Zeitschrift: Arbido
Herausgeber: Verein Schweizerischer Archivarinnen und Archivare; Bibliothek Information Schweiz
Band: 18 (2003)
Heft: 3

Artikel: Digitale Archivierung im Bundesarchiv - ein Erfahrungsbericht
Autor: Keller-Marxer, Peter
DOI: https://doi.org/10.5169/seals-769888

Digitale Archivierung im Bundesarchiv – ein Erfahrungsbericht

Peter Keller-Marxer
Schweizerisches Bundesarchiv Bern
Leiter der Fachstelle ARELDA und Gesamtprojektleiter des eGovernment-Projektes ARELDA

Das Schweizerische Bundesarchiv (BAR) archiviert seit 1982 digitale Unterlagen der Bundesverwaltung. Die heute im BAR archivierte digitale Datenmenge beträgt sieben Terabyte (TB; ein Terabyte sind ca. 1000 Gigabyte), was der Datenmenge entspricht, die sich auf 12 000 handelsüblichen CD-R speichern lässt. Im Jahr 2003 wird die Datenmenge um ca. weitere neun TB anwachsen; ab 2004 wird mit einer jährlichen Zuwachsrate von 20 TB pro Jahr gerechnet.

Für laufende Übernahme, Aufbereitung und Qualitätssicherung der digitalen Daten verfügt die Fachstelle ARELDA (Archivierung elektronischer digitaler Daten und Akten) im BAR über eine autonome, besonders abgesicherte Informatikinfrastruktur, die zusammen mit dem Informatik-Dienstleistungszentrum des Departements des Innern betrieben wird. Die Archivdaten sind auf Magnetbändern des Typs AIT-2 (Advanced Intelligent Tape der Firma Sony) gespeichert, die in mehreren Bandrobotern verwaltet werden.

Projekt ARELDA: Suche nach langfristigen Lösungen

In der Tagesarbeit der laufenden Archivierungen werden «Ad hoc»-Lösungen erarbeitet, bei denen mit relativ grossem personellem und technischem Aufwand in jedem Einzelfall Datenformate und Schnittstellen mit den abliefernden Bundesstellen vereinbart werden. Die Aufbereitung, Qualitätssicherung und Integration der Daten ins Archiv folgen Analogien und aus Erfahrungen gewonnenen «best practices» vorangegangener Übernahmen. Dabei existieren jedoch weder eine umfassende Datenarchitektur noch standardisierte Verfahren, klare Normen oder automatisierbare Prozesse.

Mit dieser Verfahrensweise können zwar kurzfristig Überlieferungslücken verhindert werden; bei stark wachsender Menge und Heterogenität der Datenbestände lässt sich damit aber langfristig kein technisch, personell und finanziell plan- und verwaltbares digitales Archiv führen.

Das Bundesarchiv hat deshalb im Jahr 2000 den Anlauf genommen, mit dem fachlich wie finanziell ambitiösen Projekt ARELDA[1] Lösungen zu entwickeln und zu institutionalisieren, welche die digitale Archivierung im BAR langfristig sichern sollen. Dies bedeutet, dass die zu implementierenden Methoden, Normen, Konzepte und Prozesse auch dann Bestand haben und die Planbarkeit und Finanzierbarkeit des Archivs sowie den kontinuierlichen Datenunterhalt garantieren sollen, wenn die dem System unterliegende technische Basis erwartungsgemäss alle zehn Jahre obsolet und ausgewechselt wird.

Das Projekt ARELDA ist der gleichnamigen Fachstelle des BAR angegliedert und gehört zu den fünf Schlüsselprojekten der eGovernment-Strategie des Bundes[2]. Die erste Etappe 2001–2004 des bis 2008 laufenden Projektes ARELDA wird hauptsächlich durch die Interdepartementale Koordinationsgruppe Informationsgesellschaft[3] (KIG) des Bundesrates finanziert.

