# 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

IBS Phenomics Data and Informatics Workshop - Canberra 2010

# Outline

# Motivation

## Challenges in phenomics research data management

- Data is huge
  - APN: estimated 1.8TB + 30% growth/year
  - TPA: 0.5TB/week $\sim$ 25 TB/year

# Motivation

## Challenges in phenomics research data management

- Data is huge
  - APN: estimated 1.8TB + 30% growth/year
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies emerge

# Motivation

## Challenges in phenomics research data management

- Data is huge
  - APN: estimated 1.8TB + 30% growth/year
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies emerge
- Data needs ***context***
  - Scientific, administrative & other metadata

# Motivation

## Challenges in phenomics research data management

- Data is huge
  - APN: estimated 1.8TB + 30% growth/year
  - TPA: 0.5TB/week $\sim$ 25 TB/year
- New platforms/processes/technologies emerge
- Data needs ***context***
  - Scientific, administrative & other metadata

## Repositories for the management of data

- Not all questions answered

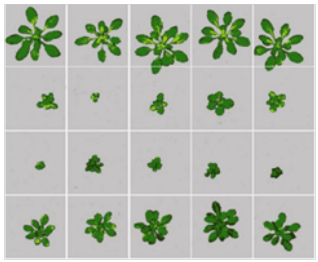
# Clients & Collaborators

## Australian Integrated Biological Science Facilities

# Clients & Collaborators

## Australian Integrated Biological Science Facilities

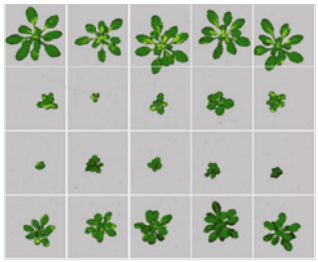
- Australian Plant Phenomics Facility (APPF)
  - High-throughput (TPA) & high-resolution (HRPPC) centers

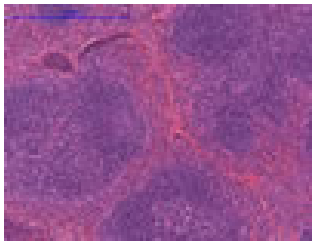


# Clients & Collaborators

## 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

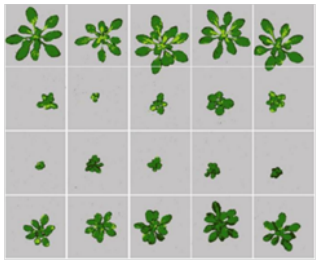
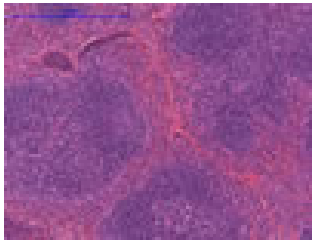
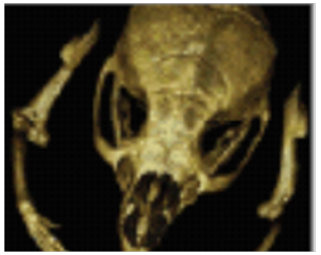




# Clients & Collaborators

## 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 data collection & management

# 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 |
| ... | ... |

# Data Management Requirements

## Data generation

| | |
|---|---|
| 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

| | |
|---|---|
| Data distribution | ? |
| Data sharing | ? |
| Data publishing | ? |
| Access control | ? |
| Archival & versioning | ? |
| Structure & metadata | ? |
| Data discovery & analysis | ? |

# PODD: an ontology-driven repository

## Goals

- Supports data captured by different platforms
- Supports different data formats
- Supports effective data management
  - Metadata, distribution, access control, discovery, etc.

