



DIGITALE LANGZEITARCHIVIERUNG

Auch in der öffentlichen Verwaltung fallen Daten und Unterlagen immer häufiger in elektronischer Form an. Diese müssen nach dem Ablauf der Aufbewahrungsfristen den zuständigen Archiven übergeben und dort archiviert werden. Dies ist die Aufgabe der digitalen Langzeitarchivierung. Entsprechende Systeme befinden sich vielerorts gerade erst in der Entstehung.

| von ANDREAS RAQUET

Konnten sich staatliche und kommunale Archive in Zeiten papiergebundener Akten vorwiegend auf die fachliche Erschließung und Nutzbarmachung der übergebenen Unterlagen konzentrieren, sehen sie sich im Zeitalter digitaler Informationen auch erheblichen technologischen Herausforderungen gegenübergestellt.

INFORMATIONSVIELFALT BEHERRSCHEN

In der klassischen Archivierung werden Akten in der Regel in papiergebundener Form angeliefert. Elektronische Akten hingegen werden in den unterschiedlichsten Datenformaten angeboten. Darüber hinaus müssen Multimediadaten wie digitale Fotos, Audio- und Videostreams oder auch ganze Datenbanken übernommen werden. Auf Dauer lassen sich diese steigende Vielfalt und Heterogenität nicht beherrschen. Daher besteht die erste Herausforderung darin, einen beherrschbaren Satz an Dateiformaten zu definieren, der für die Einlagerung im digitalen Archiv geeignet ist. Dabei spielt nicht nur die Langlebigkeit der Formate eine Rolle, sondern auch die Unabhängigkeit von Softwareprodukten und -herstellern, um entsprechende Formate dauerhaft nutzbar zu machen.

Doch selbst wenn solche archivfähigen Formate identifiziert wurden, können die abliefernden Stellen nicht gezwungen werden, diese auch zu liefern. Vielmehr muss das Archivierungssystem selbst in der Lage sein, Fremdformate ohne Verlust von Primär- und Metadaten in die Archivformate zu konvertieren.

Es bedarf Prüf- und Konvertierungsfunktionen, die nicht nur die Formate bei Bedarf umwandeln, sondern auch Metadaten möglichst umfassend extrahieren und in das definierte Metadatenformat des Langzeitarchivs übertragen.

DAUERHAFTE SPEICHERUNG

Eine weitere Herausforderung betrifft die Speicherung der Daten. Diese erfolgt nicht mehr in einem klimatisierten Magazin, sondern auf elektronischen Datenträgern, und zwar – theoretisch – in großem Umfang, dauerhaft, sicher und bezahlbar. Theoretisch, denn leider erfüllt kein einziges heute verfügbares Speichersystem diese Anforderungen. Die erforderliche Speicherkapazität erreicht schnell Größenordnungen von hunderten Terabyte oder geht gar bis in den Petabyte-Bereich. Die große Anzahl benötigter Datenträger erhöht die Ausfallwahrscheinlichkeit und reduziert so die ohnehin nicht ausreichende statistische Lebensdauer (vgl. Tabelle rechts „Haltbarkeit von Datenträgern“). Daher sind aufwendige Redundanzmechanismen gefordert. Häufig werden Speicherhierarchien aus Plattensystemen und Bandlaufwerken eingesetzt, um den nötigen Grad an Performance und Sicherheit zu erreichen. Damit wird das Storage-System typischerweise zum größten Kostentreiber eines digitalen Langzeitarchivs. Die geforderte Langlebigkeit auf unbegrenzte Zeit lässt sich trotzdem nicht erreichen. Dazu bedarf es organisatorischer Prozesse, die dafür sorgen, dass die Storage-Systeme immer wieder ausgetauscht und Daten regelmäßig umgespeichert werden.