Die Situations- und Schwachstellenanalyse, welche der Lancierung des eGovernment-Projektes ARELDA vorausging, führte zu einer methodischen Neuorientierung gegenüber den bisherigen, seit 1992 unter demselben Namen geführten Projektarbeiten. Dies betrifft vor allem die Feststellung, dass sich fehlendes oder ungenügendes Informatikwissen im Archivbereich in zweifacher Hinsicht als wesentliches Hemmnis bei der Lösungssuche erweist. Einerseits verhinderte die fehlende informatiktechnische Fachkompetenz eine Wirkungsanalyse auf der konzeptionellen Ebene: Technische Lösungsvorschläge externer Beratungsfirmen konnten nicht nach ihrer Wirksamkeit oder Tauglichkeit zur Umsetzung der archivischen Ziele beurteilt werden, welche primär aus dem bisherigen Verständnis papiergebundener Archive heraus formuliert wurden.

Entsprechend frustrierend verlief die «marktübliche» Zusammenarbeit mit externen Auftragnehmern aus der IT-Branche. «Marktüblich» heisst, dass es in der Regel nicht nötig ist, dass der Auftraggeber ein Verständnis für die technische Lösung seines Problems haben muss, solange er die fachlichen Anforderungen genügend präzis definieren kann. Genau dies war aber im Fall der digitalen Langzeitarchivierung nicht möglich, da die Probleme selber ursächlich auf einer technischen Ebene liegen und die im «konventionellen» Archivbereich gewohnten Konzepte von Verstehbarkeit, Zugänglichkeit, Bestandserhaltung, Konservierung, Integrität, Authentizität, Vermittelbarkeit etc. im digital-technischen Kontext weitgehend erst neu verstanden und definiert werden müssen.

Neben einer fehlenden *Wirkungsanalyse* hat sich somit die Unfähigkeit zur *Problemanalyse* auf der ursächlichen Ebene als zweiter Hemmfaktor erwiesen. Es war deshalb ein strategischer Entscheid, die Vorgehensmethodik in ARELDA neu so zu orientieren, dass ein Schwerpunkt in der Erarbeitung und Institutionalisierung einer eigenen Fachkompetenz «Archivinformatik» des BAR liegt. Dies hat dazu geführt, dass die Hälfte der heute in der Fachstelle ARELDA beschäftigten acht Personen Berufsinformatiker sind.

Die Erfahrungen mit dieser Politik sind durchwegs positiv. Im Bereich der *Problemanalyse* hat sich vor allem die Möglichkeit eines eigenen experimentellen Prototypings (Software und Pilotsysteme) als wesentliches Instrument zur Reduktion von Projektrisiken und -kosten erwiesen, indem anhand von eigenen Prototypen kritische Teilsysteme auf ihre Machbarkeit und Plausibilität hin untersucht, Kernfunktionalitäten schrittweise in Pilotsystemen entwickelt, konsistente und vollständige Anforderungen definiert und «vorgedachte Lösungen» für externe Auftragnehmer spezifiziert werden können.

Im Bereich der *Wirkungsanalyse* wurde die Erfahrung gemacht, dass es durchaus gerechtfertigt ist, von der Unverzichtbarkeit einer fachspezifischen Archivinformatik zu sprechen. Ein wachsendes IT-Marktsegment bietet heute vollmundig Archivlösungen an, die jedoch mit der Aufgabe der digitalen Langzeitarchivierung, wie sie sich etwa

---

1 *http://www.bundesarchiv.ch/webserver-static/docs/d/arelda_expose_0301_d.pdf*

2 *http://www.admin.ch/ch/d/egov/egov/strategie/strategie.html*

3 *http://www.admin.ch/ch/d/egov/egov/kig/kig.html* und *http://www.infosociety.ch*

einem Nationalarchiv stellt, wenig zu tun haben. Vielmehr orientieren sich diese Lösungen mehrheitlich an einem Zeithorizont von zehn Jahren (typische gesetzliche Aufbewahrungsfrist für Buchhaltungsunterlagen privater Firmen) und setzen voraus, dass der Käufer die vollständige Kontrolle über die Unterlagenproduktion besitzt, also die Typen und Parameter der zu archivierenden Unterlagen bereits bei ihrer Erzeugung auf die Funktionalitäten des anzuschaffenden Archivierungsprodukts ausrichten kann.

