Langzeiterhaltung digitaler Publikationen

Archivierung elektronischer Zeitschriften (E-Journals)

Dr. Gunnar Fülle Tobias Ott
pagina GmbH, Tübingen

nestor-materialien 4

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

This page is intended to be blank.

Langzeiterhaltung digitaler Publikationen

Archivierung elektronischer Zeitschriften (E-Journals)

Dr. Gunnar Fuelle
Tobias Ott

pagina GmbH, Tübingen

nestor – materialien 4

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

http://www.langzeitarchivierung.de

Projektpartner

Bayerische Staatsbibliothek, München
Bundesarchiv
Computer- und Medienservice / Universitätsbibliothek der Humboldt-Universität zu Berlin
Die Deutsche Bibliothek, Leipzig, Frankfurt am Main, Berlin (Projektleitung)
Generaldirektion der Staatlichen Archive Bayerns, München
Institut für Museumskunde, Berlin
Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Betreuer dieser Veröffentlichung: Die Deutsche Bibliothek (Hans Liegmann)

URN: urn:nbn:de:0008-20051024019
http://nbn-resolving.de/urn:nbn:de:0008-20051024019

Eine der vorrangigen Aufgabe für Archivbibliotheken ist die Übernahme der Verantwortung für die Langzeiterhaltung elektronischer Zeitschriften.

Im Unterschied zu gedruckten Publikationen werden digitale Publikationen kommerzieller Verlage ganz überwiegend nicht mehr an die Kunden übergeben. Vertriebsmodelle physischer Publikationsobjekte wurden durch Lizenzierungsmodelle ersetzt, in denen die zu nutzenden Objekte in der Verantwortung der Rechteinhaber verbleiben. Die Verbreitung von physischen Zeitschriftenbeständen über ausgedehnte geographische Räume war bislang einer der Garanten für den dauerhaften Erhalt und den beständigen Zugang zu den Inhalten. Unter den Bedingungen des digitalen Publizierens muss der gewohnte, sich implizit ergebende Vorteil der "redundanten Datenspeicherung" durch neue Organisationsformen, Regelungen von Verantwortlichkeiten und technische Verfahrensmodelle errungen werden. Das Risiko des Datenverlustes ist bei der jetzigen, so oft vorteilhaft wirkenden Einmalspeicherung digitaler Inhalte, entschieden zu groß. Hinzu kommt, dass die Langzeitarchivierung digitaler Publikationen mit dem Ziel beständiger Verfügbarkeit Herausforderungen mit sich bringt, die in der Zukunft nur von spezialisierten und vertrauenswürdigen Archivsystemen erfüllt werden können.

Die vorliegende Expertise konzentriert sich mit technischem Fokus auf die speziellen Fragestellungen, die sich bei der Langzeiterhaltung elektronischer Zeitschriften ergeben. Entsprechend dem aktuellen Entwicklungsfortschritt bei der Implementierung von Archivierungssystemen wird insbesondere die komplexe Aufgabe behandelt, wie die digitalen Ressourcen vom Produzenten in die zukunftssichere Umgebung eines digitalen Langzeitarchivs überführt werden können.

Mit der Firma pagina GmbH (Tübingen) konnte eine kompetente Auftragnehmerin gefunden werden, die das Thema unter Berücksichtigung internationaler Standardisierungsbestrebungen beleuchtet. Die formulierten Empfehlungen sind nach Auffassung des Projekts "nestor - Kompetenznetzwerk Langzeitarchivierung" geeignet, die praktische Gestaltung von Archivierungsprozessen einer für Wissenschaft und Forschung bedeutsamen Publikationskategorie zu unterstützen.

für die Partner des Projekts nestor - Kompetenznetzwerk Langzeitarchivierung

Hans Liegmann
Projektleiter nestor
Die Deutsche Bibliothek (Leipzig, Frankfurt am Main, Berlin)

Die Vorgabe zur Erstellung der vorliegenden Expertise lautete**:**

Elektronische Zeitschriften (E-Journals) sind ein wichtiger Publikationstyp mit hoher Bedeutung für die Kommunikation in Wissenschaft und Forschung. Die Erhaltung der langfristigen Verfügbarkeit dieses Typs digitaler Ressourcen ist deshalb eines der vorrangigen Ziele wissenschaftlicher Bibliotheken. Bevor Aktivitäten zur Langzeiterhaltung in einem dedizierten Depotsystem durchgeführt werden können, ist der Transfer aus der Publikationsumgebung (dem „Originalserver“) in die Hoheitsumgebung des Archivs erforderlich. Dieser Transfer ist je nach technischer Ausprägung der digitalen Objekte von unterschiedlicher Komplexität. Der Transfer statisch vorliegender Objekte kann durch Abholung (Web-Harvesting, Mirroring) oder Lieferung (FTP, Datenträger) bewältigt werden. Dynamische Objekte werden erst auf Nutzeranforderung aus Datenhaltungssystemen (z.B. Content-Management-Systemen) entnommen und „on the fly“ zur Präsentation aufbereitet. Ohne die gesonderte Definition einer auf die Belange des Archivs zugeschnittenen Transferschnittstelle (Datenformat und Protokoll) können sie nicht transportiert werden.

Inhalte und Erwartungen an die Expertise

- möglichst vollständige Ermittlung der deutschen sowie der wichtigsten internationalen E-Journal-Produzenten (Verlage und verlegende Stellen)
- Mengenbestimmung des Publikationstyps E-Journal (Artikel/Jahr)
- Übersicht zur technischen Typisierung der Objekte (statisch, dynamisch)
- Ermittlung existierender Transferwege
- Erarbeitung von Vorschlägen zur Gestaltung des Datentransfers zwischen Produzenten und Archiven unter Nachnutzung existierender Vorarbeiten, insbesondere für den Bereich der dynamischen Objekte

Neben dem technischen Aspekt der Transferprozedur ist es auch Gegenstand der Expertise, die Möglichkeiten eines Metadatentransfers für E-Journal-Artikel unter Anwendung von existierenden Standards zu evaluieren und eine Empfehlung (ggf. unter Nennung bestehender Defizite) abzugeben. Hinweise darauf, welche Metainformationen prinzipiell mitgeliefert werden sollten, sind sinnvoll.

Expertise

Langzeiterhaltung digitaler Publikationen
Archivierung elektronischer Zeitschriften
(E-Journals)

von
Dr. Gunnar Fuelle
und
Tobias Ott

pagina GmbH, Tübingen
Oktober 2005

This page is intended to be blank.

Inhaltsübersicht

This page is intended to be blank.

Einleitung

Ziel des Projektes nestor (Network of Expertise in Long-Term Storage of Digital Resources) ist der Aufbau eines Kompetenznetzwerks zur Langzeitarchivierung und Langzeitverfügbarkeit digitaler Quellen für Deutschland in einer dauerhaften Organisationsform sowie die Abstimmung über die Übernahme von Daueraufgaben. nestor ist ein Teilprojekt des Vorhabens Neue Dienste, Standardisierung, Metadaten des Bundesministeriums für Bildung und Forschung.

Die Aufgaben von nestor umfassen die Schaffung von Problembewusstsein, die Bildung eines Netzwerkes zur Bereitstellung von bisher verstreutem technischen, organisatorischen und juristischen Wissen, den Ausbau der Kooperation, die Entwicklung von Technologien und Standards sowie die Konzipierung permanenter Organisationsformen.

Darunter fallen u.a. die Erarbeitung von Kriterien für vertrauenswürdige digitale Archive, Zertifizierungsverfahren für Archivserver, Auswahlverfahren für die Archivierung digitaler Quellen, Grundsätze für die Langzeitarchivierung sowie die Einbindung der Museen und Archive. Konferenzteilnahme, Gremienarbeit und einige Workshops sind geplant. Das Kompetenznetzwerk bietet Synergieeffekte durch Nachnutzungsmöglichkeiten und best practice-Informationen. Zugleich ist nestor ein Forum, in welchem sich über Standards und die nachhaltige Übernahme von Daueraufgaben verständigt wird.

Die pagina GmbH Gesamtherstellung wissenschaftlicher Werke ist seit 1973 Partner der Verlage bei Datenaufbereitung und -ausgabe für Print- und digitale Medien.

Dr. Gunnar Fuelle ist bei der pagina GmbH als Projektleiter und Berater für Technologie und Change Management tätig, zuvor war er wissenschaftlicher Mitarbeiter an der Humboldt-Universität zu Berlin.

Tobias Ott ist Geschäftsführer der pagina GmbH und Lehrbeauftragter für »Elektronisches Publizieren« und »Grundlagen Medienvorstufe« (Satz) an der Hochschule der Medien Stuttgart.

Quellen: *www.langzeitarchivierung.de (nestor-Website). – www.pagina-tuebingen.de*

Gegenstand, Ziel und Methodik

Gegenstand dieser Studie ist eine Analyse derjenigen technischen Gegebenheiten bei Produktion und Publikation wissenschaftlicher E-Journals, die für die Langzeitarchivierung relevant sind, sowie die Formulierung von Empfehlungen für die Nutzung und Verbesserung dieser Gegebenheiten. Organisatorische, rechtliche und wirtschaftliche Aspekte werden nicht behandelt.

Die Studie umfasst drei aufeinander bezogene Teile:

Teil 1 gibt eine Einführung in die Problematik der Langzeitarchivierung mit Fokus auf die Langzeitarchivierung von wissenschaftlichen E-Journals. Vorhandene Konzepte, Methoden und Formate für die Langzeitarchivierung von E-Journals werden im Überblick dargestellt und auf ihre Vor- und Nachteile untersucht. Für den institutionellen und organisatorischen Hintergrund wird vorausgesetzt, dass die Langzeitarchivierung auf nationaler Ebene durch Einrichtungen öffentlicher Träger erfolgt, wobei eine abgestimmte, aber dezentrale und arbeitsteilige Struktur angestrebt wird. Konzepte einer Archivierung durch die Produzenten selbst werden in der Studie nicht thematisiert.

Teil 2 hat Empfehlungen zur Entwicklung eines standardisierten Datenpaketes für die Übergabe von E-Journal-Daten vom Produzenten an Archive zum Ziel (Submission Information Package; SIP). Grundlage ist die Evaluierung vorhandener und in Entwicklung befindlicher Packaging-Standards mit Blick auf die Besonderheiten von E-Journals.

Teil 3 analysiert die technischen Gegebenheiten bei der Produktion von E-Journals und zeigt, wie die Produzentenseite die Thematik Langzeitarchivierung von E-Journals einschätzt.

Mittels einer Fragebogenversendung wurden bei den wichtigsten deutschen Verlagen und E-Journal-Produzenten an deutschen Universitäten und Forschungsinstituten Daten über den Umfang der E-Journal-Produktion und das Vorliegen von technischen Voraussetzungen für die Langzeitarchivierung (Formate, dynamische Elemente, Transferwege) erhoben sowie das Interesse an der Thematik ermittelt.

Die Umfrage wurde aufgrund der vorgegebenen zeitlichen und ressourcenmäßigen Begrenzungen bei der Erhebung von Produzenten- und Journaldaten auf Produzenten wissenschaftlich-technischer E-Journals beschränkt und zwar auf die Bereiche Science/Technology/Medicine (STM) und Geisteswissenschaften (Humanities).

Entsprechend des Zieles von nestor, ein Kompetenznetzwerk für die Langzeitarchivierung digitaler Quellen in Deutschland zu entwickeln, konzentriert sich die Studie auf deutsche Produzenten.

Teil 1

Definition: Langzeitarchivierung digitaler Publikationen

Unter Langzeitarchivierung wird hier nicht alleine die Bewahrung von Dokumenten als solche über einen nicht fixierten Zeitraum, sondern auch und vor allem die dauerhafte Verfügbarmachung der intellektuellen Inhalte von Dokumenten verstanden. Das Ziel der Bewahrung dieser Inhalte auch über potenzielle technologische Brüche in der Zukunft hinweg wird als gleichrangig mit dem Ziel der tatsächlichen Nutzbarkeit angesehen. Dies versteht sich vor dem Hintergrund, dass bei digitalen Daten eine reine Substanzerhaltung nicht sinnvoll ist – ohne den technologischen und organisatorischen Kontext, der ihre Nutzbarkeit garantiert, ist die in der digitalen Substanz enthaltene Information nicht verwertbar.

Ziel ist also, jederzeit einen Datenstrom zur Verfügung stellen zu können, der die im digitalen Ursprungsdokument enthaltenen Informationen vollständig und unverfälscht wiedergibt.

Dennoch schließt die Langzeitarchivierung nicht notwendigerweise die unveränderte Bewahrung des ursprünglichen Publikationsformates ein. Sollte das Publikationsformat der Langzeitverfügbarmachung seiner Inhalte nicht entgegenstehen, kann die Transformation in ein anderes Format notwendig werden.

Die vorzugsweise Verwendung langzeitstabiler Dokumentformate sowie Migration und Emulation als Mittel zur Überwindung technologischer Brüche werden daher als gleichermaßen mögliche und sich ergänzende Strategien betrachtet.

Quellen: *Research Libraries Group, Trusted digital repositories: Attributes and responsibilities. An RLG-OCLC Report, 2002 (http://www.rlg.org/longterm/repositories.pdf). – Ute Schwens, Hans Liegmann, Langzeitarchivierung digitaler Ressourcen, in: Grundlagen der praktischen Information und Dokumentation, Bd. 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, München 2004, S. 567- 570. (auch verfügbar unter http://www.langzeitarchivierung.de/downloads/digitalewelt.pdf).*

Definition: E-Journal

Die exakte Definition des Untersuchungsgegenstandes E-Journal bzw. elektronische Zeitschrift ist nicht trivial, da schon für den Begriff »Zeitschrift« keine allgemein zutreffende Definition existiert. Im Gegensatz zur Tageszeitung, die sich durch die Merkmale Periodizität, Publizität, Disponibilität, Aktualität und Universalität auszeichnet, gelten für Zeit-

schriften zwar die ersten drei Merkmale, die letzteren beiden aber nicht notwendigerweise.

Auf der organisatorischen Ebene verfügen Zeitschriften wie Zeitungen über eine Redaktion und einen Herausgeber.

Weiterhin findet man häufig die Unterscheidung, dass Zeitschriften im Gegensatz zu Zeitungen weniger auf Nachrichten, sondern schwerpunktmäßig auf analytische oder investigative Hintergrundberichterstattung zu aktuellen Themen ausgelegt sind.

Meist widmet sich eine Zeitschrift einem bestimmten Themenbereich (Fachzeitschriften, Special Interest Journals). Publikumszeitschriften (General Interest Journals) sind Ausnahmen von dieser Regel.

E-Journals unterscheiden sich vor allem durch das elektronische Verbreitungsmedium von konventionellen Zeitschriften, wobei in der Regel das Internet als Distributionskanal dient.

Von den oben genannten Merkmalen einer Zeitschrift entfällt bei E-Journals das Merkmal der Periodizität. Die Periodizität der Printzeitschriften ist durch den physikalischen Informationsträger »Heft« begründet, der aus wirtschaftlichen Gründen einen Mindestumfang aufweisen muss und einen bestimmten Zeitraum zur Herstellung benötigt. E-Journals, die lediglich eine elektronische Version einer ansonsten identischen Printzeitschrift darstellen, folgen zwar in der Regel der Periodizität der Printversion. Allerdings gehen die Herausgeber solcher E-Journals im Interesse höherer Aktualität vermehrt dazu über, Artikel schon vor dem Erscheinen der Printversion als E-Version zur Verfügung zu stellen (»Online first«). E-Journals ohne parallele Printausgabe sind gar nicht an eine periodische Erscheinungsweise gebunden.

Unter Anwendung dieser Definitionsversuche soll für die vorliegende Studie ein E-Journal als elektronisches Publikationsforum mit einer zeitschriftenähnlichen Organisation (Qualität sichernde Redaktion und verantwortlicher Herausgeber) gelten, das schwerpunktmäßig der fortlaufenden Veröffentlichung von Artikeln analytisch-investigativen Inhalts dient.

Quellen: *Walther Umstätter, Digitales Lehr- und Handbuch der Bibliothekswissenschaft, -The Digital Textbook of Library Science-, http://www.ib.hu-berlin.de/~wumsta/infopub/textbook/definitions/di3.html, Stand 25.7.2005. – Holger Rada, Von der Druckerpresse zum Web-Server – Zeitungen und Magazine im Internet, Berlin 1999 (Einleitung verfügbar unter http://www.uni-tuebingen.de/uni/nmw/forschung/1/rada/einleitung.pdf). – Hans J. Koelsch (Springer-Verlag Heidelberg), Vortrag »Von der Bibliothek an den Schreibtisch – Die Zukunft wissenschaftlicher Monografien und Zeitschriften« in der Heidelberger Print Media Academy 14. Februar 2002 (Abstract unter http://www.uni-heidelberg.de/presse/news/Abstracts_Feb_20021.pdf).*

Dringlichkeit der Langzeitarchivierung digitaler Ressourcen

Für die Langzeitarchivierung elektronischer Ressourcen gibt es noch keine etablierten Verfahren. Ursache für diesen Zustand ist die Vielfalt von Konzepten der digitalen Informationsrepräsentation, die ständig wachsende Zahl verschiedenster Datenträger- und Datenformate und der dazugehörigen Soft- und Hardware, die nach nur wenigen Jahren von leistungsfähigeren Nachfolgeversionen abgelöst wird.

Werden elektronische Ressourcen wie konventionelle Ressourcen archiviert – nämlich durch einfache Ablage der Originaldaten auf ihrem Originaldatenträger – sind sie spätestens nach ein oder zwei Jahrzehnten nicht mehr oder nur mit unverhältnismäßig hohem Aufwand nutzbar. Die technische Infrastruktur für den Zugriff auf Datenträger und Daten steht schon nach einem solchen relativ kurzen Zeitraum in der Regel nicht mehr zur Verfügung, Lesegeräte und Anwendersoftware wurden von Nachfolgeversionen abgelöst, die mit den archivierten Formaten nicht mehr kompatibel sind. Hinzu kommt das Problem der teilweise sehr begrenzten Haltbarkeit der Datenträger.

Eine langfristige Archivierung und Verfügbarmachung elektronischer Ressourcen setzt daher zweierlei voraus. Erstens, dass das Problem der begrenzten Haltbarkeit der Datenträger gelöst wird und zweitens, dass für alle archivierten Daten jederzeit die passende Infrastruktur verfügbar gemacht werden kann – sei es durch die Anpassung der Daten an neuere Systeme oder neuerer Systeme an die archivierten Daten.

Die Umsetzung dieser Aufgaben ist in fast jeder Hinsicht aufwendiger als konventionelle Archivierungsstrategien und wurde bislang nicht in zufriedenstellender Weise gelöst.

Gleichzeitig nimmt die Zahl der elektronischen Dokumente ebenso beständig zu wie ihre Akzeptanz in Wirtschaft, Wissenschaft, Verwaltung, Unterhaltung und Kultur. In all diesen Bereichen werden zunehmend digitale Dokumente produziert, für die kein analoges Äquivalent mehr zur Verfügung steht. Umso dramatischer ist es, dass für diese Dokumente keine zuverlässigen Archivierungsmethoden zur Verfügung stehen.

Es ist nicht nur absehbar, sondern in vielen Fällen schon jetzt der Fall, dass wegen fehlender oder ungeeigneter Archivierung wichtige elektronische Daten nicht mehr verwendbar sind oder mit viel Aufwand rekonstruiert werden müssen.

Die Dringlichkeit, dieses Problem zu lösen, wächst exponentiell an. Daher ist eine Schärfung des Problembewusstseins bei Produzenten wie bei Archivaren elektronischer Dokumente notwendig, mit dem Ziel der Vereinbarung von Standards, ohne die eine kostenverträgliche Langzeitarchivierung digitaler Ressourcen nicht möglich ist.

Besonderes Augenmerk gilt hier dem Publikationstyp E-Journal, da neben die klassische Zweitverwertung von Printjournalen als E-Version zunehmend »E-Only«-Journale treten, die wissenschaftliche Forschungsergebnisse nur noch in digitaler Form veröffentlichen.

Quellen: *Uwe M. Borghoff, Peter Rödig, Jan Scheffczyk, Lothar Schmitz: Fehlt der Wissensgesellschaft bald das Gedächtnis? mesh – Magazin für Wissens- und Informationsdiskurs, 12/2003*

Konzepte der Langzeitarchivierung digitaler Ressourcen

Mit allen anderen Informationsressourcen teilen elektronische Ressourcen das Problem der begrenzten Haltbarkeit ihrer Trägermaterialien. Anders als bei alterungsbeständigem Papier, das bei optimaler chemischer Zusammensetzung und richtiger Lagerung mehrere hundert Jahre hält, ist über die langfristige Haltbarkeit von magnetischen oder optischen Datenträgern wenig bekannt. Die Haltbarkeit der Kunststoffe von optischen oder magnetischen Datenträgern dürfte bestenfalls einige Jahrzehnte betragen, die Magnetisierung schwächt sich zudem mit der Zeit von alleine ab.

Allgemeine Ressourcen für die folgenden Abschnitte: Uwe M. Borghoff, Peter Rödig, Jan Scheffczyk, Lothar Schmitz, Langzeitarchivierung: Methoden zur Erhaltung digitaler Dokumente, Heidelberg 2003. – Hans Liegmann, Langzeitverfügbarkeit digitaler Publikationen, 2001 (http://www.uni-muenster.de/Forum-Bestandserhaltung/konversion/digi-liegmann.shtml). – nestor-Website: www.langzeitarchivierung.de. – Info-Website des Projektes »Langzeitarchivierung« an der Universität der Bundeswehr München: http://ist.unibw-muenchen.de/Inst2/Research/LZA/. – Website zur Langzeitarchivierung digitaler Ressourcen der Library of Congress, Washington: http://www.digitalpreservation.gov/. – Website zur Langzeitarchivierung digitaler Ressourcen der National Library of Australia, Project PADI: Preserving Access to Digital Information: http://www.nla.gov.au/padi/. – Website des internationalen Bibliotheksprojektes NEDLIB: http://www.kb.nl/coop/nedlib/.

»Auffrischung« und Datenbanksysteme

Allerdings ist bei elektronischen Daten die Möglichkeit, ohne großen Aufwand – und wenn nötig automatisch – eine Kopie auf einem neuen oder neuartigen Datenträger anzufertigen, ein erheblicher Vorteil gegenüber konventionellen Ressourcen. Durch diese Möglichkeit des »Auffrischens« der Träger stellt die begrenzte Haltbarkeit der Trägermaterialien bei elektronischen Ressourcen ein gut lösbares Problem dar.

Eine weitere Möglichkeit, dieses Problem anzugehen, ist die Trennung der Daten von ihrem Originalträger (Diskette, CD, DVD etc.) und ihre Übernahme in ein datenbankgestütztes Archivierungssystem, verbunden mit professionellen Sicherheits- und Backupstrategien, die im Prinzip eine automatische Auffrischung für komplette Datenbestände darstellen.

Auffrischung und datenbankgestützte Archivierung basieren auf bewährten und leicht verfügbaren Technologien und finden bereits Anwendung oder stehen vor dem Einsatz.

Langzeitstabile Datenträger

Eine Alternative zu Auffrischung und datenbankgestützter Archivierung ist die Verwendung von Datenträgern, deren Material und Codierungssystem extrem langzeitstabil sind. Das trifft beispielsweise auf die wahlweise metallische oder keramische Rosetta-Disk der Firma Norsam zu, die analoge und digitale Daten durch mikromechanische Änderungen der Oberflächenstruktur speichert. Die Rosetta-Disk ist extrem unempfindlich gegenüber Umwelteinflüssen und daher ohne besondere Vorkehrungen unbegrenzt haltbar. Zudem lassen sich bei Nutzung als Träger digitaler Daten auf der Fläche einer CD 165 GB speichern. Eine breite Anwendung ist aufgrund der hohen Kosten allerdings noch nicht in Sicht.

Quelle: *http://www.norsam.com/rosetta.html. – Heminger, Alan R.. – Robertson, Steven B., Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents, 1998, (http://www.ercim.org/publication/ws-proceedings/DELOS6/rosetta.pdf)*

Technische Infrastruktur

Anders als die meisten traditionellen Informationsquellen haben elektronische Ressourcen den Nachteil, dass sie nicht ohne hochkomplexe Hilfsmittel lesbar sind. Ihre Nutzung setzt die Existenz einer entwickelten technischen Infrastruktur voraus, bestehend aus einem System von Hard- und Software, das in der Lage sein muss, den Datenträger zu lesen und die ausgelesenen Daten in der intendierten Form verfügbar zu machen. Aber nicht nur zur Wiedergabe, auch für die Auffrischung von Datenträgern oder die datenbankgestützte Archivierung muss die entsprechende hochentwickelte Infrastruktur vorhanden sein.

In historischen Zeiträumen gedacht stellt die grundsätzliche Weiterexistenz der technischen Infrastruktur ein potentielles Problem dar. Ein Zivilisationsbruch mit einem Verlust dieser Infrastruktur würde zu verheerenden Wissensverlusten führen. Will man für derartige Fälle vorsorgen, bietet sich die Nutzung von Datenträgern an, deren Material und Codierungssystem so langzeitstabil sind, dass sie auch in Jahrhunderten noch ausgelesen werden können. Die einfachste Methode wäre das Ausweichen auf langlebige und direkt lesbare Printmedien oder Mikrofilme von diesen und deren stabile Einlagerung, was freilich nur für »ausdruckbare« Daten möglich ist (z.B. Sicherungsverfilmung von Archiv- und Bibliotheksgut im Rahmen des Schutzes der Kulturgüter bei bewaffneten Konflikten). Alternativen wie extrem langzeitstabile digitale Datenträger befinden sich im Entwicklungsstadium bzw. sind aus Kostengründen noch nicht in Stückzahlen anwendbar (z.B. die Rosetta-Disk).

Allerdings bedarf es keines Kulturbruchs, um den Verlust der Lesbarkeit elektronischer Informationsquellen herbeizuführen. Die technische Infrastruktur im Bereich der elektronischen Datenverarbeitung entwickelt sich sehr schnell. Sowohl Speichermedien und dazu-

gehörige Lesegeräte als auch die benötigte Software sind durch die kurzen Innovationszyklen der IT-Industrie nach wenigen Jahren veraltet und werden nicht mehr produziert. Wird dieser Tatsache nicht Rechnung getragen, können Daten nach wenigen Jahrzehnten gar nicht oder nur mit unverhältnismäßig großem Aufwand wieder lesbar gemacht werden.

Im Folgenden sollen die wesentlichen Lösungsansätze für das Infrastrukturproblem vorgestellt werden.

Quellen: *http://www.zivilschutz-online.de*

Museale Archivierung

Eine auf den ersten Blick nahe liegende Lösung des Problems ist die Archivierung der Inhalte auf originalen Datenträgern gemeinsam mit der benötigten Hard- und Software. Bei genauerer Betrachtung erweist sich diese Lösung jedoch als nicht praktikabel. Probleme sind neben dem enormen Platzbedarf ein hoher Administrationsaufwand, Einschränkungen bei der Verfügbarkeit für die Nutzer, mit der Zeit zunehmende hardwaretechnische Probleme (Ersatzteile etc.) und die begrenzte Verfügbarkeit der Datenträger für die Auffrischung.

Migration

Wenn die Hard- und Software, mit der archivierte Daten erzeugt und Datenträger beschrieben wurden, nicht mehr zur Verfügung steht, muss neue verwendet werden. Bei der Migration werden daher die zu archivierenden Daten nicht im Original belassen, sondern beim Verschwinden der dazugehörigen Infrastruktur auf neue Datenträger kopiert und in neue Datenformate konvertiert.

Dieser Weg ist jedoch sehr aufwendig, da alle vorhandenen Daten regelmäßig konvertiert werden müssen. Der Aufwand wächst also mit dem Datenbestand und kann allenfalls durch Automatisierung in den Griff bekommen werden. Noch entscheidender ist das Problem der Datenvielfalt – für jedes Datenformat muss für jeden Migrationsschritt ein Konverter geschrieben werden.

Wenn Quell- und Zielformat nicht vollständig kompatibel sind, birgt dieser Weg zudem das hohe Risiko, dass eine ungewollte Veränderung der Daten bei der Konvertierung stattfindet. Das bedeutet ein hohes Risiko der Informationsverfälschung sowie des Verlustes der Lesbarkeit durch die Anwendungssoftware – und damit im Extremfall den Verlust des betroffenen Datenbestandes.

Je besser die zu migrierenden Datenformate dokumentiert sind, umso eher lassen sich solche Veränderungen vorhersagen und gegebenenfalls korrigieren. Allerdings ist es um die Dokumentation bei den weit verbreiteten proprietären Datenformaten kommerzieller Softwarehersteller zum Teil schlecht bestellt. Hier wäre eine individuelle Überprüfung des Konvertierungserfolgs für jede Quelle notwendig, was freilich den Aufwand extrem steigern würde, da der Vorteil der Automatisierung durch den Kontrollaufwand zunichte gemacht würde.

Zum eigentlichen Aufwand der Datenkonvertierung und Erfolgskontrolle kommt die Planung hinzu, da gerade bei kommerziellen Formaten Zeitpunkt und Aufwand der nächsten Konvertierung nur durch eine ständige Beobachtung der Herstelleraktivitäten bestimmt werden können.

Das Konzept der Migration stößt zudem sofort an seine Grenzen, wenn für eine Software keine Nachfolgeversionen mehr entwickelt werden und in der Folge entwickelte Rechner und Betriebssysteme die Software nicht mehr unterstützen.

Aufwand und Risiken lassen sich erheblich minimieren, wenn nur Datenformate verwendet werden, die für die Langzeitarchivierung optimiert sind und daher nur wenig Planungs-, Konvertierungs- und Kontrollaufwand bei der Migration verursachen. Dabei ist vor allem wichtig, dass die Formate möglichst wenig auf Besonderheiten einzelner Hard- und Softwareplattformen ausgerichtet sind, sondern Fähigkeiten nutzen, die vielen Plattformen gemeinsam sind, dass sie vollständig dokumentiert sind und dass sie eine weite Verbreitung aufweisen. Damit sind für die Migration besonders plattformübergreifende Datenformate geeignet, für die offene internationale Standards existieren.

Quellen: *VdW Arbeitskreis »Elektronische Archivierung«, Bericht: Zweite Tagung des VdW-Arbeitskreies »Elektronische Archivierung« am 17./18. November 2003 im Siemens-Forum in München (http://www.wirtschaftsarchive.de/akea/m_akea_bericht2003.htm)*

Weitere aktuelle Quellen und Links unter der Themen-Webpage »Migration« der National Library of Australia, Project PADI: Preserving Access to Digital Information (http://www.nla.gov.au/padi/topics/21.html).

Emulation

Bei der Emulation wird nicht mehr verfügbare Hard- und Software durch neuere Hard- und Software nachgeahmt (emuliert). Dieser Weg wird von der Computerindustrie aus Gründen der Kompatibilität neuer Produkte mit noch weit verbreiteter älterer Software und mit anderen Systemen schon seit längerem erfolgreich gegangen.

Der Vorteil für die Langzeitarchivierung liegt auf der Hand: im Idealfall muss lediglich die vom Markt verschwindende Hardware durch neuere Systeme softwaretechnisch emuliert

werden, und schon können alle Programme und alle Daten, die für diese Hardware geschrieben wurden, unverändert weiter verwendet werden.

Da Hardware in der Regel mit Blick auf ältere noch in Gebrauch befindliche Software meist über mehrere Generationen abwärtskompatibel ist, können mit einer Hardwareemulation ganze Softwaregenerationen bedient werden.

Selbst wenn der Aufwand für die Programmierung einer Hardware-Emulation relativ hoch ist, liegt er doch unter dem Aufwand für die Migration aller betroffenen Datenbestände. Zudem entfällt das Risiko einer unerwünschten Modifikation der Originaldaten durch fortwährende Konvertierung.

Schließlich können auch Formate wiedergegeben werden, die wegen fehlender Weiterentwicklung der Software nicht migrieren können. In Fällen, wo Informationsgehalt und Präsentationssoftware unlösbar eng und individuell miteinander verbunden sind, entziehen sich Objekte der Behandlung durch Migration. Eine CD-ROM-Anwendung, die für eine bestimmte Betriebssystemumgebung produziert wurde, kann mit dieser so verflochten sein, dass eine nachträgliche Umsetzung auf andere Systembedingungen nicht mit vertretbarem Aufwand möglich ist.

Die Emulation setzt voraus, dass die Daten von den veraltenden Datenträgern auf neue übertragen werden, die von aktueller Hardware gelesen werden können. Dies wird im Rahmen der Datenträgerauffrischung jedoch ohnehin der Fall sein bzw. entfällt bei der Übernahme der Daten in ein Datenbanksystem (welches allerdings auch in größeren Abständen »aufgefrischt« werden muss). Auch die Software, die zur Wiedergabe der Daten benötigt wird, muss in einer Form vorgehalten werden, die für das neue System lesbar ist.

Die Problematik der Emulation liegt daher vor allem in der Organisation der Wiederherstellung der Plattform, die für die Verwendung archivierter Daten erforderlich ist. Je komplexer das Datenformat aufgebaut und je spezieller die intendierte Anwendung ist, desto komplexer und spezieller wird die Kombination von Hard- und Software sein, die die Daten so lesbar macht, wie es von deren Erzeuger beabsichtigt war. Dazu kann nicht nur eine spezielle Hardwarekonstellation erforderlich sein, sondern auch eine ganz spezifische Betriebssystemversion mit einer Reihe von Erweiterungen sowie die eigentliche Anwendersoftware in einer bestimmten Version, ggf. mit Plug-Ins etc. Gerade in der Medienindustrie, von der typischerweise viele zu archivierende Inhalte erzeugt werden, existieren Kompositformate, deren korrekte Umsetzung auf Ausgabegeräten so spezielle Konfigurationen erfordert, dass oft nur die persönliche Inaugenscheinnahme des Ergebnisses durch den Produzenten die Korrektheit der Umsetzung gewährleisten kann.

Solche Konfigurationen durch Emulation zuverlässig rekonstruieren zu können, stellt eine enorme Herausforderung dar.

Zur Lösung dieses Problems existieren zwei Ansätze. Der Ansatz der Datenkapselung setzt voraus, dass die gesamte Wiedergabe-Software, einschließlich des Betriebssystems, zusammen mit den zu archivierenden Daten als ein Datencontainer gespeichert wird. Alternativ wird Software gesondert archiviert. Bei der Archivierung von Daten wird durch Metadaten angegeben, welche Software für die Datenwiedergabe erforderlich ist. In beiden Fällen müssen weitere Metadaten zur Emulatorspezifikation alle Informationen enthalten, die zur Herstellung der benötigten Hardware-Emulation erforderlich sind. Schließlich sind Metadaten erforderlich, die die Herstellung der Gesamtkonstellation aus emulierter Hard- und archivierter Software ermöglichen. Eine wichtige Herausforderung ist daher die Sicherung der Lesbarkeit und Verwendbarkeit der Metadaten über lange Zeiträume, und zwar unabhängig von der Emulation, die ja die Metadaten selbst nicht betreffen kann.

Ist für ein Datenformat erst einmal eine zuverlässig emulierbare Umgebung herstellbar und getestet, sind anders als bei der Migration in der Zukunft keine weiteren Schritte erforderlich, die an den archivierten Daten ausgeführt werden müssen. Lediglich die weitere Emulierbarkeit der erforderlichen »historischen« Hardware auf Hardware immer neuer Generationen muss ebenso sichergestellt werden wie die Verwendbarkeit der Metadaten.

Für alle diese Ansätze existieren allerdings noch keine umfassend anwendbaren Lösungen. Einsetzbare Methoden zur vollständigen Spezifikation von Hardware- und Softwarearchitekturen stehen in absehbarer Zeit auch nicht zur Verfügung.

Grundsätzliche Grenzen der Emulation werden erreicht, wenn die technische Weiterentwicklung dazu führt, dass Eingabe- oder Ausgabegeräte zur Mensch-Maschine-Kommunikation, die von »historischer« Software vorausgesetzt werden, nicht mehr verfügbar sind. Diese Geräte müssen dann wenn möglich ebenfalls emuliert werden oder nach dem Konzept des Technikmuseums (mit den entsprechenden Nachteilen) verfügbar gehalten werden.

Eine erfolgreiche Emulationsstrategie wird umso eher verfügbar sein, je überschaubarer die Zahl der zu emulierenden Konstellationen ist. Auch für die Emulation gilt daher, dass sie leichter umzusetzen ist, wenn die Zahl der zu archivierenden Datenformate mit der von ihnen benötigten Soft- und Hardware überschaubar bleibt, also Standardformate genutzt werden können. Standarddatenformate sind bei der Emulation jedoch nur der erste Schritt, ebenso werden Standards für die Spezifikation von Hard- und Softwarearchitekturen und -konfigurationen benötigt.

Quellen: *Dominik Bodi, Emulation als Methode zur Langzeitarchivierung digitaler Dokumente, 24. Mai 2000 (http://www2-data.informatik.unibw-muenchen.de/Lectures/FT2000/Digitale-Bibliotheken/handout5.pdf). – S. Granger, Emulation as a digital preservation strategy, 2000 (http://www.dlib.org/dlib/october00/granger/10granger.html). – Joint Informations Systems Comittee/National Science Foundation, Emulation Options for Digital Preservation: Technology emulation as method for long-term access and preservation of digital resources (http://www.leeds.ac.uk/cedars/JISCNSF/index.htm). – Koninkliijke Bibliotheek and RAND-Europe, Emulation Testbed for Digital Preservation (http://www.konbib.nl/coop/nedlib/index.html?/coop/nedlib/results/WP4-E-factsheet.html). – Jeff Rothenberg, Ensuring the Longevity of Digital Documents, Scientic American, Vol. 272, No. 1, Januar 1995, S. 42-47*

(http://www.clir.org/programs/otheractiv/ensuring.pdf). – Jeff Rothenberg, Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Council on Library and Information Resources, Januar 1999 (http://www.clir.org/pubs/reports/rothenberg/contents.html). – Jeff Rothenberg, Tora Bikson, Carrying Authentic, Understandable and Usable Digital Records Through Time, RAND-Europe, August 1999. – Jeff Rothenberg, An Experiment in Using Emulation to Preserve Digital Publication, Koninkliijke Bibliotheek Den Haag, April 2000 (http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf).
Weitere aktuelle Quellen und Links unter der Themen-Webpage »Emulation« der National Library of Australia, Project PADI: Preserving Access to Digital Information (http://www.nla.gov.au/padi/topics/19.html).

Langzeitstabile Datenformate:

Allgemeines

Wie in den vorangehenden Abschnitten deutlich wurde, sind die Migration von technisch veraltenden Datenformaten zu aktuellen und die Emulation von technisch veraltenden Nutzungsumgebungen durch aktuelle komplexe und damit aufwändige und fehleranfällige Prozesse. Wenn irgend möglich, sollten sie überflüssig gemacht werden oder zumindest Häufigkeit und Umfang ihrer Anwendung sowie die damit verbundenen Risiken minimiert werden.

Dies ist nur möglich, wenn Datenformate verwendet werden, die Migration und Emulation nur dann notwendig machen, wenn es zu einem grundlegenden Technologiewandel kommen sollte, kaum aber, wenn sich die Informations-Technologie auf den Bahnen weiterentwickeln sollte wie bisher – und wie bisher absehbar.

Das Interesse an solchen Datenformaten ist besonders auf Seiten derer hoch, die die gesetzliche oder kulturelle Verpflichtung zur Archivierung von Daten haben – und damit den Aufwand leisten müssen, der zur Herstellung einer Langzeitverfügbarkeit notwendig ist.

Auch die Produzenten von Daten sollten sich der Vorteile der Verwendung derartiger Datenformate bewusst sein, zumindest dann, wenn ihnen an der Langzeitverfügbarkeit ihrer Daten gelegen ist. Denn: sollte das Publikationsformat der Langzeitverfügbarkeit der intellektuellen Inhalte nicht entgegenkommen, kann die Transformation in ein anderes Format unumgänglich werden – mit allen damit verbundenen Risiken.

Neben sehr speziellen Datenformaten, die nur mit dem Programm eines einzigen Herstellers auf einer bestimmten Rechnerplattform laufen (und damit grundsätzlich nicht langzeitstabil sein können), gibt es Formate, die von vornherein auf größtmögliche Kompatibilität mit den verschiedensten Systemen ausgelegt sind (auch Austauschformate genannt). Das geschieht in der Regel dadurch, dass diese Formate nur auf solche Basisfunktionen zurückgreifen, die bei allen Hard- und Software-Konstellationen gleich oder ähnlich sind. Diese Basisfunktionen sind es auch, die bei künftigen Systemen mit großer Wahrscheinlichkeit ebenfalls vorhanden sein dürften. Darauf aufbauende Datenformate können daher sehr

wahrscheinlich über lange Zeiträume ohne Transformationen mit vielen gängigen Systemen genutzt werden.

Hier entfällt die Notwendigkeit von Migration oder Emulation entweder ganz oder diese Schritte sind einfacher, sicherer und kostengünstiger durchzuführen. Lediglich eine Auffrischung der Datenträger oder die Übernahme in ein Datenbanksystem ist in jedem Fall notwendig.

Ein Nachteil solcher Formate ist oft, dass sie weniger leistungsfähig sind als Formate, die die Möglichkeiten eines Systems voll ausnutzen und dafür auf anderen Systemen nicht lauffähig sind. Allerdings ist im Zeitalter des Internet, das von Nutzern mit unterschiedlichsten Hardware-Plattformen und verschiedensten Software-Konfigurationen gleichermaßen genutzt wird, die Tendenz zu plattformübergreifenden Formaten erheblich gewachsen.

Langzeitstabile Datenformate sollten nicht alleine größtmögliche Kompatibilität mit den verschiedensten Hardware/Software–Konfigurationen aufweisen. Mindestens ebenso wichtig ist, dass die Dokumentation des Formates öffentlich verfügbar ist, idealerweise handelt es sich um eine von der ISO zertifizierte Norm. Nur so ist die Unabhängigkeit von Geschäftsinteressen und Dokumentationssorgfalt der kommerziellen Entwickler eines Formats gegeben. Die Herstellung von Software zur Auswertung von Daten in dem fraglichen Standard kann auch dann auf zuverlässige und für jedermann frei verfügbare Dokumentationen aufsetzen, wenn keine kommerzielle Umsetzung mehr zur Verfügung steht.

Wünschenswert ist weiterhin, dass die fraglichen Formate eine weite Verbreitung gefunden haben, das Gleiche sollte auf Software zutreffen, die diese Formate erstellen und verarbeiten kann. Nur wenn die Produzenten ihre Daten bereits in langzeitstabilen Dokumentformaten an die Archivierungsstelle liefern, kann der fehleranfällige und aufwendige Schritt der Konvertierung der gelieferten in langzeitstabile Datenformate entfallen.

Im Folgenden sollen zuerst die wichtigsten langzeitstabilen Formate vorgestellt werden, die für die Archivierung von Textinhalten und Grafiken in Frage kommen. Im Anschluss erfolgt ein Ausblick auf die sogenannten Multimedia-Formate, die im Prinzip alle Dateninhalte umfassen, die über Text und Grafik hinausgehen. Die Frage der Speicherung wissenschaftlicher Primärdaten wird im Zusammenhang mit den Mulitmedia-Formaten angesprochen.

Bei allen langzeitstabilen Formaten ist zu beachten, dass es sich um professionelle Formate handelt, die mit sehr vielen Möglichkeiten auch viele Fehlerquellen bei der Erstellung aufweisen, und zwar in einer Weise, dass Folgen und Umfang der Fehler auf den ersten Blick nicht sichtbar sind. Es ist daher unabdinglich, im Zweifelsfall professionelle Hilfe zu suchen – sei es in der eigenen Organisation oder bei einem externen Dienstleister – um bei der Umwandlung unwiederbringlicher Originaldaten in ein langzeitstabiles Format Daten- oder Qualitätsverluste auszuschließen.

Welche Datenformate bei der Produktion von E-Journals eine besondere Rolle spielen, wird in Teil 3 dieser Studie im Rahmen der Umfrage und ihrer Auswertung erörtert.

Quellen: *exzellente Informationsbasis für alle Datenformate zur Langzeitarchivierung ist die Website »Digital Formats for Library of Congress Collections« (http://www.digitalpreservation.gov/formats/)*

Langzeitstabile Formate für textbasierte Informationen: SGML, XML und HTML

Das Paradebeispiel für langzeitstabile Formate ist das Trio SGML, XML und HTML. Der Ursprung von SGML liegt im Bestreben von Publishern im Industrie- und Verlagsbereich, ein Standard-Datenformat für Textinhalte zu schaffen, das der Verwendung in verschiedenen Publikationssystemen (Print, Elektronische Publikationen) ebenso entgegenkommt wie der Auswertbarkeit durch Rechnersysteme (z.B. für intelligente Suchanfragen).

Solche Daten nennt man »medienneutral«. Medienneutrale Daten müssen in einer Form vorgehalten werden, die für alle gewünschten Publikationsformen als Quelldatenformat dienen kann.

Eine weitere wichtige Anforderung ist die Plattformneutralität: die Daten müssen auf allen Computersystemen verwendbar sein. Ebenso wichtig ist Herstellerunabhängigkeit, was meint, dass das Format nicht an Hard- oder Software eines einzigen Herstellers gekoppelt ist. Idealerweise handelt es sich um ein freies Format, dessen vollständige Dokumentation öffentlich zugänglich ist und das keinerlei lizenzrechtlichen Beschränkungen unterliegt. Auf das Format soll mit beliebigen Programmiersprachen zugegriffen werden können.

Doch auch Daten, die auf allen Plattformen lesbar, öffentlich dokumentiert und lizenzrechtlich frei sind, sind damit noch nicht notwendigerweise sinnvoll recherchierbar oder in sinnvoller Weise in verschiedenen Medien ausgebbar. Eine einfache Textdatei beispielsweise wird häufig den Anspruch der Systemunabhängigkeit erfüllen. Aber sie kann nur per einfacher Volltextsuche durchsucht und nur durch aufwendige manuelle Arbeit formatiert werden, da schon einfachste Gliederungselemente wie z.B. Überschriften nicht als solche in den Daten gekennzeichnet sind und damit nicht automatisiert verarbeitet werden können.

Man kann also noch einen weiteren Anspruch an die Daten hinzufügen, nämlich die Möglichkeit, diese gemäß ihren spezifischen Inhalten und Strukturen automatisch auszuwerten.

Es gibt mithin mehrere Gründe, in medienneutralen Daten nicht die Formatierung eines Dokumentes, sondern die Struktur und die Art des Inhalts eindeutig zu beschreiben. Der

wichtigste ist sicherlich, dass die Struktur eines Dokumentes mit seinem Inhalt verbunden ist, während die Formatierung nichts Absolutes ist.

