



Kooperativer Aufbau eines Langzeitarchivs digitaler Informationen

Der Workflow zur Speicherung digitaler Daten in einem kooperativen Modell

> Dr. Thomas Wollschläger Deutsche Nationalbibliothek











- Das kopal-Projekt: Grundlagen, Lösungskonzept, Struktur
- Einzuspielende Daten für kopal
- Die Workflows der Archivierung
- Herausforderungen an das Projekt

#### Auftrag der DNB



- Rasch wachsender elektronischer Publikationssektor
- Zunehmender Teil der kulturellen Aktivitäten / Publikationen aller Art im Netz
- Dieser "Nachlass" ist auch gesetzlich fixiert Aufgabe von Nationalbibliotheken und vergleichbaren Einrichtungen
  - → Gesetzesnovellierung in Deutschland 2006: Sammelauftrag für nicht-trägergebundene digitale Materialien ("unkörperliche Medienwerke"), bezieht sich auch auf Web-Archiving

#### Das kopal-Projekt



- Förderprojekt des Bundesministeriumes 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
  - Deutsche Nationalbibliothek (DNB) (Gesamtprojektleitung)
  - Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
  - Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
  - IBM Deutschland GmbH (IBM)

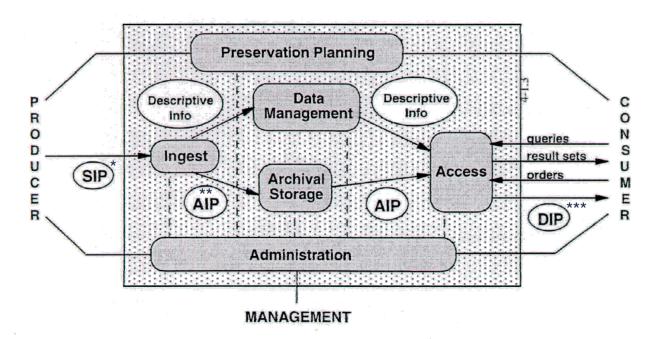


- Kernssystem DIAS von IBM
  - Basiert auf bewährter Standardsoftware
  - Zukunftsfähige Trennung von Speicherkonzept und Datenverwaltung
  - Erfahrungs- und Technologieaustausch mit der Nationalbibliothek der Niederlande und anderen internationalen Partnern
- Lokale Software koLibRI von DNB und SUB
  - Generische freie Softwarebibliothek zur Ein- und Anbindung unterschiedlicher Mandanten
  - Automatische Erstellung von technischen Metadaten
  - Konfigurierbare Workflows
- Universelles Objektformat
- OAIS-Referenzmodell konsequent implementiert

#### OAIS Modell als Grundlage



- Funktionale Anforderungen, auf denen DIAS beruht:
- Die relevanten Business Prozesse basieren auf dem OAIS Modell (=Reference Model for an "Open Archival Information System" – ISO-Standard ISO 14721)
- OAIS dient als Framework für die DIAS-Lösung
- Der Kern von DIAS besteht aus 6 Prozessen, die mit Elektronischen Publikationen (EPs) in den drei Standardformaten arbeiten:
  - Ingest Process
  - Archival Storage
  - Data Management
  - Access
  - Administration
  - Preservation