Für die Produkte kann in der Regel nicht plausibel gemacht werden, wie heterogene Archivbestände – vernünftigerweise spricht man heute von Daten in der Grössenordung von 100 TB – nach der typischen «Lebenszeit» eines herstellerspezifischen Produkts (15 Jahre) ohne Verlust an Authentizität und Information *und* mit finanzierbarem Aufwand in neue Systeme überführt werden können.

Dabei sollte nicht vergessen werden, dass die «technologische Kurzlebigkeit» im Informatikmarkt *den* wesentlichen Ertragsfaktor darstellt und das Denken in Zeiträumen von mehr als zehn Jahren in diesem Bereich aus Erfahrung hoch spekulativ ist und deshalb als unattraktiv gilt.

Es ist aber gerade die Beschäftigung mit diesen Fragen, die das Fachspezifische der Archivinformatik ausmacht und ermöglicht, die Wirksamkeit von Lösungen zu beurteilen, Lösungen kontinuierlich zu betreiben und Technologiefolge- und Risikoabschätzungen für das langfristige Management eines digitalen Archivs zu erstellen.

Charakteristika digitaler Archivierung

Die Diskussion um «Verfügbarkeit und Zuverlässigkeit» von informatik-technologisch bereitgestellter Information (information availability & reliability), die im IT-Bereich sehr en vogue ist, lässt sich aus der Welt der «lebenden Informationen» zwanglos in den Archivbereich fortsetzen: Archivierung bedeutet die zeitlich unbeschränkte Verfügbarkeit von Informationen, d.h. die Erfüllung von vier Grundanforderungen:

- *Persistenz:* Fähigkeit eines digitalen Archivobjekts, länger zu existieren als jede es umgebende technische Ausrüstung. «Existieren» meint hier implizit auch die Zugänglichkeit (Lesbarkeit) des Objektes.
- *Physische Integrität:* sichere und unversehrte Aufbewahrung, d.h. Vollständigkeit und Unbeschädigtheit eines digitalen Archivobjektes auf Bit-Ebene über die Zeit.
- *Authentizität:* «intellektuelle Integrität»; Authentifizierung (der Autorenschaft und Provenienz) und Zuverlässigkeit (der enthaltenen Evidenz). Dies beinhaltet als Voraussetzung auch die intellektuelle Verstehbarkeit.
- *Kontinuität:* gleichzeitige, korrelierte (d.h. zueinander in Bezug gesetzte) und kontinuierliche Präsenz der obigen Charakteristika in einem parametrisierten (d.h. qualitativ und quantitativ definierten und damit messbaren) Prozess über die Zeit.

Allerdings führt der Versuch einer deterministischen, durch quantitative Parameter messbaren Auslegung dieser Charakteristika für digitale Unterlagen zu heute noch ungelösten Fragestellungen[4, 5, 6, 7]. Die Probleme entstehen massgeblich dadurch, dass die Konzepte nicht unabhängig sind: Persistenz bedingt Integrität (aber nicht umgekehrt), aber keine Authentizität; Authentizität bedingt Integrität, aber nicht Persistenz etc.

Das Konzept «Kontinuität» birgt durch die zeitliche Komponente einen weiteren Komplexitätsgrad, wie folgendes triviales Beispiel zeigt. Im Gegensatz zu einem authentischen, handschriftlich unterzeichneten Papierdokument, welches 400 Jahre im Archiv lagert, ist die Authentizität eines mit einer heute gültigen digitalen Signatur versehenen digitalen Dokuments nicht konstant: Eine digitale Signatur (deren rechtliche Gültigkeit ohnehin auf wenige Jahre beschränkt ist) und Verschlüsselungen werden bei der Archivierung entfernt oder mindestens bei der ersten Migration/Konversion der Unterlage zerstört. In diesen Fällen muss das Archiv die «originale» Authentizität durch eine ersetzen, für deren Gleichwertigkeit es selber einsteht: Archive übernehmen die Rolle eines Substituts, der originale Authentizität digitaler Unterlagen immer neu definieren, erzeugen und gewährleisten muss.

