PODD: An Ontology-driven Data Repository for Collaborative Phenomics Research

Yuan-Fang Li

liyf@itee.uq.edu.au

The eResearch Lab, School of ITEE
The University of Queensland

12th International Conference on Asian Digital Libraries – Gold Coast, Australia

Outline

1 Introduction

2 Related Work

3 The PODD Ontology

4 Conclusion

Motivation

Challenges in phenomics research data management

- Data is huge
  - TPA: 0.5TB/week $\sim$ 25 TB/year

Motivation

Challenges in phenomics research data management

- Data is huge
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies quickly emerge

Motivation

Challenges in phenomics research data management

- Data is huge
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies quickly emerge
- Data needs ***context***
  - Scientific, administrative & other metadata

Motivation

Challenges in phenomics research data management

- Data is huge
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies quickly emerge
- Data needs ***context***
  - Scientific, administrative & other metadata

Repositories for the management of data

- Not all questions answered

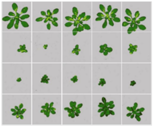
Clients & Collaborators

Australian Integrated Biological Science Facilities

Clients & Collaborators

Australian Integrated Biological Science Facilities

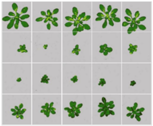
- Australian Plant Phenomics Facility (APPF)
    - High-throughput (TPA) & high-resolution (HRPPC) centers

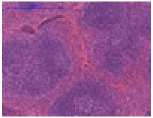


Clients & Collaborators

Australian Integrated Biological Science Facilities

- Australian Plant Phenomics Facility (APPF)
  - High-throughput (TPA) & high-resolution (HRPPC) centers
- Australian Phenomics Network (APN)
  - Mouse models, *deep* imaging and measuring platforms

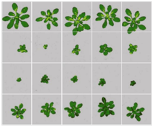


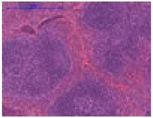


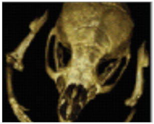
Clients & Collaborators

Australian Integrated Biological Science Facilities

- Australian Plant Phenomics Facility (APPF)
  - High-throughput (TPA) & high-resolution (HRPPC) centers
- Australian Phenomics Network (APN)
  - Mouse models, *deep* imaging and measuring platforms
- Atlas of Living Australia (ALA)
  - Biodiversity information portal







Data Management Requirements

Data capturing

Flow Cytometry	FACS data
Histopathology	Zeiss slide images
Plant imaging	Lemnatec images, Flourogroscan images, 3D imaging
Infrared imaging	FLIR images
Chemical measurements	Chlorophyll content, Stomatal conductance
Visual observation	Manual reports (plant, mouse phenotypes)
. . .	. . .

Data Management Requirements

Metadata capturing

Project	Project proposal, project plan
Investigation	Objectives, design
Materials	Lines/genotypes, samples, growth conditions
Devices	Specs, settings, versions
Processes	Workflows, protocols, variations
Measurements	Data, images
Analysis	Observations, results
. . .	. . .

Data Management Requirements

Data management tasks

- Data distribution & sharing
- Data publishing
- Access control
- Archival & versioning
- Data discovery & analysis
- Data integration

PODD: an ontology-driven repository

Goals

- Acquisition and storage of large volumes of data
  - Distribution, access control, versioning, etc.
- Data conxtualization
  - Logical organization
  - Provenance tracking
  - Discovery & integration
- Prepare for change
  - Changes in domain model

PODD: an ontology-driven repository

Goals

- Acquisition and storage of large volumes of data
  - Distribution, access control, versioning, etc.
- Data conxtualization
  - Logical organization
  - Provenance tracking
  - Discovery & integration
- Prepare for change
  - Changes in domain model

Approach

- An ontology-driven approach
- **Ontologies as the domain model**
- Benefits: flexibility & extensibility

