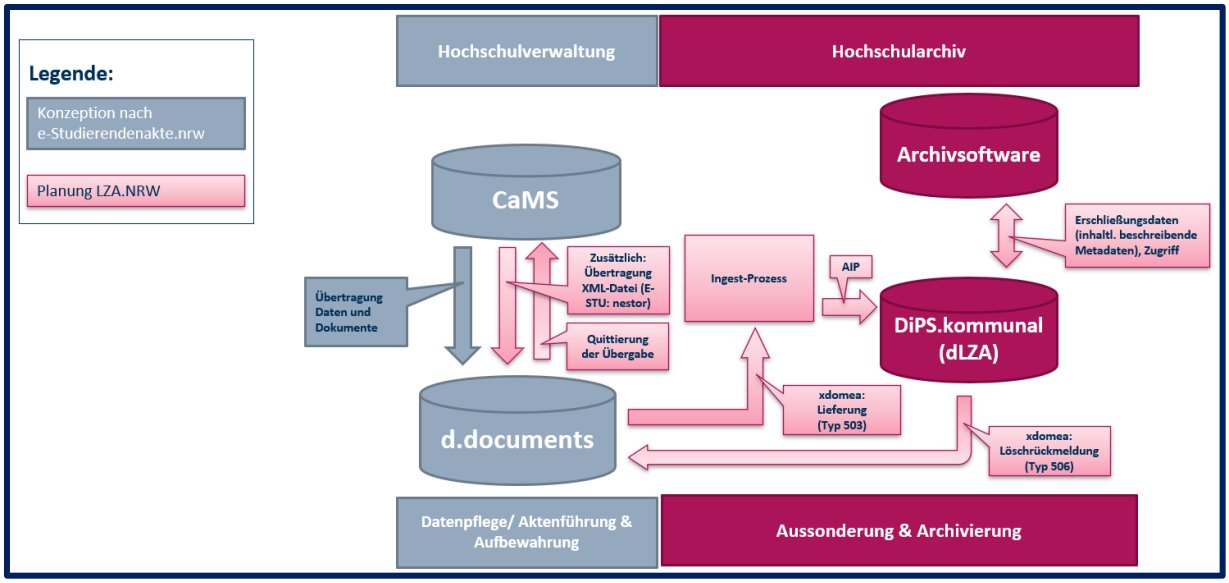


Abbildung 3: Aussonderungskonzeption der eStudierendenakten und -daten

Erstellung und Aufbau der XML-Datei

Bevor die archivwürdigen Studierendendaten zur Langzeitarchivierung an das DMS übergeben werden, prüfen die CaMS, ob noch Prozesse offen sind. Falls ja, wird eine Archivierungssperre gesetzt. Nach der Schlussverfügung (z. B. Exmatrikulationsdatum) im DMS und der Aufhebung der Sperre kann der automatische Datentransfer stattfinden. Da eine Lese- und Schreibsperre für die Studierendendaten-XML besteht, müssen die Daten bereits in den CaMS in der geforderten Weise strukturiert werden.

Die archivwürdigen Daten der Studierenden gliedern sich, orientiert am nestor-Archivstandard, in drei Bereiche: Personenstammdaten, Studierendenverlaufsdaten und Studienabschlussdaten. In der XML-Datei (ein Muster folgt auf der nachfolgenden Seite) sind beispielsweise verschiedene Adressen wie Semester- und Heimatadressen erfasst, die zu den Stammdaten zählen. Ebenfalls zu den Stammdaten gehören das Immatrikulations- und Exmatrikulationsdatum sowie die Angaben zur Hochschulzugangsberechtigung.

Der Studienverlauf ist nach Semestern organisiert und beinhaltet die Studiengänge und die jeweiligen Fächer. Dabei werden Details wie die Art des Studiums, Studienwechsel und Fachvertiefungen festgehalten. Auch die Anzahl der Urlaubssemester und deren Gründe werden dokumentiert.

Im Anschluss an den Studienverlauf werden die Studienabschlüsse aufgeführt, einschließlich der erbrachten Studienleistungen, Fachbezeichnungen und Bewertungen. Neben den internen Abschlüssen finden auch externe Hochschulabschlüsse Eingang in die XML-Datei.

Die als XML-Datei gelieferten archivwürdigen CaMS-Daten werden unterhalb der Akte abgelegt und dienen als „Aktendeckel“. Sie können als eigene Dokumentart „Archivdokument“ eingerichtet werden, sodass nur das Archiv Lese- und Schreibzugriff hat.

```
−<studierendendaten>
−<personenstammdaten>
−<name>
<!-- Pflichtfeld; einziger Name-->
<nachname/> <!-- Pflichtfeld; Vornamen/einziger Name -->
<vorname/>
<!-- optional; und andere frühere Namen-->
 <geburtsname/> −<!- optional; Prüfung, dass vorangestellte, nachgestellte oder mittige Namenszusätze jeweils nur einmal vorkommen -->
−<namenszusaetze>
   <namenszusatz typ="vorangestellt"/>
 −<!- optional; Beispiel: Dr., Prof., Freifrau, Freiherr
  -->
   <namenszusatz typ="mittig"/>
   <!-- optional; Beispiel: von, zu, ben -->
   <namenszusatz typ="nachgestellt"/>
   <!-- optional; Beispiel: M.A., M.Sc., M.LL., PD, PHD -->
</namenszusaetze>
<kuenstler_ordensname/> <!-- optional -->
</name>
<geschlecht/> <!-- optional -->
−<lebensdaten>
<geburtsdatum/> <!-- Pflichtfeld -->
<geburtsort/> <!-- Pflichtfeld --> <geburtsland/> <!-- optional -->
<todesdatum/> <!-- optional -->
<todesort/> <!-- optional -->
</lebensdaten>
−<staatsbuergerschaften> <!-- optional, mehrere Staatsbürgerschaften möglich --> <staatsbuergerschaft/>
 −<!- soll zwischen 1., 2. Staatsbürgerschaft unterschieden werden?
  -->
 <staatsbuergerschaft/> </staatsbuergerschaften>
<familienstand/> <!-- optional -->
<religionszugehoerigkeit/> <!-- optional -->
<matrikelnummer/> <!-- Pflichtfeld -->
<immatrikulationsdatum/> <!-- Pflichtfeld -->
−<exmatrikulation> <!-- Pflichtfelder -->
 <grund/>
 <datum/>
</exmatrikulation>
−<hochschulzugangsberechtigungen> <!-- Pflichtfeld -->
−<hochschulzugangsberechtigung><!-- Pflichtfeld; muss mind. einmal vorhanden sein -->
   <art/>
   <!-- Pflichtfeld -->
   <abschlussdatum/>
   <!-- Pflichtfeld; kann auch nur das Jahr beinhalten -->
   <ort/> <!-- optional -->
   <land/> <!-- optional -->
   <note/>
   <!-- optional -->
                </hochschulzugangsberechtigung>
               </hochschulzugangsberechtigungen>
      −<vorbildungen>
 −<vorbildung> −<!- optional; kann sich wiederholen; Beispiel: Studienkolleg, Praktikum, berufspraktische Tätigkeit -->
   <art/>
   <!-- Pflichtfeld;  -->
   <datum/>
   <!-- Pflichtfeld; kann auch Zeitraum enthalten -->
   <ort/> <!-- optional -->
         <note/>
```

```
    <!-- optional -->
            </vorbildung>
          </vorbildungen>
−<anschriften>
<!-- mindestens eine Anschrift -->
<!-- optional-->
−<anschrift typ="heimatanschrift">
    <strasse/>
    <hausnummer/>
    <postleitzahl/>
    <ort/>
    <land/>
    <!-- optional -->
    <anschriftenzusatz/> <!-- optional -->
</anschrift>
<!-- optional -->
<!-- optional -->
−<anschrift typ="semesteranschrift">
    <strasse/>
    <hausnummer/>
    <postleitzahl/>
    <ort/>
    <land/>
    <!-- optional -->
    <anschriftenzusatz/> <!-- optional -->
</anschrift> <!-- optional -->
−<anschrift typ="">
 −<!- falls nicht unterschieden wird zwischen Heimat- oder Semesteranschrift
   -->
```

Abbildung 4: Muster der XML-Datenstruktur archivwürdiger Studierendendaten (Projekt LZA.NRW)

Der Einstieg in die digitale Langzeitarchivierung: Hybridakten

Bei der Einführung von eAkten reicht es nicht aus, sich nur auf die Archivierung der digitalen Inhalte zu konzentrieren. Hochschulinterne Regelungen wie Schriftgut- und Prüfungsordnungen sowie gesetzliche Vorgaben können erfordern, dass bestimmte Dokumente bzw. Nachweise weiterhin in Schriftform vorliegen müssen. Ein Beispiel hierfür sind Abschlusszeugnisse in Studierendenakten, die nach wie vor in Papierform ausgestellt werden und nur mit entsprechenden Sicherheitsmerkmalen wie Unterschriften und Stempeln der Hochschule Gültigkeit besitzen. Um die Authentizität und Überprüfbarkeit digitaler Dokumente sicherzustellen, sind elektronische Siegel und Signaturen notwendig. Diese lassen sich jedoch derzeit nicht in DiPS.kommunal verarbeiten. Vorhandene elektronische Siegel und Signaturen müssen somit vor einer Übernahme in DiPS.kommunal aufgelöst werden. Ihre zeitlich begrenzte Gültigkeit würde darüber hinaus regelmäßiges Nachsiegeln erforderlich machen. Aus archivfachlicher Sicht gilt zudem die Auffassung, dass die Funktionalitäten des elektronischen Langzeitarchivsystems ohnehin die Rechts- und Revisionssicherheit der darin gespeicherten Daten gewährleisten.

Die Einführung von eAkten bringt somit für die Archive auch die Herausforderung mit sich, Hybridakten verwalten zu können, die sowohl digitale Inhalte als auch Papierrestakten umfassen. Um die Vollständigkeit der Akten sicherzustellen (und gleichzeitig eine doppelte Datenhaltung ausschließen zu können), ist es notwendig, die digitalen und analogen Bestandteile einer Akte sinnhaft miteinander zu verknüpfen, beispielsweise über einen gemeinsamen

Identifikator sowie Verweise im AFIS. Diese Verknüpfung ist zudem entscheidend für die Auffindbarkeit der einzelnen Aktenteile.
Die Umstellung erfordert eine Neugestaltung der Arbeitsabläufe im Records Management. Dabei geht es nicht nur darum, Ressourcen effizient einzusetzen, sondern auch neue Arbeitsschritte zu integrieren, die durch die hybride Datenhaltung und die Anforderungen an die digitale Langzeitarchivierung entstehen. Hierzu zählen etwa die Entwicklung von Verfahren zur Überprüfung der Datenintegrität sowie die kontinuierliche Anpassung an technologische Veränderungen.
Die Einführung von eAkten steht an vielen Hochschulen derzeit im Vordergrund. Gleichzeitig bindet die Entwicklung von Scanstrategien, die möglichst viele Unterlagen digital zugänglich machen sollen, bereits erhebliche Kapazitäten. Parallel dazu muss eine Langzeitarchivierungsstrategie für eAkten(-teile) entwickelt werden, die von Anfang an in diese Prozesse integriert ist. Nur so kann eine nachhaltige und zukunftssichere Verwaltung sowohl der digitalen als auch der analogen Inhalte gewährleistet werden.

Fazit und Ausblick

Die Bedeutung von Studierendenakten für die Hochschulen in Nordrhein-Westfalen geht weit über ihren historischen Wert hinaus: Sie sichern die Nachvollziehbarkeit administrativer Entscheidungen und bieten Rechtssicherheit. Die Digitalisierung stellt dabei sowohl eine Herausforderung als auch eine Chance dar: Einerseits müssen Hochschulen den gesetzlichen Anforderungen gerecht werden und ihre Akten künftig digital führen. Andererseits bietet die digitale Transformation die Möglichkeit, Archivierungsprozesse von Anfang an mitzudenken und in Zusammenarbeit mit Archivexpert:innen nachhaltige Lösungen zu entwickeln.
Bei der Archivierung von eStudierendenakten ist vieles zu beachten – von den rechtlichen Rahmenbedingungen über die Inhalte in Form von Primär- und Metadaten bis hin zur Art und Weise, in der verschiedene datenführende Systeme wie Fachverfahren und das DMS miteinander kommunizieren. In diesem Zusammenhang muss, zumindest übergangsweise, auch die Verwaltung von Hybridakten bedacht werden.
Die Wahl einer geeigneten Langzeitarchivierungslösung ist dabei ebenso essenziell wie die Bereitstellung von Personalressourcen und die Neugestaltung von Arbeitsabläufen, beispielsweise im Bereich des Records Managements. Mit den eAkte-Projekten der Digitalen Hochschule NRW wird eine strukturierte und rechtssichere digitale Ablage und Verarbeitung sowie Archivierung von eAkten vorangetrieben, die für die Zukunft der Hochschulverwaltung von zentraler Bedeutung sind.

Für Hochschulen ist es entscheidend, die digitale Archivierungsstrategie von Anfang an sorgfältig zu planen und kontinuierlich anzupassen. Die effektive Umsetzung dieser Strategie wird nicht nur die Verwaltung von Studierendendaten optimieren, sondern auch sicherstellen, dass alle archivierten Daten langfristig verfügbar und geschützt bleiben. Die Gründung neuer Hochschularchive und die rechtzeitige Implementierung von Langzeitarchivierungslösungen sind daher ausschlaggebend, um auch den langfristigen Herausforderungen der digitalen Transformation erfolgreich zu begegnen.

Bibliografie

Archivgesetz NRW (2010), *Gesetz über die Sicherung und Nutzung öffentlichen Archivguts im Lande Nordrhein-Westfalen (Archivgesetz Nordrhein-Westfalen – ArchivG NRW) vom 16.3.2010,* zuletzt geändert am 16. September 2014, in: GV. NRW. S. 188.

Becker, T. et al. (2009), *Dokumentationsprofil für Archive wissenschaftlicher Hochschulen.* Saarbrücken: Universität des Saarlandes.

Brühbach, N. (2010), 'Das Referenzmodell OAIS' in Neuroth, H. et al (Hrsg.), *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. Kapitel 4. Das Referenzmodell OAIS – Open Archival Information System.* Göttingen: Niedersächsische Staats- und Universitätsbibliothek, Kapitel 4.2, http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_474.pdf (17.9.2024).

Datenschutzgesetz NRW (2018), *Datenschutzgesetz Nordrhein-Westfalen (DSG NRW)* vom 17. Mai 2018, in: GV. NRW. S. 244.

Digitale Hochschule NRW (2024), *eAkte Projekte*, https://e-akte.dh.nrw/unsere-leistungen/e-akte-projekte-1 (17.9.2024).

E-Government-Gesetz NRW (2016), *Gesetz zur Förderung der elektronischen Verwaltung in Nordrhein-Westfalen (E-Government-Gesetz Nordrhein-Westfalen – EGovG NRW)* vom 8. Juli 2016, zuletzt geändert am 1. Februar 2022, in: GV.NRW. S. 551.

Kaltenbach, M. et al (2022), *E-Studierendenakte.NRW. Hochschulmaster zur Einführung einer E-Studierendenakte.* Stand 30.11.2022, https://ilias.huef-nrw.de/ilias/goto.php?target=file_39053_download&client_id=huefilias (18.10.2023).

Koordinierungsstelle für IT-Standards (KoSIT) (Hrsg.) (2021), *xdomea 3.0.0 – Spezifikation. XÖV-Standard für den IT-gestützten Austausch und die IT-gestützte Aussonderung behördlichen Schriftgutes,* https://www.xrepository.de/api/xrepository/urn:xoev-de:xdomea:kosit:standard:xdomea_3.0.0:dokument:Spezifikation_xdomea_3.0.0 (18.10.2023).

LWL/ Stadt Köln (2022), *Benutzerhandbuch DiPS.kommunal*.

nestor-AG Archivstandards (Hrsg.) (2023), *nestor-materialien 25. Archivierung von Studierendendaten aus Fachverfahren*, https://d-nb.info/1294122746/34 (18.10.2023).

IV.

E-MAIL, WEBARCHIVIERUNG, SOCIAL MEDIA

Multimodale Ansätze der Webarchivierung: Einblick in das Konzept des Erzbischöflichen Archivs Freiburg

Tony Franzky

Webarchivierung ist aufgrund der fortwährenden technischen, kulturellen und gesellschaftlichen Weiterentwicklung des Internets bei zeitgleicher Kurzlebigkeit formathafter Standards eine der herausforderndsten Bereiche der Digitalen Archivierung. Zudem müssen Archive zeitnah reagieren und sich diesen Aufgaben stellen, da sich Inhalte wegen der hohen Fluktuation von Trends, Nutzungsgewohnheiten und steten Disruptionsprozesse des Webs schnell transformieren oder dauerhaft verloren gehen.

Das Erzbistum Freiburg betreibt im Bereich webgestützter Anwendungen ein Contentmanagementsystem mit fast 500 Mandanten. Hinzu kommen diverse Social-Media-Kanäle unterschiedlicher medialer Domänen, Intranetinhalte und teilwebgestützte Fachverfahren. Um sowohl diesem quantitativen Umfang als auch den mediumbedingten Herausforderungen und archivfachlichen Anforderungen gerecht zu werden, hat das Erzbischöfliche Archiv Freiburg ein skalierbares, multimodales Konzept zur Websitearchivierung entwickelt, um die Internetpräsenz des Erzbistums möglichst hochautomatisierbar bei geringer Datenlast in eine funktionale Überlieferung zu überführen.

Ausgangslage im Erzbistum Freiburg

Das Erzbischöfliche Archiv Freiburg ist laut Anordnung über die Sicherung und Nutzung der Archive der katholischen Kirche (KAO) verpflichtet, „Unterlagen, die das Wirken der Kirche dokumentieren, der Rechtssicherung dienen oder von bleibendem Wert für Wissenschaft, Forschung oder kirchliche Bildungsarbeit“ (KAO, §3(4)) sind, zu archivieren. Dies gilt explizit auch für elektronische Unterlagen. Diese „sind, sofern sie laufenden Aktualisierungen unterliegen, ebenfalls der Archivierung anzubieten“ (KAO, §6(3)). Hieraus ergibt sich der begründete Sammlungsauftrag für das Erzbischöfliche Archiv Freiburg, auch Websites zu archivieren.

Das Erzbistum Freiburg hat seit spätestens 1998 die ersten Inhalte im Internet veröffentlicht, vereinzelte Kirchgemeinden auch deutlich früher. Eine eigene Webpräsenz mit einem systematisierteren Angebot wurde 2003 etabliert. Dies hat sich schließlich zu einem eigenen Contentmanagementsystem (CMS) mit aktuell über 500 Mandanten weiterentwickelt (Franzky, 2024). Die Bandbreite der Webauftritte ist dabei sehr divers. Neben den Seelsorgeeinheiten betreiben auch verschiedene Bildungswerke, Kirchliche Verbände, Dekanate und weitere

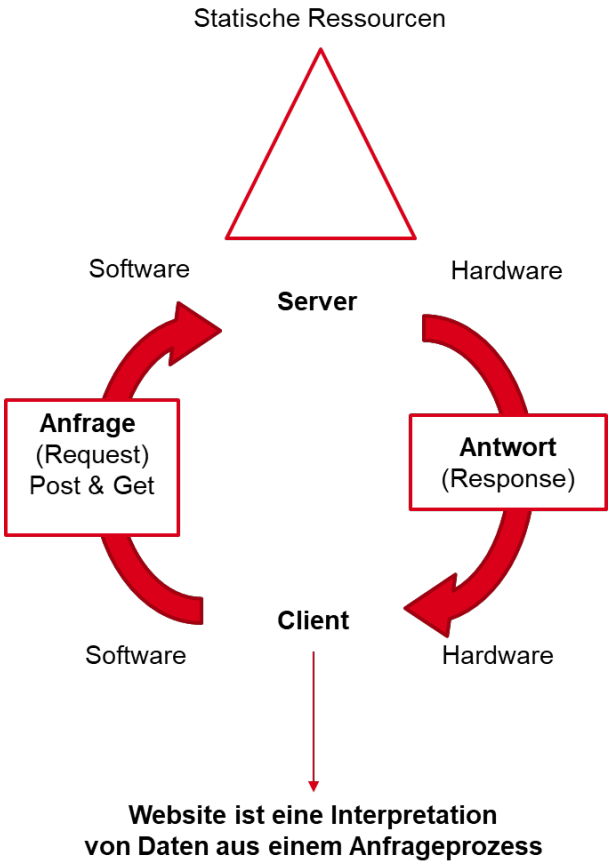
Struktureinheiten eigene Webangebote. Hinzu kommt eine unbekannte Anzahl von Webauftritten, die nicht über das bistumseigene CMS gehostet und verwaltet werden, sondern beispielsweise auf Privatinitiativen zurückgehen oder von lokalen Kirchgemeinden getragen werden. Daneben werden auch auf Webtechnologie basierende Produkte wie ein Intranet, Apps sowie Social-Media-Kanäle betrieben.
Da die Ausgangslage sehr vielfältig ist, war es notwendig, ein adaptives Modell zu etablieren, um sowohl den über das eigene CMS verwalteten Webinhalten als auch den individuell betriebenen Webauftritten Rechnung tragen zu können.

Websites und Pflichtexemplare

Vereinzelt scheint es noch beharrliche Lehrmeinung zu sein, dass Websites im engeren Sinne Publikationen darstellen und daher Autor:innen bzw. Seitenbetreiber:innen gegenüber z.B. Bibliotheken pflichtexemplarpflichtig sind, weshalb die Sicherung von Websites eher in der Hand der Betreiber:innen bzw. Produzierenden liegt. Diese Position verkennt jedoch relevante Aspekte zur Erhaltung digitalen Kulturguts ebenso wie die soziale Praxis bei der Entstehung von Webinhalten.
Zwar gibt es z.B. in vielen Landesbibliotheksgesetzen eine Pflichtexemplarregelung für elektronische Publikationen, doch wurde diese fast immer von physischen Druckwerken her gedacht, bei denen sich Form und Inhalt wechselseitig bestimmten und das entstehende Produkt institutionalisierten Verlags- und Vervielfältigungstätigkeiten unterlag. Inzwischen können Werke jedoch relativ frei ihr Medium wechseln, und Medien ihrerseits müssen nicht mehr physisch gebunden sein. Gleichzeitig ist der redaktionelle Prozess zur Erstellung von Webinhalten dezentral und teilweise informell organisiert. Hinzu kommt, dass genuine Web-Inhalte (im Gegensatz zu digitalen Publikationen) nicht notwendigerweise persistent verortet oder dauerhaft verfügbar sein müssen, nicht an Auflagen oder Versionierungen gebunden sind, ihre Distributions- und Verbreitungswege teilweise unkontrolliert sind und sie im Zweifelsfall nicht einmal inhaltlich abgeschlossen sein müssen. Darüber hinaus können Inhalte aufgrund ihres Umfangs als Akzidenzen eingestuft und damit von Pflichtexemplarregelungen ausgenommen sein. Zudem liegen Web-Inhalte aufgrund ihrer technischen Struktur nicht notwendigerweise in Dateiformaten vor, die langfristig stabil gespeichert oder auch in Zukunft geöffnet und dargestellt werden können. All diese Eigenschaften stellen den ursprünglichen Werkbegriff hinter dem Begriff „Pflichtexemplar“ stark in Frage und nehmen damit die kulturellen Gedächtnisinstitutionen stärker in die Verantwortung, sich proaktiv mit der Archivierung von Websites auseinanderzusetzen.

Was ist eine Website?

Technisch gesehen ist eine Website eine menschenlesbare, transmediale Interpretation einer Serverantwort (Response) durch ein technisches Endgerät. Dem geht eine Anfrage (Request) voraus. Die dabei übertragenen Daten können sich aus unterschiedlichen Quellen speisen. So können sie beispielsweise aus Datenbanken oder statischen Ressourcen des Servers stammen oder von Webseiten Dritter aus eingebettet oder verlinkt sein.



Die genaue Darstellung einer Website hängt stark vom verwendeten Endgerät auf der Clientseite ab (Smartphone, Tablet, Desktop-PC, ...) und der darauf installierten Software (Webbrowser, eventuell installierte Plugins, verwendete Laufzeitumgebungen, ...) ab. Zudem passt die moderne Webentwicklung den Inhalt und das Design einer Seite teilweise bereits serverseitig individuell an das anfragende Endgerät an. Man spricht hier auch von responsivem Webdesign. Diese Umstände erschweren jedoch die Archivierung an verschiedenen Stellen. So können neben den inhaltlichen Aspekten, die als Webcontent vorliegen, auch eine Reihe von gestalterischen oder funktionalen Aspekten eine sammlungsbegründende Rolle spielen, ebenso wie übertragene Metadaten, das visuelle Erscheinungsbild, das Verhalten der Website (Look & Feel) und anderes mehr.

Darüber hinaus steht aus archivischer Sicht die Speicherung der übermittelten Daten in langzeitstabilen Dateiformaten im Fokus. Gerade bei derart volatilen Technologien wie Websites ist dies jedoch eine nicht ganz einfache Aufgabe. Darüber hinaus sollte die Weiterverarbeitung der Daten durch das Archivpersonal möglichst unkompliziert und ohne zusätzlichen technischen Aufwand möglich sein. Beispielsweise, um digitale Archivalien für Nutzungsszenarien aufzubereiten oder in bestehende Archiv-IT-Infrastrukturen einzubinden (z.B. hinsichtlich der Recherchierbarkeit).

Ein weiterer, nicht zu unterschätzender Aspekt ist, dass je nach Website auch die Gefahr besteht, dass unerwünschte Inhalte wie Spam, Adware oder sogar Malware mit archiviert werden. Beispielsweise dann, wenn Webseiten Dritten die Möglichkeit bieten, eigene Inhalte, Downloads oder Links bereitzustellen. Dies kann beispielsweise durch Kommentarfunktionen oder den Betrieb eines Forums der Fall sein.

Multimodale Methoden der Webarchivierung

Um diese unterschiedlichen Probleme zu adressieren, haben sich verschiedene Ansätze der Webarchivierung entwickelt und etabliert. Dazu gehören die generelle Speicherung des Server-Outputs auf dem Endgerät (Crawling, Speichern unter), der Datenbankschnitt, die Emulation von Websites sowie das Abfotografieren oder Abfilmen von Seiten. Diese Methoden seien im Folgenden kurz beschrieben.

Einmaliges Speichern (manchmal als statisches Verfahren bezeichnet wie bei Frech & Grossmann, 2024) oder *systematisches Speichern / Crawlen* (manchmal als dynamisches Verfahren bezeichnet, ebd.): Dabei werden ein oder mehrere Serveraufrufe durchgeführt und die Serverantwort (in der Regel ein kompilierter HTML-strukturierter Bitstream mit verschiedenen verknüpften Ressourcen) gespeichert. Dies kann direkt auf einem Dateisystem als Einzeldateien erfolgen oder in einem Containerformat wie z.B. WARC-Dateien. Die Erzeugung neuer Zeitschichten kann dabei auch inkrementell erfolgen. Beim inkrementellen Crawling werden z. B. Daten, die sich über verschiedene Zeitpunkte hinweg nicht ändern (z. B. Hintergrundbilder oder Begrüßungsseiten), nur einmal gespeichert und nicht zu jedem Crawlzeitpunkt. Inkrementelles Crawling wird jedoch nicht von allen Programmen unterstützt.

Vorteile dieser Methode sind, dass Inhalte und verknüpfte Dateien erhalten bleiben und im Falle eines systematischen Crawlings auch eine Vielzahl von Links und Verknüpfungen. Externe Inhalte und eingebettete Elemente (wie z. B. Kartenmaterial oder eingebettete Videos von Streaming-Plattformen) können jedoch schwierig zu verarbeiten sein. Das Crawlen dieser Informationen kann dabei auch ein rechtliches Problem darstellen. Darüber hinaus kann diese

Variante einen höheren Aufwand bei der Aufbereitung der Daten erfordern, z. B. um eine offline lesbare Kopie mit gleichem Funktionsumfang zu erzeugen.

Zudem liegen Daten, die heruntergeladen werden, nicht automatisch in langzeitarchivfähigen Formaten vor, da prinzipiell jedes Dateiformat im Internet zumindest zum Download angeboten werden kann. Moderne Browser akzeptieren selbst bei eingebetteten üblichen Medien wie Bildern oder Audiodateien eine Vielzahl auch archivisch ungeeigneter Formate.

Wie bereits erwähnt, hängt das Verhalten und das visuelle Erscheinungsbild einer Website sehr stark vom aufrufenden Browser ab. Gerade im Bereich Design und Seitenverhalten entwickeln sich technische Standards weiter und HTML-Tags oder eingebettete Technologien können auch rasch veralten. Dies war in der Vergangenheit bereits häufig der Fall. Beispiele hierfür sind veraltete Features und Tags in HTML 5 (siehe: Web Hypertext Application Technology Working Group, 2024) oder auch veraltete Laufzeitumgebungen und Plattformen für multimediale Inhalte wie Flash (am 31. Dezember 2020 eingestellt, Adobe Systems Software, 2021), Shockwave (eingestellt am 9. April 2019, Adobe Systems Software, 2019) oder die clientseitige Verwendung von VBScript (eingestellt am 13. August 2019, Microsoft, 2019). Streng genommen müsste also auch ein aktueller Browser gesichert werden, um den bisherigen Stand von Technologie zu dokumentieren und aktuelle Darstellung und Verhalten einer Webseite auch später noch reproduzieren zu können.

Datenbankabzug: Die meisten modernen Webanwendungen basieren auf im Hintergrund arbeitenden Datenbanken. Eine Kopie dieser Daten ermöglicht es, relevante Inhalte (z. B. Benutzerdaten, angelegte Unterseiten, Postings etc.) zu exportieren und in der Regel in einem Tabellenformat zu speichern. Dies hat den Vorteil, dass die Informationen bereits gut strukturiert vorliegen, durchsuchbar sind und in der Regel ohne großen Aufwand in einem langzeitstabilen Format gesichert werden können. Unterschiedliche Zeitschichten von Inhalten können je nach Datenbankstruktur sauber voneinander getrennt werden, so dass eine redundanzfreie Überlieferungsbildung je nach Datenbankstruktur möglich ist. Nachteil dieser Methode ist jedoch, dass die visuellen Elemente und das Design einer Website verloren gehen. Zudem ist ein Zugriff auf die Datenbank einer Website erforderlich, der in der Regel einen Administrationszugang voraussetzt.

Quellcodekopie / Emulation: Hierbei handelt es sich um eine Strategie, bei der alle für den Betrieb einer Website erforderlichen Ressourcen (Datenbanken, Quellcode, Laufzeitumgebungen, Serverinfrastruktur usw.) übernommen werden mit dem Ziel, sie zu einem späteren Zeitpunkt wieder in Betrieb nehmen zu können. Diese Strategie ist recht aufwändig und erfordert umfangreiche Kenntnisse in Webtechnologien, Serveradministration und

Emulationsmöglichkeiten, da wie beschrieben auch die Serverumgebung als Ganzes erhalten werden muss. Dies kann z. B. bei lizenzierten Produkten oder proprietärer Software (z. B. Oracle Datenbanken, Microsoft Server) schwierig sein. Darüber hinaus kann es aufwändig sein, veraltete Softwareversionen (z. B. alte Versionen von PHP, MySQL etc.) auch in naher Zukunft technisch lauffähig und sicher zu halten. Ähnlich wie beim Crawling kann es zudem technisch schwierig sein, das heutige Design und Verhalten der Website auf zukünftigen Browsergenerationen abzubilden, da auch dieser Ansatz der Weiterentwicklung unterliegt. Dieser Ansatz kann jedoch, ähnlich wie der Datenbankabzug, nur realisiert werden, wenn ein voller administrativer Zugriff auf die Datenbank einer Webseite besteht.

Bildschirmfotos / Screenshots: Bei dieser Methode werden Bildschirmfotos von einer Website angefertigt. Dies kann systematisch oder punktuell geschehen. Der Vorteil dieser Methode ist, dass die visuellen Informationen einer Webseite erhalten bleiben. Leider ist eine einfache Recherchierbarkeit nur dann gegeben, wenn zusätzliche Metadaten erfasst und den Screenshots beigefügt werden. Die Speicherung in langzeitstabilen Dateiformaten ist jedoch kein Problem. Aufgrund der weiten Verbreitung von Bildformaten ist auch die Erstellung von Nutzungsderivaten oder eine eventuelle Weiterverarbeitung relativ unkompliziert möglich und in gängigen Anwendungen, wie z. B. digitalen Lesesälen, implementierbar.

Bildschirmaufzeichnung / Screencasts: Ähnlich wie bei der Screenshot-Methode werden auch hier die Daten über die visuelle Darstellung am Bildschirm generiert. In diesem Fall wird eine Website besucht und verschiedene Funktionalitäten angeklickt und genutzt. Dies wird mittels einer Software als Video aufgezeichnet und gespeichert. Später kann dann durch dieses Video das Verhalten der Webseite nachvollzogen werden. Es wird jedoch nicht möglich sein, zu einem späteren Zeitpunkt selbst mit einer Seite zu interagieren. Soll dieses nachbearbeitet werden, erfordert dies ein gewisses technisches Wissen, z. B. im Bereich des Videoschnitts. Zudem ist dieses Vorgehen bedingt skalierbar, da das systematische Besuchen und „Durchklicken“ einer Website in der Regel Echtzeit passieren müssen.

Betrachtet man nun die fünf genannten Methoden gespiegelt auf die inhaltlichen Ressourcen und funktionalen Aspekten einer Website, so zeigt sich, dass die verschiedenen Ansätze der Webarchivierung den unterschiedlichen Aspekten der Webarchivierung unterschiedlich gut gerecht werden.

		Methoden der Websitearchivierung				
		Speichern / Crawling	Daten-bank-schnitt	Emulation	Screens-hot	Screencast
Archivierbare Ressourcen	Inhalt	x	x	x	(x)	(x)
	Downloadbare Dateien	x	(x)	x		
	Eingebettete Inhalte	(x)	(x)	(x)		
	Links / Verknüpfungen	x	(x)	x		
	Metadaten	x	(x)	x		
	Visuelle Erscheinung	(x)		(x)	x	x
	Verhalten der Seite	(x)		(x)		x
Fachl. Anforderungen	Langzeitstabile Formate		x		x	x
	Gefahr von Malware	x	(x)	x		
	Bearbeitbarkeit / Nutzungsderivate		x		x	(x)
	Durchsuchbarkeit		x	(x)		
	Inkrementelles Speichern	(x)	(x)			
	Speicherbedarf	mittel bis hoch	gering bis mittel	sehr hoch	gering	mittel

Dies führte zur weiteren Überlegung, dass keine einzige derzeitige Methode der Websitearchivierung in der Lage ist, das gesamte Erscheinungsbild einer Website (Inhalt, visuelles Verhalten, Verlinkung, etc.) abzubilden und gleichzeitig allen archivfachlichen Anforderungen (langzeitstabile Formate, Recherchierbarkeit, Metadatenübernahme) gerecht zu werden. Darüber hinaus spielen auch ressourcenökonomische Überlegungen eine zentrale Rolle. Gerade Verfahren wie das nicht-inkrementelle Crawling oder die versionsweise Übernahme von Quellcode erzeugen enorme Redundanzen, die gerade bei sich ggf. nur geringfügig ändernden Seiteninhalten in keinem Verhältnis stehen.

Daraus ergibt sich der Ansatz, multimodale Verfahren zur Archivierung von Websites einzusetzen, um einerseits möglichst viele Aspekte einer Website auch für künftige Generationen zu sichern und andererseits unterschiedlichen archivfachlichen Anforderungen gerecht zu werden, um damit dem in der KAO formulierten Sammlungsauftrag für elektronische Unterlagen, zu entsprechen.

Exkurs: Das Problem langzeitstabiler Dateiformate in WARC-Containern

Das Web Archive Format (WARC) ist ein Containerformat, das die Möglichkeit bietet, verschiedene Webressourcen zu bündeln und zu aggregieren. Es wird von einer Vielzahl von Archiven empfohlen und verwendet. Das Format stellt eine Weiterentwicklung des ARC-Formats dar und wurde 2009 ISO standardisiert (Weimer & Schoger, 2021). Die aktuelle Version ist als ISO 28500:2017 verfügbar (für weitere Informationen siehe auch Library of Congress, 2024).

Das Format löst teilweise das Problem verteilter Ressourcen in der Webarchivierung, indem alle Informationen (egal ob Datenbankinhalte, statische Ressourcen oder verlinkte Inhalte) inklusive Metadaten akkumuliert werden. Dadurch wird zwar ein hohes Maß an Kontextsicherung erreicht. Das Problem langzeitstabiler Archivformate für die Webarchivierung wird damit aber keineswegs gelöst. So kann prinzipiell jeder Dateityp in einen WARC-Container geschrieben werden, wodurch nicht gewährleistet ist, dass archivierte obsolete Dateiformate zu späteren Zeitpunkten noch interpretierbar sind. Da Websites wiederum prinzipiell beliebige Dateiformate enthalten, ist hier ein Migrationsschritt dringend erforderlich.

In den Spezifikationen von WARC selbst sind auch Conversion Records vorgesehen (ISO 28500:2017), die die Hinterlegung von migrierten, langzeitstabilen Dateirepräsentationen ermöglichen. In der Praxis werden diese jedoch kaum umgesetzt und im IIPC GIT als unterspezifiziert problematisiert (Ato, 2018).

Zudem sind WARC-Container, wenn sie nicht inkrementell erzeugt werden, wie andere Crawls sehr redundant. Hinzu kommt, dass sie nur mit technischer Fachkenntnis weiterverarbeitet werden können, was in der Gesamtschau der Websitearchivierungsmethoden also dringend zu berücksichtigen ist bei der Wahl der Zielformate und der Lösung des Problems langzeitstabiler Dateiformate.

Das Bewertungsmodell

Um hier eine fundierte archivfachliche Bewertung vornehmen zu können, hat das Erzbischöfliche Archiv begonnen, alle Webauftritte zu erfassen und dabei sowohl technische Besonderheiten wie Login-Systeme, Newsletter, eingebundene Social-Media-Kanäle, RSS-Feeds etc. als auch administrative Aspekte wie verantwortliche Personen und Redakteur:innen zu erheben und mit einer Bewertungseinschätzung zu versehen. Die Bewertungsentscheidung wird nach dem Vieraugenprinzip getroffen. Sollte diese mit archivwürdig ausfallen, besteht die Möglichkeit, über drei Kategorien die konkrete Art und Häufigkeit der Archivierung zu steuern.

Für die eigentliche Webseitenarchivierung selbst stehen technisch folgende Optionen zur Verfügung:

- Nicht-inkrementelles Crawling (über einen Heritrix-Crawler)
- Spidering (Sonderform des Crawlings: Erfassung der Linkstruktur einer Seite ohne Speicherung weiterer Inhaltsdaten)
- Erstellung von systematischen Screenshots (über ein selbst entwickeltes Tool)
- Erstellung eines Datenbankabzugs (sofern die Diözese auch Betreiberin der Website ist)
- Vollständiger Quellcodeabzug (wird nicht wahrgenommen)

Alle als nicht archivwürdig bewertete Seiten (Kategorie V) werden nicht gesichert. Die als archivwürdig eingestuften Websites werden in folgende Kategorien eingeteilt.

- Kategorie A-1: Website wird regelmäßig mit einer der verfügbaren Methoden gesichert. Bevorzugt werden regelmäßige Datenbankabzüge oder Screenshots.
- Kategorie A-2: Website wird regelmäßig mit mindestens einer der verfügbaren Methoden gesichert. In der Regel Datenbankabzüge und Screenshots. Initial wird zudem Crawl durchgeführt.
- Kategorie A-3: In Abhängigkeit von Dynamik und Novitätscharakter der Website werden zwei der Methoden regelmäßig eingesetzt. Zusätzlich werden in größeren Zeitabständen Crawls durchgeführt.

Die Häufigkeit der jeweiligen Aktionen richtet sich nach Entwicklung, Lebenszyklus und Einschlägigkeit der jeweiligen Website für das Sammlungs- und Dokumentationsprofil des Erzbischöflichen Archivs Freiburg. Die Archivierungsfrequenz kann zwischen anlassbezogen, monatlich, jährlich oder bis zu alle zwei Jahre variieren.

Vorteile und Entwicklungsfelder des multimodalen Ansatzes

Das Modell selbst ist gut skalierbar und anpassbar an unbekannte Webseitenformen, projektbezogene Auftritte und private Initiativen. Flankiert wird dies mit Extraansätzen für die jeweiligen Social-Media-Kanäle, die im Jahrestakt gesichert werden. Das Modell ist in einem hohen Maße automatisierbar und produziert (bis auf Initialcrawls) langzeitstabile Formate. Herzstück sind dabei der Bewertungskatalog der Websites sowie die vorangegangene Erfassung und Bewertung der Inhalte, die im Wesentlichen die Steuerung für die Ressourcenverteilung übernehmen. Die Erstellung und Listung kann in dieser Phase viel Zeit binden. Eine Änderung der Bewertungsentscheidung von Seiten für eine modifizierte Archivierung ist vorgesehen, sofern es etwa zu einer thematischen oder technologischen Verschiebung von Einzelseiten kommt.

Sehr stark multimediale Inhalte können aufgrund technologischer Grenzen mitunter nur schwierig oder herausgelöst verarbeitet werden. Nichtinkrementelles Crawlen einer Videoseite etwa würde eine große Datenlast und Redundanz ohne Mehrwert erzeugen. Zudem ist für das

vollständige Entfalten der gesamten Strategie auch zuweilen die Volladministrationsmöglichkeit von Inhalten nötig (z. B. redaktioneller Zugang auf das CMS), da sonst Teile des multimodalen Ansatzes, wie etwa ein Datenbankabzug, nicht realisiert werden können. Dies ist jedoch besonders bei Drittanbietern oder privaten Initiativen sowie verwaisten Seiten nicht immer gegeben.

Bekannte Schwierigkeiten bei den einzelnen Webarchivierungsmethoden können durch den multimodalen Ansatz stellenweise ausgeglichen werden. Generelle Schwierigkeiten der Einzelnen Webarchivierungsmethoden bestehen jedoch nach wie vor. So haben etwa Screenshot-basierte Archivierungen oft das Problem mit eingeblendeten Cookie Consents, Paywalls oder Loginsystemen. Crawling-basierte Verfahren haben nach wie vor das Problem nicht langzeitgeeigneter Dateiformate. Hinzu kommen allgemeine Aspekte wie Bot Detection Tools oder Captchas auf Webseiten, die ein systematisches Abrufen von Informationen verhindern sollen.

Darüber hinaus kann dieser Ansatz insbesondere für kleinere Institutionen schwierig umzusetzen sein, da diese häufig vor dem Problem stehen, Webseiten überhaupt archivieren zu können. Das Anfertigen von Screenshots zusätzlich zu Crawls und Datenbankabzügen kann bei sehr begrenzten Ressourcen und geringem Automatisierungsgrad einen untragbaren Mehraufwand bedeuten, der angesichts der vielfältigen Aufgaben nicht zu bewältigen ist. Dennoch ist dieser Ansatz im Hinblick auf Speicherressourcen und Zuverlässigkeit der Überlieferungsbildung sehr performant.

Bibliografie

Ato (2018), ‘Conversion’ records are underspecified, https://github.com/iipc/warc-specifications/issues/40 (30.9.2024).

Adobe Systems Software (2019), *Shockwave*, https://helpx.adobe.com/de/support/programs/support-options-free-discontinued-apps-services.html#shockwave (30.9.2024).

Adobe Systems Software (2021), *Flash wurde eingestellt*, https://www.adobe.com/de/products/flashplayer/end-of-life-alternative.html (30.9.2024).

Anordnung über die Sicherung und Nutzung der Archive der katholischen Kirche (2014), *Kirchliche Archivordnung – KAO*, https://www.kirchenrecht-ebfr.de/document/141/search/kirchliches%2520archiv%2520recht (30.9.2024).

Franzky, T. (2024), ‘Das Erzbistum im WWW’, *Archiv-Blog: Der Blog des Erzbischöflichen Archivs Freiburg*, https://www.ebfr.de/archivblog/detail/nachricht/id/191572-das-erzbistum-im-www/?cb-id=12327217 (30.09.2024).

Frech, A. und Grossmann, Y. V. (2024), ‘Das Internet vergisst doch – Handreichung für die Archivierung von wissenschaftlichen Webseiten’. doi: 10.5281/ZENODO.10556375.

ISO 28500:2017 (2017), *Information and documentation — WARC file format*, https://www.iso.org/standard/68004.html (30.9.2024).

Library of Congress (2024), *Sustainability of Digital Formats: Planning for Library of Congress Collections – WARC, Web ARChive file format*, https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml (30.9.2024).

Microsoft (2019), *An update on disabling VBScript in Internet Explorer 11*, https://blogs.windows.com/msedgedev/2019/08/02/update-disabling-vbscript-internet-explorer-windows-7-8/ [30.09.2024].

Web Hypertext Application Technology Working Group (2024), *HTML - Living Standard, § 16 Obsolete features*, https://html.spec.whatwg.org/multipage/obsolete.html [30.09.2024].

Weimer, K. und Schoger, A. (2021), ‘Das Dateiformat WARC für die Webarchivierung’, *Nestor Thema* 15, urn:nbn:de:0008-2021042614.

Archivierung von Social Media Data durch DSGVO-konformen Abruf: Ein Praxisbericht

Dominik Feldmann

Allgemeine Grundlagen

Öffentliche Diskussionen und öffentliche Meinungsbildung sind heutzutage ohne Social Media und den Einfluss von Facebook, Instagram, TikTok, X und Co. kaum denkbar. Aus Sicht einer kommunalen Verwaltung besitzen Social-Media-Kanäle noch weitere Möglichkeiten wie eine Information der Bevölkerung über verschiedene Themen oder eine (werbende) Öffentlichkeitsarbeit für die eigene Stadt. Nicht nur deswegen dürfte es über die grundsätzliche Archivwürdigkeit von Social-Media-Accounts heutzutage keine großen Diskussionen mehr geben. Vielmehr stellt sich aus Sicht eines Kommunalarchivs die Frage, wo man anfängt und wo man aufhört. Neben den Social-Media-Auftritten der eigenen Behörden können im Rahmen einer ergänzenden Überlieferungsbildung ebenso Kanäle von Stadträten und Stadträtinnen, anderen Personen des öffentlichen Lebens oder von Vereinen, Bürgerinitiativen und anderen Gruppierungen oder Institutionen in den Fokus rücken.

Nachdem die Frage der Archivwürdigkeit unstrittig ist und auch die rechtlichen Aspekte der Archivierung von Social-Media-Accounts im Auftrag des Archivs der sozialen Demokratie mittlerweile diskutiert und dargelegt worden sind (Walz und Marquet, 2022), ist es weiterhin vor allem die Frage des „Wie“, die die Archivwelt beschäftigt. Wie können und sollen Social-Media-Kanäle sinnvoll übernommen und archiviert werden?

Hierfür existieren bisher verschiedene Möglichkeiten, die wiederum abhängig von der jeweiligen Plattform sind. In der Regel lassen sich die Lösungsansätze jedoch in zwei Gruppen einteilen: Entweder wird ein gewisses technisches/IT-fachliches Knowhow benötigt oder es wird mit einem Dienstleister zusammengearbeitet, der eine entsprechende finanzielle Entlohnung fordert. Dass beide Arten erfolgsversprechend sein können, wurde bereits von verschiedenen Archiven gezeigt. So scheint beispielsweise Twint eine geeignete Lösung für die Archivierung von Twitter/X-Accounts zu sein (Worm, 2021). Das Stadt- und Stiftsarchiv Aschaffenburg hat dagegen mit einem Dienstleister sehr beachtliche Ergebnisse erzielt (Schuck, 2022), die bereits für Twitter auch online einsehbar sind. Anders sieht es dagegen mit einem nicht erfolgreichen Versuch aus, Faceboook-Daten über die von der Social-Media-Plattform bereitgestellte Schnittstelle Graph-API zu archivieren. Zwar erzielte das Projekt Laurentius in Kooperation mit dem Archiv für Christlich-Soziale Politik zu Beginn Erfolge (Köhn, 2013; Burkert, 2013), doch nach

einer Änderung der API Seitens Facebook wurde das Projekt nach Kenntnisstand des Autors eingestellt.

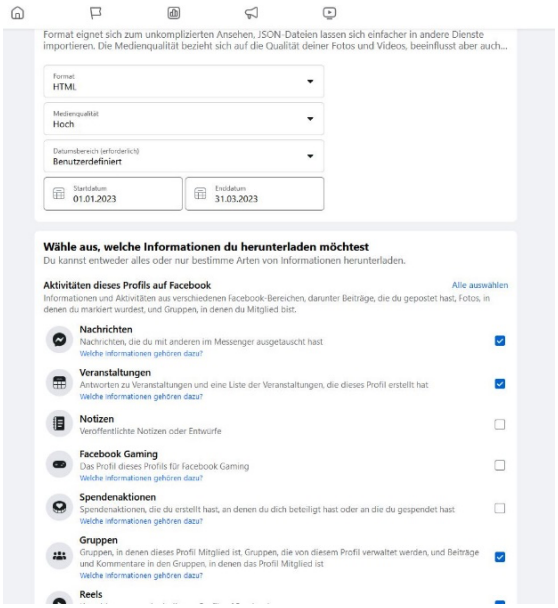
Aufgrund des traditionell knappen Budgets eines Archivs, der mangelnden Unterstützung der städtischen IT – die jedoch nachvollziehbar begründet worden ist – sowie der Unsicherheit über mögliche Veränderungen in API von Social-Media-Betreibern, hat sich das Stadtarchiv Augsburg auf die Suche nach einer alternativen Form des Übernehmens von Social-Media-Daten gemacht. Im Fokus stand dabei zunächst der gesamtstädtische Facebook-Account der Stadtverwaltung Augsburg, welcher von der Hauptabteilung Kommunikation betreut und bespielt wird. In den Blickpunkt des Stadtarchivs rückte die Möglichkeit, dass Social-Media-Unternehmen die eigenen Daten einem User kostenlos zur Verfügung stellen müssen. Diese Anforderung lässt sich auf die Datenschutzgrundverordnung (DSGVO) zurückführen. In den Artikeln 12 bis 23 werden die „Rechte von betroffenen Personen" bei Datenverarbeitungen festgelegt. Innerhalb dessen regelt DSGVO Art. 20 das „Recht auf Datenübertragbarkeit". In dem Artikel heißt es: „Die betroffene Person hat das Recht, die sie betreffenden personenbezogenen Daten, die sie einem Verantwortlichen bereitgestellt hat, in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten."

Dieser gesetzlichen Vorgabe kommen Facebook bzw. der Mutterkonzern META nach, indem sie eine kostenlose Downloadfunktion der eigenen Facebook-Seite anbieten. Dadurch können auch Verwaltungen kostenlos und ohne großen Aufwand die Daten der eigenen Facebook-Seite durch META aufbereitet erhalten. Der Vollständigkeit halber sei erwähnt, dass META ebenfalls eine API anbietet, um die Daten im JSON-Format zur Verfügung zu stellen. Dieser Weg wurde bewusst nicht gewählt, da dieser deutlich mehr technisches Wissen erfordert sowie die bereits erwähnte Gefahr einer Veränderung der API besteht, da die Bereitstellung dieser Schnittstelle nicht rechtlich vorgeschrieben ist und als freiwillige Leistung gesehen werden kann. Das Stadtarchiv Augsburg wollte mittels des städtischen Facebook-Auftritts exemplarisch in der Praxis überprüfen, welche Vor- und welche Nachteile ein Abruf gemäß DSGVO-Vorgaben für eine fachgerechte Archivierung mit sich bringt.

DSGVO-konformer Datenabruf in der Praxis

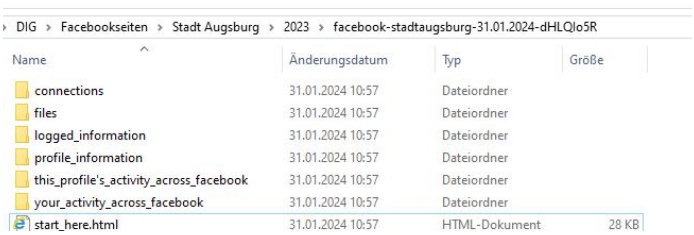
Damit das Stadtarchiv überhaupt in der Lage war, Daten abrufen zu können, musste zunächst ein Zugriff auf den Account bestehen. Daher ist gemeinsam mit der für den Auftritt zuständigen

Hauptabteilung Kommunikation ein Funktionsuser eingerichtet worden, welcher einen Vollzugriff mit Administratorenrechten besitzt.



Die Daten selbst müssen über die Privatsphäreeinstellungen abgerufen werden. Dabei fordert Facebook auf, verschiedene Angaben zu machen. Zunächst muss der Zeitraum bestimmt werden, für den die Daten zur Verfügung gestellt werden. Das Stadtarchiv hat sich für vierteljährliche Schnitte entschieden, damit die Datenportionen überschaubar groß bleiben. Ebenfalls wird abgefragt, ob die Daten in JSON oder HTML bereitgestellt werden sollen. Da es möglichst einfach und praktikabel sein sollte, fiel die Wahl auf HTML. Außerdem fragt Facebook ab, welche Daten des Accounts ausgegeben werden sollen. Hierfür werden die Daten in 23 Kategorien eingeteilt. Diese reichen von normalen Beiträgen über Likes oder Follower bis hin zu Facebook Gaming. An dieser Stelle kann demnach bereits ein Bewertungsvorgang vorgenommen werden. Nicht alle 23 Kategorien erscheinen als archivwürdig. So ist beispielsweise Facebook Gaming für einen Social-Media-Kanal einer öffentlichen Verwaltung irrelevant und wird nicht genutzt. Ein Datenabruf ist somit unnötig.

Wird der Auftrag zur Datenbereitstellung abgeschickt, dauert es zwischen zwei Minuten und 24 Stunden, bis diese für den Download bereitstehen. Dies hängt von der Größe der Datenpakete und den Serverauslastungen ab. Sobald ein Download stattfinden kann, erfolgt eine Information via E-Mail. Es muss also nicht ständig im Account nachgesehen werden. Anschließend kann eine zip-Datei heruntergeladen werden, nach deren Entpacken die Daten in einer Ordnerstruktur liegen, deren Unterordner im Wesentlichen den 23 Kategorien entsprechen. Über die *start_here.html*-Datei lässt sich das heruntergeladene Facebook-Profil als Homepage über den Browser öffnen. Dabei kann durch die Kategorien navigiert werden. Das heißt, dass die klassische Timeline- und Chronikdarstellung von Facebook aufgebrochen wird und die Kategorien getrennt voneinander als eigene Informationsbereiche betrachtet werden müssen. So finden sich beispielsweise unter den Beiträgen keine



Struktur der von Facebook bereitgestellten nach Entpacken der ZIP-Datei. Die start_here.html öffnet die die Ansicht der HTML-Datei

Kommentare mehr. Diese werden in einem extra Bereich angezeigt. Gleiches gilt auch für Messages und alle anderen Informationen aus den 23 Kategorien. Der Konnex zwischen den Kategorien ist somit nicht sofort gegeben, sondern lässt sich lediglich über das Datum vollziehen. Wenn man also nach der Betrachtung eines Postings auch noch die dazugehörigen Kommentare lesen möchte, muss in die Kategorie der Kommentare unter dem entsprechenden Datum navigiert werden. Die Informationen sind also nicht verloren, aber an unterschiedlichen Stellen gespeichert.

Bei den Kommentaren zu einem Posting kommt erschwerend hinzu, dass nur die Kommentare des eigenen Profils, jedoch nicht die Kommentare anderer User ausgegeben werden. Das kann zu der etwas merkwürdigen Situation führen, dass eigene Kommentare als Antwort auf einen vorherigen Kommentar eines anderen Users vorhanden sind, jedoch der Kommentar des anderen Users fehlt. Der Hintergrund der fehlenden Kommentare von anderen Usern ist die DSGVO. Denn es handelt sich bei der Bereitstellung der Daten um einen DSGVO-konformen Abruf, sodass Daten anderer Personen nicht zur Verfügung gestellt werden dürfen.[1]

Deshalb wurde für die Seite der Stadt Augsburg eine statistische Stichprobe erhoben, wie groß der Überlieferungsverlust ohne die Kommentare von Dritten ist. Beispielsweise hat die Stadt Augsburg im 2. Quartal des Jahres 2023 64 Facebook-Beiträge veröffentlicht, die insgesamt eine Reichweite von über 445.000 Usern besaßen. Unter diesen Beiträgen gab es lediglich 380 Kommentare (inklusive der eigenen Kommentare der Stadt Augsburg), was 5,9 Kommentare pro Beitrag macht. Berücksichtigt man, dass ein Beitrag zu den Augsburger Stadtfarben allein 138 Kommentare erhielt und rechnet diesen Beitrag heraus, sind es sogar nur 3,8 Kommentare pro Beitrag. Noch dramatischer fällt die Statistik aus, wenn man bedenkt, dass die Reichweite bei 445.000 Usern lag. Das bedeutet, dass lediglich jeder 1171ste User, der einen Beitrag gesehen, diesen auch kommentiert hat. Von einem großen Austausch zur Meinungsbildung zwischen der Stadt Augsburg und seinen Followern oder gar irgendeiner Art von Repräsentativität kann in diesem Falle also nicht gesprochen werden. An dieser Stelle sei angemerkt, dass Daten über die Reichweiten, Interaktionen und Aufrufe der Beiträge über das Business-Center ebenfalls kostenlos abgerufen werden können. Diese können die rein inhaltliche Überlieferung sinnvoll ergänzen.

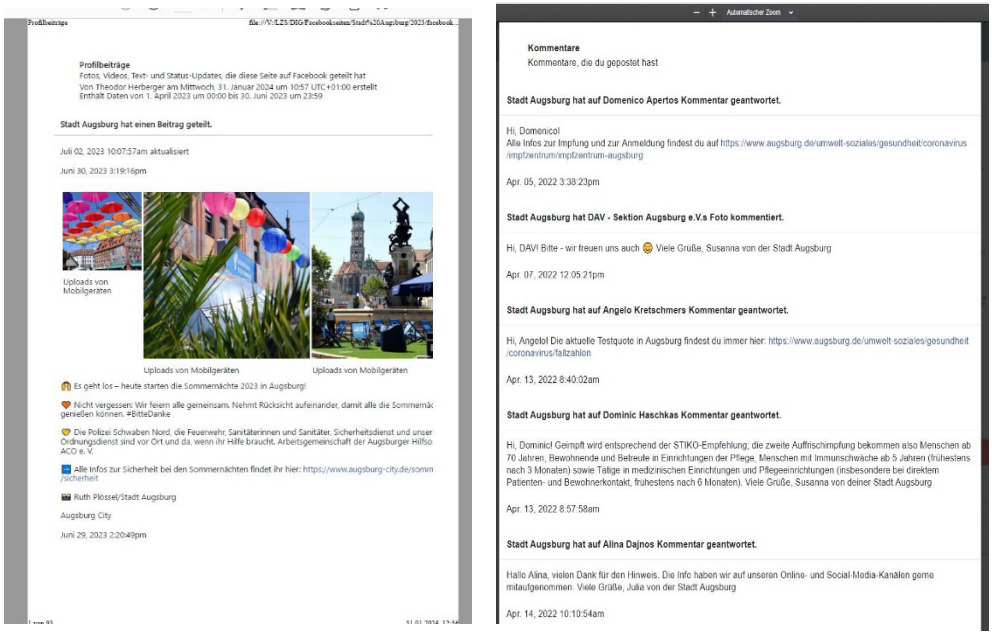
Wie geschildert, erfolgt der Export der Daten im HTML-Format. Bei HTML handelt es sich jedoch nicht um ein langzeitarchivierungsfähiges Format. Der gängige Standard für Webseiten

[1] Das Archiv der sozialen Demokratie ist in seiner rechtlichen Überprüfung zu dem Schluss gekommen, dass auch Kommentare anderer User unter bestimmten Voraussetzungen archiviert werden dürfen. Diese können bei Facebook beispielsweise über die Graph-API zur Verfügung gestellt werden. Allerdings scheint auch hier der Konnex zu den ursprünglichen Beiträgen in der Timeline des Accounts zu fehlen (Walz und Marquet (2022), S. 30).

ist WARC.[2] Ohne den großen Aufwand zu betreiben zu wollen, aus HTML eine WARC-Datei zu erstellen, entstand die Idee, aus HTML eine PDF/A-Datei zu generieren. Dies ist mittels der Speicher- und Druckfunktionen eines jeden Browsers problemlos möglich.