Die Auslegung aller Konzepte muss grundsätzlich die technologischen Obsoleszenzen berücksichtigen:

- Speicher-, Datei- und Datenformate kommen innert weniger Jahre aus der Mode und werden obsolet.
- Herstellerspezifische Software zum Verarbeiten/Lesen der Daten ist innert weniger Jahre nicht mehr verfügbar. Neuere Versionen derselben Software können mit Vorgängerversionen erzeugte Daten nicht mehr fehlerfrei lesen.
- Datenträgertechnologien, mit denen Daten gespeichert werden, und die dazugehörigen Lesegeräte verschwinden innert weniger Jahre vom Markt.

Obsoleszenzen verhindern es in jedem Fall, dass digitale Unterlagen in einer völlig original-identischen Form über mehr als wenige Jahre erhalten werden können, weil dazu auch die sie erzeugenden Software- und häufig auch Hardware-Umgebungen original miterhalten und -betrieben werden müssen. Sie erzwingen vielmehr, dass die Unterlagen vor der Archivierung in der einen oder anderen Weise für die Langzeitarchivierung «präpariert» werden, womit zwingend ein Verlust an Authentizität und Information verbunden ist. Damit stellt sich die Frage nach der Definition der genannten Konzepte von Persistenz, Integrität, Authentizität und Kontinuität in dieser präparierten Archivumgebung, und zwar grundsätzlich für jeden Unterlagentyp unterschiedlich.

Prinzipien der Lösungsentwicklung ARELDA

Die Wahl der grundlegenden (strategischen) Prinzipien für die Lösungsentwicklung in ARELDA hat sich an folgenden Fragen und Feststellungen orientiert:

- Unter welchen Bedingungen lassen sich die erwähnten vier Konzepte am ehesten definieren sowie messbar und langfristig kontrollierbar implementieren?
- Vorgehen im Bewusstsein, ein unvollständig gelöstes Problem an die nächste Generation weiterzugeben, jedoch in (aus heutiger Sicht!) handhabbarer und finanzierbarer Form. («Stafetten-Prinzip»)
- Wo können fachliche Synergien mit bestehenden Lösungsansätzen in verwandten Gebieten gefunden werden?
- Unter welchen Bedingungen lässt sich ein plausibles Risikomanagement[8] betreiben?

Zum letzten Punkt ist anzumerken, dass es grundsätzlich drei mögliche Arten gibt, digitale Archivbestände zu verlieren:

- *physischer Verlust* durch hohes Risiko von Fehlmanipulationen oder Defekten;
- *logischer Verlust* durch irreversible Obsoleszenzen;

[4] Authenticity in a Digital Environment; Council on Library and Information Resources, Washington, D.C. (Hrsg.), 2000, *http://www.clir.org/pubs/reports/pub92/pub92.pdf*

[5] Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust; Clifford Lynch, CLIR, 2000, *http://www.clir.org/pubs/reports/pub92/lynch.html*

[6] Preserving the Authenticity of Contingent Digital Objects; Anne J. Gilliland-Swetland and Philip B. Eppard, D-Lib Magazine Volume 6, Number 7/8, 2000, *http://www.dlib.org/dlib/july00/eppard/07eppard.html*

[7] Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report, Research Libraries Group, Mountain View, CA, May 2002, *http://www.rlg.org/longterm/repositories.pdf*

[8] Risk Management of Digital Information, Gregory W. Lawrence et al., Council on Library and Information Resources, Washington, D.C, 2001, *http://www.clir.org/pubs/reports/pub93/pub93.pdf*

- *operativer Verlust* durch inflationären, nicht mehr finanzierbaren Aufwand für manuelle Interventionen bei Verwaltung und Migrationen durch zunehmend heterogenere und technisch unvollständig dokumentierte Datenbestände.

