Nachhaltige Nutzung digitaler Dokumente

Diplomarbeit

im Fach
Informationsvermittlung
Studiengang Informationsmanagement

Fachhochschule Stuttgart – Hochschule der Medien
Fachbereich Information und Kommunikation

Claudia Neumann

Erstprüferin: Prof. Margarete Payer
Zweitprüfer: Prof. Dr. Wolfgang von Keitz

Angefertigt in der Zeit vom 01. April 2003 bis 30. Juni 2003

Stuttgart, Juni 2003

Kurzreferat

Mit der steigenden Anzahl digitaler Dokumente steigt auch der Bedarf an geeigneten Archivierungssystemen. Die Aufgabe dieser Archivierungssysteme ist es, digitale Dokumente nicht nur zu erhalten, sondern auch ihre nachhaltige Verfügbarkeit zu gewährleisten. Für die Bewältigung dieser Herausforderung müssen sowohl technische als auch organisatorische Probleme gelöst werden. Die vorliegende Arbeit beschreibt diese Probleme und stellt Lösungsmöglichkeiten vor.

Schlagwörter

Nachhaltigkeit ; Digitale Dokumente ; Elektronische Ressourcen ;
Langzeitarchivierung

Abstract

With the exponential growth of digital documents the requirement for appropriate preservation systems is growing too. The function of such preservation systems is not only the preservation of digital documents, they also have to guarantee durable access. To manage this challenge preservaton systems have to solve technical problems as well as organisational problems. In this diploma thesis those problems and possible solutions will be described.

Key Words

Sustainability ; Digital Documents ; Electronic Ressources ; Long-Term-Preservation

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die rasante Entwicklung der Informationstechnologie im Hard- und Softwarebereich, einhergehend mit der Entwicklung des Internets und hier vor allem dem World Wide Web (WWW) eröffnet immer mehr Anwendern die Möglichkeit, digitale Dokumente auf einfache Weise zu erzeugen und kostengünstig weltweit zu publizieren.

Mit der wachsenden Anzahl der Publikationen steigt gleichzeitig das Risiko des Informationsverlustes. In diesem Zusammenhang wird deshalb bereits vom Verlust des digitalen Erbes oder auch einem „digitalen Alzheimer“ (Roetzer 2003) gesprochen.

Die Institutionen, die mit der Aufgabe der Sammlung, Bewahrung und Bereitstellung von dieses kulturellen Erbes für die Nachwelt betraut sind, sind die Bibliotheken und Archive. Sie stehen bei der Erfüllung ihres Auftrages, im Hinblick auf die wachsende Flut digitaler Publikationen vor neuen Herausforderungen.

Als besonders schwierig für die Entwicklung von Lösungsstrategien, die es ermöglichen, digitale Dokumente nicht nur zu archivieren, sondern ihre Benutzbarkeit auch in der Zukunft zu gewährleisten, erweist sich die besondere Vielschichtigkeit der Problematik. So gilt es, neben ausschließlich technischen Verfahren auch organisatorische und rechtliche Probleme zu lösen. Unter dem Begriff der *Langzeitarchivierung* beschäftigen sich deshalb seit einigen Jahren zahlreiche nationale und internationale Initiativen mit der Lösung dieser Problematik.

1.1 Motivation

Die vorliegende Arbeit beschäftigt sich aus der Sicht der Bibliothken und Archive mit der Problematik der Langzeitarchivierung. Sie will einen Überblick über die zu lösenden Probleme verschaffen und bereits vorhandene Lösungen aufzeigen. Das Augenmerk wurde hier auf die technischen und organsatorischen Bereiche gelegt. Die ebenfalls interessanten, rechtlichen, sozialen und finanziellen Aspekte konnten in dieser Arbeit nicht berücksichtigt werden.

1.2 Zielsetzung

Das Ziel der Arbeit ist es, einen Überblick über den aktuellen Stand der Langzeitarchivierung digitaler Dokumente in öffentlichen Einrichtungen zu geben.

1.3 Vorgehensweise

Nach der Begriffsbestimmung wird auf die Charakteristika digitaler Dokumente und ihre Unterscheidung zu analogen Dokumenten eingegangen. Anschließend sollen mit der Darstellung der Gegenwärtigen Situtation die einzelnen Problembereiche verdeutlicht werden. Abschließend werden Lösungsansätze aus organisatorischer und technischer Sicht dargestellt.

2 Digitale Dokumenttypen

Digitale Dokumente bieten gegenüber den herkömmlichen, analogen Dokumenten Vorteile bezüglich der Darstellung, Verarbeitung, Verteilung und Speicherung von Informationen. Doch häufig sind es gerade diese, für den Anwender attraktiven Eigenschaften, die bei der Archivierung und nachhhaltigen Nutzung von digitalen Dokumenten Probleme verursachen.

In diesem Kapitel soll deshalb zunächst dargestellt werden, wodurch sich analoge von digitalen Dokumenten unterscheiden, wie digitale Dokumente enstehen und welche Eigenschaften sie besitzen. Abschließend werden unterschiedliche digitale Dokumenttypen vorgestellt.

2.1 Begriffsbestimmung

2.1.1 Dokument

Der Begriff Dokument kann nach Brugger folgendermaßen definiert werden:

> „Ein Dokument ist somit eine Einheit für die Darstellung, Organisation, Verteilung und Archivierung von Information, sei es textueller oder graphischer Form. Dokumente können auf einer Vielzahl von Medien dargestellt werden.“ (Brugger 1998)

Bis zur Entstehung und Anwendung der digitalen Technologie konnten Dokumente nur auf analogen Datenträgern dargestellt werden. Neben Papier, das der vermutlich am häufigsten verwendete analoge Datenträger ist, gelten auch Steintafeln oder Münzen als Beispiele für analoge Datenträger. Mit der Weiterentwicklung der analogen Technik enstanden neue Aufzeichnungsmöglichkeiten und neue Datenträger. Es wurden Ton- und Videoaufzeichnungen möglich, die auf Datenträgern wie Schallplatten oder Magnetbändern gespeichert und mit den entsprechenden Abspielgeräten wiedergegeben werden konnten.

Erst durch die Anwendung digitaler Technologie können Dokumente sowohl auf anlogen Datenträgern (Ausdruck auf Papier) als auch auf digitalen Datenträgern wie beispielsweise der *Harddisk* (Fesplatte), der *Floppy Disk* (Diskette), der *CD-Read-Only-Memory* (CD-ROM), der *Digital Versatile Disk* (DVD) oder dem *Digital Audio Tape* (DAT) dargestellt und gespeichert werden. Während für analoge Dokumente verschiedene Wiedergabegeräte benötigt werden, kann bei digitalen Dokumenten meist der

Personal Computer (PC) als universelles Abspielgerät eingesetzt werden (vgl. Endres und Fellner 2000, S. 15).

2.1.2 Digitales Dokument

Endres und Fellner definieren digitale Dokumente wie folgt:

> „Ein *digitales Dokument* ist eine in sich abgeschlossene Informationseinheit, deren Inhalt digital codiert und auf einem elektronischen Datenträger gespeichert ist, so daß er mittels eines Rechners genutzt werden kann.“ (Endres und Fellner 2000, S. 15)

Einschränkend fügen Endres und Fellner jedoch hinzu, daß aus ihrer Sicht nicht alle digital codierten Informationseinheiten auch als digitale Dokumente bezeichnet werden können. Nach dieser Auffassung handelt es sich bei *digitalen Objekten* und *digitalen Betriebsmitteln* (digital resources) nicht um Dokumente. So kann ein Dokument aus mehreren digitalen Objekten bestehen, die für sich alleine jedoch keine Dokumente darstellen. Dies ist zum Beispiel bei Abbildungen oder dem Inhaltsverzeichnis der Fall (vgl. Endres und Fellner 2000, S. 15).

Der Ausdruck digitales Betriebsmittel wird von Endres und Fellner (2000) verwendet, *„... wenn etwa im Zusammenhang mit dem Internet auch Betriebsmittel inbegriffen sind, die man nicht als Dokumente ansehen kann.“* (Endres und Fellner 2000, S. 15)[1]

Derartig spezifische Unterscheidungen sind in der englisch-sprachigen Literatur selten zu finden. Hier werden die allgemeinen Begriffe *electronic publications*, *electronic records*, *digital resources*, *digital objects* oder *digital documents* oftmals willkürlich verwendet, solange nicht von einem speziellen Dokumenttyp die Rede ist.

Eine Ausnahme stellt die Bezeichnung *electronic resources* dar. Im Rahmen der Bestrebungen, weltweit einheitliche, bibliographische Standards zu definieren entstand unter der Federführung der *International Federation of Library Associations and Institutions* (IFLA)[2] eine *International Standard Bibliographic Description for Electronic Resources* (ISBD(ER)). Derzufolge electronic resources nach IFLA (2003) auch eindeutig definiert sind:

> „Electronic resources consist of materials that are computer-controlled, including materials that require the use of peripheral (e.g. a CD-ROM player) attached to a computer; the items may or may not be used in the interactive mode. Included are two types of resources: data (information in

[1] Da Endres und Fellner in ihrem Buch *Digitale Bibliotheken* auch im Zusammenhang mit dem Internet keine konkreten Beispiele für digitale Betriebsmittel nennen, bleibt unklar, worauf sich der Unterschied zum englischen Begriff digital resource bezieht.

[2] http://www.ifla.org

> the form of numbers, letters, graphics, images and sound, or a combination thereof) and programs (instructions or routines for performing certain tasks including the processing of data). In addition, they may be combined to include electronic data and programs (e.g. online services, interactive multimedia).“

In der vorliegenden Arbeit sollen die Bezeichnungen digitales Dokument und elektronische Resource als Bezeichnung für alle Arten von digitalen Dokumenttypen gelten. Ist von einem speziellen Dokumenttyp die Rede, so geht dies aus dem Text hervor.

2.2 Erzeugung digitaler Dokumente

Die Erzeugung von digitalen Dokumenten kann auf zwei Arten erfolgen. Erstens durch die Produktion von originär digitalen Dokumenten. Zweitens durch die Konvertierung analoger Dokumente in eine digitale Form, der sogenannten *Digitalisierung*.

Diese Unterscheidung spielt für den Archivierungsprozess eine wichtige Rolle. Während analoge und digitalisierte Dokumente als abgeschlossene Einheiten, *„die einen hohen Grad an Endgültigkeit erreicht haben“* (Endres und Fellner 2000, S. 119) betrachtet werden können, zeichnen sich originär digitale Dokumente durch die Möglichkeit der ständigen Veränderbarkeit aus. Betrachtet man die elektronischen Ausgaben von Tageszeitungen oder Nachrichtendiensten, wird schnell deutlich, welche Vorteile und Nachteile mit dieser Eigenschaft verbunden sind. Was aus Benutzersicht der Vorteil gegenüber gedruckten Versionen ist, nämlich die schnelle oder sogar zeitgleiche Übertragung von Informationen, stellt für die Archivierung, insbesondere die Erschließung eines der größten Probleme dar (siehe Kapitel 4 auf Seite 24).

Abbildung 2.1 auf der nächsten Seite übernommen aus: Brugger (1998) veranschaulicht den Produktionsprozess digitaler Dokumente. Der Einfachheit halber wird von Textdokumenten ausgegangen. Damit sind Dokumente gemeint, die vorwiegend Text beinhalten aber auch Abbildungen oder Tabellen enthalten können.

Editieren: Zunächst erzeugt oder verändert der Autor mit Hilfe eines Editierprogrammes ein Dokument. Dieses erzeugte Dokument besteht aus dem *Dokumentinhalt* (Text und Bilder) sowie der *logischen Struktur* und den *logischen Elementen*.

Die logische Struktur zeigt, wie der Inhalt eines Dokuments strukturiert ist, also in welche logischen Elemente der Inhalt gegliedert ist. Beispiele für logische Elemente sind Kapitel, Abschnitte, Überschriften aber auch Zitate, Fußnoten oder andere besondere Funktionen von Textstellen. Der Inhalt und die logische Struktur ergeben zusammen das *logische Dokument*. Das logische Dokument kann auch als die Sicht des Autors und Lesers verstanden werden.

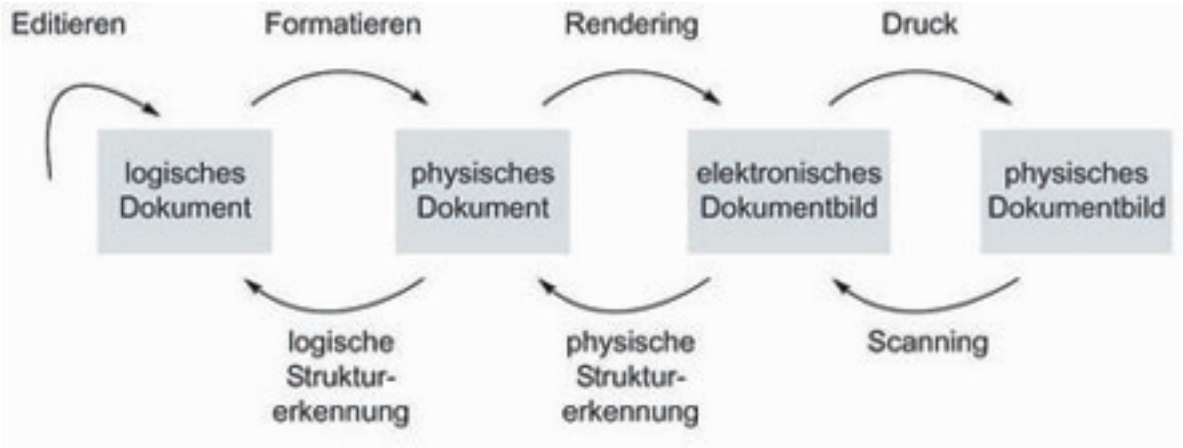

Abb. 2.1: Erzeugung und Erkennung von Dokumenten

Formatieren: Durch Formatierung wird anschließend das *physische Dokument* erzeugt. Dieser Vorgang entspricht der Tätigkeit des Schriftsetzers. Hier wird das Erscheinungsbild des Dokuments nach typographischen Regeln festgelegt. Dies geschieht zum Beispiel durch Auswahl der Schriftart, der Schriftgröße und der Ausrichtung des Textes. Brugger (1998) beschreibt physische Dokument als die Sicht des Setzers oder Graphikers, der einen Text angemessen darstellt.

Rendering: In einem weiteren Schritt wir das *elektronische Dokumentenbild* erstellt. Dieser Vorgang wird als *Rastern* oder *Rendering* bezeichnet. Es handelt sich hierbei um die Berechnung aus dem physischen Dokument. Die Bildauflösung wird dabei in der Regel dem Ausgabemedium angepaßt.

Druck: Als letzer Schritt in der Dokumenterzeugung erfolgt die Erstellung des *physischen Dokumentbildes* durch den Druck.

Die hier beschriebene Form der Dokumenterzeugung bezieht sich auf das Konzept des *generic coding* oder auch *generic markup*, der Trennung von Informationsgehalt und äußerer Form eines Dokumentes. Im Gegensatz hierzu stellt der formatorientierte Ansatz*visual markup* der *WYSIWYG*[3]-Programme, die graphische Darstellung eines Dokuments in den Vordergrund (vgl. Behme und Mintert 2000, S. 34ff.).

Generic Markup bildet die Grundlage für die Auszeichnungssprache *Standard Generalized Markup Language* (SGML) sowie den Teilmengen von SGML, der *Hypertext Markup Language* (HTML) und der *Extensible Markup Language* (XML), die in Abschnitt 4.4.1 auf Seite 45 näher beschrieben wird.

[3] „What you see is what you get“

2.2.1 Born digital

Wie bereits zu Beginn dieses Kapitels beschrieben wurde, können digitale Dokumente in ausschließlich digitaler Form, oder als digitalisierte Form von ursprünglich analogen Dokumenten vorliegen.

Die englische Sprache verwendet hierfür die bezeichnenden Begriffe *born digital document* und *digitised document*. Hedstrom (1995) beschreibt ein born digital document als ein Dokument, *„... that begin its life in digital form."*

2.2.2 Digitalisierung

Im Gegensatz zu born digital beschreibt der Begriff Digitalisierung den Vorgang, bei dem ein analog vorhandenes Dokument die sogenannte *Primärform* (zum Beispiel ein gedrucktes Buch) in eine digitale Datei die *elektronische Sekundärform* konvertiert wird. Diese Konvertierung kann durch verschiedene Verfahren erfolgen. Entweder durch das direkte *einscannen* der Vorlage oder durch die *Verfilmung* der Vorlage und das anschließende Scannen des Mikrofilms (vgl. Harloff 2001).

Nach Brugger (1998) kann das Scannen als der umgekehrte Weg der in Abbildung 2.1 auf der vorherigen Seite dargestellten Dokumenterzeugung betrachtet werden. Ein physisches Dokumentbild wird digitalisiert und als Bild dargestellt. Dieser Vorgang wird daher auch als *Imaging* oder *digital Imaging* bezeichnet (CDL 2001).

Bezüglich des Zeitpunkts der Digitalisierung kann nach Harloff (2001) zwischen der nachträglichen Digitalisierung *Retrodigitalisierung* und der Digitalisierung, die zeitgleich mit der Publikation der analogen Form erfolgt, unterschieden werden.

Für die Retrodigitalisierung können nach Harloff (2001) der selbst Sitts (2000) zitiert, hauptsächlich zwei Gründe genannt werden:

- *„Die Bestandserhaltung, also den Schutz der Primärform und die Sicherung ihres Inhalts durch Langzeitarchivierung der elektronischen Sekundärform."*
- *„Die Benutzung, also die Verfügbarmachung des Inhalts der Primärform durch Bereitstellung der elektronischen Sekundärform im Internet."*

Jedoch weist Harloff (2001) zit. n. Leskien (2000) darauf hin, daß eine Retrodigitalisierung mit dem Ziel der Langzeitarchivierung mittlerweile kritisch betrachtet werden muß, *„...da die Frage der Sicherstellung einer dauerhaften Verfügbarkeit elektronischer Dateien sich als sehr schwierig erwiesen hat."*

Da sich Mikrofilme durch eine besonders lange Haltbarkeit auszeichnen (siehe auch Tabelle 3.1 auf Seite 16) wird deshalb zur Langzeitarchivierung immer häufiger die Verfilmung der Vorlagen und anschließende Digitalisierung der Mikrofilme empfohlen. Als Vorteil erweist sich bei diesem Verfahren, daß sowohl ein analoges Dokument für die Langzeitarchivierung als auch eine elektronische Sekundärform für die Benutzung erzeugt werden (vgl. Harloff 2001).

2.3 Eigenschaften digitaler Dokumente

Neben der Unterscheidung, auf welche Weise ein digitales Dokument entstanden ist, können digitale Dokumente nach Bide (2000) grundsätzlich in folgende Kategorien eingeteilt werden:

Offline-Dokumente: Hierzu gehören alle Dokumente, die sich auf physischen Speichermedien befinden wie zum Beispiel Magnetbändern, Disketten oder CD-ROMs[4].

Hybride-Dokumente: Gemeint sind ofline-Dokumente, die Verweise zu online Materialien enthalten.

Online-Dokumente: Hier handelt es sich um Dokumente, die über das Internet oder proprietäre Netzwerke zugänglich sind.

2.3.1 Statische, kumulative und dynamische Dokumente

Online-Dokumente können zusätzlich nach der Anordnung der im Dokument enthaltenen Informationen unterschieden werden (Brugger 1998; Bide 2000). Mit der Entstehung digitaler Dokumente sind nach Brugger (1998) zur klassischen, sequentiellen Form neue Formen hinzugekommen, die als *statische* und *dynamische* Dokumente bezeichnet werden können.

So sind bei herkömmlichen, auf Papier gedruckten Textdokumenten die Informationen im wesentlichen in einer eindimensionalen Leseordnung angeordnet, wobei zwischen den einzelnen Seiten hin- und hergeblättert werden kann. Der Inhalt bleibt jedoch vom Zeitpunkt der Publikation und während der gesamten Lebensdauer des Dokuments unverändert. Hingegen kann sich der Inhalt bei dynamischen Dokumenten ständig verändern.

Bide (2000) fügt dieser Unterscheidung noch eine dritte Kategorie hinzu: *cumulative resources* die nachfolgend als *kumulative Dokumente* bezeichnet werden sollen. Folgende Beschreibung soll die Unterschiede verdeutlichen:

statische Dokumente: Neben den bereits erwähnten analogen Papierdokumenten gehört hierzu auch die digitale Form *Hypertext*.

Der Begriff Hypertext wurde 1960 von TED NELSON geprägt und basiert auf der Idee der Verknüpfung von Textstellen durch Verweise (vgl. Behme und Mintert 2000, S. 38). Nach Endres und Fellner handelt es sich bei Hypertext um einen Text, *„... aus dem heraus Verweise (engl.: links) zu anderen Texten oder Textteilen automatisch aktiviert werden können.“* (Endres und Fellner 2000, S. 172).

[4]Ebooks gehören nach Bide (2000) nicht zu dieser Kategorie, da sie online verkauft aber offline gelesen werden. Es handelt sich um eine Mischform die keiner der genannten Kategorien eindeutig zugeordnet werden kann.

Zusammen mit dem bereits in Kapitel 2.2 erwähnten Prinzip des generic coding bildet Hypertext die Grundlage des *World Wide Web* (WWW), das auf dem Prinzip der *Verlinkung* von Textteilen innerhalb eines Dokuments oder zu anderen Dokumenten beruht. Somit kann man das *surfen* im WWW auch als ein Springen von Hyperlink zu Hyperlink betrachten, bei dem sich eine gewisse Dynamik entwickelt. Der Inhalt der Dokumente dabei aber immer noch statisch bleibt.

