

Langzeitarchivierung

REINHARD ALTENHÖNER, MARKUS BRANTL, KLAUS CEYNOWA

Digitale Langzeitarchivierung in Deutschland – Projekte und Perspektiven

Foto: DNB



Reinhard Altenhöner

Foto: privat



Markus Brantl

Foto: BSB, I. Gessner



Klaus Ceynowa

Erhalt digitaler Datenbestände

Der Beitrag beschreibt die derzeit maßgeblichen, kooperativen Projekte und Konzepte der digitalen Langzeitarchivierung in Deutschland im Spannungsfeld öffentlich geförderter und kommerzieller Lösungsansätze. Im ersten Teil werden Stand und Entwicklungsperspektiven der Langzeitarchivierung in der Deutschen Nationalbibliothek vorgestellt. Im Mittelpunkt der Darstellung stehen die Erfahrungen und Ergebnisse aus Kooperationen mit kommerziellen und öffentlich-rechtlichen Partnern sowie die Bestimmung des spezifischen Beitrags der Deutschen Nationalbibliothek zur Infrastruktur der digitalen Langzeitarchivierung insgesamt. Der zweite Teil stellt den Weg der Bayerischen Staatsbibliothek mit ihrem Münchener Digitalisierungszentrum und dem Bibliotheksverbund Bayern von der explorativen zu einer auf den Regelbetrieb der Verarbeitung großer Datenvolumina angelegten digitalen Langzeitarchivierung und den damit verbundenen organisatorischen und technischen Herausforderungen dar. Es wird beschrieben, wie fortgeschrittene Digitalisierungstechnologien, zum Beispiel der Einsatz von Scanrobotik, und groß angelegte Projekte der Massendigitalisierung einerseits und die Anforderungen an Softwaresysteme zur Langzeitarchivierung andererseits sich gegenseitig bedingen.

This article describes the state of the art of leading initiatives in Germany in the field of digital long-term preservation. The first part outlines the ongoing projects on long-term preservation at the German National Library. It is focused on the experiences and results with regard to cooperations with commercial partners and publicly funded institutions and depicts the specific contribution of the German National Library to a nationwide infrastructure for digital long-term preservation. The second part describes the transition of the Bavarian State Library with its Munich Digitization Center and the Library Network of Bavaria from the exploratory project phase to the implementation of a software technology in daily operating processes for the digital long-term preservation of large amounts of data. The article shows how advanced digital technologies, for example, the use of scanning robots, and large-scale projects of mass digitization on the one hand and the demands on software systems for long-term preservation on the other hand, are mutually dependent.

EINLEITUNG

Parallel zur zunehmenden Bedeutung digitaler Informationsproduktion und Informationsnutzung hat sich seit nunmehr rund 15 Jahren der Erhalt der Langzeitverfügbarkeit digitaler Objekte als eine der Kernaufgaben von Einrichtungen der Informationsinfrastruktur herausgebildet. Insbesondere elektronische Publikationen, oft in multimedialer Form, sind heute ein integraler Bestandteil des Arbeitens in Wissenschaft und Forschung, in Bildung und Ausbildung. Sie stellen eine wesentliche Komponente im Bestand der Bibliotheken dar, wobei der Bestandsbegriff hier weiter zu fassen ist als im traditionellen Verständnis: Er geht über den physischen Besitz der Publikation auf einem Datenträger hinaus und beinhaltet auch das Recht des Zugriffs auf Netzpublikationen im Internet. Diese Publikationen sind entweder ausschließlich elektronisch produziert und publiziert (»born digital«), oder sie sind

durch Digitalisierung analoger Publikationen entstanden (Retrodigitalisate).

Die Gewährleistung der Verfügbarkeit und Nutzbarkeit digitaler Datenbestände über lange Zeiträume unterliegt Risiken, sowohl auf organisatorischer als auch auf technischer Ebene, zu nennen sind hier vor allem:

- Die begrenzte Lebensdauer einzelner Datenträger: Einerseits verfallen die Trägermedien physisch, andererseits veralten sie, mit der Konsequenz, dass oft keine geeigneten Abspielumgebungen mehr zur Verfügung stehen.
- Die Gefährdung der Lesbarkeit und speziell der Interpretierbarkeit der Daten durch die permanente Fortentwicklung der Dateiformate und der zugehörigen Anwendungsprogramme.
- Die Flüchtigkeit von Netzpublikationen, die häufig bereits nach kurzer Zeit nicht mehr auffindbar oder nur noch unter einer veränderten Internetadresse erreichbar sind.
- Die zeitlich nur begrenzte Verfügbarkeit insbesondere elektronischer Zeitschriften aufgrund befristeter Lizenzverträge mit den anbietenden Verlagen.

Die Gesamtheit aller Maßnahmen zum Erhalt eines digitalen Datenbestandes und seiner Nutzbarkeit über lange Zeiträume hinweg nennt man digitale Langzeitarchivierung. Diese Maßnahmen umfassen die gezielte Auswahl des archivwürdigen Materials, den Transfer in ein digitales Archiv, die geeignete Erschließung und persistente Identifizierbarkeit der Objekte, die Archivspeicherung, die Durchführung der Langzeiterhaltungsmaßnahmen (Preservation, Planning and Action), die Bereitstellung der digitalen Daten für die Nutzung sowie die Migration (Konvertierung veralteter Dateiformate in aktuelle, standardisierte Formate) und die Emulation (Nachbildung alter Betriebssysteme und Anwendungsprogramme auf neuen Rechnern), die dem Erhalt der Daten selbst und ihrer Interpretierbarkeit dienen.¹

In Deutschland sind vor allem die Aktivitäten der Deutschen Nationalbibliothek (DNB) in Kooperation mit der Staats- und Universitätsbibliothek (SUB) Göttingen einerseits und der Bayerischen Staatsbibliothek (BSB) und des Bibliotheksverbundes Bayern (BVB) andererseits zu nennen, die sich in ihrer Ausrichtung und ihrer Entwicklung unterscheiden. Im Folgenden werden die Initiativen dieser Einrichtungen und ihre Zu-

kunftsplanungen auf dem Feld der digitalen Langzeitarchivierung vorgestellt und ihr Beitrag zu einer verteilten Langzeitarchivierungs-Infrastruktur auf regionaler, nationaler und internationaler Ebene diskutiert.

1. LANGZEITARCHIVIERUNG AN DER DEUTSCHEN NATIONALBIBLIOTHEK

Das BMBF-Projekt *kopal* als Basis für die digitale Langzeitarchivierung in der DNB

Nach einigen Vorläuferaktivitäten vor allem im europäischen Kontext – wegweisend hier das NEDLIB-Projekt mit seinen spezifischen konzeptionellen Ergänzungen des noch immer maßgeblichen und aktuell in der Revision befindlichen ISO-Standards »Reference Model for an Open Archival Information System« (OAIS) – unternahm die DNB 2003 konkrete Schritte zur Implementierung eines digitalen Langzeitarchivs. Dieses Langzeitarchiv sollte das abstrakte OAIS-Modell umsetzen und die nötigen technischen Voraussetzungen für die Langzeitverfügbarkeitsstrategien der Migration und Emulation zur Verfügung stellen. Während bei der Migration digitale Objekte in Formate überführt werden, die in aktuellen Ablaufumgebungen anzeigbar sind, wird bei der Emulation eine historische Ablaufumgebung auf einer aktuellen Rechnerplattform technisch so weitgehend simuliert, dass das alte Format anzeigbar und nutzbar wird.

Nach einer intensiven Marktsichtung wurde das von der niederländischen IBM für die Nationalbibliothek der Niederlande entwickelte DIAS-System als diesen Anforderungen entsprechend ausgewählt. Das DIAS-System basiert auf Standardsoftware-Komponenten der IBM, die weltweit vieltausendfach im Einsatz sind, angereichert um spezifische Funktionen. Allerdings waren sich die DNB und die SUB Göttingen, die für die weiteren Projektschritte eine Kooperation eingingen, darüber im Klaren, dass DIAS noch funktionale Defizite vor allem in zwei Bereichen aufwies:

- Service- und Betriebskonzept: Das zukünftige System sollte von verschiedenen Partnern genutzt werden können und auch in einer verteilten Betriebsinfrastruktur funktionieren. Das erforderte eine Abkehr von dem »Inhouse«-Modell der niederländischen Implementierung, in dem die Speicherung, Prozessierung und Zugänglichmachung von Daten in einem geschlossenen System an einem Standort stattfindet.
- Außerdem sollten die verfügbaren Metainformationen zu archivierten Objekten so reichhaltig sein, dass sie für gezielte Migrationsaktivitäten auch im Massenverfahren differenziert genutzt werden können. Diesem Ansatz liegt die Einsicht zugrun-

de, dass nur technisch validierte Objekte und stabile Informationen zu ihrem jeweiligen Status den Zugriff dauerhaft sicherstellen können. Die sichere Speicherung der Daten selbst, die naturgemäß die entscheidende Grundlage jeder Aktivität im Bereich der Langzeitarchivierung darstellt, bildet nur eine wesentliche Voraussetzung für erfolgreiche Maßnahmen zur Gewährleistung der langfristigen Verfügbarkeit der digitalen Objekte. Möglich werden solche Aktivitäten aber erst durch die Kenntnis der technischen Eigenschaften der archivierten Objekte, die entsprechend angereichert sein müssen. Auf dieser Basis können Erhaltungsprozesse (wie etwa die Migration aus einem gefährdeten Format in ein aktuell als sicher betrachtetes Format) maschinell angestoßen werden.

