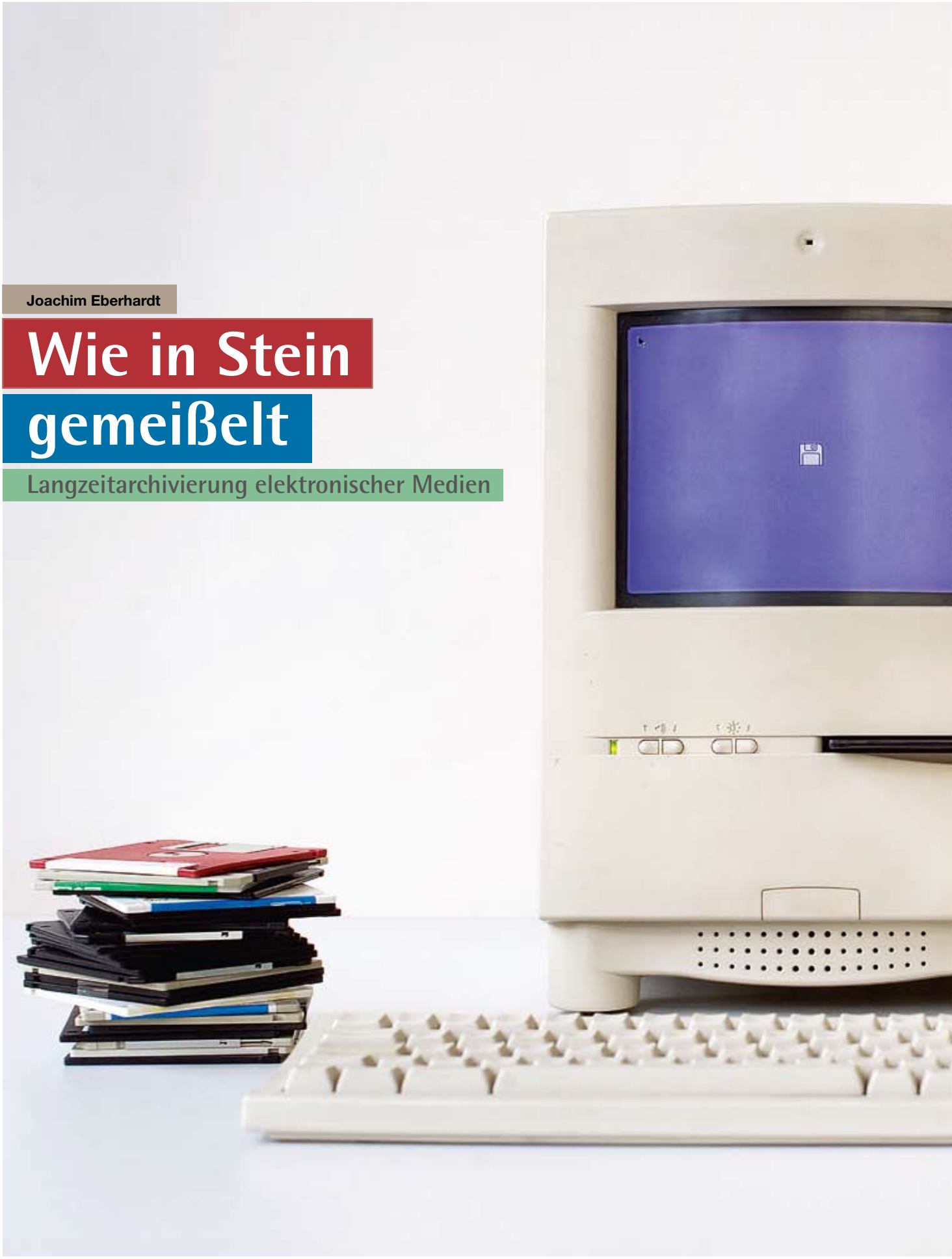
Joachim Eberhardt

Wie in Stein gemeißelt

Langzeitarchivierung elektronischer Medien



„Langzeitarchivierung“ von Medien ist ein junges Wort – und das lässt vermuten, dass das damit bezeichnete Phänomen auch jünger ist. Das Wort ist darum auffällig, weil es wie ein Pleonasmus wirkt. Ist Archivierung nicht per se für eine lange Zeit gedacht? Die Antwort ist: im Prinzip schon. Aber Langzeitarchivierung wird gerade in einem Zusammenhang verwendet, in dem die Vorstellung, etwas müsse über lange Zeit aufbewahrt werden, neu ist, weil die betroffenen Medien selbst noch „neue“ Medien sind. Die folgende Darstellung gibt einen pointierten Überblick über das Problem und stellt einige Lösungsansätze vor.

