

Fachbeitrag

Bettina Kann und Stefan Majewski

Konsortiale digitale Langzeitarchivierung für Wissenschaft, Kultur und Gedächtnisinstitutionen

Consortial Services in Digital Long-term Preservation for Scholarly, Cultural and Memorial Institutions

<https://doi.org/10.1515/abitech-2023-0048>

Zusammenfassung: Mit der fortschreitenden Digitalisierung entwickelte sich in den letzten Jahrzehnten ein steigendes Bewusstsein für die Herausforderungen bei der langfristigen Erhaltung der Informationen. Dies ist auch im vollen Umfang in den Kultur- und Gedächtnisinstitutionen angekommen, die in unterschiedlichem Ausmaß in der Lage sind, eigenständig diesen Anforderungen nachzukommen. Die Österreichische Bibliothekenverbund und Service GmbH bietet seit Juni 2022 „Digitale Langzeitarchivierung“ als konsortialen Dienst an. Die Eckpunkte eines konsortial geeigneten Produkts wurden in einer europaweiten Ausschreibung gemeinsam mit der Österreichischen Nationalbibliothek als Erstkunden festgelegt. Die Implementierung des Dienstes auf Basis des Produkts Libsafe von Libnova erfolgte ebenfalls gemeinsam mit der Österreichischen Nationalbibliothek in einem mehrmonatigen Projekt.

Der Artikel behandelt die für einen konsortialen Betrieb notwendigen Anforderungen, deren technische Umsetzung, erforderlichen Komponenten und Funktionalitäten, sowie organisatorische und wirtschaftliche Aspekte. Insbesondere die Rolle einer Verbundzentrale als Dienstleister wird herausgearbeitet. So werden die Möglichkeiten beschrieben, zentral Kompetenzen, Wissen und Infrastruktur zu bündeln, um einem breiten Spektrum an Institutionen digitale Langzeitarchivierung (LZA) zu ermöglichen. Der Service richtet sich daher ausdrücklich auch an Institutionen, die nicht Mitglieder des österreichischen Bibliothekenverbundes sind. Die enge Kooperation mit dem Kunden steht dabei im Zentrum. Der Dienst umfasst den kompletten Betrieb der erforderlichen Hard- und Software sowie Konfiguration des Langzeitarchivierungssystems, Datenanalyse, Einrichten geeigneter Archivierungspläne und Workflows.

Schlüsselwörter: digitale Langzeitarchivierung, konsortialer Dienst, Österreichischer Bibliothekenverbund

Abstract: With the advancement of digitalization, there has been growing awareness of the challenges of long-term availability and preservation in recent decades. This has also fully arrived at cultural and memory institutions, which can meet these requirements on their own to varying degrees. The Austrian Library Network Service GmbH has been offering “Digital Long-Term Preservation” as a consortial service since June 2022. The key elements of a suitable product for a consortial setting were jointly established with the Austrian National Library as the first customer in a European procurement process. The service based on Libsafe by Libnova has also been implemented with the Austrian National Library in a several-month project.

The article deals with the requirements necessary for consortial operation, their technical implementation, required components and functionality, as well as organizational and economic aspects. In particular, the role of a consortium central agency as a service provider is highlighted. The possibilities of pooling competencies, knowledge, and infrastructure to make long-term preservation possible for a wide range of institutions are described. The service is therefore also aimed at institutions that are not members of the Austrian Library Network. Close cooperation with the customer is at the center of the offering. The service comprises the complete operation of the necessary hardware and software, the configuration of the long-term preservation system, data analysis, and suitable preservation plans and workflows.

Keywords: digital longterm preservation, consortial service infrastructure, Austrian Library Network

1 Konsortiale digitale Langzeitarchivierung für Wissenschaft, Kultur und Gedächtnisinstitutionen

1.1 Ausgangslage

Mit der fortschreitenden Digitalisierung entwickelte sich in den letzten Jahrzehnten ein steigendes Bewusstsein für die Herausforderungen bei der langfristigen Erhaltung digitaler Informationen. Dies ist auch im vollen Umfang in den Kultur- und Gedächtnisinstitutionen angekommen, die sich in unterschiedlichem Ausmaß, aber letztlich fast alle dieser Aufgabe stellen müssen. Neben den großen Einrichtungen wie der Österreichischen Nationalbibliothek (ÖNB), die einen gesetzlichen Auftrag zur Archivierung digitaler Publikationen hat, und dem Österreichischen Staatsarchiv, das mit digitalen Akten konfrontiert ist, sind auch Einrichtungen, wie beispielsweise Fachhochschulen mit einer digitalen Abgabepflicht für Abschlussarbeiten oder universitäre Institute mit digitalen Forschungsdaten betroffen. Sie alle stehen vor der gleichen Herausforderung: dafür Sorge zu tragen, dass die gesammelten Daten auch in Zukunft noch benutzbar sind. Die Frage, welcher Zeithorizont anzunehmen ist, kann seriös nicht mit einem konkreten „Mindesthaltbarkeitsdatum“, sondern ehrlicherweise nur mit einem „so lange und so gut wie möglich“ beantwortet werden.