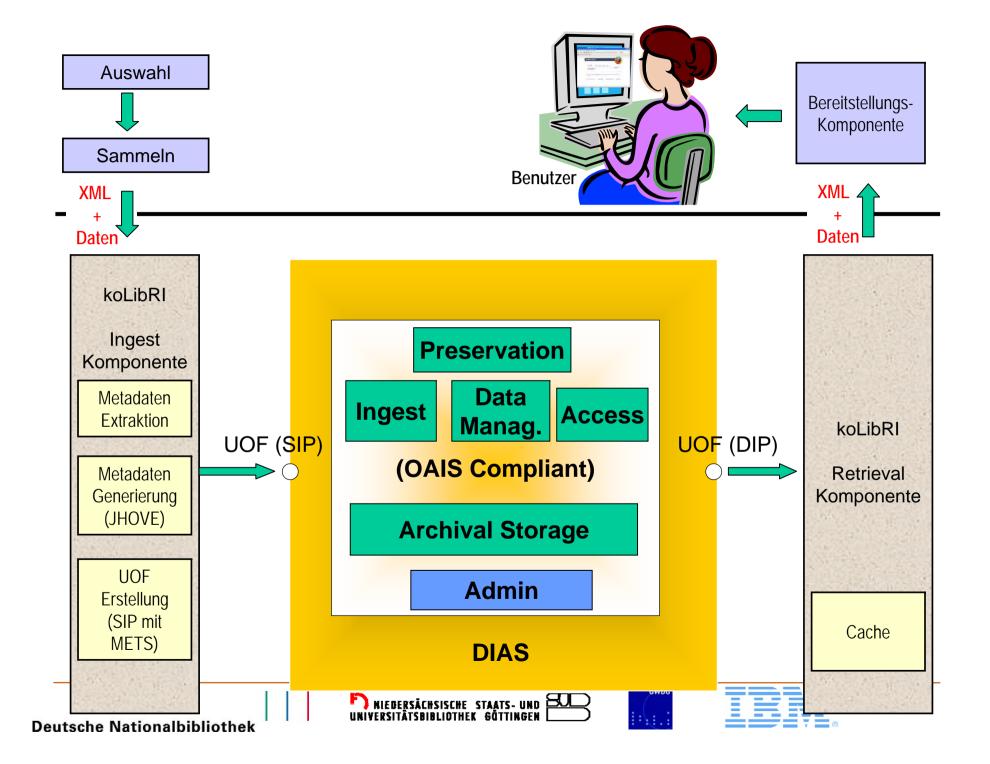


<sup>\* =</sup> SIP: Submission Information Package

Figure 4-1: OAIS Functional Entities

<sup>\*\* =</sup> AIP: Archival Information Package

<sup>\*\*\* =</sup> DIP: Dissemination Information Pachage

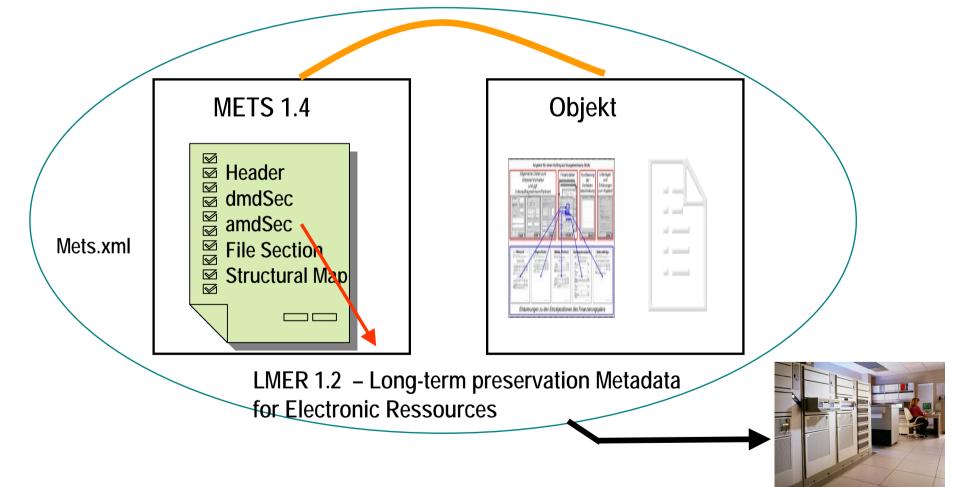


## Erstellung der Archivobjekte



#### **Universelles Objektformat**

**Submission Information Package** 



#### Beispiel für mets.xml in kopal



```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<mets OBJID="" TYPE="Kopal Submission Information Package" PROFILE="DDB" xmlns="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink"</p>
  xmlns;xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns;mets="http://www.loc.gov/METS/" xmlns;imerFile="http://www.ddb.de/LMERfile"
  xmlns:lmerObject="http://www.ddb.de/LMERobject" xmlns:lmerProcess="http://www.ddb.de/LMERprocess" xsi:schemaLocation="http://www.loc.gov/METS/
  http://www.loc.gov/standards/mets/mets.xsd http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version14/mets.xsd
  http://www.ddb.de/LMERfile http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/LMERobject http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/LMERobject http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/LMERobject http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/standards/lmer-file.xsd http://www.ddb.de/sta
  object.xsd http://www.ddb.de/LMERprocess http://www.ddb.de/standards/lmer/lmer-process.xsd">
  <metsHdr CREATEDATE="2006-01-12T14:50:16" RECORDSTATUS="PRODUCTION">
  - <agent ROLE="ARCHIVIST" TYPE="ORGANIZATION">
        <name>Die Deutsche Bibliothek</name>
        <note>Automatisch generierte Metadaten. Für weitere Informationen: kopal.langzeitarchivierung.de</note>
   </metsHdr>
- <amdSec ID="AmdSec-0001">
   - <techMD ID="TechMD-LMER-Object">
     - <mdWrap ID="TechMD-LMER-Object-MdWrap" MIMETYPE="text/xml" MDTYPE="OTHER" OTHERMDTYPE="ImerObject" LABEL="LMERobject">

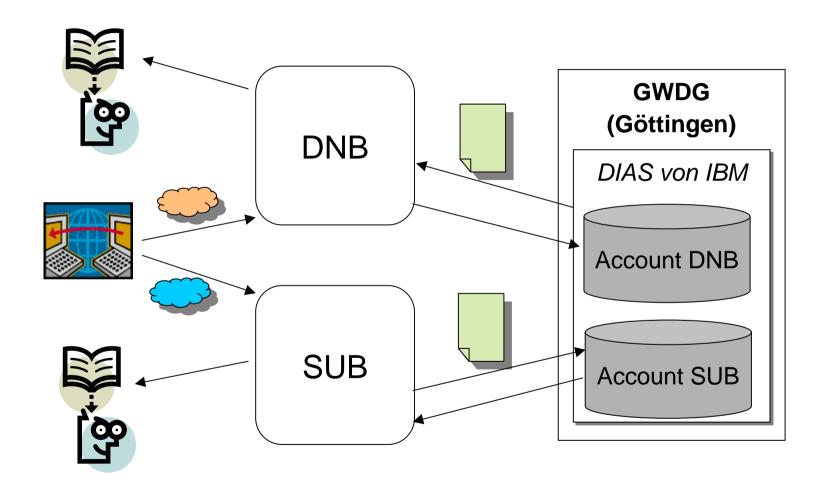
<
              </
              <lmerObject:objectVersion>1</lmerObject:objectVersion>
              dmerObject:masterCreationDate>2006-01-12T14:50:14//merObject:masterCreationDate>
              <merObject:metadataCreationDate>2006-01-12T14:50:14</merObject:metadataCreationDate>
              <lmerObject:metadataRecordCreator>KOPAL DIAS