# PODD: an ontology-driven repository

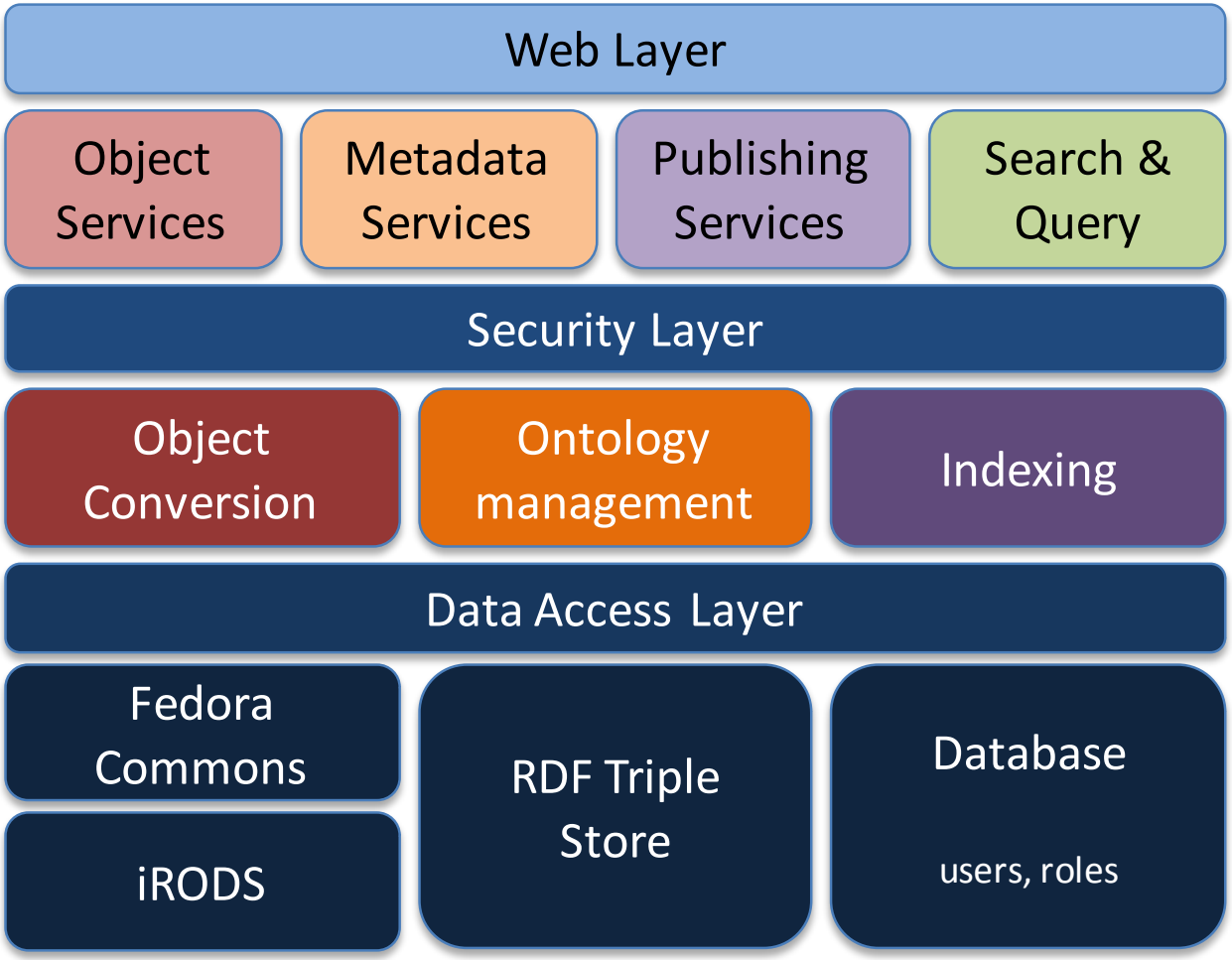
## Goals

- Supports data captured by different platforms
- Supports different data formats
- Supports effective data management
  - Metadata, distribution, access control, discovery, etc.

