– Übernahme/Workflows –

Der Workflow zur Speicherung digitaler Daten in einem kooperativen Modell (kopal)

THOMAS WOLLSCHLÄGER

Das elektronische Publizieren im Internet verändert die Aufgaben der Bibliotheken, die für den Erhalt des kulturellen Erbes verantwortlich sind. Ziel des Projekts kopal ist die Entwicklung eines kooperativ erstellten und betriebenen Archivsystems für digitale Dokumente als Lösung für ihre Langzeiterhaltung und Langzeitverfügbarkeit.

1. Herausforderungen für die digitale Langzeitarchivierung

Das rasante Wachstum der elektronischen Veröffentlichungen in Menge und Vielfalt hat in den letzten Jahren die Aufgaben von Bibliotheken und Archiven massiv erweitert. Ein zunehmender Teil der kulturellen Aktivitäten und Publikationen aller Art findet seinen Platz im Internet. Die Sammlung und Archivierung diesen „Nachlasses" ist – auch gesetzlich fixiert – Aufgabe von Nationalbibliotheken und vergleichbaren Einrichtungen. Die Novellierung des „Gesetzes über die Deutsche Nationalbibliothek" im Jahre 2006 erweitert deren Sammelauftrag auch auf nicht-trägergebundene digitale Materialien („unkörperliche Medienwerke"). Dies bezieht sich auch auf den Bereich Web-Archiving.

Für die elektronischen Materialien werden entsprechend geeignete und dimensionierte Archive benötigt. Bisherige Systeme haben jedoch die Anforderungen an „vertrauenswürdige Archive" nicht oder nur teilweise erfüllt. Ein geeignetes Archivsystem für Online-Publikationen muss vor allem zwei Probleme adressieren können:
Als Grundlage müssen zum einen die binären Daten erhalten werden, denn kein existierender Datenträger ist ewig oder auch nur langfristig genug haltbar. Zum anderen muss angesichts des rasanten Technologiewechsels der Zugriff auf ältere Datenformate weiterhin gewährleistet sein.

Lösungen bieten die Verfahren Migration und Emulation. Da beide Verfahren Vor- und Nachteile haben, sollte in der Praxis ein Archivsystem zur Langzeiterhaltung und Langzeit-Verfügbarmachung digitaler Publikationen eine Kombination beider Verfahren ermöglichen.

2. Lösungsprinzipien des kopal-Projekts und generelle Workflows in der Projektarchitektur

Das Ziel von kopal ist daher der Aufbau einer nachnutzbaren technischen und organisatorischen Infrastruktur zur Sicherung der Langzeitverfügbarkeit elektronischer Publikationen. Es ist ein Förderprojekt des Bundesministeriums für Bildung und Forschung über eine Laufzeit von drei Jahren (bis Mitte 2007, Fördervolumen: 4,2 Mio.). Den Kern des kopal-Archivsystems bildet das von IBM in Zusammenarbeit mit der Königlichen Bibliothek der Niederlande (KB) entwickelte Digital Information Archiving System (DIAS). Innerhalb des Projekts werden digitale Materialien aller Art der Deutschen Nationalbibliothek (DNB) und der Niedersächsischen Staats- und Universitätsbibliothek (SUB Göttingen) in das Archiv eingestellt. Im Wesentlichen beruht kopal dabei auf den drei Prinzipien Kooperation, Universalität und Nachnutzbarkeit.

Im Rahmen einer nationalen Kooperation arbeiten im Projekt zwei sehr unterschiedliche Bibliotheken zusammen. Dies betrifft sowohl den jeweiligen Auftrag (National- bzw. Universitätsbibliothek), als auch Schwerpunkte bei den bisherigen Aktivitäten. Der technische Betrieb des Systems ist ausgelagert und bei der „Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen" (GWDG) angesiedelt. Der Partner IBM Deutschland ermöglicht eine professionelle Anpassung der Software und bietet eine langfristig stabile Unterstützung. Im internationalen Rahmen arbeiten die Projektpartner eng mit der Königlichen Bibliothek der Niederlande zusammen und entwickeln gemeinsam Anforderungen an künftige Weiterentwicklungen von DIAS.