Eine erweiterte Form der statischen Dokumente stellen die sogenannten *Multimediadokumente* oder auch *Hypermedia* dar. Sie besitzen *„... zusätzlich zur räumlichen eine zeitliche Dimension...“* (vgl. Brugger 1998) und können neben statischen Informationen auch Animationen wie Sprache oder Videosequenzen enthalten. So ist beispielsweise ein Textdokument, das Bilder mit Verweisen zu Ton- oder Videosequenzen enthält, die durch den Benutzer aktiviert werden können ebenfalls ein statisches Dokument, da sich der Inhalt während der gesamten Lebensdauer des Dokumentes nicht verändert.

kumulative Dokumente: Bei dieser Kategorie handelt es sich um Dokumente, die bestimmte gleichbleibende Elemente enthalten, denen jedoch ständig neue Inhalte hinzugefügt werden. Für Bide (2000) sind diese Dokumente das elektronische Pedant zu fortlaufend erscheinenden Druckerzeugnissen.

dynamische Dokumente: Dabei handelt es sich um Dokumente, deren Inhalt jederzeit veränderbar ist. Diese Veränderungen erfolgen in der Regel durch Benutzereingaben oder selbstständige Berechnungen, weshalb diese Form von Dokumenten auch als *interaktive Dokumente* bezeichnet wird. Beispiele hierfür sind elektronische Formulare die vom Benutzer ausgefüllt werden können.

2.3.2 Vorteile digitaler Dokumente

Neben den bereits beschriebenen, möglichen Unterscheidungen von digitalen Dokumenten, nennen Endres und Fellner (2000) außerdem eine Vielzahl von Gründen, weshalb ein digitales Dokument einem analogen Dokument überlegen ist:

Speicherkapazität: Im Gegensatz zu analogen Medien können auf digitalen Datenträgern Informationen auf wesentlich kleinerem Raum gespeichert werden. So kann beispielsweise eine DVD den Inhalt von etwa 5000 Büchern übernehmen (vgl. Endres und Fellner 2000, S. 16).

Schnelligkeit der Übertragung: Während die Beschaffung eines analogen Dokuments sehr zeitaufwendig sein kann, können digitale Dokumente in kürzester Zeit im Netz lokalisiert und heruntergeladen werden. Die Geschwindigkeit hängt dabei von den technischen Gegebenheiten ab.

Gleichzeitige Nutzung desselben Exemplars: Der Zugriff kann mit Hilfe einer entsprechenden technischen Infrastruktur durch mehrere Nutzer gleichzeitig erfolgen, da das Dokument immer vorhanden ist.

Selektive Informationsverteilung: Durch die kostengünstigere Erzeugung können Informationen mit Hilfe digitaler Dokumente in kleinerem Umfang und kürzeren Zeitabständen verteilt werden.

Weltweite Verfügbarkeit: Standortabhängigkeiten bei der Bereitstellung und der Nutzung von Dokumenten entfallen, da der Zugriff weltweit erfolgen kann.

Weiterverarbeitbarkeit: Mit Hilfe entprechender Software können digitale Dokument beliebig bearbeitet werden zum Beispiel durch Einfügen, Entfernen oder Korrigieren von Inhalten.

Erschließbarkeit: Digitale Dokumente können mit Hilfe von Metadaten erschlossen werden, die eine direkte Suche im Dokument ermöglichen. Zum Beipiel im Titel oder im Text.

Integrierte Darstellung verschiedener Medien: In digitalen Dokumenten können zur Unterstützung der Informationsvermittlung Texte und Grafiken mit bewegten Bildern, Tonaufzeichnungen kombiniert werden. Diese Möglichkeit wird bei der Produktion von *e-learning* Programmen genutzt.

2.3.3 Nachteile digitaler Dokumente

Endres und Fellner (2000) weisen jedoch auch auf die Nachteile digitaler Dokumente hin, von denen einige nachfolgen kurz beschreiben werden:

Abhängigkeit von technischen Hilfsmitteln: Um digitale Dokumente nutzen zu können wird die entsprechende Hard- und Software benötigt. Da diese Hilfsmittel einer ständigen technischen Weiterentwicklung unterliegen, ist deren Aktualisierung mit Kosten verbunden.

Leichte Veränderbarkeit: Durch die bereits genannte Weiterverarbeitbarkeit besteht die Gefahr, daß veränderte Dokumente oder Plagiate in Umlauf gebracht werden, deren Authentizität nicht mehr festzustellen ist.

Gefahr von Beschädigung und Verlust: Die Daten digitaler Dokumente können durch Beschädigung oder Verlust des Datenträgers nicht mehr gelesen werden (siehe auch 3.1 auf Seite 15).

Risiken bei der Übertragung über offene Netze: Im Gegensatz zu analogen Dokumenten besteht bei der Übertragung digitaler Dokumente über offene Netze ein erhöhtes Risiko der Einsichtnahme durch Dritte. Dies kann zum Beispiel durch

Verschlüsselungsverfahren verhindert werden, ist aber mit erheblichem Aufwand verbunden.

Aufwand für Langfrist-Archivierung: Für die Langfrist-Archivierung müssen erhebliche technische und organisatorische Maßnahmen getroffen werden, deren Inhalt Gegenstand dieser Arbeit ist.

2.4 Dokumenttypen

Leider wird bei der Verwendung des Begriffs *Dokumenttyp* in der Literatur nicht immer deutlich, was unter einem Dokumenttyp zu verstehen ist. So bieten beispielsweise (vgl. Endres und Fellner 2000, S. 109) die folgende, sehr allgemeine Definition an:

> „Unter *Dokumenttyp* verstehen wir eine Klasse oder Kategorie von Dokumenten mit gleichen Eigenschaften.“ (vgl. Endres und Fellner 2000, S. 109)

Auch die nachfolgende Erläuterung trägt mehr zu Verwirrung als zu Klarheit bei, wenn Endres und Fellner (2000) schreiben: *„Bei konventionellen Bibliotheken dominieren zwei Dokumenttypen, Bücher und Zeitschriften. Der im Wort Bibliothek vorkommende Begriff Buch (griechisch: biblos) ist nur einer von vielen Dokumenttypen, dazu noch ein Dokumenttyp mit vielen Einschränkungen. Als logische Einheit kann ein digitales Dokument unter Umständen aus mehreren Objekten bestehen. Was im Einzelfalle darunter zu verstehen ist, hängt vom Dokumenttyp ab.“* (vgl. Endres und Fellner 2000, S. 109)

Eine wesentlich eindeutigere Beschreibung ist bei Mintert (1999) zu finden, der sich mit der Anwendung von XML beschäftigt. Aus seiner Sicht läßt sich der Dokumenttyp wie folgt definieren:

> „Der Dokumenttyp beschreibt eine Klasse von Dokumenten, die sich in ihrem strukturellen Aufbau gleichen.“ Mintert (1999)

Die Bedeutung des strukturellen Aufbaus, als Merkmal eines Dokumenttyps wird schnell anhand der Dokumenttypen Brief und Buch deutlich. Nach Mintert (1999) können dem Dokumenttyp Brief alle Dokumente zugeordnet werden, die eine Absenderadresse, eine Empfängeradresse, eine Anrede, einen Hauptteil und eine Grußformel enthalten. Ebenso enthält der Dokumenttyp Buch verschiedene strukturelle Merkmale, wie Titel, Autor, Verzeichnisse, Kapitel und so weiter. Dieses Verständnis des Begriffs Dokumenttyp soll auch für die vorliegende Arbeit gelten.

Entsprechend den Möglichkeiten der digitalen Technologie sind bereits neue, digitale Dokumenttypen entstanden, die parallel zu den traditionellen Dokumenttypen existieren.

2.5 Digitale Dokumenttypen

Im Dezember 2000 verabschiedetet der U.S. amerikanische Kongreß unter dem Titel *„Preserving our digital heritage“* die Ausarbeitung eines *National Digital Information Infrastructure and Preservation Program* (NDIIPP) zur systematischen Archivierung elektronischer Ressourcen. Unter der Leitung der *Library of Congress* (LoC) wurde mit einem Budget von 5 Millionen US-Dollar ein nationales Programm mit Maßnahmen zur Sammlung und Erhaltung elektronischer Ressourcen geplant.[5] Beispielhaft für elektronische Ressourcen wurden die nachfolgend aufgeführten, digitalen Dokumenttypen definiert und unter dem Aspekt der Langzeitarchivierung untersucht (vgl. NDIIPP 2002):

- large Web sites
- electronic books (ebooks)
- electronic journals
- digitally recorded sound
- digital film
- digital television

Das NDIIPP (2002) verweist jedoch auch darauf, daß es sich bei den unter Abschnitt 2.4 auf der vorherigen Seite genannten Dokumenttypen keinesfalls um die einzigen neuen digitalen Typen von Dokumenten handelt. Es werden Dokumenttypen, mit noch höheren technischen Funktionalitäten und anwenderspezifischen Eigenschaften entstehen, die traditionelle Dokumenttypen weitgehend verdrängen.

Führend werden dabei aus Sicht des NDIIPP elektronische Resourcen sein, die Dokumente auf Abruf als Ergebnis einer Datenbankabfrage produzieren. Ein erstes, bereits exisitierendes Beispiel hierfür ist das *Geographic Information System* (GIS). Das NDIIPP (2002) geht davon aus, daß dieser digitale Dokumenttyp die Produktion gedruckter Landkarten ersetzen wird. Ein weiteres Beispiel für diese Entwicklung sieht das NDIIPP (2002) in der Verbreitung des *Global Positioning Systems* (GPS). Vermutlich wird dieses System gedruckte Straßenkarten, Stadtpläne aber auch Karten der Volkszählungen, die demographische Angaben enthalten, ersetzen.

Diese Beispiele der *digitalen Kartographie* machen deutlich, daß die Entstehung neuer Dokumenttypen Bibliotheken und Archive mit stets neuen Fragestellungen konfrontiert. Bereits heute werden bei der LoC Überlegungen bezüglich der kartographischen

[5] Im Februar 2003 wurde das NDIIPP vom U.S. amerikanischen Kongress gebilligt und weitere 20 Millionen US-Dollar Dollar dafür freigegeben. Insgesamt ist für dieses Programm ein Budget von 100 Millionen US-Dollar vorgesehen. Die restlichen 75 Millionen US-Dollar soll die LoC durch private Spenden aufbringen vgl. Roetzer (2003).

Sammlungen angestellt. Was geschieht mit den existierenden Sammlungen? Wie sieht die Sammlung kartographischer Materialien in Zukunft aus? Werden statt gedruckten Karten umfangreiche Datensätze und entsprechende Suchabfragen gesammelt und archiviert? Welche technische Infrastruktur ist hierfür notwendig? Wie erfolgt der Zugang und die Benutzung solcher Materialien? (vgl. NDIIPP 2002)

Ein weiteres Beispiel für die Entstehung neuer, digitaler Dokumenttypen sowie die Schwierigkeiten ihrer Archivierung stellt der Nachlaß des 1995 verstorbenen Schriftstellers Thomas Strittmeier dar, den das Deutsche Literaturarchiv in Marbach im Jahr 2000 erworben hat.

Da Strittmeier vermutlich zu den ersten deutschen Autoren gehörte, die ausschließlich am Computer schrieben, handelt sich bei seinem Nachlaß neben Zeitungsausschnitten und handschriftlich korrigierten Computerausdrucken seiner Texte vor allem um einen Computer vom Typ Atari SM 124 aus den achtziger Jahren und zwei bis drei Dutzend Disketten mit diversen Dateien (vgl. Bernard 2003, S. 16).

Vor allem die Fragen der Darstellbarkeit und der Lesbarkeit, als Voraussetzungen für eine literaturwissenschaftliche Erschließung erhalten bei einem digitalen Nachlass eine völlig neue Bedeutung. Eine Interpretation der Texte ist ohne technische Hilfsmittel nicht mehr möglich. Bernard (2003) schreibt dazu:

> In konventionellen Nachlassbearbeitungen fielen Entzifferung und Forschung immer zusammen; von Nietzsches Konvoluten bis zu Robert Walsers Mikrogrammen war der betreuende Philologe im Zweifel der einzige, der die Handschrift zu lesen vermochte. Künftig wird sich diese Konstellation ändern. EDV-Spezialisten und Geisteswissenwschaftler müssen sich die Lektüre gewissermaßen teilen; der Interpretation der literarischen geht die der Computer-Sprache voraus.“ (Bernard 2003, S. 16)

Erschwerend kommt außerdem hinzu, daß eine exakte Wiedergabe des Originaltextes im Fall eines digitalen Nachlasses niemals garantiert werden kann. Die Darstellung und Lesbarkeit von Dokumenten, die mit älterer Hard- und Software erstellt wurden, gelingt meist nur mit hohem technischen Aufwand. Der Preis dafür ist in den meisten Fällen der Verlust der originalen Darstellung wie beispielsweise Markierungen von Textstellen oder spezielle Schriftarten- und größen. Aus diesem Grund sieht (Bernard 2003, S. 16) hier bereits eine neue Stufe der Archivierung literarischer Dokumente: *„... von der Reliquie des Manuskripts über das profanere Typoskript hin zu einer Art unsichtbarem „Kryptoskript“ der veralteten Diskette oder Festplatte.“* Daß auch an diesem Punkt Einigungen zwischen Technikern und Literaturwissenschaftlern nötig sein werden, macht ebenfalls der Fall Strittmeier deutlich. Für die EDV-Spezialisten waren die Disketten nach erfolgreicher Sicherung nur noch unnötiger Abfall. Schon fast auf dem Weg zur Entsorgung wurden sich die Literaturwissenschaftler *„... der Bedeutung dieses letzten Rests an Materialität bewußt.“* (Bernard 2003, S. 16). Nun bilden den einzig materiellen Teil des Nachlasses, der auch ausgestellt werden kann, das Computergehäuse und die Disketten.

Das Literaturarchiv in Marbach nimmt dieses Exempel zum Anlaß, Richtlinien für den Umgang mit Computernachlässen zu erarbeiten. Ebenso wurden Regelungen für die wissenschaftliche Nutzung von Strittmeiers Nachlaß getroffen. So können kürzere Texte ausschließlich in gedruckter Version gelesen werden. Für längere Manuskripte sind PCs mit besonderen Schutzvorrichtungen vorgesehen (ohne Diskettenlaufwerk und Internetzugang) (vgl. Bernard 2003, S. 16).

Bei der Erschließung zukünftiger Computernachlässe werden Archive zusätzlich mit der *„... vermutlich schwierigsten Frage der Editionsphilologie im 21. Jahrhundert...“* (Bernard 2003, S. 16) konfrontiert: Wie soll auf das Verschwinden von Briefausgaben reagiert werden? Strittmeier hat keine e-mails geschrieben. Trotzdem ist zu diskutieren, wie gegenwärtige Dichterkorrespondenzen gesichert werden können. In Marbach hatte der Chef der EDV-Abteilung bereits einen Vorschlag zu Lösung dieses Problems: *„... man sollte einer Anzahl von Autoren einen Account in Marbach einrichten. Sämtliche E-Mails gingen dann als Blindkopie direkt ans Archiv.“* (Bernard 2003, S. 16).

Auch das NDIIPP (2002) sieht bisher nicht allzu viele Möglichkeiten, neu entstehende Dokumenttypen zu sichern: *„ All we can do is acquire some of these new formats and track closely what their custodial needs are and how users interact with them“* NDIIPP (2002). Aus diesem Grund lautet die Empfehlung des NDIIPP, in Zukunft den Schwerpunkt nicht nur auf die Beobachtung technischer Entwicklungen[6] zu legen, *„... but also a format and genre watch and a careful watch of users“* NDIIPP (2002).

[6] engl.: *technology watch*

3 Gegenwärtige Situation

Die Beispiele des vorhergehenden Kapitels haben bereits deutlich gemacht, daß die Aufgabe, digitale Dokumente nicht nur zu archivieren, sondern auch ihre Nutzung in der Zukunft zu gewährleisten Bibliotheken und Archive vor vielfältige Herausforderungen stellt. Während für die Erhaltung analoger Dokumente Strategien und Methoden entwickelt wurden, die bereits erfolgreich Anwendung finden, wie beispielsweise die Möglichkeit, säurehaltigem Papier nachträglich die Säure zu entziehen und es somit vor dem Zerfall zu bewahren, gestaltet sich die Erhaltung von digitalen Dokumenten weitaus schwieriger. Die zu bewältigenden Probleme sind komplex und in der Regel zu neu, um sie mit bisher gemachten Erfahrungen lösen zu können.

Unter dem Titel *„Taking a Byte out of History: The Archival Preservation of Federal Computer Records“* veröffentlichte das U. S. House of Representatives Committee on Government Operations bereits 1990 einen Bericht, der zahlreiche Beispiele für Verluste digitaler Informationen enthält (vgl. Rothenberg 2001). Als Hauptursachen für die Datenverluste werden die begrenzte physikalische Lebensdauer der Trägermedien sowie die schnelle Veralterung von Hard- und Software genannt.

Neben diesen, ausschließlich technisch bedingten Ursachen, exisitiert auch eine Vielzahl nichttechnischer Ursachen, die zu Datenverlusten führen können. Zusammenfassend können deshalb als Gründe für Datenverluste genannt werden:

- Lebensdauer von Speichermedien
- Technologischer Wandel
- Vielzahl von Formaten und Standards
- Fehlbedienung und Probleme der Organisation

3.1 Lebensdauer der Speichermedien

Jahrzehntelang wurde die Lebensdauer digitaler Speichermedien, die als *„beliebig oft kopierbar und fast endlos haltbar“* (Schmundt 2000) galten, überschätzt. Wie es tatsächlich um die Haltbarkeit digitaler Speichermedien bestellt war, wurde häufig erst erkannt, wenn die oftmals seit Jahren gespeicherten Daten benötigt oder geprüft wurden. Dann nämlich stellte sich heraus, daß die als dauerhaft archiviert geglaubten Informationen durch beschädigte oder zerstörte Speichermedien nicht mehr zugänglich waren.

Die meisten Erfahrungen auf diesem Gebiet liegen mit Magnetbändern vor. Sie wurden bis zur Einführung von optischen Speichermedien zur Archivierung verwendet. Magnetbänder bestehen aus einem sogenannten Stützfilm, der mit einer Magnetschicht, die magnetische Signale aufnehmen kann, beschichtet ist. Diese Magnetschicht wiederum besteht aus magnetischen Partikeln und einem Bindemittel, das die Partikel zusammenhält aber auch die Aufgabe hat, eine glatte Oberfläche und somit einen reibungslosen Transport des Bandes durch das Laufwerk zu gewährleisten (vgl. Betts und Schmidt 1999)

Sowohl der Stützfilm als auch die magnetischen Partikel und das Bindemittel sind mögliche Quellen für Zerfall und Zerstörung. Erschwerend kommt nach Betts und Schmidt (1999) hinzu, daß neben unterschiedlichen *Bänderformaten* wie zum Beispiel U-matic, VHS, S-VHS, acht Millimeterfilme und Betacams auch unterschiedliche *Bänderarten* existieren, die sich wiederum durch das Material der Beschichtung (Eisenoxid oder Chromoxid) sowie die Art der Beschichtung (aufgedampft oder pulverbeschichtet) unterscheiden.

Entscheidend für die Langlebigkeit von Magnetbändern ist jedoch nicht nur die Qualität des Materials. Auch die Pflege und die Aufbewahrung in klimakontrollierten Räumen spielen eine große Rolle bei der Erhaltung der auf den Magnetbändern gespeicherten Daten. So kann nach Rothenberg (2001) entsprechend den Lagerbedingungen die Lebensdauer von Magnetbändern zwischen 2 und 30 Jahren variieren.

Dem optischen Speichermedium Compact Disk (CD) hingegen attestiert Rothenberg (2001) eine Lebensdauer von 5 bis 59 Jahren. Die nachfolgende Tabelle 3.1 übernommen aus: Grote (2000) zit. n. McLean und Davis (1999) gibt einen Überblick über die Lebensdauer unterschiedlicher Speichermedien[1].

Medium	**Jahre**
CD-ROM	5 bis 200
Zeitungspapier	10 bis 20
VHS-Band	10 bis 30
Digitalband	10 bis 30
Magnetband	10 bis 30
Mikrofilm	10 bis 500
Kodakcrome Dias	100
Säurefreies Papier	100 bis 500
HD-Rosetta	1000 plus
Ägypt. Steinschrift	2200 plus

Tabelle 3.1: Lebensdauer von Datenträgern

[1]Das Speichermedium HD-Rosetta wird in Abschnitt 3.6.1 auf Seite 20 eingehend beschrieben.

Wenn auch Voraussagen bezüglich der physikalischen Lebensdauer von Speichermedien in Jahren kritisch zu betrachten sind, so wird doch hinsichtlich der Archivierung und Erhaltung das Problem deutlich: digitale Speichermedien haben begrenzte Lebenszeiten, die größtenteils noch nicht bekannt sind. Betts und Schmidt (1999) formulieren diese Tatsache so:

> „Wir werden es nie wirklich wissen, wie lange die heutigen Speichermedien Daten zuverlässig behalten können, bevor nicht ein oder zwei Jahrzehnte verstrichen sind. Und wir werden nicht sehen, ob die Daten verdorben sind oder lückenhaft, bis wir versuchen, sie zu lesen."

Auf der Grundlage vorhandener Erfahrungen, konnten für die Aufbewahrung magnetischer Speicher bereits etliche Standards geschaffen werden. Dies ist bei den dazu vergleichsweise immer noch neuen, digitalen Datenträgern nicht der Fall. Tabelle 3.2 übernommen aus: Gschwind u. a. (2000) zeigt die bereits vorhandenen Standards[2] für digitale Datenträger. Es handelt sich dabei sowohl um Standards für die Aufbewahrung, als auch um Richtlinien zur Durchführung von Untersuchungen bezüglich der Lebensdauer, die vom *American National Standards Institute* (ANSI) entwickelt wurden[3].

