

Digitale Archivierung: Unser Ansatz für kleine und mittelgrosse Archive



Dieses Papier zeigt unser Konzept zur langfristigen Erhaltung digitaler Objekte für kleine und mittelgrosse Archive auf. «Langfristig» bedeutet in diesem Zusammenhang, dass Daten über viele Generationen von Hardware, Betriebssystemen und Anwendungen hinweg verfügbar gehalten werden sollen. Es kommen somit grundsätzlich andere Strategien zur Anwendung als z. B. bei der Erstellung von Backups, welche nur für eine vergleichsweise kurze Zeit aufbewahrt werden.



Der von uns entwickelte Ansatz zur Langzeitarchivierung stützt sich auf das Referenzmodell OAIS (Open Archival Information System, ISO 14721). Es wird ein Lifecycle-Modell verfolgt, d. h. die digitalen Objekte werden zu einem bestimmten Zeitpunkt aus den Produktivsystemen ausgekoppelt und an ein Langzeitarchiv abgeliefert, wo sie nicht mehr verändert werden. Das Langzeitarchiv übernimmt als Institution unter anderem die Aufgabe, mit geeigneten Strategien und Werkzeugen die Authentizität der digitalen Objekte zu schützen und zu dokumentieren.



Wo möglich und sinnvoll verwenden wir bei der digitalen Archivierung ähnliche Denkweisen und Vorgehen wie bei der Archivierung von Papierakten.



Phase 1: Was gehört ins Langzeitarchiv?

Meist ist nur ein kleiner Teil der im Tagesgeschäft benötigten Unterlagen langfristig aufzubewahren. Es müssen standardisierte Richtlinien und Checklisten für die Bewertung der Akten erstellt werden, vorzugsweise zusammenhängend mit einem Aktenplan. Oberstes Selektionsziel ist es, die Entscheidungsprozesse einer Organisation zu dokumentieren. Es sollen nur Unterlagen langfristig archiviert werden, die wirklich aussagekräftig sind, und Mehrfachüberlieferungen sollen ausgeschlossen werden. Mit klar dokumentierten Selektionsstrategien wird sichergestellt, dass die Nachwelt nicht mit Daten überflutet wird, dass am Ende aber auch keine wichtigen Unterlagen fehlen. Wir orientieren uns am folgendem Grobraster und bewahren Unterlagen auf, wenn sie...



- über beweiskräftigen, rechtlich relevanten Inhalt verfügen,
- Entscheidungsprozesse dokumentieren oder als Grundlagen für wichtige Entscheidungen dienen,
- Kompetenzen und Aufgabenbereiche belegen,
- die Kernprozesse einer Organisation nachzeichnen (Hauptaufgaben und -projekte).

Phase 2: Ablieferung und Migration

Die für die Langzeitarchivierung vorgesehenen Unterlagen müssen aus der produktiven Umgebung ausgekoppelt werden. In manchen Fällen ist der Aufwand dazu gering (z. B. für Dateien auf einem Fileserver oder eine Sammlung von CD), in anderen Fällen schwieriger (z. B. Daten aus einer Fachanwendung zusammen mit Metadaten so exportieren, dass es sich im Sinne der Diplomatik tatsächlich um «Records» handelt). Besondere Massnahmen erfordern in dieser Phase auch obsolete Medien, die kaum mehr gelesen werden können (8"-Disketten, alte Tapes, ...).

Falls es in der Produktivumgebung nicht längst geschehen ist, müssen die Daten in geeigneter Weise organisiert werden (entsprechend der Dossierbildung im Papierarchiv). Einzelne oder mehrere Dateien bilden nun logische Einheiten, sogenannte «digitale Objekte».

Zur langfristigen Aufbewahrung der Daten verwenden wir Formate, die zu diesem Zweck besonders geeignet sind. Wir migrieren die Daten gewissermassen «auf Vorrat». Für die Auswahl der Dateiformate spielen die Standardisierung bei anerkannten Gremien, eine offene Spezifizierung und weite Verbreitung eine wichtige Rolle. Je nach Datentyp kann dies z. B. TIFF, PDF/A, XML oder ODF (Open Document Format) sein. Wir beschränken uns auf möglichst wenige Dateiformate, so dass die Komplexität für spätere Migrationen gering bleibt.