NUTZEN STIFTEN

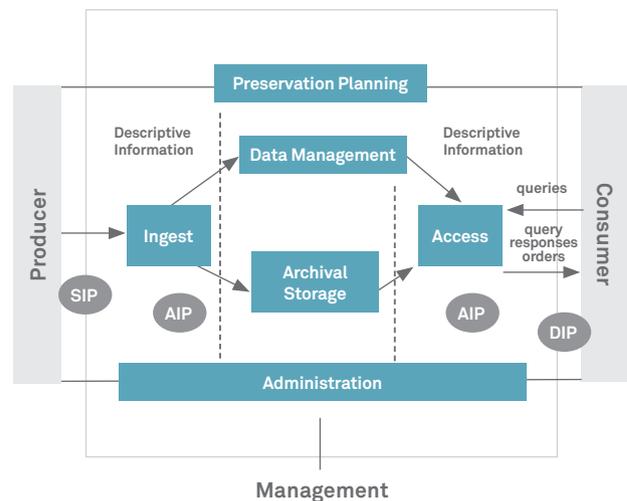
Da das eingelagerte Archivgut nur dann wieder genutzt werden kann, wenn es auch wieder auffindbar ist, werden sowohl eine Recherchemöglichkeit als auch eine Möglichkeit, die gefundenen Archivalien für den Nutzer bereitzustellen, benötigt.

Dies kommt der eigentlichen Archivarbeit am nächsten: Die Daten müssen erschlossen und geordnet werden, sodass sie über komfortable Recherchefunktionen wiedergefunden werden können. Das passiert auch heute bereits bei papiergebundenen Archivalien nicht mehr mittels klassischer Findbücher, sondern softwaregestützt mittels sogenannter Nachweissysteme. Dort werden Metadaten durch das Archiv gepflegt, sodass darüber ein Durchsuchen der Bestände möglich wird.

Digitale Archive bieten hierbei erhebliche Vereinfachungspotenziale. Vielfach sind Metadaten in den Quellsystemen der abliefernden Stellen vorhanden und müssen nur noch in das Archiv übertragen werden. So beinhalten viele Dokumentenformate Metadaten über den Autor oder Titel, Bildformate enthalten Daten über die Auflösung eines Digitalisats oder die Brennweite eines Fotos. Zum Teil werden Metadaten sogar in extrahierter Form bereitgestellt, z. B. im XDOMEA2-Format. Darüber kann

Datenträger	Geschätzte Haltbarkeit in Jahren	Herstellerangaben zum Langzeitverhalten	Fehlerkorrektur
HDD	5	MTBR (Mean Time Between Failures) – Fehlerrate POH (Power-on-hours) – Betriebsdauer AFR (Annualized Failure Time) – Ausfallrate	ja
SSD, Flash-Speicher	10	Datensicherheit bis 10 Jahre Anzahl der Lösch- und Schreibzyklen pro Zelle begrenzt auf 10.000 - 100.000	ja
CD ± RW, DVD ± RW	10-30	–	nein
CD ± R, DVD ± R	10-30	–	nein
CD-ROM, DVD, BD	10-50	–	nein
Magnetband	30-50	–	nein
UDO	30-50	Datensicherheit bis 30 Jahre	nein
HVD	50-100	Datensicherheit bis 50 Jahre	ja
Mikrofilm	500	–	nein

Quelle: nestor Handbuch, Kapitel 10.3



Quelle: Reference Model for an Open Archival Information System (OAIS), Ausgabe 2, Juni 2012, Seite 4-1

z. B. die Verortung im Aktenplan vollständig automatisch übernommen werden.

Die Schwierigkeit besteht auch hier in der Vielfalt der Formate, die an das Archiv geliefert werden. Die Automatisierungspotenziale können nur gehoben werden, wenn das digitale Langzeitarchiv ein hohes Maß an Flexibilität bis hin zu individuellen Programmierungen bei der Übernahme bietet.

