Flexible Scientific Data Management for Plant Phenomics Research

Peter Ansell[1], Robert Furbank[2], Kutila Gunasekera[1], Jianming Guo[2], David Benn[3], Gareth Williams[3], Xavier Sirault[2]

[1] eResearch Group, School of Information Technology and Electronic Engineering, University of Queensland, Brisbane, Australia
[2] CSIRO Plant industry, High Resolution Plant Phenomics Centre, Canberra, Australia
[3] CSIRO IM&T Advanced Scientific Computing and Research Data Services, Melbourne, Australia

Abstract. In this paper, we expand on the design and implementation of the Phenomics Ontology Driven Data repository [1] (PODD) with respect to the capture, storage and retrieval of data and metadata generated at the High Resolution Plant Phenomics Centre (Canberra, Australia). PODD is a schema-driven Semantic Web database which uses the Resource Description Framework (RDF) model to store semi-structured information. RDF allows PODD to process information about a range of phenomics experiments without needing to define a universal schema for all of the different structures. To illustrate the process, exemplar datasets were generated using a medium throughput, high resolution, three-dimensional digitisation system purposely built for studying plant structure and function simultaneously under specific environmental conditions. The High Performance Compute (HPC), storage and data collection publication aspects of the workflow and their realisation in CSIRO infrastructure are also discussed along with their relationship to PODD.

Keywords: eResearch, Semantic Web, RDF, OWL, Data collection citation, BagIt, Data Access Portal

1 Introduction

Since the genomics era, biology has become a data-driven science. Advances in robotics, automation and imaging, in combination with high performance computing have permitted the rapid production of large and complex biological datasets. Currently, high volumes of heterogeneous image data, physiological and morphological measurements are being acquired by a range of new phenotyping platforms located in purpose built phenomics centres across the world. These large datasets of phenotypic characteristics such as growth rate, plant architecture, photosynthetic performance, yield must be stored and correlated with genotypes. These factors provide evidence of genetic variation in natural and derived genetic populations (e.g. germplasm collections, association genetic

panels, recombinant inbred lines). They also enable a deeper understanding of the dynamic relationship between phenotype, genotype and environment which is necessary to continue delivering the increase in productivity necessary for feeding the world.

The vast array of phenotypic data collected from a variety of phenomics platforms must be combined with metadata explaining how the raw data was collected. This combination of raw data and metadata are then delivered to a range of analysis pipelines, which transform the raw data into aggregated multi-phase datasets, each phase representing a new aggregation or inference from the original raw data. This reduction process converts the raw multi-dimensional data into information which is conceptually interpretable by a human being, i.e. new knowledge. The additional metadata describing the steps taken are recorded to give context to the data.

To make sense of this large amount of information, sophisticated storage, archiving, searching and analysis capabilities are required. To date solutions to this problem have been handled essentially by private companies, and no suitable solution exists in the public domain. Lack of systems, both to manage linked metadata, and controlled vocabularies to describe plant growth and experimental conditions, have severely hampered sharing of plant phenomics data, comparison of results between laboratories and the capacity to carry out meta-analysis of existing data sets.

Thus, to support publicly-funded phenomics activities in Australia, the Phenomics Ontology Driven Data repository (PODD) has been developed as a repository for data produced by the variety of plant imaging and phenotyping platforms available at the High Resolution Plant Phenomics Centre, as well as for recording the contextual metadata associated with plant genotypes, treatments and environmental conditions [1].

In this paper, we describe the workflow management that the High Resolution Plant Phenomics Centre (HRPPC) has implemented for keeping track of its phenomics data, metadata and experimental processes. This complex challenge was addressed by building a multi-disciplinary group of information technology experts and embedding users of phenomics technologies into it. The result of the approach is a state of the art computational and data mining environment, optimised for data access, data discovery and data sharing, which also provides the flexibility for linking genomic information through the use of RDF triples. In this context, we also describe the role of the CSIRO Data Access Portal (DAP) [2] to annotate and store raw and processed datasets. DAP also provides long term secure storage for data collections and the ability to search for, control access to, and cite them via Digital Object Identifiers. PODD manages the mapping of collections located in DAP to PODD projects, providing for the storage of large images and documents unsuited to RDF databases. Figure 1 shows the relationship between components and key data flows.
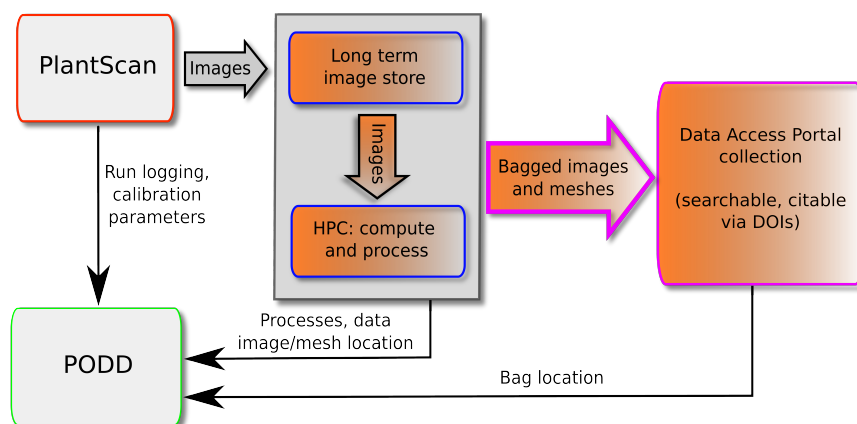
Fig. 1. HRPPC component relationships and data flow