4000 Jahre Mediengeschichte

Doch zunächst lohnt sich ein kleiner Blick zurück auf die gut 4000 Jahre währende Geschichte der schriftlichen Medien. Medien werden seit ihrer Erfindung über ihren Entstehungszusammenhang hinaus bewahrt – sei’s zufällig, sei’s absichtlich. Gesetze und Verträge zum Beispiel sind Texttypen, die nur dann Rechtssicherheit garantieren, wenn sie auch später noch – im Zweifelsfall – konsultiert werden können. So gehört zu den frühesten heute noch lesbaren schriftlichen Zeugnissen der Codex Hammurabi, der um 1800 vor Christus in Mesopotamien in Basalt gehauen wurde. Die Entscheidung der Schreiber für den harten, nicht zu spröden Stein war auch eine Entscheidung dafür, den Text auf Dauer verfügbar zu halten.

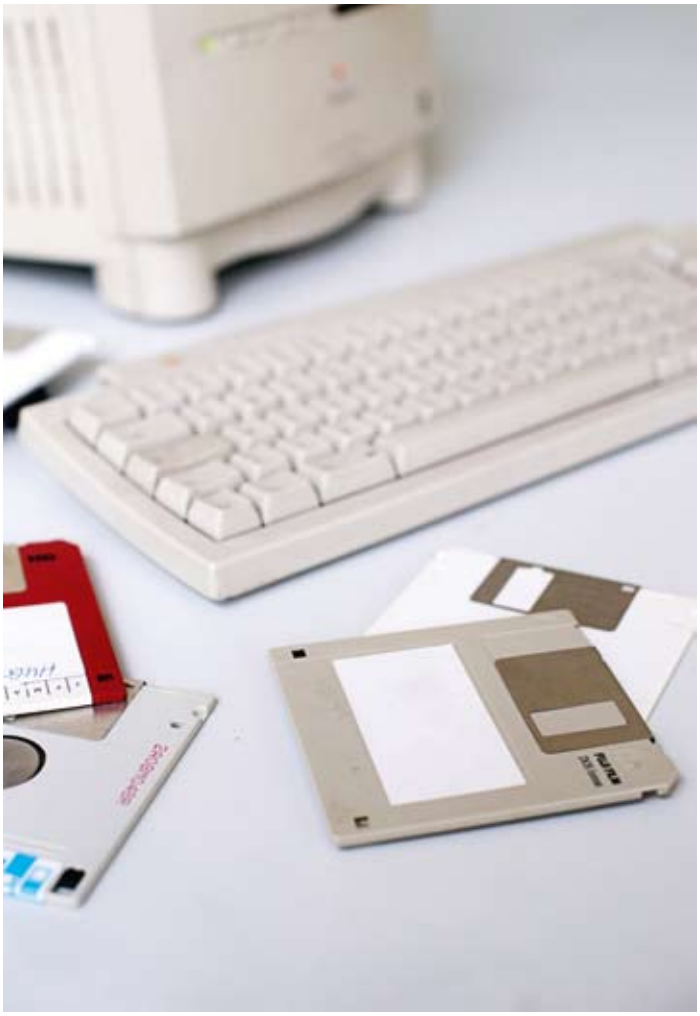
Die dann folgende Geschichte der textlichen Medien ist eine Geschichte des Bemühens, die Benutzung zu vereinfachen: Das gebundene Buch ist besser zu nutzen als die Schriftrolle oder gar der Stein, der Druck leichter und genauer herzustellen als die Handschrift. Das auf Holz basierende Papier ist billiger als das Hadernpapier und der elektronische Satz flexibler und billiger als der Bleisatz. Keine dieser Änderungen geschah mit Rücksicht darauf, ob sich die geschaffenen Medien danach besser würden archivieren lassen. Es überrascht daher nicht, dass einige dieser Verbesserungen unter archivischem Gesichtspunkt Nachteile haben. Holzhaltiges Papier etwa wurde mit einem säurehaltigen Bindemittel hergestellt, und diese Säure arbeitet im Papier bis zu seinem Zerfall – der sich schneller oder langsamer vollzieht, abhängig davon, wie das Druckwerk gelagert wird. Ideal sind: konstante Temperatur um 18 Grad, Luftfeuchtigkeit um 50 Prozent, keine Sonneneinstrahlung.

