PODD
An Ontology Driven Architecture for Extensible Phenomics Data Management

Gavin Kennedy
PODD Project Manager
High Resolution Plant Phenomics Centre
Canberra, Australia
What is Plant Phenomics?

**Phenome = Genome X Environment**

- Genomics is accelerating gene discovery but how do we capitalise on these data sets to establish gene function and development of new genotypes for agriculture?
- High throughput and high resolution analysis capacity now the factor limiting discovery of new traits and varieties
- Accelerated systematic analysis of the phenotypes that arise from the combination of genome and environment

“In the next 50 years we must produce more food than we have consumed in the history of mankind”

Megan Clarke, CSIRO CEO 2009
Imagine a plant breeder walking his trials, logging plant performance from distributed sensors with his mobile phone or logging on to Phenonet from home to view his wheat in real time.
Plant Phenomics: A Case Study

High Resolution Plant Phenomics Centre (HRPPC)

Multiple platforms for “deep phenotyping”

High resolution phenotyping of model and crop plants

Scale from small model plants to crop fields

Scale from a stand of trees to the plant cell

Phenomics is Characterised by Multiple Heterogeneous High Volume High Resolution Data Generation Platforms

www.plantphenomics.org.au
The Phenomics Ontology Driven Data repository

A research data and metadata repository.

Managing Phenomics Data from Multiple Heterogeneous High Volume High Resolution Data Generation Platforms

A methodology for managing and publishing research data outputs.

A semantic web data resource.

www.plantphenomics.org.au
A word on the Semantic Web

The **Semantic Web** is a "web of data" that enables machines to understand the semantics, or meaning, of information on the World Wide Web.
Semantic Web technologies (i.e. Resource Description Framework, or RDF) allow us to not only describe data files, but to link them across the web.
RDF (Resource Description Framework) is a simple format for making knowledge assertions.

Composed of subject – predicate – object statements.

Gavin (subject) Likes (predicate) Coffee (Object)
Or
Investigations (subject) Have (predicate) Materials (object)

With these statements we make assertions about a domain and thus model domain knowledge.
The PODD Project

Web based data management system.

Manage multiple **heterogeneous** data resources.

Data objects are **annotated** with **metadata**.

**Metadata** is **Ontology** based to support data discovery and analysis. **Metadata** maintained in format to support the semantic web.
Annotating with Metadata

Metadata
“Data about data”
Information that supports understanding and contextualisation of the data.

Data

<table>
<thead>
<tr>
<th>Day</th>
<th>Top view</th>
<th>Side view</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Metadata
Format of data
Experimental Protocols
Time and date of capture
Platform captured on
Organism imaged
Growth stage of organism
Genotype of organism
Treatments
Analysis Protocol

Metadata solves Heterogeneity
Annotating with Ontologies

Ontology

“Common Vocabulary”

A set of terms used to describe a common domain.
Using Ontologies

Annotation

This image shows the wheat plant on the left has increased “salt tolerance (TO_0006001)”

Discovery

Show me all experiments associated with the term “salt sensitivity (TO_0000429)”

Modelling a domain

OBI_0000050: “platform”
“A platform is an object_aggregate that is the set of instruments and software needed to perform a process.

Annotation means we describe a domain object or process.

Modelling means we categorize a domain object or process, and its relationship to other domain objects or processes.
To share common understanding of the structure of information among people or software agents

To enable reuse of domain knowledge

To make domain assumptions explicit

To separate domain knowledge from the operational knowledge

To analyse domain knowledge

*Ontology Development 101: A Guide to Creating Your First Ontology
Natalya F. Noy and Deborah L. McGuinness
Modelling Phenomics Data with an Ontology

In the PODD Ontology we model everything as objects:

- Experiments (Investigations)
- Plants (Materials)
- Treatments
- Environments
- Measurement Platforms
- Temporal Events
- Raw Data (Data)

We then define the relationships between objects:

- Investigation has Material
- Material has Observations
- Material references Genotype
- Data references Material
The extensible PODD Ontology

- Project
  - Project Plan
  - Investigation
  - Platform
  - Analysis
  - Event
    - Genotype
  - Treatment Material
    - Sex
    - Observation/Phenotype
    - Measurement
      - Measurement Parameter
  - Material
  - Container
  - Data
    - Environment
    - Design
      - Gene
      - Sequence
  - Archive Data
    - Measurement
    - Observation/Phenotype
    - Treatment
The PODD ontology is written in an ontology schema language called **OWL/RDFS**.

