Braucht es in der Schweiz eine Stelle für die Langzeitarchivierung und Sicherung digitaler Daten?

Rudolf Gschwind und Sergio Gregorio, Imaging & Media Lab (IML), Universität Basel

Einführung

Das Imaging & Media Lab digitalisiert seit vielen Jahren analoge Quellen und befasst sich auch mit der Aufbewahrung der dabei entstandenen digitalen Daten. Durch die Zusammenarbeit mit verschiedenen Institutionen und Ämtern (Memoriav, Kulturgüterschutz, Gedächtnisinstitutionen usw.) ist die aktuelle Situation auf der nationalen Ebene bestens bekannt, unter anderem auch durch die seit einigen Jahren angebotenen Workshops zur digitalen Archivierung, eine Art Hilfe zur Selbsthilfe und Informationsnetzwerk zwischen den Institutionen. Aus diesen Kontakten und aus dem immer dringender werdenden Bedürfnis nach nachhaltiger Aufbewahrung digitalen Archivguts ist die Idee entstanden, ein Kompetenzzentrum für die Langzeitarchivierung ins Leben zu rufen.

Braucht es in der Schweiz überhaupt eine Stelle für die Langzeitarchivierung digitaler Daten? Diese Frage kann vorbehaltlos bejaht werden. Ungeachtet, ob die Benennung treffend ist, stellen sich vor allem folgende Fragen: Wofür und mit welcher Aufgabe?

Die digitale Welle ist schon längst über uns hinweggeschwappt. Nicht nur kulturelle Institutionen, sondern die gesamte Gesellschaft und Wirtschaft ist mit digitaler Information konfrontiert und in hohem Masse davon abhängig. Diese existentiell wichtige Abhängigkeit nimmt mit dem stetigen Zuwachs digitaler Daten zu auf Kosten analoger Informationen und Träger (Papier, Fotografien usw.).[1] Digitale Information muss demzufolge auch digital aufbewahrt werden.[2]

In der konkreten Umsetzung treten jedoch umgehend Schwierigkeiten auf. Während das Erstellen digitaler Daten, deren Nutzung und Anwendung mit Hilfe eines Computers sehr einfach ist, verhält es sich bei der Aufbewahrung digitaler Informationen – im Gegensatz zu herkömmlichen, analogen Informationen – nahezu umgekehrt proportional.

Digitale Langzeitarchivierung

Die langfristige Aufbewahrung digitaler Daten ist eine komplexe und nach wie vor anspruchsvolle Aufgabe. Das grundsätzliche Problem liegt in der Informationsspeicherung im Digitalen, die für die Archivierung in einem vereinfachten Modell in zwei distinkte Schritte aufgeteilt wird.

1. Schritt: Erzeugen der für die Archivierung relevanten Daten - das Erstellen des *digitalen Archivpaketes* auf der logischen Objektebene.

   Ein Archivar ist auch im digitalen Kontext im Grunde genommen nur an der Botschaft (Bedeutung, Information) des Archivguts interessiert. Diese wird gedanklich als *konzeptuelles Objekt* begriffen und liegt digital als einheitliche Folge von Bits vor. Diese als Bitstrom bezeichnete Folge wird von der entsprechenden Software als Dateiformat erkannt und definiert das *logische Objekt*.

[1] Zurzeit wird mit einer Verzehnfachung alle fünf Jahre gerechnet.
[2] http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf (06.08.2008).

Information, Metadaten	Dossiers, Fotos, E-Mail
Software, Dateiformate	Anwendungsprogramme, Dateien

Logische Ebene

Die Umwandlung eines konzeptuellen Objektes in ein archivierbares Dateiformat kann, je nach Anforderung und Vorgaben, sehr aufwändig sein.