Dass das Papier zerfällt, wurde erst in der zweiten Hälfte des 20. Jahrhunderts als Problem bemerkt, und die Methoden zu seiner Bewältigung sind beispielhaft auch in unserem Zusammenhang. Der erste Schritt war, die Papierproduktion umzustellen auf säurefreies Papier: die Sicherung der Zukunft. Der zweite, Verfahren zur sogenannten Entsäuerung zu entwickeln, um die gefährdeten Medien physisch zu bewahren. Der dritte, die Information zu retten, falls das Medium nicht bewahrt werden konnte. Am leichtesten geschah und geschieht dies durch Abfotografieren, beispielsweise durch Mikroverfilmung oder – seit den 90er Jahren – Digitalisierung. Man nennt das Medienwechsel, und der ist vor allem dann eine gute Idee, wenn man für das neue Medium die Probleme schon gelöst hat, welche die Aufgabe des alten nötig machten. Bei der Mikroverfilmung zeigte sich bald, dass das nicht der Fall war: Auch hier wurden zunächst Filme eingesetzt, die zersetzende Chemikalien enthielten. Inzwischen sind diese abgelöst durch solche, für die Hersteller eine Lebensdauer von 500 Jahren angeben. Ob sie aber wirklich so lang halten werden, wissen wir nicht; schließlich hatte noch niemand Gelegenheit, die Alterung zu beobachten. Hoffen wir das Beste!

Elektronische Medien und Daten, Sicherung und Verfügbarkeit

Die Deutsche Forschungsgemeinschaft (DFG) bemerkte erstmals 1995 in einem Positionspapier „Elektronische Publikationen im Literatur- und Informationsangebot wissenschaftlicher Bibliotheken“, dass in der Wissenschaft vermehrt direktes elektronisches Veröffentlichen zu beobachten sei und darum die Langzeitsicherung elektronischer Publikationen künftig zu den Aufgaben von Bibliotheken gezählt werden müsse. Sie benannte – in ihrer trockenen Diktion – zwei wesentliche Charakteristika der elektronischen Publikationsform, nämlich: 1. „Eine dauerhafte, unauflösliche Verbindung zwischen Information und ihrem Träger besteht nicht“; 2. die Veröffentlichungen können „zentral an einer Stelle vorgehalten und ohne Abnutzung des Originals in orts- und zeitunabhängigen Zugriffen gelesen bzw. kopiert werden“. Beides lässt sich leicht konkretisieren: 1. Texte, Bilder, Präsentationen, Filme etc. mögen zwar auf Diskette, CD-ROM oder Flash-Speicher in einer Bibliothek ankommen; bewahrt werden müssen aber nicht diese physischen Medien, sondern nur die Daten darauf, also eben Texte, Bilder etc. 2. Via Internet, unabhängig vom physischen Datenträger, auf dem sie gespeichert waren, lassen sich diese Informationen am besten nutzen.

Die DFG hat so schon von Anfang an das Umkopieren von Daten als ein Hilfsmittel für die Langzeitsicherung vorgesehen. Die Frage liegt nahe, ob man nicht zusätzlich auch Lesegeräte bereithalten sollte für die originalen physischen Medien – auch wenn diese veralten (z.B. 5-1/4-Zoll-Disketten). Und zusätzlich zu diesen Lesegeräten die entsprechende Betriebssoftware und die Programme, die nötig sind, um mit den Datenformaten etwas anzufangen?



