Digital Long-Term Archiving at GWDG and other archiving systems

Dagmar Ullrich

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Am Fassberg, 37077 Göttingen

Fon: 0551 201-1827 Fax: 0551 201-2150
dagmar.ullrich@gwdg.de wwwuser.gwdg.de/~dullric/

Storing and Preserving Data

Two different tasks of digital Long-Term Archiving

- A digtial archive must do both -

Long-Term Preservation

Providing the long-term usability of the data content.

Bitstream Preservation

Storing data over large periods of time.

Securing the integrity of the „Bits n Bytes„.

The main task of GWDG lies in the field of Bitstream Preservation.

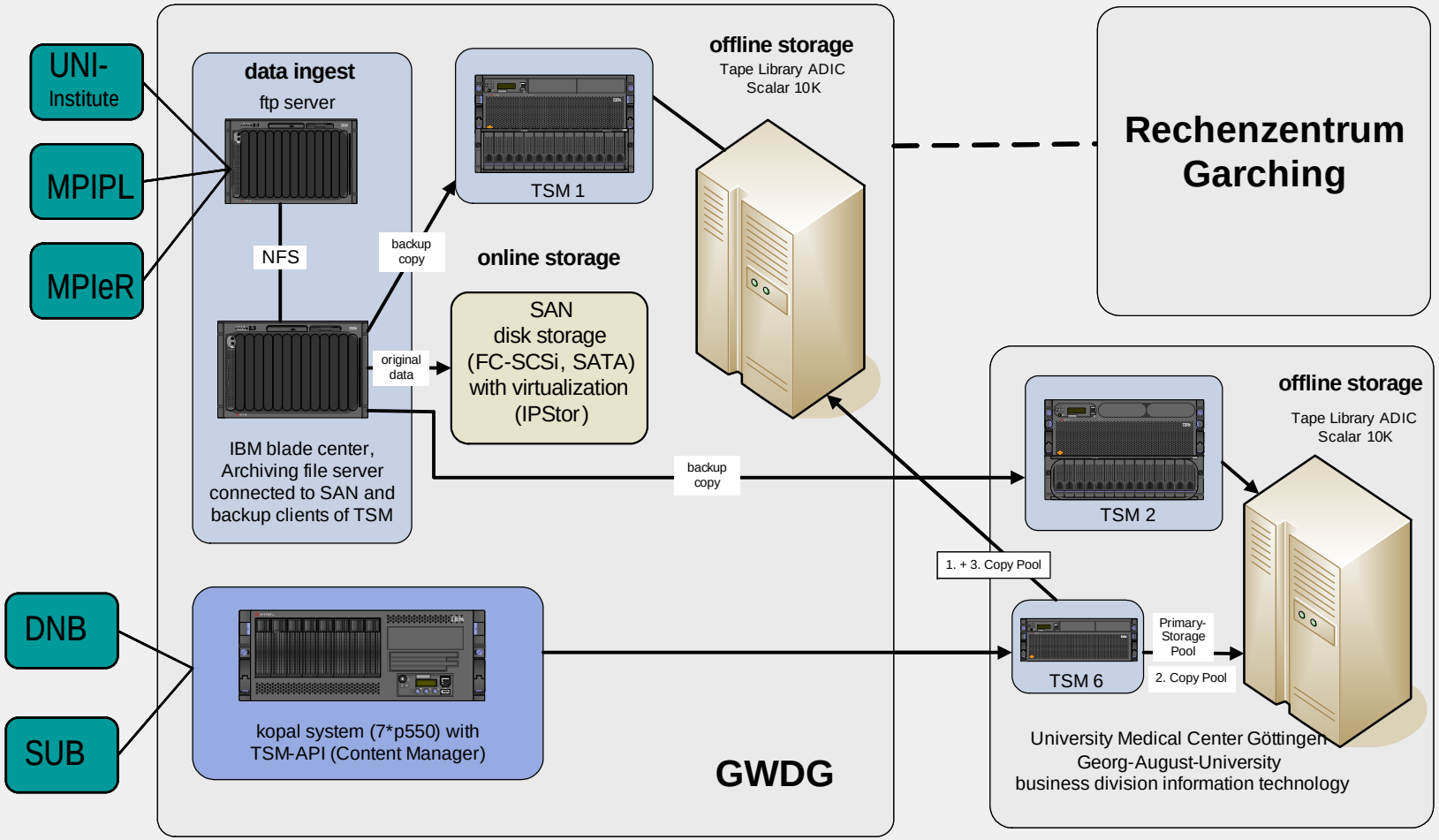
Tasks of Bitstream Preservation

- Evaluating storage media and technologies
  - life expectancy, cost, performance
  - long life expectancy of media is not sufficient, since the access technology also has a limited life expectancy.
- Developing an appropriate storage strategy considering properties, conditions and life cycles of data
    - size and amount of single files
    - data/file-structure
    - ingest, change and access needs
    - data integrity and authenticity
    - redundancy and security (different locations for storage)
- Technology shifts
  - Highly complex migration processes (often for large amounts of data) must be planned, proceeded and terminated.