Outline

1. Introduction
2. Related Work
3. The PODD Ontology
4. Conclusion

Related Work

FuGe – Functional Genomics Experiment

- *Material*, *Protocol*, *Data*, etc.
- Can be extended to support phenomics
- × Defined in UML & mapped to database schemas – difficult to extend for new concepts

Related Work

FuGe – Functional Genomics Experiment

- *Material*, *Protocol*, *Data*, etc.
- Can be extended to support phenomics

× Defined in UML & mapped to database schemas – difficult to extend for new concepts

OBI – Ontology for Biomedical Investigations

- *“An integrated ontology for the description of life-science and clinical investigations.”*
- Comprehensive: 2,600+ classes, 10,000+ axioms

× Complex, computationally ($\mathcal{SHOIN}(D)$)

Related Work

Web Ontology Language (OWL)

- Precise, open & extensible – exactly what we need!
- Provides core language constructs & vocabularies for expressing complex ontologies – *data models*
- APIs, query engines & automated reasoners available

Fedora Commons

- Mature open-source digital repository software
- Modular & extensible
- Widely used

Outline

1 Introduction

2 Related Work

3 The PODD Ontology

4 Conclusion

The PODD Ontology

Modeling essentials

- Domain concepts – *OWL classes*
- Inter-concept relations – *OWL predicates* & *OWL restrictions*
- Concrete domain objects – *OWL Individuals*
- Comments, descriptions – *OWL annotations*

The PODD Ontology

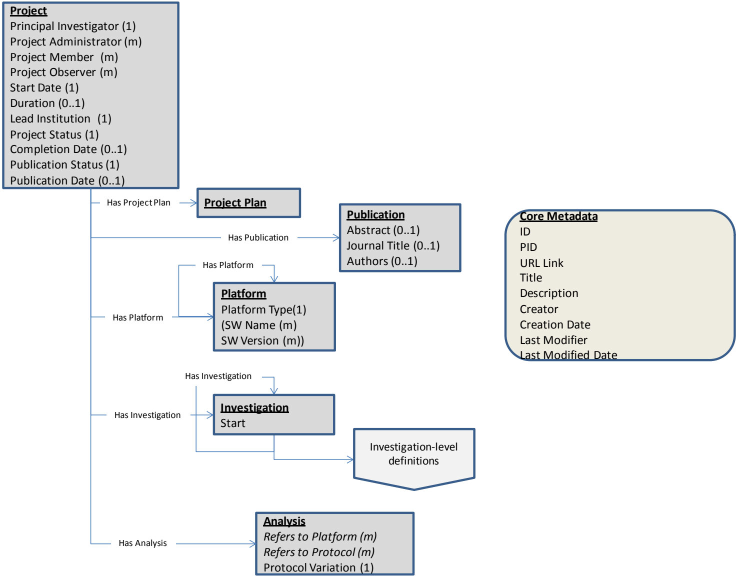
Modeling essentials

- Domain concepts – *OWL classes*
- Inter-concept relations – *OWL predicates* & *OWL restrictions*
- Concrete domain objects – *OWL Individuals*
- Comments, descriptions – *OWL annotations*

Benefits

- Extensibility through inheritance
- Reuse & integration through **ontology mapping** & **ontology annotation**
    - Gene Ontology, Plant Ontology, etc.

The PODD Ontology – Overview



Project
Principal Investigator (1)
Project Administrator (m)
Project Member (m)
Project Observer (m)
Start Date (1)
Duration (0..1)
Lead Institution (1)
Project Status (1)
Completion Date (0..1)
Publication Status (1)
Publication Date (0..1)

Has Project Plan

Project Plan

Has Publication

Publication
Abstract (0..1)
Journal Title (0..1)
Authors (0..1)

Has Platform

Has Platform

Platform
Platform Type(1)
(SW Name (m)
SW Version (m))

Has Investigation

Has Investigation

Investigation
Start

Investigation-level definitions

Has Analysis

Analysis
Refers to Platform (m)
Refers to Protocol (m)
Protocol Variation (1)

