H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.3

Kapitel 11.3

Speichersysteme mit Langzeitarchivierungsanspruch

nestor

nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, K. Huth
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.

URL für Kapitel 11.3 „Speichersysteme mit Langzeitarchivierungsanspruch“ (Version 2.3):
urn:nbn:de:0008-20100305225
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-20100305225

Gewidmet der Erinnerung an Hans Liegmann (†), der als Mitinitiator und früherer Herausgeber des Handbuchs ganz wesentlich an dessen Entstehung beteiligt war.

11.3 Speichersysteme mit Langzeitarchivierungsanspruch

Karsten Huth, Kathrin Schroeder und Natascha Schumann

Einführung

Dieser Beitrag gibt einen Überblick über Speichersysteme mit Archivierungsanspruch. Dabei stehen weniger die technischen Ausprägungen im Mittelpunkt, als vielmehr die allgemeinen Bedingungen, beispielsweise die Entstehungsgeschichte, denn oftmals sind diese Systeme aus Projekten zu konkreten Anwendungsszenarien heraus entstanden. Außerdem soll die generelle Strategie der Langzeitarchivierung dargestellt werden. Die Auswahl ist nicht vollständig, es wurde versucht, die gängigsten Systeme zu berücksichtigen.

Als Beispiele für Lösungen aus Projekten bzw. für konkrete Anwendungsfelder werden DIMAG (Digitales Magazin des Landesarchivs Baden-Württemberg), BABS (Bibliothekarisches Archivierungs- und Bereitstellungssystem der Bayerischen Staatsbibliothek), das Digitale Archiv des Bundesarchiv und PANDORA (Preserving and Accessing Networked Documentary Resources in Australia) vorgestellt. Aus dem Bereich der Institutional Repositories Software werden DigiTool von Ex Libris und Fedora (Flexible Extensible Digital Object and Repository Architecture) erläutert und abschließend Portico, kopal (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen) und LOCKSS (Lots of Copies Keep Stuff Safe) dargestellt.

DIMAG

DIMAG steht für das Digitale Magazin des Landesarchivs Baden Württemberg[16]. Es wurde konzipiert für verschiedene Formen von digitalen Archivalien, seien es elektronische Akten aus Behördensystemen, Statistiken aus Behörden oder Datenbanken. Die Software des DIMAG wurde vom Landesarchiv in Eigenregie entwickelt. Das System setzt auf offene Softwareprodukte (LINUX, PHP, MySQL und Apache), so dass die Architektur weitestgehend unabhän-

16 Keitel; Lang; Naumann: Konzeption und Aufbau eines digitalen Archivs: Von der Skizze zum Prototypen In: Erfahrungen mit der Übernahme digitaler Daten. Bewertung, Übernahme, Aufbereitung, Speicherung, Datenmanagement – Veröffentlichungen des Archivs der Stadt Stuttgart Bd. 99 Im Internet unter http://www.landesarchiv-bw.de/sixcms/media.php/25/aufsatz_labw_aufbau.pdf

gig von kommerziellen Anbietern ist. Gesichert werden die Daten auf einem RAID-Festplattensystem. Durch die Offenheit des RAID für einen Datentransfer auf andere Medien erhält sich das Archiv die Möglichkeit, den Speicher um eine Tapelibrary zu erweitern. Auch eine Konversion ausgewählter Datenobjekte für eine Belichtung auf Mikrofilm ist denkbar.

Das Produktivsystem steht in Ludwigsburg, Sicherheitskopien gehen an das Hauptstaatsarchiv in Stuttgart und das Generallandesarchiv in Karlsruhe. Das Speichersystem prüft stetig die Integrität und Authentizität der Daten anhand von gespeicherten Hashwertdateien. Abgelegt werden die Daten innerhalb des DIMAG in einem speziell geordneten Filesystem. Dieses Filesystem ist auch dann verfügbar, wenn das Archiv die Kontrolle über die laufende DIMAG-Software verlieren sollte. In dem Filesystem werden sowohl alle Metadaten als auch alle Inhaltsdaten gespeichert. Damit sind die Metadaten für den Fall eines Datenbankverlustes gesichert. Natürlich werden die für eine Recherche relevanten Teile der Metadatensätze in eine Datenbank importiert.