Wer heute alte elektronische Medien im Besitz hat und die Daten darauf benutzbar halten will, hat nur drei Möglichkeiten: Entweder richtet er ein „Hardwaremuseum“ ein, mit originaler Hardware (die irgendwann ausfallen wird) und Software (die auf Medien gespeichert sind, die auch irgendwann ausfallen werden). Oder er bietet den alten Daten mit geeigneter Emulationssoftware auf heutigen Computern eine simulierte historische Datenumgebung. Oder er überträgt die Daten in ein anderes Format. Denkbar ist hier auch der Medienwechsel zum Analogen, beispielsweise die sogenannte Ausbelichtung digitaler Bilder auf Mikrofilm.

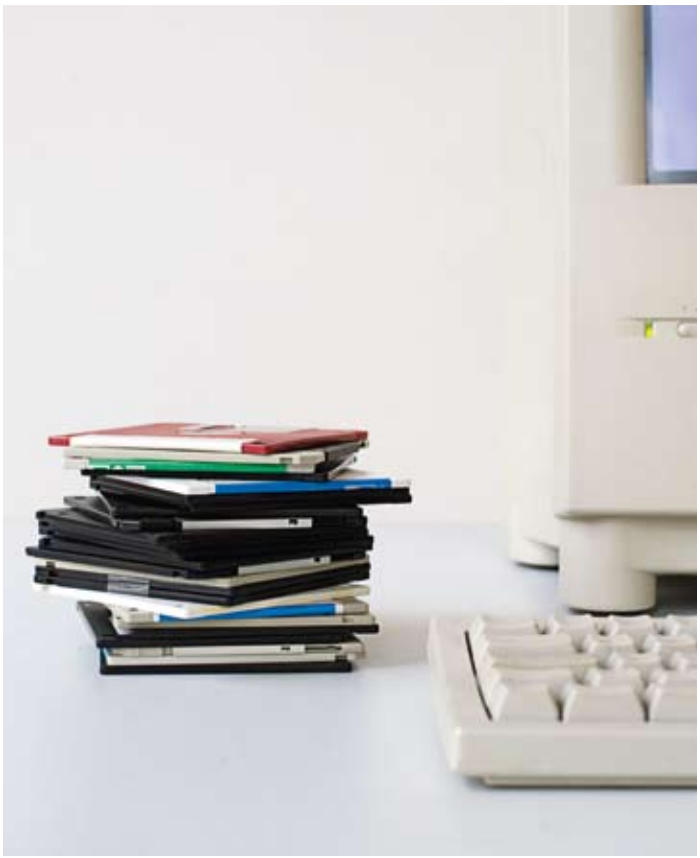
Die ersten beiden Strategien sind der Versuch, das Trägermedium (bzw. das Trägerformat) zu erhalten. Aber genauso wie das Verfahren der Entsäuerung im Bereich der gedruckten Bücher (entgegen seinem Namen) die Bücher weder entsäuert noch deren Zerfall stoppt (sondern ihn nur verzögert), erwirken sie nur einen Aufschub, bis die Daten nicht mehr lesbar sind. Zukunftsträchtig ist daher allein die dritte Strategie: Ihr Ziel ist es eben nicht, die Medien zu erhalten, sondern die Benutzbarkeit der gespeicherten Informationen. Es ist daher präziser, wenn man nicht vom Ziel der Langzeitarchivierung von Medien spricht, sondern davon, bei Daten die Langzeitverfügbarkeit zu sichern. Werfen wir nun einen Blick darauf, welchen konkreten Problemen die Universitätsbibliothek gegenübersteht und wie damit am besten umzugehen ist.