Nummer	**Medium**
IT9.23-1998	Storage of Polyester Based Magnetic Tape
IT9.25-1998	Storage of Optical Disk Media
IT.21-1996 (ISO 15525)	Methods for Estimating Effects of Temperature and Relative Humidity on Life Expectancy on Compact Disks (CD-ROM)
IT9.26-1997	Methods for Estimating Effects of Temperature and Relative Humidity on Life Expectancy on Magneto-Optic (MO) Disks
IT9.27 (in Bearbeitung)	Methods for Estimating Effects of Temperature and Relative Humidity on Life Expectancy on Recordable Compact Disks (CD-R)

Tabelle 3.2: ANSI Standards

Unter den digitalen Speichermedien hat sich in den letzten Jahren vor allem die CD in ihren verschiedenen Formen[4] zur Archivierung von Daten durchgesetzt. Daher beschäftigen sich auch zahlreiche Untersuchungen mit der Lebensdauer dieses Datenträgers. Gschwind u. a. (2000) verweisen jedoch darauf, daß die Durchführung der

[2] Im Jahr 2000
[3] Für ausführliche Beschreibungen der Standards vgl. Gschwind u. a. (2000).
[4] CD-ROM, CD-R, CD-RW

Tests für die Lebenserwartung von CDs sehr komplex sind. Zudem werden bei den bisher standardisierten Tests nur die Faktoren Temperatur und relative Luftfeuchtigkeit getestet.

Für den Kauf von CDs empfehlen Gschwind u. a. (2000) deshalb, den Hersteller sorgfältig auszuwählen und dabei Testergebnisse aus dem Internet[5] oder entsprechenden Zeitschriften zu berücksichtigen[6].

3.2 Technologischer Wandel

Die Verfügbarkeit und Lesbarkeit von digital gepeicherten Daten hängt jedoch nicht nur von der physikalischen Lebensdauer der Speichermedien ab. Eine weitere, wesentliche Ursache für Datenverluste oder den Verlust des Zugriffs auf die gespeicherten Daten stellt die schnelle Veralterung der Hard- und Software dar, die für die Interpretation und die Darstellung der digital vorhandenen Informationen benötigt werden.

Als prominentes Beispiel für einen durch technologischen Fortschritt bedingten Datenverlust gilt die US-Volkszählung aus dem Jahr 1960. Bereits 16 Jahre nach der Zählung war das zum Lesen der Daten benötigte Univac Typ II-A-Bandlaufwerk ein Museumsstück (vgl. Betts und Schmidt 1999). Diese Tatsache führte zu der paradoxen Situation, daß die auf Papier vorhandenen Daten der Volkszählung von 1860 noch immer vollständig verfügbar waren, während auf die 1960 digital gespeicherten Daten nicht mehr zugegriffen werden konnte (vgl. Betts und Schmidt 1999; Kodak 2001). Nur durch *„wesentliche Rettungsaktionen“* konnten Daten in ein neues Speicherformat überführt und somit wieder zugänglich gemacht werden (vgl. Betts und Schmidt 1999; Rathje 2002).

Schmundt (2000) schildert den Fall der Penn State University in den USA. Dort mußte festgestellt werden, daß die Daten von 3000 Studenten nicht mehr lesbar sind, was zur Folge haben kann, daß bereits erlangte Qualifikationen nicht mehr nachgewiesen werden können. *„So wäre eine Examensarbeit, geschrieben 1990 in Wordstar und gespeichert auf einer 5 1/4 Zoll Diskette heute kaum noch rekonstruierbar - es sei denn, es gäbe noch einen Ausdruck auf Papier.“* (Schmundt 2000)

An ein Wunder grenzt laut Betts und Schmidt *„...nur schon einen PC zu finden, der ein Wordperfect-4.0 File auf einer 5,25-Diskette lesen kann und weder Fußnoten noch Formatierung verliert.“* (Betts und Schmidt 1999)

Für weitere Datenverluste in großen Mengen wird nun vermutlich auch das Verschwinden der 3,5-Zoll-Diskette sorgen, die, nachdem sie 1981 von Sony auf den Markt gebracht wurde, 20 Jahre lang weltweit als das beliebteste Speichermedium galt (jth 2003). Nach Computerhersteller Apple verzichtet nun auch der Hersteller Dell bei seinen Computern auf Diskettenlaufwerke. Bei der Begründung für diesen Schritt beruft sich auf das Verhalten der Anwender: *„Wenn Sie die Leute fragen, ob sie ein Laufwerk*

[5] htp://www.cd-info.com
[6] vgl. dazu auch test (2003)

brauchen, sagen alle ja. Wenn Sie fragen, wann sie das letzte Mal eine Diskette benutzt haben, liegt das länger als ein Jahr zurück“ (vgl. jth 2003).

3.3 Vielzahl von Formaten und Standards

Ein weiteres, großes Problem für die Langzeitarchivierung stellt die Vielzahl der bei der Erstellung und Speicherung digitaler Dokumente verwendeten Datenformate und Dokumentenstandards dar.

1998 führten Margaret Hedstrom und Sheon Montgomery im Auftrag der *Research Libraries Group* (RLG) eine Umfrage bei den Mitgliedseinrichtungen der RLG durch, die Aufschluß über die Archivierungspraxis digitaler Dokumente in den einzelnen Einrichtungen geben sollte. Teilnehmer dieser Umfrage, deren Ergebnisse unter dem Titel *Digital Preservation Needs and Requirements in RLG Member Institutions* 1999 von der RLG veröffentlicht wurden, waren insgesamt 54 internationale Archivierungseinrichtungen darunter Museen, Universitätsbibliotheken und Nationalbibliotheken.

Ein Teil der Umfrage befaßte sich mit den in den Einrichtungen vorhandenen Datenformaten. Gemeint waren sowohl die Datenformate, in denen digitale Dokumente erworben wurden als auch Speicherformate, die von den Einrichtungen für die Archivierung digitalisierter Dokumente verwendet werden. Die Auswertung dieses Teils der Umfrage ergab folgendes Ergebnis:

> „Twenty-four different file formats were reported by the respondents through a combination of checking off formats listed in the survey instrument and listing additional formats that had not been identified“ (Hedstrom und Montgomery 1999).

Diese Vielzahl von Formaten erschwert die Archivierung, die Zusammenarbeit und den Austausch von Daten erheblich.

3.4 Fehlbedienung

Gschwind u. a. (2000) weisen auf eine weitere, meist unbeachtete aber nicht unbedeutende Ursache für Datenverluste hin: die Fehlbedienung. Fehlbedienungen können sowohl im Umgang mit Daten auf einem Rechner als auch in der falschen Handhabung von Datenträgern erfolgen.

Hierzu gehören beispielsweise das versehentliche Löschen von Daten, mangelnde Kontrolle von bereits geschriebenen Daten aber auch Verschmutzung von Datenträgern durch Fingerabdrücke, Staub oder Klebemarken (vgl. Gschwind u. a. 2000).

3.5 Organisatorische Probleme

Das Datenverluste auch nichttechnische, sondern organisatorische Ursachen haben können, beweist der sogenannte *„NASA-Effekt“* (vgl. Rathje 2002). Hierbei handelt es sich um Datenverluste bei der amerikanischen Weltraumbehörde NASA[7] Bereits Mitte der 1990er Jahre waren *„... mehr als 1,2 Millionen Magnetbänder mit Daten aus 30 Jahren Raumfahrt nicht mehr benutzbar, teilweise wegen mangelnder Zuordnung zu den jeweiligen Weltraummissionen und Projekten[...]die Bänder waren nicht oder nur notdürftig beschriftet“ (Rathje 2002)*[8]. Derartige Datenverluste beruhen nach Rathje (2002) auf organisatorische Defiziten wie beispielsweise *„... der Nichteinhaltung einfacher Archivierungsgrundsätze“.*

3.6 Neue Speichertechnologien

Zur Lösung des Problems der dauerhaften Speicherung von Informationen wird an der Entwicklung neuer Speichermedien gearbeitet. Wenn auch über die Lebensdauer oftmals nur vage oder gar keine Aussagen getroffen werden können, ist zumindest eine deutliche Steigerung der Speicherkapazität festzustellen. Nachfolgend werden einige der wichtigsten Entwicklungen neuer Speichertechnologien vorgestellt.

3.6.1 HD-Rosetta

Das analoge Speichermedium HD-Rosetta (*High Density Rosetta*) wurde in Zusammenarbeit von den *Los Alamos Laboratories*[9] und *Norsam Technologies*[10] entwickelt. Seinen Namen erhielt das Speichermedium in Anlehnung an den 1799 in Ägypten entdeckten Rosetta Stein, dessen Inschriften in verschiedenen Sprachen die Entschlüsselung rätselhafter Hieroglyphen ermöglichten (vgl. Landwehr 2001).

Es handelt sich bei HD-Rosetta um eine Nickelplatte, mit einer Seitenlänge von 2 Zoll (= 5,08 cm) und einer Dicke von 1/4 Zoll (=0,63 cm)[11].

Die Informationen werden in digitaler Form auf die Größe von Mikronen verkleinert und mit einer *Focused Ion Beam* (FIB) Maschine auf die Nickelplatte graviert. Das Lesen der Informationen erfolgt mit einem Elektronen- oder Lichtmikroskop. Dabei hängt die Speicherkapazität vom Typ des Mikroskops ab. Bei Verwendung eines Elektronenmikroskops können ca 196.000 DIN A4 Seiten auf einer HD-Rosetta gespeichert werden, während es bei Verwendung eines Lichtmikroskpos 5000 bis 18000 Seiten sind (vgl. Norsam 2001).

[7] NASA = National Aeronautics and Space Administration.
[8] Der selbst Schmundt (2000) und Archimedes (1999) zitiert
[9] http://www.lanl.gov
[10] http://www.norsam.com/hdrosetta.html
[11] 1 Zoll = 2,54 cm

Ein eigens von Norsam Technologies entwickeltes Lesegerät arbeitet zusätzlich mit einem integrierten Koordinatensystem. Somit kann der Benutzer eine Seitenzahl eingeben, die anhand ihrer Koordinaten mit Hilfe spezieller Software gefunden und auf einem Bildschirm zur Ansicht oder zum Ausdruck angezeigt werden kann. Zu berücksichtigen sind die Kosten: alleine das Lesegerät kostet bereits 10 000 Dollar (vgl. Schmundt 2000).

Praktische Anwendung findet HD-Rosetta bereits in einem Projekt, das von der *Long Now Foundation* initiiert wurde[12]. Bei diesem Projekt wird davon ausgegangen, daß innerhalb des nächsten Jahrhunderts 50 bis 90 Prozent aller Sprachen verschwinden werden. Die meisten dieser Sprachen sind nur wenig oder gar nicht dokumentiert. Ziel ist es, ein Archiv mit 1000 der ungefähr 7000 auf der Welt vorhandenen Sprachen und den dazugehörigen linguistischen Informationen zu schaffen.

Dieses Archiv soll gemäß dem *Open Source* Gedanken als frei zugängliche Informationsresource für Forschungen auf dem Gebiet der Linguistik weltweit zur Verfügung stehen. Aus diesem Grund wird als weiteres Ziel des Projekts die Schaffung eines *Linux of Linguistics* (vgl. Rosetta 2003) genannt.

Ein weiteres Beispiel, für die Verwendung von HD-Rosetta ist die Speicherung der Handschriften von Abraham Lincoln. Dieses Projekt wurde in Zusammenarbeit mit der *Library of Congress* (LOC) durchgeführt (vgl. Norsam 2001). Schmundt (2000) verweist außerdem auf den Einsatz von HD-Rosetta bei der *New York Times*.

Die Lebensdauer von HD-Rosetta wird von Norsam (2001) mit *„... at least 1,000 years“* (Norsam 2001) beziffert. Als Gründe hierfür werden von Norsam die analoge Form des Datenträgers und die Haltbarkeit des Materials genannt.

3.6.2 Iridium-CD

Diese Lösung wurde nach Lupprian (2002a) bereits vor etlichen Jahren vorgeschlagen und ähnelt dem Verfahren von HD-Rosetta. Digitale Daten werden als analogisierter Bitstrom (eine Kette von Nullen und Einsen) auf einen Iridium[13]-Träger geätzt.

Zusätzlich werden die für die Dekodierung des Bitstroms notwendigen Informationen in extrem verkleinerter Form als Klartext auf den Träger gebracht. Das Lesen der Informationen erfolgt mit einem Elektronenmikroskop. Mit Hilfe der beigefügten Informationen sollte ein Programmierer einen sogenannten *Viewer* schreiben können, mit dem die Daten auf einem Ausgabemedium gelesen werden können (vgl. Lupprian 2002a).

[12] http://www.rosettaproject.org

[13] Bei Iridium handelt es sich um ein Platinmetall, daß sehr korrosionsbeständig ist und zudem eine hohe chemische Widerstandsfähigkeit aufweist (vgl. Meyers 2001, S. 260).

3.6.3 Holographischer Speicher

Dieses Verfahren wurde von dem Mineralogen THEO WOIKE am Institut für Kristallographie der Universität Köln[14] entwickelt. Als Speichermedien dienen kleine Kristalle, die extra für diesen Zweck gezüchtet werden. Die Kristalle sind etwa fünf Zentimeter groß und drei Millimeter dick.

Landwehr (2001) verweist auf die Möglichkeit, auf diesen Kristallen Datenmengen in der Größenordnung von bis zu 100 Peta Byte[15] zu speichern. Schnelle Zugriffszeiten, die mit denen von Halbleiterspeichern vergleichbar sind (vgl. Meyers 2001, S. 57), ermöglichen das Abspeichern und Auslesen in kurzer Zeit. So könnten beispielsweise mehrere digitalisierte Spielfilme so schnell ausgelesen werden, *„... daß der Film ruckelfrei angesehen werden kann“* (Lupprian 2002a). Der Prototyp eines speziellen Lesegeräts hierfür existiert bereits.

Ermöglicht wird dies durch ein spezielles Speicherverfahren. An einem Speicherort des Kristalls können mit Hilfe eines Lasers bis zu 10 000 verschiedene holographische Bilder gespeichert werden. Wird der Kristall nur geringfügig gedreht, entsteht ein neuer Einfallswinkel für den Laser und somit neuer Speicherplatz, der beschrieben werden kann. Diese Verfahren bietet nach Meyers (2001) außerdem den Vorteil, daß Verschmutzungen oder lokale Fehler nicht zu Informationsverlusten führen. Die so gespeicherten Daten sollen bis zu 100 Jahre haltbar sein (vgl. Landwehr 2001; Lupprian 2002a).

Eine kostengünstigere und in großen Mengen herstellbare Alternative zu dieser holographischen Methode bietet die sogenannte Mikroholographie. Anstelle von Kristallen werden hierbei als Speichermedien Photopolymere verwendet. Auf dieser Grundlage können sogenannte *holographische Disks* geschaffen werden, die sich äußerlich von den bisher bekannten CDs und DVDs nicht unterscheiden werden und mit den entsprechenden Abspielgeräten kompatibel sind. Nach Zaun (2001) wird die erste CD dieser Art ein Speichervolumen von 150 GB[16] besitzen und ab 2006 auf dem Markt erhältlich sein[17].

3.6.4 Millipede

Das Speichermedium *Millipede* wurde an einem IBM Forschungslabor in Kalifornien entwickelt. Landwehr (2001) beschreibt diese rein mechanische Technologie, die auf dem Prinzip der Rastertunnel-Mikroskopie aufbaut, als ein Verfahren, bei der mit einem unendlich feinen Instrument Atome verschoben werden. Dieser Vorgang ist vergleichbar mit dem Abtasten einer Vinylplatte durch eine feine Nadel. Da dies relativ

[14] http://linux23.kri.uni-koeln.de/grwoike

[15] 100 Peta Byte = 1000 Terabyte = 1000 Millionen Megabyte

[16] Dies entspricht etwa 200 handelsüblichen CDs mit einer durchschnittlichen Speicherkapazität von 650 MB (vgl. Zaun 2001).

[17] Zur Lebensdauer dieses Speichermediums konnten keine Angaben gefunden werden.

lange dauert, werden tausende solcher abgetasteten Einheiten zusammengenommen und in einen Chip eingebaut. Parallel werden die Daten auf einen Speicherchip, den Millipede, übertragen. Die Lebensdauer von Millipede ist unklar, jedoch können mit diesem Verfahren Speicher mit fast unendlicher Kapazität hergestellt werden (vgl. Landwehr 2001).

4 Organisatorische Konzepte

Wie bereits in vorhergehenden Kapiteln erwähnt wurde, sind es nicht nur technische Probleme, die bei der Langzeitarchivierung elektronischer Ressourcen zu lösen sind. Einen zentralen Punkt stellt in diesem Zusammenhang die Erschließung digitaler Dokumente, insbesondere die Erschließung von Internet-Ressourcen dar. Dieses Kapitel beschäftigt sich mit Möglichkeiten zur Lösung dieser Problematik.

Dazu werden zunächst Grundlagen der Erschließung erläutert. Anschließend wird auf die Bedeutung von Metadaten für die Erschließung elektronischer Ressourcen eingegangen. Beispielhaft werden zwei Metadatenstandards für die Erschließung von elektronischen Ressourcen vorgestellt. Neben den Möglichkeiten zur Gewährleistung der Authentizität von digitalen Dokumenten wird in diesem Kapitel außerdem die Bedeutung von Datenformaten und Auszeichnungssprachen für die Archivierung digitaler Dokumente untersucht. Abschließend wird das OAIS Modell als Beispiel für ein Organisationskonzept zur Archivierung elektronischer Ressourcen vorgestellt.

4.1 Erschließung digitaler Dokumente

4.1.1 Grundlagen der Erschließung

Unter *Erschließung* versteht man die Beschreibung einer Ressource, mit dem Zweck, sie für den Benutzer zugänglich zu machen. Da dies in Form von Nachweisen in Katalogen geschieht, kann die Erschließung auch als Katalogisierung bezeichnet werden. Das Recherchieren und Auffinden in erschlossenen Beständen wird als *Information Retrieval* bezeichnet (vgl. Engster 2002).

Die Erschließung von Informationsressourcen gilt als klassischer Aufgabenbereich von Bibliotheken. Hierzu gehört sowohl die Beschreibung der formalen Gegebenheiten einer Ressource, die *Formalerschließung* (vgl. Payer 1999) als auch die Beschreibung des Inhalts, die *Inhaltserschließung*.

Grundsätzlich erfolgt die Erschließung indem bestimmte Elemente der Vorlage erfaßt und teilweise nomiert werden. Das geschieht nach den Vorschriften der jeweiligen Regelwerke und zusätzlich durch Nutzung von Systematiken und Thesauri. Für die maschinelle Erfassung in bibliographischen Datenbanken benötigt man zusätzlich Regeln zur Belegung der Felder, die sogenannten Formate.

Beispiele für formale Regelwerke sind die *Anglo-American cataloguing rules 2. ed., Revision 1998* (AACR2R) auf internationaler Ebene oder die *Regeln für die alpha-*

betische Katalogisierung in wissenschafltichen Bibliotheken(RAK-WB) auf nationaler Ebene (vgl. Payer 1999). Für die inhaltliche Erschließung können die *Regeln für den Schlagwortkatalog* (RSWK) als Beispiel für ein Regelwerk genannt werden.

Da es sich bei der Erschließung um einen sehr komplexen Vorgang handelt, erfolgt die Durchführung in konventionellen Bibliotheken im wesentlichen intellektuell und, wenn keine Fremdaten übernommen werden können, mit hohem Zeitaufwand.

Engster (2002) verweist zwar auf die Möglichkeiten der maschinellen Erschließung die teilweise bereits in digitalen Bibliotheken genutzt werden, macht aber auch auf die noch bestehenden Probleme beispielsweise bei der maschinellen Erschließung, und dem Retrieval von Bildern, Video- oder Tonsequenzen aufmerksam.

4.1.2 Die Problematik der Erschließung von digitalen Dokumenten

Exkurs: Das World Wide Web in Zahlen[1]

- Das WWW ist das größte Dokument, das jemals verfaßt wurde.
- Das WWW enthält über 4 Milliarden[2] Seiten, die öffentlich zugänglich sind.
- Im sogenannten *„deep Web“*[3] existieren zusätzlich ca. 550 Milliarden Dokumente, die miteinander verknüpft sind.
- Obwohl das WWW ist in 220 Sprachen von Autoren aus der ganzen Welt geschrieben ist, sind davon 78 Prozent der Texte in Englisch verfaßt.
- Bei der Anzahl der im WWW öffentlich zugänglichen Seiten handelt es sich um eine Sammlung, die 50mal größer ist als die bisher gesammelten Textdokumente der LoC.
- Jeden Tag kommen im WWW ca. 7 Millionen neue Seiten hinzu.
- Gleichzeitig verschwinden ständig Inhalte. Die durchschnittliche Lebensdauer einer Webseite beträgt nur 44 Tage. 44 Prozent der 1998 vorhanden Webseiten konnten 1999 nicht mehr gefunden werden.

[1] vgl. Lyman (2002)

[2] 1 Milliarde = 1000 Millionen

[3] Nach Lyman (2002) muß zwischen dem *„surface“* Web und dem *„deep“* oder *„dark“* Web unterschieden werden. Das surface Web beinhaltet alle HTML Seiten, die über eine URL erreichbar sind. Hierzu gehören auch kommerzielle Seiten, die durch Paßwörter oder Verschlüsselung geschützt sind. Dieser Bereich des WWW wird auch als *„private“* Web bezeichnet. Das deep Web beinhaltet große Datenmengen, wie beispielsweise Datenbanken der Klima- und Raumfahrtbehörden. Anfragen in diesen Datenbanken erzeugen Webseiten *on the fly*, die wiederum im surface Web erscheinen. Deshalb kann das deep Web auch als Informationsarchitektur für die Inhalte des surface Web verstanden werden. Das deep Web ist schätzungsweise 500 Mal größer als das surface Web.

– Eine durchschnittliche Webseite enthält 15 Links zu anderen Seiten oder Objekten und 5 Objekte wie zum Beispiel Tonsequenzen oder Bildmaterial.