Die ersten Versuche waren zwar erfolgreich, allerdings entstand das Problem, dass der blaue Facebook-Header mit Logo oben auf jeder PDF-Seite erschien und teilweise die Schrift überdeckte. Ein Informationsverlust war die Folge. Über einen Eingriff in den Quellcode über die Entwicklungstools, die sich in jedem Browser befinden, konnte der Header entfernt werden. Hierfür sind leider grundlegende HTML-Kenntnisse notwendig. Diese Maßnahme war jedoch nur zu Beginn der Tests notwendig. Mittlerweile hat META die Zurverfügungstellung der Daten so geändert, dass bei der Erzeugung der PDF/A-Dateien kein Header mehr irgendwelche Informationen überblendet. Ein Eingriff in den Quellcode ist nicht mehr notwendig. Es entsteht dadurch eine PDF/A-Datei, in der alle Beiträge des Accounts chronologisch dargestellt werden. Die mit der Konvertierung einhergehende Veränderung des Bitstreams ist an dieser Stelle zu vernachlässigen, da die signifikanten Eigenschaften, die im Bereich des Text- und Bildinhalts zu definieren sind, sich nicht verändern.

Die Konvertierung einer HTML-Seite in PDF/A hat zur Folge, dass zwar ein langzeitarchivierungsfähiges Format entsteht, jedoch aus einem dynamischen Social-Media-Account eine statische Datei wird. Für die meisten Beiträge auf Facebook ist dies kein Problem, da sie aus Text und Bild bestehen. Der Zusammenhang zwischen den geposteten Bildern und dem Text bleibt bestehen, sodass Text und Bild zusammen zu sehen sind. Schwieriger wird es jedoch, wenn zum Text ein Video vorhanden ist. PDF/A-Dateien können als statisches Dateiformat keine Videos abspielen. Diese werden bei der Umwandlung von HTML zu



Beiträge und Kommentare als getrennte PDF/A-Dateien

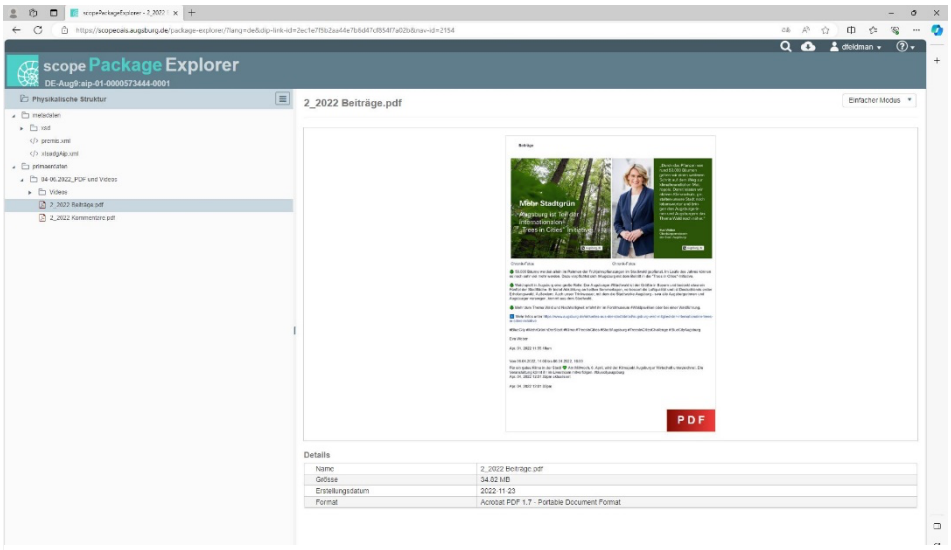
[2] Vgl. hierzu die Analysen der KOST. Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen, https://kost-ceco.ch/cms/willkommen.html, zuletzt eingesehen am 26.07.2024.

PDF/A durch eine weiße Lücke im Dokument ersetzt. Das bedeutet jedoch nicht, dass die Videos nicht vorhanden sind. Denn in dem bereitgestellten Datenpaket befinden sich in einzelnen Ordnern sowohl die Video- als auch die Fotodateien. Ein Informationsverlust ist somit nicht vorhanden, jedoch ist wie beim Verhältnis von Beitrag und Kommentaren der Konnex zwischen Beitrag und Video nicht mehr vorhanden und nur über die Datumsangaben und Dateinamen zu vollziehen.

Um den Nutzenden später überhaupt einen Konnex zwischen diesen Inhalten zu ermöglichen, ist der Pre-Ingest von elementarer Bedeutung. Denn die in ein digitales Archiv zu ingestierenden SIP müssen vom Archiv mittels der gängigen Tools selbst gebaut werden. Ziel muss es ein, dass alle als archivwürdig definierten Kategorien am Ende in einem AIP vorhanden sind. Nur so können Nutzende später über den Access auch alle Informationen erhalten. Als einfaches Beispiel hat das Stadtarchiv Augsburg innerhalb einer einfachen AIP-Struktur die Beiträge, die dazugehörenden Kommentare sowie die Videos. Die Fotos sind nicht extra einzeln in das SIP/AIP übernommen worden, da diese in den Beiträgen enthalten sind. Grundsätzlich wäre dies jedoch möglich. Es könnten sogar Überlegungen angestellt werden, sowohl die Fotografien als auch die Videos gesondert in die jeweiligen Foto- und Filmsammlungen aufzunehmen.

Zusammenfassung

Für den DSGVO-konformen Abruf von Social-Media-Seiten anhand des Beispiels von Facebook ergeben sich für die Archivierung einige Vor- und Nachteile. Positiv zu beurteilen ist, dass es sich beim Datenabruf nach DSGVO um eine kostenlose Variante ohne die Nutzung eines externen Dienstleisters handelt. Auch technisches Wissen ist nicht vonnöten, da die Handhabung sehr einfach und mit gängiger Software durchzuführen ist. Ebenso ist man nicht auf API von weltweiten Konzernen angewiesen. Bei den erstellten PDF/A-Dateien handelt es sich um ein langzeitarchivierungsfähiges Format, welches darüber hinaus per Volltextrecherche durchsuchbar ist. Diese Funktion ist im Hinblick auf



Darstellung der Facebookseite als DIP

eine zukünftige Benutzung von großem Vorteil. Nachteilig ist, dass die Archivierung ähnlich zu Webseiten nicht im WARC-Format stattfindet und durch die Konvertierungen eine dynamische Seite statisch wird. Ebenso ist die Trennung von Kommentaren und Videos vom eigentlichen Beitrag und die Zuweisung über das jeweilige Datum unkomfortabel. Gemischt ist die Lücke in der Überlieferungsbildung zu betrachten, die dadurch entsteht, dass nur die eigenen Kommentare und nicht die Kommentare anderer User ausgegeben werden. Sicherlich ist ein soziales Netzwerk auf Kommunikation untereinander aus. Das Beispiel für die offizielle Seite der Stadt Augsburg zeigt jedoch, dass dies nicht zwingend immer im Mittelpunkt steht oder gegeben ist.

Ob Archive diesen Weg des DSGVO-konformen Datenabrufs gehen wollen, hängt demnach stark von den zu archivierenden Social-Media-Seiten, den Zielen und den vorhandenen Ressourcen des jeweiligen Archivs ab. Die jeweiligen Vor- und Nachteile sollten projektabhängig abgewogen werden.

Bibliografie

Burkert, Philipp Carl (2013), *COMDOK Laurentius Social Media Archivierung: Vortrag beim nestor-Workshop „Webarchive und Social Media“ am 16.10.2013*, https://www.langzeitarchivierung.de/Webs/nestor/DE/Veranstaltungen_und_Termine/Praktikertag/2013webarchivierungb.html (25.7.2024).

Köhn, Katharina (2013), *Parteienstiftungen und Social Media. Facebook-Archivierung mit OWA und Laurentius: Vortrag beim nestor-Workshop „Webarchive und Social Media“ am 16.10.2013*, https://www.langzeitarchivierung.de/Webs/nestor/DE/Veranstaltungen_und_Termine/Praktikertag/2013webarchivierungb.html (25.7.2024).

KOST. Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen, https://kost-ceco.ch/cms/willkommen.html, zuletzt eingesehen am 26.07.2024.

Schuck, Johannes (2022), ‘Schritt für Schritt auf neuen digitalen Wegen: Webseiten- und Social-Media-Kanal Archivierung im Stadt- und Stiftsarchiv Aschaffenburg’, *Archivpflege in Westfalen-Lippe*, 2022, 96, S. 17-20.

Walz, Annabel / Marquet, Andreas (Hrsg.) (2022), *Sicher Sichern? Social-Media-Archivierung aus rechtlicher Perspektive im Archiv der sozialen Demokratie*, Bonn.

Worm, Peter (2021), *Ein neuer Ansatz für die Langzeitarchivierung von Twitter-Accounts,* 9. März 2021, aktualisiert 26.06.2024, https://archive20.hypotheses.org/10031 (25.7.2024).

EMILiA:
Entwicklung einer E-Mail-Archivierungssoftware für kulturelle Gedächtnisinstitutionen

Elisabeth Klindworth und Nico Beyer

Einleitung

Archive sind das Langzeitgedächtnis unserer Gesellschaft. Sie spielen eine unverzichtbare Rolle bei der Überlieferung des kulturellen Erbes. Darüber hinaus sind sie maßgeblich an der Dokumentation politischer, kultureller, wissenschaftlicher und wirtschaftlicher Prozesse beteiligt. Archive leisten somit einen signifikanten Beitrag zur Sicherstellung der Nachvollziehbarkeit gesamtgesellschaftlichen Handelns sowie zur Vorbeugung von Desinformation. Grundlage hierfür sind sowohl Unterlagen von Einrichtungen des öffentlichen Rechts als auch ausgewählte Dokumente aus privater Hand. Um auch aktuelle Geschehnisse abbilden zu können, müssen in Archiven zunehmend digitale Dateien berücksichtigt werden, da andernfalls irreversible Überlieferungslücken drohen.

Im Jahr 2023 sollen weltweit pro Tag mehr als 347 Milliarden E-Mails gesendet und empfangen worden sein (The Radicati Group, Inc., 2023, S. 3). Diese beeindruckende Zahl verdeutlicht die zentrale Rolle, die E-Mails in der modernen Kommunikation einnehmen. Eine vernünftige Gesamtlösung für die Überlieferung historisch oder juristisch bedeutsamer E-Mail-Konten, die alle relevanten archivfachlichen Standards berücksichtigt und sich in allen Gedächtnisinstitutionen effektiv einsetzen ließe, existiert bislang noch nicht. Aktuell kommen in vielen Einrichtungen vor allem Behelfslösungen wie zum Beispiel die Aufbewahrung der ursprünglichen E-Mail-Formate, die Speicherung im PDF-Format oder sogar das Ausdrucken von E-Mails auf Papier zum Einsatz. In anderen Gedächtnisinstitutionen existieren zum jetzigen Zeitpunkt noch gar keine Lösungsansätze. Ein wesentlicher Grund hierfür ist das Fehlen geeigneter Werkzeuge für die Bearbeitung großer und unstrukturierter Datenmengen sowie die meist überschaubaren personellen Ressourcen.

Das in diesem Beitrag vorgestellte EMILiA-Projekt will diese Lücke schließen. Hauptziel des Projekts ist es, eine Software zu entwickeln, die von der Übernahme über die technische sowie inhaltliche Aufbereitung bis hin zur Nutzung alle archivfachlichen Arbeitsschritte abdeckt und diese so weit wie möglich automatisiert. Hierbei kommen sowohl herkömmliche Methoden als auch Ansätze aus dem Bereich der künstlichen Intelligenz zum Einsatz. Wichtige

archivfachliche Entscheidungen werden nach wie vor von den Anwender:innen der Software getroffen. Die Automatisierungsmethoden sind lediglich ein Werkzeug für die Bewältigung der riesigen Datenmengen. Um zu gewährleisten, dass die archivierten Postfächer trotz der sich rasch verändernden technischen Rahmenbedingungen lesbar bleiben und zeitnah von Forschenden genutzt werden können, muss eine Vielzahl von technischen, rechtlichen, inhaltlichen und organisatorischen Hindernissen überwunden werden.

Dieser Beitrag beleuchtet in einem kurzen Rückblick, wie das EMILiA-Projekt entstanden ist, fasst die wesentlichen Herausforderungen der E-Mailarchivierung zusammen, stellt vor, wie das EMILiA-Projektteam diesen begegnet und schließt mit einem Ausblick auf anstehende Schritte. Dadurch möchte der Beitrag den Austausch der Archive zu konkreten Anforderungen an die E-Mailarchivierung und deren praktische Umsetzung fördern. Da geplant ist, die Software EMILiA anderen Archiven anzubieten, nimmt das Projektteam jederzeit gerne Feedback und Ideen für weitere Funktionalitäten entgegen.

Rückblick

Die wichtigste Grundlage des derzeitigen Projekts sind Vorarbeiten, die seit 2015 im Archiv der Max-Planck-Gesellschaft entstanden sind. Hintergrund der damaligen Bemühungen war der Umstand, dass die Abgabe analoger Unterlagen durch die Organe und Institute der Max-Planck-Gesellschaft stark rückläufig war. Nachforschungen des Archivs ergaben, dass sich ein signifikanter Teil der langfristig relevanten Kommunikation vollständig in den digitalen Raum verlagert hatte. In den 2010er-Jahren, wie auch heute noch, lag der Fokus der meisten Archive in Deutschland und international auf der Archivierung von E-Akten verbunden mit der Vorstellung, dass archivwürdige E-Mails ebenfalls in E-Akten ins Archiv kommen (vgl. z. B. Das Bundesarchiv, 2018).

Das Archiv der Max-Planck-Gesellschaft entschied sich dennoch, eine spezifische Lösung für die Archivierung von E-Mails anzustreben. Dafür waren im Wesentlichen zwei Gründe ausschlaggebend. Zum einen bilden Vor- und Nachlässe von Wissenschaftler:innen der MPG neben dem Verwaltungsschriftgut eine zweite wichtige Säule der Überlieferungsbildung. Die Wissenschaftler:innen arbeiten nicht mit Verwaltungsakten. Für sie ist die E-Mail ein zentrales Medium des Austauschs und der Organisation ihrer Tätigkeit. Letztlich handelt es sich bei der E-Mail-Technologie um nicht weniger als den direkten Nachfolger der herkömmlichen Briefkorrespondenz, welche eine der am häufigsten konsultierten Quellengattungen im Archiv der Max-Planck-Gesellschaft darstellt. Man stelle sich nur das Unverständnis der heutigen

Archivnutzer:innen vor, wenn beispielsweise die Bewertungsentscheidung über die persönlichen Briefe eines Albert Einstein in Richtung einer Kassation ausgefallen wäre.

Zum anderen ist davon auszugehen, dass E-Mails für Forschende der historisch orientierten Geisteswissenschaften als digitale Quelle einen Wert an sich haben, weil man sie in verschiedenen technischen Umgebungen weiterverarbeiten kann. Insbesondere im Hinblick auf neue (quantitative) Auswertungsmethoden wie zum Beispiel Netzwerkanalysen, Themenmodellierungen oder verschiedenste Möglichkeiten der Datenvisualisierung kann es sich lohnen, die technischen Funktionalitäten von E-Mails sowie die dazugehörigen Zusammenhänge so gut wie möglich zu erhalten.

Da 2015 in der Fachcommunity weder praktische Erfahrungen noch zufriedenstellende Softwarelösungen für die Archivierung von E-Mails existierten, machte sich das Archiv auf die Suche nach geeigneten Partnerinstitutionen für eine Eigenentwicklung. Das Ergebnis dieser Suche war die bis heute anhaltende Kooperation mit dem Fachbereich Informatik der Freien Universität Berlin. Im Rahmen verschiedener Universitätsprojekte wurden in den folgenden zwei Jahren die konzeptionellen Grundlagen für die Entwicklung einer Softwarelösung für die sichere und verlustfreie Übernahme von E-Mail-Postfächern geschaffen. Diese wurden 2017 schließlich vom Informatiker Felix Gericke, der über die Projekte als Mitarbeiter für das Archiv der Max-Planck-Gesellschaft gewonnen werden konnte, in Form einer Spezifikation und eines Prototyps in die Tat umgesetzt. Die Entwicklung erfolgte anhand von Echtdaten des Archivs, also mehreren E-Mailaccounts von Wissenschaftler:innen und Direktor:innen der Max-Planck-Gesellschaft. Ein Jahr später testeten verschiedene Kommunal- und Landesarchive, darunter das Landesarchiv Baden-Württemberg und das Stadtarchiv Stuttgart, den Prototypen.

Aktuell (im Jahr 2024) wird das Entwicklungsvorhaben durch das Förderprogramm „ProValid“ der Investitionsbank Berlin finanziert. Diese Förderung erlaubte nicht nur eine Intensivierung der Entwicklungsarbeit, sondern auch die personelle Aufstockung des Projektteams. Neben Felix Gericke arbeiten nun auch Alexander Hinze-Hüttl als Informatiker und Nico Beyer als Archivar im Projekt. Beide Informatiker hatten zuvor bereits ihre universitären Qualifikationsarbeiten über Teilaspekte der E-Mail-Archivierung geschrieben. Somit konnte das Team im Förderzeitraum ohne lange Einarbeitungsphase sofort mit der Entwicklungstätigkeit beginnen. Inzwischen liegt ein zweiter Prototyp mit erweiterten Funktionen vor, der vielversprechende Ergebnisse liefert. Von Beginn an wurde die Zielvorstellung verfolgt, EMILiA anderen Archiven als Lösung für die E-Mailarchivierung anzubieten.

Ausgangspunkt

Inhaltliche Herausforderungen

Aufgrund von Spam, Rundmails und einer Vielzahl an Nachrichten, die keine Informationen von bleibendem Wert enthalten, fallen E-Mail-Postfächer im Vergleich zu herkömmlichen Briefkorrespondenzen häufig sehr umfangreich aus. Die E-Mail-Konten, die bisher für Testzwecke übernommen werden konnten, umfassten im Durchschnitt rund 50’000 Mails. Der bisher größte Account enthielt sogar weit über 130’000. Würde man den Inhalt all dieser Nachrichten ausdrucken, würden dabei rund 600’000 A4-Seiten entstehen. Die 258’180 Anhänge sind bei dieser Rechnung noch nicht einmal berücksichtigt. Nur ein Bruchteil dieser Datenmenge ist von bleibendem Wert. Vor dem Hintergrund dieser Zahlen liegt es auf der Hand, dass eine inhaltliche Bewertung und Erschließung von E-Mail-Konten selbst in Institutionen mit umfangreichen personellen Ressourcen händisch nicht zu bewerkstelligen ist. Eine hinreichende archivfachliche Aufbereitung ist daher nur mithilfe einer Automatisierung zu erreichen.

Technische Herausforderungen

Erschwerend kommt eine ganze Reihe von technischen Besonderheiten hinzu, die bei der archivfachlichen Bearbeitung berücksichtigt werden müssen. Die gängigsten Formate für die Speicherung von E-Mail-Postfächern sind die Container-Formate PST und MBOX (vgl. Hanson und Eggert, 2005; Microsoft, 2022). Für die Speicherung von Einzelformaten werden meist die Formate EML und MSG verwendet. Das PST-Format ist ein proprietäres Dateiformat der Firma Microsoft, das nicht ohne spezielle Software gelesen werden kann. In Exchange-Umgebungen mit domänenseitig verwalteten E-Mail-Konten werden häufig globale Adressbücher verwendet. Bei einer Übernahme von PST-Dateien kann es daher vorkommen, dass die E-Mail-Adressen verloren gehen und lediglich die selbst gewählten Bezeichner aus den Adressbüchern mitgeliefert werden. Beim MBOX-Format werden hingegen alle Nachrichten eines Ordners in einer einzigen Datei gespeichert und durch Trennzeichen separiert. Diese Struktur führt dazu, dass für jede Bearbeitung die gesamte MBOX-Datei eingelesen werden muss, was bei großen Postfächern sehr ineffizient sein kann. Weiterhin genügt eine einzige virenbelastete oder anderweitig kompromittierte Datei, damit der gesamte Container als potenziell gefährlich gekennzeichnet werden muss. Ein weiteres Problem besteht darin, dass sich die Prüfsumme der Datei nach jeder inhaltlichen Bearbeitung, wie beispielsweise einer Kassation, zwangsläufig verändert, was die Sicherstellung der Integrität und Authentizität erschwert. Bei der archivfachlichen

Bearbeitung müssen nicht nur die E-Mail-Formate, sondern auch die Formate der Anhänge berücksichtigt werden. Die bislang untersuchten Testdaten enthielten die verschiedensten Dateien, die sich von herkömmlichen PDF-Dokumenten über verschiedenste Text- und Bilddateien bis hin zu einer breiten Palette von veralteten oder proprietären Formaten erstreckten. Des weiteren können fehlende oder fehlerhafte Headerinformationen zu Problemen bei der Datenverarbeitung und die Vielzahl der unterschiedlichen Zeichenkodierungen zu Darstellungsfehlern führen. In den bisher bearbeiteten Konten befand sich außerdem eine erhebliche Anzahl von verschlüsselten und signierten Nachrichten. Weiterhin können E-Mail-Anhänge gefährliche Dateien enthalten, die im Zuge der Archivierung unschädlich gemacht werden müssen.

Rechtliche Herausforderungen

Eine weitere Problemdimension besteht darin, dass E-Mails, genau wie andere Unterlagen auch, personenbezogene Daten umfassen oder urheberrechtlich relevant sein können. Eine zeitnahe Nutzbarmachung größerer E-Mail-Bestände kommt daher nur in Frage, wenn Werkzeuge für eine automatisierte Anonymisierung oder Pseudonymisierung bereitgestellt werden. Zudem stellt sich die Frage, in welcher Form E-Mails sinnvoll für eine Nutzung bereitgestellt werden können. Weniger problematisch ist hingegen die rechtliche Absicherung von Übernahmen, da sich die hierfür notwendigen Vereinbarungen nicht grundsätzlich von denen unterscheiden, die bereits aus dem Nachlassbereich bekannt sind. Rein dienstlich genutzte Konten können, je nachdem, in welchem rechtlichen Kontext ein Archiv agiert, sogar unter die Anbietungspflicht fallen.

Lösungskonzept

Die Softwarelösung wird sich aus den drei Teilmodulen „Übernahme“, „Bewertung & Erschließung“ und „Nutzung“ zusammensetzen, die jeweils unterschiedliche Teilbereiche des OAIS-Referenzmodells abdecken. Lediglich die langfristige Speicherung der Datenpakete wird außerhalb des Systems erfolgen. Der modulare Aufbau garantiert, dass sich die Softwarelösung optimal an die Bedürfnisse der jeweils nutzenden Institution anpassen lässt. EMILiA soll sich sowohl von Archivar:innen als auch von Nutzer:innen intuitiv und ohne Programmierkenntnisse bedienen lassen. Um dies zu ermöglichen, wird eine grafische Oberfläche zur Verfügung gestellt, von der aus alle Arbeitsschritte angestoßen und überwacht werden können. Da sich einige Prozesse aufgrund der erforderlichen Rechenleistung über einen längeren Zeitraum erstrecken können, verfügt EMILiA über ein Benachrichtigungssystem, welches die Anwender:innen automatisch informiert, sobald ein Datenpaket für den nächsten Bearbeitungsschritt bereit ist.

Abbildung 1: Überblick über den Funktionsumfang von EMILiA (Grafische Darstellung: Nico Beyer)

Übernahme

Das Übernahmemodul ermöglicht den abgebenden Personen und Institutionen mithilfe einer intuitiven und barrierefreien grafischen Oberfläche einen sicheren und reibungslosen Transfer ihrer Daten an das jeweils zuständige Archiv. Durch den Einsatz von auf dem neusten Stand befindlichen Verschlüsselungsverfahren wird der Schutz sensibler Informationen gewährleistet. Darüber hinaus werden die E-Mails in diesem Modul in ihre Urformate zurückgeführt, was sowohl die weitere technische Bearbeitung als auch die spätere Nutzung der Daten erheblich vereinfacht. Das Ergebnis dieses Prozesses ist ein archivtauglicher BagIt-Container, in dem sich eine Strukturdatei, die E-Mails im TXT- oder HTML-Format, die dazugehörigen XML-Metadatendateien sowie die Anhänge befinden (Kunze et al., 2018).

Bewertung

Eine der wichtigsten archivfachlichen Aufgaben besteht in der Auswahl der überlieferungswürdigen Dokumente. Aufgrund der riesigen Datenmengen, die bei der Archivierung von E-Mail-Postfächern anfallen, ist eine rein manuelle Herangehensweise undenkbar. Das zweite Teilmodul bietet daher sowohl konventionelle Methoden als auch Ansätze aus dem Bereich des maschinellen Lernens, mit denen die Identifikation historisch relevanter Nachrichten vereinfacht werden kann. Mithilfe von Naive Bayes Classifiern und dem Deep-Learning-Modell BERT werden Spam- und Phishing-Nachrichten effektiv erkannt und markiert (Devlin et al., 2019). Durch einen Abgleich der Prüfsummen ist es zudem möglich, Dubletten auszusortieren. Jeder Arbeitsschritt wird im Einklang mit den Metadatenstandards METS und PREMIS dokumentiert. Langfristig ist zudem eine Funktion für die globale Deduplizierung angedacht.

Insbesondere in Organisationen, in denen viel mit Mailinglisten gearbeitet wird, könnte dieser Ansatz zu einer signifikanten Einsparung von Speicherplatz führen.

Erschließung

Das zweite Teilmodul bietet verschiedene Ansätze für die automatisierte Extraktion von Metadaten und die Bereitstellung von Zusatzinformationen. Neben Freitextfeldern wird EMILiA Funktionen für die automatische Erfassung von Personen, Orten und Organisationen beinhalten. Für die Erkennung einfacher Zusammenhänge werden reguläre Ausdrücke identifiziert. Komplexere Analysen werden mithilfe der neuronalen Netze SpaCy und Flair durchgeführt (Honnibal et al., 2020; Akbik et al., 2019). Eine initiale Spracherkennung sorgt dafür, dass die Daten mit dem jeweils passenden Modell verarbeitet werden. Für den Fall, dass bei der Vorverarbeitung eine Entität übersehen wurde, gibt es eine Funktion für die manuelle Markierung von Textabschnitten. Sowohl in der grafischen Oberfläche als auch in den Metadaten ist erkennbar, ob eine Entität manuell oder automatisch markiert worden ist. Darüber hinaus ist es möglich, automatisch erkannte Begriffe zu verifizieren. Weiterhin wird dokumentiert, welches KI-Modell für die Erkennung verwendet wurde. Auf diese Weise wird sichergestellt, dass stets ersichtlich ist, welche Information von einer künstlichen Intelligenz und welche von einem Menschen generiert wurde.

Perspektivisch soll außerdem eine Themenmodellierung durchgeführt werden, mit der E-Mail-Threads automatisch klassifiziert werden können. Hinsichtlich der Anbindung an bestehende Archivsoftware ist vorgesehen, E-Mail-Konten auf Postfachebene in der Datenbank des jeweils genutzten Archivinformationssystems zu verzeichnen und mit einem Link zur EMILiA-Datenbank zu versehen. Dort erfolgt dann die detaillierte Suche. Die Verzeichnung einzelner E-Mails im AFIS wurde zwar diskutiert, aufgrund der riesigen Datenmengen aber schnell verworfen.

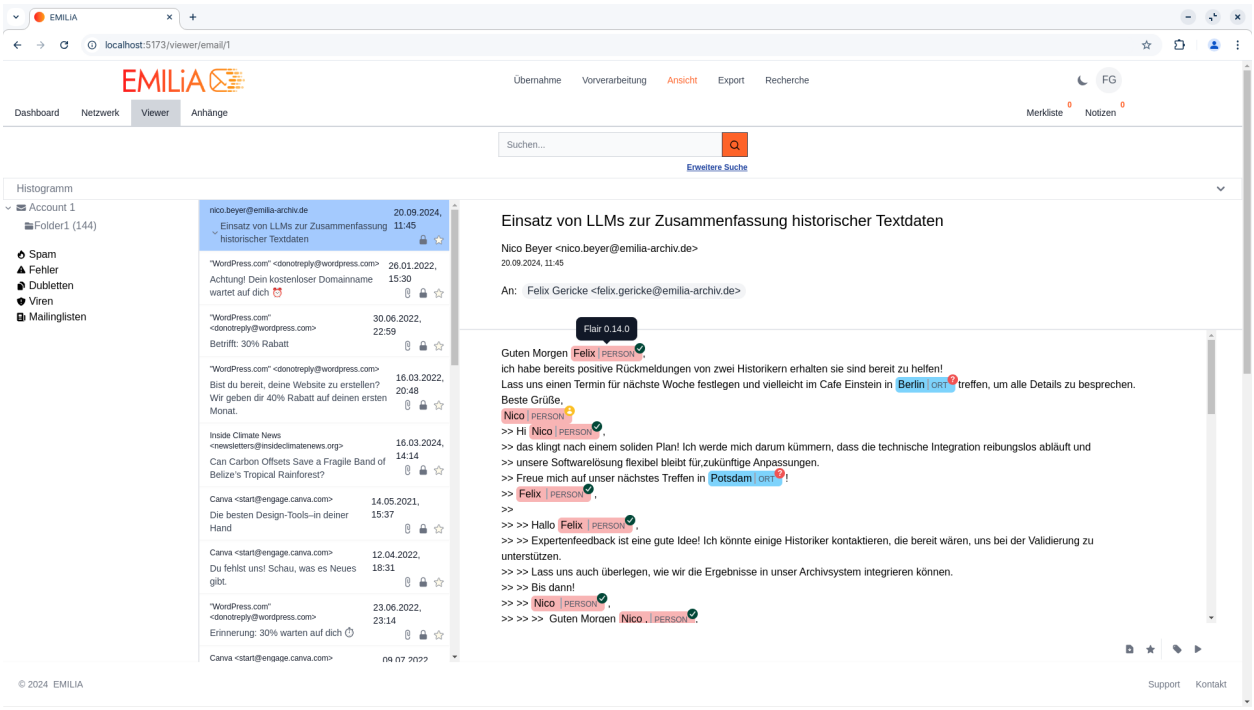


Abbildung 2: Bildschirmfoto des funktionstüchtigen Viewer-Prototyps (Bildschirmfoto: Nico Beyer)

Darstellung und Nutzung

Die Auswertung der Daten kann mithilfe einer grafischen Weboberfläche durchgeführt werden. Neben verschiedenen Filter- und Sortierfunktionen sind eine Volltextsuche und ein interaktives Histogramm für die exakte Eingrenzung des Suchzeitraums geplant. Darüber hinaus wird die inhaltliche Zusammensetzung der archivierten Konten mithilfe verschiedener Datenvisualisierungen dargestellt. Unter anderem finden sich hier Angaben zur Menge und Art der Nachrichten und Anhänge, das Ergebnis der Spracherkennung sowie eine grafische Darstellung des Korrespondenznetzwerks auf Grundlage der verwendeten E-Mail-Domains. EMILiA soll nicht nur ein Werkzeug für die Archivierung, sondern auch ein Katalysator für neue Erkenntnisse in den Sozial- und Geisteswissenschaften sein. Da bei der Entwicklung nicht nur die Anforderungen der Archive, sondern auch die der Archivnutzer:innen berücksichtigt werden sollen, wird aktuell eine an Forschende adressierte Umfrage durchgeführt.

Umgang mit sensiblen Informationen

Alle Berechnungen können lokal und offline durchgeführt werden, um die Vertraulichkeit von Daten zu gewährleisten. Die Schutzfrist kann bei der Übernahme von der zuständigen Archivfachkraft festgelegt werden und wird nach Ablauf der Frist automatisch deaktiviert. Eine der wichtigsten Funktionen von EMILiA ist die automatische Erfassung und optionale Anonymisierung sensibler Informationen im Zuge der Bereitstellung. Nur durch diesen Schritt ist eine

zeitnahe und rechtskonforme Nutzung möglich. Technologische Grundlage hierfür ist die im Rahmen der inhaltlichen Erschließung durchgeführte Entity Recognition.

Entwicklungsstand und anstehende Herausforderungen

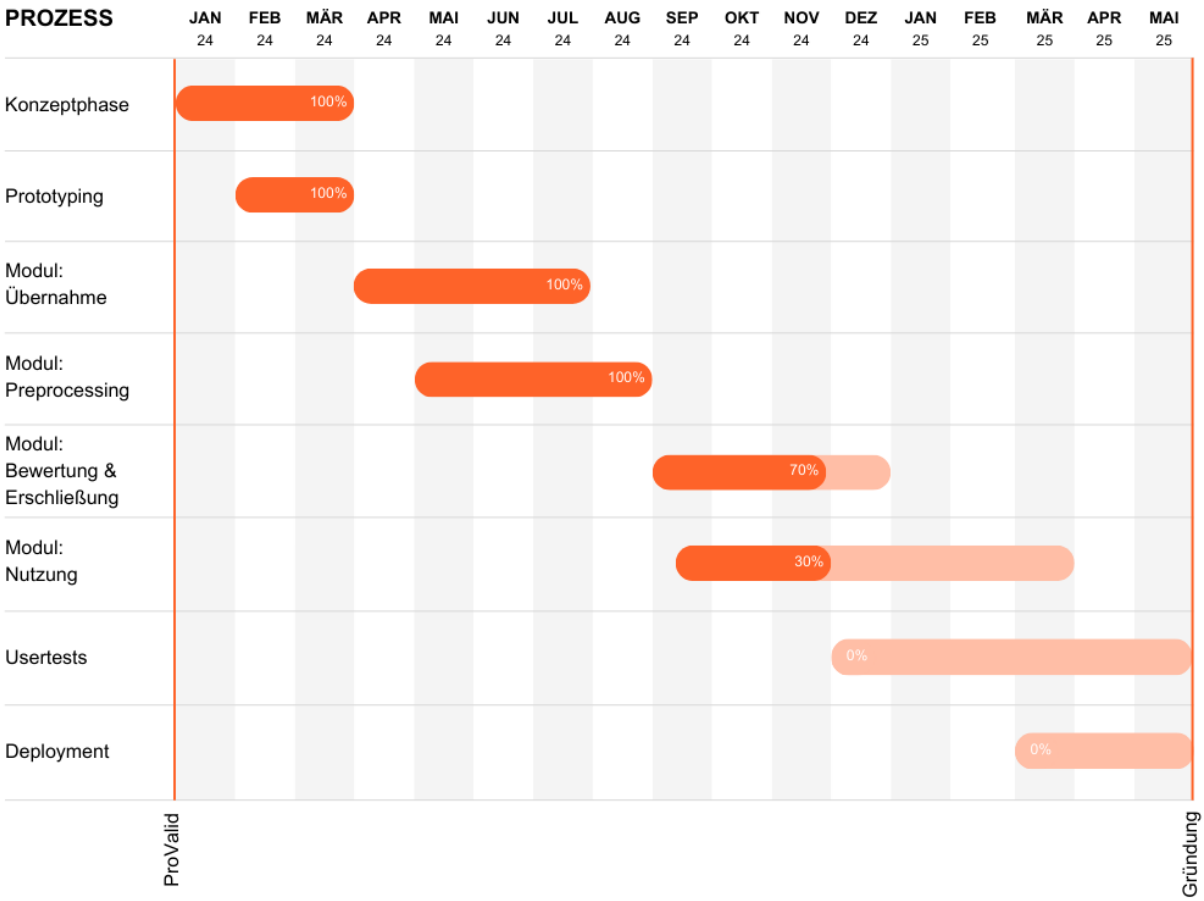


Abbildung 3: Vereinfachte Planung bis zur Gründung (Grafische Darstellung: Nico Beyer)

Die Ermittlung der Anforderungen, die Konzeptphase sowie das Prototyping sind vollständig abgeschlossen. Das Übernahmemodul und die Vorverarbeitung sind fertiggestellt und konnten erfolgreich mit Echtdaten der Max-Planck-Gesellschaft getestet werden. Viele der notwendigen Schnittstellen für das Teilmodul „Bewertung & Erschließung“ sind bereits vorhanden. Die Funktionen für die Erkennung von Spam, Dubletten und Mailinglisten sowie Entity Recognition funktionieren reibungslos. Dasselbe gilt für die automatische Anonymisierung. Die KI-gestützte Zusammenfassung von Threads ist grundlegend implementiert, muss aber noch mit Echtdaten evaluiert und gegebenenfalls mithilfe weiterer Trainingsdaten angepasst werden. Die ersten Ergebnisse sind aber durchaus vielversprechend. Aktuell wird hauptsächlich an der grafischen Oberfläche für das Gesamtsystem und dem Teilmodul „Nutzung“ gearbeitet. Im Anschluss wird ein Schwerpunkt auf Usertests, Bugfixing, die Optimierung der

Softwareergonomie sowie die Verbesserung der Performance gelegt werden. Einer der wichtigsten anstehenden Arbeitsschritte ist die Zusammenführung des Gesamtsystems sowie die Verknüpfung der Front- und Backendmodule. Fast alle Prozesse funktionieren bereits auf Kommandozeilenebene. Einige Module müssen aber noch mit den jeweiligen Eingabemöglichkeiten der grafischen Oberfläche verbunden werden. Für den Umgang mit verschlüsselten E-Mails liegen bislang noch keine finalen Lösungskonzepte vor. Aktuell sammelt das Team aktiv Informationen für die Umsetzung eines möglichst reibungslosen Transfers der bewerteten und erschlossenen Datenpakete in die gängigen Langzeitarchivlösungen. Um die Bearbeitung von Anfragen so unkompliziert wie möglich zu gestalten, wird eine Recherchedatenbank implementiert, die eine Volltextsuche über mehrere Konten hinweg ermöglicht. Da EMILiA in der finalen Version die Möglichkeit einer kollaborativen Bearbeitung von Beständen bieten soll, muss außerdem ein Rollen- und Rechtemanagement implementiert werden, mit dem Accounts erstellt und verwaltet werden können. Derzeit werden für den Installationsprozess noch erweiterte IT-Kenntnisse benötigt. Da nicht jedes Archiv jederzeit Zugriff auf entsprechende Ressourcen hat, muss die Inbetriebnahme stark vereinfacht werden. Obwohl die grafische Oberfläche so intuitiv wie möglich gestaltet wurde, macht die Komplexität des Bearbeitungsprozesses schließlich die Erstellung einer detaillierten Dokumentation unabdingbar.

Ausblick

Sobald die Entwicklung abgeschlossen ist, soll eine Ausgründung in Form eines Startup-Unternehmens erfolgen, damit die Software aktiv entlang existierender Bedarfe weiterentwickelt und langfristiger Support gewährleistet werden kann. Aktuell ist eine Finanzierung durch ein Lizenzmodell angedacht, bei dem der Preis von der Größe der jeweils nutzenden Institution abhängig ist. Auf diese Wiese soll sichergestellt werden, dass die Softwarelösung in möglichst vielen Einrichtungen eingesetzt werden kann. Nach der Entwicklung möchte das Projektteam einen umfassenden technischen Support und praxisnahe Schulungen anbieten. Die fertige Softwarelösung muss darüber hinaus fortwährend an die aktuellen technologischen Rahmenbedingungen angepasst und sicherheitstechnisch auf dem neusten Stand gehalten werden. Um diesen Service gewährleisten zu können, soll Mitte 2025 ein Startup-Unternehmen gegründet werden. Langfristig möchte das EMILiA-Team weitere Softwareprojekte im Bereich der automatisierten und standardkonformen Verarbeitung digitaler Textdaten umsetzen. Jedes Softwareprodukt ist nur so gut wie die Anforderungen, die seine Anwender:innen an es formuliert haben. Das EMILiA-Projektteam möchte Sie mit diesem Beitrag herzlich dazu einladen, über den Fortgang

der Entwicklung auf dem Laufenden zu bleiben und sich mit dem Team über Ihre Bedarfe auszutauschen.

Bibliografie

Das Bundesarchiv (2019), *Umgang mit E-Mails in elektronischen Akten*, https://www.bundesarchiv.de/assets/bundesarchiv/de/Downloads/Erklaerungen/beratungsangebote-grundl-sgv-umgang-mit-e-mails-in-elektronischen-akten-juni-2022.pdf (29.10.2024).

Devlin, J. et al. (2019), *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*, https://doi.org/10.48550/arXiv.1810.04805 (29.10.2024).

Hanson, T., Eggert, L. (2005), *The application/mbox Media Type. RFC 4155*, https://www.rfc-editor.org/rfc/rfc4155.html (29.10.2024).

Honnibal, M. et al. (2020), *spaCy: Industrial-strength Natural Processing in Python*, https://zenodo.org/records/10009823 (29.10.2024).

Kunze, J. A. et al. (2018), *The BagIt File Packaging Format (V1.0). RFC 8493*, https://datatracker.ietf.org/doc/rfc8493/ (29.10.2024).

Microsoft (2022), *[MST-PST]: Outlook Personal Folders (.pst) File Format*, https://learn.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/141923d5-15ab-4ef1-a524-6dce75aae546 (29.10.2024).

The Radicati Group, Inc. (2023), *Email Statistics Report, 2023–2027*, https://www.radicati.com/wp/wp-content/uploads/2023/04/Email-Statistics-Report-2023-2027-Executive-Summary.pdf (29.10.2024).

Langzeitarchivierung von E-Mails an der ETH Zürich

Claudia Briellmann und Fabian Schneider

Einleitung

E-Mails haben für Archive und Langzeitarchive in den letzten Jahren massiv an Bedeutung zugenommen. Die Papierablage wurde längst durch eine digitale Ablage ersetzt, die Kommunikation erfolgt seit Jahren überwiegend elektronisch. Der E-Mail-Verkehr, seit langem der zentrale Kommunikationskanal, wird für den Austausch von Entscheidungen, für die Verbreitung von Informationen, für Absprachen oder die Terminfindung genutzt. Mit anderen Worten: In den E-Mails liegen Datenschätze verborgen. Für das Hochschularchiv sind die E-Mails, die im Kontext der ETH Zürich versendet oder empfangen wurden archivwürdig, weil sie die Geschäftstätigkeit der ETH Zürich dokumentieren und dies bis in die 90er-Jahre des letzten Jahrhunderts zurück. Besondere Bedeutung erhalten diese Daten außerdem dadurch, dass es sich um genuin digitale Dateien (Born Digitals) handelt.

Das Projekt „E-Mail-Archiv 2.0“

Beim Projekt „E-Mail-Archiv 2.0“ handelte es sich um ein von den Informatikdiensten der ETH Zürich (ID) geleitetes Projekt, das in Zusammenarbeit mit dem Hochschularchiv der ETH Zürich (HSA) sowie der Gruppe Forschungsdatenmanagement und Datenerhalt (FDD) der ETH-Bibliothek durchgeführt wird. Im Folgenden wird ausschließlich auf die Aspekte der archivischen Bewertung, Übernahme, Erschließung und Langzeitarchivierung von E-Mails eingegangen.

Bisher wurden alle Objekte in Microsoft Outlook Konten von ETH-Angehörigen (E-Mails, Kalender, Termine oder Notizen), welche nicht innerhalb von 30 Tagen aus dem E-Mail-Postfach gelöscht wurden, automatisch im sogenannten „Vault“ zwischengespeichert (ETH, 2021). Dabei handelt es sich um keine digitale Langzeitarchivierungslösung, sondern um ein „E-Mail-Archivierungssystem“, wie es von Karin Schwarz (2010, S. 558-559) definiert ist. Ein solches System eignet sich für die mittelfristige, aber nicht für die dauerhafte Aufbewahrung von E-Mails. Es erfolgt vorgängig auch keine archivische Bewertung.

Im Rahmen des Projekts wurden E-Mail-Konten wichtiger Stellen an der ETH Zürich vom HSA bewertet und die als archivrelevant eingestuften Konten übernommen und langzeitarchiviert. Archivwürdig sind diejenigen Konten, welche die geschäftsrelevanten Vorgänge an der ETH Zürich dokumentieren. Einerseits wurde eine retrospektive Bewertung der bereits im Vault

gespeicherten Outlook-Objekte vorgenommen, andererseits wurde prospektiv festgelegt, welche E-Mail-Konten in welchem zeitlichen Abstand an das Hochschularchiv abgeliefert werden. In enger Zusammenarbeit mit der Gruppe FDD und den ID wurden die Arbeitsabläufe und Prozesse des Exports und der Übernahme ins ETH Data Archive definiert. Diese sollen im Idealfall für andere E-Mail-Bestände, zum Beispiel aus Privatnachlässen, nachgenutzt werden können.

Bewertung von E-Mail-Beständen

Im digitalen Zeitalter, in dem immer mehr Daten entstehen und die Kosten für Speicher verhältnismäßig gering sind, stellt sich die Frage, wie relevant die Bewertung vor der Archivierung noch ist. Früher ging es noch darum, möglichst viel zu archivieren, allerdings waren die Bestände auch kleiner. Digitale Bestände sind viel umfangreicher als die traditionellen analogen Bestände und die Bewertung wird ein zunehmend aufwändiger Prozess. Geoffrey Yeo (2019, S. 45) stellte deshalb die bewusst provokante Frage: „*Can we – should we – try to keep everything?*".

Die Task Force on Technical Approaches for Email Archive (2018, S. 3-14) beschrieb die Problematik der Archivierung von E-Mails sehr gut: E-Mails gewinnen immer mehr an Bedeutung als historische Quellen und doch gibt es noch keine allgemeine Lösung, wie man E-Mails bewertet, übernimmt, erschließt, langzeitarchiviert und für die Nachnutzung zugänglich macht. Es muss deshalb ein Weg gefunden werden, ausschließlich relevante E-Mails zu archivieren. Denn es ist, wie bei den meisten anderen behördlichen Unterlagen, davon auszugehen, dass ein Teil der E-Mail-Konten archivwürdig ist, ein anderer Teil jedoch nicht (vgl. Knobloch, 2016, S. 221).

Im Rahmen der Bewertung von E-Mails aus dem Vault sprach sich das Hochschularchiv mit der Schulleitung der ETH Zürich sowie mit dem ETH-Rat bezüglich der Bewertungskriterien ab. Die Möglichkeit, die E-Mails aus einem RMS ins Hochschularchiv zu übernehmen bot sich nicht, da bisher kein Geschäftsverwaltungssystem an der ETH Zürich im Einsatz ist. Die Entscheidung fiel deshalb auf den Capstone Approach:

Bei diesem handelt es sich um eine von den National Archives und Records Administrations (NARA) entwickelte Bewertungsmethode, welche eine Alternative zur sehr aufwändigen Einzelbewertung darstellt (NARA, 2013). Dabei werden ausschließlich E-Mail-Konten von bestimmten Personen, sogenannten Capstones (auf Deutsch nach Benauer (2020, S. 105) „Schlusssteine"), innerhalb der Verwaltung als archivwürdig definiert und übernommen. Das Hauptargument für diesen Ansatz ist, dass die meisten geschäftsrelevanten und somit als archivwürdig eingestuften E-Mails von diesen Personen gesendet oder empfangen werden. Durch

die Übernahme der vollständigen Konten ohne zeitaufwändige Einzelbewertung lassen sich erheblich Ressourcen sparen. Der Capstone Approach bewegt sich im Rahmen des Macro Appraisal und bringt viele Vorteile mit sich: Zum einen bleibt so das Provenienzprinzip der Bestände erhalten und es kann eine repräsentativere Überlieferung erreicht werden, wenn komplette E-Mail-Konten übernommen werden. Die überlieferten E-Mail-Postfächer sind zudem nach Benauer homogener und hochwertiger in der Qualität. Dies führt auch dazu, dass der Evidenzwert dieser E-Mails erhalten bleibt, da die interne Ordnungsstruktur und die Metadaten der Postfächer erhalten bleibt (vgl. Artefactual Systems and the Digital Preservation Coalition, 2021, S. 7; Benauer, 2020, S. 103-105; NARA, 2015, S. 7-8). Schließlich spricht sich auch Benauer (2020, S. 106) dafür aus, dass beim Capstone Approach eine höhere Rechtssicherheit garantiert werden kann.

Wie die NARA in ihrem Whitepaper festhalten (2015, S. 8), schließt der Capstone Approach die Übernahme von E-Mails aus einem RMS nicht aus. Dies war für das Hochschularchiv wichtig, da bereits geplant war, in den nächsten Jahren die Geschäftsverwaltung an der ETH Zürich durch ein RMS zu ergänzen. Der Capstone Approach soll in diesem Zusammenhang dazu dienen, diejenigen E-Mails zu archivieren, die keinen Eingang in ein RMS gefunden haben.

Erschließung von E-Mails

Eine Schwierigkeit bei der Erschließung von E-Mails ist der Umstand, dass E-Mails innerhalb eines Postfachs unterschiedlich gefiltert und dargestellt werden können. Die Ordnung eines Postfachs kann variabel eingestellt werden, so dass keine klare Struktur für die Erschließung vorgegeben ist. Zusätzlich können Ordner innerhalb eines Postfachs angelegt werden, welche einen Hinweis auf den Kontext der einzelnen E-Mails geben können. Diese Ordnerstruktur entsteht jedoch meist ohne die Hilfe von Archivarinnen oder Archivaren und kann kein klassisches Records Management System ersetzen. Als Archivarin oder Archivar muss also in der Regel damit gearbeitet werden, was vorgefunden wird. Außerdem ist bei solch großen Datenmengen die Erschließung auf Einzelstückebene kaum möglich oder auch nicht sinnvoll, zumal innerhalb von E-Mail-Postfächern mit Volltext gesucht werden kann, womit die Auffindbarkeit ohnehin in einer Form gegeben ist, die bei einer detaillierten Erschließung kaum zu erreichen wäre (vgl. Bunn, 2021, S. 26; Task Force on Technical Approaches to Email Archives, Andrew W. Mellon Foundation und Digital Preservation Coalition, 2018, S. 7-8; Zhang, 2012, S. 185-186).

In mehreren Beiträgen von Spezialisten für E-Mail-Archivierung, wie Jenny Bunns *DPC Technology Watch Report* (2021) oder dem Report *The future of email archives* der Task Force on Technical Approaches to Email Archives (2018) wird deshalb eine möglichst niederschwellige

Art der Erschließung für E-Mails vorgeschlagen: Der „More Product, Less Process"-Ansatz von Dennis Meissner und Mark Greene (2005). Dieser eignet sich besonders für die Erschließung von umfangreichen digitalen Beständen, die an sich bereits viele Metadaten enthalten, die nicht zuerst manuell erfasst werden müssen (Belovari, 2019, S. 197-198).
Der übergeordnete Zweck der Erschließung der E-Mail-Konten soll es sein, dass diese innerhalb des Archivs für Recherchezwecke sowie von Forschenden für wissenschaftliche Zwecke genutzt werden können. Letzteres wird jedoch noch einige Jahre auf sich warten lassen, da davon ausgegangen werden muss, dass sich innerhalb der E-Mail-Konten besonders schützenswerte Personendaten befinden. Diese unterliegen nach dem Bundesgesetz über die Archivierung (Stand 2023) einer 50-jährigen Schutzfrist. Da die ältesten E-Mails aus dem Vault Ende der 1990er-Jahre entstanden sind, werden die entsprechenden E-Mails noch mehrere Jahre nicht einsehbar sein. Trotzdem sollte die Auffindbarkeit und Durchsuchbarkeit sowohl für den internen Gebrauch als auch für gewährte Einsichtsgesuche gewährleistet werden. Dazu muss ein Ansatz gewählt werden, der einen möglichst großen Output mit möglichst wenig Aufwand generiert.
Dies führt konkret zu einer Erschließung, die auf der Bestandsebene das E-Mail-Archiv beschreibt und darunter, nach Serien geordnet, die einzelnen Funktionsträger aufführt. Diese Serien wiederum umfassen Jahresdossiers, die kaum weitere Informationen enthalten, es aber ermöglichen, die Schutzfristen auf ein Jahr genau zu setzen und die Durchsuchbarkeit auf Jahresebene zu ermöglichen. Diese Erschließung ist mit verhältnismäßig wenig Aufwand verbunden und doch klar genug strukturiert, dass man sich darin zurechtfindet.

Langzeitarchivierung

Mit der Langzeitarchivierung der E-Mails soll die Authentizität sowie die Nachnutzbarkeit bzw. Auffindbarkeit über einen unbestimmten Zeitraum hinweg gewährleistet werden. Dabei stellt sich zunächst die Frage nach geeigneten Dateiformaten, in welchen die E-Mails archiviert werden sollen. Oft gibt es einen Zielkonflikt: Archivtaugliche Dateiformate eignen sich für die Archivierung, aber nur bedingt für die Nachnutzung, während für die Nachnutzung geeignete Formate nicht den Anforderungen archivtauglicher Dateiformate entsprechen. Durch die Konvertierung lassen sich Dateiformate normalisieren und Informationen können in zweckbestimmten Dateiformaten archiviert werden, wobei dadurch das Risiko eröffnet wird, dass die originalen Dateien versehentlich verändert werden. Zudem gehen bei den meisten Konvertierungen bestimmte Eigenschaften oder gar Informationen im Zielformat verloren. Was die Archivierung von E-Mails betrifft, dürften die Originaldaten neben der Gewährleistung der Authentizität

vorerst auch eine optimale Grundlage für die Nachnutzung bieten, da mit den Originaldaten auch sämtliche Metadaten mitgespeichert werden und in der originalen Software genutzt werden können (z.B. Volltextsuche über das ganze Postfach).

Dateiformate

Die Wahl der Dateiformate für die zu archivierenden Daten gilt allgemein als Schlüsselfaktor für die Langzeitarchivierung. Eine gute Übersicht über Kriterien, welche archivtaugliche Formate aufweisen sollten, bietet Ludwig im *nestor Handbuch* (2010, S. 146-148) in Kapitel 7 über Formate bzw. im Unterkapitel 7.3 „Auswahlkriterien". Demnach sind Dateiformate dann für die Langzeitarchivierung empfehlenswert, wenn sie unter anderem offen spezifiziert sowie weit verbreitet sind und möglichst wenig Abhängigkeiten zu spezifischen Softwareprodukten aufweisen. Die Spezifikation des Dateiformats erlaubt neben dem Aufbau von Formatwissen gegebenenfalls die Extraktion und Speicherung technischer Metadaten und die Validierung von Dateien gegen die jeweilige Spezifikation. Eine weite Verbreitung führt einerseits zu tendenziell besserer Unterstützung der entsprechenden Dateiformate, andererseits gilt das Risiko von Obsoleszenz als reduziert. Dieses wird auch durch die Wahl von Dateiformaten, die unabhängig einer bestimmten proprietären Software gelesen und interpretiert werden können reduziert.
Im Rahmen des Projekts „E-Mail Archiv 2.0" wurden einzelne E-Mails im MSG- und zusätzlich die ganzen Postfächer im PST-Dateiformat abgeliefert. Als Folge der häufigen Verwendung von Microsoft Office Programmen sind die Outlook-Dateiformate MSG für einzelne E-Mails und PST für Postfächer sehr weit verbreitet. Es handelt sich außerdem um Formate, die offen spezifiziert sind (Microsoft, 2022a und 2022b). Während sowohl die Verbreitung als auch die zugänglichen Dateiformatspezifikationen insgesamt gute Voraussetzungen für die Langzeitarchivierung darstellen, bleiben Bedenken bezüglich ihrer Abhängigkeit von Outlook, einer proprietären Software[1]. Textbasierte Formate wie EML für einzelne E-Mails und MBOX für ganze Postfächer sind gänzlich unabhängig einer proprietären Software lesbar, was in Bezug auf die Langzeitarchivierung für diese Dateiformate spricht (vgl. Task Force on Technical Approaches to Email Archives, 2018, S. 59). Es gibt dennoch gute Argumente für die Bewahrung der originalen Dateien im MSG- bzw. PST-Format. Vor dem Hintergrund der teilweise sehr umfangreichen Ablieferungen wäre die Konvertierung nach EML bzw. MBOX äußerst aufwändig. Da es sich bei den E-Mails um besonders wichtige Daten handelt, müsste zudem eine lückenlose Qualitätskontrolle durchgeführt werden, die mit nicht leistbarem Aufwand verbunden

[1] Laut der Library of Congress (Library of Congress, 2024) wurden zumindest zwei Open-Source-Programmbibliotheken für das Auslesen und Manipulieren von PST-Dateien entwickelt. Diese wurden im Rahmen des Projektes aber nicht weiter getestet.

wäre. Mit der Erhaltung der Originaldaten kann dieses Problem umgangen und gleichzeitig die Authentizität der Daten gewährleistet werden, da sie ab dem Quellsystem unverändert übernommen und archiviert werden.

Zusätzlich zur Erhaltung der Originaldaten als PST wurde beschlossen, sämtliche E-Mails, die als MSG-Dateien vorliegen ins PDF/A zu konvertieren. Damit werden die Daten in einem für die Langzeitarchivierung de facto standardmäßig verwendeten Dateiformat archiviert und die Nachnutzung der einzelnen E-Mails ist unabhängig von Outlook möglich.

Metadatenextraktion

Für die Nachnutzung einzelner E-Mails ist die Extraktion von Metadaten auf Ebene des einzelnen E-Mails und die Übernahme dieser Metadaten in das Langzeitarchivsystem notwendig, um die Daten besser durchsuchbar zu machen.[2] Wir nutzen dazu das Toolkit Apache Tika.[3] Dabei ist es aber weder sinnvoll noch zielführend, alle Metadaten von Tikas Ausgabe zu übernehmen, da teilweise die identischen Informationen in unterschiedlicher Form ausgegeben werden. Zudem erfordern einige Metadaten tiefgehende technische Kenntnisse und sind für die Nachnutzung im Langzeitarchivsystem irrelevant. Die wichtigen Metadaten mussten entsprechend ausgewählt, extrahiert und den vom Langzeitarchivsystem genutzten Metadatenstandard Dublin Core zugeordnet werden.[4]

TIKA	DC-Feld Rosetta
Message:From-Name	FILE - Creator (DC)
Message:From-Email	FILE - Creator (DC)
Message-To	FILE - Contributor (DC)
Message-Recipient-Address	FILE - Contributor (DC)
Message-Cc	FILE - Contributor (DC)
Message-Bcc	FILE - Contributor (DC)
dc:title	FILE - Description (DC)
dcterms:created	FILE - Date (DC)
Message:Raw-Header:X-MS-Has-Attach	FILE - Description (DC)
meta:mapi-message-class	FILE - Type (DC)

Abbildung 1: Screenshot der Mappingtabelle. Die linke Spalte enthält die ausgewählten, von Tika extrahierten Metadatenfelder, die rechte Spalte die Dublin-Core-Felder im Langzeitarchivsystem Rosetta, welche für die Speicherung der Metadaten auf Dateiebene genutzt werden.

[2] Das Langzeitarchivsystem Rosetta bietet in der derzeit im ETH Data Archive eingesetzten Version keine brauchbare Volltext-Suchfunktionalität.
[3] Siehe Website von Apache Tika: https://tika.apache.org/ (18.9.2024).
[4] In diesem Zusammenhang spricht man von einem Mapping. Es definiert, in welche Felder im Zielsystem die Metadaten aus dem Quellsystem gespeichert werden. Es kann sich dabei um unterschiedliche Felder in Ziel- und Quellsystem sowie um unterschiedliche Metadatenstandards handeln.

PDF/A-Konvertierung

Für die Konvertierung der MSG-Dateien nutzten wir den bei uns schon länger im Einsatz stehende 3-Heights Document Converter.[5] Der Document Converter erlaubt es, die Inhalte, Metadaten und Anhänge in eine PDF/A-Datei zu konvertieren. Die Konvertierung des Anhangs führt allerdings nicht selten zu Fehlern, weshalb es sinnvoll sein kann, Anhänge von der Konvertierung auszuschließen, wodurch die E-Mails ohne die Anhänge konvertiert werden. Um die Fehlerrate zu eruieren, wurden testweise mehrere tausend E-Mails in einem zweistufigen Verfahren mit dem 3-Heights Document Converter konvertiert: Im ersten Durchlauf mit Berücksichtigung des Anhangs konnten rund 80% der E-Mails konvertiert werden. Im zweiten Durchlauf, in welchem nur die rund 20 Prozent der E-Mails verarbeitet wurden, welche im ersten Durchlauf nicht konvertiert werden konnten, wurde die Konvertierung der Anhänge ausgeschlossen. Davon konnten nahezu alle restlichen E-Mails konvertiert werden. Insgesamt konnten so 99,8% der E-Mails aus dem Test-Set konvertiert werden. Wir haben deshalb beschlossen, die Konvertierung im zweistufigen Verfahren vorzunehmen. Nicht konvertiere Anhänge werden allerdings nicht separat extrahiert, da dies wiederum die Komplexität des Prozesses erhöht und somit mehr Ressourcen beansprucht hätte. Vor dem Hintergrund, dass sämtliche Originaldateien mitarchiviert werden, ist es für uns vertretbar, die nicht konvertierbaren Anhänge nicht separat zu archivieren. Außerdem ist in den Metadaten zu den PDF-Dateien ersichtlich, ob das E-Mail ursprünglich einen Anhang besaß oder nicht.