DAS „BIG PICTURE“ VERSTEHEN

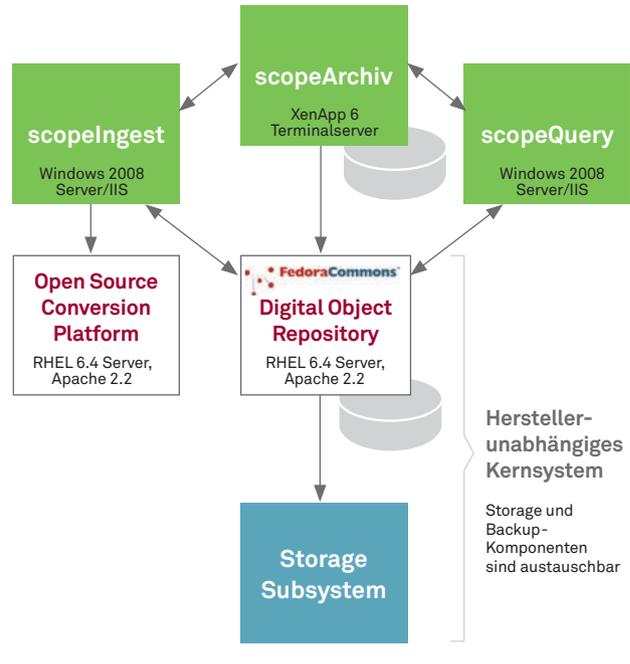
Bereits dieser Überblick zeigt: Digitale Langzeitarchivierung ist keine Aufgabe, die sich durch den Kauf einer Software von der Stange erledigen lässt. Vielmehr gilt es, das Thema als Ganzes zu verstehen und die Realisierung auf dieser Grundlage sorgfältig und von langer Hand zu planen. Eine sehr gute Ausgangsposition dafür bietet das Referenzmodell Offenes Archiv-Informationssystem (OAIS). Dieses beschreibt – auf grober fachlicher Ebene – die wesentlichen funktionalen Einheiten eines digitalen Langzeitarchivs, deren Funktionen und die darin verarbeiteten Informationsobjekte. Es enthält keinerlei Angaben zur Implementierung, aber es steckt den Rahmen ab und liefert ein gemeinsames Begriffsbild. Das Referenzmodell ist in der ISO-Norm 14721:2012 standardisiert.¹

Die funktionalen Einheiten des Referenzmodells können nun konkretisiert und durch entsprechende IT-Lösungen unterfüttert werden, die dann einzeln konzipiert und umgesetzt werden müssen. Umfangreiche Hilfestellung liefert das Kompetenznetzwerk nestor. Hierbei handelt es sich um einen Zusammenschluss von Archiven, Bibliotheken und Museen mit Experten zum Thema digitale Langzeitarchivierung. Das Netzwerk veranstaltet u. a. Workshops, publiziert Whitepapers und gibt das „nestor-Handbuch“ heraus – eine mehr als 600 Seiten umfassende Enzyklopädie zum aktuellen Stand der digitalen Langzeitarchivierung. Auch eine deutsche Übersetzung des OAIS-Referenzmodells ist bei nestor frei erhältlich.²

EIN BEISPIEL AUS DER PRAXIS

Auf dieser Grundlage wurde in den vergangenen Jahren unter Mitwirkung des städtischen IT-Dienstleisters it@M, des Direktoriiums der Stadt München sowie von msg das digitale Langzeitarchiv der Landeshauptstadt München konzipiert. Die Inbetriebnahme ist für das vierte Quartal 2014 vorgesehen.

Die Umsetzung basiert auf fachlichen Vorarbeiten der Arbeitsgruppe „Digitale Langzeitarchivierung“ im Stadtarchiv München, die bis in das Jahr 2006 zurückgehen. Am 25. Oktober 2011 wurde die Implementierung schließlich als IT-Vorhaben durch den Stadtrat genehmigt und mit der Konzeption begonnen. Die



Systemarchitektur digitales Langzeitarchiv München (vereinfachte Darstellung)

zentralen Softwarekomponenten wurden 2012/2013 europaweit ausgeschrieben. Den Zuschlag erhielt die Softwaresuite scopeOAIS des Schweizer Herstellers scope Solutions. Die Implementierung begann im Herbst 2013.

In der ersten Stufe wurden drei Fachverfahren als Pilot an das Archiv angebunden. In den kommenden Jahren werden dann sukzessive bis zu 40 weitere Fachverfahren aus 13 Referaten der Stadt München hinzukommen.

Wie im OAIS-Referenzmodell vorgesehen, werden die Informationen in Form von SIP (Submission Information Packages) an das Langzeitarchiv übergeben, konkret an die Komponente scopelngest. Diese erlaubt die Implementierung von Übernahmeworkflows für die von abliefernden Stellen übermittelten SIP. Dabei lassen sich flexibel verschiedene Validierungs- und Konvertierungsschritte konfigurieren und die Inhalte so umpaketieren, wie sie später im Archiv gespeichert werden sollten (AIP – Archival Information Package). Metadaten werden automatisch in die Metadatenstandards xISAD(G) und PREMIS transformiert. Über eine Open-Source-Konvertierungsplattform können jederzeit zusätzliche Validierungs- und Konvertierungsfunktionen in den Übernahmeprozess eingebunden werden.

