Digitale Archive

Das Phantasma der Unvergänglichkeit

Die Funktion des Archivs ist eine unmögliche. Der Aufgabe, die Archivalien zu bewahren, indem man sie vor dem zerstörerischen Zugriff ihrer Nutzer bewahrt, ist unvereinbar mit der Aufgabe, die eingelagerten Materialien allgemein zugänglich zu machen. Jeder Griff nach dem Manuskript hinterläßt Fingerabdrücke, das Anfertigen von Faksimiles beansprucht die brüchigen Kostbarkeiten, jeder Blick bleicht die Schätze.

So ganz anders und vielversprechend benehmen sich da digitale Speicherinhalte. Sie lassen sich mit Hilfe raffinierter Verfahren so gut gegen Abschreibefehler schützen, daß man sagen kann, sie ließen sich verlustfrei kopieren und vervielfältigen. Die Unterhaltungsindustrie weiß davon ihr klagend' Lied zu singen, denn sie hat seinerzeit in Form der Audio-CD Musik verkauft, die sich endlos und verlustfrei kopieren läßt, und zwar von Jeder und Jedem im Besitze eines CD-Brenners.

Digitale Datenhaltung könnte also die Lösung des archivarischen Dilemmas sein, könnte dem Metier, das so nach Staub und Moder duftet, neuen, modernen, effizienten Auftrieb geben, Speicherungsprobleme lösen, wie auf so vielen anderen Gebieten des spätmodernen Lebens.

Ein Grund für die Anrufung des Digitalen mag die Tatsache sein, daß die digitalen Medien nach und nach die alten, analogen abzulösen scheinen. Es gibt zwar Retro-Moden – die Leidenschaft für die gute alte Schallplatte aus Vinyl bei den jüngeren Musikkonsumenten etwa gehört dazu –, aber insgesamt scheint das Digitale handstreichartig eine analoge Bastion nach der anderen einzunehmen, auch die der Archive. Es wäre dann mithin normal, auch auf diesem Feld digitale Tendenzen zu erwarten.

Doch über solche Normalität eines digitaltechnischen Imperialismus hinaus gibt es allerorten signifikante Ewigkeits-Vorstellungen zum Digitalen, die den Schluß von »digital« auf »Archiv« nahelegen. Es ist allerdings, um die Pointe vorwegzunehmen, ein Kurz-Schluß, ein Phantasma und kein technisches Faktum, um das es hier geht. So ist etwa die Rede von der Immaterialität der computerisierten Daten, es gelte »für

den digitalen Raum, daß in virtuellen Welten … gerade die Möglichkeit des Defekts und Verfalls«[1] fehle, womit zwar der mainstream-Diskurs über das Digitale richtig wiedergegeben wird, was aber dennoch technisch nicht zutrifft.

Das Digitale, wir werden es noch sehen, ist ein archivarischer Albtraum, und zwar aus vielerlei Gründen, die noch zur Sprache kommen sollen.

Ein digitales Archiv von überhaupt allem

Kommen wir aber zunächst zu Aspekten der reinen Quantität, zu technischen Randbedingungen digitaler Speicherung als Teil einer Archiv-Strategie: Was halten Sie davon, einmal grundsätzlich überhaupt alles archivieren zu wollen?

Fangen wir klein an, mit dem Gedächtnis eines Menschen etwa. Ein gewisser kognitiver Psychologe mit dem Namen Thomas Landauer aus Boulder, Colorado, hat geschätzt, daß der Informationsgehalt eines typischen menschlichen Langzeitgedächtnisses zwischen 150 und 225 MB umfasse.[2] Das ist immerhin der Text von etwa ein paar Hundert Paperbacks. Manchmal beschleicht mich das Gefühl, bei mir müsse es sehr viel weniger sein, aber auch dann, wenn das Gedächtnis gut sein sollte, bei Ihnen vielleicht: es paßt ohne Probleme mittlerweile auf briefmarkengroße Speicher-Karten, wie ich sie in meine digitale Kamera schiebe, um Urlaubs-, also Erinnerungsphotos zu machen. Ein gängiger PC müßte nur etwa ein Hundertstel seines Festplatten-Speichers opfern, um das Gedächtnis einer typischen Kleinfamilie abzuspeichern. Sofern dieses Bild überhaupt stimmt, wohlgemerkt, sofern also Erinnerung und Speicher vergleichbar wären. Rechnen wir diesen Wert auf die gesamte Menschheit hoch, so ergeben sich 1.350 PetaByte, also kein Problem, wie wir gleich sehen werden. Auch was ein PetaByte ist, wird gleich geklärt.

Nehmen wir uns als nächstes vor, was – anders als ein menschliches Gedächtnis – ohnehin schon als Datenmasse externaliert ist: die Texte,

1 Wolfgang Ernst: Das Rumoren der Archive, Berlin: Merve 2002, S. 28.
2 Thomas Landauer: »How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-term Memory«, in Cognitive Science 10 (1986), S. 477-493.

Bilder und Töne, die die Menschheit über technische Medien ständig absondert.

Sehen wir einmal nach.[3]

Die Library of Congress hält 20 Millionen Bücher, macht 20 Tera–Byte (Tera ist 10^{12}, eine Million Millionen), 13 Millionen Fotos, macht 13 TB, 4 Millionen Karten und Pläne, macht 200 TB, 500.000 Filme zu etwa 500 TB, 3,5 Millionen Klangdokumente: 2.000 TB. Zusammen ergibt das etwa 3.000 TB oder 3 PetaByte, Peta heißt 10^{15}, eine Milliarde Millionen.

Nehmen wir noch mehr hinzu: alles Schriftliche, alle Fotos, alle neu ausgestrahlten Fernsehsendungen, der Hörfunk, alle publizierte Musik und vor allem: alle geführten Telefonate, die einige spezielle staatliche Stellen sowieso gern aufzeichnen würden, machen auf der ganzen Erde pro Jahr ca. 4.600 PB oder 4,6 ExaByte, Exa: 10^{18}, eine Milliarde Milliarden. Eine ganze Menge. Aber: das ist weniger als der jährlich von der Industrie gefertigte und verkaufte elektronische Speicher.