Diese Fakten machen deutlich, daß mit der Entstehung elektronischer Ressourcen vor allem der Internet-Dokumente im WWW der Prozess der Erschließung in seiner bisherigen Form in Frage gestellt wird. Die traditionell angewandten Verfahren können den Eigenschaften elektronischer Ressourcen oftmals nicht mehr gerecht werden. So stellt die Bayerische Staatsbibliothek fest:

> „Die Überlegungen zur Erschließung von Internet-Dokumenten tangieren grundsätzliche Veränderungen im Aufgabenfeld von Bibliotheken. So manche Definition, die lange Gültigkeit hatte, muss überdacht werden. Kataloge z. B. weisen heute nicht mehr nur das nach, was Bibliotheken besitzen, sondern auch das, wozu sie Zugang ermöglichen. Auch Begriffe wie „Werk“ und „Ausgabe“ müssen genauer definiert und ggf. revidiert werden. Wird eine oft veränderte Online-Datei ständig zu einer neuen Ausgabe oder gar zu einem neuen Werk?“ (BSB 2002)

Die BSB (2002) verweist zwar auf das bereits bestehende Regelwerk RAK-NBM für die Katalogisierung von *Nicht-Buch Materialien* macht aber auch deutlich, daß die Katalogisierung nach RAK-NBM einen hohen Aufwand erfordert, und *„daß das Regelwerk trotz dieses Aufwands den neuen technischen Möglichkeiten und Erfordernissen in den Bereichen Indexierung und Retrieval bei Netzpublikationen nur bedingt gerecht wird.“* (BSB 2002)

Auch die bereits existierenden Formate werden von Endres und Fellner (2000) aufgrund ihrer Vielzahl auf internationaler und nationaler Ebene und ihrer komplexen Anwendung als nicht geeignet für die Beschreibung von digitalen Dokumenten kritisiert. So wird beipielsweise die Arbeit mit dem Format *US-MARC*[4] als so komplex beschrieben, *„... daß nur Spezialisten sie anhand eines 10 cm dicken Handbuchs vornehmen können.“* (Endres und Fellner 2000, S. 299) und weiter heißt es an dieser Stelle: *„ Diese groteske Situation wird dadurch noch verschlimmert, daß deutsche Bibliothekare sich auf ein Metadaten-Format (MAB)*[5] *geeinigt haben, daß nichts mit MARC gemein hat.“* (Endres und Fellner 2000, S. 299).

Erschwerend hinzu kommt außerdem der zeitliche Aspekt. Mit herkömmlichen Verfahren kann die steigende Anzahl zu erschließender elektronischer Ressourcen, speziell der Internet-Dokumente von Bibliotheken nicht mehr bewältigt werden. Byrum (2002) schreibt dazu: *„Der unglaubliche Anstieg von Netzpublikationen in einer Vielzahl von komplexen und variablen Formaten erhöht jedoch mit seinen zahlreichen und schwierigen Herausforderungen die Gefahr des bibliografischen Chaos.“*

[4] US-Machine Readable Catalog. Dieses umfassende Format stammt von der LoC. Eine internationale Variante davon ist der UNI-MARC (vgl. Endres und Fellner 2000, S. 299).

[5] Maschinelles Austauschformat für Bibliotheken

4.2 Metadaten

Bei Metadaten, wörtlich *„Daten über Daten“*, handelt es sich um strukturierte Daten, die zur Beschreibung von Informationsresourcen dienen. Gilliland-Swetland (2000) beschreibt Metadaten als *„... the sum total of what one can say about any information object at any level of aggregation.“* Metadaten bilden somit die Grundlage für alle Arten von strukturierten Beschreibungen, weshalb aus bibliothekarischer Sicht festgestellt werden kann: *„All library service functions are based on metadata“* (Van der Werf 1998).

Wie bereits am Beispiel der Erschließung deutlich wurde, dienen Metadaten in erster Linie dazu, eine Informationsressource nachzuweisen und zugänglich zu machen. Bei analogen Informationsressourcen geschieht dies durch die Katalogisierung.

Mit der Entstehung von elektronischen Ressourcen haben sich gleichzeitig auch neue Möglichkeiten der Darstellung und des Zugriffs auf Informationen, sowohl auf lokaler als auch auf globaler Ebene, entwickelt. Auch in diesem elektronischen Umfeld dienen Metadaten dazu, die digital vorhandenen Ressourcen zu erfassen, zu speichern und zugänglich zu machen. Mit Hilfe von Metdaten kann dem Benutzer in digitalen Informationssystemen eine Vielzahl von Zugangspunkten und Suchmöglichkeiten angeboten werden. Die Voraussetzung hierfür ist jedoch eine hochgradige Strukturierung der elektronischen Ressourcen mit Hilfe von Metadaten, denn *„...the more highly structured an information object is, the more structure can be exploitet for searching, manipulation, and interrelating with other information objects.“* (Gilliland-Swetland 2000)

Um diese Strukturierung gewährleisten zu können, beschäftigen sich seit Mitte der 1990er Jahre zahlreiche Projekte mit der Entwicklung von Metadatenstandards und Metadatadenschemata[6]. Ziel dieser Bemühungen ist es, Alternativen zu den komplexen Regelwerken und Formaten der konventionellen Erschließung zu entwickeln. Unterschiedlichste Anwendergruppen aus verschiedensten Fachgebieten sind daran interessiert, *„... Internetressourcen schnell, einfach, eindeutig und international nachzuweisen“* (Payer 2002). Somit sind bereits zahlreiche Standards zur Beschreibung elektronischer Ressourcen entstanden. Der am meisten verbreitete Metadatenstandard ist der Dublin Core[7].

[6] Für eine Beschreibung der wichtigsten Projekte vgl. (Day 2001a).

[7] Eine Beschreibung zahlreicher Metadatenstandards ist unter http://www.ifla.org/II/metadaten.htm und unter http://www.diffuse.org/oii/en/metadata.html zu finden.

In einem Umfeld, in dem der Benutzer unmittelbaren Zugang zu Informationsressourcen über ein Netzwerk hat, haben Metadaten deshalb folgende Aufgaben (vgl. Gilliland-Swetland 2000):

- Bestätigung der Authentizität und des Grades der Vollständigkeit des Inhalts.
- Erfassung und Dokumentation des Kontexts in dem der Inhalt steht.
- Identifizierung und Nutzung der strukturellen Verbindungen die zwischen den Informationsressourcen bestehen.
- Bereitstellung von verschiedenen intellektuellen Zugangspunkten für unterschiedlichste Benutzer.
- Bereitstellung von Informationen, die auch in einem physikalischen Referenzwerk vorhanden wären.

Um diese Aufgaben erfüllen zu können sind verschiedene Arten von Metadaten mit unterschiedlichen Funktionen notwendig, die in Tabelle 4.1 übernommen aus: Gilliland-Swetland (2000) dargestellt werden.

Type	Definition	Examples
Administrative	Metadata used in managing and administering information resources	-Aquisition information -Rights and reproduction tracking -Documentation of legal access requirements -Location information -Selection criteria for digitization: -Version control and differentiation between similar information objects -Audit trails created by record keeping systems
Descriptive	Metadata used to describe or identify information resources	-Cataloging records -Finding aids -Specialized indexes -Hyperlinked relationships between resources -Annotations by users -Metadata for recordkeeping systems generated by records creators
Preservation	Metadata related to the preservation management of information resources	-Documentation of physical condition of resources -Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	-Hardware and software documentation -Digitization information, e.g., formats, compression ratios, scaling routines -Tracking of system response times -Authentication and security data, e.g., encryption keys passwords
Use	Metadata related to the level and type of use of information and resources	-Exhibit records -Use and user tracking -Content re-use and multi-versioning information

Tabelle 4.1: Typen von Metadaten und ihre Funktionen

Im Gegensatz zu analogen Dokumenten, die bei der Aufnahme in einen bibliothekarischen Bestand einmalig erfaßt werden, durchlaufen elektronische Ressourcen innerhalb von digitalen Informationssystemen verschiedene Phasen. Dies kann Veränderungen der formalen Gegebenheiten der Ressource zur Folge haben. Um diese Veränderungen für eine spätere Nutzung nachvollziehbar machen zu können, muß eine elektronische Ressource in jeder Phase mit den dafür geeigneten Arten von Metadaten erschlossen werden. Abbildung 4.1 übernommen aus: Gilliland-Swetland (2000) stellt den Lebenszyklus eines digitalen Dokuments in einem digitalen Informationssystem, wie beispielsweise einer digitalen Bibliothek, dar.

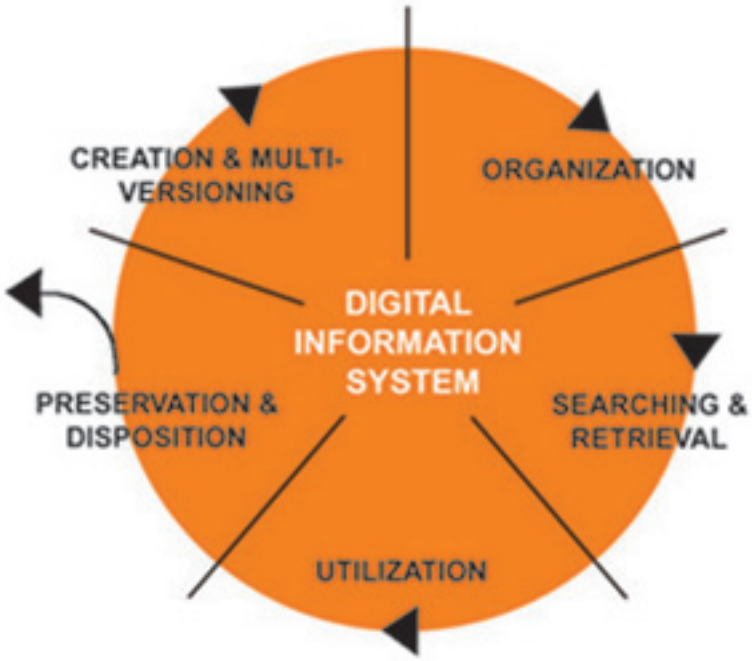

Abb. 4.1: Lebenszyklus eines Dokuments in einem digitalen Informationssystem

Creation and multi-versioning: Das Dokument wird in einen digitalen Bestand aufgenommen. Dazu muß es, wenn es sich nicht um ein born-digital Dokument handelt, zunächst digitalisiert werden. Oftmals werden bereits zu diesem Zeitpunkt verschiedene Versionen derselben Ressource zur Archivierung, Nutzung und Verteilung erzeugt[8]. An dieser Stelle können bereits administrative und deskriptive

[8]Bei der Digitalisierung wird in diesem Zusammenhang auch von der Erzeugung sogenannter *Digitaler Master* oder *Digital master files* gesprochen. Ein *Digitaler Master* ist die digitalisierte Form eines analogen Objekts in höchster Qualität. Aufgabe des *Digitalen Masters* ist es, so genau wie möglich die Informationen des Originalobjekts zu repräsentieren. Der *Digitale Master* dient als Vorlage für die Erstellung weiterer Kopien, die für Benutzer, beispielsweise im Internet oder in einem elektronischen Katalog, zugänglich gemacht werden.

Metadaten vergeben werden.

Organization: Das digitale Dokument wird in den Bestand eingearbeitet. Dies geschieht durch die Erschließung und Indexierung mit Hilfe deskriptiver Metadaten.

Searching and retrieval: Das gespeicherte digitale Dokument soll für Benutzer retrievalfähig sein. Hierfür werden Metadaten zur Ausführung von Retrievalprogrammen erzeugt.

Utilization: Die gefundenen digitalen Dokumente können durch die Benutzer bearbeitet oder vervielfältigt werden. Metadaten für Anmerkungen der Benutzer, Zugangsbeschränkungen und Versionskontrollen werden erzeugt.

Preservation and disposition: Für Gewährleistung der Verfügbarkeit muß das Dokument technisch aktualisiert werden. An dieser Stelle werden auch nicht mehr zugängliche Dokumente gelöscht. Metadaten zur Dokumentation der Archivierungsprozesse werden erzeugt.

Metadaten können entweder Bestandteil eines Dokuments sein oder in einer seperaten Datenbank gespeichert werden und über einen Link auf die beschriebene Ressource verweisen. Das Speichern der Metadaten in einer Datenbank erlaubt eine einfachere Pflege und Weiterleitung der Metadaten.

4.2.1 Dublin Core

Die *Dublin Core Metadata Initiative* (DMCI) hat ihre Anfänge im Oktober 1994 genommen. Vertreter des *Online Computer Library Center* (OCLC) und des *National Center for Supercomputing Applications* (NCSA) diskutierten über die Schwierigkeiten, Informationsressourcen im WWW zu finden[9] (vgl. DCMI 1995). Diese Diskussion gab den Anlaß, 1995 in Dublin Ohio einen Workshop mit dem Titel *„OCLC/NCSA Metadata Workshop“* zu veranstalten. Dort wurde diskutiert, wie das Retrieval von Informationsressourcen im WWW verbessert werden kann. Das Ergebnis dieses Workshops ist das *Dublin Core Element Set*, kurz *Dublin Core* (DC) genannt, daß nach dem Ort seiner Entstehung benannt wurde. Mittlerweile ist DC sowohl von der *International Organization for Standardization* (ISO) als auch von der *National Information Standards Institute* (NISO) als Standard anerkannt[10] und wurde in über 20 Sprachen übersetzt.

[9] Das WWW enthielt zu dieser Zeit etwa 500 000 adressierte Objekte (vgl. DCMI 1995).
[10] ISO 15836 und Z39.85

DC Element	Beschreibung
Title	Name der Ressource
Creator	Autor
Subject	Schlagwörter, die den Inhalt der Ressource beschreiben. (Empfohlen wird die Verwendung von kontrolliertem Vokabular)
Description	Inhaltliche Beschreibung
Publisher	Verleger, Herausgeber
Contributor	Sonstige beteiligte Körperschaften oder Personen
Date	Datum
Type	Ressourcenart (Empfohlen wird die Verwendung von kontrolliertem Vokabular z. B. Dissertation)
Format	Physikalisches oder datentechnisches Format (z. B. text/html oder image/jpg)
Identifier	Identifikation (z. B. URL, ISBN)
Source	Quelle
Language	Sprache
Relation	Beziehung zu anderen Ressourcen
Coverage	Räumliche bzw. zeitliche Ausdehnung/Erscheinungsweise
Rights	Informationen über Urheberrechte/Nutzungsbedingungen

Tabelle 4.2: DC Elemente

Dublin Core simple

Die DCMI definiert DC als *„... a metatdata pidgin for digital tourists: easily grasped, but not necessarily up to the task of expressing complex relationships or concepts.“*

Bei DC handelt es sich um einen einfachen, kostengünstigen Metadatenstandard, der aus 15 Elementen besteht, die so angelegt sind, daß ein Autor sie selbst in sein Dokument einfügen kann, z.B. im Header seines HTML Dokuments. Diese ursprüngliche Form des DC wird daher auch als *DC simple* bezeichet[11]. Die 15 Elemente des DC simple sind in Tabelle 4.2 übernommen aus: ISO (2003) beschrieben. Wenngleich sich DC besonders bei Produzenten von Internet-Ressourcen als Metadatenstandard durchgesetzt hat, gibt es aus bibliographischer Sicht einige Kritikpunkte an DC und anderen Metadaten-Strukturen, die sich in den letzten Jahren entwickelt haben. Byrum (2002) nennt es charakteristisch für die verschiedenen Schemata, *„... dass sie Strukturen für*

[11] Die aktuelle Version ist Dublin Core 1.1.

Herkunftsinformationen der Veröffentlichung bieten, aber wenig Anhaltspunkte für die Beschreibung.“

Jedoch ist bei dieser Kritik das ursprüngliche Ziel von DC zu berücksichtigen. Payer (2002) verweist darauf, daß DC ursprünglich als einfaches Schema für einfache Internetressourcen entwickelt wurde und dafür auch ausreichend ist. Auch ISO (2003) macht auf die Stärken und Schwächen von DC aufmerksam: *„The simplicity of Dublin core can be both a strength and a weakness. Simplicity lowers the cost of creating metadata and promotes interoperability. On the other hand, simplicity does not accommodate the semantic and functional richness supported by complex metadata schemes.“*

Dublin Core qualified

Mit der Schaffung von DC qualified versucht die DCMI auch bibliothekarischen Ansprüchen an die Erschließung gerecht zu werden. DC qualified hat hierfür zwei Klassen von sogenannten *qualifiern* eingeführt:

Element Refinements: Dieser qualifier dient zur spezifischeren Beschreibung eines Elements ohne die Bedeutung des Elements zu verändern.

Encoding Schemes: Dieser qualifier beschreibt, nach welchen Regelwerken der Feldinhalt angegeben werden soll. Beispiel Date: 2003-06-23[12].

In Tabelle 4.3 auf der nächsten Seite übernommen aus: DCMI (2000) wird die Anwendung von DC qualified am Beispiel der Elemente Date und Relation dargestellt.

Das Dumb-down Prinzip

Die Verwendung von DC qualified basiert auf dem sogenannten *Dumb-down Prinzip*. Dieses Prinzip besagt, daß ein Anwender alle qualifier ignorieren kann und eine Ressource trotzdem anhand der Elemente des DC simple auffindbar bleiben muß. Nach der DCMI (2000) dient der Einsatz von DC qualified *„... only to refine, not extend the semantic scope of an element.“*

Interoperabilität

Mit DC wurde ein Metadatenstandard entwickelt, der es Autoren von Internetressourcen ermöglicht, selbst Metadaten zu vergeben. Dies wird wohl auch in Zukunft

[12] Dieser Form der Angabe des Datums sind die *W3C Encoding rules for dates and times* zugrunde gelegt, die wiederum auf dem ISO-Standard 8610 basieren. Weitere Beispiele für Encoding schemes sind der *Uniform Resource Identifier* (URI) und eigene Encoding schemes der DCMI, wie zum Beispiel DCMI Period. Angaben zu den DCMI Encoding schemes sind unter http://dublincore.org/usage/terms/dcmitype zu finden.

DC Element	Element Refinement(s)	Element Encoding Scheme(s)
Date	Created Valid Available Issued Modified	DCMI Period W3C-DTF
Relation	Is Version of Has Version Is Replaced By Replaces Is Required By Requires Is Part of Has Part Is Referenced By References Is Format Of Has Format	URI

Tabelle 4.3: Beispiele für DC Qualifiers

hauptsächlich durch die Anwendung von DC simple geschehen. Byrum (2002) merkt dazu an, *„... dass Autoren und Verlage kein großes Interesse an der Komplexität bibliografischer Beschreibungen, standardisierter Zugangsmöglichkeiten und sachlicher Erschließung haben.“*

Umgekehrt wird selbst DC qualified traditionelle bibliothekarische Regelwerke, wie AACR2 oder RAK, nicht ersetzen können (vgl. Payer 2002). Trotzdem muß für eine vollständige bibliographische Beschreibung nicht auf die Vorteile von DC verzichtet werden, da DC ein hohes Maß an Interoperabilität bietet.

Für die Einbindung und Verwertung bereits vorhandener DC Metadatensätze existieren bereits zahlreiche Möglichkeiten und Vorschläge. Bereits realisiert werden Umsetzungen von DC in verschiedene Formate wie MARC und MAB[13] (vgl. Payer 2002). Neben der Einbindung von DC in HTML kann DC auch in XML implementiert werden. Das DC Metadatenschema kann auch mit Elementen aus anderen Metadatenschemata gemischt werden. Dies ist dann sinnvoll, wenn es sich um Anwendungen handelt, die Elemente aus dem DC benötigen oder umgekehrt.

Byrum (2002) weist außerdem auf die Notwendigkeit der Zusammenarbeit zwischen Bibliotheken und Produzenten von Internetressourcen hin. Besonders großen Einrichtungen wie Nationalbibliotheken schlägt Byrum (2002) vor, Metadaten-Normdateien

[13]Informationen zur Umsetzung in andere Formate sind unter http://www.ifla.org zu finden.

erstellen zu lassen und damit die Produzenten von Netzpublikationen anzuregen, brauchbare Metadaten in ihre Produkte zu übernehmen. DC selbst bietet für DC simple einen sogenannten Metadaten Generator an, der aus den vom Autor angegebenen Informationen DC Metadatensätze generiert.

4.2.2 ISBD(ER)

Die *International Standard Bibliographic Description* (ISBD) ist das Ergebnis einer Konferenz für Katalogisierungsexperten, die von der IFLA 1969 in Kopenhagen organisiert wurde. Auf dieser Konferenz wurde beschlossen, daß ein international gültiger Standard für die bibliographische Beschreibung eingeführt werden soll. Ihre Hauptaufgabe sieht die ISBD darin, diesen Standard weltweit zu verbreiten und somit für einheitliche bibliographische Beschreibungen zu sorgen. Diese Vereinheitlichung vereinfacht den internationalen Austausch bibliographischer Beschreibungen und erleichtert die Übernahme von bibliographischen Beschreibungen in elektronische Form.

Die ISBD bildet zusammen mit den Regelwerken für die Katalogisierung die Voraussetzung für eine komplette bibliographische Erschließung. In der Regel werden die aus der ISBD benötigten Elemente in das entsprechende Regelwerk eingearbeitet. Die ISBD legt fest, welche Elemente in einer bibliographischen Beschreibung angegeben werden müssen, in welcher Reihenfolge diese Elemente angegeben werden müssen und durch welche Zeichensetzung sie voneinander abgegrenzt werden.

Mittlerweile existieren sieben ISBDs:

- ISBD(M) für Monographien
- ISBD(CM) für Kartographische Materialien
- ISBD(S) für Serien
- ISBD(NBM) für Nicht-Buch-Materialien
- ISBD(A) für Monographien vor dem 18. Jahrhundert
- ISBD(PM) für Noten
- ISBD(ER) für elektronische Ressourcen

Bei der ISBD(ER) handelt es sich um die überarbeitete Fassung der ISBD(CF) (=Computer files). Sie beinhaltet demzufolge Regeln für die bibliographische Beschreibung von elektronischen Ressourcen. Die auf der nächsten Seite abgebildete ISBD(G)[14] übernommen aus: Van der Werf (1998) zeigt, wie eine ISBD aufgebaut ist. Die Auflistung macht die Komplexität der ISBD deutlich. Payer (2002) stellt fest, daß sich bei elektronischen Ressourcen, vor allem bei Internettexten, die Anwendung der ISBD(ER)

[14](G) steht für general.