2004 wurde dann das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte, dreijährige Forschungsprojekt »Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen« (*kopal*) initiiert. Ziel des Projekts war die Entwicklung und praktische Erprobung eines kooperativ nutzbaren Modells für ein digitales Langzeitarchiv.² Die kooperative Nutzung bezog sich dabei auf das separierte Einspielen und Abrufen von Datenbeständen unterschiedlicher Institutionen unter Nutzung einer gemeinsamen Infrastruktur. Im Ergebnis sollte jede archivierende Institution nur ihre zuvor eingespielten Daten abrufen können, analog zu einem Bankschließfach. Diese sogenannte »Mandantenfähigkeit« ermöglicht einerseits institutionell unterschiedliche Festlegungen (es gibt z. B. unterschiedliche Migrationspfade), eröffnet aber andererseits Synergiepotentiale, weil Systemaufwendungen nur einmal getätigt werden müssen.

Partner des Projekts waren neben der Deutschen Nationalbibliothek, die die Gesamtleitung hatte, die SUB Göttingen, die IBM Deutschland und die Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG). Aufgabe von IBM war es, das fest in der IT-Infrastruktur der niederländischen Nationalbibliothek integrierte DIAS-System in seinem Kern herauszulösen und mit transparenten Schnittstellen zu versehen. Das System selbst wurde bei der GWDG als Rechenzentrumsprovider installiert, der für den stabilen Betrieb und insbesondere den Erhalt der Daten (z. B. durch regelmäßige Konsistenzprüfungen und Umkopiermaßnahmen) sorgt. Um das System an die unterschiedlichen Infrastrukturen und Workflows der DNB und der SUB Göttingen (sowie potenzieller weiterer Partner) anbinden zu können, wurde zudem eine flexible und konfigurierbare Softwareumgebung benötigt. Dazu entstand im Projekt unter dem Namen »kopal Library for Retrieval and Ingest« (koLibRI) eine Open

Erhaltungsprozesse

DIAS-System

transparente Schnittstellen

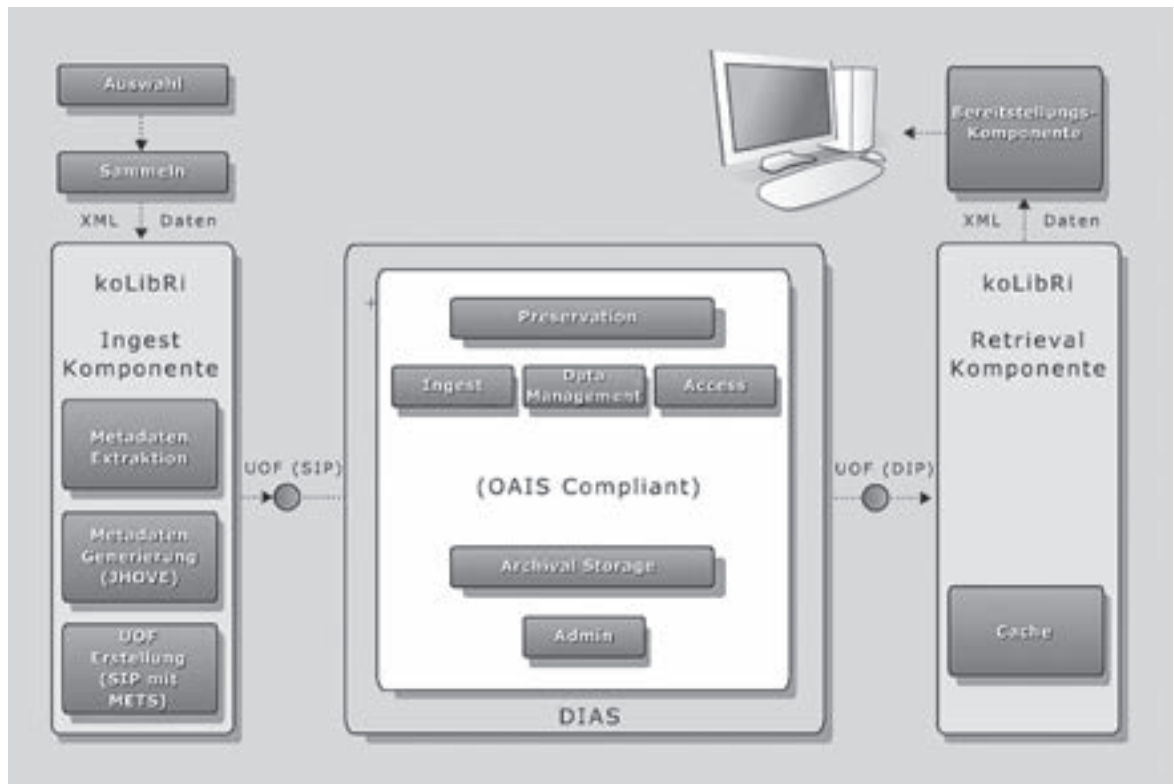


Abb. 1: kopal-Gesamtarchitektur

Source-Software-Bibliothek, deren Aufgabe die Steuerung des Ingests ist (Aufbereitung und Validierung der Objekte, Extraktion technischer Metadaten, Zugriffssteuerung).³ Die von der IBM verantwortete Weiterentwicklung von DIAS bezog sich über die Änderungen der Gesamtarchitektur hinaus auf die Unterstützung des erweiterten Metadatenmanagements (Migrationsunterstützung) und die Standardisierung der Einlieferungschnittstelle für Archivpakete und der Administrationschnittstelle generell.

Die Funktionsweise des Gesamtsystems ist dadurch gekennzeichnet, dass den archivierten Dateien technische Informationen mitgegeben werden, die über eine Datenbank für spätere Migrationsschritte ausgewertet werden können. Wenn beispielsweise alle Objekte im Archiv, die im Format PDF in der Version 1.3 codiert sind, in eine neuere Version konvertiert werden müssen, weil die Version 1.3 von gängigen Repräsentationsumgebungen nicht mehr unterstützt wird, müssen alle Objekte, für die diese Bedingung zutrifft, identifiziert und vom System bereit gestellt werden. Extern erfolgt dann die Konversion der Daten und die neue Version wird wieder in das Archiv zurückgespielt. Ob die alte Version ersetzt wird oder aus Gründen der Sicherheit und der Transparenz erhalten bleibt, ist eine Frage, die der Archivar entscheidet. Fällt die Entscheidung zu Gunsten des Erhalts der Ausgangsversion, wird diese ebenfalls gespeichert. In

beiden Fällen muss das System in der Lage sein, die durchgeführte Operation – in unserem Fall also die Migration – zu dokumentieren und den neuen Informationsstand zur neuen Versionsstufe vorzuhalten. Das im DIAS verwendete Metadatenformat ist das »Universal Object Format«, das in einer Dokumentation öffentlich zur Verfügung steht.⁴

DIAS ist also der Kern der Lösung, aber nicht das komplette Archivsystem für eine einzelne Bibliothek oder andere Kultureinrichtung. Erst im Zusammenspiel mit zahlreichen weiteren Komponenten, zum Beispiel spezifischer Lieferschnittstellen für Verlage, Recherche- und Präsentationssysteme der Bibliotheken für ihre Endnutzer sowie differenzierter Auslieferungsstellen, entsteht funktional das vollständige Langzeitarchivsystem. Die Entwicklung im Projekt kopal umfasste insofern nur einen Baustein der Langzeitarchivierungs-Infrastruktur einer Gedächtnisorganisation, nämlich die eigentliche Archivierung und das dazu gehörige Metadatenmanagement. Demgegenüber erfüllt die koLibRI-Softwarebibliothek Aufgaben wie die sichere exklusive Kommunikation mit DIAS, die Erzeugung der Archivpakete inklusive der technischen Metainformationen und die Anbindung an vorhandene Datenspeicher wie lokale Repositories und Bereitstellungssysteme. Außerdem deckt koLibRI mit weiteren Komponenten auch die Implementierung von Migrationsläufen ab.

»Universal Object Format«

Lieferschnittstellen

Implementierung von Migrationsläufen

Mitte 2007 lief das Projekt *kopal* mit der Unterzeichnung eines Memorandums aus, das die weitere Zusammenarbeit der Projektpartner regelte und vor allem den Willen aller Beteiligten zur Fortführung der Entwicklungsarbeiten dokumentierte. Aufgrund technischer Umsetzungsprobleme mit einer Softwareerweiterung von DIAS waren allerdings deutliche Verzögerungen bei der Überführung des Gesamtsystems in den regulären Betrieb unvermeidlich. 2009 erfolgten abschließende Tests und 2010 die finale Abnahme der neuen Version auf dem voll funktionalen Testsystem. Gegenwärtig wird auf einer neuen erweiterten Produktivumgebung die Einbindung in die Einspielaufläufe der DNB für abgelieferte Objekte vorbereitet.