Das Projekt erfüllt den Anspruch an ein universell nutzbares Archivsystem, indem zum einen die Sicherstellung der langfristigen Verfügbarkeit durch Migration und Emulation unterstützt wird. Zum anderen gibt es dabei in kopal keinerlei Einschränkungen für die Art des Materials, welches in das

– Übernahme/Workflows –

Archiv eingespielt werden kann (Text, Bilder, Audio, Video), und für die möglichen Dateiformate. Obwohl das kopal-System für den Projektzeitraum eine begrenzte Gesamtkapazität hat, ist die Größe des einzelnen Archivobjekts nicht begrenzt. Jeder der Partner ist völlig frei in der Auswahl und Regelfestlegung beim Einspielen der von ihm gesammelten Objekte.

Die Nachnutzung von kopal durch weitere Institutionen, die eine Langzeitarchivierung benötigen, ist ausdrücklich erwünscht. Die kopal-Solution ist von vorneherein auf unterschiedliche Bedürfnisse ausgerichtet. Um die Nachnutzbarkeit zu gewährleisten, werden etablierte Standards genutzt. Der Transfer der Objekte in ein digitales Archiv über standardisierte Formate, Transportwege und Schnittstellen ist dabei ein wichtiges Erfordernis. Bereits das Kernssystem DIAS von IBM basiert auf bewährter Standardsoftware (wie der DB2-Datenbank, dem Content Manager und dem Tivoli Storage-Manager) und weist eine zukunftsfähige Trennung von Speicherkonzept und Datenverwaltung auf. Das OAIS-Referenzmodell für digitale Archivierung ist in DIAS konsequent implementiert. Die Erweiterbarkeit von DIAS für neue Nutzerinstitutionen und präzise definierte Import- und Exportschnittstellen ermöglichen es, eine Archivnutzung in verschiedenste Workflows unterschiedlichster Institutionen zu integrieren.

Das kopal-System erlaubt eine getrennte Nutzung durch verschiedene Institutionen (Mandantenfähigkeit). Vergleichbar getrennten Schließfächern, verfügt jede Institution über einen eigenen Bereich, in den nur sie Daten einstellen, verändern und abrufen kann. Die Institutionen können per Fernzugriff auch über sichere Internetverbindungen auf das kopal-System (derzeit auf den Speicherstandort der GWDG in Göttingen) zugreifen *(siehe Abb. 1)*. Upgrades des DIAS-Systems an sich sind jedoch für alle Mandanten gemeinsam gültig. Die Mandantenfähigkeit und der mögliche Fernzugriff sind die Kernvoraussetzungen für die Nachnutzbarkeit des kopal-Systems durch andere Gedächtnisorganisationen.