Rechte sichern

Die Universitätsbibliothek stellt den Angehörigen der Universität inzwischen eine Menge Texte und Informationen via Internet zur Verfügung. In der Regel besitzt sie diese Daten nicht selbst auf eigenen Rechnern, sondern hat nur den Zugang zu ihnen auf den Servern der Verlage lizenziert oder gekauft. Konkretes Beispiel sind die Elektronischen Versionen von Büchern des Springer-Verlags, für die die Universitätsbibliothek – übrigens auf Wunsch der Studierenden und aus Studienbeiträgen – Zugriffsrechte gekauft hat. Was passiert, wenn Springer sich aus dem Verlagsgeschäft zurückzieht oder Pleite geht? Für einen solchen Fall muss natürlich vertraglich vorgesorgt werden; damit die Bibliothek die Daten erhält, wenn der Verlag diese nicht mehr selbst vorhalten kann.

Vielfalt verwalten

Mehr Einfluss hat die Bibliothek auf die Sicherung der Langzeitverfügbarkeit, wenn sie die Datenträger selbst besitzt. Das ist in jüngerer Zeit häufiger der Fall. Beispielsweise kauft die Bibliothek Hörbücher auf CD, Filme auf DVD, Texte, Programme und Datenbanken auf CD-ROM oder DVD-ROM; in Zukunft vielleicht auch auf Flash-Speicher. Die einfachste Methode der Sicherung der Daten wäre ihre 1:1-Kopie auf einen gleichen Datenträger, also die Kopie einer CD auf CD; eine bessere Methode noch ist die Kopie auf einen leichter zu handhabenden Datenträger, z.B. die Kopie einer CD-ROM auf eine Festplatte. Allerdings spielt bei derlei Maßnahmen auch das Juristische eine Rolle. Das einschlägige Urheberrecht sagt nämlich, einen Datenträger dürfe man zu Archivierungszwecken kopieren – sofern man dafür nicht einen vom Hersteller eingebauten Kopierschutz umgehen müsse. In diesem Fall bleibt eigentlich nur eine legale Lösung: der Kauf eines Updates.



Die obige Aufzählung der Datenträger umfasst eine Gruppe von ganz unterschiedlichen Medien, die entsprechend eine Menge von Daten unterschiedlichster Formate enthalten. Gibt es eine allgemeine Strategie, wie mit solchen heterogenen Beständen und den vielfältigen Formaten umzugehen ist?

Das deutsche Bibliothekswesen hat seit einigen Jahren mit dem Nestor-Projekt (www.langzeitarchivierung.de) ein Kompetenznetzwerk der Langzeitsicherung aufgebaut. Die erste Frucht dieses Projekts war Ende 2007 KOPAL: ein elektronisches Archivsystem, das beliebige Daten sichern kann. Das System funktioniert so: Die zu sichernden Daten werden in das System gespeichert, zusätzlich mit standardisierten Angaben zum verwendeten Datenformat. KOPAL erhält regelmäßig Informationen über die Entwicklung von Datenformaten und sorgt selbst für die Datenkonversion in gebräuchlichere Formate, falls ein Datenformat unlesbar zu werden droht, weil die Software ausstirbt, die es liest. In der Theorie liest sich das gut, und jede deutsche Bibliothek kann das KOPAL-Archiv nutzen, indem sie dort Speicherplatz mietet. In der Praxis ist die KOPAL-Strategie aufwändig und entsprechend teuer – und bewährt hat sie sich auch noch nicht, dafür ist sie schlicht zu jung. Hier lohnt es sich, noch ein wenig abzuwarten.

Vereinfachen und auffindbar machen

Für eine andere Gruppe von Daten können wir uns allerdings das Warten nicht leisten: Seit es in den Prüfungsordnungen die Möglichkeit gibt, Dissertationen in elektronischer Form abzugeben, hat die Universitätsbibliothek die Dienstleistung übernommen, diese zu sichern und verfügbar zu halten. Hier ist sie in der Pflicht und muss daher schon jetzt entsprechende Konzepte entwickeln. Sie hat sich dafür entschieden, wie andere Universitätsbibliotheken auch, die elektronische Fassung der Dissertation nur im PDF-Format zu akzeptieren, und nicht beispielsweise als Word-, Open Office oder LaTeX-Datei. Das abzugebende Format ist standardisiert. Sollte das Format einmal obsolet werden, genügt es, eine einzige Lösung für die Konversion zu finden. Standardisierung ist das oberste Gebot der Langzeitverfügbarkeit!