Insgesamt hat sich gezeigt, dass die Aufteilung von Zuständigkeitsbereichen mehrerer Kooperationspartner und die Verständigung über technische Standards zwar aufwändig ist, aber in der Praxis funktioniert. Unstrittig vergrößert sich der Bedarf, Übergabeschnittstellen technisch und organisatorisch sehr bewusst zu gestalten und Möglichkeiten zur Absicherung und Abnahme erfolgreicher Übergaben zu vereinbaren. Genau hier liegt aber auch ein wesentlicher Vorteil einer derart verteilt arbeitenden Struktur, da hier Qualitätssicherungsmaßnahmen explizit als Bestandteil einer Prozesskette vereinbart werden. Allerdings hat sich auch gezeigt, dass wechselseitige Abhängigkeiten entstehen, bei denen das jeweils »schwächste Glied« zu einem hemmenden Faktor werden kann. An drei Aspekten soll dies hier konkret erläutert werden:

► Während in der Außenperspektive zu Beginn des Projekts die Standardsoftware-Komponenten der IBM als kritischer Faktor vor allem auch unter Kostengesichtspunkten diskutiert wurden, erwies sich in der Projektlaufzeit und danach vor allem die DIAS-Entwicklung selbst als Problem. Es gelang über das Projekt hinaus nicht, eine weltweite Präsenz von DIAS mit vielen Installationen aufzubauen. Der Status eines offiziellen Produkts der IBM wurde nicht erreicht, sondern DIAS blieb ein »Asset«, das heißt eine kundenspezifische Entwicklung, die vom Interesse des jeweiligen Kunden gesteuert wird. Natürlich steht die IBM im Bedarfsfall mit ihrem Know-how zur Verfügung, und die üblichen Qualitätssicherungsmechanismen greifen, wenn es Probleme gibt. Selbstverständlich wurde auch ein entsprechender Wartungsvertrag abgeschlossen. Ohne Zweifel aber entwickelt sich abhängig von der Menge und dem Volumen derartiger Anforderungen das Know-how der Firma auf diesem Gebiet nicht mehr weiter, ja droht zu versickern. Ein Asset erfährt seine Weiterentwicklung primär nur dann, wenn der Kunde entsprechende

Anforderungen formuliert. Im Gegensatz zu einem vollwertigen Produkt gibt es kein Eigeninteresse der Firma, das Produkt weiterzuentwickeln und eine Ausdehnung am Markt anzustreben. Entsprechend wird es schwieriger, mit der generellen technischen Entwicklung Schritt zu halten. Verliert also der kommerzielle Partner aus welchen Gründen auch immer sein Interesse an der wirtschaftlichen oder technischen Weiterführung einer Software, entsteht potentiell eine riskante Situation.

► Eine wichtige Aufgabe im Projekt lag in der Etablierung eines Geschäftsmodells für die weitere Verbreitung der Softwarelösung und der Präzisierung eines Vertriebs- und Servicemodells für *kopal*-Kunden. Dabei hat sich gezeigt, dass die Formulierung eines »Business Case« aus der Konstellation eines Forschungs- und Evaluierungsprojektes heraus schwierige und enorm zeitaufwändige Klärungsprozesse innerhalb des Konsortiums und insbesondere innerhalb der weltweit agierenden Firma IBM auslöste. Angesichts der noch sehr offenen Marktsituation 2006/2007 für Systeme zur OAIS-kompatiblen Langzeitarchivierung kam diese Anforderung möglicherweise für alle beteiligten Partner zu früh. Hinzu kam die geforderte Verknüpfung eines kommerziellen Softwaremodells mit den Ansprüchen eines Open-Source-Softwareprojekts, die sich als nicht immer leicht vermittelbar herausstellten. Das letztendlich erarbeitete Modell war ambitioniert, ging aber an den Anforderungen der Kunden insbesondere aufgrund seiner Komplexität vorbei.⁵

► Unterschiedliche Eigeninteressen einzelner gleichberechtigter Partner können grundsätzlich zu Interessenskonflikten führen, die sich hemmend auf ein System und seine Weiterentwicklung auswirken. Um dies zu vermeiden, wurde das DIAS-System insbesondere im Bereich des Metadatenmanagements bewusst sehr generisch ausgelegt. Auf diese Weise können auch unterschiedliche Vorstellungen zur Bedeutung bestimmter Informationen hinterlegt werden (beispielsweise sah die SUB Göttingen Bedarf dafür, auch im Archiv ein basales bibliographisches Metadatenset mitzuführen; dieser Ansatz wurde realisiert, ohne dass die DNB ihm gefolgt wäre). Ähnliches gilt für den modularen Ansatz von koLibRI, der – der Bedarfslage einzelner Partner folgend – die Implementierung unterschiedlicher, technisch voneinander getrennter Features erlaubt, die dann der Entwicklergemeinschaft wieder zur Verfügung stehen. Allerdings entstehen Zwänge zur Einigung hinsichtlich Betriebsinfrastruktur und Investitionsbedarf, und hier zeigen sich auch die Grenzen eines vollständig

kommerzielle Partner

Übergabeschnittstellen

Generik

kooperativen Modells: Für die Nationalbibliothek existiert ein gesetzlicher Auftrag, der zur Herstellung einer produktiven, performant laufenden Infrastruktur zwingt; für die SUB Göttingen gibt es einen derartigen Auftrag zumindest nicht institutionell, sondern nur eingebettet in den Hochschulkontext insgesamt. Dies führt zu vermehrtem Abstimmungsbedarf und zwingt unter Umständen auch zur Reorganisation des Kooperationsmodells.

Folgerungen und Konsequenzen im Sinne einer Zwischenbilanz

Risikoanalyse

Die Arbeit im *kopal*-Projekt zeigt zunächst, dass die Abhängigkeit von einem kommerziell motivierten Partner in die Risikoanalyse vorab sehr bewusst einbezogen werden muss. Verliert die Firma das Interesse an diesem Marktsegment, müssen entsprechende Absicherungen und valide Auffanglösungen existieren. Diese können vertraglicher und funktionaler Art sein, im Fall von *kopal* wurde dies durch die Offenlegung aller relevanten Spezifikationen und die Festlegung einer »Notentladeschnittstelle« abgesichert. Die Open-Source-Software-Komponenten werden in einer gemeinsamen Entwicklungsplattform gehalten und öffentlich gestellt, ein kooperativ getragener Bissupport ist eingerichtet. In jedem Fall aber muss die langzeitarchivierende Einrichtung ihr Know-how angemessen weiterentwickeln. Dies kann bis zur Offenlegung des Software-Codes reichen.

Eine konkrete Konsequenz für die DNB war, dem in der internationalen Diskussion formulierten Grundsatz, aus Sicherheitsgründen in zwei unterschiedlichen Systemwelten zu archivieren, zumindest insofern zu folgen, als die Anforderungen an die Qualität des Speicher- und Archivierungsmanagements innerhalb der Einrichtung selbst deutlich gesteigert wurden (Redundanz, lokale Verteilung, Identifizierungssysteme). Dieser noch laufende Ausbau der technischen Infrastruktur wäre ohne den Know-how-Gewinn aus der intensiven fachlichen Beschäftigung mit den erforderlichen Technologien im Rahmen des *kopal*-Projekts nicht möglich gewesen.

Grundsätzliche Konsequenzen zur Herangehensweise an Vorhaben im Bereich der Langzeitarchivierung ergeben sich aus der Perspektive der DNB insbesondere für die Softwarearchitektur und die darauf aufsetzenden Services. Systemanforderungen wie die Modularität der einzelnen, für sich selbstständigen Komponenten und die Entkoppelung der Services voneinander mit dem Ziel, sie flexibel auch in unterschiedlichen Umgebungen einsetzen zu können, sind maßgebliche Bausteine für den Erfolg. Diese Erkenntnisse sind auch in eine europäische Arbeitsgruppe eingebracht wor-

den, die ihre Anforderungen an ein Langzeitarchivsystem der nächsten Generation zusammengestellt hat, die auch für die Nationalbibliothek den Ausgangspunkt für eine sicherlich in der Zukunft anstehende Ablösung der DIAS-Schicht bilden. Dabei kommt der DNB zu Gute, dass DIAS nur eng begrenzte und definierte Funktionen in einer Gesamtstruktur übernimmt, so dass der Austausch der betreffenden Komponenten (im Wesentlichen des Metadatenmanagements) eine handhabbare Größenordnung darstellt.⁶

Die technische Weiterentwicklung von Verfahren der Langzeitarchivierung ist auf allen Stufen zwingend erforderlich. Viele Fragestellungen beim Umgang mit komplexen digitalen Objekten (eingebettete Dateien, Multimedia-Daten, verteilt liegende Informationen, lebende Dokumente, aggregierte Forschungsdaten, Datenbanken, interaktive Objekte, eigenständige Applikationen) sind nicht gelöst. Vor diesem Hintergrund sucht die DNB Möglichkeiten zur anwendungsbezogenen Weiterentwicklung von Wissen und Werkzeugen, etwa durch Teilnahme an zahlreichen nationalen und europäischen Projekten und durch eigene Aktivitäten. Neben technischen Fragestellungen sind es organisatorische, planerisch-konzeptionelle und rechtliche Themen sowie die Standardisierung und der Aufbau einer Zertifizierungsinfrastruktur, die die Arbeit prägen. Vor diesem Hintergrund spielt in der Weiterentwicklung des Arbeitsfeldes Langzeitarchivierung die Schaffung einer Koordinations- und Vernetzungsinfrastruktur auf nationaler und internationaler Ebene eine herausragende Rolle. Konkret manifestieren sich diese Aktivitäten im Kompetenznetzwerk nestor und in europäischen Vernetzungsinitiativen.⁷

Das DP4lib-Projekt (Digital Preservation for Libraries)

Ein kooperatives, arbeitsteiliges Vorgehen bildet das entscheidende Erfolgsmerkmal im Umgang mit dem komplexen Handlungsfeld der digitalen Langzeitarchivierung. Die Projekterfahrungen der DNB zeigen, dass eine wesentliche Aufgabe – neben den technischen und organisatorischen Fragestellungen – in der Entwicklung einer auch unter Kosten- und Servicegesichtspunkten kalkulierbaren Dienstestruktur liegt, die gerade solchen Partnern angeboten werden kann, die einen umfassenden Einstieg in die digitale Langzeitarchivierung nicht anstreben, gleichzeitig aber entsprechende Aufgaben – etwa auf Grund einer Vorgabe ihres Unterhaltsträgers – wahrnehmen müssen.