Mithin: wenn wir wollten, könnten wir – alles speichern.

Wenngleich kritische Stimmen angesichts der enormen menschlichen Mitteilsamkeit, vor allem am Telefon, fragen:

> So who wins the war here – a handful of cybrarian archivists, or the entire chattering human race?[4]

»Schnatternde menschliche Rasse« nennt man uns, das haben wir nun davon.

Das World Wide Web als Archiv

Doch man muß gar nicht zu den Sternen, also zum gesamten Weltgeist, greifen, um auf erhebliche Probleme zu stoßen: Nehmen wir uns vor, was ohnehin digital da und weltweit verfügbar ist: das Internet, speziell das World Wide Web!

Ist es selbst vielleicht schon ein digitales Archiv? Es liegt ja bitweise zugreifbar vor.

3 Vgl. Hartmut Krech: »Der Weltgeist: 1350 Petabyte«, in: Die Zeit 46, 5.11.1998.
4 Dead Media Working Note 26.6.

Technisch umfaßt das Internet die Gesamtheit aller Datenleitungen, aller Computer, die Relais-Funktionen übernehmen, aller Datenpakete, die transportiert werden, und aller Server und Clients, die Informationen anbieten und abfordern.

Wären Internet-Pakete solche wie bei der Paketpost, so würden sie, zusammengenommen, den Bestand des Internet ausmachen. Doch: Internet-Pakete sind vergänglich. Damit alle die Daten, die zu übertragen sind und die gelegentlich verschiedene Wege ausprobieren müssen, um dann schließlich hoffentlich beim Empfänger anzukommen, damit diese Daten nicht sinnlos herumliegen und alles verstopfen, hat man ihnen einen Selbstzerstörungs-Mechanismus eingebaut, das TTL-Feld, das heißt »Time To Live«. Normalerweise trägt diese digitale Lebenserwartung anfangs den Wert 255, und bei jeder Passage über eine Computer-Relais-Station, bei jedem *hop*, wird dieser Wert um Eins dekrementiert. Ist er Null, wird das Paket gelöscht.[5]

Das Internet ist offenbar eine Mischung aus reinem Transport und temporärer Speicherung, die bei den Paketen nie länger als ein paar Handvoll Sekunden dauert. Es kommt für die Frage nach der Lebensdauer von Internet-Dokumenten also auf die Endgeräte an, denn: die Übertragungspakete verschwinden von selbst. Was nicht mehr auf den Servern liegt, kann nicht mehr erreicht werden, es schlägt der berüchtigte »Error 404, document not found« zu.

Das Netz selbst ist also offenbar als Archiv untauglich.

Lassen Sie uns nachsehen, wie es um die Daten auf den Servern selbst bestellt ist.

Ein Blick auf die Statistik eines typischen Servers – hier desjenigen der eigenen Universität – besagt, daß während des überwachten Zeitraums rund 2,5% aller Anforderungen an WWW-Seiten nicht bedient werden konnten, weil es sie nicht mehr gab. Sie sind also verschwunden. Gelöscht. Nicht mehr für wichtig oder erhaltenswert befunden, vielleicht durch Neueres ersetzt.

Das wirft die Frage auf, wie hoch die Lebenserwartung einer WWW-Seite ist, bis sie von den Servern verschwunden ist. Tage, Monate, Jahre?

Brewster Kahle weiß, wieviele WebSites stehen, und er weiß auch, wie lange sie am Himmelszelt des Internet glänzen: noch Ende 1996

5 Andrew S. Tanenbaum: Computernetzwerke, München: Addison-Wesley 2000, S. 438.

sollten es 400.000 WebSites mit insgesamt 1,5 TeraByte Umfang sein, im März 2000 waren es 13,8 TB[6], also wie vorhersehbar etwa 10 Mal so viele. Und wie lange bleibt so ein Dokument typischerweise auf seinem Server erreichbar? Auch hier weiß Brewster Kahle Antwort. Seine Verfahren registrieren, wenn sich etwas auf einem Server verändert, wenn Daten verschwinden. Die Veränderungsrate der Dateien 1996 von 600 GB im Monat läßt den Schluß zu, daß die mittlere Lebenserwartung eines Dokuments im WWW ganze 75 Tage beträgt, denn bei 600 GB pro Monat muß man zweieinhalb Monate oder 75 Tage warten, bis im Mittel 1,5 TeraByte, also alles, weg ist. Nach 75 Tagen sind die meisten Seiten also nicht mehr dort, wo sie einmal vorzufinden waren. Referenzen, Zitate, Bezüge auf sie liefern dann den »Error 404, document not found«. Zweieinhalb Monate sind wahrhaftig nicht viel, vergleicht man das mit der Langlebigkeit analoger Medien, etwa den viele Tausend Jahre alten frühesten Texten des Menschheit, den Proto-Keilschrift-Tafeln aus Mesopotamien. Selbst Bücher auf säurehaltigem Papier kommen gut dagegen weg.

Was heißt das für die Inkunablen des Internet, die »Wiegendrucke«, Erstlinge, Erstauflagen?

Sie sind verschwunden. Die Frühzeit des WWW ist verloren. Es gibt die Dokumente nicht mehr, weil niemand sie bewahrt oder archiviert hat. Ständig verschwinden Dokumente im digitalen Nirwana, und niemand scheint das aufhalten zu können.

Fast niemand. Brewster Kahle versucht es. Seine Organisation heißt »The Internet Archive«, und sie macht Schnappschüsse des gesamten WWW, weshalb er auch die Kenndaten des WWW besitzt. Die digitale Momentaufnahme des Web von 1997, etwa 2 TeraByte Daten. Man kann nur hoffen, daß diese zwei TeraByte nicht die ganze Zeit draußen im Freien herumstehen, sie werden es ohnehin schon schwer genug haben, die nächste Zukunft zu überdauern.

