COPO - Linked Open Infrastructure for Plant Data

F Shaw[1], A Etuk[1], A Gonzalez-Beltran[2], P Rocca-Serra[2], D Johnson[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. Collaborative Open Plant Omics (COPO) is a brokering service between plant scientists and public repositories, enabling management, aggregation and publication of research outputs. COPO consolidates access to services and disparate information sources via web interfaces and Application Programming Interfaces (APIs). Users will deposit and view open access data, as well as seamlessly pull such data into analysis environments. Subsequent accessions and associated metadata will be tracked in COPO, thus creating a provenance trail from data to publication.

1 Introduction

In plant science, high throughput "-omics" technologies have resulted in more and larger datasets. Researchers are realizing 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 in private infrastructure. This is may be due to unfamiliarity with services and technology, lack of standards and common metadata, or a lack of funding to support archiving. The large number and size of datasets make them difficult to store, let alone download, making cloud-based analysis tools essential. 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 meta data capture and data management from the end-user. The ISA (Investigation/Study/Assay) infrastructure (www.isa-tools.org) provides the interoperability between metadata formats required for deposition to repositories. Logical groupings of artifacts (e.g. experimental meta data 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 data objects can be deposited directly into public repositories (such as the European Nucleotide Archive, Figshare and F1000) through COPO interfaces.

2 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.

3 Platform in Development

The COPO framework is being built using Python, Django, MongoDB, JSON-LD, ISATools, jQuery and Bootstrap technologies. A single sign-on (SSO) mechanism provided via ORCiD, allows COPO to track service integration and rich user profile data. Anonymous users are able to search the COPO index for research artifacts. 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, subsequently making 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 artifacts (PDFs, images, figures, supplementary data, etc) is also supported.

4 Future Work

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