Diesem Ziel ist das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt »DP4lib – Digital Preservation for Libraries« verpflichtet. Basis des Vorhabens sind dabei die Ergebnisse des Projekts

Speicher- und Archivierungsmanagement

Modularität

kopal.⁸ Aufgrund seiner flexiblen Servicestruktur bietet *kopal* ein solides Fundament für die organisatorische und technische Integration institutioneller Teilnehmer. *kopal* kennt neben dem Betrieb einer eigenen kompletten Langzeitarchivierungslösung in einer definierten institutionellen Umgebung die Nachnutzung des schon bestehenden Systems durch sogenannte »Mandanten« und »Teilnehmer«. Mandanten »mieten« ein eigenes »Schließfach«, also einen separaten, virtuell abgetrennten Bereich mit eigenen Verwaltungsmöglichkeiten im *kopal*-System. Damit geht die Verpflichtung einher, die laufende Pflege und Erhaltung der eingebrachten Daten eigenverantwortlich zu betreiben. *kopal*-Teilnehmer hingegen überlassen die genannten Aufgaben einem *kopal*-Mandanten, der im Verbund mit Anderen die genannten Services rund um die digitale Langzeitarchivierung anbietet, individuell anpasst und die Teilnehmer umfassend begleitet und berät. Zwischen diesen Modellen sind in der Praxis unterschiedliche Realisierungsformen denkbar, die individuell nach Anforderung und Leistungsvermögen der jeweiligen Partner ausgehandelt werden können.

Notwendig ist es zunächst, die durch die verschiedenen Partner (Bibliotheken, Verbundzentralen, Dokumentationszentren etc.) gegebenen unterschiedlichen Nutzungsszenarien zu evaluieren, auf ihre technische Abbildbarkeit hin zu untersuchen und dann in modularer Weise so umzusetzen, dass die entstehenden Dienste nahtlos in die jeweilige Infrastruktur eingebunden werden können. Neben technischen Anforderungen, die die Systemkommunikation zur Unterstützung automatisierter Abläufe im Blick haben, stellen sich auch eine ganze Reihe von organisatorischen Fragen, die bis zur Entwicklung geeigneter Geschäftsprozesse und Geschäftsmodelle reichen. Vor diesem Hintergrund ist im Projekt vorgesehen, ein generisches, auch in anderen Kontexten nachnutzbares Vorgehensmodell zu entwickeln. Dies steht dann als konkreter Handlungsleitfaden für alle Informationsinfrastruktur-Einrichtungen zur Verfügung, die vor der Herausforderung stehen, die Langzeitverfügbarkeit ihrer digitalen Sammlungen sicherzustellen. Wesentliche Arbeitsmodule des Projekts auch im Sinne der Nachnutzbarkeit sind:

- Identifikation, Spezifizierung und Anpassung sämtlicher notwendiger Geschäftsprozesse der Langzeitarchivierung inklusive aller administrativen Funktionen: Hier ist die Balance zwischen den individuellen Ausprägungen und den modellbildenden, allgemein gültigen Abläufen zu finden.
- Aufbau eines lückenlosen Berichtswesens und eines funktionsfähigen Controllings über die durchgeführten Aktivitäten der Langzeitsicherung.

- Aufbau eines tragfähigen Betriebs- und Kostenmodells für die angebotenen Dienste zur digitalen Langzeitarchivierung.
- Qualitätssicherungsmaßnahmen zur Gewährleistung des Erhalts der Dateiintegrität über einen prinzipiell unbegrenzten Zeitraum.
- Konkretisierung und Durchführung von Migrationsläufen im Rahmen des Preservation Planning.
- Ausformulierung rechtlich-vertraglicher Regelungen im Sinne einer Mustervereinbarung.

Das übergreifende Ziel der dargestellten Projektinitiativen der DNB ist es, ein arbeitsteiliges, kooperativ zusammenwirkendes technisches Netzwerk für die Langzeitarchivierung aufzubauen, das Synergien nutzt, Aufwände ressourcenschonend verteilt, allen Beteiligten angemessene Partizipationschancen bietet und gleichzeitig und vor allem die – zunehmend multimedialen – Bestände und ihre Verfügbarkeit dauerhaft sichert. Die Langzeitarchivierung kann, so die Projekterfahrung der DNB, immer weniger als ein monolithischer Block verstanden werden, der als geschlossenes Softwaresystem in den Kontext einer Einrichtung transferiert und gegebenenfalls angepasst wird. Diesem Problem ist nur mit modularen Softwarebausteinen, die transparent verknüpft werden, zu begegnen. Locker miteinander gekoppelt, können diese Werkzeuge flexibel in sich wandelnde Abläufe eingepasst und so ein sehr viel höheres Maß an Produktivität und Arbeitsteilung erreicht werden. Dies gilt insbesondere für die Automatisierung von Abläufen, die in der DNB verschiedene Prozessabläufe insbesondere in der Selektion und Erschließung digitaler Objekte erfassen. Nach dem Verständnis der DNB wird die Langzeitarchivierung in Zukunft eine solche Koppelung unterschiedlicher technischer Services voraussetzen. Ihre Nutzung und Integration in die betrieblichen, zunehmend automatisierten Abläufe in einer Einrichtung bedarf eines erheblichen technologischen Know-hows. Vor diesem Hintergrund und zur Minimierung des Risikos einseitiger Abhängigkeiten von einzelnen Partnern wird seitens der DNB die Strategie einer verteilten Zuständigkeit verfolgt bei Sicherstellung der Transparenz und Zugänglichkeit aller Verfahren bis hinunter zur Ebene des Softwarecodes.

Netzwerk für die Langzeitarchivierung

Nutzungsszenarien

wesentliche Arbeitsmodule

2. LANGZEITARCHIVIERUNG AN DER BAYERISCHEN STAATSBIBLIOTHEK

Digitalisierung und digitale Langzeitarchivierung als landesweite Kooperationsaufgabe

Für die Bayerische Staatsbibliothek (BSB) als zentrale Landes- und Archivbibliothek des Freistaates Bayern und als eine der international bedeutendsten Uni-

versal- und Forschungsbibliotheken zählt die dauerhafte Sicherung und Nutzbarmachung ihrer rasch wachsenden digitalen Datenbestände zu den gegenwärtig und zukünftig wichtigsten Herausforderungen. Bereits 1999 hat sie begonnen, in mehreren explorativen Projekten zur Langzeitarchivierung die Anforderungen zu identifizieren, Lösungen zu konzipieren, konkrete Erfahrungen mit deren Umsetzung zu sammeln und Kooperationen auf nationaler und internationaler Ebene einzugehen. Aufgabenschwerpunkte sind zum einen die Koordinierung, Sicherung und laufende Optimierung der Prozesse zur Langzeitarchivierung des Bestandes der BSB und der bayerischen wissenschaftlichen Bibliotheken, zum anderen die Beteiligung an Forschung und Entwicklung zur Langzeitsicherung. Konkret übernimmt die BSB Verantwortung für die Langzeitarchivierung von digitalen Informationsobjekten aus verschiedenen Bereichen. Das Sammlungs- und Archivierungsprofil umfasst:

- Eigene sowie in Kooperationsprojekten erstellte Retrodigitalisate; hier verfügt die BSB mit rund 470.000 digitalisierten Werken über den derzeit größten digitalen Bestand aller Bibliotheken im deutschsprachigen Raum (Stand Dezember 2010).
- Publikationen von bayerischen Verlagen (in Vorbereitung auf ein novelliertes Pflichtstückegesetz).
- Bayerische amtliche Veröffentlichungen (aufgrund des zum 1. Januar 2009 in Kraft getretenen Erlasses für die Abgabe amtlicher Veröffentlichungen an Bibliotheken).
- Lizenzpflichtige elektronische Medien (Zeitschriften, E-Books, Datenbanken) in lokalen, konsortialen und Nationallizenzen.
- Elektronische Medien in den Sondersammelgebieten aus nationaler und internationaler Produktion, insbesondere Internetressourcen.