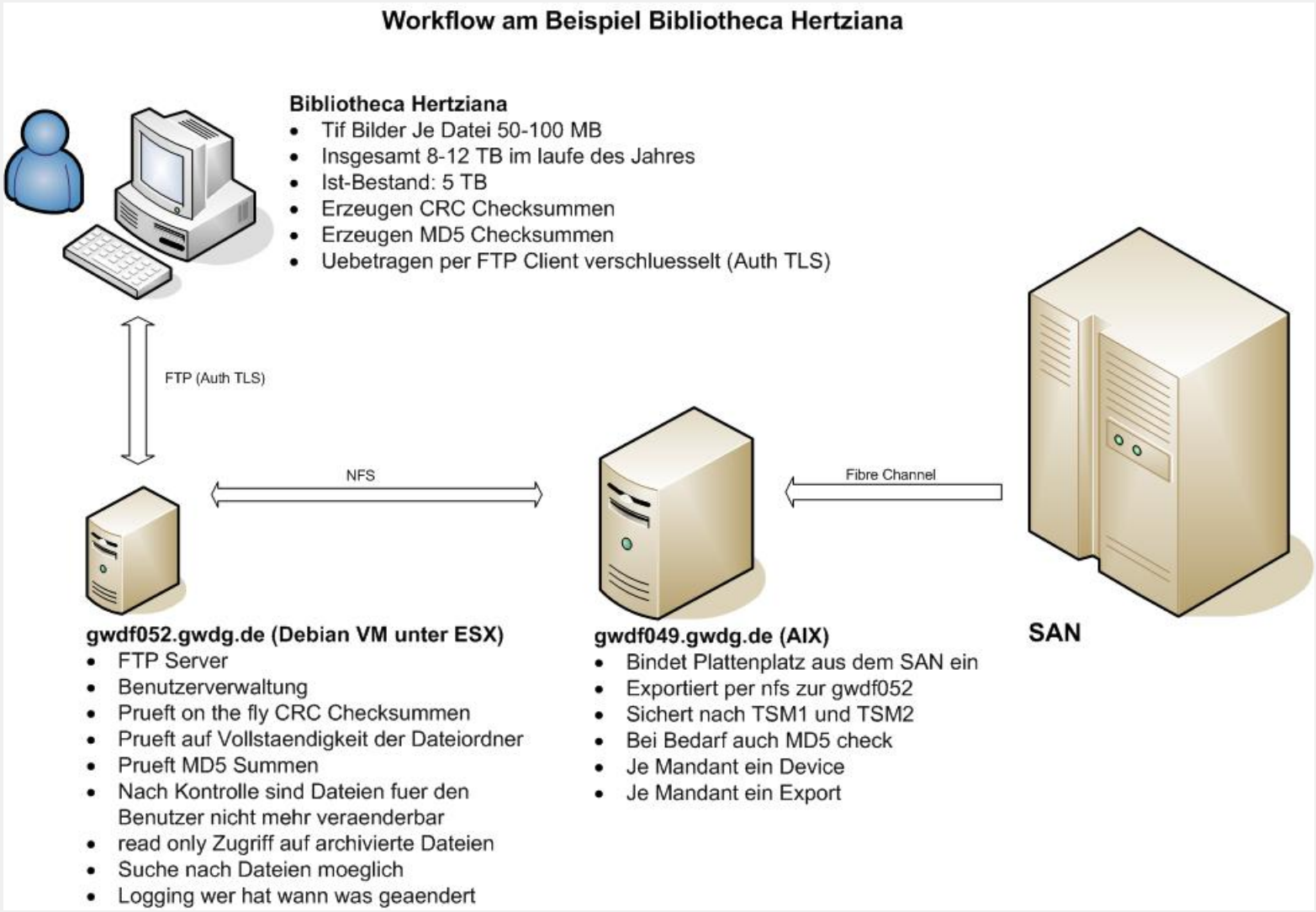
Digital Long-Term Archiving at GWDG

- Bitstream Preservation for MPIs and University Göttingen
  - The President of the MPS put GWDG and RZG (Rechenzentrum Garching) in charge of bitstream preservation for scientific data from MPIs.
  - Data includes world cultural heritage
    -Audio and video files of endangered languages
    -Digitized literature and images of architecture
- Safeguarding good scientific practice
  - Storing data over 10 years (system of MPIbpc)
- Long-Term Preservation
  - Hosting of kopal system (currently for DNB / SUB) „Co-operative Development of a Long-Term Digital Information Archive”