Using the PODD ontology as our template ...

We then create instances of the metadata objects and their relationships (in RDF).

And attach the data files to the relevant objects.
The PODD Ontology Drives Repository Functions

**Presentation**: Rendering of the web pages for: object creation; editing, display, etc.

**Management**: Objects stored as per the model in the ontology.

**Validation**: Ontology reasoning performed against objects.

**Discovery**: SPARQL Queries or metadata searches.

Refine the Domain? Extend the Ontology.

Change the Domain? Change the Ontology.

www.plantphenomics.org.au
# Phenomics Data Management Requirements

| Capture          | Multiple heterogeneous proprietary data generation platforms.  
|                 | Move data off localised servers/workstations  
|                 | Automate data & metadata capture |
| Manage          | Centralise data access  
|                 | Organise data at the project level |
| Store           | Manage data files at remote data stores/data clouds. |
| Secure          | Secure data at the project level.  
|                 | Protect IP. |
| Annotate        | Provide context and meaning to the data |
| Distribute      | Delivering data to clients  
|                 | Sharing data with collaborators |
| Publish         | Public Online Access.  
|                 | Data publication identifiers. |
| Discover        | Online Access.  
|                 | Semantic Web Enabled. |

*Based on analysis of HRPPC Information Systems Requirements, March, 2009*
Phenomics Data Management Requirements

Annotate
Provide context and meaning to the data

Context and meaning through metadata and annotation.

Metadata through PODD ontology objects and relationships between objects.

Extended annotation through external ontologies (e.g. Gene Ontology, Trait Ontology)

Trait Ontology
- Stress Trait (80 objects)
  - Abiotic Stress Trait (40 objects)
    - Stem Strength
    - Leaf Rolling
    - Water Stress Trait (8 objects)
  - Chemical Stress Sensitivity (32 objects)
    - Nutrient Sensitivity (12 objects)
      - Micronutrient Sensitivity (3 objects)
      - Macronutrient Sensitivity (3 objects)
    - Salt Sensitivity (6 objects)
      - Salt Tolerance (6 objects)
      - Potassium Chlorate Tolerance
  - Pesticide Sensitivity
  - Oxidative Stress (20 objects)
- Biotic Stress Trait (40 objects)

Project: Salt Tolerance comparison in Uni of Adelaide Wheat Breeding Lines (Public Access) Principal Investigator: Mark Tester, Lead Institution: University of Adelaide
Object: Material, Title: Salt_Tol_1_1

Project: Salt Tolerance comparison in Uni of Adelaide Wheat Breeding Lines (Public Access) Principal Investigator: Mark Tester, Lead Institution: University of Adelaide
Object: Material, Title: Salt_Tol_1_2

...
Phenomics Data Management Requirements

<table>
<thead>
<tr>
<th>Publish</th>
<th>Public Online Access. Data publication identifiers.</th>
</tr>
</thead>
</table>

Projects are secure, only authorised users can access them, until...

Projects can be published, granting public (read only) access to the data.

PODD objects are identified by persistent publishable URLs (e.g. DOIs or ANDS PIDS). http://hdl.handle.net/10378.3/14058
Resources


Mouse Phenomics Production Instance: http://podd.australianphenomics.org.au

PODD Project Website: http://projects.arcs.org.au/trac/podd


Contact: Gavin.Kennedy@csiro.au
Ph: 0413 337 819

www.plantphenomics.org.au
The PODD Collaboration

PODD Project Manager
Gavin Kennedy

University of Queensland eResearch Lab:
Faith Davies (Developer)
Simon McNaughton (Developer)
Peter Ansell (Developer)
Jane Hunter (eResearch Lab Leader)

APN
Philip Wu (Developer)
Martin Hamilton (Developer)
Adrienne McKenzie (APN Head of Network Services)

Monash University
Yuan-Fang Li (Developer, Guru)

NeAT
Andrew Treloar (Deputy Director ANDS)
Paul Coddington (Projects Manager, ARCS)

ALA
Donald Hobern (Director, ALA)

This work is part of a National eResearch Architecture Taskforce (NeAT) project, supported by the Australian National Data Service (ANDS) through the Education Investment Fund (EIF) Super Science Initiative, and the Australian Research Collaboration Service (ARCS) through the National Collaborative Research Infrastructure Strategy Program.