Das komplexe Themenfeld der digitalen Langzeitarchivierung wird hierbei in kooperativem Zusammenwirken zwischen dem Münchener Digitalisierungszentrum der BSB, dem Referat »Virtuelle Bibliothek Bayern« der Verbundzentrale des Bibliotheksverbundes Bayern (BVB) und dem Leibniz-Rechenzentrum bearbeitet. Die Grundidee hierbei ist, dass jeder Partner genau das tut, was er am besten kann. Während das Leibniz-Rechenzentrum die Speicherung und langfristige Archivierung der digitalen Daten im eigentlichen Sinne und die regelmäßig durchzuführenden Erhaltungsmaßnahmen im Bereich der Speicherschicht übernimmt, können sich die Fachabteilungen der BSB und des BVB auf die organisatorischen, funktionalen und fachlichen Aspekte der digitalen Langzeitarchivierung konzentrieren. Die konkrete Rollenverteilung gestaltet sich hierbei folgendermaßen:

- Das Münchener Digitalisierungszentrum / Digitale Bibliothek (MDZ) wurde 1997 als Abteilung der BSB mit einer Anschubfinanzierung der DFG als eines von zwei nationalen Kompetenzzentren für die Digitalisierung von Bibliotheksbeständen gegründet. Es ist heute die zentrale Innovations- und Produktionseinheit der BSB für die Entwicklung, Erprobung und Inbetriebnahme neuer Produkte und Prozesse rund um das facettenreiche Thema »Digitale Bibliothek«. Seit seiner Gründung hat das MDZ mehr als 200, nach Inhalt und Organisation höchst unterschiedliche, explorative Projekte erfolgreich umgesetzt – von der Retrodigitalisierung von historischen Drucken und Bildmaterialien über Audio, 3D⁹, Electronic Publishing, den Aufbau virtueller Fachportale bis hin zur digitalen Langzeitarchivierung. Dabei wurden neue Verfahren und Methoden für die Erstellung, Verwaltung, Speicherung, Langzeitarchivierung und Bereitstellung digitaler Objekte im Internet entwickelt. Des Weiteren übernimmt das MDZ übergreifende Aufgaben als nationaler Serviceprovider, z. B. für Digitalisierungstechnologien und für Entwicklungsaufgaben im Rahmen der Open-Source-Community im Bereich »Digitale Bibliothek«.
- Die Verbundzentrale des Bibliotheksverbundes Bayern (BVB) ist organisatorisch eine Abteilung der Bayerischen Staatsbibliothek und fungiert als IT-Kompetenzzentrum für alle verbundweiten Fachanwendungen. Der zunehmenden Bedeutung verbundweit aufzusetzender digitaler Angebote und Services wurde im Jahre 2006 durch die Einrichtung des Referats »Virtuelle Bibliothek Bayern« Rechnung getragen. Nicht zuletzt aufgrund der begrenzten IT-Ressourcen an den einzelnen Bibliotheken und Hochschulrechenzentren hat sich die Nachfrage nach zentraler, verbundweit nutzbarer Infrastruktur für die Speicherung und Präsentation digitaler Objekte in den zurückliegenden Jahren signifikant erhöht. Darüber hinaus erzeugt das rasche Wachstum der digitalen Ressourcen der Hochschulbibliotheken (Retrodigitalisate, Repositories, Forschungsprimärdaten etc.) gegenwärtig einen wachsenden Handlungsdruck in Richtung einer bayernweiten Infrastruktur zur digitalen Langzeitarchivierung.
- Das Leibniz-Rechenzentrum (LRZ) ist das Rechenzentrum für die Münchener Universitäten und die Bayerische Akademie der Wissenschaften. Es ist für das Münchener Wissenschaftsnetz verantwortlich und stellt Höchstleistungsrechner für die Wissenschaft in Deutschland und international zur Verfügung. Insbesondere betreibt es zentrale Speicher-

systeme mit Online-Plattenspeichern (Network Attached Storage und Storage Area Network) und robotergestützten Magnetbandsystemen zur Datensicherung und Archivierung.¹⁰ Für Bibliotheken und Bibliotheksverbände sind der Aufbau und der nachhaltige Betrieb einer professionellen Speicher- und Archivierungsumgebung in den künftig absehbaren Größenordnungen wirtschaftlich und technisch nur schwer realisierbar. Die Kooperation mit dem LRZ ermöglicht hier aufgrund seiner Expertise im Betrieb von großen Onlinespeichern und seiner skalierbaren, robotergesteuerten Archivierungs- und Backupumgebung beträchtliche Synergiegewinne gerade im Bereich der digitalen Langzeitarchivierung.

— Von der »Boutique«-Digitalisierung zur Massendigitalisierung

Die Erschließung, Bewahrung und Zugänglichmachung des ihr anvertrauten, über die Jahrhunderte aufgebauten schriftlichen Kulturerbes gehört zu den Kernaufgaben der BSB. Unter den Handlungsbedingungen des Internets, das den Zugang zu Informationen grundsätzlich zu jedem Zeitpunkt und von jedem Ort aus ermöglicht, leitet sich hieraus für die BSB das strategische Ziel ab, ihre einzigartigen, historischen Bestände für den schnellen, zeitunabhängigen und ubiquitären Netzzugriff kostenfrei bereitzustellen. Bei Einbeziehung aller urheberrechtsfreien Bestandsegmente bedeutet dies die Digitalisierung von rund 1,2 Millionen Werken. Die Finanzierung dieses ehrgeizigen Vorhabens erfolgt derzeit primär aus Drittmitteln der DFG, des Freistaates Bayern und der Europäischen Union, vor allem im Bereich Handschriften, Alte Drucke und Spezialsammlungen vom 6. bis zum 16. Jahrhundert, aber auch im Rahmen spezieller Förderprogramme für das 20. und 21. Jahrhundert sowie durch eine 2007 begründete Public-Private-Partnership der BSB mit Google zur Digitalisierung des gesamten urheberrechtsfreien Bestandes des 17. bis späten 19. Jahrhunderts.

Bis 2005 war die Auswahl der Projekte primär am Bestandsprofil der BSB als international agierende Forschungsbibliothek sowie an ihren Funktionen als zentrale Landes- und Archivbibliothek orientiert. Das MDZ nahm bevorzugt sogenannte »Boutique«-Digitalisierungsprojekte in Angriff, also die Digitalisierung besonderer und herausragender Bestände sowie kleinerer themenspezifischer Sammlungszusammenhänge – beispielsweise zur Geschichte und Kulturgeschichte Bayerns im Rahmen des kulturwissenschaftlichen Internetportals »Bayerische Landesbibliothek Online« (BLO)¹¹, aus den DFG-Sondersammelgebieten der BSB

sowie von ausgewählten Handschriften und frühneuzeitlichen Drucken aus den weltweit unikalen Beständen der Bibliothek. Mit der Jahreswende 2005/2006 wurden an der BSB die technischen und organisatorischen Weichen für eine Ausweitung des digitalen Bestandsaufbaus gestellt und der Übergang in die Phase der Massendigitalisierung eingeleitet. Mit dem – weltweit erstmaligen – Einsatz von Scanrobotik wurde in einem groß angelegten, DFG-geförderten Projekt mit der Digitalisierung von 40.000 deutschen Drucken des 16. Jahrhunderts aus dem Bestand der BSB begonnen. Das österreichische Start-Up-Unternehmen Treventus hatte 2006 ein völlig neues, innovatives Scanrobotersystem vorgestellt, das in einer Entwicklungspartnerschaft mit dem Institut für Buch- und Handschriftenrestaurierung der BSB und dem MDZ sukzessive an die spezifischen Merkmale Alter Drucke unter besonderer Berücksichtigung der hohen konservatorischen Anforderungen angepasst werden konnte. In den ersten 18 Monaten der Projektlaufzeit konnten bereits 12.000 Titel des 16. Jahrhunderts mit ca. 1,7 Millionen Seiten digitalisiert und bereitgestellt werden.¹²

Als weitere Beispiele von Massendigitalisierungsvorhaben an der BSB sind insbesondere die ebenfalls DFG-geförderten Projekte zur Inkunabeln-Digitalisierung und zur Digitalisierung von Sondersammelgebietsbeständen zu nennen. Im ersten Projekt werden seit 2008 mit einer Projektlaufzeit von 48 Monaten rund 9.700 ausgewählte Inkunabeln digitalisiert. Der geschätzte Gesamtumfang liegt bei rund 2 Millionen Seiten. Mit Stand Dezember 2010 waren hier 4.080 Titel mit 1.023.000 Seiten digitalisiert. Das zweite Projekt ist ein Kooperationsprojekt mit Wissenschaftsverlagen, die der Bayerischen Staatsbibliothek die Rechte zur Digitalisierung und freien Internetbereitstellung geschichtswissenschaftlicher Literatur des 20. Jahrhunderts abgetreten haben. Hier konnten in 24 Monaten rund 4.700 Titel mit 1,4 Millionen Seiten digitalisiert werden. Im Vergleich mit diesen Zahlen bewegt sich die Public-Private-Partnership der BSB mit Google, sieht man auf den Umfang der erzeugten digitalen Datenbestände, in einer anderen Kategorie: Hier werden in wenigen Jahren rund 1 Million Werke des 17. bis 19. Jahrhunderts mit einem Gesamtvolumen von 300 Millionen Seiten digitalisiert. Google übernimmt dabei auf eigene Kosten die komplette Produktion, im Gegenzug erhält die BSB Kopien der Daten, die von Google erzeugt werden (»Library Digital Copy«). Diese werden über die Webangebote der BSB, über den bayerischen Verbundkatalog sowie über internationale Angebote wie Europeana und WorldCat kostenfrei bereitgestellt.¹³ Mit Stand Dezember 2010 stehen aus sämtlichen Digitalisierungsprojekten der BSB so rund

Scanrobotik

Massendigitalisierungsvorhaben

Public-Private-Partnership

455.000 Werke aus dem Zeitraum von 600 bis ca. 1900 online bereit (das entspricht einem Datenvolumen von rund 275 Terabyte), die im Archivspeicher des Leibniz-Rechenzentrums langfristig aufbewahrt werden.¹⁴

Zur Gewährleistung einer Workflow-seitig standardisierten und für alle Digitalisate einheitlichen Produktion wurde seit 2003 am MDZ die »Zentrale Erfassungs- und Nachweisdatenbank« (ZEND) entwickelt, ein Software-Tool mit vielfältigen und differenzierten Dokumenten-Management-Komponenten. Die ZEND steuert und unterstützt den gesamten Produktionsprozess, so unter anderem die Vorbereitung, das Scannen, die Erschließung in unterschiedlicher Tiefe, die Generierung der unterschiedlichen Metadattentypen, die Speicherung und Langzeitarchivierung sowie die Bereitstellung und Präsentation im WWW. Durch die Einführung der ZEND konnte der Produktionsaufwand für die Digitalisierung eines Werkes deutlich verkürzt werden, was eine notwendige Voraussetzung für die Massendigitalisierungsprojekte darstellt. So werden heute pro Woche bis zu 1,5 Millionen Images von Buchseiten über die ZEND verarbeitet.

