

Susanne Dobratz, Inka Tappenbeck

## Thesen zur Zukunft der digitalen Langzeitarchivierung in Deutschland



*Der Artikel beschreibt Strategien und Techniken zur Langzeiterhaltung digitaler Information unter dem Aspekt verfügbarer Standards und deren Anwendungskontexten. Zukünftiger Handlungsbedarf wird vor dem Hintergrund bereits gewonnener Erfahrungen benannt und erläutert. Ferner werden Vorschläge gemacht, die darlegen, wie die Langzeitarchivierung digitaler Dokumente unter den in Deutschland gegebenen Bedingungen verteilter Aufgabenwahrnehmung zu gestalten wäre.*

Theses for long term archiving of digital information in Germany

*The article describes strategies and techniques for long term archiving of digital information under the aspect of existing standards and their context of application. Future call for action is described and explained against the background of practical experience. Further proposals are made which outline, how long term archiving of digital documents could be arranged under the conditions of a distributed structure as given in Germany.*

Remarques sur l'archivage de longue durée de l'information digitale en Allemagne

*L'article décrit les stratégies et techniques pour l'archivage de longue durée de l'information digitale en prenant en considération les standards déjà existants et le contexte de leur application. En vue des expériences déjà faites les auteurs décrivent et discutent le besoin futur d'agir. De plus elles proposent comment réaliser l'archivage de longue durée des documents digitaux en Allemagne sous la condition des devoirs pas centralisés.*

### 1 Probleme und Aufgaben

Mit den digitalen Medien ist eine Vielzahl neuer elektronischer Publikationsformen entstanden: digitale Sekundärveröffentlichungen, aber auch genuin elektronische Publikationen, von denen einige sogar ausschließlich als reine Netzpublikationen existieren. Der Bestand dieser Dokumente ist an elektronische Trägermedien wie z.B. Server und deren Festplatten oder CD-ROMs gebunden, oder gar an die Existenz des Internet selbst. Die Anzahl und gesellschaftliche Relevanz dieser digitalen Ressourcen nimmt weltweit kontinuierlich zu; gleichzeitig ist die Frage nach den Möglichkeiten und Bedingungen ihrer zuverlässigen Archivierung heute noch weitgehend unbeantwortet. Dies gilt sowohl für die Sicherung der Datenspeicherung (*Trägermedium*) als auch den zukünftigen Zugriff auf die in ihnen enthaltenen Informationen (*Datenformate*) und deren dauerhafte Nutzbarkeit (*Erschließung*). Die Faktoren, die einer einfachen Lösung dieser Aufgaben entgegenstehen, sind vielfältig: Datenträger zerfallen, der rasante Technologiewechsel erschwert den Zugriff auf ältere Träger und Datenformate. Vor allem aber fehlen verbindliche technische und organisatorische *Standards* für die Archivierung digitaler Ressourcen, die Rahmen und Grundlage für die Bewältigung dieser Herausforderungen bilden könnten.

Zur Zeit bemüht sich eine Reihe führender deutscher Informationseinrichtungen unter Federführung Der Deutschen Bibliothek um die Schaffung solcher Grundlagen für die Sicherung der Langzeitarchivierung digitaler Do-

kumente in Deutschland. Ein Teil der hierzu erforderlichen konzeptionellen Vorarbeit ist von der Arbeitsgruppe „Langzeitarchivierung“ des Arbeitskreises „Infrastrukturen für Digitale Bibliotheken“<sup>1</sup> im Rahmen des Digital Library Forums<sup>2</sup> bereits geleistet worden. Die Arbeitsergebnisse dieser Gruppe sind in einem Grundlagenpapier niedergelegt<sup>3</sup> und stellen eine wichtige Basis für eine weitere koordinierte Herangehensweise an nationale Archivierungsbestrebungen dar. Die folgenden Ausführungen erläutern die zentralen Thesen dieses Grundlagenpapiers.