2 Phenomics Ontology Driven Data repository

2.1 Semantic science for phenomics data management

Scientists have focused on including semantics into datasets, typically using the foundations of RDF and OWL, from two main directions. Some focus on defining ontologies based on hierarchies of scientific concepts and properties, while others have focused on mapping complex scientific datasets to RDF using syntax transformations without initially defining the semantic meaning of the results. In reality, most efforts fall somewhere in the middle, with ontological annotations attached to some data points while other nearby data points are syntactically represented using RDF, without links to ontologies of scientific concepts.

Increasingly however, providers of scientific datasets are focusing on enhancing their datasets using curated scientific concepts from ontologies. For example, scientists have used the Gene Ontology [3] to link well known concepts to represent common elements across genomics datasets, while the Plant Ontology [4] allows the description of plant based datasets.

2.2 Redesign of the Phenomics Ontology Driven Data repository

The PODD repository relies on semantic web technologies to manage phenomics data and metadata. Although both ontologies and mappings are essential, in PODD it was necessary to build the system with a relaxed ontological vocabulary. This enables scientists to sparsely populate their datasets and sparsely link to community defined upper ontologies as necessary. This allows scientists to continue to maintain projects containing curated scientific concepts alongside raw experimental data. The PODD repository was redesigned based on an evaluation of the original software [1] that found it was not able to scale sufficiently

to suit the HRPPC needs due to design and implementation deficiencies. The major design differences to the software implemented by [1] are that projects are no longer the only supported top object type, and projects are not stored in multiple parts, as that approach was not able to scale as was originally hypothesised.

A PODD project in PlantScan™ contains top level branches describing the various parts of a scientific project. These include a branch for raw data, along with separate branches for results, analysis, and publications related to the project. In the case of raw data, the semantics are not necessarily clear and are not easily defined by the automated platforms collecting the data. The scientist may later semantically link the data with results, conclusions, and external ontologies. For example, a scientist may annotate the data objects representing images of a plant with a link to a trait that is defined in the Plant Ontology. They may also annotate the image with a link to a trait that is defined inside of the project, such as when the trait is novel and not represented in a community ontology.

2.3 Semantic validation

PODD validates scientific project descriptions using independently configurable constraints based on OWL (Web Ontology Language) ontologies. Although PODD currently solely supports OWL for constraint verification, it could be easily extended in other cases to use different systems such as N3, RDFS, SPARQL, or SPIN as rules languages [5].

OWL is used to determine whether projects are both internally consistent, with all objects having an explicit RDF type, and whether they are consistent with the ontologies that they import. For example, any OWL object property that has been defined to link from image acquisition runs to images defines the provenance of an image.

General scientific properties and phenotype specific properties are defined in optional extension ontologies as illustrated in Figure 2. These are used by scientists to annotate their projects with concepts specific to their field, without requiring other scientists using the same PODD installation to use phenotype properties to annotate their projects.