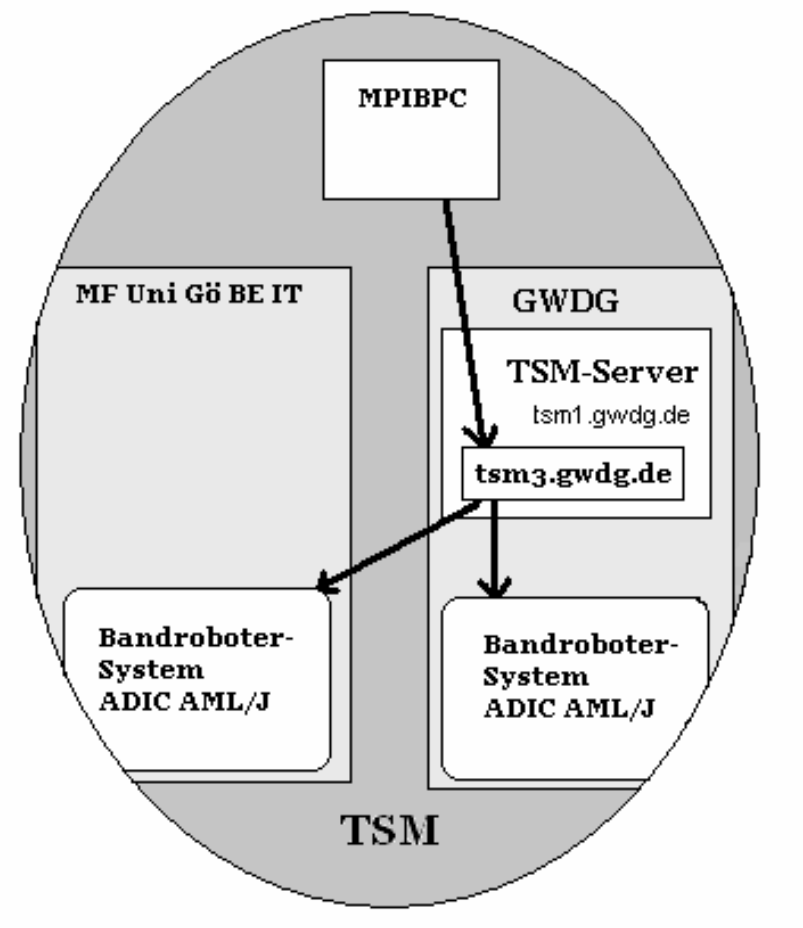
dLTA – infrastructure at GWDG



Bitstream Preservation for MPIs



Safeguarding good scientific practice GWDG / MPI f. biophysical chemistry