2. Schritt: Archivierung der binären Daten, des Bitstroms auf der physischen Objektebene.

Information wird mit Hilfe von Computern in Binärcode umgewandelt und auf digitalen Datenträgern gespeichert (Aufbewahrung des Bitstroms). Dieser maschinell effizient verarbeitbare Code kann von einem Menschen ohne die Unterstützung eines Computers, der als «technischer Mediator» fungiert, nicht interpretiert werden. Auch wenn Binärcode auf Papier gedruckt wird, kann ein Mensch nichts damit anfangen. Binärcode kann folglich nur, beziehungsweise *muss* mit Hilfe von Maschinen dekodiert werden.

Hardware, Datenträger	CPU, Speicher, Magnetband, optische und magnetische Disks

Physische Ebene

Diese «Bitstrom-Schicht» (Hardwareschicht) ist aus verschiedenen Gründen kurzlebig (technologischer Wandel, obsolete Datenträger, Datenträgerverschleiss usw.). Datenträger sind für die Nutzung mit Computern optimiert und unglücklicherweise nicht besonders stabil. Magnetbänder und Festplatten halten im besten Fall nur einige Jahrzehnte. Noch viel schlimmer wirkt sich der technische Fortschritt aus. Nach wenigen Jahren kommen neue Generationen von Geräten auf den Markt. Technischer Fortschritt, Marktwirtschaft und Konkurrenz haben zur Folge, dass Hard- und Software im Schnitt alle 5-10 Jahre erneuert werden müssen (Technologiewechsel). Datenträger werden obsolet und können schon nach wenigen Jahren nur schwer oder gar nicht mehr gelesen werden. Dabei macht es keinen Unterschied, ob der Datenträger defekt, altersbedingt nicht mehr gelesen werden kann oder ob es keine Lesegeräte mehr gibt: Die Information ist in jedem Fall unwiderruflich verloren.

Der einzige Weg, der zurzeit aus dieser Sackgasse führt, ist, auf die Kerneigenschaften der digitalen Codes einzugehen. Dieser kann beliebig oft und ohne Informationsverlust kopiert werden, vorausgesetzt es treten beim Kopiervorgang keine Fehler auf. Binärer Code muss in regelmässigen Zeitintervallen migriert werden.[3] Dies muss ausnahmslos und ohne Unterbruch erfolgen. Da man davon ausgehen kann, dass das Datenvolumen konstant zunimmt, wird jede Migration aufwändiger und komplexer, was wiederum mit steigenden Kosten verbunden ist. Eine Migration verursacht hohe Fixkosten und benötigt

[3] Im Durchschnitt alle fünf Jahre: Vollständiger Austausch der Technologie und Übertragen der Daten auf neue Datenträger.

Spezialkenntnisse, damit Daten nicht verloren gehen. Höhere juristische, finanzielle und sicherheitsrelevante Risiken müssen ebenfalls mitberücksichtigt werden.

Diese beiden Schritte, das Erzeugen des Archivpaketes (logische Ebene) und der langfristige Erhalt des Bitstroms (physische Ebene) werden in ihrer Bedeutung und in Bezug auf den erforderlichen Aufwand in der Praxis oft sehr unterschiedlich wahrgenommen.

Für ein Archiv, das zum Beispiel Akten und Buchhaltungsdaten digital archivieren möchte, ist der Aufwand für die Erstellung des Archivpaketes meistens sehr gross (Workflow, Metadaten, Organisation, Konversion in ein Archivformat, z.B. in PDF/A usw.). Die dabei anfallende Datenmenge ist vergleichsweise gering, (Textdaten, PDF, bitonale Scans, komprimierte Dateiformate, ZIP), so dass die eigentliche Archivierung des Bitstroms vom Aufwand und den Kosten her im Vergleich zum ersten Schritt als vernachlässigbar erscheint.

Für ein Archiv, das audiovisuelle Daten speichert, z.B. digitale Bilder einer Sammlung wertvoller Fotografien, stellt sich das Archivierproblem gerade umgekehrt dar. Die Erzeugung der digitalen Daten ist vergleichsweise einfach und mit einmaligen Kosten verbunden. Im Gegenzug entstehen bei audiovisuellen Informationen grosse Datenmengen, da die Daten für die Archivierung nicht komprimiert werden sollen.[4] Aufwand und Kosten für die Archivierung dieses Bitstroms fallen hier schnell und erheblich ins Gewicht. Die Kosten sind im Wesentlichen durch die Datenmenge, durch die gewünschte Sicherheit (Anzahl gespiegelte Datensätze an geografisch verteilten Orten) und durch die gewünschte Zugriffszeit vorgegeben.