Nun ist, leider, PDF nicht gleich PDF. Das Format wird von seinem Erfinder, der Firma Adobe, stetig weiterentwickelt, unter anderem, um interaktive Elemente aufzunehmen oder die Möglichkeit zu verbessern, Bilder einzubinden. Nötig ist daher eine präzisere Standardisierung. Das Zielformat ist künftig das sogenannte PDF/A, das 2005 von der Internationalen Organisation für Normung als ISO-Standard formuliert wurde. Ein einfaches PDF kann Speicherplatz sparen, indem es z.B. nicht die graphische Gestalt von Schriftzeichen enthält, sondern nur die Information, welches Schriftzeichen in welcher Schrifttype angezeigt werden soll. Wird ein solches PDF auf einem Rechner geöffnet, auf dem die gewünschte Schrifttype gar nicht vorhanden ist, verwendet das PDF-Leseprogramm dafür Zeichensätze, die es stattdessen auf dem Rechner findet. Dies kann dazu führen, dass einfache PDFs auf verschiedenen Rechnern unterschiedlich angezeigt werden – ein Authentizitätsproblem. PDFs mit nicht westlichen Zeichensätzen können auf westlichen Rechnern gar als komplett unverständliches Kauderwelsch erscheinen. Beim PDF/A ist das ausgeschlossen, da alle zur Anzeige notwendigen Informationen, auch z.B. die gewünschte Gestalt der Schriftzeichen, in der Datei enthalten sind. Ein PDF/A-Dokument sieht überall gleich aus: Damit ist die Zitierbarkeit gesichert. Allerdings ist dieses Format jünger als der Hochschulschriftenserver der Universitätsbibliothek. Die UB wird also in absehbarer Zeit die nicht als PDF/A-abgegebenen Dateien in das PDF/A Format konvertieren.

Für die reine Datensicherheit steht das Regionale Rechenzentrum Erlangen gerade, indem es die Inhalte des Schriften-

servers regelmäßig auf Magnetbänder sichert. Als Alternative oder Ergänzung solcher Backup-Verfahren ist in den letzten Jahren das sogenannte LOCKSS-Verfahren (www.lockss.org) bekanntgeworden, das an der Stanford University Library entwickelt worden ist. LOCKSS steht für Lots Of Copys Keep Stuff Safe und ist gleichzeitig der Name der verwaltenden Software und der teilnehmenden Gemeinschaft. LOCKSS sorgt automatisch dafür, dass mehrere Kopien eines elektronischen Dokuments auf den Servern der teilnehmenden Institutionen verteilt werden, so dass der Ausfall einer Speichereinrichtung für die Datensicherheit kein Problem darstellt. Brennt das Stanforder Rechenzentrum, sind die digitalen Hochschulschriften immer noch auf zig anderen Servern zu finden, und das System weiß, auf welchen. Damit ist zugleich ein anderes Problem gelöst, nämlich das der konstanten Webadresse. Die deutschen Bibliotheken setzen hier auf ein anderes Verfahren, nämlich die Vergabe von URNs. Im Unterschied zur URL, die sich ändern kann – denken Sie an die vielen toten Links, auf die Sie schon geklickt haben –, ist der „Uniform Resource Name“ eine Art Meta-Adresse, die eine Anfrage zu einer zentralen Datenbank schickt, wie denn nun die konkrete passende URL tatsächlich lautet. Die Universitätsbibliothek vergibt die URNs für die Hochschulschriften und sorgt dafür, dass der URN-Resolving-Dienst der Deutschen Nationalbibliothek weiß, welche URL gegenwärtig jeweils dazu gehört.

Retrodigitalisierung: Nichtproprietäre Formate

URNs vergibt die Bibliothek auch für ihre Retrodigitalisate. Mit diesem Ausdruck bezeichnet man meist die elektronische Darstellung der in konventionellen Medien enthaltenen Information. Die einfachste Form, ein Buch zu retrodigitalisieren ist, es in eine Textverarbeitung abzuschreiben. Da das aber viel Arbeitskraft kostet, kommt es weniger häufig vor als beispielsweise die digitale Fotografie von Büchern oder Handschriften. Lassen es die Schriftarten der Vorlagen zu, dann kann man über die Digitalbilder der Buchseiten einen Texterkennungsprozess laufen lassen, um eine elektronische Version des Textes zu erhalten. Ergebnis der Retrodigitalisierung sind entsprechend Bilder und, gelegentlich, Texte.