Core Metadata
ID
PID
URL Link
Title
Description
Creator
Creation Date
Last Modifier
Last Modified Date

The PODD Ontology – An Example

Example

The ***Project*** concept

- The top-level concept
- Constraints on inter-object relations & attributes

The PODD Ontology – An Example

Example

The ***Project*** concept

- The top-level concept
- Constraints on inter-object relations & attributes

$$\begin{aligned} \mathit{Project} &\sqsubseteq = 1\ \mathit{hasProjectPlan} \sqcap \forall\ \mathit{hasProjectPlan.ProjectPlan} \\ &\sqsubseteq \geq 1\ \mathit{hasInvestigation} \sqcap \forall\ \mathit{hasInvestigation.Investigation} \\ &\sqsubseteq = 1\ \mathit{hasStartDate} \sqcap \forall\ \mathit{hasStartDate.xsd{:}date} \\ &\sqsubseteq \leq 1\ \mathit{hasPublicationDate} \sqcap \forall\ \mathit{hasPublicationDate.date} \end{aligned}$$

- **Extensibility** from inheritance of OWL classes & predicates

PODD Ontology – Roles

Ontologies *drive* repository functions

Presentation

- Object creation, editing, display, etc.

Storage

- Object (de)serialization to/from ontologies

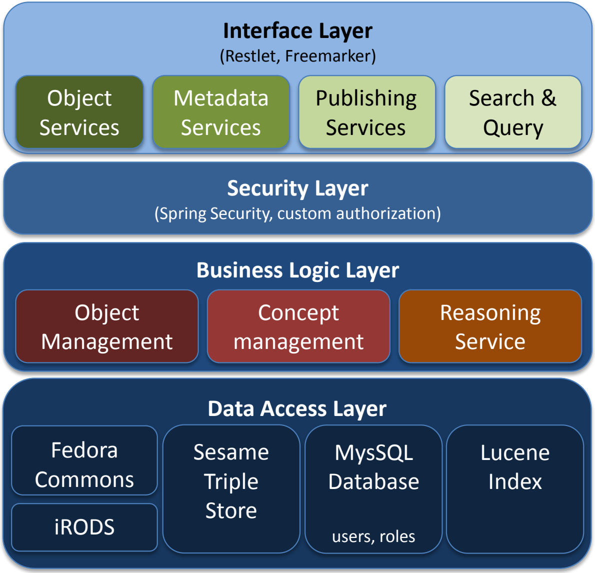
Validation

- Validation based on concept constraints

Discovery

- Queries using SPARQL
- Full-text search

The PODD Repository: The High-level Architecture



Interface Layer

(Restlet, Freemarker)

Object Services

Metadata Services

Publishing Services

Search & Query

Security Layer

(Spring Security, custom authorization)

Business Logic Layer

Object Management

Concept management

Reasoning Service

Data Access Layer

Fedora Commons

iRODS

Sesame Triple Store

MySQL Database

users, roles

Lucene Index

Outline

1 Introduction

2 Related Work

3 The PODD Ontology

4 **Conclusion**

Conclusion

To recap

- Large amounts of data need to be managed
  - There is a need for data archival, storage & discovery
- Current approaches lacking/inadequate/inflexible
  - Emerging processes, platforms, technologies require a extensible conceptual framework
- An ontology-driven architecture as the foundation of PODD
  - Ontologies as the domain model
  - Extensible & open

Conclusion

Where we are now

- PODD ontology for phenomics research
- Development of basic repository functionality
- Development of PODD web interface

Conclusion

Where we are now

- PODD ontology for phenomics research
- Development of basic repository functionality
- Development of PODD web interface

What’s next

- Development of batch data import/export processes
- Development of object discovery services
- Integration with Shibboleth authentication
- Exposing data for discovery
- Integrating with other data sources

Thank You!

Acknowledgment

Faith Davies, Gavin Kennedy, Jane Hunter @ eResearch Lab, School of ITEE, UQ