Diese Problematik zeigt sich auch auf der Softwareebene, z.B. bei neuen Formaten. Der Wechsel erfolgt zwar langsamer, ist aber gleichsam problematisch.
Zum heutigen Zeitpunkt wird die eigentliche Langzeitarchivierung am wenigsten als Herausforderung angesehen. Dabei gilt:

- Ein **Datenverlust** hat **immer** einen **Informationsverlust** zur Folge!
- Ein Informationsverlust kann aber auch eintreten, obwohl die Daten noch vollumfänglich vorhanden sind!
- Aus technischer Sicht beginnt die Sicherung der langfristigen Verfügbarkeit eines digitalen Objektes mit dem **Erhalt des ursprünglichen Bitstroms**. Ohne diese Daten sind alle weiteren Verfahren zwecklos.

Kompetenzzentrum – Storage Service Stelle

Was sollte demzufolge die Aufgabe einer Stelle für die Langzeitarchivierung digitaler Daten sein?

Die Idee eines Kompetenzzentrums ist aus den Erfahrungen des IML und aus den intrinsischen Eigenschaften digitaler Information entstanden. Dringendster Handlungsbedarf sieht das IML auf der untersten Stufe (Bitstromerhalt, Hardware). Zuallererst geht es darum, den Binärcode langfristig aufzubewahren und zu sichern. Dies ist erwiesenermassen das Fundament der digitalen Langzeitarchivierung.

Das Kompetenzzentrum sollte demzufolge eine Storage Service Stelle sein, die Binärcode langfristig und sicher aufbewahrt (Persistenz).

[4] 5000 hochauflösende Scans können ohne weiteres 1 Terabyte Daten erzeugen.

Die Hauptaufgaben der Stelle wären:

> Zu jedem Zeitpunkt die Daten, d.h. den Binärcode, zurückzuliefern, die zu einem früheren Zeitpunkt abgeliefert wurden.
> Für die korrekte Sicherung und Migration der Daten zu sorgen.
> Die Sicherung und Datenerhaltung soll im Sinne eines «Dark Archive» betrieben werden, d.h. nur ein restriktiver Zugriff auf die Daten und kein öffentlicher Zugang (Open Access).
> Höchstmögliche Datensicherheit auch im Katastrophenfall[5]: «Digitale Rückversicherung».

Weitere Aufgaben:

> Beratung bei Formatfragen und Umgang mit Metadaten, damit der aufbewahrte Binärcode auch in Zukunft einen Sinn ergibt. Letztlich ist der Besitzer (Archivar) der Daten dafür zuständig, dass er mit der Information noch etwas anfangen kann (langfristig die Nutzbarkeit der Daten sicherstellen).
>
> Ausbildung
>
> Unterstützung und aktive Mithilfe bei Software- und Formatmigrationen
>
> Weitere

Eine solche Stelle kann nur mit dem erforderlichen Fachwissen (Kompetenz) betrieben werden. Dabei ist irrelevant, ob es sich um Text-, Bild-, Ton- oder Bewegtbilddaten handelt. Auf der Bitebene gibt es keine Unterschiede. Aber genau auf dieser Ebene erfolgt der Datenverlust am schnellsten.

Die Nutzung und der schnelle Zugriff auf die Daten sollen nicht Aufgabe dieser Stelle sein. Dies ist Aufgabe des Datenbesitzers (Museum, Archiv, Bibliothek etc.). Der Vorteil liegt darin, dass der Besitzer sich auf die Nutzung konzentrieren kann und sich nicht zusätzlich um die aufwändige Langzeitarchivierung kümmern muss (z.B. dreifach gespiegelte Datensätze an verschiedenen Orten).

