Reinhard Altenhöner[1]

Daten für die Zukunft – Das BMBF-Projekt Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen (kopal) und seine Hintergründe

Seit Sommer 2004 führt Die Deutsche Bibliothek zusammen mit mehreren Partnern mit kopal ein vom Bundesministerium für Bildung und Forschung finanziertes Projekt zur Etablierung einer technischen Plattform für die Langzeitarchivierung in Deutschland durch. Die Tatsache, dass das Projekt mittlerweile in die eigentliche Entwicklungs- und Umsetzungsphase eingetreten ist, gibt Anlass, das Vorhaben und den aktuell erreichten Stand in größerem Zusammenhang vorzustellen.

Data into the future – the project of the Federal Ministry of Education and Research „Co-operative Development of a Long Term Digital Information Archive" (kopal) and its background

As of summer 2004, DDB and several partners are conducting a project called kopal, financed by the Bundesmi-nisterium für Bildung und Forschung, with the objective to establish a technology platform for long-term archiving in Germany. Because kopal has already entered the development and implementation phase, we feel the time is ripe for us to present the project and its current realization to a wider public.

Des dates pour l'avenir – Le projet du Ministère fédéral de la formation et de la recherche „Établissement coopératif d'archives à long terme des informations numérisées" (kopal) et ses arrière-plans

Depuis l'été 2004 Die Deutsche Bibliothek (Francfort/Leipzig) avec quelques partenaires poursuit le projet kopal financé par le Ministère fédéral de la formation et de la recherche en vue de créer une plate-forme technique destinée à la préservation archivistqiue en Allemagne. Le fait que le projet est entré entre-temps dans la phase de développement et de réalisation propre, donne lieu de présenter le projet et l'état actuel dans un contexte plus large.

Seit Sommer 2004 führt Die Deutsche Bibliothek zusammen mit mehreren Partnern (der Staats- und Universitätsbibliothek Göttingen [SUB], der Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen [GWDG] und der Firma International Business Machines Corporation Deutschland [IBM]) ein groß angelegtes, vom Bundesministerium für Bildung und Forschung finanziertes Projekt zur Etablierung einer technischen Plattform für die Langzeitarchivierung in Deutschland durch. Der vorliegende Beitrag schildert in knapper Form die Ausgangslage und die Rahmenbedingungen, unter denen Projekte zur Langzeitarchivierung generell stehen. Darüber hinaus wird die Tatsache, dass das Projekt selbst mittlerweile in die eigentliche Entwicklungs- und Umsetzungsphase eingetreten ist, dazu genutzt, das Vorhaben, seine konkreten Arbeitsbedingungen und die aktuell erreichte Situation kurz vorzustellen.

1. Die steigende Anzahl, Heterogenität und wissenschaftliche Relevanz digitaler Ressourcen haben die Möglichkeiten und Bedingungen ihrer zuverlässigen und dauerhaften Archivierung und Verfügbarkeit in den Blickpunkt von Gedächtnisinstitutionen wie Bibliotheken, Archive und Museen gerückt. Allein für die wissenschaftlichen Journale wird erwartet, dass bis zum Jahr 2010 65 % bis 95 % von ihnen in digitaler Form vorliegen werden – schon dies belegt die Bedeutung ihrer langfristigen Zugänglichkeit.
Hinsichtlich des gedruckten Publikationssektors gibt es eine klare Situation: Bibliotheken hatten und haben den Auftrag, das Schaffen einer Nation für die Nachwelt im Sinne der Gedächtnisüberlieferung eines Zeitalters verfügbar zu erhalten, mindestens aber einen relevanten Ausschnitt aus diesem Schaffen. In Deutschland liegen diese Funktionen zum einen bundesgesetzlich umfassend geregelt bei Der Deutschen Bibliothek, zum anderen bei einer größeren Zahl von Regionalbibliotheken, die im allgemeinen aufgrund eigener landesgesetzlicher Regelungen für die jeweilige Region begrenzte Pflichtexemplarrechte haben. Das angesprochene Problem, digitale Publikationen langfristig verfügbar zu halten, betrifft sowohl die Haltbarkeit der verwendeten Datenspeicher als auch die Sicherung des künftigen Zugriffs und der Benutzbarkeit der nur noch digital publizierten Informationen. Datenträger zerfallen, rasante Technologiewechsel erschweren den Zugriff auf ältere Träger und Objektstrukturen, Datenformate und Betriebssystemumgebungen geraten in Vergessenheit.
Prinzipiell müssen Aktivitäten zur Langzeitarchivierung digitaler Information auf alle Arten digitaler Ressourcen zielen: sowohl auf diejenigen, die unter den klassischen Kategorien „Dokumente", „Publikationen" oder auch „Informationenobjekte" gefasst werden können, als auch auf diejenigen, die in Form wissenschaftlicher Primärdaten Grundlage für zukünftige Auswertungen darstellen und dauerhaft bereitgehalten werden müssen. Diese

[1] Das Projekt und auch darauf sich beziehende Publikationen werden von zahlreichen Personen getragen; ich danke besonders den Kollegen Hans Liegmann und Tobias Steinke aus Der Deutschen Bibliothek für vielfältige Unterstützung.