-Archiving of data for a 10 years time periode (Safeguarding good scientific practice)

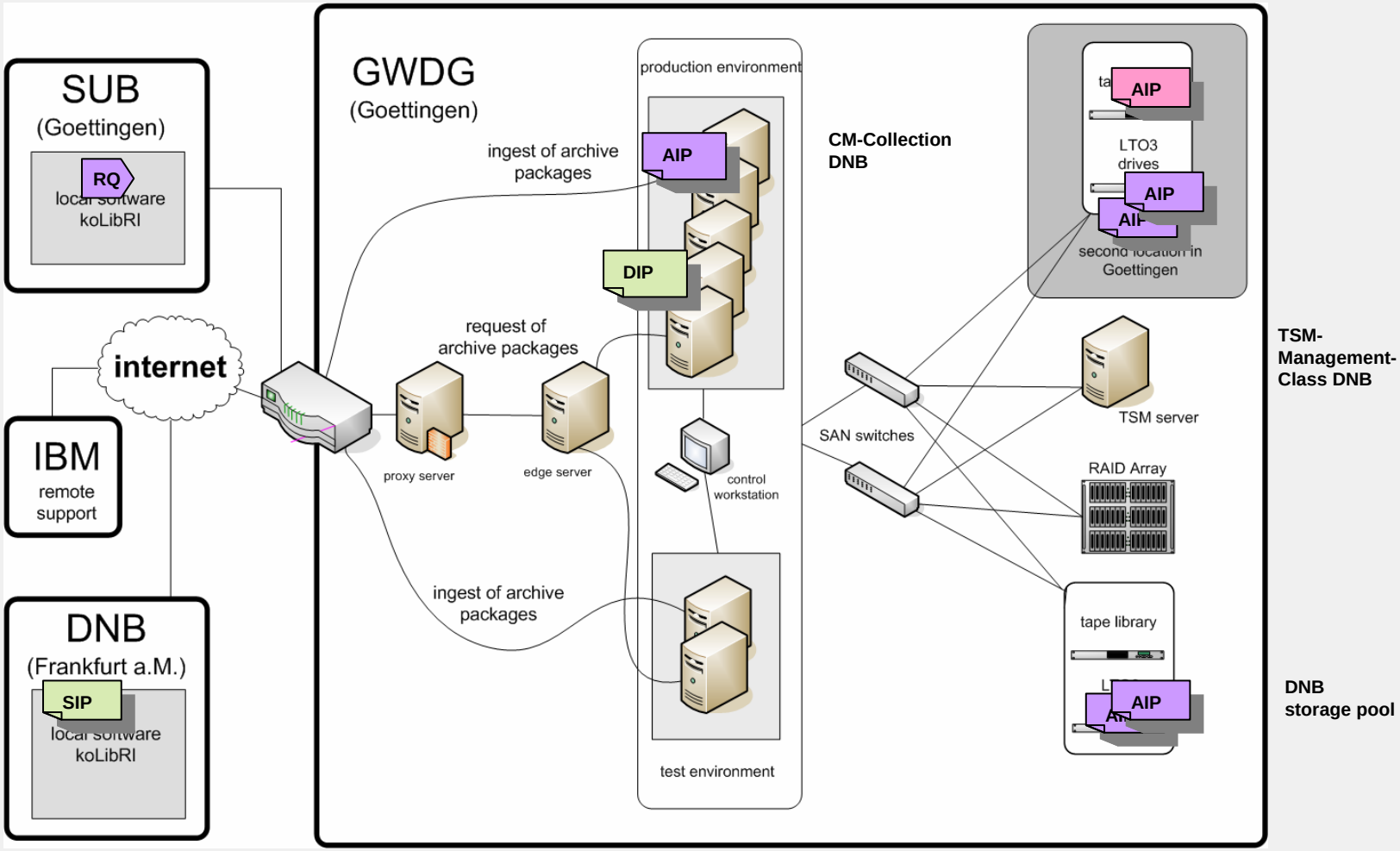
-Redundant data storage at two different locations in Göttingen