1 Eine frei verfügbare Dokumentation kann unter <http://public.ccsds.org/publications/archive/650x0m2.pdf> abgerufen werden.

2 Nähere Informationen unter <http://www.langzeitarchivierung.de>



„Die von Beginn an enge und fruchtbare Zusammenarbeit von Stadtarchiv und IT, insbesondere das Verständnis der IT-Kollegen für archivische Arbeitsprozesse und für die archivischen Anforderungen an ein System zur digitalen Langzeitarchivierung, haben wesentlich zum erfolgreichen Verlauf des Projekts beigetragen.“

Dr. Daniel Baumann, Archivar, Stadtarchiv München

Gespeichert werden die Daten in Form von AIP über das digitale Repository Fedora Commons. Dieses verwaltet die Objekte mit samt ihren Metadaten und stellt mithilfe der Open-Source-Suchmaschine Solr eine Volltextsuche über das gesamte Archiv bereit.

Die Komponente scopeArchiv bietet leistungsfähige Funktionen zur Erschließung der Archivalien. Jede Verzeichnungseinheit, die jeweils ein AIP beschreibt, lässt sich frei in der Archivtekonik einsortieren. Zudem lassen sich beliebig konfigurierbare Metadaten mittels ebenfalls konfigurierbarer Formulare erfassen. Eine komfortable Suchfunktion erlaubt es, sowohl in den Metadaten als auch im Volltext zu recherchieren. Auf Wunsch können AIP direkt aus scopeArchiv heraus aus dem Storage-System abgerufen und in aufbereiteter Form zur Anzeige gebracht werden.

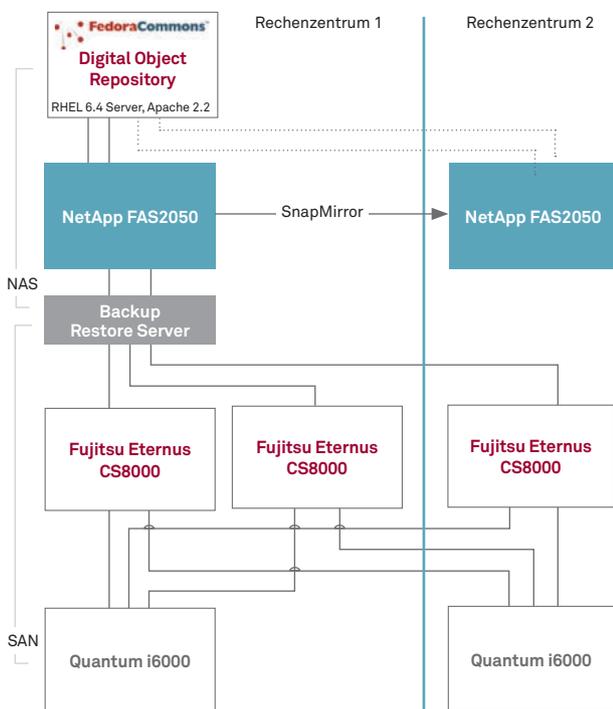
In der Pilotimplementierung wurden Dokumente und Bildbestände aus zwei bestehenden Fachverfahren über deren proprietäre Schnittstellen übernommen. Als dritte Schnittstelle wurde ein Ingest-Workflow für Übernahmen aus XDOMEA2.2-Aussonderungen implementiert. Darüber können zukünftig Fachverfahren angebunden werden, die diesen Standard unterstützen.

Das Kernstück der Lösung bildet ein hochgradig redundantes, individuell für das Langzeitarchiv konzipiertes Storage-System aus zwei via SnapMirror asynchron gespiegelten NetApp FAS2050 sowie einer auf drei Knoten verteilten Fujitsu Eternus CS8000 VTL. Physische Backups werden über Quantum i6000 Bandroboter auf LTO5- und LTO6-Bändern in dreifacher Ausfertigung gesichert. Das gesamte Storage-System ist aus Disaster-Recovery-Gründen über zwei Rechenzentren örtlich verteilt. Zum Schutz der gespeicherten Daten kommt der SnapLock-Mechanismus zum Einsatz. In der ersten Stufe ist das System auf die Speicherung von 5 TB ausgelegt – es wird die Grenze von 100 TB jedoch schon in wenigen Jahren überspringen.