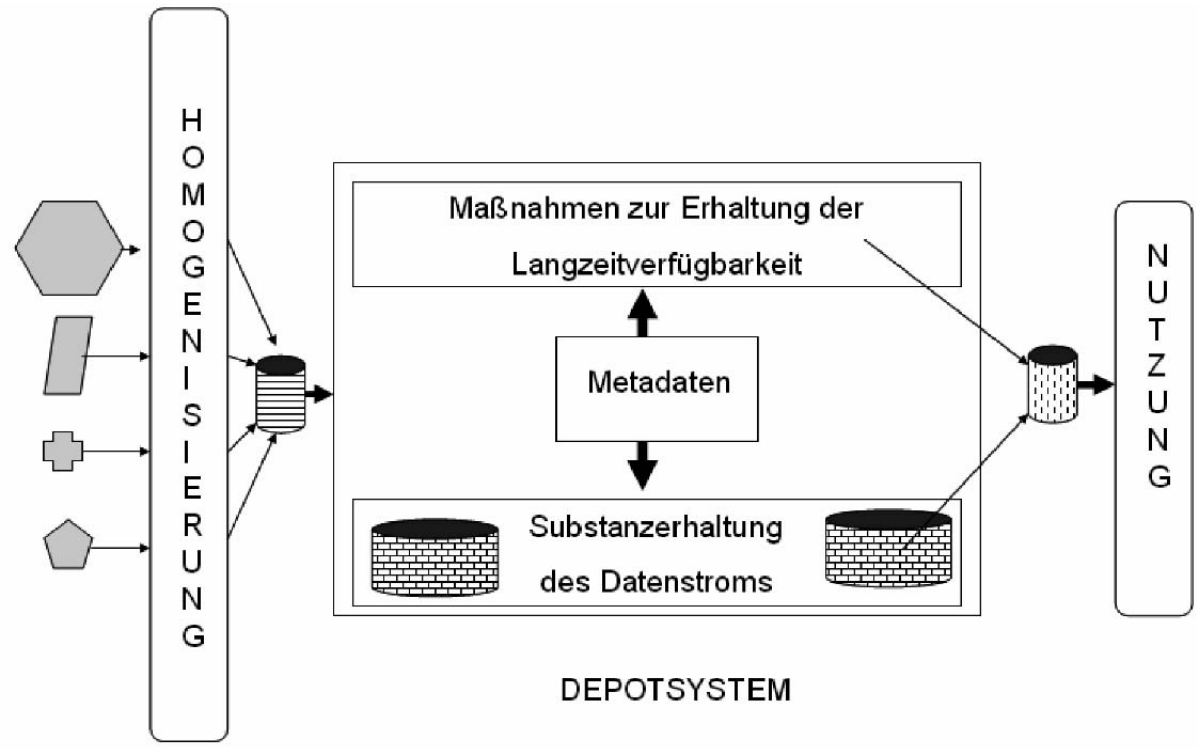

Abb. 1: Quelle: Hans Liegmann.

digitalen Ressourcen umfassen demnach alle heute bekannten medialen Typen wie Text, Bild, Bewegtbild, Ton sowie deren Mischformen; die Vertriebswege können sowohl trägergebunden sein als auch trägerlos über Netzwerke laufen.

2. Es kommt hinzu, dass vor dem Hintergrund der wachsenden Verlagerung des Publikationslebens in elektronische Medien einerseits bislang gültige Kategorien der Zuordnung einzelner Objekte zu bestimmten Dokumentenklassen ungültig werden, zum anderen aber auch Abgrenzungen anderer Art fallen: Beispielweise greifen bei Netzpublikationen räumliche Zuordnungen von ablieferpflichtigen Produzenten zu einer Bibliothek ebenso wenig wie die Definition von Zeitschriften als formale Hülle für die Publikation einzelner Aufsätze.
Ein ganz entscheidender Unterschied aber ist, dass die Selbstverständlichkeit, mit der man auf die immanente Eigenschaft des jeweiligen Objekts, langfristig verfügbar zu sein, vertrauen kann, obsolet wird: Gedruckte Entitäten weisen per se physisch eine hohe Haltbarkeit auf, während digitale Objekte besonderer Vorkehrungen zur Sicherung ihrer Langzeitverfügbarkeit bedürfen. Diese Entwicklung und die Erkenntnis der besonderen Problemlage im Umgang mit den elektronischen Publikationen sind inzwischen Gegenstand intensiver, häufig international geführter Diskussionen und umfangreicher Aktivitäten geworden.

3. Unter „Langzeitarchivierung“ verstehen wir die erfolgreiche Gewährleistung der Langzeitverfügbarkeit einer digitalen Ressource in einer nicht vordefinierten Zeitspanne. Die zurzeit mit begrenztem Mitteleinsatz erreichbaren Zeitvorgaben von fünf, zehn oder 20 Jahren müssen dabei deutlich überschritten werden. Bereits heute sind viele elektronische Publikationen, die in früher üblichen Dateiformaten vorliegen, bedingt durch die technische Entwicklung gar nicht mehr oder nur unter erheblichem technischen Aufwand benutzbar. Publikationen in älteren Formaten und Programmen oder auf veralteten Datenträgern können nur noch mit großem Aufwand auf Systemen der Gegenwart lesbar und nutzbar gemacht werden. Beispielsweise stellt bereits heute die Nutzung von Ressourcen der ersten Home-Computer-Generationen (z. B. C64, Atari, AMIGA) ein ernsthaftes Problem dar, weil geeignete Lesegeräte kaum mehr existieren und zum Teil auch die Datenformate nicht mehr erkannt werden.
Die wesentlichen Inhaltsbestandteile einer digitalen Publikation der Gegenwart sollen also auch zu einem Zeitpunkt der ferneren Zukunft noch nutzbar und zugänglich sein, auch wenn wir heute noch nicht wissen, wie der weitere Verlauf des technischen Fortschritts aussehen wird.
Digitale Daten sind letztlich nur eine Aneinanderreihung von Nullen und Einsen. Erst technische Geräte (Computer) geben einen für Menschen verständlichen Sinn. Sie benötigen dazu wiederum umgebungsabhängige Verarbeitungsprozeduren (Programme). Für den künftigen Zugriff auf die Inhalte werden Verfahren benötigt, welche die Leistungen dieser Systemumgebungen (Hardware und Software) in Zukunft bereitstellen.
Das obige Bild soll die Position eines Depotsystems in einer Umgebung heterogener Publikations- und Dokumenttypen auf der einen und den Dienstleistungen für die Nutzung auf der anderen Seite zeigen. Ohne eine wesentliche Verminderung der Heterogenität der Objekte (z. B. durch Standardisierung oder Konvertierung unter Einhaltung der Authentizität) ist die Eingangsschnittstelle des Depotsystems nicht beherrschbar. Nach erfolgreicher Übernahme des Objekts steuert das System über die Metadaten einerseits den Erhalt der Substanz und führt andererseits gezielte Maßnahmen (Migration / Emulation, s.u.) durch, die gemeinsam die Langzeitverfügbarkeit des jeweiligen