Das Filesystem des DIMAG baut sich aus festgelegten Knoten auf. Unter der Tektonik (= hierarchische Ordnungssystematik der Bestände eines Archivs) des Landesarchivs befindet sich der Knoten „digitales Objekt“, der wiederum mehrere Unterknoten enhalten kann. Diese Unterknoten werden Repräsentationen genannt. Jede Repräsentation enthält dieselbe Information, ist aber technisch verschieden (z.B. eine Repräsentation als Microsoft Office Format und die zweite Repräsentation als PDF/A Format). Repräsentation Nummer eins ist immer das Format, in dem das digitale Objekt an das Archiv übergeben wurde. Auf der Ebene „digitales Objekt“ protokolliert eine XML-Datei die technische Übernahme und die weitere Bearbeitung im Archiv. Unter einem Knoten „Repräsentation“ werden die primären Dateien abgelegt. Die Metadaten zu jedem Knoten und jeder Primärdatei werden jeweils in einer eigenen XML-Datei abgelegt. Alle Metadaten- und Primärdateien werden durch errechnete Hashwerte in eigenen MD5-Dateien gesichert.

Alle Rechte an der DIMAG-Software liegen beim Landesarchiv Baden-Württemberg. Bislang wird das System nur vom Landesarchiv betrieben.

BABS

Das Akronym BABS steht für das Bibliothekarische Archivierungs- und Bereitstellungssystem der Bayerischen Staatsbibliothek (BSB). Unter dem Namen wurde 2005 ein kooperatives Projekt zwischen der Bayerischen Staatsbibliothek und dem Leibniz-Rechenzentrum (LRZ) begonnen, das zum Ziel hatte,

eine organisatorisch-technische Infrastruktur für die Langzeitarchivierung von Netzpublikationen aufzubauen[17]. In BABS werden Retrodigitalisate aus der Produktion des Münchner Digitalisierungszentrums (MDZ) und seit 2008 auch die Bibliothekskopien aus der Public-Private-Partnership der BSB mit Google archiviert sowie auch elektronische Publikationen weiterer Produzenten – amtliche Veröffentlichungen, wissenschaftlich relevante Websites, freiwillige Ablieferungen kommerzieller Verlage etc.

Die Funktionalitäten Ingest, Data Management und Access werden einerseits von dem am MDZ entwickelten Electronic Publishing System ZEND (Zentrale Erfassungs- und Nachweisdatenbank) für die Retrodigitalisate, andererseits von dem Digital Asset Managementsystem DigiTool (siehe auch weiter unten) der Firma Ex Libris für elektronische Publikationen bereitgestellt.

Die Aufgabe des Archival Storage übernimmt das robotergesteuerte Archiv- und Backupsystem mit dem Softwarepaket Tivoli Storage Manager der Firma IBM am Leibniz-Rechenzentrum.

Derzeit (Stand: Januar 2009) wird in BABS ein Datenvolumen von 99,2 TB archiviert.

In einem weiteren Projekt (BABS2) soll die bestehende Infrastruktur nun zu einem vertrauenswürdigen und skalierbaren digitalen Langzeitarchiv ausgebaut werden, um den Herausforderungen rasch wachsender Datenmengen sowie gesetzlicher Verpflichtungen (Erlass über die Abgabe Amtlicher Veröffentlichungen an Bibliotheken, Pflichtstückegesetz) gewachsen zu sein.

Digitales Archiv

Das Digitale Archiv[18] ist die Archivierungslösung des Bundesarchivs. Potenzielle Nachnutzer sind alle Bundesbehörden.

Mit dem Digitalen Archiv können Daten und Metadaten aus disparaten Systemen der Behörden kontrolliert, fehlerfrei und effizient archivtauglich aufbereitet sowie in das Bundesarchiv überführt werden. Eine Pilotanwendung ist erfolgreich getestet worden, der Produktivbetrieb wurde im Oktober 2008 aufgenommen. Die Lösung wurde mit Hewlett Packard (HP) als Generalunternehmer und dem Partner SER geschaffen.

Der Gesamtprozess von der abgebenden Stelle bis in das Storage-System orientiert sich strikt an dem Standard DIN ISO 14721:2003 (Open Archival Information System - OAIS[19]). Technisch nutzt der Prozess zwei Komponenten,

17 BABS-Website: www.babs-muenchen.de

18 http://www.bundesarchiv.de/aktuelles/fachinformation/00054/index.html

19 http://public.ccsds.org/publications/archive/650x0b1.pdf

eine Workflowkomponente für die weitgehend automatisierte Eingangsbearbeitung (Standard-Archivierungsmodul - SAM) und eine Archivierungskomponente mit einer skalierbaren Storage-Lösung für die revisionssichere Speicherung des elektronischen Archivguts.
Kosten und Nutzen:

- Entlastung der Behörden von nicht mehr laufend benötigten Unterlagen
- Aufbau einer zentralen IT-Infrastrukur für die langfristige Speicherung
- komfortable Rückgriffsmöglichkeiten auf archivierte Unterlagen