Digitale Langzeitarchivierung ist ebenso wie analoge Archivierung keine statische Angelegenheit, sondern ein aktiver Prozess. Dieser umfasst dabei viele einzelne Schritte: von der Auswahl des geeigneten Datenformats, der Speicherstrategie, der Sicherstellung der Datenintegrität über das Erstellen von Metadaten (deskriptive, technische, administrative), die den gesamten Lebenszyklus der digitalen Objekte dokumentieren, bis hin zu Migrationsmaßnahmen hinsichtlich Speichersystemen und Dateiformaten. Die Anschaffung geeigneter Hard- und Software, die diese Schritte unterstützen, ist dabei nur einer der Faktoren. Wesentlicher ist die mittelfristige Investition in ausgebildetes Personal, das das nötige Knowhow hat, die relevanten Prozesse durchführen zu können. Alles zusammen ergibt den Umstand, dass nicht alle Einrichtungen die Ressourcen im notwendigen Ausmaß aufbringen können.

Bereits mehrmals wurde in den letzten Jahren daher die Anregung an die Österreichische Bibliothekenverbund und Service GmbH (OBVSG) herangetragen, digitale Langzeitarchivierung als konsortialen Service zu etablieren. 2019 ergab sich dann die Chance, gemeinsam mit der Österreichischen Nationalbibliothek eine Ausschreibung für ein

digitales Langzeitarchivierungssystem durchzuführen, da das bisher in Verwendung stehende System am Ende seines Lebenszyklus angekommen war und somit die Anschaffung eines zeitgemäßen digitalen Langzeitarchivierungssystems dringlich wurde. ÖNB und OBVSG entschieden sich daher, gemeinsam ein konsortial einsetzbares System anzuschaffen und im Rahmen des Implementierungsprozesses den Dienst „Digitale Langzeitarchivierung“ aufzubauen. Die ÖNB sollte in diesem Kontext als erster Kunde fungieren, der durch die langjährige Beschäftigung mit der Thematik inhaltliche Schwerpunkte setzen konnte.

1.2 Ausschreibung

Die Vorarbeiten zur Ausschreibung begannen im April 2019 mit dem Erstellen der entsprechenden Unterlagen. Besonders inhaltliches Augenmerk lag dabei auf dem Prüfhandbuch, das die wesentlichen zu erfüllenden Kriterien des neuen Systems enthielt. Das Prüfhandbuch umfasste Muss- und Sollkriterien, wobei die Nichterfüllung eines Musskriteriums zum Ausschluss des Bieters führte. Da der gesamte Dienst konsortial aufgebaut werden sollte, beinhaltete das Prüfhandbuch ein eigenes Kapitel zum Thema „Mandantenfähigkeit“ und „konsortiale Funktionalitäten“. Auf Vorgaben zur konkreten technischen Umsetzung der geforderten Funktionalitäten wurde weitgehend verzichtet, um potenzielle Bieter nicht von vornherein auszuschließen. Anstelle von konkreten technischen Implementierungsvorgaben wurden Kennwerte zur Skalierbarkeit und Grundsätze der Skalierungsmechanismen als Kriterien festgelegt. Andererseits wurde inhaltlich darauf geachtet, dass potenzielle Bieter bereits Erfahrung in der Bibliotheks-, Archiv- und Museenlandschaft mitbringen, da die Anwendungsfälle der digitalen Langzeitarchivierung im Kontext der Gedächtnisinstitutionen andere sind als zum Beispiel im Finanzsektor. Zu den absoluten Musskriterien zählte unter anderem die Anwendung des OAIS-Standards (*Open Archival Information System*-Standards) als Referenzmodell, eine Userverwaltung mit Rollenvergabe, die Protokollierung des gesamten Lebenszyklus eines Objekts, die Automatisierbarkeit von Workflows via API und anderen Schnittstellen, Formaterkennung und -validierung, Unterstützung bei Risikoanalyse und Datenmigration inkl. Versionierung u.v.m.

Nach mehrtägigen Teststellungen, die während der Covid-19-Pandemie remote durchgeführt wurden, und ebenfalls per Videokonferenz durchgeführten Verhandlungsrunden erging der Zuschlag am 30. Oktober 2020 an die spanische Firma Libnova mit dem Produkt Libsafe. Libnova ist seit 2009 auf dem Markt und setzt einen klaren Schwerpunkt auf Produkte zur digitalen Langzeitarchivierung. Die

Software Libsafe selbst hat damit überzeugt, dass sie einerseits bereits um wesentliche Komponenten zur Langzeitarchivierung verfügt, andererseits offen genug ist, um eigene Entwicklungen an das System anzudocken.

1.3 Konsortiale Implementierung

Grundgedanke hinter den Überlegungen zur konsortialen Implementierung ist die Effizienzsteigerung für alle teilnehmenden Einrichtungen. Diese lässt sich im Wesentlichen auf die Bereiche Infrastruktur (Hard- und Software), Workflows und Personal (Know-how) umlegen.

Einige der grundsätzlichen Herausforderungen der digitalen Langzeitarchivierung sind unabhängig von der jeweiligen Institution und ihrer Anforderungen und Rahmenbedingungen. Eine konsortiale Langzeitarchivierungslösung bietet also die Chance, zumindest diese Teile der gemeinsamen Herausforderungen zu bündeln und somit gemeinsam zu lösen. Insbesondere Themen, die einen weiten Entscheidungsspielraum erlauben, fachlich nicht zwingend vorgegeben sind und damit eine relativ große Varianz bieten, stechen hier hervor. Die Ausgestaltung der Lieferpakete (SIPs = Submission Information Packages) sind hierfür ein gutes Beispiel. In den relevanten OAIS-Spezifikationen finden sich lediglich Vorgaben, dass diese Art von Paket für die Anlieferung vorhanden sein muss. Die konkrete Ausgestaltung der Implementierung ist aber im Wesentlichen nicht vorgegeben.