Anders gesagt: ein Dokument bestimmten Inhalts kann auf unterschiedlichste Weise formatiert werden, ohne dadurch seine Struktur zu verlieren. Wenn also der Text und seine formale wie inhaltliche Struktur eindeutig codiert sind, wird sich jede systematische Formatierung seiner Elemente daraus herstellen lassen.

Die Struktur eines Textes in einer objektivierten und für den Rechner lesbaren Form im Dokument abzuspeichern, löst eine ganze Reihe Probleme: Zunächst wird der Text automatisiert weiterverarbeitbar, und zwar in beliebiger Typografie und für die unterschiedlichsten Ausgabeformen. Damit wird dem Anspruch auf Medienneutralität Rechnung getragen. Da die Struktur explizit in den Daten codiert ist, sind die Daten durch Suchanfragen erschließbar, die die Struktur mit berücksichtigen und so viel genauer formuliert werden können.

Schließlich lassen sich in einem Text sehr viel mehr Informationen abspeichern, als zur rein typografischen Umsetzung benötigt werden, z.B. Verwaltungsinformationen.

Ein erster solcher medienneutraler Standard wurde 1986 geschaffen: Die »Standard Generalized Markup Language« (SGML) trägt bereits im Namen, was sie leisten soll: Eine Auszeichnungssprache zu sein, die als internationaler Standard alle textuell wiedergebbaren Inhalte strukturieren kann.

Eine SGML-Datei stellt eine ganz normale Textdatei dar, die sich auf allen Plattformen und mit einfachster Software öffnen lässt. Die Auszeichnung oder Markierung (engl. Markup) der Inhalte mit weitergehenden Informationen erfolgt ebenfalls in Textform, wobei die Auszeichnungen mit Begrenzungszeichen (engl. Delimiter) von den Inhalten abgegrenzt werden. Ein Text kann in SGML z.B. so als Überschrift gekennzeichnet werden:

```
<ueberschrift>Text der Überschrift</ueberschrift>
```

Die Zeichen »<« und »>« sind die Begrenzer der Auszeichnung, das Zeichen »/« in der zweiten Auszeichnung zeigt an, dass diese Auszeichnung das Ende von »ueberschrift« markiert. Die mit den Delimitern gekennzeichneten Auszeichnungen nennt man auch Tags. Auszeichnungen in SGML können auch ganze Abschnitte umfassen, z.B. als »kapitel« oder »abschnitt« markiert, wobei Abschnitte in Abschnitte geschachtelt auftreten können, um so Unterabschnitte zu repräsentieren. Jeder Abschnitt kann wiederum seine eigene Überschrift tragen etc.

Auszeichnungen in SGML können so gestaltet werden, dass sie selbsterklärend und auch ohne Software verständlich sind. Eine so gestaltete SGML-Datei kann auch als Papierausdruck gelesen und verstanden werden. SGML-Software kann die Auszeichnungen verwen-

den, um den markierten Text z.B. in einem bestimmten Layout anzuzeigen oder spezifisch nach Textstellen in Überschriften zu suchen.

SGML ist dabei nicht selbst eine Auszeichnungssprache (d. h. es werden keine bestimmten Auszeichnungen vorgegeben), sondern es definiert die Syntax, in der eigene Auszeichnungssprachen definiert werden können. Jede Art von Dokument soll entsprechend seiner Eigenart nach Inhalt und Struktur erschlossen werden können. SGML löst dieses Problem dadurch, dass keine Struktur und keine Elemente vorgegeben werden, sondern eine Sprache zur präzisen Beschreibung von Dokumentstrukturen bereitgestellt wird. So kann der Anwender selbst festlegen, welche Textelemente mit welchen Tags ausgezeichnet werden sollen. Diese Möglichkeiten werden von Anwendergruppen (Industriezweige, Verlagsverbünde) genutzt, die für ihre Inhalte passende Auszeichnungen definieren, um diese Inhalte leichter austauschen und automatisiert verarbeiten zu können.

Allerdings hat SGML einen Vorteil, der ihm in der Praxis zum Nachteil gereicht, nämlich seine extreme Flexibilität und Komplexität. SGML lässt dem Nutzer so viele Möglichkeiten und bietet so komplexe Konstruktionen an, dass es sehr aufwendig ist, Software zu programmieren, die den Standard voll umsetzt.

Daher hat sich SGML nur begrenzt durchsetzen können. Anwendung findet SGML insbesondere dort, wo große Datenmengen codiert werden müssen und die finanziellen Ressourcen nicht zu knapp sind. So trifft man auf SGML in Bereichen wie der Flugzeug-, der Automobil- und anderer Großindustrien, im Militär – aber nur wenigen großen Verlagshäusern.

Nur in einer konkreten Anwendung fand SGML bisher über alle Nutzergruppen hinweg weltweite Verbreitung: als Hypertext Markup Language (HTML), die Seitenbeschreibungssprache für Internet-Browser, festgelegt vom World Wide Web Consortium, dem Internet-Standardisierungs-Gremium (W3C). HTML ist also eine konkrete Anwendung von SGML – allerdings eine extrem beschränkte. Wenige festgelegte Auszeichnungen dienen dem einzigen Zweck, Text und Grafiken im Webbrowser zu formatieren bzw. zu positionieren und Web-Seiten mit Links zu verbinden. Mit HTML können medienneutrale oder inhaltlich orientierte Auszeichnungen nur sehr begrenzt vorgenommen werden.

Diese Grenzen von HTML, die auch dem Datenaustausch im Internet zunehmend hinderlich wurden, führten dazu, dass das W3C 1998 mit der eXtensible Markup Language (XML) eine neue Auszeichnungssprache definierte. XML ist wie SGML keine Sprache mit einem festen Repertoire von Auszeichnungen, sondern eine Sprache zur präzisen Beschreibung von Dokumentstrukturen mittels frei festlegbarer Auszeichnungen. Sie ist also im Gegensatz zum ganz und gar vordefinierten HTML erweiterbar (extensible). Andererseits sind bei XML viele Merkmale, die bei SGML frei bestimmbar waren, in XML konkret festgelegt.

Der Verzicht auf manche »exotische« Komponente hat die Verarbeitbarkeit von XML-Dateien im Vergleich zu SGML enorm vereinfacht. Während die ISO-SGML-Norm über 500 Seiten umfasst, kommt die Druckfassung des XML-Standards mit 26 Seiten aus.

Der erwünschte Effekt trat ein: durch die Vereinfachung der Programmierung von Anwendungssoftware entstand bald eine breite und kostengünstige Basis für die praktische Nutzung des neuen Standards. Diese Basis hat zusammen mit der vergleichsweise einfachen Anwendbarkeit der Sprache XML zu ihrer enormen Verbreitung beigetragen. XML hat SGML de facto verdrängt, nur in wenigen Fällen wird SGML noch verwendet. Der Vorteil der Verbreitung von XML wiegt die Einschränkungen gegenüber SGML leicht auf.

Inzwischen arbeiten fast alle relevanten Anwendungen im Publishing-Bereich mit XML, angefangen mit Microsoft Office über Satzprogramme bis hin zu Online-Publishing-Systemen. Aber XML ist nicht immer gleich XML – das XML, das in Microsoft Word verwendet wird, enthält wie schon HTML vorrangig Formatierungsinformationen, während die medienneutrale Verwendung Informationen über Inhalt und Struktur benötigt. Dennoch ist die XML-Fähigkeit der meisten Programme eine Voraussetzung für die breite Nutzung des Formats. Um beim Beispiel Microsoft Office zu bleiben – mit PlugIns kann man Microsoft Word nun auch dazu bringen, als Editor für inhalts- und strukturorientiertes XML zu dienen.

Viele Verlage haben schon auf SGML/XML-Datenhaltung umgestellt oder sind dabei. An erster Stelle stehen hier Verlage, deren Inhalte sich gut für die Weiterverwertung in elektronischen Medien wie CD-ROM oder Websites eignen: Wissenschafts- und Fachverlage, die in großem Umfang Referenzwerke publizieren. Aber auch die Produktion von wissenschaftlichen Zeitschriften gehört zu den vorrangig nach SGML/XML umgestellten Workflows.

Für die medienunabhängige Codierung von Textinhalten eignet sich XML in ausgezeichneter Weise. Es erfüllt dabei zugleich alle Kriterien, die an ein langzeitstabiles Datenformat gestellt werden können.

Quellen: *XML- und HTML-Spezifikationen unter www.w3c.org. – SGML-Spezifikation ISO 8879:1986 unter www.iso.org. – XML and Digital Preservation, Testbed Digitale Bewaring, September 2002, (http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-en.pdf)*

Langzeitstabile Formate für Pixelgrafiken (1):

TIFF

TIFF (Tagged Image File Format) ist ein Datenformat zur Speicherung von Bilddaten. Es wurde ursprünglich von Aldus (1994 von Adobe übernommen) und Microsoft für gescannte Bitmapbilder entwickelt. Es ist eines der wichtigsten Formate zum Austausch von Daten in der Druckvorstufe. Die Dokumentation ist frei verfügbar. ISO-normiert sind die TIFF-

Spezifikationen für Digitalfotografie (TIFF/EP; ISO 12234-2:2001) und die medienunabhängige Bildverarbeitung (TIFF/IT; ISO 12639:2004). Eine sehr große Verbreitung hat die Basisversion von 1992 (TIFF 6.0).

TIFF ist plattformunabhängig. Software zur Erstellung und Verarbeitung von TIFF ist für alle Hard- und Softwareplattformen verfügbar, alle professionellen Grafik- und Satzprogramme akzeptieren TIFF-Dateien.

TIFF eignet sich in besonderem Maße für den Druck, da es sowohl Farbmanagementinformationen, Farbseparation und den Beschneidungspfad für Bildmotive ohne Hintergrund speichern kann. Für die Archivierung ist die verlustfreie Qualität des TIFF-Bildes gefragt. TIFF berücksichtigt verschiedene Verfahren zur Datenkomprimierung.

In einer TIFF-Datei können mehrere Bilder abgelegt werden. Das können verschiedene Versionen desselben Bildes sein, z.B. ein Vorschaubild (Thumbnail) und das Originalbild oder mehrere Bilder mit jeweils einem Vorschaubild. Dabei unterstützt es sowohl verlustlose als auch verlustbehaftete Kompressionsverfahren.

Ein Nachteil des TIFF-Formates ist seine Komplexität, die dazu führt, dass es oft von Programmen mit einer fehlerhaften Implementierung nicht richtig verarbeitet wird. Die Vielfalt möglicher gültiger TIFF-Dateien kann zudem von keinem einzelnen Programm vollständig unterstützt werden. In der Spezifikation des Datenformats ist deswegen mit Baseline TIFF eine Untermenge gültiger TIFF-Dateien definiert, die jedes TIFF-fähige Programm verarbeiten können sollte.

TIFF-Dateien sind auf eine Größe von 4 GB beschränkt, eine Grenze, die hochauflösende wissenschaftliche Grafiken z.B. aus der Astronomie inzwischen überschreiten. Weiterhin können TIFF-Dateien nicht gestreamt werden, d.h. es muss vor einer Anzeige erst die ganze Datei oder ein erheblicher Teil davon geladen werden.

Eine Dokumentation des Formats wird von Adobe kostenlos als PDF-Datei zur Verfügung gestellt. Die aktuelle Version ist TIFF 6.0 vom 3. Juni 1992. Sie wird ergänzt durch TIFF Technical Notes. Dabei handelt es sich um Erweiterungen, die TIFF einzelne Fähigkeiten hinzufügen, u.a. das Deflate-Verfahren zur verlustlosen Datenkompression.

Aufgrund seiner Verbreitung und Plattformneutralität ist TIFF als langzeitstabiles Datenformat geeignet, wobei aber die Baseline-Einschränkungen eingehalten werden sollten.

Die DFG empfiehlt über die Beachtung der Baseline-Spezifikationen hinaus folgende Einschränkungen zu beachten:

- Farbbilder nicht als »Palette-color images« (Pseudofarben) zu speichern, obwohl dies von Baseline-TIFF unterstützt wird.

- Bitonale Bilder immer und ausnahmslos unter Verwendung der verlustfreien (Fax-)Komprimierung Gruppe 4 (Standard der ehemaligen CCITT, heute ITU) zu speichern, obwohl dies technisch gesehen keine Baseline-Option ist.
- Colormetrie-Informationen mitzuspeichern, wenn möglich.

Für Farb- und Graubilder sollte die verlustfreie LZW-Kompression verwendet werden. Diese Option wird derzeit leider relativ selten unterstützt, weil potentiell Lizenzgebühren für die Software, die dieses Verfahren nutzt, anfallen. Da mit LZW-Kompression eine Reduktion der Datenmenge um bis zu 50 % erreicht werden kann, sollte diese Option für Projekte mit großen Datenmengen dennoch in Erwägung gezogen werden.

Quellen: *TIFF 6-Dokumentation von Adobe Inc.: http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf (englisch). – TIFF ISO-Normen ISO/IEC 12234-2:2001, ISO/IEC 12639:2004 unter www.iso.org. – Praxisregeln des Förderprogramms für »Kulturelle Überlieferung« der DFG, 2004 (www.dfg.de/forschungsfoerderung/ formulare/download/12_151.pdf).*

Langzeitstabile Formate für Pixelgrafiken (2):

PNG

Aufgrund seiner technischen Eigenschaften und seines Status als ISO-Norm (seit 2004) und W3C-Recommandation kommt auch das neue Format PNG (Portable Network Graphics) für die digitale Langzeitarchivierung in Frage. Bisher ist PNG jedoch noch nicht soweit verbreitet wie TIFF oder EPS.

Die Tatsache, dass PNG und das von PNG verwendete Kompressionsverfahren im Gegensatz zu TIFF vollständig frei von Lizenzansprüchen ist, spricht für seine Verwendung als Datenformat für Archivierungszwecke.

Technisch ist PNG ein plattformübergreifendes Datenformat und enthält einen verlustfreien LZ77-Komprimierungsalgorithmus.

PNG eignet sich sowohl für digitale Master als auch digitale Nutzungsformen. Wie auch GIF kann PNG Pixel aus einer Farbpalette mit bis zu 256 Einträgen verarbeiten. Für die Archivierung interessant ist aber eher die Unterstützung hoher Farbtiefen (16 Bit für Graustufenbilder und bis zu 48 Bit für Farbbilder gegenüber 8 Bit bei Graustufen und 24 Bit bei Farbe im Baseline-TIFF-Format). PNG kann Informationen zu Farbmanagement, Farbseparation und den Beschneidungspfad für Bildmotive ohne Hintergrund speichern (Gamma-Faktor, Alpha-Kanal und K-Wert, ab Version 1.2 können auch ICC-Profile eingebettet werden, LAB-Fähigkeit).

Anders als TIFF unterstützt PNG einen linearen und schrittweisen Bildaufbau (bereits bei 20-30% der übertragenen Bilddaten ist der Bildinhalt erkennbar). Damit ist es als Web-Format verwendbar.

PNG ermöglicht das Abspeichern zusätzlicher Information in der Grafikdatei, zum Beispiel Autoren- und Urheberhinweise.

Bei der Verwendung von PNG für Archivzwecke sollte nur die verlustfreie Kompression verwendet, auf Palettenfarben verzichtet und so viele technische Bildinformation wie verfügbar mit gespeichert werden.

Quellen: *PNG-Spezifikation des W3C: http://www.w3.org/TR/PNG/. – ISO/IEC-Standard 15948:2004 (www.iso.org). – Handbuch: PNG, The Definitive Guide von Greg Roelofs, O'Reilly 2002. – aktuelle Informationen auf der informellen PNG (Portable Network Graphics) Home Site http://www.libpng.org/pub/png/*

Langzeitstabile Formate für Pixelgrafiken (3):

GIF, BMP, JPEG, JPEG 2000

Die weit verbreiteten Grafikformate GIF, BMP und JPEG sind nicht für die Langzeitarchivierung geeignet (Ausnahme: JPEG 2000).

GIF (Graphics Interchange Format; Service Marke der CompuServe Inc.) ist ein Format für die Bildschirmdarstellung von Pixelgrafiken und ermöglicht lediglich die Verwendung von maximal 256 Farben aus einer Palette, sein Komprimierungsmechanismus ist lizenzrechtlich nicht frei (bis 2003/2004).

BMP (Bitmap) ist ein ursprünglich proprietäres Grafikformat von Microsoft Windows, das in verschiedenen Varianten vorliegt. Es ist auf die Bildschirmdarstellung von Pixelgrafiken beschränkt und bietet bei weitem nicht die Möglichkeiten der Prepress-Formate TIFF und PNG. Am weitesten verbreitet ist Version 3 (frühere gibt es nicht). Microsoft hat mit Windows 95 und Windows 98 neuere Versionen 4 und 5 des BMP-Formates eingeführt, die Alphakanäle und Farbkorrektur ermöglichen und als Containerformat für PNG und JPEG-Dateien verwendet werden können. Diese neuen Formate sind jedoch nur sehr selten als eigenständige Dateien anzutreffen und werden kaum von Anwendungen unterstützt; sie finden eher als internes Format in Windows-Programmen Verwendung.

JPEG (Joint Photographic Experts Group) verwendet eine verlustbehaftete Kompression, was ein Ausschlusskriterium für die Anwendung in der Langzeitarchivierung ist.

JPEG 2000 ist dagegen ein äußerst mächtiges Grafikformat, das auch als ISO-Norm vorliegt. Bei der Entwicklung von JPEG 2000 wurden ausdrücklich die Kriterien berücksichtigt, die die Langzeitarchivierung an ein Format stellt. Allerdings hat JPEG 2000 den Status eines tatsächlichen Standards noch nicht erreicht, seine Verbreitung und Akzeptanz ist noch gering. Sollte sich das ändern, ist JPEG 2000 ein starker Kandidat für die Langzeitarchivierung von Pixelgrafiken.

Quellen: *www.jpeg.org. – JPEG 2000 - ISO/IEC-Standard 15444 (www.iso.org). – aktuelle BMP-Spezifikation von Microsoft: http://msdn.microsoft.com/library/default.asp?url=/library/en-us/gdi/bitmaps_2w1f.asp. – John Miano, The Pro-*

grammer's Guide to Compressed Image Files: Jpeg, Png, Gif, Xbm, Bmp, Addison-Wesley 2003. – Kurz-Information: Grafikformate und deren Konvertierung, Website des LEIBNIZ-RECHENZENTRUM München (http://www.lrz-muenchen.de/services/software/grafik/grafikformate/)

Langzeitstabile Formate für Vektor- und kombinierte Grafiken (1):

EPS

Neben TIFF ist EPS (Encapsulated Postscript; Adobe Systems) das zweite professionelle Standardformat für Grafiken in der Druckvorstufe. Als Austauschformat ist es seit vielen Jahren etabliert. Der Standard ist vom Hersteller vollständig dokumentiert und nicht lizenzbelastet. EPS ist eine Untermenge von Adobes ebenfalls frei dokumentierter Seitenbeschreibungssprache PostScript und Austauschformat für diese. Viele Grafik- und Layout-Programme auf allen wichtigen Betriebssystemen unterstützen EPS.

EPS dient zur Speicherung von Vektorgrafiken, Rastergrafiken mit Halbtönen, formatiertem Text und ganzen Seitenlayouts einschließlich Schriften. Verfügbare Farbmodi sind RGB, Lab, CMYK, Duplex, indizierte Farben und Graustufen. EPS arbeitet mit Farbtiefen von 1, 4, 8 und 24 Bit/Pixel. Im Gegensatz zu PostScript beschreibt EPS pro Datei immer nur eine Seite. Es sind daher einige PostScript-Befehle, insbesondere druckerspezifische, nicht zulässig.

Durch die frei verfügbare Dokumentation, die große Verbreitung und die Systemunabhängigkeit ist EPS ebenso wie TIFF für die Langzeitarchivierung geeignet.

Allerdings ist eine korrekte Darstellung und Bearbeitung nur über ein Programm möglich, das die Vektorinformationen in EPS verarbeiten kann, was aber mit allen professionellen Zeichenprogrammen möglich ist. Viele Bildbetrachtungs-Programme können das für die Druckausgabe konzipierte EPS aber nicht darstellen. Daher kann in EPS ein Vorschaubild integriert werden, wobei auch plattformspezifische Formate erlaubt sind, was allerdings dem Ziel der Plattformneutralität zuwiderläuft. Als Format für die schnelle Online-Übertragung ist EPS daher ebenfalls nicht geeignet.

EPS enthält anders als TIFF auch keinen integrierten Komprimierungsalgorithmus.

EPS hat gegenüber TIFF jedoch den Vorteil, dass enthaltene Vektorgrafiken und Schriften in der Größe skaliert werden können, ohne dass die Genauigkeit leidet. Texte und Grafiken können zudem aus EPS-Daten extrahiert werden. EPS ist daher besonders als Langzeitarchivierungsformat für Vektorgrafiken geeignet, insbesondere wenn diese in Kombination mit Texten auftreten (z.B. Charts, Pläne).

Bei der Erstellung von EPS sollte auf die folgenden Punkte geachtet werden:

- Alle notwendigen Daten wie z.B. Schriften müssen in die EPS-Datei eingeschlossen (inkludiert) werden.
- Auf die Abspeicherung von EPS-Daten in Form von Binärdaten (8-Bit-EPS) sollte verzichtet werden, stattdessen sollten sie als ASCII-EPS (7-Bit-EPS) gespeichert werden. Binäre EPS-Dateien sind zwar kleiner, können aber nicht mit allen Systemen gelesen werden.
- Auf die Einbindung eines Vorschaubildes sollte verzichtet werden, da hier je nach Plattform unterschiedliche Formate verwendet werden, was zu Problemen auf der jeweils anderen Plattform führen kann. (Die Einbindung von Vorschaubildern im EPS-spezifischen, plattformneutralen EPSI-Format wird leider von vielen Programmen nicht unterstützt und führt zudem zu einer deutlichen Vergrößerung des Speicherbedarfs).
- Die Bounding-Box (das die Grafik umschließende Rechteck) muss korrekt angegeben werden.
- DSC-Kommentare sollten weitgehend genutzt werden (DSC, Document Structuring Conventions; Angaben zu technischen Spezifika des EPS).
- Keine geräteabhängigen Optionen verwenden (Rastereinstellungen, Transferfunktionen, Überdruckvorschau, Schwarzaufbau, ICC als Gerätefarben)
- Da die Erzeugung von EPS-Daten mittels Druckertreibern oft zu Problemen führt, sollten stattdessen die EPS-Speicheroptionen professioneller Grafikprogramme verwendet werden.

Bei der Wahl von langzeitstabilen Formaten sollte auch der Aspekt einer möglichst einfachen Nutzbarkeit eine Rolle spielen. Hier schneidet EPS aufgrund des Problems der fehlenden Darstellbarkeit durch Standardviewer nicht gut ab. PDF, das als ebenso mächtiges Format für skalierbare und extrahierbare Schriften und Vektorgrafiken in Frage kommt, kennt keine Viewer-Probleme und kann daher als flexiblere Alternative in Betracht gezogen werden (zu PDF und damit verbundenen potentiellen Problemen siehe weiter unten).

Quellen: *PostScript- und EPS-Spezifikation: http://partners.adobe.com/asn/tech/ps/specifications.jsp. – Peter Vollenweider, EPS-Handbuch, Hanser 1997.*

Langzeitstabile Formate für Vektor-Grafiken:

SVG

SVG – Scalable Vektor Graphics – ist eine auf XML basierende Auszeichnungssprache, die seit 2001 als W3C Recommendation vorliegt (seit 2003 als Version 1.1, 1.2 ist in Arbeit). Mit SVG können skalierbare Vektorgrafiken und Vektoranimationen auf der Grundlage von XML kodiert werden.

Als internationaler Standard und W3C-Norm kommt SVG auch als Format für die Archivierung von Vektorgrafiken in Frage. Auch Texte und Pixelgrafiken lassen sich einbinden. Die Orientierung von SVG auf die Nutzung im Internet bedeutet allerdings, dass SVG bei weitem nicht so viele Formatierungsmöglichkeiten bietet wie EPS oder PDF.

Zudem hat SVG bei weitem noch nicht den Verbreitungsgrad der etablierten Formate, ist aber auf dem Vormarsch.

Solange es um Speicherung und Archivierung von Vektorgraphiken geht, die bereits als SVG vorliegen, spricht nichts gegen die Archivierung der SVG-Daten.

Als Alternative bietet sich auch hier PDF an, das Vektorgrafiken und Schriften ebenfalls in skalierbarer und extrahierbarer Form speichern kann.

Quelle: *SVG-Spezifikation unter www.w3c.org. – J. D. Eisenberg, SVG Essentials, O'Reilly 2002.*

Langzeitstabile Formate für Seitenbeschreibung und beliebige Grafiken:

PDF

PDF (Portable Document Format) ist ein Datenformat zur systemübergreifenden Seitenbeschreibung, das von Adobe Systems entwickelt wurde. PDF ist neben PostScript der internationale De-facto-Standard für die Erzeugung von Druckvorlagen.

PDF ist ursprünglich ein proprietäres Datenformat, es basiert zu großen Teilen auf dem Seitenbeschreibungs-Format PostSript, ebenfalls von Adobe Systems. Das Format ist im PDF Reference Manual von Adobe vollständig dokumentiert. Dadurch können durch Drittentwickler beliebige PDF-Werkzeuge bereitgestellt werden. Die aktuelle PDF-Version ist 1.4.

PDF als ISO-Norm

Seit 2001 hat die PDF-Variante PDF/X (X für exchange), die speziell dem Dokumentenaustausch dienen soll, den Status einer ISO-Norm (ISO 15930). PDF/X wurde seitdem beständig weiterentwickelt und liegt inzwischen als Version 3 vor (PDF/X-3; ISO 15930-6:2003).

Eine weitere Variante speziell für Zwecke der Langzeitarchivierung (PDF/A) befindet sich auf dem Weg zur ISO-Norm, den sie 2006 erreicht haben soll (ISO/CD 19005-1 Electronic document file format for long-term preservation – Use of PDF 1.4 (PDF/A); Publication target date: 2006-09-01).

PDF als Web-Format

Seinen hohen Verbreitungsgrad verdankt PDF vor allem dem Adobe Reader von Adobe, einem kostenlosen PDF-Viewer, der sich auch als PlugIn in Internet-Browser einbinden lässt und für alle Soft- und Hardwareplattformen zur Verfügung steht. Dadurch hat PDF weltweit eine herausragende Bedeutung bei der Publikation von formatierten Print-Inhalten über das Internet erhalten. Eine PDF-Druckvorlage ist immer zugleich für die Webpublikation verwendbar – lediglich die Dateigröße muss minimiert werden, was bei PDF unproblematisch ist, wenn man Qualitätsverluste etwa bei Pixelgrafiken in Kauf nimmt. Das Anzeigen von Printseiten auf einem PC-Bildschirm ist zwar nicht die optimale Lösung für Online-Publishing-Zwecke, aber kostengünstig und im Falle von PDF in der Darstellungsqualität (vor allem auch bei Ausdrucken) sehr zuverlässig.

PDF-Features

PDF kann von einer Vielzahl von Layout- und Grafikprogrammen erzeugt werden; Dutzende von Tools ermöglichen die Herstellung von PDF aus den unterschiedlichsten Quellen. PDF-Dateien geben die Dokumente der Ursprungsprogramme einschließlich aller Schriften (auch nicht-europäische), Farben, Grafiken und Vektorgrafiken präzise wieder. Da PDF über effektive Komprimierungsmechanismen verfügt, sind PDF-Dateien meist deutlich kleiner als ihre Quelldateien.

PDF-Dateien können beliebig viele Seiten Umfang haben, die maximale Seitengröße beträgt 4 x 4 Meter. Schriften und Vektorgrafiken können ohne Qualitätsverlust bis 6400% vergrößert werden. So finden auch extrem große Übersichten oder Pläne auf einer PDF-Seite Platz. Über die Textsuche im einzelnen Dokument oder die Volltextrecherche innerhalb einer PDF-Dokumentensammlung lassen sich sehr einfach Fundstellen auffinden.

Textpassagen, Tabellen und Bilder aus PDF-Dokumenten können leicht in anderen Anwendungsprogrammen durch Kopieren und Einfügen der jeweiligen Elemente weiterverarbeitet werden. Es ist auch möglich, PDF-Dateien direkt zu bearbeiten, z.B. Texte zu ändern oder Grafiken zu verschieben.

PDF ermöglicht die Anlage von Lesezeichen, mittels derer Anwendungen hierarchische Inhaltsverzeichnisse zur einfachen Navigation durch große Dokumente erzeugen können; ebenso gibt es die Möglichkeit Links und Anker in PDF anzulegen.

Schließlich lassen sich in PDF auch Ton- und Filmdaten unterbringen sowie Formulare mit Skripting-Funktionen erstellen oder Kommentare »anheften«.

Mittels des optionalen Dokumentenschutzes mit 40 oder 128 Bit Verschlüsselung kann der Ersteller des Dokuments gezielt die Rechtevergabe des betreffenden Dokuments steuern. Er kann z.B. verhindern, dass Benutzer das Dokument abändern, ausdrucken oder Inhalte über

die Zwischenablage kopieren können. PDF-Dateien können auch mit digitalen Signaturen »unterschrieben« werden.

PDF kann auch als Format zur verlustfreien Abspeicherung von einzelnen Grafiken verwendet werden und zwar nicht nur von Pixelgrafiken, sondern wegen der Skalierbarkeit von Vektorgrafiken und Schriften auch als Alternative zu EPS und SVG.

Tagged PDF

In der neuesten Version ermöglicht PDF die logische Strukturierung der abgespeicherten grafischen und Textinhalte auf eine ähnliche Weise, wie dies in XML geschieht, nämlich durch Tags. Tagged PDF ermöglicht so eine Erschließung der Inhalte, die über ihre rein grafische Repräsentation hinausgeht. Zum einen kann Tagged PDF dafür verwendet werden, die Inhalte je nach Ausgabegerät (Standardrechner, Handheld, Handy) anders zu formatieren, zum anderen können Inhalte besser durchsucht werden, und schließlich können sich Sehbehinderte Tagged PDF vorlesen lassen. Der letztere Punkt hat enorme Bedeutung, da seit 1998 in den USA gesetzlich vorgeschrieben ist, dass von Bundesbehörden veröffentlichte Inhalte auch für Behinderte voll zugänglich sein müssen (Accessibility Act).

PDF-Probleme

PDF ist alles in allem eine extrem komplexe Technologie, was dazu führt, dass beim Austausch von PDF-Daten nicht immer bei allen Beteiligten und Ausgabegeräten das erwünschte Ergebnis zustandekommt. Schriften oder hochauflösende Grafiken müssen z.B. nicht zwingend in eine PDF-Datei eingebunden werden, sondern können auch lediglich referenziert werden, so dass die Datei von einem Rechner, auf dem die betreffende Schrift oder Grafik fehlt, nicht richtig dargestellt wird oder es bei der Druckausgabe zu Fehlern kommt.

PDF/X

Diese Situation hat zu Bemühungen geführt, das PDF-Format wieder soweit einzuschränken, dass es für einen bestimmten Zweck zuverlässig nutzbar ist. Für die Druckvorstufe ist die ISO-Norm PDF/X (X für eXchange) eine solche Lösung. Neben dem Ausschluss von Ton-, Film- und Skripting-Komponenten sowie der Festlegung, dass alle benötigten Daten Bestandteil der PDF-Datei sein müssen, verlangt PDF/X die Einhaltung von Regeln bei Verarbeitungsanweisungen (Farbprofile, Bemaßung u.ä.). So kann man eine PDF-Datei auf PDF/X-Konformität prüfen und sichergehen, dass jeder andere, der PDF/X-konforme Software verwendet, auch das erwünschte Ergebnis erhält (Blind-Exchange-Fähigkeit).

PDF/A

PDF/A verfolgt das Ziel, den Aufbau und Inhalt von PDF-Daten so zu spezifizieren, dass sie die Ansprüchen der Langzeitarchivierung erfüllen. Die US-amerikanischen Verbände »Association for Suppliers of Printing, Publishing and Converting Technologies (NPES)« und die »Association for Information and Image Management, International (AIIM International)« arbeiten seit einigen Jahren auf dieses Ziel hin: Sie wollen mit PDF/A einen internationalen Standard für die digitale Archivierung von Schriftgut definieren. Wie schon oben erwähnt, soll PDF/A 2006 den Status einer ISO-Norm erhalten.

Archive, Behörden und Verwaltungen sollen ebenso von diesem Standard profitieren wie Gerichte, Bibliotheken, Zeitungsverlage und Industrieunternehmen.

PDF/A soll ein mehrseitiges Dokumentenformat definieren, das eine Mischung aus Texten, Bilddaten und Vektorgrafiken enthalten kann. Ebenso sollen die Eigenschaften und Fähigkeiten der Systeme definiert werden, die für das Lesen, Reproduzieren und die Volltextsuche verwendet werden.

Ziel ist es sicherzustellen, dass PDF/A-Dateien alle zur Darstellung des Inhalts notwendigen Daten enthält und dass diese Daten selbst wieder internationalen Standards entsprechen und unabhängig von Ausgabegeräten sind. Das trifft auch auf die Metadaten-Ebene zu. PDF/A soll alle Metadaten enthalten können, die zur Beschreibung und Auswertung des Dokumentes notwendig sind. Dabei soll PDF/A noch konsequenter vorgehen als PDF/X.

Fazit

Eine Empfehlung für PDF als Format zur Langzeitarchivierung kann derzeit nicht gegeben werden. Allenfalls die ISO-normierten Varianten PDF/X und das noch in Entwicklung befindliche PDF/A kommen hierfür in Frage. PDF/X hat nicht primär die Langzeitarchivierung, sondern die Kompatibilität von Daten in der Druckvorstufe zum Ziel, die endgültige Form von PDF/A ist noch abzuwarten. Eine Empfehlung für diese PDF-Typen als Formate zur Langzeitarchivierung birgt zudem das Risiko in sich, dass die Einschränkung auf diese Sonderformen nicht hinreichend Beachtung findet.

Zudem stellt sich im Bereich der Langzeitarchivierung wissenschaftlicher Inhalte die Frage, ob ein Kompositformat, das die korrekte Wiedergabe von formatierten Inhalten vor allem in Print- und erst danach auch in elektronischen Medien zum Ziel hat, die richtige Wahl ist. Im wissenschaftlichen Bereich geht es meist nur sekundär um die typografisch korrekt umgesetzte und layoutete Form, es geht primär um die Auswertbarkeit der Inhalte. Auf diesem Gebiet sind semantisch strukturierbare Paketformate, bei denen unterschiedliche Inhaltsbestandteile wie Texte, Grafiken und ggf. wissenschaftliche Quelldaten im jeweils optimalen Datenformat separat abgelegt und untereinander durch ein Referenzsystem verknüpft sind, mit großer Wahrscheinlichkeit die bessere Lösung.

Quellen: *Hersteller-Sepzifikation: http://partners.adobe.com/asn/tech/ps/specifications.jsp. – ISO-Standards: PDF/X -3: ISO/IEC 15930-6:2003. – PDF/A: ISO/CD 19005-1 Electronic document file format for long-term preservation. Use of PDF 1.4 (PDF/A). Publication target date: 2006-09-01). – The Association for Suppliers of Printing, Publishing and Converting Technologies: www.npes.org. – The Association for Information and Image Management: www.aiim.org. – John Mark Ockerbloom, Archiving and Preserving PDF Files, In: RLG DigiNews, Vol. 5 No. 1, February 2001 (http://www.rlg.org/legacy/preserv/diginews/diginews5-1.html#feature2).*

Langzeitstabile Formate für Multimedia-Daten

Mit dem Schlagwort »multimedial« wird in der Regel der Umstand bezeichnet, dass etwas (eine Organisation, eine Technik, ein Gerät etc.) mehrere Medien beherrscht. Dabei ist meist die Beherrschung von Techniken gemeint, die über das klassische Printmedium mit seinen Komponenten Text und zweidimensionales Bild sowie über seine digitalen Wiedergabeformen hinausgehen. Es geht also auf den ersten Blick um Audiodaten und um bewegte, dreidimensionale oder Panoramabilder. In vielen Wissenschaften sind aber auch mehrdimensionale Datensätze Bestandteil der Dokumentation wissenschaftlicher Arbeit, deren Integration in wissenschaftliche Publikationen durch die digitale Form möglich wird.

Beispiele für multimediale Daten sind:

- Audio-Dateien
- Animationen
- Videos
- Virtuelle Räume
- dreidimensionale (beispielsweise chemische) Strukturen
- Spektren und Chromatogramme (multidimensionale Messdaten)

Insgesamt stehen für die Codierung und Speicherung derartiger Inhalte mehrere Dutzend Formate zur Verfügung. Wir beschränken uns im Folgenden auf die wesentlichen, allgemein relevanten Anwendungen, nämlich Audio- und Videodaten, Animationen und virtuelle Räume sowie Tabellen- und Datenbankdaten.

Quellen: *exzellente Informationsbasis für alle Datenformate zur Langzeitarchivierung ist die Website »Digital Formats for Library of Congress Collections« (http://www.digitalpreservation.gov/formats/)*

Audiodaten

Für Audiodaten gibt es eine Vielzahl von Formaten. Beschränkt man sich jedoch auf solche, die eine verlustfreie Kompression ermöglichen und als internationale Norm und/oder als verbreiteter Standard vorliegen, wird die Zahl sehr überschaubar.

Verlustlos speichern zunächst die proprietären Formate WAV (Microsoft und IBM) und AIFF (Apple Macintosh).

WAV (eigentlich RIFF WAVE) hält sich an das von Microsoft und IBM definierte »Resource Interchange Format« (RIFF) für Multimediadaten. WAV ist eigentlich der Sammelbegriff für verschiedene Unterformate, von denen PCM (Pulse Code Modulation) das gebräuchlichste ist und meist mit WAV gleichgesetzt wird. Bei WAV (PCM) handelt es sich um eine unkomprimierte Aufzeichnung von Soundsamples. WAV (PCM) Dateien sind eine Eins-zu-Eins-Kopie des Originals in voller CD-Qualität (44.1 kHz, 16 Bit, Stereo). Eine Minute in CD-Qualität benötigt etwa 10 MB Speicherplatz, die Dateien sind also verhältnismäßig groß, da keine Kompression stattfindet. Das Äquivalent bei Apple Macintosh-Rechnern ist AIFF (Audio Interchange File Format). Beide Formate lassen sich trotz ihrer proprietären Herkunft auf vielen Plattformen abspielen und kommen für die Langzeitarchivierung in Frage. Die Datenformate sind öffentlich dokumentiert.

LPAC (Lossless Predictive Audio Compression; Technische Universität Berlin), FLAC (Free Lossless Audio Codec; SourceForge.net) und Monkey's Audio (www.monkeysaudio.com) sind populäre Formate zur verlustfreien Kompression von WAV (PCM) Dateien (derer es auch noch ein Dutzend weitere gibt).

Die ISO-Norm für verlustlose Audio-Daten-Speicherung ist ISO/IEC 14496-3:2001/AMD 4, Audio Lossless Coding (ALS), zugleich Teil des MPEG-4-Standards (Mulitmedia-Standard der Moving Picture Experts Group). MPEG 4 ALS verwendet LPAC als Komprimierungsmethode. Die ISO-Norm befindet sich im Stadium Working Draft 3 (http://mpeg.nist.gov/mpeg/docs/68_Munich/wg11/w6435.zip; http://www.nue.tu-berlin.de/forschung/projekte/lossless/mp4als_d.html), die Publikation wird für 2005 erwartet. Es muss sich jedoch erst zeigen, ob sich diese Norm auch als Standard durchsetzt. Leider ist auch der LPAC-Quellcode bisher nicht frei verfügbar.

Bis sich zeigt, ob sich MPEG-4-ALS als Standard durchsetzen kann, empfiehlt sich die Archivierung der unkomprimierten WAV- oder AIFF-Daten oder ihre Speicherung mit einer verlustlosen Kompressionsmethode, deren Quellcode offen gelegt und die für möglichst viele Systeme verfügbar ist (z.B. FLAC).

Video und audiovisuelle Daten

Analoge Videodaten müssen digitalisiert werden, wobei die analogen Steuersignale in Einzelbilder (Frames) umgewandelt werden (Framegrabbing). Diese Einzelbilder können in verschiedenen Formaten gespeichert werden. Im Prinzip sind dabei alle Pixelgrafikformate möglich. In der professionellen Studiotechnik gibt es seit Jahren bewährte digitale Standardformate, die speziell für das Abspeichern von Videoframes verwendet werden. Im TV-Bereich ist das die Norm CCIR 601 (= ITU-R BT.601-4), der alle professionellen digitalen Band-Formate wie D1, D5 und Digital Betacam entsprechen. Im Bereich High-End-Video (HDTV – High Definition-TV und Digital Cinema) ist das der Norm ANSI/SMPTE 268M-1994 entsprechende dpx (Digital Moving Picture Exchange) Standard.

Diesen professionellen Formaten ist gemeinsam, dass die unkomprimierten Daten einen Speicherbedarf haben, der auch bei heutigen Festplattengrößen kaum handhabbar ist und bei der Bearbeitung spezielle Workstations erfordert. Bei Videodaten in CCIR 601 mit einer horizontalen Auflösung von 720 Punkten und einer vertikalen Auflösung von 576 Zeilen (DVD in PAL) ist ein unkomprimiertes Einzelbild 830 KB groß, eine Minute Video benötigt 1,26 GB. Eine Minute Video in dpx kann bis zu 72 GB Speicherplatz benötigen. Größere Datenmengen in diesen Formaten werden daher auf Magnetbändern gespeichert.

Bei der Übernahme von unkomprimierten digitalen Videodaten von Bändern in andere Speichersysteme wird daher meist eine Kompression vorgenommen. Bei verlustloser Kompression, die bei Videodaten nach der Huffman-Methode vorgenommen wird, können die Daten jedoch nur um den Faktor 2 (maximal 5) komprimiert werden.

Semiprofessionelle Kamera- und Aufnahmesysteme speichern in der Regel keine unkomprimierten Videodaten, sondern verwenden bereits bei der Aufzeichnung je nach Hersteller verschiedene Methoden zur Datenkompression, die bei den meisten Systemen mit einem Qualitätsverlust verbunden sind (Komprimierung 1:2 bis 1:10). Aber auch diese Daten sind meist noch zu umfangreich, um ohne weitere Kompression auf gängigen Computersystemen verarbeitet zu werden.

Für die in der Praxis notwendigerweise hohe aber verlustbehaftete Komprimierung von Videodaten ist seit 1994 MPEG-2 der etablierte Standard (= ISO 13818). Mit MPEG-2 können Video-Ausgangsdaten unterschiedlichster Qualität und Auflösung bei geringem Qualitätsverlust hoch komprimiert werden, natürlich in Verbindung auch mit mehrkanaligen Audiodaten (bei Kompression mit minimalem Qualitätsverlust ist die Kompressionsrate ca. 1:11; bei höherem Verlust bis zu 1:100). MPEG-2, ursprünglich für professionelle TV- und Studioanwendungen entwickelt, ist vor allem durch die Verwendung als Format für Video-DVDs bekannt geworden. Durch das Wachstum von Rechnerleistung und Speicherkapazitäten bei Standardrechnerkonfigurationen für den Büro- und Heimgebrauch ist MPEG-2 jetzt auch in diesem Bereich Standard. Es gibt daher eine Vielzahl von Programmen für alle Rechnersysteme, die MPEG-2 schreiben und darstellen können. Da MPEG-2 auch streaming-

fähig ist – eine Breitband-Internet-Anbindung vorausgesetzt -, wird es zunehmend auch als Video-Format im Internet verwendet. Der neue MPEG-4-Standard soll MPEG-2 nicht ersetzen, sondern ist für die Optimierung von Streaming-Video bei geringerer Bandbreite konzipiert (z.B. Video auf Handys).

Einen Kompromiss zwischen verlustloser und MPEG-2-Komprimierung stellt Motion JPEG (MJPEG) dar, bei dem jedes Bild separat als JPEG-Bild komprimiert wird. Mit MJPEG komprimierte Videos haben im Gegensatz zu MPEG-Videos eine von der Bewegung des Bildes unabhängige Qualität. Durch die individuelle Kompression der Bilder ist es bei diesem Format möglich, ein Video bildgenau zu schneiden, was bei MPEG nur beschränkt möglich ist. Es gibt jedoch zahlreiche herstellerabhängige MJPEG-Varianten, die zum Teil nicht untereinander kompatibel sind, was gegen die Verwendung dieses Formats für Archivierungszwecke spricht. Auch vom JPEG 2000 gibt es eine M-Variante für die Speicherung von Bewegtbildern, deren Verbreitungsgrad allerdings gering ist (als Norm ISO 15444-3:2002).

Die vielen weiteren Verfahren zur stärkeren Komprimierung von Videodaten sind für die Archivierung weniger interessant, da sie mit dem Ziel der besseren Übertragung der Videodaten über das Internet einen Qualitätsverlust in Kauf nehmen, der für Archivierungszwecke nicht tragbar ist.

Umfassender Standard für den Austausch von Videodaten und dazugehörigen technischen und inhaltlichen Metadaten in einem Datencontainer ist im professionellen Bereich MXF (Media Exchange Format), das auch als SMPTE-Normpaket 377M – 394M vorliegt. MXF ist kompatibel mit dem bisherigen Standard AAF (Advanced Authoring Format) und wird industrieweit unterstützt. Für die Codierung von Video- und Audiometadaten im XML-Format steht die Norm MPEG-7 (ISO/IEC TR 15938-8:2002/Amd 1) zur Verfügung, die als Ergänzung der übrigen MPEG-Standards, also auch MPEG-2, konzipiert ist.

Für die Langzeitarchivierung von semiprofessionell erzeugten Video-Daten ist die Archivierung im MPEG-2-Format optimal, gegebenenfalls mit MPEG-7-Metadaten. Sollen unkomprimierte HQ-Daten archiviert werden, wird die Verwendung der entsprechenden Standardformate (CCIR 601, dpx) in Verbindung mit verlustfreier Huffman-Komprimierung und MXF-Austauschcontainern der naheliegendste Weg sein.

Animationen und virtuelle Räume

Für animierte und interaktive Präsentationen (animierte Anleitungen, Trickfilme, Spiele, Formulare) im Internet hat der proprietäre Standard-Flash von Macromedia eine starke Position (Version 1 von 1996, aktuell Version 7 (Flash MX 2004)). Mit Flash werden vektorbasierte Grafiken erstellt und animiert, auch Pixelgrafiken, Text und Sound können integriert werden. Objekte einer Flash-Animation können auf (Mouse-)Eingaben reagieren und so skriptgesteuerte Aktionen veranlassen. Das Flash-Datenformat ist zwar binär und unter-

liegt der Kontrolle durch Macromedia, aber das Format ist dokumentiert und alternative Tools sind auf dem Markt.