3 CSIRO Data Access Portal

CSIRO's Research Data Service (RDS) has developed the Data Access Portal (DAP), an open source web application that enables research data to be discovered, managed and shared. [2]

Researchers can describe a data collection, deposit data, choose a license, and add attribution details. Access to a collection's description and/or data can be restricted to CSIRO or a set of individuals (within CSIRO or partner organisations) or it can be made public, becoming searchable by anyone via the Internet. In the case where a collection and its data are public, a Digital Object
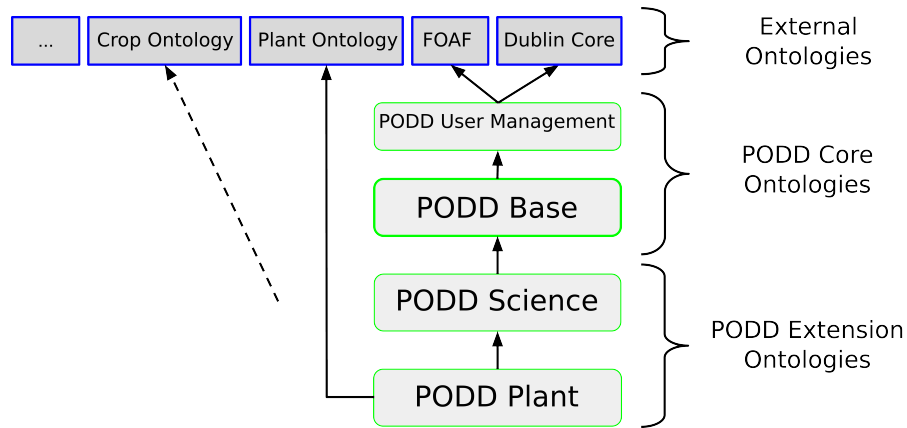
Fig. 2. PODD ontology hierarchy

Identifier (DOI) is issued and can be used to formally cite the collection in a publication.

4 The PlantScan™ digitisation platform

4.1 BagIt

BagIt is defined by an Internet Engineering Task Force (IETF) document as an "hierarchical file packaging format for storage and transfer of arbitrary digital content"[6]. A payload manifest details content and MD5 or SHA hashes for content integrity verification. Data file related metadata can be stored in pre-defined files as key-value pairs.

For PlantScan™, file-level metadata includes plant barcodes, batch numbers, and plant type, although the BagIt specification does not mandate a particular archiving strategy, with the focus being upon the directory structure, special files, and integrity checking. BagIt-conforming tools [7] [8] were assessed and where necessary, improvements were implemented and tested to ensure that the tools were fit for purpose in the CSIRO Advanced Scientific Computing (ASC) HPC environment.

4.2 Bag preparation for a DAP collection

CSIRO ASC shared facilities [9] are used to process the raw PlantScan™ data to derive data products (meshes). Raw data and meshes are collected using the BagIt format [6] and stored in the ASC archival system. ASC High Performance Compute (HPC) hosts (systems with high processor count and large memory) are taken advantage of to create and verify bags more rapidly than would be possible on conventional computer systems. CSIRO's HRPPC makes use of DAP

to store collections of PlantScan™ raw images and processed mesh data as bags. Currently, one bag is equivalent to a single batch scanned on the PlantScan™ local software system, which usually means the same kind of plant with different genotypes scanned under one experiment configuration profile.

Raw data from PlantScan™ local storage (HRPPC-Store) and data processed on HPC hosts are transferred to ASC bulk storage where image and mesh files are organised in folders by batch, then barcode number, then subfolders for each image file type, including RGB images, IR images, and LiDAR (Light Detection and Ranging Sensors, and their related meshes. Bag creation is carried out via an allocated ASC HPC job. The metadata required for a DAP publication is created and the bag transferred to the DAP staging area via SFTP (SSH File Transfer Protocol). After publication of the DAP collection, the data from PlantScan™ for the given project becomes discoverable via DAP. In addition, experiment reports, published papers, and sensor configurations can either be made accessible via a DAP collection's "related materials" links, other metadata fields, or within the collection's data (e.g. bag).

4.3 Heterogeneous data streams

PlantScan™ is a medium throughput high resolution phenotyping platform, which brings together a number of imaging sensors–light detection and ranging, far-infrared imaging, and multi-wavelength imaging–to non-invasively measure plant growth and function using in-silico approaches. Raw data is captured with its contextual information (e.g. system configuration, time of acquisition, batch number and project) and is stored in a purpose-built database as the data is being generated. The various data streams are collated and used to produce full 3D representation of each plant with overlaid spectral information. The metadata collected during image acquisition are necessary inputs for the computer vision techniques which are used to create the 3D representation of the plant. The 3D meshes are then automatically segmented in order to semantically identify the different parts of the plants [10]. A longitudinal 3D matching pipeline for plant mesh parts is then used to evaluate temporal changes at the whole plant and/or organ level.