PANDORA

PANDORA[20], das australische Web-Archiv, wurde 1996 von der Australischen Nationalbibliothek ins Leben gerufen und wird inzwischen von neun weiteren Bibliotheken bzw. Gedächtniseinrichtungen getragen. Es beinhaltet eine Sammlung von Kopien von Online Publikationen, die in Bezug zu Australien stehen. Dabei stehen v.a. Regierungsdokumente, wissenschaftliche Zeitschriften sowie Proceeding-Bände im Fokus. In der Regel sind die archivierten Publikationen frei zugänglich. Allen Ressourcen wird automatisch ein Persistent Identifier zugewiesen.

Archiviert wird nicht nur der Inhalt, sondern auch das „Look and Feel", sofern das möglich ist.

Die Architektur von PANDORA besteht aus dem Archivierungssystem PANDAS, dem Speichersystem DOSS, einem Bereitstellungssystem sowie einer Suchmaschine. Die Strategien zur Langzeitarchivierung beinhalten sowohl die technische Erhaltung durch Hardware und Software als auch, je nach Format, Migration und Emulation.

DigiTool

DigiTool[21] von Ex Libris ist ein Digital Asset Management Tool zur Verwaltung von digitalen Inhalten. Es wird von etlichen Institutional Repositories genutzt. Neben der Verwaltung von digitalen Objekten kann es auch zur Archivierung genutzt werden. Grundlage bildet das OAIS-Referenzmodell. Unterstützt werden Persistent Identifier und die Erstellung von Metadaten unter Verwendung

20 http://pandora.nla.gov.au/
21 http://www.exlibrisgroup.com/category/DigiToolOverview

des Metadatenstandards METS[22]. Mit DigiTool können unterschiedliche Dokumentenarten und Formate verwaltet werden sowie der Ingest-Prozess nach OAIS durchgeführt werden. DigiTool ermöglicht die Integration unterschiedlicher Sammlungen und bietet verschiedene Suchmöglichkeiten.

Im Januar 2009 wurde mit Rosetta[23] von Ex Libris ein eigenes Archivierungssystem gelauncht. Dieses ist direkt als Angebot für Nationalbibliotheken, Museen und weitere Gedächtnisorganisationen als Archivierungssystem gedacht. Das System wurde zusammen mit der Nationalbibliothek von Neuseeland entwickelt. Es hat eine verteilte Architektur und ist skalierbar. Kopien zum Gebrauch und die Dokumente für die Langzeitarchivierung werden getrennt gehalten. Es ist OAIS konform und orientiert sich an Richtlinien für vertrauenswürdige Archive.

FEDORA

Fedora[24] steht für Flexible Extensible Digital Object and Repository Architecture. Entwickelt wurde es an der Cornell University und an der University of Virginia Library. Zunächst als Projekt gefördert, wird Fedora seit 2007 als Non-Profit-Organisation geführt und ist als Open Source Software lizenziert. In erster Linie ist Fedora eine Repository Anwendung, die auch für Archivierungszwecke genutzt werden kann. Es bietet eine Metadatenbasierte Verwaltung der Daten und Unterstützung beim Ingestprozess.

Neben beschreibenden Metadaten werden auch technische Metadaten erfasst, die mittels JHOVE[25] und aus der Formatregistry PRONOM[26] gewonnen werden. PREMIS[27] und weitere LZA relevanten Metadaten können integriert werden. Fedora ist OAIS-konform und unterstützt die Migration. Alle Objekte erhalten Persistent Identifier und es erfolgt eine automatische Versionierung.

22 http://www.loc.gov/standards/mets/
23 http://www.exlibrisgroup.com/category/ExLibrisRosettaOverview
24 http://www.fedora-commons.org/
25 http://hul.harvard.edu/jhove/
26 http://www.nationalarchives.gov.uk/pronom/
27 http://www.loc.gov/standards/premis/

PORTICO

PORTICO[28] kommt ursprünglich aus dem wissenschaftlichen Bereich und hat sich zum Ziel gesetzt, wissenschaftliche e-Journale in Zusammenarbeit mit Verlagen und Bibliotheken dauerhaft zu archivieren. Gespeichert wird der Inhalt in der Form, in der er veröffentlicht wurde, nicht aber veränderte oder korrigierte Fassungen. Ebenso wenig werden Kontextinformationen, z.B. das „Look and Feel" gespeichert. In den Quelldateien können Grafiken, Text oder andere Ressourcen enthalten sein, die den Artikel ausmachen.
Nach Lieferung der Originaldatei wird diese in ein eigenes Format migriert. Dieses Format basiert auf dem „Journal Publishing Tag Set". Die Archivierungsmethode von PORTICO basiert in erster Linie auf Migration, das heisst, die Dateien werden, wenn nötig, in ein aktuelleres Format umkopiert. Zusätzliche Dienste werden nicht angeboten.
Portico dient als Sicherheitsnetz, das heißt, die Ressourcen werden nur im Notfall herausgegeben und sind nicht für den täglichen Gebrauch gedacht. Die Kosten werden einerseits von den Autoren und andererseits von den Bibliotheken getragen.