Bei Konvertierungen ist immer eine Qualitätskontrolle notwendig, wobei die Handlungsfähigkeit bei unerwünschten Ergebnissen meist nicht gegeben ist. In unserem Fall waren wir mit Problemen wie unschönen Umbrüchen im PDF bei sehr breiten Tabellen konfrontiert, ein Umstand, der so hingenommen werden musste. Audiovisuelle Datenanhänge können selbstredend nicht ins PDF/A konvertiert werden und wurden entsprechend ausgelassen.

Workflow

Was den Export der E-Mails aus dem Vault betrifft, werden diese von den ID auf ein eigens für diesen Anwendungsfall eingerichteten Share auf einem Network Attached Storage (NAS) mit eingeschränkten Zugriffsberechtigungen exportiert. Für jedes Postfach wird für den Zeitraum eines Jahres einerseits die Postfach-Datei (PST) sowie separat die darin enthaltenen E-Mails (MSG-Dateien) exportiert. Dieser Prozess wurde während des Projekts für alle vordefinierten Postfächer rückwirkend durchgeführt, weshalb in der Anfangsphase besonders viele Daten

[5] Der 3-Heights Document Converter wird nicht mehr explizit auf der Herstellerseite aufgeführt: https://www.pdf-tools.com/ (13.9.2024).

exportiert wurden. Nach der initialen Übernahme wird dieser Prozess für die vom HSA als archivrelevant definierten Postfächer jährlich wiederholt.

Abb. 2 gibt einen groben Überblick über die wesentlichen Prozesse, die im Rahmen der Langzeitarchivierung für das Projekt „E-Mail-Archiv 2.0“ durchgeführt werden:

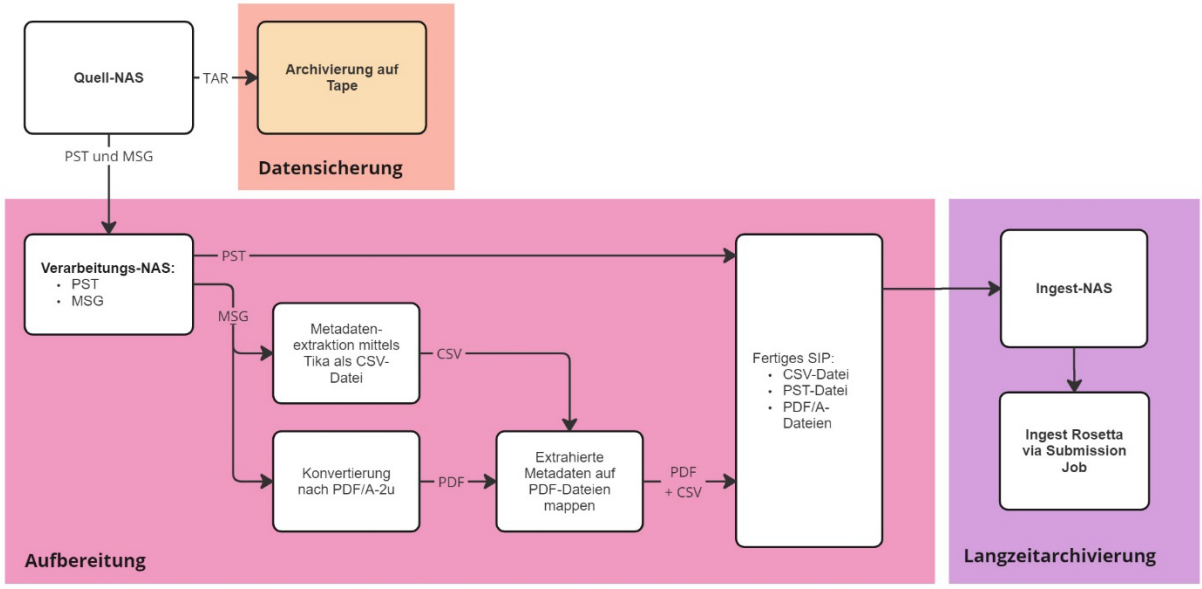


Abbildung 2: Übersicht der verschiedenen Phasen und der ablaufenden Prozesse im Projekt E-Mail-Archiv 2.0

Datensicherung

Im Rahmen des Projekts wurden zuerst die Altdaten verarbeitet. Da es sich dabei um einen sehr großen Bestand handelt, wurden die Originaldaten auf Tape gesichert. Diese zusätzliche Kopie sollte einen zusätzlichen Schutz für die Authentizität und Datenintegrität bieten, solange sich die Daten während der Aufbereitungsphase noch nicht im Langzeitarchivsystem befanden. Ein Script erstellte dabei Tar-Archive geeigneter Größe, welche den Vorgaben der ID für Dateien auf dem Tape-Speicher entsprechen.[6] Bei den jährlichen Ablieferungen kann auf diese zusätzliche Sicherung verzichtet werden, da es sich um wesentlich geringere Datenmengen handelt, welche direkt verarbeitet werden können.

Aufbereitung

In der Phase der Aufbereitung werden die MSG-Dateien einerseits für die Extraktion von Metadaten mittels Apache Tika sowie für die PDF/A-Konvertierung mit dem 3-heights Document Converter genutzt und danach gelöscht. Die relevanten Metadaten werden per Script (vgl.

[6] Dateien für Tape-Speicher sollten 10 GB nicht unterschreiten (ETH, 2019).

Abb. 2) mit den entsprechenden PDF/A-Dateien zusammengeführt und in eine CSV-Datei geschrieben, die für den Ingest-Vorgang ins Langzeitarchiv genutzt wird.

Das fertige Submission Information Package (SIP) besteht aus der in der Aufbereitungsphase erstellten CSV-Datei mit den für den Import ins Langzeitarchiv notwendigen Metadaten. Neben den extrahierten Metadaten aus den E-Mails enthält es manuell erfasste deskriptive Metadaten zum Postfach[7] sowie langzeitarchivsystemspezifische Metadaten, welche die Zuordnung zu den richtigen Workflows und Prozessen im Langzeitarchivsystem gewährleisten. Die Ordnerstruktur des SIPs ist für den CSV-Ingest vorgeschrieben. Während sich die Metadendatei im SIP-Ordner im Unterordner („contents“) befinden muss, müssen sich die zu archivierenden Dateien (PDF/A-Dateien und die entsprechend originale PST-Datei) im Ordner „streams“ befinden, welcher seinerseits wiederum einen Unterordner von „contents“ darstellt.

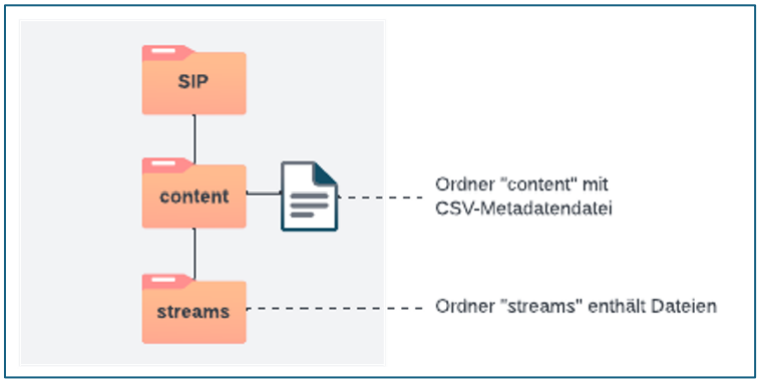


Abbildung 3: SIP-Ordner-Struktur für den CSV-Ingest

Langzeitarchivierung

Das im ETH Data Archive eingesetzte Langzeitarchivsystem Rosetta[8] kann mit einem sogenannten Submission Job so konfiguriert werden, dass der für die Ablage der SIPs vorgesehene Speicherpfad in beliebig definierbaren Abständen nach neuen SIPs überprüft wird. Somit werden neue SIPs automatisiert in das Langzeitarchivsystem übernommen. Dabei wird neben einem Integritäts- und Virencheck auch eine Formatidentifikation durchgeführt. Die mit Apache Tika extrahierten Metadaten werden indexiert und sind somit im Langzeitarchivsystem für die Suche nach einzelnen E-Mails nutzbar.

[7] Diese Metadaten entsprechen den im AIS (Archival Information System) erfassten Metadaten. Darunter fallen unter anderem Titel, Entstehungszeitraum sowie der im AIS generierte Digital Objekt Identifier (DOI).
[8] Rosetta wird in den nächsten Jahren nicht mehr weiterentwickelt und durch den Nachfolger „Specto Preservation“ abgelöst werden. Auf der Herstellerseite findet sich zwar noch Material zu Rosetta, bei den Produkten wird aber nur noch „Specto Preservation“ aufgeführt: https://exlibrisgroup.com/products/specto-preservation-product/

Fazit

Im Rahmen des Projekts „E-Mail-Archiv 2.0“ wurden über fast zwei Jahre 42 Postfächer, verteilt auf 230 Jahrgänge übernommen, was insgesamt 2'247'634 E-Mails ausmacht. Dies entspricht einem Speicherplatzbedarf von 1.1 TB. Während des Projekts gab es diverse Herausforderungen, von denen zum Glück die meisten gelöst werden konnten: So haben wir uns dazu entschieden, Postfächer, welche Viren enthalten, dahingehend zu kennzeichnen und nicht einzelne E-Mails aus der PST-Datei zu löschen. Auch mussten wir unsere zeitliche Einschätzung neu kalibrieren, da sowohl Kopiervorgänge als auch der Ingest von sehr vielen Einzeldateien in Rosetta deutlich mehr Zeit in Anspruch nahmen als die gleichen Arbeitsschritte bei einer großen Datei mit dem gleichen Datenvolumen gebraucht hätte. Zudem gab es einige wenige Postfächer, die viel mehr E-Mails enthielten, als wir ausgehend vom gelieferten Testset angenommen hatten,[9] wodurch auch die Anzahl der zu verarbeiteten E-Mails viel höher war, als ursprünglich angenommen. Die Systemlast war außerordentlich hoch, was auch an der Indexierung der vielen Metadaten auf Ebene des einzelnen E-Mails liegen dürfte. Andere Punkte konnten nicht umgesetzt werden, wie zum Beispiel der Umgang mit verschlüsselten E-Mails. Selbst der Austausch mit der E-Mail-Archivierungs-Community konnte dahingehend keine zufriedenstellende Antwort liefern.

Der Abschluss des Projekts war trotz aller Schwierigkeiten ein großer Erfolg und die E-Mail-Archivierung ist mittlerweile in abgewandelter Form in den Standard-Workflow für die digitale Archivierung im Hochschularchiv der ETH Zürich integriert worden. Da dort nun der Foxit PDF-Compressor[10] verwendet wird, muss die PDF/A-Konvertierung noch auf den neuen Konverter umgestellt und getestet werden. Zudem stehen bereits die nächsten Herausforderungen an: Die Umstellung auf Exchange Online und die Einführung eines Records Management Systems.

Bibliografie

Artefactual Systems and the Digital Preservation Coalition (2021), *Preserving Email, DPC Technology Watch Report*. Digital Preservation Coalition.

Belovari, S. (2019), ‘Expedited Digital Appraisal for Regular Archivists: An MPLP-Type Appraisal Workflow for Hybrid Collections’, *Journal of Archival Organization*, 16(4), S. 197-219.

Benauer, M. (2020), ‘E-Mails, ihr Wert und ihre Bewertung’, *Scrinium*, 74, S. 87–115.

Bunn, J. (2021), *Born digital archive cataloguing and description, DPC Technology Watch Report*. Digital Preservation Coalition.

ETH Zürich (2019), *Long Term Storage (LTS)* [Online], https://ethz.ch/content/dam/ethz/associates/services/Service/IT-Services/files/catalogue/storage/lts/long-term-storage-lts.pdf (11.9.2024).

ETH Zürich (2021), *Standard Dienstleistungsvereinbarung (Service based SLA) für „Mail und Groupware“*.

Knobloch, C. (2016), ‘Überlegungen zur Übernahme und Archivierung von E-Mail-Konten’ in *Digitale Archivierung. Innovationen – Strategien – Netzwerke. Tagungsband zur 19. Tagung des Arbeitskreises*

[9] 49% aller E-Mails gehörten nur zu 10% der Postfächer.

[10] Siehe Hersteller-Seite unter: https://www.foxit.com/de/ (13.9.2024).

„Archivierung von Unterlagen aus digitalen Systemen“ (Mitteilungen des Österreichischen Staatsarchivs 59/2016). Wien, S. 221-231.

Library of Congress (2024), *Sustainability of Digital Formats: Planning for Library of Congress Collections* [Online], https://www.loc.gov/preservation/digital/formats/fdd/fdd000377.shtml (30.9.2024).

Ludwig, J. (2010), ‘Auswahlkriterien’ in *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.3*. S. 146–148.

Microsoft (2022a) *[MS-OXMSG]: Outlook Item (.msg) File Format* [Online], https://learn.microsoft.com/en-us/openspecs/exchange_server_protocols/ms-oxmsg/b046868c-9fbf-41ae-9ffb-8de2bd4eec82 (18.9.2024).

Microsoft (2022b), *[MS-PST]: Outlook Personal Folders (.pst) File Format* [Online], https://learn.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/141923d5-15ab-4ef1-a524-6dce75aae546 (18.9.2024).

Meissner, D./Greene, M.A. (2010), ‘More Application while Less Appreciation: The Adopters and Antagonists of MPLP’ in *Journal of Archival Organization*, 8 (3–4), S. 174-226.

National Archives and Records Administration (NARA) (2013), ‘Guidance on a New Approach to Managing Email Records’ in *NARA Bulletin*, 2013-02.

National Archives and Records Administration (NARA) (2015), *White Paper on The Capstone Approach and Capstone GRS*.

Schwarz, K. (2010), ‘E-Mail-Archivierung’ in *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.3*. S. 550-563.

Task Force on Technical Approaches to Email Archives, Andrew W. Mellon Foundation, Digital Preservation Coalition (Hgg.) (2018), *The future of email archives: A report from the Task Force on Technical Approaches for Email Archives*. Washington, DC (CLIR publication 175).

Yeo, G. (2019), ‘Can we keep everything? The future of appraisal in a world of digital profusion’ in Brown, C. (Hrsg.), *Archival Futures*. London, S. 45–63.

Zhang, J. (2012), ‘Original Order in Digital Archives’ in *Archivaria*, 74, S. 167-193.

V.

RECORDS MANAGEMENT, ÜBERNAHME UND ERSCHLIESSUNG

Sonderfall Universität: Ein Nachlass aus der Cloud

Christine Rigler

An einem Universitätsarchiv werden nicht nur Verwaltungsunterlagen und Institutsbestände, sondern auch Unterlagen von Einzelpersonen, meist Angehörigen des wissenschaftlichen Personals, verwahrt. Im Umfang variieren diese Bestände von kleinen Konvoluten bis zu umfangreichen Privatarchiven. Materialien aus persönlicher Provenienz sind rechtlich anders gestellt als Verwaltungsunterlagen und Sammlungen, die im Eigentum der Universität stehen und daher einer Anbietungspflicht unterliegen. Vor- und Nachlässe sind Privatbesitz und werden freiwillig, oft im Rahmen einer Schenkung, übergeben. Die Abgabe erfolgt entweder zu Lebzeiten der betroffenen Personen (Vorlass) oder nach dem Tod durch Rechtsnachfolger (Nachlass). Vor- und Nachlässe sind individuelle, gemischte Bestände, die Schriftgut, Bildmaterial, audiovisuelle Aufzeichnungen, verschiedenste Arten von Sammlungen und Erinnerungsgegenständen nicht nur in materieller, sondern auch in digitaler Form enthalten können.

Digitale Aufzeichnungen als Archivgut manifestieren einen tiefgreifenden Kulturwandel. Die Etablierung digitaler Technologien verändert nicht nur einzelne Arbeitsabläufe, sondern greift massiv in Organisationsstrukturen ein. Der technologisch basierte Wandel führte besonders in Einrichtungen und Betrieben, die nach öffentlich-rechtlichen Kriterien geführt werden, zu einer teilweisen Personalisierung der Arbeitsumgebung. Diese Entwicklung hat auch Auswirkungen auf das Selbstverständnis der Arbeitnehmer.

Allen Mitarbeiterinnen und Mitarbeitern einer Universität etwa wird eine persönliche Arbeitsumgebung zur Verfügung gestellt, die durch ein Passwort geschützt ist, das jeder/jede für sich selbst festlegt und geheim hält. Bemerkenswert ist daran vor allem, dass auch hierarchisch übergeordnete Personen diesen Bereich, der üblicherweise zumindest einen E-Mail-Account und ein persönliches Laufwerk zur Dateiablage beinhaltet, nicht einfach einsehen können. In Österreich weisen sowohl die Kammer für Arbeiter und Angestellte als auch die Wirtschaftskammer (zwei der wichtigsten berufsständischen Körperschaften) in ihrer Rechtsauskunft im Internet darauf hin, dass es dem Arbeitgeber nicht erlaubt ist, die E-Mails von Mitarbeitern einzusehen, wenn die private Nutzung des E-Mail-Accounts nicht ausdrücklich untersagt ist. (Arbeiterkammer, no date; Wirtschaftskammer, 2023). Dasselbe gilt für Dateien, die auf persönlichen Laufwerken abgelegt werden. Damit soll verhindert werden, dass aus Versehen private Inhalte, die nicht als solche erkannt werden, von Unbefugten gelesen werden. Aus diesem Grund empfehlen Organisationen, die eine private Nutzung der betrieblichen Infrastruktur

gestatten, ihren Mitarbeitern, private Inhalte in eigens gekennzeichnete Ordner zu verschieben. Denn es könnte ein Ereignis eintreten, zum Beispiel ein Cyberangriff oder ein plötzlicher Todesfall, dass der Zugriff auf den passwortgeschützten Arbeitsbereich doch notwendig wird.
Das unberechtigte Lesen von E-Mails, die durch ein Passwort geschützt sind, gilt nach österreichischem Recht als Verletzung des Briefgeheimnisses (Mosing, 2001, S. 7). Am Arbeitsplatz wäre eine transparente Gestaltung der E-Mail-Korrespondenz durch die Funktion Carbon Copy (CC), also den Einschluss beliebig vieler Empfänger, zwar jederzeit leicht möglich, ist aber freiwillig. Da eine E-Mail innerhalb der Organisation kein Sekretariat und keine Poststelle passieren muss wie ein physischer Brief, findet sehr viel schriftliche Kommunikation direkt bilateral und von Außenstehenden unbemerkt statt. Es ist dadurch auch leichter geworden an Personen heranzutreten, die früher organisatorisch abgeschirmt waren. Die Abschirmung besteht nach wie vor, aber gerade die persönliche E-Mail-Box wird von Führungskräften oft ganz bewusst als direkter Kommunikationskanal offengehalten.

Besonderheiten von Vor- und Nachlässen

Vor- und Nachlässe waren immer schon Archivbestände, für die eine Verflechtung privater und berufsbezogener Inhalte geradezu typisch ist. Darin liegt ihre Bedeutung als historische Quelle: dass wir über Aufzeichnungen verfügen, die uns einerseits faktische Information (im Sinne unveröffentlichter Tatsachen, Erkenntnisse oder Werke) bieten, aber auch tiefere Einblicke in individuelle Denk-, Arbeits- und Lebensweisen bestimmter Personen ermöglichen. Der Philosoph Wilhelm Dilthey prägte in diesem Zusammenhang bereits 1889 das Bild der geistigen *Werkstatt*. Als einer der ersten warb er dafür, Handschriften und Briefe neben den gedruckten Werken in öffentlichen Einrichtungen zu archivieren. Dilthey sah in den von ihm propagierten „Archiven für Literatur“[1] eine Chance für die Weiterentwicklung der Forschung. Das Archiv wirkt in dieser Vorstellung wie ein Labor, in dem die Vielfalt der Quellen zur Erprobung neuer Methoden anregt (Dilthey, 1889, S. 365–367). Interessant ist, dass dieser frühe Verfechter der Nachlasserhaltung gegen einen Einwand argumentieren musste, der uns in Zeiten der Digitalisierung wohlbekannt ist, nämlich die Befürchtung, dass die Menge des zu archivierenden Materials zu groß werden könnte: „Wer kennt nicht die Klage, Erhaltung und Druck solcher Papiere diene nur einem gelehrten Interesse; ja schließlich breche in diesen ungeheuren Papiermassen und ihre Vervielfältigung durch den Druck das neue alexandrinische Zeitalter über uns herein“ (Dilthey, 1889, S. 363).

[1] Unter „Literatur“ verstand Dilthey nicht nur die Dichtkunst, sondern „alle dauernd wertvollen Lebensäußerungen eines Volkes, die sich in der Sprache darstellen: also Dichtung wie Philosophie, Historie wie Wissenschaft“ (Dilthey, 1889, p. 365).

Die Grenze zwischen Institutseigentum und Privatbesitz ist bei Materialsammlungen, die sich bei Universitätsangestellten im Zuge ihrer Forschungs- und Lehrtätigkeit ansammeln, nicht immer eindeutig zu bestimmen. Wissenschaftliche Errungenschaften gelten als geistige Eigenleistungen von Einzelpersonen oder Personengruppen, sind aber zugleich von Finanzierung abhängig. Arbeit- und Fördergeber, die im öffentlichen Auftrag agieren, bestehen zu Recht auf einer gesellschaftlichen Verwertbarkeit der geförderten Leistungen und verlangen die Freigabe der Forschungsdaten. In diesem Spannungsfeld muss ein Ausgleich gefunden werden.
Auch an persönliche Archive ist ein Anspruch auf Authentizität zu stellen. Als ursprünglichen Zustand kann man den natürlichen Entstehungszusammenhang betrachten, in dem sich Unterlagen zu Leben und Werk einer Person zusammenfinden. Dieser authentische Ur-Zustand ist keine starre Größe, sondern veränderlich und es ist bestenfalls möglich, ihn in einer Momentaufnahme zu erhalten oder zu dokumentieren. Wenn nun eine Person beschließt, ihre Unterlagen selbst zu Lebzeiten abzugeben, findet ein Auswahl- und Ordnungsprozess zum Zweck der Übergabe statt. Man inszeniert quasi eine Überlieferung von sich selbst als Person, nach eigenen Vorstellungen. Es entsteht ein Übergabe-Zustand, der insofern authentisch ist, als er von der bestandsbildenden Person selbst geschaffen wurde, der aber bereits eine Abweichung darstellt. Im anderen Fall, wenn eine Person stirbt, ohne solche Vorkehrungen getroffen zu haben, gerät die Hinterlassenschaft unweigerlich in andere Hände. Ist das Umfeld ein privates, dann ist der Zustand zum Zeitpunkt der Übergabe an ein Archiv in Bezug auf seine Authentizität in jedem Fall kritisch zu hinterfragen. Eingriffe Dritter sind immer in Betracht zu ziehen, auch dann, wenn sie keine erkennbaren Spuren hinterlassen (Grond-Rigler, 2018, S. 163–167).
All dies gilt natürlich auch für jene Vor- oder Nachlassteile, die nicht in materieller Form, sondern digital vorliegen. Im Wissenschaftsbereich nehmen die digitalen Anteile derzeit generationsbedingt zu. Jene Wissenschafter:innen, die derzeit das Pensionsalter erreichen, sind noch keine Digital Natives, mussten sich aber – mehr oder weniger freiwillig – digitale Techniken aneignen. In den Beständen dieser Übergangsgeneration, und vor allem, wenn sie einen langen Zeitraum abdecken, wird sich der digitale Aufbruch besonders gut nachvollziehen lassen.

Fallgeschichte

Diese allgemeinen Beobachtungen und Überlegungen zur digital gestützten geistigen Arbeit im universitären Umfeld möchte ich mit einem Fallbeispiel fortführen. 2023 verstarb ein pensionierter Universitätsprofessor nach längerer Krankheit. Wie es an der Universität Graz üblich ist, hatte er im Ruhestand seinen Universitätsaccount weiterhin nützen können. Dieser Account schließt die Nutzung eines webbasierten Softwarepakets ein, das Mitarbeitern kostenfrei

angeboten wird und auch einen Clouddienst umfasst. Das Softwarepaket kann auf allen Geräten installiert und verwendet werden und steht explizit nur für die private Nutzung zur Verfügung. Die Anmeldung muss jedoch mit dem Universitätskonto erfolgen. Auch unser Universitätsprofessor nutzte dieses Arrangement. Er verwendete die Software und speicherte Dateien in der für den privaten Gebrauch reservierten Cloud. Mit seinem Tod stellte sich eine Pattsituation ein: Auf Seiten der Erben bildete der Universitätsaccount eine Schranke für den Zugriff auf diese Dateien; die Universität wiederum verfügt zwar über die technischen Möglichkeiten zur Umgehung des Passwortschutzes, ist aber rechtlich nicht dazu befugt, vermeintlich private Inhalte ihrer Mitarbeiter einzusehen.

Zu diesem Zeitpunkt war an der Universität noch keine Vorgangsweise für eine derartige Situation festgelegt worden, daher wurde anlassbezogen ein individuelles Prozedere entwickelt. Die Öffnung des Accounts erfolgte in Anwesenheit der Erben sowie eines Mitarbeiters der IT-Abteilung und zweier Jurist:innen, von denen eine auch die Funktion der Datenschutzbeauftragten innehatte. Die Dateien wurden in dieser Gruppe gemeinsam und von allen Beteiligten zum ersten Mal gesichtet. Es zeigte sich, dass sowohl private als auch universitätsbezogene Inhalte vorhanden waren. Die Erben bekamen Gelegenheit, private Dateien für sich zu kopieren, die dann aus dem Bestand gelöscht wurden. Die Vertreter der Universität achteten zugleich darauf, dass von den Erben keine betriebsinternen Informationen (zum Beispiel Prüfungsunterlagen, Institutsprotokolle, Unterlagen zu Gremiensitzungen) entnommen und die Rechte Dritter nicht verletzt wurden. All dies geschah in einer einzigen Sitzung, die laut Protokoll zweieinhalb Stunden dauerte. Bei diesem Zusammentreffen wurde auch letztgültig entschieden, was an das Archiv gehen sollte, nämlich sowohl jene Unterlagen, die als privater Nachlass zu klassifizieren sind, als auch jene, die als Universitätseigentum zu betrachten sind. Für die Übernahme der Nachlassmaterialien schließen die Erben einen Schenkungsvertrag mit der Universität.

Fazit

Das Universitätsarchiv ist nun im Besitz einer Sammlung von 13.464 Ordnern und 101.954 Dateien im Umfang von 130 GB, die ein Universitätsprofessor hinterlassen hatte. Die Übergabe an das Archiv erschien sowohl den Vertretern der Universität als auch den Erben aufgrund der komplexen Rechtslage als die beste Möglichkeit, diesem elektronischen Vermächtnis zu weiterer Nutzung zu verhelfen. Andernfalls wäre im Zuge des Accountbeendigungsverfahrens, das bei Verlassen der Universität zur Anwendung kommt, alles restlos gelöscht worden. Eine genaue Analyse und Bewertung der betreffenden Dateien steht noch aus. Bemerkenswert erscheint jedenfalls der Umstand, dass weder die Erben noch die Vertreter der Universität ohne die

Anwesenheit der jeweils anderen Partei auf die Dateien zugreifen konnte und daher auch niemand die Gelegenheit hatte, unbeobachtet etwas zu verändern. Das Mehraugenprinzip hat eine Minimierung der Eingriffe in den Bestand sichergestellt und die digitalen Unterlagen sind in einem Zustand ins Archiv gelangt, der dem vom Bestandsbildner hinterlassenen Zustand sehr nahe ist. Dieses Ausmaß an Authentizität ist sonst nur gewährleistet, wenn eine Person ihre Unterlagen selbst als Vorlass an das Archiv übergibt. Zwar wurden einzelne private Unterlagen aus inhaltlichen Gründen gelöscht, es fand aber keine ordnungsgeleitete Durchsicht und Bereinigung statt, wie wir an den vorhandenen Doubletten (zum Beispiel Verdoppelungen ganzer Ordner) sehen können. Die Ordnungsstruktur, die immer als Teil einer auktorialen Inszenierung des Bestandsbildners verstanden werden muss, blieb unverändert. Vielleicht werden Bestände wie dieser in der Zukunft für sozialgeschichtliche Studien zur Computerisierung von Universitätsmitarbeitern herangezogen. Es scheint fast so, als hätte der verstorbene Kollege selbst etwas Derartiges im Sinn gehabt, denn es wurde uns berichtet, dass er die elektronischen Unterlagen vor seinem Tod noch bewusst durcharbeitete. Es fällt zum Beispiel auf, dass einige Ordner erhalten blieben, deren Inhalt aber vollständig gelöscht worden war – so als hätte er die ursprüngliche Ordnungsstruktur nicht zerstören wollen.

Die strenge Gesetzgebung zum Schutz personenbezogener Daten wird aus Archivsicht oft als Limitierung empfunden, in unserem Beispiel hat sie zur Qualität der Überlieferung aber durchaus beigetragen. Der Datenschutz kam in seinem eigentlichen Wortsinn zur Anwendung. Es wurden nicht nur die Menschen hinter den Daten, sondern auch die Daten an sich geschützt.

Bibliografie

Arbeiterkammer (no date), *Überwachung am Arbeitsplatz – Kontrolle der E-Mail und Internetnutzung,* Portal der Arbeiterkammern [online]. https://www.arbeiterkammer.at/beratung/arbeitundrecht/Arbeitsklima/Big_Brother_am_Arbeitsplatz.html (5.6.2024).

Dilthey, W. (1889), ‚Archive für Literatur', Deutsche Rundschau 58, S. 360–375. https://archive.org/details/deutscherundscha58stutuoft/page/360/mode/2up (9.7.2024).

Grond-Rigler, C. (2018), ‚Im Dialog mit der Nachwelt: Auktoriale Inszenierung in Vorlässen' in Dallinger, P.-M., Hofer, G., Judex, B. (eds.) *Archive für Literatur: Der Nachlass und seine Ordnungen.* Berlin, Boston: De Gruyter, S. 163–180. https://doi.org/10.1515/9783110594188-011

Mosing, M. W. (2001), *Briefgeheimnis im Strafrecht und E-Mail in Ö und D. Ein Microvergleich*, it-law.at [online]. https://www.it-law.at/publikation/briefgeheimnis-und-e-mail-oe-und-d-aus-strafrechtlicher-sicht/ (9.7.2024).

Wirtschaftskammer (2023), *Internet und E-Mail am Arbeitsplatz. Welche Kontrollmaßnahmen sind erlaubt* (2023) Wirtschaftskammer Oberösterreich [online]. https://www.wko.at/ooe/arbeitsrecht-sozialrecht/internet-und-e-mail-am-arbeitsplatz-welche-kontrollmassn (9.7.2024).

OAIS-konforme Softwarearchitektur für eine Plattformlösung

Frank Obermeit

Dieser Beitrag informiert über den Abschluss eines mehrjährigen Projektes zur Entwicklung einer Cloud-fähigen Plattformlösung „Archive Digital as a Service“. Mit dieser Plattformlösung können OAIS-Referenzmodell-konforme Fachanwendungen implementiert, integriert und migriert werden.

Grundsätzliche Anforderungen

Die grundsätzlichen Anforderungen des Projektes wurden erfüllt:

- Flexible Prozessgestaltung
- Benutzerfreundliche Anwendung
- Umsetzung des Rechte- und Rollenkonzepts mit der eigenen Nutzerverwaltung
- Unterstützung von allen Schutzbedarfskategorien
- Skalierbarkeit
- Mandantenfähigkeit.

Qualitative und quantitative Projektanforderungen

Die Anforderungen an die elektronische Archivierung wachsen qualitativ und quantitativ. Qualitativ, weil archivische Fachanwendungen für unterschiedliche Fachbereiche benötigt werden. Dabei soll die Software effizient und hochwertig entwickelt werden. Diese Fachanwendungen sollen selbstverständlich wartbar bleiben, nachnutzbar und robust sein. Quantitativ, da der Datenumfang zunimmt und nur durch Automation bewältigt werden kann. Dazu müssen der Softwareentwicklungsprozess und Softwarebetrieb standardisiert und letztendlich routiniert werden. Dies ist nur mit einer ausgereiften und zukunftssicheren Software-Architektur erreichbar. Seit 2018 beschäftige ich mich mit einer Softwarearchitektur für Archive, mit dem Ziel, die Softwareentwicklung und den Softwarebetrieb effizient zu unterstützen. Cloud-Technologien haben sich durchgesetzt und ermöglichen, diese Ziele, mit einer nach der Softwarearchitektur entwickelten Plattformlösung, zu erreichen. Damit wird ein mehrjähriges Projekt abgeschlossen.

Softwarearchitektur

Warum wird eine Software-Architektur benötigt? Softwareprojekte zur Entwicklung von Fachanwendungen müssen einige grundsätzliche Probleme bewältigten. Einerseits müssen die umzusetzenden Anforderungen ausreichend definiert werden und andererseits muss explizit festgelegt werden, welche Themenbereiche wegen des personellen und finanziellen Budgets nicht ganzheitlich umgesetzt werden können. Die entstehenden Fachanwendungen bleiben also per se unzulänglich und oftmals inkompatibel gegenüber anderen Fachanwendungen. Diese grundsätzliche Problematik kann nur durch eine quasi-Standardisierung des Software-Entwicklungsprozesses aufgelöst werden.

Auf der AUdS-Tagung 2019 in Prag habe ich die grundsätzliche Struktur einer OAIS-konformen-Softwarearchitektur in einem Satz zusammengefasst: „Die Archival Storages werden mit REST gekapselt, die Prozessmodellierung und Prozessverarbeitung erfolgt mit BPMN, alles wird über eine webbasierte Benutzeroberfläche dem Nutzer zur Verfügung gestellt und die vorhandene Rechteverwaltung wird integriert."

Entstanden ist 2020 eine Software inklusive Architekturkonzept, die diese Anforderungen umsetzt und die Basis für die Plattformlösung bildet. Wesentliche Softwarebestandteile sind eine IAM-Komponente, eine Workflow-Komponente, die REST-basierte Kommunikation zwischen allen Softwarebestandteilen und die integrierende Benutzeroberfläche für die Benutzerinteraktionen. Diese Software deckt also genau die Anforderungen ab, die bei Software-Entwicklungsprojekten oftmals vernachlässigt werden müssen.

Cloud-basierte-Plattformlösung – Archive Digital as a Service

Die Plattformlösung „Archive Digital as a Service" wurde Container-basiert aufgebaut, was eine schnittstellenorientierte Softwareentwicklung erzwang. Dies ermöglichte eine elegante Überführung in eine Cloud-basierte-Plattformlösung. Kubernetes bildet den de-facto-Standard für die „Orchestrierung" von Containern. Die quantitative Anforderung der Mandantenfähigkeit löst Kubernetes durch die sogenannten Namespaces in Kombination mit „Ingress"-Regeln. Die quantitative Anforderung der technisch notwendigen Skalierbarkeit wird durch die beliebige Erweiterbarkeit mit Master- und Worker-Nodes erreicht. Auf den Worker-Nodes werden die produktiven Container – also die Fachanwendungen – ausgeführt. Als Worker-Nodes können sowohl Unix- als auch Windows-Systeme eingesetzt werden.

Kubernetes gehört zu den wesentlichen Cloud-Produkten, die bei Providern gebucht oder lokal betrieben werden können. Es muss zukünftig containerbasiert und schnittstellenorientiert entwickelt werden. Alle anderen Anforderungen wie Orchestrierbarkeit, Mandantenfähigkeit und

Skalierbarkeit sind Kubernetes-inhärente Fähigkeiten, zumal Kubernetes Betriebssystem-übergreifend nutzbar ist. Die Plattformlösung kann entsprechend den eigenen Fähigkeiten bzw. dem eigenen Budget lokal betrieben oder bei IT-Providern gebucht werden.

Perspektiven der Plattformlösung

Die IT-Dienstleister der Bundesländer bieten bereits Kubernetes als Service an. Perspektivisch werden sich Cloud-basierte Anwendungen durchsetzen. Die Plattformlösung unterstützt bei der Softwareentwicklung und dem Softwarebetrieb und bietet Archiven einen unkomplizierten Weg in die Cloud-basierte-IT.

Softwareentwicklung und Softwarebetrieb – Bestandteile der Plattformlösung

Die Plattformlösung ist ganzheitlich, weil damit Fachanwendungen nicht nur betrieben, sondern auch effizient entwickelt werden können.

Software-Integration und Migration – Bestandteil der Plattformlösung

Bestehende Fachanwendungen können in die Plattformlösung integriert werden, allerdings müssen diese zunächst containerisiert werden. Dies ist aber unproblematisch. Anschließend könnten die Fachanwendungen sukzessiv migriert werden. Als erstes würden die Unzulänglichkeiten bei der Authentifizierung und Autorisierung durch die Integration der IAM-Komponente bereinigt werden. Bei Kubernetes bildet der Pod, in dem sich 1 bis n Container befinden, die kleinste Arbeitseinheit. Angestrebt wird eine 1:1-Beziehung zwischen Pod und Container und dies wäre die nächste Softwaremigrationsaufgabe. Wahrscheinlich erfolgt dabei eine Umstellung auf REST-Services und diese bilden wiederum die Grundlage für die Integration der Workflow-Komponente.

Das Migrationsprojekt endet mit einer schlanken Fachwendung, die wesentlich einfacher gepflegt werden kann, sprich: Fachliche Anforderungen können effizient umgesetzt werden. Die Anwendungsentwickler können sich auf die Umsetzung fachlicher Anforderungen konzentrieren. Bereits durch die Integration wird die Installation standardisiert und der Konfigurationsaufwand auf wenige zu ändernde Parameter reduziert.

Beispiel für die Neuentwicklung

Die Fachanwendung X-BA zur xdomea-basierten Archivierung[1] wurde im Jahr 2022 entwickelt und zeigt die Vorteile der Plattformlösung. Was bot dabei die Plattformlösung?

[1] https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2022.html

- IAM-Komponente (Keycloak)
- Workflow-Komponente (Camunda 7)
- XML-Datenbank inkl. REST-Services (BaseX – XQuery)
- Integrierende Benutzeroberfläche für die Entwicklung und den Betrieb (APEX)

Was musste zusätzlich entwickelt bzw. integriert werden?

- REST-Services wurden mit XQuery entwickelt
- Workflows und Entscheidungstabellen wurden mit BPMN und DMN modelliert
- Fachanwendung wurde mit der integrierenden Benutzeroberfläche entwickelt
- Rechte und Rollen wurden konzipiert und konfiguriert und in Keycloak konfiguriert.

Beispiel für die Integration einer bestehenden Fachanwendung

Bei der Integration zeigen sich die Kubernetes-Vorteile und bei der Migration sind die Vorteile der Plattformlösung offensichtlich. Welche Voraussetzungen mussten für die Integration geschaffen werden (DIMAG-KM)?

- Docker - Containerisierung
- K8S - Pod- und Deployment-Konfiguration
    - MariaDB
    - SFTP
    - Apache, Ingestlist, (LoadTektonik)
    - PhpMyAdmin
- K8S - Job (Cron-Jobs)
- K8S - Services
- K8S - Ingress

Wie erfolgte die Integration (experimentelle Lösung 2021) der IAM-Komponente?

- Zwei bestehende PHP-Dateien wurden angepasst.
- Eine PHP-Funktion wurde für die IAM-Integration entwickelt.

Workflow-Modellierung

In meinen letzten Vorträgen habe ich darauf hingewiesen, dass die Workflows auch von Archivaren selbstständig gepflegt werden könnten, wenn die Workflows entsprechend der Methodik des prozessgesteuerten Ansatzes, entwickelt von Prof. Dr. Volker Stiehl, modelliert werden. Erfolgreich wurde diese Methodik bei der Fachanwendung X-BA angewendet.

Bestandteile von KaaS (Kubernetes as a Service)

Die Konfiguration einer lokalen Kubernetes (K8S) - Installation besteht aus:

- Photon OS 5 (VMWare) – Betriebssystem für Kubernetes
- Cloud-Init, Ansible, Helm (CI/CD)
- Kubernetes 1.27 (verfügbar 1.27/28/29) (1 Master- und 1-n Worker-Nodes)
- Cert-Manager
- Dashboard (Kubernetes) (K8S-Administration und -Monitoring)
- Ingress-Controller (HAProxy)
- Load Balancer (HAProxy)
- Storage (NFS-Provisioner)
- Portainer (K8S- Administration und -Monitoring)
- Private Registry (Harbor) (Docker-Compose-Service auf einer separaten Photon OS 5 VM).

Bestandteile der Plattformlösung (Archive Digital as a Service)

- Ansible, Helm (CI/CD)
- Landingpage (NGINX)
- IAM-Komponente (Keycloak)
- Mail (für Tests während der Softwareentwicklung) (Roundcube)
- XML-Datenbank inkl. REST-Services (BaseX - XQuery)
- REST-Services (Jetty)
- Workflow-Komponente (Camunda 7)
- Integrierende Benutzeroberfläche für Entwicklung und den Betrieb (APEX)
- DIMAG-KM-Konfiguration (Software muss bei DIMAG bezogen werden)

Bestandteile der Fachanwendung X-BA (xdomea-basierte Archivierung)

- Ansible, Helm (CI/CD)
- REST-Services (XQuery- und Jetty- basiert)
- WebDAV-Services
- APEX-Applikation
- Keycloak-Konfiguration

Leitlinien der Softwareentwicklung

Die von mir in 2021 zitierten Leitlinien für die Softwareentwicklung wurden eingehalten: „Nutze Technologien mit offenen Standards zum Bau moderner Web-Anwendungen. Use open-standards technologies to build modern web apps.“[2] „Baue für die Änderung und nicht für die Ewigkeit. Build to Change Instead of Building to Last.“[3]

Veröffentlichung

Im Oktober 2024 wurde die Plattformlösung Interessierten vorgestellt und im März 2025 als Open Source Produkt bereitgestellt.

[2] https://developer.ibm.com/technologies/web-development/ [nicht mehr gültig]
[3] https://www.tutorialspoint.com/software_architecture_design/key_principles.htm

Prüfkatalog für strukturierte Unterlagen

Elia Peng

Einleitung

Im Zuge der zunehmenden Digitalisierung sieht sich die öffentliche Verwaltung mit der Herausforderung konfrontiert, Unterlagen effizient, rechtskonform und langfristig zu verwalten. In der Stadtverwaltung Zürich wurde zur Bewältigung dieser Aufgabe eine umfassende Records Management Policy (RM-Policy) eingeführt. Diese Policy schafft den rechtlichen und organisatorischen Rahmen für eine rechtskonforme, sichere und effiziente Verwaltung städtischer Unterlagen, sowohl in digitaler als auch in physischer Form. Ein zentrales Instrument zur Umsetzung der Policy ist dabei der Prüfkatalog für strukturierte Unterlagen, der entwickelt wurde, um die Konformität der in der Stadtverwaltung Zürich eingesetzten Fachapplikationen mit der städtischen RM-Policy zu überprüfen.

Ausgangslage und Ziel

Die Grundlage für den Prüfkatalog bildet ein Stadtratsbeschluss aus dem Jahr 2015 (STRB Nr. 670 vom 10. Juli 2015), durch welchen die RM-Policy eingeführt wurde. Die RM-Policy dient der Sicherstellung, dass alle städtischen Unterlagen, unabhängig von ihrer Form (digital oder physisch), den gesetzlichen Anforderungen entsprechen und effizient organisiert und verwaltet werden. Aufgrund des anfänglich größeren Bedarfs, den Umgang mit unstrukturierten Unterlagen zu verbessern, wurde für die Umsetzung ein zweistufiger Ansatz gewählt. Die erste Phase, die bis Ende 2022 lief, konzentrierte sich auf unstrukturierte Unterlagen und in der zweiten Phase, die bis Ende 2025 abgeschlossen werden soll, wird der Anwendungsbereich der RM-Policy auch auf strukturierte Unterlagen ausgeweitet. Die strukturierten Unterlagen umfassen dabei Daten und Informationen, die in Fachapplikationen beziehungsweise datenbankgestützten Systemen verwaltet werden, wie zum Beispiel reine Datenbankanwendungen, Fallführungssysteme oder auch Geoinformationssysteme.

Anforderungen der Records Management Policy

Die städtische RM-Policy basiert auf international anerkannte Normen wie ISO 15489, ISO 16175, ISO 30300, ISO 30301 sowie ISO 30302. Die Anwendung dieser Normen stellt sicher, dass die Stadt Zürich ein systematisches Vorgehen zur Verwaltung aller geschäftsrelevanten Unterlagen verfolgt. Die in der ISO-Norm 15489 definierten Grundsätze der

Authentizität, Zuverlässigkeit, Integrität und Benutzbarkeit der Unterlagen sind somit auch in der RM-Policy verankert. Dadurch soll sichergestellt werden, dass sowohl strukturierte als auch unstrukturierte Unterlagen echt, glaubwürdig, vollständig, zugänglich und vor unberechtigtem Zugriff geschützt sind.

Um die in der RM-Policy definierten Grundsätze effektiv umzusetzen, müssen die Organisationseinheiten der Stadt Zürich eine Reihe von Instrumenten einsetzen. Diese Instrumente unterstützen die Einhaltung der Records Management-Grundsätze und stellen eine konsistente Anwendung der RM-Policy über alle Organisationseinheiten sicher. Dazu gehören ein strukturiertes Ordnungssystem zur Kategorisierung und Verwaltung der Unterlagen, interne Organisationsvorschriften, die den Aufbau und Ablauf des Records Managements regeln, und Qualitätsmanagementmaßnahmen zur Sicherstellung und kontinuierlichen Verbesserung der Qualität des Records Managements. Die Integration dieser Instrumente in die täglichen Arbeitsprozesse ist eine Kernforderung der RM-Policy. Daher forderte die verbindliche Policy die Organisationseinheiten auf, die genannten Instrumente bis Ende 2022 auf unstrukturierte Unterlagen anzuwenden. Bis Ende 2025 soll nun das Ordnungssystem auch die Aufgaben der Fachapplikationen dokumentieren, während die Organisationsvorschriften den Umgang mit strukturierten Unterlagen explizit festhalten und das Qualitätsmanagement die Handhabung dieser Unterlagen mitberücksichtigen sollen.

Eine weitere wesentliche Anforderung der RM-Policy ist die Steuerung der Lebensphasen. Die Lebensphasensteuerung soll, soweit sinnvoll und anwendbar, im Rahmen der Umsetzung der zweiten Phase der städtischen RM-Policy auch in Fachapplikationen ermöglicht werden. Dieses Prinzip ist wichtig für die Durchführung und die Nachvollziehbarkeit von Amts- oder Geschäftshandlungen.

Anwendung der Records Management Policy auf strukturierte Unterlagen

Die städtische RM-Policy schafft somit eine einheitliche Grundlage, um sicherzustellen, dass alle in der Stadtverwaltung Zürich anfallenden Unterlagen korrekt verwaltet werden. Der Prüfkatalog für strukturierte Unterlagen ist dabei das zentrale Werkzeug, um die Einhaltung dieser Policy auch bei Fachapplikationen sicherzustellen beziehungsweise um die Überprüfung der Systeme hinsichtlich ihrer Konformität mit der RM-Policy zu ermöglichen.

Der Prüfkatalog soll die Organisationseinheiten bei der Einhaltung der RM-Policy unterstützen und das Risiko von Datenverlusten und Datenschutzverletzungen minimieren. Die Sicherstellung der Compliance mit rechtlichen und regulatorischen Anforderungen war daher ebenfalls ein integraler Bestandteil für die Motivation zur Entwicklung des Prüfkatlogs. Historisch

bedingte heterogene IT-Landschaften und unterschiedliche Vorgehensweisen bei der Beschaffung von Fachapplikationen vor der Einführung des ISDS-Prozesses (Informationssicherheit und Datenschutz) haben nämlich zu Inkonsistenzen geführt, welche die Einhaltung der RM-Policy erschweren und das Risiko von Datenverlusten und Sicherheitsverletzungen erhöhen. Der Prüfkatalog soll diese Inkonsistenzen minimieren und sicherstellen, dass alle Fachapplikationen der Stadtverwaltung Zürich ein Minimalset an festgelegten Records Management-Anforderungen erfüllen und somit auch der städtischen RM-Policy entsprechen.

Die Methodik zur Entwicklung des Prüfkatalogs umfasste mehrere Schritte. Basierend auf den Anforderungen der RM-Policy und der Analyse der bestehenden Prozesse, wie beispielsweise des ISDS-Prozesses, wurde ein umfassender Prüfkatalog entwickelt. Ein wesentlicher Bestandteil der Methodik war zudem die Einbeziehung von Feedback verschiedener Stakeholder. Diese interdisziplinäre Zusammenarbeit ermöglichte es, unterschiedliche Perspektiven zu integrieren und ein umfassendes Verständnis des Records Managements zu gewinnen. Pilotprojekte spielten ebenfalls eine entscheidende Rolle bei der Methodik. Der Prüfkatalog wurde an ausgewählten Fachapplikationen einzelner Organisationseinheiten getestet, um seine Anwendbarkeit in der Praxis zu überprüfen. Er wurde so konzipiert, dass er möglichst allgemeingültige Prüfpunkte verwendet, um dem breiten Spektrum an unterschiedlichen Organisationseinheiten gerecht zu werden. Die Ergebnisse dieser Pilotprojekte lieferten wertvolle Erkenntnisse über die Stärken und Schwächen des Prüfkatalogs und ermöglichten notwendige Anpassungen und Verbesserungen, die den realen Anforderungen der Stadtverwaltung Zürich entsprechen.

Um der zweiten Phase der städtischen RM-Policy zu entsprechen, sollen die Organisationseinheiten den Prüfkatalog in Form eines Self-Assessments auf ihre im Bereich Kerngeschäft eingesetzten Fachapplikationen einsetzen. Die Selbstüberprüfung beruht dabei einerseits auf der Tatsache, dass die Organisationseinheiten selbst für die Records Management-Konformität ihrer Fachapplikationen verantwortlich sind, und anderseits darauf, dass eine große Anzahl von Fachapplikationen im Einsatz ist. Die Überprüfung anhand des Prüfkatalogs konzentriert sich daher auch nur auf jene Fachapplikationen, die für die Abwicklung der Kerngeschäfte zentral sind. Diese Auswahl stellt sicher, dass die Ressourcen auf Fachapplikationen fokussiert werden, deren Konformität mit der RM-Policy den größten Einfluss auf die operationelle Integrität und rechtliche Compliance der Stadtverwaltung Zürich hat. Darüber hinaus soll der Prüfkatalog auch bei der Einführung neuer oder bei der Ablösung alter Fachapplikationen angewendet werden. Die Prüfpunkte sollen in dieser Hinsicht gewährleisten, dass die Fachapplikationen von Beginn an den Anforderungen der städtischen RM-Policy entsprechen.

Um die effektive Anwendung des Prüfkatalogs im Self-Assessment sicherzustellen, werden Workshops angeboten, in denen der Prüfkatalog im Detail vorgestellt und allfällige Fragen geklärt werden. Ziel dieser Workshops ist es, den Organisationseinheiten der Stadt Zürich die notwendigen Kenntnisse zu vermitteln, um den Prüfkatalog selbstständig anzuwenden und die Ergebnisse dem Stadtarchiv zurückzumelden. Sollten die städtischen Organisationseinheiten dabei potenzielle Defizite erkennen, müssen sie diese im Rahmen der Umsetzung der zweiten Phase der RM-Policy beheben.

Prüfkatalog für strukturierte Unterlagen

Die Aufgabe der Organisationseinheiten der Stadt Zürich besteht nun darin, für jede ihrer Fachapplikationen, die sie im Bereich Kerngeschäft einsetzen, die einzelnen Prüfpunkte des Prüfkatalogs zu beantworten sowie bei einem allfällig erkannten Defizit den Handlungsbedarf anzugeben. Dabei folgt der Prüfkatalog einem Ampelsystem, bei welchem die Organisationseinheiten angehalten sind, die Erfüllung der einzelnen Prüfpunkte entsprechend farblich zu markieren. Grün bedeutet, dass die Anforderungen vollständig erfüllt sind, orange weist auf potenzielle Probleme hin, die weiterer Überprüfung und gegebenenfalls Maßnahmen bedürfen, und rot signalisiert, dass die Anforderungen nicht erfüllt werden und Handlungsbedarf bestehen könnte. Diese farbliche Markierung ermöglicht eine schnelle Identifikation und erleichtert es den Organisationseinheiten, gezielte Maßnahmen zur Verbesserung einzuleiten. Die ausgefüllten Prüfkataloge sind in der Folge an das Kompetenzzentrum Records Management, das beim Stadtarchiv Zürich angesiedelt ist, zurückzusenden, welches die Ergebnisse aus dem Self-Assessment überprüft. Die städtischen Organisationseinheiten sind abschließend noch in der Pflicht, die potenziell identifizierten Mängel unter Berücksichtigung wirtschaftlicher Aspekte sowie einer fundierten Risikobeurteilung zu beheben.

Der Prüfkatalog deckt insgesamt sechs zentrale Prüfbereiche ab, die allesamt wesentliche Aspekte der Verwaltung von strukturierten Unterlagen betreffen. Die untergeordneten Prüfpunkte wurden so gestaltet, dass sie sowohl die technischen als auch die organisatorischen Anforderungen der städtischen RM-Policy abdecken und sicherstellen, dass die Fachapplikationen die langfristige Rechtskonformität und Datenintegrität gewährleisten. Im Folgenden werden die übergeordneten Prüfbereiche und deren spezifische Prüfpunkte detailliert vorgestellt.

Schutz vor Einsichtnahme durch Unberechtigte

Der erste Prüfbereich des Prüfkatalogs konzentriert sich auf den Schutz sensibler Daten vor unbefugtem Zugriff. Fachapplikationen, die personenbezogene oder (besonders) schützenswerte Informationen speichern, müssen sicherstellen, dass nur autorisierte Personen auf diese

Daten zugreifen können. Der Prüfkatalog stellt hier spezifische Fragen zu den technischen und organisatorischen Maßnahmen, die den Zugriff auf die Daten einschränken. Besonders ältere Fachapplikationen, die vor der Einführung des ISDS-Prozesses entwickelt wurden, erfordern diesbezüglich besondere Aufmerksamkeit. Diese Systeme müssen überprüft und gegebenenfalls aktualisiert werden, um den modernen Anforderungen der Informationssicherheit und des Datenschutzes zu genügen.

Authentizität, Zuverlässigkeit und Integrität

Ein zentraler Aspekt der RM-Policy ist die Sicherstellung der Authentizität, Zuverlässigkeit und Integrität der gespeicherten Daten in den Fachapplikationen. Diese drei Kriterien sind entscheidend für die Nachvollziehbarkeit und Rechtssicherheit der strukturierten Unterlagen. Der Prüfkatalog fragt in diesem Zusammenhang nach der Fähigkeit der Fachapplikation, alle Änderungen an den Daten zu protokollieren, nach der Möglichkeit, die genaue Herkunft der Daten (wer, wann, was) zu identifizieren und nach Mechanismen, die sicherstellen, dass Daten nicht manipuliert oder verfälscht werden können. Die vollständige und transparente Nachvollziehbarkeit aller Änderungen ist dabei unerlässlich, um sicherzustellen, dass die Daten ihre Beweiskraft behalten und als zuverlässig gelten.

Benutzbarkeit

Ein weiterer wichtiger Punkt des Prüfkatalogs ist die langfristige Benutzbarkeit der in den Fachapplikationen gespeicherten strukturierten und unstrukturierten Unterlagen. Hierbei geht es um die Frage, ob die Unterlagen auch nach vielen Jahren noch lesbar und nutzbar sind. Daher untersucht der Prüfkatalog, ob die Fachapplikationen Maßnahmen zur Sicherstellung der Lesbarkeit von Datenformaten ergreifen, ob sie potenzielle Risiken durch veraltete Dateiformate erkennen und ob es auf technischer oder organisatorischer Ebene Migrationsstrategien gibt für den Fall, dass Dateiformate veralten. Dieser Prüfbereich betrifft insbesondere Fallführungssysteme, in denen zu den einzelnen Fällen auch zusätzliche Dokumente abgelegt werden. Es ist nämlich unerlässlich, dass die Unterlagen auch langfristig noch interpretiert und genutzt werden können, insbesondere im Hinblick auf die Archivierung der Unterlagen.

Steuerung der Lebensphasen

Ein zentraler Bestandteil des Records Managements und der städtischen RM-Policy ist die Steuerung der Lebensphasen von Unterlagen. Auf Fachapplikationen bezogen bedeutet das, dass diese in der Lage sein müssen, die verschiedenen Lebenszyklen der Unterlagen zu verwalten – von der Erstellung über die Nutzung bis hin zur Archivierung oder Löschung der Unterlagen. Der Prüfkatalog überprüft daher, ob Fachapplikationen die Möglichkeit bieten, Dossiers oder

Datensätze abzuschließen, beispielsweise durch eine Statusänderung, um nachträgliche Änderungen zu verhindern und die Integrität der Unterlagen zu gewährleisten. Es muss also sichergestellt werden können, dass abgeschlossene Datensätze oder Dossiers nicht mehr verändert werden können. Insbesondere bei Fachapplikationen, die langfristige Geschäftsprozesse unterstützen, besteht häufig das Problem, dass Datensätze nicht ordnungsgemäß abgeschlossen oder gelöscht werden können. Daher sollen Fachapplikationen auch die Möglichkeit bieten, Aufbewahrungsfristen zu hinterlegen, um die korrekte Archivierung oder Löschung von Unterlagen zu gewährleisten. Idealerweise wird hierfür eine Ablieferungsschnittstelle nach dem Standard eCH-0160 verwendet. Falls dies nicht möglich ist, müssen für eine allfällige Archivierung von archivwürdigen Unterlagen alternative Schnittstellen oder Verfahren definiert werden, um die Archivierung sicherzustellen.
Hinsichtlich der Steuerung der Lebensphasen sind jedoch Einzelfallprüfungen notwendig, um festzustellen, ob sie in den jeweiligen Fachapplikationen technisch und fachlich überhaupt möglich ist. Die städtische RM-Policy soll nämlich nur dort angewendet werden, wo dies sinnvoll und anwendbar ist. Einige Fachapplikationen, wie beispielsweise Geoinformationssysteme, überschreiben Daten kontinuierlich und unterstützen daher keine herkömmliche Lebensphasensteuerung.

Interne Regelungen im Umgang mit der Fachapplikation
Der fünfte Prüfbereich des Prüfkatalogs befasst sich mit den internen Regelungen der städtischen Organisationseinheiten im Umgang mit den Fachapplikationen. Die Organisationseinheiten müssen klare Vorgaben dafür haben, wie die in den Fachapplikationen verwalteten Daten angelegt, bearbeitet und gelöscht werden. Daher wird mittels des Prüfkatalogs erfragt, ob es überhaupt interne Richtlinien für den Umgang mit den Fachapplikationen gibt. Dies soll gewährleisten, dass die Verwaltung der Unterlagen systematisch und konsistent erfolgt.

Qualitätsmanagement zum Umgang mit der Fachanwendung
Der letzte Prüfbereich betrifft schließlich das Qualitätsmanagement, das ein wesentlicher Bestandteil des Records Managements ist und sicherstellen soll, dass auch die strukturierten Unterlagen jederzeit den Anforderungen der städtischen RM-Policy entsprechen. Daher muss jede Fachapplikation einem kontinuierlichen Qualitätsmanagement unterliegen, welches sicherstellen soll, dass die in der Fachapplikation verwalteten Unterlagen vollständig, aktuell und korrekt sind. Der Prüfkatalog erfragt daher, ob eine regelmäßige Überprüfung der Datenqualität stattfindet. Übergeordnetes Ziel ist hierbei die Gewährleistung einer hohen Datenqualität.

Bisherige Erfahrungen

Die ersten Erfahrungen mit der Anwendung des Prüfkatalogs in den Organisationseinheiten der Stadtverwaltung Zürich waren überwiegend positiv. Die städtischen Organisationseinheiten schätzen die klare Struktur des Katalogs und die einfache Möglichkeit, potenzielle Defizite in ihren Fachapplikationen eigenständig zu identifizieren und zu beheben. Dabei hat sich das Self-Assessment-Verfahren als äußerst wirkungsvolles Instrument erwiesen, um eine systematische Überprüfung derjenigen Fachapplikationen zu ermöglichen, deren Konformität mit der städtischen RM-Policy den größten Einfluss auf die Rechtssicherheit der Stadtverwaltung Zürich haben.