ISBD(G) Description areas and elements

1. Title and statement of responsibility area
title proper
[general material designation]
= parallel title
: other title information
/ statement of responsibility
; subsequent statement of responsibility

2. Edition area
edition statement
= parallel edition statement
/ statements of responsibility relating to the edition
, additional edition statement

3. Material (or type of publication) specific area
specific material designation

4. Publication, distribution, etc., area
place of publication, distribution, etc.s̈ ; subsequent place
: name of publisher, distributor, etc.
[]statement
, date of publication, distribution, etc.
(place of manufacture
: name of manufacturer
,) date of manufacture

5. Physical description area
specific material designation and extend of item
: other physical details
; dimensions of item
+ accompanying material statement

6. Series area
(title of series statements
/ responsibility statements
, ISSN of series
; numbering within series
enumeration and/or sub-series statements)

7. Note area
additional information on the physical make-up ot the item or its contents

8. Standard number and terms of availability area
standard number
key title of ISSN
: terms of availability and/or price
()qualification

meist nicht lohnt. So erübrigt sich beispielsweise die Beschreibung der Vorlageform bei Internettexten gänzlich, da man hier die Vorlage mit einem Klick sehen kann (vgl. Payer 2002). Payer (2002) schlägt deshalb vor, bei Internettexten das Inhaltsverzeichnis und die Titelseite einzuscannen und nur noch *„... wenige, insbesondere normierte Elemente zu erfassen.“*

4.3 Gewährleistung der Authentizität digitaler Dokumente

Für die Auffindbarkeit von elektronischen Ressourcen ist, neben der Strukturierung mit Hilfe von Metadaten, auch eine dauerhafte Standortangabe der Ressource eine wichtige Voraussetzung. Leider erweist sich der *Uniform Resource Locator* (URL) als Standortangabe einer Internetressource oftmals als instabil. Webseiten werden durch Serverwechsel oder Überarbeitungen nicht mehr auffindbar, der Benutzer erhält die Fehlermeldung „404-Site Not Found“ nach dem er eine URL aufgerufen hat. Um eine dauerhafte Auffindbarkeit von Internetressourcen gewährleisten zu können, wurden die sogenannten *Persistent Identifier* (PI) entwickelt.

4.3.1 Persistent Identifier

> „Persistent Identifers sind eindeutige, standortunabhängige Identifikatoren für digitale Objekte, mit denen gleichzeitig der dauerhaft Zugriff auf digitale Ressourcen gewährleistet wird.“ Schroeder (2003)

Der Einsatz von Persistent Identifiern bietet Vorteile gegenüber der Verwendung von URLs (vgl. Schroeder 2003; HeBIS 2002a):

- Dauerhafte und zuverlässige Erreichbarkeit und Zitiermöglichkeit von elektronischen Ressourcen durch stabile Links in Dokumenten, Nachweisdiensten und Katalogen.
- Effizientere Recherche nach elektronischen Ressourcen, da Persistent Indentifier weltweit eindeutig sind.
- Effektivere Verwaltung von internen elektronischen Ressourcen, zum Beispiel bei Änderung des Standorts im Informationssystem oder für eindeutige Identifizierung von Dokumenten von denen mehrere Kopien vorhanden sind.
- Die wissenschaftliche Arbeit wird durch international einheitliche Regelungen unterstützt.

Um diese Vorteile der Persistent Identifier nutzen zu können, müssen jedoch verschiedene Voraussetzungen erfüllt sein. Hierzu gehören (vgl. HeBIS 2002b):

- Die Schaffung von Standards für die Struktur und Syntax eines Persistent Identifiers. Beispiele hierfür sind der *Uniform Resource Name* (URN) und der *Digital-Object-Identifier* (DOI).
- Registrierungsagenturen für Persistent Identifier.
- Systeme für die Verwaltung von Persistent Identifiers.
- Sogenannte *Resolving* Mechanismen, die geänderte URLs automatisch dem gültigen Persistent Identifier zuordnen.

4.3.2 Uniform Resource Name (URN)

Bei URN handelt es sich um einen nicht kommerziellen Persistent Identifier Standard dessen Syntax 1997 als Internet Standard RFC2141 von der *Internet Engineering Task Force* (IETF)[15] definiert wurde. (vgl. Van der Werf 1998; HeBIS 2002b). Die Betreuung des URN-Systems wird von der LoC koordiniert. Ein URN besteht aus Elementen, die in nachfolgender Tabelle übernommen aus: HeBIS (2002a, b) dargestellt werden:

Abkürzung	Bezeichnung	Beispiel
URN	Uniform Ressource Name	Kennzeichnung aller PI des URN-Schemas
NID	Namespace Identifier	NBN = National Bibliography Number
SNID	Subnamspace Identifier (wiederholbar)	SNID1:DE = Deutsche Bibliothek SNID2:HEBIS = Bibliotheksverbund (Kürzel) SNID3:30 = Universitätsbibliothek (Sigel)
NSS	Namespace Specific String	Nummerischer oder alphanummerischer String mit Prüfziffer

Tabelle 4.4: Die Elemente des Uniform Resource Name

1999 wurde im Rahmen der *Conference of Directors of National Libraries* (CDNL) festgelegt, daß alle am URN-System beteiligten Nationalbibliotheken ihre zu vegebenden Persistent Identifier mit dem Namespace NBN einleiten. Die Deutsche Bibliothek in Frankfurt am Main verwaltet als Teilmenge der NBN den Subnamespace „de.“. An einem Beispiel aus der HeBIS-Verbundatenbank[16] sieht die Umsetzung eines URN in der Praxis wie folgt aus (vgl HeBIS 2002b): urn:nbn:de:hebis:30-0000000759.

[15] http://www.ietf.org/home.html
[16] HeBIS = Hessisches Bibliotheksinformationssystem.

URN-Verwaltung bei der Deutschen Bibliothek

In Deutschland führte die Deutsche Bibliothek (DDB) 2001 im Rahmen eines Projekts Persistent Identifier für Online-Dissertationen ein. Für diese Persistent Identifier übernimmt die DDB auch die zentrale Verwaltung. Projektpartner sind Universitätsbibliotheken in ganz Deutschland, darunter die Stadt- und Universitätsbibliothek/Senckenbergische Bibliothek Frankfurt am Main.

Die URNs werden dezentral von den teilnehmenden Projektbibliotheken erzeugt. Der entsprechende Subnamespace wird von der DDB zur Verfügung gestellt. Die Verwaltung und Registrierung der erzeugten URNs erfolgt wiederum zentral durch die DDB. Die Meldung der vergebenen URNs erfolgt zusammen mit der Übergabe der Metadaten der Online-Dissertationen. Hierfür stellt die DDB ein *„Anmeldeformular für Hochschulschriften“*[17] im Internet zur Verfügung. Die DDB archiviert die angemeldeten Dissertationen und nimmt sie zusammen mit den URNs in ihren Katalog auf (vgl. HeBIS 2002a).

URN-Resolving

Resolving bezeichnet den Prozess, der geänderte URLs nach wie vor gültigen URNs automatisch zuordnet. Im Fall der DDB werden URL Adressänderungen von den Projektbibliotheken an die DDB weitergeleitet. Über entsprechend programmierte Skripte wird das Resolving in Gang gesetzt. Zusätzlich erfolgt nach dem Resolving ein Linkcheckverfahren, daß regelmäßig die Erreichbarkeit von URLs prüft und eine Fehlermeldung ausgibt, wenn eine URL nicht mehr erreichbar ist. Die Projektbibliotheken müssen dann eine Nachmeldung des gültigen URL vornehmen. Die Änderungsmeldungen und das Linkcheckverfahren gewährleisten, daß keine URL-Adressänderung unbemerkt bleibt. Somit ist die Langzeitverfügbarkeit und Referenzierbarkeit der elektronischen Ressource gesichert (vgl. HeBIS 2002a).

Persistent URL (PURL)

Bislang können Browser Persistent Identifier nicht interpretieren und von sich aus in einen gültigen URL umsetzen. Damit der Endnutzer trotzdem durch Eingabe oder anklicken des URN zu dem dazugehörigen Dokument gelangt, wird ein *Persistent URL*(PURL)[18] benötigt. Im OPAC der DDB wird durch das Anklicken des URN ein Skript aufgerufen, das den URN in einen Persistent URL umsetzt (vgl. HeBIS 2002b).

Ein PURL besteht aus einem Persistent Identifier und einer Adresse für den Resolving-Dienst, der den PI in eine gültige URL auflöst und den Nutzer an diese weiterleitet.

[17] Das Formular ist unter http://www.ddb.de/professionell/anmeldeformulare/hochschulschrift.htm zu finden.

[18] http://www.purl.org

Der PURL muß deshalb der URN vorangestellt sein. Dies kann in der Praxis folgendermaßen aussehen (vgl. HeBIS 2002a, b):

http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:hebis:30-0000000759

4.3.3 Digital Object Identifier (DOI)

Bei DOI[19] handelt es sich um ein kommerzielles PI-System, daß von der *International DOI Foundation* (IDF) betrieben wird. Als Reaktion auf die Schwierigkeiten mit e-publishing und e-commerce im Internet initiierte die *Association of American publishers* (AAP) die Entwicklung eines Systems, mit dem Urheberrechte geschützt und kommerzielle Transaktionen erleichert werden sollten (vgl. Van der Werf 1998). Die Zielsetzung der DOI ist es, Kunden und Content-Anbieter ortsunabhängig miteinander zu vernetzen und e-commerce zu erleichtern, sowie eine Basis für eine automatisierte Verwaltung von Urheber- und Lizenzrechten zu schaffen (vgl. HeBIS 2002b).

Durch DOI wird eine eindeutige Identifikation von elektronischen Ressourcen ermöglicht. Damit die Inhalte dieser Ressourcen identifizierbar sind, muß jeder DOI mit Metadaten verknüpft werden. Diese Metadaten geben Auskunft über die Ressource und werden außerdem für den Resolving-Prozess benötigt. Ein einmal vergebener DOI bleibt unverändert, während sich die Metadaten ändern können. Dies ist zum Beispiel bei einem Wechsel der Eigentumsrechte einer Ressource der Fall. In einem zentralen Verzeichnis, dem DOI-Directory werden die Anbieter und die Ressourcen registriert. Das DOI-Directory fungiert als Vermittler zwischen Anbieter und Nutzer und muß daher immer die aktuelle gültige URL des Anbieters enthalten (vgl. Payer 1997).

Ein DOI besteht aus einem *Präfix* und einem *Suffix* (vgl. HeBIS 2002b). Präfix und Suffix werden durch einen Schrägstrich getrennt.

Präfix: Der Präfix eines DOI beginnt immer mit einer 10 als Namensraum, gefolgt von einer Nummer für die Content-Einheit, zum Beispiel ein Verlag oder eine Produktlinie.

Suffix: Der Suffix dient als Identifikator für den jeweiligen Content und kann Einheiten beliebiger Größe und beliebiger Dateitypen bezichnen (Buch, Artikel, Abstract, Bild, Video oder Software).

Ein DOI kann beispielsweise so aussehen (vgl. HeBIS 2002b):

10.1007/s00468-002-0161-y

DOI-Verwaltung

Die DOI Präfixe werden von Registrierungsagenturen wie beispielsweise der ISBN-Agentur Preußische Staatsbibliothek in Berlin vergeben. Diese Registrierungsagentu-

[19] http://www.doi.org

ren legen außerdem Metadatenstandards fest und betreiben entsprechende Datenbanken zur Verwaltung der Metadaten. Für die Vergabe von DOI-Präfixen, die Registrierung von DOIs und das DOI Retrieval werden von den Registrierungsagenturen Gebühren erhoben (vgl. HeBIS 2002b).

Die IDF hat die Aufgabe, die internationale DOI Entwicklung zu überwachen und legt fest, welche Registrierungsagenturen und Technologieanbieter Lizenzen für die Vergabe von DOI Präfixen erhalten. 2002 hatte die IDF 70 Mitglieder. Es handelt sich dabei um Unternehmen aus Asien, Europa und den USA wie beispielsweise den Springer Verlag oder Elsevier Science. Die IDF finanziert sich über jährliche Mitgliedsbeiträge, die nach Mitgliedskategorien gestaffelt sind. So beträgt beispielsweise der jährliche Mitgliedsbeitrag in der höchsten Kategorie „Charter-Members“ 35 000 US Dollar (vgl. HeBIS 2002b).

DOI-Resolving

Das DOI Resolving erfolgt wie der PURL-Mechanismus im OPAC de DDB über den Aufruf eines Resolving-Dienstes gefolgt von der Angabe des DOI. Im Fall der IDF wäre das: http://dx.doi.org/<entsprechender DOI>. Der gültige URL wird aus dem DOI-Directory ermittelt, die Verbindung erfolgt direkt zum gewünschten Dokument oder zu einer Liste, die auf das Dokument bezogene Informationene enthält, wie beispielsweise Rezensionen oder Bezugsmöglichkeiten (vgl. HeBIS 2002b).

Van der Werf (1998) sieht in der Kostenpflichtigkeit von DOI Vor- und Nachteile. Einerseits werden die mit der Teilnahme erforderlichen Verpflichtungen durch den Kostendruck eher erbracht, andererseits schrecken die Mitliedsbeiträge und anfallenden Gebühren kleinere Verlagshäuser und publizierende Einrichtungen von der Benutzung der DOI ab.

4.4 Datenformate und Standards

Datenformate und Standards sind für die Langzeitarchvierung von digitalen Dokumenten von zentraler Bedeutung. Die steigende Anzahl von neuen Produktentwicklungen im Bereich der Hard- und Software bringt auch die Entstehung neuer Datenformate und Standards mit sich. Die meisten dieser Datenformate und Standards sind jedoch insbesondere für die Langzeitarchivierung ungeeignet, da ihre Funktionsfähigkeit in der Regel von Hard- und Software abhängig ist.

Datenformate

Das Datenformat oder auch Dateiformat[20] legt fest, wie die Daten auf einem Speichermedium gespeichert werden. Die Daten werden in einer bestimmten Struktur gespei-

[20] engl: *Logical format*

chert, die ein späteres Öffnen wieder ermöglicht. Diese Struktur hängt von verschiedenen Faktoren ab. Hierzu gehören die Art des Programmes, der zu speichernde Inhalt, der Hersteller der Software, das Betriebssystem und der verwendete Zeichensatz (vgl. Grabmeyer und Schimmer 2000). Die unzähligen, exisitierenden Datenformate können grob in Text-, Bild-, Graphik-, Audio- und Videoformate eingeteilt werden. Auf eine Beschreibung einzelner Datenformate an dieser Stelle wird verzichtet[21].

Für die Speicherung der Datenformate gibt es zwei Möglichkeiten. Wird ein Datenformat im binären Code gespeichert, kann es nur von einem bestimmten Programm verarbeitet werden und ist für den Menschen nicht lesbar. Ein Datenformat kann aber auch in für den Menschen lesbaren Zeichen gespeichert werden. Je nach verwendetem Zeichensatz handelt es sich dann dabei um das ASCII [22]- Format, das 7 Bit umfaßt, oder das ANSI-Format, das 8 Bit umfaßt (vgl. Grabmeyer und Schimmer 2000; Dickschus und Brebeck 2000).

Nach Gschwind u. a. (2000) muß ein für die Langzeitarchivierung geeignetes Datenformat *„...plattformunabhängig gelesen und beschrieben werden können.“*

Standards

Standards können hinsichtlich ihrer Entstehung in zwei Arten unterschieden werden. In *de-facto-Standards* und *de-jure-Standards*. Bei de-facto -Standards handelt es sich um eine Regel, die in der Öffentlichkeit weit verbreitet ist. *„Sie wird aufgrund ihrer Verbreitung und nicht durch ein Gremium zum Standard erhoben.“* (Grabmeyer und Schimmer 2000) De-jure-Standards entstehen aufgrund der Zustimmung eines offiziellen Normierungs- oder Standardgremiums (vgl. Grabmeyer und Schimmer 2000). Drei der wichtigsten Standardisierungsgremien werden an dieser Stelle vorgestellt.

International Organization for Standardization (ISO)

Die ISO ist die Dachorganisation der konventionellen, nationalen Standardgremien mit Sitz in Genf[23]. In Deutschland ist das *Deutsche Institut für Normung* (DIN) ISO-Mitglied[24]. Neben der ISO ist auch die 40 Jahre fürher gegründete *International Electronical Comission* (IEC)[25] für den technischen Bereich von Bedeutung. Wegen der Zusammenarbeit von ISO und IEC tragen viele Standards die Bezeichnung ISO/IEC (vgl. Behme und Mintert 2000, S. 477). Im Zusammenhang mit der Entwicklung des Internet sind zwei ISO-Standards zu erwähnen:

SGML (ISO 8879): Mit SGML hat die ISO die erste Metasprache zur Definition von Auszeichnungssprachen definiert (vgl. Grabmeyer und Schimmer 2000).

[21] Eine Beschreibung von Datenformaten findet sich bei Grabmeyer und Schimmer (2000).
[22] American Standard Code for Information Interchange.
[23] http://www.iso.ch
[24] http://www.din.de
[25] http://www.iec.ch

HTML (ISO/IEC 15445:2000): Es handelt sich bei diesem Standard um eine Variante von HTML 4, die üblicherweise auch als ISO-HTML bezeichnet wird (vgl. Behme und Mintert 2000).

Internet Engineering Task Force (IETF)

Die IETF kümmert sich zuammen mit der *Internet Society* (ISOC) um die offene Weiterentwicklung des Internet. Die IETF entwickelt dabei keine Standards für Datenformate, sondern setzt sich mit der Infrastruktur des Internets auseinander. Die dabei entstehenden Veröffentlichungen sind die *Requests for Comments* (RFC) wie beispielsweise der RFC für die Entwcklung der URNs. Werden in RFCs Datenformate anderer Gremien behandelt, so geschieht dies meist in Form von Verbesserungsvorschlägen, die dann wiederum von anderen Standardorganisationen aufgenommen werden (vgl. Grabmeyer und Schimmer 2000).

World Wide Web Consortium (W3C)

Dieses Konsortium wurde 1994 gegründet und hat seinen Sitz in Genf[26]. Die Aufgabe des W3C ist die Normung von HTML und dem WWW. Das W3C gibt dafür die maßgeblichen Spezifikationen für die Weiterentwicklung von HTML und des WWW heraus. Zu den Mitgliedern des W3C zählen viele Firmen der Softwarebranche, aus Deutschland zum Beispiel Siemens und SAP. Die Interessen an der Weiterentwicklung des WWW sind daher auch meist kommerzieller Natur (vgl. Behme und Mintert 2000, S. 478). Steyer (1998) sieht außerdem in der Tatsache, daß viele der Mitglieder auf dem Internet-Markt als Konkurrenten auftreten die Ursache für die geringen Fortschritte bei der Weiterentwicklung einheitlicher Strukturen, wie beispielsweise bei der Sprachstandardisierung von HTML (vgl. Steyer 1998, S. 33).

Zur Entstehung von De-facto-Standards

Wie bereits erwähnt wurde, erhalten de-facto-Standards ihre Legitimierung nicht durch Standardisierungsgremien, sondern durch die Öffentlichkeit. Die Anerkennung eines de-jure-Standards durch ein Standardisierungsgremium hat nicht unbedingt zur Folge, daß sich dieser Standard auch durchsetzen wird (vgl. Grabmeyer und Schimmer 2000). Dies liegt vor allem an daran, daß die Verbreitung eines offiziell verabschiedeten Standards stark von der Unterstützung der großen Programmhersteller ist[27].

Umgekehrt gelingt es Softwareanbietern immer wieder, Formate als Standard durchzusetzen, die *„.. in keiner Weise standardisiert sind und teilweise nicht einmal offen*

[26] http://www.w3.org

[27] Als Beispiel für einen an der Akzeptanz von Programmherstellern gescheiterten Standard gilt HTML 3.0 vgl. dazu auch Grabmeyer und Schimmer (2000); Steyer (1998).

zugänglich.“ (Grabmeyer und Schimmer 2000). Dies ist beispielsweise bei Anwenderprogrammen wie den WWW-Browsern zu beobachten.

Gelingt es einer Firma, ein proprietäres Datenformat zum de-facto-Standard zu machen, kann dieses Unternehmen Lizenzgebühren für die Nutzung des Datenformats in Anwendungen anderer Firmen verlangen (vgl. Grabmeyer und Schimmer 2000).

Ein Unternehmen kann aber auch Editierprogramme zum Erstellen von Inhalten in ihrem Datenformat verkaufen, und die dazugehörigen Anzeigeprogramme kostenlos abgeben. Diese Vorgehensweise praktiziert die Firma Adobe bei Ihrem Datenformat *Portable Document File* (PDF) und die Firma Macromedia bei ihrem Datenformat Flash (vgl. Grabmeyer und Schimmer 2000). Diese Verkaufspolitik bewirkt, daß Anwender für die Erzeugung von PDF-Dateien meist auch ein Editierprogramm der Firma Adobe verwenden, auch wenn andere Anbieter Editierprogramme für dasselbe Datenformat entwickeln und anbieten.

Ebenfalls eine Form der Verkaufsförderung stellt die Verfügbarkeit von Anzeigeprogrammen nur für bestimmte Betriebssysteme dar. Diese Verkaufspolitik, wird vor allem von der Firma Microsoft praktiziert (vgl. Grabmeyer und Schimmer 2000).

Derartige Erzeugungen von Produktabhängigkeiten erschweren die Langzeitarchivierung digitaler Dokumente erheblich. Hinzu kommt, daß proprietäre- und de-facto-Standards durch Nichtoffenlegung auch für Archivierungszwecke nicht verfügbar sind.

Verfügbarkeit von Standards

Grundsätzlich läßt sich die Verfügbarkeit von Standards für die Öffentlichkeit in drei Varianten unterscheiden, die vor allem im informationstechnischen Bereich zu finden sind (vgl. Grabmeyer und Schimmer 2000):

- Proprietäre Standards: Die Vermarktung und Entwicklung erfolgt durch Firmen oder Konsortien. Diese Standards dienen komerziellen Zwecken und werden in der Regel nicht offengelegt.
- Öffentliche Standards: Die Entwicklung erfolgt über internationale Standardisierungsgremien. Diese Standards dienen nichtkommerziellen Zwecken, die Anwendungsprogramme sind kostenlos. Standardisierungsprozesse dauern relativ lange, da die Bedürfnisse möglichst vieler Anwender berücksichtigt werden sollen.
- Offene Standards: Die Entwicklung und Veröffentlichung erfolgt durch die Anwender selbst. Durch die ständige Veränderung des Standards, läßt sich eine breite Unterstützung für solche Formate kaum realisieren.