Gegen den proprietären Flash-Standard hat das W3C zwei offene Alternativen gesetzt, SVG, von dem schon oben die Rede war, sowie SMIL. Das vektorbasierte SVG kann in ähnlichem Umfang wie Flash animiert und interaktiv gestaltet werden. Da SVG auf XML beruht, ist es durchsuchbar und leicht zu transformieren. Bisher hat SVG auf dem Gebiet der animierten und interaktiven Präsentationen nur eine beschränkte Verbreitung gefunden, da es sich gegen das viel früher angebotene und sehr weit verbreitete Flash durchsetzen muss. Es ist aber eine zunehmende Unterstützung von SVG durch Webbrowser und Grafikanwendungen zu beobachten.

Für die Erstellung interaktiver audiovisueller Präsentationen im Internet dient die W3C-Recommendation SMIL (ausgesprochen wie engl. smile; Synchronized Multimedia Integration Language, Version 1.0 von 1998 und 2.0 von 2001). SMIL gestattet es, durch eine XML-basierte Syntax die räumlichen und zeitlichen Beziehungen zwischen Medienobjekten genau zu spezifizieren und dem Nutzer ggf. die Interaktion mit der SMIL-Anwendung zu ermöglichen. Die SMIL-Multimedia-Anwendung kann dabei alle Arten von Medienobjekttypen – (Streaming-)Audio und Video, Grafiken, Text usw. – enthalten. Zur Verbindung von Medienobjekten werden diese räumlich und auf einer gedachten Zeitachse angeordnet. Die Verbreitung von SMIL ist derzeit noch stärker begrenzt als die von SVG.

In SMIL und SVG verwendete Medienobjekte werden anders als bei Flash nicht zu einer einzelnen Datei kompiliert, in die alle Medienobjekte eingebettet sind. Medienobjektdateien werden lediglich mittels ihres URI referenziert. Damit verbleiben die Medienobjekte außerhalb des SMIL- bzw. SVG-Dokumentes als individuelle Objekte auf einem Server und werden zur Laufzeit angefordert. Werden SMIL oder SVG-Animationen archiviert, müssen daher alle referenzierten Medienobjekte mitarchiviert werden.

VRML (Virtual Reality Modeling Language) ist eine Beschreibungssprache des Web3d-Consortiums für 3D-Szenen (erste Version 1994). In der Version von 1997 ist sie ISO-Norm (ISO 14772). VRML definiert Geometrien, Ausleuchtungen, Animationen und Interaktionsmöglichkeiten von 3D-Szenen. Ursprünglich wurde VRML als 3D-Standard für das Internet entwickelt, seine Anwendung erstreckt sich über 3D-Modelle und Simulationen in Industrie und Forschung bis zu Computerspielen. VRML ist inzwischen als ein Austauschformat von 3D-Modellen etabliert. Die Mehrzahl der 3D-Modellierungsprogramme ermöglichen den Im- und Export im VRML-Format.

Eine VRML-Darstellung wird in Echtzeit generiert. Die VRML-Anwendung berechnet jedes einzelne Bild aus den Geometriedaten sowie dem Verhalten und den Bewegungen des Benutzers in seiner Rolle als »Besucher«. Komplexe VRML-Szenen stellen hohe Anforderungen an die Hardware. Mit zunehmender Rechnerleistung werden immer realistischere Welten möglich.

VRML-Dateien tragen die Dateierweiterung ».wrl« (world), und sind als ASCII codiert, jedoch kein XML oder SGML. Zur Steuerung der VRML-Szenen kann die Programmiersprache Java verwendet werden.

VRML wird vom Web3d-Consortium als X3D (Extensible 3D) weiterentwickelt, wobei es sich bei X3D um eine XML-basierte Sprache handelt. Zusätzlich zu X3D-XML wird an einem binären Format gearbeitet. X3D steht kurz vor der Veröffentlichung als ISO-Norm 19775. Es kann mit Sicherheit davon ausgegangen werden, dass X3D eine ähnlich breite Unterstützung erhalten wird wie VRML.

Tabellenverarbeitungs-, Datenbank- und wissenschaftliche Formate

Zur Vermeidung der Nutzung von proprietären Formaten wie Microsoft Excel gibt es mehrere Wege. Die Speicherung von Tabellenverarbeitung als einfache ASCII-Textdaten mit Feldtrennern (CSV, Comma Separated Value oder TSV, Tab Separated Value) ist sicher der Weg mit der größten Abwärtskompatibilität, würde aber den Verlust von Metadaten bedeuten. In dieser Hinsicht sind die komplexeren ASCII-Austauschformate SYLK (Symbolic Link Format; dokumentiertes Austauschformat vom Microsoft) oder DIF (Data Interchange Format von Lotus, jetzt IBM) eher geeignet. Diese Formate stellen jahrzehntelang bewährte de-facto-Standards dar und werden von jeder Tabellenbearbeitung als Im- oder Exportformat unterstützt, haben jedoch ihre Grenzen (so kann immer nur ein Tabellenblatt abgespeichert werden).

Als XML-basierter Standard steht das von der OASIS (Organization for the Advancement of Structured Information Standards) vertretene Open Office XML zur Verfügung, das der ISO zur Evaluation als ISO-Norm vorgelegt wurde (Stand September 2004) und auch ein komplexes Tabellenverarbeitungsmodul enthält, das Daten anderer Systeme importieren und in das eigene XML-Format verwandeln kann.

Die dazugehörige OpenOffice Software ist kostenlos unter den Lizenzen LGPL (GNU Lesser General Public License) und SISSL (Sun Industry Standards Source License) für Windows und Linux/Unix-Systeme verfügbar.

Falls die Informationen mehr als zwei Dimensionen oder eine hierarchische Struktur haben, ist die Ablage in einfachen Spreadsheet-Formaten nicht mehr möglich. In dieser Domäne kommen im Wesentlichen nur die freien und im wissenschaftlichen Bereich weit verbreiteten systemunabhängigen Formate netCDF (network Common Data Form der US-amerikanischen University Corporation for Atmospheric Research) und HDF (Hierarchical Data Format des US-amerikanischen National Center for Supercomputing Applications) in Frage.

Wissenschaftliche Primärdaten als Grundlage wissenschaftlichen Arbeitens enthalten oft Informationen, deren Nutzen erst viel später erkannt wird. Zudem sind sie ein wichtiger

Schlüssel zum detaillierten Nachvollziehen von Forschungsergebnissen und damit auch der Identifizierung von methodischen Fehlern oder gar Fälschungen.

Die Archivierung von wissenschaftlichen Primärdaten kann aus verschiedenen Perspektiven betrachtet werden. Eine davon ist der Publikationskontext, der sich in der Regel aus der Veröffentlichung der Ergebnisse in einem Artikel in einem Fachjournal ergibt. Es ist daher eine Option, die Primärdaten, die hinter einer Publikation stehen, im Zusammenhang mit dieser Publikation zu archivieren. Das Thema »Archivierung wissenschaftlicher Primärdaten« kann daher im Kontext der LZA von E-Journals an Relevanz gewinnen. Die Feststellung, welche spezielleren wissenschaftlichen Formate für die Langzeitarchivierung relevant sind, kann nur im Dialog mit den wissenschaftlichen Fachgesellschaften getroffen werden. Hilfreich wäre gewiss auch eine Sensibilisierung der Datenproduzenten für dieses Thema.

Metadaten

Bei der Vielzahl von Inhaltstypen, Medientypen und Datenformaten, die bei der Langzeitarchivierung digitaler Inhalte verarbeitet werden müssen, ist es zu einer effizienten Verarbeitung erforderlich, über Informationen zu diesen Daten zu verfügen, ohne die Daten zum Erhalt dieser Informationen öffnen zu müssen. Daten über Daten nennt man auch Metadaten.

Das klassische Beispiel für Metadaten sind diejenigen Kataloginformationen, die einen Gegenstand in einem Archiv erst auffindbar machen. Diese Metadaten, die z.B. bei Printpublikationen über Titel, Autoren, Erscheinungsjahr etc. meist in einer normierten Form Auskunft geben, werden auch beschreibende Metadaten genannt.

Für die Langzeitarchivierung digitaler Inhalte sind weitere Metadaten notwendig, die über technische, administrative und urheberrechtliche Aspekte Auskunft geben.

(Es sei angemerkt, dass die in diesem Abschnitt verwendete Systematik der Metadaten keine normative ist. Verschiedene Standards verwenden für dieselbe Art der Metadaten oft unterschiedliche Bezeichnungen oder weitere Unterteilungen.)

Metadatenstandards

Für die Codierung dieser Daten gibt es eine Reihe von Standards, unter denen sich in den letzten Jahren vor allem XML-basierte durchgesetzt haben. Standards erleichtern nicht nur den Datenaustausch, sondern sie ermöglichen auch erst die automatische Bearbeitung der Daten im Archiv. Standardisierte Metadaten im XML-Format können automatisch ausgelesen werden. So kann ein beim Archiv eingegangenes E-Journal mittels seiner beschreiben-

den Metadaten automatisch in den Katalog aufgenommen werden. Ebenso kann den technischen Metadaten entnommen werden, ob zu einem E-Journal-Artikel Daten gehören, die wegen der technischen Entwicklung nicht mehr ohne weiteres dargestellt werden können und deshalb in ein aktuelles Format transformiert werden müssen, was ebenfalls automatisch geschehen kann.

Im Rahmen der Übermittlung eines E-Journal-Artikels vom Produzenten an das Archiv ist es vorteilhaft, wenn möglichst viele Metadaten übermittelt werden können.

Da es den Rahmen dieser Expertise sprengen würde, wird hier darauf verzichtet, die verschiedenen Standards im Einzelnen vorzustellen. Zum besseren Verständnis der nachfolgenden Erörterungen soll jedoch eine kurze Begriffsbestimmung stattfinden.

Beschreibende Metadaten

Diese Metadaten entsprechen im Wesentlichen dem, was man unter bibliographischen Angaben versteht. Bei klassischen bibliographischen Angaben findet jedoch meist keine Trennung in allgemein beschreibende und technische Aspekte statt, hier sind auch mitunter Angaben über das Format eines gedruckten Bandes, die Heftung, das Material u.ä. enthalten. Dagegen enthalten beschreibende Metadaten für digitale Inhalte keine Auskünfte über technische Aspekte wie Datenformate u.ä. Dort sind solche technischen Informationen Bestandteil der technischen Metadaten.

Technische Metadaten

Technische Metadaten enthalten Informationen, die darüber Auskunft geben, mit welcher Hard- und Software Daten geöffnet werden müssen, um korrekt bearbeitet werden zu können.

Beispielsweise können Datenformate oft nicht automatisch erkannt werden. Wer schon einmal als Nutzer eines PC Daten von einem Apple Macintosh erhalten hat, kennt eventuell das Problem, dass Dateien wegen unterschiedlicher Speicher- und Namenskonventionen zunächst nicht erkannt und geöffnet werden konnten, obwohl geeignete Software auf dem Rechner vorhanden ist. Fehlt die benötigte Software, kann ein Rechner selbst dann nichts mit den Daten anfangen, wenn die Daten auf dem gleichen Hardware-System und unter dem gleichen Betriebssystem erzeugt wurden. Und selbst dann, wenn eine Datei erfolgreich geöffnet werden kann, heißt das nicht notwendigerweise, dass ihre Auswertung oder Darstellung optimal erfolgt. Bei Grafikdaten oder Seitenbeschreibungssprachen wie PDF sind gegebenenfalls Angaben über das Farbprofil oder über Besonderheiten der Erstellungsgeräte notwendig, um eine optimale Verarbeitung zu ermöglichen.

Zur Identifikation eines Datenformats und Übermittlung weiterer technisch relevanter Daten werden daher Metadaten verwendet, die nicht Bestandteil der Datei sind, welche die eigentlichen Inhalte trägt. Es gibt bei verschiedenen Datenformaten zwar auch die Möglichkeit, technische Metadaten gemeinsam mit den Inhalten in einer Datei abzulegen, was aber zu Problemen führen kann, wenn das Datenformat nicht bekannt ist. In solchen Fällen wird für die Langzeitarchivierung eine externe Speicherung der Metadaten sinnvoll sein. Welche Metadaten notwendig sind, hängt vom jeweiligen Medientyp und Datenformat ab.

Ein interessantes Szenario, bei dem eine zentrale Stelle für wichtige Datenformate regelmäßig überprüft, ob Bestandserhaltungsmaßnahmen notwendig werden und welche dafür am besten geeignet sind, schlagen Hunter und Choudhury vor (http://metadata.net/newmedia/Papers/JCDL2004_paper.pdf). Mittels eines regelmäßigen Vergleichs der technischen Metadaten archivierter Objekte mit den Daten dieser zentralen Stelle über das Internet wird die Notwendigkeit von Bestandserhaltungsmaßnahmen signalisiert. Bei diesem Szenario kann der Aufwand, der für eine regelmäßige Überprüfung der verschiedenen Datenformate auf Notwendigkeit und Art von Bestandserhaltungsmaßnahmen erforderlich ist, auf mehrere Parteien, die diese zentrale Überprüfungsstelle gemeinsam betreiben, verteilt und damit akzeptabel gemacht werden. Hier käme der Standardisierungsvorteil besonders zum Tragen.

Administrative Metadaten

Administrative Metadaten geben Auskunft über verwaltungstechnische Vorgänge, die die Daten durchlaufen oder über Beziehungen, die zwischen ihnen durch solche Vorgänge entstanden sind. Beispielweise würde das Datum der letzten Transformation der Daten zu den administrativen Metadaten zählen oder auch die Angabe, welche Datei, die nach einer Transformation ggf. in verschiedenen Versionen vorhanden ist, das Original ist und welche aus dem Original erzeugt wurden. Zu den administrativen Metadaten gehören auch Angaben zur Sicherung der Unversehrtheit und Authentizität der Daten, zum Beispiel Prüfsummen, anhand derer erkannt werden kann, ob eine Datei nach einem Bearbeitungsvorgang noch unverändert ist.

Welche Angaben hier notwendig sind, hängt von der konkreten Gestaltung der Datenverwaltung in einem Archiv ab.

Urheber- und nutzungsrechtliche Metadaten

Diese Art der Metadaten informiert über die vielfältigen Aspekte der rechtlichen Beschränkung von Nutzung und Zugang zu den Daten. Hierunter fallen Copyright-Angaben ebenso wie Sperrvermerke, wenn beispielsweise die digitale Ausgabe eines Printjournals erst nach einem vom Produzenten definierten Zeitraum verfügbar gemacht werden darf.

In diesen Bereich fallen auch die Angaben zum Digital Rights Management, das durch technische Mittel verhindern will, dass ein Nutzer mit den Daten Vorgänge ausführt, die der Produzent ausschließen möchte, beispielsweise die Vervielfältigung oder Weitergabe.

Welche Angaben hier notwendig sind, hängt von der konkreten Gestaltung der vertraglichen Regelungen zwischen Produzenten und Archiv ab.

Quellen: *Die wichtigsten Metadaten-Standards mit ihren Homepages werden auf folgenden Websites aufgelistet: http://metadata.net. – http://www.ifla.org/II/metadata.htm. – http://www.loc.gov/standards/mets/*

Besonderheiten von E-Journals

Besonderheiten des Aufbaus des Medientyps E-Journal werden ausführlich in Teil 2 im Abschnitt »Informationseinheiten in E-Journals als SIP« behandelt. In E-Journals auftretende Datenformate sind unter anderem Gegenstand von Teil 3 und werden dort in den Abschnitten »Datenformate I: Textdaten und Grafiken«, »Datenformate II: Multimediale Elemente« und »Datenformate III: Dynamische Elemente« behandelt.

Teil 2

Das OAIS-Referenzmodell

Neben Standards zur Speicherung von Daten und Metadaten sind für die Langzeitarchivierung auch Standards für den Umgang mit diesen Daten notwendig. Dabei sind Fragen zu klären wie: in welcher Form und auf welchem Weg sollen zusammengehörige Dateien eines E-Journal-Artikels zum Archiv gelangen? Was geschieht mit den Daten dort, bevor sie archiviert werden? Welche Abläufe sind im Archiv erforderlich, um die Daten verfügbar zu halten? In welcher Form werden die Daten vom Archiv an Nutzer ausgegeben usw. usf.?

Um für diese Abläufe ein Grundkonzept zur Verfügung zu stellen, das grundlegende Definitionen für Abläufe und die dazugehörige Terminologie enthält, hat die NASA die Entwicklung eines Modells für digitale Archive angestoßen, das zu einem internationalen Standard geworden ist. Unter Beteiligung einer Reihe von bedeutenden Forschungszentren legte das internationale Beratungskomitee für Weltraumdatensysteme (Consultative Committee for Space Data Systems; CCSDS) 1999 den Entwurf eines Referenzmodells für ein Offenes Archivisches Informationssystem vor (OAIS). Dieser Entwurf ist inzwischen zur ISO-Norm gediehen.

Er soll hier soweit vorgestellt werden, wie es für die Konzeption eines Datenübertragungspaketes vom Produzenten an das Archiv notwendig ist.

Quellen: *OAIS ISO/IEC-Norm 14721:2003. – Spezifikation des Consultative Committee for Space Data Systems (zugleich Quelle der in diesem Abschnitt verwendeten Grafiken):*
http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf. – CCSDS-Website für OAIS mit Links auf vielfältige Informationen zu OAIS: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

OAIS Archivdefinition

Ein *archive* im Sinne des OAIS ist ein Archiv, das

- eine Organisation von Menschen und Systemen ist und das verantwortlich ist
- für die Aufbewahrung von Informationen
- über lange Zeiträume und
- die Verfügbarmachung dieser Informationen für eine bestimmte Nutzergemeinschaft.

Entspricht diese Definition auch der in Deutschland üblichen Definition eines Archivs, so ist sie doch im angelsächsischen Sprachraum notwendig, da das Englische mit dem Wort *archive(s)* sowohl die Registratur bezeichnet (also allgemein eine Schriftgut verwaltende Einrichtung) als auch das Archiv im engeren Sinne (also eine Einrichtung, die für die langfristige Aufbewahrung von Schriftgut u.ä. zuständig ist). Aber auch die Abgrenzung von rein technisch orientierten Archivlösungen im Gegensatz zu einer verantwortlichen Organisation ist relevant.

Der Begriff »open« bedeutet hier, dass die Entwicklung des Systems sich im öffentlichen Raum abspielt und nicht Eigentum eines Unternehmens ist, meint jedoch nicht den uneingeschränkten öffentlichen Zugang zu einem Archiv.

Das Konzept des Referenzmodells

Das OAIS-Referenzmodell ist eine schematische Darstellung von archivalischen Abläufen

- zur Beschreibung und zum Vergleich von Aufbau und Arbeitsweise von Archiven
- zum besseren Verständnis archivalischer Elemente und Verfahren für die langfristige Erhaltung und Nutzbarmachung digitaler Informationen
- für nichtarchivalische Einrichtungen, um am Erhaltungsprozess sinnvoll teilnehmen zu können
- für den Vergleich der Datenmodelle digitaler Informationen, die von Archiven aufbewahrt werden, und für die Diskussion, wie Datenmodelle und Informationen sich im Lauf der Zeit verändern können
- für die Entwicklung von auf das OAIS bezogenen weiteren Standards

die die langfristige Erhaltung auch von nichtdigitalen Informationen einbeziehen kann.

Das *OAIS-Referenzmodell*

- ist logisch strukturiert und unabhängig von konkreten Implementationen oder Archivierungsstrategien (Migration, Emulation usw.),
- lässt sich mit UML (Unified Modeling Language; ein Standard zur Darstellungen von logischen Abläufen) grafisch darstellen, und zwar von der Ebene einfachster Übersichten bis zum komplexen Detailplan
- kann aus der Sicht der Funktionalität wie auch des Informationsflusses dargestellt werden
- betrachtet den Informationsfluss als Abfolge von aufgabenorientierten Informationspaketen

- ist primär für die Verarbeitung digitaler Informationen gedacht, kann jedoch auch auf nichtdigitale Informationen angewendet werden und erlaubt damit die Verarbeitung hybrider Unterlagen.

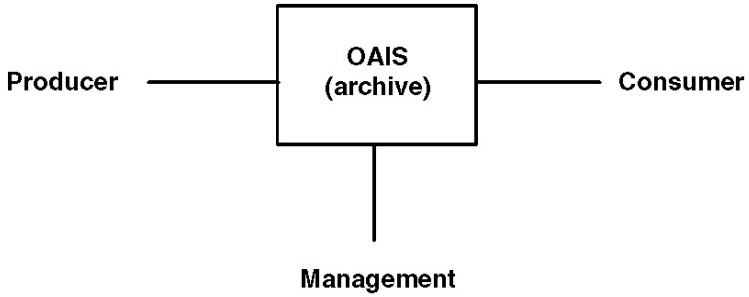


Abb. OAIS-Umgebung

Die Abbildung zeigt das OAIS-Archiv in seiner Umgebung (Environment). Die Produzenten (Producer) von Dokumenten liefern diese an das Archiv, das wiederum stellt die Dokumente den Nutzern (Consumer) zur Verfügung. Die Beziehungen des Archivs zu Lieferanten und Nutzern werden durch ein Regelwerk gesteuert, das vom Archivträger (Management) festgelegt und dessen Einhaltung von ihm kontrolliert wird. Produzenten können z.B. Verlage sein, das Management die Verwaltung einer wissenschaftlichen Bibliothek, die für das Archiv zuständig ist, und die Consumer die Benutzer dieser Bibliothek.

OAIS-Funktionsbereiche

Ein OAIS-Archiv besteht aus sieben Funktionsbereichen:

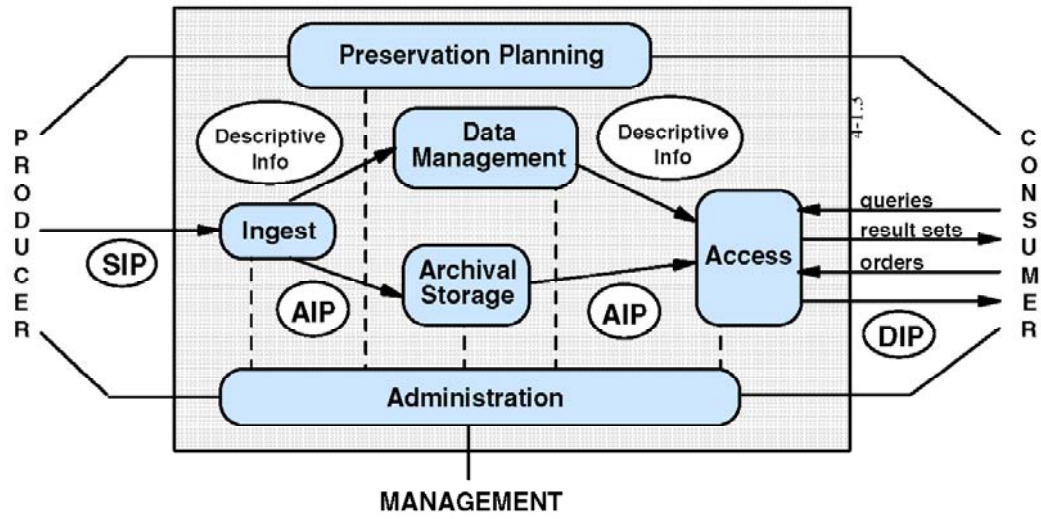


Abb. OAIS-Funktionsbereiche

1. Übernahme (Ingest)

Bei der Übernahme (Ingest) werden Informationseingangspakete (Submission Information Packages, SIP) vom Produzenten (Producer) eingeliefert und in Informationsarchivierungspakete (Archival Information Packages, AIP) umgewandelt, die den archivinternen Regeln entsprechen (z.B. den Regeln für Datenformate, für die Form der Datendokumentation etc.). Diese Pakete können unterschiedlichster Natur sein, es kann sich je nach Art des zu archivierenden Gutes um »echte« Pakete mit Printprodukten, CD-ROMS, Disketten, Magnetbändern etc. handeln oder um rein digitale Datenpakete, die per Datenfernübertragung eingeliefert werden.

2. Datenverwaltung (Data Management)

Die zum Eingangspaket gehörige Beschreibung (Descriptive Information) wird dabei in die Datenverwaltung (Data Management) übernommen, z.B. für die Katalogdatenbank des Archivs. Die Datenverwaltung verwaltet neben den beschreibenden Informationen, die die Dokumente identifizieren, auch die Beziehungen zwischen Dokumenten und dazugehörigen Archivierungspaketen und stellt Informationen für den Nutzerzugang (Access) zur Verfügung.

3. Archivspeicher (Archival Storage)

Bei erfolgreicher Übernahme wird das Archivierungspaket in den Archivspeicher (Archival Storage) eingestellt. Der Archivspeicher ist für die Aufbewahrung und Erhaltung der Archivierungspakete zuständig. Er sorgt bei digitalen Daten beispielsweise für Backups, prüft regelmäßig die Integrität der Daten, verfügt über Wiederherstellungsmechanismen bei Datenverlusten und stellt die Daten für die Nutzung zur Verfügung.

4. Archivverwaltung (Administration)

Die Archivverwaltung (Administration) ist für die Steuerung der Gesamtabläufe im OAIS und seiner Außenbeziehungen zuständig. Dazu zählt u.a. die Festlegung des Aufbaus der Informationspakete und der Regeln für Kommunikation und Datenaustausch mit Produzenten und Nutzern, die Steuerung der internen Abläufe sowie die Koordination von Aufbau und Wartung der technischen Infrastruktur.

5. Bestandserhaltungsplanung (Preservation Planning)

Zu den Aufgaben der Archivverwaltung gehört auch die Planung von Maßnahmen zur Bestandserhaltung (Preservation Planning). Diese umfasst nicht nur die Sicherung der Be-

stände in der Form, in der sie ursprünglich als Archivierungspaket übernommen wurden, sondern auch die Sicherung der Nutzbarkeit der Daten. Die Archivverwaltung muss also in regelmäßigen Abständen die Verwendbarkeit der archivierten Daten durch die Nutzer prüfen und bei absehbaren Problemen, wie z.B. anstehende Generationswechsel von Hard- oder Software, entsprechende Maßnahmen ergreifen (z.B. Migration der Daten oder Emulation der Software).

6. *Nutzerzugang (Access)*

Der Nutzerzugang (Access) nimmt Anfragen der Nutzer (Queries) entgegen, gibt nach Auswertung der Kataloginformationen (Descriptive Information) in der Datenverwaltung die Suchergebnisse (Result Sets) zurück und liefert bei Anforderung (Orders) die gewünschten Dokumente in Form von Informationsverteilungspaketen (Dissemination Information Packages, DIP) an den Nutzer aus. Verteilungspakete werden durch die Verbindung von Archivierungspaketen mit den Kataloginformationen erstellt, wobei DIPs nur die für den Nutzer relevanten Informationen enthalten.

7. *Allgemeine Dienste (Common Services)*

Hinzu kommen schließlich die allgemeinen Dienste (Common Services), deren Aufgaben im Wesentlichen aus dem Aufbau und der Wartung der technischen Infrastruktur bestehen (Systemadministration, Netzwerkverwaltung, Sicherheitsmaßnahmen).

Daten und Informationen

Die zu archivierenden Objekte werden im OAIS-Modell als »Informationen« angesehen, wobei zwischen »Informationen« und »Daten« unterschieden wird. Die Abbildung zeigt die Erzeugung von Informationen aus Daten. Gemäß der OAIS-Definition von Informationen sind Daten (Data Objects) noch keine Informationen. So ist z.B. eine Datenbank mit Zahlenwerten ohne Angaben, was diese Zahlen bedeuten, unverständlich und wertlos. Erst die inhaltliche Dokumentation (Representation Information) der Datenbank ermöglicht eine sinnvolle Interpretation der Daten. Repräsentationsinformation ist mithin die Information darüber, welche Daten wo und wofür stehen. Daraus folgt die Formel »Datenobjekte interpretiert unter Nutzung ihrer Repräsentationsinformation ergeben Informationsobjekte«:

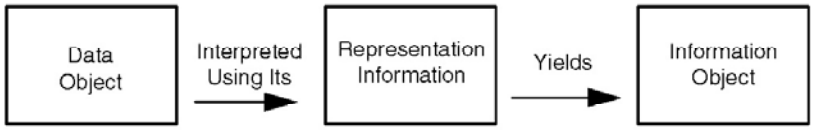


Abb. OAIS –Daten und Informationen

Information Packages

Informationen werden im OAIS-Modell stets in Form von Paketen (Packages) gehandhabt, die es ermöglichen, den eigentlichen Inhalt, also die zu archivierenden Informationsobjekte (Content Information, CI) mit Informationen zu verbinden, die für die weitergehende Einordnung und Nutzung des Inhalts erforderlich sind (Preservation Description Information; PDI):

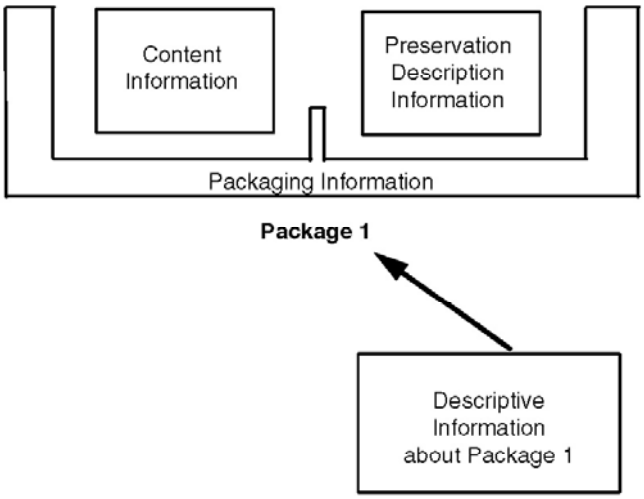


Abb. OAIS – Aufbau eines Information Packages

Die Preservation Description Information wird in vier Teile unterteilt:

- Quellenangaben (Provenance)
- Kontextangaben, wie z.B. die Zugehörigkeit einer Grafik zu einem Text, der in einem anderen Information Package enthalten ist (Context)
- Referenzinformationen, wie z.B. eine Bezeichnung des Objektes (IDs, Objektnummern etc.), mittels derer von anderen Objekten auf die Grafik verwiesen wird (Reference)
- Unversehrtheitsangaben, wie z.B. eine Checksumme, anhand der überprüft werden kann, ob die Grafik z.B. nach einer Übertragung der Datei von einem Speicherort zum anderen noch unverändert ist (Fixity)

Content Information und Preservation Description Information werden in einer Hülle (Pakkaging Information) zusammengefasst. Um das Paket finden zu können, muss auch eine Beschreibung des Pakets selbst (Descriptive Information) vorhanden sein.

Um dies an einem Beispiel verständlich zu machen, stelle man sich als Content Information eine Grafikdatei (Data Object) und eine Textdatei mit der dazugehörigen Bildlegende (Re-

presentation Information) vor. Die Angaben über Quellen, Referenzen usw., also die Preservation Description Information können wiederum in einer Textdatei, z.B. in XML-Form, abgelegt sein.

Nehmen wir weiterhin an, Grafikdatei und Bildlegendendatei befänden sich in einem gemeinsamen Ordner namens Content_Information, die XML-Datei mit der Preservation Description Information in einem eigenen Ordner namens Preservation_Description_Information. Diese beiden Ordner wiederum sind abgelegt in einem gemeinsamen Überordner namens Package_1. All das wird in eine ZIP-Datei verpackt. Die Information, die die ZIP-Datei über ihren Inhalt enthält, also Typ, Namen, Größe der Dateien und Verzeichnisstrukturen, ist die Packaging Information (diese Information kann natürlich ebenso gut in einer gesonderten Datei z.B. im XML-Format aufgezeichnet werden).

Um dieses Datenpaket im Archiv zu finden, bedarf es schließlich auch seiner inhaltlichen Katalogisierung. Die Informationen, die dazu benötigt werden, werden Descriptive Information genannt. Diese können sich z.B. in einer Textdatei befinden, die zusammen mit der ZIP-Datei an das Archiv geliefert wird. Dort können diese Angaben z.B. in eine Datenbank übernommen werden.

Das OAIS-Modell unterscheidet drei Typen von Information Packages, je nachdem, ob es sich um Daten handelt, die vom Produzenten in das Archiv eingeliefert werden, die im Archiv aufbewahrt werden oder die an den Nutzer aus dem Archiv ausgeliefert werden (die folgende Abbildung stellt die Taxonomie der Information Packages dar):

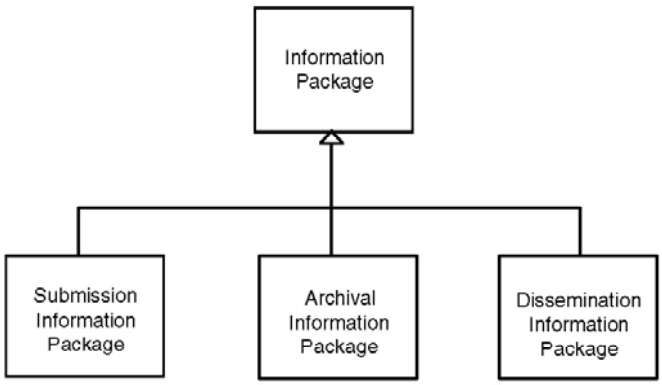


Abb. OAIS – Taxonomie von Information Packages

Wenn sie auch alle dem Modell eines Information Packages entsprechen, können sie sich im konkreten Fall in der technischen und organisatorischen Form und im Inhalt unterscheiden:

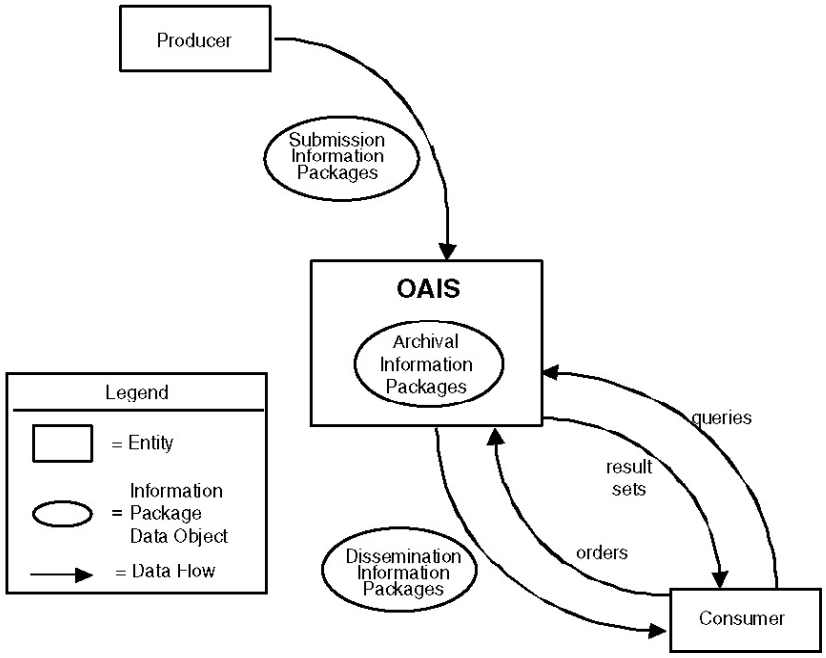


Abb. OAIS –Typen von Information Packages im Kontext

So kann beispielsweise das ins Archiv eingelieferte Package (Submission Information Pakkage; SIP) als gepackte ZIP-Datei eingeliefert werden, aber im Archiv (Archival Information Package; AIP) in Form von Datenobjekten in einer Datenbank aufbewahrt und in wieder anderer Form, z.B. als HTML-Datei mit verknüpften Grafik-Dateien, an den Nutzer ausgeliefert werden (Dissemination Information Package; DIP). Auch können Teile der Preservation Description Information oder der Descriptive Information erst im Archiv hinzugefügt werden (z.B. Checksummen oder bibliographische Systematisierung), welche wiederum nicht an den Nutzer mit ausgeliefert werden müssen.

Während das OAIS-Modell lediglich feststellt, dass es diese drei Typen gibt, liegt es bei jedem Anwender, die konkreten Parameter und Vorgehensweisen zu definieren.

Ingest im Detail

Im Folgenden sollen die zentralen Prozesse der Annahme und Übernahme von Daten in das OAIS (der Ingest) genauer beschrieben werden, da sie für die Konzeption eines SIP besonders wichtig sind. Das OAIS-Konzept schreibt nicht vor, dass diese Prozesse vollautomatisch ablaufen müssen. Manuelle und teilautomatisierte Ausführung ist ebenfalls möglich.

Entscheidend ist, dass der Ablauf der Prozesse konzeptuell und im konkreten Anwendungsfall auch im Detail definiert ist. Ein SIP sollte so gestaltet sein, dass es diese Prozesse grundsätzlich ermöglicht. Im Falle von digitalen Daten ist es natürlich naheliegend, die SIPs so zu gestalten, dass die Ingest-Prozesse so weit automatisiert ablaufen wie möglich. Das ist jedoch eine Zusatzanforderung zum OAIS-Modell.

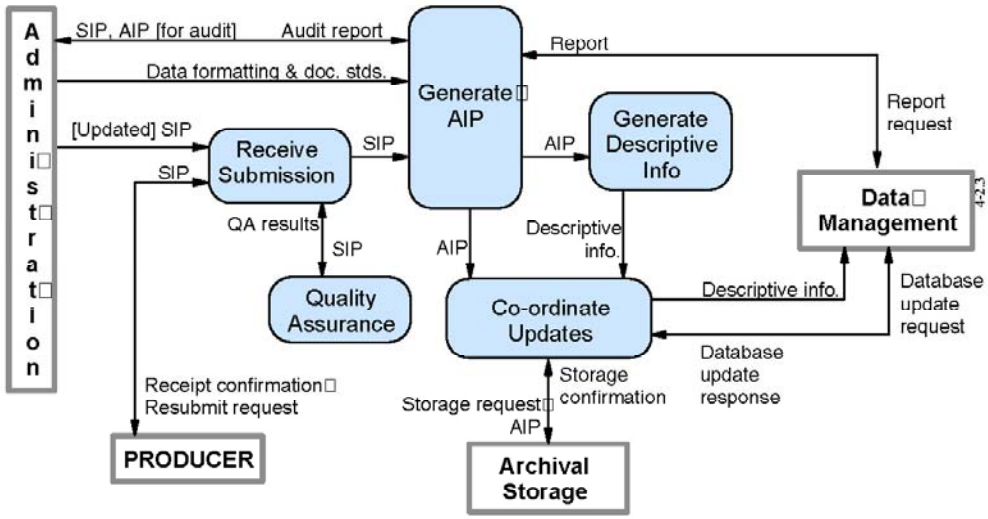


Abb. OAIS-Prozesse

Einsendungsempfang (Receive Submission)

Der Einsendungsempfang stellt die benötigte Speichertechnik zur Verfügung, um SIPs vom Produzenten (oder der Verwaltung) entgegenzunehmen. Digitale SIPs können durch Datenfernübertragung (z.B. FTP), durch das Überspielen von eingesendeten Datenträgern (Bändern, Disketten) oder durch direkte Aufnahme in das System (z.B. von CD-ROMS in ein CD-basiertes Archiv) übernommen werden. Für nichtelektronische Ressourcen können bewährte Aufnahmeverfahren verwendet werden.

Eingegangene Einsendungen können schon beim Einsendungsempfang auf Vollständigkeit, Unversehrtheit und Einhaltung von Regeln für die Einsendung (z.B. Verwendung bestimmter Datenformate oder Metadatenformate) überprüft werden. Fällt die Prüfung negativ aus, kann die Übernahme verweigert und der Produzent um die Einsendung eines neuen, korrigierten SIP gebeten werden. Bei positivem Ergebnis kann eine Empfangsbestätigung versendet werden.

Der Einsendungsempfang kann zudem mit rechtsverbindlicher Übergabe der Aufbewahrungszuständigkeit an das Archiv einhergehen.

Qualitätssicherung (Quality Assurance)

Die Qualitätssicherung ermöglicht die Kontrolle des Transfers des angenommenen SIP in die Bearbeitungsphase. Bei digitalen Einsendungen kann dieser Schritt die Bildung von Checksummen für jede Datei oder die Erzeugung von Logfiles beinhalten, die den Transfer von Dateien oder Medien und Lese-/Schreibfehler bei der Übernahme und Weitergabe dokumentieren.

AIP-Generierung (Generate AIP)

Die AIP-Generierung transformiert ein oder mehrere SIPs in ein oder mehrere AIPs, die den Standards des Archivs für Datenformate und Dokumentation entsprechen. Dieser Schritt kann Konvertierungen von Datenformaten, Datenrepräsentationen oder die Reorganisation der Content Information der SIPs umfassen. Die AIP-Generierung kann Anfragen an das Datenmanagement senden, um die Descriptive Information zu erhalten, die sie für die Erstellung eines AIP benötigt. Weiterhin übergibt sie die Datenpakete an die Archivverwaltung zur letzten Überprüfung der Einhaltung von Archivstandards (Audit) und empfängt einen Audit-Report. Die Überprüfung der Archivstandards kann weit über eine rein technische Prüfung hinausgehen und die Einbeziehung externer Fach-Gremien erfordern. Bei Nichteinhaltung kann das SIP wieder an den Einsendungsempfang übergeben werden, der den Produzenten informiert und gegebenenfalls zur Nachbesserung auffordert. Bei positivem Audit-Ergebnis erfolgt die endgültige Übernahme ins Archiv.

Generierung der Descriptive Information (Generate Descriptive Information)

Die Generierung der Descriptive Information extrahiert beschreibende Informationen aus den AIPs und sammelt gegebenenfalls weitere beschreibende Informationen von anderen Quellen, um diese für die Koordinierung von Updates und das Datenmanagement zur Verfügung zu stellen. Die Descriptive Information schließt z.B. Metadaten für Such- und Updatefunktionen ein oder z.B. auch Vorschaubilder von archivierten Objekten für das »Blättern« (browsing) durch einen Datenbestand.

Koordinierung von Datenmanagement-Updates (Coordinate Updates)

Dieser Prozess ist für die Übergabe der AIPs an den Archivspeicher und der Descriptive Information an das Datenmanagement zuständig. Dabei wird die bestehende Descriptive Information des zu übernehmenden AIP um Angaben zum Speicherort erweitert und dem Datenmanagement übergeben. Falls das neue AIP sich im Kontext anderer AIP befindet, müssen die entsprechenden Descriptive Informations ebenfalls aktualisiert werden. Um ein Update handelt es sich also insoweit, als dass ein AIP weitere Teil-AIPs enthalten kann, die

sukzessive hinzugefügt werden (beispielsweise werden die Informationen zu planetaren Forschungsmissionen der NASA im OAIS-Archiv Planetary Data System nach Zielobjekten gespeichert, so dass ein entsprechendes AIP bei Eingang von neuen Daten zu dem Objekt um neuere Informationen in Form von Sub-AIPs erweitert wird).

Das SIP-Konzept im Detail

Das Submission Information Package ist das Paket, in dem die zu archivierenden Informationen vom Produzenten zum OAIS gelangen. Die Form und der Inhalt im Einzelnen werden vom Archiv und den Produzenten gemeinsam in einem Submission Agreement definiert. Der Grad der Standardisierung kann dabei ganz unterschiedlich ausfallen. Im einen Extrem können Inhalt und Form des Packages ganz ohne Abstimmung mit dem OAIS vom Produzenten festgelegt werden. Das erschwert unter Umständen die Archivierung und die spätere Nutzbarmachung, da eine Automatisierung der Verarbeitung bei individuellen Datenpaketen wegen des unverhältnismäßig hohen Aufwandes nicht oder nur eingeschränkt möglich ist. Daher werden Archiv und Produzenten bei größeren und/oder in ähnlicher Struktur immer wieder anzutreffenden Datenmengen bestrebt sein, möglichst exakte Regeln für den Aufbau der entsprechenden SIPs festzulegen. Dies ermöglicht im anderen Extrem eine zuverlässige vollautomatische Weiterverarbeitung der Daten in allen weiteren OAIS-Modulen.

Wie schon oben erwähnt, kann ein SIP Content Information und Preservation Description Information enthalten, die sich jeweils aus Data Objects und Representation Information zusammensetzen. Zusammengehalten wird das Package von der Packaging Information und beschrieben von der Descriptive Information.

Es ist bei SIPs aber keineswegs zwingend notwendig, dass alle diese Bestandteile in jedem SIP vorkommen oder gar, dass das SIP eine vollständige, inhaltlich in sich abgeschlossene Informationseinheit bildet. Abhängig von der Art der zu archivierenden Daten können SIPs auch nur Teilinformationen beinhalten, die erst vom Archiv zu einem vollständigen Archival Information Package zusammengesetzt werden. Die Teilinformationen können von Fall zu Fall ganz unterschiedlich ausfallen. So ist es beispielweise möglich in manchen SIPs nur Data Objects zu liefern und in anderen nur die dazugehörige Representation Information und in wieder anderen nur die Preservation Description Information.

Solche Teillieferungen können natürlich nicht willkürlich erfolgen. Ihr Aufbau muss zwischen OAIS und Produzenten abgestimmt sein. Dabei muss vor allem gewährleistet sein, dass über die immer notwendige Packaging Information und Descriptive Information die Teildaten korrekt zusammengeführt werden können.

Die zu einem SIP gehörige Descriptive Information kann gemeinsam mit dem SIP oder auch separat an das OAIS gelangen. Im einfachsten Fall handelt es sich um eine einfache Textinformation, ebenso ist auch als XML strukturierte, komplexe Information möglich.

Das SIP kann vom Produzenten geschickt werden oder vom OAIS beim Produzenten abgeholt werden (sog. harvesting). Bei digitalen Daten wird die Übertragung als Data Submission Session bezeichnet. Auch für den technischen und zeitlichen Ablauf dieser Sessions müssen Vereinbarungen zwischen Archiv und Produzenten getroffen werden, um eine fehlerfreie und vollständige Übertragung der Daten zu gewährleisten.

Im OAIS werden beim Ingest SIPs zu AIPs konvertiert, wobei es je nach der Art der zu archivierenden Daten sowohl möglich ist, aus einem SIP genau ein AIP zu generieren, oder aus mehreren SIPs ein AIP oder aus einem SIP mehrere AIPs. Dabei muss auch die Preservation Description geändert werden. Beispielweise könnte ein SIP eines E-Journals alle Artikel eines Jahrgangs enthalten, die im Archiv in mehrere AIPs – eines pro Artikel – aufgeteilt werden, oder auch umgekehrt.

Informationseinheiten in E-Journals als SIP

Jahrgang und Nummer

Bei der Anwendung des SIP-Konzeptes auf den Publikationstyp E-Journal stellt sich die Frage, welche Einheiten innerhalb dieses Publikationstyps in ein SIP gefasst werden sollen. Soll ein Jahrgang, eine Nummer oder ein Artikel als zu archivierende Einheit gelten? Bei E-Journals, die nur eine digitale Zweitverwertung eines identischen Printproduktes darstellen, welches sich schließlich auch als Jahrgangsband in den Bibliotheken wiederfindet, scheinen Jahrgang oder Nummer geeignete Einheiten zu sein.