Man kann den Internet-Archiv-Service auch als normaler *user* in Anspruch nehmen: unter dem Namen »Wayback Machine«[7] kann man sich HomePages aus vergangenen Tagen anzeigen lassen. Besonders ergiebig ist dieser Dienst übrigens nicht, oft gibt es nur die Leitseiten, manchmal auch diese nicht mehr.

6 http://www.archive.org/
7 http://www.archive.org/web/web.php

Zu den üblichen Verlusten kommt hinzu, daß vieles im Web erst bei der Abfrage entsteht, etwa die Bahn- oder die Telefonauskunft oder aktuelle Preis- und Produktlisten oder alles was mit Content-Management-Systemen gemacht wird, weil die Seiten dynamisch erst zur Laufzeit aus Datenbanken erzeugt werden. Diese sind den Suchrobotern ohnehin unzugänglich und können von ihnen nicht archiviert werden.

Fazit: das WWW taugt bei einer Dokument-Lebensdauer von zweieinhalb Monaten nicht zum Archiv. Das Web dennoch zu archivieren ist ein heroischer Akt, der seine Frühzeit ohnehin nicht mehr retten kann. Und auch die meisten der zeitgenössischen Websites lassen sich so nicht dem Vergessen entreißen, sie gehen ständig und unwiederbringlich verloren.

Doch: Wo aber Gefahr ist, wächst das Rettende auch. Das wußte nicht nur Hölderlin, das wissen auch die klugen Menschen etwa der Deutschen Bibliothek, die, wie alle guten Archivare oder Bibliothekarinnen immer wissen wollen, wo alle ihr Schätze stehen, selbst dann, wenn jene das Regal gewechselt haben. Übertragen auf das World Wide Web heißt das, es ist Buch zu führen über Objekte und, separat davon, über deren Adresse im Web. Ersteres soll Bestand haben, persistent sein, letzteres darf fließen wie offenbar alles im Web. Statt der üblichen URL, der gängigen Web-Adresse, die im Mittel nach 75 Tagen ungültig wird, muß etwas Dauerhaftes her: ein Katalog von angemeldeten Objekten, die Namen haben, URNs, Uniform Resource Names, die man der Deutschen Bibliothek meldet, die dann noch aufgelöst werden müssen in die URLs, die unsere Web-Browser verarbeiten können.[8]

Das funktioniert, und das ist das Mindeste, was man machen muß, um Archiv-Dokumente im Web verfügbar zu halten.

Lebensdauer digitaler Speichermedien

Nehmen wir einmal an, alles, was archiviert werden sollte, wäre tatsächlich schon von irgendeinem Cyberspace-Librarian, einem »cybrarian«, digital gespeichert. Sind die Kisten voller Bänder, Platten und CDs dann ein sanftes Ruhekissen für den guten Menschen, dem der Erhalt der Kulturgüter so am Herzen liegt?

8 http://www.persistent-identifier.de/

Lassen wir Augenzeugen berichten (ich übersetze aus dem Amerikanischen):

> In Taiwan habe ich Disketten gesehen, die voller Pilze und Schimmel waren (grün und haarig). In Missouri habe ich Disketten-Hüllen gesehen, die von der Hitze im Inneren eines Autos völlig verzogen waren. In Utah habe ich Disketten voller Flugsand gesehen. Der Besitzer sagte mir, es habe ein seltsames Kratzgeräusch gegeben, als sie seinem Laufwerk den Lesekopf zu Schrott geschliffen haben.[9]

Aber selbst vorsichtigeren Zeitgenossen, ja selbst den sagenhaften Raketenwissenschaftlern bei der NASA, mit Technologie doch auf Du und Du, ist schon Schlimmes widerfahren:

> ›Der Inhalt von 1,2 Millionen Magnetbändern, die drei Jahrzehnte amerikanische Raumfahrt dokumentieren, ist hinüber‹, so schreibt Dr. Michael Friedewald vom Fraunhofer-Institut für Systemtechnik und Innovationsforschung in Karlsruhe.

Die Aufbewahrung in Lagerhäusern hat den wertvollen Tapes nicht gut getan. Die Trägerfolie löst sich auf, die Bänder zersetzen sich.[10] Bei Audio-Bändern kennt man das als Quietschen beim Abspielen, als ein Zeichen dafür, daß bald alles vorbei sein wird.

Wie lange Medien halten, hängt davon ab, wie sorgsam sie aufbewahrt werden. Aber auch bei größter Vor- und Umsicht ist ihnen nur eine gewisse Spanne beschieden. Etwa so:

Bänder halten zwischen 2 und 30 Jahren. CD-ROM (die silbrigen, industriell gefertigten) 5 bis 100 Jahre, ebenso wie magneto-optische Platten. Ahnliches kann von anderen Medien gesagt werden, die nur ein Mal beschrieben werden.

Kurz und knapp: »Computerbänder, Videobänder und Tonbänder halten ungefähr so lange wie ein Chevy oder ein Pudel.«[11] Dasselbe kann von digitalen Medien im allgemeinen behauptet werden. Dann schlägt *data rot* zu, die kalte Datenrotte. Für Menschen mit starken Nerven gibt es eine WebSite[12], die den Sound von »dying disks« – »ster-

9 www.phlab.missouri.edu/~ccgreg/tapes.html
10 Digital-Alzheimer, macmagazin 10/2000, S. 134.
11 www.phlab.missouri.edu/~ccgreg/tapes.html

benden« Festplatten – zu Gehör bringt. Wer solches hört, ohne daß ihr oder ihm Blut in den Adern gefriert, hat noch nie einen Plattencrash erlitten.

Und doch ist das noch nicht einmal der Schlimmste, denn nicht nur die Träger altern, auch die Lesegeräte kommen in die Jahre und sterben einfach aus. Haben Sie noch ein 5 1/4-Zoll-Laufwerk? Wo kann ich meine Lochkarten von vor zwanzig Jahren einlesen lassen? Kennt jemand noch das Format eines Schneider-Schreibcomputers?