Trotz der positiven Rückmeldungen gibt es aber auch einige Herausforderungen, insbesondere im Bereich der Löschung oder Archivierung der Unterlagen. Viele Fachapplikationen bieten nämlich keine ausreichenden Mechanismen zur Löschung von nicht archivwürdigen Unterlagen nach Ablauf der Aufbewahrungsfristen, was insbesondere im Hinblick auf Datenschutzvorgaben problematisch ist. Ein weiteres Problem betrifft die Archivierung der Unterlagen. Viele Fachapplikationen verfügen nicht über standardisierte Schnittstellen zur Ablieferung von archivwürdigen Unterlagen ans Stadtarchiv, was individuelle Lösungen erfordern wird.

Fazit

Der Prüfkatalog für strukturierte Unterlagen hat sich als effektives Instrument für die Sicherstellung der Records Management-Konformität von Fachapplikationen in der Stadtverwaltung Zürich etabliert. Er bietet eine strukturierte und systematische Vorgehensweise, um potenzielle Defizite in Fachapplikationen zu identifizieren und notwendige Verbesserungen einzuleiten. Trotz bestehender Herausforderungen, insbesondere bei der Löschung nicht archivwürdiger und der Archivierung archivwürdiger Unterlagen nach Ablauf der Aufbewahrungsfrist, hat der Prüfkatalog seine Effektivität bereits unter Beweis gestellt und leistet einen wesentlichen Beitrag zur Sicherstellung der Rechtskonformität in der Stadtverwaltung Zürich.

Archivierung aus der Cloudplattform einer Landesgesundheitsbehörde: Zusammenwirken von archivfachlicher und informationstechnischer Seite

Bernhard Homa und Isabell Schönecker

Bei der Digitalen Archivierung verändern sich die klassischen Vorgehensweisen aus der analogen Welt spürbar, wie anhand eines Fallbeispiels aus dem Niedersächsischen Landesarchiv demonstriert werden soll. Dies betrifft im Wesentlichen drei Themenbereiche:

1) Das erforderliche Zusammenwirken von archivfachlicher und informationstechnischer Seite im Archiv für eine erfolgreiche Übernahme von elektronischem Archivgut.
2) Die in gleicher Weise erforderliche Zusammenarbeit mit der anbietenden beziehungsweise abgebenden Stelle, um überhaupt elektronisches Archivgut identifizieren, bilden und übernehmen zu können.
3) Die Veränderung klassischer Termini und Konzepte der Archivwissenschaft, wie sie bisher für die Überlieferungsbildung aus analogem Registraturgut üblich waren.

Die archivfachliche Seite

Im Niedersächsischen Landesarchiv (NLA) ist die Zuständigkeit für die Bewertung von Registraturgut ganz typisch nach Ressorts und regionalem Sprengel (Behördensitz) aufgeteilt. Zugleich existiert seit einigen Jahren für die Digitale Archivierung ein eigenes Team, welches sich um die spezifischen Details der Bildung elektronischen Archivgutes kümmert (zur Organisation des Landesarchivs s. Niedersächsisches Landesarchiv, 2024). Der prinzipielle Ansatz des NLA besteht darin, dass bei der Anbietung elektronischer Unterlagen immer mindestens zwei Personen beteiligt sind, nämlich einmal eine archivfachliche Kraft aus der für die jeweilige Provenienzstelle zuständigen Abteilung und eine aus dem Team DIMAG.

Vorab noch kurz einige Worte zu der Behörde beziehungsweise Provenienzstelle, die den folgenden Ausführungen zugrunde liegt. Das Niedersächsische Landesgesundheitsamt (NLGA) ist eine seit 1995 bestehende obere Landesbehörde im Ressort des Niedersächsischen Sozialministeriums (zur Organisation s. Niedersächsisches Landesgesundheitsamt, 2024). Ihre Schwerpunkte liegen in den drei Bereichen

1) Infektionsmedizin und Krankenhaushygiene;
2) Umweltmedizin;

3) Öffentliches Gesundheitswesen (zum Beispiel Gesundheitsberichterstattung zu Schulkindern oder Krebserkrankungen).

Die Tätigkeit beinhaltet ganz überwiegend wissenschaftliche Untersuchungen und die daraus folgende Beratung beziehungsweise Informationsweitergabe an die Landespolitik, die kommunalen Gesundheitsämter, Einrichtungen des Gesundheitswesens sowie die interessierte Öffentlichkeit – im Ansatz ganz ähnlich dem Robert-Koch-Institut (RKI) auf Bundesebene. Es handelt sich beim NLGA also nicht um eine Behörde der Eingriffsverwaltung und auch nicht der Leistungsverwaltung, zumindest soweit damit einzelfallbezogene Vorgänge gemeint sind (zu den unterschiedlichen Verwaltungstypen s. Krumme, 2018).

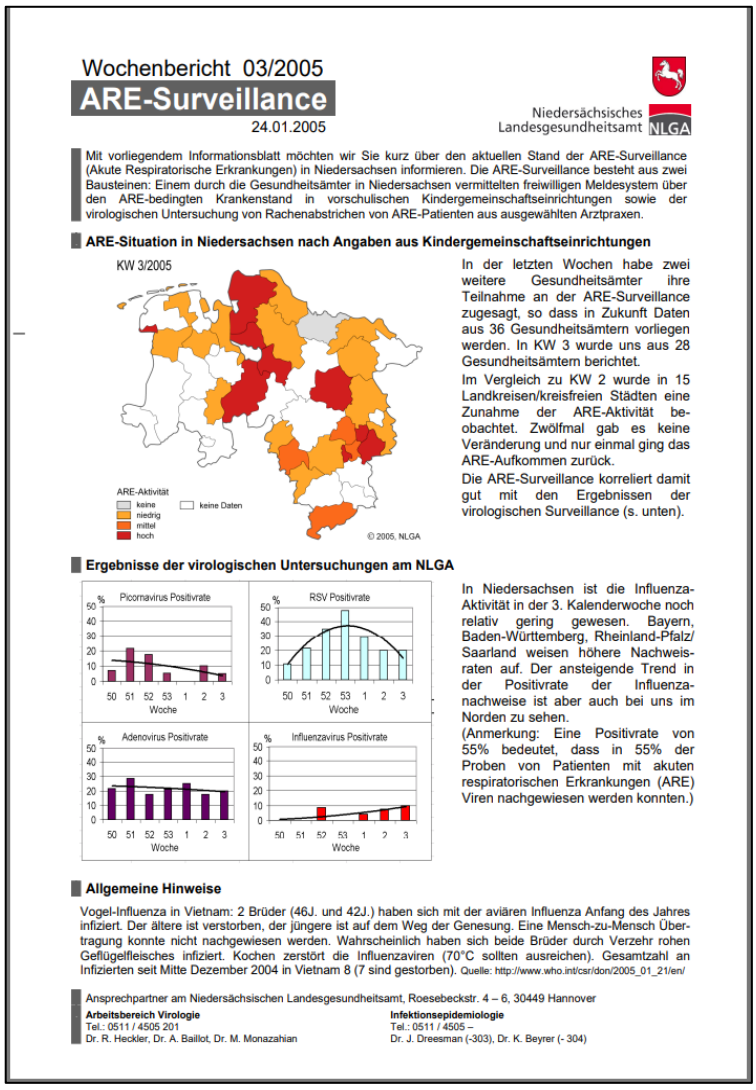
Übernahmen aus dem NLGA hat es bis ins Jahr 2023, trotz vereinzelter Kontaktaufnahmen seit den 2000er Jahren, nicht gegeben, die Kommunikation war längere Zeit unterbrochen. In den Jahren 2021/2022 konnten dann aber über die Behördenleitung – im NLGA hatte es zuvor einen Führungswechsel gegeben – endlich aufklärende Vorgespräche mit allen NLGA-Abteilungsleitungen und in der Folge auch Anbietungen in die Wege geleitet werden. Nebenbei bemerkt, zeigte sich hier ein typisches Muster: hat man mal die Leitung der anbietenden Stelle für das eigene Anliegen sensibilisiert, wird das Leben bei der analogen wie elektronischen Archivierung erheblich leichter.

Diese Vorgespräche mit der abgebenden Stelle erwiesen sich als eminent wichtig. Zunächst einmal ergab sich daraus, dass in den Fachabteilungen so gut wie keine Aktenüberlieferung vorhanden ist. Aufgrund der Gutachten- und Beratungsfunktion der Behörde dominieren vielmehr wissenschaftliche Publikationen, Infomaterialien oder teilweise, wie etwa bei der Beratung der Gesundheitsämter, bloße Sachbearbeiterablagen in E-Mail-Konten. Selbst wo noch „Akten“ im klassisch-analogen Sinne geführt werden, handelt es sich eher um Informationssammlungen und nicht um Akten zur Steuerung von Geschäftsprozessen, denn die Federführung für den jeweils beratungsbedürftigen Sachverhalt liegt bei der das NLGA um Hilfe bittenden Stelle (zur generellen Problematik der Aktenführung im elektronischen Zeitalter s. z.B. Ernst, 2017).

Da nun Beratung und Begutachtung zu den Kernaufgaben des NLGA gehören und es wie erwähnt sonst kaum Überlieferung gab, wurden auch bei fehlender Federführung derartige Unterlagen für grundsätzlich archivwürdig erklärt, zumal diese in komprimierter und serieller Form eben nur im NLGA zu finden sind. Die Informationsmaterialien und Gutachten sind häufig auch publiziert worden, sodass man fragen kann, ob es sich überhaupt um Archiv- und nicht eher um Bibliotheksgut handelt: die diesbezüglich vom NLGA bereitgehaltenen Unterlagen waren und sind eben keine unikalen Originale im streng archivwissenschaftlichen Sinne. An-

dererseits besitzt nur weniges davon eine ISBN (Bibliothekscode), vielfach handelt es sich um eine Art „graue Behördenliteratur“, deren Inhalte bewahrungswürdig sind, die aber ohne Archivierung seitens des NLA kaum erhalten bleiben dürften. Strenge archivwissenschaftliche Dogmatik hätte hier also nicht weitergeführt.

Nun mögen die bisher genannten Punkte vor allem dem Charakter der Behörde geschuldet sein und weniger mit der Archivierung digitaler Unterlagen zusammenhängen, doch verschärft letztere die schon angesprochenen Problemstellungen. Das NLGA hat nämlich von Anfang an deutlich gemacht, dass ein Großteil seiner Unterlagen und Publikationen nur noch rein digital vorhanden ist, und dies seit mindestens zwanzig Jahren – ein typisches Beispiel sind etwa die Monatsberichte zu Atemwegserkrankungen (s. Abb. 1). Maßgebliches Medium hierfür ist, neben der Webseite, die im Aufsatztitel angesprochene Cloudplattform „ÖGD-Intern“, Sogenannte „Fachverfahren“ sind zwar auch vorhanden, deren Inhalte wurden aber vom Landesarchiv überwiegend als nicht archivwürdig eingeschätzt.


Wochenbericht 03/2005

ARE-Surveillance

24.01.2005

Niedersächsisches Landesgesundheitsamt NLGA

Mit vorliegendem Informationsblatt möchten wir Sie kurz über den aktuellen Stand der ARE-Surveillance (Akute Respiratorische Erkrankungen) in Niedersachsen informieren. Die ARE-Surveillance besteht aus zwei Bausteinen: Einem durch die Gesundheitsämter in Niedersachsen vermittelten freiwilligen Meldesystem über den ARE-bedingten Krankenstand in vorschulischen Kindergemeinschaftseinrichtungen sowie der virologischen Untersuchung von Rachenabstrichen von ARE-Patienten aus ausgewählten Arztpraxen.

ARE-Situation in Niedersachsen nach Angaben aus Kindergemeinschaftseinrichtungen

KW 3/2005

ARE-Aktivität: keine, niedrig, mittel, hoch; keine Daten

© 2005, NLGA

In der letzten Wochen habe zwei weitere Gesundheitsämter ihre Teilnahme an der ARE-Surveillance zugesagt, so dass in Zukunft Daten aus 36 Gesundheitsämtern vorliegen werden. In KW 3 wurde uns aus 28 Gesundheitsämtern berichtet.

Im Vergleich zu KW 2 wurde in 15 Landkreisen/kreisfreien Städten eine Zunahme der ARE-Aktivität beobachtet. Zwölfmal gab es keine Veränderung und nur einmal ging das ARE-Aufkommen zurück.

Die ARE-Surveillance korreliert damit gut mit den Ergebnissen der virologischen Surveillance (s. unten).

Ergebnisse der virologischen Untersuchungen am NLGA

Picornavirus Positivrate; RSV Positivrate; Adenovirus Positivrate; Influenzavirus Positivrate

% 50 40 30 20 10 0

50 51 52 53 1 2 3

Woche

In Niedersachsen ist die Influenza-Aktivität in der 3. Kalenderwoche noch relativ gering gewesen. Bayern, Baden-Württemberg, Rheinland-Pfalz/Saarland weisen höhere Nachweisraten auf. Der ansteigende Trend in der Positivrate der Influenzanachweise ist aber auch bei uns im Norden zu sehen.

(Anmerkung: Eine Positivrate von 55% bedeutet, dass in 55% der Proben von Patienten mit akuten respiratorischen Erkrankungen (ARE) Viren nachgewiesen werden konnten.)

Allgemeine Hinweise

Vogel-Influenza in Vietnam: 2 Brüder (46J. und 42J.) haben sich mit der aviären Influenza Anfang des Jahres infiziert. Der ältere ist verstorben, der jüngere ist auf dem Weg der Genesung. Eine Mensch-zu-Mensch Übertragung konnte nicht nachgewiesen werden. Wahrscheinlich haben sich beide Brüder durch Verzehr rohen Geflügelfleisches infiziert. Kochen zerstört die Influenzaviren (70°C sollten ausreichen). Gesamtzahl an Infizierten seit Mitte Dezember 2004 in Vietnam 8 (7 sind gestorben). Quelle: http://www.who.int/csr/don/2005_01_21/en/

Ansprechpartner am Niedersächsischen Landesgesundheitsamt, Roesebeckstr. 4 – 6, 30449 Hannover

Arbeitsbereich Virologie
Tel.: 0511 / 4505 201
Dr. R. Heckler, Dr. A. Baillot, Dr. M. Monazahian

Infektionsepidemiologie
Tel.: 0511 / 4505 –
Dr. J. Dreesman (-303), Dr. K. Beyrer (- 304)

Abbildung 1: Monatsbericht des NLGA zu Atemwegserkrankungen (ARE-Surveillance)

ÖGD-Intern – Abkürzung für „Öffentlicher Gesundheitsdienst-Intern“ – dient entsprechend dem gesetzlichen Auftrag des NLGA der Informationsweitergabe und Beratung der Behörden und Einrichtungen des Gesundheitswesens in Niedersachsen. Es handelt sich um eine Dateisammlung, die in einer festgelegten Ordnerstruktur die entsprechenden Informations-Materialien bereitstellt (s. Abb. 2).

Die für Archive typische Eingangsfrage „Wer ist für die Plattform federführend zuständig?“ wurde vom NLGA in etwa so beantwortet: „Alle und niemand“. Tatsächlich befüllt nämlich jede Abteilung ihren fachlichen Teil

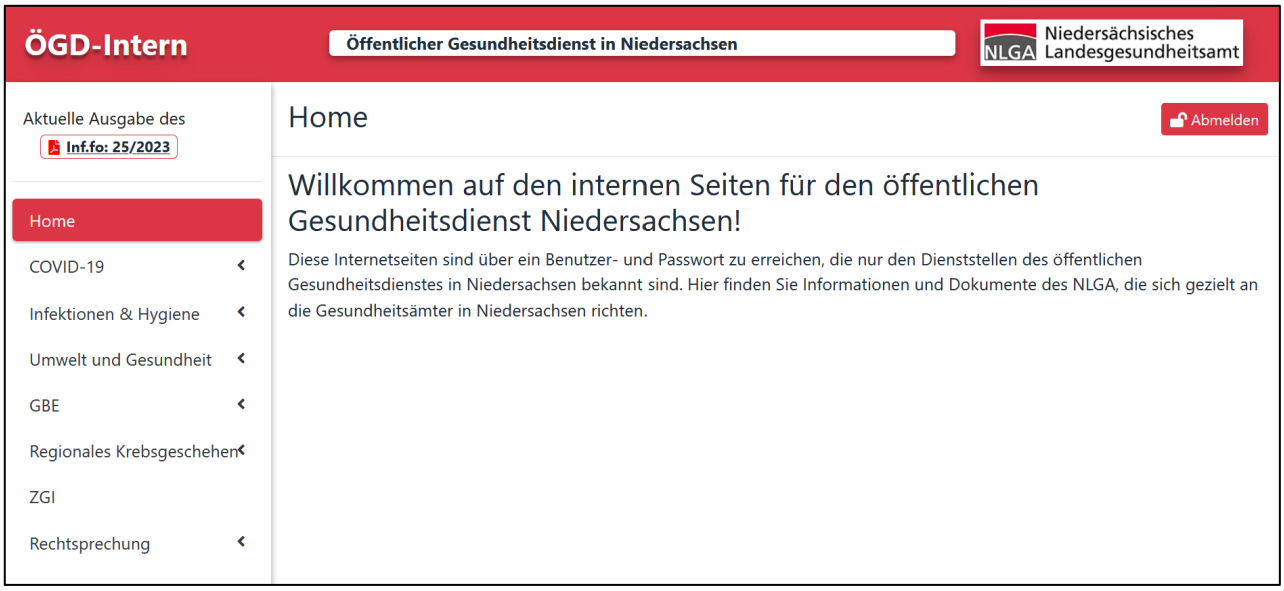


Abbildung 2: Startseite ÖGD-Intern

eigenständig – freilich in unterschiedlicher Intensität – die technische Betreuung übernimmt die IT des NLGA. Um hier überhaupt die konkreten Modalitäten der Archivierung klären zu können, wurde eine Ad-hoc-Arbeitsgruppe gebildet. Diese bestand auf Seiten des NLGA aus Personen aus der Pressestelle, dem Präsidialbüro und der IT, auf Seiten des NLA aus einem Facharchivar von der für das NLGA zuständigen Abteilung Hannover sowie Fachleuten für die digitale Archivierung vom Team DIMAG.

Diese Bildung von Ad-hoc-Arbeitsgruppen in der geschilderten Zusammensetzung ist, zumindest bei der erstmaligen Archivierung, durchaus nicht untypisch: denn ein eingespieltes Verfahren gibt es ja meist noch nicht und auf beiden Seiten – nämlich bei der abgebenden Stelle wie auch im Archiv – sind sowohl fachliche als auch informationstechnische Fragen zu klären. Jedenfalls im vorliegenden Fall hat sich diese Vorgehensweise als für den Übernahmeprozess sehr hilfreich erwiesen und erscheint auch generell empfehlenswert.

Der entsprechende Austausch erbrachte dann rasch die relevanten Informationen und Leitplanken für Bewertung und Übernahme:

1) Die Cloudplattform ÖGD-Intern enthält zwar keine sensiblen, sehr wohl aber nicht öffentlich zugängliche Daten (passwortgeschützter Zugang). Insofern war klar, dass hier im Gegensatz zu den auf der extern zugänglichen Homepage veröffentlichten Publikationen zumindest allgemeine Schutzfristen vergeben werden müssen (§ 5 Abs. 2 NArchG).
2) Die Cloudplattform bildet immer den aktuellen Stand ab, es gibt keine Versionierung und veraltete Dateien werden, wenn auch meist erst nach längerer Zeit, gelöscht. Der zentrale Newsletter „inf.fo“ ist dagegen durchgängig seit seinem erstmaligen Erscheinen im Jahr 2001 vorhanden.

3) Da es sich um eine Informationsplattform des gesamten Öffentlichen Gesundheitsdienstes in Niedersachsen handelt, finden sich dort auch Unterlagen anderer Provenienz, etwa von den Kreisgesundheitsämtern oder dem Landesverband Niedersachsens der Ärztinnen und Ärzte des öffentlichen Gesundheitsdienstes e.V. Diese Überlieferung wurde vom zuständigen Facharchivar als archivwürdig eingeschätzt und auch übernommen, da eine Verteilung oder Nachfrage bei allen betroffenen Provenienzstellen einen unverhältnismäßigen Verwaltungsaufwand bedeutet hätte. Das Landesarchiv stellt sich diesbezüglich auf den Standpunkt: wer Unterlagen in einer staatlichen Cloud zur Verfügung stellt, muss auch mit der staatlichen Archivierung rechnen.

Die informationstechnische Seite

Die Archivwürdigkeit von Unterlagen wird im NLA grundsätzlich durch die fachlich für einen Bestand zuständige Person festgelegt. Das ändert sich auch bei digitalen Daten nicht, jedoch stehen diese häufig in verschiedenen Formen – und dies meint hier nicht ausschließlich die Dateiformate – zur Verfügung. Das Team DIMAG berät daher im Rahmen von Vorgesprächen zwischen der abgebenden Stelle und der zuständigen Person unter anderem bei der Frage, in welcher Struktur die archivwürdigen Daten übernommen werden können.

Die Plattform ÖGD-Intern ist im Grunde eine zugriffsbeschränkte Website, auf der Informationen bereitgestellt werden, welche auch im Rahmen der Webarchivierung gesichert werden könnten. Da die Archivierung von Websites ganz eigene Hürden mit sich bringt und der Erhalt der Optik sowie der Benutzerführung für eine Archivierung nicht maßgeblich war, bestand hierfür keine Notwendigkeit. Davon abgesehen wurde ÖGD-Intern für einen reinen Wissenstransfer konzipiert – das relevante und archivwürdige sind also die vermittelten Informationen in Form von Dateien. Das Team empfahl daher die der Plattform zu Grunde liegende Fileablage zu übernehmen, welche mit einer gut strukturierten Dateisammlung zu vergleichen ist.

Die vorteilhafte Ausgangslage, dass die inhaltliche Strukturierung der Website auch nahezu identisch auf der Fileablage abgebildet wurde, konnte direkt für die Übergabepakete (SIP) übernommen werden. Das SIP enthielt somit zu jedem Themenbereich einen eigenen Ordner mit entsprechenden Unterordnern, je nach Strukturierung in ÖGD-Intern (s. Abb. 3).

Abbildung 3: ÖGD-Intern: Unterseite mit Informationsmaterialien und Links

Das NLGA übersandte alle Dateien aus ÖGD-Intern als ZIP-Datei über eine Landes-Cloud an das Archiv. Bei der Eingangsprüfung und Validierung wurde dann festgestellt, dass der „inf.fo“ Newsletter (Themenbereich Infektion & Hygiene) mit den dazugehörigen Anhängen in voneinander getrennten Ordnern übergeben wurde. Dadurch ergab sich bei der AIP-Bildung die Aufgabe zusammengehörige Dateien wieder zu vereinen – dies wahrt die Authentizität und gewährleistet eine bessere Nachvollziehbarkeit.

Es wurde daher das Ziel formuliert, pro Jahrgang ein Archivpaket (AIP) zu bilden, in dem jeder Bericht sowie die dazugehörigen Anhänge in einem eigenen Unterordner liegen. Im ersten Schritt stand das Problem im Raum, dass in der Fileablage die Berichte und Anhänge getrennt übergeben wurden und vor allem die älteren Dateien nicht über gemeinsame Kennzeichen (etwa Datum im Dateinamen) verfügten, um sicher zugeordnet werden zu können. Eine manuelle Zuordnung war angesichts der Menge der Dateien nicht sinnvoll. Das gesamte SIP umfasste über 3000 Dateien, wovon über die Hälfte auf den „inf.fo“ Newsletter entfielen. Es wurde daher ein Python-Skript geschrieben, welches in der HTML-Datei von ÖGD-Intern die Berichte und Anhänge identifiziert und diese dann in der übernommenen Fileablage sucht (s. Abb. 4). Wenn die betreffenden Dateien gefunden wurden, wurden diese in einen neu angelegten Ordner verschoben. Durch dieses Vorgehen konnten – als netter Nebeneffekt – auch fehlende Anhänge ermittelt werden, die zwar auf der Website eingestellt, aber nicht in der Fileablage vorhanden waren. Einige davon konnten dem Archiv nachgereicht werden und das NLGA konnte die be-

troffenen Dateien wieder online stellen. So ein nachträglicher Austausch mit der abgebenden Stelle ist nicht unüblich. Vor allem bei ersten Übernahmen können erst im Laufe der Bearbei-

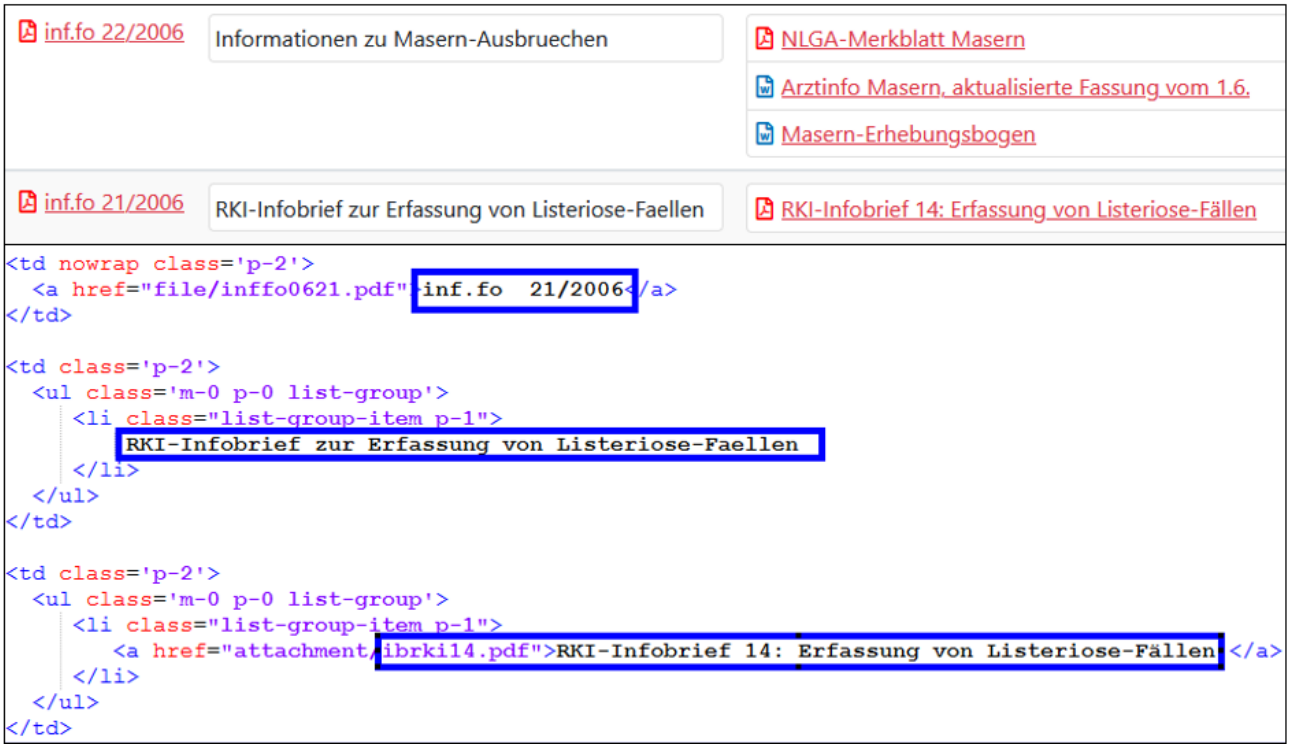


Abbildung 4: Auszug zu den in ÖGD vorliegenden Information (Website mit Inhaltsbezeichnung und Auflistung der Anhänge sowie dazugehöriger, vereinfachter HTML-Code)

tung weitere (Verständnis-)Fragen zu den Daten auftauchen oder Unregelmäßigkeiten auffallen, die dann zu klären und dokumentieren sind.

Während der Erstellung des Skripts und der intensiven Auseinandersetzung mit den Daten ist die Idee gewachsen, dass die Inhaltsbezeichnungen zu den Berichten für die Erschließung mit übernommen werden sollten. Wenn ohnehin Informationen aus der HTML-Datei extrahiert werden, können diese auch direkt nachgenutzt werden! Das Skript wurde dementsprechend um das Auslesen der HTML-Tags zum Inhalt erweitert (s. Abb. 4).

Eine rein händische Erschließung wäre beispielsweise nicht weit über den einfachen Titel „Infektionsepidemiologisches Forum (inf.fo) Jahrgang 2005“ hinausgegangen. Die Gliederung im Bestand wiederum wurde der Struktur von ÖGD-Intern nachempfunden (s. Abb. 5).

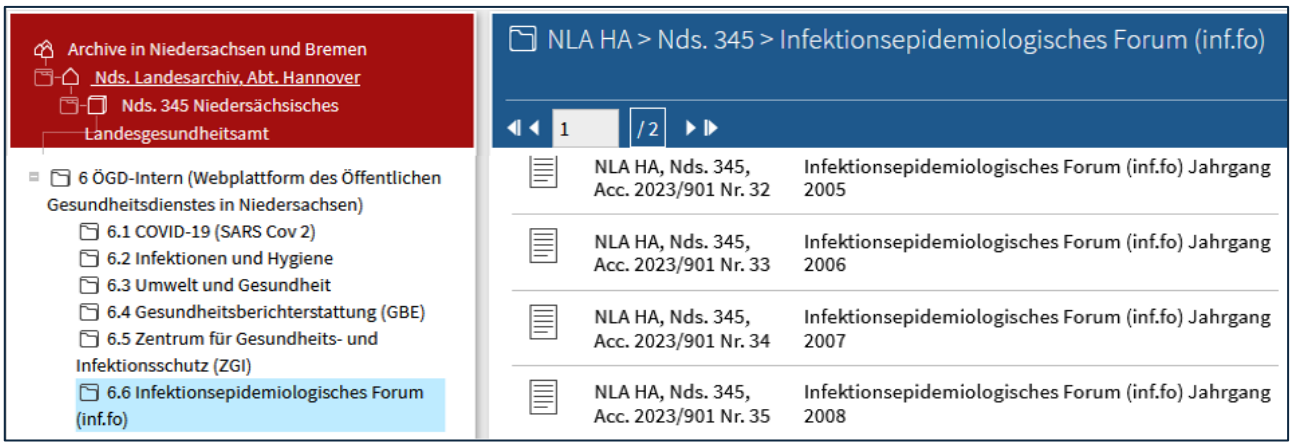


Abbildung 5: Bestand NLA HA Nds. 345 Niedersächsisches Landesgesundheitsamt mit Auszug aus der Bestands-gliederung und einfachem Titel

Durch die Inhaltsbezeichnungen konnten mehr Stichworte für die Suche generiert werden, die eine höhere Auffindbarkeit der Archivalien ermöglichen. Je nach Nutzungsszenario können so interessierte Personen bereits über die Erschließungsdaten erste Auswertungen durchführen, zum Beispiel Identifizierung von Worthäufungen wie „Norovirusausbruch“ usw. (s. Abb. 6).

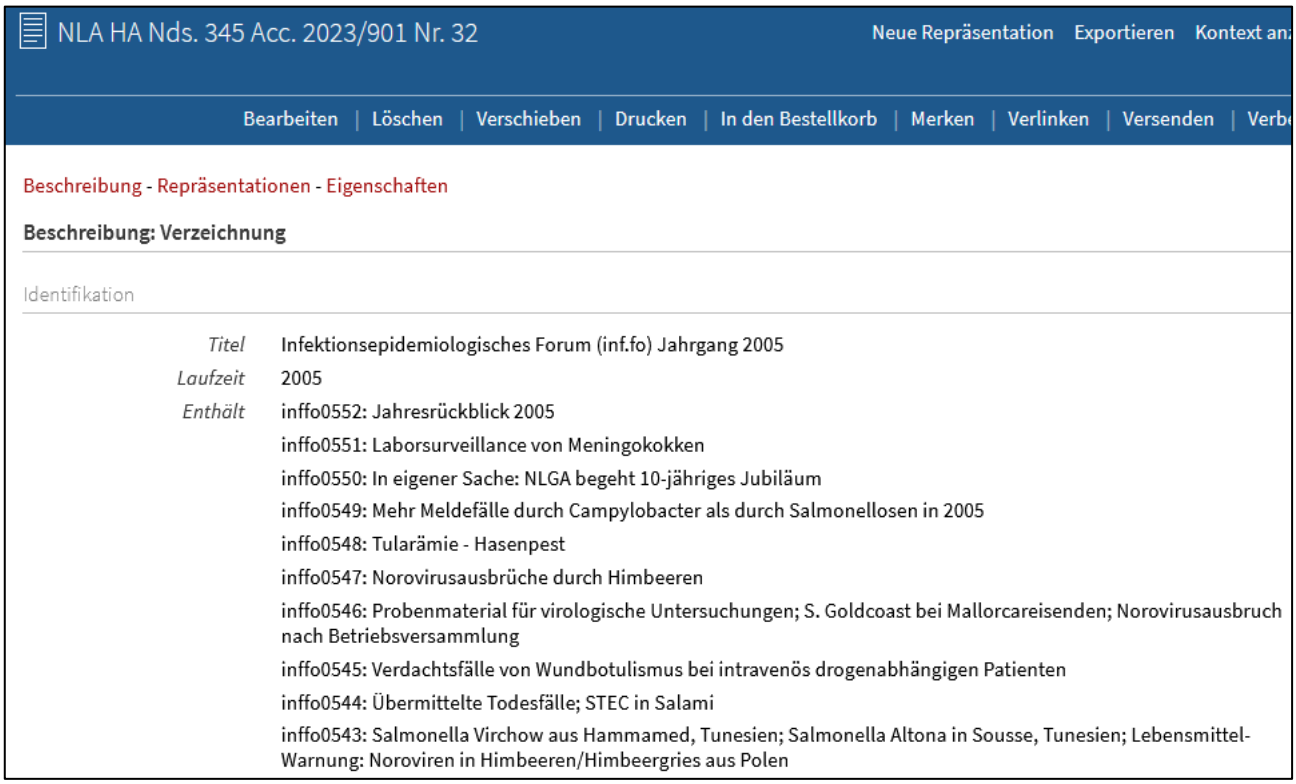


Abbildung 6: Tiefenerschließung eines Jahrgangs von inf.fo Newslettern

Konkret wurden durch das Skript die Inhaltsbezeichnungen pro Bericht für einen Jahrgang in ein Excelfeld exportiert und anschließend mit dem DIMAG IngestTool ausgelesen. Der Export wurde in dieser Weise umgesetzt, da einerseits im NLA grundsätzlich die AIP-Bildung und zugleich die Erschließung vor dem Ingest der Daten erfolgen und zum anderen digitale Übernahmen standardmäßig über eine Exceltabelle beschrieben werden. Es ist natürlich auch jedes andere Exportformat, welches zur Weiterverarbeitung geeignet ist, denkbar.
Zeitgleich kann zur AIP-Bildung und Erschließung auch eine Feinbewertung durch die fachlich zuständige Person erfolgen. Insbesondere können in diesem Schritt Dateien aus inhaltlichen Aspekten oder auch Dubletten nachkassiert werden. Die AIP-Bildung umfasst im NLA zusätzlich eine Prüfung der Dateiformate in der ersten Repräsentation eines Archivales. Wenn keine archivfähigen Formate vorliegen, werden die Dateien in den betroffenen Archivpaketen in eine zweite Repräsentation konvertiert und anschließend gekoppelt ingestiert. Dabei werden zeitgleich die Erschließungsdatensätze im Archivinformationssystem Arcinsys angelegt und die Daten im DIMAG-Kernmodul magaziniert.
Diese intensive Aufbereitung der zu archivierenden Daten hat sich nur beim „inf.fo“ Newsletter angeboten, da nur zu diesem auch zusätzliche Inhaltsinformationen auf der Website eingestellt waren. Der damit verbundene Aufwand hat sich aber schon allein dadurch gelohnt, weil der Newsletter grundsätzlich als archivwürdig eingestuft wurde und somit das Skript und das Vorgehen bei der AIP-Bildung für kommende Abgaben nachgenutzt werden kann. Das in der hier skizierten Übernahme nicht die mitunter angestrebte Vorstellung „SIP ist gleich AIP“ umgesetzt wurde, bedeutet keinen Nachteil. Die Daten im SIP waren gut strukturiert, sodass mit ihnen unkompliziert weitergearbeitet werden konnte. Auch hätte die abgebende Stelle nicht ohne weiteres die Bezüge zwischen „inf.fo“ Bericht und Anhang herstellen können.
Für künftige Übernahmen aus dem NLGA wurde eine jährliche Anbietung der im letzten Jahr neu hinzugekommenen Dateien vereinbart. Dadurch soll eine Kontinuität der Überlieferung gewahrt werden. Die vom NLA aus ÖGD-Intern übernommenen Dateien werden vor Ort nicht gelöscht.

Fazit

Dank des hier vorgestellten Vorgehens konnten 2023 erstmalig Unterlagen des NLGA sowohl in analoger als auch elektronischer Form archiviert werden. Einige wichtige Erkenntnisse seien nochmals kurz zusammengefasst:

1) Die archivisch oft ungeliebten Dateisammlungen (s. zur Thematik generell Bacia et al., 2021) können bei fehlender Aktenüberlieferung hohe Relevanz gewinnen, sowohl aus in-

haltlichen Gründen (Informationswert) als auch für die Nachvollziehbarkeit des Behördenhandelns (Evidenz).

2) Es ergeben sich dann allerdings oft Redundanzprobleme, etwa wenn Unterlagen sowohl ausgedruckt als auch digital vorhanden sind. Wenn kein einfacher Abgleich ohne Suche nach der sprichwörtlichen „Nadel im Heuhaufen“ möglich ist, sollte das im Sinne der Überlieferungssicherung hingenommen werden. Denn in der aktuellen Übergangszeit wird es längerfristig ohnehin eine parallele Übernahme sowohl analogen als auch digitalen Schriftgutes von Registraturbildnern geben, sodass Redundanzen sich nicht immer vermeiden lassen.
3) Klassische archivwissenschaftliche Begriffe und Konzepte werden im Zeitalter der Digitalen Archivierung brüchig oder erhalten neue Bedeutung. Für „Original“ ist das seit langem eine Binsenweisheit, doch gilt dies auch für anderes: „Unikales“ Archivgut gibt es in vorliegendem Fall nicht, weil die Dateien ja erst mal auch im NLGA verbleiben – sollten diese dort allerdings irgendwann gelöscht werden, mag das anders aussehen. Doch auch das zentrale Konzept der „Provenienz“ wird diffus: wie gezeigt ist die Cloudplattform nur eine Hülle, unter der sich ganz verschiedene Provenienzen, sogar innerhalb derselben Behörde, tummeln – und wo sich dann die ganz praktische Frage stellt, ob auch Fremdprovenienzen mit archiviert werden sollen, oder überhaupt dürfen.
4) Intensive Kommunikation und Kooperation mit der anbietenden Stelle sind zur Vorbereitung unerlässlich, wobei seitens anbietender Stelle wie Archiv sowohl fachliche als auch informationstechnische Kompetenzen beigezogen werden sollten. Der Anpassungsprozess und Mehraufwand darf nicht unterschätzt werden: es sind mehr Schritte und mehr Beteiligte – zumindest in der Anlaufphase – erforderlich, und auch im Regelbetrieb ist archivseitig der gewohnte simple Dreisatz „Kommen – Bewerten – Liefern lassen“ nicht mehr haltbar.
5) Umgekehrt bietet aber die Maschinenlesbarkeit von digitalem Registraturgut viele neue Möglichkeiten der Bewertung und nachgelagerten Erschließung, die zumindest teilweise den zuvor geschilderten Mehraufwand kompensieren können.

Bibliografie

Bacia. J. et al. (2021), *Bewertung schwach strukturierter Unterlagen. Mit Beiträgen aus dem Arbeitskreis „Archivische Bewertung“ im VdA – Verband deutscher Archivarinnen und Archivare e.V.* Köln: Historisches Archiv der Stadt Köln.

Ernst, K., (2017), „Welche Zukunft hat die Akte?“, in: VdA (Hrsg.), *Transformation ins Digitale. 85. Deutscher Archivtag in Karlsruhe*. Fulda: Selbstverlag des VdA, S. 67–75.

Niedersächsisches Landesarchiv (2024), *Aufgaben und Organisation des Niedersächsischen Landesarchivs*. https://nla.niedersachsen.de/startseite/landesarchiv/aufgaben-des-niedersachsischen-landesarchivs-203071.html (11.04.2024).

Niedersächsisches Landesgesundheitsamt (2024), *Niedersächsisches Landesgesundheitsamt* [Online]. https://www.nlga.niedersachsen.de/startseite (11.04.2024)
Krumme, J.-H. (2018), *Verwaltung*. https://wirtschaftslexikon.gabler.de/definition/verwaltung-47011/version-270283 (11.04.2024).

OCFL Native Archive System: Neue Technologien und Kollaboration im Digitalen Langzeitarchiv

Jürgen Enge

Einleitung

Für viele sogenannte Memo-Institutionen[1] wird die digitale Langzeitarchivierung (dLZA) zu einer immer zentraleren und unverzichtbareren Aufgabe. Nach Jahrzehnten der Digitalisierung und der Annahme von Born-digital Medien, Inhalten und Akten gilt es, die geschaffenen sowie angesammelten Werte auch dauerhaft aufzubewahren und möglichst lange authentisch zugänglich zu machen. Gleichzeitig ist der Gürtel oft eng geschnallt und es stehen nicht immer die nötigen personellen, technologischen und/oder finanziellen Ressourcen zur Verfügung.

Eine mögliche Lösung kann in Konzepten der Zusammenarbeit gefunden werden. Arbeitsteilung, Lernen von- und miteinander sowie die gemeinsame Nutzung von Infrastrukturen wirken insbesondere dann hilfreich, wenn Synergien genutzt werden und der Austausch auf Augenhöhe stattfindet. Das folgend beschriebene Beispiel der kooperativen Archivierung nutzt gocfl. gocfl[2] ist eine Open Source Software, die digitale Daten in OCFL-Kapseln verpackt, welche dann direkt an die Systeme, die vor dem Ausführen der Operationen bestimmt wurden, übergeben werden können. Damit ist gocfl die Basis für die verteilte Speicherung der OCFL-Objekte. Der Text geht auf die Aufgabenverteilung und Zuständigkeiten im Anwendungsfall ebenso ein wie auf die technischen und inhaltlichen Voraussetzungen. Bevor jedoch genauer die Umsetzung sowie Herausforderungen, Vorteile und Erfahrungen vorgestellt werden, lohnt ein Blick auf die unterschiedlichen Ausgangslagen an der Universitätsbibliothek Basel und der Zentral- und Hochschulbibliothek Luzern. Zudem werden die Prinzipien und Werte erläutert, die als zentral betrachtet werden und den Entscheid für OCFL als probates Mittel für das Feld der nachhaltigen dLZA charakterisieren. Es folgt der Umsetzungsansatz aus Basel und Luzern, bevor der Text mit einer Zusammenfassung mit Ausblick endet.

Ausgangslage

Als universitäre Hochschulbibliotheken der Schweiz weisen die UB Basel und die ZHB Luzern gewisse typologische Ähnlichkeiten auf. Allerdings variieren die konkreten Leistungsaufträge,

[1] In der Schweiz spricht man in Anlehnung an den Begriff der sogenannten „Memopolitik“ (Emanuel Amrein 2008) gern von Memo-Institutionen, wenn allgemein Einrichtungen adressiert werden, die sich mit der Aufbewahrung, Erhaltung und Vermittlung von Kulturgütern beschäftigen. Hierzu gehören i.d.R. auch Bibliotheken.
[2] https://github.com/ocfl-archive/gocfl/

die organisatorischen Strukturen, die personelle Ausstattung und die aktuelle Haltung gegenüber digitalen Archivierungsaufgaben:

Die UB Basel war in den letzten Jahren mit der Inbetriebnahme einer docuteam[3]-Infrastruktur beschäftigt, die im Februar 2024 produktiv ging. Gleichzeitig befinden sich etwas mehr als 200'000 ZIP-Kapseln mit Retro-Digitalisaten aus dem Corpus von e-Rara und e-Manuscripta auf Zwischenspeichern. Darin sind wiederum fast 3 Mio. TIFF-Dateien und ca. 700'000 Volltexte enthalten. Hinzu kommen hier nicht genauer zu beziffernde Publikationen aus der Universität, diverse Spezialsammlungen, die auch AV-Medien beherbergen, sowie weitere digitale Quellen aus unterschiedlichen UB-Projekten. Zusammengefasst sind das ca. 150 TB Rohdaten. Der „unsichtbare Elefant" im Raum ist die grundlegende Frage: Wie weiter? Beibehalten des alten Infrastrukturkonzepts, auch wenn das erst jetzt in den kommenden Monaten/Jahren produktiv wird, oder doch einen aktuelleren Ansatz wagen?

Auch in Luzern fragt man sich Grundsätzliches. Trotz mehrerer Abklärungsversuche konnte man sich in den letzten Jahren noch auf keine spezifische Archivlösung einigen. Ähnlich wie auch in anderen Einrichtungen steigt dadurch der Druck, auch wenn die digitalen Bestände im Vergleich zu Basel überschaubarere wirken. Während gewisse Budgets für externe Entwicklung lanciert werden könnten, sind die eigenen Personalressourcen im Haus überschaubar. Es gibt keine allzu nennenswerten Erfahrungen oder Knowhow zum operativen Betrieb von dLZA-Infrastrukturen. Aber die Kolleg:innen haben viel Motivation und andere IT-Kompetenzen. Daher gilt es zu entscheiden: Ausschreibung einer Archivlösung und Suche eines kommerziellen Anbieters oder Kooperation mit einem/r „Branchenpartner:in"?

Prinzipien und Werte

Jenseits von institutionellen Analogien und Differenzen ist eine Verständigung über grundlegende gemeinsame Ziele und Werte für eine gute Zusammenarbeit von Vorteil. Wie eingangs angedeutet, lassen sich diese, auch im Anwendungsbeispiel, aus den aktuellen technischen Rahmenbedingungen ableiteten.

Der vorliegende Ansatz geht von vier Annahmen aus, die als besonders wichtig für die künftige dLZA eingeschätzt werden:

- *„AIP first"*

Das Open Archival Information System (OAIS) lässt den Prozess der digitalen Verwaltung von Daten mit dem sogenannten Archival Information Package (AIP) als Kernstück erkennen.[4] Analog hierzu geht der vorliegende Ansatz davon aus, dass AIPs systemunabhängig gebildet

[3] https://www.docuteam.ch/
[4] https://de.wikipedia.org/wiki/OAIS

werden und zugänglich sein sollten. Die Funktionalität des dLZA-Systems sollte sich daran orientieren und ihre Prozesse mithin dem AIP bzw. dessen Bildung unterstellen.

- *„Distributed Resources“*

Verschiedene Stakeholder sollen am gemeinsamen System partizipieren können, ohne auf zentrale Speichersysteme zurückgreifen zu müssen. So können sie mit eigenen oder ausgelagerten Speicherinfrastrukturen teilnehmen, ohne auf ihre Daten notwendig auch über eine gemeinsame Infrastruktur zu speichern. Das *distributed-resources*-Prinzip impliziert damit auch, dass verteilte Prozesse und Ressourcen verwaltet werden können. Dass dies mit möglichst minimalistischen (Zentral-)Strukturen erfolgen sollte, ergibt sich aus dem Effizienzgebots schrumpfender finanzieller und ökologischer Mittel.

- *„Cloud native Technology“*

Das System sollte in möglichst vielen sowie den zentralen Prozessen auf aktuellen Cloud-Technologien aufsetzen. Die technologische Basis sollte dennoch keine bestimmte Infrastruktur vorschreiben. Die Archiv-Cloud sollte vielmehr sowohl public Cloud-Anbieter ermöglichen wie auch private und lokal betriebene Infrastrukturen.

- *„Zero Trust Architecture“*

Auch im Archivumfeld wirkt es heute geboten, technisch auf „Zero Trust Strategien“ umzurüsten. Die Systemumgebungen (Virtualisierung, Netzwerk, Speicher, …) müssen mithin einen sehr hohen Sicherheitsstandard erfüllen und sollten so ausgelegt sein, dass bspw. Hackerangriffe überstanden werden können.[5] Damit die Archiv-Infrastruktur hier mithalten kann, muss auch sie Zero-Trust-Prinzipien umsetzen.

Neben diesen Vorgaben ist es wichtig, das System möglichst einfach zu halten. Nach jahrelangem Wachstum der Systeme und ihrer Komplexität durch das Hinzufügen immer neuer Features ist es an der Zeit, zurück zu schlanken Systemen mit wenig Abhängigkeiten zu gelangen.[6] Fehlende Eigenschaften können aufgrund der Modularität aktueller Systeme dann gut hinzugefügt werden, wenn diese z. B. als sogenannte MACH-Architekturen[7] (s. u.) mit Microservice-Komponenten umgesetzt sind. So erzeugen die Kernsysteme weniger Komplexitäten.

Anforderungen für die Zusammenarbeit im Verbund

Übertragen auf die kooperative Zusammenarbeit zwischen der UB Basel und der ZHB Luzern lassen sich aus diesen Anforderungen folgende Verantwortlichkeiten ableiten:

[5] Das System besteht aus einfachen Komponenten, welche auch automatisiert mit Hilfe von Software Bill of Materials (SBOM) auf sicherheitsrelevante Aspekte geprüft werden können.
[6] Auch der zukünftige Standard METS 2.x (https://github.com/mets/METS-schema?tab=readme-ov-file#mets-2x) setzt auf Vereinfachungen
[7] MACH steht für M - Microservice first, A – API-basiert, C – Cloudbasiert, H – Headless. Vgl. hierzu https://en.wikipedia.org/wiki/MACH_Alliance.

- In die Eigenverantwortung der Partner fallen:
  - Verantwortlichkeit für die eigenen Bestände und die zu archivierenden Datenobjekte
  - Bereitstellung, Betrieb und Qualitätskontrolle(n) der Pre-Ingest-Workflows
  - Betrieb einer eigenen Ingest-Workbench
  - Verantwortlichkeit für die Bereitstellung und den Betrieb der Speicherinfrastruktur(en) – auch wenn diese z. B. an Dritte/Cloud-Anbieter ausgelagert wird
- Zu den kooperativen Aufgaben gehört:
  - Transfer der AIPs ins gemeinsame Archivmanagementsystem – inklusive der Verteilung auf die zuvor von den Partnern definierten Speicherstandorte
  - Aufbereitung der Daten im Sinne des Archivmanagement- und Erhaltungsmetadaten – inklusive der Bereitstellung eines mandantenfähigen Reporting-GUI[8]
  - Preservation Planning sowie Planung und Durchführung von systematischen Migrationsmaßnahmen (optional)
  - Erarbeitung und Pflege von Tools und Knowhow für die stete Weiterentwicklung entlang der konkreten Anforderungen und dem sukzessiv stattfindenden technologischen Wandel

Mechanismen, Systeme und Workflows

Den institutionellen Wünschen und Rahmenbedingungen stehen konkrete technologische Möglichkeiten gegenüber, die im Folgenden zusammengefasst werden. Zum besseren Verständnis werden die Erfordernisse entlang einer Begriffsklärung ausformuliert.

OCFL

Das Oxford Common File Layout (OCFL)[9] liefert eine standardisierte Spezifikation für das Erstellen von Archivinformationspaketen (AIP). OCFL-Container oder Kapseln sind dabei sowohl menschen- als auch maschinenlesbar. OCFL kann mehrgliedrige Objektstrukturen auf Dateiebene vollständig beschreiben. Es ist erweiterbar und unterstützt die Versionierung. Es setzt sowohl eine umfassende als auch eine integrierte Selbstdokumentation der im Archivpaket enthaltenen Dokumente operativ und gemäß der Spezifikationsstruktur um. Damit liefert OCFL eine sehr gute Grundlage, um ausgehend vom Standard (als Beschreibung) Archivsysteme und Workflows auch jenseits der bisher dominierenden Produktlogiken zu denken.

Zu den Stärken von OCFL gehören darüber hinaus:

[8] GUI = Graphical User interface.
[9] https://ocfl.io.

- *Vollständigkeit*

  Das Repository kann aus den im Repository gespeicherten Dateien (selfcontained) wiederhergestellt werden.
- *Parserfähigkeit*

  Um sicherzustellen, dass die Inhalte auch ohne die ursprüngliche Software verstanden werden können, sind sie sowohl für Menschen als auch für Maschinen lesbar.
- *Robustheit*

  Aufgrund seiner stringenten Einfachheit weist OCFL eine große Resilienz gegenüber technischen Fehlern, Korruption und Migration zwischen Speichertechnologien auf. Bei guter Umsetzung ist auch die Migration zwischen verschiedenen Archivsystemen möglich.
- *Versionierung*

  Änderungen können in die OCFL-Objekte nativ integriert werden. Zudem ist eine Deduplizierung auf Dateiebene möglich. Damit wird die Objekthistorie direkt im AIP vorgehalten.
- *Speichervielfalt*

  Inhalte können auf verschiedenen Speicherinfrastrukturen gespeichert werden. OCFL setzt nur eine Ordnerstruktur voraus, ist aber ansonsten offen für alle Arten von Speichersystemen inklusive Objektspeichern, wie sie üblicherweise in der Cloud zum Einsatz kommen.

gocfl

Um OCFL im Kontext der dLZA als Systemkomponente umzusetzen, wurde die go-basierte Software gocfl[10] implementiert. Gocfl erzeugt AIPs und liefert die Basis für die Dienste des kooperativen Verbundsystems OCFL-Native-Archive-System (ONAS). Hierzu erweitert das kommandozeilenbasierte Werkzeug OCFL-kompatible Strukturen, die entweder verzeichnisbasiert oder als ZIP-Container gespeichert werden.

Hervorzuheben sind die Erweiterungen, mit denen gocfl die klassischen Verpackungsroutinen zu einem archivfähigen Werkzeug erweitert. Dieses ermöglich die Validierung, das Reporting, die Extraktion von technischen Metadaten sowie die menschenlesbare Darstellung der im Archivpaket enthaltenen Inhalte.

Die go-Routine zur AIP-Erzeugung enthält folgende Funktionalitäten:

- *Formaterkennung*

  Die Formaterkennung extrahiert die technischen Metadaten der im Archivpaket enthaltenen Daten. Technisch setzt die Softwarebibliothek auf dem sogenannten „Indexer“ auf (Enge/Kramski 2016, 229-236), der ursprünglich für das Deutsche Literaturarchiv (DLA)

[10] https://github.com/ocfl-archive/gocfl

in Marbach implementiert wurde. Für gocfl wurde es auf den aktuellen technischen Stand gebracht. gocfl nutzt dabei unter anderem die Softwarebibliotheken Siegfried,[11] Tika,[12] ffmpeg[13] sowie Image Magick[14].

- Generierung von *METS/PREMIS*

  Der Bauplan zum Erstellen der METS/PREMIS-Beschreibung lehnt sich an die Definition der E-ARK AIPs[15] an. Sie wurden erstmals 2017 vom Digital Information LifeCycle Interoperability Standards Board (DILCIS) publiziert. Die letzte Aktualisierung vor der Textlegung erfolgte im Mai 2024.
- Generierung von grundlegenden *semantischen Metadaten*

  Sofern die die Erzeugung semantischer Metadaten im Zuge des Pre-Ingests, des Ingests oder durch die Formatstruktur automatisiert möglich und auslesbar ist, überträgt gocfl diese bei der Generierung der AIPs an das Archivmanagementsystem (s. u.).
- Metadaten des *Dateisystems*

  gocfl dokumentiert ferner die Ablagestruktur der Dateien im Dateisystem und gibt sie als ausgeschriebene Metadaten aus. Dies wurde gewünscht, da die Ablagestruktur insbesondere bei born-digitalen Vor- und Nachlässen häufig wichtige inhaltliche Information(en) enthält.
- *Umbenennung* von Pfad- und Dateinamen mit Sonderzeichen

  gocfl benennt Pfad- und Dateinamen im AIP um, wenn sie Sonderzeichen enthalten, die nicht nachhaltig sind. Diese angepassten Sonderzeichen werden vereinfacht und in UTF8-konforme Zeichenketten umgewandelt. Zudem werden die Anpassungen in der OCFL-Kapsel dokumentiert.
- Generierung von *Thumbnails*

  Um einen schnellen, visuellen Überblick über die AIP-Inhalte zu erhalten, erzeugt gocfl bei einer Reihe von definierten Dateiformaten Vorschaubilder in Größe eines Thumbnails. Unterstützt werden bisher bspw. alle gängigen Bild- und Videoformaten. Für Audiodateien werden Spektrogramme sowie für PDFs sogenannte Poster erzeugt.
- Automatische *Migration* (Prototyp)

  Beim Ingest können ferner automatische Transkodierungen von definierten, nicht nachhaltigen Dateiformaten durchgeführt werden. Diese werden dann als neue Version abgelegt und in den METS/PREMIS Daten aufgeführt.

[11] https://openpreservation.org/blogs/siegfried-pronom-based-file-format-identification-tool/.
[12] https://tika.apache.org/.
[13] https://www.ffmpeg.org/.
[14] https://www.imagemagick.org/.
[15] https://dilcis.eu/specifications/aip.

Sämtliche Funktionalitäten sind als OCFL-Erweiterung ausgeführt und legen zusätzliche, einfach verständliche Metadaten am dafür konfigurierten Ort im OCFL-Objekt ab. Weitere Metadaten-Standards (z.B. RiC) können problemlos als OCFL-Extension implementiert werden.

gocfl-Reporting

Die Schilderungen verdeutlichen, dass das gocfl-Tool dem Reporting besondere Aufmerksamkeit widmet. Mit Blick auf die Reports der erzeugten OCFL-Kapseln wird zwischen einem vollumfänglichen Report und einer druckbaren Kurzversion unterschieden. Der umfangreiche Report kann im Browser angeschaut werden. Generell unterstützt das Reporting den AIP-First Ansatz. Dieser soll sicherstellen, dass auch alleinstehende AIPs (d.h. ohne Archivierungssystem) einfach verständlich sind. Die folgenden Abbildungen vermitteln einen Eindruck davon, wie die gocfl-Reports konkret aussehen:

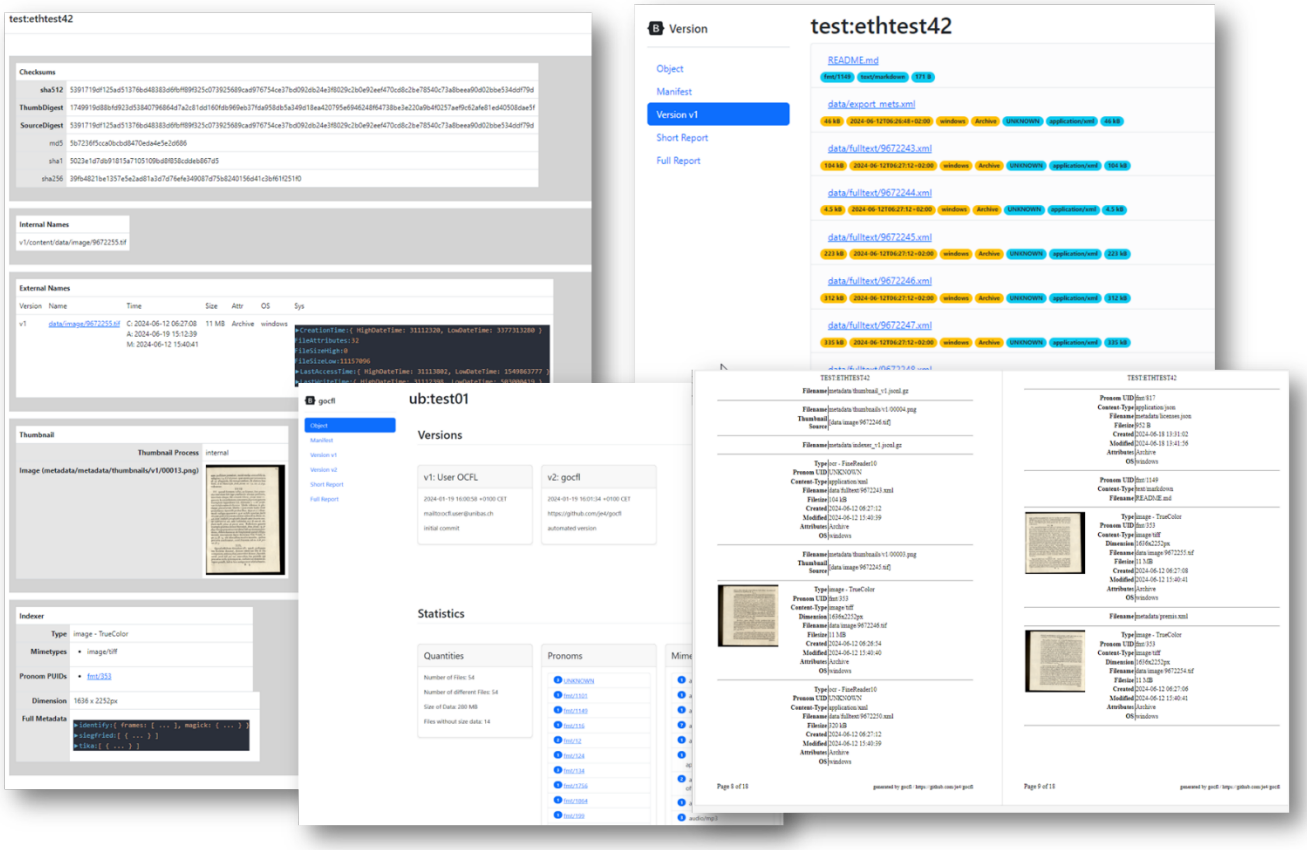


Abbildung 1: OCFL Objekt in der Dokumentationsansicht

Im Hintergrund werden links der Report sowie rechts die Verzeichnisstruktur des OCFL-Pakets dargestellt. Während links die unterschiedlichen Analyseergebnisse des webbasierten Reports zu einer Datei angezeigt sind, vermittelt der Screenshot rechts die Struktur des OCFL-Pakets mit seinem Manifest sowie den unterschiedlichen Versionen des Objekts. Auch ist der Einstieg

in die Webansicht (Fullreport) und die Kurzversion des PDFs möglich. Die beiden Abbildungen im Vordergrund zeigen dann die Übersicht des Objekts sowie einen Auszug des visuellen Reports, der eine Zusammenfassung der Dateiinformationen bereitstellt.