Artikel als zentrale Informationseinheit

Die zentrale Informationseinheit in Journals ist jedoch der Artikel, also ein kurzer bis mittellanger Beitrag, der einen in sich geschlossenen inhaltlichen Zusammenhang bildet (Fortsetzungsartikel wären daher als ein mehrteiliger Artikel zu verstehen). Artikel können den unterschiedlichsten Inhalt und Aufbau aufweisen, teilweise mit Abbildungen (und bei E-Journals auch anderen nichttextuellen Bestandteilen), mitunter auch mit Anhängen. Der Begriff Artikel kann also weit gefasst sein und wird in wissenschaftlichen Journals nicht nur Forschungsberichte bezeichnen. Auch die meisten anderen wesentlichen Informationseinheiten in einem Journal haben Artikelcharakter, etwa Personalia, Forschungsankündigungen, Kongressankündigungen und -berichte, Editorials, Rezensionen etc.

Im Gegensatz zu klassischen Print-Journals und solchen E-Journals, die Print-Journals in digitaler Form eins zu eins widerspiegeln, gibt es bei Online-only-Journals Artikel, die fortlaufend erscheinen, ohne an eine übergeordnete Ordnungseinheit gebunden zu sein. Hier gibt es also keinen Jahrgang oder keine Nummer als geschlossene Einheit.

Auch die E-Journals, die Inhalte von Print-Journals online verfügbar machen, tun dies in der Regel artikelweise.

Aus allen diesen Gründen – weil sie die zentrale Informationseinheit darstellt und mitunter die einzige Einheit ist – liegt es nahe, die Einheit Artikel als die Einheit zu verwenden, aus der jeweils ein SIP erstellt wird.

Potenziell problematisch sind hier artikelbezogene Informationen, die nicht auf der Artikelebene, sondern auf der Journalebene angeordnet sind. Solche Fälle sind beispielsweise Informationen zu Autoren (Kurzvita, Adressen u.ä.), die als Anhang am Ende eines Journals stehen oder Errata, die meist ebenfalls separat angelegt sind. Eine weitere besondere Inhaltskategorie sind Register. Ob und in welcher Form derartige Inhalte auch bei E-Journals vorkommen, wie sie z.B. bei E-Journals gehandhabt werden, die die digitale Ausgabe eines Printjournals sind, müsste ermittelt werden. In jedem Fall sollte es eine Möglichkeit geben, auch solche Zusatzinformationen erhalten zu können; bei Errata ist dies unbedingt notwendig.

Eine Option, die bei der digitalen Publikation möglich wird, ist die Einbeziehung des Peerreview-Vorganges. Es wäre denkbar, beispielsweise bei einem als Preprint-Artikel erscheinenden Beitrag die eingehenden Peer-reviews zu publizieren oder das Peer-reviewing als Diskussionsthread online zu gestalten. Die dynamische oder statische Einbindung des Peerreviewing ist natürlich auch bei rein digitalen Publikationen möglich.

Journalbezogene Informationen

Bei vielen Journals gibt es eine Reihe von Informationen, die nicht der Einheit Artikel zugeordnet sind, sondern den übergeordneten Ordnungseinheiten wie Jahrgang oder Nummer, wie z.B. die Zusammensetzung des wissenschaftlichen Beirats oder die editorischen Regeln zur Zeit der Publikation, oder auch die Zugehörigkeit des Artikels zu einem Themenheft mit speziellem Editorial etc.

Diese Informationen sind zweifellos ebenfalls archivierungswürdig, da E-Journals nicht nur als Publikationsplattform für wissenschaftliche Forschungsergebnisse und übrige wissenschaftliche Kommunikation relevant sind, sondern da sie auch selbst Forschungsgegenstand sein können. Die wechselnde Zusammensetzung eines Editorial Board kann beispielsweise aufschlussreich für bestimmte Fragen der Wissenschaftsgeschichte oder -soziologie sein.

E-Journal-SIPs müssen also auch eine Rekonstruktion der temporären Entwicklung (man könnte auch sagen »Geschichte«) der Gesamtentität E-Journal ermöglichen, und zwar unabhängig davon, ob dieses E-Journal die E-Version eines Printjournals ist oder ein Online-only-Journal.

Besonders problematisch sind dabei die journalbezogenen Informationen, als die hier alle Informationen gelten sollen, die nicht spezifisch einem Artikel zugeordnet sind. Um diese digital zu archivieren, bieten sich zwei grundverschiedene Ansätze an.

Zum einen kann man diese Informationen im SIP jedem einzelnen Artikel beifügen. Schließlich sind die journalbezogenen Informationen nur deshalb nicht den einzelnen Artikeln zugewiesen, weil sie für alle gleichermaßen gelten und man sie so platzsparend für alle Artikel gemeinsam abdrucken kann. Sachlich spräche also nichts dagegen, diese Informationen für jeden Artikel zu wiederholen. Das würde jedoch einen hohen Grad an Redundanz mit sich bringen.

Ein anderer Ansatz ist die Speicherung dieser Informationen als separate Entitäten, gewissermaßen als Sonderfälle von Artikeln. Ein Editorial zu einer Ausgabe eines Journals würde also ebenso als ein Artikel angesehen werden wie die Liste der Herausgeber oder die der wissenschaftlichen Beiräte. Der Zusammenhang zwischen diesen journalbezogenen und »echten« Artikeln – und damit eventuellen übergeordnete Einheiten (z.B. Journalausgabe) – kann dann temporär hergestellt werden (etwa: »Welche Publikationsrichtlinien galten zu der Zeit der Veröffentlichung des Artikels?«) oder über die bibliographischen Metadaten (etwa: »Welche Artikel gehörten noch zu dem Sonderheft, das als Kontext des Artikels angegeben ist; gibt es einen Artikel vom Typ Editorial, der in diesem Sonderheft erschienen ist?« etc.).

Das Problem bei journalbezogenen Informationseinheiten – speziell solchen wie Herausgeber- und Beiratslisten – ist jedoch, dass sie nicht als Artikel wahrgenommen werden, weil sie oft über längere Zeit unverändert bleiben und bei Print-Journals dann auch in identischer Form wiederholt abgedruckt werden. Das kann bei E-Journals auch für allgemeine Statements gelten, die bei einem Print-Journal als Editorials einzelner Ausgaben angesiedelt wären. Will man also solche Informationseinheiten als separate SIPs behandeln, muss den Produzenten bewusst sein, dass jede Änderung an einer dieser Informationseinheiten in der Erstellung eines entsprechenden Änderungs-SIP resultiert.

Standardisierung und Offenheit

Für eine effektive – und das heißt automatisierte – Verarbeitung von SIPs im OAIS ist eine möglichst hohe Regelgebundenheit von Form, Inhalt und Übertragung der SIPs notwendig. Artikel in wissenschaftlichen E-Journals werden in sehr großer Zahl produziert, die Not-

wendigkeit der Standardisierung ist also besonders hoch. Damit stellt sich die zentrale Frage nach der Standardisierbarkeit von E-Journal-Artikeln.

Artikel in wissenschaftlichen Print-Zeitschriften unterliegen durch das Medium Papier ganz bestimmten Beschränkungen. Die Inhalte können sich ausschließlich aus Text (hierzu zählen auch Tabellen und ähnliche Strukturen) und aus Abbildungen zusammensetzen.

E-Journals, die Print-Journals in digitaler Form eins zu eins widerspiegeln, beschränken sich ebenfalls auf die Datentypen Text und Grafik. In der Regel wird daher das Druckvorstufen-Format für die Print-Ausgabe, PDF, auch für die Online-Verwertung verwendet (in stärker komprimierter Form, um die Ladezeiten zu vermindern). Für die Erstellung eines SIPs stellen sich hier keine besonderen Schwierigkeiten, da hier ebenfalls PDF verwendet werden kann, alternativ die Textbestandteile als XML und die Abbildungen in einem der langzeitstabilen Standardformate.

Komplexer wird es bei Online-Artikeln, die stärker von den technischen Möglichkeiten des Mediums Internet Gebrauch machen und auch nichtstatische grafische Komponenten, Audiodaten etc. verwenden. Auch die Möglichkeit, wissenschaftliche Primärdaten zusammen mit dem Artikel, der ihre Auswertung darlegt, zu veröffentlichen, besteht. Schließlich gibt es die Möglichkeit, an Artikel Diskussionsbeiträge anzuhängen. Zwar wird von diesen Möglichkeiten nur sehr begrenzt Gebrauch gemacht, besteht doch die Mehrzahl der E-Journals derzeit noch aus Pendants zu Print-Ausgaben. Aber die immer weiter zunehmende Akzeptanz des Mediums Internet und die wachsenden technischen Möglichkeiten wie Breitband-Anschlüsse dürften Entwicklungen jenseits des klassischen Print-Artikels befördern. Welche Formen die Publikationen der Open-Source-Bewegung annehmen werden, die Online-only-Publishing in weit stärkerem Maße verwendet als es Wissenschaftsverlage tun, lässt sich ebenfalls nicht absehen.

Ein SIP-Konzept für E-Journals erfordert daher einen Kompromiss von Standardisierung und Offenheit gegenüber neuen Artikelstrukturen und Datenformaten. Idealerweise ist die Aufnahme von neuen Strukturen und Formaten ebenfalls standardisiert, so dass der Abstimmungsbedarf beim Auftreten neuer Artikelbestandteile möglichst gering gehalten werden kann.

Fazit

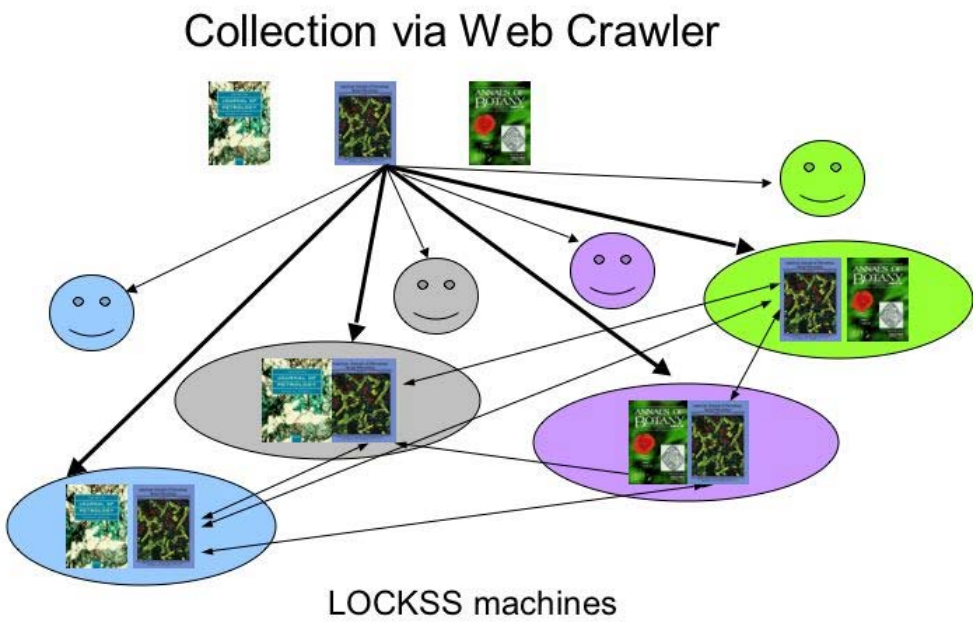
Für die weiteren Erörterungen wird mithin vorausgesetzt, dass E-Journal-SIPs artikelorientiert sind. Neben strukturierten Texten müssen beliebige andere Datentypen berücksichtigt werden. Informationen, durch die Artikel zu eventuellen übergeordneten Einheiten zugeordnet werden können, müssen als Metadaten der Artikel vorgehalten werden.
E-Journal-SIPs müssen auch eine Rekonstruktion der temporären Entwicklung der Gesamt-

entität E-Journal ermöglichen. Ob umfangreichere journalspezifische Informationen ebenfalls als Artikelmetadaten vorgehalten werden oder als separate Entitäten in separaten SIPs enthalten sein sollen, ist noch zu klären. Die Standardisierung hat hohe Priorität, auch die Aufnahme von neuen Strukturen und Formaten in ein SIP sollte standardisiert erfolgen.

LOCKSS – eine OAIS-Implementation für E-Journals

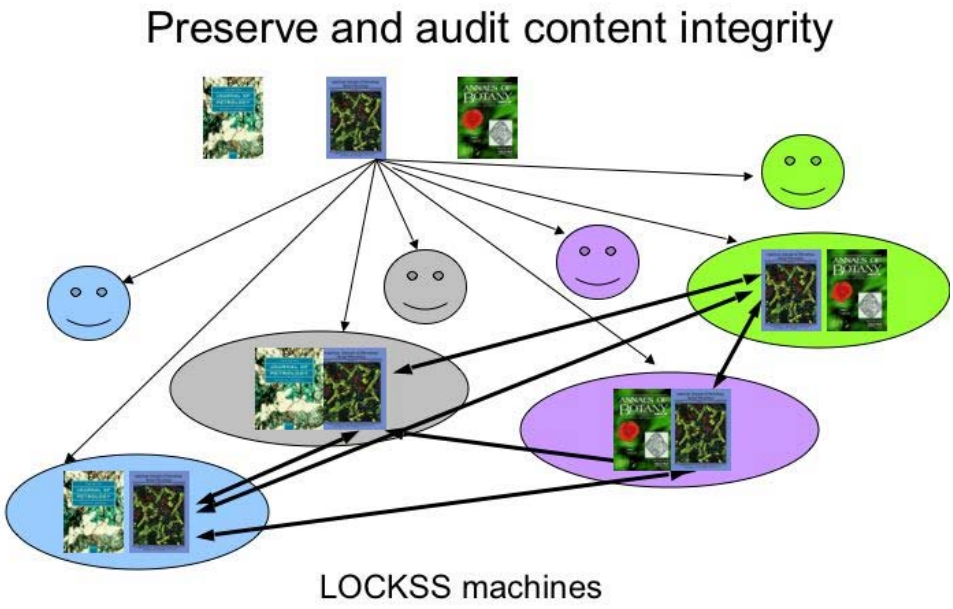
LOCKSS ist eine Initiative der US-amerikanischen Stanford University von 1999, der sich inzwischen über 50 wissenschaftliche Bibliotheken weltweit angeschlossen haben. Das System ist seit April 2004 aktiv. LOCKSS steht für »Lots of Copies Keep Stuff Save« und stellt eine verteilte Low-Cost-Lösung für die Archivierung webpublizierter Inhalte dar.

LOCKSS präsentiert sich als eine Software, die auf Standardrechnerkonfigurationen lauffähig ist und von Produzenten freigegebene Web-Inhalte automatisch auf Änderungen überprüft und immer die aktuellste Version herunterlädt (Web Crawler):



Auf diese Weise werden Web-Sites und mit ihnen verbundene Objekte (z.B. PDF-Dateien mit Artikelinhalten) auf den LOCKSS-Rechner geladen und können von dort den Nutzern eines Archivs zu Verfügung gestellt werden. Vor der Freigabe der Inhalte kann über ein Administratoren-Tool eine Qualitätssicherung und Einordnung in das Katalogsystem des Archivs erfolgen.

Zudem gleichen LOCKSS-Rechner an verschiedenen Standorten ihre Inhalte miteinander ab und aktualisieren ihre Inhalte gegenseitig oder reparieren Fehler:



Die Freigabe erfolgt dadurch, dass die Produzenten eine spezielle HTML-Seite auf ihrer Website einrichten, die eine Zugangsberechtigungserklärung für LOCKSS, beschreibende Metadaten (Dublin Core) und Links auf die freigegebenen Inhalte enthält. Dann teilen sie einem LOCKSS-Betreiber die Adresse dieser Seite mit, so dass LOCKSS die freigegebenen Inhalte samt Metadaten finden und abspeichern kann. Diese Website wird dann von allen LOCKSS-Systemen angesteuert, so dass viele Kopien entstehen, die sich zudem untereinander abgleichen.

Dieses System ist ein klassisches Pull-System mit den entsprechenden Vor- und Nachteilen (dazu siehe unten Abschnitt Transfer-Methoden).

LOCKSS enthält ein vorläufiges Formal statement of Conformance to ISO 14721:2003. Die Übereinstimmung von LOCKSS mit dem OAIS-Standard stellt sich wie folgt dar:

- Content Information. Die Content Information besteht aus Datenfolgen, die mit den Informationen angereichert sind, die zur Übertragung über das Internet und die Darstellung in Webbrowsern benötigt werden (HTTP-Header und MIME-Typangaben).
- Preservation Description Information (Provenance, Context, Reference and Fixity). Die Provenance und Reference wird bei LOCKSS durch die weltweit eindeutige und verlinkbare Webadresse (URL) angegeben, von der der Content geholt wurde. Der Context

wird durch die Links zwischen den Beiträgen hergestellt. Fixity wird durch den Abgleich zwischen verschiedenen LOCKSS-Servern gesichert.

- Packaging Information. Die Packaging Information wird durch Informationen im LOCKSS-PlugIn vorgehalten (Instanzen von Java-Klassen basierend auf XML-Dateien).
- Submission Information Package (SIP). Das SIP besteht bei LOCKSS aus den Web-Zugangsseiten des Produzenten, die dieser für LOCKSS bereitstellt und den Inhalten selbst.
- Archival Information Package (AIP). Das AIP besteht aus Instanzen von Java-Klassen basierend auf XML-Dateien.
- Dissemination Information Package (DIP). Die Daten werden in der gleichen Form für den Nutzer bereitgestellt, in der sie vom Produzenten zur Verfügung gestellt wurden.
- Die Descriptive Information besteht aus den URLs, unter denen die Inhalte ursprünglich von Produzenten bereitgestellt wurden und durchsuchbaren Informationen wie Metadaten und Volltext.
- LOCKSS ist ein einfaches und effektives System zur Archivierung von Web-Inhalten. Seine Grenze liegt allerdings darin, dass die eingesammelten SIPs für sich stehen und eine weitere Zuordnung von SIPs zueinander und damit eine übergreifende Systematisierung der Inhalte nur sehr begrenzt möglich ist. Zudem ist es auf die Möglichkeiten beschränkt, die sich durch die Präsentation der Inhalte mit den Techniken des Internet ergeben. Eine umfassende Lösung – die LOCKSS freilich gar nicht darstellen will – kann sich mit beiden Grenzen nicht zufrieden geben.

Quellen: *Offizielle LOCKSS-Website: http://lockss.stanford.edu/*

Packaging Standards

Für das Packen von Datenpaketen, bestehend aus aufeinander bezogenen Dateien unterschiedlichen Datentyps und dazugehörigen Metadaten, gibt es bereits eine Reihe von Standards. Daher soll im Folgenden analysiert werden, wie diese Standards aufgebaut sind und inwieweit sie den beschriebenen, spezifischen Anforderungen des Publikationstyps E-Journal-Artikel entgegenkommen.

Metadata Encoding and Transmission Standard (METS)

Der seit 2001 auf Initiative der amerikanischen Digital Library Federation entwickelte Standard METS hat das Ziel, ein Metadaten-Framework für Information Packages nach dem

OAIS-Standard anzubieten. Framework bedeutet, dass METS nicht für alle Metadaten-Arten (beschreibende, technische, administrative Metadaten etc.) eine eigene Lösung anbietet, sondern dass es einen Rahmen bildet, mit dem verschiedene etablierte Metadaten-Standards so verbunden werden, dass alle für OAIS-Information-Packages relevanten Informationen abgelegt und aufeinander bezogen werden können. Verantwortlich für METS zeichnet das METS Editorial Board, das vor allem aus Vertretern von wissenschaftlichen Großbibliotheken aus dem angelsächsischen Sprachraum zusammengesetzt ist. Deutschland ist hier durch Markus Enders vom Göttinger Digitalisierungs-Zentrum der Staatsbibliothek in Göttingen vertreten.

METS ist als speziell für den Archivgebrauch entwickeltes Packaging Format in der Archivwelt gut bekannt und wird lt. METS-Website bereits in einer Vielzahl archivalischer Projekte angewendet. Daher soll auf diesen Standard im Folgenden recht detailliert eingegangen werden, um eine Ausgangsbasis für Vergleiche mit anderen Packaging Formats zu schaffen.

METS bedient sich des XML-Standards. Der Dokumenttyp »METS-Datei« wird durch ein W3C XML-Schema festgelegt. Bei der Abfassung dieser Expertise lag METS in der Version 1.3 vor (seit Mai 2003 verfügbar), Version 1.4 befindet sich in Vorbereitung.

Mit METS werden beschreibende, technische, administrative und strukturelle Metadaten erfasst. Es kann aber auch die Content Information selbst in eine METS-Datei aufgenommen werden. Ein SIP, das auf METS-Basis erstellt wurde, besteht also aus einer Datei oder einem Set von Dateien: der (oder den) METS-Dateien und gegebenenfalls den Dateien, die die Content Information enthalten.

METS-Dateien können, wenn nötig, aufeinander verweisen und so Hierarchien bilden. Beispielsweise kann eine METS-Datei für einen Jahrgang einer Zeitschrift stehen und lediglich Verweise auf METS-Dateien zu den einzelnen Ausgaben beinhalten, die wiederum auf die METS-Dateien für die einzelnen Artikel verweisen.

Da METS lediglich ein Framework darstellt, muss seine konkrete Anwendung durch ein sogenanntes Profil spezifiziert werden.

Quellen: *Offizielle METS-Website: http://www.loc.gov/standards/mets/*

Aufbau des METS-XML-Frameworks

Das XML-Element für METS ist mets, es kann durch mehrere Attribute genauer bestimmt werden und sieben Unterelemente enthalten, die weitere Detailinformationen aufnehmen:

```
<mets ID="A12345" OBJID="B12345" LABEL="Example Document">
      [Unterelemente]
</mets>
```

Das Attribut ID enthält eine technische Dokumentnummer, anhand derer das METS-Dokument im Verarbeitungssystem erkannt werden kann, OBJID (Object ID) enthält die Inventarnummer des digital zu archivierenden Quell-Objektes, LABEL ist eine kurze Beschreibung, anhand derer auch ein menschlicher Bearbeiter direkt erkennen kann, welches METS-Dokument er vor sich hat.

Ein METS-Dokument kann des Weiteren bis zu sieben verschiedene Abschnitte enthalten. Diese Abschnitte sind der Kopfteil mit Metadaten über das METS-Dokument selbst (Header), es folgen deskriptive Metadaten, administrative Metadaten, Dateiinformationen, Strukturinformationen, Linkinformationen und Aktionsinformationen.

Das heißt: wenn ein METS-Dokument Informationen zu einem Information Package enthält, das aus mehreren Bestandteilen (Dateien oder Dateigruppen) besteht, werden zuerst alle deskriptiven Metadaten, dann alle administrativen Metadaten, dann alle Dateiinformationen abgebildet usw. Die verschiedenen Informationen zu ein und demselben Teil des Information Packages stehen also getrennt voneinander. Die Zuordnung von bestimmten Metadaten zueinander und zu bestimmten Dateien erfolgt daher über eindeutige Identifikationsnummern (IDs), die jeder Eintrag erhält, für den solche Zuordnungen vorkommen können.

METS Header

Der Header kann, muss aber nicht vorhanden sein. Der METS Header enthält Metadaten, die das METS-Dokument selbst beschreiben. Dazu zählen das Erstellungsdatum, das Datum der letzten Änderung, der Bearbeitungsstatus, Angaben über den Ersteller und die Bearbeiter der METS-Datei, auch weitere Identifikatoren, unter denen das Dokument gefunden werden können soll (z.B. Signaturen) usw.

Ein METS Header könnte beispielsweise so aussehen:

```
<metsHdr CREATEDATE="2003-07-04T15:00:00" RECORDSTATUS="Complete">
  <agent ROLE="CREATOR" TYPE="ORGANIZATION">
    <name>The British Library</name>
  </agent>
  <agent ROLE="ARCHIVIST" TYPE="INDIVIDUAL">
    <name>Ann Butler</name>
  </agent>
</metsHdr>
```

metsHdr ist das XML-Element, das den gesamten Header beinhaltet. Bestimmte Standardinformationen sind in Attributen dieses Elementes abgelegt. So zum Beispiel das Datum der Erstellung (CREATEDATE="2003-07-04T15:00:00") und der Bearbeitungsstatus des METS-Dokumentes (RECORDSTATUS="Complete").

Weiterhin sind innerhalb von metsHdr mittels der agent-Elemente zwei »Agenten« genannt, also Personen oder Organisationen, die das METS-Dokument erstellt oder modifiziert haben. Auch die agent-Elemente werden über Attribute näher bestimmt: ihr Typ wird angegeben (TYPE="INDIVIDUAL") und die Rolle, die sie spielen (ROLE="ARCHIVIST"). Das name-Element innerhalb von agent nennt schließlich den Namen des Agenten.

Für die Werte dieser Attribute gibt das XML Schema von METS eine Liste vor, aus der der Wert stammen muss. So wird hier Wildwuchs verhindert. Das Attribut ROLE kann die Werte ARCHIVIST, CREATOR, CUSTODIAN, DISSEMINATOR, EDITOR, IPOWNER und OTHER annehmen; das TYPE-Attribut die Werte INDIVIDUAL, ORGANIZATION oder OTHER.

Descriptive Metadata

Descriptive Metadata können, aber müssen nicht vorhanden sein. In den Descriptive Metadata finden sich Katalogisierungsinformationen im weitesten Sinn. Sie können aus einem Verweis auf externe Informationen bestehen, z.B. auf einen Eintrag im OPAC des Archivs, oder selbst einen Katalogisierungsdatensatz darstellen oder eine Kombination aus Verweis und Datensatz.

METS bietet kein eigenes Konzept für deskriptive Metadaten. Stattdessen können in einem METS-Dokument XML-codierte Metadaten international etablierter Standards wie z.B. MARC, TEI oder DC verwendet werden.

Das XML-Element für Descriptive Metadata ist dmdSec. dmdSec steht für Descriptive Metadata Section. dmdSec kann auch mehrfach hintereinander vorkommen, wenn mehrere Sets von Metadaten angegeben werden sollen, beispielsweise für ein Set von Daten, die ein Information Package bilden. Daher muss das Element dmdSec ein Attribut mit einer ID haben, um mehrere dmdSec voneinander unterscheiden zu können.

Da innerhalb einer Descriptive Metadata Section nur zwei prinzipielle Möglichkeiten der weiteren Füllung dieses Elementes vorgesehen sind, nämlich der Verweis auf externe Metadaten oder die Füllung mit XML-Metadaten eines anderen Standards, gibt es auch nur zwei Unterelemente von dmdSec, mdRef und mdWrap.

mdRef steht für Referenzen auf externe Metadaten zur Verfügung. Im Attribut MDTYPE wird der Typ der Metadaten angegeben, also der Standard, nach dem die externen deskriptiven Metadaten gehalten sind. Mit einer Reihe weiterer Attribute können technische Angaben zum Typ und zur Verortung der externen Metadaten gemacht werden. Dabei richtet sich METS nach vorhandenen Standards des W3C. So nutzt METS für die Verweise den XLink-Standard und für die Angabe des Datentyps MIME-Typen.

Eine Beispiel für eine METS Descriptive Metadata Section mit einem Verweis auf externe Metadaten:

```
<dmdSec ID="dmd01">
      <mdRef LOCTYPE="URL" MIMETYPE="text/xml" MDTYPE="DC"
      xlink:href="http://www.library.org/catalogue/dc/dc_set_12345
      6.xml" />
</dmdSec>
```

Das mdRef-Element im Beispiel gibt mittels Attributen an, dass die externen Metadaten mit Hilfe einer URL, also einer Internet-Adresse, auffindbar sind (LOCTYPE="URL"), dass es sich bei diesen Daten um eine XML-Datei handelt (MIMETYPE="text/xml"), dass diese XML-Daten sich nach dem Standard *Dublin Core* (DC) richten (MDTYPE="DC") und unter welcher Web-Adresse sich die XML-Datei mit den Metadaten befindet (xlink:href="http://www.library.org/catalogue/dc/dc_set_123456.xml").

Interne Metadaten werden in das Element mdWrap (für Metadata Wrapper) eingeschlossen. Dieses Element enthält wiederum entweder ein xmlData-Element, das im XML-Format vorliegende Metadaten umschließt oder ein binData-Element, das verwendet wird, um binäre Metadaten zu umschließen. Binäre Metadaten müssen dabei mittels des Base64-Standards zeichenkodiert sein.

Ein Beispiel für eine Descriptive Metadata Section mit internen Metadaten in XML-Form nach dem Standard *Dublin Core*:

```
<dmdSec ID="dmd02">
<mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core Metada-
ta">
         <xmlData>
          <dc:title>Alice's Adventures in Wonderland</dc:title>
          <dc:creator>Lewis Carroll</dc:creator>
          <dc:date>between 1872 and 1890</dc:date>
          <dc:publisher>McCloughlin Brothers</dc:publisher>
          <dc:type>text</dc:type>
         </xmlData>
        </mdWrap>
</dmdSec>
```

Ein Beispiel für eine Descriptive Metadata Section mit internen Metadaten in binärer Form als MARC-Datensatz:

```
<dmdSec ID="dmd003">
      <mdWrap MIMETYPE="application/marc" MDTYPE="MARC" LA-
      BEL="OPAC Record">
       <binData>MDI00Ddjam0gIDIyMDA10DkgYSA0NU0wMDAxMDA...(etc.)
       </binData>
      </mdWrap>
</dmdSec>
```

Administrative Metadata

Administrative Metadaten können, aber müssen nicht vorhanden sein. Das XML-Element für diesen Abschnitt eines METS-Dokumentes ist amdSec (für Administrative Metadata Section). Dieser Abschnitt umfasst technische, urheberrechtliche, quellenspezifische und versionsrelevante Verwaltungsinformationen zu den Dateien, die die Content Information enthalten.

Für jeden Typ Verwaltungsinformation gibt es einen eigenen Abschnittstyp.

Die Technical Metadata (techMD) beinhalten Informationen zum Datenformat, Versionsangaben und anderen Parametern, die benötigt werden, um die Datei lesen zu können.

Die Intellectual Property Rights Metadata (rightsMD) geben urheberrechtliche, Copyright- und Lizenzinformationen.

Die Source Metadata (sourceMD) betreffen deskriptive und administrative Metadaten über die analoge Quelle, von der ein digitales Archivobjekt stammt.

Die Digital Provenance Metadata (digiprovMD) umfassen Informationen über die Versionen, in denen eine Datei vorliegt, und über die Transformationen oder Migrationen, denen die Datei seit der Digitalisierung des Objektes unterworfen war.

METS bietet kein eigenes Konzept für administrative Metadaten. Stattdessen können in einem METS-Dokument XML-codierte Metadaten international etablierter Standards wie z.B. NISO verwendet werden. Ebenso kann auch auf extern vorliegende Verwaltungsinformationen verwiesen werden.

Demzufolge können die Verwaltungsabschnitte die gleichen Elemente enthalten, wie sie auch in der Descriptive Metadata Section vorkommen, nämlich mdWrap und mdRef.

Jeder Abschnittstyp kann mehrfach auftreten, um auch Information Packages gerecht zu werden, die mehrere Dateien oder Dateigruppen enthalten, für die jeweils unterschiedliche Verwaltungsinformationen vorliegen. Die Abschnitte müssen daher ein ID-Attribut aufweisen, damit sie einfach voneinander unterschieden werden können.

Ein Beispiel für technische Metadaten in einem Verwaltungsdatenabschnitt zu einer TIFF-Grafik, die als XML dem NISO-Standard entsprechend abgelegt sind:

```
<amdSec>
 <techMD ID="amd-tech-001">
  <mdWrap MIMETYPE="text/xml" MDTYPE="NISOIMG" LABEL="NISO Img.
Data">
   <xmlData>
    <niso:MIMEtype>image/tiff</niso:MIMEtype>
    <niso:Compression>LZW</niso:Compression>
    <ni-
so:PhotometricInterpretation>8</niso:PhotometricInterpretation>
    <niso:Orientation>1</niso:Orientation>
```

```
    <niso:ScanningAgency>NYU Press</niso:ScanningAgency>
 </xmlData>
  </mdWrap>
 </techMD>
<amdSec>
```

File Section

Die File Section (XML-Element fileSec) führt alle Dateien (XML-Element file) auf, die Teil des Information Package sind. Mehrere Dateien werden dabei in Gruppen (XML-Element fileGrp) zusammengefasst, wobei jede Gruppe für eine Version der Dateien steht. So würden beispielsweise Scans von Printseiten zu einer Gruppe zusammengefasst, die dazugehörigen Thumbnails zu einer weiteren Gruppe und die per OCR erzeugte XML-Version des Textes der Printseiten zu einer dritten Gruppe. Die Dateien können sich entweder außerhalb des METS-Dokumentes befinden, welches dann die Dateien nur auflistet, oder der Dateiinhalt kann in Form von XML oder Base64-Binär-Information in die METS-Datei aufgenommen werden.

Im folgenden Beispiel einer fileSec finden sich insgesamt fünf Dateien. Jeweils zwei TIFF- und GIF-Grafiken, die die Seiten 1 und 2 eines gescannten Printdokumentes enthalten, sowie eine XML-Datei, die den Text des Dokumentes enthält. Der Inhalt dieses digitalisierten Printdokumentes liegt also in drei verschiedenen Versionen vor, so dass drei fileGrp-Elemente benötigt werden:

```
<fileSec>

  <fileGrp ID="VERS1">
    <file ID="FILE001" MIMETYPE="image/tiff" CREATED="2001-06-10"
         DMDID="dmd001" AMDID="amd-tech-001 amd-digiprov-001"
         GROUPID="grp001" >
      <FLocat LOCTYPE="URL" xlink:href="http://.../page1.tiff" />
    </file>
    <file ID="FILE002" MIMETYPE="image/tiff" CREATED="2001-06-10"
         DMDID="dm001" AMDID="amd-tech-001 amd-digiprov-001"
         GROUPID="grp002" >
      <FLocat LOCTYPE="URL" xlink:href="http://.../page2.tiff" />
    </file>
  </fileGrp>

  <fileGrp ID="VERS2">
    <file ID="FILE003" MIMETYPE="image/gif" CREATED="2001-05-17"
         DMDID="dmd001" AMDID="amd-tech-002 amd-digiprov-002"
         GROUPID="grp001" >
    <FLocat LOCTYPE="URL" xlink:href="http://.../page1.gif" />
```

```
    </file>
    <file ID="FILE004" MIMETYPE="image/gif" CREATED="2001-05-17"
         DMDID="dm001" AMDID="amd-tech-002 amd-digiprov-002"
         GROUPID="grp002" >
    <FLocat LOCTYPE="URL" xlink:href="http://.../page2.gif" />
    </file>
  </fileGrp>

  <fileGrp ID="VERS3">
    <file ID="FILE005" MIMETYPE="text/xml" CREATED="2001-05-18"
         DMDID="dmd001" AMDID="amd-tech-003"
         GROUPID="grp001 grp002" >
      <FLocat LOCTYPE="URL" xlink:href="http://.../fulltext.xml"
      />
    </file>
  </fileGrp>

</fileSec>
```

Das XML-Element file nimmt Basisinformationen zu der Datei in Form von Attributen auf. Dazu zählen die Datei-ID (die nichts mit dem Dateinamen zu tun haben muss, sondern innerhalb des METS-Dokumentes frei wählbar ist), der Datentyp (MIMETYP) und das Erstellungsdatum (CREATED) sowie ggf. weitere Informationen zu Dateigröße, Checksummen u.ä.

Die Attribute DMDID (für *Descriptive Metadata ID*) und AMDID (für *Administrative Metadata ID*) stellen die Verbindung von der Datei zur Descriptive Metadata Section und zur Administrative Metadata Section her. Das heißt, dass die jeweiligen Metadaten zu einer Datei und zu ihrem Inhalt in derjenigen Metadata Section zu finden sind, die die ID hat, die diese Attribute angeben.

Im Beispiel haben alle file-Elemente dieselbe DMDID. Das ergibt durchaus Sinn, denn da alle Dateien sich auf dasselbe Archivierungsobjekt beziehen, können die deskriptiven Metadaten für alle identisch sein und müssen nur einmal vorgehalten werden. Es gibt also einen Abschnitt in der Administrative Metadata Section mit der ID »dm001«, der für alle Dateien gleichermaßen zutrifft.

Bei den Angaben zu den Administrative Metadata (Attribut AMDID) ist es im Beispiel so, dass Dateien eines Formats (TIFF, GIF und XML) jeweils identische Angaben haben. Der Grund dafür liegt darin, dass gleiche Formate gleiche technische Metadaten haben können. Zudem muss durch die Digital Provenance Metadata angegeben werden, welche Dateien aus welchen konvertiert wurden. Daraus ergibt sich, dass das Attribut AMDID mehrere Werte haben kann, denn für eine Datei können ja in verschiedenen Abschnitten der Administrative Metadata Angaben vorliegen, wobei jeder Abschnitt eine eigene ID hat. Im Beispiel sind das die Verweise auf je zwei Abschnitte, nämlich technische und Digital Provenance Metadata (AMDID="amd-tech-002 amd-digiprov-002").

Schließlich können durch das Attribut GROUPID Gruppen von Dateien gebildet werden, die bis auf das Format identisch sind. Im Beispiel haben jeweils die Grafikdateien, die dieselbe Seite des Printwerkes abbilden, die gleiche GROUPID, während die XML-Datei zu beiden Gruppen gehört, da sie den Inhalt beider Seiten vereint. So können inhaltlich zusammengehörige Dateien schnell identifiziert und zusammen ausgewertet werden.

Das XML-Element file kann zwei Unterelemente enthalten, FLocat und FContent.

FLocat gibt den Ort an, unter dem die Datei zu finden ist, im einfachsten Fall den Dateinamen oder auch komplexere Angaben wie die Internet-Adresse der Datei. FLocat ist genauso aufgebaut wie das Element mdRef aus der Metadaten-Sektion, es benutzt die gleichen W3C-Standards zur Angabe von Typ und Ort einer externen Ressource. Ein file-Element kann beliebig viele FLocat-Elemente enthalten. So kann zur Sicherheit auf mehrere identische Kopien der betreffenden Datei an verschiedenen Orten verwiesen werden.

FContent enthält den Inhalt einer Datei in XML-Form oder als Base64-Binär-Information. FContent ist genauso aufgebaut wie das Element mdWrap aus der Metadaten-Sektion, es kann also ein Unterelement xmlData oder binData enthalten, in dem dann der XML- oder Binärinhalt der Datei zu finden ist. Jedes file-Element kann maximal ein FContent-Element aufweisen, das auch mit einem oder mehreren FLocat-Elementen kombiniert werden kann.

Structural Map

Die Structural Map (XML-Element structMap) stellt das Zentrum eines METS-Dokumentes dar. Durch die Structural Map werden die hierarchischen Zusammenhänge der Dateien verdeutlicht. Beispielsweise können Grafikdateien, die jeweils eine gescannte Seite eines Buches beinhalten, so in Kapitel u.ä. zusammengefasst werden.

Mit Hilfe der Structural Map kann beispielsweise eine Navigationsstruktur aufgebaut werden, mittels derer ein Nutzer eines Information Package auf die enthaltenen Daten gezielt zugreifen kann.

Für ein Information Package können bei Bedarf mehrere Structural Maps verwendet werden, die jeweils unterschiedliche Strukturgesichtspunkte berücksichtigen (etwa die logische Struktur und die physikalische Struktur eines Objektes).

Der Aufbau einer Structural Map ist vergleichsweise einfach. Jeder Abschnitt wird durch ein div-Element (für division) repräsentiert, wobei div-Elemente wiederum weitere div-Elemente enthalten können, die entsprechend für Unterabschnitte stehen. Die div-Elemente tragen Attribute, die sie näher bestimmen: ID, ein LABEL mit einer Bezeichnung, die Angabe des TYPE. Weitere Attribute sind ORDER und ORDERLABEL, wobei ORDER für die absolute Position des Abschnittes steht und ORDERLABEL für eine eventuelle abweichende Zählung (so kann ORDER die tatsächliche Seitenzahl einer Seite in einem Buch an-

geben, ORDERLABEL dagegen die Paginierung der Seite). Auch die schon von file her bekannten Attribute AMDID und DMDID stehen für div zur Verfügung, so dass Metadaten für ganze Abschnitte angegeben werden können (wodurch Angaben für jede Datei überflüssig werden können).

Innerhalb der div-Elemente können Verweise auf Dateien stehen, in denen weitere Informationen zum Inhalt des betreffenden Abschnittes zu finden sind. Dabei kann es sich um Verweise auf file-Elemente in der File Section derselben METS-Datei handeln (XML-Element fptr; für File Pointer) oder um Verweise auf andere METS-Dateien, in denen weitere Informationen über den Inhalt des betreffenden Abschnittes stehen (XML-Element mptr, für METS Pointer). Im letzteren Fall wird die Information der Structural Map der METS-Datei, auf die verwiesen wurde, logisch an der Stelle eingefügt, an der sich der Verweis befindet.

Beispielsweise kann die Structural Map einer METS-Datei, die ein digitalisiertes Buch beschreibt, div-Elemente für jedes Kapitel und Unterkapitel enthalten, in denen schließlich fptr-Elemente auf die files mit den eingescannten Seiten des Buches verweisen. Ein anderer Ansatz wäre eine METS-Datei für die Ausgabe einer Zeitschrift, bei der die Structural Map div-Elemente enthält, in denen mptr-Elemente auf andere METS-Dateien verweisen, die sich auf die eigentlichen Artikel beziehen.

Nachstehend ein Beispiel für die Structural Map einer METS-Datei, die ein Buch strukturiert, das aus Kapiteln besteht und von dem es für jede Seite eine TIFF-Datei und eine XML-Datei gibt. Das erste div-Element steht für das Kapitel, das zweite, das sich innerhalb des ersten befindet, steht für die Seite, das fptr-Element verweist auf die Dateien, indem es ihre ID angibt. Es wird dabei vorausgesetzt, dass es in der File Section des METS-Dokumentes file-Elemente mit diesen IDs gibt, aus denen sich weitere Details zu den betreffenden Dateien entnehmen lassen.

```
<structMap>
  <div ID="kap1" TYPE="chapter" LABEL="Einleitung"
     ORDER="1" ORDERLABEL="I." >
    <div ID="p1" TYPE="page" ORDER="5" ORDERLABEL="I" >
      <fptr FILEID="file0001" />
      <fptr FILEID="file0022" />
    </div>

    ... weitere div-Elemente für die übrigen Seiten ...

  </div>

  ... weitere div-Elemente für die übrigen Kapitel ...

<structMap>
```

Das erste div-Element ist ein Kapitel (TYPE="chapter"), und zwar das erste (ORDER="1"), die Kapitelnummerierung im Buch ist »I.« (ORDERLABEL="I."). Es handelt sich dabei um die Einleitung (LABEL="Einleitung"). Das zweite div-Element steht innerhalb des ersten, wie die Seiten innerhalb des Kapitels stehen, sein Typ ist »page«, und zwar ist es absolut die 5. Seite des Buches (Titelei mitgezählt) und trägt die Paginierung »I«. Die beiden durch fptr-Elemente angegebenen Dateien, die den Inhalt der Seite tragen, haben die IDs »file0001« und »file0002«. An dieser Stelle könnte auch ein Verweis auf die ID einer Dateigruppe stehen (fileGrp), wenn mehrere Dateien zusammen den Inhalt des div-Elementes ergeben.

Innerhalb von fptr können auch noch weitere XML-Elemente stehen, die angeben, wo genau in der Datei sich die gesuchte Information befindet. Das ist beispielsweise notwendig, wenn für ein Buch pro Seite eine TIFF-Grafik vorhanden ist, aber nur eine einzige XML-Datei, die den Text des gesamten Buches enthält. Dann kann mittels fptr-Elementen für jede Seite angegeben werden, welche TIFF-Datei ihr entspricht und welcher Abschnitt innerhalb der Gesamt-XML-Datei. Die genaueren Angaben über Informationsausschnitte innerhalb einer Datei sind natürlich davon anhängig, um welche Dateiart es sich handelt. Bei Grafiken können beispielsweise die Koordinaten des Ausschnittes angegeben werden, in dem die betreffende Information zu finden ist, bei einer XML-Datei dagegen kann man die ID des betreffenden Abschnittes angeben. (Es sei an dieser Stelle darauf verzichtet, hierzu weitere Details anzuführen. Sie können der METS-Dokumentation entnommen werden).

Bei dem obigen Beispiel wird in METS davon ausgegangen, dass die beiden angegebenen Dateien den gleichen Inhalt in alternativer Form enthalten. Allerdings gibt es auch die Möglichkeit anzugeben, dass mehrere Dateien erst zusammen den Inhalt des div-Elementes ergeben, und zwar entweder hintereinander (z.B. mehrere Dateien für ein langes Musikstück) oder gleichzeitig (z.B. Bild- und Tondaten einer Videoaufnahme). (Auch hier sei für die Details auf die Dokumentation verwiesen.)

Nachstehend ein Beispiel für die Structural Map einer METS-Datei, die eine Ausgabe einer Zeitschrift strukturiert, wobei hier nicht Dateiverweise erfolgen, sondern Verweise auf andere METS-Dateien für die einzelnen Artikel. Diese Verweise erfolgen mittels des mptr-Elementes, das als Verweis auf eine externe Ressource genauso aufgebaut ist wie die schon vorgestellten mdRef- und FLocat-Elemente. Allerdings ist in diesem Beispiel keiner der Standardtypen für die Ortsangabe verwendet, sondern ein anderer (LOCTYPE="OTHER") und zwar die Angabe eines Datenbank-Records (OTHERLOCTYPE="db-record"), dessen Nummer dann mit dem Attribut xlink:href angegeben wird.

```
<structMap>
  <div TYPE="journal-issue" LABEL="Wonders of Nature; Vol.
     2003/1" >

    <div TYPE="journal-article" LABEL="WoN 2003/1: Article: A New
       Species, by P. Toole" >
      <mptr LOCTYPE="OTHER" OTHERLOCTYPE="db-record"
      xlink:href="mets125" />
    </div>

    <div TYPE="journal-article" LABEL="WoN 2003/1: Article: Flying
    Dogs
       in Namibia, by K. C. Smith" >
      <mptr LOCTYPE="OTHER" OTHERLOCTYPE="db-record"
      xlink:href="mets123"/>
    </div>

    ... weitere mptr-Elemente für die übrigen Artikel ...

  </div>
<structMap>
```

Structural Links

Durch Structural Links können in METS Hyperlinks zwischen div-Elementen der Structural Map hergestellt werden. Damit ist es beispielsweise möglich, inhaltliche Beziehungen zwischen Abschnitten sichtbar zu machen, ohne dass die Links in der Content Information ausgewertet werden müssten. So können etwa »Siehe-auch«-Bezüge schon auf der METS-Ebene hergestellt werden.