-MD5 checksum

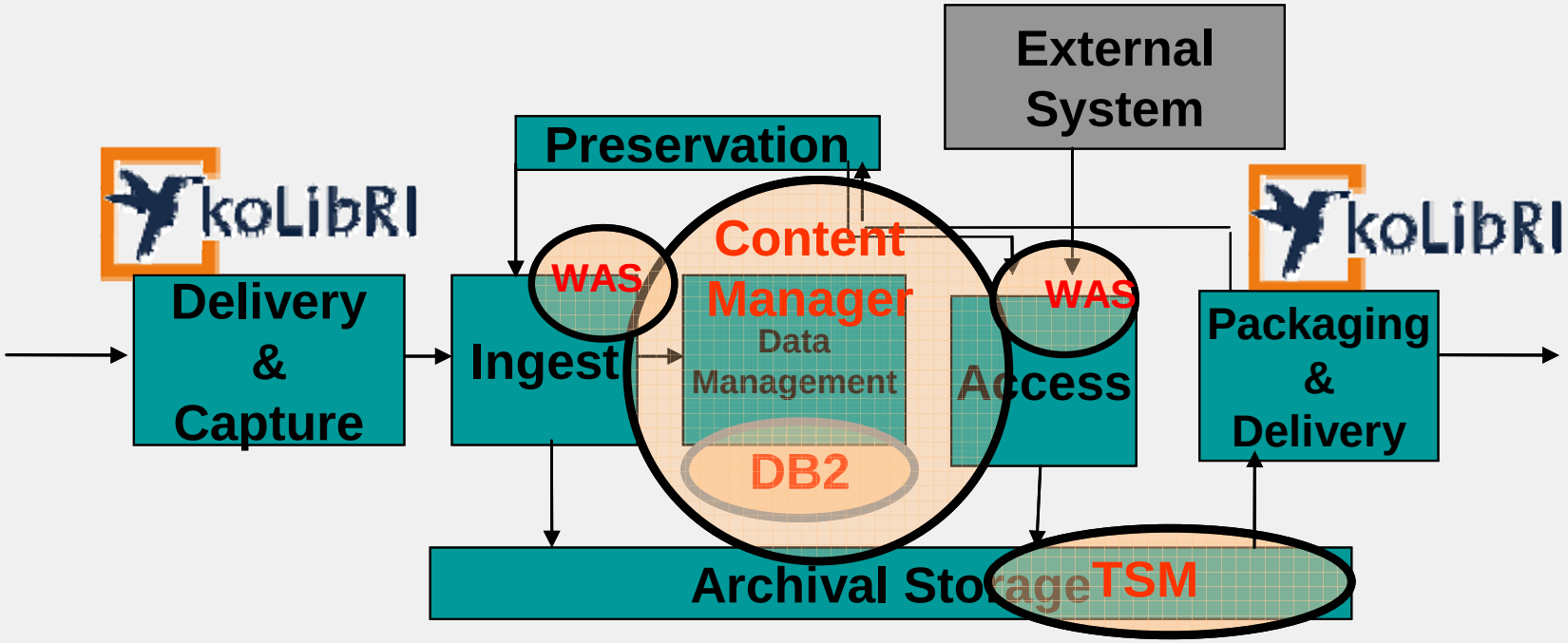
-HTML-Interface

-File with content describing metadata

kopal-System at GWDG: package workflow



DIAS and koLibRi Software



Other archiving systems

- Bitstream-Preservation (Archival Storage):
  - LOCKSS
  - Distarnet
- Long-Term Preservation
  Systems with Ingest, Retrieval, Data Management and (hopefully) components that support long term preservation
  - BABS
  - Portico
  - eDoc-Server Humboldt Universität Berlin
  - DSpace
  - Fedora
  - MyCoRe

Bitstream Preservation: Archival Storage

- **LOCKSS** (www.lockss.org) **Lots of copies keeps stuff safe**
  - Stanford University, LOCKSS Alliance (user community of LOCKSS)
  - P2P-Architecture
  - Redundant data storage at many different locations each supplied with a LOCKSS node
  - Erroneous data is identified by comparison with content of other nodes and automatically replaced
  - Open Source
  - Mediamigration by automatic replacement of data if a node is unavailable

- **Distarnet** (www.distarnet.ch) **Distributed Archival Network**
  - Universität Basel, imaging & media lab
  - P2P-Architecture
  - Redundant data storage at many different locations each supplied with a distarnet node
  - Erroneous data is identified by comparison with content of other nodes and automatically replaced
  - Data encryption
  - Licensing
  - Mediamigration by automatic replacement of data if a node is unavailable

BABS (www.babs-muenchen.de)

Bibliothekarisches Archivierungs- und Bereitstellungssystem

- Bavarian State Library / Leibniz-Rechenzentrum
- Components:
  - DigiTool (DB: Oracle) from Exlibiris for asset management
  - TSM-HSM for archival storage (redundant at LRZ/RZG)
- dLTA Features:
  - Persistent Identifier (URN)
  - dLTA relevant Metadata used
  - Emulation and migration support

Portico (www.portico.org)

- JSTOR, Library of Congress,