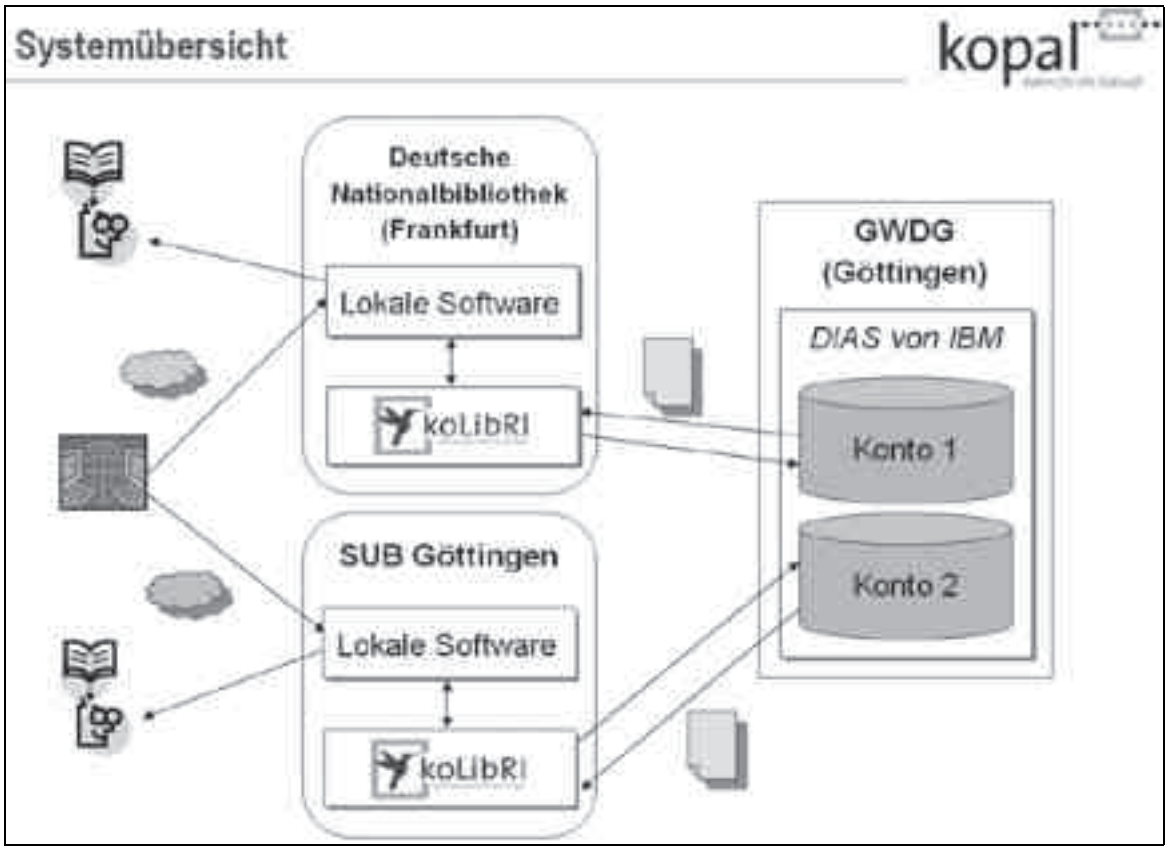

Abb. 1: Das kopal-Gesamtsystem mit Fernzugriff auf den Speicherstandort

– Übernahme/Workflows –

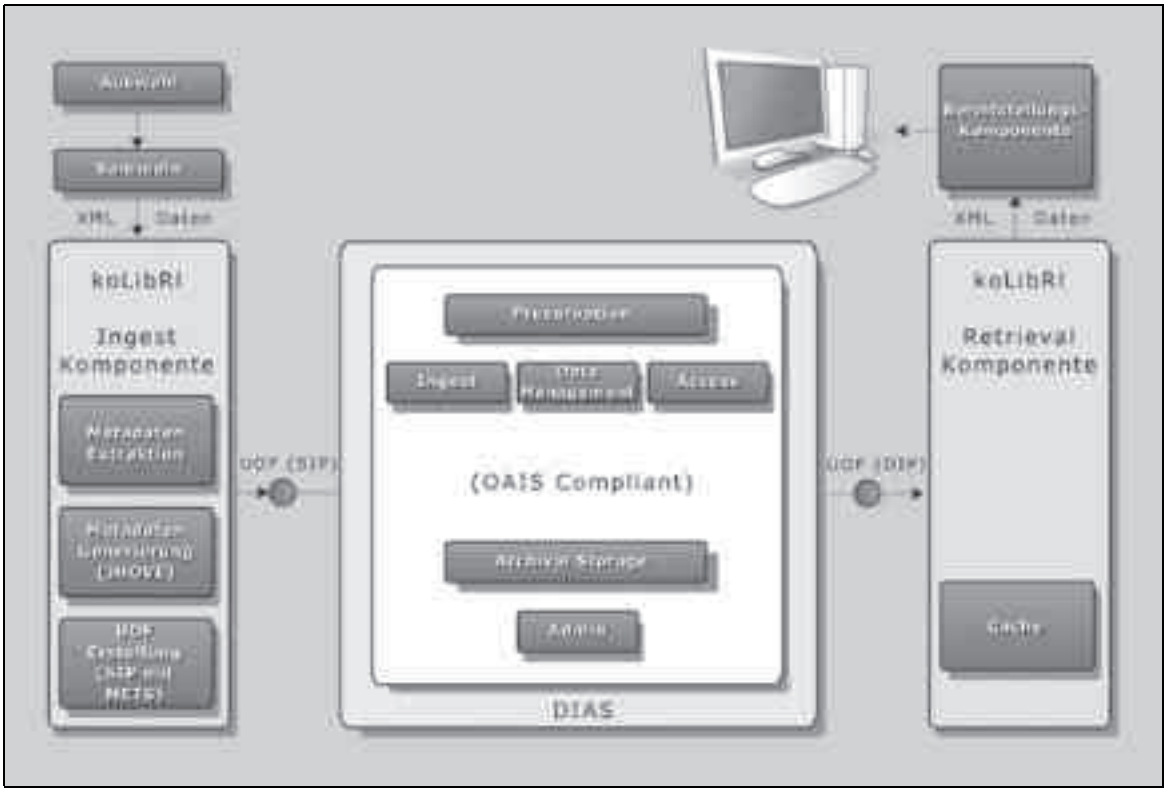

Abb. 2: Funktionen der koLibRI-Software um den OAIS-konformen Archivkern

Außerdem wird die DIAS-Software durch flexible Module erweitert. Hierfür haben die DNB und die Niedersächsische Staats- und Universitätsbibliothek Göttingen auf DIAS abgestimmte Softwareprodukte erstellt, die als „kopal Library for Retrieval and Ingest" (koLibRI) unter einer Open Source Lizenz veröffentlicht werden. Bei diesen kopal-Tools geht es hauptsächlich um den Bereich des Einspielens von Objekten in das DIAS sowie den Zugriff auf die archivierten Objekte *(siehe Abb. 2)*. koLibRI ist eine generische freie Softwarebibliothek (basierend auf JAVA) zur Ein- und Anbindung unterschiedlicher Mandanten und der automatischen Erstellung von technischen Metadaten. koLibRI besitzt konfigurierbare Workflows. Das Projekt kopal hat dafür das „Universelle Objektformat" vorgelegt, mit dem digitale Objekte zusammen mit Metadaten archiviert und zwischen Institutionen und Systemen ausgetauscht werden können.

Die Systementwicklung wird dabei so offen angelegt, dass eine Ausdehnung der kooperativen Nutzung auf weitere Nachnutzer aus dem Kreis aller „Gedächtnisorganisationen" (Bibliotheken, Archive und Museen) möglich ist. Bereits seit März 2006 steht für Testzwecke in einer Beta-Version ein vorläufiger Release der neu erstellten koLibRI-Software zur Verfügung, der zu Beginn des Jahres 2007 als Beta-2-Release aktualisiert worden ist. Mit dem Projektabschluss im Juli 2007 werden ein nachnutzbares System und ein endgültiges Release der vollständig entwickelten KoLibRI-Software bereitgestellt.