Structural Links sind sehr einfach aufgebaut. Innerhalb des XML-Elementes structLink befindet sich lediglich eines oder eine Reihe von smLink-Elementen (Structural Map Linking). Ein solches Element gibt durch die Attribute from und to die IDs der div-Elemente an, zwischen denen es Hyperlinks gibt. Wie genau dabei die Verlinkung ist, hängt von der Granularität der div-Elemente ab. Ist beispielsweise in einem Buch das Kapitel die kleinste div-Einheit, können nur Link-Beziehungen zwischen ganzen Kapiteln sichtbar gemacht werden; entspricht dagegen die kleinste div-Einheit einem Absatz, einer Zeile oder gar einem Wort, sind die METS-Links entsprechend exakt.

```
<structLink>
  <smLink from="div1" to="div234" />
  <smLink from="div456" to="div164" />
</structLink>
```

Das genaue Verhalten der Structural Map Links kann entsprechend dem Xlink-Standard über weitere Attribute genauer definiert werden.

Behaviour Section

Die Behaviour Section definiert sogenannte Behaviours für Objekte oder Objektgruppen. Behaviours sind Verarbeitungs-, Darstellungs- und Übertragungsmethoden, die für ein Objekt zur Verfügung stehen. Dieser Abschnitt des METS-Modells soll dazu dienen, verarbeitungsorientierten Systemen einen Zugriff auf Datenobjekte zu gewähren, die in METS-codierten Information Packages vorliegen. Für diesen Zweck kann dem METS-Dokument eine verarbeitungsorientierte Structural Map hinzugefügt werden, der dann die in der Behaviour Section definierten Methoden zugeordnet werden.

Der verarbeitungsorientierte Ansatz ermöglicht es, für ein in einem Archiv gefundenes Objekt anzugeben, was der Nutzer damit tun kann (Bei Grafiken z.B. »Vergrößern«, bei Texten, die als Scan und Transkription vorliegen, »Faksimile anzeigen«, »Blättern«, »Durchsuchen«; etc.).

Dieser Abschnitt wurde vor allem in METS eingefügt, um diesen Standard mit dem FEDORA-System kompatibel zu machen, einem im angelsächsischen Archivbereich weit verbreiteten Digital Repository-System, das verarbeitungsorientiert arbeitet (www.fedora.info).

METS Profile

Wie die vorstehenden Abschnitte verdeutlicht haben, ist METS ein sehr flexibles Framework. Bis auf die Structural Map sind alle METS-Elemente fakultativ und können teilweise auch mehrfach vorkommen. Für Metadaten können beliebige XML- oder binäre Standards verwendet werden, die Content Information kann innerhalb oder außerhalb der METS-Datei aufbewahrt werden, METS-Dateien können für sich stehen oder auch Hierarchien bilden, die Anwendung der Structural Map ermöglicht eine Vielzahl an konkreten Umsetzungen etc.

Alle diese Möglichkeiten in der konkreten Anwendung auf digital zu archivierende Objekte beliebig zu nutzen, hieße, eine organisatorisch und technisch kaum zu bewältigende Vielfalt zuzulassen, die die Effektivität, die das Arbeiten mit digitalen Archivobjekten ja letztlich ermöglichen soll, konterkarieren würde. Mithin ist in der Praxis eine Einschränkung der Möglichkeiten von METS wünschenswert, die es erleichtert, METS-Dokumente zu erstellen und weiterzuverarbeiten. Solche genauer definierten METS-Umsetzungen werden als Profile bezeichnet.

Für unterschiedliche Arten von Archivobjekten können unterschiedliche Umsetzungen von METS sinnvoll sein. Daher bietet es sich an, genauer definierte METS-Profile nach Objekttypen einzurichten, wie zum Beispiel Monographien, Zeitschriften, Sammelbände, Karten, Tondokumente etc. Da diese Objekte von unterschiedlichen Archiven zum Teil unterschiedlich behandelt werden, ist auch hinsichtlich archivspezifischer Eigenheiten eine Anpassung für die praktische Nutzung notwendig. Es wäre zwar für einen effektiven Datenaustausch zwischen verschiedenen Institutionen wünschenswert, wenn alle METS-Nutzer die gleichen Metadaten-Standards und die gleichen Regeln für bestimmte Archivobjekte anwenden würden, aber das ginge an der archivalischen Realität vorbei. Allerdings sollten alle möglichen Anstrengungen unternommen werden, um Divergenzen in der Anwendung des Standards so gering wie möglich zu halten.

Die Anpassung von METS erfolgt dadurch, dass für alle die Punkte, bei denen METS verschiedene Optionen anbietet, angegeben wird, welche der verfügbaren Optionen genutzt werden sollen. Das beginnt damit festzulegen, welche der METS-Elemente überhaupt genutzt werden sollen und wie ihre Nutzung im Detail geschehen soll. Für die Metadaten-Elemente, die lediglich den Rahmen für die Anwendung anderer Standards bilden, muss angegeben werden, welcher Standard an welcher Stelle zum Einsatz kommen soll. Das Resultat ist ein eingeschränktes XML-Schema, das nur noch die Elemente und Attribute (sowie ggf. bestimmte Werte der Attribute) zulässt, die für den Objekttyp benötigt werden, sowie eine Anwendungsdokumentation, die die gewünschte Nutzung der erlaubten XML-Elemente beschreibt.

Da bei einem Framework die Einrichtung von Profilen ein naheliegender Schritt ist, sieht METS hierfür ein standardisiertes Verfahren vor. Die genauere Einrichtung von METS auf konkrete Anwenderbedürfnisse erfolgt über eine XML-Konfigurationsdatei, für die ein eigenes XML-Schema existiert. Diese Datei enthält nicht nur technische Angaben, sondern auch organisatorische, wie z.B. eine Kurzbeschreibung, die Kontaktadresse der für das Profil zuständigen Person oder Institution, eine Beschreibung der Beziehungen zwischen ähnlichen oder aufeinander bezogenen Profilen etc. METS Profile können über die Library of Congress beim METS Editorial Board registriert werden. Bevor für einen Objekttyp ein eigenes Profil entworfen wird, empfiehlt es sich, bereits registrierte Profile zu konsultieren. Es ist vorgesehen, registrierte Profile online verfügbar zu machen, derzeit erhält man sie auf Anfrage beim METS Editorial Board.

METS in der Anwendung auf E-Journals

Wie könnte nun ein METS Profil in der Anwendung auf E-Journals aussehen? An dieser Stelle empfiehlt es sich, auf geleistete Vorarbeiten zurückzugreifen, und das Beispiel des E-Journal-SIP des E-Journal-Archives (EJAR) der Harvard University in Cambridge, M.A., zu

analysieren. Dieser SIP-Entwurf wurde im Rahmen des Projektes »Design of an E-Journal Archive« der Andrew W. Mellon Foundation im Jahr 2001 von Mitarbeitern der Harvard University Library erstellt. Dabei wurden größere wissenschaftliche E-Journal-Produzenten zur Mitarbeit eingeladen, so dass das Projekt von der Harvard University Library und den drei Publishing-Partnern Blackwell Publishing, John Wiley und University of Chicago Press, die zusammen weit über 1000 E-Journals publizieren, umgesetzt wurde.

Quellen: *EJAR-SIP Specification: http://hul.harvard.edu/~stephen/SIP_Spec.doc und http://www.diglib.org/preserve/harvardsip10.pdf. – allgemeine Informationen zum Projekt: http://xml.coverpages.org/harvardEJournalArchive.html. – Beispieldaten: http://www.oasis-open.org/cover/Harvard-SIP-Examples20011219.txt.*

Das E-Journal-Konzept

Das Konzept des Harvard-EJAR-SIPs basiert auf einer Strukturdefinition von E-Journals, die die Existenz von zwei Ebenen voraussetzt: die Ebene der Ausgabe (Issue) und des Beitrags (Item). Die Issue wird als eine vom Herausgeber definierte Sammlung von Items verstanden. Item umfasst alle Bestandteile mit Artikelcharakter. Zur Issue gehören journalbezogene Bestandteile wie Titelseite, Inhaltsverzeichnis, Editorial Board etc. Die Einheit, die als SIP eingereicht wird, ist die Issue.

Aufbau des SIP

Die Einlieferung der Daten erfolgt durch Einsendung des SIP durch den Produzenten per FTP-Upload auf einen passwortgesicherten Server des Archivs, bei größeren Datenmengen besteht die Alternative, Datenträger einzusenden.

Das SIP wird in Form einer ZIP-Datei eingereicht, deren Namen die ISSN des Journals ist, für das das SIP relevant ist. Die ZIP-Datei enthält Verzeichnisstrukturen, die für die Verarbeitung der gelieferten Daten erforderlich sind.

Für jede Issue und für jedes Item ist je eine Metadatendatei in METS-Format (issue-md.xml und item-md.xml) beigefügt. Für die Datenlieferung ist eine feste Ordnerstruktur vorgeschrieben. Der oberste Ordner ist mit der ISSN benannt, die vorhanden sein muss, damit ein E-Journal akzeptiert wird. Der darunterliegende Issue-Ordner enthält die issue-md.xml-Datei, Dateien mit dem Issue-bezogenen Content sowie Unterordner, die die Item-Daten enthalten. Der Issue-Ordner wird nach den Konventionen für ein SICI Item Segment benannt (Serial Item and Contribution Identifier Standard). Item-Ordner sollten mit dem digitalen Identifier benannt sein, unter dem ihr Inhalt bei einer Registrierungsagentur registriert ist (DOI, PURL o.ä.).

Jeder Item-Ordner enthält die item-md.xml-Datei und die zum Item gehörenden Content-Dateien.

Alle Dateien folgen bestimmten Namens- und Formatkonventionen. EJAR akzeptiert als normative Formate XML für Textinhalte, PDF als Kompositformat sowie wenige Standardformate für Grafik-, Audio- und Videodaten. Andere Formate werden ebenfalls aufgenommen, aber ohne Garantie für ihre zukünftige Nutzbarkeit. Soweit sich nichtnormative Datenformate automatisch in ein normatives Format konvertieren lassen, wird diese Konvertierung beim Ingest vorgenommen. Für die XML-Daten gibt es weitere Vorgaben, es existiert eine spezifische DTD.

Falls eine XML-Datei mit dem Item-Inhalt geliefert wird, ist eine zusätzliche Link-Datei (item-links.xml) erforderlich, in der die in der Content-XML-Datei enthaltenen Links separat aufgeführt werden. Dies ist eine Besonderheit des EJAR-XML-Konzeptes, die es ermöglichen soll, Links separat zu bearbeiten, beispielweise um Wissensnetze (Knowledge Bases) zu erzeugen.

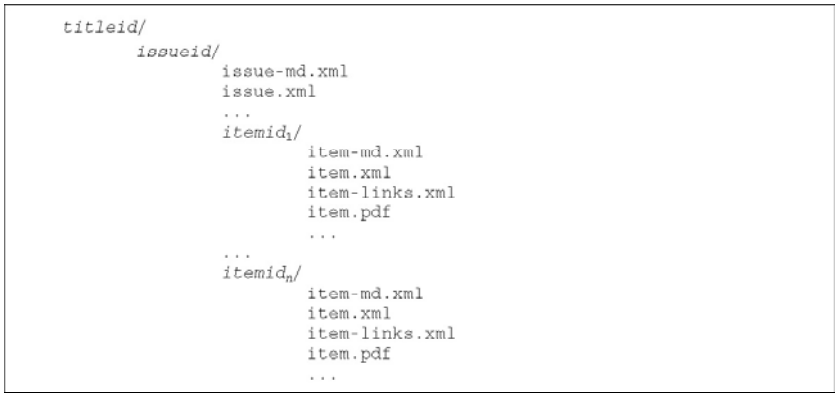


Abb. Harvard-EJAR-SIP – Konventionen zur Ordnerstruktur und Dateibenennung

Für die einzelnen Metadatenabschnitte ist zunächst kein spezielles XML-Metadatenformat vorgesehen, es wurde beim Design des SIP davon ausgegangen, dass die anzuwendenden Metadaten-Standards in einem separaten Auswahlverfahren bestimmt werden.

Die METS-Datei auf Issue-Ebene (issue-md.xml) enthält die Metadaten, die File Section und die Structural Map für die Content-Dateien der Issue-Ebene. Die Issue-Ebene kann in der Structural Map in einzelne Abschnitte (div-Elemente) unterteilt werden, soweit dies inhaltlich erforderlich ist (z.B. Artikel, Rezensionen, Ankündigungen, Personalia etc.). Soweit diese Ebenen Issue-spezifische Inhalte haben, enthält das betreffende div-Element File Pointer (fptr) auf Dateien, die in der File Section der Issue-METS-Datei aufgelistet sind. Das sind alle die Dateien, die sich im gleichen Ordner befinden wie issue-md.xml. Die Ebenen, die Item-spezifische Inhalte haben, sind dagegen mit METS-Pointern (mptr) gefüllt, die auf die Item-Level-METS-Dateien in den Unterordnern verweisen. Da alle diese METS-Dateien den gleichen Namen haben (issue-md.xml), erfolgt die Differenzierung durch relative Pfadangaben im xlink:href-Attribut (Beispiel: xlink:href="../itemid1/item-md.xml").

Item-Level-METS-Dateien enthalten die Metadaten, die File Section und die Structural Map für die Content-Dateien der Item-Ebene. Eine Differenzierung des Artikelinhalts auf METS-Ebene und damit weitere div-Unterteilungen sind nicht vorgesehen.

Fazit

Das Harvard-EJAR-SIP stellt ein einfaches, METS-basiertes SIP für Journalinhalte dar. Es basiert im Wesentlichen auf dem Grundsatz, dass E-Journals digitale Entsprechungen von Print-Journalen sind.

Besonderheiten der digitalen Form werden insbesondere auf der Ebene der audiovisuellen Daten berücksichtigt, für die jeweils ein normatives Format angeboten wird. Ebenso wird die Möglichkeit der Verwendung von Hyperlinks in digitalen Daten und deren potenzielle Weiterverwertung für Wissensnetze dadurch berücksichtigt, dass Links in einer separaten XML-Datei vorgehalten werden.

Eine Einschränkung besteht darin, dass das SIP issue-basiert ist. Der Fall, dass es bei Online-only-Journals diese Einheit nicht gibt, ist nicht explizit vorgesehen, so dass hier ein Workaround notwendig wäre. Das Konzept sieht ebenfalls kein Journal-Level vor, so dass offen bleibt, wo journalspezifische Informationen von Online-only-E-Journals, die kein Issue-Level aufweisen, abgelegt werden. Der Hintergrund für diese Selbstbeschränkung ist, dass das Harvard-EJAR-SIP in Kooperation mit drei ausgewählten größeren Publishern entstanden ist, die keine Issue-losen Publikationen veröffentlichen und auch keine konkreten Vorhaben in dieser Richtung unterhielten.

Weiterhin gibt es keine Vorkehrungen für Objekte auf Journal- oder Issue-Ebene, die für die Artikel relevant wären (Errata u.ä.).

Auf der Ebene der Formate für die Content Information wählt das Harvard-EJAR-Projekt mit der Definition eines eigenen XML-Formats und der Ergänzung bzw. Alternative PDF einen Weg, der sicher effektiv und zuverlässig ist, der aber bei der Öffnung für weitere – vor allem auch kleinere – Produzenten problematisch sein dürfte. Bei weitem nicht alle E-Journals liegen auch als PDF vor – und die Umstellung der Produzenten auf ein vom Archiv vorgegebenes XML dürfte in vielen Fällen nicht möglich sein.

Allerdings handelt sich dabei nicht um grundsätzliche Einschränkungen von METS. Das Framework METS ist offen und flexibel genug, um hier Lösungen zu ermöglichen.

Digital Item Declaration Language (DIDL, MPEG-21)

DIDL, die Digital Item Declaration Language, ist eine XML-basierte Beschreibungssprache für digitale Datenpakete, die Bestandteil des MPEG-21-Standards ist. MPEG-21 ist der fünfte Standard der Moving Pictures Experts Group (MPEG). Während es sich bei MPEG-1, MPEG-2 und MPEG-4 um Standards zur Datenkodierung handelt, ging man bereits mit MPEG-7 in eine andere Richtung. MPEG-7 spezifiziert keinen Code, sondern ist ein XML-basierter Beschreibungsstandard für Multimedia-Daten. Mit MPEG-21 will die MPEG nun ein Multimedia-Framework schaffen, welches den Austausch von digitalen Daten unter Praxisbedingungen ermöglicht. Das heißt, es sollen alle relevanten technischen, organisatorischen und rechtlichen Aspekte Berücksichtigung finden, die bei einem solchen Austausch relevant sein können. Dementsprechend umfangreich fällt der Standard aus.

MPEG-21 ist wie alle anderen MPEG-Standards zugleich auch eine ISO/IEC-Norm (ISO/IEC 21000). Er ist relativ neu, den Status einer veröffentlichten ISO-Norm haben noch nicht alle Teile von MPEG-21 erreicht, die ersten Teile erhielten 2003 diesen Status, darunter DIDL. Weitere Teile wurden 2004 veröffentlicht oder befinden sich noch im letzten Entwurfsstadium.

Der Standard ist bisher in der Digital Library-Gemeinschaft noch nicht breit wahrgenommen worden, was sicher einerseits daran liegt, dass es ihn noch nicht lange gibt und andererseits, dass eine genaue Dokumentation nicht frei im Internet verfügbar ist, da ISO-Normen erst nach offizieller Veröffentlichung erhältlich sind. Schließlich hat sich MPEG-21 nicht primär die Datenarchivierung auf die Fahnen geschrieben und ist als Teil des komplexen MPEG-Standard-Geflechts auch nicht so leicht fassbar wie der derzeitige Favorit unter den Packaging Standards, METS.

Dennoch dürfte DIDL von Interesse auch für Archivierungszwecke sein. Dies legt auch der positive Ausgang einer entsprechenden Evaluierung der Los Alamos National Laboratory Digital Library nahe, die zur der Entscheidung führte, DIDL-basierte Packages als Standard in dieser Einrichtung zu verwenden.

Quellen: *MPEG ISO/IEC-Norm: www.iso.org. – Website der Moving Picture Experts Group (MPEG): http://www.chiariglione.org/mpeg/. – Jeroen Bekaert, Patrick Hochstenbach, Herbert Van de Sompel, Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library, In: D-Lib Magazine, Volume 9 Number 11, November 2003 (http://www.dlib.org/dlib/november03/bekaert/11bekaert.html).*

Das Konzept von MPEG-21

Kernkonzept von MPEG-21 ist das Digital Item. Ein Digital Item ist eine Entität zur Verteilung von Informationen im Multimedia-Framework in Form einer Zusammenfassung von Ressourcen und Metadaten. In der OAIS-Terminologie entspricht einem Digital Item ein Information Package.

Digital Items werden von sogenannten Usern verwendet, wobei für die MPEG als User sowohl Distributoren wie Konsumenten zählen. Das MPEG-21 Multimedia-Framework bietet den Usern sechs Grundfunktionalitäten an. Sie können Digital Items erstellen, verändern, anbieten, konsumieren, suchen und anpassen. Mit dem letzten Punkt ist gemeint, dass das MPEG-21 Multimedia-Framework die automatische Anpassung der Digital Items an Besonderheiten des Users, des Endgerätes und des Netzwerks unterstützt. Aspekte wie beschreibende und technische Metadaten spielen dabei ebenso eine Rolle wie die Einbeziehung von Digital Rights Management.

MPEG-21 bietet aber nicht nur ein Regelwerk für die Beschreibung von Digital Items, sondern auch für die Interaktion von Usern und der technischen Infrastruktur. Daher beinhaltet es auch XML-basierte Standards für die Beschreibung von User-Profilen, von Profilen für technische Geräte und für Übertragungskanäle. MPEG-21 unterstützt damit alle praktischen Bedürfnisse, die aus der Nutzung des Internet für den Datenaustausch entstehen.

Die Regeln für den Aufbau und die Beschreibung von Digital Items sind so allgemein und flexibel wie möglich gehalten, um neue Funktionalitäten einfach hinzufügen zu können. Der MPEG-21-Standard bietet für verschiedene Metadatenbereiche Framework-Lösungen an, die die Einbindung anderer Standards ermöglichen. So wird die Identifikation eines Digital Item und seiner Bestandteile durch die Digital Item Identification Language (DII) geregelt. Jedes Digital Item kann mit Rechten versehen werden, hierzu dient die MPEG-21 Rights Expression Language. Um Urheberrechte zu kodieren, kann das MPEG-21-Framework »Intellectual Property Management and Protection« verwendet werden. Digitalen Objekten können durch Digital Item Processing auch Verarbeitungsmethoden zugeordnet werden, vergleichbar den Behaviours von METS. Um Digital Items je nach Endgerät bzw. Netzwerk automatisch anzupassen, steht das MPEG-21-Modul »Digital Item Adaptation« zur Verfügung. Schließlich soll für Digital Items mit MPEG-21 auch ein Datenformat festgelegt werden, das als Container die verschiedenen Metadaten und die Content Information kapselt (noch nicht abgeschlossener Teil des Standards).

Voraussetzung für die Umsetzung dieser umfassenden Aspekte ist in jedem Fall eine Verbindung von Content Information und Metadaten, die derartige Anpassungen automatisierbar macht.

Digital Item Declaration (DID)

Ein Digital Item ist ein mittels eines DIDL-XML-Dokumentes strukturiertes digitales Objekt, das aus der Verbindung von Metadaten mit Ressourcen besteht. Ressourcen entsprechen der Content Information nach dem OAIS Standard.

Bei Ressourcen handelt es sich normalerweise um digitale Informationen, wie beispielsweise Text, Bilder oder Videos. Jedoch können theoretisch auch physische Objekte als Res-

source auftreten. Ein Digital Item könnte beispielsweise auch eine Beschreibung eines Hauses enthalten, welches durch seine Adresse referenziert wird. Wie bei METS können neben Ressourcen, auf die aus der XML-Datei lediglich verwiesen wird, auch Ressourcen als Text, XML oder zeichenkodierte Binärdaten direkt in einem DIDL-Dokument enthalten sein.

Ein Digital Item wird mittels der Digital Item Declaration Language beschrieben, die auf XML basiert und durch ein W3C XML Schema definiert ist.

Wie METS bildet DIDL ein Framework, das ein Set von wenigen Basiselementen zur Verfügung stellt, deren konkrete Verwendung und Füllung für verschiedene Objekte und Anwendungszwecke vom Anwender genauer definiert werden muss.

Ein DIDL-Dokument (XML-Element DIDL) enthält mehrere Hauptabschnitte (XML-Element Container), die der Trennung in signifikante Abschnitte dienen. Die Container enthalten wiederum Item-Elemente, die auch ineinander geschachtelt auftreten und damit eine Hierarchie bilden können. Item-Elemente enthalten Komponenten (XML-Element Component), die Gruppen von Ressourcen darstellen (XML-Element Resource). Ressourcen können Verweise auf Dateien enthalten oder direkt Daten in Text-, XML- oder binärer Form (Base-64-zeichencodiert). Über Verweise ist es möglich, bestehende Digital Items (oder deren Elemente) durch Verlinkung in ein Digital Item einzubinden. So kann – wie bei METS – auch ein verteiltes Digital Item erstellt werden.

An den verschiedensten Punkten kann nun das Metadaten-Element Descriptor positioniert werden, das wiederum eines oder mehrere Statement-Elemente enthalten kann, die einzelne Metadaten-Abschnitte darstellen. Statement-Elemente können wieder beliebigen Inhalt aufweisen oder auf externe Ressourcen verweisen.

DIDL schreibt nicht vor, wie diese Elemente im Detail zu verwenden und zu füllen sind. Ob beispielsweise Metadaten in einem eigenen Container stehen, oder den Container mit den Content Information Items teilen oder jeweils in diesen stehen oder etwa ein eigenes Item bilden, steht dem Nutzer prinzipiell völlig frei. Ebenso ist offen, wie im konkreten Fall mehrteilige Content Information untergebracht wird oder Strukturen abgebildet werden.

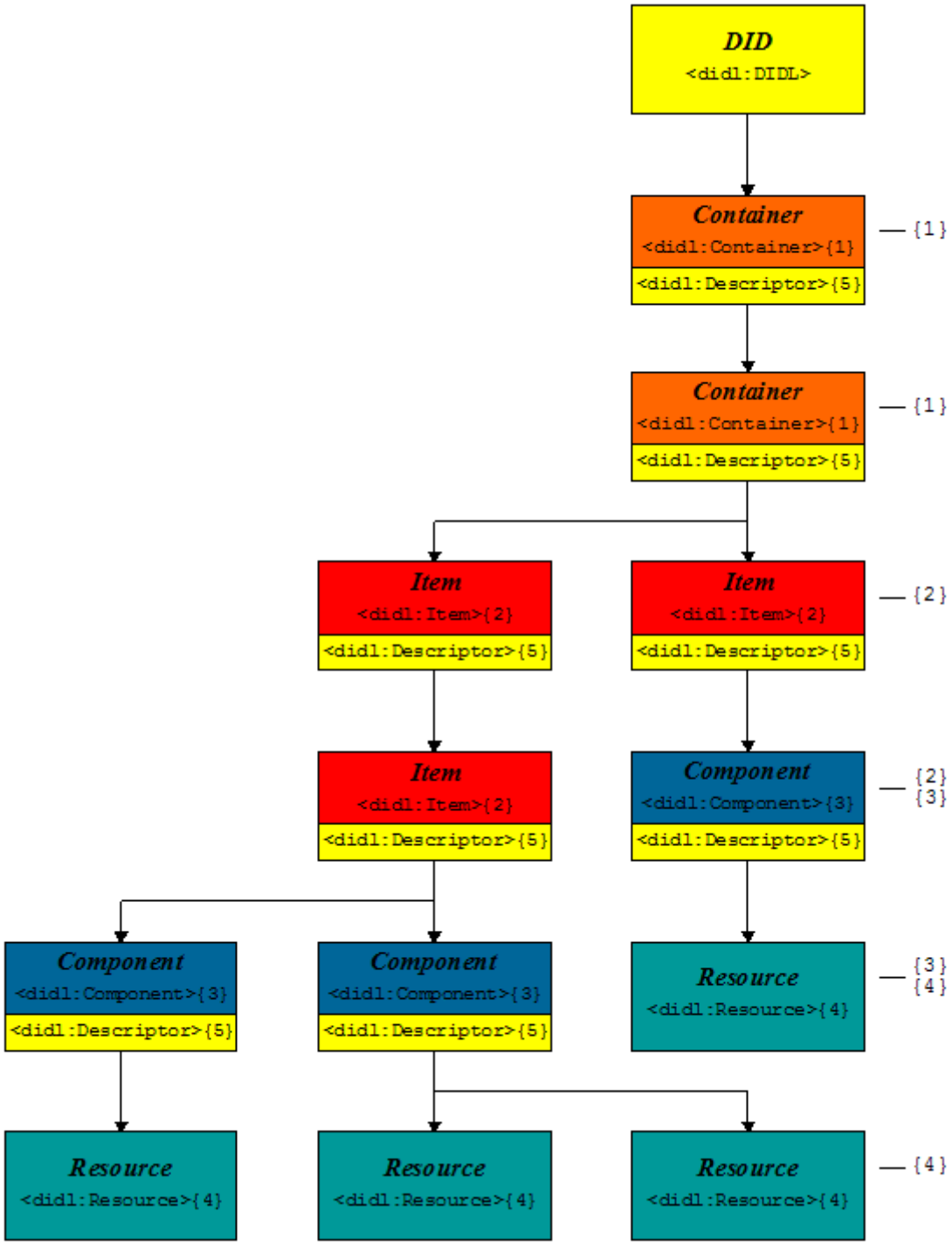


Abb. Konzeptioneller Aufbau einer DIDL-Instanz

Das Framework von DIDL ist also noch weit lockerer als das von METS. Die Verwendung der Metadaten-Elemente ist nicht genauer spezifiziert, so dass – anders als bei METS, wo ja bereits bestimmte Metadatentypen spezifiziert sind – diese erst festgelegt werden müssen.

Metadaten müssen bei METS in einer speziellen Sektion abgelegt werden, auf die von wenigen anderen Elementen aus verwiesen werden kann. In DIDL ist dieses Verfahren ebenfalls möglich, allerdings können Metadaten auch an fast jeder beliebigen anderen Stelle abgelegt werden.

Während METS mit der Structural Map eine eigene feste Strukturkomponente aufweist, fehlt diese bei DIDL und müsste – wenn benötigt – mit einer weiteren vom Nutzer zu definierenden Komponente umgesetzt werden. Hierfür würden sich die entsprechenden XML-Standards zur Beschreibung von Beziehungen eignen (z.B. RDF – Ressource Description Framework oder DAML + OWL). Wo und wie diese Komponente im DIDL-Dokument untergebracht wird, ist ebenfalls Entscheidung des Nutzers.

Genau diese Offenheit stellt die Stärke einer DID dar. So kann ein Statement nicht nur Metadaten enthalten, sondern es kann beispielsweise an den verschiedensten Stellen eingesetzt werden, um IDs mittels der Digital Item Identification Language oder Digital Rights-Informationen zuzuordnen. Auch die Möglichkeiten, Verweise zwischen den verschiedensten Elementen herzustellen, sind bei DIDL sehr offen gestaltet.

Fazit

Die extrem hohe Offenheit und Flexibilität, die DIDL bietet, ermöglicht es, alle Funktionen, die METS beinhaltet, auch mit DIDL-Mitteln zu realisieren. Auch eine von METS abweichende Art der Strukturierung ist möglich. DIDL ist mithin der offenere Standard.

Der Aufwand, ein Profil für DIDL zu schaffen, ist aufgrund der hohen Offenheit und Flexibilität größer, als der, eines für METS zu erstellen. Auch ist dadurch die Wahrscheinlichkeit, dass zwei Anwender vergleichbare Profile für vergleichbare Objekte erstellen, deutlich geringer als bei METS. Die Abstimmung der Nutzung ist daher ein wichtiges Thema, aber eine entsprechende Koordinierungsstelle existiert noch nicht.

Beim DIDL-Projekt der Los Alamos National Laboratory Digital Library wurden beispielsweise an einigen Punkten selbst definierte XML-Elemente eingesetzt, um DIDL und andere Standards zu verbinden. Dieses Vorgehen ist kritisch zu bewerten, da dadurch das Ziel, hohe Interoperabilität durch die ausschließliche Nutzung von Standards zu erreichen, konterkariert wird.

Die extrem hohe Offenheit und Flexibilität von DIDL kann sich daher als ein Schwachpunkt erweisen, insofern sie zu Anwendungen führen kann, die zwar alle dem DIDL-Schema entsprechen, aber untereinander inkompatibel sind. Der Aufwand, solche DIDL-

Implementationen aufeinander abzustimmen kann ebenso hoch ausfallen wie die Abstimmung völlig verschiedener Standards. Im Extremfall wäre also DIDL lediglich das gemeinsame Etikett von im Prinzip grundverschiedenen Umsetzungen.

Der entscheidende Vorteil von DIDL liegt in der Einbettung in den MPEG-21-Standard, der wiederum als Teil der Familie der übrigen MPEG-Standards zu sehen ist. Für viele Aspekte, die für METS völlig offen gehalten sind, bietet es sich bei DIDL an, sich aus dem Baukasten der übrigen MPEG-21-Module zu bedienen und so innerhalb eines Frameworks zu bleiben. Dieser Baukasten dürfte genügend Möglichkeiten bieten, eine komplette OAIS-Anwendung innerhalb des MPEG-21-Standards zu realisieren.

Die Beurteilung der Eignung von DIDL sollte daher auf eine höhere Ebene gehoben werden als die, auf der sich diese Studie bewegt. Es müsste die Eignung des gesamten MPEG-21-Standards für die Archivierung digitaler Objekte evaluiert werden.

Dem steht derzeit noch im Wege, dass sich Teile des Standards noch in Entwicklung befinden. Auch der Grad der Akzeptanz dieses Standards muss sich erst erweisen. Die MPEG-Standards haben bisher durchweg eine hohe Verbreitung im Bereich Multimedia gefunden und es ist möglich, dass dies auch für MPEG-21 zutrifft.

Grundsätzlich ist DIDL als Packaging Standard für E-Journals geeignet, da es ausreichend offen ist, um alle denkbaren Aspekte abzudecken. Eine Implementation eines E-Journal-SIP gibt es unseres Wissens nach derzeit noch nicht, ließe sich aber auf Grundlage des von der Los Alamos National Laboratory Digital Library vorgeschlagenen DIDL-Profils realisieren.

Im Archiv-Bereich hat METS bereits eine hohe Akzeptanz, aber es wird aufgrund seiner Zielsetzung auch auf diesen Bereich beschränkt bleiben. Sollte sich MPEG-21 durchsetzen, wäre sein Vorteil die Geltung auch außerhalb des Archivbereichs und die Einbettung in ein weiteres Framework zur Gestaltung von Information Packages.

Eine Notwendigkeit, aus Sicht der E-Journal-Archivierung DIDL als Standard für SIPs zu empfehlen, besteht derzeit nicht. Die Entwicklung der weiteren Gestaltung, Umsetzung und Verbreitung dieses Standards sollte in jedem Fall weiter sehr genau beobachtet werden. Steht genügend Zeit zur Verfügung, um die weitere Entwicklung von MPEG-21 abzuwarten, wäre der Aufbau eines Archivs im Rahmen dieses Standards eine mögliche Option. Ein späterer Umstieg auf MPEG-21 ist aber aufgrund der Offenheit des Frameworks ebenfalls immer möglich.

IMS Content Packaging Specification / SCORM

Ein weiterer Standard für Information Packages entstammt dem Bereich des Web-Based Training. Auch hier geht es darum, komplexe digitale Objekte unterschiedlichster Art ver-

bunden mit einer Reihe Metadaten auszutauschen und in ganz unterschiedlichen Umgebungen (LMS – Learning Management Systems) abzuspielen. Außerdem sollen einmal erstellte Lerninhalte auch bei weiterentwickelter Technologie verwendbar bleiben. Das Interesse ist also dem der Langzeitarchivierung nicht unähnlich.

Diese Ziele werden auch von der europäischen Initiative ARIADNE (Authoring and Distribution Network for Europe), dem US-amerikanischen Instructional Management Systems Global Learning Consortium (IMS) und der vom US-Verteidigungsministerium initiierten Advanced Distributed Learning Initiative (ADL) verfolgt.

Quellen: *Websites der beteiligten Organisationen: www.adlnet.org. – www.ariadne-eu.org. – www.imsglobal.org. – weitere Informationen von E-Learning-Anbietern: http://www.l3s.de/elan/kb3/index.php?id=128. – www.rhassociates.com/scorm.htm*

Aufbau des IMS Framework

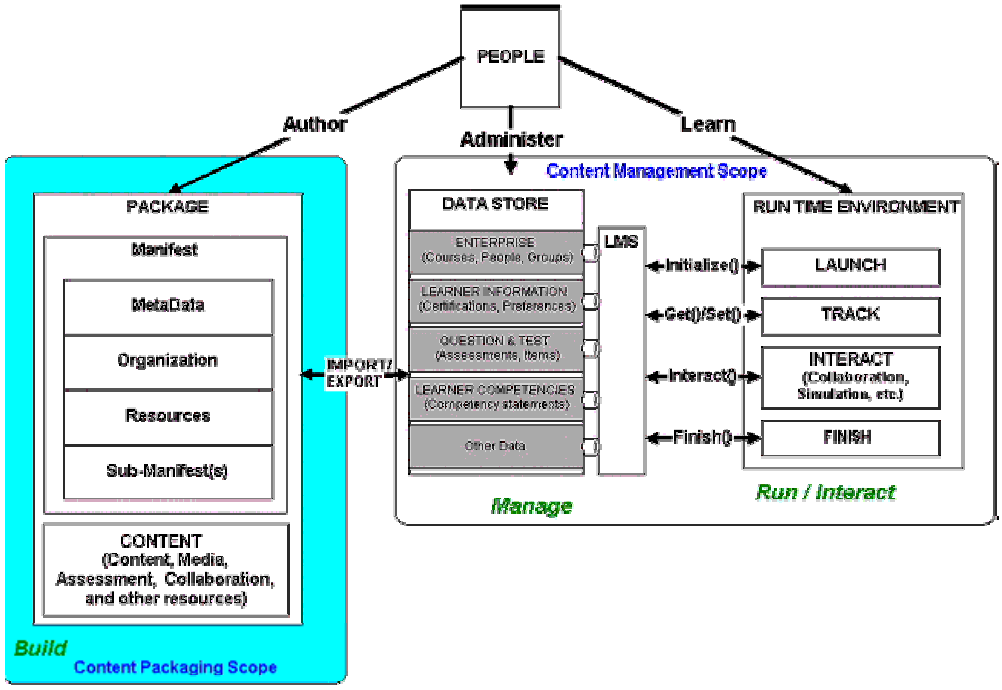


Abb. IMS Framework

IMS und ARIADNE entwickelten in den vergangenen Jahren gemeinsam ein XML-basiertes Information Package-Modell, das unter dem Namen IMS Content Packaging Specification firmiert. ADL entwickelte bereits vorher das Sharable Content Object Reference Model (SCORM). SCORM umfasst ein Set von Spezifikationen, die die verschiedenen Aspekte des Web-Based Training abdecken. SCORM verwendet dabei die IMS Content Packaging Specification.

Das IMS Content Package

Ein IMS Content Package wird durch ein W3C XML-Schema definiert. Die aktuelle Version ist 1.1.4 vom 4. Oktober 2004. Das IMS Content Package besteht äußerlich aus einem Pakkage Interchange File (PIF), das ein komprimiertes Archiv (z.B. in ZIP-Form) darstellt. Es ist auch möglich, statt einer Archiv-Datei einen unkomprimierten Datenträger zu verwenden, dessen Inhalt dem einer IMS-konformen Archiv-Datei entsprechen muss. In der Archiv-Datei bzw. auf dem Datenträger befindet sich eine XML-Datei mit dem Standardnamen imsmanifest.xml, die die Metadaten und die Strukturinformationen des Packages enthält, gegebenenfalls vom Anwender hinzugefügte DTDs oder XSDs, die für Erweiterungen IMS Content Package notwendig sind, sowie weitere Dateien beliebigen Formats, die die Content Information enthalten.

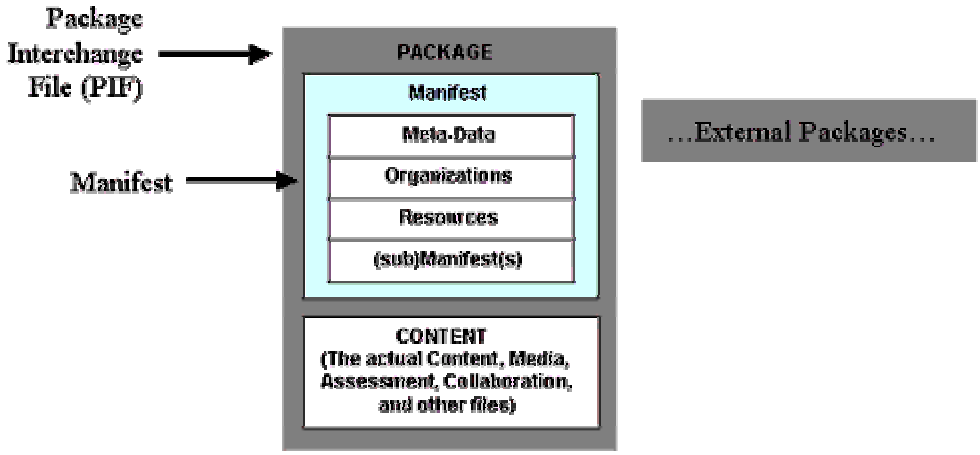


Abb. IMS Content Package

Die Manifest-XML-Datei besteht aus drei Abschnitten und gegebenenfalls Submanifesten, die selbst so aufgebaut sind wie ein Hauptmanifest. Das Haupt-XML-Element ist manifest.

IMS Content Packages sind METS-Packages sehr ähnlich. Der optionale Abschnitt Metadaten (XML-Element metadata) entspricht im Wesentlichen dem Metadaten-Abschnitt von METS. Das heißt, er kann beliebige, vom Nutzer genauer zu bestimmende XML-Inhalte enthalten, aber – anders als METS – keine binären Inhalte. Mittels der Elemente schema und schemaversion kann angegeben werden, welches XML-Schema den eingefügten Metadaten zugrunde liegt. Anders als bei METS gibt es keine weitere Standardunterteilung des Metadaten-Abschnittes. Zudem kann ein Metadaten-Abschnitt nicht nur als Teil des Gesamtpackages auftreten, sondern auch in den anderen Unterabschnitten des Manifests. Ein IMS Package kann also mehrere Metadaten-Abschnitte an den verschiedensten Punkten aufweisen, wobei sich die Metadaten jeweils auf den Teil beziehen, in dem sie positioniert sind. Externe Metadaten sind derzeit nicht vorgesehen, können aber dadurch referenziert

werden, dass man anstelle der Metadaten einen Link auf solche einfügt. Dies wäre aber kein Bestandteil des IMS Content Packaging Standard, sondern eine Erweiterung, die das IMS-Konzept nicht standardmäßig unterstützt:

```
<metadata>
   <schema>ADL SCORM</schema>
   <schemaversion>CAM 1.3</schemaversion>
   <adlcp:location>Lesson01.xml</adlcp:location>
</metadata>
```

Der obligatorische Abschnitt Organizations (XML-Element organizations) entspricht prinzipiell dem Structure-Abschnitt von METS. Er kann eine oder mehrere Strukturbeschreibungen enthalten (XML-Element organization). Ein Organization-Element enthält Item-Elemente, die ineinander geschachtelt auftreten können und so die Hierarchie wiedergeben. Organization- und Item-Elemente können zur näheren Beschreibung Title- und Metadata-Elemente enthalten. Auf Ressourcen (Content Information) kann über das Attribut identifierref der Item-Elemente verwiesen werden. Über dieses Attribut können auch Unter-Manifest-Abschnitte referenziert werden, die weitere Details enthalten, aber Teil desselben Manifests sein müssen. Externe Referenzen sind nicht vorgesehen.

Außerdem können alle genannten XML-Elemente auch beliebigen anwenderdefinierten XML-Inhalt aufweisen, so dass über diesen Umweg Referenzen – z.B. auf externe Items oder auf Metadatenabschnitte – möglich würden.

```
<organizations>
      <organization identifier="TOC1">
            <title>Default Organization</title>
            <metadata>
             [any metadata]
            </metadata>
            <item identifier="ITEM1" identifierref="RESOURCE1">
                  <title>Lesson 1</title>
                  <item identifier="SUBITEM1" identifier-
                  ref="RESOURCE1-1">
                        <title>Task 1</title>
                  </item>
                  <item identifier="SUBITEM2" identifier-
                  ref="RESOURCE1-2">
                        <title>Task 2</title>
                  </item>
            </item>
            <item identifier="ITEM2" identifierref="SUBMANIFEST1">
```

```
                <title>Lesson 2</title>
            </item>
        </organization>
</organizations>
```

Der obligatorische Abschnitt resources (XML-Element resources) entspricht im Wesentlichen dem File-Abschnitt von METS. Er kann eine oder mehrere Ressourcengruppen enthalten (XML-Element resource). Ein resource-Element enthält ein oder mehrere File-Elemente, die Ort und Name der Dateien enthalten. File-Elemente können nur auf Dateien innerhalb des PIF oder auf externe Dateien verweisen, aber nicht wie das Äquivalent bei METS Dateiinhalte aufnehmen. Allerdings kann auch ein resource-Element oder ein File-Element beliebigen, anwenderdefinierten XML-Inhalt aufweisen, so dass über diesen Umweg direkte Dateninhalte möglich würden.

Alle File-Elemente eines resource-Elements bilden zusammen die Ressource (beispielsweise eine HTML-Datei mit dazugehörigen Stylesheets, Grafiken und Scripten). Mittels des href-Attributes von resource kann angegeben werden, welches die Hauptdatei der Ressource ist. Zudem kann ein resource-Element das Element dependency enthalten, das dann verwendet wird, wenn eine Gruppe von Dateien für mehrere resource-Elemente relevant ist. Mit dependency kann aus betreffenden resource-Elementen auf die ID anderer resource-Elemente verwiesen werden (beispielsweise wenn mehrere HTML-Dateien auf die gleiche Gruppe von Grafiken für Buttons oder Scripte zugreifen sollen).

Fast alle XML-Elemente innerhalb des Abschnitts resources können auch eigene Metadata-Elemente enthalten.

Nachstehend ein Beispiel für einen resources-Abschnitt:

```
<resources>
      <resource identifier="R_A1" type="webcontent"
      href="sco06.html">
            <metadata>
             [any metadata]
            </metadata>
            <file href="sco06.html" />
            <file href="scripts/APIWrapper.js" />
            <file href="scripts/Functions.js" />
            <dependency identifierref="R_A4" />
            <dependency identifierref="R_A5" />
            <dependency identifierref="R_A6" />
      </resource>
      <resource identifier="R_A2" type="webcontent"
      href="sco1.html">
```

```
            <metadata>
             [any metadata]
            </metadata>
            <file href="scol.html" />
            <file href="scripts/APIWrapper.js" />
            <file href="scripts/Functions.js" />
            <dependency identifierref="R_A5" />
      </resource>
      <resource identifier="R_A4" type="webcontent"
      href="pics/distress_sigs.jpg">
            <file href="pics/distress_sigs.jpg" />
      </resource>
      <resource identifier="R_A5" type="webcontent"
      href="pics/distress_sigs_add.jpg">
            <file href="pics/distress_sigs_add.jpg" />
      </resource>
      <resource identifier="R_A6" type="webcontent"
      href="pics/nav_aids.jpg">
            <file href="pics/nav_aids.jpg" />
      </resource>
</resources>
```

Fazit

Die IMS Content Packaging Specification ist der METS -Spezifikation vom Ansatz und Aufbau her sehr ähnlich. Die genauere Analyse hat gezeigt, dass IMS weniger komplex ist als METS und dass eine Reihe von Möglichkeiten, die METS bietet, bei IMS erst durch Erweiterungen möglich werden. Die IMS Content Packaging Specification bietet dabei keine Eigenschaften, die über METS hinausgehen und IMS als Content Packaging Standard für E-Journals geeigneter erscheinen lassen würden. Das Gesamtframework von IMS und SCORM ist speziell auf die Bedürfnisse des Web-based Training orientiert, so dass auch der Framework-Kontext dieser Content Packaging Specification keinen Anlass gibt, diese für ein E-Journal-SIP-Konzept in die engere Wahl zu ziehen.