In den kommerziellen Interessen der Anbieter sowie der Nichtoffenlegung ihrer Standards sieht auch Hedstrom (2002) die Hauptgründe für die Bevorzugung offener Standards gegenüber De-facto-Standards bei der Archivierung. Während offene und öffentliche Standards bereits offengelegt und verfügbar sind, entwickeln sich de-facto-Standards oftmals zu proprietären Standards Zudem sind proprietäre Standards nach

Ansicht von Hedstrom (2002) besonders aufgrund ihrer Herstellerabhängigkeit nicht für die Archivierung digitaler Dokumente geeigenet. *„If a digital preservation strategy depends on proprietary standards, then the long-term future of digital resources in that proprietary standards is dependent upon the longevity of the firm that owns the standard or on its continued market dominance.“* (Hedstrom 2002).

Für Hedstrom (2002) ist es außerdem von Bedeutung, daß die Entwicklung offener Standards auf dem Konsens, verschiedener Interessengruppen basiert. Hedstrom (2002) sieht deshalb generell bei der Entwicklung von Standards die Ideallösung in der Zusammenarbeit von Institutionen und Einzelpersonen und schreibt dazu: *„... institutions that are building digital libraries and individuals with particular expertise in this area would participate avtively in standards developments that affect the ability to preserve digital information and lower the cost of doing so.“*.

4.4.1 XML

Bereits in den vorhergehenden Abschnitten wurde deutlich, daß die Langzeitarchivierung und die Langzeitverfügbarkeit von elektronischen Ressourcen ein hohes Maß an Strukturierung und Standardisierung der digitalen Daten erfordert. Die hierfür verwendeten Systeme sollen jedoch nicht proprietär, hardwareunabhängig und an unterschiedliche Bedürfnisse anpaßbar sein.

Die Metasprache XML scheint diese Anforderungen erfüllen zu können. XML ist eine Untermenge der Metasprache SGML. Sowohl SGML als auch XML beruhen auf dem bereits seit den 1960er Jahren bekannten Verfahren des Generic Markup, der Trennung von Inhalt und äußerer Form eines Dokuments. Markups sind ursprünglich handschriftliche Markierungen. Diese Markierungen wurden im Druckwesen als typographische Informationen für das Layout eines Manuskripts, wie zum Beispiel Seitenformat oder Zeichensätze eingesetzt. Die Markups wurden von einem Layouter handschriftlich in das Manuskript eingefügt und beim anschließenden Setzen berücksichtigt (vgl. Behme und Mintert 2000, S. 36).

Mit der Entwicklung des WWW Ende der 1980er Jahre entwickelte sich auch HTML als ein Standard für die einfache, kostengünstige und plattformübergreifende Verteilung von Informationen im Internet. HTML basiert auf SGML, jedoch handelt es sich bei HTML nicht wie bei XML um eine Metasprache, sondern um eine Auszeichnungssprache. Mit der Weiterentwicklung des WWW reichte HTML für die steigenden Anforderungen an die Übertragung und Präsentation von Informationen allein nicht mehr aus. Ein Grund hierfür liegt in der begrenzeten Anzahl von Elementtypen bei HTML. Insgesamt stehen für die Dokumentauszeichnung im HTML-Standard lediglich 70 Elementtypen zur Verfügung. Die zunehmende Nichteinhaltung des HTML-Standards führt außerdem zu Problemen bei der Darstellung von Dokumenten (vgl. Winter 2001).

Da SGML aufgrund seiner hohen Komplexizität als Standard für das WWW nicht einsetzbar war, bemühte sich das W3C einen neuen Standard zu entwickeln, mit dem

die Vorteile von SGML auf einfache Weise genutzt werden konnten. Als Ergebnis wurde der erste XML Entwurf im Rahmen der SGML´96 Konferenz in Boston vorgestellt (vgl. Behme und Mintert 2000, S. 41). 1998 wurde die erste XML-Spezifikation veröffentlicht und über das Internet frei zugänglich gemacht. Der aktuelle XML-Standard ist XML 1.0 (Second Edition) (vgl. Winter 2001).

Aufgrund des Prinzips des Generic Markup enthalten XML-Tags keinerlei Informationen, wie der Inhalt dargestellt werden soll. Ein einmal erstelltes XML-Dokument kann mit Hilfe von Stylesheets[28] in beliebigen Variationen dargestellt werden[29].

Neben dem Generic Markup ist die *Dokument-Typ-Definition* (DTD) ein wichtiges Merkmal von XML (vgl. Mintert 1999). Im Gegensatz zu der bei HTML begrenzeten Anzahl von Elemente können Anwender in XML beliebige Elemente selbst definieren, die in der DTD deklariert werden. Viele Vorteile von XML hängen damit zusammen, daß man einen Arbeitschritt in Bezug auf die DTD durchführt und das Ergebnis auf alle Dokumente dieses Typs angewendet werden kann. Somit genügt es einmalig eine DTD für einen bestimmten Dokumenttyp zu erstellen. Behörden und Verlage, die mit XML arbeiten, stellen ihren Mitarbeitern beziehungsweise Autoren teilweise bereits Standard DTDs zur Verfügung. Somit ist ein einheitliches Format gewährleistet. Hierdurch wird die Weiterverarbeitung erleichtert und Fehlerquellen werden reduziert. Die in den Elementen einer DTD enthaltenen Metadaten sind außerdem alle in XML retrivalfähig.

Diese Funktionalität ermöglicht in Zukunft auch die bibliographische Erschließung mit Hilfe von *„... einfachen und klaren Document Type Definitions, die bibliographische und inhaltliche Daten leicht extrahierbar machen.“* (Payer 2002)

XML bietet mit der *XML Linking Language* (XLink)und der *XML Pointer Language* (XPointer) außerdem eine Erweiterung des Hypertext-Konzeptes. Im Gegensatz zu der Syntax von DTDs liegt für das Linking noch keine Spezifikation vor. Die gegenwärtig vorliegenden Entwürfe müssen dem W3C noch zur Abstimmung vorgelegt werden, bevor sie zum Standard ernannt werden können (vgl. Behme und Mintert 2000, S. 103).

XLink

Eine Besonderheit bietet die in XLink mögliche Form der *erweiterten Links*. Nach der Definition aus dem XLink-Entwurf wird ein erweiterte Link wie folgt beschrieben:

> „ Ein erweiterter Link ist ein Link, der eine beliebige Anzahl von Ressourcen verbindet. Die beteiligten Ressourcen können irgendeine Kombination von lokalen und entfernten Ressourcen sein. Sind alle entfernte, handelt es sich um einen Out-of-line-Link. Ist auch nur eine Ressource lokal, handelt es sich um einen Inline-Link.“ (Behme und Mintert 2000, S. 109)

[28] Stylesheets sind Vorlagen, in denen die Darstellung eines Inhalts festgelegt ist.

[29] Stylesheets für XML können mit der speziell entwickelten *Extensible Style Language* (XLS) erstellt werden.

Erweiterte Links ermöglichen es, Links von Read-only Medien wie beispielsweise einer auf CD-Rom gespeicherten Datei zu anderen Stellen einzurichten. Sie bieten zudem die Möglichkeit, Links zu und von Daten aus zu erstellen, die selbst kein Linking unterstützen (vgl. Behme und Mintert 2000, S. 110). Weiterhin soll XLinking die Möglichkeit bieten, auf mehrere optionale Ziele anstatt ein Zieldokument zu verweisen. XLinking ermöglicht außerdem die Einrichtung von *multidirektionalen* Links, das heißt ein Link führt vom Zieldokument aus wieder zurück. In HTML war dies bisher nur in eine Richtung möglich.

XPointer

XPointer bietet neben den in XLink beschriebenen Verbesserungen noch zusätzlich Möglichkeiten, auf einzelne Dokumentteile zuzugreifen. Mit XPointer kann auf beliebige Elemente und Attribute innerhalb eines Dokuments zugegriffen werden. Zum Beispiel auf bestimmtes Kapitel oder bestimmte Textteile.

Die Vorteile von XML auf einen Blick

Zusammenfassend kann gesagt werden, daß die zu Beginn dieses Abschnitts genannten Anforderungen durch den XML-Standard durchaus erfüllt werden (Mintert 1999):

- Der XML-Standard ist offen zugänglich. Das XML-Datenformat ist unabhängig von bestimmter Software oder einem bestimmten Softwareanbieter.
- Der XML-Standard ist nicht von einem Betriebssystem oder bestimmter Hardware abhängig.
- Der XML-Standards ist aufgrund des Generic Markup unabhängig von einem Ausgabeformat oder einem Ausgabegerät.
- Der XML-Standard ist kein starrer Standard, sondern kann auf die jeweiligen Bedürfnisse zugeschnitten werden *(Enabling Standard)*.

4.5 Das Open Archival Information System (OAIS) Modell

1999 stellte das amerikanische *Consultative Committee for Space Data Systems* (CCSDS) unter Beteiligung bedeutender Forschungseinrichtungen und nationaler Weltraumforschungszentren den Entwurf eines Referenzmodells für ein Open Archival Information System vor[30]. Das OAIS wurde von einer Initiative der ISO entwickelt. Die Koordinati-

[30]Das Dokument kann unter http://www.ccssds.org.documents/pdf/CCSSDS-650.0-B-1.pdf im Internet abgerufen werden.

on wurde von der CCSDS übernommen. Aufgrund der bereits geschilderten Datenverluste bei der NASA ist es nicht verwunderlich, daß die NASA selbst, die Entwicklung dieses Organisationsmodells zur Langzeitarchivierung digitaler Dokumente angeregt hat (Lupprian 2002b).

Das OAIS-Referenzmodell dient lediglich als Grundlage für die Einrichtung von Archiven für elektronischen Ressourcen. Das CCSDS (2002) weist aus diesem Grund in seinem Bericht ausdrücklich darauf hin, daß es sich bei diesem Modell nicht um eine Anleitung für die Einrichtung von Archiven für digitale Ressourcen handelt: *„This reference model does not specify a design or an implementation. Actual implementations may group or break out functionality differently.“* (CCSDS 2002)

Ziel des OAIS-Referenzmodells ist in erster Linie die Schaffung einer einheitlichen Terminologie, sowie die Entwicklung von Konzepten zur Beschreibung und zum Vergleich von Datenmodellen und Archivierungsarchitekturen. Im weiteren geht es darum, die relevanten Einheiten und ihre Abhängigkeiten innerhalb eines Archivierungssystems zu identifizieren, sowie die Schlüsselfunktionen eines Archivierungssystems zu verdeutlichen. Die Vorschläge des OAIS-Referenzmodells können außerdem als Grundlage für Standardisierungsvorhaben dienen (OCLC 2002).

Das OAIS-Referenzmodell wurde im Rahmen einer Empfehlung als ISO-Standard, für die Anwendung über die Weltraumforschung hinaus, einer intensiven Überprüfung unterzogen. Basierend auf den Ergebnissen dieser Untersuchung, wurde 2001 eine überarbeitete Version des Referenzmodells veröffentlicht (OCLC 2002). Das CCDS definiert in seinem Bericht ein OAIS wie folgt:

> „An OAIS is an archive consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a **Designated Community**.“

Abbildung 4.2, übernommen aus: (CCSDS 2002) zeigt die Funktionseinheiten und ihr Zusammenspiel innerhalb des OAIS. Das Referenzmodell enthält eine Reihe von Funktionen eines Archivs, wie *Ingest* (Übernahme), *Archival Storge* (Archivspeicher), *Data Management* (Datenmanagement), *Access* (Nutzung) und *Administration* (Verwaltung) (vgl. Van der Werf 2000; Lupprian 2002b), die nachfolgend erläutert werden.

Das OAIS-Referenzmodell definiert zudem drei Arten von *Information Packages* (IP) das *Submission Information Package* (SIP), das *Archival Information Package* (AIP), und das *Dissemination Information Package* (DIP).

Ein IP kann als Datenpaket verstanden werden, das zwei Arten von Informationen enthält. Informationen über den Inhalt *(Content information)* und Informationen, die zur Beschreibung der Archivierung dienen *(Preservation Description Information)* (PDI). Die Content information und die PDI bilden zusammen den Inhalt eines Datenpaketes der auch als *Packaging Information* bezeichnet wird. Der Nachweis zur Auffindbarkeit diese Datenpaketes ist in der sogenannten *Descriptive Information* enthalten (vgl. Van der Werf 2000, S. 15).

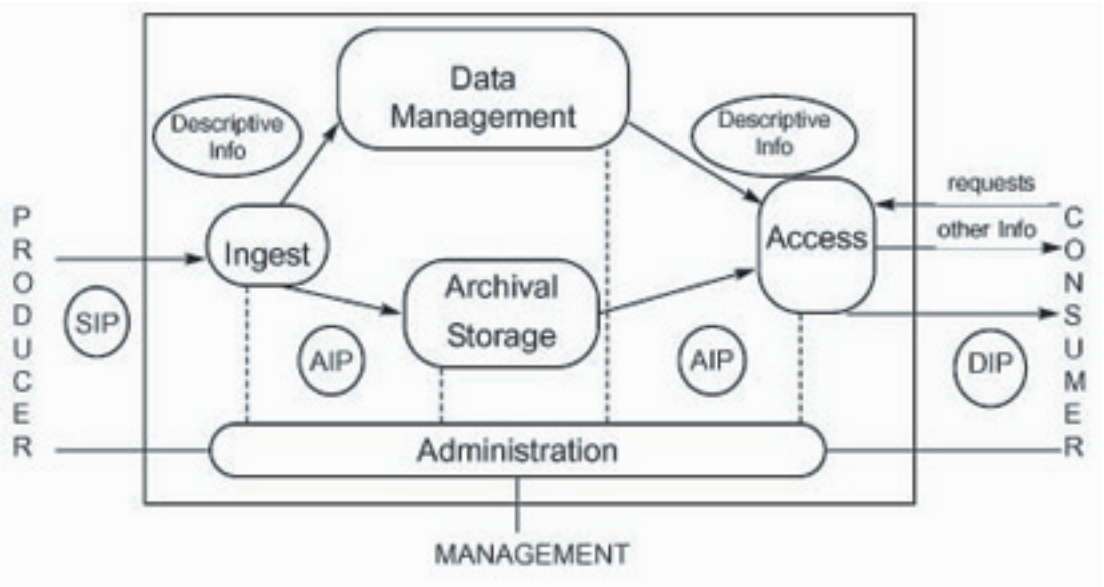

Abb. 4.2: OAIS-Modell

- SIP: enthält Informationen, die vom Produzenten der elektronischen Ressource zur Verfügung gestellt werden sollen.
- AIP: enthält Inforamtionen, die für die Archivierung benötigt werden, wie zum Beispiel die bibliographische Beschreibung der Ressource, spezielle Archivierungsinformationen, die nötige Software zur Darstellung der Ressource.
- DIP: Informationen für die Nutzung und Verteilung. Enthält einen Teil oder alle Teile des AIP.

Übernahme (Ingest): SIPs werden angenommen und geprüft. Aus dem SIP wird ein AIP erzeugt,das den archivinternen Regeln entspricht. Aus dem AIP werden beschreibende Metadaten für die Findmitteldatenbank des Archivs extrahiert. Das AIP wir dann in den Arichspeicher eingestellt, die Datenverwaltung wird benachrichtigt.

Archivspeicher (Archival Storage): Der Archivspeicher dient für die Aufbewahrung und Erhaltung des eingestellten AIP. Hier werden Backups erzeugt, es wird regelmäßig die Integrität der Daten geprüft. Für Notfälle werden Wiederherstellungsmaßnahmen bereitgestellt.

Datenverwaltung (Data Management): Hier efolgt die Verwaltung der beschreibenden Informationen des Archivbestandes sowie der Datenbank. Außerdem werden in diesem Bereich Anfragen von Nutzern entgegengenommen und bearbeitet.

Verwaltung (Administration): Steuerung der Gesamtabläufe im OAIS und seiner Beziehungen nach außen. Hierzu gehört die Konfirguration der Hard- und Software.

Nutzung (Access): Anlaufstelle für Nutzer. Annahme von Anfragen, Überprüfung der Zugriffsrechte. Erzeugung von DIPs und Übergabe an den Nutzer (Lupprian 2002b)

Eine besondere Bedeutung kommt im OAIS-Referenzmodell den Daten und Informationen zu. Das OAIS definiert ein *„... Information Object as Data Object interpreted using its Representation Information“* (Van der Werf 2000, S. 14). Für die erfolgreiche Archivierung eines Datenobjektes ist es für ein OAIS schwierig, das Datenformat eines Datenobjektes zu identifizieren und zu verstehen. Eine weitere Schwierigkeit stellt die dazugehörige *Respresenting Information* dar, die sich in der Software zur Interpretation und Darstellung des Datenobjekts befindet (vgl. Van der Werf 2000, S. 14).

Das Konzept der Unterscheidung von Datenobjekt und der das Datenobjekt repräsentierenden Information, ist ein zentraler Punkt im OAIS-Modell. Die Langzeitarchivierung des Datenformates und der zur Interpretation benötigten Informationen ist von größter Wichtigkeit (vgl. Van der Werf 2000, S. 15). Abbildung 4.3 übernommen aus: (Van der Werf 2000) veranschaulicht den Prozess der Gewinnung von Information aus Daten. Das OAIS-Modell beinhaltet jedoch keinerlei Vorschläge zu technischen Ar-

Abb. 4.3: Informationsgewinnung aus Daten

chivierungsstratregien wie beispielsweise der *Migration* oder der *Emulation* und möglichen Auswirkungen auf das Modell. Es sieht zwar die *transformation* von digitalen Inhalten vor, dies führt jedoch in jedem Fall zu einer neuen Version des Originaldokuments. Als Argument für die in diesem Bereich fehlenden Vorschläge, wird angeführt, daß eine Archivierungeinrichtung keinerlei Kenntnis über Inhalte hat oder benötigt. Es genügt lediglich, die Inhalte der im Archivspeicher gespeicherten AIPs regelmäßig auf den aktuellen Stand zu bringen. Dies geschieht durch einen Mechanismus, der sich in der Verwaltungseinheit befindet. *„However, the Reference Model does not clarify, if and in what way this function belongs to a preservation process.“* (Van der Werf 2000, S. 16)

Das OAIS Modell wurde bereits in die Praxis umgesetzt. Die *National Archive Record Administration* (NARA) in Washington setzt OAIS für die Archivierung digitaler Unterlagen ein (Lupprian 2002b, vgl.). Das Projekt *Networked European Deposit Library* (NEDLIB) hat das OAIS-Modell für den Aufbau eines Depotsystems für digitale Dokumente eingesetzt (Van der Werf 2000; Hedstrom 2002; Lupprian 2002b, vgl.)

4.6 Die Networked European Deposit Library (NEDLIB)-Initiative

Mit der Problematik der Lanzeitarchivierung digitaler Dokumente befassen sich zahlreiche, nationale und Internationale Initiativen und Projekte. Da es im Rahmen dieser Arbeit nicht möglich ist, diese Projekte einzeln vorzustellen, wird an dieser Stelle beispielhaft das Projekt NEDLIB aus den Niederlanden vorgestellt.

In den Niederlanden besteht keine gesetzliche Abgabepflicht für analoge oder elektronische Publikationen. Aus diesem Grund hat die Niederländische Nationalbibliothek *Koninklijke Bibliothek*(KB) eigene Initaitven zur Sammlung und Bewahrung von Publikationen entwickelt. Die Sammelgebiete der KB sind in erster Linie auf Bücher und serielle Veröffentlichungen spezialisiert. Hierzu gehören auch Multimediadokumente jedoch keine Filme, Spiele, Software oder Tonaufnahmen. Ein zukünftiges Sammelgebiet werden jedoch Datenbanken sein, ebenso ist die KB daran interessiert, in Zukunft Teile des Internets zu archivieren (vgl. Beagrie 2002).

Die KB hat ca. 350 Mitarbeiter und ist an das Ministerium für Ausbildung, Kultur und Wissenschaft angeschlossen. Das jährliche Budget der KB betrug im Jahr 2002 36,5 Millionen Euro.

Zu den bisherigen Initiativen der KB gehört die Entwicklung eines *digital archive store projects*. Es handelt sich dabei um ein nationales Abkommen, in dem eine freiwillige Abgabepflicht für Verlage verankert ist. Weitere Initiativen der KB sind beispielsweie eine Studie zur Langzeitarchivierung in Zusammenarbeit mit IBM sowie etliche Digitalisierungsprojekte (vgl. Beagrie 2002).

NEDLIB ist ein dreijähriges Projekt, das im Januar 1998 begann und im Januar 2001 endete. NEDLIB wurde von der Europäischen Kommission unterstützt. Das Projekt wurde ins Leben gerufen, um die technischen und organisatorischen Bereiche, die bei der Entwicklung einer Archivbibliothek für elektronische Publikationen eine Rolle spielen. Die Projektpartener waren acht Nationalbiobliotheken, darunter die Deutsche Bibliothek, ein nationales Archiv, zwei Organsiationen aus dem Bereich der Informationstechnik und drei Verlage. Die Projektleitung hatte die KB (Beagrie 2002, vgl.).

Das Projekt hatte folgende Ergebnisse (vgl. Beagrie 2002):

- Richtlinien für „Best practices“, technische Standards und technische Lösungen, Methoden und Vefahren für die praktische Implementierung.
- Test von Softwareprodukten für den Einsatz in Archivbibliotheken.
- Ein Modell für ein Archivierungssystem das Aufnahme, Speicherung, Zugang und Langzeitarchivierung elektronischer Ressourcen unterstützt.
- Die Erweiterung des OAIS-Standards um eine Funktion für die Planung von Langzeitarchivierungsvorhaben

– Eine Serie mit sieben Reports.