## Approach

- Define domain ontologies – focus of this talk
- Design an appropriate architecture
- Develop an ontology-aware repository

# PODD: The High-level Architecture



# Outline

# Related Work

## FuGe – Functional Genomics Experiment

- *Material*, *Protocol*, *Data*, etc.
- Can be extended to support phenomics
- × May be difficult to extend for new concepts

# Related Work

## FuGe – Functional Genomics Experiment

- *Material*, *Protocol*, *Data*, etc.
- Can be extended to support phenomics
- × May be 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 vocabularies for expressing complex ontologies – *data models*
- Main language constructs
  - Classes, e.g, `Human`
  - Predicates , e.g., `hasParent`
  - Individuals, e.g., `Aristotle`
- APIs, query engines & automated reasoners available

# Outline

# The Ontology Approach

## Benefits

- Greater extensibility
  - New concepts/relations can be easily added/modified

# The Ontology Approach

## Benefits

- Greater extensibility
  - New concepts/relations can be easily added/modified
- Better reuse & integration
  - Ontologies are *open*
  - Other ontologies can be integrated on multiple levels

# The Ontology Approach

## Benefits

- Greater extensibility
  - New concepts/relations can be easily added/modified
- Better reuse & integration
  - Ontologies are *open*
  - Other ontologies can be integrated on multiple levels
- Balance between expressivity & reasoning complexity
  - Formal semantics enables automated query & analysis
  - Off-the-shelf tools available

# The PODD Ontology

## Modeling essentials

- Domain entities
  - Abstract concepts
  - Concrete objects
- Domain entities defined using OWL *ontologies*

# The PODD Ontology

## Modeling essentials

- Domain entities
  - Abstract concepts
  - Concrete objects
- Domain entities defined using OWL *ontologies*

## Modeling in OWL

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

# The PODD Ontology – An Example

## Example

The ***Project*** concept

- The top-level concept
- Constraints on inter-object relations & attributes – like everything else

# The PODD Ontology – An Example

## Example

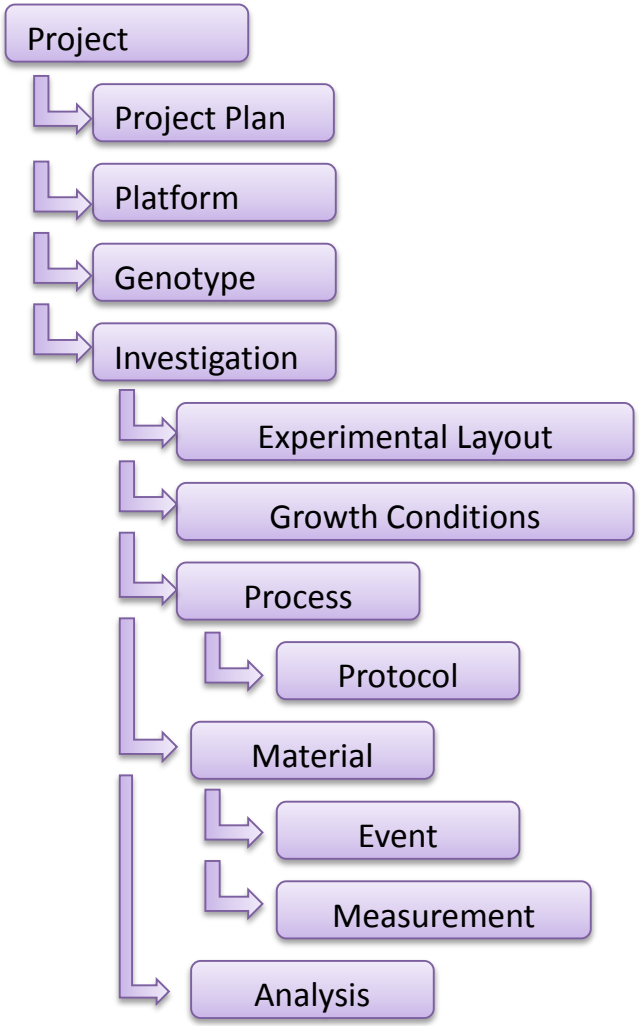
The ***Project*** concept

- The top-level concept
- Constraints on inter-object relations & attributes – like everything else