Basis dieser Überlegungen ist ein spezifisches Verständnis des Terminus „Langzeitarchivierung“. Unter „Langzeitarchivierung“ wird von der Arbeitsgruppe primär die erfolgreiche Gewährleistung der *Langzeitverfügbarkeit* einer Ressource verstanden. Das bedeutet, dass die wesentlichen Inhaltsbestandteile einer digitalen Publikation auch in ferner Zukunft noch nutzbar sein sollen. Diese Definition setzt sich bewusst von einem rein konservatorischen Standpunkt ab. Es wird primär die Nutzbarkeit des Inhaltes eines digitalen Dokumentes in den Vordergrund gestellt, weniger die Bewahrung der Unversehrtheit seiner ursprünglichen Form.

1 Susanne Dobratz (HU Berlin), Hans Liegmann (DDB), Inka Tappenbeck (SUB Göttingen).

2 <<http://www.dl-forum.de>>.

3 Dobratz, Susanne; Liegmann, Hans; Tappenbeck, Inka: Langzeitarchivierung digitaler Dokumente. In: Zeitschrift für Bibliothekswesen und Bibliographie 48 (2001) 6, S. 327-332.

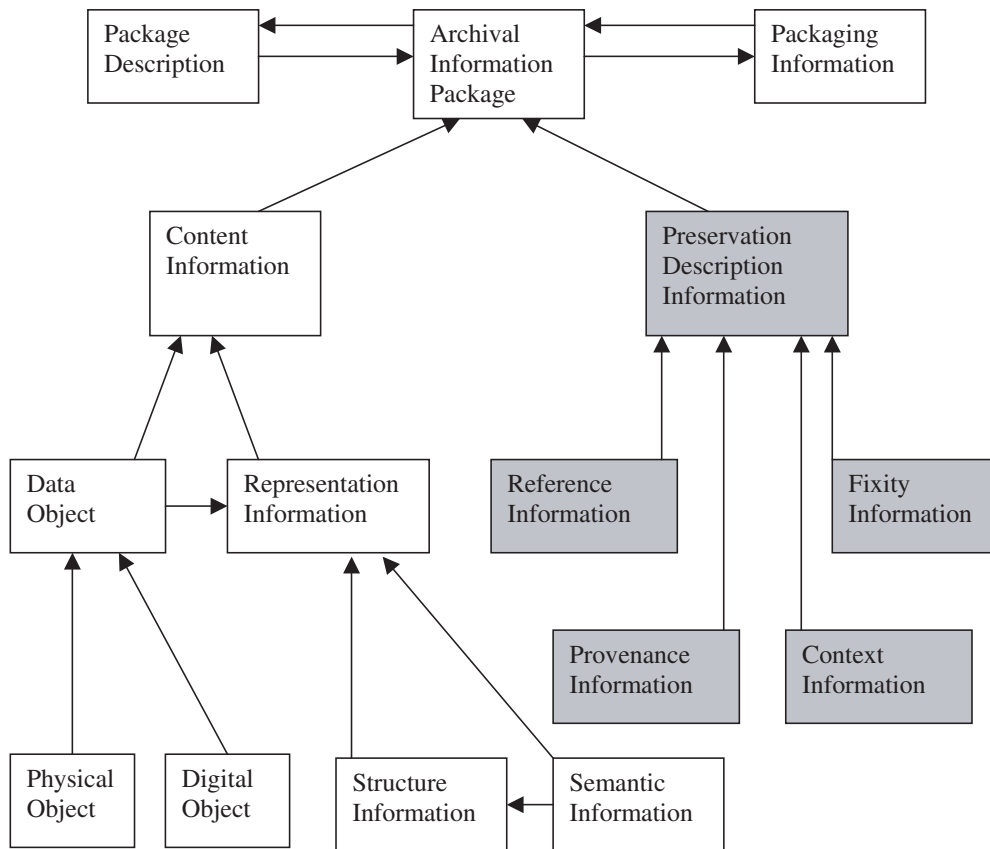


Abb. 1 Reference Model for an Open Archival Information System  
 <<http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>>