Schätzt man somit die Lebensdauer von Datenformaten anhand der Lesegeräte, die damit etwas anfangen können, so landet man bei noch sehr viel niedrigeren Werten, die bei fünf bis zehn Jahren liegen. Danach hilft nur noch ein Computermuseum mit geschickten Technikerinnen oder Technikern, die ohne Ersatzteile, die die Industrie natürlich nicht mehr liefern kann, durch Basteln die alten Geräte am Laufen halten.

Glücklicherweise ist *retro computing* zum Sport einiger Unverzagter geworden, die Spaß daran haben, mit Computern zu spielen, in die man noch hineinblicken kann.[13]

Eine Interessengruppe namens »Dead Media« bringt ihre »Dead Media Working Notes«[14] heraus, in denen natürlich nur von toten Medien die Rede ist, etwa der pneumatischen Post in Paris, Hummels Telediagraphen, den »Peek-a-Boo-Index-Cards«, aber auch den obsoleten Computern, die nach Gordon Moores Gesetz alle eineinhalb Jahre durch ihre schnelleren Nachfolger ersetzt werden. Dort, wo tatsächlich alte Formate und die Anmutung der alten Hardware gebraucht werden, hilft nur noch eines: man muß das alte Zeug auf neuem Gerät simulieren, oder, wie Informatiker sagen: emulieren.

Dead Media Activist Bruce Sterling merkt an (übersetzt aus dem Amerikanischen):

> Diese Entwicklung ist für Dead Media Studies interessant, weil die rasche Folge, durch die elektronische Komponenten obsolet werden, immer ein Kainsmal der elektronischen Medien war. Simulation und Emulation toter Hardware wird weiter an Bedeutung zunehmen, solange der Friedhof toter Multimedien nur so wimmelt von Opfern des Mooreschen Gesetzes.[15]

12 http://kiza.kcore.de/technology/harddisks.shtml
13 Detlef Borchers: »Der Glanz von Gestern«, in: Süddeutsche Zeitung 233, 10.10.2000, S. V2/15.
14 www.well.com/user/jonl/deadmedia/NOTES26-28.txt.

So sieht sich die amerikanische Air Force auch gezwungen, spezielle Vorsorge zu treffen, um beim Generationswechsel elektronischer Schaltungen, mit denen die modernen Kampfjets ja vollgestopft sind, nicht auch gleich neue Flugzeuge bauen zu müssen. Denn die Hardware und die darauf implementierte Software muß einwandfrei laufen, damit der Vogel am Himmel bleibt, auch wenn die Chips schon längst nicht mehr hergestellt werden. Also gibt es Ersatzteil-Probleme.

> Die Teileknappheit rührt größtenteils von der kurzen kommerziellen Lebensspanne digitaler elektronischer Komponenten, verglichen mit dem langen Wartungsleben von Waffensystemen. Eine digitale Komponente z. B. mag eine Lebenszeit von 18 Monaten haben, während ein Waffensystem, das diese Komponente verwendet, oft Jahrzehnte im Einsatz ist.[16]

Das kommt sehr teuer. Und es führt uns wieder zurück zu unserem eigentlichen Thema, den digitalen Archiven, die ohne heftigste Anstrengungen auf dem Feld einer aufwendigen Daten-Archäologie sehr schnell digitalem Vergessen anheimfallen.

Daten-Archäologie

Jeff Rothenburg ist durch ein Diktum bekannt geworden, das da lautet:

> Digital documents last forever – or five years, whichever comes first.[17]

Und er weiß, wovon er spricht, denn sein in die Ewigkeit, das heißt in die nächsten fünf Jahre, greifendes Urteil ist Frucht einer ausführlichen und sehr überzeugenden Studie zum Thema digitaler Dokumentarchivierung. Das Resultat lautet:

> … there is – at present, no way to guarantee the preservation of digital information.

15 Ebd.
16 Ebd.
17 Jeff Rothenberg: Avoiding Technological Quicksand, 1998, nach www.clir.org/pubs/reports/rothenberg/

Wenngleich Garantien nicht abzugeben sind, so gibt es doch eine Strategie, die, wenn verfolgt, Hilfe verspricht, und von der auch schon die Rede war:

> The best way to satisfy the criteria for a solution is to run the original software under emulation on future computers.

Die Originalsoftware unter einer Emulation hoffnungslos veralterter Betriebssysteme längst verrotteter Hardware muß immer wieder zum Laufen gebracht werden, um Funktion und Anmutung obsoleter digitaler Dokumente wiederherzustellen.

Zunächst ist jedoch über die Jahrzehnte der Bitstrom der digitalen Daten zu erhalten, umzukopieren auf je neue Speichermedien, als Maßnahme gegen den den Verschleiß von Trägermaterial und Gerätschaft, zu ergreifen etwa alle ein bis zwei Jahre. Anschließend hat man dafür zu sorgen, daß die Daten auch korrekt interpretiert werden. Wenn man nicht weiß, wie der Inhalt eines Mediums zu interpretieren ist, ist man noch nicht viel weiter. Neben dem *data rot* war auch fehlende Beschriftung eine der Ursachen für die massiven Datenverluste der NASA. Metadaten sind anzubringen. Sie beschreiben, was wie zu interpretieren ist. Schlägt man die Emulations-Strategie ein, müssen die Metadaten beschreiben, unter welchem Betriebssystem und auf welcher Hardware die Software lief, die die Daten einstmals interpretierte. Rothenburg schreibt dazu:

> This point cannot be overstated: in a very real sense, digital documents exist only by virtue of software that understands how to access and display them; they come into existence only by virtue of running this software.

Anzulegen ist also auch ein Archiv von Betriebssystem-Emulationen in allen relevanten Versionen und eine Sammlung von Software, die die Dokumente interpretieren kann. Ständig frisch umkopierte Dokumente könnten so auf neuesten Computern unter Betriebssystem-Emulationen von Originalsoftware angezeigt, mithin archiviert und verwendet werden.

Nur dann, wenn wir diesen Aufwand treiben, werden digitale Dokumente archiv-fähig. Sie sehen, das ist nichts mehr für Privatleute, hier sind staatliche Institutionen gefragt, die eine solche außerordent-

lich aufwendige Arbeit kontinuierlich leisten. Der Archiv-Begriff, der ja auf das Amtshaus des Archonten zurückgeht, der die Macht über die Regierungsdokumente ausübt, zeigt seine ursprüngliche Bedeutung.

