

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

kopal 
Daten für die Zukunft

kopal - Aufbau eines kooperativ entwickelten und betriebenen Archivsystems:
Ein Erfahrungsbericht am Beispiel Hochschulbibliotheken

Frank Klaproth, SUB Göttingen



- Projekthintergrund
- Aktueller Stand der kopal-solution
- koLibRI tools
- Praxiserfahrung am Beispiel SUB - GDZ
- Ausblick und Perspektiven

- Förderprojekt des Bundesministeriums für Bildung und Forschung (BMBF)
- Charakter: Forschungs- und Innovationsprojekt
- Fördervolumen: 4,2 Mio €, Laufzeit: 1.7.2004 – 30.6.2007
- Projektziel: Aufbau einer von Gedächtnisorganisationen nachnutzbaren technischen und organisatorischen Infrastruktur zur Sicherung der Langzeitverfügbarkeit elektronischer Publikationen
- Basis: DIAS (Digital Information Archiving System) entwickelt von IBM für die Koninklijke Bibliotheek, die Nationalbibliothek der Niederlande
- Projektpartner
 - Die Deutsche Bibliothek (DDB) (Gesamtprojektleitung)
 - Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
 - Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
 - IBM Deutschland GmbH (IBM)

- Technische Lösung für die langfristige Verfügbarkeit von digitalen Informationen benötigt:
 - Hardwareumgebungen zum sicheren Speichern der digitalen Information (Festplattensysteme, Bandroboter, Netzwerkkomponenten, usw.)
 - Softwareumgebungen zum Verwalten der Hardwareumgebungen, der digitalen Objekte, der deskriptiven und technischen Metadaten der digitalen Objekte und deren Formate und Eigenschaften innerhalb der kopal-solution (DIAS).
 - Softwareumgebungen zum Extrahieren von Metadaten, für die Unterstützung der ingest und retrieval-Prozesse (koLibRI).
 - Softwareumgebungen für Migration und Emulation (über koLibRI)
 - Policies für die Langzeitarchivierung

- Testumgebungssystem und Produktivsystem für Projektpartner im Einsatz (ingest)
 - UOF-Schema in DIAS implementiert
 - Systemtechnische Installationen und Konfigurationen bei der GWDG abgeschlossen
 - Serverkonzept, Storagekonzept, Remotezugriff, Mandantenfähigkeit
- koLibRI erste Version veröffentlicht (OSS)
 - Website, Forenbeiträge

- Weiterentwicklung von DIAS-Komponenten
 - Administration Interface (AI)
- Weiterentwicklung von koLibRI tools
 - Workflow DB, migration tool
- Einbindung/Anpassung von vorhandenen workflows für digitale Objekte für ingest und retrieval Prozesse bei den Projektpartnern
- Aufnahme von Tests mit projektexternen Institutionen

- koLibRI – kopal Library for Retrieval and Ingest
 - Open Source Software
 - Java Class-Bibliothek
 - Workflow Datenbank
 - Einbettung von Jhove Framework (eventuell andere)
 - Sammlung von Metadatenextraktionsprogrammen
 - Softwaremodul zur Erstellung des generischen UOF-Formats
 - Softwaremodul zum ingest des UOF und der digitalen Objekte
 - Softwaremodul zum retrieval der digitalen Objekte
 - Softwaremodul zum Einbinden von Migrationstools

- GDZ – Göttinger Digitalisierungszentrum an der SUB Göttingen
- Produktions- und Präsentationsumgebung für Digitalisate.
- Primär TIFF – Image Produktion aus Retrodigitalisierung gedruckter Materialien.
- Produktionsmenge z.Z. ca. 6 mio TIFF Images.
- Integrierter Workflow von der Auswahl, Produktion bis hin zur Publikation vorhanden.
- Viele Projektpartner mit Datenaustauschwunsch.
- Fragestellung der Langzeitverfügbarkeit seit geraumer Zeit vorhanden.



- Entscheidung für die GDZ-Situation mit kopal-solution als Langzeitarchiv im „back end“.
- D.h. in einem Stufenmodell werden
 - 1. Die digitalen TIFF-Objekte ausgewählt.
 - 2. Die TIFF-Objekte in einem preprozess mit Hilfe von koLibRI-tools analysiert (einfache Fehler, Formatkorrektheit, ...).
 - 3. Die TIFF-Objekte mit Hilfe von koLibRI und einem Jhove-TIFF Modul analysiert und dabei die technischen Metadaten (z.B. Pixelanzahl und Dimensionen) extrahiert.
 - 4. Mit Hilfe eines koLibRI moduls werden die deskriptiven (bibliographischen) Metadaten aus dem Bibliothekskatalog extrahiert.
 - 5. Das SIP (submission information package) wird „geschnürt“

- Zu 5. Identifier vom GDZ wird übernommen, interner Identifier wird vergeben, usw.
- 6. Mit Hilfe von koLibRI wird das SIP zum DIAS ingestet.
- 7. Die Datenkontrolle erfolgt nun im environment beim technischen Host (Speichermedienmigrationen, Bit stream preservation)
- 8. Z.B. im Fall eines Verlusts der TIFF-Objekte im Präsentationssystem des GDZ wird auf das Objekt in der kopal-solution zurückgegriffen. access und retrieval des DIP (dissemination information package).

- Erfahrungen beim GDZ-workflow beinhalten:
 - Notwendig ist eine Selektions policy für die Langzeitarchivierung
 - Bereitstellungsweg der TIFF-Images möglichst einfach gestalten (CD-ROMs, Verzeichnisfreigaben)
 - TIFF-Header ist nicht gleich TIFF-Header (Standardauslegung)
 - TIFF-Produktion nicht fehlerfrei
 - Qualitätskontrolle wird auf LZA-Aspekte ausgeweitet
 - QA (Metadaten, Formatfragen) muss vor ingest erfolgt sein!
 - koLiRI-Module wurden in ihrer Stabilität bestätigt
 - Community-Arbeit für TIFF-Metadatenextraktion (Jhove) notwendig
 - Dateningest hin zum DIAS läuft problemlos

- Zum Projektende werden zahlreiche digitale Objekte von den Partnern ingestet sein (mehr Erfahrungen)
- Die Anbindung und Einbindung von projektexternen Partnern und deren Inhalte, sowie die Erweiterung der notwendigen workflows steht im Fokus. Auch zwischen LZA-Archiven
- Ein Service- und business Modell wird entwickelt
- Preservation Management
 - Migration tool wird exemplarisch entwickelt
 - File registration (für verschiedenste Formate) muss möglichst global (international) erstellt und kooperativ mit Inhalten gefüllt werden. Anbindung an diese Registry wäre dann ein Ziel der kopal-solution

- Weitere Informationen zum Projekt finden Sie unter <http://kopal.langzeitarchivierung.de/>

Vielen Dank für Ihre Aufmerksamkeit!