Für die Wahl des Datenformats der Bilder ist der wichtigste Punkt: Standardisierung. Der zweitwichtigste Punkt: Kompressionsfreiheit. Denn bei Datenkompression geht oft Information verloren. Darum speichert die Universitätsbibliothek ihre Retrodigitalisate als unkomprimierte Einzelbilder im TIFF-Format. Solche Bilder benötigen allerdings recht viel Speicherplatz, was sich hier ausbuchstabieren lässt als etwa 15 Megabyte für das farbige Bild einer A5-Seite in einer Auflösung von 300 Bildpunkten pro Zoll. Das kann man ohnehin zur Ansicht im Internet niemandem zumuten, daher werden für die Internetpräsentation eine Reihe von abgeleiteten Dateien hergestellt in niedrigerer Bildqualität, die ohne Schwierigkeiten und schnell im Webbrowser angezeigt werden können. Für die Langzeitsicherung sind diese „Derivate“ nicht von Belang, da sie jederzeit nach Bedarf neu hergestellt werden können; gesichert werden muss nur der digitale Master, das TIFF-Bild.

Gelingt es, durch Texterkennungsverfahren auch Textdaten zu erhalten, dann stellt sich die Frage, wie diese am besten zu sichern sind. Texterkennungsprogramme können in der Regel den Text in einem von einer Textverarbeitung akzeptierten Format ausgeben, z.B. im bekannten Word-Format DOC. Das ist allerdings zur Langzeitsicherung ungeeignet, weil es ein geschlossenes Format ist, für das man idealer Weise eine einzige bestimmte Software braucht, die es liest, und von der man nicht weiß, ob es sie in 100 Jahren noch gibt. Bibliotheken ziehen daher XML-basierte Formate wie z.B. TEI-XML vor, die zwei Vorteile bieten: Sie lassen sich erstens in jedem beliebigen Editor öffnen, weil sie nur aus Zeichen einer definierten Zeichenmenge bestehen, und beschreiben zweitens auch die Struktur eines Textes.

Würde sich ein XML-Format auch für die elektronischen Hochschulschriften eignen? Im Prinzip schon – aber man darf nicht übersehen, dass das PDF/A mehr leistet, als bloß den Text wiederzugeben, der aufgeschrieben wurde. Es gibt den Text in genau der Form wieder, in der er gestaltet wurde und bietet zugleich Mittel, die Manipulation der Datei zu verhindern bzw. zu bemerken – eben diese Eigenschaften machen das Format geeignet für die zitierbare Veröffentlichung im Internet. XML-basierte Formate haben ihre Stärken dort, wo die Struktur des digital gebotenen Texts ebenfalls von Interesse ist, also z.B. bei digitalen textkritischen Editionen. PDF/A und XML erfüllen unterschiedliche Funktionen und bringen dafür unterschiedliche Fähigkeiten mit.

Gemeinsame Strategien

Das Bemühen um die Langzeitverfügbarkeit elektronischer Daten kann nur als Gemeinschaftsunternehmen Erfolg haben. Standardisierung von Metadaten, Weiterentwicklung von Datenformaten, dauerhafte Webadressen: das sind Mechanismen und Werkzeuge, die von vielen gemeinsam verwendet werden müssen, um sowohl technisch erprobt als auch wirtschaftlich zu sein. Für die Universitätsbibliothek bedeutet dies, dass sie sich in ihren Aktivitäten am Kontext des Bayerischen Bibliotheksverbundes und an den von der Deutschen Nationalbibliothek getragenen nationalen Strategien orientiert – und diese einsetzt, wenn sie sich für den Routinebetrieb eignen.

Dr. Joachim Eberhardt ist Stellvertretender Direktor der Lippischen Landesbibliothek Detmold und war vorher in der Universitätsbibliothek Erlangen-Nürnberg für Digitalisierung zuständig.