Für eine Einzelinstitution läge hier nahe ein Paketformat zu wählen, das dem eines Quellsystems entspricht. Eine einfache, aber langfristig problematische Variante, denn wir gehen davon aus, dass Quellsysteme mittelfristig ausgetauscht werden. Dies ist leichter umsetzbar mit einem stabilen, homogenen, aber möglichst nicht applikations-spezifischen Basisformat. Ferner sind die Schritte im Lade-prozess und der weiteren Verarbeitung abhängig von der Struktur der SIP. Hier kann auf zwei Ebenen gewonnen werden: Erstens, dass bei einer Änderung des Quellsystems viele Schritte der Prüfungen und Prozesse beim Laden nicht angepasst werden müssen, wenn ein definiertes Ladeformat unabhängig von der Quellapplikation vorliegt. Zweitens wirkt sich die Struktur des SIP-Formats auf die weitere Verarbeitung im System aus und müsste entsprechend angepasst werden, sollte das SIP-Format geändert werden. Ein konsortial definiertes Basisformat bietet hier also die Chance auf einen möglichst einheitlichen und ressourcenschonend zu wartenden Workflow.

Die Etablierung und Weiterentwicklung geeigneter Verfahren zur digitalen Langzeitarchivierung setzen ein spezielles Know-how voraus, das in vielen Einrichtungen nicht

immer im erforderlichen Ausmaß vorhanden ist. Je nach Ausgangssituation benötigt man Spezialist*innen für Text-, Bild-, Audio-, 3D- oder Videoformate, aber auch schwieriger zu behandelnde spezielle Datenformate, die beispielsweise bei wissenschaftlichen Forschungsprojekten anfallen. Dazu kommt notwendigerweise Wissen über Standards und deren konkrete Implementierung, über Metadatenformate und Langzeitarchivierungsprozesse. In allen Bereichen Expertise aufzubauen, wird für einzelne Institutionen nur schwer möglich sein, hier wird deutlich, dass Kooperationen auf allen Ebenen nicht nur ressourcenschonender, sondern auch zielführender sein werden. Neben dem reinen Austausch von Know-how und Best-Practices zur digitalen Langzeitarchivierung bietet eine Kooperation auch neue Möglichkeiten für die Erforschung und Nutzung von digitalen Inhalten und somit die Chance, die Wertschöpfung der digitalen Sammlungen zu erhöhen.

Die konsortialen Aspekte müssen nicht nur auf der Ebene des Software-Stacks abgebildet sein, sondern auch von der Infrastrukturebene unterstützt werden. Insbesondere das Zusammenspiel von Rechenknoten, verschiedenen Speicherorten und institutionellem Zugriff müssen gut abgestimmt sein. Diese Ebene wurde exemplarisch für die Österreichische Nationalbibliothek sowie für einen Testmandanten der OBVSG durchgeführt und bildet den Kern einer Mandantenumgebung. Aus den gewonnenen Erfahrungen zeigt sich, dass sowohl auf fachlicher Ebene von der Bündelung der Kompetenzen als auch auf Infrastrukturebene von einer konsortialen Implementierung profitiert wird.

2 Systemanforderungen

2.1 Systemarchitektur

Auch die von der Langzeitarchivierungslösung unterstützte Systemarchitektur war eines der wesentlichen Anforderungsbereiche des Ausschreibungskatalogs. Hier wurde insbesondere auf frühere Erfahrungen zurückgegriffen, um Systemtypen auszuschließen, die nicht die gewünschten Skalierungsmöglichkeiten und Datenanbindungen vorsehen. Dies soll gewährleistet werden, indem gefordert wurde, dass es sich zwingend um ein verteiltes System handeln muss. Insbesondere sollte hiermit sichergestellt werden, dass auf Änderungen des Ressourcenbedarfs temporär oder dauerhaft reagiert werden kann. Die Systemarchitektur soll so gestaltet sein, dass temporäre sehr rechenintensive Vorgänge vorübergehend mit entsprechend zusätzlichen Ressourcen versehen werden können. Beispiele für Prozesse, die dadurch leichter möglich sein