CCSDS Packaging Standard

Auch das Consultative Committee for Space Data Systems (CCSDS), federführend zuständig für den OAIS-Standard, arbeitet seit 2001 an einem Entwurf für einen XML-basierten Pakkaging Standard. Ziel ist die Definition eines SIP, mit dem die unterschiedlichsten Daten an das Archiv geliefert werden können.

Im Rahmen dieser Arbeiten wurden die folgenden Kernanforderungen definiert:

- Abwärtskompatibilität mit dem bisher vom CCSDS empfohlenen Standard SFDU (Standard Formatted Data Units)
- die Fähigkeit, mit beliebigen Binärdaten umgehen zu können
- die Fähigkeit, Beziehungen zwischen Anwendungsdaten und Metadaten festlegen zu können
- Erweiterbarkeit
- Interoperabilität

Als mögliche Vorbilder wurden die Konzepte der Packaging Standards von METS, IMS und MPEG-21 evaluiert und die Entwicklungsarbeiten unter das Motto »Adopt techniques others using and do simple things simply" gestellt.

Parallel mit der Entwicklung eines Packaging Standards erfolgt die Entwicklung eines »Producer-Archive Interface Methodology Abstract Standard" (PAIMAS), der die Interaktion zwischen Archiv und Produzenten regeln soll.

Ursprünglich war vorgesehen, den Packaging Standard 2004 bis zum Status eines CCSDS-White Book (Preliminary Draft Recommendation that is distributed only within a Panel) voranzubringen und einen SIP-Prototypen zu vorzustellen. Allerdings haben mehrmonatige Verzögerungen dazu geführt, dass der Prototyp erst im März 2005 zur Verfügung stehen wird. Die Entwicklung des Standards befindet sich also noch immer in einem nichtöffentlichen Stadium, und es wird sicher noch einige Zeit bis zu seiner Publikation vergehen.

Eine Bewertung des CCSDS Packaging Standards hinsichtlich seiner Eignung für ein E-Journal-SIP ist mithin noch nicht möglich. Aufgrund der Komplexität und Heterogenität der Daten, die im Bereich der Weltraum-Forschung anfallen und für die der Packaging Standard des CCSDS geeignet sein muss, ist eine so große Offenheit und Flexibilität dieses Standards notwendig, dass zweifellos auch ein Profil für E-Journals in diesem Rahmen realisierbar sein dürfte.

Wie auch beim DIDL-Standard wäre hier eine Evaluierung der Eignung des gesamten Frameworks und der weiteren damit verbundenen Standards für die Langzeitarchivierung von

nicht raumfahrtspezifischen Inhalten notwendig. Aufgrund der Tatsache, dass nicht alle relevanten Teile des Frameworks als verabschiedete und veröffentlichte Standards vorliegen, ist eine solche Evaluierung derzeit noch nicht möglich.

Da das CCSDS bestrebt ist, seine Standards auch in den Rang von ISO-Normen zu erheben, könnte hier eine ISO-Alternative zu MPEG-21 entstehen. Während aber MPEG-21 einen viel umfassenderen Anspruch hat, nämlich den Austausch von Multimediadaten überhaupt zu standardisieren, geht es dem CCSDS um die Archivierung wissenschaftlicher Primärdaten. Es ist daher wahrscheinlich, dass die CCSDS-Standards schon enger an archivalische Bedürfnisse angepasst sind als MPEG-21.

Bei gegenwärtigen Stand der Dinge müssen wir uns auf die Aussage beschränken, dass mit dem CCSDS Packaging Standard ein weiterer ernstzunehmender Kandidat für die Gestaltung eines E-Journal SIP im Entstehen ist, dessen konkrete Ausformung und Anbindung an andere relevante Standards weiter beobachtet werden sollte.

Quellen: *Offizielle CCSDS-Website: http://public.ccsds.org/*

ONIX

ONIX ist ein XML-basierter Standard für den Datenaustausch zwischen Verlagen und Buchhandel, der inzwischen auch von einigen Bibliotheken verwendet wird. Da ONIX auch die Einbeziehung von beliebigen Textinhalten sowie Bild-, Video- und Audiodaten ermöglicht, liegt es nahe zu evaluieren, ob ONIX möglicherweise auch als Format für ein E-Journal-SIP in Frage kommt.

Angestoßen wurde die Entwicklung von ONIX von der Association of American Publishers, weiterentwickelt und vertreten wird es vom ONIX International Steering Committee der internationalen Dachorganisation für Standardisierung im Buchhandel EDItEUR mit Sitz in London. Weiterhin gibt es nationale Koordinierungsstellen, in denen Vertreter von Verlagen und Buchhandel die Anwendung in nationalen Kontexten abstimmen sowie Vorschläge zu Weiterentwicklung und Erweiterung des Standards an das Steering Committee übermitteln.

Die aktuelle ONIX-Version ist 2.1 Revision 2 von 2004.

Quellen: *Offizelle ONIX-Website: www.editeur.org*

ONIX als Datenaustauschformat im Buchhandel

ONIX ist ein Akronym für ONline Information eXchange. Der Anlass für die Entwicklung von ONIX war die Notwendigkeit, für den Online-Buchhandel im Internet mehr Informa-

tionen über die Titel zur Verfügung zu stellen als nur einfachste bibliographische Daten. Für die Präsentationen auf den Webseiten der Online-Shops werden webfähige Abbildungen der Titelseite und ggf. der Autoren, Klappentexte, Rezensionsauszüge, Inhaltsübersichten, Textbeispiele, Autorenbiografie und ähnliches benötigt. Um die Übermittlung und Verwendung dieser Daten effizient zu gestalten, wurde ein Standard benötigt und mit ONIX entwickelt.

Der Vorteil einer Standardisierung von bibliographischen und anderen vertriebsrelevanten Informationen kommt nicht nur im Kontext des Online-Buchhandels zum Tragen, sondern ist geeignet, den gesamten Informationsaustausch zwischen Verlagen und Buchhandel zu effektivieren. Interne, meist miteinander inkompatible Formate der Verlage und Buchhandlungen durch ein einheitliches System zu ersetzen, ermöglicht einen automatisierbaren Informationsaustausch, der die Übernahme von Produktdaten von Geschäftspartnern außerordentlich erleichtert. Da auch das Verzeichnis lieferbarer Bücher und die großen Barsortimenter ONIX als Standard übernommen haben, ist der Anreiz für Verlage, ebenfalls auf ONIX umzusteigen, sehr hoch.

Auch für Bibliotheken ist es vorteilhaft, diesen Standard zu akzeptieren, da sie so Katalogdaten aus ONIX-Datensätzen der Verlage erzeugen können. Beispielsweise akzeptiert Die Deutsche Bibliothek Titelmeldungen im ONIX-Format.

ONIX als Standard für Information Packages

Der ONIX-Standard von EDItEUR wird auch als »ONIX for Books« bezeichnet, da sich auch eine Version für Video- und Audioprodukte in Entwicklung befindet. »ONIX for Books« ist jedoch nicht nur für Bücher geeignet, sondern für Printprodukte im weitesten Sinne, also auch für Reihen und Periodika. Für Periodika existiert eine Variante »ONIX for Serials«. Bei Periodika ist es möglich, auch einzelnen Beiträge jeweils ein ONIX-Dokument zuzuordnen, so dass prinzipiell auch das Medium E-Journal mitbedient werden kann. Auch eine Kategorie e-publications existiert. Eine Offenheit anderen, nicht textorientierten Medien gegenüber bringt ONIX anders als METS, DIDL oder IMS jedoch nicht mit.

ONIX ermöglicht neben der Ablage von Metadaten auch die Kodierung von Textinhalten sowie Bild-, Video- und Audiodaten. Eine gründliche Analyse der ONIX-DTD und der Anwendungsdokumentation hinsichtlich der zentralen Aspekte für die Eignung als Standard auch für Information Packages umfasst die Bereiche Metadaten, Strukturbeschreibung, Offenheit und Erweiterbarkeit sowie Mittel zur Herstellung von Verbindungen oder zur direkten Aufnahme von Content Information.

Metadaten

ONIX bringt einen eigenen Standard für beschreibende Metadaten mit. Dieser umfasst Dutzende von Elementen und Elementgruppen zur Beschreibung aller erdenklichen beschreibenden und vertriebstechnischen Informationen unter Beachtung der Besonderheiten des Buchhandels in den verschiedenen Mitgliedsländern von EDItEUR. Eine Kompatibilität mit archivalischen Standards steht bei ONIX nicht im Vordergrund. Es existiert eine Konkordanz zu MARC, die jedoch in keine Richtung vollständig ist.

Technische Metadaten beinhaltet ONIX nur sehr rudimentär. So können für den gesamten Bereich der sogenannten MediaFiles, also Bild-, Audio- und Videofiles nur sechs Datenformate angegeben werden (GIF, JPEG, PDF, TIF, RealAudio und MP3). Die hier zur Verfügung stehenden Dateitypen sind auf Produkt- und Logodaten sowie Softwaredemonstrationen und Audio- und Videoausschnitte beschränkt.

Strukturbeschreibung

ONIX weist keinen Bereich zur Strukturbeschreibung auf.

Offenheit und Erweiterbarkeit

Erweiterbarkeit ist ein zentrales Thema, wenn es um Packaging Standards geht. Bei ONIX hingegen spielt sie keine bedeutende Rolle. ONIX versteht sich als ein Standard, der für alle XML-Elemente eine eindeutige Beschreibung bietet, so dass jeder ONIX-Nutzer mit jeder ONIX-Datei sofort umgehen kann, ohne sich um Profile oder etwaige Erweiterungen kümmern zu müssen.

Änderungen oder Erweiterungen dürfen nur dann vorgenommen werden, wenn ihre Nutzung sich auf rein interne Anwendung innerhalb eines Unternehmens oder einer Organisation beschränkt – also der eigentliche Zweck von ONIX, der Datenaustausch zwischen verschiedenen Partnern der Vertriebskette, ausgeschlossen bleibt.

Die mangelnde Erweiterbarkeit ist auch technologisch bedingt. Obwohl für ONIX nicht nur eine DTD, sondern auch ein XML Schema vorhanden ist, ist es nicht möglich, beliebiges anwenderdefiniertes XML an dafür freigegebenen Stellen einzufügen. Solche Erweiterungen wären mit Schema-Mitteln zwar realisierbar, die XML-Daten würden aber nicht mehr der DTD entsprechen.

Für eine kleine Zahl von ONIX-Elementen ist das Einfügen von XHTML, HTML und beliebigem XML- und SGML-Code zwar laut Standard erlaubt. Allerdings zeigt sich bei genauerer Lektüre von DTD und Dokumentation, dass nur echter XHTML-Code eingefügt werden kann, da die XHTML-DTD als Modul in die ONIX-DTD integriert ist. Sobald anderer HTML-, XML- oder SGML-Code eingefügt werden soll, muss dieser maskiert werden. Das heißt, dass

die Delimiter der Tags mit Ersatzzeichenfolgen ausgetauscht werden müssen, damit die ONIX-DTD hier keine Fehler meldet. Somit handelt es sich um keinen intakten Code mehr, der ohne besondere Vorkehrungen (Rückgängigmachen der Maskierung) verwendet werden kann.

Einbindung von Content Information

ONIX hat nicht den Anspruch, Inhalte oder Referenzen zu Inhalten zu transportieren, sondern diese nur zum Zweck der Verkaufsförderung, z.B. als Leseproben, also lediglich als Ergänzung zu den Metadaten zur Verfügung zu stellen.

Links auf Textinhalte sind nicht möglich. Wie oben beschrieben, können auch keine anwenderdefinierten XML-Inhalte außer XHTML integriert werden, ohne diese zu maskieren. Die Integration dieser Inhalte erfolgt zudem in Elementen, die ausschließlich vertriebsrelevante Inhalte aufnehmen sollen. Allenfalls das Element Text käme in Frage. Diesem Element kann über das Element TextTypeCode ein Code zugewiesen werden, der den Inhaltstyp des Textes angibt. Diese Typen sind bis auf FullText aber wiederum Inhalte, die vertriebsrelevant sind (ONIX Code List 33). Der Kontext des Elementes Text macht diese Intention vollends deutlich: es gehört zur Element-Gruppe 15 »Descriptions and other supporting text«.

Das Gleiche trifft auf die sogenannten MediaFiles zu, also Bild-, Audio- und Videofiles. Hier sind nur Links auf externe Ressourcen möglich. Die zur Verfügung stehenden Typen sind zudem auf Produkt- und Logodaten sowie Softwaredemonstrationen und Audio- und Videoausschnitte beschränkt .

Fazit

ONIX ermöglich in der Tat die Einbeziehung von beliebigen Textinhalten sowie Bild-, Video- und Audiodaten. Ebenso ist ONIX auf E-Journals anwendbar. Allerdings beschränkt sich ONIX ganz auf den Austausch von beschreibenden Produkt-Metadaten und begleitenden, vertriebsrelevanten Informationen. Die Integration oder Referenzierung von Content Information ist nur in diesem Rahmen und in extrem eingeschränktem Maße möglich. Mittel zur Strukturbeschreibung fehlen völlig, technische Metadaten sind nur rudimentär vorhanden. Auch Erweiterungen, die hier Abhilfe bieten würden, sind weder konzeptionell noch technisch vorgesehen. Die Nutzbarkeit von ONIX im archivarischen Kontext beschränkt sich daher auf den Austausch von beschreibenden Metadaten im weitesten Sinne. Eine Eignung als Standard für Information Packages liegt nicht vor.

Packaging Standards – Fazit

Von den fünf analysierten Standards haben sich im Vergleich drei Kandidaten als geeignet herausgestellt.

METS ist ein Standard, der alle Anforderungen erfüllt, gleichzeitig offen und flexibel sowie archivierungsorientiert gestaltet ist. METS kann unterschiedlichste Daten-, Metadaten- und Strukturinformationen aufnehmen und ist für die Integration beliebiger Metadatenstandards offen. Ein großes Plus für METS ist das formalisierte Konzept von Profilen verbunden mit einer zentralen Profil-Verwaltungsstelle. So kann jeder interessierte Anwender auf schon vorhandene Profile anderer Anwender zurückgreifen oder wenn nötig seine individuelle METS-Implementierung entwickeln und zugleich auf eine optimale Abstimmung mit vorhandenen Profilen achten. Zugleich weist METS einen großen Bekanntheitsgrad und eine hohe Akzeptanz im internationalen Archivumfeld auf. Anwendungen für E-Journals existieren bereits. Der Standard ist für die Anwendung auf E-Journal-SIPs voll geeignet.

MPEG-21 DIDL ist ein neuer Standard, der der umfassenden Standardisierung des Austauschs von komplexen Datenpaketen im Internet und anderen Datenfernübertragungswegen dienen soll. Aus diesem Grund ist DIDL extrem offen und flexibel gestaltet, so dass es unterschiedlichste Daten-, Metadaten- und Strukturinformationen aufnehmen kann. Die Offenheit von DIDL kann allerdings bei mangelnder Koordination zu inkompatiblen Implementationen führen. Die Anwendung auf E-Journal-SIPs ist möglich, wobei weitere Standards für Metadaten- und Strukturinformationen herangezogen werden können und müssen. Besonders interessant ist die Einbindung in die MPEG-Standard-Familie, die sich einer hohen Akzeptanz erfreut. Der Aufwand, ein DIDL-basiertes E-Journal-SIP zu definieren ist nur dann gerechtfertigt, wenn weitere Argumente für MPEG als Gesamtframework für eine digitale Archivlösung sprechen. Da insbesondere der Teil MPEG-21 – Metadaten für Multimedia-Objekte, zu dem DIDL gehört, noch nicht abgeschlossen ist, ist hier eine weitere Beobachtung der Entwicklung notwendig, bevor ein abschließendes Urteil möglich ist.

Die gleichen Erwägungen gelten für den Packaging Standard des CCSDS, der sich in einem späten Entwicklungsstadium befindet und in ein bis zwei Jahren den Status einer ISO-Norm erreicht haben soll. Dabei hat der CCSDS Packaging Standard keinen so umfassenden Anspruch wie DIDL, sondern stammt aus Archivkreisen und wird daher speziell auf Archivierungsbedürfnisse abgestimmt sein.

Der IMS Packaging Standard ist mit METS vergleichbar, die Spezifikation bietet jedoch keine Eigenschaften, die über METS hinausgehen und IMS als Content Packaging Standard für E-Journals geeigneter erscheinen lassen würden.

Grundsätzlich nicht geeignet ist ONIX, dessen Zielsetzung zu speziell auf den Austausch von Produktinformationen und zu wenig auf den Austausch von komplexen Inhalten mit den dazugehörigen Metainformationen orientiert ist. Ebenso fehlt ONIX die Offenheit, um diese Mängel durch Erweiterungen zu beheben.

Nach dem aktuellen Stand der Dinge empfehlen wir die Verwendung von METS. Ob sich DIDL durchsetzen kann und ob die MPEG-Standard-Familie für Archivlösungen geeignet ist, müssen zukünftige Analysen ergeben. Beides gilt auch für den noch unvollendeten CCSDS-Standard und das dazugehörige Framework. Als Alternative für METS erscheint uns jedoch DIDL attraktiver als CCSDS zu sein, da mit DIDL die Einbindung eines Standards möglich wäre, der bei großer Akzeptanz weit über die Archivwelt hinaus von Relevanz wäre. Da ein Umstieg von METS auf DIDL in jedem Fall möglich ist, würde die Verwendung von METS einen späteren Wechsel auf DIDL nicht behindern.

Beispiele für ein E-Journal SIP

Hier empfehlen wir die Heranziehung des Harvard-EJAR-Profils und die Kontaktaufnahme mit den zuständigen Bearbeitern. Da diese sich der Schwächen der Begrenzung auf ein issue-orientiertes SIP bei der Konzeption bewusst waren, ist anzunehmen, dass es auch bei EJAR Lösungsansätze für ein feiner granuliertes SIP gibt. Hier könnte eine Kooperation zu effektiveren und kompatibleren Lösungen führen als die Erstellung eines ganz eigenen Entwurfs. Vorausgesetzt muss werden, dass das Harvard EJAR eine hohe Produzentenbeteiligung erfordert, die bei einem deutschen Projekt dieser Art ebenfalls notwendig wäre. Hier sind die anstehenden gesetzlichen Regelungen relevant (Novelle des Archivgesetzes).

Transfer der Information Packages

Eine weitere entscheidende Frage ist die nach der Art der Übermittlung der Information Packages vom Produzenten an das Archiv. Hier ist zunächst zwischen Push- und Pull-Lösungen zu unterscheiden. Push bedeutet, dass der Produzent die Daten zum Archiv »schiebt«, während bei Pull das Archiv die Daten abholt. Aus rein technischer Sicht gibt es hier keine grundlegenden Unterschiede. Ob die Daten von einem Ort über internetspezifische Techniken auf einen anderen Ort hochgeladen oder andersherum heruntergeladen werden, ist nicht bedeutsam. Bei gleichem Aufwand, der vor und nach dem Transportvorgang der Daten für die Sicherung der Datenqualität getroffen wird, wird kein Unterschied zu bemerken sein.

Bedeutsam ist jedoch von der organisatorischen Seite her die Frage, wer den Vorgang auslöst und wer dabei für welche Aspekte der Datenübergabe welche Verantwortung übernimmt.

Pull-Lösungen stellen eine primäre Aktivität des Archivs dar und sind daher in der Praxis auch mit der Übernahme anderer Aktivitäten durch das Archiv zugunsten eines geringeren Aufwandes beim Produzenten verbunden. Bei Push-Lösungen ist dagegen eine größere Aktivität des Produzenten erforderlich, was sich auch in seiner stärkeren Beteiligung an anderen Aspekten der Datenübergabe fortsetzt.

Pull-Lösung / Geringe Produzentenbeteiligung

Pull-Lösungen erfordern keine oder geringe Aktivität des Produzenten, so dass sie dort verwendet werden, wo die Produzentenbeteiligung minimiert werden soll, weil die Produzenten nur dann zur Mitarbeit bereit sind, wenn sie möglichst wenig Aufwand haben (Beispiel LOCKSS). Daher geht hier die Aktivität beim Transfer primär vom Archiv aus, was auch Folgen für die Strategien zur Problemlösung hat.

Der Produzent teilt dem Archiv lediglich mit, wo es die Daten abholen kann, so dass das Archiv für den Transfer zuständig ist und die Daten herunterlädt.

Pull-Lösungen und die damit verbundene Strategie der Aufwandsminimierung für den Produzenten haben jedoch erhebliche Nachteile.

Zum einen ist so der Zugriff auf datenbankgestützte Inhalte, die erst auf Anfrage des Nutzers erzeugt werden, nicht möglich oder mit entsprechendem Aufwand verbunden, der den Vorteil des Pull-Systems konterkarieren würde.

Zum anderen findet zugunsten der Aufwandsminimierung für den Produzenten keine Qualitätsprüfung bei diesem statt, so dass die Datenqualität nur vom Archiv überprüft wird. Probleme würden erst dort erkannt. Die Behebung durch den Produzenten setzt in jedem Fall eine mehrfache Kommunikation zwischen Archiv und Produzent voraus. Die mehrfache Kommunikation ergibt sich daraus, dass nicht nur das Archiv die Fehler mitteilen, sondern auch der Produzent dem Archiv mitteilen muss, wann es die korrigierten Daten neu abholen kann.

Aus diesen Gründen ist beim Pull-Konzept nur eine minimierte Qualitätsüberprüfung möglich, was auch einen Verzicht auf einen hohen Standardisierungsgrad mit sich bringt.

Auch Probleme bei der Auffindung der Daten (Daten nicht am angegebenen Ort) müssen durch Aktivität des Archivs und mehrfache Kommunikation zwischen Archiv und Produzent gelöst werden.

Schließlich ist der Umgang mit journalbezogenen Daten hier nicht leicht zu lösen, da der Auslöser, der dazu führt, dass nicht nur neue Artikel, sondern auch Informationen über das Herausgeberkollegium u.ä. wegen Änderungen neu geholt werden müssen, nur über eine stärkere Einbindung des Produzenten möglich ist.

Pull-Lösungen empfehlen sich also dort, wo aus Mangel an Mitteln eine Archivierung ohne großen Aufwand und unter Verzicht auf hohe Qualitätssicherungsansprüche durchgeführt werden soll.

Push-Lösung / Hohe Produzentenbeteiligung

Bei Push-Lösungen stellt der Produzent die Datenpakete zusammen und lädt sie an einen angegebenen Ort zum Archiv hoch (Beispiel Harvard EJAR). Da er damit als der aktive Part die Verantwortung für die Korrektheit der Daten in einer stärkeren Weise übernimmt als der Produzent bei Pull-Lösungen, ist hier auch das Interesse an der Überprüfung der Korrektheit größer. Dies geht einher mit der Bereitschaft zu einer größeren Aktivität im Interesse einer optimalen Archivierung.

Idealerweise verfügt der Produzent über Tools zur Kontrolle (eine Überprüfung beim Produzenten z.B. durch vom Archiv online zur Verfügung gestellte Prüfmechanismen ist in dieser Hinsicht einer Überprüfung beim Produzenten durch eigene Mittel gleichgestellt). Wenn er ein nicht korrektes SIP vom Archiv mit Korrekturkommentaren zurückerhält, obliegt es ihm, die Korrekturen vorzunehmen und erneut den Upload zu starten. Eine Kommunikation mit dem Archiv und eine ausführende Aktivität des Archivs ist hier anders als beim Pull-Konzept nicht notwendig.

Probleme bei der Auffindung der Daten entfallen.

Datenbankgestützte Inhalte, die erst auf Anfrage des Nutzers erzeugt werden, stellen kein Problem dar, da hier die zu archivierenden Einheiten vom Produzenten erstellt werden, der intern einen vollen Zugriff auf seine Datenbanken hat und entsprechende SIPs leicht und automatisch erzeugen kann.

Journalbezogene Daten können vom Produzenten als SIP eingeschickt werden, wenn sich Änderungen ergeben haben.

Bei Lösungen mit hoher Produzentenbeteiligung ist es auch eher möglich, Standards zu definieren und einzuhalten, was alle weiteren Bearbeitungsschritte im Archivsystem erheblich erleichtert und eine effektive Langzeitarchivierung erst möglich macht.

Vorausgesetzt ist freilich die Bereitschaft der Produzenten zur Mitarbeit, sei sie aus eigenem Interesse vorhanden oder durch gesetzliche Regelungen bewirkt.

Fazit

Technisch besteht kein signifikanter Unterschied darin, ob der Produzent die SIPs zum Archiv hochlädt (Push-Prinzip) oder ob sich das Archiv die Daten herunterlädt (Pull-Prinzip). Aus der Workflow-Sicht entlastet das Pull-Prinzip den Produzenten, erschwert aber die Qualitätssicherung. Zudem kann es keine datenbankgenerierten Inhalte laden, es sei denn, es wird ein größerer Zusatzaufwand getrieben. Das Pull-Prinzip ist daher für Lösungen geeignet, bei denen die Beteiligung auch solcher Produzenten, die nur einen sehr geringen Aufwand zugunsten der Langzeitarchivierung leisten können, wichtiger ist als eine optimierte Qualitätssicherung. Das Push-Prinzip ermöglicht dagegen eine optimale Datenqualität und bietet den Zugang zu datenbankgenerierten Inhalten, setzt aber eine aktive und intensivere Beteiligung des Produzenten voraus.

Teil 3

Umfrageauswertung

Die im Rahmen dieser Studie durchgeführte Umfrage ermöglicht es, die Produzentensicht in die Untersuchungen mit einzubeziehen und einen Überblick über den technischen Ist-Stand sowie die strategischen Planungen bezüglich der Langzeitarchivierung in den jeweiligen Häusern zu erhalten.

Die Umfrage erhebt dabei nicht den Anspruch, statistisch-repräsentative Ergebnisse vorzulegen. Die Methodik der schriftlichen Befragung lässt dies bei der relativ kleinen und hinsichtlich der zu erhebenden Fakten sehr heterogen strukturierten Grundgesamtheit nicht zu. Gleichzeitig widerspricht der geringe Rücklauf der Fragebogen den Vorgaben einer repräsentativen Stichprobe.

Datenbasis und Methodik der Umfrage

Insgesamt wurden die 100 größten wissenschaftlichen Verlage sowie sämtliche öffentlichen Hochschulen Deutschlands zu der Befragung eingeladen. Aufgrund des geringen Rücklaufs der Fragebogen wurde aus dieser Grundgesamtheit eine weitere qualifizierte Auswahl erstellt, die insgesamt etwa 75 Institutionen – Verlage und Hochschulen – umfasst. Der Rücklauf der Fragebogen aus dieser zweiten Aussendung lag zum Auswertungsstichtag bei etwa 40% der ausgesendeten Bögen. Auch wenn diese Menge nicht hinreicht, um statistisch präzise quantifizierbare Aussagen zu liefern, so zeigen die Ergebnisse doch deutliche Tendenzen.

Die insgesamt 50 Fragen gliedern sich in folgende Bereiche:

- Allgemeine statistische Fragen (Produzentenkategorie, Produktionsumfang)
- Stellenwert der Langzeitarchivierung
- Datenformate I: Textdaten und Grafiken
- Datenformate II: Multimediale Elemente
- Datenformate III: Dynamische Elemente

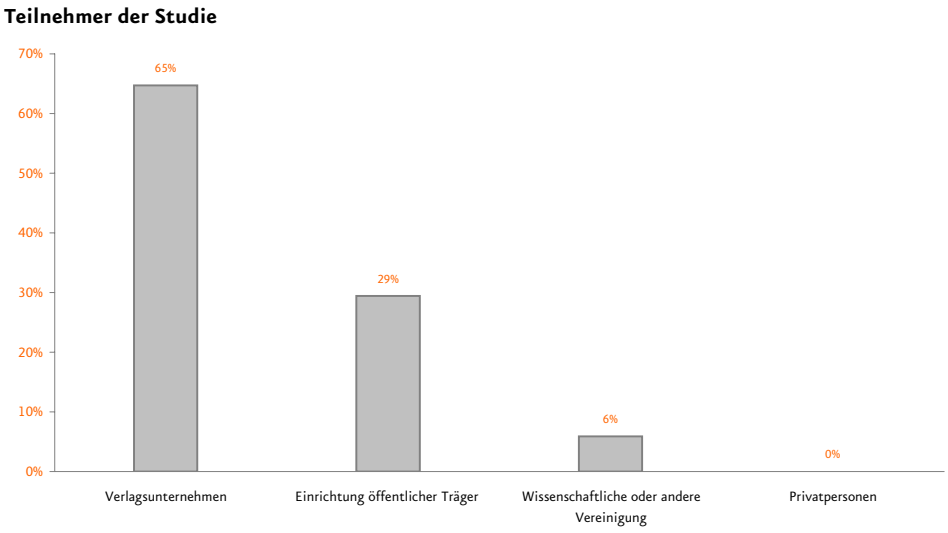
- Bereitstellungsform der Inhalte
- Bibliographische und organisatorische Daten (Meta-Daten)
- Dokument-Identifizierung und Linking
- Digitale Signaturen
- Bereitschaft zur Umsetzung von Anforderungen für die Langzeitarchivierung

Die folgende Auswertung der Ergebnisse folgt dieser Gliederung. Der Umfrage-Bogen ist im Anhang abgedruckt.

Allgemeine statistische Aussagen

Teilnehmer der Studie

Die ausgewerteten Fragebogen kamen zu ca. 65% von Verlagen, die damit die deutlich größte Fraktion bilden. Etwa 30% der Fragebogen wurden von Einrichtungen öffentlicher Träger ausgefüllt (in diese Kategorie fallen auch Einrichtungen öffentlicher Träger mit Verlagscharakter), der Rest von wissenschaftlichen Vereinigungen, die ohne feste Bindung an einen öffentlichen Träger existieren.



Da bei der Auswertung der Fragen kaum signifikante Unterschiede zwischen diesen Produzentenkategorien erkennbar waren, beziehen sich die nachfolgenden Ausführungen, wenn nicht explizit anders vermerkt, immer ungewichtet auf die Gesamtheit dieser drei Kategorien.

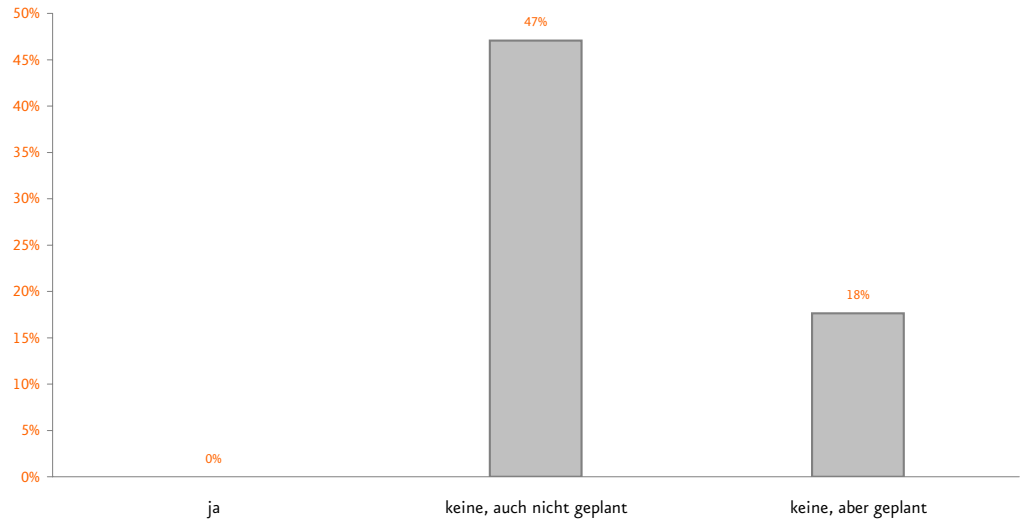
Produktionsvolumen und Aufbau der E-Journals

Die Zahl der publizierten E-Journals der an der Studie beteiligten Institutionen beläuft sich auf über 300 Titel. Mit Titel ist hierbei jeweils ein eigenständiges elektronisches Publikationsorgan gemeint (einschließlich der Titel, die parallel in Printform erscheinen). Nicht gemeint ist die Zahl der jährlich erscheinenden Nummern. In diesen Titeln werden jährlich insgesamt ca. 37.000 E-Journal-Artikel veröffentlicht, was einem Durchschnitt von etwa 90 Artikeln pro Publikationsorgan und Jahr entspricht. Hinsichtlich der Menge der Titel pro Produzent sowie der Anzahl der Artikel pro Titel und Jahr unterscheidet sich die Gruppe der Verlage signifikant von den übrigen E-Journal-Produzenten: Über 90% der erfassten Titel werden von Verlagen publiziert.

Über 60% der befragten Produzenten publizieren ihre Titel kostenpflichtig – das entspricht im Wesentlichen der Produzentenkategorie der Verlage, die mit ihrem Online-Angebot wirtschaftliche Ziele verfolgen.

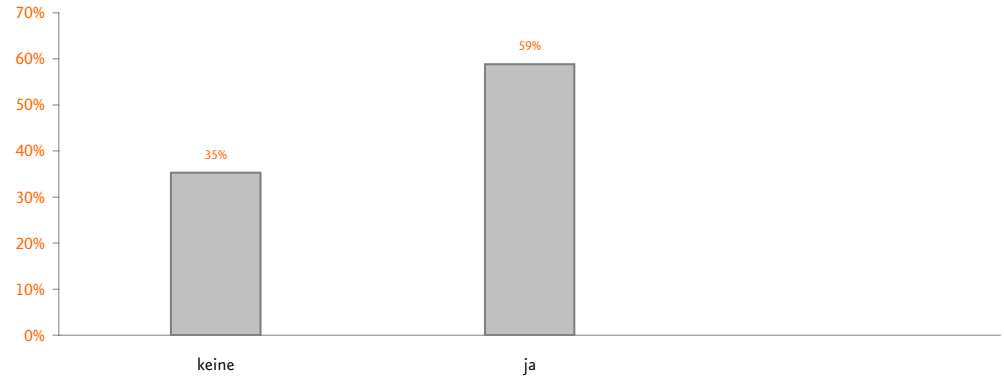
Von den E-Journals, die ein Supplement zu einer gedruckten Zeitschrift darstellen, enthielt zum Zeitpunkt der Umfrage keines Artikel, die ausschließlich online erhältlich sind, also in der gedruckten Fassung nicht vorkommen. Knapp 20% der Befragten gaben jedoch an, dies zu planen. Nur die Hälfte der Befragten war sich hingegen sicher, auch zukünftig keine Artikel online stellen zu wollen, die nicht auch Bestandteil des Printproduktes sind. Dieser Aspekt ist in sofern von besonderer Bedeutung, als dass es bei Übereinstimmung von Print- und Online-Version einer Zeitschrift legitim sein könnte, ausschließlich eine Version zu archivieren, um dem Anspruch der Langzeitverfügbarkeit der Inhalte gerecht zu werden.

Artikel, die ausschließlich online publiziert werden



Es wird noch die Frage zu stellen sein, welche Teile eines E-Journals archiviert werden sollen. In diesem Zusammenhang ist die inhaltliche Zusammensetzung der E-Journals von besonderer Bedeutung. Insgesamt zwei Drittel der Befragten gaben an, neben den eigentlichen Artikeln noch weitere Inhalte in den E-Journals zu publizieren. Fast durchgängig wurden hier die Editorials genannt, daneben (in absteigender Häufigkeit) Rezensionen, Ankündigungen, Kongress- und Konferenzabstracts, Leserbriefe, Diskussionsbeiträge sowie vereinzelt Personalia, Miszellen, Veranstaltungskalender und Produktinformationen.

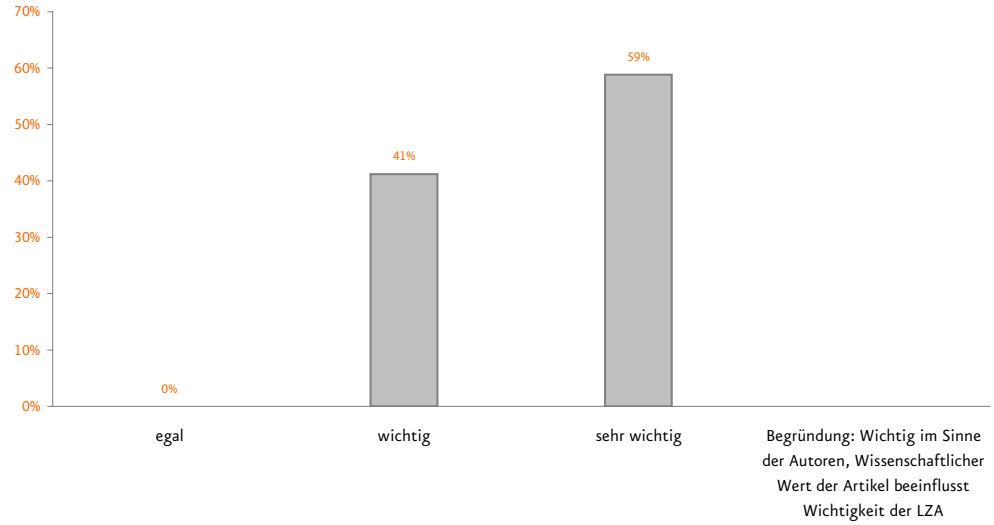
Zusätzliche Inhalte neben den Artikeln



Stellenwert der Langzeitarchivierung

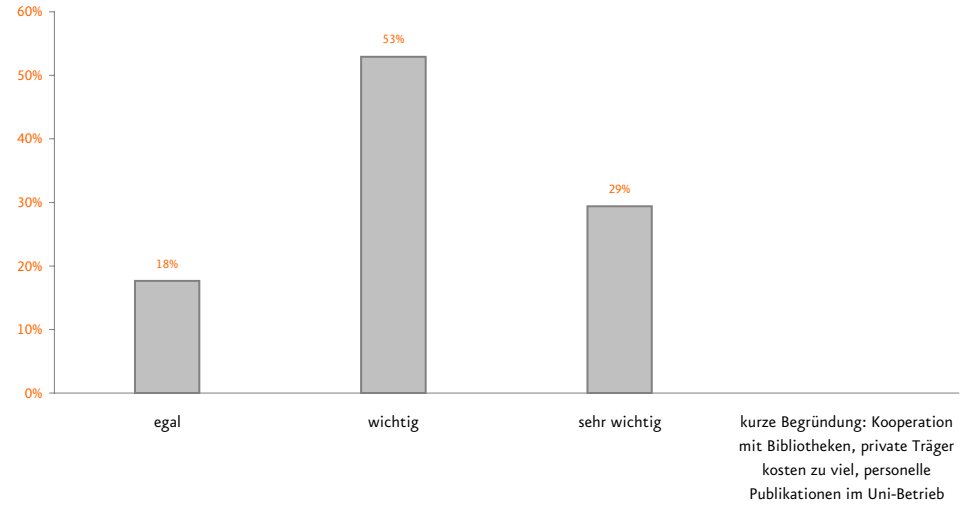
Die Fragestellungen dieses Komplexes behandeln die vom Produzenten empfundene Bedeutung der Langzeitarchivierung und ermitteln gleichzeitig das Vorhandensein eigener Langzeitarchivierungssyssteme. Damit soll zunächst festgestellt werden, wie weit ein Problembewusstsein für den Bereich »Langzeitarchivierung« besteht. Daneben hat dieser Fragenkomplex auch das Ziel, einen ersten Soll-Ist-Vergleich zwischen Wunsch und Wirklichkeit vornehmen zu können.

Wie wichtig ist die LZA der E-Journals?



Ausnahmslos alle Befragten beantworteten die Frage nach der Langzeitarchivierung der E-Journals mit »wichtig« bis »sehr wichtig«. Gleichzeitig verfügen aber nur etwa 35% über ein eigenes Archivierungssystem. Eine Tendenz zur Zunahme in diesem Bereich durch geplante Neuinvestitionen ist nicht erkennbar. Bei genauerer Betrachtung der beiden großen Produzentenkategorien (Verlage/öffentliche Träger) wird deutlich, dass die Mehrzahl der Archivierungssysteme von den öffentlichen Trägern unterhalten werden: zwei Drittel aller E-Journal-Produzenten im Umfeld von Universitäten und anderen öffentlichen Forschungseinrichtungen können auf ein eigenes Archivierungssystem zurückgreifen, während nur knapp 20% der Verlage – und zwar nur die größten – über ein eigenes System verfügen.

Wie wichtig ist die LZA durch öffentliche Träger?



Diese Zahlen legen den Schluss nahe, dass im universitären Umfeld zur Langzeitarchivierung von Daten häufig auf eine bestehende technische Infrastruktur, auch die anderer wissenschaftlicher Einrichtungen (Rechenzentren etc.) zurückgegriffen werden kann, während im Verlagsbereich diese Möglichkeiten in der Regel nicht bestehen. Freilich stellt sich hier die Frage, ob die Archivierungsmöglichkeiten der Forschungseinrichtungen über das Anlegen bloßer Sicherheitsbackups hinausgehen. In der Regel wird das nicht der Fall sein, so dass hier die an eine sinnvolle LZA zu stellenden Ansprüche bei weitem nicht erfüllt sind.

Die Verlage als die zweite und größere Gruppe von E-Journal-Produzenten schätzen dabei aber den Stellenwert der Langzeitarchivierung keineswegs als geringer ein. Hier steht aber der Wunsch nach Archivierung der Daten – bzw. die Erkenntnis der Notwendigkeit – im krassen Missverhältnis zur vorhandenen Infrastruktur.

Daraus erklärt sich auch der Ruf nach einer durch öffentliche Träger finanzierten Stelle, die die Langzeitarchivierung sicherstellt:

Über 80% der Befragten halten die Gewährleistung der Langzeitarchivierung durch Einrichtungen öffentlicher Träger für »wichtig« bis »sehr wichtig«. Lediglich kleine Produzenten (<8 Titel) stehen dieser Frage gleichgültig gegenüber.

Fazit

Das Problembewusstsein bei den E-Journal-Produzenten für die Langzeitarchivierung ist sehr stark. Gleichzeitig verfügt nur ein Bruchteil der Verlage über die (technischen) Möglichkeiten hierzu, während Herausgeber von E-Journals an Universitäten häufig auf eine bestehende Infrastruktur zurückgreifen können. Hier wird der Ruf nach der Gewährleistung der Langzeitarchivierung durch Einrichtungen öffentlicher Träger laut. Die Auswertung weiterer Fragenkomplexe weiter unten wird zeigen, dass im Zuge dessen eine große Bereitschaft zur technischen Unterstützung einer solchen Stelle vorhanden ist, etwa durch die Einhaltung neuer, vereinheitlichter Standards zur Datenarchivierung.

Viele der Fragenkomplexe der Befragung ergaben ein auffällig eindeutiges Bild, das sich häufig mit der Einschätzung und Branchenkenntnis der Bearbeiter dieser Studie deckt. Dennoch muss trotz aller Sorgfalt bei der Erstellung des Fragebogens und der Vermeidung von Suggestivfragen davon ausgegangen werden, dass bei einzelnen Fragen, etwa der nach der generellen Bedeutung der Langzeitarchivierung oder auch der von multimedialen Elementen, die Teilnehmer der Studie so geantwortet haben, wie sie glaubten, dass es von ihnen erwartet wird. Dieser in der Literatur vielfach beschriebene Effekt lässt sich nicht ganz verhindern, so dass er umgekehrt bei der Bewertung der Ergebnisse als relativierendes Element Berücksichtigung finden muss.

So ist es naheliegend, bei einer Studie, die die Langzeitarchivierung zum Thema hat, die Frage nach der Wichtigkeit ebendieses Themas mit »hoch« zu beantworten – auch wenn es in der eigenen Arbeit vielleicht derzeit überhaupt keine Rolle spielt. Um diesem Effekt entgegenzuwirken, wurden in der Umfrage Fragen allgemeinen Inhalts mit solchen nach der konkreten Arbeitssituation ebenso gemischt wie offene und geschlossene Antwortformen. Auch das Abfragen von Begründungen für Antworten wurde als bewährtes Mittel verwendet, um (über die intensivere Beschäftigung des Befragten mit der Fragestellung) weniger prädeterminierte Antworten zu erhalten.

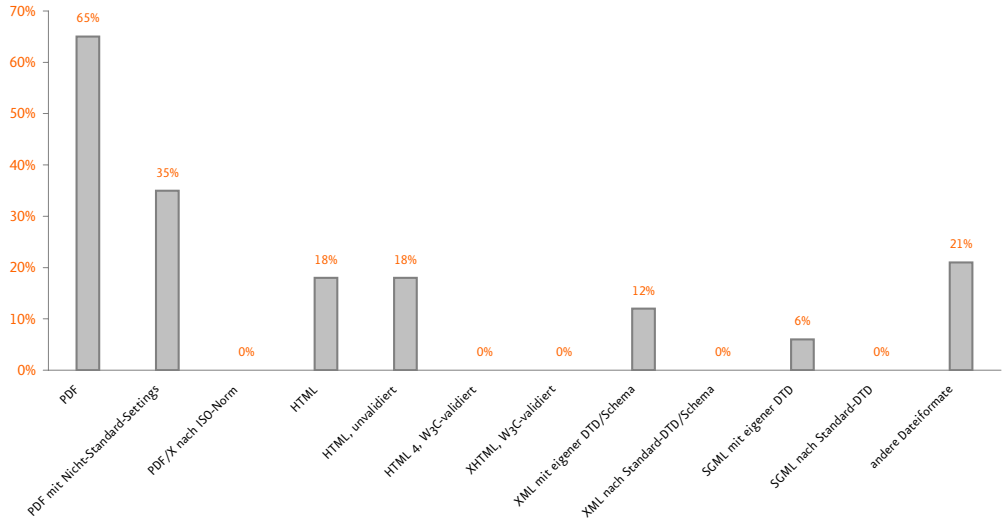
Datenformate I: Textdaten und Grafiken

Nachdem die ersten Fragenkomplexe eher allgemeine Fragen zu Vorhandensein und Menge der publizierten E-Journals sowie der Grundhaltung hinsichtlich der Langzeitarchivierung behandelt haben, wendet sich der Blick im Folgenden auf die E-Journals selbst. Hier steht zunächst eine Bestandsaufnahme der Ist-Stituation in technischer Hinsicht an. Im Kontext der Langzeitarchivierbarkeit befinden sich weniger die Inhalte als vielmehr deren technische Vorhaltung im Mittelpunkt des Interesses.

