WISSEN

LANGZEIT-ARCHIVIERUNG IN DER WISSENSCHAFT

Daten für die Ewigkeit

Damit kann kein digitales Medium mithalten: In Stein gemeißelte Informationen – im Bild der berühmte Stein von Rosette – halten Tausende von Jahren.
Quelle: British Museum Press

Digital gespeichertes Wissen birgt hohe Verlustrisiken. Viele Medien halten kaum länger als zehn Jahre. Eine Reihe von Institutionen versucht, das Problem in den Griff zu bekommen. Daraus ergeben sich auch Tipps für die persönliche Archivierung.

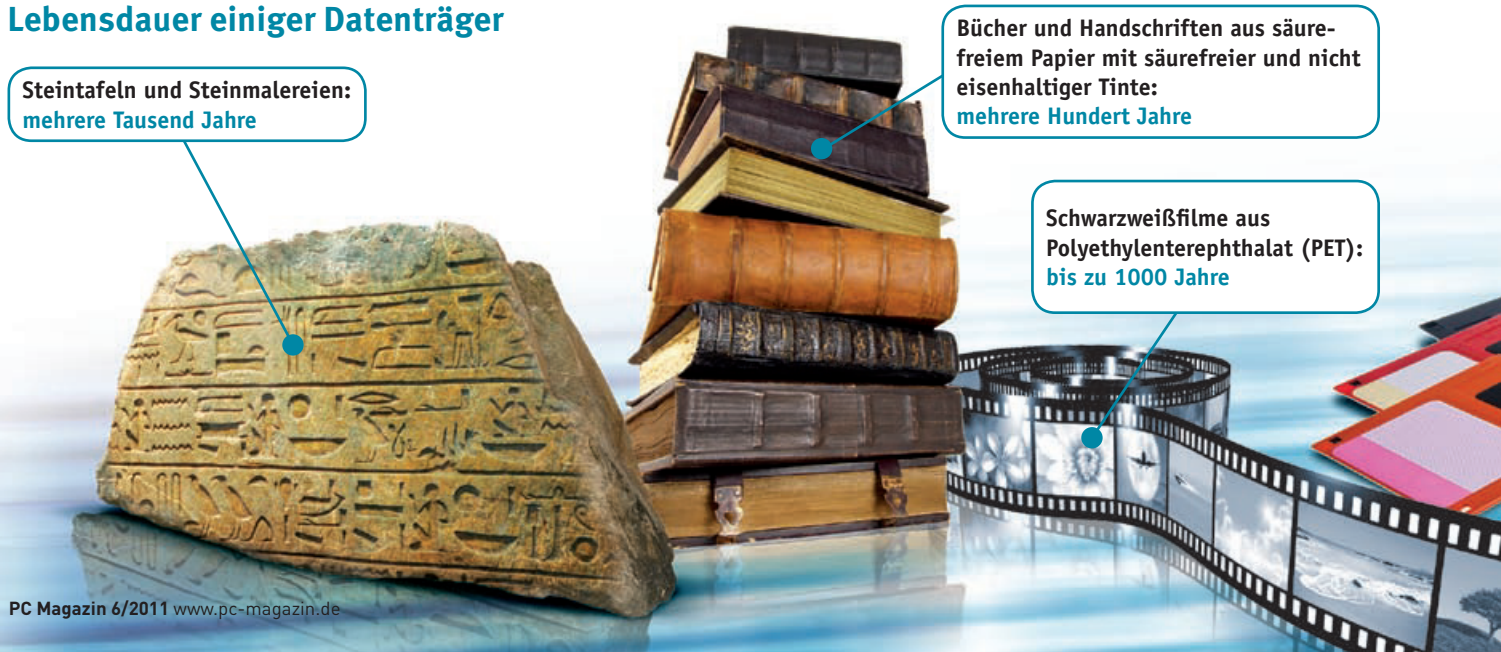
VON **DR. KLAUS MANHART**

Ein großer Teil des Wissens der Menschheit ist heute digital gespeichert – und droht unwiederbringlich verloren zu gehen. Sind die Daten auf CD gespeichert, greifen chemische oder physikalische Einflüsse das Trägermaterial an. Nach 25, vielleicht 80 Jahren, sind sie unlesbar. Das gilt selbst bei optimalen Lagerbedingungen – wie das Deutsche Musikarchiv erfahren musste. Dort sind bereits 200 Musik-CDs, die zwischen 1983 und 1986 zur Archivierung eingegangen sind, unbrauchbar – nach knapp 25 Jahren zerstört durch aggressive Lacke des Labelaufdruckes.
Selbstgebrannte CDs und DVDs schaffen oft nur fünf bis zehn Jahre. Das US-Nationalarchiv warnt gar in seiner FAQ, CDs und DVDs könnten bereits nach zwei bis fünf Jahren nicht mehr lesbar sein. Bänder verlieren ihre Magnetisierung nach zirka 20 bis 30 Jahren. Und wie lange Festplatten halten, hängt extrem von ihren Einsatzbedingungen ab. Sicher ist: Bei keinem digitalen Speichermedium ist garantiert, dass es nach mehr als zehn Jahren noch gelesen werden kann.

Sind .DOC-Dateien in 10 Jahren noch lesbar?

Doch nicht nur die Haltbarkeit des Mediums ist ein Problem. Printprodukte unterliegen nur dem zeitlich bedingten Materialverfall. Bei digitalen Dokumenten kommt auch der technologische Fortschritt hinzu.