Archivmanagementsystem

Flexibilität, Erweiterbarkeit und Anpassungsfähigkeit wird auch vom Archivmanagementsystem selbst gewünscht. Daher orientiert sich die Systematik zum Archivmanagementsystem im vorliegenden Fall am Ansatz gegenwärtiger MACH-Architekturen. Das Akronym steht für **M**icroservice-based, **A**PI-first, **C**loud-based und **H**eadless. Sowohl das Aufsetzen auf Mikroservices als auch die Regelung der systeminternen Kommunikation über definierte Schnittstellen (APIs) garantieren Flexibilität und Erweiterbarkeit bei den verwendeten Software-, Identifikations- und Kontrollkomponenten. Auch die Dienste Dritter lassen sich leicht und definiert einbinden. Während letztes die grundsätzliche Cloudfähigkeit des Systems erfordert, erlaubt es die Ausrichtung als entkoppeltes (headless) Management-System, unterschiedliche bedarfsgerechte Dienste nahtlos zu integrieren. Insgesamt liefert der MACH-Ansatz damit eine robuste Grundlage für eine flexible Softwarearchitektur, die als moderne Organisationsform eine große technologische Bandbreite an Möglichkeiten bereitstellt.

gocfl-basiertes Umsetzungsbeispiel der UB Basel

Die Komponenten des Archivmanagementsystems der UB Basel setzen den skizzierten, OCFL-basierten Ansatz um. Wie die folgende Grafik (Abb. 2) verdeutlich, wird zwischen dem sogenannten Pre-Ingest, der eigentlichen ONAS-Routine und dem Abschnitt Management & Reporting unterschieden.

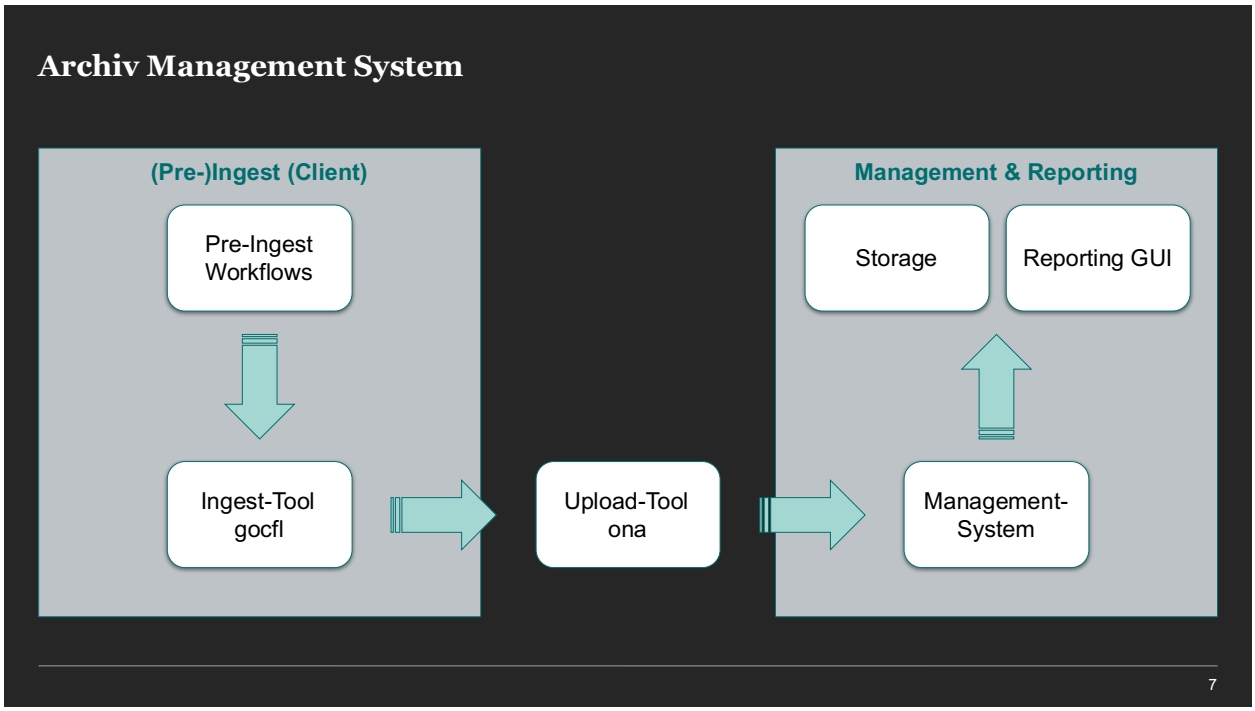


Abbildung 2

Da der (Pre-)Ingest-Workflow stark von der jeweils vorhandenen Ablagestruktur und den Inhaltstypen abhängt, muss er üblicherweise individuell, das heißt pro Einrichtung oder pro erfassende Einheit leicht modifiziert implementiert werden. Am Ende dieser vorbereitenden Schritte stehen OCFL-kompatible Datenobjekte, welche nach der ONA-Routine der gocfl-Operationen archiviert werden.

Blickt man eine Ebene tiefer auf die operative Umsetzung, zeigt sich im Falle der UB Basel folgende Umsetzung eines ONAS (vgl. Abb. 3).

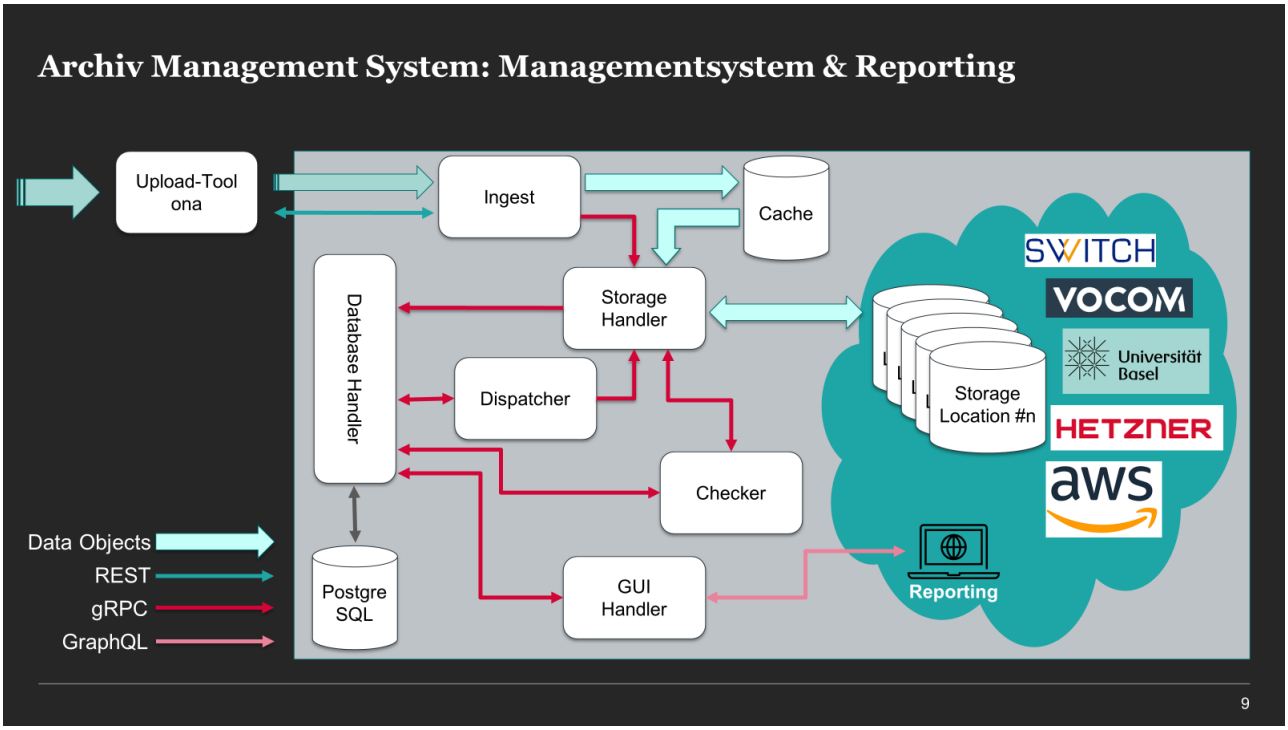


Abbildung 3

Nach dem (Pre-)Ingest wird der OCFL-Container mit Hilfe des ONA-Werkzeugs dem Ingest-Service des OCFL Native Archive übergeben. Mit Hilfe des Storage-Handlers werden die Datenobjekte dann je nach Notwendigkeit an die verschiedenen Speicherorte kopiert und die Konsistenz geprüft. Die interne Kommunikation zwischen den Microservices erfolgt performant und sicher mit Hilfe von gRPC und mTLS. Der Zugriff auf die Daten ist nur durch den Storage-Handler möglich, welcher gesondert gesichert werden muss.

ONAS und gocfl

Vor dem Hintergrund der zuvor erläuterten Workflows und Automatismen kann das eigentliche ONAS (OCFL-Native-Archive-System) am Beispiel des kooperativen Archivverbunds zwischen UB Basel und ZHB Luzern wie folgt charakterisiert werden: Als OCFL-basiertes Archivmanagementsystem baut ONAS die Objektstrukturen der AIPs gemäß dem OCFL-Vorgaben und dokumentiert diese. Die Dokumentation ist sowohl innerhalb der OCFL-Objekte quasi als Selbstdokumentation enthalten als auch in den beiden automatisch erstellten Reports, also einer umfassenden (Browser-)Langversion und einem PDF-kompatiblen Kurzreport.

Berücksichtigt sind im OCFL-Objekt auch Versionierung und/oder Anpassungen wie etwa verstetigte Dateinamen, bei denen nicht-nachhaltige Sonderzeichen in UTF8-kompatible Zeichenketten umgewandelt wurden, oder migrierte Inhalte. Diese zusätzlichen Metadaten werden in der Logik des OCFL-Standards im Inventory abgelegt. In der Anwendungsform der UB Basel

sind zudem die beiden im Archivumfeld hoch geschätzten Beschreibungsstandards METS und Premis sowie E-ARK AIP integriert.

Bezogen auf die kooperative Zusammenarbeit der beiden Partnerinstitutionen sind die beiden Prinzipien der Transparenz und der Flexibilität hervorzuheben: Jede Organisation hat jederzeit einen Überblick über ihre Daten. Transparenz der Information bedeutet, dass jeder Stakeholder folgende Fragen über seine Daten unabhängig vom anderen beantworten kann:

- Welche Objekte wurden gespeichert?
- Wie sehen die AIPs der Objekte aus?
- Wo sind die OCFL-Derivate der APIs abgelegt?
- Wann wurden die OCFLs zuletzt überprüft (Prüfsumme)?
- Wie viel Speicherplatz wird aktuell benötigt?
- Welche Medientypen befinden sich in den Objekten?

Flexibilität in allen technischen Bereichen meint hingegen, dass jeder der Stakeholder

- den/die Speicherdienstleister oder Hersteller frei auswählen kann (aktuell S3 Protokoll)
- die Daten bei Speicherausfall oder einer Abschaltung automatisch auf Alternativsysteme transferieren kann
- Pre-Ingest Prozesse spezifisch implementierbar und nicht zuletzt
- Ingest-Pipelines mit bewährten Tools modular bestimmt werden können.

All das ist möglich, ohne dass der/die anderen Kooperativpartner des Verbundes davon unmittelbar tangiert werden.

Weiterentwicklung

Mögliche Erweiterungen der bestehenden gocfl-Routinen sind wie folgt vorgesehen:

- Neue Partner – im Verbund und für Dienstleistungen
- Feinabstimmung im Bereich (Pre-)Ingest und Monitoring
- Synergien schaffen bei Erarbeitung und Pflege von Tools und Knowhow
- Format-Migrationsstrategien können, aber müssen nicht gemeinsam erarbeitet und umgesetzt werden.

Finanzierungskonzept

Mit Blick auf das Finanzierungskonzept wurden folgende Prinzipien angewandt:

- Möglichst niedriger Kostensockel für Betrieb durch ein günstiges, opensource-basiertes Managementsystem, das ohne allzu großen Schulungsaufwand betrieben werden kann.
- Erweiterbarkeit durch Partner:innen für Dienste oder Erweiterungen als bedarfsgerechtes Entwickeln auf eigene oder verteilte Kosten

- Speicherkosten abhängig von der Wahloption der Partner:innen (Kosten, Betrieb (Cloud).

Zusammenfassung und Ausblick

Zusammenfassend lassen sich die wichtigsten Punkte wie folgt verallgemeinern: Durch die Verwendung von OCFL und die Nutzung von gocfl wird sowohl die Ablage- als auch die AIP-Struktur des Archivs nachvollziehbar und für Maschinen aber eben auch für Menschen und damit für künftige Generationen verständlich. Zugang und Wiederherstellung der Archivalien sind ohne den Einsatz von Spezialwerkzeugen möglich. Auf der Verwaltungsebene ist das System mandantenfähig, was neue Kooperations- und Kollaborationsoptionen ermöglicht. Die unterschiedlichen Parteien (Mandanten) können eigene Speichersysteme einbringen oder für ihre Lagerung auswählen oder je nach Kontext auch geteilte. Da der Einstieg in die Nutzung jederzeit ausbaufähig ist, können z. B. neue Mitglieder eines Verbundes auch erst einmal mit kleinem Featuresatz beginnen und sich einen Überblick über die Komplexität und die institutionsintern anfallenden Aufwände machen, bevor sie ihre gesamten Workflows umstellen. Wünschen eine Partei oder mehrere Partner eines Verbundes Spezialfeatures, lassen sich diese integrieren, ohne dass die anderen tangiert würden. So kann jede/r eine für sie/ihn optionale Umsetzungsversion nutzen. Die Zusammenarbeit der Metadaten im Archiv ist verständlich und transparent strukturiert. Aufgrund der Einfachheit der AIPs und Metadaten ist es nicht schwierig, eigene Sichtungsfrontends zu implementieren und somit nach außen weiterhin als wiedererkennbare, eigenständige Institution aufzutreten. Nicht zuletzt besteht jederzeit absolute Kostentransparenz und die Ausgaben lassen sich auf jenes Minimum reduzieren, das in Anbetracht der gewünschten Speicherqualität erforderlich ist. Die Kosten werden nicht durch die Kosten der Systemanbieter in die Höhe getrieben. Mit Blick auf die Community-Komponente des Systems sei zuletzt angemerkt, dass für Spezial-Erweiterungen, die dauerhaft allen zur Verfügung stehen sollen, aufgrund der Spezifikationen eine Voll-Integration als OCFL-Extension möglich ist.

Bibliografie

Amrein, E. (2008), *Memopolitik. Eine Politik des Bundes zu den Gedächtnissen der Schweiz. Bericht des Bundesamtes für Kultur*. Bern: BAK.

Enge, J. / Kramski, H. W. (2016), „Exploring Friedrich Kittler's Digital Legacy on Different Levels. Tools to Equip the Future Archivist", in: Swiss National Library (Hg.), *iPRESS 2016*. Bern: SNB, S. 229–236.

Xdomea-Aussonderungsmanager: Open-Source-Lösung des Landesarchivs Thüringen zur Bewertung und Übernahme von E-Akten

Christine Träger

Seit 2017 ist das Datenaustauschformat xdomea gemäß Beschluss des IT-Planungsrats für die Aussonderung von elektronischen Akten in der öffentlichen Verwaltung in Deutschland verbindlich anzuwenden. Xdomea stellt für die im zwei- und vierstufigen Aussonderungsverfahren üblichen Arbeitsschritte definierte Nachrichten bereit, die zwischen einem Dokumentenmanagementsystem und einem Archivsystem ausgetauscht werden sollen (Anbietungsnachricht, Bewertungsnachricht, Abgabenachricht sowie Empfangsbestätigungen). Mit entsprechender Vorkenntnis und Mühe ist xdomea als XML-basierter Standard zwar menschenlesbar, die Erstellung und Verarbeitung der Nachrichten ist jedoch maschinell gedacht, insbesondere um die gemäß XML und xdomea-Spezifikation vorgegebene Syntax und Struktur der Nachrichten sowie eine zuverlässige, automatisierte Verarbeitung zu gewährleisten.
Für die Steuerung des gesamten Aussonderungsprozesses und die Erstellung der benötigten xdomea-Nachrichten auf Seite der abgebenden Stelle bieten die meisten DMS umfassende Funktionen an. Für die archivseitige Anzeige, Verarbeitung und Erzeugung von xdomea-Nachrichten ist das Archiv hingegen auf sich gestellt. Im besten Fall betreibt das Archiv ein Digitales Magazin, das Funktionen zur Bewertung und Übernahme von E-Akten bereitstellt. Mit der vom Landesarchiv Thüringen im Herbst 2023 entwickelten Webanwendung xdomea-Aussonderungsmanager (kurz: x-man) können sowohl mit als auch ohne vorhandene Langzeitarchivierungslösung elektronische Akten medienbruchfrei bewertet und ins Archiv übernommen werden.

Datenaustausch zwischen DMS und x-man

Unter Verwendung des Austauschstandards xdomea ist es nicht erforderlich, DMS und Archivsystem für den Datenaustausch über eine direkte Schnittstelle zu verbinden. Vielmehr kann die Kommunikation zwischen beiden Systemen über ein Transferverzeichnis (bspw. webDAV) erfolgen. X-man überwacht das hinterlegte Verzeichnis stetig, holt neue Anbiete- und Abgabenachrichten automatisch von dem Verzeichnis ab, prüft die xdomea-Nachrichten technisch und zeigt sie anschließend für die Archivarin, den Archivar in x-man zur Bearbeitung an, ohne das

sendende DMS zu kennen. Auch die vom Archiv im Aussonderungsprozess zu erstellende Bewertungsnachricht wird nach Abschluss der Bewertung ohne weiteres Zutun der Archivarin in der Anwendung erzeugt und per Knopfdruck im Austauschverzeichnis für das DMS zur Abholung bereitgestellt. Auf diese Weise kann x-man mit jedem beliebigen xdomea-konformen DMS Aussonderungsnachrichten austauschen.

Im aktuellen Entwicklungsstand unterstützt x-man die xdomea-Versionen 2.3 bis 3.1 (Stand: Oktober 2024). Die Anwendung erkennt dabei die einer Nachricht zugrunde liegende xdomea-Version automatisch. Die bestehende Rückwärtskompatibilität soll soweit möglich auch zukünftig bestehen bleiben. Weichen die Spezifikationen des Standards zukünftig zu stark voneinander ab, wird es jedoch erforderlich sein, einige derzeit unterstützte ältere xdomea-Versionen aufzugeben.

Funktionsumfang von x-man

Der Aussonderungsablauf mit x-man richtet sich nach den Vorgaben des Standards xdomea für das zwei- und vierstufige Aussonderungsverfahren, das sich wiederum an den Abläufen der Aussonderung papiernen Schriftguts orientiert. X-man kann alle im Prozess vorgesehenen xdomea-Nachrichten empfangen und senden (s. Abb. 1).

Abbildung 4: xdomea-Nachrichten im vierstufigen Aussonderungsverfahren

Bei der Konzeption und Entwicklung der Anwendung wurde besonderes Augenmerk daraufgelegt, dass Archivarinnen und Archivare ohne tiefere IT-Kenntnisse mit Hilfe des Programms die Bewertung und Übernahme von E-Akten eigenständig durchführen können. Auch auf eine sichere Nutzerführung wurde geachtet. So kann das Programm bspw. jederzeit verlassen werden, ohne dass Informationen verloren gehen. Die Speicherung von Eingaben muss nicht aktiv ausgelöst werden und erfolgt fortwährend im Hintergrund.

Übersicht offener Aussonderungsvorgänge

In der Administration des Programms können Usern abgebende Stellen zugeordnet werden. Dadurch ist es möglich, eine eingehende Nachricht entsprechend der Geschäftsverteilung an die für die abgebende Stelle zuständige Archivarin zuzuweisen. X-man kann über den Eingang einer neuen Nachricht per E-Mail informieren (bspw. die zuständige Archivarin, eine Poststelle). Eine regelmäßige manuelle Überprüfung der Anwendung auf neue Eingänge ist dadurch nicht erforderlich. Die E-Mail-Benachrichtigung weist Zeitpunkt und Art des Eingangs (bspw. Zeitstempel der Anbietung) sowie die abgebende Stelle aus. Es ist auch möglich, der E-Mail-Benachrichtigung die eingegangene xdomea-Nachricht anzuhängen, damit diese als Eingangsnachweis in der Registraturbildnerakte veraktet werden kann. Auf ein separates Anschreiben der abgebenden Stelle zur Anbietung oder Übergabe kann dadurch verzichtet werden.

Alle offenen Aussonderungsvorgänge der zuständigen Archivarin sind auf der Startseite des xdomea-Aussonderungsmanagers in eine Übersicht zusammengefasst (s. Abb. 2). Anhand der Prozess-ID können alle xdomea-Nachrichten, die zu einer Aussonderung gehören, in Zusammenhang gebracht werden. Dadurch kann bspw. eine neu eingehende Abgabenachricht dem schon in x-man bestehenden Aussonderungsvorgang mit der Anbietenachricht zugeordnet werden. Der Status einer Aussonderung kann in der Übersicht anhand des Zeitstempels zu einem Arbeitsschritt visuell schnell erfasst werden. Solange die Bewertung noch nicht abgeschlossen und versendet wurde, wird in der Übersicht angezeigt, für wie viele Akten bzw. Vorgänge von den angebotenen Schriftgutobjekten bereits eine Bewertungsentscheidung getroffen wurde (bspw. 10/240). Grün markiert werden abgeschlossene Vorgänge, die bereits zur Löschung aus x-man vorgemerkt sind. Rot markiert werden Vorgänge, in denen Fehler aufgetreten sind. Die Benutzeroberfläche aktualisiert sich bei eingehenden Nachrichten und abgeschlossenen Prozessen automatisch.

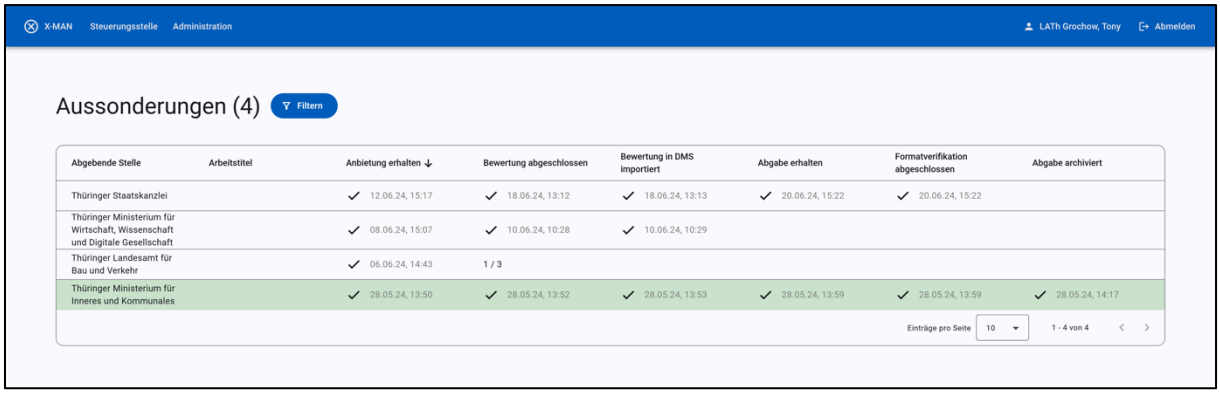


Abbildung 5: Übersicht der offenen Aussonderungsvorgänge auf der Startseite

Um während der Bearbeitung mit x-man die verschiedenen parallel vorliegenden Aussonderungen einer Behörde leichter auseinander halten zu können, kann die Archivarin für jede Anbietung bzw. Abgabe einen Arbeitstitel vergeben, der in der Übersicht der offenen Aussonderungen auf der Startseite mit angezeigt wird. Der Arbeitstitel kann frei gewählt werden (bspw. Schlagwort „Lotteriewesen“ oder „Kabinettsprotokolle“) und wird nicht mit archiviert.

Stufigkeit von E-Akten

Vereinzelt gibt es in der Thüringer Landesverwaltung den Bedarf, Akten mit mehr Gliederungsebenen als der klassischen dreistufigen Akte (Akte, Vorgang, Dokument) zu führen. Für Liegenschaftsakten wurde bspw. ein eigener, fünfstufiger Aktentyp definiert. Xdomea erlaubt Akten mit bis zu fünf Stufen (bspw. Akte, Teilakte, Vorgang, Teilvorgang, Dokument), kann jedoch technisch nicht verhindern, dass mit dem Standard auch Akten ausgetauscht werden, die mehr als fünf Stufen besitzen. X-man erkennt deshalb die jeweils vorliegende Strukturtiefe automatisch und ermöglicht die hierarchisierte Anzeige jeder beliebigen Stufigkeit. Aussonderungen mit Akten, die mehr als fünf Stufen aufweisen, werden vom System jedoch mit einem Warnhinweis versehen. Da solche Nachrichten streng genommen nicht xdomea-konform sind, soll vor der weiteren Bearbeitung der Aussonderung entschieden werden können, ob die Anbietung akzeptiert oder abgelehnt wird.

Archivische Bewertung

Die Angaben aus der xdomea-Anbiete- bzw. -Abgabenachricht werden in der Benutzeroberfläche von x-man menschenlesbar und in strukturierter Form angezeigt (s. Abb. 3). Auf der linken Seite können die Akten, Vorgänge, Dokumente und ggf. weitere Schriftgutobjekte in einer Baumansicht aufgefächert werden. Für jede Hierarchieebene (Anbietung, Akte, Vorgang, Dokument usw.) wurde ein Minimalset an Metadaten festgelegt, das nach Klick auf das entsprechende Schriftgutobjekt im Baum auf der rechten Seite angezeigt wird. Angezeigt werden können dabei alle Metadaten, die in der xdomea-Datei enthalten sind, eine Erweiterung des derzeit in die Oberfläche übertragenen Minimalsets ist auf Wunsch der Archivarinnen möglich. Im Bereich Status kann außerdem der aktuelle Bearbeitungsstand der geöffneten Aussonderung aufgerufen werden, der auch in der Übersicht auf der Startseite dargestellt wird.

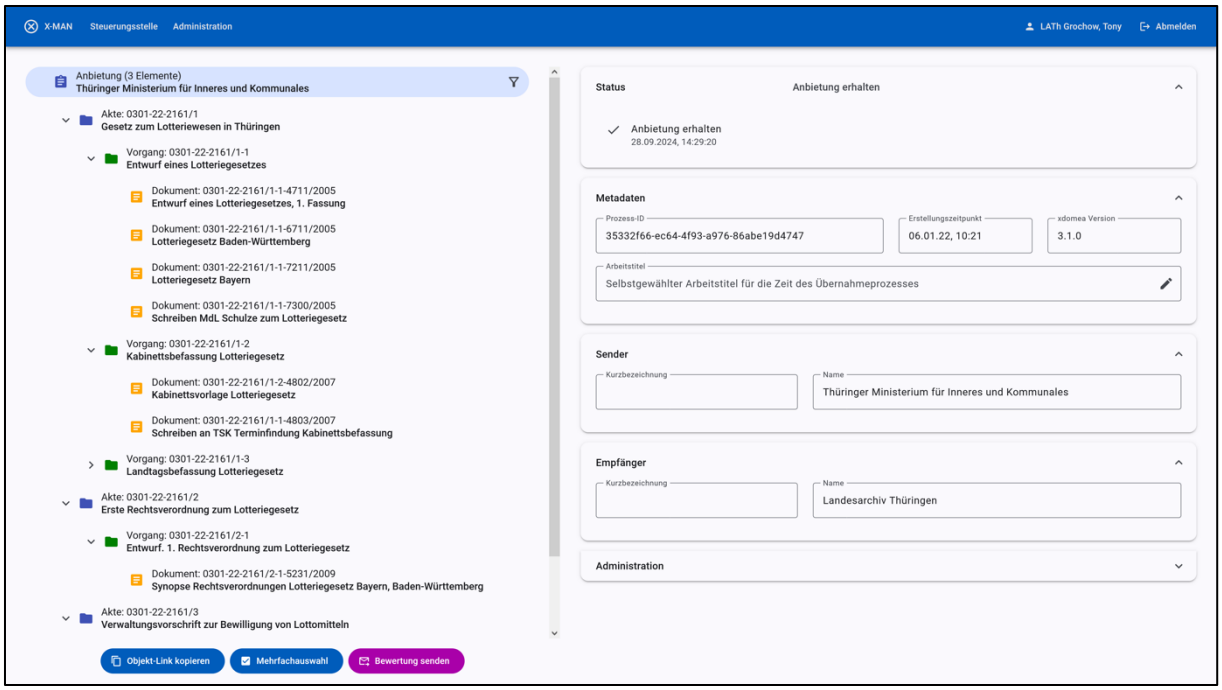


Abbildung 6: Detailansicht einer Anbietung

Die Bewertung einer Anbietung kann von der Archivarin ohne Medienbruch innerhalb der Anwendung vorgenommen werden. Standardmäßig erfolgt die Bewertung auf der Wurzelebene (Aktenebene). Bei Bedarf kann in den Umgebungsvariablen von x-man konfiguratorisch eingestellt werden, dass die Bewertung auf allen Ebenen zugelassen wird (Akte, Teilakte, Vorgang, Teilvorgang). Je nach Konfiguration wird unterhalb der behördlichen Metadaten einer Hierarchieebene (bspw. Akte) der Bewertungsbereich angezeigt (s. Abb. 4). Sofern die abgebende Stelle einen behördlichen Bewertungsvorschlag für eine Akte im DMS hinterlegt hat und dieser in die xdomea-Anbietenachricht übertragen wurde, wird der Bewertungsvorschlag in einem eigenen Feld in diesem Bereich angezeigt. Ansonsten ist das Vorschlagsfeld leer.

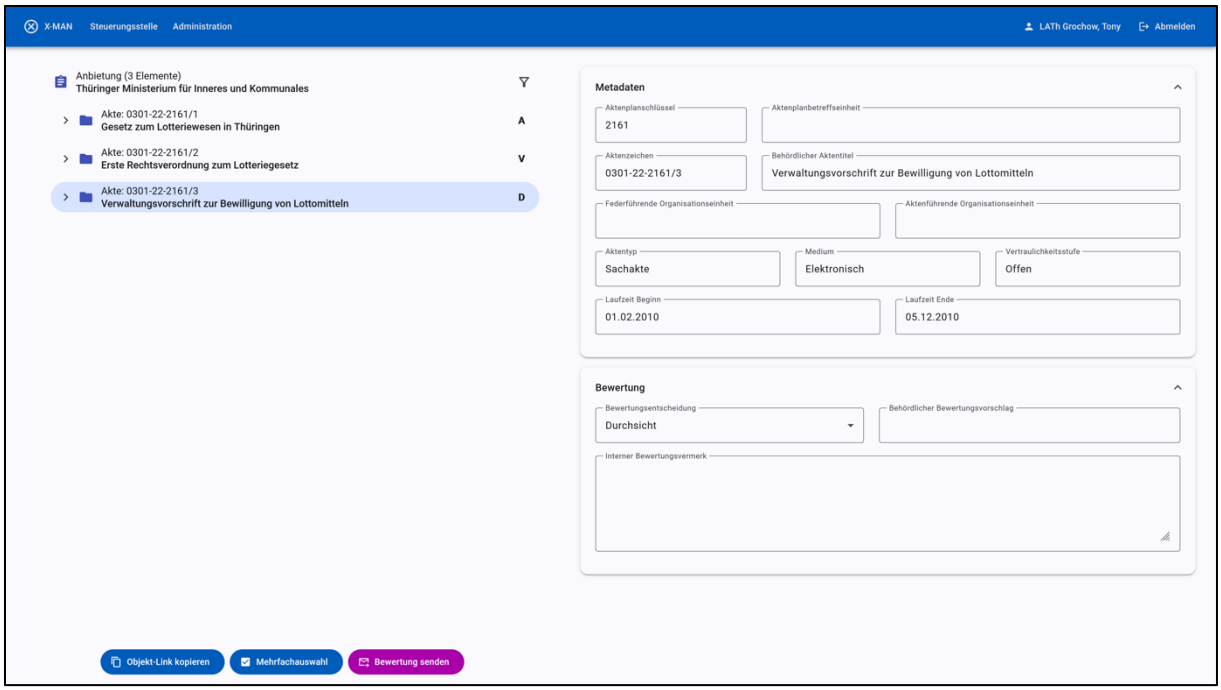


Abbildung 7: Bewertungsentscheidung in der Baumansicht

Die Hinterlegung der Bewertungsentscheidung erfolgt im Feld *Bewertungsentscheidung* über eine Auswahlliste mit den üblichen Werten *Archivieren* (A), *Durchsicht* (D) und *Vernichten* (V). Die eingestellte Bewertung wird in Form der jeweiligen Abkürzung untermittelbar in die Baumansicht übertragen, um der Archivarin einen schnellen Überblick zu ermöglichen, welche Schriftgutobjekte bereits bewertet wurden (s. Abb. 4). Im Textfeld *Interner Bewertungsvermerk* kann eine Begründung für die Bewertungsentscheidung hinterlegt werden. Die Angaben werden nicht in das Archivpaket übertragen, sondern ausschließlich in den Bewertungsbericht aufgenommen, der zur Dokumentation der Bewertung dient. Die Angaben können einen separaten Bewertungsvermerk ersetzen oder als Gedankenstütze für einen solchen dienen.

Xdomea erlaubt im Bewertungsverzeichnis auch die Rückgabe des Wertes *Durchsicht* (D), mit denen die Archivarin E-Akten kennzeichnet, die für die Bewertung eingesehen werden müssen. Dadurch würde mit Abschluss der Bewertung und Erzeugung der xdomea-Nachricht 0502 keine abschließende Bewertungsentscheidung für alle Schriftgutobjekte vorliegen. D-Positionen werden in der Regel vom DMS gemeinsam mit den archivwürdigen E-Akten an das Archiv übergeben und im Archiv nachbewertet. Die nachträgliche Vervollständigung eines solchen vorläufigen Bewertungsergebnisses ist jedoch im xdomea-Ablauf nicht vorgesehen. Die Rückmeldung des endgültigen Bewertungsergebnisses an die abgebende Stelle müsste außerhalb des Prozesses, bspw. per E-Mail erfolgen. Auch die automatische Vollständigkeitsprüfung einer Abgabe anhand der Bewertungsnachricht wird durch D-Positionen gestört. Darüber hinaus werden in diesem Fall E-Akten in das Digitale Magazin übernommen und gespeichert, die unter

Umständen nicht archivwürdig sind und nachträglich wieder aus dem Archiv entfernt werden müssen, es sei denn, es werden in den Übernahmeprozess Funktionen zur Nachbewertung implementiert. Aufgrund des in Thüringen archivgesetzlich zur Verfügung stehenden Bewertungszeitraums von einem Jahr und dem Bestreben nach medienbruchfreien, automatisierten und einfach gehaltenen Prozessen, wurde entschieden, den Wert *Durchsicht* in x-man nur als Arbeitsinstrument für den Zeitraum der Bewertung zuzulassen. Vor Abschluss der Bewertung und Übersendung dieser an die abgebende Stelle müssen D-Positionen händisch zu A-Positionen bzw. automatisch zu V-Positionen geändert werden. Dies entspricht auch dem Standardvorgehen bei der Bewertung papiernen Schriftguts, bei der D-Positionen nur zur übergangsweisen Markierung der Listeneinträge genutzt werden, die in der Altregistratur eingesehen werden sollen.

Für die Bewertung umfangreicher Anbietungen mit Hunderten von Akten stehen in der Baumansicht Filterungsmöglichkeiten zur Verfügung, die anhand praktischer Erfahrungen in den nächsten Jahren weiter ausgebaut werden sollen. So kann derzeit bspw. nach Aktenplanschlüsseln und Laufzeit gefiltert werden. Auch kann die Ansicht auf die noch nicht bewerteten Schriftgutobjekte reduziert werden.

Nach Abschluss der Bewertung erstellt x-man einen Bewertungsbericht (s. Abb. 5), in dem nicht nur dokumentiert ist, wann die Anbietung eingegangen ist, von wem die Bewertung durchgeführt und wann sie an die abgebende Stelle übersandt wurde, sondern auch eine Liste aller angebotenen Schriftgutobjekte (oberste Hierarchieebene) mit der jeweiligen Bewertungsentscheidung und ggf. angebrachten Bewertungsvermerken enthalten ist. Der Bericht dokumentiert so nicht nur die Bewertungsentscheidungen der Archivarin, er dient auch als Beleg für die Übersendung der Bewertung an die abgebende Stelle. Es besteht die Möglichkeit, den Bericht in der Anwendung herunterzuladen. Zudem wird der Bericht der zuständigen Archivarin bzw. dem konfigurierten Empfängerkreis zur Veraktung automatisch per E-Mail zugesandt. Anlage der E-Mail ist neben dem Bewertungsbericht auch die von x-man erzeugte xdomea-Bewertungsnachricht 0502.

Abbildung 8: Auszug aus dem Bewertungsbericht

Formaterkennung und -validierung übernommener Dateien mit BorgFormat

Nach Eingang einer xdomea-Abgabenachricht wird zunächst geprüft, ob im System eine zugehörige Bewertung vorliegt (im zweistufigen Aussonderungsverfahren wäre dies nicht der Fall) und eine Vollständigkeitsprüfung der Abgabe durchgeführt. Dabei wird nicht nur festgestellt, ob alle als archivwürdig bewerteten E-Akten in der Abgabe enthalten sind, sondern auch, ob zusätzliche Schriftgutobjekte in der Abgabe vorliegen oder verwiesene Primärdateien fehlen.

Im Anschluss daran startet automatisch die Formaterkennung und -validierung für die in der Abgabe enthaltenen Dateien. X-Man nutzt für die Formatverifikation das Open-Source-Programm *BorgFormat* des Landesarchivs Thüringen (für nähere Informationen siehe den Beitrag von Tony Grochow zu Borg in diesem Tagungsband).

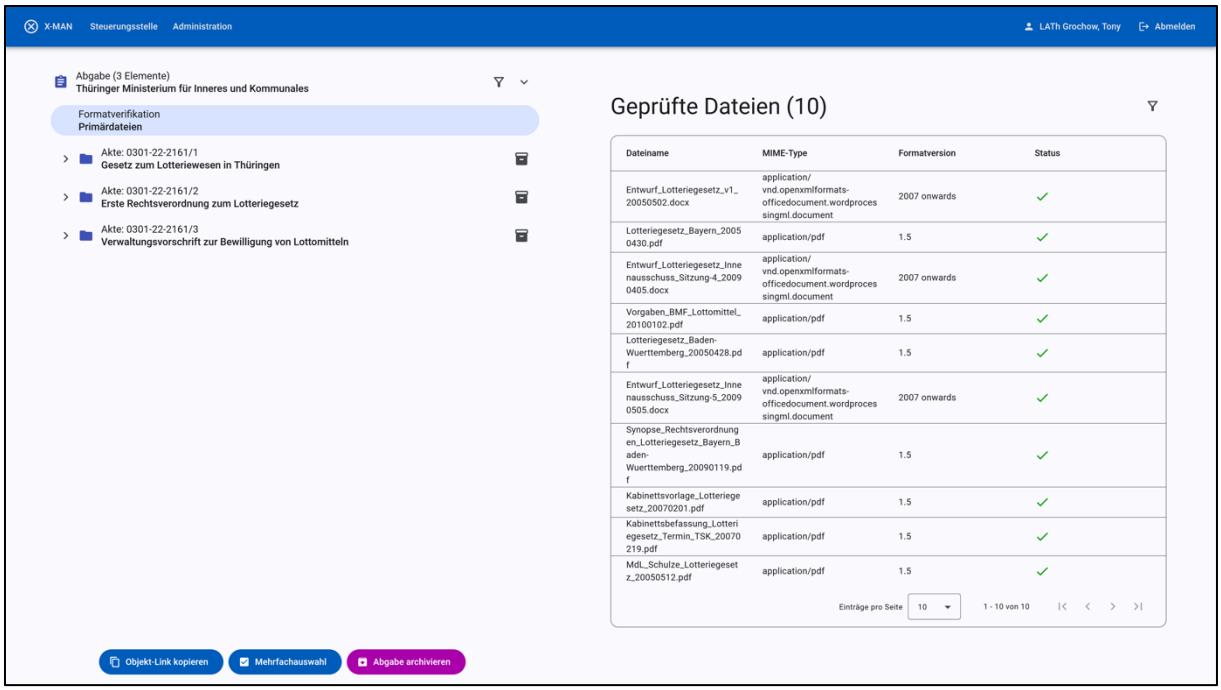


Abbildung 9: Gesamtergebnis der Formatverifikation mit Borg

Die Ergebnisse der Formaterkennung und -validierung können unmittelbar an der Abgabenachricht aufgerufen und angesehen werden (s. Abb. 6). Mit Hilfe des von Borg ermittelten Gesamtstatus einer Datei kann die Archivarin schnell überblicken, ob problematische Dateiformate enthalten sind, die vor der Archivierung von der Steuerungsstelle geprüft werden müssen. Eine tiefere Befassung der Archivarin mit den Details der Formatverifikation ist zwar mit Klick auf einzelne Ergebnisse möglich, aber nicht erforderlich.

Paketierung und Archivierung einer Abgabe

Die Aufteilung einer Abgabe in Archivpakete ist nach individuellen Bedingungen möglich. Standardmäßig wird für jede Akte ein separates Archivpaket gebildet (s. die Kartonsymbole in Abb. 7). Sofern Vorgänge die Wurzelebene der Nachricht bilden (Vorgangsbildung ohne zugehörige Akte), wird ebenfalls pro Vorgang ein Archivpaket angelegt. Lose Dokumente auf der Wurzelebene werden in ein Archivpaket zusammengefasst. Ab xdomea-Version 4.0.0 sind lose Dokumente auf der Wurzelebene nicht mehr gestattet.

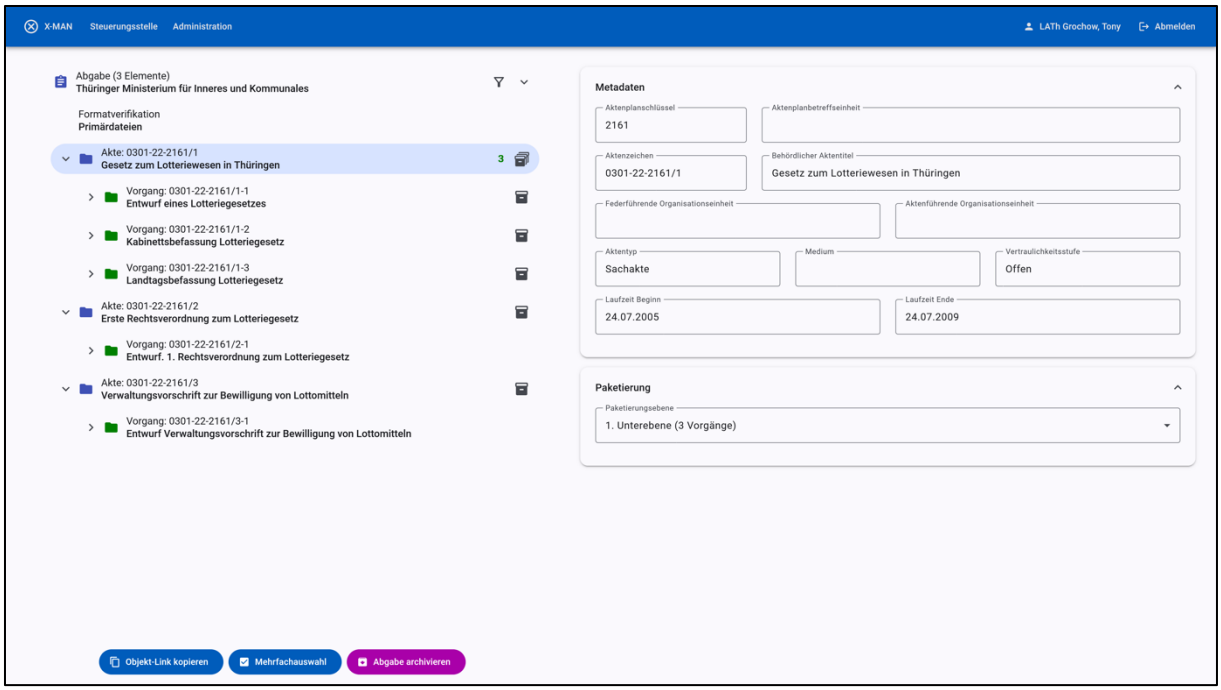


Abbildung 10: Paketierung der Abgabe auf Akten- und Vorgangsebene

Es besteht zudem die Möglichkeit, die automatische Paketierung vor der Archivierung der Pakete anzupassen und für die gewünschten Schriftgutobjekte auf einer tieferen Ebene (bspw. nach Teilakten oder Vorgängen) vorzunehmen.

X-man kann an den Archivspeicher eines Digitalen Magazins angebunden werden, so dass die mit x-man übernommenen archivwürdigen E-Akten nach der Paketierung zu Archivpaketen unmittelbar im Digitalen Magazin gespeichert werden, ohne dort noch einmal den Ingestprozess durchlaufen zu müssen. Eine Schnittstelle zum Kernmodul des DIMAG-Verbundes wird mit dem Programm bereits ausgeliefert. Die Übergabe der Archivpakete an das DIMAG-Kernmodul erfolgt in einer BagIt-Struktur, um die Integrität der Daten bei der Übertragung sicherzustellen. Zudem ist es möglich, die mit x-man geschnürten Archivpakete anstatt in ein Digitales Magazin auf ein lokales Verzeichnis zu speichern. Dadurch können elektronische Akten auch dann bereits sachgerecht ausgesondert, bewertet und auf einem Zwischenspeicher im Archiv gesichert werden, solange noch keine Langzeitarchivierungslösung in Betrieb ist.

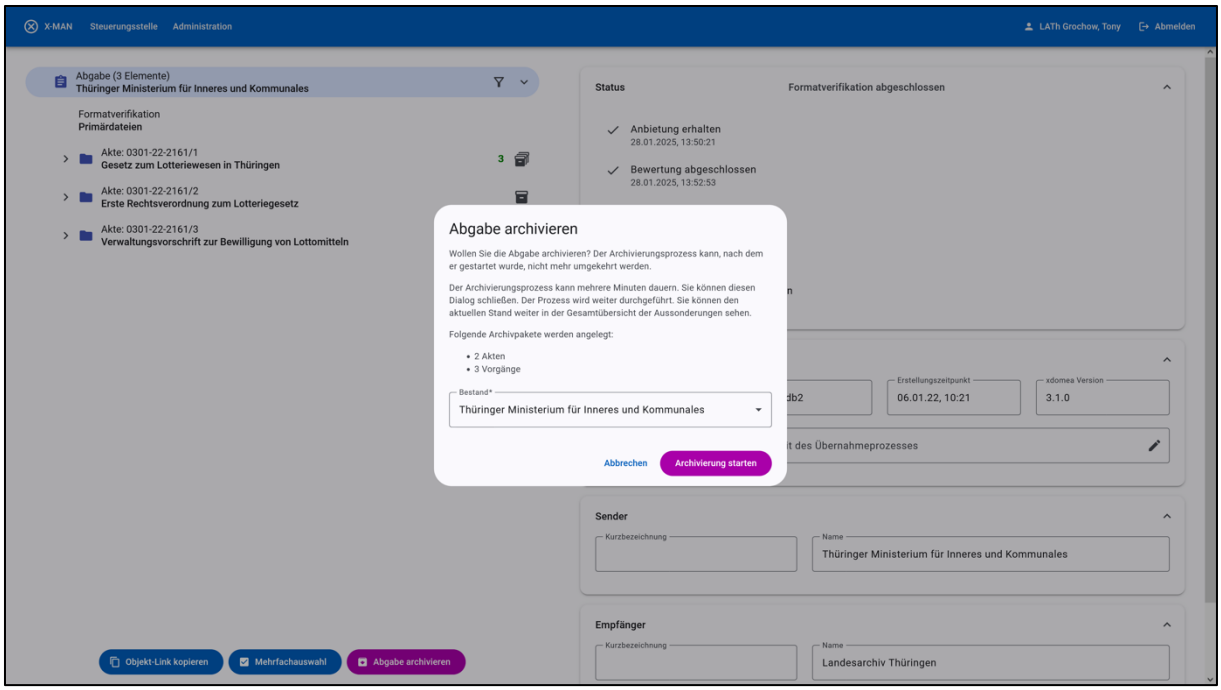


Abbildung 11: Sicherheitsabfrage bei der Speicherung der Archivpakete

Die Zuordnung der Archivpakete zu einem Bestand im Digitalen Magazin erfolgt standardmäßig nach den für die abgebende Stelle in x-man hinterlegten Bestandsangaben. Im Archivierungsdialog kann der Bestand bei Bedarf manuell geändert werden (s. Abb. 8). Mit Auslösen der Archivierung der Abgabe wird die xdomea-Abgabenachricht (0503) auf die Archivpakete zugeschnitten. Jedes Archivpaket enthält im Anschluss eine eigene xdomea-Abgabenachricht, die nur die Informationen zu den im Archivpaket enthaltenen Schriftgutobjekten enthält. Die zerteilten Nachrichten sind weiterhin xdomea-konform und validierbar.

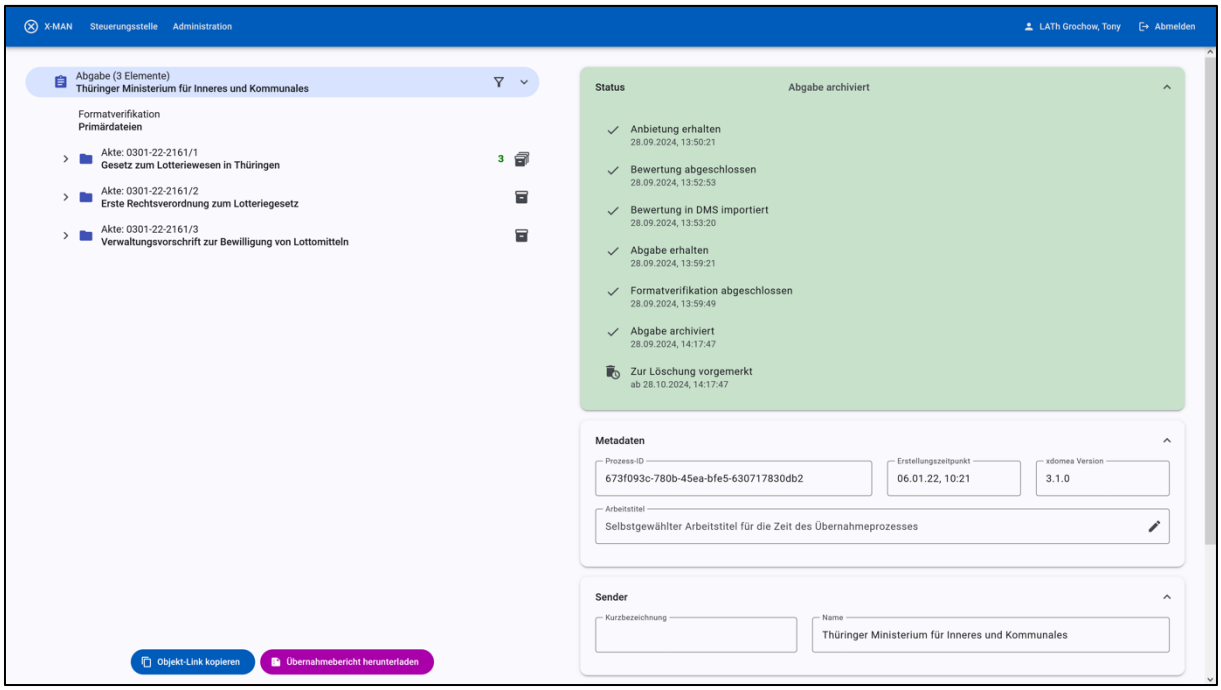


Abbildung 12: Status "Zur Löschung vorgemerkt"

Sobald die Speicherung der gebildeten Archivpakete abgeschlossen ist, werden die xdomea-Nachrichten, die zu dieser Abgabe gehören, in den Status „zur Löschung vorgemerkt“ gesetzt (s. Abb. 9). X-man stellt abschließend per E-Mail und zum Download einen Übernahmebericht bereit, der die Bildung der Archivpakete dokumentiert.

Problembehandlung und Protokollierung

Im Hintergrund prüft x-man die ein- und ausgehenden xdomea-Nachrichten automatisch auf Konformität zum Standard, Vollständigkeit und Plausibilität. Wird ein Fehler festgestellt, wird die betroffene Nachricht in der Anwendung als fehlerhaft markiert und zunächst für die weitere Bearbeitung gesperrt.

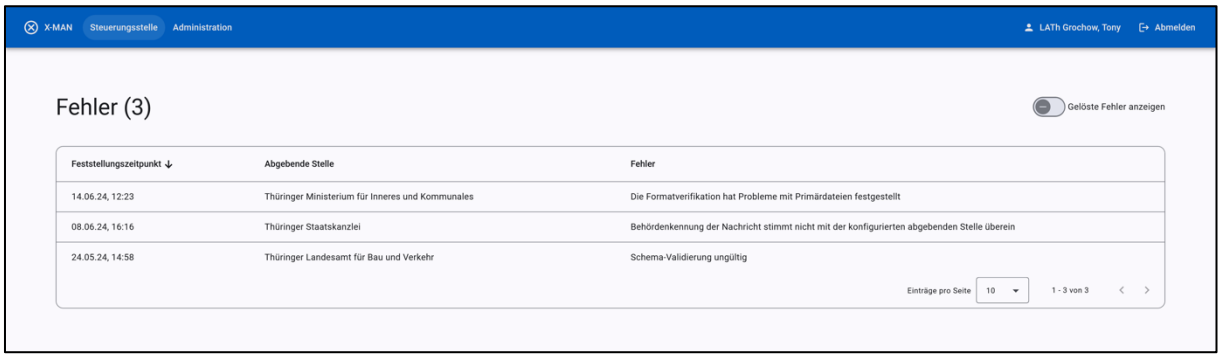


Abbildung 13: Ansicht der Steuerungsstelle

Mit dem administrativen Zugriff der Steuerungsstelle können die Fehlermeldungen eingesehen werden (s. Abb. 10 und 11). Je nach Problemfall werden im Kontextmenü verschiedene Lösungen zur Behebung angeboten. So kann eine fehlerhafte xdomea-Nachricht bspw. automatisiert gelöscht und neu eingelesen werden oder ein Problem kann aufgrund der Geringfügigkeit des Fehlers auf Ignorieren gesetzt und somit die Bearbeitung wieder aktiviert werden. Auch fehlgeschlagene Prozesse werden der Steuerungsstelle gemeldet. Schlägt die Formatverifikation für einzelne Dateien fehl oder ist die Archivierung im Archivspeicher nicht erfolgreich, können die Prozesse für die betroffenen Dateien und Pakete über die Steuerungsstelle erneut angestoßen werden.

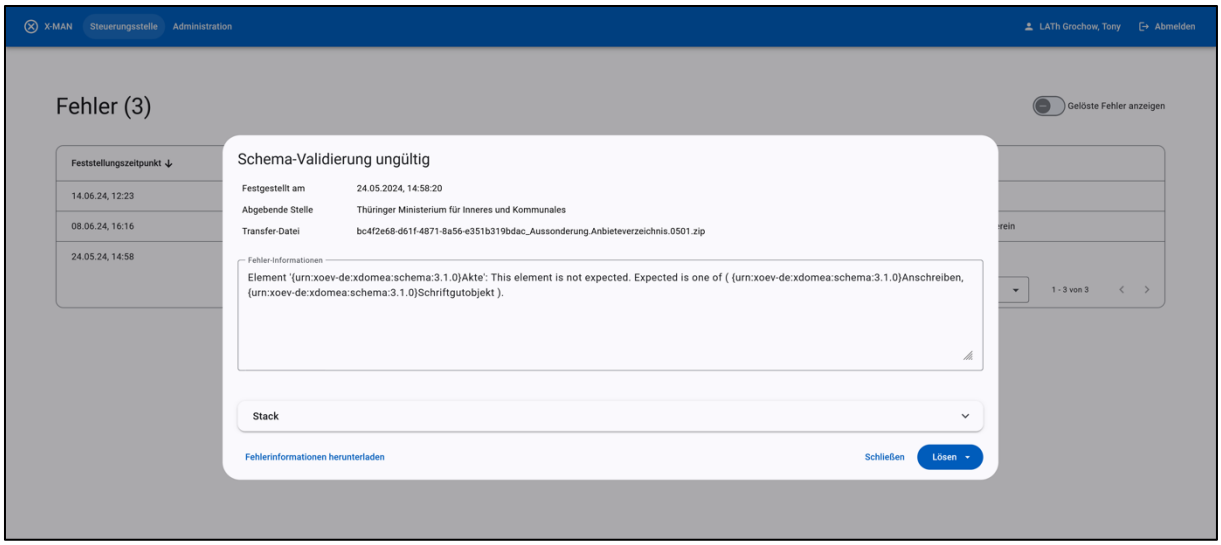


Abbildung 14: Fehlermeldung zu einer ungültigen xdomea-Nachricht

Die wichtigsten Ereignisse und Fehler werden dem Archivpaket in einer Textdatei beigelegt. Sofern die Schnittstelle zum DIMAG-Kernmodul genutzt wird und die Archivpakete dorthin gespeichert werden, werden die Protokollinformationen in die Protokolldatei des Kernmoduls integriert. In diesem Fall entfällt die separate Protokolldatei von x-man im Archivpaket.

Testbetrieb mit VIS (PDV)

Seit Inkrafttreten des Thüringer Gesetzes zur Förderung der elektronischen Verwaltung (Thüringer E-Government-Gesetz – ThürEGovG) 2018 ist es Behörden und Einrichtungen des Freistaats Thüringen möglich, Akten ausschließlich elektronisch zu führen. Einige Landesbehörden setzten DMS bereits in den Jahren davor zur doppelten Aktenführung oder als Registraturprogramm ein. In Umsetzung der gesetzlichen Vorgaben wird allen Landesbehörden seit Ende 2021 zur Führung der elektronischen Verwaltungsakte das einheitliche und zentral im Thüringer Landesrechenzentrum betriebene Dokumentenmanagementsystem VIS der Firma PDV zur Verfügung gestellt. Für das Landesarchiv Thüringen ergibt sich daraus eine behördenübergreifend einheitliche Aussonderungsschnittstelle für Verwaltungsakten.

Für die initiale Einrichtung des VIS-Aussonderungsmoduls für die Thüringer Landesverwaltung wurde Ende 2023 im Thüringer Landesrechenzentrum eine Teststellung mit aussonderungsreifen Produktivdaten des Thüringer Ministeriums für Inneres und Kommunales eingerichtet. Im Zusammenspiel zwischen VIS-Aussonderungsmodul und x-man wird seitdem geprüft, ob der Workflow fehlerfrei abläuft, an welchen Stellen noch Anpassungsbedarf besteht und welche Konfigurationen des VIS-Aussonderungsmoduls für die Produktivsetzung der Aussonderung zugrunde gelegt werden sollen. So ist bspw. zu entscheiden, welche Dateiformate

im DMS in welche Archivierungsformate konvertiert und ob alle (technischen) Versionen eines Dokumentes an das Archiv mit übergeben werden sollen.
Im Rahmen der Tests ist unter anderem aufgefallen, dass PDV bei der Dateibezeichnung der xdomea-Container und -Nachrichten vom Standard abweicht, um den häufig auftretenden Problemen mit zu langen Speicherpfaden in Windows-Systemen zu begegnen. Statt *UUID_Aussonderung.Aussonderung.0503* wird eine verkürzte Dateibezeichnung *UUID_0503.zip* verwendet. Da die Nummer des xdomea-Nachrichtentyps (bspw. 0503) eindeutig ist, ist der vollständige Nachrichtenname nicht notwendig, um den Nachrichtentyp zuordnen zu können. Um dem Problem langfristig zu entsprechen, soll ein Änderungsantrag zu den Dateibezeichnungen bei der AG xdomea gestellt werden.