Ein wichtiges Ziel der Langzeitarchivierung im definierten Sinne bleibt dabei die Sicherung der *Authentizität* (im Sinne der Vertrauenswürdigkeit) des archivierten Dokuments. Idealerweise sollte im Prozess der Langzeitarchivierung die volle Originalität des digitalen Dokuments bewahrt werden. Realiter ist dieses Ziel jedoch aufgrund der im digitalen Kontext gegebenen Bedingungen häufig nicht vollständig erreichbar. Die Qualitätsstufe des Authentizitätsanspruches muss daher in Abhängigkeit von der Beschaffenheit der zu archivierenden Dokumente und der verfügbaren technischen Archivierungsmöglichkeiten definiert werden. So ist zum Beispiel zu klären, ob die Einbeziehung von Referenzen auf externe Quellen ein wichtiges Kriterium für die Authentizität bestimmter Dokumente ist, und – falls diese Frage positiv zu beantworten wäre – mittels welcher Archivierungsstrategie der Zugriff auf diese externen Quellen im Prozess der Archivierung gemeinsam mit dem Dokument bewahrt werden kann. Es ist also jeweils für bestimmte Klassen von Dokumenten zu klären, welche Erhaltungsstrategie ihnen prioritär zu erhaltenen Eigenschaften angemessen ist.

Im Prozess der Planung solcher Erhaltungsstrategien sind u.a. drei wichtige Arbeitsschritte zu vollziehen:

1. Da ein nationaler Alleingang in der globalen Informationsgesellschaft ein sicherer Misserfolgswahrscheinlichkeit wäre, ist es erstens wichtig, eine Bestandsaufnahme, Analyse und Auswertung der *internationalen Entwicklungen* vorzunehmen und zu prüfen, welche

der bereits existierenden Lösungsvorschläge der deutschen Situation angemessen sein könnten.

2. Die *Entwicklung von Norm-Standards* ist unbedingt erforderlich. Diese sollten in Übereinstimmung mit den sich aktuell im internationalen Rahmen abzeichnenden Standardisierungsinitiativen erarbeitet werden.
3. Der Aufbau einer dezentralen und *kooperativen Infrastruktur für die Archivierung* digitaler Dokumente in Deutschland, die nicht nur Zuständigkeiten klar definiert sondern auch effektive und effiziente Kooperationsstrukturen etabliert, ist notwendig.

## 2 Standardisierung I: Das Open Archival Information System

Die Sichtung der aktuellen internationalen Archivierungsbestrebungen macht deutlich, dass es vor allem ein Modell ist, das die Standards der Zukunft erheblich prägen wird: Das Reference Model for an Open Archival Information System, kurz *OAIS* (Red Book, Version Juli 2001)<sup>4</sup>. Dieses Organisationsmodell (siehe Abb. 1) identifiziert die zentralen Funktionen und Abläufe eines Archivsystems, bietet eine Terminologie und ein Strukturkonzept für Archivierungsmetadaten an, ist neutral gegenüber unterschiedlichen Archivierungstechniken

4 <<http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>>.

(Migration, Emulation etc.) und ermöglicht aufgrund seiner Containerstruktur eine dezentrale Implementierung. Die zentrale Einheit der OAIS stellt das *Archival Information Package* (AIP) dar. Es setzt sich zusammen aus der *Content Information* (CI), die das Archivobjekt selbst sowie Angaben zu seiner Darstellung enthält, und der *Preservation Description Information* (PDI), in deren vier Containern die Archivierungsinformationen untergebracht sind: Der *Reference-Container* enthält beschreibende Informationen zum Archivobjekt, die vor allem für das Retrieval wichtig sind. In der *Provenance-Information* wird der Archivierungsprozess selbst beschrieben (Akteur, Zeitpunkt und Art der Modifikation des Datenobjekts, technische Werkzeuge etc.). Der *Context-Container* enthält Informationen zu formalen und inhaltlichen Beziehungen des Archivobjekts zu anderen Dokumenten; in der *Fixity-Information* schließlich findet die Authentizitätssicherung (z.B. durch die Vergabe digitaler Signaturen) statt. Dabei ist wichtig, dass jeder Container mit einem eigenen Identifier versehen ist und isoliert von allen anderen verwaltet werden kann, womit eine wesentliche Bedingung für den Aufbau eines verteilten Archivsystems gegeben ist.