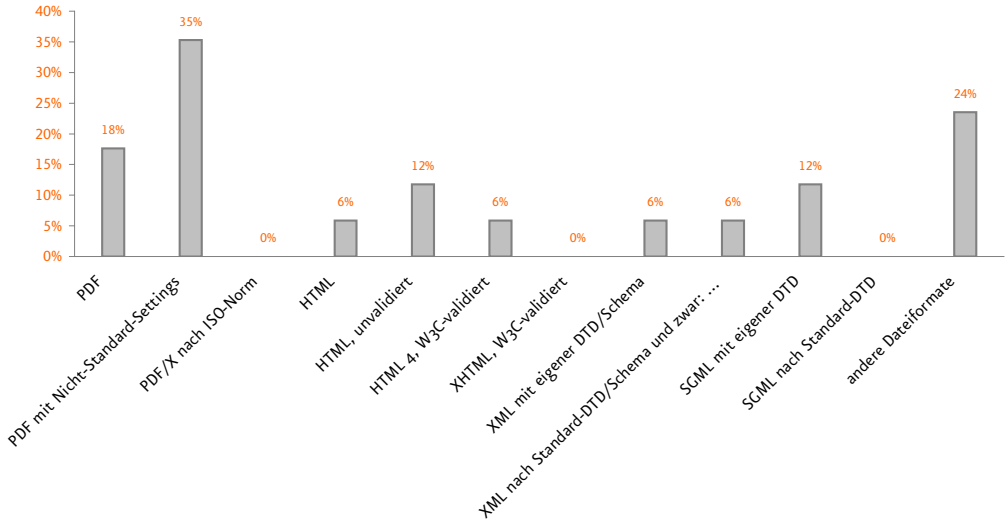
Die These, E-Journals bestünden heute zum überwiegenden Teil aus statischen Text- und Grafikelementen, wurde durch die Umfrage eindeutig belegt. Diese Elemente sollen zunächst untersucht werden. Weitere Fragenkomplexe werden sich mit multimedialen und dynamischen Elementen auseinander setzen.

Die detaillierte Betrachtung betrifft sowohl Fragen zum Publikationsformat und Dokumentenschutz als auch solche zum Format der Datenerstellung und internen Vorhaltung. Diese Unterscheidung ist in sofern wichtig, als dass das veröffentlichte Format nicht zwangsläufig auch das zu archivierende ist.

In welchen Formaten publizieren Sie E-Journals?



Interner Formatgebrauch



In diesem Kontext ist noch eine weitere Unterscheidung vorzunehmen: Die nachfolgenden Untersuchungen beziehen sich nicht auf die Plattform, von der aus die Daten der E-Journals heruntergeladen oder innerhalb derer sie durchsucht und gelesen werden können, sondern auf die E-Journals selbst. Die präzise Trennung zwischen E-Journal und seinem Publikationskanal ist schwierig aber notwendig, denn sie impliziert eine Aussage zu den als archivierungswürdig betrachteten Datenteilen. Die Aufgabenstellung dieser Studie legt nahe, dass die Inhalte und nicht deren Bereitstellungsplattformen als die als archivierungswürdig betrachteten Datensammlungen angenommen werden sollen.

PDF

Die Bestandsaufnahme der heute verwendeten Publikationsformate ergibt ein sehr eindeutiges Bild: Sämtliche der an der Umfrage teilnehmenden Produzenten von E-Journals verwenden hierfür PDF. Für etwa 70% der Befragten ist PDF sogar das ausschließlich eingesetzte Format.

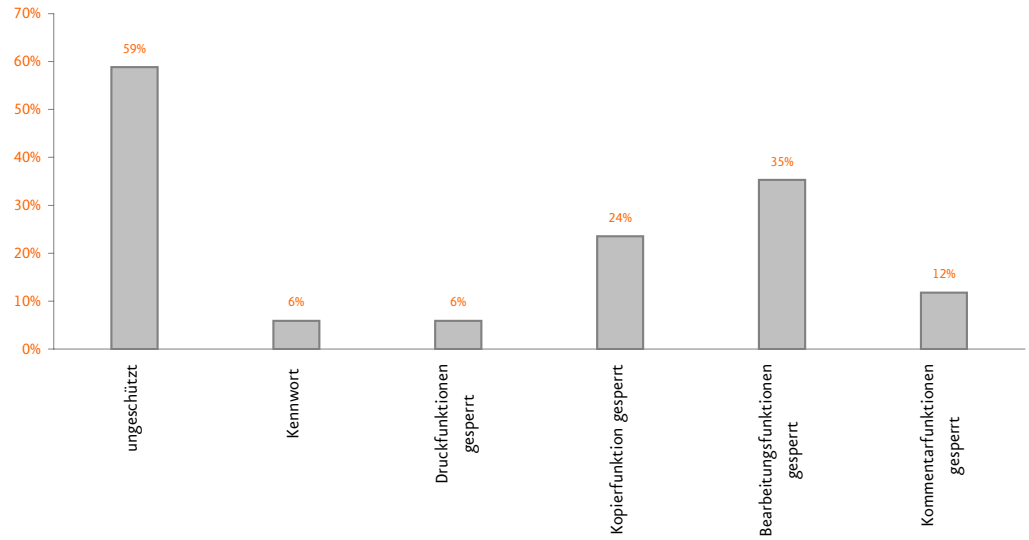
Mit »Publikationsformat« ist hierbei das Datenformat gemeint, in dem die E-Journals (oder einzelne Artikel daraus) nach Abschluss der redaktionellen und herstellerischen Arbeit zur Distribution vorliegen.

PDF verfügt über eine Reihe leistungsfähiger Schutzmechanismen, über die verschiedene Funktionen (Drucken, Kopieren, Bearbeiten) innerhalb der Anwendungssoftware gesperrt werden können. Solche Mechanismen können der Archivierbarkeit in anderen Systemen im Wege stehen. Etwa 60% der Befragten gaben an, die von ihnen erzeugten PDF-Dateien

ohne Einschränkung der Nutzerrechte und ohne Passwort, also gänzlich ungeschützt zu publizieren. Die übrigen Teilnehmer der Studie gaben fast alle an, die Bearbeitbarkeit des PDFs zu sperren. Überwiegend wird auch die Kopierfunktion gesperrt, während die Druckfunktion und die Möglichkeit, eigene Kommentare anzulegen, nur in Einzelfällen unterbunden wurde. Auch die Möglichkeit, das PDF mit einem Kennwortschutz zu versehen, kommt nur vereinzelt zur Anwendung. Etwa ein Viertel der Befragten gab an, darüber hinaus den Einsatz anderer technischer Maßnahmen zum Schutz der Dokumente vor unerwünschter Nutzung oder Weiterverbreitung zu planen oder einzusetzen. Explizit genannt wurden digitale Wasserzeichen und die Begrenzung der Anzahl der Ausdrucke des Dokuments.

Das Thema »PDF als Archivformat«, dessen technische Aspekte bereits in Teil 1 ausführlich besprochen wurden, muss also eines der zentralsten Themen im Bereich der Langzeitarchivierung bleiben.

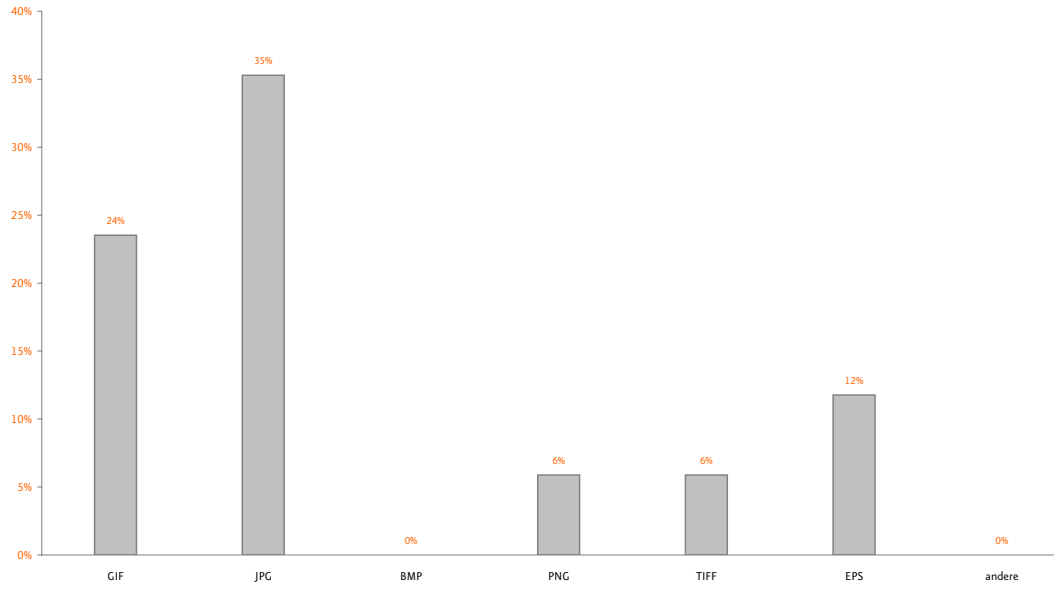
Wie schützen Sie Ihre PDF-Publikationen?



HTML

Neben PDF kommt als Publikationsformat fast ausschließlich HTML zum Einsatz. HTML ist (wie in Teil 1 erörtert) ein Format, das nur mit Einschränkungen den Ansprüchen der Langzeitarchivierung gerecht wird. Wie die Umfrageergebnisse nahelegen, kommt ein weiterer problematischer Aspekt hinzu: es wird in aller Regel nicht in »nacktem« HTML publiziert, sondern der Funktionsumfang der Web-Browser wird über Skripte erweitert. Auch die Möglichkeit, Text über Stylesheets zu generieren (Trennzeichen, Überschriften o.ä.), muss in diesem Kontext bedacht werden, da die entsprechenden Zeichen nicht Bestandteil der Datei im engeren Sinne sind.

Dateiformate für Grafiken beim Einsatz von HTML oder XML zur Online-Publikation



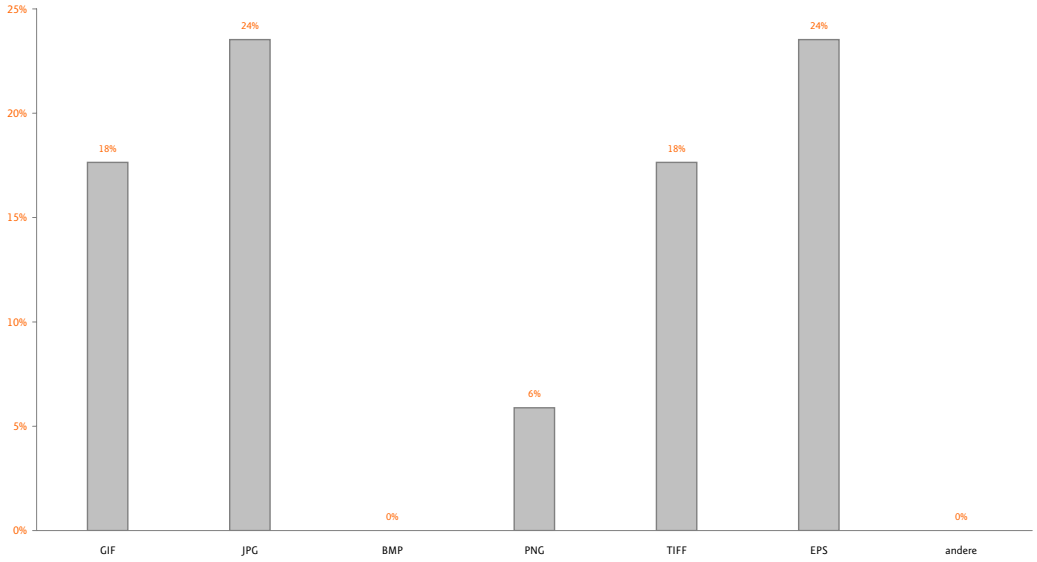
Vor allem jedoch bieten die immer stärker eingesetzten Content Management Systeme und andere Textdatenbanken die Möglichkeit, HTML dynamisch zu generieren. Diese dynamischen Teile auf der Nutzerseite, d.h. im fertig generierten Publikationsformat zu archivieren, wirft große Probleme auf (siehe dazu weiter unten: »Bereitstellungsform der Inhalte«).

All diese Aspekte machen es erforderlich, die HTML-basierten E-Journals genauer zu analysieren.

Zwei Drittel der Befragten, die HTML als Publikationsformat einsetzen, gaben an, Textteile per Stylesheet zu generieren. Hier also ist hinsichtlich der Archivierbarkeit der (Quell-) HTML-Dateien Vorsicht angeraten. Einheitlich hingegen stellen sich die jeweiligen E-Journals bezüglich der internen Datenlage dar: Wenn E-Journals auch andere Publikationsobjekte als Artikel enthalten, so liegen diese überwiegend in der gleichen Form wie die Artikel vor – das gilt sowohl für PDF-basierte als auch für HTML-basierte E-Journals.

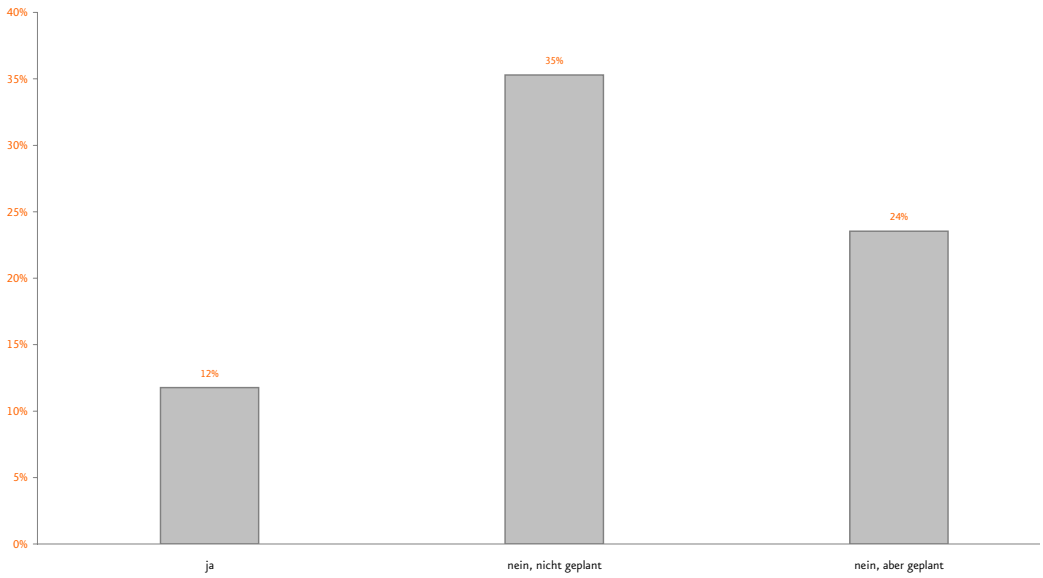
Als »Mischform« kommt darüber hinaus noch der Fall vor, dass eine PDF-basierte Zeitschrift das Editorial und ähnliche Journalteile in HTML vorhält und nur die einzelnen Artikel als PDF zum Download vorhält. Relativ einheitlich stellt sich auch die Situation bei den publizierten Grafiken dar. Während intern beim Produzenten die unterschiedlichsten Grafikformate nebeneinander zum Einsatz kommen (neben GIF und JPG vor allem TIFF und EPS), so kommen zur Online-Publikation fast ausschließlich GIF und JPG zum Einsatz. Diese Grafiken werden derzeit von einem Drittel der Produzenten in mehreren Auflösungen publiziert (z.B. als Vorschaugrafiken mit einem Link auf die hochauflösende Version), die restlichen zwei Drittel gaben an, den Einsatz mehrerer Auflösungen zu planen.

Dateiformate für Grafiken beim Einsatz von HTML/XML - intern



Lediglich für die Gruppe der ausschließlich in PDF publizierenden Produzenten kommt ein Veröffentlichen von Grafiken in verschiedenen Auflösungen nicht in Frage.

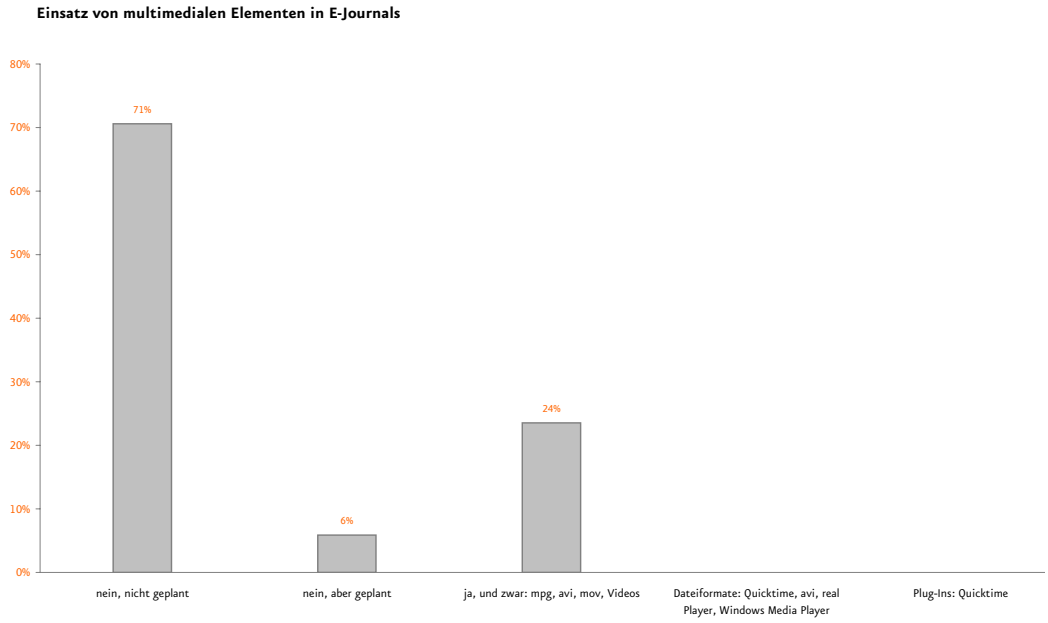
Nutzung von Grafiken in unterschiedlicher Aufllösung



Diese Ergebnisse überraschen nicht, folgen sie doch im Wesentlichen den technischen Gegebenheiten bei Verarbeitung und Publikation in den jeweiligen Formaten. Dennoch ist die Bestätigung dieser Annahmen ein wichtiges Ergebnis, um darausf Archivierungsmodelle zu entwickeln.

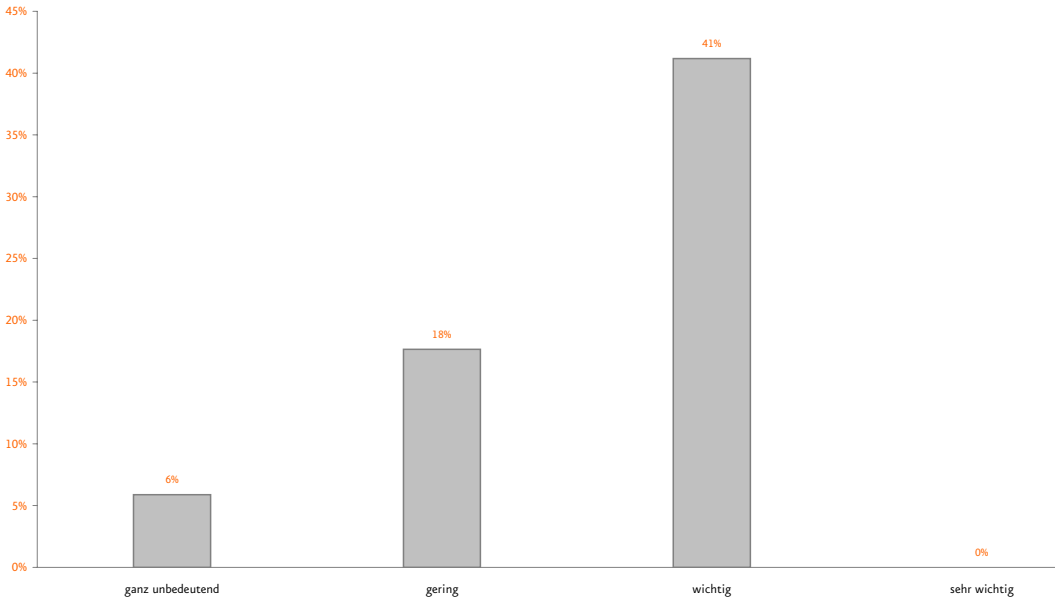
Datenformate II: Multimediale Elemente

Ein zentraler Gesichtspunkt elektronischen Publizierens ist der multimediale Aspekt. Anders als bei gedruckten Titeln können neben Texten und zweidimensionalen, statischen Abbildungen auch andere Elemente, z.B. dreh- und zoombare 3-D-Modelle, 360-Grad-Ansichten, Animationen, Videos, Audiodaten etc. in die Publikation eingebunden werden und einen deutlichen Mehrwert darstellen. Für die Fragen der Archivierbarkeit ist es daher wichtig, eine Bestandsaufnahme auch dieser Teile vorzunehmen sowie die Bedeutung der multimedialen Elemente für die elektronische Zeitschrift zu kennen.

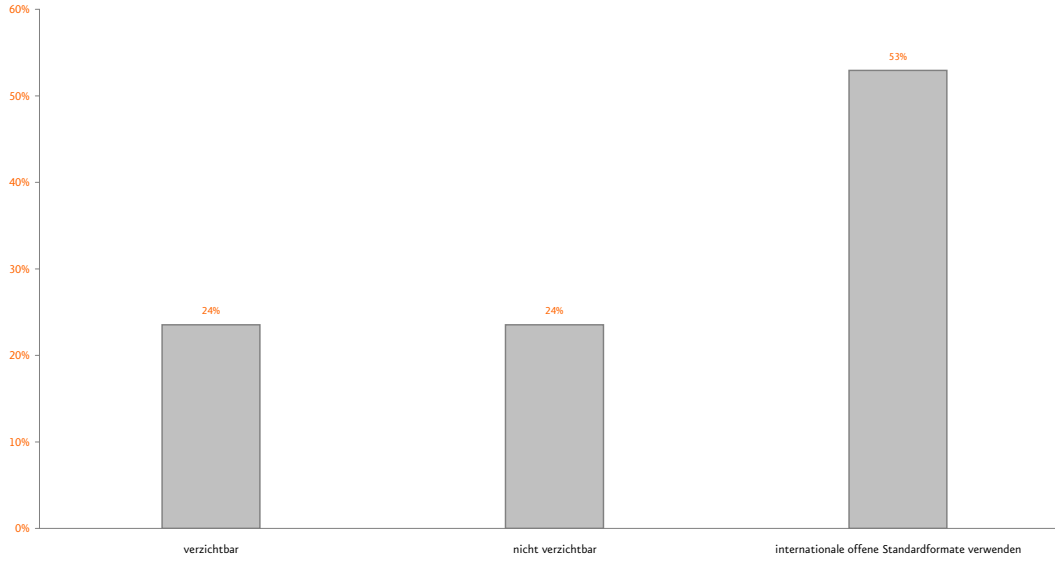


Von den Teilnehmern der Umfrage gab ein Viertel an, dass multimediale Elemente in ihren E-Journals eingebunden seien. Als einzige Medienform kommt hierbei das Video zum Einsatz, in der Regel als .mpg oder .avi-Datei. Dabei sieht nur Hälfte dieser Produzentengruppe multimediale Elemente als wesentlichen Bestandteil des E-Journals an, die andere Hälfte hält diese nur für Beiwerk, auf das ggf. bei der Langzeitarchivierung auch verzichtet werden könnte.

Zukünftige Bedeutung von MM-Elementen



Sind MM-Standards & dynamische Elemente verzichtbar, da Sie Probleme bei der LZA verursachen?

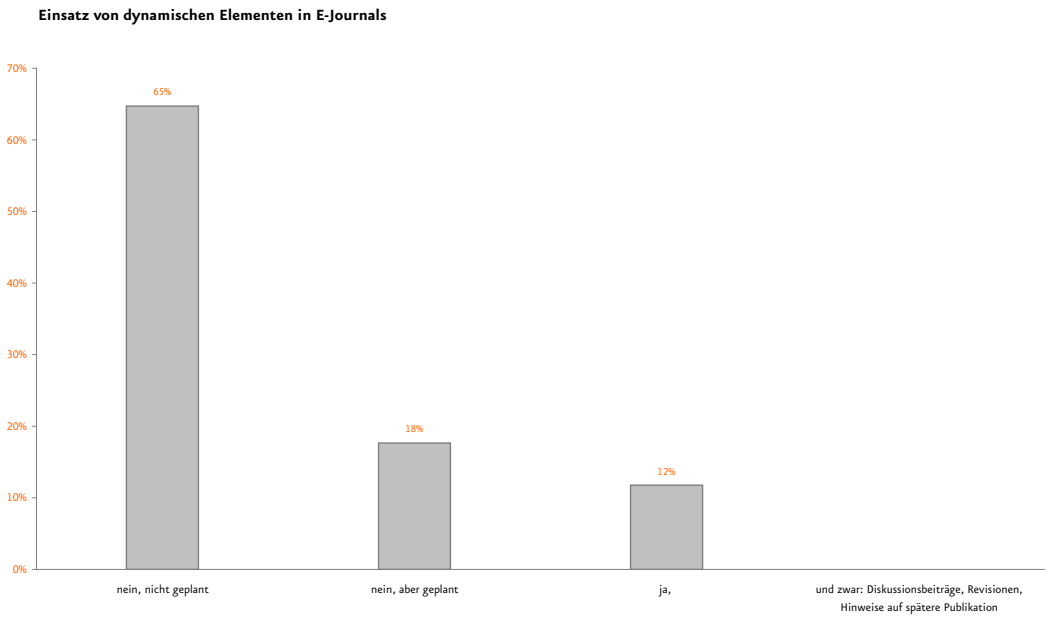


Nur etwa 12% der Produzenten setzen also multimediale Elemente (und zwar ausschließlich Videos) ein, die sie selbst auch für archivierungswürdig halten. Dem entgegen steht ein anderes Ergebnis der Umfrage, das nachdenklich macht: Während über 70% der Befragten angaben, auch zukünftig keine multimedialen Elemente in ihre E-Journals einbauen zu wollen, schätzt nur ein Teilnehmer der Studie deren zukünftige Bedeutung als unbedeutend ein. 18% der Befragten schätzen die Bedeutung als gering, der Rest misst den multimedialen Elementen zukünftig eine wichtige Rolle bei. Die Diskrepanz zwischen der überwiegend ablehnenden Haltung einerseits und der Vermutung einer steigenden Bedeutung andererseits lässt vermuten, dass die Bedeutung in anderen Bereichen der Branche höher eingeschätzt wird als im eigenen.

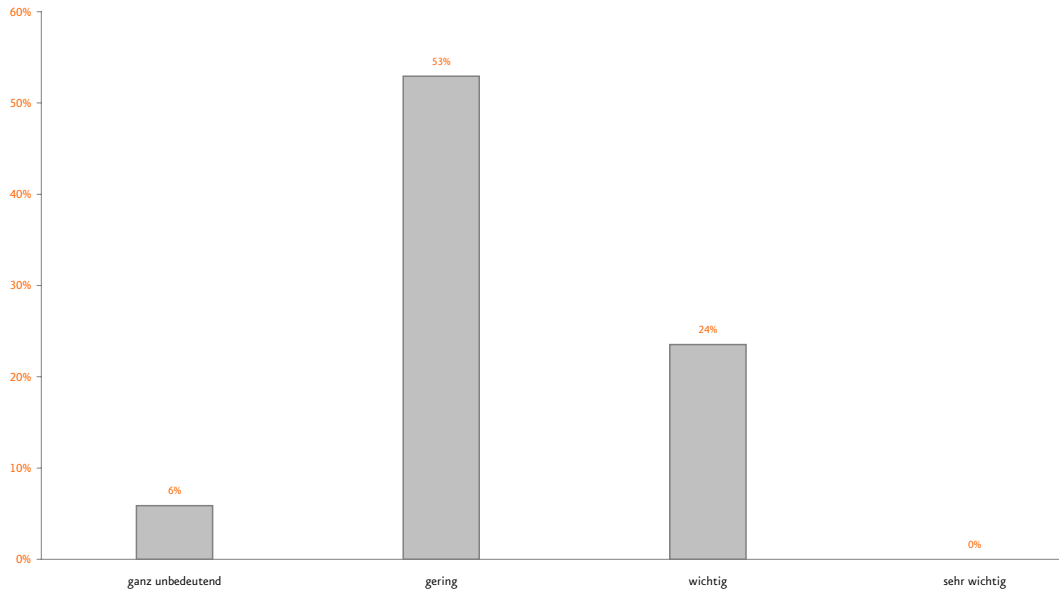
Datenformate III: Dynamische Elemente

Während schon beim Fragenkomplex der multimedialen Elemente eine deutliche Tendenz zu erkennen war, die neuen Technologien eher zögerlich einzusetzen, so ist dies bei den dynamischen Elementen noch klarer. Mit »dynamischen Elementen« sind hier Elemente gemeint, die nicht mit der Publikation abgeschlossen sind, sondern sich danach noch verändern, so z.B. von Lesern erstellbare Diskussionsbeiträge, die einem Dokument zugeordnet sind oder Diagramme, die aus regelmäßig aktualisierten Daten immer neu erstellt werden (Börsenkurse etc.). Da es für diese Elemente keinen Status der »Fertigstellung«, sozusagen des »digitalen Imprimatur« gibt, sind sie besonders schwierig zu archivieren.

Bei den gleichen Fragestellungen wie zum Komplex der multimedialen Elemente ergab sich folgendes Bild: Nur zwei der Befragten gaben an, bereits dynamische Elemente in ihren E-Journals einzusetzen, von denen wiederum nur einer diese als wesentlichen Bestandteil des E-Journals einstufte. Konkret enthalten diese Hinweise auf spätere Publikationen und Diskussionsbeiträge. Als Format kommt hier HTML bzw. wiederum PDF zum Einsatz.



Zukünftige Bedeutung von dynamischen Elementen



Die zukünftige Bedeutung von dynamischen Elementen wird von über 50% als »gering« eingestuft, nur knapp ein Viertel der Befragten messen den dynamischen Elementen zukünftig eine wichtige Rolle bei.

Von Seiten der E-Journal-Produzenten selbst werden also offensichtlich die technischen Möglichkeiten der elektronischen Publikationswege nicht nur nicht ausgeschöpft, sondern ihrer Bedeutung nach auch relativ gering bewertet. Der Charakter der E-Journals wird also weniger durch die Multimedialität und Interaktivität bestimmt. Es handelt sich vielmehr vor allem um einen neuen Vertriebskanal für Informationen, die in recht traditioneller Weise den Rezipienten dargeboten werden.

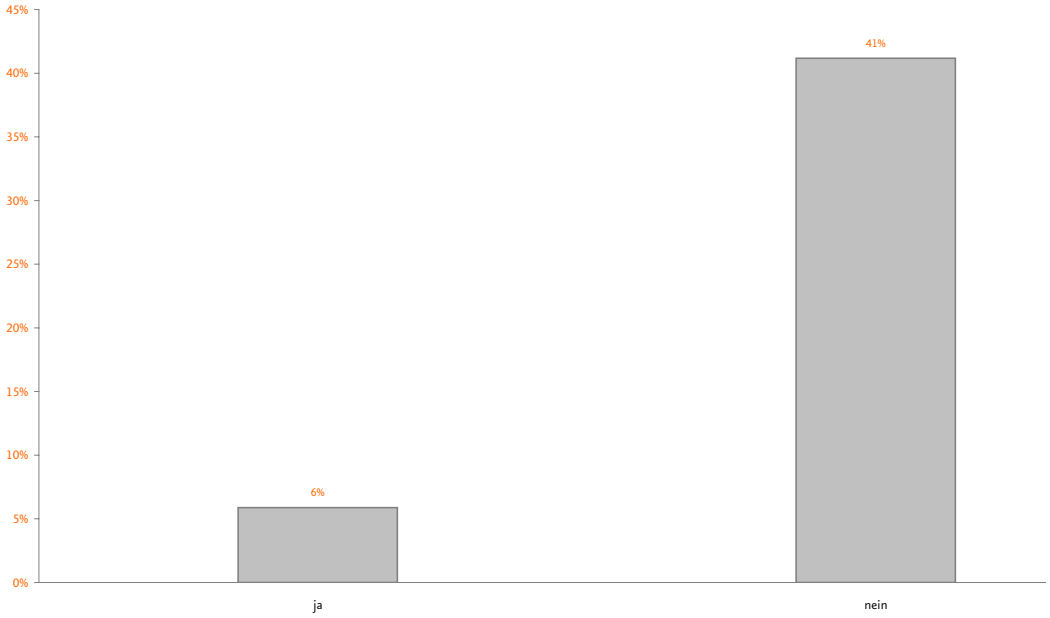
Bereitstellungsform der Inhalte

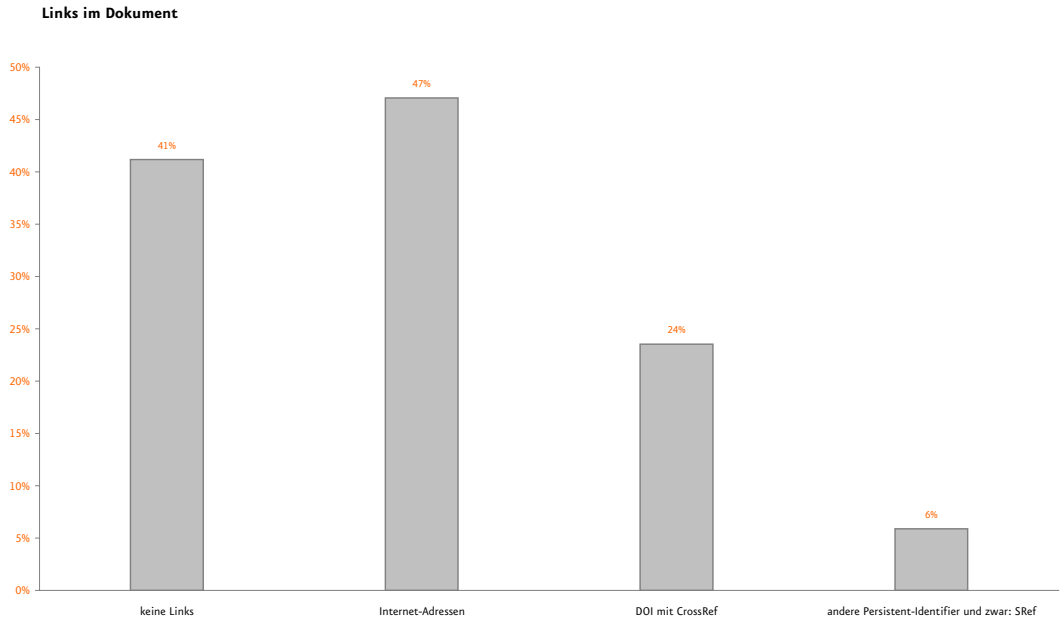
Auch hinsichtlich der Bereitstellungsform der Inhalte überwiegen die »klassischen« Methoden der statischen Präsentation: Über 70% der Befragten gaben an, die Inhalte als statische Dateien zur Nutzung zur Verfügung zu stellen.

Lediglich ein Teilnehmer der Untersuchung gab an, seine Dokumente dynamisch zu generieren, d.h. erst bei der Abfrage durch den Endnutzer aus einer Datenquelle, etwa einer Datenbank, zu erzeugen. In diesem einen Fall variiert die Ausgabe auch je nach Nutzer, d.h. es findet eine Personalisierung der Ausgabe statt. In allen anderen Fällen, auch in denen, in denen Teile der Publikation dynamisch generiert werden, hat der Nutzer keinen Einfluss auf das Ausgabeergebnis.

Bereitstellungsform der Inhalte

Einfluß des Nutzers auf das Ausgabeergebnis





Von der Gruppe der Produzenten, die die Inhalte als statisch Dateien publizieren, gab keiner an, eine Umstellung auf eine generierte Ausgabe zu planen. Auf absehbare Zeit wird also das Gros der elektronischen Zeitschriften in statischen Einzeldokumenten abgelegt und verwaltet sein.

Diese Aussage ist für die weiteren Überlegungen hinsichtlich der Langzeitarchivierbarkeit von großer Bedeutung: Stellt doch die Archivierung und Verfügbarhaltung von dynamisch generierten, gar personalisierten Inhalten die archivierende Institution vor ganz andere Herausforderungen als die Vorhaltung von Inhalten in einer bekannten – nämlich der Präsentationsform.

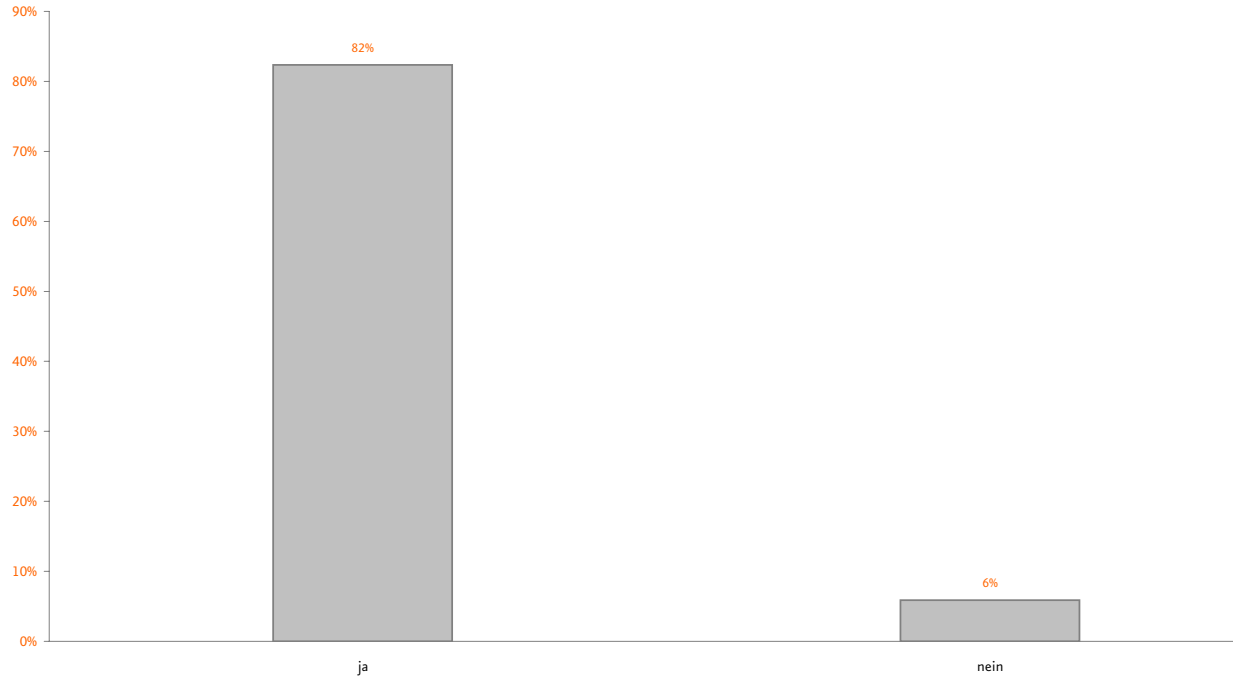
Produzenteninteresse und Langzeitarchivierung

Die Kooperation der Produzenten ist notwendig für eine erfolgreiche Standardisierung des E-Journal-Ingest. Es kann wohl vorausgesetzt werden, dass die Autoren wissenschaftlicher Inhalte ein Interesse an der Langzeitarchivierung ihrer Forschungsergebnisse haben. Zeitschriften, bei denen die Langzeitarchivierung der Artikel ungeklärt ist, dürften einen echten Nachteil im Wettbewerb um die besten Autoren haben.

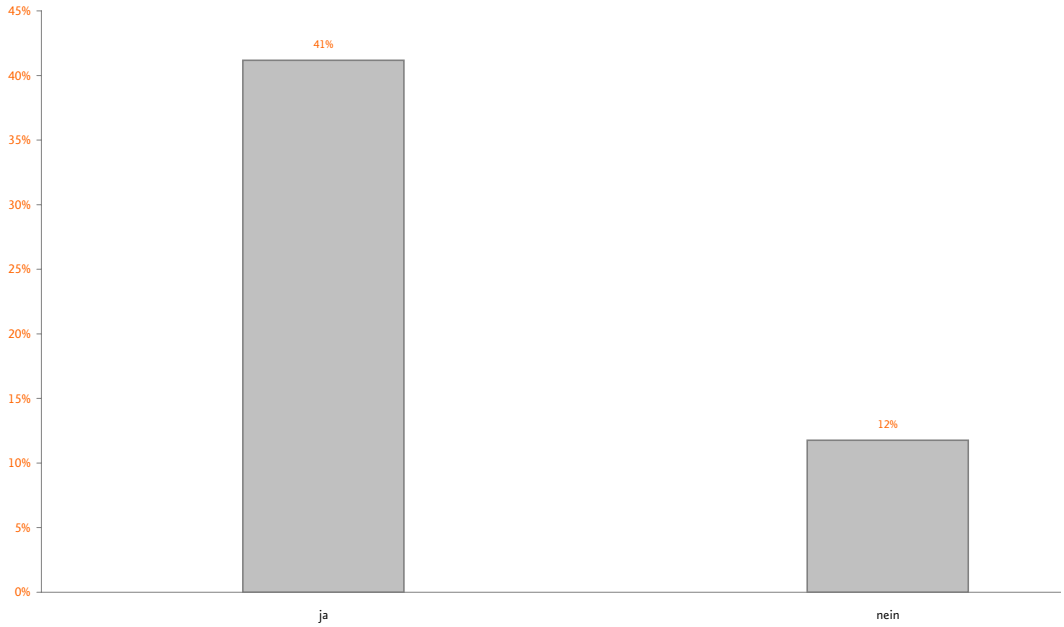
Relevanz der digitalen Langzeitarchivierung

Bei der Publikation in einer Zeitschrift, die parallel in Print- und elektronischer Form erscheint (typisch z.B. für Zeitschriften der wissenschaftlichen Großverlage), können die Autoren das Kriterium der Langzeitverfügbarkeit auch dann als erfüllt ansehen, wenn die digitale Ausgabe der Journals keinen besonderen Maßnahmen zur Langzeitarchivierung unterworfen ist.

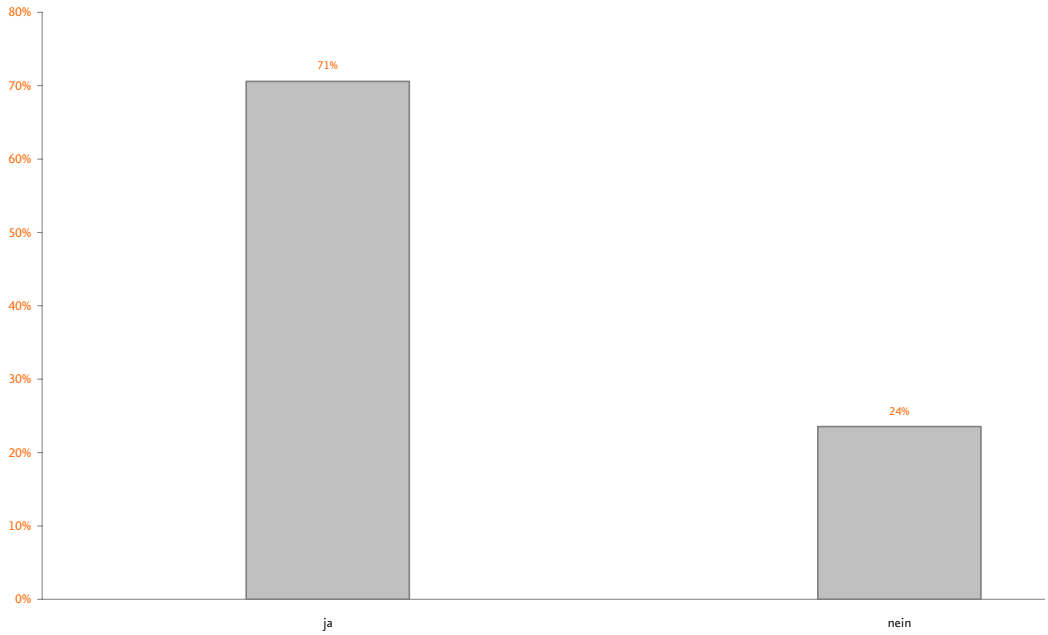
Bereitschaft Standardformate zur Verfügung zu stellen zur Erleichterung der LZA durch öffentlichen Institutionen

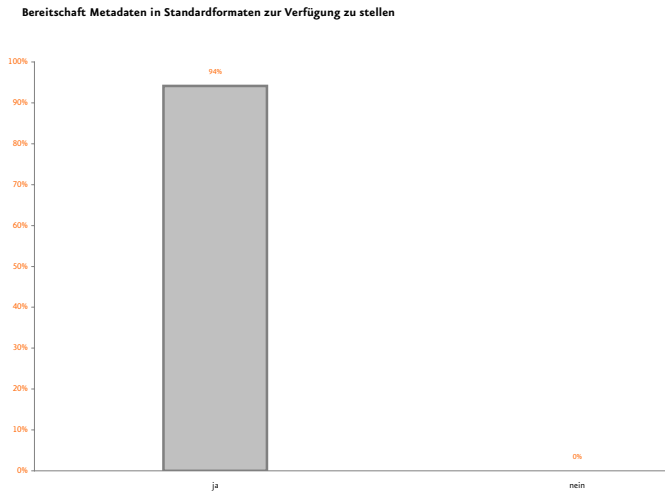


Bereitschaft XML-Daten zur Verfügung zu stellen

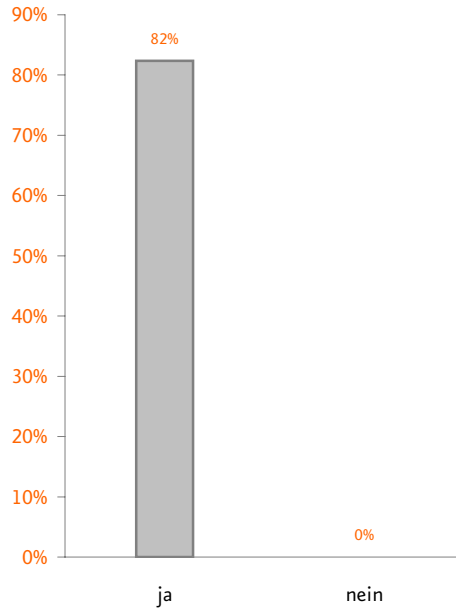


Bereitschaft Daten auf XML umzustellen





Bereitschaft einen Standard für den Transfer zu nutzen



Die Langzeitverfügbarkeit der Inhalte kann durch die traditionelle Archivierung der Printexemplare als grundsätzlich gewährleistet gelten. Eine zusätzliche Langzeitarchivierung der digitalen Version hätte zwar den Vorteil, dass der im Vergleich zur Printversion viel schnellere und komfortablere Zugang zu den Inhalten durch digitale Archivsysteme auch für die Zukunft gewährleistet ist; sie stellt aber keine absolute Notwendigkeit dar.

Zu einem eigenen Thema wird die Langzeitarchivierung digitaler Inhalte erst dann, wenn bei Online-only-Publikation keine traditionelle Archivierung von Printexemplaren durch öffentliche Einrichtungen mehr stattfindet. Die Langzeitverfügbarkeit kann dann alleine durch die Vorhaltung der digitalen Objekte gewährleistet werden.