Ein Aspekt ist die Formatfrage. Wer kann heute sicherstellen, dass es Formate wie *.DOC* oder *.XLS* in zehn oder zwanzig Jahren noch gibt? Irgendwann verschwindet jedes Format oder wird gravierend geändert. Dann sind ältere Dokumente wegen fehlender Kompatibilität nicht mehr nutzbar.

Lebensdauer einiger Datenträger



Und es gibt noch ein drittes Problem: Im Gegensatz zu Printprodukten können elektronisch gespeicherte Informationen ohne geeignetes Lesegerät überhaupt nicht dargestellt werden. Aktuelle Rechner haben kein Diskettenlaufwerk mehr. Schon heute ist es schwierig, eine in den achtziger Jahren auf 3½-Zoll oder gar 5¼-Zoll-Disketten gespeicherte Diplomarbeit zu lesen. In zehn Jahren gibt es mit großer Wahrscheinlichkeit keine CD- und DVD-Laufwerke mehr.

Dass die Lesegeräte über viele Jahre hinweg verfügbar sind, garantiert niemand. Das musste schon die NASA in den 90er Jahren leidvoll erfahren, als sie auf Daten der Saturnmission der Raumsonde *Pioneer* nicht mehr zugreifen konnte. Trotz redundanter Speicherung auf verschiedenen Datenträgertypen waren keine entsprechenden Lesegeräte mehr vorhanden.

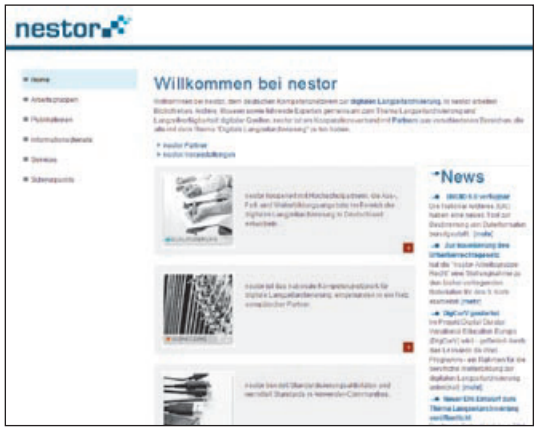
Digitales Erbe bewahren

Da digitale Daten inzwischen zentraler Bestandteil der kulturellen und wissenschaftlichen Überlieferung sind, haben Hochschulen, wissenschaftliche Rechenzentren, Museen und vor allem Bibliotheken ein besonderes Interesse an dem Thema Langzeit-Archivierung.

Die letzten Ignoranten wollte die UNESCO 2003 mit ihrer *Charter on the Preservation of the Digital Heritage* wach rütteln. Die Charter betont in Artikel 1 den dauerhaften Wert und die Bedeutung vieler digitaler Materialien als Teil des kulturellen Erbes, das für künftige Generationen geschützt und bewahrt werden muss. Zum digitalen Erbe gehören neben Texten, Fotografien, Musik, Filmen und Multimediawerken auch Webseiten und elektronisches Verwaltungsschriftgut.