### 3 Standardisierung II: Dokumentenstandards

Um die Kosten für Emulations- und Migrationsmodelle zu senken und im Prozess der Archivierung inhaltliche Dokumentstrukturierungen zu bewahren, ist die Nutzung von präsentationsunabhängigen Auszeichnungssprachen wichtig. Hierbei sollte nicht auf bloße de-facto Firmenstandards aufgesetzt werden, deren zukünftiges Schicksal ungewiss und von unkalkulierbaren Marktbebewegungen beeinflussbar ist. Vielmehr ist die breite Anwendung von Norm-Standards wie Standard General Markup Language, (SGML, ISO 88879), eXtensible Markup Language (XML), eXtensible Style Language (XSL) und eXtensible Linking Language (XLink) weiter zu fördern. Dasselbe gilt für Standard-Dokumentdefinitionen (DTD) wie z.B. MathML für mathematische Formeln, CML für chemische Formeln, SVG für Vektorgrafiken, die ISO DocBook DTD, den Open e-Book Standard oder auch die xDiML – für Dissertationen Online.

### 4 Standardisierung III: Metadaten

Für die Entwicklung eines verteilten, arbeitsteilig geführten Archivsystems ist die Verwendung von Metadaten zur Beschreibung und Verwaltung der angebotenen Ressourcen unerlässlich. Erst so wird eine einheitliche Beschreibung der verschiedenen Dokumente und folglich auch ein Retrieval nach gemeinsamen Prinzipien möglich. Eine wichtige Rolle spielt hier das von der Dublin Core Metadata Initiative (DCMI)<sup>5</sup> seit 1995 entwickelte Dublin Core Metadata Element Set, kurz DCMES<sup>6</sup>, das folgendermaßen aussieht:

dc.title	dc.contributor	dc.source
dc.creator	dc.date	dc.language
dc.subject	dc.type	dc.relation
dc.description	dc.format	dc.coverage
dc.publisher	dc.identifier	dc.rights

Dieses Metadaten(austausch)format besteht aus 15 Kategorien, die eine einheitliche Beschreibung verschiedener Objekttypen und damit auch ein einheitliches Retrieval ermöglichen. Diese Beschreibungskategorien sind über die Bibliotheks- und Fächergrenzen hinweg anwendbar und haben bereits heute eine weltweite Akzeptanz und Verbreitung gefunden. Zur tieferen Beschreibung der ausgewiesenen Dokumente sind von der DCMI zusätzlich zu diesem Core-Set sogenannte Qualifier<sup>7</sup> verabschiedet worden.

## 5 Problemstellungen und Handlungsbedarf

### 5.1 Archivierungsfreundliche Dokumentformate

Die Kodierung der Dokumente spielt eine entscheidende Rolle bei der Anwendung verschiedener Archivierungskonzepte, denn jede Konvertierung birgt die Gefahr einer Informations- und Authentizitätsverfälschung. Daher sollten alle an der Publikationskette Beteiligten, angefangen beim Autor, auf eine minimale Dokumentenkonvertierung hinarbeiten. So sollte der Autor, unterstützt durch entsprechende Werkzeuge, bereits strukturierte, konvertierbare Urformen, d.h. „archivierungsfreundliche“ Dokumente liefern können. Um dieses Ziel zu realisieren, sind Konzepte zur Autorenausbildung im Umgang mit neuen Medien und in der Erstellung multimedialer Dokumente zu erarbeiten. Weiterhin sind standardisierte Werkzeuge und Schnittstellen zu häufig genutzten Werkzeugen für die Dokumentenerstellung zu schaffen. Auch dynamische Komponenten (z.B. Annotations- und Zitationsmodelle) müssen von solchen Standarddokumentformaten und -werkzeugen unterstützt werden. Und nicht zuletzt ist die Verbreitung von Datenmodellen und Technologien voranzutreiben, die ein medienneutrales Publizieren ermöglichen.

### 5.2 Transferstandards und -protokolle

Im Regelfall müssen Dokumente für die Archivierung vom Ort ihres Entstehens bzw. ihrer Verbreitung in ein Depotsystem überführt werden. An die Gestaltung dieses Transfers ergeben sich mehrere Anforderungen: Die Anwendung eines verbindlichen Transferformats soll verhindern, dass das Objekt schon bei der Überführung in ein Archiv einen ersten Migrationsschritt durchmachen muss. Die Dokumente sollten daher an der Eingangsschnittstelle des Archivsystems in standardisierter Form vorliegen und ihre archivierungsrelevanten Metadaten durch automatisierte Verfahren generiert werden können. Geschlossene Transfercontainer, die alle relevanten Daten und Metadaten des Dokuments enthalten, sowie Verfahren zur automatisierten Übermittlung dieser Container könnten den Archivierungsprozess weiter optimieren.