Langzeitarchivierung zwischen Exploration und Praxis: Das BABS-Projekt

Nach ersten Schritten, wie z. B. der Migration von 3.000 CDs aus der Produktion des MDZ im Jahre 2004 auf das Archivierungssystem des Leibniz-Rechenzentrums (LRZ), wurde die Kooperation mit dem LRZ im Bereich der digitalen Langzeitarchivierung durch weitere gemeinsame Projekte vertieft. Zur Sammlung erster systematischer Erfahrungen mit der Konzeption und Umsetzung der Langzeitarchivierung von Netzpublikationen und Digitalisaten hat das MDZ in Kooperation mit dem LRZ und mit Unterstützung der DFG in den Jahren 2005 bis 2007 eine organisatorisch-technische Infrastruktur zur Langzeitarchivierung als Pilotsystem aufgebaut, das »Bibliothekarisches Archivierungs- und Bereitstellungssystem« (BABS).¹⁵ Dazu wurden in der Bibliothek und im LRZ vor allem Workflows für Materialauswahl, Ablieferung und Sammlung, Erschließung, Archivspeicherung und Bereitstellung von E-Books und E-Journals spezifiziert und erprobt. Das technische Konzept für BABS integriert in der Bayerischen Staatsbibliothek und am Leibniz-Rechenzentrum bereits vorhandene technische Komponenten zu einer integrierten Infrastruktur: die ZEND, das Digital Asset Management System »DigiTool« der Firma Ex Libris, das im Rahmen von BABS für die Verwaltung und Bereitstellung von Netzpublikationen von Drittanbietern eingesetzt wird, sowie das Archiv- und Backupsystem als Archivspeicher beim Leibniz-Rechenzentrum.

Im weiteren Ausbau der BABS-Infrastruktur wurde die bestehende organisatorisch-technische Infrastruktur zu einer vertrauenswürdigen und skalierbaren Lösung für die Langzeitarchivierung ausgebaut, um dem rasant steigenden Zuwachs von Dateien und Datenvolumina aus der Massendigitalisierung (rund 300 Terabyte in über 500 Millionen Dateien)¹⁶ gewachsen zu sein. Insbesondere wurden folgende Workflows der digitalen Langzeitarchivierung konzipiert und implementiert:

- Ein komfortables und sicheres Ablieferungsverfahren für die amtlichen Druckschriften der bayerischen Ministerien und Mittelbehörden über eine Webschnittstelle: Damit kommt die BSB der Bekanntmachung der Bayerischen Staatsregierung vom 1. Januar 2009 nach, in der die Pflichtablieferung amtlicher Veröffentlichungen auf Netzpublikationen ausgeweitet wurde. Die übermittelten elektronischen Dokumente werden gesichtet, katalogisiert und langzeitarchiviert. Zu den amtlichen Druckschriften zählen auch die Amtsblätter der Bayerischen Ministerien, die über eine zentrale Publikationsplattform, die von der BSB aufgebaut und betrieben wird, genuin elektronisch verkündet werden. Die »Verkündungsplattform Bayern«¹⁷ hat derzeit mit über 14.000 Abonnenten und 119 bereits publizierten Amtsblättern eine Vorreiterrolle im Bereich E-Government in Deutschland. Die Langzeitarchivierung der Amtsblätter ist integraler Bestandteil der Verkündungsplattform.
- Ein neu konzipierter Workflow für die Webarchivierung der Internetressourcen der virtuellen Fachbibliotheken der BSB: Bereits über 40.000 Websites sind in den insgesamt sechs Virtuellen Fachbibliotheken der BSB erfasst, die nun sukzessive über ein integriertes Verfahren von der Genehmigungseinholung, dem Harvest-Prozess über die Archivierung bis hin zur Katalogisierung im OPAC der BSB physisch auf den Servern der BSB vorgehalten und zugänglich gemacht werden. Für dieses Verfahren implementierte die BSB das WebCurator Tool (Open Source), das in Zusammenarbeit von der British Library und der Nationalbibliothek von Neuseeland entwickelt wurde.
- Ein Arbeitspaket zur Qualitätssicherung der Archivobjekte in Bezug auf Integrität und Authentizität: Dazu wurden das bestehende Metadatenprofil angepasst und durchgängig Checksummen gebildet. Die technische Infrastruktur wurde um ein großvolumiges, erweiterbares Onlinespeichersystem ergänzt, aus dem die Daten nach Abschluss aller qualitätssichernden Maßnahmen in das Archiv- und Backupsystem verschoben werden.

Einen weiteren Schwerpunkt in der Projektarbeit stellte schließlich die Entwicklung von strategischen Plänen für die Langzeitarchivierung (Preservation Planning) von Archivobjekten unterschiedlicher Dateiformate (z. B. TIFF und PDF) dar.¹⁸

Digitale Langzeitarchivierung im Routinebetrieb: Das Rosetta Digital Preservation System

Mit den im BABS-Projekt geschaffenen technischen, organisatorischen und fachlichen Infrastrukturen waren die Voraussetzungen gegeben, die Implementierung und den Routineeinsatz einer anspruchsvollen Softwarelösung zur digitalen Langzeitarchivierung für die BSB und den Bibliotheksverbund Bayern in Angriff zu nehmen. Mit Blick auf die internationale Marktsituation 2009 bot sich mit dem aus dem »National Digital Heritage Archive«-Programm in Neuseeland hervorgegangenen Produkt »Rosetta Digital Preservation System« der Firma Ex Libris, das anerkannte Standards der digitalen Langzeitarchivierung umsetzt und explizit für einen multi-institutionellen Einsatz ausgelegt ist, die aussichtsreichste Chance zur Einführung einer landesweit einsetzbaren, nachhaltigen Infrastruktur.

Die initiale Entwicklung der Rosetta-Software erfolgte in einer Kooperation der Nationalbibliothek von Neuseeland mit Ex Libris. Unterstützt wurde die Entwicklungsarbeit durch eine international besetzte Peer Review Group.¹⁹ Derzeit zählen neben der Nationalbibliothek von Neuseeland unter anderem das National Library Board von Singapur, die Bibliothek der ETH Zürich, das Goportis-Konsortium und die »Kirche Jesu Christi der Heiligen der Letzten Tage« zu den Anwendern der Software. Diese Anwender und weitere interessierte Institutionen bilden im Rahmen des sogenannten »Rosetta Charter Customer Program« ein Anwenderkonsortium, das aktiv an der weiteren Produktentwicklung von Rosetta mitwirkt.

Unterstützt durch Sondermittel des Bayerischen Staatsministeriums für Wissenschaft, Forschung und Kunst konnte 2010 operativ mit einem Pilotprojekt zur Einführung von Rosetta im Bibliotheksverbund Bayern begonnen werden. In einer ersten Projektphase wird Rosetta für drei Workflows der Bayerischen Staatsbibliothek implementiert:

➤ **Retrodigitalisate:** Schwerpunkte sind hierbei zum einen das an die etablierten Digitalisierungsworkflows anschließende vollautomatische Laden der Digitalisate in das Langzeitarchiv über eine sogenannte »Submission Application«, die Standard-Programmierschnittstellen von Rosetta anspricht, zum anderen die Skalierbarkeit von Rosetta insbesondere hinsichtlich des zu verarbeitenden Tagesdurchsatzes.

➤ **Elektronische Pflichtablieferung:** Wesentliche Schwerpunkte sind dabei die Modellierung von Periodika-Strukturen im Einklang mit den Erfordernissen eines digitalen Archivs, die Erprobung webbasierter, komfortabler und sicherer Ablieferungsverfahren für Produzenten sowie das automatische Anlegen von bibliographischen Datensätzen im Verbundkatalog, ausgehend von den Metadaten, die von den Produzenten mitgeliefert werden.

➤ **Webarchive:** Hier steht das automatische Laden periodischer Zeitschnitte ausgewählter Websites im Vordergrund (das eigentliche Web-Harvesting wird vom vorgelagerten Web-Curator-Tool ausgeführt).

Im abteilungsübergreifenden Projektteam der BSB, bestehend aus Mitarbeitern des MDZ und des Referates »Virtuelle Bibliothek Bayern« der Verbundzentrale des BVB, werden die vorliegenden Erfahrungen aus der Retrodigitalisierung, der digitalen Langzeitarchivierung sowie dem Betrieb von Verbundanwendungen gebündelt. Das Projekt wird – in Fortführung der Kooperation im BABS-Projekt – gemeinsam mit dem Leibniz-Rechenzentrum durchgeführt. Nach der Erstinstallation des Systems wurde Rosetta innerhalb einer mehrmonatigen Projektphase gemeinsam mit Ex Libris konfiguriert und an die Bedürfnisse der BSB unter Berücksichtigung des künftigen Verbundeinsatzes angepasst.