Sicher ist: Die mit der Bewahrung von Kulturgut betrauten Institutionen brauchen Strategien gegen den drohenden Verlust von Information. Mittlerweile gibt es viele verschiedene Projekte und Initiativen, die sich in Europa und den USA damit befassen, wie sich digitale Daten als Quellen für Wissenschaft und Forschung langfristig verfügbar halten lassen. In Deutschland ist dies vor allem das *Kompetenznetzwerk Langzeit-Archivierung und Langzeitverfügbarkeit digitaler Ressourcen* — kurz: Nestor. Das Netzwerk bündelt das Know-how und die Kompetenzen für die digitale Langzeit-Archivierung. Vertreter von „Gedächtnisinstitutionen" – Archive, Bibliotheken, Museen, Rechenzentren – arbeiten dort unter Federführung der Deutschen Bibliothek in Frankfurt an einem nationalen Konsens zur Organisation der Langzeit-Archivierung in der Bundesrepublik.

Dabei nimmt sich Nestor aller Aspekte der langfristigen Bewahrung an – das heißt für die Archivare die Erhaltung über Generationen von technischen Systemplatt-



Das Nestor-Netzwerk bündelt das Know-how und die Kompetenzen im Bereich der digitalen Langzeit-Archivierung.

Vorsintflutliche Datenträger? Daten auf Disketten sind heute schon kaum mehr lesbar.

Tipps zur Langzeit-Archivierung

Die vorgestellten Projekte sind vor allem auf die wissenschaftliche Archivierung ausgelegt. Eine unmittelbare Übertragung auf andere Bereiche ist derzeit kaum möglich. Dennoch schälen sich einige Strategien heraus, die auch im Unternehmen oder Privatbereich angewendet werden sollten.

Hier die wichtigsten Faustregeln:

1 Verwenden Sie möglichst wenige Datenformate.

2 Verwenden Sie möglichst offene und standardisierte Dateiformate. Also statt *.DOC, XLS* oder *BMP* besser das OpenDocument-Format *ODF, PDF* oder *TIFF*. Diese gelten auch als besonders langlebig.

3 Sie sollten das Archiv regelmäßig aktualisieren und Archive mindestens doppelt anlegen.

4 Halten Sie sich bei der Vergabe von Dateinamen an den 8.3-Standard: Acht für den Namen, drei für den Dateityp, etwa: *beispiel.pdf*

5 Bewahren Sie das Backup räumlich getrennt vom Original auf.

6 Alle drei bis fünf Jahre sollte das komplette Archiv plus Backup vom alten Datenträger auf einen neuen migrieren.

7 Verwenden Sie keine CDs oder DVDs zur Archivierung, sondern statt dessen Festplatten oder, im Unternehmensbereich, Magnetbänder. Festplatten sollten staubdicht lagern und gelegentlich gestartet werden.

8 Im professionellen Bereich bieten sich zur Verwaltung des Archivs Dokumentenmanagementsysteme an.

Weitere konkrete Tipps finden Sie in der FAQ von *langzeitarchivierung.de*

Quelle: Treventus

Scan-Robotor wie von der österreichischen Firma Treventus digitalisieren Bücher und andere gebundenen Druckerzeugnisse automatisch und schonend.

formen und Nutzern hinweg. Wo sich die internationale Fachwelt bei der Suche nach langfristigen Aufbewahrungsstrategien befindet, haben Projektmitarbeiter in dem kostenfreien PDF-Handbuch *Kleine Enzyklopädie der digitalen Langzeit-Archivierung* zusammengefasst.

Archivbeispiel Leibniz-Rechenzentrum

Auch in den wissenschaftlichen Rechenzentren ist Langzeit-Archivierung ein großes Thema. Primärdaten aus Studien und Experimenten müssen laut den Richtlinien der Deutschen Forschungsgemeinschaft (DFG) mindestens zehn Jahre aufbewahrt werden, einzelne Einrichtungen wie Unikliniken wollen ihre Daten 30 Jahre und länger halten.