5 <<http://dublincore.org>>.

6 <<http://dublincore.org/documents/dces>>.

7 <<http://dublincore.org/documents/dcmes-qualifiers/>>.

### 5.3 Authentizität und Integrität

Authentizität und Integrität archivierter Dokumente sind im Verlauf von Archivierungsprozessen besonderen Gefahren ausgesetzt, wenn die ursprüngliche Bitsequenz aufgrund technischer Gegebenheiten verändert werden muss (Migration). Um Authentizität und Integrität der Dokumente dennoch sicherzustellen, müssen in der Zusammenarbeit aller an der „information chain“ Beteiligten organisatorische und technische Strukturen entwickelt werden, die innerhalb dieses Prozesses soviel Sicherheit wie möglich gewährleisten. Dazu gehört der Aufbau und die Nutzung von Public Key Infrastrukturen (PKI) in Universitäten, Bibliotheken und Archiven. Auch die Vergabe digitaler Signaturen und Zeitstempel spielt eine große Rolle beim Aufbau vertrauenswürdiger digitaler Archive. Dabei muss den besonderen Herausforderungen dynamischer Publikationen auch unter dieser Problemstellung begegnet werden.

### 5.4 Technische Archivierungskonzepte

Für die Langzeitarchivierung digitaler Dokumente existiert aktuell kein optimales technisches Konzept. Daher ist die Marktbeobachtung („technology watch“) im Bereich der Archivierungstechnologien (Weiterentwicklung der technischen Plattformen, Datenträgertechnologie etc.) eine permanente Aufgabe. Für die automatisierte Migration von Massendaten müssen nachnutzbare technische Verfahren gefunden werden, wobei auch die Rolle von Metadaten bei der Steuerung von Konversionsprozessen zu beachten ist. Nicht zuletzt sind für Archivsysteme, insbesondere dann, wenn sie in verteilter Verantwortung betrieben werden sollen, Qualitätskriterien und Zertifizierungsverfahren zu entwickeln, um die Einhaltung der technischen Standard-Anforderungen zu gewährleisten.

### 5.5 Metadatenmanagement

Informationen zu Nutzungsrechten und archivierungsrelevanten Daten gewinnen mehr und mehr an Bedeutung für die Verwaltung digitaler Dokumente. In diesem Zusammenhang bietet sich das „Reference Model for an Open Archival Information System“ (OAIS) als Strukturmodell für die Metadatenentwicklung an. Um eine möglichst weitgehende Kompatibilität der Metadaten zu gewährleisten, sollten diese als Erweiterung des Dublin Core Metadata Element Sets (DCMES) realisiert werden. Damit auf diese Metadaten dann auch in standardisierter Form zugegriffen werden kann, ist ein Normstandard für die Implementierung von Metadaten in XML/RDF zu erarbeiten. Um die immer notwendiger werdende Kooperation zwischen Produzenten, Vermittlern und Vertreibern wissenschaftlicher Information zu unterstützen, müssen ferner standardisierte Workflows zur dezentralen Erstellung und Pflege der Metadaten erarbeitet werden und in praktischen Tests erprobt und optimiert werden.