Das Gesamtsystem wird auf insgesamt sechs virtuellen Servern unter RedHat Linux 6.4 und, wo erforderlich, unter Windows Server 2008 betrieben. Der Client-Zugriff erfolgt über Citrix XENApp 6 Terminalserver. Die Virtualisierung erfolgt mittels VMWare ESXServer 5.

Zur Drucklegung dieses Artikels befand sich das digitale Langzeitarchiv München im fachlichen Systemtest. Die Aufnahme des Wirkbetriebs ist für das vierte Quartal 2014 geplant. Für das Jahr 2015 ist die Anbindung von fünf weiteren Fachverfahren vorgesehen, die jedoch zum Teil noch um Funktionalität zur Aussonderung erweitert werden müssen.

Zeitgleich wird auch der funktionale Umfang des Systems weiter ausgebaut. Derzeit steht das Langzeitarchiv nur den abliefernden Stellen und dem Stadtarchiv selbst zur Verfügung. In der nächsten Ausbaustufe soll auch der Bereich „Access“ aus dem OAIS-Referenzmodell implementiert werden. Darüber können dann auch Nutzer des Archivs im Lesesaal oder über



Storage-Subsystem des digitalen Langzeitarchivs München

Der Begriff der digitalen Langzeitarchivierung

Der Begriff der Archivierung wird in der IT bereits mehrfach verwendet, z. B. im Bereich vom Storage Appliances oder im Dokumentenmanagement. Das OAIS-Referenzmodell beschreibt ein digitales Langzeitarchiv als „eine Organisation, in der Menschen und Systeme mit der Aufgabenstellung zusammenwirken, digitale Informationen dauerhaft über einen langen Zeitraum zu erhalten und einer definierten Nutzerschaft verfügbar zu machen“. Hier die zur Abgrenzung wichtigsten Charakteristika:

- Gesamtsystem bestehend aus Hardware, Software und Prozessen
- Einlagerung erst nach Ablauf der gesetzlichen Aufbewahrungsfrist
- Speicherung auf unbegrenzte Zeit
- Zweck ist die historische Nachvollziehbarkeit

Hochsicherheitssystem Langzeitarchiv

Die in einem digitalen Langzeitarchiv eingelagerten Daten sind in der Regel nicht mehr in anderen Systemen vorhanden und – nach längerer Aufbewahrung – auch nicht aus anderen Quellen rekonstruierbar. Bei einem Datenverlust sind sie unrettbar verloren. Daher ist der Schutzbedarf bezüglich der Verfügbarkeit, Integrität und Authentizität der Daten sehr hoch.

Aber auch bezüglich der Vertraulichkeit gelten höchste Anforderungen. Schließlich übernimmt das Archiv Daten aus fast allen Fachverfahren, darunter auch solche mit sehr sensiblen Daten, z. B. Strafregisterauszüge oder ärztliche Gutachten. Daher gehört das Langzeitarchiv zu den sicherheitskritischsten Systemen in der IT.

das Internet nach Archivalien suchen, diese bestellen oder auch direkt abrufen, sofern dies keine Schutzfristen oder Urheberrechte verletzt.

ZUSAMMENFASSUNG UND FAZIT

Die digitale Langzeitarchivierung ist weiterhin ein anspruchsvolles Thema. Fertige Lösungen gibt es dafür nicht. Für viele Aspekte stehen jedoch bereits Teillösungen zur Verfügung – sei es im Bereich der Nachweissysteme oder auch im Storage-Bereich – die zu leistungsfähigen Gesamtlösungen integriert werden können. Voraussetzung dafür ist ein konzeptionelles Gesamtverständnis der fachlichen Anforderungen und der technischen Möglichkeiten. Hier bieten das Referenzmodell OAIS sowie das Kompetenznetzwerk nestor wertvolle Hilfestellung.

Für die IT-technische Implementierung eines digitalen Langzeitarchivs muss eine Projektlaufzeit von ungefähr drei Jahren bis zur ersten Ausbaustufe veranschlagt werden. Fertig ist das Archiv allerdings nie, denn es werden dauerhaft neue Fachverfahren angeschlossen und immer wieder Pflegeaktivitäten zur Bewahrung des digitalen Archivguts anfallen. ●

ANSPRECHPARTNER – ANDREAS RAQUET

Lead IT-Consultant

Public Sector

- +49 151 16739288
- andreas.raquet@msg-systems.com