kopal

Im Rahmen des Projekts kopal[29] (Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen), an dem die Deutsche Nationalbibliothek, die SUB Göttingen, die GWDG Göttingen und IBM Deutschland beteiligt waren, wurde ein digitales Langzeitarchiv auf Basis des DIAS-Systems von IBM entwikkelt. Die im kopal-Projekt entwickelte Open Source Software koLibRI[30] (kopal Library for Retrieval and Ingest) ermöglicht das Erstellen, Einspielen und Abfragen von Archivpaketen (Objekt und zusätzliche Metadaten). Da die Arbeitsabläufe je nach Einrichtung variieren, erlaubt koLibRI, diese je nach Bedarf zu konfigurieren. Das Modell ist flexibel und bietet unterschiedliche Nutzungsmodelle. Da das kopal-System mandantenfähig ist, bietet es sich als zentrale Lösung für unterschiedliche Institutionen an. Der Kern, das DIAS-System, ist beim Dienstleister GWDG gehostet und wird mit Hilfe der koLibRI-Software von den Mandanten im Fernzugriff angesprochen. Zur Gewährung der Langzeitverfügbarkeit unterstützt das kopal-System durch entsprechende Metadaten

28 http://www.portico.org/
29 http://kopal.langzeitarchivierung.de/index.php.de
30 http://kopal.langzeitarchivierung.de/index_koLibRI.php.de

und einen Migrationsmanager die Dateiformatmigration.

Für das kopal-System bestehen drei verschiedene Nutzungsoptionen. 1. Als „kopal-Mandant" erhält eine Einrichtung einen eigenen Bereich des Archivsystems, den sie selbstständig verwaltet. Der Serverbetrieb bleibt allerding ausgelagert. 2. Eine Institution lässt ihre digitalen Daten durch einen „kopal-Mandanten" archivieren. 3. Eine Einrichtung installiert und konfiguriert ihr eigenes kopal-basiertes Archivsystem.

LOCKSS

LOCKSS[31] steht für Lots of Copies Keep Stuff Safe. LOCKSS ist eine Kooperation mehrerer Bibliotheken. Initiiert wurde das Projekt von der Stanford University. Inzwischen sind mehr als 150 Bibliotheken beteiligt, das heisst, sie haben eine LOCKSS-Box in Gebrauch. Das ist ein Rechner, der mit der (Open Source) LOCKSS- Archivierungssoftware ausgestattet wird. Die zu archivierenden Ressourcen werden über einen Webcrawler geharvestet. Die Inhalte werden regelmäßig mit denen der anderen Boxen abgeglichen. LOCKSS bietet Zugang zu den Daten und auch zu den Metadaten. Außerdem bietet es eine Verwaltungsebene, die die Mitarbeiter zur Erfassung und zum Abgleich nutzen können. Nachdem der Herausgeber dem Harvesten zugestimmt hat, gibt er die exakte Harvesting-Adresse an.

Die Boxen kommunizieren miteinander und im Falle eines Datenverlustes bei einer Bibliothek springen die anderen ein, um ein nutzbares Exemplar zur Verfügung zu stellen.

Der Zugriff auf die Ressourcen kann auf zwei Arten erfolgen: Entweder wird im Falle der Nichterreichbarkeit auf der Ursprungsseite auf eine archivierte Kopie weitergeleitet oder es wird eine Infrastruktur implementiert, die einen Zugang via SFX erlaubt.

LOCKSS ist format-unabhängig und für alle Arten von Webinhalten nutzbar. Neben dem Inhalt wird ebenso das „Look and Feel" gespeichert. Als Strategie zur Sicherung der Verfügbarkeit der Objekte wird Formatmigration genutzt.

Eine Erweiterung gibt es mit dem Projekt CLOCKSS[32] (Controlled LOCKSS), das als „Dark Archive" nur im Notfall Zugriff auf die archivierten Objekte erlaubt.

---

31 http://www.lockss.org/lockss/Home

32 http://www.clockss.org/clockss/Home