3. Bisheriger Projektverlauf und Dateneinspielung von kopal

Nach einer Pilotphase zur System-Evaluierung und Entwicklung des Universellen Objektformats befindet sich kopal derzeit in der Entwicklungsphase. Erklärtes Ziel war, zunächst anhand einer Vielzahl von Objekten in verschiedenen Formaten die praxisnahe Nutzung der kopal-Lösung aufzuzeigen. In einem nächsten Schritt waren die in kopal entwickelten Arbeitsabläufe prinzipiell an die Erfordernisse eines künftig in der Routine laufenden Einspielbetriebs in den Bibliotheken anzupassen. Derzeit werden bis zum Ende der Laufzeit des Projektes Komponenten für Administration und Prozess-Monitoring im Archivsystem fertiggestellt sowie die Voraussetzungen geschaffen, um in kopal

Migrations- und Emulationsprozesse durchzuführen, die eine langfristige Interpretierbarkeit der archivierten Dokumente sicherstellen.
Ende 2005 wurden das bei der GWDG installierte System und die neu entwickelten Tools erfolgreich getestet. Im Jahre 2006 wurde das aus den Parametern des erfolgreich getesteten Referenzsystems abgeleitete eigentliche Produktivsystem von kopal aufgesetzt. Im August 2006 konnte dann die erste Stufe des Produktivbetriebs erfolgreich aufgenommen werden. Dabei haben die Projektpartner DNB und SUB allein im ersten Durchgang über 40.000 zu archivierende Dokumente in das bei der GWDG gehostete System eingespielt. Bei Projektende werden diese Datenbestände den Grundstock der dauerhaften Archivierung der elektronischen Materialien der Bibliotheken bilden.

Weitere im Projektantrag dezidiert genannte einzuspielende Objektklassen bilden E-Journal-Artikel, Digitalisate und CD-ROMs bzw. DVDs. Letztere Objekte müssen vor dem Einspielen in nicht trägergebundene digitale Formen überführt werden; hierbei soll die Archivierung möglichst als Images nach ISO 9660 erfolgen. In der DNB wurden hierzu exemplarische Mengen solcher Publikationen ausgewählt. Durch kopal erfolgten die Erstellung von Images einschließlich Analysen auftretender Problemfälle und das Einspielen dieser Objekte in das Langzeitarchiv. Anhand dieser Objekte erfolgten auch wesentliche Anstöße für die Umgestaltung der Workflows in der Bibliothek.

4. Das Management der Workflows in der Deutschen Nationalbibliothek

Bisher existierten für die Verarbeitung von Online-Dokumenten mehr oder weniger halbautomatische Verfahren, außerdem gab es verschiedene Workflows für Netzpublikationen, Online-Dissertationen und weitere Materialien. Ziel ist die Schaffung eines automatischen, einheitlichen Verfahrens mit der Übergabe der Archivobjekte an kopal (Ingest) bzw. beim Access die Übergabe von kopal an Arbeitsplatzrechner oder das neu entstehende Bereitstellungssystem. Dabei sind zahlreiche Abteilungen in der DNB involviert: das (Pica-)ILTIS-Team, die Portal-Gruppe, die Abteilung Erwerbung/Formalerschließung, die Benutzungsabteilung und nicht zuletzt die (externen) Ablieferer. Letztere benötigen für die Anbindung an die Prozesse in der DNB bzw. Zulieferung in der Regel Unterstützung.

Der Workflow von elektronischen Materialien auf Datenträgern (d.h., CD- bzw. DVD-Veröffentlichungen) bildet, wie bereits erwähnt, einen derjenigen, die aufgrund der Anforderungen seitens Archivsystem und künftiger Bereitstellung anzupassen sind. Nach der Erstellung der Images und Analysen wurde ein Änderungs- und Ergänzungsvorschlag für den Geschäftsgang dieses Materials vorgelegt. Dieser Workflow wird derzeit überarbeitet. Ebenso wird der Workflow für genuin online vorliegende Netzpublikationen unter Einbeziehung der Anforderungen der Langzeitarchivierung neu gestaltet und der Workflow auf die Schnittstellen des Archivsystems angepasst. Für sogenannt fortlaufende Publikationen (vor allem elektronische Zeitschriften-Artikel) entsprechen die künftigen Archivobjekte oft nicht der aktuellen Abbildung im Online-Katalog. Bibliografische Metadaten von Archivobjekten müssen künftig in ILTIS (Pica) abgebildet werden. Dazu müssen eine Festlegung von Erschließungsvarianten und ein Mapping von Archivobjekten auf Katalogobjekte erfolgen. Das URN-Management in der DNB wurde bereits erweitert. Da jedes Objekt zum Einspielen in das Archiv einen Persistent Identifier benötigt, erfolgt für bereits gesammelte Objekte ohne URN eine Retro-Vergabe der URN. Alle neuen Objekte müssen mit URN geliefert werden bzw. bei Eingang/Bearbeitung einen URN erhalten, was dem künftigen Verfahren entspricht.

