Daten für die Ewig

Was haben Ihre Urlaubsfotos mit dem kulturellen Erbe der Menschheit zu tun? An beiden nagt unerbittlich der Zahn der Zeit. Auf den Bildern aus Irland fehlt nach 20 Jahren ausgerechnet das so charakteristische Grün, die winterliche Gebirgslandschaft ist in penetrantes Rosarot gehüllt und das letzte Weihnachtsfest verblasst schneller als jede Erinnerung. In unseren Bibliotheken und Archiven bedrohen derweil Schimmel, Tintenfraß und Säure unersetzliches Schriftgut. Elektronische Medien können zwar Backups speichern, schaffen aber auch neue Probleme: Keiner weiß, wie lange die Daten beispielsweise auf einer CD wirklich halten.

Der private Anwender kann hier von Profis lernen, für die Langzeit-Archivierung ein Dauerbrenner ist. Jedes Jahr entstehen weltweit inzwischen mehr als eine Milliarde Gigabyte an Informationen, von denen gerade einmal 0,003 Prozent gedruckt werden. Der Löwenanteil der neu generierten Informationen besteht aus Bildern, Animationen, Filmen, Ton, Grafiken und anderen Multimedia-Daten. Ein Teil davon ist wichtig – auch über den Tag hinaus.

„Elektronische Medien sind nicht archivierbar“, meinte Clifford Stoll, amerikanischer Astronom und Spezialist für Datenschutz und Computersicherheit, 1996 in seinem Buch „Die Wüste Internet. Geisterfahrten auf der Datenautobahn“. Historiker sind nicht die einzigen, denen diese Zukunftsvision nicht gefällt. „Wir dürfen nicht zulassen, dass wir kollektiv vergessen“, mahnt Dr. Elisabeth Niggemann, Generaldirektorin der Deutschen Bibliothek. Unserer Gesellschaft drohe der digitale Alzheimer, wenn nicht schleunigst etwas dagegen unternommen werde.

Neben der Aufnahme elektronischer Publikationen beginnen Bibliotheken auch damit, Bücher einzuscannen und so einen elektronischen Fundus an älteren Werken aufzubauen. „Retro-Digitalisierung“ heißt das Zauberwort, mit dem kostbare Exemplare geschont und zugleich öffentlich zugänglich gemacht werden sollen. Deutschlands erste Digitalisierungscenter in Göttingen und München haben unter anderem eine Gutenberg-Bibel gescannt und die älteste bekannte Handschrift des Talmud digitalisiert.

Andere Interessen verfolgen US-Firmen wie Google und Amazon. Mit groß angelegten Digitalisierungsprojekten haben sie den Massenmarkt im Visier, um ihre Angebote um digitalisierte Werke zu ergänzen. Am ambitioniertesten ist „Google print“: Die Suchmaschine wird künftig auch immer mehr

Foto: picture-alliance; photothek.net; Picture-Press, NASA VRS; Gutenberg-digital; K. Flemming, Corbis

keit

Google scannt die altehrwürdige Bibliothek in Oxford. Doch wie lange halten die digitalisierten Werke? Sind unsere Daten auf CDs, Platten und Bändern in einigen Jahren überhaupt noch lesbar? *Von Manfred Flohr*

DAS NASA-SYNDROM

Wenn die Amerikaner 2018 wieder auf den Mond wollen, fangen sie bei Null an: Die alten Daten sind futsch. Rechts: Die sorgsam gescannte Gutenberg-Bibel sollen kommende Generationen in digitaler Form lesen können.

Treffer in Büchern landen und einzelne Seiten oder ganze Bücher online zur Verfügung stellen.

Google hat bereits mit dem Scannen der Bestände renommierter Universitätsbibliotheken begonnen. In Stanford und Michigan werden alle 15 Millionen Bände digitalisiert, Harvard und die New York Public Library steuern einen Teil ihres Fundus bei. Von diesem Herbst an ist die Bodleian Library in Oxford als erstes Google-Projekt in Europa dran. Verzögert wird das Projekt durch bislang noch nicht ausgeräumte Copyright-Einwände von Verlagen und Autoren.