Im internationalen Umfeld existieren bereits öffentliche Bitstrom-Archivierstellen. Zu erwähnen sind:

Die **GWDG** (Gesellschaft für wissenschaftliche Datenverarbeitung) in Göttingen, eine gemeinsame Einrichtung des Landes Niedersachsen und der Max-Planck-Gesellschaft, betreibt eine solche Service Stelle und schreibt auf ihrer Webseite:[6]

Die GWDG bietet die Langzeitarchivierung wissenschaftlich oder kulturell bedeutender nicht-reproduzierbarer Daten an. Die eingesetzten Speicherverfahren werden regelmässig an den technischen Wandel auf diesem Gebiet angepasst. Die zu sichernden Daten werden hierbei auf aktuelle Speichermedien migriert, um sie vor einem Verlust durch veraltete Technik oder schadhafte Datenträger zu schützen. Die Archivierung der Daten kann auf diese Weise theoretisch unbegrenzt lange erfolgen.

Das Angebot der GWDG für eine solche Datenarchivierung umfasst den physischen Erhalt der archivierten Dateien (Bitstream Preservation). Die Sicherstellung der langfristigen Interpretierbarkeit der Daten verbleibt dabei in der Verantwortung des Eigentümers. Die GWDG gewährleistet derzeit nicht das Vorhandensein einer für die Darstellung erforderlichen Hard- und Softwareumgebung (Langzeitverfügbarkeit). Daher werden auch zum jetzigen Zeitpunkt keine Vorgaben hinsichtlich des Datenformats oder erforderlicher zu-

[5] Dies betrifft nicht nur Naturkatastrophen oder Kriegsereignisse. Fehlende finanzielle Mittel, die eine Migration verhindern, wirken sich in gleichem Masse desaströs aus.
[6] http://www.gwdg.de (06.08.2008).

sätzlicher Angaben zur Interpretationsumgebung gemacht. Die GWDG ist jedoch an verschiedenen Projekten zur Langzeitarchivierung beteiligt und strebt durch diese Aktivitäten eine künftige Ausweitung dieses Serviceangebotes an. Die Daten werden redundant an geographisch getrennten Standorten gehalten.[7]

Das **Leibniz Rechenzentrum** in München führt die Langzeitarchivierung für die Bayerische Staatsbibliothek durch.[8] Ziel des Projektes ist ein «*Exemplarischer Aufbau einer organisatorischen und technischen Infrastruktur für die Langzeitarchivierung von Netzpublikationen einer Universalbibliothek in Kooperation mit einem Rechenzentrum und verschiedenen Produzenten.*»

In den USA ist das **SDSC** (San Diego Supercomputer Center) zu erwähnen, das ebenfalls einen Langzeitarchivier-Service betreibt.[9] Auf der Webseite findet man folgende Kurzbeschreibung: *The centralized, long-term data storage system at SDSC is the High Performance Storage System (HPSS). SDSC manages one of the world's largest productions of HPSS, which currently stores more than 5 PB of data (as of February 2008) with a total system capacity of 25 PB of data. Data has been added at an average rate of 100 TB per month from August 2005 on*.

Schlussfolgerung

In Anbetracht der Bedürfnisse in Gesellschaft und Privatwirtschaft sowie der aktuellen Situation auf der nationalen Ebene ist aus heutiger Sicht die Schaffung eines grossangelegten digitalen Archivs (Kompetenzzentrum), das die langfristige Lesbarkeit und die Sicherheit garantiert, der einzige erfolgreiche Weg, um digitale Informationen langfristig aufzubewahren. Archivieren bedeutet im Kontext der Langzeitarchivierung digitaler Daten: Können die Daten zurückgelesen werden? Bevor über Formate und Metadaten gesprochen werden kann, geht es zunächst darum, den Bitstrom zu erhalten und dafür zu sorgen, dass dieser permanent lesbar bleibt. Ist dies aus irgendwelchen Gründen nicht mehr der Fall, haben alle anderen Aspekte keine Relevanz mehr.