Geplante Weiterentwicklung

X-man wird vom Landesarchiv Thüringen kontinuierlich weiterentwickelt. Seit Frühjahr 2024 wird das Entwicklungsprojekt von einer archivinternen Arbeitsgruppe unterstützt, der neben der Fach-IT des Landesarchivs auch Archivarinnen und Archivare aus dem Bereich der Überlieferungsbildung angehören. Die Arbeitsgruppe soll auch nach Produktivsetzung der E-Akten-Aussonderung der Thüringer Landesverwaltung die Anforderungsumsetzung und Weiterentwicklung des Programms fachlich begleiten.
Für die nächste Ausbaustufe des Programms ist neben einer Schnittstelle zu dem im Landesarchiv Thüringen eingesetzten Archivinformationssystem AUGIAS-Archiv zur automatischen Lieferung von Metadaten für die Verzeichnung auch die Implementierung der neuesten xdomea-Version 4.0.0 vorgesehen. Darüber hinaus wird x-man 2025 für die Verarbeitung von Austauschnachrichten im XJustiz-Standard erweitert, um auch die medienbruchfreie Anbietung, Bewertung und Übernahme elektronischer Gerichtsakten zu ermöglichen. Aus dem xdomea-Aussonderungsmanager wird dann ein XÖV-Aussonderungsmanager.

Kostenfreie Nutzung von x-man

Gemäß § 4 Abs. 3 ThürEGovG ist der Quellcode neuer Software, die von der öffentlichen Verwaltung oder speziell für diese entwickelt wurde, unter einer geeigneten Freie-Software- und Open-Source-Lizenz zu stellen und zu veröffentlichen, sofern keine sicherheitsrelevanten Aspekte entgegenstehen. X-man wurde daher im Mai 2024 zur kostenfreien Verwendung für andere Archive und Einrichtungen freigegeben. Die Open-Source-Lizenz GNU General Public License Version 3 (GPLv3) ermöglicht nutzenden Einrichtungen auch eine selbstständige Anpassung und Weiterentwicklung des Programms. Die Server-Komponente von x-man wird in

einer Docker-Umgebung auf einem Linux-System betrieben und mit Umgebungsvariablen konfiguriert. Installation und Betrieb des Programms sind daher in einem professionellen IT-Betrieb, bspw. in einem Rechenzentrum vorgesehen. Nähere Informationen zu x-man, der Quellcode des Programms und Hinweise zur Installation und Benutzung sind im GitHub des Landesarchivs Thüringen unter https://github.com/Landesarchiv-Thueringen/x-man zugänglich.

Fachliche, technische und betriebliche Fragen zu x-man können an das Team des Digitalen Magazins im Landesarchiv Thüringen unter digitales.magazin@la.thueringen.de gerichtet werden.

Datenbankarchivierung in der Tschechischen Republik

Martin Rechtorik

Einführung

Die Datenbankarchivierung ist ohne Frage eine absolute Notwendigkeit für unsere Gegenwart und Zukunft.[1] Es muss immer zuerst definiert werden, um welche Art von Datenbank und welche Art von Archivierung es sich handelt, denn die Archivierung von Datenbanken kann auf viele Arten erfolgen und hängt immer davon ab, welche Informationen in der Datenbank archivierungswürdig sind und daher archiviert werden sollten. Aus diesem Grund gliedert sich dieser Artikel in drei Teile und jeder Teil stellt spezifische Methoden und Werkzeuge zur Verwaltung der Unterlagen und Dateien vor, mit denen in den letzten Jahren praktische Erfahrungen gesammelt wurden. Es ist wichtig, einen gut durchdachten Plan zu haben, um sicherzustellen, dass alle Informationen ordnungsgemäß gespeichert, verwaltet und bei Bedarf abgerufen werden können (Databases for 2080, 2022).

Dokumentenmanagementsysteme

In der Tschechischen Republik sind Dokumentenmanagementsysteme (weiter nur DMS) seit 2012 für die öffentliche Verwaltung verpflichtend. Sie basieren auf mehreren bindenden Grundsätzen. Erstens ist es Klassifizierung, dies ist eine obligatorische Funktion von DMS. Sie ermöglicht die strukturierte Ablage und Verwaltung von Unterlagen. Zweitens gibt es eine Pflicht zur Aktenbildung. Jedes DMS muss die Erstellung von Akten unterstützen. Die dritte Pflicht ist eine Erstellung von SIP-Paketen: Die Erstellung von SIP-Paketen ist eine Aufgabe, die DMS übernehmen. Archive validieren nur obligatorische Inhalte. Der letzte Auftrag sind Metadaten und Verarbeitungsgeschichte: Jedes SIP-Paket muss Informationen über die Verarbeitung und Geschichte aus den Originalsystemen enthalten. Dies gewährleistet die Integrität und Nachvollziehbarkeit der Unterlagen. Die Einführung von DMS in der öffentlichen Verwaltung ist ein wichtiger Schritt zur Digitalisierung von Arbeitsprozessen und zur Verbesserung der Verwaltungseffizienz.

[1] Ein großes Dankeschön an meine geschätzte und freundliche Kollegin Elisabeth Klindworth für die Zeit, die sie mit dem Korrekturlesen und der Klärung von Fristen verbracht hat, an meinen Kollegen Zbyšek Stodůlka für seine Hilfe und Beratung bei der Vorbereitung der AUdS 2024, an meine Kollegin Isabel Taylor, die mir die Möglichkeit bot, an der AUdS teilzunehmen, und an meine Frau Zuzana, die mir bei den deutschen Übersetzungen sehr geholfen hat. Der Autor wurde unterstützt durch das Projekt SGS-2022-027: Anwendung von Mathematik und Informatik in der Geomatik.

Datenbanken als Informationssysteme

Im Laufe der Zeit erwiesen sich standardisierte DMS als unnötig aufwändig in der Verwaltung, sodass Informationen zunehmend in spezialisierten Fachverfahren verwaltet wurden. Diese sind eng auf eine oder wenige Arten von Agenden ausgerichtet und bilden dann nur wenige Agenden im Aktenplan ab. Die Informationen werden in diesen Systemen benutzerfreundlicher verarbeitet und nur die endgültigen Entscheidungen werden an das DMS zurückgegeben. In zunehmendem Maße werden DMS nur noch als Poststelle genutzt. Fachverfahren als auf bestimmte Aufgaben spezialisierte Datenbanken sind hingegen intuitiver und effizienter geworden und ihre Anzahl wächst stetig, da immer mehr Behörden und Organisationen die Vorteile der rein elektronischen Verwaltungsarbeit erkennen. In einer Datenbank wird jede Datei nur einmal vorgehalten, was bedeutet, dass Daten nur einmal eingegeben und dann für verschiedene Zwecke wiederverwendet werden können.

In Datenbanken gibt es keine „Unterlagen" im Sinne der traditionellen Vorstellung von Dokumenten auf Papier. Dokumente existieren oft lediglich als Eingangsdaten für die weitere Extraktion von Informationen. Stattdessen werden „Unterlagen" oder Datensätze in einer Datenbank auf der Grundlage einer Anfrage erstellt. Diese Anfragen können so einfach sein wie das Abrufen eines einzelnen Datensatzes oder so komplex wie das Zusammenführen mehrerer Tabellen und die Durchführung verschiedener Berechnungen (Kroenke, Auer, 2013).

Datenbanken von Ministerien und zentralen Behörden

Die Archivierung von Datenbanken ist in mehrfacher Hinsicht möglich, aber wir müssen akzeptieren, was die Behörden exportieren können oder wollen. Eine der Möglichkeiten ist ein SIARD Format zu nutzen. Dies ist eine hervorragende Lösung, die vor allem für Szenarien verwendet kann, in denen diese Art von Datenbanken wie Register funktionieren und in denen keine kontinuierliche Löschung von Daten erfolgt. Wir haben das SIARD Format für Schulregister, ARIS-System (Rechtorik, 2022) oder PEVA Software verwendet.

Das Schulregister

Das Schulregister besteht aus zwei Teil-Datenbanken. Der erste Teil ist das Register der Schulen und Bildungseinrichtungen, in dem alle schulischen Einrichtungen in der Tschechischen Republik eingetragen sind. Im zweiten Teil sind nur Informationen über Privatschulen und Privateinrichtungen eingetragen. Die Benutzeroberflächen zeigen für die Endnutzer der Datenbanken jedoch nur eine Auswahl der Daten an, die in der Datenbank enthalten sind. Eine vollständige Übersicht der in den Datenbanken gespeicherten Daten vermittelt das dahinterliegende

Datenmodell. Über die Benutzeroberflächen wird also lediglich ein Teil der vorhandenen Daten tatsächlich nutzbar gemacht.

Es handelt sich um ein zentrales System, das zu Beginn des neuen Jahrtausends im Auftrag des Bildungsministeriums entwickelt und im Jahr 2005 in Betrieb genommen wurde. Es erfasst alle Änderungen, die Schulen und Schuleinrichtungen betreffen. Die Eintragung ins Register kann als rechtliche Bestätigung der Existenz einer Einrichtung im Sinne des Bildungsgesetzes bezeichnet werden. Ohne Eintragung ins Schulregister können die im Bildungsgesetz beschriebenen Dienstleistungen in der Tschechischen Republik nicht erbracht werden.

Das Register selbst ist in den §§141 bis 159a des oben genannten Gesetzes geregelt:

a) Im Falle einer Schule oder Einrichtung: Schulart, Name, Sitz, Identifikationsnummer, Rechtsform, Abteilungskennzeichen der juristischen Person

b) Für den Gründer der Schule oder der Einrichtung: wie a)

c) Für eine schulische Rechtsperson (d. h. Privatschulen und -einrichtungen): wie a)

d) Liste der Unterrichtsfächer

e) Höchstzahl der Schüler und Studenten für Schulen, für Einrichtungen z. B. die Zahl der Betten in Unterkünften oder die Zahl der Gäste in Kantinen

f) Höchstzahl der Schüler oder Studenten, die in den einzelnen Bildungszweigen zugelassen sind

g) Angabe des Ortes, an dem der Unterricht stattfindet

h) Unterrichtssprache, falls anders als Tschechisch

i) Name, Vorname und Geburtsdatum des Schulleiters/Gründungsdatum der Schule oder Einrichtung

j) Name, Vorname und Wohnort des gesetzlichen Vertreters

k) Zeitraum, für den die Schule oder die Einrichtung eingerichtet ist

l) Datum der Eintragung und Datum der Aufnahme der Schule oder Einrichtung

m) Elektronische Adresse der Schule oder der Einrichtung

Zusätzlich zu den gesetzlich vorgeschriebenen Informationen enthält das Register oft noch andere Angaben, wie die Adresse einer Website, einen Identifikator für ein nationales System für Datenmeldungen (das Data-Box-Informationssystem) und möglicherweise andere zusätzliche Informationen, die von den Beamten als relevant eingestuft wurden.

Wie der vorangehende Text deutlich macht, handelt es sich um Informationen, die den Anforderungen des Archivgesetzes entsprechen. Daher kann das Register als zukünftiges Archivmaterial von außerordentlicher Bedeutung bezeichnet werden, da es den Forschenden einen

umfassenden Überblick über den Stand und die Entwicklung des Bildungswesens in der gesamten Tschechischen Republik zu Beginn des 21. Jahrhunderts gibt.

Abbildung 1: Das ursprüngliche Aussehen der Webanwendung

Abbildung 2: Das ursprüngliche Aussehen der Webanwendung

Für die Archivierung hat sich das Nationalarchiv für das SIARD-Archivformat entschieden, das sich als die am besten geeignete Option erwiesen hat, da das Informationssystem auf der Grundlage einer relationalen Microsoft SQL Server-Datenbank entwickelt wurde. Neben dem SIARD-Format ist auch die Archivierung mit CSV-Flat-Textdateien möglich, die Datenvalidierung ist jedoch technisch anspruchsvoll und zeitaufwändig. Für die Archivierung können auch als XML-Dateien geöffnete Exporte verwendet werden, die mehrmals wöchentlich und immer in aktueller Form veröffentlicht werden. Die Archivierung über diese Exporte ist jedoch nachweislich weniger komfortabel. Die offenen Daten sind zwar relativ informationsreich, aber die relationale Datenbank ist viel reichhaltiger und umfangreicher und müsste einen völlig unnötigen Datenzuwachs bewältigen.

Um die im SIARD-Paket archivierten Daten nutzen zu können, muss das Paket zunächst entpackt werden, da es sich um ein Zip-Paket handelt. Anschließend können die archivierten Daten in eine beliebige relationale Datenbank importiert oder einzelne XML-Tabellen mittels Parsing durchsucht werden. Die Archivierung im SIARD-Format ermöglicht somit nicht nur die Wiederholung der ursprünglichen SELECT-Abfragen und die Anzeige von Informationen, sondern auch die erneute Ausstellung und Bestätigung auf der Grundlage der in der neuen relationalen Umgebung gespeicherten Vorlagen. Diese werden in Datenfeldern im HTML-Format gespeichert. Eine zusätzliche Archivierung der Informationen aus der Schulregisterdatenbank als PDF oder in Papierform ist nicht notwendig. Eventuell vorhandene Papierakten werden vernichtet. Die vorhandenen Vorlagen lassen ferner vermuten, dass das Schulregister auf der Grundlage älterer, vermutlich nur interner Datenbank entwickelt wurde, da die Vorlagen auch als Grundlage für verschiedene Entscheidungen aus den 1990er-Jahren dienen können.

Auf Initiative des Nationalarchivs wurde die Funktionalität von Select-Abfragen hinzugefügt, so dass es möglich ist, auf verschiedene Weise nach Schulen zu recherchieren. Nach der erfolgreichen Übernahme ist es nötig, das spezielle Werkzeug für die Datenbankverarbeitung zu nutzen, in unserem Fall es ist dbDIPview. dbDIPview ist ein Programm, das vom Slowenischen Nationalarchiv entwickelt wurde (Domajnko, 2022). Ein Nutzer oder eine Nutzerin kann nicht nur die strukturierten Dateien zugänglich machen und in Form einer CSV-Datei herunterladen oder als Einrichtungs-Karte im PDF-Format ausdrucken. Dies ist vor allem dann nützlich, wenn im Archiv recherchiert werden soll, von wann bis wann die Schule ins Register eingetragen wurde oder über Ausbildungsprogramme, Fremdsprachen, Kapazität, die Person des Schulleiters oder der Schulleiterin. Der erste Snapshot wurde bereits 2012 übernommen, und als wir vor kurzem einen weiteren Snapshot anforderten, haben wir auch überprüft, dass alle Archiv-Metadaten in Form von Select-Abfragen gültig sind.

Popis zobrazení 1a: Výpis podrobností a historie školy/školského zařízení

Základní informace o názvu školy, sídle.
Pomocí sloupce "Obory" lze získat informace o studijních oborech (pouze školy) a pomocí "Historie" pak získat informace o historii školy/zařízení zachycenou v rejstříku.

Název školy	Právní forma	Sídlo	IČO	Obory	Historie
Křesťanské gymnázium	školská právnická osoba	Kozinova 1000, Praha 10 - Hostivař, 10200	60162961	9035BDEC-478F-4A33-9FC3-6C80FB641EF7	600006476

Kontaktní údaje
Sloupec Web obsahuje odkazy na webové stránky, které škola nahlásila do evidence, ty již nemusí být funkční. Z tohoto důvodu doporučujeme využít webarchiv Národní knihovny, který je dostupný zde<

Telefon	Fax	Web	E-mail 1	E-mail 2
271 750 632	271 750 632	www.krestanskegymnazium.cz	info@krestanskegymnazium.cz	
271 750 632	271 750 632	www.krestanskegymnazium.cz	reditel@krestanskegymnazium.cz	info@krestanskegymnazium.cz

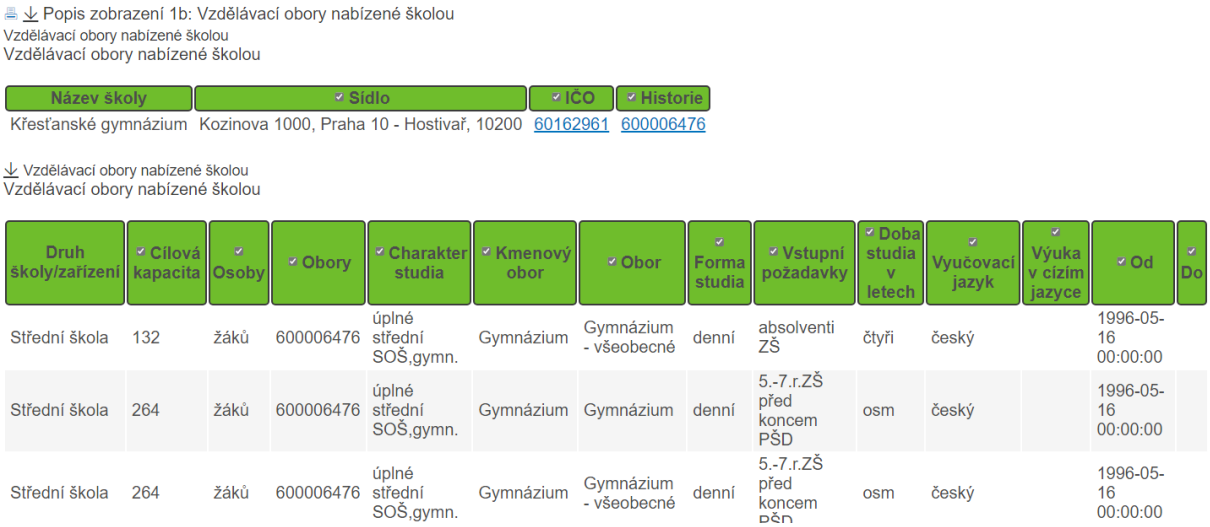
Školní zařízení
Detail zařízení provozovaných v rámci jedné právnické osoby.

Školní zařízení	Sídlo	Telefon	Fax
Středisko volného času - dům dětí a mládeže	Kozinova 1000, Praha 10 - Hostivař, 10200	271 750 632	271 750 632
Střední škola	Kozinova 1000, Praha 10 - Hostivař, 10200	271 750 632	271 750 632
Školní klub	Kozinova 1000, Praha 10 - Hostivař, 10200	271 750 632	271 750 632

Zřizovatel
Pomocí sloupců "Právnická osoba/Fyzická osoba" získá badatel podrobnosti ke zřizovateli školy.

Typ zřizovatele	Vazba - Právnická osoba	Vazba - Fyzická osoba
církev	34D0CF30-9CCE-4B15-8B2D-03F0A1C65E82	

Abbildung 3: Die Form der Ausgabe der Archiv-Webanwendung



Popis zobrazení 1b: Vzdělávací obory nabízené školou

Vzdělávací obory nabízené školou
Vzdělávací obory nabízené školou

Název školy	Sídlo	IČO	Historie
Křesťanské gymnázium	Kozinova 1000, Praha 10 - Hostivař, 10200	60162961	600006476

Vzdělávací obory nabízené školou
Vzdělávací obory nabízené školou

Druh školy/zařízení	Cílová kapacita	Osoby	Obory	Charakter studia	Kmenový obor	Obor	Forma studia	Vstupní požadavky	Doba studia v letech	Vyučovací jazyk	Výuka v cizím jazyce	Od	Do
Střední škola	132	žáků	600006476	úplné střední SOŠ,gymn.	Gymnázium	Gymnázium - všeobecné	denní	absolventi ZŠ	čtyři	český		1996-05-16 00:00:00	
Střední škola	264	žáků	600006476	úplné střední SOŠ,gymn.	Gymnázium	Gymnázium	denní	5.-7.r.ZŠ před koncem PŠD	osm	český		1996-05-16 00:00:00	
Střední škola	264	žáků	600006476	úplné střední SOŠ,gymn.	Gymnázium	Gymnázium - všeobecné	denní	5.-7.r.ZŠ před koncem PŠD	osm	český		1996-05-16 00:00:00	

Abbildung 4: Die Form der Ausgabe der Archiv-Webanwendung

Linked Data Systeme und Soziale Netzwerke

Im Jahr 2023 ist es schnell gegangen. Wir haben die Übernahme vom Zentralen Melderegister auf Interessenkollision getestet. Es handelt sich um Informationen in einigen CSV Format Dateien und auch ARES. Dies ist ein Informationssystem, das Dateien über Betriebe und Unternehmen enthält. Die aktuellen Dateien sind offene und werden täglich aktualisiert und herausgegeben in mehr als 1 Million XML Dateien. Aber es gibt nicht nur eine Datenbank, sondern mehrere, und keine von ihnen hat vollständige Datensätze. ARES ist ein Linked Data System von Daten über Dienste. Und ein drittes Beispiel – eine neue elektronische Sammlung von

Gesetzen. Mit dem Innenministerium haben wir vereinbart, dass jedes Jahr strukturierte Dateien in XML Format und Gesetze in PDF/A Format (PDF/A) übernommen werden.

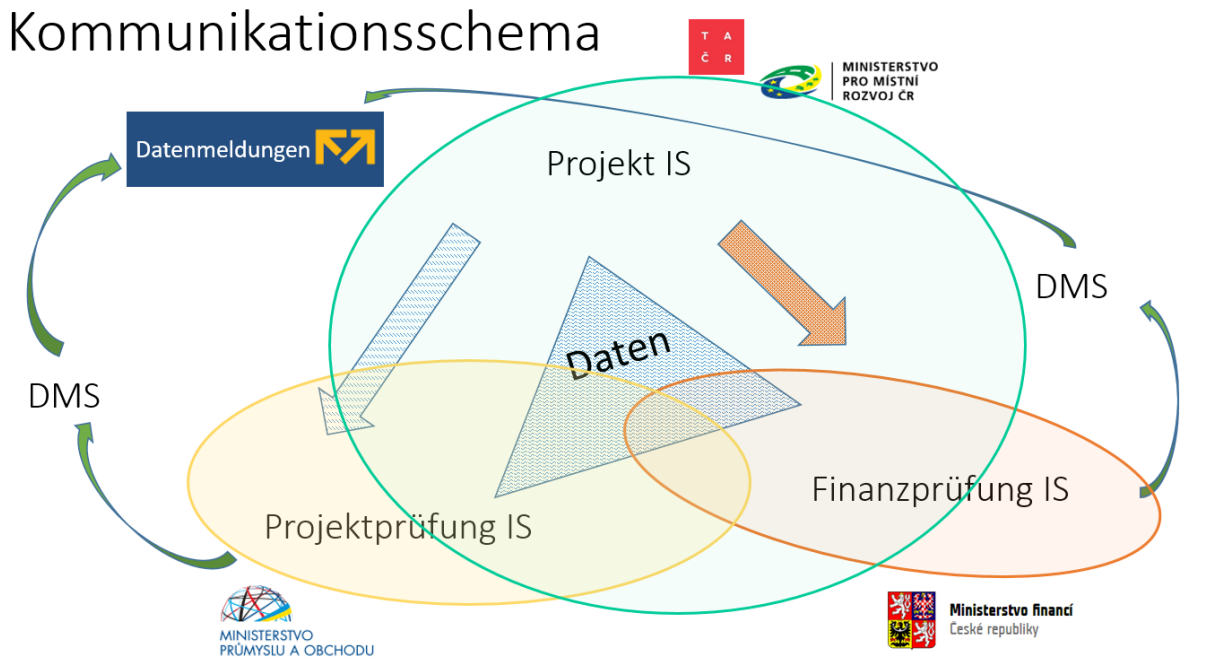


Abbildung 5: Einfaches Datenfluss-Kommunikationsschema zwischen Systemen mit Linked-Daten

Was also sollte eine digitale Archivarin oder Archivar in einem solchen Fall von Linked Data Systemen übernehmen? Wir können SIP-Pakete übernehmen, wir können ganze Datenbanken im SIARD-Dateiformat herunterladen. Zum Beispiel können wir nur statistische Informationen im CSV-Format exportieren oder Snapshots von Datenbanken erwerben. All das können wir heute tun, aber noch wichtiger ist, dass wir in der Lage sein müssen, diese Daten wieder zu verbinden, denn wenn sie nicht verbunden sind, haben sie nur einen geringen Archivierungswert.

Die Archivierung der detaillierten statistischen Daten ist immer noch Gegenstand von Debatten im Nationalarchiv. Angesichts der Möglichkeiten der künstlichen Intelligenz wird sie sich aber früher oder später als absolut notwendig erweisen. Wenn Statistiken auf der Grundlage realer Daten präzise erstellt wurden, können sie zur Schaffung eines neuen Datensatzes dienen, dem verschiedene Algorithmen zur Analyse vorgelegt werden können, ohne dass die Gefahr des Missbrauchs sensibler Informationen besteht. Auf diese Weise stellt beispielsweise das Gesundheitsministerium in Zusammenarbeit mit Universitäten Informationen für die Modellierung der zu erwartenden Entwicklung der Alzheimer-Krankheit oder anderer Krankheiten in der Bevölkerung bereit.

Neben Datenbanken haben wir auch Erfahrung mit Dateien aus einigen sozialen Netzwerken. In den Jahren 2021 bis 2023 haben wir fast 1 Million Tweets geharvestet. Dieses Thema hat Michael Held perfekt zusammengefasst (Held, 2022). Das Sammeln von Beiträgen auf Twitter (jetzt X) war eine äußerst interessante Erfahrung (Rechtorik, 2022a und 2023). Obwohl das Sammeln in der SQLite Datenbanken jetzt aus finanziellen Gründen eingestellt wurde, hat es uns in anderen Bereichen sehr geholfen – insbesondere bei der Überlegung, wie wir unsere Prozesse mehr und besser automatisieren können. Für die Archivierung wurde ein benutzerdefiniertes Tool entwickelt, das in der Lage ist, Daten aus der SQLite-Datenbank in das CSV-Format zu migrieren, die erforderlichen Metadaten hinzuzufügen und so ein DDVExt-Paket für das Tool dbDIPview zu erstellen. In Verbindung mit anderen Tools haben wir z. B. ein Harvesting der Konten der Kandidaten für die Präsidentschaftswahlen 2023 durchgeführt.
Der wertvollste Teil sind eigentlich die Fotoalben von verschiedenen offiziellen und inoffiziellen Aktivitäten der Behörden, die datiert, beschrieben, im JPEG-Format und in einer für den LCD-Bildschirm geeigneten Qualität vorliegen. Diese Dateien werden niemals in DMS geladen.

Tweet in detail

ID: 1102851589855948802
Datum: 2019-03-05 08:41:23
Tweet: If you have anything interesting to share in the field of #information_governance and/or #digital_preservation, submit your presentation proposal for the next #DLM_Forum AGM in Bern (21-22 May) by Friday, 8 March 2019! More information on our website: https://t.co/tblIEKAy9E https://t.co/kOio32lm3l
Hashtag: information_governance; digital_preservation; DLM_Forum

Media Content

Media:

Abbildung 6: Tweet im Archiv: Form der Ausgabe durch die Archiv-Webanwendung

Schulunterlagen

Auch im Bildungswesen können wir heute auf Datenbanken treffen. Aktuell verwenden mehr als 95% der Schulen in der Tschechischen Republik ein Informationssystem, und praktisch alle Dateien werden nur digital gespeichert, ob es sich um Noten oder entschuldigte Fehlzeiten handelt. Wir erhalten diese Informationen sowohl vom Bildungsministerium als auch vom Systemanbieter. Die Schulen sind das Hauptarchivgut für Bezirksarchive, denn es besteht eine große Nachfrage nach dieser Art von Archivmaterial. Was früher an Schulunterlagen auf Papier festgehalten wurde, muss heute als Datei exportiert werden. Wir müssen nur noch sagen, welche Dateien, wie oft und in welchem Format archiviert werden müssen.

2022 wurde eine detaillierte archivübergreifende Untersuchung der als Archivmaterial ausgewählten Dokumente durchgeführt. Auf Grundlage der so gewonnenen Informationen wurde eine Methodik für die Archivverwaltung des Innenministeriums mit einem Vorschlag zur Archivierung von Daten aus Informationssystemen, die in Schulen verwendet werden, ausgearbeitet, so dass diese Informationen, die typischerweise in Archiven stark nachgefragt werden, in Zukunft zur Verfügung stehen werden. Es stellte sich unter anderem heraus, dass die Archivierung der Schulunterlagen auf Papier zu teuer ist, nur schon Archivboxen sind teurer als Hardware. In der Praxis gibt es aber eigentlich keine Möglichkeit, eine so große Anzahl von Schulen mit einem Projektansatz zu bearbeiten. Es gibt einfach zu viele Schulen.

Aus diesem Grund haben wir eine Umfrage der verwendeten Systeme gemacht, die zwei Teile umfasste. Im Rahmen des ersten Teils besuchten wir eine Reihe von Schulen, um die Funktionsweise und die Datenexportmöglichkeiten der verwendeten Systeme zu verstehen, und wir befragten auch Archive, um herauszufinden, welche Art von Unterlagen in der Praxis am häufigsten als Archivmaterial ausgewählt werden. Eine positive Erkenntnis war nicht nur, dass die verwendeten Systeme den Export von Informationen ermöglichen, die seit mehr als 200 Jahren üblich sind, sondern auch, dass die Praxis der Auswahl von Archivgut in der Tschechischen Republik praktisch einheitlich ist. So wählt die überwiegende Mehrheit der Archivare stets die in Anlage Nr. 2 des Archivgesetzes, Artikel 17, gekennzeichneten Unterlagen als Archivgut aus. Darüber hinaus werden nur Stichproben ausgewählt, und zwar nicht systematisch. Besonders erwähnenswert ist ein Archiv, das die Schulen zwingt, Druckvorlagen bei Schreibwarenhändlern zu kaufen und die Informationen von Hand mit einem Stift in Informationsqualität für Archivzwecke abzuschreiben. Ich war von dieser Information ziemlich überrascht, und als ich sicherheitshalber telefonisch im Frühling 2022 nachgefragt habe, hat mir eine sehr nette Kollegin fröhlich geantwortet: „Da ist die Welt noch in Ordnung." Es kann nur hinzugefügt werden, dass die Zukunft zeigen wird, ob dies eine kluge Lösung oder eine völlig unsinnige Anforderung

war. Die Ergebnisse dieses Teils wurden in Form eines Online-Workshops veröffentlicht, an dem praktisch alle für Schulen zuständigen Archive teilnahmen, d. h. fast 100 Archive. Dies war der Workshop mit der größten Beteiligung des Jahres 2022.

Auf der Grundlage der Umfrage und der Ergebnisse des Workshops wurde auf Anweisung des Innenministeriums eine spezielle Arbeitsgruppe mit Mitgliedern aus verschiedenen Archiven gebildet, die einen Vorschlag für die Archivierung von Daten aus Schulsystemen in digitaler Form ausarbeiten sollte. Die Arbeit gliederte sich in 1) Verfahren im Umfeld des Nationalen Archivportals und 2) detaillierte Erhebungen des Dateienexports von Schulinformationssystemen, die überprüft wurden und wobei festgestellt wurde, dass viele von ihnen genau die Informationen enthalten, die seit über 200 Jahren auf Papier niedergeschrieben sind, wobei der einzige große Unterschied in der Anzahl der Seiten besteht. Während früher nur zwei Seiten, d. h. ein Blatt, ausgereicht haben, werden heute ähnliche Informationen auf vier bis acht Blätter gedruckt. Wenn also der Schuldrucker vergisst, beidseitig zu drucken, füllen sich die Archive achtmal schneller als heute. Als letztes war es die wissenschaftliche Nutzung von Archivgut im Hinblick auf die „Designated Communities" und die Bedürfnisse der staatlichen Verwaltungsorgane.

Anhand dieser thematischen Aufteilung wurde eine Methodik ausgearbeitet, die nun zur Genehmigung durch das Innen- und das Bildungsministerium ansteht. Obwohl die Methodik mit der Empfehlung schließt, die exportierten Informationen im PDF/A-Format zu archivieren, wird für die Zukunft eine Umstellung auf das XML-Format erwartet, da die Daten in diesem Format wesentlich kleiner sind und kaum migriert werden müssen. Das größte Problem ist jetzt der Verwaltungsaufwand, der unser Ziel beschränkt, das ganze Vorgehen völlig zu automatisieren. Die Schulen sollen jedes Jahr eine PDF/A-Datei mit den Informationen des Vorjahres ans Archiv übergeben, damit ist das Problem gelöst, aber dafür muss die Automatisierung solch trivialer Prozesse gesetzlich verankert werden.

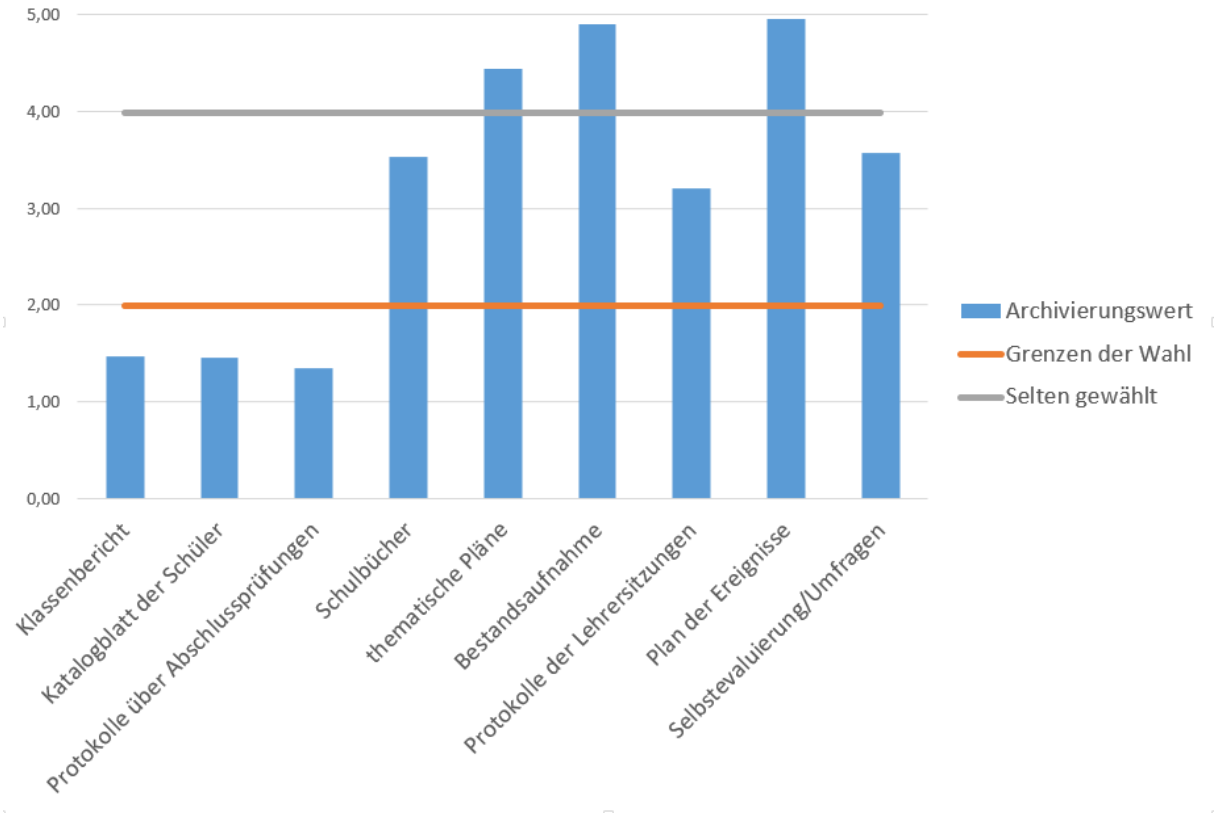


Abbildung 7: Die Grenzen für die Auswahl von Archivgut der Schulunterlagen. Alle Schulunterlagen unterhalb der orangefarbenen Achse werden immer ausgewählt, die unterhalb der grauen Achse werden selten ausgewählt.

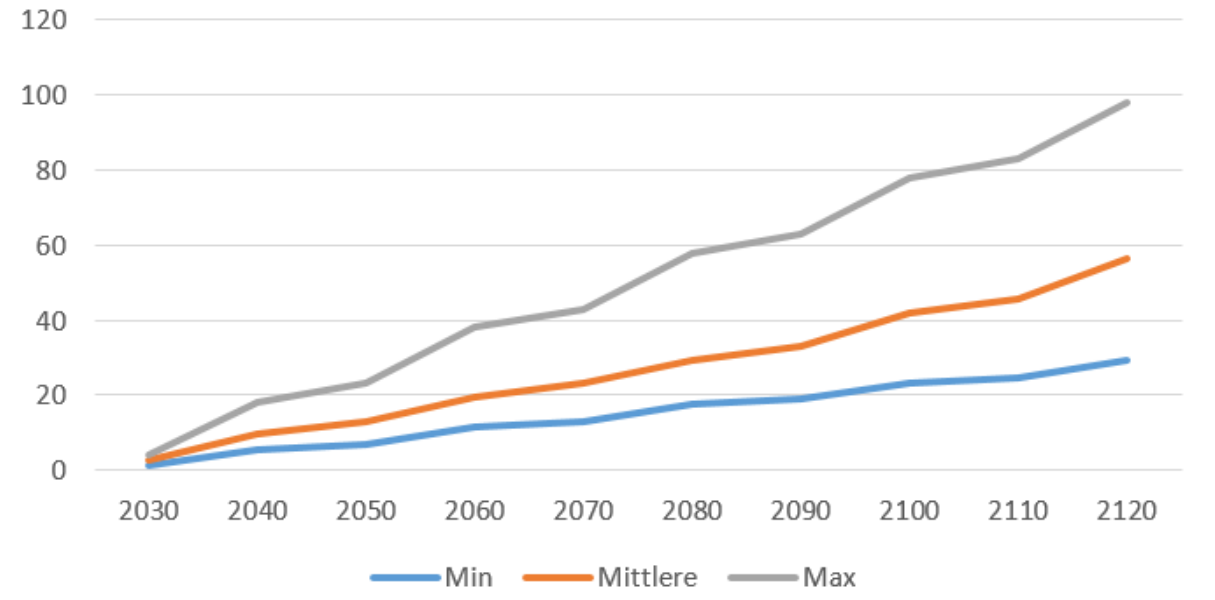


Abbildung 8: Datenwachstum für Schulunterlagen in der Tschechischen Republik. Die Prognose basiert auf Stichproben aus den Systemen und langfristigen Schülerzahlen aus der amtlichen Statistik.

Zusammenfassung

Dieser breit angelegte Ansatz für die Auswahl von Datenbanken mag auf den ersten Blick fehlerhaft erscheinen, aber die derzeitige Unterstützung durch den Gesetzgeber in der Tschechischen Republik ist nicht ausreichend. Andererseits zeichnet sich immer deutlicher ab, dass die SIARD-Archivalien ohnehin durchsucht werden, wobei den Forschern mit Hilfe der Graphdatenbank interessante Informationen in Form von Linked Data präsentiert werden (Lund, 2023). Außerdem gibt es, wie Trevor Owens (2018) beschreibt, keinen einzigen richtigen Ansatz, sondern das Ziel kann auf unterschiedliche Weise erreicht werden. Ebenso interessant ist seine Diskussion mit Chris Adams und Richard Lehane, in der Trevor Owens darauf hinweist, dass die Archivierung ganzer Datenbanken unweigerlich zur unnötigen Archivierung ungültiger Daten führt, und dass die vorgeschlagene Option von Richard Lehane der Verwendung von Open-Source-SQLite (Digital Preservation Q&A, 2014) angesichts der Beschränkungen dieser Datenbank und der darin verwendeten dynamischen Datentypen fast nicht realisierbar ist.

In jedem Fall sollte betont werden, dass der allgemein verwendete Begriff „Datenbankarchivierung“ ungenau ist. In den allermeisten Fällen geht es bei der Archivierung nur um die Archivierung von Daten aus einer Datenbank, es sei denn, die ursprüngliche Anwendung wird aus einem besonderen Grund ebenfalls archiviert, was aber ebenfalls ein Werkzeug für eine spätere Emulation erfordert. Dies bedeutet jedoch nicht, dass die Archivierung der Daten und die damit verbundene Archivierung der verwendeten Logik und der Interpretation der Informationen durch die Anwendung schlecht ist. Im Gegenteil, wenn wir wissen, dass nur die Daten und die Logik Gegenstand der Archivierung sind, dann ist die Verwendung verschiedener Methoden wie CSV, XML oder SIARD, ergänzt durch Screenshots der Originalanwendung, der richtige Ansatz, da Originalanwendungen oft mit Lizenzgebühren, proprietärer Software und vielen anderen Problemen verbunden sind.

Eine der schwierigsten Fragen ist, wer die „Designated Communities“ für archivierte Datenbanken sein wird. Auf jeden Fall können wir unterschiedliche Niveaus ihrer Computerkenntnisse erwarten. Die größte uns bekannte Gruppe stellen Forscherinnen und Forscher dar, die Mehrheit Menschen, die auf einfache Weise Informationen finden wollen. Darüber hinaus gibt es Forschende, die tiefer in die Materie eintauchen und die Datensätze genauer analysieren. Was mich betrifft, so ist das immer noch unsere Komfortzone, aber es ist offensichtlich, dass wir von unseren Nutzerinnen und Nutzern neue Niveaus erwarten müssen. Ich spreche von Datenbankexperten, die archivierte Daten auf eine andere Art und Weise nutzen wollen. Nicht nur, um darin zu suchen, sondern um neue Informationen zu erhalten, um Daten mit Graphdatenbanken zu verbinden oder neue Spezialkarten erstellen, usw. Und das ist noch nicht alles. Die

höchste Stufe stellen die Werkzeuge der künstlichen Intelligenz dar, die Sandboxen abfragen, detaillierte Statistiken extrahieren und synthetische Daten erstellen. Das wird eine Herausforderung sein.

Die Datenbankarchivierung ist entscheidend für die Zukunft, aber die SIARD Datei allein wird nicht ausreichen. Die Bedeutung offener Daten, die auf Interoperabilität und digitaler Archivierung beruhen, darf nicht unterschätzt werden. Es ist wichtig, die Gesamtergebnisse nicht zu vergessen, da die Wiederherstellung dieser Ergebnisse mit einem hohen Zeit- und Technologieaufwand verbunden sein kann. Die Entwicklung von Werkzeugen der künstlichen Intelligenz bietet ein großes Potenzial, aber digitale Archive werden auch in Zukunft mit Problemen konfrontiert sein, die durch den uninformierten Benutzer verursacht werden, der ein Mensch ist und bleiben wird.

Bibliografie

Archivgesetz, https://www.e-sbirka.cz/sb/2004/499/2022-04-27?f=t%C5%99%C3%ADdn%C3%AD%20v%C3%BD&zalozka=text (08.07.2024).

ARES-System. https://ares.gov.cz/ (08.07.2024).

CSV-Flat-Textdateien. https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml (Zugriff 08.07.2024); http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=45; (08.07.2024); http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1600 (08.07.2024).

Data-Box-Informationssystem. https://www.e-sbirka.cz/sb/2008/300?zalozka=text (08.07.2024).

dbDIPview, Database DIP viewer. https://github.com/dbdipview/dbdipview/ (08.07.2024).

DDVExt-paket. https://github.com/dbdipview/dbdipview/tree/master/testing/TestAndDemo2 (08.07.2024).

Digital Preservation Q&A (2014). https://qanda.digipres.org/24/should-keep-database-content-original-format-export-flaten (08.07.2024).

Domajnko, Boris (2022), *Database archiving with dbDIPview: A brief overview. Databases for 2080 Workshop*, urn:nbn:de:101:1-2022071903. S. 47–49. https://nbn-resolving.org/urn:nbn:de:101:1-2022071903 (08.07.2024).

Elektronische Sammlung von Gesetzen. https://www.e-sbirka.cz/ (08.07.2024).

Held, Michael (2022), „Tweets in Archiv. Herausforderungen einer (post-)modernen Bewertung“, *Scrinium*, 76: 122–143.

Kroenke, D., Auer, D., J. (2013), *Database Concepts*. 6th Edition. London: Pearson Education.

Lund, Audun (2023), *SIARD Suite Presentation. Das Schweizerische Bundesarchiv (BAR)*. https://www.dlmforum.eu/index.php/all-events/143-webinars-2023 - webinar6 (08.07.2024).

Melderegister. https://cro.justice.cz/

Olson, Jack (2009), *Database Archiving. How to keep lots of data for a very long time*. Berkeley: Elsevier USA.

Owens, Trevor (2018), *The Theory and Craft of Digital Preservation*. Baltimore: Johns Hopkins University Press.

PEVA. Die Datenbank Archivbestände und Sammlungen in der Tschechischen Republik. pro Archiváře – NARP (nacr.cz) (08.07.2024).

PDF/A. https://www.pdflib.com/pdf-knowledge-base/pdfa/the-pdfa-standards/?gclid=EAIaIQobChMIr_Sgtqrv8gIVE5_VCh3zfAzhEAMYASAAEgKiOvD_BwE; https://www.loc.gov/preservation/digital/formats/fdd/fdd000030.shtml ; https://kost-ceco.ch/cms/pdf-a-2.html ; https://kost-ceco.ch/cms/pdf-a-1.html ; kost-ceco | PDF/A-3 (08.07.2024).

Rechtorik, Martin (2022), „DB archiving nad use at Czech authorities“. In: *Databases for 2080 Workshop*, urn:nbn:de:101:1-2022071903, S. 44–46. https://nbn-resolving.org/urn:nbn:de:101:1-2022071903 (08.07.2024).

Rechtorik, Martin (2022a), *Harvesting data from Twitter, 20. April, 2022*. DLM Forum - Webinars in 2022 (08.07.2024).

Rechtorik, Martin (2023), *From Twitter to SIP/DIP*. DLM Forum - DLM Forum Members' Meeting in Ljubljana, 10-11 May 2023 (08.07.2024).

SQLite. https://www.loc.gov/preservation/digital/formats/fdd/fdd000461.shtml; GitHub - Digitalia-Xamk/Twitter-study: Collect tweets to sqlite (08.07.2024).
Schulregister. https://rejstriky.msmt.cz/rejskol/, Rechtsvorschriften: 561/2004 Sb., 1. 1. 2024, aktuální znění, informativní znění systému e-Sbírka (e-sbirka.cz) (08.07.2024).
SIARD format. https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml; http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2006; lhttps://kost-ceco.ch/cms/siard_de.html; https://github.com/DILCISBoard/SIARD; https://github.com/keeps/dbptk-developer/releases; https://github.com/sfa-siard/SiardGui/releases (08.07.2024).
Tschechischer nationaler Standard für DMS, Submission Information Packages. Národní standard pro elektronické systémy spisové služby - Ministerstvo vnitra České republiky (mvcr.cz) (08.07.2024).
Workshop: Databases for 2080 – Preserving database content for the long term. https://www.landesarchiv-bw.de/de/aktuelles/termine/72973 (08.07.2024).
XML, eXtensible Markup language file. https://www.loc.gov/preservation/digital/formats/fdd/fdd000075.shtml; https://www.nationalarchives.gov.uk/PRONOM/fmt/101, (08.07.2024).

VI.

DATENAUFBEREITUNG, AUTOMATISIERUNG

Kenne deine Daten:
Wie frei verfügbare KI-Modelle bei der Analyse von großen Datenmengen die Erschließung unterstützen können

Martin Vogel

Die Entwicklung von Künstlicher Intelligenz (KI) hat in den letzten Jahren einen starken Aufschwung mit einer schnell wachsenden Anzahl von KI-Modellen für unterschiedliche Bereiche erlebt. Dieses Papier beschreibt, wie Metadaten für die archivische Bewertung und Erschließung von KI-Modellen aus den Bereichen Audio, Computer Vision und Natural Language Processing erzeugt werden können.

Die Anzahl der verfügbaren KI-Modelle steigt ständig an. Laut der Webseite Hugging Face (Hugging Face, 2024b) sind aktuell 1'012'232 KI-Modelle verfügbar. Diese KI-Modelle können für eine Vielzahl von Anwendungsfällen verwendet werden, wobei die Ergebnisse stark von den verwendeten Daten, die analysiert werden sollen, abhängt.

In diesem Papier werden Modelle aus den Bereichen Audio, Computer Vision und Natural Language Processing untersucht. Insbesondere werden folgende Anwendungsfälle betrachtet:

- Bilderkennung (Objekterkennung / Segmentierung)
- Transkription
- Detektieren von Namen, Personen und Orten
- Erstellen von Inhaltsangaben für Dokumente

Ausgewählte Bereiche und verwendete KI-Modelle

Abb. 1 zeigt die in diesem Papier betrachteten Bereiche, um die Anwendungsfälle durchzuführen. Um die Analysen durchzuführen, wurden die Modelle facebook/detr-resnet-50 (Hugging Face, AI at Meta 2024a), nvidia/segformer-b0-finetuned-ade-512-512 *(*Hugging Face, NVIDIA 2024), OpenAI / Whisper (large) *(*OpenAI, Whisper 2024), Spacy / de_core_news_lg *(*Hugging Face, spaCy 2024) und *Falconsai/text_summarization (*Hugging Face, Falcons.ai 2024) verwendet.

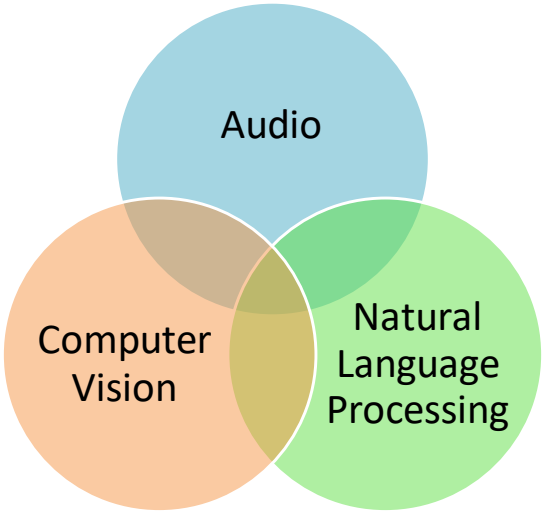


Abbildung 1: Ausgewählte Bereiche der KI-Modelle

Audio:

- Transkription: gesprochene Worte in Textform übersetzen

Computer-Vision:

- Objekterkennung: automatische Erkennung von Objekten in einem Bild
- Segmentierung: automatische Erkennung von Objekten und Regionen in einem Bild

Natural Language Processing:

- Zusammenfassung: automatische Zusammenfassung von Texten
- Token Classification: automatische Klassifizierung von Worten in einem Text

KI-Modelle auf unterschiedliche Datentypen anwenden

Die Analyse der Daten wurden auf folgender Testumgebung durchgeführt:

Testumgebung:

- 4 CPUs
- 24 GB RAM
- Keine Grafikkarte (GPU Unterstützung)
- Online: Hugging Face (Hugging Face 2024a)

Objekterkennung:

- Input: Bilddokumentation der Altstadt während der Sanierung (Niedersächsisches Landesarchiv, 2017)
- Modell: facebook/detr-resnet-50 (Hugging Face, AI at Meta 2024a)
- Output: Liste mit erkannten Objekten

Segmentierung:

- Input: Bilddokumentation der Altstadt während der Sanierung (Niedersächsisches Landesarchiv, 2017)
- Modell: nvidia/segformer-b0-finetuned-ade-512-512 *(*Hugging Face, NVIDIA, 2024)
- Output: Liste mit erkannten Objekten

Transkription:

- Input: „Warum Umweltschutz“ von Prof. Dr. Adolf Brauns vom 13.11.1975 (Niedersächsisches Landesarchiv, 2013)
- Modell: OpenAI / Whisper (large) (OpenAI, Whisper 2024)
- Output: Textdatei mit der Transkription

Natural Language Processing:

- Input: Ergebnis der Transkription (Textdatei)
- Modell: Spacy / de_core_news_lg *(*Hugging Face, spaCy 2024)
- Output: Liste mit erkannten Personen, Orte, Organisationen

Zusammenfassung:

- Input: Ergebnis der Transkription (Textdatei)
- Modell: Falconsai/text_summarization *(*Hugging Face, Falcons.ai 2024)
- Output: Textdatei mit einer Zusammenfassung der Transkription

Das Beispiel für die Segmentierung wurde online durchgeführt.

Objekterkennung in Bildern

Im folgenden Abschnitt wird untersucht, wie das KI-Modell (Hugging Face, AI at Meta, 2024a) Objekte in Bilddateien erkennen kann. Das Bild zeigt die Aufnahme eines Straßenzugs mit Gebäuden, mehreren Autos, einer Straße, Bürgersteigen und noch weiteren Objekten.

Abbildung 2: Bild eines Straßenzugs einer Stadt

Nach der Analyse wurden folgenden Objekte mit einer Wahrscheinlichkeit von > 90% gefunden:

Abbildung 3: Gefundene Objekte mit (Hugging Face, AI at Meta 2024a)

Detected truck with confidence 0.933 at location [490.53, 654.98, 696.96, 785.54] Detected car with confidence 0.99 at location [489.19, 654.63, 698.07, 784.66]

Das KI-Modell (Hugging Face, AI at Meta, 2024a) findet einen Truck (der nicht existiert) mit der Wahrscheinlichkeit von 93,3 % und ein Auto mit der Wahrscheinlichkeit von 99% an der

gleichen Stelle. Weitere Objekte (Gebäude, Straßen, etc.) werden nicht erkannt. Warum der Truck mit so einer hohen Wahrscheinlichkeit gefunden wurde, konnte nicht ermittelt werden.

Segmentierung von Bildern

Das gleiche Bild wird nun mit dem KI-Modell (Hugging Face, NVIDIA 2024) analysiert. Die Segmentierung des Bildes liefert detailliertere Ergebnisse als die reine Objekterkennung mit (Hugging Face, AI at Meta, 2024a).

Abbildung 4: Segmentierung des Bildes mit (Hugging Face, NVIDIA, 2024)

Abbildung 5: Ergebnisse der Segmentierung mit (Hugging Face, NVIDIA, 2024)

Das Auto wird auch in diesem Modell erkannt. Zusätzlich werden noch Gebäude, Himmel, und weitere Objekte erkannt. Der Truck, welcher fehlerhaft bei der Objekterkennung detektiert wurde, wird bei der Segmentierung nicht erkannt. Mit diesem Ergebnis lässt sich der Inhalt des Bildes genauer beschreiben.

Vergleich der Objekterkennung und Segmentierung

Das verwendete KI-Modell für die Objekterkennung (Hugging Face, AI at Meta, 2024a) ist darauf trainiert, Objekte zu finden. Aktuell kann es 91 Objekte (Hugging Face, AI at Meta, 2024b) erkennen. Es ist nicht darauf trainiert, zusammenhängende Flächen zu finden, welche Objekte enthalten können. Somit eignet sich dieses Modell nur dafür, Objekte innerhalb von Bildern zu erkennen. Das KI-Modell (Hugging Face, NVIDIA, 2024) dagegen verwendet einen anderen Ansatz. Es erkennt zusammenhängende Flächen (Segmente), separiert diese und kann diese dann mit den trainierten Objekten, die selbst wiederrum gelabelte Segmente sind, vergleichen. Dieses Beispiel zeigt, dass es sinnvoll ist, unterschiedliche KI-Modelle auf gleiche Daten anzuwenden und die Ergebnisse miteinander zu vergleichen.

Um einen ersten schnellen Überblick über seine Daten zu bekommen, eignet sich eine Zero Shot Analyse, d. h. die Daten werden mit unterschiedlichen ausgewählten KI-Modellen analysiert. Anhand der Analyseergebnisse werden im zweiten Schritt weitere KI-Modelle ausgewählt, die die Daten genauer und detaillierter auswerten können.

Aufbau einer Pipeline für unterschiedliche KI-Modelle

Im folgenden Beispiel (s. Abb. 6) soll gezeigt werden, wie eine Verkettung (Pipelining) von einzelnen KI-Modellen aufgebaut werden kann. Als Input wird (Niedersächsisches Landesarchiv, 2013) verwendet. Das Pipelining wurde in den Entwicklungsumgebung Jupyter Notebook (Jupyter, 2024) durchgeführt.

Transkription von Audiodateien

Als Input dient eine Audiodatei. Die Audiodatei wird mit dem KI-Modell (OpenAI, Whisper, 2024) transkribiert. Das Ergebnis der Transkription wird mit dem KI-Modell (Hugging Face, spaCy, 2024) auf Personen, Orte und Organisationen analysiert und gelabelt. Danach wird aus der Transkription mit dem Modell *(*Hugging Face, Falcons.ai, 2024) eine Zusammenfassung erzeugt.

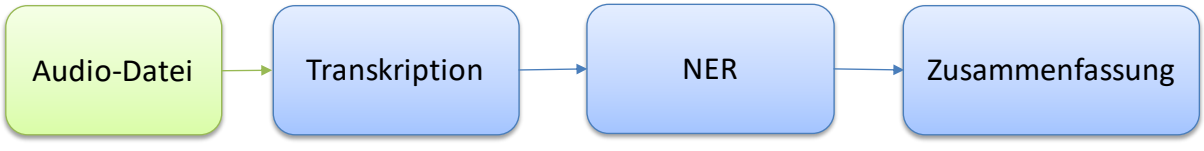


Abbildung 6: Aufbau der Pipeline für KI-Modelle

Der folgende Python Code führt in Jupyter Notebook die Transkription aus:

```
!pip install git+https://github.com/openai/whisper.git

from datetime import datetime
```

```
import whisper
print("Start", datetime.now())
base_model_medium = whisper.load_model("large")
whisper.DecodingOptions(fp16 = False, language = 'de')
result_medium = base_model_medium.transcribe("/home/jovyan/work/au-
dio/auds2024/wo_41_c_nds_zg._2013_088_nr._380.wav")
print("Ende", datetime.now())
print(result_medium["text"])
```

Auf der Testumgebung hat die Transkription des 34minütigen Beitrags ca. zwei Stunden gedauert. Das Ergebnis der Transkription (Auszug, insgesamt umfasst die Transkription 1466 Wörter) ist folgender Text:

„Im Jahre 1970 eröffnete das Senckenberg-Museum in Frankfurt am Main eine Umweltausstellung und brachte gleichzeitig eine Broschüre mit dem Titel Umwelt 2000 heraus. Eigentlich erst seit diesem Jahre wurde bei uns in der Bundesrepublik der Umweltschutz aktueller. Auf dem Büchermarkt in Mitteleuropa erschienen die verschiedensten Titel. Schutz unseres Lebensraumes, Vorträge an der Eidgenössischen Hochschule in Zürich im Jahre 1970, Belastete Landschaft, Gefährdete Umwelt, herausgegeben von Olschowi 1971, Die Erde hat keinen Notausgang von Dahmen 1971, Umwelt heute von Schwabe 1973. Und schließlich der ausgezeichnet bebildete Ökologieband in Jimmocks Tierleben mit einem Artikel. Abschnitt über die Umwelt des Menschen im Jahre 1973. In diesem Zusammenhang dürfen nicht unerwähnt bleiben …“

Die Qualität der Transkription ist in Ordnung und eignet sich für eine erste Analyse des Inhalts der Audio-Datei.

Erkennen von Personen, Orte und Organisationen mit Natural Language Processing

Das Ergebnis der Transkription wird durch das KI-Modell (Hugging Face, spaCy, 2024) auf Personen, Orte und Organisationen analysiert.

Der folgende Python Code führt die NER-Analyse in Jupyter Notebook durch:

```
!pip install textacy
!python -m spacy download de_core_news_lg

import spacy
from spacy import displacy
import srsly
from pathlib import Path

nlp = spacy.load("de_core_news_lg")
d1 = nlp(result_medium["text"])
displacy.render(d1, style="ent", jupyter=True)
```

Das Ergebnis der NER-Analyse ist folgendes:

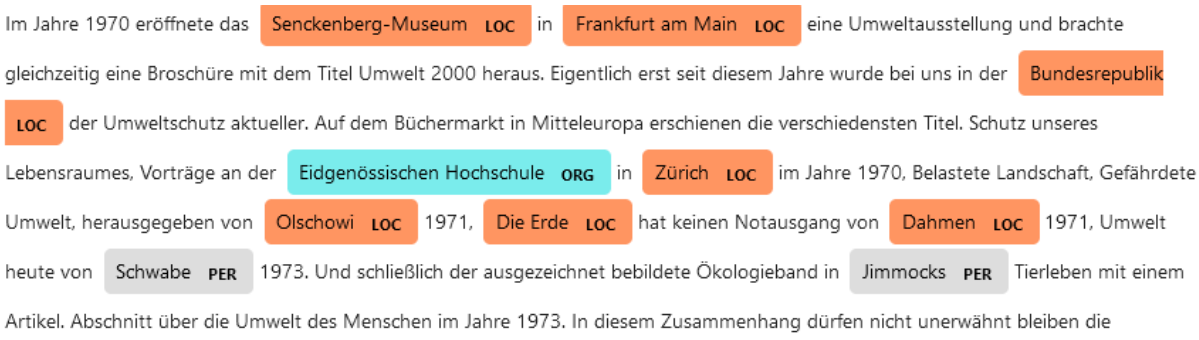


Abbildung 7: Ergebnisse der NER-Analyse

Es werden Personen (PER), Organisationen (ORG) und Orte (LOC) erkannt und grafisch in Jupyter Notebook dargestellt (Abb. 7).