Roboter scannen ganze Bibliotheken

Zum Teil lässt Google noch von Hand scannen, der Großteil der Arbeit wird aber Kopier-Robotern überlassen, die selbstständig ganze Bücher durchblättern und kopieren. Bis 2015 dürften die fünf Bibliotheken erfasst sein. Die Daten müssen aber nicht nur sicher gespeichert, sondern auch so aufbereitet werden, dass sie bei Bedarf tatsächlich gefunden werden. Ohne die so genannten Metadaten, die Autor, Verlag, Erscheinungsdatum, Stichworte und andere Informationen enthalten, wäre dies unmöglich.

Ein spektakuläres Beispiel dafür, wie schnell der digitale Alzheimer eintreten kann, liefert die NASA: Wenn die Amerikaner 2018 wieder vier Astronauten zum Mond schicken, ist die erste Mondlandung ein halbes Jahrhundert zuvor einer von vielen weißen Flecken in der Geschichte der Raumfahrt. Unfreiwillig hat die NASA gezeigt, wie viele Möglichkeiten es gibt, digitale Daten ins Nirwana zu schießen. Mitte der 1990er Jahre wurde bemerkt, dass 1,2 Millionen Magnetbänder mit Daten aus 30 Jahren Raumfahrt nicht mehr benutzbar sind – teilweise wegen mangelnder Zuordnung zu den jeweiligen Missionen.

» Wir können uns keinen digitalen Alzheimer leisten.

Dr. Elisabeth Niggemann, Deutsche Bibliothek

Bis zu 20 Prozent der Informationen, die 1976 während der Viking-Mission zum Mars gesammelt wurden, sind weg, weil Speichermedien unlesbar geworden sind. 1979 wurden die von der Raumsonde „Pioneer“ vom Saturn übertragenen Daten auf Magnetbänder archiviert. Obwohl die Daten auf vier verschiedenen Datenträgern gespeichert waren, waren sie zwei Jahrzehnte später nicht mehr lesbar – die NASA hatte für keines der Medien mehr die passenden Lesegeräte.

Jetzt haben die NASA und Google eine Zusammenarbeit vereinbart, in deren Rahmen Google der Weltraumbehörde beim Verwalten großer Datenmengen helfen soll. Google errichtet dazu einen neuen Campus auf dem Gelände des NASA-Forschungszentrums in Silicon Valley, der doppelt so groß ist wie Googles Hauptquartier.

Oft sind es nicht ausrangierte Geräte, sondern alte Software, aufgrund derer Daten nicht mehr entziffert werden können. Selbst wenn alle Bits auf dem Datenträger noch lesbar sind, erschließt sich der Inhalt nur mit dem passenden Programm – Nullen und Einsen sind da wie Hieroglyphen. Stephen Abrams ist Direktor des Digital Library Programms an der Harvard Universität. Dass er die Journalistenfrage, wie viele verschiedene Datenformate es überhaupt gebe, nicht einmal annäherungsweise beantworten konnte, hat ihn sichtlich gewurmt. In einem →

VERGÄNGLICHE BILDER

Erinnerungsfotos einst und heute: Das älteste erhaltene Foto aus dem Jahr 1827 (oben) wurde auf Asphalt belichtet. Das 170 Jahre später entstandene Bild (unten) hat der Besitzer auf CD-ROM gespeichert – leider als JPEG. Das komprimierte Dateiformat macht aus einem einzigen falsch gesetzten Bit einen groben Bildfehler.

Vortrag am folgenden Tag lieferte er Zahlen aus dem Internet: Eine Seite, die sich rühmt, „jedes Format in der Welt“ zu haben, listet 3.189 Extentions auf.