### 5.6 Funktionale und organisatorische Erfordernisse

Ausgehend von bereits geltenden Zuständigkeiten bei der Archivierung gedruckter Publikationen durch Die

Deutsche Bibliothek und andere Pflichtexemplarbibliotheken müssen nun auch für die digitale Welt gezielte koordinierende Überlegungen zur verteilten Archivierung angestellt werden. Dabei sind neben der Aufteilung der inhaltlichen Zuständigkeiten auch sicherheitstechnisch erforderliche Redundanzen in die Überlegungen einzubeziehen. Auch ist eine wirksame Abstimmung über die Art der gemeinsam zu entwickelnden und verteilt einzusetzenden technischen Instrumentarien nötig. Um eine solche Aufgabenteilung für die Archivierung digitaler Publikationen in Deutschland zu etablieren, sollte eine Struktur aufgebaut werden, die die Aufgabe der Koordination der verteilten Archivierungsaktivitäten wahrnimmt. Ein Beispiel für eine solche Struktur findet sich in Großbritannien in Form der *Digital Preservation Coalition*<sup>8</sup>. Um die angestrebte verteilte Archivierung digitaler Dokumente in den verschiedenen beteiligten Institutionen zu realisieren, ist ferner eine technische und organisatorische Unterstützung entsprechender Publikationsmodelle und Geschäftsgänge in Universitäten, Bibliotheken und Archiven erforderlich.

### 5.7 Rechtliche Fragen

Die Entwicklung von Modellen zur aktiven und passiven Verwaltung von Rechten („Digital Rights Management“) wird im digitalen Bereich immer dringlicher. Erforderlich sind vor allem semantische und technische Standards zur Beschreibung der rechtlichen Eigenschaften eines Dokuments und zur Abwicklung von Transaktionen (Bezugsrechte und -modalitäten). Weiterhin ist es von zentraler Bedeutung, eine rechtliche Basis für alle Bereiche der digitalen Archivierung zu schaffen. Dies betrifft etwa die Klärung der Frage von im Rahmen von Archivierungsprozessen erlaubten Dokumentmodifikationen.

## 6 Fazit

Eine wesentliche Vorbedingung für die Etablierung einer Archivierungsstruktur für digitale Publikationen in Deutschland ist die Stärkung der öffentlichen Bewusstseinsbildung für die Relevanz dieser Thematik. Die erfolgreiche Tradierung unseres kulturellen Erbes, das in immer stärkerem Maße in digitaler Form vorliegt, hängt in entscheidendem Maße auch davon ab, ob die erforderlichen Finanzmittel zur Schaffung der hierzu erforderlichen technischen, organisatorischen, theoretischen und infrastrukturellen Bedingungen bereit gestellt werden.

Grundlage der Bemühungen innerhalb Deutschlands sollte dabei eine umfassende Analyse der aktuellen internationalen Entwicklungen und Standardisierungstendenzen im Bereich der digitalen Archivierung sein. Diese Entwicklungen kommen zur Zeit vor allem aus dem angloamerikanischen Raum (USA, England, Australien). Um die Anschlussfähigkeit der deutschen Archivierungsaktivitäten an diese Entwicklungen zu gewährleisten und diese vor dem Hintergrund der spezifischen Bedürfnisse und Gegebenheiten der deutschen Informationslandschaft mitzugestalten, ist eine stärkere Partizipation an diesen Initiativen notwendig.

8 <<http://www.jisc.ac.uk/dner/preservation/dpcintro.htm>>.

Ferner sollten die bisher in Deutschland begonnenen Archivierungsinitiativen stärker als bisher koordiniert werden, um Doppelentwicklungen zu vermeiden und Synergieeffekte sinnvoll nutzen zu können. Die praktische Effektivität und Effizienz der entstehenden Konzepte einer verteilten digitalen Archivierung sollten unter Beteiligung aller am Informationsprozess beteiligten Parteien – vom Autor bis zum Archiv – in Projekten getestet werden.

Nicht zuletzt kann erst die Einrichtung einer wirksamen Koordinationsstruktur, die angesichts der Komplexität und Vielschichtigkeit der anstehenden Aufgaben selbst wiederum nur dezentral und kooperativ zu denken ist, den Erfolg dieser Bestrebungen insgesamt gewährleisten.

**Anschrift der Autorinnen:**

Susanne Dobratz  
Humboldt-Universität zu Berlin  
Universitätsbibliothek/Rechenzentrum  
Arbeitsgruppe „Elektronisches Publizieren“  
Dorotheenstr. 1  
D-10099 Berlin  
dobratz@rz.hu-berlin.de

Dr. Inka Tappenbeck  
Niedersächsische Staats-  
und Universitätsbibliothek  
D-37073 Göttingen  
tappenbeck@sub.uni-goettingen.de