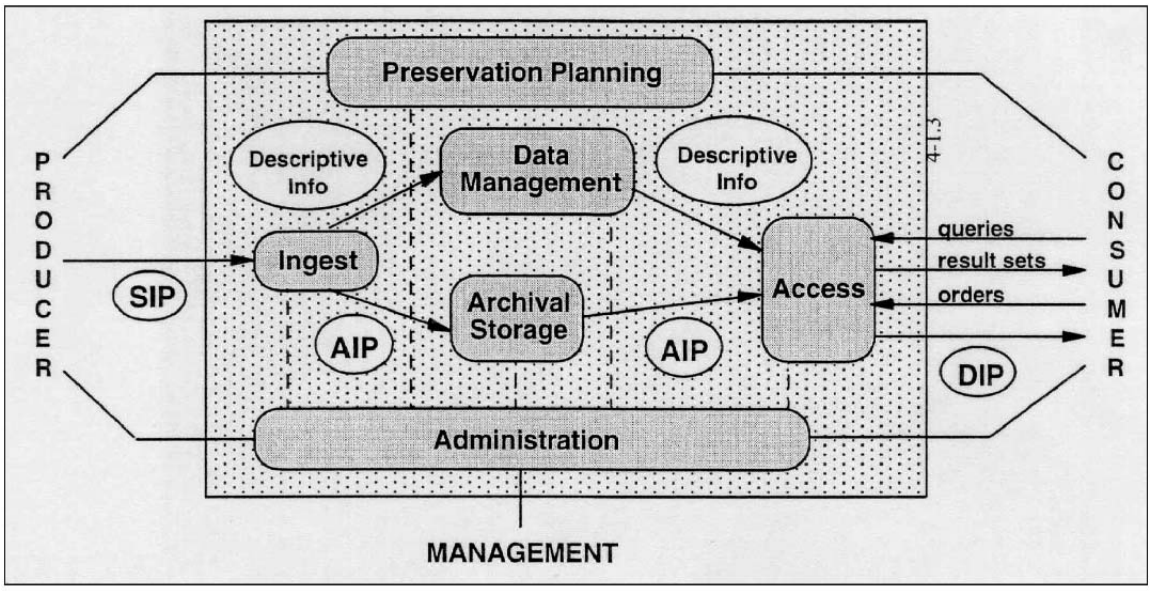

Abb. 2: OAIS-Modell.

Objekts für eine Benutzungssituation sicherstellen – im Idealfall die Bereitstellung des Objekts auf einem gesichert nutzbaren Formatstand mit all den Informationen, die für den Aufbau einer angemessenen Benutzungsumgebung erforderlich sind.

4. In einer eher abstrakten Sicht besteht hinsichtlich der geforderten Grundfunktionen eines solchen Archivierungssystems international weitgehend Einigkeit. Der ursprünglich aus dem Kontext der Raumfahrt stammende ISO-Standard **14721:2003** – Reference Model for an Open Archival Information System (OAIS)[2] – beschreibt die Infrastruktur eines digitalen Archivs in Form eines Referenzmodells. Durch die Abgrenzung und eindeutige Benennung von Funktionsmodulen, Schnittstellen und einer Typologie von Informationsobjekten ist es gelungen, eine über die Grenzen der Anwendergemeinschaften Archiv, Datenzentren und Bibliotheken hinweg geltende allgemeine Sicht auf die Kernfunktionen eines digitalen Archivs zu schaffen. Damit ist eine wertvolle und allgemein anerkannte Grundlage für die Beauftragung, Planung und Implementierung produktiver Systeme entstanden[3].
OAIS beschreibt eine Anzahl von Funktionsmodulen, die dem Datenfluss und den Arbeitsabläufen eines Archivs entsprechend angeordnet sind: Eingangsbearbeitung (ingest), Metadatenverwaltung (data management), Objektspeicherung (archival storage), Erhaltung der Langzeitverfügbarkeit (preservation planning), Administration und Bereitstellung (access).
Erklärungsbedürftig ist sicherlich die „Abkürzungsfamilie" SIP, AIP und DIP: Unterschieden wird im Referenzmodell zwischen dem data object (Datenobjekt) und dem information object (Informationsobjekt). Danach ist ein Objekt – beispielsweise ein Programm – nur dann für mich als Nutzer nachvollziehbar, wenn ich die entsprechende Programmiersprache kenne („knowledge") und gleichzeitig über weiterführende Nachschlagemöglichkeiten zu dieser Sprache verfüge („representation information").
Im OAIS-Modell ist das „information package" der entscheidende Begriff, der aus der „content information" (enthält das Objekt selbst inkl. der „representation information") und der „preservation description information", also den für die Archivierung erforderlichen Informationen, besteht. Die Metadaten wiederum beschreiben den Inhalt des „information package".
Im modellhaften Archivsystem bezeichnet daher das „submission information package" (SIP) quasi die Eingangsschnittstelle, in der das Objekt mit deskriptiver Information zu einem archive information package (AIP) konvertiert wird. Das AIP wird dann in den eigentlichen Archivprozess übergeben, während die Metadaten in das Datenmanagementsystem eingehen. Im System greifen verschiedene Prozesse zur Sicherstellung der Langzeitverfügbarkeit. Für das dissemination information package (DIP) kehren sich die Prozesse auf eine Anforderung hin um und die eingelagerten Objekte werden zur Verfügung gestellt.
Das OAIS-Modell ist international sicher eines der meist genannten Konstrukte, wenn es darum geht, Anforderungen an Systeme zur Langzeitarchivierung festzulegen. Besonders wichtig in dem Modell sind die zahlreichen „Aussenbezüge" (Monitoring) zum Beispiel mit einer Funktion wie „Monitor Designated Community", die sicherstellen soll, dass aktuelle Informationen über die Nutzerbedürfnisse gesammelt werden. Mit der Funktion „Monitor Technology" wird die Entwicklung digitaler Technologien in der Außenwelt des Systems beobachtet. Es sind diejenigen Entwicklungen frühzeitig zu identifizieren, die schädliche Auswirkungen auf die Benutzbarkeit der im System gespeicherten Objekte haben können.