Zusammenfassung von Texten

Auf das Ergebnis aus der Transkription wird das KI-Modell (Hugging Face, Falcons.ai, 2024) angewandt, um eine Zusammenfassung zu erzeugen.

Der folgende Python Code führt die-Zusammenfassung in Jupyter Notebook durch:

```
from transformers import pipeline

summarizer = pipeline("summarization", model="Falconsai/text_summarization")
print(summarizer(result_medium["text"]), max_length=1200, min_length=30,
do_sample=False))
```

Das Ergebnis der Zusammenfassung lautet:

„Im Jahre 1970 eröffnete das Senckenberg-Museum in Frankfurt am Main eine Umweltausstellung und brachte gleichzeitig eine Broschüre mit dem Titel Umwelt 2000 heraus. In diesem Zusammenhang dürfen nicht unerwähnt bleiben die Materialien zum Umweltprogramm der Bundesregierung 1971 und die Berichte des Club of Rome, die Grenzen des Wachstums 1972 und Menschheit am Wendepunkt 1974. Heute hat fast jede Tageszeitung dann und wann eine Seite mit der Überschrift Umweltschutz."

Durch die Verwendung mehrerer KI-Modelle in einer Pipeline können bessere Ergebnisse erzielt werden. Durch die Spezialisierung, die jedes Modells hat, können verbesserte Ergebnisse erzielt werden. Die Ergebnisse aus der Pipeline eignen sich für einen ersten Überblick des Inhalts der Audiodatei sowie der erkannten Personen, Orte und Organisationen.

Welche Online-Plattformen gibt es für KI-Modelle?

Es gibt mittlerweile eine Vielzahl von Anbietern für frei verfügbare KI-Modelle. Zwei der größten Anbieter sind die Plattformen Hugging Face (Hugging Face, 2024a) und Kaggle (Kaggle, 2024). Diese beiden Plattformen bieten eine Vielzahl von KI-Modellen mit Beispielen für Jupyter Notebook an. Die meisten KI-Modelle können auch kommerziell genutzt werden.

Anforderungen für den lokalen Einsatz von KI-Modellen

Heutzutage ist es nicht notwendig, teure Hardware zu beschaffen, um KI-Modelle lokal nutzen zu können. Ein handelsüblicher Laptop- oder Desktop-PC mit einer normalen NVIDIA Grafikkarte (8–12 GB VRAM) und 16 Gigabyte Arbeitsspeicher RAM reichen aus, um erste Tests durchzuführen und Erfahrungen zu sammeln. Auf diesen PCs wird die Software Docker Desktop (Docker, 2024a, 2024b) installiert, um eine isolierte Umgebung für die Ausführung von KI-Modellen zu erstellen. Die meisten KI-Modelle auf Hugging Face werden in sogenannten Jupyter Notebooks bereitgestellt. Diese Notebooks enthalten kleine Programmbeispiele in Python, die es ermöglichen, die KI-Modelle schnell und ohne Programmierkenntnisse zu nutzen.

Zusammenfassung und Ausblick

KI-Modelle können heute mit überschaubarem finanziellem Aufwand datenschutzkonform genutzt werden. Die Qualität der Modelle verbessert sich stetig und es kommen täglich neue Modelle hinzu. KI-Modelle aus den unterschiedlichen Bereichen können bei großen Datenmengen Metadaten über Daten generieren, die für die Bewertung und Erschließung genutzt werden können. Die Ergebnisse der eingesetzten KI-Modelle waren in Ordnung, um Metadaten für die analysierten Daten zu erstellen. Die Anwendungsfälle dienten zur Erkenntnisgewinnung, wie KI-Modelle angewandt werden können. Inwieweit die gewonnenen Metadaten für die Erschließung eingesetzt werden, muss daher weiter eruiert werden. Daher ist es zukünftig wichtig, weitere Forschung im archivischen Bereich durchzuführen, um die Genauigkeit, Qualität und Zuverlässigkeit der KI-Modelle weiter zu verbessern und die Ergebnisse für die Bewertung und Erschließung nutzen zu können.

Bibliografie

Docker, 2024a, *Docker: Accelerated Container Application Development*, https://www.docker.com/ (29.09.2024).

Docker, 2024b, *Docker Hub Container Image Library | App Containerization*, https://hub.docker.com (29.09.2024).

Hugging Face, 2024a, *The AI community building the future,* https://huggingface.co (29.09.2024).

Hugging Face, 2024b, *Models - Hugging Face*, https://huggingface.co/models (29.09.2024).

Hugging Face, AI at Meta, 2024a, *Modell facebook/detr-resnet-50*, https://huggingface.co/facebook/detr-resnet-50 (29.09.2024).

Hugging Face, AI at Meta, 2024b, *Modell facebook/detr-resnet-50*, https://huggingface.co/facebook/detr-resnet-50/blob/main/config.json (29.09.2024).
Hugging Face, Falcons.ai, 2024, *Modell: Falconsai/text_summarization*, https://huggingface.co/Falconsai/text_summarization (29.09.2024).
Hugging Face, NVIDIA, 2024, *Modell: nvidia/segformer-b0-finetuned-ade-512-512*, https://huggingface.co/nvidia/segformer-b0-finetuned-ade-512-512 (29.09.2024).
Hugging Face, spaCy, 2024, *Modell: Spacy / de_core_news_lg*, https://huggingface.co/spacy/de_core_news_lg (29.09.2024).
Jupyter, 2024, *jupyterhub/jupyterhub*, https://hub.docker.com/r/jupyterhub/jupyterhub (29.09.2024).
Kaggle, 2024, *Kaggle: Your Machine Learning and Data Science Community*, https://www.kaggle.com (29.09.2024).
Niedersächsisches Landesarchiv, NLA WO 41 C Nds Zg. 2013/088 Nr. 380 (2013): *Tonbandaufnahme „Warum Umweltschutz" von Prof. Dr. Adolf Brauns vom 13.11.1975* https://www.arcinsys.niedersachsen.de/arcinsys/detailAction.action?detailid=r11440416 (29.09.2024).
Niedersächsisches Landesarchiv, NLA AU Rep. 227/24 acc. 2017/23 Nr. 58 (2017): *Bilddokumentation der Altstadt während der Sanierung*, https://www.arcinsys.niedersachsen.de/arcinsys/detailAction.action?detailid=v10623371 (29.09.2024).
OpenAI, Whisper, 2024, *Modell: OpenAI / Whisper (large)*, https://github.com/openai/whisper (29.09.2024).

Automatisierte Tiefenerschließung von Digitalen Topographischen Karten

Antje Lengnik

Das Niedersächsische Landesarchiv (NLA) archiviert seit 2019 digitale Geobasisdaten der Landesvermessungsverwaltung. Die verschiedenen Produkte zeichnen sich unter anderem durch sehr große Datenmengen aus, die bewältigt werden müssen. Das betrifft nicht nur die Speicherung, sondern auch die archivische Erschließung, bei der die Möglichkeiten der klassischen händischen Verzeichnungsarbeit schnell an ihre Grenzen stoßen. Im folgenden Beitrag soll anhand der Übernahme der Digitalen Topographischen Karten eine Möglichkeit aufgezeigt werden, wie digitale Massendaten mithilfe von aufbereiteten Metadaten automatisiert tiefenerschlossen und leicht recherchierbar gemacht werden können.

Ausgangslage

Um was für Daten handelt es sich bei den Digitalen Topographischen Karten (DTK)? Die DTK sind ein Produkt aus der ATKIS-Familie (Amtliches Topographisch-Kartographisches Informationssystem) und stellen die digitale Fortführung der analogen topographischen Blattschnitte dar. Es handelt sich dabei um eine vereinfachte Darstellung der Erdoberfläche, bei der Objekte und Flächen durch Symbole und Farben repräsentiert sind. Als Rasterdaten werden sie aus dem Digitalen Landschaftsmodell (Basis-DLM) abgeleitet und in Form von Bildkacheln abgelegt. Je nach Maßstab zeigen diese einzelnen Bilddateien einen unterschiedlich großen Teil der Landesfläche. Ergänzt werden diese Bilder jeweils mit einer Textdatei, die die zugehörigen Koordinaten enthält (Abb. 1).

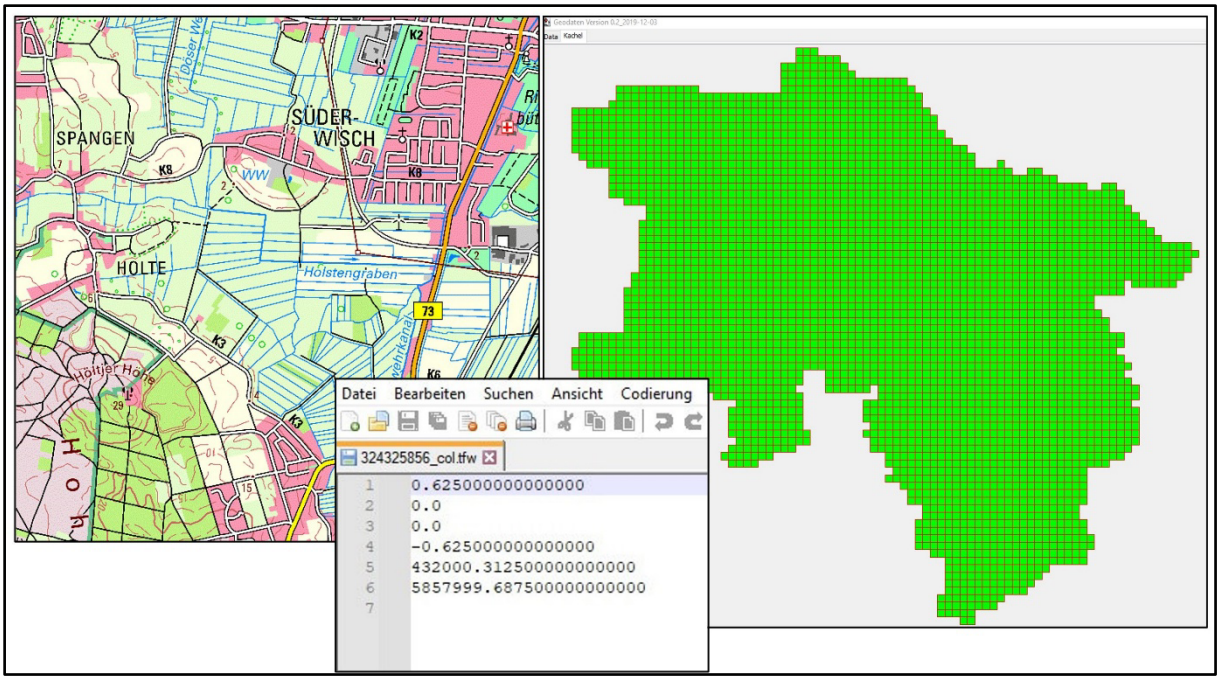


Abbildung 1

Das NLA orientiert sich bei der Archivierung von Geobasisdaten an den Leitlinien zur bundesweit einheitlichen Archivierung von Geobasisdaten, die erstmals 2015 von einer gemeinsamen Arbeitsgruppe aus Vertretenden der Landesarchive und der Vermessungsverwaltungen veröffentlicht und 2022 überarbeitet wurden. Auf dieser Grundlage wurden seit 2019 drei verschiedene Maßstäbe in drei Zeitschnitten archiviert, bis das Produkt DTK im Jahr 2023 von der Landesvermessungsverwaltung eingestellt wurde. Insgesamt handelt es sich dabei um ungefähr 35.000 Datensätze, die jeweils aus der Kartenkachel als tif-Datei und den Koordinaten als tfw-Datei bestehen. Verknüpft sind diese beiden einzelnen Dateien durch die so genannte Kachelnummer, eine eindeutige neunstellige Referenznummer, die sich aus den Koordinaten der Kartenkachel zusammensetzt und im Dateinamen enthalten ist. Diese Kachelnummer darf niemals verloren gehen, da sie der einzige Anhaltspunkt ist, einer tif-Datei die richtige tfw-Datei zuzuordnen. Abgesehen von dieser Nummer enthielten die Übernahmen zunächst aber keine weiteren Metadaten, die sich für die Erschließung verwenden ließen.

Tiefenerschließung

Nachdem die Daten 2019 erstmals ins Landesarchiv übernommen worden waren, mussten zunächst Überlegungen angestellt werden, wie die Daten am sinnvollsten zu Archivpaketen formiert werden. Für dieses Kartenprodukt wurde davon ausgegangen, dass Nutzende am häufigsten den Bedarf haben, einzelne Kacheln zu sichten. Eine Zusammenfassung aller Kacheln pro Maßstab und Zeitschnitt in einem AIP war daher nicht zielführend. Da die möglichen

Forschungsfragen an diese Quellen sehr breit gefächert sein können, half es für die Benutzung auch nicht, die Kartenkacheln nach Regionen in AIPs zusammenzufassen. Der einzige sinnvoll erscheinende Ansatz war daher die Erschließung jeder Kartenkachel als separates AIP. Für Forschungsfragen, die die Auswertung der gesamten Landesfläche beinhalten, steht zudem ein anderes Produkt der Landesvermessungsverwaltung, das Basis-DLM, zur Verfügung. Wie bereits erwähnt, wurden die Kacheln jedoch ohne weitere beschreibende Metadaten abgegeben, sodass wir auf diese Weise ca. 35.000 Verzeichnungsdatensätze erzeugt hätten, die sich nur in der Kachelnummer unterscheiden und ansonsten keine weiteren geographischen Metadaten enthalten (Abb. 2).

Abbildung 2

Es wäre für Nutzende damit unmöglich, die für sie relevanten Kacheln in unserer Recherchedatenbank ausfindig zu machen. Es musste also eine Lösung gefunden werden, die Erschließung so anzureichern, dass eine Suchmöglichkeit nach geographischen Namen besteht. Eine manuelle Tiefenerschließung war angesichts der Menge an Datensätzen nicht realisierbar, weshalb es eine technische Umsetzung brauchte. Die Idee war, softwaregestützt einen Abgleich der Kacheln mit Wohnorten und anderen geographischen Gebieten durchzuführen.

Es wurde erneut Kontakt zur Landesvermessungsverwaltung aufgenommen, um herauszufinden, ob weiteres Material vorhanden war, mit dem sich die Erschließung anreichern ließe. Letztendlich konnten folgende Daten zusätzlich gewonnen werden:

- Das Ortsverzeichnis von Niedersachsen mit allen aktuellen Orten und Wohnplätzen sowie sogenannten „Volkstümlichen Siedlungsnamen“ untergegangener Wohnplätze mit Georeferenzierung (verfügbar als Open Data).
- Die Verwaltungsgrenzen Niedersachsens als Shape-Dateien mit allen Gemeinde- und Gemarkungsgrenzen (verfügbar als Open Data).

- Das Kachelraster für jeden Maßstab als Shape-Datei mit einer Übersicht, welche Kartenkachel welches Gebiet abdeckt.
- Die niedersächsischen Nordseegebiete als Shape-Datei, in der Wattgebiete, Sandbänke, Leuchtfeuer und Leuchttürme als bounding boxes oder Punktkoordinaten georeferenziert sind.

All diese zusätzlichen Inhaltsinformationen versprachen eine sehr umfangreiche Tiefenerschließung jeder einzelnen Kartenkachel, sofern es gelang, diese miteinander zu verknüpfen. Zunächst wurde ein Konzept erstellt, wie der spätere Erschließungsdatensatz aussehen und welche Informationen dort aufgenommen werden sollten. Anschließend ging es darum, diese Verknüpfung über ein technisches Mapping herzustellen, um dem geographischen Gebiet jeder Kachel die richtigen Orte und Grenzen zuzuordnen.

Technische Umsetzung

Auf Grundlage der von der Landesvermessungsverwaltung zusätzlich gelieferten Daten entwickelte unser Fachinformatiker ein dazu passendes Tool. Darin wurden die einzelnen Metadaten-Layer über das Kachelraster gelegt und von der Software mithilfe der Georeferenzierung Schnittmengen erkannt (Abb. 3).

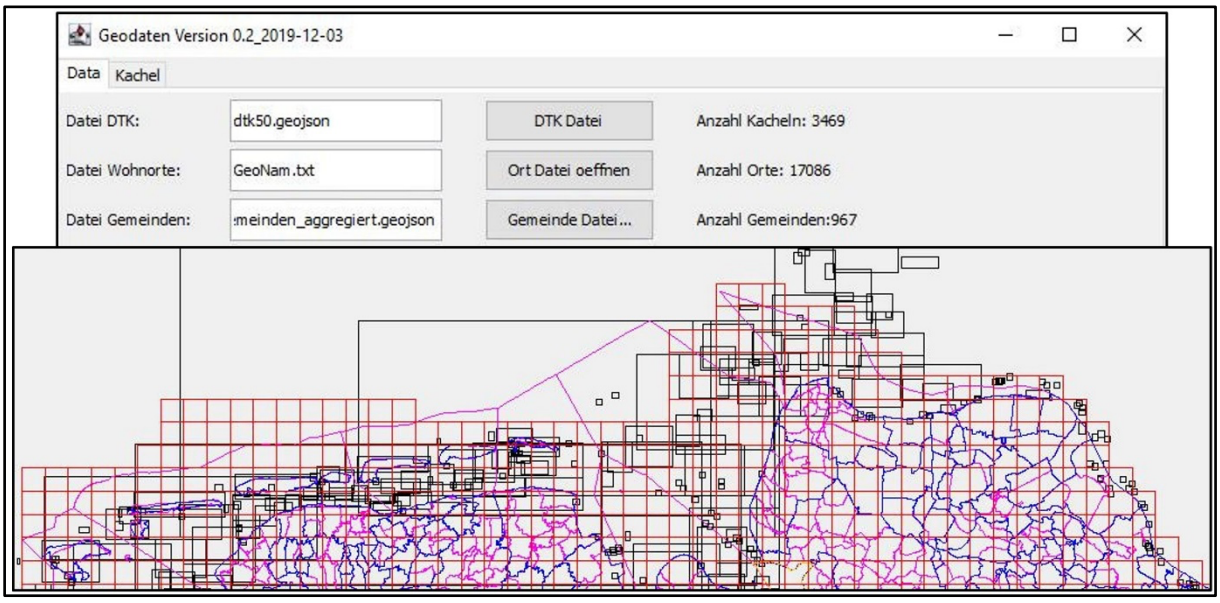


Abbildung 3

Das Ergebnis wurde in eine csv-Datei exportiert, in der für jede einzelne Kartenkachel aufgeführt wurde, welche Wohnplätze in ihrem Gebiet liegen, welche Verwaltungsgrenzen geschnitten werden und ob Seegebiete abgebildet sind (Abb. 4).

dtk100_Kachel_UTM_DTK2020_2023-07-11_13_55_09_angepasst.csv

```
UU_NO;Wohnplätze;Gemeindegebiete;Gemarkungen;Nordsee_Gebäude;Nordsee_Flächen
"s323405808";"keine";"Getelo, Itterbeck";"Getelo (Gemeinde: Getelo), Itterbeck (
"s323405816";"Balderhaar, Wohnplatz (Mitgliedsgemeinde: Wielen, Samtgemeinde: Ue
"s323405824";"Groß Ekenhorst, Volkstümlicher Siedlungsname (Mitgliedsgemeinde: L
"s323405832";"Agterhorn, Stadt-/Gemeindeteil (Mitgliedsgemeinde: Laar, Samtgemei
"s323485808";"Getelo, Gemeinde (Mitgliedsgemeinde: Getelo, Samtgemeinde: Uelsen)
"s323485816";"Achterende, Wohnplatz (Mitgliedsgemeinde: Itterbeck, Samtgemeinde:
"s323485824";"Echteler, Stadt-/Gemeindeteil (Mitgliedsgemeinde: Laar, Samtgemein
"s323485832";"Bahnhof Laarwald, Wohnplatz (Mitgliedsgemeinde: Laar, Samtgemeinde
"s323565808";"Baukamp, Wohnplatz (Mitgliedsgemeinde: Lage, Samtgemeinde: Neuenha
"s323565816";"Achteresche, Wohnplatz (Mitgliedsgemeinde: Esche, Samtgemeinde: Ne
"s323565824";"Arkel, Wohnplatz (Mitgliedsgemeinde: Hoogstede, Samtgemeinde: Emli
"s323565832";"Alexisdorf, Wohnplatz (Mitgliedsgemeinde: Ringe, Samtgemeinde: Eml
"s323645784";"Bardel, Stadt-/Gemeindeteil (Gemeinde: Bad Bentheim, Stadt), Elder
"s323645792";"Achterberg, Stadt-/Gemeindeteil (Gemeinde: Bad Bentheim, Stadt), G
"s323645800";"Am Birkenvenn, Wohnplatz (Gemeinde: Nordhorn, Stadt), Bahnübergang
"s323645808";"Altendorf, Volkstümlicher Siedlungsname (Gemeinde: Nordhorn, Stadt
"s323645816";"Alte Piccardie, Stadt-/Gemeindeteil (Mitgliedsgemeinde: Osterwald,
"s323645824";"Adorf, Stadt-/Gemeindeteil (Gemeinde: Twist), Füchten, Wohnplatz (
"s323645832";"Hesepertwist, Stadt-/Gemeindeteil (Gemeinde: Twist), Neuringe, Sta
"s323645840";"Hebelermeer, Stadt-/Gemeindeteil (Gemeinde: Twist), Schöninghsdorf
"s323645848";"Lindloh-Schwartenberg, Stadt-/Gemeindeteil (Gemeinde: Haren (Ems),
"s323725784";"keine";"Bad Bentheim, Ohne, Samern, Schüttorf";"Gildehaus (Gemeind
"s323725792";"Bad Bentheim, Stadt (Gemeinde: Bad Bentheim, Stadt), Bad Bentheim,
"s323725800";"An der Schule, Wohnplatz (Mitgliedsgemeinde: Engden, Samtgemeinde:
"s323725808";"Klausheide, Stadt-/Gemeindeteil (Gemeinde: Nordhorn, Stadt)";"Emsb
"s323725816";"Hustede, Volkstümlicher Siedlungsname (Gemeinde: Wietmarschen), Lo
"s323725824";"Dalum Feld, Wohnplatz (Gemeinde: Geeste), Dalumer Rull, Wohnplatz
"s323725832";"Heseper Torfwerk, Wohnplatz (Gemeinde: Geeste), Korde, Wohnplatz (
"s323725840";"Abbemühlen, Wohnplatz (Gemeinde: Meppen, Stadt), Auf der Heide, Wo
"s323725848";"Altenberge, Stadt-/Gemeindeteil (Gemeinde: Haren (Ems), Stadt), Al
"s323805784";"Hermeling, Wohnplatz (Mitgliedsgemeinde: Ohne, Samtgemeinde: Schüt
"s323805792";"Brameier, Wohnplatz (Mitgliedsgemeinde: Ohne, Samtgemeinde: Schütt
"s323805800";"Ahlde, Stadt-/Gemeindeteil (Gemeinde: Emsbüren), Auf dem Hörtel, W
```

Abbildung 4

Das DIMAG IngestTool ermöglicht es, beim Ingest der Daten in das digitale Magazin gleichzeitig die entsprechenden Verzeichnungsdatensätze in Arcinsys anzulegen und mit beliebigen Informationen anzureichern. Die erzeugte csv-Datei wurde mit dem IngestTool eingelesen und über die Kachelnummer eine Verknüpfung jeder Zeile zum passenden AIP hergestellt. Das Anlegen und Befüllen von knapp 35.000 Verzeichnungen erfolgte anschließend im Rahmen des Ingest automatisch.

Titel	Digitale Topographische Karte 1:50.000 Kachelnummer: 324605852
Laufzeit	2015
Enthält	Wohnplätze (2020): Aldrup, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Altona, Wohnplatz (Gemeinde: Goldenstedt), Denghausen, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Garmhausen, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Hanstedt, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Kleinenkneten, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Lohmühle, Bauerschaft (Gemeinde: Wildeshausen, Stadt), Vor Lohmüllers Felde, Wohnplatz (Gemeinde: Wildeshausen, Stadt), Wollringshof, Volkstümlicher Siedlungsname (Gemeinde: Wildeshausen, Stadt) Gemeindegebiete (2019): Goldenstedt, Visbek, Wildeshausen Gemarkungen (2019): Wildeshausen (Gemeinde: Wildeshausen, Stadt), Goldenstedt (Gemeinde: Goldenstedt), Visbek (Gemeinde: Visbek) Watt, Sandbänke, Leuchtfeuer, Leuchttürme (2019): keine
Benutzungshinweise	Die Legende/Zeichenerklärung befindet sich unter der Signatur NLA HA Nds. 128 Acc. 2019/73 Nr. 1.

Abbildung 5

Vergleicht man den Erschließungsdatensatz ohne den Enthält-Vermerk (Abb. 2) mit dem angereicherten Datensatz (Abb. 5), sieht man eine deutliche Verbesserung der Datenqualität. Nun ist die Verzeichnung mit einer Vielzahl an geographischen Begriffen versehen, nach denen über die Stichwortsuche recherchiert werden kann, sodass Forschende die Möglichkeit haben, die für ihre Fragestellung passenden Kartenausschnitte schnell zu finden.

Ergänzt wird die Suche anhand von Stichworten seit 2024 auch durch die Funktion der Geosuche[1]. Dabei handelt es sich um eine an Arcinsys angeschlossene webbasierte Kartenanwendung, die auf OpenStreetMap basiert. Nutzende können sich hier verschiedene Layer, zum Beispiel Verwaltungsgrenzen, Archivsprengel oder auch die Kachelraster der DTK, über dem Gebiet Niedersachsens einblenden lassen. Die Layer sind mit Objekten in Arcinsys verknüpft, sodass zum Beispiel beim Klick auf eine bestimmte Kartenkachel im Raster direkt der dazugehörige Verzeichnungsdatensatz angesteuert werden kann (Abb. 6). Dies vereinfacht die Suche in dem großen Kartenbestand zusätzlich.

[1] https://www.arcinsys.niedersachsen.de/geosuche/ (02.09.2024).

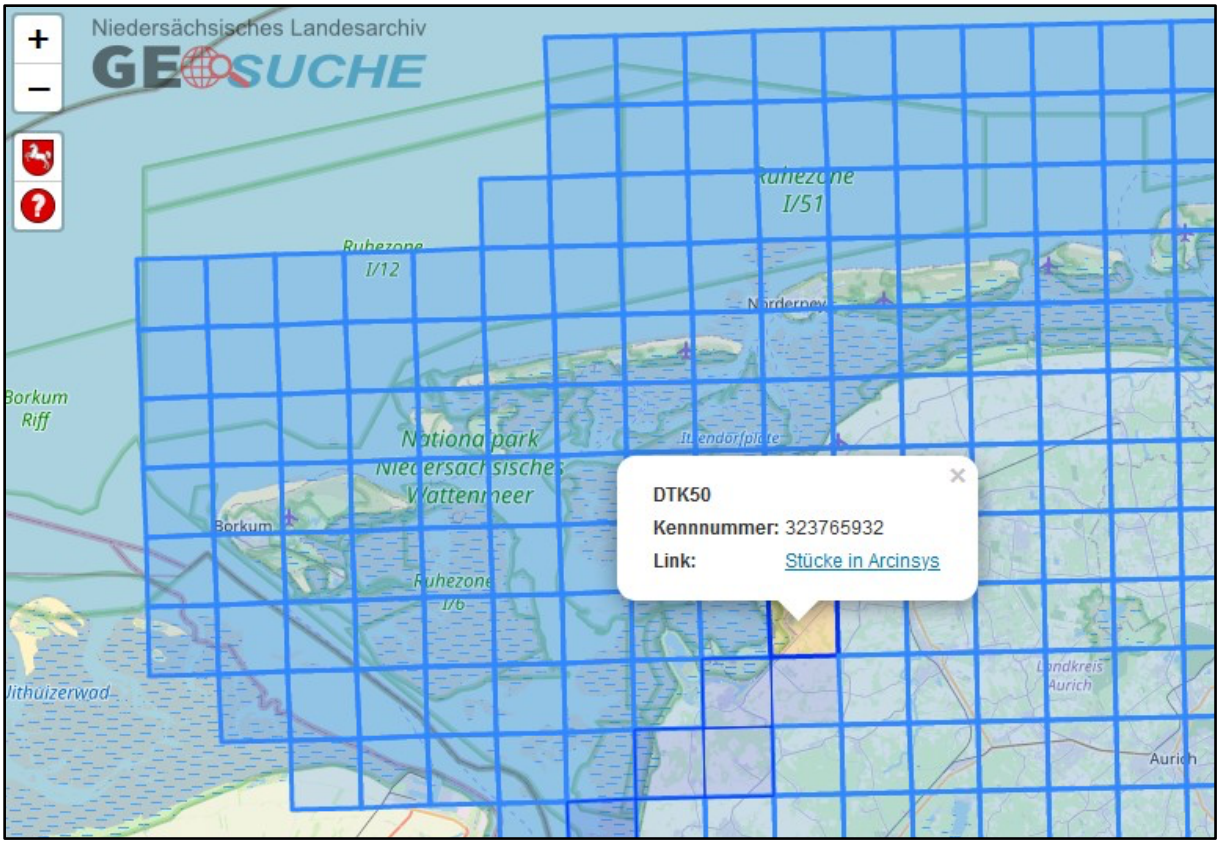


Abbildung 6

Nachdem die DTK Mitte 2024 als Open Data eingestuft wurden, wird derzeit an einer weiteren Vereinfachung der Nutzung gearbeitet. Ähnlich wie Digitalisate von analogem Archivgut sollen zeitnah auch die Kartenkacheln selbst mit dem Verzeichnungsdatensatz in Arcinsys verknüpft und über einen Online-Viewer betrachtet und heruntergeladen werden können. Dadurch entfällt der aktuell noch notwendige Schritt der Bestellung und ermöglicht auch eine ortsunabhängige Nutzung der Daten.

Fazit

Es hat sich gezeigt, dass digitales Archivgut neue Möglichkeiten der archivischen Tiefenerschließung bietet. Viele hilfreiche Metadaten, gerade im Bereich der Geodaten, stehen zunehmend als Open Data zur Verfügung und können zur Anreicherung verwendet werden. Wer in der eigenen Institution nicht über das nötige Fachwissen zum Programmieren eines Tools verfügt, kann für solche Projekte auch Kooperationen beispielweise mit Universitäten anstreben. Für das NLA hat sich die Mühe in jedem Fall ausgezahlt, da ohne die Aufbereitung der Erschließung zwar ein riesiger Satz an Daten übernommen worden, dieser aber wegen der schlechten Recherchierbarkeit kaum benutzbar gewesen wäre.

Kriterien für den Umgang mit unterschiedlichen Formaten in Dateiablagen im Archiv der sozialen Demokratie

Andreas Marquet und Annabel Walz

Einleitung

Digitale Unterlagen werden dem Archiv der sozialen Demokratie (AdsD) der Friedrich-Ebert-Stiftung (FES) seit einigen Jahren in unterschiedlichen Formaten und Formen angeboten. Allen technischen Entwicklungen zum Trotz stellen Dateiablagen dabei weiterhin das vorrangige Übernahmeszenario dar. Dateiablagen sind keineswegs die einzigen Systeme, in denen archivwürdiges Schriftgut entsteht und vorgehalten wird. In der Regel muss man eine umfassende Überlieferung weiterer Systeme, mindestens jedoch von E-Mail-Servern berücksichtigen. Dateiablagen müssen daher in Beziehung zur vorhandenen Infrastruktur und Arbeitslogik der abgebenden Person oder Organisation gesetzt werden.[1]

Als privates Archiv berät das AdsD seine Hinterleger:innen aktiv beim Records Management, kann jedoch auf keine verbindlichen Vorgaben zur Anwendung von Ordnungssystemen oder gar zur Führung von E-Akte-Systemen verweisen. Die bisherigen Erfahrungen weisen denn auch eine große Spannbreite in Umfang und Struktur der übernommenen digitalen Unterlagen auf, wobei eine Differenzierung zwischen der Überlieferung einzelner Personen und derjenigen größerer Organisationseinheiten geboten ist.

Die hierbei auftretenden Probleme sind grundsätzlich bekannt und bereits von verschiedenen Archivar:innen beschrieben worden (Wendt/Westphal, 2017; Taylor, 2016; Miegel et al., 2017; Naumann, 2017a). Strukturierungen, die zum Teil schwach und unvollständig durchgesetzt sind, wenig aussagekräftige Benennungen von Ordnern und Dateien, Redundanzen, die wegen des unklaren führenden Entstehungskontexts nicht oder nur mit erheblichem Aufwand aufgelöst werden können und vielfältige Dateiformate sind hierunter zu subsumieren. Trotz dieser – erweiterbaren – Liste an Schwierigkeiten suchen Archivar:innen nach geeigneten Lösungen, da File-Systeme oftmals die einzige Überlieferung digitaler Unterlagen darstellen und daher, auch aus grundsätzlichen Erwägungen, nicht ignoriert werden können.

[1] Aus der Parallelität verschiedener Systeme können für Archivar:innen und für Nutzer:innen praktische Probleme erwachsen, auf die in einem in Vorbereitung befindlichen Beitrag auf Grundlage des Vortrags auf dem Deutschen Historikertag 2024 eingegangen wird (Marquet, 2025).

Stand der Praxis

Im praktischen Umgang mit Dateiablagen werden i. d. R. einzelne Arbeitsschritte eines Workflows im Pre-Ingest nach organisatorischen, technischen und rechtlichen Anforderungen kombiniert – freilich ohne, dass diese notwendig klar gegeneinander abgegrenzt werden könnten. Die einzelnen Bearbeitungsschritte reichen von der (verifizierten) Übernahme der Dateisammlungen über Virenchecks, Analysen, technische Aufbereitung und ebenfalls technische Bewertung bis zu fachlicher Bewertung, Vorerschließung, Erschließung und Paketierung für den Ingest. Diese Prozessschritte können aus verschiedenen einzelnen Arbeiten bestehen, die Naumann (2017b) in drei Phasen zusammengefasst hat: 1. Analyse, 2. Nachbewertung und SIP-Formierung, 3. SIP-Erzeugung für digitales Archiv und AFIS (Birn/Naumann, 2019). Jüngst hat Leitzbach (2023) eine praktische Umsetzung vorgestellt, die auch unabhängig von dem Szenario des DIMAG-Anschlusses gewinnbringend ist. Die kanadische Nationalbibliothek und das kanadische Nationalarchiv bearbeiten Dateiablagen in einem vergleichbaren Workflow im Rahmen des Pre-Ingest, wenn auch insofern verkürzt, als die Bewertung hiervon ausgenommen ist (Tompkins, 2020). Um den Arbeitsaufwand zu minimieren, hat Gillner (2023) vorgeschlagen, gezielt Dokumente abzufragen, anstatt ganze Dateiablagen zu übernehmen. Je nach Umfang eines solchen Vorgehens wären auch solche Übernahmen geeignet, im weiteren Verfahren gemäß dem Phasenmodell bearbeitet zu werden. Wenngleich Vorgehensweisen im Einzelnen also variieren, kann dennoch grundsätzlich ein großer Bearbeitungsaufwand konstatiert werden.

In verschiedenen Beiträgen wurden beispielhafte Bearbeitungen übernommener File-Systeme so konkret beschrieben, dass einzelne Software-Tools und ihr jeweiliger Nutzen ebenfalls diskutiert wurden.[2] So hat Taylor (2016) den Bearbeitungsaufwand für eine schwach strukturierte Dateiablage quantifiziert und zur Diskussion über die Bewertung dieser Überlieferungsform beigetragen. Jaeger/Kobold (2017) haben bei ihrem Vorgehen automatisierbare Verfahren und den vertretbaren Einsatz von Arbeitszeit berücksichtigt, dabei freilich auch technische Begrenzungen bei der Ermittlung von Systemdateien hinnehmen müssen. Auch Belovari (2017) hat exemplarisch versucht, eine Dateiablage mit möglichst effizientem Ressourceneinsatz zu bewerten, und verschiedene Tools verwendet. Die Automatisierung einzelner Arbeitsschritte auf Basis eigens in der Programmiersprache Python geschriebener Skripte hat Lenartz (2020) entwickelt und hier zugleich mit der Verflachung der tief verschachtelten Ordnerstruktur sowie der Umsetzung eines Samplings Lösungen für bekannte Probleme erprobt. Der Ansatz ist auf eine

[2] Beiträge, die gezielt einzelne Tools auf ihren Nutzen für die digitale Archivierung evaluieren sind u. a. Naumann, 2017a und 2017b; Birn, 2017; Huth/Beyer, 2017; Näser/Herschung, 2017; Klein et al., 2017.

Sammlung von Fotografien angewandt worden, kann jedoch teilweise auf die Anforderungen von Schriftgutablagen konzeptionell übertragen werden.

Neben der Automatisierung einzelner Arbeitsschritte weist Sloyan (2016) zudem auf die Adaption bekannter Konzepte und Verfahrensweisen aus der Archivierung analoger Unterlagen hin. Die Klassifikation des Risikos, dass digitale Unterlagen schützenswerte personenbezogene Daten enthalten, verdeutlicht das Potenzial, neben technischen auch konzeptionelle Ansätze zu berücksichtigen.

Technische Voraussetzungen im AdsD

Die Umsetzung des Workflows zur digitalen Übernahme im AdsD findet unter einer Reihe technischer Rahmenbedingungen statt. So kann der Transfer von Dateien aus dem System der Hinterleger:innen ins AdsD remote über sftp bzw. das Hochladen in eine von der FES betriebene Cloud erfolgen oder alternativ über externe Datenträger. Das Ziel der digitalen Übernahme ist der Ingest via scopeArchiv in ein Fedora-Repository. Da die Konfigurierung des Ingests in scopeArchiv wenig flexibel ist, folgen aus dieser Zielvorgabe einige Setzungen: Dazu zählt, dass für Fileablagen nur der Ingest gemäß dem Schweizer Standard eCH-0160 möglich ist.[3] Für das hier behandelte Thema am folgenreichsten ist, dass scopeArchiv eine Erhaltungsplanung nur anhand von mit DROID[4] ermittelten PRONOM-Identifiern (PUID) ermöglicht (Brown, 2006). Es muss also im Pre-Ingest sichergestellt werden, dass beim Ingest eine Formatidentifikation mit DROID erfolgen kann.

Die hauptsächliche Arbeitsumgebung im AdsD sind Windows-Rechner. Es stehen aber auch einzelne Rechner mit Linux-Betriebssystem bereit, auf denen frei Software installiert werden darf. Um die Hürde zu senken, dass möglichst viele Kolleg:innen im Bereich der digitalen Übernahme mitarbeiten können, ist der Einsatz von auf Windows laufenden Programmen wünschenswert.

Unter Berücksichtigung dieser Ausgangsvoraussetzungen wurde zunächst ein Workflow für die Bearbeitung digitaler Unterlagen konzipiert, der auf verschiedenen Tools basiert: TotalCommander[5] für verifizierte Kopien, DROID für die Formatidentifikation, archifiltre[6] für einen Überblick über die Ordnerstrukturen und der Package Handler[7] für die Erstellung von SIPs gemäß eCH-0160. Die Schritte zwischen den Tools sollten nach Möglichkeit über Skripte

[3] In der aktuell im AdsD eingesetzten scopeArchiv-Version wurde bislang nur der Ingest von eCH-0160 Version 1.1 erprobt. (Verein eCH, 2015) Dieser Standard ist mittlerweile abgelöst (Verein eCH, 2024).
[4] https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ (9.10.2024).
[5] https://www.ghisler.com/deutsch.htm (9.10.2024).
[6] https://archifiltre.fabrique.social.gouv.fr/archifiltre-doc/ (9.10.2024).
[7] https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/package-handler.html (9.10.2024).

automatisiert werden. Für die Skripte wurde zunächst PowerShell eingesetzt, weil diese Skripte, wie alle genannten Tools, auf den Windows-Rechnern unkompliziert ausgeführt werden konnten.

Dieser erste Workflow zum Umgang mit digitalen Dateiablagen bestand aus folgenden Schritten:

(1) Aufspielen auf einen Quarantänerechner mit zweimaliger Virenprüfung im Abstand von vier Wochen
(2) Hashsummen-verifizierter Dateitransfer auf Speicher mit mehrfach redundanter Sicherung und Anlegen verifizierter Arbeitskopie (TotalCommander)
(3) De-Komprimierung von in der Ablage vorgefundenen Paketen auf Basis einer Container-Formaterkennung durch DROID (z.B. zip, gz, 7z)
(4) Allgemeine Formatidentifikation (DROID)
(5) Löschung von Dateien, die basierend auf ihrem PUID als kassabel eingeordnet werden können, wie Systemdateien oder temporäre Dateien (Powershell-Skript, das als Input Listen erhält, die auf Basis von gefilterten DROID-Ergebnislisten erstellt werden)
(6) Intellektuelle Analyse der Dateiablage und Bewertung auf Ordner/Datei-Ebene (archifiltre)
(7) Löschung der in (6) zur Kassation freigegebenen Ordner/Dateien (PowerShell-Skript, erzeugt durch archifiltre)
(8) Packen von SIPs gemäß eCH-0160 (Package Handler).

Im Package Handler werden Hashsummen angelegt, anhand derer das SIP beim Ingest in scopeArchiv validiert werden kann. Außerdem können hier Dateinamen, auch massenhaft, normalisiert und Pfadlängen gekürzt werden. Die Originalinformationen werden im SIP so gespeichert, dass sie im AIP und im Erschließungsmodul in scopeArchiv zugänglich sind. Die fertigen SIPs können dann in scopeArchiv ingestiert werden, wo die Dateiformate per DROID identifiziert und, soweit möglich, mit JHOVE[8] validiert werden. Je nach Konfiguration werden die Dateien bestimmter Formate nach festgelegten Regeln in andere Formate migriert. Der Ingest-Prozess wird gemäß dem Standard Preservation Metadata: Implementation Strategies (PREMIS) dokumentiert.

Probleme und Formatkatalog als Lösungsansatz

Bei der praktischen Erprobung dieses Workflows traten allerdings an verschiedenen Stellen Probleme auf. Zum einen wurde schnell deutlich, dass die Werkzeuge für die Bewertung unzureichend sind. Die Identifikation kassablen Materials in wenig strukturierten Ablagen erfordert

[8] https://jhove.openpreservation.org/ (9.10.2024).

einen viel zu hohen Zeitaufwand. Zum anderen wuchs mit jeder analysierten Dateiablage der Backlog an Dateien, bei denen sich aus (dem Versuch) der Formatidentifikation Probleme ergeben. Für zahlreiche Formate bietet scopeArchiv keine Migrationsstrategie, weshalb geprüft und festgelegt werden muss, ob und wie sie vor dem Ingest migriert werden müssen. Bei identifizierten Dateien könnte die Entwicklung einer Migrationsstrategie grundsätzlich auch zu einem späteren Zeitpunkt erfolgen.[9] Anders bei Dateien, die nicht (eindeutig) identifiziert werden konnten: Hier muss vor dem Ingest eine Lösung gefunden werden.

Beide Erkenntnisse waren dem Prinzip nach nicht überraschend, warfen aber die Frage auf, wie diese Masse an auftretenden Problemen handhabbar gemacht werden kann. Nebeneinander gestellt ergab sich außerdem die Frage, inwiefern Bewertungsfragen und Fragen der Priorisierung im Umgang mit Formaten in der Digitalen Langzeitarchivierung (DLZA) gekoppelt betrachtet werden können: So fiel auf, dass bestimmte Formate Indizien für kassable Inhalte sind. Ein Beispiel dafür sind Ordner, in denen in erster Linie Dokumente Dritter in Form von heruntergeladenen Webseiten abgelegt sind und wo folglich html- oder css-Seiten gehäuft auftreten. Bei Dateien, die aufgrund ihres Kontexts bereits als kassabel einzuordnen sind, kann wiederum auch auf den Aufwand einer Formatidentifikation in problematischen Fällen verzichtet werden. Umgekehrt kann es sinnvoll sein, die sichere Formatidentifikation oder Migrationsstrategie für ein bestimmtes Format mit höherer Priorität zu behandeln, wenn bekannt ist, dass eindeutig archivwürdige Inhalte nur in diesem Format vorliegen. Konkretes Beispiel sind Protokolle wichtiger Gremien in einzelnen Organisationen, die im proprietären mmap-Format der Software MindManager geführt worden sind. Die Entwicklung einer Formatsignatur für dieses Format, das bislang nicht in PRONOM hinterlegt ist, sowie einer geeigneten Migrationsstrategie ist somit zu priorisieren.

Diese beobachteten Zusammenhänge bildeten den Ausgangspunkt dafür als ersten Schritt in Richtung eines strukturierteren Umgangs mit Dateiablagen, die mit DROID aufgefundenen Formate einer Kategorisierung zu unterziehen. Ziel war folglich die Erarbeitung eines Überblicks der Formate, die sich in den Dateiablagen fanden, eine Priorisierung, welche problematischen Formate prioritär untersucht werden sollten und welche Formate einen Hinweis auf kassable Inhalte bieten könnten.

[9] Da bei einer Formatmigration aber ein Re-Ingest des gesamten AIPs in scopeArchiv erfolgen würde, würden dadurch ggf. größere redundante Datenmengen im Endarchiv entstehen. Auch hier wäre also eine Problemlösung vor erstmaligem Ingest wünschenswert.

Umsetzung

Ein Tool, das der Analyse von DROID-Resultaten bezüglich der Anforderungen der DLZA dient, ist freud.[10] Das Tool ist entwickelt worden von den National Archives in Großbritannien, die auch DROID weiterentwickeln und als Open-Source-Werkzeug der Community zur Verfügung stellen. Das Skript freud hält als Input das Ergebnis eines DROID-Durchlaufs in Form einer csv-Datei. Der Output ist eine Excel-Tabelle, die acht Tabs enthält, in denen jeweils die Dateien aufgeführt werden, die zu einer von acht Kategorien gehören, die für die DLZA relevant sein können. Die ersten vier Kategorien führen nicht erkannte Formate, Formate, die nur anhand der Extension erkannt worden sind, Formate mit mehreren PUIDs oder Extension Mismatches auf. Alle vier Kategorien sind Fälle, in denen einen verlässliche Weiterverarbeitung in scopeArchiv nicht gewährleistet ist. Des Weiteren werden komprimierte Containerformate erkannt und, sofern beim DROID-Durchlauf auch Hashs mit SHA256 erstellt wurden, auch Duplikate, die ebenfalls einer weiteren Bearbeitung bedürfen, sowie leere Dateien, die gelöscht werden können. Schließlich wird freud neben dem DROID-Input eine CSV-Datei mitgegeben, in der eine Liste mit akzeptierten PUIDs enthalten ist. Diese Liste ist individuell konfigurierbar. Alle Dateien mit PUIDs, die nicht auf dieser Liste sind, werden ebenfalls ausgegeben.

Das in Python programmierte freud-Skript diente nun als Ausgangspunkt für den Versuch, die Formatkategorisierung zu operationalisieren. Im Laufe der Automatisierungsversuche hatte sich bereits gezeigt, dass die Verwendung von PowerShell problematisch war, weil diese Sprache weniger gut dokumentiert ist als Python und kaum Skriptvorbilder aus dem Bereich der digitalen Erhaltung zu finden sind. Es wurde deshalb beschlossen, die Skripte in Python zu schreiben.[11]

Das auf die Anforderungen des AdsD angepasste Skript heißt, gemäß seiner Aufgabe, formatcategorization.[12] Als Input dient weiterhin der DROID-Output im CSV-Format zu einer Dateiablage. Zusätzlich wird dem Skript eine CSV-Liste format-list.csv übergeben, in der nun aber nicht mehr nur eine Spalte zu akzeptierten File-Formaten enthalten ist, sondern mehrere Rubriken. Der Output des Skripts ist eine Excel-Tabelle, in der für jede Datei eine Kategorisierung vorgenommen worden ist, eine Liste mit allen für die Löschung vorgesehenen Dateien und eine Liste mit Bewertungshinweisen, die als Tags in archifiltre hinzugefügt werden können.

[10] https://github.com/digital-preservation/freud (9.10.2024).
[11] Um zukünftig trotzdem eine möglichst breite Beteiligung von Kolleg:innen bei der Bearbeitung digitaler Übernahmen zu ermöglichen, ist geplant, die Ausführung der Skript-Kategorisierung und darauf basierenden automatisierten Vorgängen auf den Dateiablagen auf einen Linux-Server auszulagern. Das Auslösen der Skripte soll möglichst einfach sein, die Ausgabe der Ergebnisse in Formaten erfolgen, die für Weiterarbeit in der Windows-Umgebung geeignet sind.
[12] https://github.com/adsd-digital/pre-ingest-workflow (9.10.2024).

In der Excel-Tabelle werden gegenüber dem ursprünglichen DROID-Output drei zusätzliche Spalten angelegt: „Category“, „Appraisal“ und „Deletion“. Die Befüllung dieser Spalten erfolgt anhand der Kategorien, die auch in freud aufgerufen werden und anhand der Rubriken, die in der format-list.csv mitgegeben werden. Der Großteil dieser Rubriken bezeichnet Kategorien bezüglich des Umgangs mit Formaten in der DLZA. Hier wird zugeordnet, ob ein Format unverändert übernommen wird bzw. beim Ingest automatisch migriert werden kann. Diese Formate sind zunächst[13] unproblematisch. Ebenfalls unproblematisch sind Formate, bei denen eindeutig festgelegt werden kann, dass sie gelöscht werden können. Bei Formaten, die in komprimierter Form vorliegen oder für die zwar eine Möglichkeit zur Migration vorliegt, die aber außerhalb des Archivsystems liegt bzw. sogar manuell erfolgen muss, besteht die Notwendigkeit zur Bearbeitung. Noch problematischer sind Dateien, von denen bekannt ist, dass sie nicht geöffnet werden können, weil sie z.B. in proprietären Formaten vorliegen, zu denen im AdsD keine Software verfügbar ist, und Dateien, die passwortgeschützt sind. Insbesondere im letzteren Fall kann, je nach Übernahme, direkt eine Kassation sinnvoll sein. Eine weitere Überkategorie sind Dateiformate, die zwar bereits bekannt sind, für die aber noch nicht entschieden ist, wie mit ihnen verfahren werden kann und soll – hier wird unterschieden zwischen solchen, bei denen bereits begonnen wurde, die Möglichkeiten zum Umgang damit zu untersuchen und solchen, die zwar bereits aufgetaucht sind, aber noch gar nicht angesehen wurden. Die Zuordnung anhand dieser Rubriken kann natürlich nur gelingen, wenn ein verlässlicher PUID vorliegt. Bei fehlendem verlässlichen PUID (nicht identifiziert, Extension Mismatch etc.) erfolgt ein entsprechender Eintrag in die Spalte „Category“.

Wenn für einen PUID angegeben ist, dass die Datei gelöscht werden kann, wird das entsprechend zusätzlich in der Spalte „Deletion“ vermerkt. Außerdem wird der Pfad zu einer Liste hinzugefügt, die separat ausgegeben wird. Diese separate Löschliste kann dazu verwendet werden, automatisiert alle aufgeführten Dateien zu löschen. Schließlich gibt es die Option, einen PUID (unabhängig von der Einordnung dieses Identifiers bezüglich des Umgangs in der DLZA) als potenziellen Hinweis für die Bewertung einzuordnen. In diesem Fall erhalten die entsprechenden Dateien einen Eintrag in der Spalte „Appraisal“. Für die Dateien, die hier vermerkt sind, wird ein JSON-Snippet erstellt. Archifiltre speichert seine Dateiablagenanalyse in JSON. In diese Datei lässt sich das erzeugte Snippet einfügen. Bei erneutem Aufruf in archifiltre lässt sich nun nach dem Bewertungs-Tag filtern. Die getaggten Dateien und ihre übergeordneten Ordner werden außerdem visuell über einen blauen Balken gekennzeichnet.[14]

[13] Weitere Probleme können allerdings bei der Validierung beim Ingest auftreten.

[14] Die farbige Markierung der Tags in archifiltre funktioniert mit Version archifiltre 3.2.2, indem die Tags an der richtigen Stelle in die in JSON gespeicherte archifiltre-Ablagen-Übersicht eingefügt werden. Bei der inzwischen

Die erzeugte Excel-Tabelle lässt sich nun für die Einordnung einer Dateiablage bezüglich des Aufwands der Formatbearbeitung im Vorfeld des Ingests verwenden. Indem nach problematischen Dateiformaten gefiltert wird, lassen sich Listen erzeugen, die für Priorisierungsfragen in der weiteren Bearbeitung dieser Formate verwendet werden können. Allerdings schließen sich hier zeitintensive[15] und komplizierte weitere Bearbeitungsprozesse an, außerdem wäre es wünschenswert, den Abgleich der Listen und die Erzeugung neuer Input-Formatlisten ebenfalls zu automatisieren. Als Werkzeug für einen Überblick über vorhandene Formate und ihre Einordnung eignet sich der Output des Skripts aber. In den bisherigen Workflow wurde entsprechend als neuer fünfter Schritt eingefügt, nach der Durchführung der DROID-Formaterkennung eine Skript-Formatkategorisierung durchzuführen (siehe Abb. 1). Das Ergebnis wird intellektuell geprüft und entsprechend entschieden, für welche Formate nun weitere Recherchen angestoßen werden, wo manuelle Migrationsprozesse angestoßen werden müssen oder Automatisierung von Migrationsprozessen erprobt werden sollte.

Als problematischer hat sich dagegen erwiesen, Bewertungshinweise und eindeutige Löschentscheidungen allein auf Basis des PUIDs zu vergeben. Es wurde deutlich, dass fundierte Hinweise oder gar Löschentscheidungen weitere Input-Parameter benötigen, insbesondere Dateiname und Erstell-/Bearbeitungsdatum sind dabei vielversprechend. Das Skript soll deshalb dahingehend angepasst werden, dass neben dem Formatkatalog eine CSV-Datei mitgegeben werden kann, in der entsprechende Kriterienkombinationen formuliert werden können. Auf deren Basis kann dann eine Zuordnung von Bewertungs- und Löschhinweise auf den DROID-Input per Skript erfolgen. Nach bisherigen Beobachtungen müssen solche Kriterienkombinationen für verschiedene Überlieferungen angepasst werden (vgl. Abschnitt hybride Überlieferung).

Wie diese Kombinationen operationalisiert werden können, muss noch erarbeitet werden. Vorerst werden noch Erfahrungen gesammelt, welche Kombinationen vielversprechend sein könnten.[16] Es erscheint sinnvoll, die technisch basierten Löschhinweise in dem bereits angeführten neuen fünften Schritt der Formatkategorisierung mitzugeben. Anhand der dabei entstehenden

---

erschienen neueren Versionen von Archifiltre (Archifiltre-Docs 4.1.x) ist das Einfügen und Anzeigen der erzeugten Tags erst nach einmaligem manuellen Anlegen eines beliebigen Tags möglich.

[15] Zeitintensiv ist insbesondere die Bearbeitung von nicht identifizierten Formaten. Da für den Ingest in scopeArchiv eine Formatsignatur in PRONOM vorhanden sein muss, muss zunächst eine entsprechende Signatur für ein Format entwickelt werden, dieses bei PRONOM eingereicht werden und dann im entsprechenden nächsten Release enthalten sein (The National Archives, 2012). Je nach Fall kann es entsprechend sinnvoll sein, nur Teile einer Überlieferung zu ingestieren und die problematischen Ordnerausschnitte vorläufig im Zwischenarchiv zu belassen.

[16] Für Löschhinweise technisch obsoleter Dateien sind oft die Dateinamen relevanter als die PUIDs. Dazu zählen z.B. Systemdateien wie Thumbs.db oder (eigentlich temporär) gespeicherte Arbeitskopien von Office-Dokumenten, die z.B. mit „~$“ oder „~.lock“ beginnen. Thumbs.db werden von DROID oft als fmt/111 identifiziert, die Arbeitskopien oft (vermutlich fehlerhaft) als x-fmt/43, fmt/473 oder fmt/494. Auch die Dateigröße (i.d.R. 162 Bytes oder weniger) ist ein Indiz. Eine Kombination der Dateinamen bzw. -anfänge mit diesen Parametern verspricht eine sichere automatisierte Löschmöglichkeit. Je nach Überlieferung können auch Apple-Systemdateien gelöscht werden, wenn in einem Windows-Dateisystem davon ausgegangen werden kann, dass die archivwürdigen Unterlagen (auch) in Windowsformaten vorhanden sind. Ob entsprechende Formate pauschal gelöscht werden können, hängt allerdings von der spezifischen Überlieferung ab (vgl. z.B. Digital Preservation Q&A, 2014).

Löschlisten kann dann im Anschluss eine Kassation dieser technisch obsoleten Dateien erfolgen. Dadurch entschlackt man die Dateiablage, bevor die intellektuelle Bewertung erfolgt. Vor dieser intellektuellen Bewertung mithilfe des Tools archifiltre soll aber ein erneuter Skriptdurchlauf erfolgen, bei dem Bewertungshinweise erzeugt werden, die dann in archifiltre zur Bewertung herangezogen werden können.

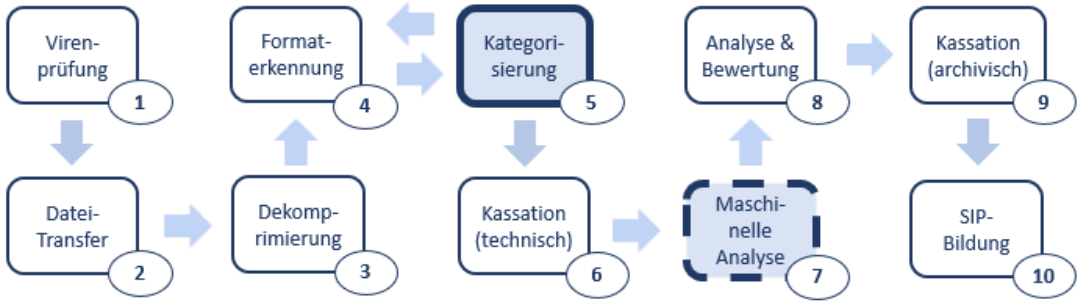


Abbildung 1: Aktualisierter Workflow

Hybride Überlieferung

Die Befassung mit Dateiformaten als Kategorie zur Bearbeitung von Dateiablagen wurde auch vor dem Hintergrund der oftmals hybrid erfolgenden Überlieferung von Unterlagen in einem weiteren Schritt auf ihre Praxistauglichkeit geprüft. Dabei wurde zunächst zwischen vereinzelten digitalen Beigaben, oftmals in Form von einzelnen Datenträgern wie Disketten und CD-ROMs, zu großenteils analogen Akten einerseits und umfangreichen Überlieferungen weit zurückreichender Dateiablagen andererseits unterschieden. Ziel ist, das führende Ablagesystem zu identifizieren und Redundanzen zu vermeiden oder zumindest zu begrenzen. Für beide Überlieferungssituationen wurde geprüft, inwiefern auf Grundlage identifizierter Dateiformate Rückschlüsse auf das führende System und das Mischverhältnis analog-digital gezogen werden können.

Vereinzelte Datenträger

Eine Stichprobe von insgesamt 12.076 Dateien mit einer Gesamtgröße von 6,7 GB wurde hierzu zunächst einer Formatidentifikation mit DROID unterzogen. Als Dateiformat war mit großem Abstand HTM / HTML am stärksten vertreten. Insgesamt 7.242 HTML-Dateien wurden durch DROID erkannt, was einem Anteil von rund 60% der gesamten Stichprobe entspricht. Anschließend folgten GIF- und JPG-Dateien. Textbezogene Dateiformate folgten hiernach

(DOC und TXT), während PDF auf Platz sechs als Container-Format nicht unbedingt ausschließlich schriftliche Überlieferung enthalten muss.