Derrida schriebt in *Mal d'Archive*:

> ... ›archive‹, sein einziger Sinn, vom griechischen archeîon: zuerst ein Haus, ein Wohnsitz, eine Adresse, die Wohnung der höheren Magistratsangehörigen, die árchontes, diejenigen, die geboten. Jenen Bürgern, die auf diese Weise politische Macht innehatten und bedeuteten, erkannte man das Recht zu, das Gesetz geltend zu machen oder darzustellen. Ihrer so öffentlich anerkannten Autorität wegen deponierte man zu jener Zeit bei ihnen zuhause, an eben jenem Ort, der ihr Haus ist (ein privates Haus, Haus der Familie oder Diensthaus), die offiziellen Dokumente. Die Archonten sind zunächst Bewahrer. Sie stellen nicht nur die physische Sicherheit des Depots und des Trägers sicher. Man erkennt ihnen auch das Recht und die Kompetenz der Auslegung zu. Sie haben die Macht, die Archive zu interpretieren.[18]

Und weiter:

> ... die technische Struktur des archivierenden Archivs bestimmt auch die Struktur des archivierbaren Inhalts schon in seiner Entstehung und in seiner Beziehung zur Zukunft. Die Archivierung bringt das Ereignis im gleichen Maße hervor, wie sie es aufzeichnet. Das ist auch unsere politische Erfahrung mit den sogenannten Informationsmedien.[19]

Insbesondere wird das die Erfahrung mit digitalen Archiven sein. Nur Macht und Geld können sie vor dem Verfall retten, die so viel anfälliger sind als ihre analogen Vorläufer. Und wer die überkommenen Dokumente so unter seiner Ägide hat, kann sie nach Belieben einsetzen, interpretieren, vorenthalten.

Sehr real vernichtet der technische Fortschritt, der unabdingbar ist, um immer mehr Dokumente in digitale Archive einstellen zu können, ganz real also vernichtet genau dieser Fortschritt das Archiv selbst: digitale Archive als Schauplätze eines *mal d'archive*, eines digitalen Archiv-Übels.

18 Jacques Derrida: Dem Archiv verschrieben – Eine Freudsche Impression, Berlin: Brinkmann + Bose 1997. S. 11.
19 J. Derrida: Dem Archiv verschrieben, S. 35.

Ein wenig läßt sich das technische Problem entschärfen, indem man digitale Dokumente möglichst in solchen Formaten abspeichert, daß der Inhalt noch lange interpretierbar bleibt, daß also schon durch den Erhalt des Bitstroms der Daten wesentliches gerettet wird. Dies versäumt zu haben, war der Fehler der NASA-Leute, die ihrer Nachwelt nur unverständliche Bitfolgen auf ihren Bändern hinterließen.

Das wichtigste dieser Formate heißt XML, eXtensible Markup Language, und es hat den Vorteil, völlig vom Erscheinungsbild der Dokumente und der Funktionalität der anzeigenden Programme abzusehen, damit unabhängig zu machen von Software und Hardware, weshalb man diese auch nicht über die Jahrzehnte und Jahrhunderte zu retten braucht. XML kodiert sehr explizit Inhalt, Struktur und Semantik der Daten, es ist lesbar von Menschen und von Programmen. Schafft man es, den Bitstrom zu erhalten, indem man immer wieder auf neue Speichermedien umkopiert, hat man die Chance, durch direkten Augenschein und durch Programme die Daten immer wieder interpretieren zu können, aber auch, wenn die Darstellungssoftware nicht mehr läuft, neu algorithmisch interpretieren zu müssen.

Man fügt den Daten Metadaten zu, kleine Schildchen gleichsam, die alle Datenatome etikettieren. Das World Wide Web-Consortium, das XML betreut, gibt in seinem Einführungskursus[20] folgendes einführende Beispiel:

```
<note>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Diese eigentlich wahrlich zu Herzen gehende Bitte Janis an Tove verliert durch ihre Explizitheit massiv an der ansonsten in Liebesdingen erforderlichen Zweideutigkeit, und das ist für unsere Archiv-Zwecke auch gut so. Absender und Empfänger bleiben genau so wenig im Dunkeln wie der Charakter des Schreibens und seine Unterteilung in Kopf und Körper. Jedem Fitzelchen sein Schildchen, sein *tag*, ersichtlich an den spitzen Klammern.

20 http://www.w3schools.com/xml/xml_whatis.asp

Welchen Fortschritt in Hinblick auf spätere Verstehbarkeit diese extensive und explizite Etikettierung hat, wird vielleicht noch deutlicher an Daten, deren Semantik sich nicht unmittelbar erschließt.

So führt eine niederländische Forscherinnen-Gruppe[21] folgendes Beispiel an, zunächst im dürren ASCII-Code,

```
26502 Martensz Matheeus Kruidenier Antwerpen 19-05-1586 B 35
```

danach in XML,