5. Digitale Ressourcen können im Prinzip auf jedem Medium gespeichert werden, das in der Lage ist, eine Folge von Nullen und Einsen geordnet aufzunehmen. Die unbe-

[2] <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>. Vgl. auch den Hinweis der ISO: <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=24683&ICS1=49&ICS2=140&ICS3>.
[3] <http://www.rlg.org/longterm/oais_schematics.html>.

schädigte Erhaltung dieser Abfolge ist die Grundvoraussetzung für alle weiteren Aktivitäten. Kann diese Voraussetzung nicht verlässlich eingehalten werden, so erübrigt sich die Diskussion um die auf ihr aufbauenden Aktivitäten zur Erhaltung der Benutzbarkeit. Die Empfindlichkeit digitaler Medien gegen den Ausfall bedeutungstragender Elemente ist unterschiedlich groß: führt ein Kratzer auf einer Audio-CD zu einem knackenden Geräusch und damit nur zu einer relativ geringfügigen Beeinträchtigung der Wiedergabequalität, so kann die Zerstörung nur eines Zeichens in einer komprimierten ausführbaren Datei den Verlust des gesamten Objektes zur Folge haben.
Die Substanzerhaltung elektronischer Publikationen hat im Wesentlichen zwei Faktoren zu beherrschen: zum einen die begrenzte Lebensdauer digitaler Speichermedien, zum anderen den rasanten technischen Fortschritt im Bereich der digitalen Speichertechnologien.
Mit der Bewältigung der begrenzten Lebensdauer digitaler Speichermedien haben wir inzwischen Erfahrungen: wir müssen die zwar kontroversen, aber immerhin als Richtwerte vorliegenden Angaben zur Lebensdauer von Datenträgern berücksichtigen und von Zeit zu Zeit prüfen, ob die gespeicherten digitalen Informationen auf der untersten Ebene (Nullen und Einsen) noch lesbar sind. Aus Sicherheitsgründen müssen wir bereits vor dem Erreichen der pessimistisch bestimmten Lebenserwartung eines Trägers die auf ihm gespeicherten Daten auf einen jüngeren Träger gleichen Typs umkopieren oder zu einer neuen Generation von Datenträgern wechseln. In Rechenzentren ist es eine eingespielte Praxis, Magnetbänder und Magnetbandkassetten zyklisch zu überprüfen bzw. vorsorglich zu ersetzen.
In weiterer Differenzierung können wir zwischen dem „Wiederauffrischen“ des Trägers (refreshing) und einem Wechsel der Trägergeneration (re-formatting) unterscheiden. Insbesondere der Wechsel des Trägers kann allerdings erhebliche Risiken für das Objekt nach sich ziehen: Das Ziel einer Sicherung des Informationsgehalts wird durch eine Veränderung des Datenstroms erkauft – mit fraglichen Einschnitten für die Authentizität des Ausgangsobjekts.
Für die Durchführung solcher substanzerhaltenden Maßnahmen benötigen wir Metadaten in einer erweiterten Qualität, die zur automatischen Prozesssteuerung eingesetzt werden können. Dies sind z. B. strukturierte und maschinell interpretierbare Angaben über Datenträgertypen, Materialarten und Produktionszeitpunkte.