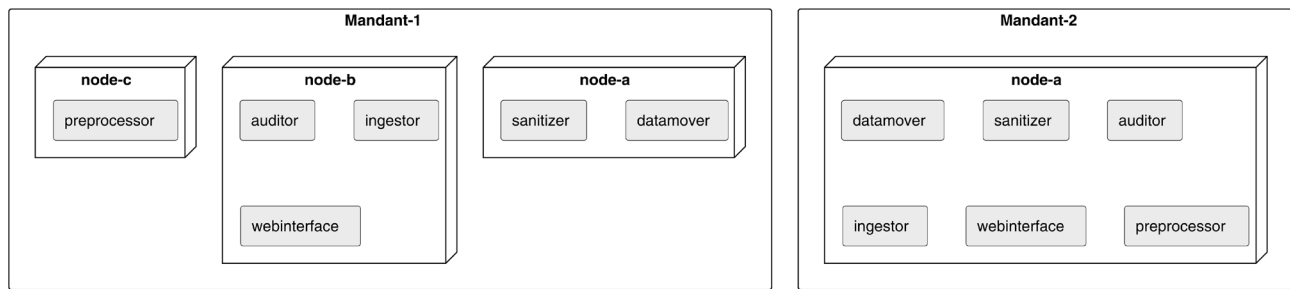


Abb. 1: Preservation diagrams

sollen, sind die Ableitung weiterer Repräsentationen, die spätere Formatidentifizierung oder auch einfache Operationen wie die Prüfsummenberechnung auf sehr großen Datenbeständen. Diese Prozesse sind zwar im OAIS-Standard Kernforderungen, aber werden in der Praxis nur sehr selten vorkommen. Hierfür permanent Serverkapazitäten vorzusehen wäre Ressourcenverschwendung. Daher muss gewährleistet sein, dass diese temporär ins System eingegliedert werden können. Ein horizontal verteilbares System soll hier eine gute Möglichkeit bieten, um auf Änderungen im Ressourcenbedarf zu reagieren. Neben nur temporären Ressourcenbedürfnissen spielt eine langfristige Erhöhung des Leistungsumfangs bei den Skalierungserfordernissen eine Rolle. Hier werden im Wesentlichen die gleichen Verteilmechanismen verwendet, jedoch wird die erhöhte Ressourcenzuordnung über längere Zeit beibehalten.

Um den Anforderungen an die Erweiterbarkeit des Systems begegnen zu können setzt Libsafe auf eine serviceorientierte Architektur, in der das Gesamtsystem auf Verarbeitungsknoten aufgeteilt wird. Diese Knoten stellen jeweils eine konfigurierbare Anzahl und Art von Diensten zur Verfügung. So können einzelne oder mehrere Knoten jeweils Ressourcen für Teilaufgaben der digitalen Langzeitarchivierung beisteuern. Diese Knoten können je nach Anforderung hinzugefügt und entfernt werden. Da alle Komponenten im Netzwerk miteinander kommunizieren, muss besonderes Augenmerk auf eine gute und für den Anwendungsfall günstige Netzwerktopologie gelegt werden. Wir gehen somit davon aus, dass Libsafe den oben beschriebenen Anforderungen gerecht wird.

2.2 Speicher

Ein konsortiales Langzeitarchivierungssystem muss verschiedene Speicheranbindungsstrategien erlauben. Im Zuge der Erstellung der Ausschreibungskriterien wurde von einem Spektrum an Bedürfnissen an die Speicheranbindung ausgegangen. Dieses sollte von Speicherorten

direkt bei dem verarbeitenden zentralen System im Rechenzentrum der OBVSG über lokale Speicherlösungen in den Institutionen, bis hin zu Speichersystemen an vertrauenswürdigen und geografisch und netzwerktechnisch von der Institution getrennten Orten reichen.¹ Zum einen lassen sich so in der Langzeitarchivierung gewünschte Sicherheitsaspekte umsetzen, zum anderen aber kann so dem Wunsch der vollen Verantwortlichkeit der Institutionen für ihre jeweiligen Daten nachgekommen werden.

Das konsortiale System ist so konfiguriert, dass über Netzwerk erreichbare Speicher über gängige Dateisystem-Protokolle, wie auch Objektspeicher, die über S3 angebunden werden, genutzt werden können. So können für die verschiedenen Sicherungskopien verschiedene Netzwerke, Netzwerkanbindungen und Speicherlösungen verschiedener Preiskategorien kombiniert werden.

2.3 Flexibilität

Ein konsortial organisiertes System wird mit einer Bandbreite an verschiedenen Materialien und zu archivierenden Strukturen konfrontiert sein. Um hier insbesondere den Bedürfnissen des Konsortiums gerecht zu werden, sind gut konfigurierbare Prozessschritte nötig. Das gewählte LZA-System Libsafe ist in dieser Hinsicht besonders flexibel. Die einzelnen Schritte wie Entpacken, weitere Vorverarbeitung oder Verbindung zu Drittsystemen sind weitgehend anpassbar. Hier sind vom Anbieter generische Mechanismen vorgesehen, um einzelne Tools einzubinden. Um diese Möglichkeit optimal zu nutzen, haben wir einige der Prozesse, insbesondere die Nutzung eines für alle Institutionen gleichen SIP-Formats und dessen Validierung, selbst implementiert.

¹ Verschiedene Speicherorte ist eine Forderung aus den Stufen der digitalen Langzeitarchivierung: <https://ndsa.org/publications/levels-of-digital-preservation/>. Zuletzt geprüft am 11.09.2023.

2.4 Mandantenfähigkeit

Die Mandantenfähigkeit wird in Libsafe durch Trennung der einzelnen Verarbeitungsknoten und Installationen je Mandanten gewährleistet. Während so je Mandant ein Aufwand entsteht, der nicht unter den Mandanten geteilt werden kann, wird dadurch ebenso eine komplette Trennung der Daten und der Verarbeitungsprozesse gewährleistet. Dieser Aspekt ist umso wichtiger, als es sich bei zu archivierenden Daten auch um personenbezogene und besonders geschützte Daten handeln kann. Dennoch können trotz dieser strikten Trennung auch Effizienzgewinne genutzt werden. Durch die flexible und modulare Gestaltung der Verarbeitungsworkflows können beispielsweise die für alle Mandanten entwickelten Werkzeuge an einer zentralen Stelle gepflegt und dann auf die Mandanteninstanzen gebracht werden. Hierbei ist auf Seiten der Installationen darauf zu achten, dass die Mandanten auf einem homogenen Softwarestand bleiben, um in der Langzeitarchivierung, aber auch in der Systembetreuung für jedes System die gleichen Werkzeuge nutzen zu können. Ferner müssen organisatorische Maßnahmen getroffen werden, damit getrennte Systeme als Gruppe mit den gleichen Werkzeugen gewartet werden können.