```
<record>
<row>
<persID>  26502  </persID>
<family_name>  Martensz  </family_name>
<first name> Matheeus </first_name>
<profession>  Kruidenier  </prefession>[22]
<origin>  Antwerpen  </origin>
<date_of_entry> 19-05-1586  </date_of_entry>
<entry_number>  B  </entry_number>
<entry_page>  35  </entry_page>
</row>
</record>
```

was sich doch, das muß man zugeben, entschieden klarer liest.

Diese *tags* darf man selbst erfinden, und so kann jeder Archivarin, jeder Archivar ihren und seinen eigenen Satz von Metadaten erfinden. Dieser Umstand schafft Freiheit und somit Probleme, und dieser versuchen die *cybrarians* durch die Standardisierung von Metadaten Herrin und Herr zu werden.

Bei den Metadaten-Standards, die Identifizierung und Suche erleichtern, gibt es gute Vorschläge, die prominentesten lauten: Dublin Core und die *Open Archive Initiative*[23] . Diese Konventionen könnten, wenn sie weite Verbreitung fänden, die Suche und das Auffinden digitaler Dokumente erheblich vereinfachen. Ihre Uni-Bibliothek und hoffentlich auch Ihr Rechenzentrum wissen, worum es geht, wenn Sie sie nach Näherem fragen. Die Verwendung solcher Standards macht Doku-

21 Annelies van Nispen/Rutger Kramer/René van Horik: »The eXtensible Past – The Relevance of the XML Data Format for Access to Historical Datasets and a Strategy for Digital Preservation«, in: D-Lib Magazine, 11.2 (2005). http://www.dlib.org/dlib/february05/vannispen/02vannispen.html

22 Der Tippfehler »</prefession>« statt »</profession>« befindet sich schon in der Originalveröffentlichung. Offenbar gibt es noch Leute, die nicht ausschließlich mit Copy & Paste arbeiten. Dank an Hubert Woltering, M.A., für den Hinweis!

23 http://www.openarchives.org/

mentarchivierung damit noch nicht zu einem Kinderspiel, aber doch wenigstens zu einem, bei dem auch akademischer Institutionen mittun können. Doch machen wir uns nichts vor: die Archivierung digitaler Daten erfordert – genau wie bei ihren analogen Vorläufern – ständige Pflege, einen großen Aufwand und: viel Geld.

Für Textdokumente kann XML eine Lösung sein, aber für Bilder und für Klänge, für Filme und alles, was gerade kein Text ist, bleibt die Kluft, daß die maschinenlesbaren Kodierungen gerade für Menschen unverständlich sind, daß das *mal d'archive* auch durch Zaubersprüche in XML nicht gebannt werden kann.

In einem fünfjährigen Projekt haben meine Kolleginnen und Kollegen und ich eine Erschließungs- und Archivierungsarbeit an der Kunst Anna Oppermanns gemacht, deren Datenformat natürlich XML lautet, deren Bildbestand aus Dateien im JPEG-Format besteht, das zwar wenigstens ein offener Standard ist, von dem wir aber noch nicht wissen, wie lange er hält.[24]

Was auf den allerersten Blick wie ein Versprechen der modernen Digitaltechnik aussieht – sehr viel speichern zu können – entpuppt sich am Ende als ein Anlaß zu umfänglichster Regelung und Verwaltung. Einen knappen Einblick in den Verhau einer Zertifizierung digitaler Repositorien gewährt etwa die Checkliste[25] der Research Libraries Group[26], die unter ihren vier Kategorien zwar auch eine technische hat, aber sehr viel mehr Augenmerk der Organisation und der Finanzierung, den Geschäftsprozessen und der Benutzbarkeit der Dokumentbewahrung widmet.

24 Christian Terstegge/Martin Warnke/Carmen Wedemeyer: »PeTAL: a Proposal of an XML Standard for the Visual Arts«, in: Vito Cappellini/James Hemsley /Gerd Stanke, Tagung EVA 2002 Florence, Florenz: Pitagora Editrice Bologna 2002, S. 94-99. http://kulturinformatik.uni-lueneburg.de/warnke/Petal_EVA_2002_Florence.doc.pdf. Martin Warnke: »Daten und Metadaten«, in: zeitenblicke 2.1 2003. http://www.zeitenblicke.historicum.net/2003/01/warnke/index.html. Uwe M. Schnede/Martin Warnke (Hrsg.): Anna Oppermann in der Hamburger Kunsthalle, Hamburg: Hamburger Kunsthalle 2004. Mit einer DVD von Martin Warnke, Carmen Wedemeyer und Christian Terstegge.

25 http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf

26 http://www.rlg.org/

Klang-Archive

Die Bewahrung von Klang-Beständen wartet mit besonderen Schwierigkeiten auf, denn im Gegensatz zu Text gibt es genuin analoge Aspekte von Klang, die Wellenform, deren Digitalisierung immer mit Verlust behaftet sein wird und über die es keinen letzthinnigen Konsens gibt. Alles, was nicht im Notenbild oder in einer MIDI-Datei aufgeht, sträubt von Berufs wegen sich gegen Digitalisierung.

Doch natürlich müssen Klangarchive auf mittlere Sicht trotzdem digitalisiert werden, denn auch Klangmedien entkommen der Hegemonie des Digitalen nicht. Einige archivarische Überlebens-Strategien lassen sich aus dem ableiten, wovon bisher die Rede war, einiges ergibt sich aus den Erfahrungen von Großprojekten zur digitalen Klangarchivierung, etwa dem der Library of Congress[27]. Dort verfolgt man eine Doppelstrategie, nämlich die der Erzeugung und des langfristigen Erhalts von digitalen Master-Digitalisaten und der Dissemination in gängigen Formaten über das Web.

Die Master-Kopien sollten in hoher Auflösung von mindestens 96 kHz gesampelt werden und eine Wortlänge von 24 Bit umfassen, um die relevanten Frequenzen einzufangen und die nötige Feinheit der Quantisierungsstufen bei der Analog-Digitalwandlung sicherzustellen. Das Dateiformat sollte offengelegt sein, damit es auch später noch Nacharbeit erlaubt, es sollte weite Verbreitung und Akzeptanz gefunden haben, damit wir noch auf Geräte hoffen können, auf denen man die Daten wieder in Klang verwandelt kann. Die Darstellungsmethode sollte transparent sein, was gegen Digital Rights Management und gegen Kompression spricht, sie sollte sich selbst dokumentieren, also wenigstens eine kurze Selbstbeschreibung enthalten, muß natürlich gut klingen und zumindest Stereo erlauben.