6. Die Substanzerhaltung des Datenstroms ist zwar die Voraussetzung für die Erhaltung der Benutzbarkeit von digitalen Objekten, ohne zusätzliche Vorkehrungen erhalten wir damit jedoch unseren Nachkommen nur eine Menge von codierten Informationen, die nur schwer oder gar nicht zu entschlüsseln sind.
Im Gegensatz zu gedruckter Information benötigen wir zur Dekodierung digitaler Objekte ein Darstellungssystem aus Hardware und Software, um den Informationsgehalt der digitalen Ressourcen zugänglich zu machen. Zwei wesentliche Strategien bestimmen die weltweite Diskussion um die Erhaltung der Benutzbarkeit:
Migration ist die periodische Übertragung digitaler Ressourcen zwischen unterschiedlichen Hardware- und Software-Konfigurationen oder Hardware- und Software-Generationen. Zweck der Migration ist es, die Integrität und die Verfügbarkeit digitaler Ressourcen trotz des stetigen Wandels der technischen Umgebung zu erhalten. Migration schließt das Kopieren zwischen Datenträgern einer Generation ein, kann jedoch auch strukturelle Eingriffe in die Objekte enthalten, um die Kompatibilität zu veränderten technischen Umgebungen herzustellen. Migration ist im Anwendungsbereich der Informationstechnologie ein bereits umfangreich erprobtes und produktiv eingesetztes Verfahren. Als Strategie zur Langzeiterhaltung ist die Migration allerdings problematisch, da die Aufwendungen für in Zukunft notwendige Migrationszyklen unbekannter Häufigkeit nicht vorausberechnet werden können.
Unter *Emulation* versteht man eine Hard- oder Softwareeinrichtung, mittels derer ein System die Funktionalität eines anderen wohl definierten Systems vollständig nachzubilden in der Lage ist. Zweck der Emulation ist es, auf einem System Daten und Programme zu verarbeiten, die ursprünglich für ein anderes, nun historisches System bestimmt waren, auf dem aktuellen System also quasi die Bedingungen des anderen Systems so exakt nach- oder abzubilden, dass die abgefragten Daten bzw. Programme tatsächlich so genutzt werden können, wie sie auf dem anderen System einmal benutzbar waren. Die Anwendung von Emulationsverfahren bei der Langzeiterhaltung digitaler Ressourcen wird vor allem dann erwogen, wenn die Migration wegen hoher Komplexität der digitalen Objekte ausgeschlossen wird. Dies kann sehr aufwändig sein und setzt vor allem eine sehr genaue Definition bzw. Beschreibung der hard- und softwareseitigen Systemanforderungen voraus. Außerdem besteht auch hier das Risiko fehlender Funktionen und von daher eines Verlustes an Information.
Die jüngste Ausprägung der Emulationsstrategie hat in Zusammenarbeit mit wissenschaftlichen Einrichtungen ein Forschungslabor der Firma IBM geschaffen: das Konzept eines „Universellen Virtuellen Computers (universal virtual computer – UVC)“. Es beruht auf der Vorstellung, dass auf den technischen Plattformen der Zukunft lauffähige Emulatorprogramme geschrieben werden, die jedoch nicht unbegrenzt viele „Alt“-Systeme emulieren, sondern lediglich ein einziges beständiges virtuelles System: den UVC. Dieser ist konzipiert als ein Computer mit einem solch grundlegenden und einfachen Aufbau, dass seine Funktionen in beliebigen Umgebungen verfügbar gemacht werden können. Sobald eine UVC-Emulation auf einem Zielsystem existiert, würde diese mit der logischen Datenbeschreibung (logical data description) eines digitalen Objektes versorgt und die Darstellung für den Nutzer übernehmen[4].
Voraussetzung für den Einsatz beider Verfahren ist die Verfügbarkeit von Kenndaten zu den archivierten Objekten, also Metadaten, denn nur aufgrund solcher technischer Informationen kann ein System in der Lage sein, gezielt mittels automatisierter Routinen Migrationsprozesse für Millionen von Objekten anzusteuern oder aufgrund einer Benutzungsanforderung eine komplexe Systemumgebung (inkl. spezieller Hardware, zum Beispiel einer Graphikbeschleunigerkarte von 1997) softwarebasiert so nachzubilden, dass das angeforderte Dokument

[4] <http://www-5.ibm.com/nl/dias/resource/uvc.pdf>.

authentisch in seiner historischen Umgebung betrachtet werden kann.

7. Mit „technischen Metadaten" sind generell Informationen gemeint, die für jedes archivierte Objekt die notwendige Darstellungsumgebung, Formatinformationen und Umstände der Entstehung sowie durchgeführte verändernde Eingriffe wiedergeben. Zur Definition der notwendigen Metadatenelemente wird derzeit eine intensive internationale Diskussion geführt, die wichtigsten Ansätze seien aus Platzgründen nur summarisch genannt:

– OCLC/RLG Preservation Metadata Working Group[5]

Hier entstand ein Rahmenkonzept für Metadaten zum Zwecke der Langzeiterhaltung.

– Fortgeführt durch PREMIS (PREservation Metadata: Implementation Strategies)[6]

Hauptziele sind die Schaffung eines implementierbaren Kerns von Langzeitarchivierungsmetadaten mit größtmöglicher Anwendbarkeitsbreite. Inzwischen liegt eine erste schriftliche Grundlage als Pre-review vor, die aber noch nicht öffentlich zugänglich ist.

– Metadata Encoding and Transmission Standard (METS)[7]

METS ist ein von der Digital Library Federation (DLF) geförderter XML-basierter Standard zur Speicherung von digitalen Objekten mit ihren Meta- und Strukturdaten. METS ist ein Containerformat für digitale Objekte und bietet die Möglichkeit, unterschiedliche Strukturen (logische, physische etc.) abzubilden sowie unterschiedliche Metadatenstandards zu berücksichtigen. Aufgrund der Flexibilität eignet sich METS dafür, einheitliche Container als submission information package (SIP) im Sinne des OAIS zu definieren.

– Metadata Standards Framework – National Library of New Zeeland[8]

Die Nationalbibliothek von Neuseeland hat 2000 ein Rahmenkonzept eines Standards für Metadaten vorgestellt. Ende 2002 entstand daraus eine konkrete Implementierung von ‚Preservation Metadata'. Diese Beschreibung von Metadaten zur Langzeitarchivierung ist momentan die konkreteste und umfassendste.