Leider hat die Finanzierung der digitalen Archivierung das unerfreulichste aller Kostenmodelle:

- Start mit hohen Initialkosten
- Jährlich wiederkehrende hohe Fixkosten, die durch die periodische Migration verursacht werden (Lohnkosten, Informatik-Infrastrukturkosten)
- Die Finanzierung muss ohne Unterbruch gewährleistet sein, sonst droht ein Datenverlust!
- Steigende Kosten bei tendenziell konstantem Datenzuwachs (um den Faktor zehn alle fünf Jahre).
- Die Entstehung eines Mehrwerts durch zukünftige Nutzungsmöglichkeiten ist heute nicht voraussehbar.

Unter Berücksichtigung aller bisherigen Ausführungen sollte die Frage deshalb lauten: Wer finanziert, wer baut und welche öffentliche Stelle betreibt den digitalen «Langzeitarchivierungsbunker»? Denn ein solcher wird als erstes benötigt, will man der Herausforde-

[7] http://www.gwdg.de/service/langzeitarchivierung (06.08.2008).
[8] http://www.lrz-muenchen.de/projekte/langzeitarchivierung (06.08.2008).
[9] http://www.sdsc.edu/us/resources/hpss (06.08.2008).

rung ***Langzeit****archivierung digitaler Daten* mit den dafür notwendigen finanziellen Mitteln und adäquaten Massnahmen begegnen.

Anhang

Eine nicht abschliessende Aufzählung der wichtigsten Akteure im Bereich der digitalen Langzeitarchivierung im internationalen Umfeld:

CASPAR Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval http://www.casparpreserves.eu (06.08.2008)
CLIR Council on Library and Information Resources http://www.clir.org (06.08.2008)
Digital Archives http://www.digitalarchives.wa.gov (06.08.2008)
DLP Indiana University Digital Library Program http://www.dlib.indiana.edu (06.08.2008)
DPC Digital Preservation Coalition http://www.dpconline.org (06.08.2008)
DPE DigitalPreservationEurope http://www.digitalpreservationeurope.eu (06.08.2008)
ECPA European Commission on Preservation and Access http://www.knaw.nl/ecpa (06.08.2008)
ERPANET Electronic Ressource Preservation and Access Network http://www.erpanet.org/index.php (06.08.2008)
JHOVE Harvard Object Validation Environment http://hul.harvard.edu/jhove/index.html (06.08.2008)
NARA National Archives and Records Administration http://www.archives.gov (06.08.2008)
Nationaal Archief NL http://www.en.nationaalarchief.nl (06.08.2008)
National Library of the Netherlands e-Depot and digital preservation http://www.kb.nl/dnp/e-depot/e-depot-en.html (06.08.2008)
PADI Preserving Access to Digital Information http://www.nla.gov.au/padi (06.08.2008)
Library of Congres http://www.digitalpreservation.gov (06.08.2008)
PALIMPSEST CoOL Conservation OnLine http://palimpsest.stanford.edu (06.08.2008)
PLANETS Preservation and Long-term Access through NETworked Services http://www.planets-project.eu (06.08.2008)
PORTICO (permanent archive of electronic scholarly journals) http://www.portico.org (06.08.2008)
NDHA National Digital Heritage Archive http://www.natlib.govt.nz/about-us/current-initiatives/ndha (06.08.2008)
NESTOR und KOPAL Kompetenznetzwerk zur digitalen Langzeitarchivierung http://www.langzeitarchivierung.de (06.08.2008) und Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen http://kopal.langzeitarchivierung.de (06.08.2008)
OCLC Online Computer Library Center http://www.oclc.org/default.htm (06.08.2008)
Österreichische Nationalbibliothek Archivierung digitaler Medien http://www.onb.ac.at/about/lza (06.08.2008)
WePreserve http://www.wepreserve.eu (06.08.2008)
SNIA 100 Year Archive Task Forum http://www.snia.org/forums/dmf/programs/ltacsi/100_year (06.08.2008)
TAPE Training for Audiovisual Preservation in Europe http://www.tape-online.net (06.08.2008)