Digital Repository Infrastructure Vision for European Research II”

Technology Watch report D4.3
part 1: report on repository interoperability
part 2: long term preservation technologies

UGent, DTU, SURF, UKOLN, UGoe
OUTLINE

- Demonstrator for long-term preservation archiving of Enhanced Publications: Eugène Durr (3TU)
- Long-term preservation archiving aspects and archiving agreement: Barbara Sierman (KB)
Academic research

- Research information
- Publications
- Research data
NL-Stakeholders

Surffoundation

- Research universities
- Universities of applied arts/sciences
- KNAW
- KB, national library
- Digital Archive and Networked Services
- 3 TU (Technical Universities)
Content of academic repositories network

- Over 160,000 full-text publications (of which 20,000 doctoral e-theses)
- Publications harvested and stored for LTP by KB
- More than 2,000 datasets [humanities-social sciences] (DANS)
- 3TU under development [sciences/technology]
- Unknown were Dutch medical research data is stored for LTP
DARE 2 DRIVER

DARE (Digital Academic REpository infrastructure) programme (2003 - 2006)
- Infrastructure and Services for open access of (text)publications (including LTP)
- Knowledge dissimination: publication process

DRIVER: Digital Repository Infrastructure Vision for European Research
- Introduction of a repository network for open access of academic publications in Europe
- ...

DRIVER II:
- Expansion of the network to more European countries
- Infrastructure and Services for open access of publications + data (including LTP
- ...

Workshop Driver II, WP4.3 M4.2 (LTP-demonstrator, The Hague, 2009-04-23)
Overview of the latest technological developments in the world of
‘Digital Repositories, Digital Libraries and beyond’
To provide theoretical and practical input for DRIVER, especially with regard to Enhanced Publications'
"Enhanced publications are envisioned as compound digital objects which can combine various heterogeneous but related web resources. The basis of this compound object is the traditional academic publication. This latter term refers to a textual resource with original work which is intended to be read by human beings, and which puts forward certain academic claims. [...] Enhancing a publication involves adding one or more resources to this ePrint. These can be the resources that have been produced or consulted during the creation of the text. [...]"

Schematische voorstelling van EP

http://driver2.dans.knaw.nl/demonstrator/html/index.html
WP 4.3 Enhanced Publications

1. Interoperability
   Packages-formats (containers); open formats

2. CRIS en digital repositories
   Integrated systems; component structure; infrastructure

3. GRID-interaction
   Distributed data, distributed processing

4. Long Term Preservation
   Digital Preservation is way of managing risks of digital objects to keep them accessible and usable over the years. We are starting to know how to tread a single object, but related objects are a new challenge
Some issues

- **Persistent Identifier**
  - NBN + resolver infrastructure
  - Digital Author Identifier (DAI)
    - Needed: link between Author and DAI
    - Enables integration of systems
    - Avoids name ambiguity and variants

- **Container**
  - DIDL
  - Dublin Core/MODS

- **Repositories**
  - By libraries
  - By authors
  - Combined with CRIS

- **Publication**
  - Ownership
  - Copyrights
  - Maintenance/flexibility/fluid publications
Dutch LTP landscape

- Publications (text+)
- KB
- Primary data (humanities)
- DANS
- Primary data (technology)
- 3TU/DELFT
IR harvester A, D, E, F= temporary storage BER = error recovery TAM = Tivoli Access Manager

DROID-service

file format controle Post-office preprocessing

Directory watcher Virus check

batchbuilder Controle fileId, Controle versionId ingest

DIDL2 EWTIJ

BER DIAS

dependent on MDObatch rights extractie

EWTIJ2 MDObatch Extra BU metadata

eDEPOT update MDO = metadata database

A, D, E, F = temporary storage

BER = error recovery

TAM = Tivoli Access Manager

DARE-proces, 2008-03-01

Workshop Driver II, WP4.3 M4.2 (LTP-demonstrator, The Hague, 2009-04-23)
Objects are characterized (KB-K) and this information is added to the metadata-xml-file

KB harvester (KB-H) retrieves the KB-DIDL-set at the EP-R, parses the DIDL and retrieves the objects. The EP-XML stays unaltered. The whole DIDL is converted into a metadata-xml-file for further processing

Objects are harvested his part of the objects (eg datasets) from the repository and processes them.

This is an existing process and could be left unaltered

Batchbuilder checks file- and file version information and assigns a unique ID for the whole batch with metadata and objects. It transforms it into a SIP.

SIP is ingested in DIAS

Schema for the processing of EP’s for LTP, 01-03-2009

Workshop Driver II, WP4.3 M4.2 (LTP-demonstrator, The Hague, 2009-04-23)

Two repositories (IR). The EP in IR-1 describes a publication in IR-2 and a dataset in IR-1 dataset in IR-1.

The EP-transformer (EP-T) harvests the EP-surrogate (ORE), decides according to a list of semantic types, which objects in the EP are dedicated for neither KB or an other LTP-A, creates a reference to them in a DIDL (an different DIDL for each LTP-A); includes the complete EP-XML-file(s) in each DIDL-container, and places those DIDL’s in different sets (one for each LTP-A) in a repository (EP-R)