The Dendro research data management platform

Applying ontologies to long-term preservation in a collaborative environment

João Rocha da Silva
joaorosilva@gmail.com

João Aguiar Castro
joaoaguiarcastro@gmail.com

Cristina Ribeiro
mcr@fe.up.pt

João Correia Lopes
jlopes@fe.up.pt

iPRES 2014, October 06 - 10 2014, Melbourne, Australia
Contents

• Research data management in the long tail
• Linked Open Data: why do we need it?
• Collaboration for easier metadata production
• The Dendro platform
• Conclusions
Research Data Management in the long tail of research

Why we need to start early
The long tail of research

2011: Science magazine reviewers are asked about their data requirements

\(\sim 1700\) replied
What is the size of the largest data set that you have used or generated in your research?

- 7.6% >1 TB
- 12.1% 100 GB–1 TB
- 32.0% 1–100 GB
- 48.3% <1 GB

Source
Is there sufficient funding for your lab or research group for data curation?

“

There are many tales of early archaeologists burning wood from the ruins to make coffee. If we fail to curate the environmental archives we collect from nature at public expense, we essentially repeat those mistakes.

”

Source
Gathering

Processing

Paper writing

Preservation, Sharing
Gathering

Processing

Paper writing

Metadata

Researcher leaves
Gathering

Processing

Paper writing

Project ends
“Where is the data?”
“How / when / by whom was the data produced?”
Researchers must participate in RDM from the start.

They are the domain experts.
Linked Open Data

What is it? Why do we need it?
Linked Open Data

- **Simplicity**
  - LOD is a very simple model for representing knowledge

- **Meaning**
  - Resources are interlinked by properties with established meaning

- **Interoperability**
  - Standard methods for querying data - SPARQL
  - Representations use standard formats - RDF, OWL
<table>
<thead>
<tr>
<th>Domain Specific</th>
<th>Analytical Chemistry Dataset</th>
<th>Fracture Mechanics Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Author Description Creation date</td>
<td>Author Description Creation date</td>
</tr>
<tr>
<td></td>
<td>Sample Count Analysed Substance</td>
<td>Initial Crack Length Specimen Type</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Generic**
- Author
- Description
- Creation date
- ...
Collaboration

For metadata useful now and in the future
Gathering

Processing

Paper writing

Preservation,
Sharing
Gathering → Sharing → Deposit → "Freeze" in repository → Collaboration → Description

Mentioned features:
- Gathering
- Sharing
- Deposit
- "Freeze" in repository
- Collaboration
- Description
Gathering
Demo
The Dendro platform

An open-source platform for Linked Open Data in research environments
• Data store fully built on Linked Data

• No relational database to preserve

• Model can grow by loading more ontologies

• External systems can retrieve resources via SPARQL
• GridFS cluster for large or numerous files

• Can work in the cloud if needed
Business Logic

Collaboration

- Flexible access control system
- Backup / Restore
- Versions history
- File type previews
- Integration
  - DSpace (SWORD)
  - ePrints (SWORD)
  - CKAN
  - Figshare
  - .......

Metadata

File Storage

Ontologies

Dendroβ
- All operations available via RESTful API using JSON
- All resources are dereferenceable (HTTP content negotiation)
- Plugin architecture allows integration with external systems
For curators

- Curators can work with researchers to build more ontologies using existing tools (e.g. Protégé)
- Established ontologies can be loaded (DC, FOAF…)
- Ontologies mature (reuse across Dendro instances)
- Data, metadata and its meaning go together
For programmers

- 100% Open-source software

- Rich API allows Dendro to be connected to almost any system (e.g. mobile apps)

Ontology-based multi-domain metadata for research data management using triple stores
Rocha da Silva, J., Ribeiro, C., Correia Lopes, J.
18th International Database Engineering & Applications Symposium (IDEAS 2014)
(pre-print available at http://dendro.fe.up.pt/)

LabTablet: semantic metadata collection on a multi-domain laboratory notebook
Amorim, R., Castro, J., Rocha da Silva, J., Ribeiro, C.
8th Metadata and Semantics Research Conference (MTSR 2014)
(pre-print available at http://dendro.fe.up.pt/)
Dendro dies, data lives on

“Database”

“Documentation”
Conclusions

- Research data management should start early
- Linked Open Data: simple, interoperable, flexible
- Collaboration support helps researchers while gathering metadata for later deposit
- Dendro: a fully open-source platform for RDM, built on Linked Open Data
- Dendro integrates with major repository platforms
Conclusions (cont’d)

• Ontologies: source of metadata descriptors
• Data model grows as more ontologies are loaded
• Curators can model and share the ontologies
• Domain ontologies evolve with reuse
João Rocha da Silva is an Informatics Engineering PhD student at the Faculty of Engineering of the University of Porto. He specializes on research data management, applying the latest Semantic Web Technologies to the adequate preservation and discovery of research data assets.

He is also an experienced freelancer iOS Developer with several Apps published on the App Store, and a self-taught DIY mechanic with a special interest in classic cars, particularly his 1987 Toyota Corolla GT Twin Cam, also known as Hachi-Roku or AE86.

Cristina Ribeiro is an Assistant Professor in Informatics Engineering at Universidade do Porto and a researcher at INESC TEC. She has graduated in Electrical Engineering, holds a Master in Electrical and Computer Engineering and a Ph.D. in Informatics. Her teaching includes undergraduate and graduate courses in information retrieval, digital libraries, knowledge representation and markup languages. She has been involved in research projects in the areas of cultural heritage, multimedia databases and information retrieval. Currently her main research interests are information retrieval, digital preservation and the management of research data.

João Correia Lopes is an Assistant Professor in Informatics Engineering at Universidade do Porto and a researcher at INESC TEC. He has graduated in Electrical Engineering in the University of Porto in 1984 and holds a PhD in Computing Science by Glasgow University in 1997. His teaching includes undergraduate and graduate courses in databases and web applications, software engineering and object-oriented programming, markup languages and semantic web. He has been involved in research projects in the area of long-term preservation, service-oriented architectures and e-Science. Currently his main research interests are e-Science and the management of research data.
Extras
Curated Dataset

Working Files

Curator

Data producers

Web Portal

SPARQL Endpoint

Free-Text Search

API

Deposit

Metadata validation

Ontology concept reuse

Sharing & evolution

Domain-Specific Lightweight Ontologies

"Mature" ontologies on the web

Specification of new metadata ontologies

CKAN Dryad

1

2

3

4

Dendro

Data reuser