4.4 Metadata

Each acquisition on PlantScan™ includes metadata (in addition to the raw data streams), such as plant genus and species, project and experiment metadata, a unique identifier for each image (Globally Unique Identifier), imaging angle, environmental temperature of the imaging chamber, location of optical and colour calibration datasets for each acquisition run, and LiDAR calibration files. The metadata associated with each acquisition is automatically generated when setting up the configuration on the platform. This information is paramount to validate and process the raw image data, and for the post-processing phases.

4.5 Data volume

Digitisation systems such as PlantScan™ generate huge amounts of data including raw image data, registration metadata, sensor configurations and plant metadata. For example, PlantScan™ generates around 500GB of raw image data, representing in excess of 200,000 database records, per day. Sufficient storage space (usually at remote locations) and fast network transfer rates are thus necessary to facilitate data movement for processing using high performance computers (HPC). Because an RDF database structure is not suitable for handling large data sets of images, it is necessary to package the raw information into elementary units with permanent addresses which could be retrieved using PODD. The CSIRO DAP [2] and ASC storage and compute facilities [9] are key resources used by PlantScan™ to process and store bulk data.

5 Semantic integration

The PODD ontology enables plant phenomics researchers to link from mesh results to the raw data that they were generated from. It also allows researchers to link from both mesh results and their recorded conclusions to shared phenomics ontologies which describe specific features of the plants. When used together, this enables scientists to trace the provenance of their results and conclusions based on well known concepts in phenomics ontologies.

Subsets of phenomics ontologies such as the Plant Ontology and the Crop Ontology were mapped into PODD by adding OWL constraints. These constraints enable PODD to verify that the use of classes and properties from these ontologies was consistent with the PODD ontology. For example, the Crop Ontology contains a class defining soil as "Sandy Loam", giving it the identifier "0000104". This was mapped into PODD to define a particular soil sample as being Sandy Loam using the triple: *poddSampleSandyLoamSoil a cropOntology* : _0000104.

6 Semantic publication

PODD provides a secure mechanism for publishing both human and machine readable descriptions of scientific experiments. It utilises the well-known DOI mechanism for publishing raw data files using DAP, and uses HTTP URIs to publish experiments using the PODD web interface.

Scientific journals increasingly require the data and provenance for articles to be available in a machine readable format. The DOI registrar that DAP uses, DataCite [11], was setup to provide unique identifiers for data items that can be attached to publications, which in turn may have their own DOIs.

By providing machine readable descriptions of scientific experiments, including semantic references to shared ontologies where possible, PODD enables the output from PlantScan™ to be interpreted and extended by others. The use of PODD URIs in other RDF documents enables scientists to extend the initial work using the Linked Data paradigm [12].

7 Conclusion

This paper described how the Phenomics Ontology Driven Data repository integrates with the PlantScan™ platform and CSIRO Data Access Portal to manage the complex workflows at the High Resolution Plant Phenomics Centre. This workflow keeps track of phenomics data, metadata and experimental processes and also provides a secure mechanism to share and publish scientific experiments in both human and machine readable formats.

References

1. Li, Y.F., Kennedy, G., Davies, F., Hunter, J.: PODD: An ontology-driven data repository for collaborative phenomics research. In Chowdhury, G., Koo, C., Hunter, J., eds.: The Role of Digital Libraries in a Time of Global Change. Volume 6102 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 179–188
2. CSIRO IM&T: CSIRO data access portal. http://data.csiro.au
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genet. **25** (2000) 25–29
4. Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F., Ware, D.: The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Research **36**(suppl 1) (2008) D449–D454
5. Fürber, C., Hepp, M.: Using sparql and spin for data quality management on the semantic web. In Abramowicz, W., Tolksdorf, R., eds.: Business Information Systems. Volume 47 of Lecture Notes in Business Information Processing. Springer Berlin Heidelberg (2010) 35–46
6. Kunze, J., Littman, J., Madden, L.: The bagit file packaging format (v0.97) (April 15 2011)
7. Summers, E.: Bagit python software. https://github.com/edsu/bagit
8. Library of Congress: Bagit java software. http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/
9. CSIRO IM&T: CSIRO advanced scientific computing. https://wiki.csiro.au/display/ASC
10. Paproki, A., Sirault, X., Berry, S., Furbank, R., Fripp, J.: A novel mesh processing based technique for 3d plant analysis. BMC Plant Biology **12**(1) (2012) 63
11. Brase, J.: Datacite - a global registration agency for research data. In: Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09. Fourth International Conference on. (2009) 257–261
12. Berners-Lee, T. http://www.w3.org/DesignIssues/LinkedData.html (2006)