Das Leibniz-Rechenzentrum in Garching bei München beteiligt sich an einer kontinuierlich wachsenden Zahl von Archivierungs-Projekten.
Foto: Christoph Rehbach

Diese Institutionen kämpfen, neben dem Medienproblem, vor allem mit riesigen Datenmengen. Beim Leibniz-Rechenzentrum (LRZ) in München z.B., dem IT-Dienstleister für alle Münchner Hochschulen, wurde bereits 2007 die 3000-Terabyte-Grenze überschritten, die Hälfte davon sind Archivdaten.

DVDs kommen bei diesen Datenmengen als Speichermedium nicht in Frage, stattdessen archiviert das LRZ auf Magnetbändern. Die Bandtechnologie, die dabei zum Einsatz kommt, ermöglicht die Speicherung von bis zu 1000 GByte auf einem Magnetband. Theoretisch sind dafür zwar auch Festplatten als Datenträger geeignet. Doch die gelten bei den großen Datenmengen wegen der hohen Energiekosten als nicht effizient genug. Im Gegensatz zu Bändern muss man nämlich in Platten auch die Wärme wieder abführen, die beim Betrieb entsteht. Für tausende von Festplatten ist eine große, teure Klimaanlage nötig. Magnetbänder hingegen brauchen im Gegensatz zu Festplatten kaum Energie und produzieren auch keine Wärme.

Die Verwaltung der Archivdaten erfolgt mit einer speziellen Archivierungs-Software, mit der sich verschiedene Policies definieren lassen. Über diese lässt sich festlegen, wie lange welche Daten aufbewahrt werden, wie viele Versionen es geben soll und wann welche Daten gelöscht werden sollen.

Mehrfach gesichert

Um digitale Daten über Jahrzehnte hinweg zu erhalten, sieht das LRZ nur eine praktikable Lösung: Die Informationen müssen nach einigen Jahren auf neue Datenträger kopiert – im Fachjargon: „migriert“ – werden. Mehr als fünf Jahre werden die Daten in der Regel am LRZ nicht auf dem gleichen Medium gehalten.

Mit dieser Strategie schlägt das LRZ zwei Fliegen mit einer Klappe: Man senkt die Wahrscheinlichkeit, dass Daten infolge mangelnder Haltbarkeit des Datenträgers verloren gehen. Und: Man bleibt auf dem aktuellen Stand der Technik der Lesegeräte.

Aufbewahrt werden die Bänder in vollklimatisierten Räumen mit konstanter Temperatur und Feuchtigkeit, die mehrfach gegen alle möglichen Katastrophen abgesichert sind. Sollte es trotz aller Vorsichtsmaßnahmen doch einmal zu einer völligen Zerstörung des Rechenzentrums kommen, bleiben immer noch die Nachbarn: Die wichtigsten Daten werden an das einige hundert Meter entfernte Rechenzentrum der Max-Planck-Gesellschaft in Garching kopiert.

Know-how und technische Infrastruktur machen das LRZ zu einem begehrten Partner bei Archivierungs-Projekten. So führt das LRZ mit der Bayerischen Staatsbibliothek mehrere Projekte durch. Zusammen mit Google erfolgt etwa die Massendigitalisierung des urheberrechtsfreien historischen Bestandes an Druckwerken – über eine Million Titel mit über 250 Millionen Seiten.

Bereits abgeschlossen ist das von der DFG geförderte Projekt *vd16digital*, bei dem zwischen 2007 und 2009 das deutsche Schriftgut des 16. Jahrhunderts eingescannt und archiviert wurde. Die Buchseiten wurden über Scan-Roboter eingelesen, die automatisch umblättern können, und anschließend als Bilddaten im TIFF-Format gespeichert – TIFF ist weit verbreitet, gilt als zukunftssicher und besitzt eine sehr hohe Farbtreue. Später sollen die TIFFs per OCR-Software eingelesen werden, um Texte per Volltextsuche zu erreichen.