– Langzeitarchivierungsmetadaten für elektronische Ressourcen (LMER)[9]

LMER, eine Entwicklung Der Deutschen Bibliothek definiert einen Kern von relevanten technischen Metadaten zur Langzeitarchivierung und fungiert XML-basiert als Austauschformat. Es ist universell ausgerichtet, sein Schwerpunkt ist die konkrete Implementierung.

8. Persistent Identifier[10] sind eindeutige, standortunabhängige Identifikatoren für digitale Objekte, durch die gleichermaßen dauerhafter Zugriff und Zitierfähigkeit unabhängig vom Speicherort gewährleistet werden. Ihre Anwendung und Nutzung als stabile Links stellt einen wichtigen Baustein für die langfristig stabile Adressierung von digitalen Objekten, auch Depotsystemen, dar. Weitere wichtige Bausteine im Kontext der Langzeitarchivierung sind format registries, Datenbanken, die möglichst umfassend und in standardisiert erfasster und auslesbarer Form Informationen zu Dateiformaten enthalten. Diese Informationen werden kollaborativ zusammengetragen und können maschinell ausgewertet werden. Neben Namen und Versionierung sind Zeichensatz und Hinweise zu Hard- und Softwareanforderungen bedeutsam. Beispiele sind PRONOM[11] oder das Global Digital Format Registry (GDFR)[12]. Angesichts der großen Mengen an Dateien, um die es in der Langzeitarchivierung geht, werden zunehmend Tools wichtig, die Teilprozesse, wie etwa das maschinelle Auslesen von Formatinformationen aus den jeweiligen Dateien, unterstützen. Diese Ansätze werden angesichts der Komplexität der Aufgabe und der großen Quantitäten (Wieviele Dateiformate gibt es eigentlich insgesamt?) kooperativ betrieben und sind als OSS angelegt[13].

9. Es steht demnach fest, dass solche Vorhaben nur eingebunden in nationale und internationale Kooperationen sinnvoll sind. Langzeitarchivierung digitaler Publikationen ist eine komplexe Aufgabenstellung, deren Lösung nur im Zusammenwirken aller am Publikations- und Archivierungsprozess beteiligten Institutionen erreicht werden kann. Das Ziel muss die kooperative, koordinierte und nachhaltige Sicherung der digitalen Informationsversorgung sein, um die gegenwärtig und zukünftig vorhandenen digitalen Ressourcen – genauso wie bisher die analogen – authentisch und dauerhaft zur Benutzung verfügbar zu halten. Auf die Existenz verschiedener nationaler Kooperationsplattformen sei hier hingewiesen[14]. Die Bündelung dieser Aktivitäten in Deutschland liegt bei nestor[15].

10. Bei der inhaltlichen und organisatorischen Vorplanung des kopal-Projekts wurden verschiedenste technische Lösungsansätze zur Langzeitarchivierung diskutiert. Eine besonders wichtige Anforderung stellte dabei die Offenheit der Plattform und das Potential zur möglichst breiten und flexiblen Nutzung des Systems dar. Auch deshalb waren von vornherein mehrere, unterschiedliche Partner an diesem Vorhaben beteiligt. Ziel einer solchen Betrachtung muss es dabei sein, ein Rahmensystem zu definieren und weiterzuentwickeln, in dem der Bestand an zu archivierenden und langzeitverfügbar zu haltenden Objekten so zur Verfügung steht, dass Einsatzszenarien für die verschiedenen genannten methodischen Herangehensweisen erprobt und optimiert werden können.

[5] <http://www.oclc.org/research/projects/pmwg/>.

[6] Ab April soll der Entwurf öffentlich kommentiert werden können; die Diskussion kann in der DC Preservation-Liste verfolgt werden, Subskription unter <http://www.jiscmail.ac.uk/cgi-bin/wa.exe?SUBED1=dc-preservation&A=1>.

[7] <http://www.loc.gov/standards/mets/>.

[8] <http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf>.

[9] <http://www.ddb.de/standards/lmer/>.

[10] <http://www.persistent-identifier.de>.

[11] <http://www.nationalarchives.gov.uk/PRONOM/default.htm>.

[12] <http://hul.harvard.edu/gdfr/>.

[13] Ein beeindruckendes Beispiel ist JHOVE – JSTOR/Harvard Object Validation Environement, <http://hul.harvard.edu/jhove/>.

[14] Australien, das Archiv: <http://pandora.nla.gov.au/>. Australien: Das Netzwerk: <http://www.nla.gov.au/padi/>. Großbritannien: <http://www.dpconline.org/>. USA: <http://www.digitalpreservation.gov/>.

[15] NEtwork of Expertise in long-term STOrage of online Resources in Germany. Vgl. <http://www.langzeitarchivierung.de>. Siehe auch Ulrich Tiedau: nestor. In: Dialog mit Bibliotheken 16 (2004) H. 2, S. 4-10.

Als Ergebnis der Marktsichtung wurde schließlich das Digital Information Archiving System (DIAS) der Königlichen Bibliothek der Niederlande und der Firma IBM als der Erfolg versprechendste Ansatz identifiziert. Aus Platzgründen können hier keine detaillierten Bewertungen weiterer Systeme gegeben werden[16], ich beziehe mich im folgenden ausschließlich auf das DIAS. Es sei außerdem für einen systematischen Überblick auf die inzwischen erschienene Studie „Vergleich bestehender Archivierungssysteme“ verwiesen, die im Rahmen des nestor-Kompetenznetzwerks erstellt wurde[17].