„Glauben Sie aber nicht, dass das schon alle sind“, warnt Stephen Abrams. Neben exotischen Formaten kommen noch verschiedene Versionen desselben Formats hinzu. Allein das PDF-Format existiert in 60 verschiedenen Varianten, die mitunter Inkompatibilitäten zeigen. „Die Industrie ist interessiert daran, dass alte Daten mit neuen Programmversionen nicht mehr lesbar sind“, kommentiert Elmar Mittler, Leiter der Göttinger Bibliothek, das Durcheinander, mit dem das Bemühen der Archivare konterkariert werde. In Harvards Archiv liegen 97 Prozent aller elektronischen Dokumente in neun Formaten vor: AIFF, ASCII, GIF, HTML, JPEG, PDF, TIFF, WAVE und XML. 90 weitere Formate bilden die restlichen 3 Prozent.

Um ältere Daten zu lesen, werden verschiedene Strategien eingesetzt. Die einfachsten sind Migration, also die Daten in moderne Formate zu überführen, und Emulation, bei der ältere Rechner auf neuen Maschinen simuliert werden. Aufwendiger ist es, ein „Museum“ mit Lesegeräten für veraltete Lochkarten, Disketten und Bänder zu unterhalten. Um das Risiko von Datenverlust zu verkleinern, speichert man auf verschiedenen Medien – oft auf Magnetbändern und zusätzlich auf CDs.

Als großes Schwarzes Loch entpuppt sich immer wieder das Internet. So verweist die Deutsche Bibliothek auf ihrer Homepage zu Ergebnissen des Projekts „Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Resourcen“ auf die entsprechende Website – doch der Link führt ins Leere: „Pagina niet gevonden / Page not found“ antwortet die Königliche Bibliothek der Niederlande auf den Mausklick.

In der Zukunft könnten Geld und Wissen fehlen

„Error 404“, die Fehlermeldung für nicht gefundene Seiten, kennt John Kunze nur zu gut. Sein Job an der California Digital Library ist die Archivierung von Webseiten. „Einfacher Text ist das einzige Format, das mit heutigen Computern noch genauso lesbar ist wie es vor 30 Jahren war“, stellt Kunze fest. Wahrscheinlich werde es auch in 30 Jahren noch lesbar sein, vielleicht sogar als einziges Format aus der heutigen Zeit. In Kalifornien speichert man daher die Originalseiten, zusätzlich wird aber auch eine Sicherung in ASCII-Format angelegt. Sollten die multimedialen Inhalte eines Tages nicht mehr darstellbar sein, bliebe wenigstens der Text erhalten.

» Unformatierter Text ist das einzige, was wir in 30 Jahren noch lesen können.

John Kunze, California Digital Library

Eine weitere Möglichkeit ist, gleich ein Rasterbild der Seiten zu erzeugen. Bessere Darstellungs-Tools als heute werden für die aktuellen Formate nie existieren. Für diese „Datenaustrocknung“ gibt es in Zukunft möglicherweise weder das Geld noch das Wissen, um es nachzuholen. Auf Papier gedruckte Dokumente sind deshalb so dauerhaft, weil sie keine komplizierten Geräte benötigen: Licht reicht zur Betrachtung aus.

In Zeiten der digitalen Fotografie hat die CD den Negativstreifen ersetzt. Nach wie vor sind Papierbilder jedoch sehr beliebt. Spätestens wenn die Bilder verblassen oder Abzüge sich verfärben, greifen Fotografen zum Speichermedium, um Ersatz zu drucken. Mitunter gibt es dann ein böses Erwachen: Wenn Papierbilder von der Zeit gezeichnet sind, könnte auch die CD bereits Schaden genommen haben. Sie kann zwar nach Herstellerangaben bis zu 100 Jahre halten, trägt aber möglicherweise schon nach dem Brennen den Keim für eine viel schnellere Zerstörung. Nicht ohne Grund kopieren professionelle Archivare ihre CDs alle fünf Jahre um.

manfred.flohr@chip.de

LINKS

www.langzeitarchivierung.de: Kompetenznetzwerk Langzeitarchivierung (Nestor)
www.ddb.de/index_txt.htm: die Deutsche Bibliothek
www.gutenbergdigital.de: Gutenberg-Bibel digitalisiert
whatis.techtarget.com/fileFormatA/0,289933,sid9,00.html: Verzeichnis von Dateiformaten
www.archive.org: Internet-Archiv

→ Teil 2 in der nächsten CHIP: Speichertechnologien der Zukunft