Die Wahl fällt dann sehr schnell auf das Format von Microsoft und IBM, das WAVE heißt, und in dem auch Audio-CDs kodiert sind. Es sind genau die Originaldaten, die die Musikindustrie auf der beliebten CD eingeführt hat, und die sich so wunderbar leicht kopieren lassen, zum Verdruß genau derselben Musikindustrie. Es ist aus dem Vorherigen klar, daß der Bitstrom dieser Daten ständig umkopiert werden muß. Welches Medium man dafür nimmt, ist noch nicht völlig klar. CDs und DVDs sind billig, aber anfällig, Bänder und Festplatten eignen

27 http://www.arl.org/preserv/sound_savings_proceedings/fleischhauer.html

sich mehr oder weniger gut, man muß also wie immer in diesem Metier auf der Hut sein und sich um die Daten geradezu liebevoll kümmern.

Zur Verbreitung der Klänge, die neuerdings, im digitalen Zeitalter auch eine Form der Erhaltung ist (»Dissemination is a method of preservation«[28]), eignen sich dann MP3 wegen der weit verbreiteten Player und Real Audio als Streaming-Format für das Web.

Nicht zu vergessen sind dann noch die Metadaten, über deren Umfang Fachleute des Feldes der Musikarchive befinden sollten, damit die digitalen Repositorien keine Datenfriedhöfe werden.

Und: bevor es zu spät ist, sollte wohl genau beschrieben werden, welche Unterschiede zu hören sind zwischen den Originalen auf der Walze, dem Draht oder der Platte und den Digitalisaten worauf auch immer. Unsere Enkel werden das wissen wollen.

Das Genom als Archiv

Gestatten Sie mir einen kurzen Exkurs in die Humangenetik, denn wenn es um große digitale Dokumente geht, muß einem auch das menschliche Genom in den Sinn kommen.

Das Genom ist zweifellos der Träger digital kodierter Information. Das Alphabet des Kodes besteht aus diskreten Zeichen, die mit den Basen Adenin, Cytosin, Guanin und Thymin identifizierbar sind. Die Länge der Zeichenkette entspricht etwa 5.000 Buch-Bänden, einer recht stattlichen Privatbibliothek, die wir in jeder unserer Zellen mit uns herumtragen.[29]

Doch natürlich stimmt die Text- und damit auch die Buch-Metapher nicht. Denn das meiste, was in der Zeichenkette steht, ist Schrott, Datenmüll. Nur ca. 3% aller Zeichen kodieren die Gene, unsere Erbanlagen, auf die es eigentlich ankommt. Beim Umkopieren der Daten auf jeweils frische Datenträger, unseren Kindern, – einem Vorgang, der, wie wir wissen, mindestens zum Erhalt einer digitalen Informationssammlung erforderlich ist – bei diesem Umkopieren verändert sich das Genom, und es schleichen sich Fehler ein, die evolutionären Fortschritt, aber auch Krankheit bedeuten können. Das Umkopieren geschieht

28 Elizabeth Cohen in »Preservation of Audio«, http://www.clir.org/pubs/reports/pub96/preservation.html

29 Matt Ridley: Alphabet des Lebens, München: Claasen 2000.

auch bei jeder Zell-Neubildung in einem Organismus, wobei es einem dem *data rot* entsprechenden Prozeß gibt: mit zunehmendem Alter und unter ungünstigen äußeren Bedingungen erhöht sich die Zahl der Kopierfehler, und was dabei entstehen kann, heißt bei Mensch und Tier: Krebs. Will man den Vergleich mit dem Internet wagen, dann wären das Zellen mit einer TTL, einer TimeToLive, von unendlich: eben eine bösartige Wucherung.

Ein weiterer Grund, warum die Schrift-Metapher in die Irre führt, liegt darin, daß die Zeichen und ihre Anordnung zwar durchaus für den Menschen lesbar sind, nämlich für die Ribosomen, die Erzeuger der Proteine, für die das Genom die Bauanleitung ist, aber durchaus nicht für das menschliche Bewußtsein, das doch ansonsten für Textinterpretation zuständig ist. Die vier Basen und ihre Kombinationen bilden keine Symbolschrift, wie wir sie aus Texten gewohnt sind. Sie wirken nur durch das Leben und Sterben selbst, nicht über Symbol-Interpretation. Denn, was vor allem dazu fehlt, das sind die Metadaten: nirgends steht geschrieben, welche Bedeutung einzelne Abschnitte der DNA haben und wie sie zu interpretieren sind, kein XML-*tag* markiert, wo genau die Augenfarbe, der Intelligenzquotient, die Länge der Wimpern beschrieben stehen. Und wahrscheinlich ginge schon die Frage nach dem Ort solcher Inschrift in die Irre. Wir stehen im Moment, der ja als der gefeiert wird, zu dem wir 99% der DNA nachbuchstabieren können, wir stehen jetzt vor der Situation, daß wir einen Dokumenttext ohne Dokumentation haben, der unter unbekannter Software auf einer nur schlecht bekannten Hardware läuft, um einmal eine andere, sicher auch sehr schlechte Metapher zu wählen.

Jedenfalls sind wir weit davon entfernt, den Bitstrom der Daten etwa verstehen zu können, und bei der Größe und Komplexität dieses Problems bin ich auch eher verzagt, daß jemals erwarten zu können. Mich erinnert diese Situation eher an den Daten-GAU bei der NASA: ein Haufen von Daten, das meiste davon Schrott, alles unbeschriftet, entzifferbar nur per *trial and error*.

Die DNA ist ein digitales Archiv des Lebens, eines, das in je neuen Versionen die Evolution dokumentiert, das aber nicht von Menschen lesbar zu sein scheint, ein Geheimarchiv größter Bedeutung, aber ohne Zutritt für uns Sterbliche, was dafür aber seine Integrität noch ein kleines Weilchen sichern kann.

Ich möchte mit einigen Überlegungen zum Verhältnis von Speicher und Gedächtnis so langsam zum Ende kommen.