Der operative Verlust steht erfahrungsgemäss eindeutig im Vordergrund. Seine Eintretenswahrscheinlichkeit hängt direkt vom zur Verfügung stehenden Budget ab, ist jedoch in jedem Fall erheblich.

Das nach der Analyse dieser Fragestellungen gewählte Prinzip lässt sich als «applikations-invariante Archivierung» umschreiben: Digitale Langzeitarchivierung über technologische Generationen hinweg bedingt, dass die Unterlagen

- von den sie erzeugenden, spezifischen Umgebungen (Software, Hardware, Speicher-, Daten- und Dateiformate) gelöst werden (unter Inkaufnahme von Verlusten an Information und Authentizität)
- und in offene, standardisierte, möglichst generische und vor allem vollständig dokumentierte Datenformate und Umgebungen überführt
- und dort in möglichst langen Migrationszyklen (mindestens 15 Jahre) unterhalten werden können
- und Funktionalität (Software, Hardware) prinzipiell nicht archiviert (sondern nur dokumentiert) wird.

Dieser Ansatz wird heute weltweit von vielen Institutionen propagiert und gilt als bester Kompromiss zwischen pragmatischem Vorgehen und langfristiger technologischer Verfügbarkeit. Die Ausbalancierung von Verlusten und generischer Umgebung stellt allerdings hohe Anforderungen.

Dieser hier nur oberflächlich dargestellte Ansatz setzt jedoch eine sorgfältige Sicherung einer archivischen Datenqualität[9] des potentiellen Archivgutes voraus:

- Auszeichnung des langfristigen Wertes der Informationen: Entwicklung von griffigen prospektiven Bewertungsinstrumenten gemäss den relevanten (rechtlichen, wirtschaftlichen, wissenschaftlichen, historischen, militärischen etc.) Kriterien;
- Identifikation der essentiellen, bedeutungstragenden Elemente;
- Erhebung aller für das langfristige intellektuelle Verständnis und technische Datenmanagement nötigen Metadaten.

Dabei sehr wesentlich: Archivische Datenqualität kann bei digitalen Unterlagen nur am Anfang des Lebenszyklus der Unterlagen gesichert werden.

«Stakeholders» der digitalen Archivierung

Verschiedene Interessengruppen haben das Fehlen von Lösungen und die Dringlichkeit des Problems der digitalen Langzeitarchivierung festgestellt. Zu den wichtigen «Stakeholders» der digitalen Langzeitarchivierung gehören heute neben den staatlichen Archiven und Bibliotheken vor allem die pharmazeutische Industrie und die naturwissenschaftlichen Science Data Centers. Die Pharmaindustrie sieht sich durch die weltweit relevante Zulassungspraxis der US-amerikanischen Food and Drug Administration (FDA) unter massivem Druck: Die gesetzliche Bestimmung «21 CFR Part 11»[10] von 1997 legt sehr weitgehende Anforderungen an die Erzeugung, das Records Management, die Einreichung und die authentische Langzeitarchivierung von elektronischen Unterlagen für Zulassungen im Life Sciences Bereich fest.

Science Data Centers verwalten und archivieren naturwissenschaftliche Messdaten aus Raumfahrt, Geosatelliten-Missionen, Experimentalphysik, Ozeanographie, Meteorologie etc., typischerweise in der Grössenordung von mehreren Petabytes (mehreren 1000 TB), mit Zuwachsraten von bis zu mehreren Terabytes pro Tag. Es besteht ein vitales Interesse, solche Daten langfristig verfügbar zu halten: ihre Erzeugung ist mit Milliarden-Investitionen verbunden und in der Regel nicht wiederholbar (z.B. Klimadaten). Zudem sollen solche Originaldaten auch in 50 Jahren oder mehr mit neuen theoretischen Modellen neu ausgewertet werden können. In den letzten Jahren haben Organisationen wie die NASA jedoch bemerkt, dass viele ihrer alten Bestände an unersetzlichen Daten nicht mehr verständlich und damit nicht mehr brauchbar sind. In der Folge haben sich einige dieser Institutionen intensiv mit den Anforderungen einer Langzeitarchivierung naturwissenschaftlicher Daten befasst.