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

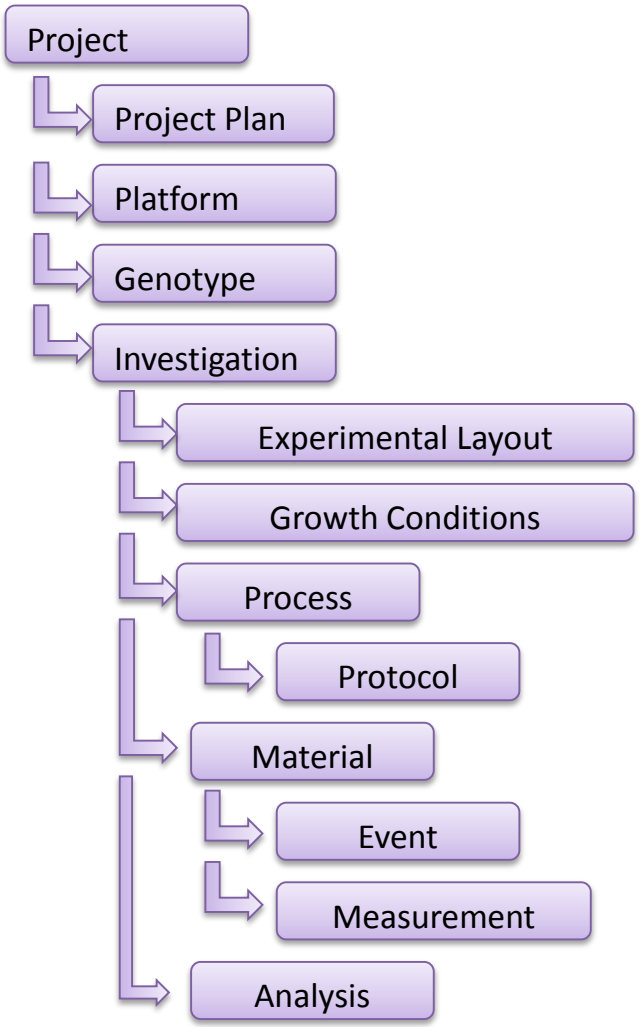
- **Extensibility** from inheritance of OWL classes & predicates

# The PODD Ontology



# The PODD Ontology

- Inspired by FuGe & OBI
- Aim: define all *essential* domain concepts, attributes & relations
- $\sim$ 30 classes,
  $\sim$ 80 predicates,
  $\sim$ 200 Axioms
- Faster reasoning, querying, etc.



# PODD Ontology – Roles

## Ontologies *drives* repository functions

**Presentation** Drives the rendering of the web pages

- Object creation, editing, display, etc.

**Management** Object (de)serialization to/from ontologies

**Validation** Ontology reasoning performed

- Validation on object type, cardinality, etc.

**Discovery** Multiple ways of finding information

- Queries using SPARQL
- Searches using Lucene

# The PODD Ontology – Object creation page rendering

**Create Project Object**

**Project Details**

Title:

Description:

Has duration:

Has start date (YYYY-MM-DD):
2010-04-19

Has lead Institution:

# Outline

# 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 extensible infrastructure
- An ontology-driven approach as the foundation of PODD
  - Extensible & open

# Conclusion

## Where we are now

- PODD ontology stabilizing
- Development of basic repository functionality
- Development of PODD web interface

# Conclusion

## Where we are now

- PODD ontology stabilizing
- 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
- Integrate with AAF federated authentication services
- Exposing data for discovery
- Integrating with other data sources

# Thank You!

**Acknowledgment**

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