Die Geschichte der Computertechnik ist die Geschichte von schrägen Metaphern und Anthropomorphismen: Charles Babbage nannte Bestandteile seiner gebauten und geplanten Maschinen in Anlehnung an die Landwirtschaft noch *mill* und *store*, aber aus dem an ein Getreidesilo erinnernden *store* wurde im öffentlichen Sprachgebrauch, wie wir wissen, *memory*, das Gedächtnis eines Künstlichen Gehirns.

Dabei ist der Neurophysiologie noch gar nicht klar, wie das Gedächtnis eines Lebewesens mit Zentralnervensystem funktioniert. Es hat etwas mit Hirnmaterie zu tun, wie die Folgen von Gehirnverletzungen zeigen, aber niemand kann unter dem Mikroskop irgendwelche Speicherplätze zeigen, an denen Gedächtnisinhalte zu lokalisieren wären. Gerhard Roth schreibt: »Das Gedächtnis ist … unser wichtigstes ›Sinnesorgan‹. Es ist zugleich aber … nur ein Glied im Kreisprozeß von Wahrnehmung, Gedächtnis, Aufmerksamkeit, Erkennen, Handeln und Bewerten.«[30] Hier ist er einig mit dem erkenntnistheoretischen Konstruktivismus eines Heinz von Foerster, aber auch mit dem späten Wittgenstein, der polemisiert: »Ein Ereignis läßt eine Spur im Gedächtnis, das denkt man sich manchmal. … Der Organismus mit einer Diktaphonrolle verglichen; der Eindruck, die Spur, ist die Veränderung, die die Stimme auf der Rolle zurückläßt. Kann man sagen, das Diktaphon (oder die Rolle) erinnere sich wieder des Gesprochenen, wenn es das Aufgenommene wiedergibt?«[31]

Gedächtnis ist mithin nicht zu isolieren, schon gar nicht technisch-konstruktiv. Gedächtnis ist weniger ein Ding als vielmehr ein Prozeß, der sich als lebendiger vollzieht und sich dabei möglichweise auf irgendwelche systemischen Zustandswechsel stützt, die im Vollzug des Erinnerns eine kodifizierende Rolle spielen könnten.

Alan Turing hat ja seine Maschine als abstraktes Modell von Computern auch mit Zuständen und Symbolen bestückt, um die Arbeit eines rechnenden Menschen zu simulieren. Dabei spielen die Symbole

30 Gerhard Roth: Das Gehirn und seine Wirklichkeit, Frankfurt/Main: Suhrkamp 1996, S. 241.
31 Ludwig Wittgenstein: Bemerkungen über die Philosophie der Psychologie, Bd. 7, I, S. 220.

noch die Rolle von »Gedächtnisstütze[n]«[32]. Die eigentliche Anthorpomorphisierung erfolgte erst später in der Künstliche-Intelligenz-Forschung, deren Credo die *symbol systems hypothesis* ist: alles, was die Welt ausmacht, sei durch Symbole kodifizierbar, die nach festen Regeln manipulierbar seien, woduch das menschliche Denken inklusive Gedächtnis nachzubilden wäre. Hier gibt es nun keinen Unterschied mehr zwischen Speicher und Gedächtnis.

Mit der tatsächlichen Situation, in der auch und gerade digital kodierte Daten dem Verfall anheim gegeben sind, kann diese KI-Metapher nicht umgehen. Weltwissen und Fakten, ihre Pendants zu Erfahrung und Gedächtnis, können zwar eventuell durch spätere Ableitungs-Prozesse obsolet und widerlegt werden, aber ein einfaches Verschwinden aufgrund von *data rot* ist für dieses Wissen und solche Fakten ebenso wenig vorgesehen wie die Einbettung des Gewußten in Handlung, Wahrnehmung und Bewertung.

Das Paradox der digitalen Archive

Die Einsicht in die Zeitlichkeit digitaler Daten, die Notwendigkeit, digitale Archive mit hohem Aufwand über Jahrzehnte hinweg zu retten, schlägt mit Gerhard Roth, den Konstruktivisten oder dem späten Wittgenstein in dieselbe Kerbe: ein Speicher mag über lange Zeit hinweg intakt bleiben, ohne zu vergehen; Gedächtnis und Erinnerung bedürfen aber genau so wie interpretierbare Daten ständiger tätiger Erneuerung.

Archive, digitale zumal, überdauern nur, wenn sie ständig benutzt werden, wenn eine erhaltende Instanz sie stets neu kodifiziert, interpretiert und bewertet, sich ihre Dokumente handelnd aneignet, sie herausgibt oder verheimlicht, damit Wissen ermöglicht und strukturiert, Handlungen provoziert oder zu unterdrücken trachtet. Nur so überstehen digitale Archive die Jahrzehnte.

Sie leben so lange, wie eine Macht sie trägt und ihren informationellen Stoffwechsel aufrecht erhält. Danach werden sie bestenfalls Mausoleen, in deren Innerem man nichts Brauchbares mehr finden wird, deren Deckel besser geschlossen bleiben, weil ihr Inhalt ohnehin Moores

32 Alan M. Turing: On Computable Numbers, with an Application to the Entscheidungsproblem. Proc. of the London Math. Society, 2.42 (1937).

Gesetz oder dem zweiten Hauptsatz der Thermodynamik zum Opfer gefallen sein wird: der Entropie, der negativen Information, zu Datenstaub werdend, an das Vergessen vergessen.

Entgegen einer geläufigen Auffassung von Immaterialität des Digitalen taugen also informatische Archivierungsverfahren nicht so ohne Weiteres für die Ewigkeit: Medien verrotten, Festplatten scheinen zu »sterben«, Chip- und ganze Rechnergenerationen lösen sich ab, Formate verschwinden so schnell, wie sie aufgetaucht sind.

Es gibt zwar Strategien, trotz *data rot* und technischer Veralterung ein Mindestmaß an Dauerhaftigkeit zu gewährleisten, aber eines funktioniert nicht mehr: das Liegenlassen und Wegschließen von Archivalien ist unter Bedingungen der Digitalität kein Schutz vor Abnutzung mehr, sondern ihr schlichtes Todesurteil.

Das Paradox der digitalen Archive, das da lautet: »Bei analogen Archivalien bleicht jeder Blick die Schätze, digitale Archivalien wollen Aufmerksamkeit um jeden Preis« oder, etwas salopper: »Rührmichnichtan trifft Betriebsnudel«, dieses Paradox ist nicht mehr aus der Welt zu schaffen.

leicht verändert erschienen in: Hedwig Pompe und Leander Scholz (Hrsg.): Archivprozesse. S. 269-281. Köln: DuMont 2002. ISBN 3-8321-6005-1.