Ohne geeignete Schnittstellen zu den bereits bestehenden technischen Systemen der BSB ließe sich ein Archivsystem nicht sinnvoll betreiben. Im Fall von Rosetta ist insbesondere die Anbindung an das Produktionssystem ZEND des MDZ und an die Bibliothekskataloge zu nennen. Aus der ZEND fließen die retrodigitalisierten Objekte, bereits fertig als Archivpakete aufbereitet, in das Archivsystem ein. Der Katalog ist von wesentlicher Bedeutung für die Belieferung der Archivobjekte mit bibliographischen Metadaten. Außerdem soll jedes Archivobjekt im Katalog nachgewiesen werden und von dort mit nur einem einzigen Klick direkt aufrufbar sein; dazu findet ein Austausch von persistenten Links statt. In einer zweiten Projektphase werden ab 2011 drei weitere bayerische Bibliotheken, die Universitätsbibliotheken Augsburg, Regensburg und Würzburg, als Partner in das Projekt einbezogen, womit die Ausweitung des Gesamtvorhabens über die BSB hinaus auf den Bibliotheksverbund Bayern eingeleitet wird.

Für die Erprobung und Einführung von Rosetta steht eine Umgebung aus vier virtuellen Applikationsservern zur Verfügung. Sowohl die auszuführenden Verarbeitungsprozesse als auch die eingehenden Anfragen werden mittels eines Load Balancers verteilt. Außerdem ist ein Datenbank-Cluster aus zwei

»Rosetta Digital
Preservation System«

künftiger Verbundeinsatz

drei Workflows

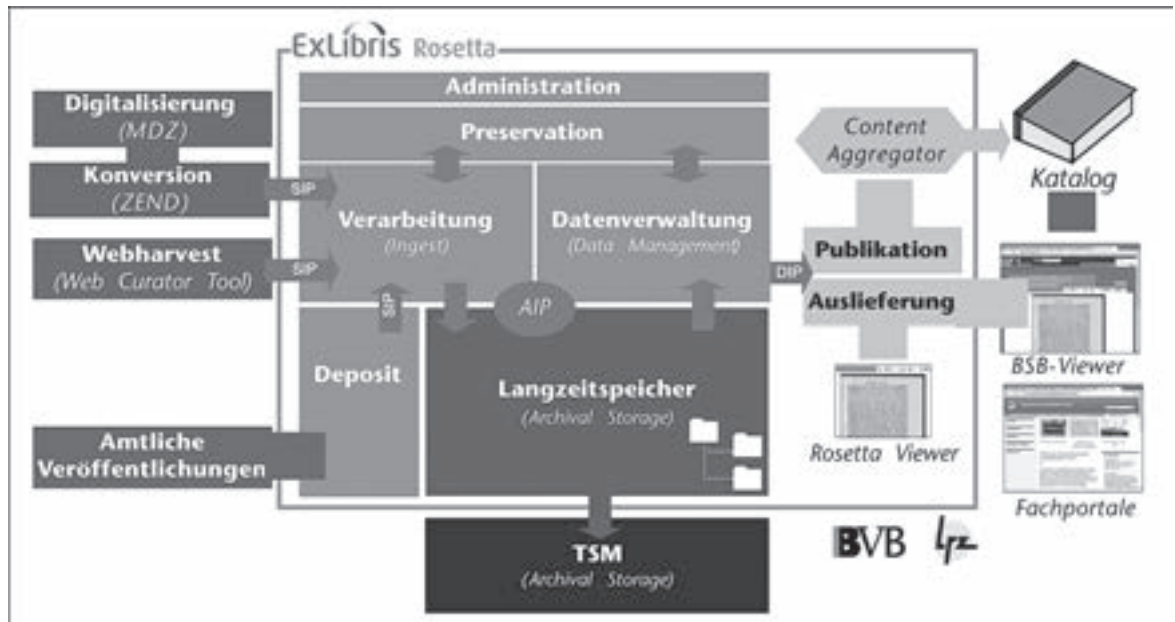


Abb. 2: Die Infrastruktur des Rosetta-Projekts an der Bayerischen Staatsbibliothek

zentrale Aspekte der Langzeitarchivierung

physischen Servern eingerichtet. Angesichts der enormen Datenmengen allein schon aus der Bayerischen Staatsbibliothek ist insbesondere die Skalierbarkeit der Anwendung von wesentlicher Bedeutung für einen erfolgreichen Produktiveinsatz. Auch dieser Aspekt wird intensiven Tests unterzogen. Die Aufnahme des Regelbetriebs von Rosetta für die genannten drei Workflows ist für den Frühsommer 2011 vorgesehen.

Für alle zentralen Aspekte der Langzeitarchivierung bietet Rosetta umfangreiche und ausdifferenzierte Funktionalitäten. Im Bereich der Ablieferung und Übergabe der Objekte an das Langzeitarchiv (Deposit/Ingest) zählt dazu insbesondere die Option, für unterschiedliche Akteure (z. B. bibliotheksinterne Organisationseinheiten und externe Produzenten) innerhalb des Systems verschiedene Rollen und Workflows definieren und zuweisen zu können. Rosetta ist so konzipiert, dass es grundsätzlich für alle digitalen Objekt- und Formattypen geeignet ist und die Aufnahme dieser unterschiedlichen Objekte in das digitale Langzeitarchiv auf vielfältigen Wegen unterstützt. Dies erfolgt in einer Kombination von automatisierten und manuellen Abläufen.

Sicherheits- und Qualitätsmanagement

Zudem bietet Rosetta einen hohen Standard im Sicherheits- und Qualitätsmanagement für den Umgang mit digitalen Objekten. Die Maßnahmen zur Prüfung und Sicherung der Integrität und Authentizität erstrecken sich auf den gesamten Lebenszyklus der digitalen Objekte, also von der Aufnahme ins digitale Archiv über die Durchführung von Erhaltungsmaßnahmen und die Verwaltung der verschiedenen Versionen bis zur Bereitstellung für den Nutzer. Sämtliche an den

Objekten durchgeführten Maßnahmen (z. B. Formatmigration) bzw. Statusänderungen (z. B. in Bezug auf Zugriffsrechte) werden dabei automatisch und dauerhaft in den Metadaten dokumentiert und bleiben damit nachvollziehbar. Weiterhin bilden die speziellen Funktionalitäten zur Planung und Durchführung der Langzeiterhaltungsmaßnahmen (Preservation Planning and Action) – d. h. vor allem Maßnahmen zum Risikomanagement (z. B. Aufbau und Betrieb einer Formatbibliothek) sowie zur Steuerung und Durchführung von Migrationen – einen zentralen Baustein der Software, der für den Erhalt der digitalen Objekte und die Sicherstellung des dauerhaften Zugangs unverzichtbar ist.

Rosetta setzt zudem auf bereits etablierten Standards und Schnittstellen (ISO-Referenzmodell OAIS, METS, Dublin Core, PREMIS) im Langzeitarchivierungs-, Bibliotheks- und IT-Sektor auf und gewährleistet so die Stabilität des Datenmodells sowie Interoperabilität beim Austausch von Meta- und Objektdaten innerhalb verschiedener und sich verändernder Systemumgebungen. Zur Vertrauenswürdigkeit des Systems gehört es, dass es den Fortbestand der archivierten Objekte über die Lebensdauer des Systems hinaus garantiert. Diese Anforderung wird von Rosetta erfüllt, indem es ein Exitszenario beinhaltet, das jederzeit und zu jedem Entwicklungsstand den vom technischen System unabhängigen Zugriff auf die archivierten Objekte und ihren Transfer in andere Systeme ermöglicht.

Die BSB nimmt im Rahmen der Implementierung von Rosetta am sogenannten »Rosetta Charter Customer Program« von Ex Libris teil, das neben dem Aus-

tausch mit anderen internationalen Anwendern ein hohes Maß an Einflussnahme auf die noch junge Produktentwicklung ermöglicht. Das Programm umfasst physische und virtuelle Treffen einer Charter Customer Advisory Group, deren erstes im Juni 2010 in München stattgefunden hat, sowie von spezialisierten Workgroups, in denen die BSB federführend (im Themenfeld Digital Cultural Heritage) oder aktiv teilnehmend (z. B. im Themenfeld Forschungsprimärdaten und Format Library) vertreten ist. Zur beratenden Begleitung der Implementierung des Rosetta Digital Preservation Systems in der BSB und des schrittweisen Rollouts in weiteren wissenschaftlichen Bibliotheken des Bibliotheksverbundes wurde zudem seitens des Bayerischen Staatsministeriums für Wissenschaft, Forschung und Kunst ein Beirat einberufen, dem neben Vertretern der Bibliotheken auch Repräsentanten der Hochschulrechenzentren Bayerns angehören. Hierdurch wird die enge Verzahnung der fachlichen mit den IT-technischen Fragen im Kontext der Einführung und des Regelbetriebs von Rosetta sichergestellt.