11. Die Königliche Bibliothek der Niederlande (KB) hat nach einer vorangegangenen europaweiten Ausschreibung im Jahr 2000 die Zusammenarbeit mit der Firma IBM zur Entwicklung eines Depotsystems für elektronische Publikationen begonnen. Nach zweijähriger Entwicklungszeit wurde Ende 2002 die erste Version von DIAS als Komplettsystem in Betrieb genommen. Die KB sieht in diesem System die praktische Umsetzung der im europäischen Projekt „Networked European Deposit Library – NEDLIB“[18] formulierten theoretischen Grundlagen und Prozessmodelle. Erster Schwerpunkt des Einsatzes von DIAS ist die Archivierung von E-Journal-Artikeln, die aus dem bisherigen Bereitstellungssystem der KB in das Depotsystem überführt werden.
DIAS ist mit Schnittstellen ausgestattet, die eine nahtlose Einbettung in die Anwendungsumgebung der KB ermöglichen. Dies gilt sowohl für das Zusammenspiel mit den außerhalb von DIAS realisierten Recherchemöglichkeiten nach digitalen Objekten, als auch für den über eindeutige und beständige Identifikatoren realisierten Objektzugriff und schließlich ebenfalls für die datenobjektbezogenen Schnittstellen zu Import- und Exportabläufen (delivery & capture, packaging & delivery).
Bereits in der ersten produktiven Version ist die modulare Erweiterbarkeit um Funktionalitäten zur Erhaltung der Langzeitverfügbarkeit in vielversprechenden Ansätzen festzustellen. KB und IBM entwickeln diese Ansätze kooperativ weiter, wobei IBM auch seine Emulationstechnik UVC (Universal Virtual Computer) einbringt.
Aus der Perspektive der Planungen versprach das System also infolge seines strikt am OAIS-Referenzmodell orientierten Aufbaus, seiner Skalierbarkeit aufgrund der integrierten, langjährig erprobten Standard-Softwareprodukte und der absehbaren Entwicklungsperspektiven unter Beteiligung der Anwendergruppen, den Anforderungen des Projektes kopal sehr nahe zu kommen. Gleichzeitig war allerdings auch klar, dass ganz wesentliche Elemente des geforderten Depotsystems noch nicht realisiert sind.

12. Der Projektablauf von kopal sah vor, zunächst im Rahmen einer Pilotphase die vorhandene Implementierungssituation und ihren Leistungsstand systematisch zu überprüfen und auf der Grundlage dieser Detailevaluation in die Entwicklung und den Aufbau der beabsichtigten produktiven Umgebung schnell einzutreten. Diese Pilotphase wurde Anfang November 2004 erfolgreich abgeschlossen.
Die sich nun anschließende Systementwicklung innerhalb der auf 32 Monate geplanten Entwicklungsphase ist so angelegt, dass eine Ausdehnung der kooperativen Nutzung um weitere Archivbibliotheken sowie Nachnutzer aus dem Kreis aller „memory institutions“ (Archive, Museen und wissenschaftlichen Datenarchive) bereits in der Projektlaufzeit möglich ist. Strukturell wird zwischen der Kernsoftware, die von der Firma IBM erstellt wurde und von dieser aufgrund der Anforderungen der Nutzer weiterentwickelt wird (DIAS-Core), und diversen, für den Betrieb des Systems erforderlichen Tools im Umfeld der Kernkomponente unterschieden. Die Kernsoftware unterliegt dabei den üblichen Regeln der kommerziellen Softwarenutzung.
Die Grundlagenentwicklung für die Umgebungstools und die laufend erforderlichen Anpassungen und Erweiterungen auf neue Objekttypen, die in das Archiv eingehen sollen, ferner Steuerungskomponenten für das Archiv selbst, werden von den Partnern entwickelt und Interessenten im Sinne des Open-Source-Modells kostenfrei zur Verfügung gestellt.
In der Entwicklungsphase soll das System insbesondere um diese Fähigkeiten erweitert werden:
– Trennung von Hardware/Software-Betrieb und Objektzulieferung/-verwaltung

Der eigentliche Betrieb des Systems, das Vorhalten und die Betreuung der technischen Infrastruktur ist Aufgabe des Projektpartners GWDG. Auf diese Weise werden anwendungsspezifische Sonderwege reduziert und nachvollziehbares Know-how zur Systembereitstellung und zu Pflegeroutinen für unterschiedliche Partner bei einem „neutralen“ Dritten gebildet.

– Mandantenfähigkeit und remote access

Eine entscheidende Voraussetzung für die Nutzung eines Systems durch mehrere Partner ist die Sicherstellung der technischen Möglichkeit, das System weitgehend unabhängig voneinander bedienen und nutzen zu können. Hierzu gehört auch die Fernbedienbarkeit des Systems, die bislang nicht gegeben ist.

– Breite Palette an Objekttypen (Texte, Bilder, Audio-Daten, Bewegtbilder ...)

[16] In die Betrachtung wurden nur Ansätze einbezogen, die über theoretische Untersuchungen hinaus praktische und umfassende Implementierungen des OAIS-Referenzmodells nachweisen konnten. Nicht berücksichtigt wurden daher zum Beispiel Herangehensweisen, die nur die Erhaltung des Datenstroms selbst unter rein technischen Gesichtspunkten im Blick haben (z. B. um rechtlichen Ansprüchen zu genügen), Versicherungen, Banken und auch das Verbundvorhaben ARCHE und partikulare Ansätze, die z. B. langzeiterhaltungsrelevante Aspekte der digitalen Signierung bearbeiten, wie z. B. das Projekt Archi-Sig, vgl. <http://www.archisig.de/>. Folgende Systeme wurden intensiver betrachtet: ARELDA / ASTOR (<http://www.bar.admin.ch/webserverstatic/docs/e/arelda_expose_0301_e.pdf>), DSpace (MIT, USA) (<http://www.dspace.org/>), Dokumenten- und Publikationsserver (Humboldt-Universität, Deutschland) (<http://edoc.hu-berlin.de/>), DigiTool (<http://www.exlibris.co.il/dtl/index.html>), Lots Of Copies Keep Stuff Save (LOCKSS) (<http://www.lockss.org>), Flexible Extensible Digital Object and Repository Architecture (FEDORA), Cornell University/ The University of Virginia (<www.fedora.de>).