Die Arbeit von NEDLIB wurde für die Implementation eines neuen Archivierungssystems *DNEP* (Desposit of Netherlands Electronic Publications) der KB übernommen.

Eine Erweiterung des OAIS-Modells

Bei der Überprüfung des OAIS-Modells für den Einsatz in NEDLIB wurde deutlich, daß im OAIS-Referenzmodell keine Einheit zur Beschreibung des Archivierungsprozesses vorhanden war. Aus diesem Grund hat NEDLIB das OAIS-Modell um ein eigenes Modul, das *Deposit System for Electronic Publications* (DSEP) zur Steuerung des Ablaufs von Archivierungsprozessen erweitert. Abbildung 5.1 übernommen aus: (Van der Werf 2000) zeigt die Funktionseinheiten des OAIS-Modells erweitert mit dem DSEP Workflow. Der DSEP Workflow umfaßt die Bereiche 3, 4, 5, 6, 7, 8, 9, 11 (vgl. Van der Werf 2000, S. 17).

Der DSEP Workflow[31]

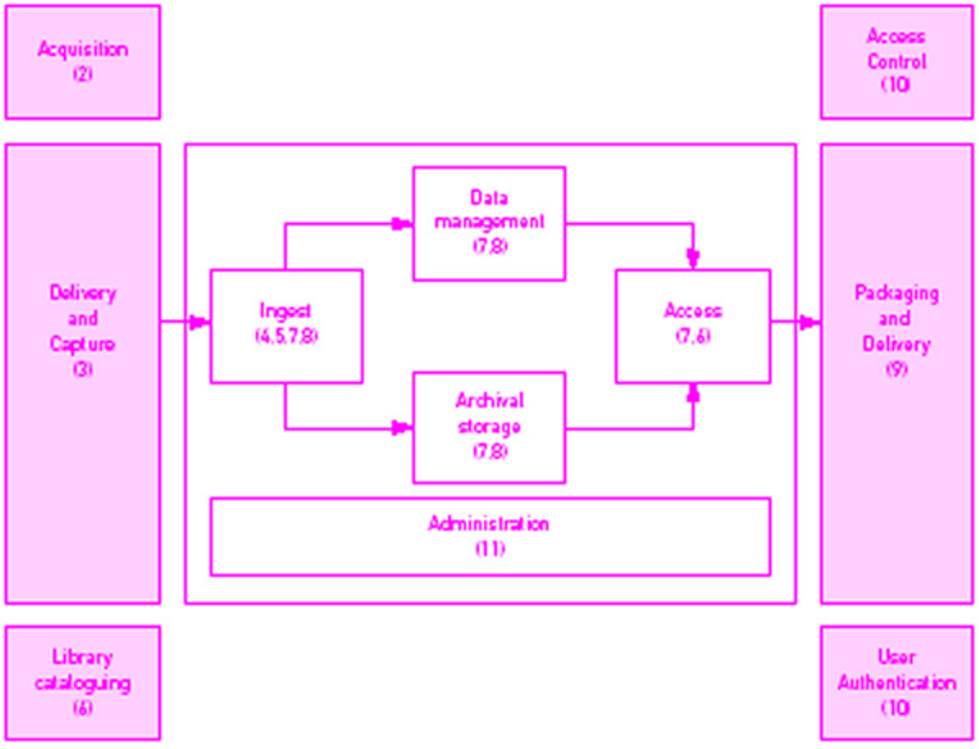

Abb. 4.4: OAIS Funktionseinheiten mit DSEP Erweiterung

[31]Die Angaben wurden übernommen aus: Van der Werf (2000)

2 Acquisition

3 Delievery and Capture: Beschreibt den elektronischen Transfer einer Kopie der elektronischen Publikation von dem Liefersystem des Verlags zur Bibliothek.

4 Registration: Die neu angekommene elektronische Publikation wird im Archivsystem überprüft.

5 Verification: Kontrollroutinen werden bei der neuen elektronischen Publikation durchgeführt um die physikalische und die logische Integrität zu bestimmen.

6 Library cataloging

7 Storage: Die neue Publikation ist im Archivsystem gespeichert. Die Speicherung stellt sicher, daß der Datenstrom der Publikation maschinenlesbar bleibt.

8 Preservation: Archivierungsstrategien zur Sicherung der Verfügbarkeit der Publikation werden durchgeführt.

9 Delivery: Die Publikation wurde aus dem Archivsystem angefordert und wird für den Nutzer bereitgetellt.

10 Access:

11 Monitoring: Der gesamte Workflow für den Umgang mit elektronischen Publikationen wird beobachtet und einer Qualitätskontrolle unterzogen.

5 Technische Konzepte

Für die Langzeitarchivierung digitaler Dokumente sind neben den organisatorischen Maßnahmen auch technische Maßnahmen zur Erhaltung des Zugriffs auf die elektronischen Ressourcen von ebenso großer Bedeutung. Hierfür existieren unteschiedliche Lösungsansätze, die kontrovers diskutiert werden. Die zur Zeit am meisten diskutierten technischen Archivierungsstrategien sind die *Migration* und die *Emulation*, die in diesem Kapitel vorgestellt werden sollen.

5.1 Technology watch

Unabhängig von der angewendeten technischen Archivierungsstrategie, muß eine Einrichtung, die elektronische Ressourcen besitzt und archiviert, permanent *technology watch* betreiben. Technology watch bedeutet, die aktuellen Entwicklungen im Bereich der Informationstechnik zu beobachten, und im Archivierungsprozess zu berücksichtigen. Durch technology watch soll vermieden werden, daß das Verschwinden oder die Neuentwicklung von Datenformaten, Software und Hardware unbemerkt bleibt, und rechtzeitige Maßnahmen zur Erhaltung der davon betroffenen, digitalen Bestände nicht mehr getroffen werden können.

Neben den routinemäßigen Erhaltungsmaßnahmen bei Datenträgern, wie beispielsweise dem regelmäßigen Umkopieren von Magnetbändern, sind technische Archivierungsmaßnahmen immer eine Reaktion auf den technischen Wandel. Die Ergebnisse des technology watch, stehen daher in engem Zusammenhang mit der Auswahl der entsprechenden technischen Archivierungsstrategie. Wobei noch anzumerken ist, daß es sich auch bei der Beobachtung der Entwicklung der Archivierungsstrategien selbst um technology watch handelt.

5.2 Refreshing

Das bereits erwähnte, regelmäßige Umkopieren wird in der Archivierung auch als *Refreshing* bezeichnet. Bei refreshing handelt es sich um eine der einfachsten technischen Archivierungsstrategien. Refreshing bedeutet, *„... to copy digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit-stream.“* (Borbinha u. a. 2000)

Insbesondere bei der Archivierung von Magnetbändern spielt das Refreshing eine große Rolle. Mittlerweile exisiteiren hierfür automatisierte Vefahren. So wird das re-

gelmäßige Refreshing der über 6000 Datenbänder des Deutschen Klimarechenzentrums Tag und Nacht von zwei Robotern durchgeführt, *„... im 15. Stock eines Hochhauses der Universität huschen sie mit ihren Metallarmen lautlos durch die Regale des zentralen Magnetbandarchivs, richten ihre roten Laseraugen auf Datenbänder, greifen Cassetten heraus und schieben sie in Laufwerke.“* (Schmundt 2000).

Das Refreshing bedeutet aber auch, daß die entsprechenden Lesegeräte mit der dazugehörigen Software und der Software für das Refreshing aufbewahrt werden müssen. Dies hätte die Entstehung eines sogannten *Computermuseums* zur Folge. Besonders bei Beständen mit sehr heterogenen Datenträgern ist diese Möglichkeit der Archivierung über kurz oder lang nicht aufrechtzuerhalten.

5.3 Migration

Migration ist die am häufigsten verbreitete Archivierungsstrategie. Sie wird definiert als:

> „Set of organised tasks designed to achieve the periodic transfer of digital materials from one hardware or software configuration to another, or from one generation of computer technology to a subsequent generation.“ (Borbinha u. a. 2000)

Obwohl die Migration als technische Archivierungsstrategie häufig angewendet wird, gibt es auch einige Kritikpunkte, die gegen dieses Verfahren sprechen. Hierzu gehört, daß es sich bei einer Migration in der Regel um eine Veränderung des originalen Datenstroms handelt. Bereits bei einem einfachen Kopiervorgang oder dem erstellen einer exakten Dublette kann es passieren, daß der Datenstrom verändert wird. Dies kann durch fehlerhafte Software, falsche Interpretation der Daten oder durch mechanische Probleme geschehen.

Ein Beispiel hierfür ist eine einfache Konvertierung eines Textdokuments im Wordformat in eine neue Version oder ein anderes Format. Fast immer ist diese Migration mit einem Verlust der Formatierung oder gar wichtiger Inhalte verbunden.

Die Folgen, die eine Veränderung des Datenstroms mit sich bringt, können unterschiedlichster Art sein. Im schlimmsten Fall kann eine Ressource nach einer Migration nicht mehr verwendet werden. Inhalte können durch Reformatierung aber auch ihre urprüngliche Bedeutung verlieren. Abbildung 5.1 übernommen aus: Rothenberg (2001) macht dies an einem Beispiel deutlich. Da die Inhalte und die Strukturen elektronischer Ressourcen immer komplexer werden, ist eine einfache Übertragung des Datenstroms auf ein neues Medium ohne Verluste nur noch bei einem geringen Prozentsatz der Ressourcen möglich (vgl. Hedstrom 2002).

Darüber hinaus wird die Migration in Bezug auf die Menge der zu archivierenden Ressourcen, sowie die dafür anfallenden Kosten, als geeignete Strategie in Frage gestellt. Die Migration ist kein einmaliger Prozess. Für jedes neue Datenformat und jedes

Abb. 5.1: Reformatierung kann die Bedeutung von Inhalten zerstören

Dokument, daß in eine neue Form konvertiert wird, muß eine neue Lösung gefunden werden (vgl. Hedstrom 2002; Rothenberg 2001).

Mit der steigenden Anzahl neuer Formate, Versionen und Betriebssystemen, werden Migrationen in immer kürzeren Abständen notwendig, um die Zugänglichkeit digitaler Daten gewährleisten zu können. Um diese anfallenden Migrationsvorgäne bewältigen zu können, ist die Einführung von Standards notwendig. Denn ohne Standards *„...each migration requires a customized approach that involves an analysis of the source file format, a selection of a target file format, and a conversion usin either off-the-shelf products or programs written specifically for the conversion.“* (Hedstrom 2002).

Auch wenn mit der Abwärtskompatibilität zunächst einige Probleme aufgefangen werden können, gibt es trotz allem keine Möglichkeit, Migration als einheitlich anwendbare Archivierungsstrategie zu implementieren. Sobald auch noch extra geschriebene Programme für eine Migration benötigt werden, ist außerdem abzuwägen, ob die dafür anfallenden Kosten in Relation zu der erhaltenden Information stehen (vgl. Hedstrom 2002).

Als problematisch hat sich die Migration bei großen Datenmengen, wie beispielsweise umfangreichen Datenbanken erwiesen, da die Transferrate bei Migrationen nicht in gleichem Maß wie die verfügbaren Speicherkapazitäten wächst uund somit immer hinterherhinkt. Selbst bei gut geplanten Migrationen, die durchgeführt werden, sobald ein neues Speichermedium eingeführt ist, ist nicht genug Zeit, alle Daten in das Zielsystem

zu migrieren, bevor dieses selbst wieder veraltet ist (vgl. Hedstrom 2002)[1].

5.4 Emulation

Das Verfahren der Emulation wird in der Informationstechnik bereits seit längerer Zeit für die Simulation von Softwareumgebungen auf dafür ursprünglich nicht vorgesehenen Rechnern. So kann zum Beispiel die Umgebung eines ATARI Rechners aus den 1980er Jahren auf einem heutigen PC simuliert und mitsamt der dazugehörigen Software angewendet werden.

JEFF ROTHENBERG hat dieses Vefahren auf die Langzeitarchivierung übertragen (vgl. Rothenberg 1995, 2001). Obwohl es sich bei der Emulation bereits um eine gut erprobte Technik handelt, wird sie in Archivierungseinrichtungen noch nicht standardmäßig eingesetzt. Rothenberg (2001)[2] Emulation kann nach Borbinha u. a. (2000) definiert werden als:

> „The imitation of a computer system, performed by a combination of hardware and software, that allows programs to run between incomatible systems."

Die Emulation basiert auf der Annahme, daß digitale Dokumente eng an die Software gebunden sind, die zu ihrer Erstellung, Bearbeitung und Betrachtung verwendet wurde. Um auf archivierte, digitale Dokumente zugreifen zu können, wäre somit das Vorhandensein dieser dazugehörigen Software die bestmögliche Lösung. Im Gegensatz zur Migration, könnte der Datenstrom des Dokuments unverändert bleiben. Dadurch können auch die bereits angesprochenen Risiken der Umformatierung umgangen werden. Um die benötigte Software auf einem Rechner laufen lassen zu können, muß das dazugehörige Betriebssystem auf dem aktuellen Rechner emuliert werden.

Damit die Emulation in der Praxis einsetzbar ist, müssen folgende Voraussetzungen erfüllt sein (vgl. Bódi 2000):

- Um auf neueren Rechnern mit mögichst geringem Aufwand einen Emulator herstellen zu können, müssen Methoden entwickelt werden, mit denen eine Spezifikation der Emulatoren möglich ist.
- Da sich zukünftige Anwender oftmals in den veralteten Betriebssystemen nicht mehr zurechtfinden, müssen einfache, für den Menschen lesbare Metadaten zur Verfügung gestellt werden, die Informationen zur Suche, zum Zugriff und zur Wiederherstellung der Daten enthalten.

[1] Dies ist bei Datenmengen, deren Größenordnung sich im mehrstelligen Terabyte bewegt, der Fall.
[2] Die Emulation wurde im Rahmen des NEDLIB Projekts getestet, die Ergebnisse liegen unter: http://www.kb.nl vor.

– Entwicklung einer sogenannten Datenkapsel, in der das archivierte Dokument, die dazugehörige Software, das Betriebssystem und die Spezifikationen für die Emulation enthalten sind.

Die Datenkapsel

Wie bereits kurz beschrieben wurde, müssen in der Datenkapsel drei wesentliche Bestandteile für die Emulation und den Zugriff auf das archivierte Dokument enthalten sein (vgl. Bódi 2000):

– Das digitale Dokument und die zugehörige Software/Betriebssystem

– Die Emulatorspezifikation

– Die entsprechenden Metadaten

Bei dem digitalen Dokument ist darauf zu achten, daß alle dazugehörigen Dateien mit in die Datenkapsel übernommen werden müssen. Ebenso muß die komplette Software mit Betriebssystem und möglicherweise benötigten Softwarebibliotheken vorhanden sein. Wichtig ist hierbei, daß all diese Daten absolut identisch mit den ursrpünglich vorhandenen Daten sind.

Genauso verhält es sich mit der Emulatorspezifikation. Sie sollte alle notwendigen Inforamtionen enthalten, die zur genauen Nachbildung der benötigten Rechnerarchitektur notwendig sind. Zwar muß jeder Emulator spezifisch für eine Rechnerarchitektur entwickelt werden, kann dann aber beliebig oft für die Emulation dieser Rechnerarchitektur verwendet und verbreitet werden. Hierin liegt ein bedeutender Vorteil gegenüber der Migration. Es müssen nicht mehr Konvertierungsprogramme für jedes einzelne Datenformat entwickelt werden (vgl. Bódi 2000).

Bei den in der Datenkapsel beigefügten Metadaten handelt es sich hauptsächlich um Angaben für den Benutzer. Am wichtigsten sind hier die Inforamtionen, wie die erforderlichen Daten aus der Datenkapsel extrahiert werden können. Weiterhin werden Metadaten benötigt, die Angaben zum Dokument, zur Dokumentation der emulierten Hardware und zu der in der Datenkapsel enthaltenen Software enthalten (vgl. Bódi 2000).

Gegenüber der Migration bietet die Emulation viele Vorteile. So wird zum Beispiel nur eine Kopie des Originaldokuments für die Weiterverarbeitung benötigt. Das heißt, im Gegensatz zur Migration existiert weiterhin ein Exemplar des Originaldokuments. Ein weiterer Vorteil ist in den niedrigeren Kosten gegenüber der Migration zu sehen. Dies hängt mit dem bereits angesprochenen geringeren Aufwand zusammen. Ist ein Emulator einmal entwickelt, entsteht in der Regel kein weiterer Entwicklungsaufwand mehr. Als Vorteil der Emulation wird auch die Erzeugung des sogenanten *„look-and-feel“* genannt. Die Vermittlung des Gefühls, daß auch bei der Benutzung des Originals auf einem Original dafür vorgesehenen Rechner vorhanden wäre.

Die Entwicklung eines Emulators kann aber auch als Nachteil der Emulation betrachtet werden. Es handelt sich hierbei um eine nicht unbedingt einfach zu lösende Aufgabe. Außerdem ist fraglich, ob Emulatoren tatsächlich alle formalen Spezifikationen und Feinheiten, die ursprünglich vorhanden waren tatsächlich erfassen können.

Hedstrom (2002) beschreibt einen Versuch, bei dem Testpersonen zum Vergleich eine migrierte und eine emulierte Version eines Computerspiels, sowie die Originalversion in der Originalumgebung vorgelegt wurde. Die Testpersonen beurteilten die migrierte Version als die dem Original am nächsten kommende. Da für die Migration der original code komplett neu geschrieben wurde, um auf einer aktuellen Plattform ausgeführt werden zu können, leistete er mehr als eine schlechte Emulation (vgl. Hedstrom 2002).

Abbildung 5.2 übernommen aus: (Rothenberg 2000) veranschaulicht den Emulationsprozess.

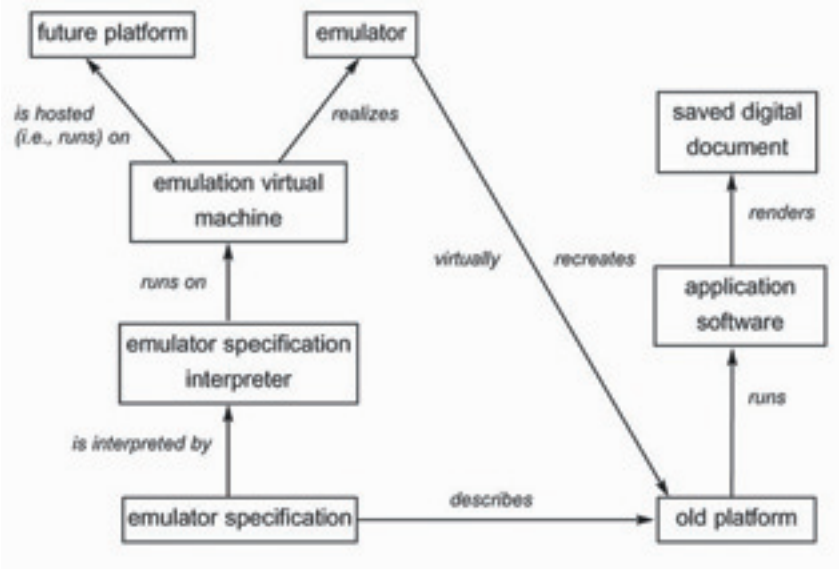

Abb. 5.2: Funktionsweise der Emulation

Zusammenfassend kann festgestellt werden, daß eine endeutig geeignete technische Archivierungsstrategie auch mit der Emulation noch nich texistiert. Für unterschiedliche Ansprüche in der Archivierung digitaler Dokumente muss auf die unterschiedlichen, dafür zur Verfügung stehenden Strategien zurückgegriffen werden. In Zukunft werden mit Sicherheit noch mehr technische Verfahren für die Langzeitarchivierung entstehen, ein neuer Ansatz ist die *UVC-basierte Emulation*[3], auf die in dieser Arbeit jedoch nicht eingegangen wird.

[3]UVC = Universal Virtual Computer, für Informationen hierzu siehe Lorie (2001).

6 Zusammenfassung und Ausblick

Zusammenfassend läßt sich feststellen, daß die Schwierigkeiten bei der Langzeitarchivierung nicht weniger werden. Ein Hauptgrund hierfür liegt sicher in der Doppelrolle, der technischen Entwicklung. Einerseits stehen mittlerweile jedem Anwender technische Möglichkeiten wie nie zuvor zur Verfügung, andererseits kann die Archivierung darauf immer nur mit teilweise großer Zeitverzögerung reagieren. Wie bereits in einigen Beispielen deutlich wurde, ist deshalb die organisatorische Struktur eines Archivs gleichbedeutend mit der technischen Struktur. Die beste technische Ausstattung nützt nichts, wenn sich Einrichtungen nicht im klaren darüber sind, was wie und zu welchem Zweck archiviert werden soll. Daß hinsichtlich der wachsenden Menge Eingrenzungen vorgenommen werden müssen, versteht sich von selbst. Ebenso wie die Tatsache, daß diese Aufgaben nicht mehr von einzelnen Einrichtungen bewältigt werden können.

Somit wird auch klar, daß in vielen Bereichen eine zunehmende Standardisierung und maschinelle Verarbeitung einkehren wird. So wird am Beispiel der Erschließung deutlich, daß es sich oftmals nicht lohnt, alle Informationen gleich aufwendig zu erschließen, sondern daß mit DC und XML Möglichkeiten zur Verfügung stehen, diesen Vorgang weniger zeit- und arbeitsaufwendig zu gestalten. Interessant ist hierbei, daß viele, der im bibliothekarischen Bereich angewendeten Vefahren aus der Informationstechnik kommen. Das sich der Einfluß der Informationstechnik auch in anderen Bereichen noch weiter verstärken wird, wurde am Beispiel des Literaturarchivs deutlich.

Wie die Zukunft der Archivierung aussehen wird, bestimmt die technische Entwicklung. Klar ist jedoch, daß sich mit der Entstehung der digitalen Dokumente auch die Bestände in Bibliotheken und Archiven ständig verändern. Dabei nicht den Überblick zu verlieren und trotz des rasanten Tempos für langfristige Verfügbarkeit der Materialien zu sorgen, wird wohl in Zukunft die Hauptbeschäftigung sammelnder Einrichtungen sein.