Als erste Mandantin wurde die ÖNB integriert und dabei wurde auch die Mandantenfähigkeit des Systems getestet. Die Dimensionierung erfolgte auf Basis der für die Ausschreibung vorgegebenen Gesamt- und Zuwachsvolumina. Es wurde schon zum Ausschreibungszeitpunkt deutlich, dass die OBVSG die Infrastrukturressourcen für den neuen Dienst erweitern wird. So zeigte sich im Bereich der vorhandenen Netzwerkhardware und Bandbreite, dass die vorhandene Kapazität zur ÖNB über das öffentliche Netz für das geplante Datenvolumen nicht ausreichend sein würde. Daher wurde eine Standleitung mit dedizierter Leitungskapazität und verschlüsselnden Übertragungsendpunkten zwischen den Rechenzentren der ÖNB und der OBVSG eingerichtet. Der für die OBVSG installierte Testmandant stellt hingegen eine Test- und Entwicklungsumgebung für die von der OBVSG entwickelten Werkzeuge dar. Diese werden in dieser Umgebung entwickelt und getestet, um anschließend in die produktiven Mandanteninstanzen übertragen zu werden.

2.5 Ingest und Submission Information Package (SIP)

Die Flexibilität des Systems ist gerade beim Import wesentlich, da die jeweiligen Quellsysteme, aus denen digitale Objekte angeliefert werden, jeweils ein eigenes Datenmo-

dell haben. Aus Gründen der Flexibilität, Systemunabhängigkeit und Nachnutzbarkeit für unterschiedliche Kunden haben wir uns entschieden, unterschiedliche Lieferquellen über ein homogenes SIP-Format einzuspielen. Im Idealfall wird unmittelbar im geforderten Format angeliefert, jedoch besteht auch die Möglichkeit, die Dateistrukturierung an das System zu delegieren. Diese Homogenisierung wird in der ersten Phase des Ingestprozesses durchgeführt. Da dieser optionale Schritt bereits Bestandteil des Ingest ist, wird er entsprechend über PREMIS-Events dokumentiert. Es entsteht hier also in keinem Fall eine undokumentierte Lücke zwischen Lieferung und Archivierung. Diesen Prozessschritt implementiert die OBVSG in Absprache mit dem jeweiligen Mandanten. Der weitere Prozess folgt für alle Pakete der regulären Verarbeitung des festgelegten Paketformats. In weiterer Folge fungiert die Struktur des so gebildeten SIP auch als Basis für das Archival Information Package (AIP).

Als Format für die SIPs wurde E-ARK² und für die Übertragung BagIt³ ausgewählt. Bei beiden handelt es sich um bereits etablierte Standards. Die konkrete Ausgestaltung des SIPs und die damit verbundenen Prozesse für den Ingest wurden exemplarisch im Rahmen der Implementierung für die Wirtschaftsuniversität Wien mit dem Ziel der Nachnutzung für weitere Mandanten entwickelt.

BagIt ist ein Format zur Verpackung und Übertragung von Daten in digitalen Archiven. Es ermöglicht, große Mengen an Daten sicher zu verpacken und übertragbar zu machen. Unter *bag* versteht man eine Sammlung von Dateien, die in einem bestimmten Format organisiert und gepackt sind. Das Format wurde an der California Digital Library sowie an der Library of Congress entwickelt und liegt aktuell in Version 1.0 als IETF-Standard vor.

E-ARK wurde im Rahmen des Projekts E-ARK entwickelt und definiert eine allgemeine Struktur für ein SIP-Format, das für eine Vielzahl von Archivierungsszenarien geeignet ist, z. B. für Dokumenten- und Bildsammlungen, Datenbanken oder geografische Daten. Die im Rahmen des E-ARK Projekts entwickelten Standards werden vom Digital Information LifeCycle Interoperability Standards Board (DILCIS Board) weiter betreut, sodass eine kontinuierliche Weiterentwicklung des Standards gegeben ist. Zusätzlich zur Spezifikation des E-ARK SIP Formats werden im DILCIS Kontext Tools entwickelt, die den Umgang mit E-ARK-SIPs erleichtern. So existiert ein Tool, das es abliefernden Institutionen ermöglicht, SIPs im E-ARK Format automatisiert zu erstellen.

² <https://earksip.dilcis.eu/>. Zuletzt geprüft am 11.09.2023.

³ <https://www.digitalpreservation.gov/documents/bagitspec.pdf>. Zuletzt geprüft am 11.09.2023.

