The links between Research Data, Scientific Analysis Workflows, Provenance and Metadata: A Researchers Perspective on RDA

Ajinkya Prabhune
Introduction: Nanoscopy

Nanoscopy Research
• Investigation on “aggressive B-cell lymphomas”
• Microscopy technique – Spectral Precision Distance Microscopy (SPDM)
• Novel imaging method → producing datasets in the range of 150-200 TB

Community Requirements
• Community specific data-processing algorithms
• Manage the continuously evolving scientific workflows
• Allow experiment reproducibility
• Automated provenance information management
Data Repository/Workflow/ Provenance/ Metadata

- Data repository for long term storage and access to scientific datasets
- Integrate scientific workflow + associated provenance information in the repository system
  - Capture workflow description and execution details for
    - enabling research reproducibility
    - tracking workflow evolution
    - assessing data quality
Generic Provenance Metadata Model (GPDM)

- Enable modelling of **prospective & retrospective** provenance information
- **Flexible**: can be modelled as per the needs of the community
- **Interoperable**: Automated conversion into OPM/PROV model
- **Extensible**: Easy to integrate vocabularies
Automated Metadata Management in Scientific Repository Systems

- **PROVENANCEGEN**: automatic generation of provenance graphs
- Metadata modelling services integrated with metadata model registry
- Building links between data, provenance and metadata

- Digital Object (DO) available in a data repository

→ Enable Scientific Data Reproducibility
Metadata WGs & IGs

Metadata Standards

• Metadata directory for storing and accessing various metadata standards
• YAML based template for submitting metadata standard
• Well documented list of tools for handling the metadata standards
• Provision an API for adding, searching, retrieving metadata standards
• Generic metadata template and metadata principles document available
Research Data Provenance

Research Data Provenance IG

- Focus on comparison and evaluation of data provenance models (OPM/PROV)
- Provide recommendation on provenance model
- Liaison with Data Citation, Data Foundation & Terminology and Metadata Standards
Repositories WGs and IGs

Repositories Platforms of Research Data IG

- Collect and analyse research data use cases in context of repository platforms
- Matrix relating use cases with functional requirements as a deliverable
- Propose a specification for generic API in future → New BOF group spawned

Domain Repositories IG

- Aim to bring together active data repositories serving scientific communities
- Provide a forum for sharing practical experience and developing joint projects
**Conclusion**

Nanoscopy Data Repository System available for scientific community

- Capable of managing the extremely large datasets
- Metadata management integrated in the repository
- Automated provenance information management enabled via PROVENCANCEGEN algorithm and GPDM

- Involvement with various WGs and IGs is an additional benefit for my research
Aligning Nanoscopy Repository System with RDA

Research Data Provenance IG
- Data Provenance Model

Data Fabric IG
- Data Management
- Data Preservation
- Data Analysis
- Data Curation/Processing
- Hardware-Infrastructure
- Reference Data Collection

Repositories WGs and IGs
- Comprehensive coverage of functional requirements
- Generic API for interoperability

Metadata SD/C WG and Metadata IG
- Metadata Model
- Metadata Management
- Metadata Store

Interactive Web Portal
Knowledge Representation

Nanoscopy Open Reference Data Repository
- Scientific Workflow
- Annotation Service
- Intelligent Search
- Data Publication Service
- Metadata Extraction
- Metadata Processing
- Metadata Modelling
- Metadata Preservation
- Data Preservation
- Data Analysis
- Data Curation
- Data Processing

Data Archive
Data Processing
Automated Metadata Management in Scientific Repository Systems

Automated Metadata Management
- **PROVENANCEGEN** algorithm for automatic generation of provenance graphs
- Metadata modelling services integrated with metadata model registry
- Integrated PID system
Data Fabric IG aims to design a flexible and dynamic ecosystem consisting of components, services, tools, infrastructure for enabling efficient, cost-effective and reproducible research.

- Data Fabric IG is the umbrella group, works together with other WGs and IGs
- Use cases submitted by various communities

<table>
<thead>
<tr>
<th>Research Area</th>
<th>Relation with WGs IGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID assignment</td>
<td>PID Information Types</td>
</tr>
<tr>
<td>Scientific data repositories</td>
<td>Repositories Platform for Research Data, Domain Repositories, Research Data Repositories Interoperability</td>
</tr>
<tr>
<td>Metadata management</td>
<td>Metadata Standards Directory, Metadata Standards Catalog and Metadata IG</td>
</tr>
<tr>
<td>Provenance data management</td>
<td>Research Data Provenance</td>
</tr>
<tr>
<td>Data management policies</td>
<td>Practical Policies</td>
</tr>
</tbody>
</table>
Introduction: RDA

WG and IGs to support the complete research data lifecycle
- Metadata WGs and IGs
- Repository IGs
- Research Data Provenance IG
- Data Fabric IG
- And more…
Scientific Workflows/ Provenance/ Metadata

Typical scientific workflow execution

Required:
Raw data repositorz
Workflow description
Provenance informai
Consolidated Motivation

Enabling efficient management of scientific research (meta)data lifecycle from the perspective of the scientific community

• Comprehensive scientific data repository system
  • Extensible architecture for integrating dynamic requirements
• Seamless integration of complex scientific workflows + associated provenance data management

Active involvement with RDA

• Firsthand feedback from domain experts
• Dedicated groups focusing on specific topics (useable/adaptable outcomes)
• Regular discussion and updates via teleconferences
Scientific Workflows/ Provenance/ Metadata

Typical scientific workflow execution

- Raw dataset ingested and available in data repository