COPO - Linked Open Infrastructure for Plant Data

F Shaw[1], A Etuk[1], A Gonzalez-Beltran[2], P Rocca-Serra[2],
P Kersey[3], R Bastow[4], S Sansone[2],
V Schneider[1], and R Davey[1]

[1] The Genome Analysis Centre, UK
[2] Oxford e-Research Centre, University of Oxford, UK
[3] European Bioinformatics Institute, Cambridge
[4] University of Warwick

Abstract. We present Collaborative Open Plant Omics (COPO), a brokering service between plant scientists and public repositories, which enables management, aggregation and publication of research outputs. COPO provides consolidated access to services and disparate information sources via web interfaces and Application Programming Interfaces (APIs). Users will be able to deposit and view open access data, as well as seamlessly pull such data into suitable analysis environments and subsequently track the outputs and associated metadata in COPO, thus creating a provenance trail from data to publication.

1 Introduction

The plant science domain has seen the advent of increasingly high throughput "-omics" technologies, resulting in more and larger datasets being produced. Researchers are realising the benefits of data sharing to promote their work and to accelerate discovery in science based on aggregated data. Many funding bodies and journals now require that data be made publicly available. Despite the opportunities that data sharing offers for recognition and reuse, many scientists still do not use public repositories, choosing instead to store data privately in their organisation's infrastructure. The reasons for this include lack of awareness of where and how to deposit data, lack of standards and common metadata, and a lack of funding to support archiving.

The large number and size of the datasets make them difficult to store, let alone download, making cloud-based analysis software highly desirable. However, submission formats to public repositories are heterogeneous, often requiring manual authoring of complex markup documents, taking scientists out of their fields of expertise.

COPO aims to streamline the process of data deposition to public repositories and data journals, by hiding much of the complexity of metadata capture and data management from the end-user. The Investigation/Study/Assay (ISA) infrastructure (www.isa-tools.org) is leveraged to provide the interoperability between metadata formats required for seamless deposition to repositories and to facilitate links to data analysis platforms. Logical groupings of artefacts (e.g. experimental metadata and results, PDFs, raw data, contextual supplementary

information) relating to a body of work are stored in COPO "collections" and represented by common open standards, which are publicly searchable. Bundles of multiple data objects themselves can then be deposited directly into public repositories through COPO interfaces.

2 Core Features

- User Interfaces
  - Web-based tools enable consolidated access to a range of data repositories
  - Allows for intuitive and intelligent labelling of data according to accepted standards
- Data Deposition, Querying and Publication
  - APIs enable deposition of data and metadata to public repositories such as the European Nucleotide Archive and Figshare
  - APIs allow querying of metadata and access to research artefacts deposited in public archives
  - APIs to allow submission of data and metadata to publication platforms such as Scientific Data and F1000
  - Prototype wizards to help users through the metadata attribution process

3 Metadata Management

The ISA model enables experimental metadata attribution and management of metadata formats, where scientific metadata comprises information about investigators, objectives, hypotheses, publications, subjects, experimental design, experimental workflow, and assays and related experimental data. ISA metadata is represented in ISA-JSON, and integrated within a broader subset of metadata, COPO-JSON, that encompasses infrastructural information relative to the platform itself. Both JSON implementations can be extended to JSON-LD linked data schemas. All JSON metadata fragments are stored in a MongoDB document-based database.

Where required, ISA converters allow traversal between representations of the same metadata, e.g. ISATab to/from ISA-JSON, and public repository formats are expressed as ISA configurations which are mapped to a COPO-JSON user interface (UI) model to power the COPO UI itself. In this way, we can quickly and easily adapt to new repositories or changes to existing repository schemas all the way from data representation to UI design.

4 Platform in Development

The COPO framework is being built using Python, Django, MongoDB, JSON-LD, ISATools, jQuery and Bootstrap technologies. We provide a single sign-on (SSO) mechanism via ORCiD which allows COPO to track service integration and rich user profile data. Anonymous users are able to search the COPO index for research artefacts, whereas deposition functionality is available to authenticated users only. The complexity of deposition services is hidden from end users, who simply fill out clean, intuitive web forms and story-driven wizards that use

the semantic level metadata to make inferences about what a user is submitting, and subsequently make suggestions based on previous submissions.

So far we have developed initial EMBL-EBI repository deposition support (European Nucleotide Archive (ENA), MetaboLights) facilitated by Aspera-powered data transfer and ISA API integration. Figshare deposition of secondary research artefacts (PDFs, images, figures, supplementary data, etc) is also supported.

5 Future Work

The large network of linked metadata that COPO will gather allows semantic meaning to be attached to research artefacts. Semantic inferences can then be made over artefacts providing a richer search experience than through text based search alone, enabling researchers to quickly find and use well-described publicly available datasets linked by a full "paper trail" of metadata. The provision of visualisation for graphs of linked metadata will aid discovery of useful connections between datasets, investigations and protocols. Support for more repositories and open publishing platforms such as GigaScience, F1000, Scientific Data and Dryad are planned, as well as integration with cloud-based analysis services such as Galaxy and iPlant.