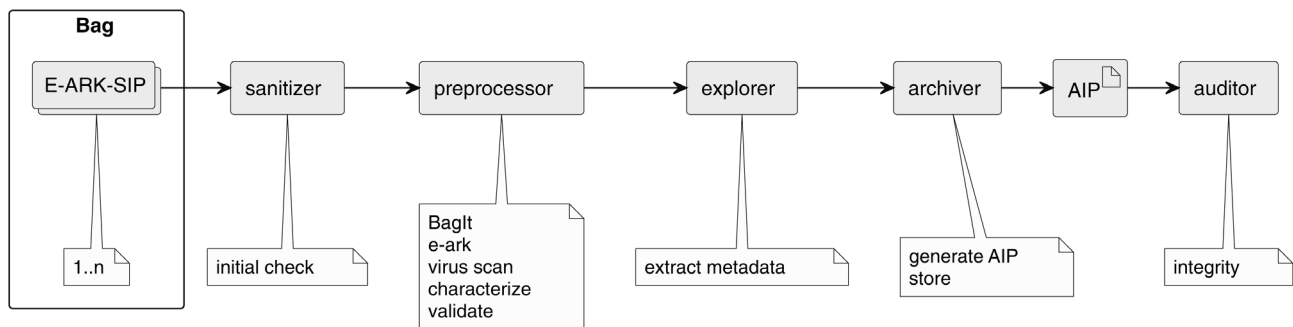


Abb. 2: Preservation diagrams

Auch wenn das LZA-System es erlaubt, Ingestprozesse manuell über die Oberfläche anzustoßen, wird ein automatisierter Import bevorzugt. So können auch große Datenmengen in einem Schwung in das System gebracht werden, was für Migrationsprojekte unumgänglich ist. Auch im laufenden Betrieb ist die Lieferung über automatisierte Prozesse weniger fehleranfällig als bei manuell gesteuerten Prozessen. Wie sieht nun ein Standard-Ingestprozess aus? Eine Grafik soll den Prozess illustrieren (s. Abb. 2).

Für das Pre-Processing wird eine standardisierte E-ARK Ordnerstruktur innerhalb des *bags* im BagIt-Format erwartet. Jeder dieser E-ARK Ordner enthält genau eine intellektuelle Einheit, die anhand des Ordnersnamens eindeutig identifizierbar ist. Der Ordnername kann z. B. ein DOI, URN oder ein anderer (idealerweise persistenter) Identifier sein. Beim Ingest findet nun ein erster Check auf strukturelle Integrität statt (ist alles, was erwartet wird, und nur das, im Paket enthalten?). Ferner wird ein Virenscheck und ein erster Prüfsummen-Check durchgeführt, um die Datenintegrität zu gewährleisten. Treten in diesem Stadium Fehler bei der Überprüfung auf, wird der Ordner insgesamt zurückgewiesen und wird nicht ins LZA-System aufgenommen. Ist die Überprüfung an dieser Stelle positiv abgeschlossen, finden weitere Prüfprozesse statt, die im System für eine Datenquelle konfiguriert werden können. Anschließend wird das übernommene Paket in das geforderte E-ARK Format umgewandelt, so nicht bereits vom Kunden entsprechend angeliefert. In den meisten Fällen wird nun die Dateiformaterkennung, hier die Ermittlung der PRONOM Unique Identifier, mittels DROID und eine Dateivalidierung mittels JHOVE durchgeführt.

An dieser Stelle werden auch solche Fehler in Dateien abgefangen, die auf den ersten Blick nicht auffallen, wie fehlerhafte oder fehlende Metadatenelemente in TIFF-Dateien, die die Datei für die Langzeitarchivierung ungeeignet machen, aber ein herkömmliches Bildbetrachtungsprogramm nicht daran hindern, die Datei problemlos anzuzeigen. Weiters werden beim Ingest Metadaten generiert

und umgewandelt, die anschließend auch im AIP archiviert werden können. Beispielsweise ist es somit möglich, alle technischen Metadaten des JHOVE-Outputs zu archivieren. Werden beim Ingest in das Archivsystem Fehler erkannt, wird das gesamte Objekt an dieser Stelle zurückgewiesen und der Fehler im Report dokumentiert. Im Idealfall kann nun das positiv geprüfte SIP in das Archival Information Package (AIP) überführt werden.

2.6 Archival Information Package (AIP)

Das Archival Information Package ist im OAIS-Modell nicht für den direkten Zugriff vorgesehen. Dennoch ist es der zentrale Dreh- und Angelpunkt im Gesamtprozess der digitalen Langzeitarchivierung. Das SIP ist die Quelle, aus dem das AIP gebildet und abgeleitet wird. Das AIP muss die Anforderung erfüllen, für sich selbst zu stehen, also alle relevanten Informationen zu dem Objekt, das heißt Nutz- und Metadaten zu beinhalten. Damit ist es auch der Ausgangspunkt für die Bildung des DIP, für dessen Bildung noch weitere Informationen, wie Daten zu Preservation-Events aus der Libsafe-Datenbank ergänzt werden. Im Zuge von Langzeitarchivierungsaufgaben, wie der Formatmigration oder der neuerlichen Validierung und Bewertung der Daten, wird dieses AIP als Ausgangsbasis verwendet. Aufgrund dieser für die Langzeitarchivierung zentralen Position ist das AIP auch das Objekt, das redundant an mehreren Speicherorten vorgehalten wird. Wie viele und welche Speicherorte hier definiert werden, wird je Mandant und Archivierungsplan festgelegt. Im Gegensatz zu den als flüchtig zu betrachtenden DIP und SIP ist das AIP also das Paketformat, das im OAIS im Zentrum der dauerhaften Erhaltung und Permanenz steht. Da das AIP aus dem SIP abgeleitet ist, folgt es im Kern dessen Strukturen. Hier profitiert ein großes Konsortialsystem von einer Vereinheitlichung des SIP, da auch das AIP dadurch mit den gleichen Werkzeugen verarbeitet werden kann.