Für jeden dieser Schritte werden ausführliche dokumentierende «Preservation Description Informations» geschrieben nach dem Standard des Premis Data Dictionary. So bleibt unsere Arbeit auch später jederzeit nachvollziehbar.

Phase 3: Einlesen und Erschliessen

Die aufbereiteten digitalen Objekte werden in ein Fedora-Repository aufgenommen (<http://www.fedora.info>). Fedora dient als Arbeitsinstrument zur Verwaltung der Nutz- und Metadaten und zur Erschliessung. Beim Einlesen erhält jedes Objekt eine eindeutige Bezeichnung, einen sogenannten «Persistent Identifier» (PID).

Die Archivpakete werden auf genau gleiche Weise erschlossen (verzeichnet und in ein Ordnungssystem integriert) wie Papierakten. Dabei richten wir uns nach den internationalen Verzeichnungsstandards ISAD(G) und ISAAR(CPF). In den Archivverzeichnissen werden auch die PID der digitalen Objekte aufgeführt, so dass ein eindeutiger Verweis auf die Nutzdaten besteht. Die Erschliessung wird soweit möglich automatisiert (z. B. Nutzung bereits bestehender oder extrahierbarer Metadaten), ist aber in der Regel weiterhin mit Handarbeit verbunden.

Mit diesen Schritten ist die Aufarbeitung der Nutzdaten und der dazugehörigen Metadaten für die langfristige Archivierung abgeschlossen. Ein digitales Objekt besteht danach aus den folgenden Elementen:

- den Nutzdaten
- den Metadaten über Inhalt und Provenienz (im EAD-Format)
- den Preservation Description Information (im Premis-Format)

Die übergreifenden Archivverzeichnisse im EAD-Standard liefern wir auch als XML-Dateien mit Stylesheets ab. Eine Konvertierung nach PDF oder Word ist jederzeit möglich.

Phase 4: Aufbau eines digitalen Langzeitarchivs

Nun müssen die erschlossenen Objekte gewissermassen in einen «digitalen Archivraum» verbracht werden. Das digitale Langzeitarchiv muss vertrauenswürdig sein, d. h. die Integrität, Vertraulichkeit und Verfügbarkeit der Daten müssen gesichert sein.

Um dieses hohe Ziel zu erreichen, braucht es einerseits einen organisatorischen Rahmen, der den Betrieb des Langzeitarchivs absichert. Andererseits müssen IT-Strategien vorhanden sein für die Replikation der Daten und gegebenenfalls für die Wiederherstellung bei einem Systemausfall. Je nach Grösse des digitalen Archivs empfehlen wir in kleinen und mittelgrossen Archivumgebungen zwei Vorgehensweisen:

- Bei einer grösseren Anzahl digitaler Objekte und einer intensiven Nutzung installieren wir ein eigenes Fedora-Repository vor Ort. Fedora ist eine robuste Middleware, welche auch mit grossen Datenmengen umgehen kann. Die Verwaltung des Langzeitarchivs gestaltet sich damit einfacher.
- Bei einer kleinen Anzahl digitaler Objekte und einer geringen Benutzungsintensität können die Daten direkt auf einem Fileserver abgelegt werden. Jedes Objekt wird in einen Ordner gepackt, welcher nach dem entsprechenden Persistent Identifier benannt ist. Dieses Vorgehen hat den Vorteil, dass zur Archivverwaltung keine spezielle Software installiert und gewartet werden muss. Dank Checksummen kann auch hier die Integrität der Daten jederzeit nachgewiesen werden.

Grundlagendokumente

- OAIS: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- Premis Data Dictionary: <http://www.oclc.org/research/projects/pmwg/premis-dd.pdf>
- Kriterienkatalog vertrauenswürdige digitale Langzeitarchive (nestor-materialien 8, Entwurf): <http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf>

Stand 5.2.2007 Tobias Wildi, t.wildi@docuteam.ch