Die Formatfrage

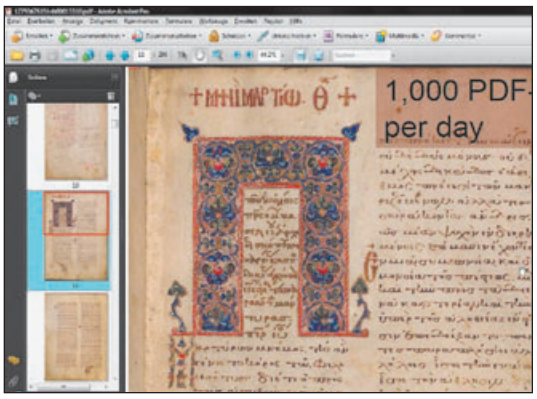
Mit dem Einlesen der Daten als TIFFs, um sie später per OCR lesbar zu machen, ist man auf der sicheren Seite. Dennoch ist es in der Praxis meist sinnvoll, Textdokumente in dafür geeigneten Formaten zu archivieren. Hier empfehlen Nestor und Institutionen wie das Bundesamt für Sicherheit in der Informationstechnik (BSI) die Beschränkung auf Standardformate.

Erste Wahl sind herstellerunabhängige Standards, die von anerkannten Organisationen wie der ISO oder dem W3C spezifiziert sind, zum Beispiel ASCII, Unicode, SVG und XSL.

Einige herstellerabhängige Formate haben sich als Quasi-Standards am Markt durchgesetzt, bestes Beispiel ist das PDF von Adobe. Die Spezifikation ist ebenfalls frei verfügbar, steht aber unter der alleinigen Kontrolle des Eigentümers.

PDF kommt eine besondere Bedeutung bei der Archivierung zu. 2005 hat die ISO das PDF/A-Format („A“ = Archive) als Standard für die Langzeit-Archivierung von Dokumenten zertifiziert. Das BSI empfiehlt dieses Format ebenfalls für die Langzeit-Archivierung. Seitdem wird dieses Format im Markt hoch gehandelt. Viele Hersteller sind bereits von den Vorteilen des PDF/A-Formates überzeugt und haben ihre Produktpalette entsprechend angepasst oder erweitert. Der Standard PDF/A (ISO Standard 19005-1) basiert auf PDF 1.4, schließt aber einige Funktionen davon aus, da sie eine langfristige Darstellbarkeit beeinträchtigen könnten. So ist die Verwendung externer oder spezifischer Ressourcen wie eingebetteter Fonts nicht erlaubt. Durch diese und andere detaillierte Vorschriften soll eine langfristige Lesbarkeit der Dokumente garantiert sein – und zwar unabhängig davon, mit welcher Anwendungs-Software und auf welchem Betriebssystem sie ursprünglich erstellt wurden.

Beim BSI finden Sie eine Liste von Formaten, die für die langfristige Archivierung von Text-, Bild-, Audio- und Videodateien geeignet sind (siehe Kasten). **pk**



Eine der Aufgaben der Bayerischen Staatsbibliothek ist die elektronische Archivierung historischer Bücher.

Die Bandbibliotheken im Daten- und Archivraum des Leibniz Rechenzentrums
Quelle: Bayerischen Akademie der Wissenschaften

LANGZEITARCHIVIERUNG IN WISSENSCHAFT UND KULTUR

Nestor-Projekt: www.langzeitarchivierung.de

Nestor-Handbuch: http://nestor.sub.uni-goettingen.de/handbuch/index.php

Kopal-Projekt: http://kopal.Langzeit-Archivierung.de

DigitalPreservationEurope (EU-Projekt): www.digitalpreservationeurope.eu

PDF/A Kompetenzcenter: www.pdfa.org/doku.php

Auswahl geeigneter Datenformate: www.bsi.bund.de/ContentBSI/grundschutz/kataloge/m/m04/m04170.html

Informationssammlung: www.uni-muenster.de/Forum-Bestandserhaltung/konversion/digi-langdat.html

Langzeit-Archivierungs-Projekte am LRZ: www.lrz.de/projekte/langzeitarchivierung

Google Projekt bei der Bayerischen Staatsbibliothek: www.bsb-muenchen.de/Massendigitalisierung_im_Rahme.1842.0.html