3. FAZIT: DIGITALE LANGZEITARCHIVIERUNG ALS KOOPERATIONSAUFGABE

Vergleicht man die Projektvorhaben der DNB und der BSB, so wird zumindest tendenziell eine Gegenläufigkeit der strategischen Annäherung an das komplexe Handlungsfeld der digitalen Langzeitarchivierung sichtbar. Die Lösungsansätze der DNB führen im Ergebnis zu einer schrittweisen Abkehr von der Integration kommerzieller Produkte hin zu einem vollständig Open-Source-basierten, Community-orientierten Konzept. Das Vorgehen der BSB hingegen ist gekennzeichnet durch den Übergang von einem primär auf Eigenentwicklungen setzenden Modell hin zur Lizenzierung einer kommerziellen Langzeitarchivierungs-Software als Kern des entwickelten Frameworks.

Bei näherem Hinsehen zeigt sich aber auch, dass beide Lösungswege in zwei wesentlichen Aspekten übereinstimmen. Die (im Falle der DNB) unter dem Titel »koLibRI« sowie (im Falle der BSB) als »BABS« und »ZEND« zusammengefassten Softwaretools und Workflow-Modelle stellen in beiden Fällen Eigenentwicklungen dar, die eine stabile, nachhaltig abgesicherte Basis der aktuellen, auf ihnen aufsetzenden Lösungsansätze bilden. Darüber hinaus betrachten beide Einrichtungen die digitale Langzeitarchivierung als eine nur kooperativ zu bewältigende Aufgabe. In beiden Projekten spielt die Zusammenarbeit mit universitären Infrastruktureinrichtungen, die anspruchsvolle und innovative Anforderungen an digitale Langzeitarchivierungs-Strategien stellen, eine wichtige Rolle

ebenso wie die Kooperation mit überregional agierenden Rechenzentren – hier eröffnen sich gerade für die Zukunft spannende Szenarien eines Cloud-basierten Datenhostings.

Deutlich wird auch, dass sich anspruchsvolle Projekte zum Ausbau digitaler Informationsangebote, vor allem in der Massendigitalisierung, beim elektronischen Pflichtexemplar und künftig im Bereich der Langfristverfügbarkeit von Forschungsprimärdaten, und die Anforderungen an eine nachhaltige technologische Lösung zur digitalen Langzeitarchivierung gegenseitig bedingen. Beide Felder verlangen zudem derart hohe Eigenleistungen in finanzieller wie personeller Hinsicht, dass auch große Informationsinfrastruktur-Einrichtungen mit einem überregionalen Leistungsauftrag hier nicht allein, sondern immer nur in institutionalisierter Kooperation erfolgreich agieren können. Dieser kooperative Weg kann durchaus mühsam sein, muss doch immer wieder neu das wohl berechnete Eigeninteresse aller Partner abgeglichen werden und in einem neuen Verständnis der jeweiligen Kooperation aufgehen. Die dargestellten Projekte zeigen hier deutlich: Nur aus dem konsequenten Zusammenwirken derart verteilter, regionaler und überregionaler »Leistungsknoten« wird sich Schritt für Schritt die zukünftige Landkarte einer nationalen Infrastruktur zur Langzeitbewahrung digitaler Informationsobjekte entwickeln.

strategische Annäherung

¹ Vgl. auch Neuroth, Heike et al. (Hrsg.): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, Version 2.3, vwh-verlag, Universitätsverlag Göttingen, 2010; Download unter http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf (Aufruf: 13.12.2010).

² <http://kopal.langzeitarchivierung.de> (Aufruf: 13.12.2010); vgl. auch Altenhöner, Reinhard: Daten für die Zukunft – Das BMBF-Projekt Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen (kopal) und seine Hintergründe. In: *Bibliothek. Forschung und Praxis* 29 (2005) 2, S. 192–198; Wollschläger, Thomas: Kopal goes live. In: *Dialog mit Bibliotheken*, 19 (2007) 2, S. 17–22.

³ http://kopal.langzeitarchivierung.de/index_koLibRI.php.de (Aufruf: 30.12.2010).

⁴ http://kopal.langzeitarchivierung.de/downloads/kopal_Universelles_Objektformat.pdf bzw. mit Beispielen angereichert: http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.de (Aufruf: 30.12.2010).

⁵ Vgl. hierzu aus Sicht der Partner DNB/SUB: Lossau, Norbert und Klaproth, Frank: *Kooperation & Nachnutzung der kopal-Lösung*. http://kopal.langzeitarchivierung.de/downloads/kopal-goes-live_kopal-Kooperation_und_Nachnutzung_Klaproth.pdf (Aufruf: 30.12.2010) und aus Sicht der IBM: Schwarz, Jürgen: *Kooperation & Nachnutzung der kopal-Lösung*. http://kopal.langzeitarchivierung.de/downloads/kopal-goes-live_IBM_DIAS_Perspektive_und_Preismodell.pdf (Aufruf: 30.12.2010).

⁶ Vgl. LTP Working Group, The Hague, 2010: www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_Long_Term_Preservation_Services_2010-08-05.pdf (Aufruf: 30.12.2010).

⁷ Vgl. dazu www.alliancepermanentaccess.eu/component/content/article/1-latest-news/62-aparsen-project-to-be-funded.html sowie die Alliance for permanent access: www.alliancepermanentaccess.eu/home.html (Aufruf: 30.12.2010).

⁸ Vgl. <http://dp4lib.langzeitarchivierung.de/> (Aufruf: 30.12.2010); vgl. auch Altenhöner, Reinhard et al.: »dp4lib – Digital preservation for Libraries«. In: *Dialog mit Bibliotheken*, 2010, 2, S. 27–30.

⁹ Zu den 3D-Handschriften siehe die WWW-Präsentation ausgewählter Spitzenstücke im Rahmen der Bayerischen Landesbibliothek Online unter www.bayerische-landesbibliothek-online.de/3D sowie

nur kooperativ zu bewältigende Aufgabe

auch deren Anwendung im gestengesteuerten Präsentationssystem für Ausstellungen unter <http://www.youtube.com/watch?v=qmMM MVnNxLI> (Aufruf: 29.12.2010).

¹⁰ Reiner, Bernd: Sicherung des Weltkulturerbes am Leibniz-Rechenzentrum. In: Akademie Aktuell 03/22 (2007), S. 18–22. www.badw-muenchen.de/aktuell/akademie_aktuell/2007/heft3/o6_Reiner_LRZ.pdf (Aufruf: 13.12.2010).

¹¹ www.bayerische-landesbibliothek-online.de (Aufruf: 13.12.2010).

¹² Brantl, Markus; Ceynowa, Klaus; Fabian, Claudia; Meßmer, Gabriele; Schäfer, Irmhild: Massendigitalisierung deutscher Drucke des 16. Jahrhunderts – Ein Erfahrungsbericht der Bayerischen Staatsbibliothek. In: Zeitschrift für Bibliothekswesen und Bibliographie 56 (2009), 6, S. 327–338. http://zs.thulb.uni-jena.de/servlets/MCRFileNodeServlet/jportal_derivate_00180423/j09-h6-auf-1.pdf (Aufruf: 13.12.2010).

¹³ Ceynowa, Klaus: Eine Million Bücher online. In: Akademie Aktuell 03/22 (2007), S. 35–39. www.badw.de/aktuell/akademie_aktuell/2007/heft3/11_Ceynowa.pdf (Aufruf: 13.12.2010).

¹⁴ Siehe den tagesaktuellen Counter auf der Homepage des MDZ: www.muenchener-digitalisierungszentrum.de/ oder www.digitale-sammlungen.de (Aufruf: 30.12.2010).

¹⁵ www.babs-muenchen.de (Aufruf: 13.12.2010).

¹⁶ Vgl. die aktuelle Statistik zur digitalen Langzeitarchivierung des MDZ unter www.digitale-sammlungen.de/index.html?c=langzeit-archivierung-zahlen&l=de (Aufruf: 29.12.2010); so betrug das zwischen November 2009 und Oktober 2010 produzierte Datenvolumen allein 95 Terabyte.

¹⁷ Vgl. www.verkuendung-bayern.de/ (Aufruf: 29.12.2010).

¹⁸ Rauber, Andreas; Brantl, Markus; Schoger, Astrid; et. al.: From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings. In: D-Lib Magazine 15 (2009). www.dlib.org/dlib/november09/kulovits/nkulovits.html (Aufruf: 13.12.2010).

¹⁹ Zum NDHA vgl.: www.natlib.govt.nz/about-us/current-initiatives/ndha (Aufruf: 29.12.2010); Präsentation von Knight, Steve: National Digital Heritage Archive, Securing the future. Prag 18. Juli 2008. www.ndk.cz/narodni-dk/publikace/prednaska-steve-knight-hlavni-programovy-architekt-projektu-national-digital-heritage-archive-novy-zeland/prezentace (Aufruf: 29.12.2010).

DIE VERFASSER

Reinhard Altenhöner ist Leiter der Abteilung »Informationstechnik« der Deutschen Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, Mail: r.altenhoener@dnb.de

Dr. Markus Brantl ist Leiter des Referates »Münchener Digitalisierungszentrum / Digitale Bibliothek« der Abteilung Bestandsaufbau und Erschließung der Bayerischen Staatsbibliothek, Ludwigstr. 16, 80539 München, Mail: brantl@bsb-muenchen.de

Dr. Klaus Ceynowa ist Stellvertretender Generaldirektor der Bayerischen Staatsbibliothek, Ludwigstr. 16, 80539 München, Mail: ceynowa@bsb-muenchen.de