Zusammenfassend kann zum Workflow-Bereich gesagt werden, dass wesentliche Voraussetzungen für die Einbindung des Archivs in die Geschäftsumgebung der Institution vorliegen bzw. gerade geschaffen werden: Das Produktionssystem wurde aufgesetzt und läuft, das produktive Einspielen von Material wurde und wird erprobt, nötige Weiterentwicklungen (z.B. noch fehlende Module zur Auswertung von Dateiformaten) wurden und werden ermittelt und Änderungen in diversen Workflows wurden angestoßen. Anstehende Aufgaben bestehen in der Übergabe des kopal-Systems vom Projektteam an eine ständige Arbeitseinheit sowie die Gestaltung des Managements von rechtzeitigem Einspielen aller Retro-Objekte und des aktuellen und künftigen Zugangs (direkt nach kopal).

– Übernahme/Workflows –

5. Herausforderungen und Perspektiven

Da die sichere Archivierung von Internetressourcen nunmehr ein wichtiges Arbeitsfeld darstellt, hat die DNB bereits vor dem Inkrafttreten des neuen Gesetzes verschiedene Möglichkeiten für automatisiertes Harvesting und entsprechende Archivierungsmechanismen getestet. Um diese Daten archivieren zu können, ist eine rechtzeitige Erhöhung des Speichervolumens von kopal zur Aufnahme aller Retro-Objekte und für den laufenden Zugang erforderlich. Angesichts großer Datenmengen und teilweise großer Einzelobjekte (z.B. digitalisierte DVDs) ist die Gewährleistung einer ausreichenden Performanz des Systems einschließlich einer entsprechenden Benutzerbetreuung (etwa für Wartezeiten, rechtliche Beschränkungen u.a.) im Blick zu behalten.

Rechtzeitig vor Projektende hat kopal außerdem Modelle für die künftige Nachnutzung der Projektergebnisse bzw. des kopal-Archiv-Systems vorgelegt. kopal bietet ein flexibel konfigurierbares und anpassbares System an. Die Komponenten von kopal können je nach Kundenanforderung vereinbart und ausgestaltet werden. Die Modalitäten hängen außerdem vom gewünschten Nutzungsmodell ab. kopal bietet hier prinzipiell drei unterschiedliche Nutzungsmodelle an:

- „kopal-Teilnehmer“: Eine Institution lässt ihre Daten „kommissarisch“ durch einen kopal-Mandanten archivieren.
- „kopal-Mandant“: Eine Institution verwaltet selbstständig einen eigenen Bereich (Schließfach) des kopal-Archivsystems, der Serverbetrieb selbst bleibt ausgelagert.
- „kopal-Eigenbetrieb“: Eine Institution betreibt unter Rückgriff auf Erfahrungen des kopal-Projekts ein eigenes vollständiges Archivsystem.

Abhängig vom individuellen Servicekonzept einer Institution sind für die jeweiligen Nutzungsmodelle eine Reihe von Kostenfaktoren wie Zahl und Komplexität der Workflows bei einer Kundeninstitution, Menge, Heterogenität und Komplexität der zu archivierenden Objekte und ihrer Metadaten oder die gewünschten Zugriffsmöglichkeiten und Schnittstellen zu berücksichtigen.