Wissenschaftliche Großverlage

Hier findet bisher in aller Regel eine parallele Print- und Online-Publikation der Inhalte statt. Ob und in welchem Zeitraum ein Übergang zu Online-only-Publikationen zu erwarten ist, lässt sich derzeit nur schwer abschätzen. Bei den wissenschaftlichen Großverlagen kann bei Online-only-Publikationen auf Eigenverantwortung und/oder auf die Bereitschaft zur Kooperation mit den öffentlichen Archiven gehofft werden. Da Großverlage durch die von ihnen verarbeiteten Informationsmengen ohnehin auf Standardisierung – wenigstens intern – setzen, dürften sich dem Anschluss ihrer Verfahren und Systeme an einen allgemeinen E-Journal-Archivierungsstandard keine prinzipiell unüberwindlichen Hindernisse in den Weg stellen – die Bereitschaft zur Kooperation vorausgesetzt. Es wäre natürlich wünschenswert, wenn auch bei parallelen Print- und Online-Publikationen wegen der oben genannten Vorteile eine Kooperation zur Langzeitarchivierung der digitalen Versionen stattfinden würde. Aufgrund der hohen Zahl der von Großverlagen publizierten Journals könnte so die standardisierte Langzeitarchivierung eines Großteils des digitalen verfügbaren Journal-Materials mit vertretbarem Aufwand gewährleistet werden.

Verlagsunabhängige Publikationsplattformen

Die zweite Produzenten-Gruppe – verlagsunabhängige Publikationsplattformen, die verstärkt auf Online-only-Publishing setzen – setzt zugleich auch verstärkt auf Standards und dürfte einer entsprechenden Kooperation mit den öffentlichen Archiven entsprechend zugeneigt sein. Da sie sich im Wettbewerb mit den Verlagsjournals um die besten Autoren befindet, wäre die Garantie einer zuverlässigen Langzeitverfügbarkeit durch Kooperation mit den öffentlichen Archiven sicher ein nicht zu vernachlässigendes Wettbewerbsargument.

Kleinproduzenten

Bei Verlagen, die nur wenige Journals in digitaler Form selbstständig publizieren, oder bei Produzenten, die E-Journals im Selbstverlag veröffentlichen, können aufgrund des geringen Publikationsvolumens das Interesse und die Möglichkeiten zur Einhaltung von Standards –

die mit einem nicht zu vernachlässigenden Aufwand verbunden ist – gering sein. Gerade bei letzteren, bei denen die Pflege der E-Journals meist ehrenamtlich neben der eigentlichen wissenschaftlichen Forschungs- und Lehrtätigkeit stattfindet, kann oft die einfache, selbstentworfene Publikationslösung die effektivste sein. In welchem Grade hier eine Bereitschaft zu erwarten ist, vorhandene Strukturen so zu modifizieren, dass Daten und Metadaten einem Standard zur Langzeitarchivierung gerecht werden, wird von Fall zu Fall verschieden sein.

Bei dieser Gruppe sind also die größten Schwierigkeiten zu erwarten – sowohl technischer also auch ökonomischer Natur.

Es kann allerdings angenommen werden, dass das grundsätzliche Interesse der Wissenschaftler an der Langzeitarchivierung ihrer Forschungsergebnisse auch bei diesen Produzenten zum Tragen kommt.

Gerade deshalb ist es besonders wichtig, diesen Produzenten die Mitarbeit bei der Langzeitarchivierung ihrer E-Journals zu erleichtern, indem man Tools und Wege zur Erstellung und Übermittlung von SIPs zur Verfügung stellt, die leicht bedienbar sind und den unterschiedlichen Voraussetzungen auf Produzentenseite gerecht werden.

Eine öffentliche akademische E-Journal-Plattform für Deutschland?

In diesem Zusammenhang stellt sich die Frage, welche Rolle öffentliche, verlagsunabhängige Publikationsplattformen spielen können, die Kleinproduzenten einen kostengünstigen, effektiven und standardkonformen Weg zur Publikation von E-Journals bieten. In den USA können auf solchen – bisher fachspezifischen – Publikationsplattformen E-Journal-Produzenten ihre Produkte kostenfrei oder zu geringen Kosten publizieren. In Deutschland wird dieser Weg derzeit von der German Medical Science-Plattform beschritten.

Die Vorteile für alle Beteiligten liegen auf der Hand. Für die Produzenten erübrigt sich die Schaffung einer eigenen technischen Infrastruktur und die Definition eigener Prozesse. Der Aufwand für die Erstellung und Pflege eines eigenen E-Journals wird deutlich geringer.

Für die Nutzer vereinfacht sich die Suche nach E-Journals, die bestimmten Mindestanforderungen an Qualität und Struktur gerecht werden. Das bedeutet für die Produzenten zugleich eine Optimierung ihrer Reichweite.

Zudem wird die Archivierung für Produzenten und Archive deutlich einfacher. Die Plattform kann so angelegt werden, dass die Einhaltung von Standards gewährleistet wird, ohne dass dabei größere Zusatzaufwände für die Produzenten entstehen. Das Vorliegen von standardgerechten Daten wiederum erleichtert die vollautomatische und zuverlässige Archivierung der Daten.

Für eine Reihe von Problemen, die im Zusammenhang mit der Produktion und Archivierung von E-Journals durch Kleinproduzenten entstehen, könnte die Schaffung einer öffentlichen, fachübergreifenden Publikations-Plattform für wissenschaftliche E-Journals Lösungen bieten. Würde eine solche Plattform von einer Arbeitsgemeinschaft wissenschaftlicher Institutionen, Bibliotheken und Archive in Zusammenarbeit mit interessierten Verlagen konzipiert und umgesetzt, wären die bestmöglichen Voraussetzungen für ihren Erfolg gegeben.

Allgemeine Zusammenfassung und Empfehlungen

Problemstellung

Für die Langzeitarchivierung elektronischer Ressourcen gibt es noch keine etablierten Verfahren. Ursache für diesen Zustand ist die Vielfalt von Konzepten der digitalen Informationsrepräsentation, die ständig wachsende Zahl verschiedenster Datenträger- und Datenformate und der dazugehörigen Soft- und Hardware, die in regelmäßigen Abständen von leistungsfähigeren Nachfolgeversionen abgelöst werden.

Es ist nicht nur absehbar, sondern oftmals schon jetzt der Fall, dass wegen fehlender oder ungeeigneter Archivierung wichtige elektronische Daten nicht mehr verwendbar sind oder mit viel Aufwand rekonstruiert werden müssen.

Die Dringlichkeit, dieses Problem zu lösen, wächst exponentiell an. Daher ist eine Schärfung des Problembewusstseins bei Produzenten wie bei Archivaren elektronischer Dokumente notwendig, mit dem Ziel der Vereinbarung von Standards, ohne die eine kostenverträgliche Langzeitarchivierung digitaler Ressourcen nicht möglich ist.

Besonderes Augenmerk gilt hier dem Publikationstyp E-Journals, da neben die klassische Zweitverwertung von Printjournalen als E-Version zunehmend »E-Only«-Journale treten, die wissenschaftliche Forschungsergebnisse nur noch in digitaler Form veröffentlichen.

Zudem erfolgt beim Übergang in Online-Medien ein Wandel der Form des Publikationstyps »Journal«, der für die Archivierung besondere Herausforderungen darstellt. Während bei der Printpublikation über die wissenschaftlichen Artikel hinausgehende Informationen (Editorial, Zusammensetzung des Editorial Board, Publishing Policies, Autoreninformationen etc.) Bestandteil der Einheit »Heft« sind, werden diese Informationen in Online-Medien getrennt von den durch sie betroffenen Artikeln und in sehr unterschiedlicher Form vorgehalten.

Die Produzenten von E-Only-Journals müssen sich den Herausforderungen der Langzeitarchivierung elektronischer Ressourcen in verstärktem Maße stellen, da hier keine Archivierung einer Printversion mehr dafür bürgt, dass die Inhalte auch in Zukunft verfügbar sein werden.

Allgemeine Konzepte zur Langzeitarchivierung

Für die langfristige Verfügbarmachung digitaler Ressourcen gibt es im Wesentlichen drei Wege. Beim Konzept des »Technikmuseums« wird mit den Datenträgern auch die benötigte Soft- und Hardware archiviert. Dieses Konzept ist aufgrund der Alterung der Datenträger und der Hardware sowie der damit verbundenen Probleme nur in Ausnahmefällen praktikabel und entfällt für umfangreiche Archivierungsvorhaben ganz.

Bei der Migration werden in ihrer Originalform nicht mehr lesbare Daten regelmäßig in neue Versionen konvertiert und ggf. auf neue Datenträger (bzw. in Datenbanksysteme) übertragen, so dass sie auch mit neuer Soft- und Hardware genutzt werden können. Dieses viel praktizierte Verfahren ist durch die massenhaften Konvertierungsvorgänge jedoch sehr aufwendig und birgt ein hohes Risiko der schleichenden Verfälschung der Daten durch ihre Veränderung bei mehrfacher Konvertierung in sich.

Bei der dritten Methode, der Emulation, werden die Daten in ihrer Originalform belassen und lediglich auf neue Datenträger (bzw. in Datenbanksysteme) übertragen. Die zu ihrer Verwendung nötige Soft- und Hardware wird, sobald sie als veraltet gelten muss, durch neue Soft- und Hardware ersetzt, die in der Lage ist, die Funktionen, die zum Abspielen der Altdaten notwendig sind, nachahmend zur Verfügung zu stellen. Dieses Konzept wird teilweise bereits verwendet, erfordert aber einen großen technologischen und organisatorischen Aufwand, der allerdings dadurch kompensiert werden würde, dass die regelmäßige massenhafte Konvertierung von Altdaten entfallen könnte. Eine Verwendung von Emulationstechniken in großem Umfang scheitert derzeit jedoch noch an fehlenden Standards für die Spezifikation von Hard- und Softwarearchitekturen und -konfigurationen.

In der Praxis wird sich eine Mixtur aus Migration und Emulation als der gangbarste Weg erweisen.

Standarddatenformate

Für Migration wie Emulation gilt, dass sie effektiver und kostengünstiger umzusetzen sind, wenn die Zahl der zu berücksichtigenden Datenformate überschaubar ist und wo möglich offene Standardformate genutzt werden.

Insbesondere die Produzenten von E-Journals sind aufgefordert, wo irgend möglich langzeitstabile Standardformate zu verwenden und zu diesem Problemfeld mit den Archiven zu kommunizieren.

Die Archive sollten umfassende Informationsmaterialien mit konkreten Empfehlungen zur Verfügung stellen, um den oftmals ehrenamtlich tätigen Produzenten von E-Journals den Umgang mit diesem Thema zu erleichtern.

Da wir uns derzeit in einer Phase befinden, in der verstärkte Bemühungen um die Entwicklung und Etablierung von langzeitstabilen Standardformaten stattfinden, die in absehbarer Zeit zur Verfügung stehen werden, muss der Kenntnisstand zu diesem Thema regelmäßig aktualisiert werden.

Metadaten-Standards

Metadaten enthalten Informationen über digitale Objekte, die deren Auffindung, Verwaltung und technischen Verarbeitung ermöglichen. Für die Codierung dieser Metadaten gibt es eine Reihe von internationalen Standards, unter denen sich in den letzten Jahren vor allem XML-basierte Standards durchgesetzt haben. Diese Standards erleichtern nicht nur den Datenaustausch, sondern sie ermöglichen auch erst die automatische Bearbeitung der Daten im Archiv.

Für die Speicherung von Metadaten sollten daher solche Standards bevorzugt werden, die einerseits einen hohen Verbreitungsgrad aufweisen und andererseits offen konzipiert sind, so dass auch zukünftig anfallende Metadatentypen abgelegt werden können.

Produzenten und Archive sind aufgefordert, zum Thema der Metadaten in Dialog zu treten, um hier rechtzeitig eine Weichenstellung hin zu einer möglichst weitgehenden Standardisierung zu ermöglichen.

Dabei ist es wichtig, nicht nur das Publikationsobjekt Artikel zu erfassen, sondern auch den bisher vernachlässigten Journal-Kontext, der oft wichtige über den Einzelartikel hinausgehende Informationen enthält, ausreichend zu berücksichtigen.

Auch hier befinden sich Standardformate in der Entwicklung, die in absehbarer Zeit zur Verfügung stehen werden, daher muss der Kenntnisstand auch zu diesem Thema regelmäßig aktualisiert werden.

Datenorganisation und -übergabe

Neben technischen Standards zur Speicherung von Daten und Metadaten sind für die Langzeitarchivierung auch Standards für die Organisation der Datenverwaltung und des Datenaustauschs notwendig.

Das verbreitete Referenzmodell des Open Archival Information System (OAIS) stellt ein standardisiertes konzeptionelles Modell zur Verfügung, das grundlegende Definitionen für Bestandteile digitaler Archive, die in ihnen stattfindenden Abläufe und ihre Kommunikation mit der Außenwelt umfasst. Dieses Modell enthält auch ein Konzept für den Aufbau von Datenpaketen zur Aufnahme in Archive zur Verfügung (Submission Information Package, SIP). In der vorliegenden Studie werden bei der Erörterung von Organisation und Übertragung von E-Journal-Daten zur Übergabe an Archive Informationseinheiten in E-Journals als Submission Information Package (SIP) nach dem OAIS-Konzept verstanden.

Auf dieser konzeptionellen Grundlage wurden verschiedene Standards zur Datenübergabe an Archive verglichen und auf ihre Eignung für die E-Journal-Archivierung überprüft.

Bei E-Journals ist zunächst die Besonderheit zu beachten, dass es neben der Informationsebene »Artikel«, die Forschungsberichte ebenso umfasst wie andere in sich abgeschlossene kurze Beiträge unterschiedlichster Inhaltstypen, auch die Journalebene gibt, also alle Informationen, die das Journal selbst und damit ggf. ganze Gruppen von Artikeln betreffen. Die Informationsebene »Journal« ist bei E-Journals mangels des physischen Rahmens »Heft« und »Band«, der bei Printjournals zur Verfügung steht, schwieriger zu handhaben. Weiterhin ist die Zahl der möglichen Datenformate für Grafiken und multimediale Elemente insbesondere in ihrer zukünftigen Verwendung nicht genau absehbar. Ein Standard zur Übergabe von E-Journal-Daten an Archive muss daher stark variierenden Gegebenheiten in der Organisation und der Art der Inhalte sowie der Datentypen gerecht werden.

Von den fünf analysierten Standards haben sich im Vergleich drei Kandidaten als geeignet herausgestellt: METS, MPEG-21 DIDL und CCSDS.

Nach dem aktuellen Stand der Dinge empfehlen wir die Verwendung von METS. Es erfüllt alle Anforderungen, die der Medientyp E-Journal stellt, gleichzeitig ist es offen und flexibel sowie archivierungsorientiert gestaltet. Zudem weist METS einen großen Bekanntheitsgrad und eine hohe Akzeptanz im internationalen Archivumfeld auf. Anwendungen für E-Journals existieren bereits.

Ob sich die grundsätzlich auch geeignete, vergleichsweise neue DIDL durchsetzen kann ist abzuwarten, und ob die noch unvollendete MPEG-Standard-Familie für Archivlösungen mit dem Ziel der Langzeitarchivierung geeignet sein könnte, müssen künftige, umfassendere Analysen ergeben. Beides gilt auch für den noch unvollendeten CCSDS-Standard und das dazugehörige Framework. Als potenzielle Alternative für METS erscheint uns jedoch DIDL attraktiver als CCSDS, da mit DIDL die Einbindung eines Standards möglich wäre, der bei großer Akzeptanz weit über die Archivwelt hinaus von Relevanz wäre. Da ein Umstieg von METS auf DIDL in jedem Fall möglich ist, würde die Verwendung von METS einen späteren Wechsel auf DIDL nicht behindern.

Der IMS Packaging Standard ist zwar mit METS vergleichbar, insgesamt jedoch weniger flexibel und damit auch weniger für die Langzeitarchivierung von E-Journals geeignet.

Grundsätzlich nicht geeignet ist ONIX, dessen Zielsetzung zu speziell auf den Austausch von Produktinformationen und zu wenig auf den Austausch von komplexen Inhalten mit den dazugehörigen Metainformationen orientiert ist. Ebenso fehlt ONIX die Offenheit, um diese Mängel durch Erweiterungen zu beheben.

Transfermethoden

Eine weitere entscheidende Frage ist die nach der Art der Übermittlung der E-Journal-Daten vom Produzenten an das Archiv.

Technisch besteht kein signifikanter Unterschied darin, ob der Produzent die Datenpakete zum Archiv hochlädt (Push-Prinzip) oder ob sich das Archiv die Daten herunterlädt (Pull-Prinzip). Aus der Workflow-Sicht entlastet das Pull-Prinzip den Produzenten, erschwert aber die Qualitätssicherung. Zudem kann es keine datenbankgenerierten Inhalte laden, es sei denn, es wird ein größerer Zusatzaufwand getrieben. Das Pull-Prinzip ist daher für Lösungen geeignet, bei denen die Beteiligung auch solcher Produzenten, die nur einen sehr geringen Aufwand zugunsten der Langzeitarchivierung leisten können, wichtiger ist als eine optimierte Qualitätssicherung. Das Push-Prinzip ermöglicht dagegen eine optimale Datenqualität und bietet den Zugang zu datenbankgenerierten Inhalten. Es setzt aber eine aktive und intensivere Beteiligung des Produzenten voraus.

Im Sinne einer optimalen Archivierung sollten die Produzenten weitgehend in die erforderlichen Prozesse eingebunden werden. Ob dies flächendeckend möglich ist, hängt nicht zuletzt von der Entwicklung der Gesetzgebung ab, die wie schon bei Printprodukten die Archivierungsaufgabe der entsprechenden Institutionen auch bei digitalen Publikationen durch eine gesetzliche Verpflichtung der Produzenten zur Mitwirkung erleichtern könnte.

Zusammenfassung der Umfrageergebnisse

Die Auswertung der ausgefüllten Fragebögen zeigt ein großes Interesse der Antwortenden an der Thematik. Dieser Eindruck wird allerdings durch die außerordentlich geringe Rücklaufquote der Fragebögen (dies trotz mehrfachen Nachfassens auch durch Die Deutsche Bibliothek) relativiert, die vermuten lässt, dass für die überwiegende Mehrzahl der Informationsanbieter das Thema Langzeitarchivierung derzeit nicht von vordringlichem Interesse ist. Die Daten aus den eingegangenen Fragebögen müssen vor diesem Hintergrund als

nichtrepräsentative, aber qualifizierte Stichprobe aus der Grundgesamtheit der Anbieter von E-Journals verstanden werden.

Die Problematik der Archivierbarkeit der Inhalte ist in Verlagen wie öffentlichen Einrichtungen bekannt und wird ernstgenommen; dennoch bleibt das Gros der Informationsanbieter offensichtlich noch weit hinter dem Stand des technisch Machbaren zurück. Von entscheidender Bedeutung ist dabei sicherlich, dass derzeit überwiegend PDF als Publikationsformat zum Einsatz kommt und damit häufig auch nur dieses Format zur Archivierung herangezogen werden kann. Neben der technischen Hürde, im Rahmen eines PDF-Workflows in der Regel kein anderes, für die Langzeitarchivierung speziell geeignetes Format zu erhalten, sind es vor allem ökonomische Gründe (gepaart mit dem Fehlen einer unmittelbaren Notwendigkeit), die der Erstellung eines Archivformates, das den Kriterien der Langzeitverfügbarkeit gerecht wird, widersprechen. Im universitären Umfeld ist eine Archivierung der E-Journals zwar generell eher gegeben als im Verlagsbereich, was aber nicht heißt, dass es sich um eine einheitliche, an Standards orientierte Form der Archivierung handelt.

Gleichzeitig ist bei den Teilnehmern an der Umfrage die Bereitschaft, sich zukünftig vermehrt Standards zu bedienen, wo diese geschaffen werden und einen wirtschaftlichen Einsatz erlauben, sehr hoch. Die Verantwortung für die Langzeitarchivierung wird dabei aber überwiegend außerhalb der eigenen Institution in öffentlicher Verantwortung gesehen.

Die Schaffung und Zurverfügungstellung von klar definierten, einfach zu verwendenden Schnittstellen und Standards sowie öffentlichen Anlaufstellen für Archivierungszwecke hat also Aussichten auf Einsatz und Erfolg, da auf diese Weise die berechtigten wirtschaftlichen Interessen der Informationsanbieter und die öffentlichen Interessen der Wahrung des Kulturerbes zusammengeführt werden.

Anhang

Abkürzungsverzeichnis

AAF	Advanced Authoring Format
ADL	Advanced Distributed Learning Initiative
AIFF	Audio Interchange File Format (Apple)
AIIM	Association for Information and Image Management
AIP	Archival Information Package
ALS	Audio Lossless Coding
AMDID	Administrative Metadata ID
ARIADNE	Alliance of remote instructional authoring and distribution networks for Europe
ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
Base64	Verfahren zur Kodierung von 8-Bit-Binärdaten
BMP	Windows Bitmap
CCIR 601	jetzt: ITU-R BT 601, Videodecoder
CCITT	Comité Consultatif International Téléphonique et Télégraphique (heute ITU)
CCSDS	Consultative Committee for Space Data Systems
CI	Content Information
CMYK	Cyan, Magenta, Yellow, Key-Color [= Black] (Farbmodus)
CSV	Comma Separated Values (Dateiformat)
DAML	DARPA Agent Markup Language
DC	Dublin Core (Metadaten-Standard)
DFG	Deutsche Forschungsgemeinschaft
DID	Digital Item Declaration
DIDL	Digital Item Declaration Language
DIF	Data Interchange Format (Lotus, jetzt IBM)
DII	Digital Item Identification Language
DIP	Dissemination Information Package
DMDID	Descriptive Metadata ID

DOI	Digital Object Identifier
dpx	Digital Moving Picture Exchange (Dateiformat)
DSC	Document Structuring Conventions
DTD	Document Type Definition, Dokumenttypdefinition für XML/SGML
EDItEUR	International group for electronic commerce in the book and serials sectors
EJAR	E-Journal-Archives der Harvard University, Cambridge MA
EPS	Adobe Encapsulated PostScript (Dateiformat)
EPSI	Adobe Encapsulated PostScript Interchange format (Dateiformat)
FEDORA	Flexible ExtensibleDigital Object and Repository Architecture
FLAC	Free Lossless Audio Codec
FTP	File Transfer Protocol
GIF	Graphics Interchange Format (CompuServe Inc.) (Dateiformat)
GROUPID	Gruppen von Metadata
HDF	Hierarchical Data Format (National Center for Supercomputing Applications)
HDTV	High Definition-TV
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICC	International Color Consortium
ID	Identifikationsnummer
IMS	Instructional Management Systems Global Learning Consortium
ISO	International Organization for Standardization
ISSN	International Standard Serial Number
ITU	International Telecommunication Union
JPEG	Joint Photographic Experts Group (Datei-/Bildkomprimierungsformat)
JPEG 2000	Joint Photographic Experts Group (Datei-/Bildkomprimierungsformat)
LAB	L*a*b*-Farbmodus
LGPL	GNU Lesser General Public License
LMS	Learning Management Systems
LOCKSS	“Lots of Copies Keep Stuff Safe”-Initiative der Universität Stanford
LPAC	Lossless Predictive Audio Compression
LZ77	Kompressionsverfahren nach Lempel/Ziv
LZA	Langzeitarchivierung
LZW	Kompressionsverfahren nach Lempel/Ziv/Welch
MARC	Mailing list ARChive
METS	Metadata Encoding and Transmission Standard

MIME	Multipurpose Internet Mail Extensions
MJPEG	Motion JPEG (Dateiformat zur Videokomprimierung)
MM	Multimedia
MP3	eigentlich: MPEG-1 Audio Layer 3 (Dateiformat zur Audiokompression)
MPEG-1	Moving Picture Experts Group 1 Audio Layer 3, auch MP3 (Dateiformat zur Audiokompression)
MPEG-2	Moving Picture Experts Group, Video-Audio-Standard
MPEG-21	Standard der Moving Pictures Experts Group (MPEG), Multimedia-Framework
MPEG-4	Moving Picture Experts Group, Multimedia Standard
MPEG-4-ALS	Moving Picture Experts Group/Audio Lossless Coding
MPEG-7	Moving Picture Experts Group, Video-Audio-Metadaten
MXF	Media Exchange Format
NASA	National Aeronautics and Space Administration
netCDF	network Common Data Form, University Corporation for Atmospheric Research
NISO	National Information Standards Organization
NPES	Association for Suppliers of Printing , Publishing and Converting Technologies, früher: National Printing Equipment Association
OAIS	Open Archival Information System
OASIS	Organization for the Advancement of Structured Information Standards
OBJID	Object Identifikationsnummer
OCR	Optical Character Recognition, Optische Zeichenerkennung
ONIX	Online Information Exchange
OPAC	Online Public Access Catalog
OWL	Ontology Web Language
PAIMAS	Producer-Archive Interface Methodology Abstract Standard
PAL	Phase Alternating Line, Fernsehnorm
PCM	Pulse Code Modulation
PDF	Portable Document Format
PDF/A	Portable Document Format / Archive
PDF/X-3	Portable Document Format / Exchange, Version 3
PDI	Preservation Description Information
PIF	Package Interchange File (Dateiformat)
PNG	Portable Network Graphics (Dateiformat)
PURL	Persistent Uniform Resource Locator

QA	Quality Assurance
RDF	Ressource Description Framework
RGB	Rot-Grün-Blau Farbmodus
RIFF	Resource Interchange File Format, Microsoft
SCORM	Sharable Content Object Reference Mode
SFDU	Standard Formatted Data Units
SGML	Standard Generalized Markup Language
SICI	Serial Item and Contribution Identifier Standard
SIP	Submission Information Package
SISSL	Sun Industry Standards Source License
SMIL	Synchronized Multimedia Integration Language
SMPTE	Society of Motion Picture and Television Engineers
STM	Science/Technology/Medicine
SVG	Scalable Vektor Graphics
SYLK	Symbolic Link Format, Microsoft
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
TIFF/EP	Tagged Image File Format/Electronic Photography
TIFF/IT	Tagged Image File Format/Image Technology
TSV	Tab Separated Values (Dateiformat)
UML	Unified Modeling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VRML	Virtual Reality Modeling Language, 3D
W3C	World Wide Web Consortium
WAV	eigentlich RIFF (Resource Interchange File Format) WAVE, Microsoft
X3D-XML	eXtensible 3D, XML-basierte Sprache
XHTML	eXtensible Hypertext Markup Language
XML	eXtensible Markup Language
ZIP	Datenkompressionsformat

nestor

<pagina>
Gesamtherstellung
wissenschaftlicher Werke

Studie und Umfrage «Langzeitarchivierung von E-Journals»

Vielen Dank, dass Sie sich an dieser Umfrage beteiligen! Sie ist Teil einer Studie zur Langzeitarchivierung von E-Journals, die von pagina im Rahmen des nestor-Projektes durchgeführt wird.

nestor ist ein **Projekt des Bundesministeriums für Bildung und Forschung,** das Voraussetzungen für eine koordinierte Langzeitarchivierung (LZA) elektronischer Ressourcen in Deutschland schaffen soll (Projektpartner sind Die Deutsche Bibliothek (Deutsche Bücherei Leipzig, Deutsche Bibliothek Frankfurt am Main, Deutsches Musikarchiv Berlin), Bayerische Staatsbibliothek, Niedersächsische Staats- und Universitätsbibliothek, Humboldt-Universität zu Berlin, Staatliche Archive Bayerns und dem Institut für Museumskunde Berlin; mehr unter www.langzeitarchivierung.de).

Ziel der Studie ist eine erste Analyse der technischen Gegebenheiten bei Produktion und Publikation wissenschaftlicher E-Journals in Deutschland, die für die Langzeitarchivierung relevant sind, sowie die Formulierung von Empfehlungen zur Nutzung und Verbesserung dieser Gegebenheiten.

Ziel der Umfrage ist, bei Ihnen, den E-Journal-Produzenten, Informationen über den aktuellen Stand der E-Journal-Produktion in Deutschland zu sammeln. Neben Art, Umfang und technischen Aspekten Ihrer Produktion interessiert uns, wieweit Sie das Anliegen einer produzentenunabhängigen Langzeitarchivierung unterstützen.

Was zählt als E-Journal? Als E-Journal gilt ein elektronisches Publikationsforum mit einer zeitschriftenähnlichen Organisation (qualitätsichernde Redaktion und verantwortlicher Herausgeber), das schwerpunktmäßig der fortlaufenden Veröffentlichung von Artikeln analytisch-investigativen Inhalts dient. Die Existenz einer parallelen Printpublikation ist also kein Ausschlusskriterium. Nicht eingeschlossen sind Diskussions- und Newsforen, Newsletter, Preprint-Server, Dissertations-Server, Expertensysteme, Materialsammlungen, Lecture Notes, Zusammenstellungen von andernorts veröffentlichten Artikeln eines oder mehrerer Autoren u.ä.

Technische Fragen: Falls Sie nicht selbst Techniker sind, benötigen Sie für die Beantwortung der technischen Fragen evtl. Unterstützung von Ihren Mitarbeitern, die die technischen Details kennen.

Rechtliche und Lizenzfragen sind nicht Teil dieser Umfrage. Bei allen Fragen, die auch lizenzrechtliche Aspekte haben (etwa Fragen nach Ihrer Bereitschaft, Daten zur Verfügung zu stellen) soll daher angenommen werden, dass lizenzrechtliche Probleme bereits geklärt sind.

Unter allen Teilnehmern verlosen wir ein Exemplar des Buches «Langzeitarchivierung – Methoden zur Erhaltung digitaler Dokumente» von Uwe Borghoff u.a. im Wert von 45 €.

Ihre Angaben werden selbstverständlich vertraulich behandelt und nur in anonymisierter Form weiterverarbeitet.

Falls Sie Fragen haben, schicken Sie mir eine E-Mail oder rufen Sie mich unter 0 70 71-98 76 18 an.

Vielen Dank für Ihre Mühe!

Tübingen, im Juni 2004

Dr. Gunnar Fuelle, pagina GmbH

Kontakt: Langzeitarchivierung, pagina GmbH Gesamtherstellung wissenschaftlicher Werke, Herrenberger Str. 51, 72070 Tübingen, Tel. 0 70 71-98 76 0, E-Mail: lza@pagina-tuebingen.de, Internet: www.pagina-tuebingen.de

nestor

<pagina>
Gesamtherstellung
wissenschaftlicher Werke

Formular zur Umfrage «Langzeitarchivierung von E-Journals»

Hinweise zum Ausfüllen

Dieses Formular ist ein PDF-Formular, das Sie mit dem Adobe Reader am Bildschirm ausfüllen und drucken können. Mit der Vollversion von Adobe Acrobat oder dem ReaderApproval können Sie Ihre Eintragungen auch zwischenspeichern und mit «Formular absenden» an uns verschicken. Natürlich können Sie das Formular auch ausdrucken, per Hand ausfüllen und uns auf dem Postweg oder per **Fax (0 70 71 - 98 76 22)** zusenden. Vielen Dank!

☐ bedeutet: bitte ankreuzen, mehrere Antworten sind möglich.

◯ bedeutet: bitte ankreuzen, nur eine Antwort ist möglich.

>> bedeutet: bitte Angaben machen bzw. kurze Antwort eingeben

Formular drucken | Eintragungen speichern* | Formular absenden*

* nur mit Vollversion von Adobe Acrobat oder ReaderApproval möglich

Ansprechpartner

Institution/Verlag	>>
Abteilung	>>
Name des Ansprechpartners	>>
Funktion des Ansprechpartners	>>
Straße Hausnr. / Postfach	>>
PLZ	>>
Stadt	>>
Telefon	>> /
E-Mail	>>

Produzenten-Kategorie

Sie sind ein...	
Verlagsunternehmen	◯
Einrichtung öffentlicher Träger *(auch Einrichtung öffentlicher Träger mit Verlagscharakter)*	◯
Wissenschaftliche oder andere Vereinigung (wenn ohne feste Bindung an einen öffentlicher Träger, sonst bitte «Einrichtung öffentlicher Träger» ankreuzen)	◯ Art der Vereinigung >>
Privatperson	◯
und publizieren E-Journals...	
kostenpflichtig	☐
kostenfrei	☐

Kontakt: Langzeitarchivierung, pagina GmbH Gesamtherstellung wissenschaftlicher Werke, Herrenberger Straße 51, 72070 Tübingen, Tel. 0 70 71-98 76 0, E-Mail: lza@pagina-tuebingen.de, Internet: www.pagina-tuebingen.de

Produktionsumfang

Gesamtzahl der von Ihnen publizierten E-Journal-Titel *(einschließlich der Titel, die parallel in Printform und als E-Version erscheinen; nicht gemeint ist die Zahl der jährlich erscheinenden Nummern)*	>>
Gesamtzahl der von Ihnen jährlich publizierten E-Journal-Artikel *(einschließlich der E-Artikel, die parallel in Printform erscheinen)*	>>
Zahl der ausschließlich online publizierten E-Journal-Titel	>>
Zahl der ausschließlich online publizierten E-Journal-Artikel pro Jahr	>>
Stellen Sie Artikel online, die nicht die Print-Version aufgenommen werden?	◯ ja ◯ nein, auch nicht geplant ◯ nein, aber geplant
Haben Ihre E-Journals neben Artikeln auch andere Inhalte? *(z.B. Editorials, Rezensionen, Miszellen, Ankündigungen, Personalien)*	◯ nein ◯ ja, und zwar >>

Stellenwert der Langzeitarchivierung

Wie wichtig ist Ihnen die Langzeitarchivierung Ihrer E-Journals?	◯ egal ◯ wichtig ◯ sehr wichtig evtl. kurze Begründung >>
Wie wichtig ist Ihnen, dass die Langzeitarchivierung Ihres/r E-Journals durch Einrichtungen öffentlicher Träger gewährleistet wird?	◯ egal ◯ wichtig ◯ sehr wichtig evtl. kurze Begründung >>
Verfügen Sie über ein eigenes Langzeitarchivierungssystem?	◯ ja ◯ nein ◯ nein, ist aber geplant

Daten-Formate I: Textdaten und Grafiken

In welchen Formaten publizieren Sie Ihre E-Journals? **Welche Formate verwenden Sie intern?**	Publikation	Intern	
Dateiformate der Dokumente (Fortsetzung auf S.4)	☐	☐	PDF (Details unbekannt)
	☐	☐	PDF mit Nicht-Standard-Settings
	☐	☐	PDF/X nach ISO-Norm
	☐	☐	HTML (Details unbekannt)

In welchen Formaten publizieren Sie Ihre E-Journals? **Welche Formate verwenden Sie intern?** (Fortsetzung von S. 3)	Publikation	Intern	
Dateiformate der Dokumente (Fortsetzung von S.3)	☐	☐	HTML, unvalidiert
	☐	☐	HTML 4, W3C-validiert
	☐	☐	XHTML, W3C-validiert
	☐	☐	XML mit eigener DTD/Schema
	☐	☐	XML nach Standard-DTD/Schema und zwar >>
	☐	☐	SGML mit eigener DTD
	☐	☐	SGML nach Standard-DTD und zwar >>
	☐	☐	andere und zwar >>
Wenn Sie zur Publikation PDF verwenden: Ist Ihr PDF geschützt, und wenn ja, auf welche Art?			☐ ungeschützt ☐ Kennwort zum Öffnen erforderlich ☐ Druckfunktionen gesperrt ☐ Kopieren aus dem Dokument gesperrt ☐ Bearbeitungsfunktionen gesperrt ☐ Kommentarfunktionen gesperrt
Planen Sie den Einsatz anderer technischer Maßnahmen zum Schutz der Dokumente vor unerwünschter Nutzung oder Weiterverbreitung (Digital Rights Management)? Wenn ja, welche?			◯ nein ◯ ja, und zwar >>
Wenn Sie SGML, XML oder HTML verwenden: Wird per Stylesheet Text generiert (z.B. Trennzeichen, Überschriften o.ä.)?			◯ ja ◯ nein
Wenn Sie kein SGML oder XML verwenden: Planen Sie die Nutzung von SGML oder XML als internes oder als Publikationsformat?			◯ nein ◯ ja, bereits konkret in Planung ◯ ja, aber erst angedacht
Wenn Ihre E-Journals auch andere Publikationsobjekte als Artikel enthalten (z.B. Editorials, Rezensionen, Miszellen), in welchem Format liegen diese Objekte vor?			☐ in gleicher Form wie die Artikel ☐ andere, und zwar >>
Wenn Sie HTML oder XML verwenden: Welche Dateiformate nutzen sie zur Online-Publikation von Grafiken?			☐ GIF ☐ JPG ☐ BMP ☐ PNG ☐ TIFF ☐ EPS ☐ andere, und zwar >>

Wenn Sie HTML oder XML verwenden: Welchen Grafik-Dateiformate verwenden Sie standardmäßig intern? («standardmäßig» meint, dass *alle* Grafiken in diesem Format vorliegen)	☐ GIF ☐ JPG ☐ BMP ☐ PNG ☐ TIFF ☐ EPS ☐ andere, und zwar >>
Publizieren Sie unterschiedliche Auflösungen derselben Grafiken? (z.B. Vorschaugrafiken mit Links auf die hochauflösende Version)	◯ ja ◯ nein, auch nicht geplant ◯ nein, ist aber geplant

Daten-Formate II: Multimediale Elemente

«Multimediale Elemente» sind andere Elemente als Texte oder unveränderliche Abbildungen, also z.B. dreh- und zoombare 3-D-Modelle, 360-Grad-Ansichten, Animationen, Videos, Audiodaten u.ä.

Enthalten Ihre E-Journals multimediale Elemente, und wenn ja, welche?	◯ nein, auch nicht geplant ◯ nein, aber geplant ◯ ja, und zwar >>
Welche Dateiformate verwenden Sie für multimediale Elemente? (Bitte wenn möglich mit Versionsangabe)	>>
Welche Plug-Ins sind zur Anzeige dieser Elemente erforderlich? (Wenn möglich mit Versionsangabe, auch selbst programmierte Plug-Ins angeben)	>>
Sind multimediale Elemente wesentlicher Bestandteil Ihrer E-Journals oder nur gelegentlich verwendetes «Beiwerk», auf das bei der Langzeitarchivierung auch verzichtet werden könnte?	◯ wesentlicher Bestandteil ◯ nur „Beiwerk“
Wie schätzen Sie die zukünftige Bedeutung von multimedialen Elementen in E-Journals ein?	◯ ganz unbedeutend ◯ gering ◯ wichtig ◯ sehr wichtig

Daten-Formate III: Dynamische Elemente

«Dynamische Elemente» sind Elemente, die nicht mit der Publikation abgeschlossen sind, sondern sich auch danach noch verändern. Dynamische Elemente sind z.B. von Lesern erstellbare Diskussionsbeiträge, die einem Dokument zugeordnet sind oder Diagramme, die aus regelmäßig aktualisierten Daten immer neu erstellt werden (z.B. B rsenkurse).

Enthalten Ihre E-Journals dynamische Elemente? Wenn ja, welche?	◯ nein, auch nicht geplant ◯ nein, aber geplant ◯ ja, und zwar >>

Welche Dateiformate verwenden Sie für dynamische Elemente? (Bitte wenn vorhanden mit Versionsangabe)	>>
Welche Plug-Ins sind zur Anzeige dieser Elemente erforderlich? (Wenn vorhanden mit Versionsangabe, auch selbst programmierte Plug-Ins angeben)	>>
Sind dynamische Elemente wesentlicher Bestandteil Ihrer E-Journals oder nur gelegentlich verwendetes «Beiwerk», auf das bei der Langzeitarchivierung auch verzichtet werden könnte?	◯ wesentlicher Bestandteil ◯ nur „Beiwerk“
Wie schätzen Sie die zukünftige Bedeutung von dynamischen Elementen in E-Journals ein?	◯ ganz unbedeutend ◯ gering ◯ wichtig ◯ sehr wichtig

Bereitstellungsform der Inhalte

Stellen Sie Ihre Dokumente - unabhängig vom Format - als fertige (statische) Dateien zur Nutzung zur Verfügung *oder* werden die Dokumente erst bei der Abfrage aus einer Datenquelle (Datenbank, CMS, SGML-, XML-Datei) generiert?	◯ statisch ◯ statisch, aber Umstellung ◯ auf generierte Ausgabe geplant ◯ generiert ◯ teils statisch, teils generiert
Wenn Dokumente generiert werden, hat der Nutzer direkt oder indirekt Einfluss auf das Ausgabeergebnis? (z.B. Personalisierung durch Anpassung an Nutzerprofil oder Möglichkeit zur Eingabe von Optionen, die die Ausgabe beeinflussen)	◯ ja, die Ausgabe variiert nach Nutzer ◯ nein, die Ausgabe ist unveränderlich

Bibliographische und organisatorische Daten (Meta-Daten)

Für welche Publikationsobjekte halten Sie bibliographische Daten vor?	☐ keine ☐ E-Journal-Titel (Gesamtwerk) ☐ Jahrgang/Volume ☐ Heft ☐ Artikel ☐ weitere (z.B. Miszellen, Rezensionen) und zwar >>
Welche bibliographischen Daten halten Sie für die Artikel vor? (z.B. Autor/en, Titel, Abstract, Jahrgang, Publikationsdatum, Publikationstyp, Erscheinungsland, Sprache, Fachbereichszuordnung, Schlagworte, ISSN usw.)	>>
Verwenden Sie Standards bei der Vergabe von bibliographischen Daten oder haben Sie eigene Regeln? (für einzelne Angaben oder für die Gesamtheit der Meta-Daten; z.B. ISSN, DIN-Abkürzungen, Internationale Dezimalklassifikation, CODEN, ONIX, DublinCore)	◯ Haben eigene Regeln ◯ Verwenden Standards, und zwar >>

In welcher Form halten Sie die bibliographischen Daten vor?	☐ als Bestandteil des einzelnen Artikels ☐ als Teil der HTML-Navigation/-Übersicht ☐ als gesonderte Datei, und zwar im Format >> ☐ in Datenbank ◯ von Nutzer durchsuchbar ◯ nicht durchsuchbar ☐ andere Form, und zwar >>
Halten Sie neben den bibliographischen auch *organisatorische Daten* vor? Wenn ja, in welcher Form? (gemeint sind z.B. Angaben über die Herausgeber, die Mitglieder des wissenschaftlichen Beirats, Schreibanweisungen, Zitationsregeln u.a.)	◯ nein ◯ ja, und zwar folgende Daten >> ◯ nur aktuell gültige Angaben ◯ auch für zurückliegende Zeiträume ◯ als Teil der HTML-Navigation/-Übersicht ◯ als gesonderte Datei, und zwar im Format >> ☐ in Datenbank ◯ von Nutzer durchsuchbar ◯ nicht durchsuchbar ☐ andere Form, und zwar >>

Dokument-Identifizierung und Linking

Internet-Adressen sind zwar eindeutig, aber nicht selten wird die Adresse von Dokumenten verändert - "tote Links" sind die Folge. Daher etablieren sich zunehmend internationale Standards, die digitalen Dokumenten weltweit eindeutige und unveränderliche Kennungen (Persistent Identifiers) zuordnen. Unter diesen Persistent Identifiers sind die Dokumente auch dann zu finden, wenn sich ihre Adresse geändert hat (mehr Infos zu Persistent Identifiers hier).

Verwenden Sie Persistent Identifiers? Wenn ja, welches System? Wenn nein, sind Sie an einer Nutzung interessiert?	◯ Verwende keine Persistent Identifiers ... ☐ plane aber Einsatz ◯ Verwende Persistent Identifiers, und zwar ☐ DOI ☐ PURL ☐ URN ☐ eigene Identifier ohne Anbindung an internationale Systeme

Enthalten Ihre Dokumente Links auf andere Dokumente? Wenn ja, welches Linking-System nutzen Sie?	☐ keine Links ☐ Internet-Adressen (URLs) ☐ DOI mit CrossRef ☐ andere Persistent-Identifier-basierte Links und zwar >>

Digitale Signaturen

Um sicherzustellen, dass Dokumente nicht verändert wurden und eindeutig einem Autor zugeordnet werden können, können Sie sie mit digitalen Signaturen versehen.

Verwenden Sie digitale Signaturen für Ihre Dokumente?	◯ nein ◯ ja, und zwar folgendes System >>

Bereitschaft zur Umsetzung von Anforderungen für die Langzeitarchivierung

Die Langzeitarchivierung durch die öffentliche Hand wird in vielen Hinsichten effektiver, nachhaltiger und kostengünstiger, wenn die Zahl der zu archivierenden Datenformate überschaubar ist und diese Formate offenen Standards entsprechen. Das gilt auch für die Handhabung von Meta-Daten.

Gibt es in Ihrer Institution Bemühungen, die Produktion von E-Journals zu standardisieren, sei es intern oder durch Mitarbeit in oder Ausrichtung an überinstitutionellen Initiativen?	◯ nein, auch nicht geplant ◯ nein, aber geplant ◯ ja, und zwar >>
Sind Sie grundsätzlich bereit, die Langzeitarchivierung durch öffentliche Institutionen zu erleichtern, indem Sie Ihre Publikationen in Standardformaten zur Verfügung stellen?	◯ ja ◯ nein
Wenn Sie bedenken, dass multimediale oder dynamische Inhalte besondere Schwierigkeiten bei der Archivierung verursachen: Würden Sie auf solche Inhalte verzichten oder sich auf Formate beschränken, für die internationale offene Standards existieren?	Ich würde auf solche Inhalte ◯ verzichten ◯ nicht verzichten ◯ nur Standards verwenden
Wenn Sie schon jetzt intern SGML oder XML verwenden: Würden Sie diese Daten zusätzlich zu den Publikationsdaten für die LZA zur Verfügung stellen?	◯ ja ◯ nein
Wären Sie bereit, Ihre Daten auf ein XML-Format umzustellen, das als Standard für E-Journals empfohlen wird bzw. würden Sie Ihre Daten zusätzlich in einem solchen Format zur Verfügung stellen?	◯ ja ◯ nein
Falls Sie kein XML oder SGML verwenden wollen, würden Sie Ihre HTML- oder PDF-Ausgabe generell oder auch nur für die LZA auf eine standardgemäße Form umstellen (XHTML, PDF/A)?	◯ ja ◯ nein

Wären Sie bereit, die Langzeitarchivierung durch öffentliche Institutionen zu erleichtern, indem Sie Metadaten in Standardformaten zur Verfügung stellen?	◯ ja ◯ nein
Wenn es einen Standard für den Transfer von E-Journals zur LZA gäbe, würden Sie diesen Standard nutzen?	◯ ja ◯ nein

Anmerkungen >>