[17] Uwe M. Borghoff u. Mitarb. Univ. d. Bundeswehr München, Fak. f. Informatik, Inst. f. Softwaretechnologie: Vergleich bestehender Archivierungssysteme, Frankfurt am Main 2005 (nestor-materialien 3).

[18] <http://www.kb.nl/coop/nedlib/>. An diesem Projekt hat Die Deutsche Bibliothek aktiv teilgenommen.

Im Rahmen des Projekts ist die Integration einer großen Zahl unterschiedlicher Formate und vor allem auch Objekte wichtig. Dies dient nicht nur der echten Austestung des Systems unter Performanz- und Komplexitätsgesichtspunkten, sondern auch der Verbreiterung der Basis an angebotenen modularen Einlieferroutinen für weitere Interessenten. Hinzu kommt der grundsätzliche Erfahrungsgewinn, der gleichfalls einer Nachnutzung des Systems von vornherein einen attraktiven Rahmen gibt.

– Abdeckung unterschiedlicher Anforderungsprofile (nationalbibliographischer Hintergrund, Integration des Sondersammelgebietssystems)

Das Projekt wird durch zwei sehr unterschiedliche Bibliotheken durchgeführt. Dies betrifft nicht nur den jeweiligen Auftrag – nationalbibliothekarischer Ansatz hier – Sondersammelgebietssystem der DFG da –, sondern auch Schwerpunkte bei den bisherigen Aktivitäten – hier Online-Dissertationen, Beilage-Disketten, Netzpublikationen, dort Digitalisate, im naturwissenschaftlich-technischen Bereich gängige Datenformate. Auf diese Weise werden sehr unterschiedliche Aspekte in das Projekt eingebracht, die die Nachnutzungsmöglichkeiten für weitere Nutzer nochmals deutlich erhöhen.

13. Ausgangsbedingung für eine erfolgreiche Umsetzung dieser Absichten ist die definitive Formulierung der für das Projekt verbindlichen Objektspezifikation, in der die für die einzelnen Objekte relevanten Informationen als Metadaten niedergelegt werden.
Das XML-basierte Objektformat wird gemäß der Spezifikation des Metadata Encoding and Transmission Standard (METS) definiert. Dieses Framework erlaubt die Definition komplexer Dokumentenmodelle sowie die Integration von Metadaten und ContentFiles. Für die Objektformatspezifikation wurde daher das Framework konkretisiert und die Nutzung einzelner Elemente standardisiert. Die Festlegung der einzelnen Metadatenfelder, die im Data Management zu halten sind und auf die im Rahmen der Administration zugegriffen werden kann, basieren in größeren Teilen auf LMER (für elektronische Ressourcen).

14. Die im Projektvorhaben entstehende Plattform wird den Ausgangspunkt für eine kooperativ betriebene Langzeitarchivierung und -bereitstellung bilden, in der das national relevante elektronische Publikationsschaffen nicht nur für die heutige Nutzung, sondern auch für die Nachwelt verfügbar gehalten wird. Auf diese Weise entsteht – ausgehend von den Anforderungen Der Deutschen Bibliothek und der SUB Göttingen – eine generisch wachsende und modular erweiterbare Struktur, die über den unmittelbaren Kreis der in aller Regel öffentlich finanzierten „gedächtnis-erhaltenden" Einrichtungen hinaus auch für den Publikationsmarkt an Bedeutung gewinnt, da auch hier sich in steigendem Maße das grundlegende Problem der Langzeitverfügbarkeit stellt. Vor diesem Hintergrund, nämlich der Überführung der klassischen Aufgabenteilung zwischen Bibliotheken und Verlagen (Publikation hier – Archivierung und Bereitstellung dort) – insbesondere im Bereich der wissenschaftlichen Publikation – in das Umfeld der elektronischen Publikation, erhält das Vorhaben sein besonderes Gewicht.
Für den Nutzer des Systems ergeben sich schon heute erhebliche, konkret greifbare Vorteile: Einerseits als Zulieferer von Publikationen (direkt oder mittelbar über Institutionen) mit der Sicherheit ihrer langfristigen, zitierfähigen Bewahrung und andererseits als Anwender auf der Suche nach speziellen Informationen und direkten Zugriffsmöglichkeiten darauf. Die Archivierung in dem von kopal entwickelten Depotsystem wird als Teil des Entstehungsprozesses digitaler Ressourcen im Interesse einer transparenten, an Nutzerbedürfnissen orientierten Infrastruktur begriffen. Mit der zunehmenden Bereitschaft zur „offenen Publikation" im Sinne der „Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities"[19] steigt noch der Bedarf für einen stabilen Hintergrunddienst, der nachhaltig die langfristige Verfügbarkeit digitaler Objekte absichert.

Anschrift des Autors:

Reinhard Altenhöner
Die Deutsche Bibliothek
Adickesallee 1
D-60322 Frankfurt am Main
Tel.: +49-69-1525-1700
Fax: +49-69-1525-1799
E-Mail: altenhoener@dbf.ddb.de

[19] Berliner Erklärung: <www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>.