Das AIP enthält auch alle Metadaten, welche vom Kunden via SIP angeliefert wurden. Diese können beispielsweise auch Informationen zur Archivtektonik im EAD-Format umfassen. Alle Metadaten werden als Dateien archiviert und sind als sogenannte *extended metadata* im LZA-System suchbar, die vorzugsweise in einem XML-basierten Format vorliegen. Aus diesen zusätzlichen Metadaten können wesentliche Merkmale extrahiert und über konfigurierte Indizes im System suchbar gemacht werden. Es ist daher nicht notwendig, alle Metadatenformate umfassend in ein gemeinsames Format zu mappen.

2.7 Zugriff auf Objekte und Dissemination Information Package (DIP)

Der LZA-Dienst selbst beinhaltet keine Rechercheoberfläche für Endbenutzer, diese Funktion muss über ein eigenes Discovery System oder ein Repositorium abgedeckt werden. Der Zugriff auf die archivierten Objekte soll im Sinne der Langzeitarchivierung nur unter kontrollierten und standardisierten Bedingungen erfolgen. Sie werden durch die Prozesse zur Bildung eines Dissemination Information Packages (DIP) sichergestellt. Diese Pakete können für die Belieferung von Drittsystemen wie Repositorien oder Discovery Systemen genutzt werden. Der Zugriff auf die archivierten Objekte über das Langzeitarchivierungssystem sollte auf jene Fälle beschränkt sein, wo im Repositorium vorgehaltene Objekte verloren oder beschädigt sind. In diesem Fall kann das archivierte Objekt aus dem LZA-System angefordert werden und steht dann außerhalb des LZA-Systems für den Kunden zur Verfügung. Um etwaige Zugriffsbeschränkungen auf das Objekt selbst sicherzustellen, werden so genannte Access Rights mit archiviert und mit ausgeliefert. Die Überprüfung und Kontrolle derselben ist wiederum Aufgabe des Repositoriums beziehungsweise jenes Systems, das Endbenutzern Zugang gewährleistet.

Das DIP ist je nach Anforderungsszenario flexibel konfigurierbar. So ist es denkbar, spezielle DIPs für den Re-Import in ein bestimmtes Zielsystem auszuliefern oder aber auch DIPs, die „nur“ ein Subset an Metadaten plus einer Repräsentation der archivierten Dateien beinhalten. Standardmäßig wird ein DIP ausgeliefert, welches die Daten in jenem Format beinhaltet, in dem sie auch angeliefert wurden.

Der Zugriff auf die Objekte und auf die operativen Bereiche des LZA-Systems werden durch ein konfigurierbares Rollen- und Berechtigungskonzept gesteuert. Dieses gewährleistet, dass nur berechtigte BenutzerInnen auf die für sie erlaubten Objekte Zugriff haben und je nach Rollenprofil entsprechende Langzeitarchivierungsprozesse steuern können.

2.8 Risikoanalyse und Reporting

Einer der wesentlichen Aspekte, mit denen Libsafe überzeugen konnte, ist die Risikoanalyse *out of the box*. Ausgehend von den IDs, die das System beim Ingest aus der PRONOM-Datenbank der britischen National Archives bezieht und anderen Faktoren, berechnet Libsafe einen *health index* (auf einer Skala von 1 bis 10), mit dem alle Dateien im System versehen werden. So ist es auf einen Blick möglich, nachzuvollziehen, welche Dateien im System bereits mit einem bekannten Obsoleszenz-Risiko behaftet sind und bei denen Handlungsbedarf bestehen könnte. Welche Handlungen letztlich aber gesetzt werden (Formatmigration, Emulation) liegt in der Verantwortung des Kunden, die Umsetzung findet dann im Rahmen eines gesonderten Projekts statt, bei dem die OBVSG gerne unterstützend zur Seite steht.

Neben der Risikoanalyse werden die wesentlichen statistischen Reports sowohl auf Mandanten- als auch konsortialer Ebene standardmäßig von Libsafe geliefert. Dazu zählen Speicherverbrauch, Anzahl der Objekte und Formate im System, Statistiken zu verschiedenen Jobs und viele weitere Aspekte. Neben den Standardreports beinhaltet der Lieferumfang von Libsafe auch die Business-Intelligence-Lösung *Metabase*,⁴ mit der vielseitige Abfragen zur Berichterstellung auf das System möglich sind.

2.9 Langzeitarchivierung als Service

Die OBVSG bietet Organisationen und Institutionen aus dem Wissenschafts- und Kultursektor ihr Know-how und die gesamte IT-Infrastruktur des digitalen Archivs als Dienstleistung an.

Der Service umfasst die folgenden Komponenten:

- Die Basis des Dienstes beinhaltet den technischen Betrieb des LZA-Systems und der dazugehörigen Infrastruktur. Dazu zählen: Betrieb der Server (Hardware, Betriebssystem), Lastverteilung der IT-Ressourcen, Betrieb der Datenbank, Betrieb der Applikation, Netzanbindung, Einspielen neuer Releases (Betriebssystem, Datenbank, Applikation).
- Archivspeicher: Entsprechend den Sicherheitsanforderungen werden die digitalen Archivpakete an mindestens zwei räumlich getrennten Standorten gespeichert, wodurch die Verfügbarkeit der Daten im Falle von Ausfällen oder Beschädigungen verbessert wird. Konkret kann eine Speicherung sowohl auf Speichersystemen

⁴ <https://www.metabase.com/>. Zuletzt geprüft am 11.09.2023.

des Kunden als auch bei der OBVSG erfolgen. In jedem Fall bleibt der Kunde vollumfänglich „Eigentümer“ der Daten.

