UPBox and DataNotes: a collaborative data management environment for the long tail of research data

Cristina Ribeiro, FEUP—UPorto & INESC LA

Based on: João Rocha da Silva, José Barbosa, Mariana Gouveia, Cristina Ribeiro and João Correia Lopes. UPBox and DataNotes: a collaborative data management environment for the long tail of research data, International Conference on Preservation of Digital Objects, iPRES 2013, Lisboa. http://purl.pt/24107/1
Contents

• Motivation and goals
• Current practice in data curation
• A data management workflow
• Proposed architecture and tools
• Conclusions and ongoing work
Research data

• Definition (OECD, 2007)
  – “research data” are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated
  – Records for scientific activity exist in many domains and have diverse nature
Data Repositories

• Why build them?
  – Data are crucial evidence for scientific production
    • Associate data with publications
  – Collaborative work requires data sharing
  – Funding agencies require data deposit
  – Data reuse in different contexts

• But also
  – Papers with associated datasets get cited more often
Data Repositories

• Why is it hard to build them?
  – No short-time benefit from the preparation effort
  – No guarantee that repositories will persist
  – No easy funding for curation
  – Projects are temporary structures, data gets lost
  – Data curators may become the bottleneck for the curation process
Current data curation workflow
UPBox and DataNotes: a collaborative data management environment for the long tail of research data

**Proposed workflow**

**Research Group**
- Collaborative deposit and annotation of research datasets

**External Researcher**
- Easy sharing with authorized personnel
- Faster data exchange
- Supports inter-institution research collaboration

**Repository**

**Curator**
- Validation of datasets for deposit
Collaboration and user engagement

• Use cases collected from our researcher pool
  – Search for datasets
  – Share datasets in a project environment

• Solution
  – Combine a private cloud with a semantic wiki
  – Offer easy upload and private sharing of datasets
  – Offer flexible metadata models
UPBox and DataNotes: a collaborative data management environment for the long tail of research data

Architecture

- File & Folder Upload, Delete
- File System Hierarchy
- RDF Metadata Export
- Folder Structures
- On-the-fly compression
- Access Control
- Collaboration
- Metadata storage
- Deposit workflow, embargo specification...
- Repository

- Directory structure annotation
- Folder structures presentation
- Description forms

MySQL
Storage

MySQL
Jena

METS SIP

METS SIP
UPBox and DataNotes: a collaborative data management environment for the long tail of research data

UPBox

- File & Folder
  - Upload, Delete

Reseacher

File System Hieararchy

1

UPBox

- Metadata
- User access control

MySQL

Storage

- RDF Metadata Export
  - Folder Structures
  - On-the-fly compression

- Access Control
- Collaboration
- Metadata storage

euroCRIS Membership Meeting

Porto, November 14-15, 2013
UPBox and DataNotes: a collaborative data management environment for the long tail of research data

DataNotes
Repository deposit

UPBox and DataNotes: a collaborative data management environment for the long tail of research data

1. UPBox
   - File & Folder
   - Upload, Delete
   - File System Hierarchy
   - Metadata
   - User access control

2. DataNotes
   - Directory structure annotation
   - Folder structures presentation
   - Description forms
   - RDF Metadata Export
   - Folder Structures
   - On-the-fly compression
   - Access Control
   - Collaboration
   - Metadata storage

3. Repository
   - Deposit workflow, embargo specification...
   - METS SIP

Repository deposit

euroCRIS Membership Meeting  Porto, november 14-15, 2013
Conclusions and ongoing work

• Tools have been tested with researchers
  – Good feedback
  – Researchers require support in the preparation of metadata models

• A service may be built on the prototype

• Current tools support researchers, they do not prepare repository ingestion

• Data prepared with these tools are available for curation
  – Descriptive metadata already associated
UPBox and DataNotes: a collaborative data management environment for the long tail of research data

More info

Available at http://dendro.fe.up.pt