  Andrew W. Mellon Foundation,
- Not a system for „self hosting“, usage against fee
- Components
  - Not known
  - archival storage redundant at different locations
- dLTA Features
  - dLTA relevant metadata used, produced at ingest
  - Format normalization at ingest +
    keeping of original file
  - Support of further migrations processes planned

eDoc-Server (edoc.hu-berlin.de)

- Humboldt University Berlin
- Components:
  - Java/PHP, Tomcat, Apache, Sybase DB
  - Archival Storage: 5 copies on disk an 4 different locations in Berlin + 2 TSM backup copies at 2 different locations in Berlin
- dLTA Features
  - Convertion of MS-Word, PS or Latex to XML as archiving format, PDF and HTML for presentation

DSpace (www.dspace.org)

- MIT Libraries, Hewlett-Packard
- Components:
  - Java/JSP using the Java Servlet Framework (Apache/Tomcat) and a relational data base (PostgreSQL or Oracle)
  - Archival storage file system based
- dLTA Features
  - Persistent Identifier (Handle System CNRI)
  - Different levels for file formats (supported, known, unknown)

MyCoRe (www.mycore.de)

- Universität Essen and MyCoRe Community
- Components:
  - Java/XML/XSL technologies can be combined with different backend systems. Open Source backends based on MySQL und Apache Lucene or commercial backend systems like Oracle, IBM DB2 und IBM Content Manager are possible.
  - Archival Storage: depends on used backend, TSM combines well with Content Manager of IBM
- dLTA Features
  - Persistent Identifier (URN)

Fedora (www.fedora-commons.org)

Flexible Extensible digital Object and Repository Architecture

- Cornell University, The University of Virginia
- Components:
  - Java-Technology (Open Source) / Apache /Tomcat with MySQL, Oracle, McKoi as possible data bases
- dLTA Features
  - Persistent Identifier (PDI)

Thank you for your attention.

Any Questions?