              <lmerObject:numberOfFiles>3</lmerObject:numberOfFiles>
           </miData>
        </mdWrap>
     </techMD>
  + <techMD ID="TechMD-File-1">
  + <techMD ID="TechMD-File-2">
  + <techMD ID="TechMD-File-3">
   </amdSec>

    <fileSec>.

  - <fileGrp ID="ASSET" ADMID="TechMD-LMER-Object">
     - <file ID="FILE0001" MIMETYPE="application/pdf" SIZE="330118" CREATED="2006-01-12T14:50:16" CHECKSUM="17673a8df8fc8ddd101fc0fbd6ed69d416608571"</p>
           CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-1">
           <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="file://./d1/973908904.pdf" />
        </file>
     - <file ID="FILE0002" MIMETYPE="text/plain; charset=US-ASCII" SIZE="48" CREATED="2006-01-12T14:50:16"</p>
           CHECKSUM="e54dbd60ec4507768eb7e4a663e19c1045e38595" CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-2">
           <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="file://./d1/ddbmd5.txt" />
        </file>
     - <file ID="FILE0003" MIMETYPE="application/octet-stream" SIZE="13531" CREATED="2006-01-12T14:50:16"</p>
           CHECKSUM="4507fffd6b0a47bc2d4e80d1b132776359e77db4" CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-3">
           <FLocat LOCTYPE="URL" xlink:type="simple" xlink:href="file://./meta.txt" />
        </file>
     </fileGrp>
   </fileSec>
```

# Projekt KOPAL: Mandantenfähigkeit





#### Verteilte Speicherung





#### Einzuspielende Daten (Projektantrag; hier: nur für DNB))



- Online-Dissertationen
  - Anzahl: derzeit > 50.000
  - Originär digitale Dokumente ab 1997
  - Datenmenge: ~ 350 GB
- Netzpublikationen (SpringerLINK)
  - Serials/Monografien, Elektronische Zeitschriften,
  - Datenmenge: ~ 300 GB
- Exemplarische CD-ROMs / DVDs
  - Anzahl: 50.000 100.000
  - Quellensammlungen, Multimedia-Anwendungen, Filme, Spiele, Datensammlungen, sonst. Publikationen
  - Datenmenge: ~ 30.000 GB 60.000 GB



- Digitalisate:
  - Exilpresse Digital + Jüdische Zeitschriften in NS-Deutschland
    - Datenmenge: ~150 GB
  - Digitalisate des Deutschen Buch- und Schriftmuseums
    - Datenmenge: (vorerst) ~10.000 GB
  - Material der HAB Wolfenbüttel
    - Datenmenge: ~ 1.500 GB bis ~12.000 GB
- Digitale Musikdaten (DMA):
  - Datenmenge: bis zu ~ 544.000 GB (von Schellackplatten bis Notendigitalisate)
- Sonstiges:
  - E-Mail Newsletter (Datenmenge vernachlässigbar)