- Etabliertes und auf internationalen Standards basierendes Archivsystem: Im Österreichischen Bibliothekenverbund wird seit jeher auf internationale Standards gesetzt. Im Bereich der digitalen Langzeitarchivierung setzen wir auf ein Archivsystem aufbauend auf dem OAIS-Referenzmodell, im Bereich der deskriptiven Metadaten auf MODS und für die Dokumentation von Archivierungsaktivitäten im Rahmen des Lifecycle Management auf PREMIS.
- Lifecycle Management und Langzeitverfügbarkeit digitaler Daten: Wie eingangs bereits erwähnt, verstehen wir unter Langzeitarchivierung einen kontinuierlichen und geordneten Prozess, der die Benutzbarkeit der archivierten Daten sicherstellt. Die dafür benötigten Prozesse sind unter anderem Risikomanagement, Metadatenmanagement, kontinuierliche Überprüfung der Datenintegrität und gegebenenfalls Formatmigration.
- Applikationsmanagement: Die OBVSG übernimmt den kompletten Betrieb inklusive Wartung und Support auf Applikationsebene.
- Entwicklung und Betreuung von Schnittstellen zum Datenimport und –export.
- Beratung und Unterstützung für die institutionseigene Langzeitarchivierungsstrategie.

Kunden ersparen sich somit Investitionen in Hard- und Software und können vom attraktiven Lizenzmodell und vom Know-how der OBVSG profitieren.

Für die Entwicklung des Kostenmodells wurden mehrere Modelle herangezogen, insbesondere Total Cost of Preservation (CDL-TCP)⁵ des 4C Project: Collaboration to Clarify the Costs of Curation,⁶ das als Grundlage diente, um alle Kostenfaktoren der digitalen Langzeitarchivierung zu berücksichtigen. Das Kostenmodell der OBVSG unterscheidet zwischen den Kosten, die anteilig pro Kunde an den Gesamtkosten zu tragen sind (konsortialer Anteil) und jenen Kosten, die jeweils kundenspezifisch anfallen. Letztere sind unter anderem abhängig vom Anteil der Dienstleistung durch die OBVSG, benötigtem physischem Speicher für AIPs, Anzahl der Lieferquellen, Komplexität der angelieferten Daten sowie dem zu archivierenden Datenvolumen. Der konsortiale Anteil wiederum beinhaltet Anteile der Anschaffungskosten der Systeminfrastruktur, an der Adminis-

tration und Wartung von Hard- und Software und an Personalkosten. Je nach in Anspruch genommenen Leistungen und den oben genannten Faktoren werden die Kosten individuell berechnet.

3 Ausblick/Fazit

Das erklärte Ziel, ein mandantenfähiges und gut in verschieden großen Institutionen verwendbares Langzeitarchivierungssystem anzuschaffen und als Dienstleistung anzubieten, spiegelte sich in allen Phasen der Einführung von Libsafe an der OBVSG wider. Bei der Implementierung wurde deutlich, dass es besonders lohnend war, einige Kerneigenschaften der Software bereits in der Ausschreibung festzuhalten. Die somit erreichte Flexibilität wird ein breites Spektrum an Bedürfnissen bedienen. Um die langfristige Wartbarkeit der hierfür nötigen Eigenentwicklungen zu gewähren, wurden in der Umsetzung organisatorische Strukturen geschaffen. Insbesondere ein homogenes SIP-Format ist hier hervorzuheben. Diese erlauben es, langfristig viele Mandanten einzubinden und gleichzeitig den Wartungsaufwand handhabbar zu halten, da viele der eigens entwickelten Komponenten auf unterschiedliche Mandanten übertragbar sind.

Nach zwei Implementierungen, in deren Rahmen jeweils Prozesse und Formate für die Datenübernahme und Speicherung entwickelt, getestet und operativ eingeführt wurden, stehen nun wesentliche Werkzeuge, Infrastrukturen und Prozesse zur Verfügung. Dies wird in Folge zu mehr Effizienz und daher Ressourceneinsparung auf allen Seiten führen. Die konsortiale Installation an der OBVSG und die Betreuung des Systems durch das Team der OBVSG stellt sicher, dass digitale Langzeitarchivierung für einzelne Einrichtungen leistbar wird. Und gerade, weil die Kernaufgaben der digitalen Langzeitarchivierung für alle betroffenen Einrichtungen in weiten Teilen ident sind, können auch Einrichtungen diesen Service nutzen und von diesem profitieren, die nicht Teilnehmerinnen am Österreichischen Bibliothekenverbund sind. Für die OBVSG wiederum bietet der neue Service die Chance, sich als Kompetenzzentrum Digitale Langzeitarchivierung zu etablieren.

⁵ <https://www.4cproject.eu/summary-of-cost-models/16-community-resources/outputs-and-deliverables/111-total-cost-of-preservation-cdl-tcp/>. Zuletzt geprüft am 11.09.2023.

⁶ <https://www.4cproject.eu/>. Zuletzt geprüft am 11.09.2023.

Autorin und Autor



Stefan Majewski
Die Österreichische Bibliothekenverbund und
Service GmbH
Raimundgasse 1/3
A-1020 Wien
stefan.majewski@obvsg.at



Bettina Kann
Die Österreichische Bibliothekenverbund und
Service GmbH
Raimundgasse 1/3
A-1020 Wien
bettina.kann@obvsg.at