Literaturverzeichnis

Archimedes 1999 : *Wir verlieren unser Gedächtnis.* Mai 1999. – URL `http://www.arte-tv-com/hebdo/archimed/19990504/dtext/sujet.html`. – Zugriffsdatum: 2003-06-18

Bódi 2000 Bódi, Dominik: *Emulation als Lnagzeitarchivierung digitaler Dokumente.* Mai 2000. – URL `http//ist.unibw-muenchen.de/lectures/FT2000/Digitale-Bibliotheken/hando%ut5.pdf`. – Zugriffsdatum: 2003-06-30

Beagrie 2002 Beagrie, Neil: National Digital Preservation Initiatives: An overview of Developments in Australia, France the Netherlands, and the United Kingdom and Related International Activity ; Appendix : 5. In: *Preserving our digital heritage: Plan for the national digital information infrastructure and preservation program.* Washington, DC : Council on Library and Information Resources ; Library of Congress, 2002. – URL `http://www.digitalpreservation.gov/ndiipp/repor/ndiipp_appendix.pdf`. – Zugriffsdatum: 2003-06-23

Behme und Mintert 2000 Behme, Henning ; Mintert, Stefan: *XML in der Praxis: Professionelles Web-Publishing mit der Extensible Markup Langugage.* 2. erweiterte Auflage. München: Addison-Wesley, 2000. – ISBN 3-8273-1636-7

Bernard 2003 Bernard, Andreas: Der unsichtbare Nachlaß. In: *Süddeutsche Zeitung* (2003), Mai, Nr. 120, S. 16

Betts und Schmidt 1999 Betts, Mitch ; Schmidt, Corinne: *Langfristig gehen Daten verloren.* November 1999. – URL `http://212.98.53.238/domino/CWArchiv.nsf`. – Zugriffsdatum: 2003-04-21

Bide 2000 Bide, Mark ; Associates: *Standards for electronic Publishing: An overview.* URL `http://www.kb.nl/coop/nedlib/results/e-publishingstandards.pdf`. – Zugriffsdatum: 2003-06-23, August 2000 (Report ; 3). – ISBN 906259147-7

Borbinha u. a. 2000 Borbinha, José L. ; Cardoso, Fernando ; Freire, Nuno: *NEDLIB Glossary.* Februar 2000. – URL `http://www.kb.nl/coop/nedlib/glossary.pdf`

Borbinha und Freire 2001 Borbinha, Josè L. ; Freire, Nuno: Deposit Collections of Digital Thesis and Dissertations. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), Nr. 3–4, S. 217–224

Brugger 1998 Brugger: *Eine statistische Methode zur Erkennung von Dokumentstrukturen*. 1998. – URL `http://www-iiuf.unifr.ch/~brugger/papers/da/da-html.html`. – Zugriffsdatum: 2003-05-13. – Dissertation

BSB 2002 Bayerische Staatsbibliothek: *Erschließung elektronischer Ressourcen: RAK-NBM oder Dublin Core?* 2002

Byrum 2002 Byrum, John D.: *Herausforderungen bei der Verzeichung von Netzpublikationen in Nationalbibliografien: Ein Überblick über Aufgaben und Lösungskonzepte*. August 2002. – URL `http://www.ifla.org/IV/ifla68/papers/117-152g.pdf`. – Zugriffsdatum: 2003-06-23. – Bibliography - Workshop im Rahmen der 68th IFLA Coucil and General Conference in Glasgow

Calanag u. a. 2001 Calanag, Maria L. ; Sugimoto, Shigeo ; Tabata, Ioichi: Digital Preservation: Some Policy and Legal Issues. In: *Digital Libraries* (2001), März, Nr. 20. – URL `http://www.dl.uli.ac.jp/Djournal/No_20/5-calanag.html`. – Zugriffsdatum: 2003-06-23. – ISSN 1345-9198

CCSDS 2002 Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System: recommendation for space data system standards / CCSDS. Washington D.C., Januar 2002 (CCSDS 650.0-B-1). – Technischer Bericht. – URL `http://www.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf`. – Zugriffsdatum: 2003-05-13. Blue-Book-Standard-Vorschlag

CDL 2001 California Digital Library: *Digital Image Format Standards*. 2001. – URL `http://www.cdlib.org/libstaff/technology/tas/Standards`. – Zugriffsdatum: 2003-06-28

Clavel-Merrin 2000 Clavel-Merrin, Genevieve: *The NEDLIB List of Terms*. URL `http://www.kb.nl/coop/nedlib/results/NEDLIBterms.pdf`. – Zugriffsdatum: 2003-06-23, December 2000 (Report ; 7). – ISBN 906259151-5

Clavel-Merrin 2001 Clavel-Merrin, Genevieve: Initatives in the Field of Long-Term Digital Preservation and the Need for a Continued Research Effort. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), Nr. 3-5, S. 184–187

Day 2001a Day, Michael: *Metadata for digital preservation: a review of recent developments*. Paper. September 2001. – URL `http://www.ukoln.ac.uk/metadata/presentations/ecdl2001-day/paper.html`. – Zugriffsdatum: 2003-05-20. – Das

Papier wurde im Rahmen eines Vortrags bei der ECDL 2001, 5th European Conference on Research and Advanced Technology for Digital Libraries in Darmstadt zur Verfügung gestellt.

Day 2001b Day, Michael: *Metadata in a nutshell*. April 2001. – URL `http://www.ukoln.ac.uk/metadata/publication/nutshell/`. – Überarbeitete Fassung vom 2001-08-21

DCMI 1995 Dublin Core Metadata Initiative: *History of the Dublin core Metadata Initiative*. 1995. – URL `http://dublincore.org/about/history`. – Zugriffsdatum: 2003-06-18

DCMI 2000 Dublin Core Metadata Initiative: *Dublin Core Qualifiers*. Juli 2000. – URL `http://dublincore.org/documents/2000/07/11dcmes-qualifiers/`. – Zugriffsdatum: 2003-06-24

DCMI 2001 Hillmann, Diane: *Using Dublin Core*. April 2001. – URL `http://dublincore.org/documents/usageguide/`

Dickschus und Brebeck 2000 Dickschus, Arthur ; Brebeck, Jürgen: *PC-Wissen: Die ganze Welt der Hard- und Software*. 4. Auflage. München: Heyne, 2000. – ISBN 3-453-16390-7

Dobratz u. a. 2001 Dobratz, Susanne ; Liegmann, Hans ; Tappenbeck, Inka: Langzeitarchivierung digitaler Dokumente. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), November/Dezember, Nr. 6, S. 327–332

Dobratz und Tappenbeck 2002 Dobratz, Susanne ; Tappenbeck, Inka: Thesen zur Zukunft der digitalen Langzeitarchivierung in Deutschland. In: *Bibliothek Forschung und Praxis* Jg. 26 (2002), Nr. 3, S. 257–261

Endres und Fellner 2000 Endres, Albert ; Fellner, Dieter W.: *Digitale Bibliotheken: Informatik Lösungen für globale Wissensmärkte*. Heidelberg: dpunkt, 2000. – ISBN 3-932588-77-0

Engster 2002 Engster, Florian: *Digitale Bibliothek: Konzeption und Implementierung mit der Greenstone Digital Library Software*, Fachhochschule Stuttgart Hochschule der Medien, Diplomarbeit, 2002

Eversberg 1999 Eversberg, Bernhard: *Was sind und was sollen Bibliothekarische Datenformate: Kapitel 10.7: Dublin Core*. Januar 1999. – URL `http://www.allegro-c.de/allegro/formate/kap107.htm`. – Zugriffsdatum: 2003-06-06

Gilliland-Swetland 2000 Gilliland-Swetland, Anne J.: *Introduction to Metadata: Setting the Stage*. Mai 2000. – URL `http://www.getty.edu/research/institute/standards/intrometadata/`. – Zugriffsdatum: 2003-06-03

Grabmeyer und Schimmer 2000 Grabmeyer, Felix ; Schimmer, Markus: *Datenbezogene Standards und Normen im Internet*. Dezember 2000. – URL `http://www.iivs.de/haag/buerger/fgrabm/studium/Datenformate.pdf`. – Zugriffsdatum: 2003-05-05. – Die Arbeit entstand im Rahmen des Seminars Electronic Business am Instiut für Wirtschaftinformatik der Universität Regensburg

Grote 2000 Grote, Andreas: Verflüchtigt: Der Zahn der Zeit nagt an digitalen Daten. In: *c't magazin für computer technik* (2000), Nr. 24, S. 114–118

Gschwind u. a. 2000 Gschwind, Rudolf ; Rosenthaler, Lukas ; Frey, Franziska: Neue Technologien und Kulturgüter / Abteilung für wissenschaftliche Photographie an der Universität Basel. Basel, Mai 2000. – Konzept. – URL `http://www.kulturgueterschutz.ch/Websitealt/dt/Publikationen/neuetechnologien.pdf`. – Zugriffsdatum: 2003-06-23. Das Konzept wurde erstellt für das Bundesamt für Zivilschutz Sektion Kulturgüterschutz

Harloff 2001 Harloff, Jan: *Grundlagen der Retrodigitalisierung von Texten und Bildern*. Referat im Rahmen der Lehrveranstaltung Wissenschafltiches Bibliothekswesen in Deutschland an der Bibliotheksschule in Frankfurt am Main am 16. 02. 2001. 2001. – URL `staff.ub.tu-berlin.de/~harloff/libint/retrodig.pdf`. – Zugriffsdatum: 2002-06-04

HeBIS 2002a Zell, Stefan: Meet the challenge: Herausforderung elektronische Ressourcen. In: *HeBIS cocktail: Neues und interessantes frisch gemixt* (2002), Nr. 1. – URL `http://www.hebis.de/hebiscocktail/2002_1/challenge02-02.pdf`. – Zugriffsdatum: 2003-06-23

HeBIS 2002b Albrecht, Rita ; Nienerza, Heike: *Roadshow Online-Ressourcen: DC, DOI, OAI, URN und andere unbekannte Größen*. September 2002. – URL `http://www.hebis.de/arbeitshilfen/schulungsmaterialien/roadshow/rs2_er-entwicklungen.pdf`. – Zugriffsdatum: 2003-06-28. – Veröffentlichung der Verbundzentrale des Hessischen Bibliotheksinformationssystems

Hedstrom 1995 Hedstrom, Margaret: *Digital preservation: a time bomb for Digital Libraries*. 1995. – URL `http://www.uky.edu/~kiernan/DL/hedstrom.html`. – Zugriffsdatum: 2003-06-23

Hedstrom 2002 Hedstrom, Margaret: *Digital preservation: Problems and Prospects*. 2002. – URL `http://www.dl.ulis.ac.jp/DLjournal/No_20/1-hedstrom/2-hedstrom.html`. – Zugriffsdatum: 2003-06-27

Hedstrom und Montgomery 1999 Hedstrom, Margaret ; Montgomery, Sheon: *Digital Preservation Needs and Requirements in RLG Member Institutions*. California: Research Libraries Group, 1999. – URL `www.rlg.org/preserv/digpres.pdf`. – Zugriffsdatum: 2003-06-23

Hodge 2000 Hodge, Gail M.: Best Practices for Digital Archiving: An Information Life Cycle Approach. In: *D-Lib Magazine* Jg. 6 (2000), Januar, Nr. 1. – URL `http://www.dlib.org/dlib/january00/01hodge.html`. – Zugriffsdatum: 2003-05-27. – ISSN 1082-9873

Huc 2001 Huc, Claude: Long Term Preservation of Digital Information in the Space Field: From the OAIS Reference Model to Practical Applications. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), Nr. 3-4, S. 188–193

IFLA 2003 International Federation of Library Associations and Institutions: *ISBD(ER) International Standard Bibliographic Description for Electronic Resources.* Januar 2003. – URL `http://www.ifla.org/VII/s13/pubs/isbder-1102.htm`. – Zugriffsdatum: 2003-06-05

ISO 2003 International Organization for Standardization: *Information and documentation: The Dublin Core metadata element set.* Februar 2003. – URL `http://www.niso.org/international/SC4/n5151.pdf`. – Zugriffsdatum: 2003-06-23

jth 2003 jth: Meilensteine: Diskette. In: *Süddeutsche Zeitung* (2003), Maerz, Nr. 61, S. 23

Keitel 2002 Keitel, Christian: *Die Archivierung elektronischer Unterlagen in der baden-württembergischen Archivverwaltung: Eine Konzeption.* Juni 2002. – URL `http://www.lad-bw.de/lad/konzeption.pdf`. – Zugriffsdatum: 2003-06-02. – Die Konzeption wurde im Rahmen eines Assesorenprojektes der baden-württembergischen Archivverwaltung erstellt

KfR 2000 Konferenz für Regelwerksfragen: Arbeitsgruppe Codes: *Arbeitsergebnisse der AG Codes.* Juli 2000. – URL `http://www.ddb.de/professionell/pdf/codc_arb_erg.pdf`. – Zugriffsdatum: 2003-06-23

Kodak 2001 Kodak: *Digital Preservation: Sicherung der digitalen Welt.* Mai 2001. – URL `http://wwwde.kodak.com/DE/de/corp/pressReleases/private/index.shtml`. – Zugriffsdatum: 2003-04-13. – Der Zugriff erfolgt nach den Angaben auf der Webseite mit: UserID:Presse Paßwort:George

Landwehr 2001 Landwehr, Dominik: Haltbarkeit von Speichermedien: Kein Backup für die Ewigkeit. In: *Infoweek* (2001), April, Nr. 14. – URL `http://www.peshawar.ch/tech/infoweek-storage-april2001.htm`. – Zugriffsdatum: 2003-06-06

Leskien 2000 Leskien, H.: Retrodigitalisierung: Eine Zwischenbilanz. In: *BFB - Bibliotheksforum Bayern* Jg. 28 (2000), Nr. 2, S. 132–153

Liegmann 2002 Liegmann, Hans: *Langzeitverfügbarkeit digitaler Publikationen.* 2002. – URL `http://www.uni-muenster.de/Forum-Bestandserhaltung/konversion/dgi-liegmann.shtml`. – Zugriffsdatum: 2003-03-10

Lorie 2001 Lorie, Raymond A.: Preserving Digital Information: An Alternative to Full Emulation. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), Nr. 3-4, S. 205–209

Lupovici und Masanés 2000 Lupovici, Catherine ; Masanés, Julien: *Metadata for long term-preservation.* URL `http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf`. – Zugriffsdatum: 2003-06-23, Juli 2000 (Report ; 2). – ISBN 90-62-59-1469

Lupovici und Masanés 2001 Lupovici, Catherine ; Masanés, Julien: Preservation Metadata - the NEDLIB's Proposal Bibliothéque Nationale de France. In: *Zeitschrift für Bibliothekswesen und Bibliographie* Jg. 48 (2001), Nr. 3–4, S. 194–199

Lupprian 2002a Lupprian, Karl-Ernst: *Ein Archiv für 1000 Jahre.* 2002. – URL `http://www.museumtheuern.de/edvtage`. – Zugriffsdatum: 2003-06-23

Lupprian 2002b Lupprian, Karl-Ernst: *Open Archival System (OAIS): Ein künftiger Standard für elektronische Archive?* 2002

Lyman 2002 Lyman, Peter: Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving: Archiving the World Wide Web ; Appendix : 2. In: *Preserving our digital heritage: Plan for the national digital information infrastructure and preservation program.* Washington, DC : Council on Library and Information Resources ; Library of Congress, April 2002, S. 53–66. – URL `http://www.digitalpreservation.gov/ndiipp/repor/ndiipp_appendix.pdf`. – Zugriffsdatum: 2003-06-28. – ISBN 1-887334-91-2

McLean und Davis 1999 McLean, Margaret (Hrsg.) ; Davis, Ben H. (Hrsg.): *Time and Bits: Managing Digital Continuity.* Getty Research Institute, 1999. – ISBN 0-89236-583-8

Meyers 2001 Meyers Lexikonredaktion (Hrsg.): *Meyers großes Taschenlexikon*, Mannheim, Leipzig, Wien: B.I. Taschenbuchverlag, 2001 (Meyers großes Taschenlexikon Bd. 10 Hin-Jah)

Mintert 1999 Mintert, Stefan: *Einführung in die Extensible Markup Language (XML).* 1999. – URL `http://www.mintert.com/xml/xmlintro.html`. – Zugriffsdatum: 2003-06-23

NDIIPP 2002 : *Preserving our digital heritage: Plan for the national digital information infrastructure and preservation program.* Washington, DC : Council on

Library and Information Resources ; Library of Congress, Oktober 2002. – URL http://www.digitalpreservation.gov/ndiipp/repor/ndiipp_plan.pdf. – Zugriffsdatum: 2003-06-28

Norsam 2001 Norsam Technologies: *HD-Rosetta Archival Preservation Services*. 2001. – URL http://www.norsam.com/hdrosetta.htm. – Zugriffsdatum: 06-06-2003

OCLC 2002 OCLC/RLG Working Group on Preservation Metadata: *Preservation Metadata and the OASIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*. Juni 2002. – URL http://www.oclc.org/research/pmwg. – Zugriffsdatum: 2003-05-13

OII 1999 Open Information Interchange: *OII Guide to Metadata*. Dezember 1999. – URL http://www.diffuse.org/oii/en/metadata.html. – Zugriffsdatum: 2003-06-10

Payer 1997 Payer, Margarete: *Grundlagen der Formalerschließung: Kapitel 2: Bibliographische Beschreibung*. Skript. Mai 1997. – URL http://www.payer.de/grundlagenfe/fegscr02.htm. – Zugriffsdatum: 2003-06-24. – Überarbeitete Fassung vom 1999-10-23

Payer 1999 Payer, Margarete: *Grundlagen der Formalerschließung: Kapitel 1: Einleitung*. Skript. Oktober 1999. – URL http://www.payer.de/grundlagenfe/fegscr01.htm. – Zugriffsdatum: 2003-06-24. – Überarbeitete Fassung vom 1999-10-18

Payer 2002 Payer, Margarete: *Die Welt neben AACR2 und RAK-WB*. Juli 2002. – URL http://www.payer.de/einzel/weltnebenrak.htm. – Zugriffsdatum: 2003-06-24. – Der Text entstand anläßlich einer Fortbildungsveranstaltung des VDB am 2002-07-09 in Stuttgart

Rathje 2002 Rathje, Ulf: Technisches Konzept für die Datenarchivierung im Bundesarchiv. In: *Der Archivar* Jg. 55 (2002), Nr. 2, S. 117–120. – URL http://www.archive.nrw.de/archivar/2002-02/heft2_02_s117_126.pdf. – Zugriffsdatum: 2003-06-23

Roetzer 2003 Roetzer, Florian: Wider das digitale Vergessen. In: *Telepolis* (2003), Februar. – URL http://www.heise.de/tp/deutsch/inhalt/te/14211/1.html. – Zugriffsdatum: 2003-06-23

Rosetta 2003 : *The Rosetta Project*. 2003. – URL http://www.rosettaproject.org. – Zugriffsdatum: 2003-06-10

Rothenberg 1995 Rothenberg, Jeff: Ensuring the Longevity of Digital Documents. In: *Scientific American* (1995), Januar, S. 42–47. – URL `http://www.informatics-review.com/classic_reviews/long.pdf`. – Zugriffsdatum: 2003-06-23

Rothenberg 2000 Rothenberg, Jeff: *Using Emulation to Preserve Digital Documents*. URL `http://www.kb.nl/coop/nedlib/results/emulation.pdf`. – Zugriffsdatum: 2003-06-30, Juli 2000. – ISBN 906259145-0

Rothenberg 2001 Rothenberg, Jeff: *Digital Information Lasts Forever- Or Five Years, Whichever Comes First*. Oktober 2001. – URL `http://www.amibusiness.com/dps/rothenberg-arma.pdf`. – Zugriffsdatum: 2003-04-15

Russell 1999 Russell, Kelly: *CEDARS: Long-term Access and Usability of Digital Resources*. 1999. – URL `http://www.ariadne.ac.uk/issue18/cedars/intro.html`. – Zugriffsdatum: 2003-05-23

Schmundt 2000 Schmundt, Hilmar: Im Dschungel der Formate. In: *Der Spiegel* (2000), Nr. 26, S. 122

Schroeder 2003 Schroeder, Kathrin: *Persistent Indentifier*. April 2003. – URL `http://www.ddb.de/professionell/persistent_identifier.htm`. – Zugriffsdatum: 2003-06-10

Sitts 2000 Sitts, Maxine. K. (Hrsg.): *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Andover, Massachusetts : Northeast Document conversation Center, 2000. – URL `http://www.nedcc.org/digital/tofc1.htm`. – Zugriffsdatum: 2003-06-23

Steyer 1998 Steyer, Ralph: *HTML 4*. 1. Auflage. Düsseldorf: Data Becker, 1998. – ISBN 3-8158-1618-1

Stoll 1996 Stoll, Clifford: *Die Wüste Internet: Geisterfahrten auf der Datenautobahn*. 3. Auflage. Frankfurt am Main: Fischer, 1996

test 2003 Einige kratzt es kaum. In: *test* (2003), April, Nr. 4, S. 40–43

Van der Werf 1998 Van der Werf, Titia: *Metadata and libraries*. 1998. – URL `http://www.kb.nl/persons/titia/publ/ticer.rtf`. – Zugriffsdatum: 2003-06-06

Van der Werf 2000 Van der Werf, Titia: *The Desposit System for electronic Publications: A Process Model*. URL `http://www.kb.nl/coop/nedlib/results/DSEprocessmodel.pdf`. – Zugriffsdatum: 2003-06-23, November 2000 (Report ; 6). – ISBN 906259150-7

Winter 2001 WINTER, Judith: *Einführung in die XML (eXtensible Markup Language)*. Dezember 2001

Zaun 2001 ZAUN, Harald: Speichermedium der dritten Dimension. In: *Telepolis* (2001), Juli. – URL `http://www.heise.de/tp/deutsch/inhalt/lis/9093/1.html`. – Zugriffsdatum: 2003-06-10

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich genannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort und Datum

Unterschrift