In einem zweiten Schritt wurde ein Abgleich von 1.061 Dateien textbasierter Dateiformate (DOC, TXT), was einem Anteil von etwa 9% der Stichprobe entspricht, mit der Papierüberlieferung in Autopsie durchgeführt. Demnach waren beinahe alle Unterlagen bereits in analoger Form vorhanden. Es handelte sich bei den Schriftstücken damit ausschließlich um Redundanzen, deren Äquivalent in der Papierakte im breiteren Entstehungszusammenhang bereits überliefert war. Ergänzend hierzu enthielten die Datenträger vereinzelt Unterlagen zur technischen Vorbereitung von Parteitagen wie etwa leere Musterformulare. Ein bleibender Wert wurde diesen Dateien nicht zugeschrieben. Schließlich enthielten einige Disketten Unterlagen, die sich in zwei verschiedenen, jedoch inhaltlich aufeinander bezogenen analogen Verzeichnungseinheiten befinden – ein Hinweis auf die Funktion der Datenträger als kondensierte und zugleich kontextarme Ergänzung der papiernen Überlieferung.

Die Analyse der HTML-Dateien erfolgte durch stichprobenartige Inaugenscheinnahme und wies ein differenziertes Bild auf. So scheinen frühe Webseiten als digital borns auf einzelnen Datenträgern gesichert worden zu sein. Diese Webseiten datieren z. T. auf die 1990er-Jahre zurück. Das Webarchiv des AdsD wurde seit 1998 auf- und ausgebaut, weshalb die hier identifizierten HTML-Dateien eine Ergänzung des Webarchivs darstellen, die freilich nicht nach systematischen Gesichtspunkten erfolgt. Dennoch sind die vereinzelten Beispiele angesichts der spärlichen Überlieferungssituation früher Webseiten eine Bereicherung.

Einige HTML-Dateien beinhalteten jedoch schriftlichen Niederschlag in Form von Textdokumenten oder E-Mail-Korrespondenz. Dies hat freilich den Vorteil, dass keine proprietären Textdateiformate in der weiteren Bearbeitung migriert werden müssen, weist jedoch zudem deutlich auf die begrenzte Aussagekraft von Dateiformaten hin. Denkbar wäre hier eine weitergehende Differenzierung hinsichtlich des Erstellungsdatums.

Gesamtes Laufwerk

Komplementär zur Stichprobe der einzelnen Datenträger wurde in einem zweiten Schritt das Gruppenlaufwerk einer operativen Fachabteilung einer Großorganisation analysiert. Die Dateiablage umfasste 122.919 Dateien mit einem Umfang von 143 GB.

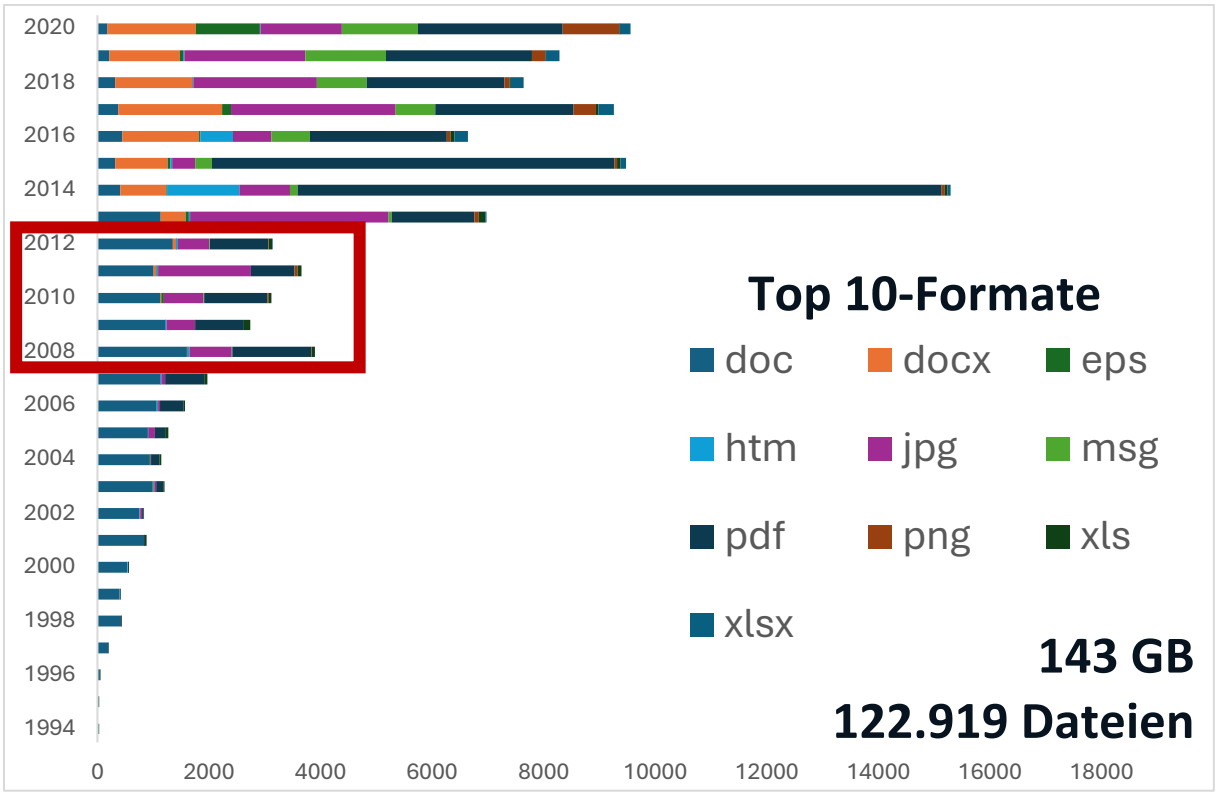


Abbildung 2: Stichprobe einzelne Dateiablage

Die Darstellung der Verteilung der Dateien nach Dateiformat und Entstehungsjahr erlaubt Schlussfolgerungen zur Genese der Dateiablage. Die hier ausgewiesenen zehn am häufigsten vorkommenden Formate addieren sich auf 100.246 Dateien, was einem Anteil von rund 82% entspricht. Ihre stärkere Ausdifferenzierung setzt etwa 2013 ein. Gegenüber den Vorjahren fällt diese Entwicklung mit einer Steigerung der Gesamtzahl der Dateien zusammen. In den folgenden Jahren bewegt sich der jährliche Zuwachs auf ähnlich hohem Niveau, vom Ausreißer 2014 abgesehen. Letzteres könnte möglicherweise mit einem forcierten Umstieg auf digitale Aktenführung verbunden sein, die zu Nachholeffekten führte. Einen ähnlichen Sprung weist, bezogen auf die Anzahl der Dateien, zudem das Jahr 2008 gegenüber den Vorjahren auf. Mit dem Format JPG kommen außerdem Bilder hinzu. Insofern kann die Arbeitshypothese aufgestellt werden, dass die zwischen den beiden hervorstechenden Jahren 2008 und 2013 liegende Zeit eine des Aufwuchses digitaler Arbeitsweisen, eine Art digitaler Sattelzeit, darstellt. Um das Ziel der Vermeidung redundanter Überlieferungen zu erreichen, erscheint es daher mit Blick auf die erforderlichen Ressourcen vertretbar, die vor 2008 liegende Zeit als Zeit der führenden Papierakte einzuordnen. Eine weitere Bearbeitung der Überlieferung sollte daher korrespondierend mit der analogen Überlieferung erfolgen resp. unterbleiben. Die Überlieferung der Phase von 2008 bis 2013 erscheint hingegen als weniger eindeutig hinsichtlich des Verhältnisses analog-digital in der Überlieferung, während alle jüngeren Unterlagen mit hoher Wahrscheinlichkeit einen

digitalen Entstehungs- und Bearbeitungskontext aufweisen. Wie oben gezeigt, sollten derartige Annahmen dennoch nicht ohne Blick auf die vorhandenen Dateiformate umgesetzt werden.

Fazit

Die Herausforderungen von Dateiablagen in der digitalen Übernahmepraxis führen zu arbeitsintensiven intellektuellen Bearbeitungsschritten im Pre-Ingest. Insbesondere die fachliche Bewertung von Dateiablagen erscheint angesichts der bekannten Probleme wie schwacher Strukturierung, nicht sprechender Dateinamen, Redundanzen etc. zeitaufwändig, wenn sie sich nicht auf Verfahren zur massenhaften Analyse oder gezielten Auswahl stützt. Für Archivar:innen wie für Nutzer:innen entstehen durch verschiedene technische Systeme und die vorhandenen parallel laufenden analogen Bestände schwer nachvollziehbare Überlieferungszusammenhänge, weshalb die Beachtung dieser Kontexte beim Umgang mit Dateiablagen notwendig erscheint. Zumindest für einen mittelfristigen Zeitraum des Übergangs wird diejenigen Archive, deren Hinterleger:innen nicht in einem klar definierten – und auch in der Praxis vollzogenen – Schnitt eine Umstellung auf die rein elektronische Aktenführung realisiert haben, diese Herausforderung begleiten.

Der vorgestellte Ansatz wählt die Dateiidentifikation als Ausgangspunkt, um den Umgang mit Dateiablagen handhabbar zu machen. Dabei ergeben sich zwei Perspektiven, die sich komplementär zueinander verhalten. Im Preservation Management zeigt sich in der Verteilung ein hohes Aufkommen gängiger Dateiformate, für die bereits Lösungen von der Formaterkennung, über Validierung bis hin zum Migrationsziel etabliert sind. Demgegenüber stehen zahlreiche Formate, die in ihrem Anteil an der Gesamtüberlieferung nach bisheriger Beobachtung überschaubar bleiben. Diese Einzelfälle bereiten aber, insbesondere, wenn es sich um seltener auftretende Formate handelt, unverhältnismäßigen Aufwand, weil es hier an Dokumentation und eindeutigen Migrationszielen und -methoden mangelt. Das vorgestellte Verfahren zum sukzessiven Ausbau des Formatkatalogs setzt bei dieser Problematik an. Gewonnene Erkenntnisse werden automatisiert bei der Analyse neuer Übernahmen berücksichtigt und geben Hinweise auf zu treffende Bewertungsentscheidungen. Es ist anzunehmen, dass der Mehrwert mit der Anzahl der Anwendungen des Instruments bis zu einem Sättigungsgrad steigen wird.

Zugleich weisen die Stichproben auf Potenzial hin, um Redundanzen bei hybriden Überlieferungen zu verringern. Freilich müssen hierfür die Überlieferungsziele sowie die analoge Überlieferungssituation berücksichtigt werden. Zumindest umfangreiche und zeitlich länger zurückreichende Dateiablagen lassen zu überprüfende Rückschlüsse auf die Entstehung des

Schriftguts zu, aus denen sich Hinweise zur Reduzierung der zu bearbeitenden digitalen Übernahme ableiten lassen.

Bibliografie

Belovari, S., 2017, „Rasche und einfache Bearbeitung von Dateisammlungen: Ein MPLP-Ansatz“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 17-29.

Birn, M. 2017, „Analyse und Datenaufbereitung von digitalen Ablagen mit TreeSize Professional und Total Commander“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 61-70.

Birn, M., Naumann, K., 2019, „Bewertung schwach strukturierter Unterlagen. Berichte und Thesen aus Baden-Württemberg“, in: *Brandenburgische Archive* 36, S. 8-14.

Brown, A., 2006, *The PRONOM PUID Scheme: A scheme of persistent unique identifiers for representation information* (= The National Archives. Digital Preservation Technical Paper 2), https://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf (9.10.2024).

Digital Preservation Q&A 2014, *What do you (or should you do) with ‘thumbs.db’ and other hidden system files?*, https://qanda.digipres.org/117/what-you-should-you-with-thumbs-and-other-hidden-system-files (9.10.2024).

Gillner, B., 2023, „Abfragen statt Anbieten: Eine alternative Praxis im Umgang mit Dateisystemen“, in: *Archiv. Theorie und Praxis* 76, S. 317-321.

Huth, K., Bayer, P., 2017, „Bytebarn – Datenbanklösung des Sächsischen Staatsarchivs zur Archivierung von Dateiverzeichnissen“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 71-78.

Jaeger, K., Kobold, M., 2017, „Zwischen Datenwust und arbeitsökonomischer Bewertung: Ein Werkstattbericht zum Umgang mit unstrukturierten Dateisammlungen am Beispiel des Bestandes der Odenwaldschule“, in: *Archivar. Zeitschrift für Archivwesen* 70, S. 307-311.

Klein, B., Steigmeier, A., Wildi, T., 2017, „docuteam packer – Informationspakete bilden und kontrolliert bewirtschaften“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 93-96.

Leitzbach, C., 2023, *Musterworkflow zur Bearbeitung von Dateisammlungen für DIMAG-Anwender*, https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2023/_jcr_content/Par/sgch_downloadlist_2069900880/DownloadListPar/sgch_download.ocFile/16_Praesentation_Leitzbach_StALb_2022-03-20.pdf (9.10.2024).

Lenartz, S., 2020, „Digital ist besser? Möglichkeiten der automatisierten Aufbereitung und Bewertung von Fileablagen mit Python am Beispiel einer digitalen Fotosammlung“, in: *Dialog Digital. Landesarchiv Baden-Württemberg*, Stuttgart, urn:nbn:de:101:1-2020052506.

Marquet, A., 2025, „Von digitalen Objekten zu Archivgut: „Uneindeutigkeit“ in der digitalen Überlieferungsbildung“, in: Kruke, A., Grothe, E. (Hg.), *Fragile Akten? Herausforderungen von (Digitaler) Überlieferungsbildung und Faktizität: Sektionsbeiträge des 54. Deutschen Historikertags, Beiträge aus dem Archiv der sozialen Demokratie*, Bonn, (in Vorbereitung).

Miegel, A., Schieber, S., Schmidt, C., 2017, „Vom richtigen Umgang mit kreativen digitalen Ablagen“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 7-16.

Näser, C, Herschung, A., 2017, „Übernahme unstrukturierter Dateisammlungen mit startext COMO“, in: Naumann, K., Puchta, M. (Hg.), *Kreative Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns*, München, S. 79-84.

Naumann, K., 2017a., „Dateisammlungen“, in: *Südwestdeutsche Archivalienkunde*, https://www.leo-bw.de/themenmodul/sudwestdeutsche-archivalienkunde/archivaliengattungen/sammlungen/dateisammlungen (9.10.2024).

Naumann, K., 2017b, „Welche Schritte erfordert die Aufbereitung von Dateisammlungen und welche Querschnitts- und Spezialwerkzeuge werden gebraucht?“, in: Naumann, K., Puchta, M. (Hg.), *Kreative*

Digitale Ablagen. Ergebnisse eines Workshops des KLA-Ausschusses Digitale Archive am 22./23. November 2016 in der Generaldirektion der Staatlichen Archive Bayerns, München, S. 44-60.

Sloyan, V., 2016, „Born-digital archives at the Wellcome Library: Appraisal and sensitivity review of two hard drives", in: *Archives and Records* 37, S. 20-36. https://doi.org/10.1080/23257962.2016.1144504.

Taylor, I., 2016, *Eine hydraartige Matrjoschka: Wie wir die Fileablage eines staatlichen Schulamtes bewertet und erschlossen haben*, https://www.sg.ch/content/dam/sgch/kultur/staatsarchiv/auds-2016/archivierung-von-unterlagen-mit-besonderen-strukturen/01_TAYLOR_Vortragsfolien%20(29.02.16).pdf (9.10.2024).

The National Archives, 2012, *PRONOM - How to research and develop signatures for file format identification*, https://cdn.nationalarchives.gov.uk/documents/information-management/pronom-file-signature-research.pdf (9.10.2024).

Tompkins, H., 2020, „Preserving the bits: Library and Archives Canada's Pre-Ingest workflow", in: *Digital Preservation Coalition*, https://www.dpconline.org/blog/wdpd/blog-heather-tompkins-wdpd (9.10.2024).

Verein eCH, 2015, *Archivische Ablieferungsschnittstelle eCH-0160, Version 1.1*, https://www.ech.ch/de/ech/ech-0160/1.1 (9.10.2024).

Verein eCH, 2024, *Archivische Ablieferungsschnittstelle eCH-0160, Version 1*.3, https://www.ech.ch/de/ech/ech-0160/1.3.0 (9.10.2024).

Wendt, G., Westphal, S., 2017, „Eine Herausforderung des Übergangs: Fileablagen als Quellen der digitalen Überlieferung", in: *Transformation ins Digitale. 85. Deutscher Archivtag in Karlsruhe, Tagungsdokumentationen zum Deutschen Archivtag*, Fulda, S. 105-113.

Borg:
Open Source-Programm des Landesarchivs Thüringen zur einfacheren Einbindung und Kombination beliebiger Formaterkennungs- und Validierungswerkzeuge

Tony Grochow

BorgFormat (kurz: Borg) ist ein vom Landesarchiv Thüringen entwickeltes Programm, das beliebige Formaterkennungs- und -validierungswerkzeuge kombiniert und die einzelnen Ergebnisse nach konfigurierbaren Regeln zu einem Gesamtergebnis zusammenführt. Borg ist unter der Open-Source-Lizenz GNU General Public License 3 im GitHub des Landesarchivs unter https://github.com/landesarchiv-Thueringen/borg veröffentlicht und kann kostenfrei genutzt werden.

Herausforderungen bei der Identifizierung von Dateiformaten

Die verlässliche Identifizierung von Dateiformaten ist ein komplexes Problem, das insbesondere aufgrund des Fehlens universeller Standards für Aufbau, Inhalt und Metadaten von Dateiformaten derzeit nicht vollständig gelöst werden kann. Grundsätzlich können Dateiformate in textbasierte und Binärformate unterteilt werden. Bei textbasierten Formaten wird jedem Byte bzw. jeder Byte-Sequenz mittels einer Zeichenkodierung ein darstellbares Zeichen zugeordnet. Zu den wichtigsten Zeichenkodierungen zählen ASCII und UTF-8. Beispiele für verbreitete textbasierte Formate sind CSV, XML und TXT (reiner Text). Im Gegensatz dazu können Binärformate beliebige Bitmuster enthalten und sind nicht auf darstellbare Zeichen beschränkt. Häufig genutzte Binärformate sind unter anderem PDF, JPEG und MP3.

Dateiformate werden typischerweise anhand bekannter Byte-Sequenzen identifiziert, die an spezifischen Positionen innerhalb der Datei auftreten. Das funktioniert am zuverlässigsten, wenn es einen vorgegebenen Datei-Header bzw. -Footer für das Format gibt. Das ist ein kleiner Block am Anfang bzw. Ende einer Datei, der die wichtigsten technischen Informationen und Metadaten zur Datei enthält. Dateiformate mit losen oder ohne jegliche Strukturvorgaben wie insbesondere textbasierte Formate sind so nur schwer identifizierbar. Dateiformate für Quellcode sind ein gutes Beispiel für Formate mit losen Strukturvorgaben. Bestimmte Byte-Sequenzen liegen zwar in der Regel vor, die Position kann jedoch variieren. TXT-Dateien besitzen keinerlei strukturelle Vorgaben, ihre Identifizierung basiert ausschließlich auf der

Übereinstimmung mit einer Zeichenkodierung und dem Fehlen von Merkmalen, die auf andere textbasierte Formate hinweisen. Für die Erkennung textbasierter Formate ist es zudem erforderlich, dass die Zeichenkodierung entweder bekannt ist oder erkannt werden kann. Die automatische Erkennung von Zeichenkodierungen ist zwar ebenfalls nicht trivial, viele textbasierte Formate legen inzwischen jedoch UTF-8 als Standard fest, wodurch die Erkennung der Zeichenkodierung entfällt.

Für eine verlässliche Identifizierung des Dateiformats ist die Formaterkennung mit den üblichen Methoden nicht ausreichend. Wie beschrieben, wird meist nur ein kleiner Teil der Datei betrachtet. Selbst wenn das Format richtig zugeordnet wurde, erlaubt das keine Rückschlüsse auf den Zustand der Datei. Diese kann beliebige Fehler enthalten. Um die Formaterkennungsergebnisse zu verifizieren und die Integrität des Dateiformats zu sichern, werden Validatoren benötigt. Ein Validator überprüft, ob Dateien mit der Spezifikation des jeweiligen Dateiformats übereinstimmen. Anstatt einige wenige Merkmale von Dateien auszuwerten, wie es bei der Formaterkennung geschieht, müssen bei der Validierung alle Aspekte geprüft werden. Den Entwicklern der Open-Source-Werkzeuge zur Formaterkennung und -validierung kann nicht genug dafür gedankt werden, dass sie diese Werkzeuge frei verfügbar bereitstellen, wodurch sich der Lösung des Problems wenigstens angenähert werden kann.

Der Gesamtprozess von Formaterkennung und -validierung wird im Folgenden vereinfacht als Formatverifikation bezeichnet. Aufgrund der Komplexität des Problems ist nachvollziehbar, dass kein Programm eine zuverlässige Formatverifikation für alle marktüblichen Dateiformate leisten kann. Es ist mittlerweile gängige Praxis, mehrere Werkzeuge zu diesem Zweck miteinander zu kombinieren. Es existieren bereits Anwendungen, die verschiedene Werkzeuge für die Formatverifikation verwenden. Ein bekanntes Beispiel hierfür ist FITS. Diese Anwendung integriert eine große Anzahl an Programmen und fasst die Ergebnisse der unterschiedlichen Werkzeuge zusammen. Das gleiche Grundprinzip nutzt auch Borg. Borg unterscheidet sich von den etablierten Anwendungen für die Formatverifikation besonders hinsichtlich Werkzeugintegration und Zusammenstellung des Gesamtergebnisses.

Technische Basis von Borg

Borg kann auf zwei verschiedene Arten verwendet werden. Die Funktionalitäten für die Formatverifikation können über eine REST-API in beliebige Anwendungen (zum Beispiel in ein Digitales Archiv) integriert werden. Zusätzlich wird eine Webanwendung bereitgestellt, mit der die manuelle Analyse von Dateien am Arbeitsplatz möglich ist.

Das Verhalten des Borg-Servers wird über eine Konfigurationsdatei gesteuert. In dieser wird definiert, wie auf die integrierten Dritt-Werkzeuge zugegriffen wird, unter welchen Bedingungen sie ausgeführt werden, welche Eigenschaften von ihnen erhoben und wie diese für das Gesamtergebnis gewichtet werden. Borg wird mit einer funktionalen Grundkonfiguration bereitgestellt, diese kann aber beliebig anhand eigener Erfahrungswerte angepasst werden.
Sowohl der Borg-Server als auch die angesprochenen Dritt-Werkzeuge werden in Docker-Containern ausgeführt. Die Containerisierung sorgt für eine konsistente Umgebung für die Ausführung und minimiert Probleme, die durch Abweichungen in Betriebssystemen oder installierten Bibliotheken entstehen können. Weiterhin wird die Verwaltung von Abhängigkeiten vereinfacht und die Portabilität der Anwendung verbessert. Insgesamt werden dadurch die Entwicklung und der Betrieb von Borg vereinfacht.
Im aktuellen Borg-Release Version 1.3.0 vom 14.11.2024 sind folgende Werkzeuge integriert:

Werkzeug	Zweck
DROID	Formaterkennung
Tika	Metadatenextraktion und Formaterkennung
JHOVE	Validator für mehrere Formate, insbes. HTML, PDF, JPEG, TIFF
veraPDF	Validator für PDF/A- und PDF/UA-Dateien
ODF Validator	Validator für OpenDocument-Formate, die von LibreOffice- und OpenOffice genutzt werden
OOXML Validator	Validator für Microsoft Office Formate

Tabelle 1: Auflistung aller integrierten Werkzeuge

Funktionsweise von Borg

Die Formatverifikation mit Borg verläuft in folgenden Schritten:

- Formaterkennung: Sobald eine Datei vollständig auf den Server übertragen wurde, werden zuerst alle Formaterkennungswerkzeuge ausgeführt. Die Ergebnisse werden für die weitere Verarbeitung gleichwertig behandelt. Es wird noch nicht versucht, ein Format für die Datei festzulegen.
- Auswahl der Validatoren: Anhand der Formaterkennungsergebnisse werden die Validatoren ausgewählt. Für jeden Validator wird konfiguratorisch vorab bestimmt, unter welchen Bedingungen er ausgeführt wird. Wenn das Ergebnis von einem Formaterkennungswerkzeug die Bedingung erfüllt, wird der Validator ausgeführt, unabhängig davon, ob andere Formaterkennungsergebnisse dem widersprechen. Beispielsweise wird das HTML-Modul von JHOVE nur ausgeführt, wenn der MIME-Type *text/html* erkannt wurde.

- Ermittlung eines vorläufigen Gesamtergebnisses aus Formaterkennung- und -validierung: Die Resultate aller eingesetzten Werkzeuge zur Formaterkennung und Validierung werden von Borg im Anschluss automatisch zu einem Gesamtergebnis aggregiert. Dabei werden ausschließlich die von den Tools extrahierten Eigenschaften berücksichtigt. Für jede Eigenschaft (z.B. MIME-Typ) werden alle von den Werkzeugen extrahierten Werte zusammengetragen. Beispielsweise könnten für ein textbasiertes Format die Werte *text/plain*, *application/xhtml+xml* und *text/html* identifiziert werden. Im nächsten Schritt erfolgt eine Bewertung dieser extrahierten Werte. Die Bewertung basiert auf einer gewichteten Abstimmung der beteiligten Tools. Jedes Werkzeug stimmt gemäß seiner in der Konfiguration vorab definierten Gewichtung für den von ihm ermittelten Wert. Werkzeuge, die denselben Wert identifiziert haben, stimmen gemeinsam ab. Die Gesamtbewertung eines Wertes ergibt sich aus der Summe der gewichteten Stimmen, die er erhalten hat, im Verhältnis zur Gesamtsumme aller Stimmen. Der Wert mit der höchsten gewichteten Gesamtpunktzahl wird somit als der am besten bewertete betrachtet.
- Anpassung der Gewichtung: Borg ermöglicht die Anpassung der Gewichtung von Werkzeugergebnissen anhand des vorläufigen Gesamtergebnisses. Für alle Werkzeuge kann definiert werden, unter welchen Umständen die Gewichtung einer extrahierten Eigenschaft automatisch geändert wird. Nachdem alle Gewichtungen entsprechend den Regeln angepasst wurden, wird ein neues Gesamtergebnis berechnet, auf die gleiche Weise wie das vorläufige Gesamtergebnis. Diese Funktion soll es ermöglichen, bekannten Schwächen von Werkzeugen entgegenzuwirken. Ein typischer Anwendungsfall ist die Abwertung einer von einem Werkzeug extrahierten Eigenschaft, weil im vorläufigen Gesamtergebnis ein Format erkannt wurde, mit dem das Werkzeug nicht gut umgehen kann. Es ist aber auch möglich, die Bewertung der Ausprägung einer extrahierten Eigenschaft zu verbessern. Beispielsweise wird die Eigenschaft Validität für das PDF-Modul von JHOVE stark abgewertet, wenn im Gesamtergebnis PDF/A als Dateiformat erkannt wurde. Das JHOVE-Modul kann nur die zu Grunde liegende PDF-Version validieren. Die Prüfung von Dateien auf Konformität mit PDF/A kann das Modul aktuell nicht leisten. Dafür wird die von veraPDF ermittelte Validität mit der vollen Gewichtung in das Gesamtergebnis eingehen, sodass sich das Ergebnis von dem geeigneten Werkzeug durchsetzt.
- Zusammenfassung des Gesamtergebnisses: Abschließend wird eine Zusammenfassung für die durchgeführte Formatverifikation erstellt. Die Zusammenfassung soll die manuelle und maschinelle Auswertung der Ergebnisse vereinfachen. Die Zuverlässigkeit des Gesamtergebnisses ist unmittelbar erkennbar und weitere notwendige Verarbeitungsschritte können

abgeleitet werden. Zusätzlich werden die unverarbeiteten Werkzeugausgaben, die Bewertung aller extrahierter Eigenschaften und die Zusammenfassung übermittelt. Es ist somit möglich, eigene Schlussfolgerungen aus den erhobenen Daten zu ziehen, ohne sich auf die Bewertung von Borg zu verlassen.

Zusätzliche Webanwendung für Standalone-Betrieb

Zusätzlich zur integrierten Nutzung von Borg wird mit dem Programm eine Webanwendung ausgeliefert, mit der die Funktionalitäten von Borg auch manuell am Arbeitsplatz genutzt werden können.

Dateiauswahl

In der Nutzeroberfläche der Webanwendung können im Bereich *Dateiauswahl* Einzeldateien oder ganze Ordner manuell ausgewählt werden (s. Abb. 1). Die Auswahl kann per Drag and Drop oder mit einem Datei-Dialog erfolgen. Der Borg-Server kann je nach Server-Kapazität und Datei-Umfang nicht alle Dateien gleichzeitig verarbeiten. In der Ansicht ist daher für jede Datei erkennbar, ob sie gerade auf den Borg-Server hochgeladen wird, analysiert wird oder sich noch in der Warteschlange befindet.

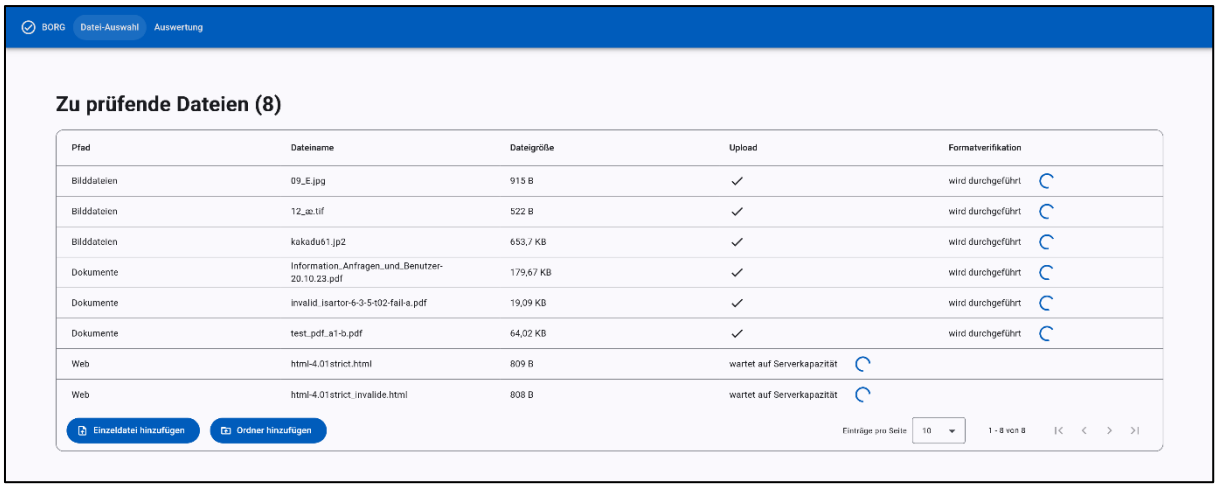


Abbildung 1: Dateiauswahl und Fortschrittsanzeige der Formatverifikation

Auswertung

Sobald die Formatverifikation abgeschlossen ist, wird die Datei aus der Übersicht der Dateiauswahl entfernt und die Ergebnisse in die Auswertung übernommen, diese zeigt für jede analysierte Datei ein kumuliertes Ergebnis der wichtigsten erhobenen Eigenschaften (s. Abb. 2). Details der Ergebnisse werden in dieser Ansicht bewusst verborgen, um einen schnellen Überblick zu ermöglichen.

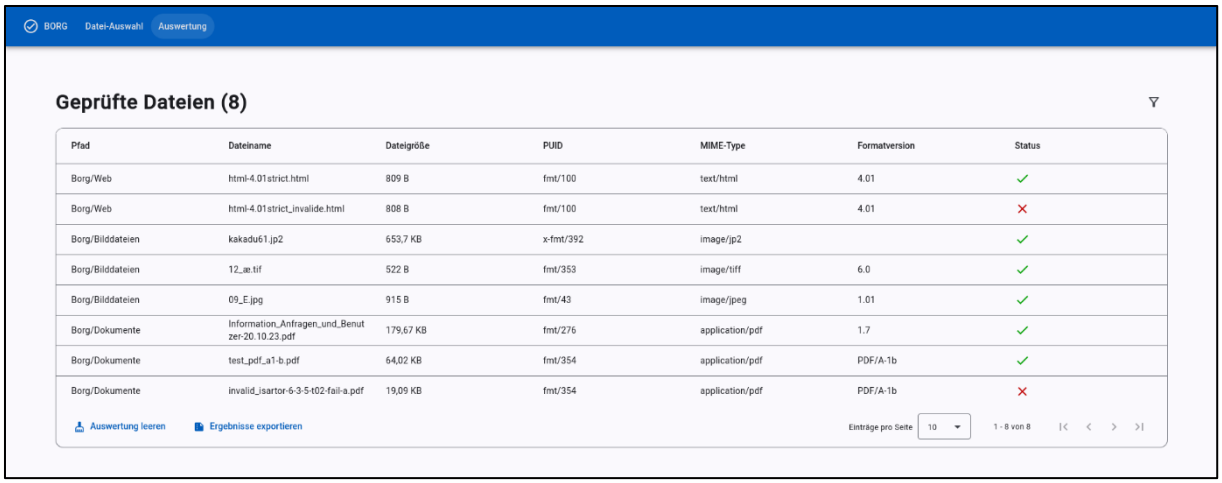


Abbildung 2: Auswertung der Formatverifikation für alle Dateien

Zusätzlich zu den Format-identifizierenden Eigenschaften wird ein Verifikationsstatus für jede Datei angezeigt. Der Status stellt das Gesamtergebnis der Formatverifikation dar. Durch den Status können problematische Ergebnisse, die eine manuelle Prüfung benötigen, einfach identifiziert werden. Folgende Fälle werden unterschieden:

Symbol	Bedeutung
✓	Es konnte mit hoher Wahrscheinlichkeit ein Format ermittelt werden, für das Format ist ein Validator verfügbar und die Datei ist valide.
✗	Es konnte mit hoher Wahrscheinlichkeit ein Format ermittelt werden, für das Format ist ein Validator verfügbar und die Datei ist invalide.
	Es konnte mit hoher Wahrscheinlichkeit ein Format ermittelt werden, es ist aber kein Validator vorhanden.
⚠	Es konnte kein Format mit ausreichend hoher Wahrscheinlichkeit ermittelt werden.
ⓘ	Es kam zu einem Fehler bei der Ausführung von Werkzeugen.

Tabelle 2: Bedeutung des Verifikationsstatus

Mit Klick auf einen Dateieintrag wird die Detailansicht der Dateiergebnisse geöffnet. In der Ansicht werden für die wichtigsten Eigenschaften die Ergebnisse aller Werkzeuge und die zugehörige Gewichtung für das Gesamtergebnis dargestellt. Optisch besonders hervorgehoben wird das Gesamtergebnis, das aus den Einzelergebnissen der Werkzeuge ermittelt wurde (s. Abb. 3).

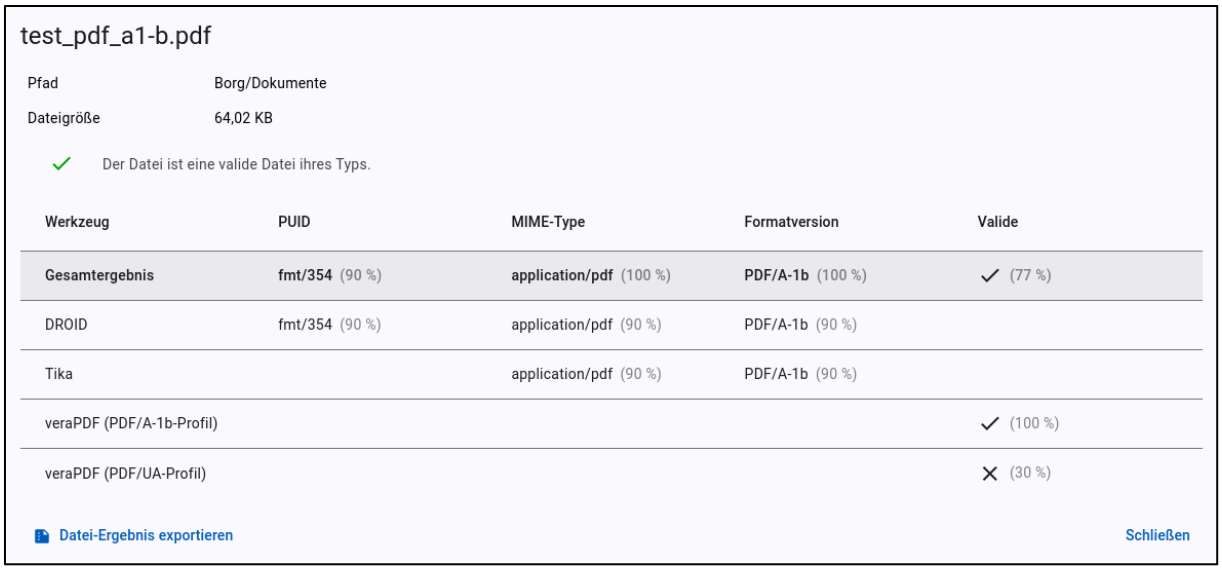


Abbildung 3: Detailansicht der Werkzeugergebnisse für eine Datei

In der Detailansicht kann zusätzlich die originale Werkzeugausgabe inklusive der extrahierten Eigenschaften geöffnet werden (s. Abb. 4 und 5).

Abbildung 4: Detailansicht der extrahierten Eigenschaften von DROID

Werkzeug-Ausgabe


```
{
  "X-TIKA:Parsed-By": [
    "org.apache.tika.parser.DefaultParser",
    "org.apache.tika.parser.html.HtmlParser"
  ],
  "dc:title": "Untitled Document",
  "Content-Encoding": "ISO-8859-1",
  "Content-Type-Hint": "text/html; charset=iso-8859-1",
  "resourceName": "8487b706-2160-4da4-b1f1-2966083f8cab_html-4.01strict.html",
  "Content-Length": "809",
  "Content-Type": "text/html; charset=ISO-8859-1"
}
```


Abbildung 5: Detailansicht der kompletten Werkzeugausgabe von DROID

Alle von Borg zusammengestellten Informationen können für einzelne oder alle Dateien in einem maschinenlesbaren Format aus der Anwendung exportiert werden.

Ausblick

Das Landesarchiv Thüringen plant Borg kontinuierlich weiterzuentwickeln, insbesondere durch die Integration weiterer Werkzeuge (z. B. Jpylyzer). Es ist bereits angedacht, das von Google entwickelte Formaterkennungswerkzeug Magika (Fratantonio et al. 2024) zu integrieren. Das Werkzeug nutzt maschinelles Lernen zur Identifizierung von Formaten. Damit unterscheidet es sich grundsätzlich von anderen Werkzeugen im Bereich Formaterkennung und könnte vor allem einen Mehrwert bei text-basierten Formaten bieten.

Bibliografie

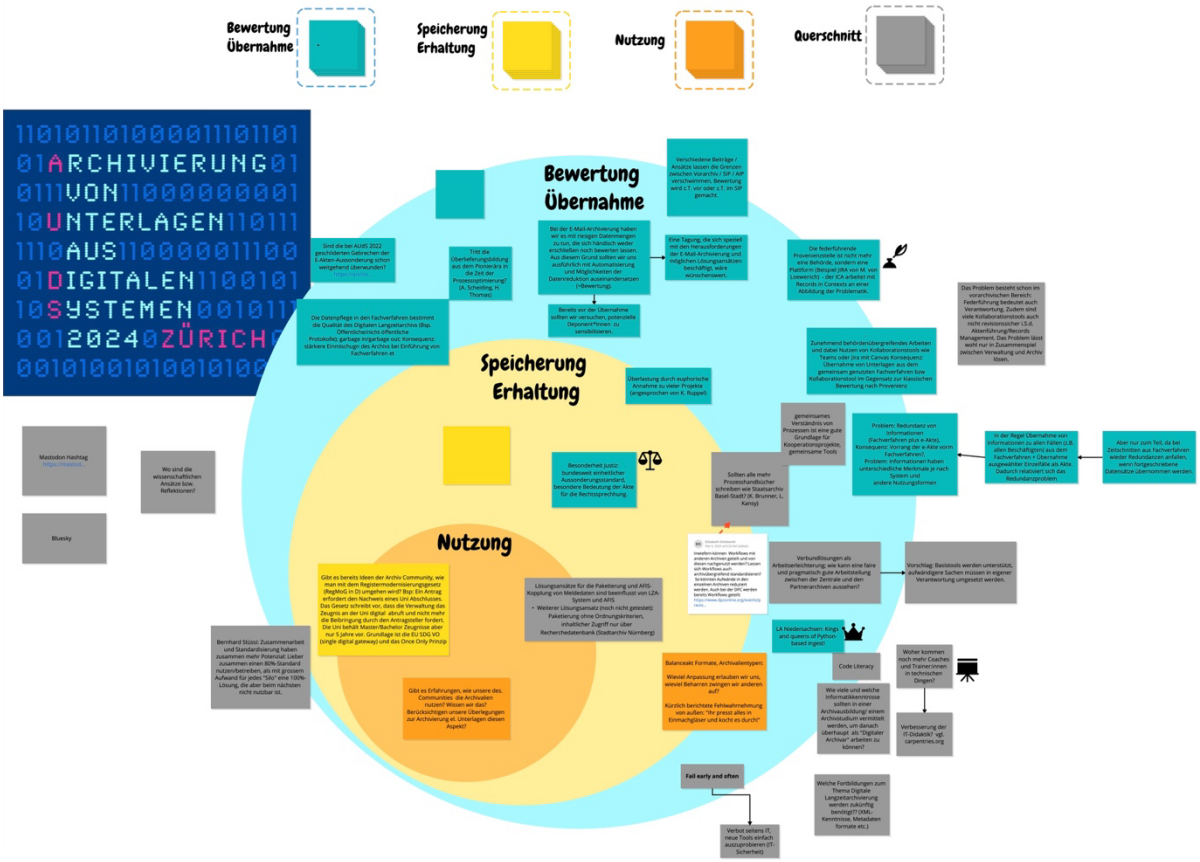
Fratantonio, Yanick et al., 2024, „Magika: AI-Powered Content-Type Detection“, arXiv [cs.CR] [Preprint], http://arxiv.org/abs/2409.13768.

FAZIT UND AUSBLICK

Kai Naumann

Im März 2024 traf sich in Zürich eine Community, die mit dem langfristigen Erhalt genuin digitaler Information in historischen Archiven und größeren Bibliotheken in Deutschland, Luxemburg, Österreich, der Tschechischen Republik und der Schweiz betraut ist. Um sie herum in der großen Politik waberten die geopolitische Zeitenwende, Fake-News, Blasenbildung, Plattformkapitalismus. Aus der deutschsprachigen Wissenschaftswelt war die Tagung beeinflusst von Bewegungen wie Open Data, FAIR Data und internationale Forschungsdateninfrastrukturen.

Die nachfolgenden Zeilen sind teils Einschätzungen des Moderators der Abschlussdiskussion, teils dem Sammlungsdokument in der Plattform Conceptboard entnommen.[1]



[1] Vgl. Diagramm der Ergebnisse des AUdS 2024, https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2024/_jcr_content/Par/sgch_downloadlist/DownloadListPar/sgch_download_1520909495.oc-File/AUdS%202024%20Board-1.pdf (28.10.2024).

Um den spontanen Eindruck des Moderators darzustellen: Die Zürcher Tagung bot eine unfassbare Vielfalt an Themen und ein vorbildliches Umfeld. Nicht nur war die Tagung mit großzügigen Ressourcen ausgestattet und hatte ein schönes Umfeld, sondern das Gastgeberland hob sich auch auf vielen Ebenen als handlungsfähig, gründlich und pragmatisch hervor. Ein gutes Beispiel sind staatliche Informationssysteme für Finanzen, Justiz, Soziales oder Geodaten. Die darin von Zentral- und Regionalverwaltungen (Kantone, Länder, Kommunen) kollaborativ geschaffenen Aufzeichnungen passen nicht zu unserer bisherigen Übung, die jede Akte genau einem zuständigen Archiv zuordnet. Gegen Anfang der 2020er-Jahre befassten sich Arbeitsgruppen in Deutschland und der Schweiz damit – und 2023 wurde bei der Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) dafür ein Koordinierungsgremium mit öffentlich einsehbarem Arbeitsprogramm eingerichtet (Liste Ebis). Im vielfach größeren Nachbarland kann die Konferenz der Leitungen der staatlichen Archive KLA (wenigstens bis zur Drucklegung dieses Beitrags) Vergleichbares nicht vorweisen.

Ein wichtiges Thema der Präsentationen waren Schnittstellen, die in viele Richtungen eine Übernahme von Informationen in unsere digitalen Archivsysteme ermöglichen, seien es Gesundheitsdaten, Justizdaten, Meldedaten, Personaldaten, Schuldaten, Studierendendaten, E-Akten aller Art, E-Mail, Social Media oder Webseiten.

Besonders zu erwähnen sind auch die Anstrengungen, quelloffene Software zu schaffen. Open-Source-Komponenten, die gemeinsam von verschiedenen Softwareprojekten benutzt werden, könnten eines Tages für diese Projekte, gleichgültig ob sie kommerziell oder öffentlich-rechtlich betrieben werden, gemeinsame Vorteile bringen (T. Grochow, F. Obermeit, C. Träger, aber auch Projekte außerhalb der Tagung wie Archifiltre, Preservation Action Registries, PRONOM oder VeraPDF, vgl. Bibliografie). Ebenso sinnvoll ist auch das Teilen von Konstruktionsdetails wie zum Beispiel das Datenmodell der Anwendung transfero des deutschen Bundesarchivs (L. Schützeichel),[2] eines OAIS am Germanischen Nationalmuseum (R. Nasarek),[3] der Basler Prozessablauf Übernahme und Erschließung (K. Brunner, L. Kansy), das Augsburger Social-Media-Verfahren (D. Feldmann) oder der Zürcher Prüfkatalog für strukturierte Unterlagen (E. Peng), der an die Ludwigsburger Handreichung (2019) erinnert.

[2] Referat „transfero – eine Anwendung zur Übernahme von genuin digitalen Unterlagen in das Bundesarchiv – ein Werkstattbericht“ von Lennard Schützeichel (nicht im vorliegenden Tagungsband).
[3] Referat „Datenmodell und Implementierung eines OAI-Systems am Germanischen Nationalmuseum Nürnberg“ von Robert Nasarek (nicht im vorliegenden Tagungsband).

Wirklich sinnvoll war aber auch ein offenes Reden über Fehler (z. B. F. Bächler und D. Rölli),[4] vor allem da oft nur hinter vorgehaltener Hand offenbart wird, dass viele, auch große Mitspieler nur mit Wasser kochen und hinter den Kulissen bei weitem nicht alles perfekt läuft. Einzelne Referate gab es auch über methodische Gedanken (M. von Loewenich, Z. Stodůlka). Beides sollte eigentlich öfter stattfinden.

Insgesamt ist die Aufgabe, digitale Akten, Datensätze, E-Mails, Geodaten oder Videos auf Dauer für Zwecke über die Lebensdauer von Hard- und Software zu erhalten, inzwischen in vielen Gebieten aus der Pilotphase heraus. Nach der Pionierära kommen nun die Mühen der Prozessoptimierung, wie A. Scheiding und H. Thomas feststellten. Dabei verschwimmen bislang in der Theorie feststehende Grenzen zwischen den Konzepten. Bewertung findet irgendwo zwischen abgebender Stelle, SIP- und AIP-Bildung statt, gleichgültig wo sie im OAIS-Standard verortet ist. Die Metadaten der Objekte sind über mehrere Systeme verteilt, so dass Archiv und Behörde sie mühsam zusammenklauben müssen (B. Gillner). Selbst wo die Bewertung ansetzt und welches Archiv zuständig ist, also das Provenienzprinzip, verschwimmt im kollaborativen Arbeiten (M. von Loewenich) – immerhin arbeitet der International Council on Archives (ICA) mit Records in Contexts (RiC) an einer Abbildung der Problematik. Alle haben dabei die IT-Sicherheitsaufgaben im Nacken, die das Entwickeln neuer Lösungen am Arbeitsplatz nicht einfacher machen. Einige haben in der Anfangseuphorie zu viele Eisen ins Feuer gelegt und kommen nun mit keinem der Projekte richtig voran.

Die logische Folgerung, nämlich innezuhalten und grundsätzlich zu fragen, ob die Kolleginnen und Kollegen fachlich hinreichend gewappnet und in ausreichender Zahl bereitstehen, um die Aufgaben wirklich zu lösen, kam etwas verhalten in der Abschlussrunde aufs Tapet. Maria Benauer (2017) hat sich inzwischen dazu in ArchivT&P geäußert. Während in der Schwesterdisziplin Forschungsdatenmanagement erhebliche Aufwände in Fort- und Weiterbildung gesteckt werden (Einführungsmaterialien, vgl. Bibliografie), wirkt die um AUdS gescharte Fachgemeinde in diesem Bereich etwas träger.

Gute Wege in die Zukunft könnten über neue Allianzen führen, sei es mit der Geschichtswissenschaft, die sich gerade mit digitalen Methoden erneuert (Digital Humanities, vgl. Bibliografie), mit den Digital Natives auf Tagungen wie re:publica oder Chaos Computer Congress und mit den übrigen Wissenschaften in der Nationalen Forschungsdateninfrastruktur und deren Pendants in Österreich und der Schweiz. Wesentlich ist auch ein Schulterschluss mit IT-Sicherheitsleuten. Diese nehmen wir oft als Spielverderber wahr, weil sie ein Ausprobieren

[4] Referat „Archivierung über Ablieferungsschnittstellen im Stadtarchiv Bern – wenige Tops, viele Flops“ von Fabienne Bächler und Daniela Rölli (nicht im vorliegenden Tagungsband).

auswärtiger Tools verhindern oder verzögern. IT-Sicherheit und auch Datenschutz sind aber Grundbedingungen für vertrauenswürdige Archive und müssen daher in unsere Prozesse integriert werden.

An der Softwareentwicklung der letzten Jahre fällt auf, dass zwar in den Verbünden (DiPS und DIMAG) viele Lösungen gründlich entwickelt und breit geteilt werden, aber in einigen anderen Bereichen das Rad gern neu erfunden wird. So gibt es neben dem bei der Tagung vorgestellten xdomea Aussonderungsmanager (C. Träger) im deutschsprachigen Raum mehr als eine Handvoll weiterer Lösungen, teils öffentlich vorgestellt und teils im internen Einsatz, die sich der Bewertung und Übernahme von E-Akten widmen. Auch EMILIA (E. Klindworth, N. Beyer) steht neben einigen anderen Lösungen aus dem Ausland, die in Deutschland bislang kaum zum Einsatz kommen. Was ist ein Erklärungsmuster dafür? Fehlt es gar nicht immer an Software, sondern an kundigen deutschsprachigen Personen, die die Verantwortung übernehmen und die teils sehr komplizierten Aufgaben der digitalen Archivierung bewältigen? Demnach bräuchte die Community eher mehr Trainingsräume und Lehrpersonen als noch mehr Softwarelösungen. Wenn Quereinsteiger zu uns kommen, um diese Lücke zu füllen, dann sollten wir uns darüber freuen, gut zuhören und abwägen, welche sinnvollen Neuerungen sie mitbringen.

Insgesamt bleibt festzuhalten: Große und kleine Mitspieler wirken derzeit sehr rege zusammen. Besonders wichtig ist dabei das Zusammenspiel der Ausbildungseinrichtungen (Berlin, Bern-Lausanne, Marburg, Potsdam, München), der Verbünde (KLA, KOST, Landschaftsverbände Rheinland und Westfalen, DIMAG, DiPS, DNS in NRW, Anwenderkreise von Archivematica, Preservica, Rosetta), der reinen Infrastruktureinrichtungen (FIZ Karlsruhe, Rechenzentren) und der Firmen im Spiel (Artefactual, Augias, Conet ISB, docuteam, DXC, ScopeArchiv, Startext u.v.a. mehr). Gemeinsam mit den übrigen beteiligten Institutionen und Einzelpersonen, auch den beruflichen Verwandten in den Bibliotheken und Museen, werden wir noch viele so inspirierende AUdS-Tagungen erleben.

Bibliografie

Archifiltre https://archifiltre.fr/ (07.10.2024)

Benauer, M. (2017), „Fuß fassen ohne stehen zu bleiben. Kompetenzrahmen und Reifegradmodelle als professionelle Kommunikationsmittel der digitalen Archivierung“, in: *ArchivT&P* 77 (2024), H. 3, S. 250-254. https://www.archive.nrw.de/landesarchiv-nrw/ueber-uns/archiv-theorie-praxis (07.10.2024)

Community Owned digital Preservation Tool Registry (COPTR), Auswahlseite zur Objektart E-Mail, https://coptr.digipres.org/index.php/Email (07.10.2024)

Digital Humanities im deutschsprachigen Raum, https://dig-hum.de (07.10.2024)

Einführungsmaterialien und Professionalisierung, Rubrik auf Forschungsdaten.info, https://forschungsdaten.info/praxis-kompakt/fdm-einfuehrungsmaterialien/, https://forschungsdaten.info/praxis-kompakt/fdm-professionalisierung/ (07.10.2024)

Liste aller ebenenübergreifenden Informationssysteme (Ebis), KOST Homepage, https://kost-ceco.ch/cms/ebuebis-liste-alle.html (07.10.2024)

Ludwigsburger Handreichung Datenträger (2019), https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2019.html (07.10.2024)
Preservation Action Registries (PAR), https://parcore.org/ (07.10.2024)
PRONOM, http://www.nationalarchives.gov.uk/pronom/ (07.10.2024)
Sandra Funck u.a. (2024), Tagungsbericht: Born digital – neue Archivaliengattungen und ihre Bearbeitung im Archiv, in: *H-Soz-Kult*, 30.07.2024, http://www.hsozkult.de/conferencereport/id/fdkn-145612 (07.10.2024).
VeraPDF, https://verapdf.org/home/ (07.10.2024)

Autorinnen und Autoren

Nico Beyer, BA: Freie Universität Berlin, im EMILiA-Projekt für den Austausch mit der Fachcommunity, die Formulierung der archivfachlichen Anforderungen und Usertests zuständig. ORCID: 0009-0003-8984-3572

Claudia Briellmann, MA, MAS ALIS: Hochschularchiv der ETH Zürich, wissenschaftliche Archivarin, u. a. im Vermittlungsdienst, der Erschließung von Verwaltungsschriftgut und der E-Mail-Archivierung.

Kerstin Brunner, lic. phil.: Staatsarchiv Basel-Stadt, Erschließung und digitale Archivierung.

Mona Bunse, MA: Universitätsarchiv der Universität Duisburg-Essen, wissenschaftliche Mitarbeiterin im Projekt LZA.NRW mit Schwerpunkten Beratung beim Einstieg in die digitale Langzeitarchivierung, Vernetzung sowie Austausch mit den Stakeholder:innen des Projekts.

Jürgen Enge, Diplominformatiker: Universitätsbibliothek Basel, Leiter IT. ORCID: 0000-0002-3148-9772

Jacqueline Fehr, dipl. Sekundarlehrerin, MBA: 1998 bis 2015 Mitglied des schweizerischen Nationalrats, seit 2015 Regierungsrätin des Kantons Zürich (Vorsteherin der Direktion der Justiz und des Innern), 2021/22 Regierungspräsidentin des Kantons Zürich.

Dominik Feldmann, Dr.: Stellvertretender Leiter des Stadtarchivs Augsburg, Leiter der Abteilung „Digitale Archivierung und Digitalisierung“.

Tony Franzky, MA: Erzbischöfliches Archiv Freiburg, Sachgebietsleiter für Digitale Archivierung mit Schwerpunkt digitale Langzeitarchivierung von elektronisch entstandenen Schrift-, Wissens- und Kulturgütern (speziell E-Akten, Websites, Fachverfahrensdaten).

Christine Friederich, Dr.: Sächsisches Staatsarchiv, Leiterin Referat 12: Archivfachliche Grundsatzangelegenheiten, elektronische Archivierung in der Abteilung Zentrale Aufgaben, Grundsatz.

Bastian Gillner, Dr.: Landesarchiv Nordrhein-Westfalen, Fachbereich Grundsätze in Duisburg, Leiter des Dezernats F 4 für E-Government und Elektronische Unterlagen; Vorsitzender des KLA-Ausschusses Records Management.

Beat Gnädinger, Dr.: Staatsarchivar des Kantons Zürich.

Tony Grochow: Landesarchiv Thüringen, Mitarbeiter der IT mit Arbeitsschwerpunkten Softwareentwicklung und Softwaretests.

Bernhard Homa, Dr.: Niedersächsisches Landesarchiv, Abteilung Hannover, u. a. stellvertretender Teamleiter Benutzung sowie zuständig für die Überlieferungsbildung im Geschäftsbereich des Sozialministeriums und der nachgeordneten Behörden des Innenministeriums.

Karsten Huth, Dokumentar: Sächsisches Staatsarchiv, Referent in der Abteilung Zentrale Aufgaben, Grundsatz.

Lambert Kansy, lic. phil., Diplomarchivar: Staatsarchiv Basel-Stadt, Leiter Informatik. ORCID: 0000-0002-5062-1097

Elisabeth Klindworth, M.A.: Archiv der Max-Planck-Gesellschaft, Zuständige für die digitale Langzeitarchivierung. ORCID: 0000-0003-1848-5870

Christian Koller, Prof. Dr., FRHistS: Direktor des Schweizerischen Sozialarchivs in Zürich, Titularprofessor für Geschichte der Neuzeit an der Universität Zürich, Dozent für Sozialgeschichte an der FernUni Schweiz. ORCID: 0000-0001-9701-0122

Antje Lengnik, Diplomarchivarin: Niedersächsisches Landesarchiv, Abteilung Zentrale Dienste, zuständig für die Digitale Archivierung mit Schwerpunkt Geobasis- und Geofachdaten.

Maria von Loewenich, Dr.: Deutsches Bundesarchiv, wissenschaftliche Archivarin im Referat B 2 (Bewertung und Erschließung analoger Unterlagen der ministeriellen Bundesverwaltung, Behördenberatung, Zwischenarchive).

Andreas Marquet, Dr., M.A., M.LIS: Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung in Bonn, Leiter des Referats Infrastrukturen und digitale Grundsatzfragen, Chief Digital Officer. ORCID: 0000-0001-7238-9033

Kai Naumann, Dr.: Landesarchiv Baden-Württemberg, Referent an der Abteilung Archivischer Grundsatz in Stuttgart mit Arbeitsschwerpunkten digitale Überlieferungsbildung und Archivrecht, Dozent an der Fachhochschule Potsdam. ORCID: 0000-0002-2799-1030

Frank Obermeit, Diplominformatiker: Landesarchiv Sachsen-Anhalt, Referent Abteilung 1: Zentrale Dienste.

Elia Peng, MA: Stadtarchiv Zürich, Technischer Records Manager.

Martin Rechtorik, MA: Národní archiv in Prag, Archivar, Mitglied des Methodenteams, zuständig für Datenbanken und räumliche Datenerhaltung, Doktorand an der Westböhmischen Universität in Pilsen. ORCID: 0009-0004-8539-237X

Christine Rigler, Mag. Dr.: Leiterin des Universitätsarchivs der Universität Graz. ORCID: 0000-0002-5050-0797

Antje Scheiding, Diplomarchivarin, M.Sc.: Sächsische Anstalt für Kommunale Datenverarbeitung (SAKD), Referentin des elektronischen Kommunalarchivs.

Fabian Schneider, B.Sc.: ETH-Bibliothek in Zürich, Abteilung Forschungsdatenmanagement und Datenerhalt, Data Archivist mit Schwerpunkt digitale Langzeitarchivierung im ETH Data Archive.

Isabell Schönecker, Diplomarchivarin: Niedersächsisches Landesarchiv, Abteilung Zentrale Dienste (Team DIMAG), v. a. für die Digitale Archivierung zuständig.

Zbyšek Stodůlka, Mgr.: Národní archiv in Prag, Digitalarchivar, Schwerpunkte in den Bereichen digitale Langzeitarchivierung, Records Management und Förderung von Transparenz in der öffentlichen Verwaltung, Lehrbeauftragter an der Karls-Universität Prag und der Masaryk-Universität in Brünn.

Henrike Thomas, B.A.: Stadtarchiv Leipzig, Bestandsreferentin der Stabsstelle e-Archiv.

Christine Träger: Landesarchiv Thüringen, Leiterin der Informationstechnologie und Projektleiterin zum Aufbau der digitalen Langzeitarchivierung.

Martin Vogel, Dipl.-Wirt.-Inf.: Niedersächsisches Landesarchiv, Abteilung Teamgruppe b: Digitale Dienste / DIMAG, Mitglied im KLA-Ausschuss „Digitale Archive“.

Annabel Walz, M.A., B.Sc.: Archiv der sozialen Demokratie der Friedrich-Ebert-Stiftung in Bonn, Referentin Digitale Langzeitarchivierung. ORCID: 0000-0001-6894-8568