#### Schnittstellen in DNB I: Ingest- / Access-Workflow



- Bisher: "halbautomatisches" Verfahren; verschiedene Workflows für Netzpublikationen, Online-Dissertationen, …
- Ziel: automatisches, einheitliches Verfahren; Übergabe an kopal (Ingest)
  - bzw. beim Access
- Übergabe <u>von</u> kopal an Arbeitsplatzrechner oder Bereitstellungssystem
- Wer: (Pica-)ILTIS-Team, Portal-Gruppe, Erwerbung/FE, Benutzungsabt., Ablieferer (hier: Unterstützung nötig)

### Workflow-Management in der DNB (II)



- Workflow von elektronischen Materialien auf Datenträgern
  - Erstellung ausgewählter repräsentativer CD/DVD-Images inkl. Problemanalysen und Einspieltests
  - Änderungsvorschlag für den GG dieses Materials vorgelegt; Workflow wird erzeit überarbeitet
- Workflow für (online-only) Netzpublikationen
  - Einbeziehung der Anforderungen der LZA in Gestaltung des neuen Workflows für NP
  - Design des Workflows auf die Schnittstellen des Archivsystems

#### Workflow-Management in der DNB (III)



- Fortlaufende Publikationen (v.a. Zss-Artikel)
  - Archivobjekte ggf. □Katalog-Einträge → bibliografische Metadaten von Archivobjekten müssen in ILTIS (Pica) abgebildet werden
  - Festlegung von Erschließungsvarianten
  - Mapping von Archivobjekten auf Katalogobjekte
- Erweiterung des URN-Mangements
  - Voraussetzung: Jedes Objekt braucht zum Einspielen einen URN
  - Retro-Vergabe f
    ür bereits archivierte Objekte ohne URN
  - Alle neuen Objekte müssen mit URN geliefert werden bzw. bei Eingang/Bearbeitung einen URN erhalten
    - = künftiges Verfahren

### Schnittstellen in DNB III: Export der Objekte zum Nutzer kopa



## Rechtemanagement

- Anforderungen müssen noch formuliert werden
  - 4-Varianten-Modell (ggf. 5, jedoch erst in Planung):
    - Inhouse anzeigbar ("Domain")
    - Inhouse + registrierte Externe ("Domain+")
    - Auch extern anzeigbar
    - Gar nicht anzeigbar
  - Information wird in (PICA-)ILTIS hinterlegt; Einträge bereits z.T.
     vorgenommen (jedoch vorerst noch ohne Domain+)
- Derzeit in Erstellung: Nutzungsoberfläche inkl. benötigter Schnittstellen im Rahmen des DNB-Portals

### Herausforderungen I: Übergang in den Regelbetrieb



- Voraussetzungen liegen vor bzw. werden geschaffen:
  - Produktionssystem wurde aufgesetzt und läuft
  - Produktives Einspielen von Material wurde und wird erprobt
  - Nötige Weiterentwicklungen (z.B. fehlende Module) wurden und werden ermittelt
  - Änderungen in diversen Workflows wurden angestoßen
- Aufgaben:
  - Übergabe des kopal-Systems vom Projektteam an ständige Arbeitseinheit
  - Management von (ggf. rechtzeitigem) Einspielen aller Retro-Objekte und des aktuellen Zugangs (direkt nach kopal)

### Herausforderungen II: Ingest und Zugriff



- Rechtzeitige Erhöhung des Speichervolumens von kopal zur Aufnahme aller Retro-Objekte und für den aktuellen Zugang
- Angesichts großer Datenmengen (künftig auch: Web-Archiving!) + großer Einzelobjekte (z.B. Digitalisierte DVDs mit ~8 GB):
  - Gewährleistung einer ausreichenden Performanz des Systems
  - Schnelle Internet-Verbindung(en); Anbindung Leipzig
  - Benutzerbetreuung (z.B. für Wartezeiten, rechtliche Beschränkungen...)

#### Derzeitige kopal-Veröffentlichungen





#### k o L i b R I

kopal Library for Retrieval and Ingest

Dokumentation

© Projekt kopal,

Deutsche Nationalbibliothek /
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Stand: Februar 2007

- Release von koLibRI 0.7
   Beta
- Kopal-"Kundenpapier"
  - Servicekomponenten von kopal
  - Nutzungsmodelle
  - Kostenfaktoren bei einer Nutzung
  - Exemplarische
     Kostenkalkulation

#### Informationen zu kopal



 Für weitere Informationen zum Projekt, den Standards und für Downloads von Dokumentationen

(inklusive "kopal-Demonstrator"):

http://kopal.langzeitarchivierung.de

- Ansonsten gerne Fragen an kopal bei der Deutschen Nationalbibliothek oder an die anderen Projektpartner
  - t.wollschlaeger@d-nb.de

Vielen Dank für Ihre Aufmerksamkeit!