Die Referate am letztjährigen, vom Centre National d'Etudes Spatiales (CNES) in Toulouse organisierten internationalen Symposium «La pérennisation et la valorisation des données scientifiques et techniques»[11] zeigte, dass die für die langfristige Verstehbarkeit solcher Daten aufgestellten Erschliessungsprinzipien jenen hiesiger Archive formal sehr ähnlich sind. Dies betrifft vor allem eine systematische und stufengerechte Beschreibung der in diesem Bereich üblichen «Verzeichnungseinheiten»: «Missionen», die mehrere «Experimente» umfassen, welche je aus verschiedenen «Instrumenten» bestehen, welche je mehrere «Messdatenreihen» erzeugen. Zu den Stufen wird das zum Zeitpunkt der Datenerzeugung aktuelle und für das langfristige Verständnis nötige Kontextwissen beschrieben: Wissensstand und Technologien, verwendete Theorien und Verfahren, Fachtermini, Masseinheiten, Datenformate etc.

Das erwachte vitale Interesse an der Langzeitarchivierung hat Organisationen aus den Bereichen Raumfahrt, Luftfahrtindustrie, Experimentalphysik, Ozeanographie, Meteorologie etc. veranlasst, ein Referenzmodell für digitale Archive zu erstellen: Das «Reference Model for an Open Archival Information System» (OAIS)[12] erschien 1999 als Empfehlung des Consultative Committee for Space Data Systems, welches von über 30 internationalen Organisationen getragen wird, unter ihnen die NASA, die europäische Weltraumagentur ESA, das französische Centre National d'Etudes Spatiales (CNES), die US-amerikanische National Oceanic & Atmospheric Administration (NOAA) und das World Data Center Panel (WDCP).

Im Jahr 2002 wurde das OAIS als internationaler Standard ISO 14721:2002 verabschiedet und es ist inzwischen auch im Bereich der Bibliotheken und Archive auf grosses Interesse gestossen. Es stellt ein rein funktional-konzeptionelles Modell dar, welches über 50 Funktionen sowie eine grobe Ontologie für digitale Archive festlegt und künftige Systeme vergleichbar machen soll. Mittelfristig wird auch eine ISO-Zertifizierung für OAIS-konforme Systeme angestrebt.

Das OAIS entfaltet seine Wirkung vor allem dadurch, dass es eine einheitliche, fachübergreifende Sprache für die Diskussion über digitale Archive anbietet und so dazu beiträgt, dass sich langsam ein fachübergreifendes Verständnis der Problematik entwickelt. Damit werden komplementäre Kompetenzen in Zukunft manche Synergien ermöglichen: dort die langjährige technologische Erfahrung in der Langzeitspeicherung sehr grosser Datenmengen (Science Data Centers), hier die langjährige Kompetenz in Kontextualisierung, Erschliessung und Metadaten (Archive). ■

contact:

E-Mail: Peter.Keller@bar.admin.ch

[9] Brauchbare, auch für Nicht-Informatiker/innen lesbare Einführung in das informationstechnische Konzept «Datenqualität» bietet: Enterprise Knowledge Management: The Data Quality Approach; David Loshin; Morgan Kaufmann / Academic Press, San Diego, 2001, ISBN 0-12-455840-2

[10] *http://www.21cfrpart11.com/pages/fda_docs/*

[11] *http://www.cnes.fr/pvdst/*

[12] *http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html* und *http://www.rlg.org/longterm/oais.html*