Embedding grid based preservation features into legacy environments

SHAMAN: Sustaining Heritage Access through Multivalent ArchiviNg

Claus-Peter Klas, Prof. Matthias Hemmje, Holger Brocks
FernUniversität in Hagen
iRODS Workshop
SHAMAN Mission

- Provide a vision and rationale to support a comprehensive Theory of Preservation that may be used to store and access potentially any type of data, based on the integration of digital library, persistent archive, and data management technologies.
- Supply an infrastructure that will provide expertise and support for users requiring the long-term preservation and reuse of data over a decades-long time span.
- Develop and implement a grid-based production system that will support the virtualization of data and services across scientific, engineering, document, and media domains. Identifying Content and Capturing of Context Demonstrate Distributed Ingestion
Representative for Work Packages

- WP 5: Data Grid Implementation
  - D-GRID
  - iRODS
  - Integration of Legacy Systems and Tools
- WP 8: Managing Shared Collections
  - Data Analysis and Advances Access
  - KOPAL within SHAMAN
  - Management
- WP 11 aka ISP 1: Document Production, Archival, Access and Reuse in the Context of Memory Institutions for Scientific and Governmental Collections
- Long-term Preservation (OAIS compliant) & Digital Libraries
Outline/Technical Challenges

- Distributed Ingest Process
  - Sources
  - New setup
  - Migration
- Embedding Legacy Software
  - KOPAL (DNB)
- Management tools based on DAFFODIL framework
- Open questions/discussion
Ingest: Sources of information

- File (System)
- OAI (Harvesting)
- Databases
- Legacy systems
  - KOPAL
  - DSpace
  - Fedora
  - Eprints

- Bulk load || Convert a source into a iRods node?
Ingest: New Setup

- No preservation environment given, only backup (hopefully)
- Fetch data from some source
- Run ingest process
  - Move/Copy information objects
  - Index process*
- Verification / Acknowledge incl. log
Ingest: Migrate

- Existing preservation environment
- Fetch data from source
  - Replace existing system -> New Setup
  - Keep system as main source
    - Run ingest process
    - Create link (URI) between original system and GRID
    - Move/Copy files, if wanted
    - Index process, maybe only on temporary data*
- Verification / Acknowledge incl. log
Index process*

- Metadata
  - Technical
  - Information specific
- Deep structure
  - Text:
    - IR functionality
    - Structure unstructured text
- Multi-media: images, sound, video
- Structured data: forms, XML documents
Management:
Quality/Requirements

- Quality: always online, must be always restoreable.
- Requirements: deep indexing, only metadata.

- Cost function: time for storage (law 10 years, forever).

User: What to pay for what quality?
Repository holder: What costs will arrive?
Management: Quality/Requirements

- # of copies (storage security)
- Always online / only restore
- Deep indexing / only metadata
- Time for storage (law 10 years, for ever)

- Design of a cost function
  - User: What to pay for what quality
  - Repository holder: What costs will arrive
DAFFODIL Framework
DAFFODIL Framework

- Service oriented architecture
- Graphical tool based user interface
  - Java Swing based
  - Web 2.0 Ajax based
DAFFODIL Framework

- Presentation Representation
  - DAFFODIL as management tool for
    - collection management
      - Ingest
      - Monitoring
    - presentation representation
  - DAFFODIL as Web 2.0 frontend for
    - information access & reuse
    - embedding multivalent access
    - information visualisation
Legacy Software: KOPAL

Deutsche Nationalbibliothek (Frankfurt)
- local software

SUB Göttingen
- local software

GWDG (Göttingen)
- DIAS by IBM
  - account 1
  - account 2
KOPAL Integration: Replace

iRODS

Ingest all data

Legacy System KOPAL

Ingest

Access only iRODS

Storage iCAT Indexer
KOPAL Integration: Only index

iRODS

Ingest

Legacy System
KOPAL

Storage
iCAT
Indexer

Ingest metadata + URI

Access unified iRODS
+ fetch original document
from legacy system
KOPAL Integration: iRODS node

Ingest object

Storage
iCAT
Indexer

Ingest

Store object in KOPAL

Access unified iRODS
+ fetch original document
from legacy system

Legacy System
KOPAL

iRODS

iCAT
Indexer

iRODS

Legacy System
KOPAL

Ingest

iRODS

Ingest object

Store object in KOPAL

Access unified iRODS
+ fetch original document
from legacy system
Open questions

• How to integrate a running legacy system with iRODS?

• Management:
  • How to we manage the information stored?
  • How do we prove, that the information is secured for the next 100 years?

• Is there already a cost function?
  • For the user: Estimate payment
  • For the service provider: Estimate cost of GRID