Bridging the Gap from Data to Publication in Plant Science

http://www.copo-project.org
copo@tgac.ac.uk
@copo_project

A Etuk[1], F Shaw[1], A Gonzalez-Beltran[2], P Rocca-Serra[2], A Abdul-Rahman[2], P Kersey[3], R Bastow[4], S Sansone[2], 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

The plant science domain has seen the advent of increasingly high throughput "-omics" technologies, resulting in larger datasets being produced more quickly and cheaply than ever before. Increasingly, researchers are realising the benefits of data sharing to promote their work, and many funding bodies and many journals now require that data produced during research be made publicly available at the time of publication. Despite the advantages that data sharing offers for recognition and data reuse, many scientists still do not use public repositories, choosing instead to store data privately in their organisation's infrastructure. The reasons for this are many and complex, but include lack of understanding of where and how to deposit data, lack of common metadata, and a lack of funding to support archiving.

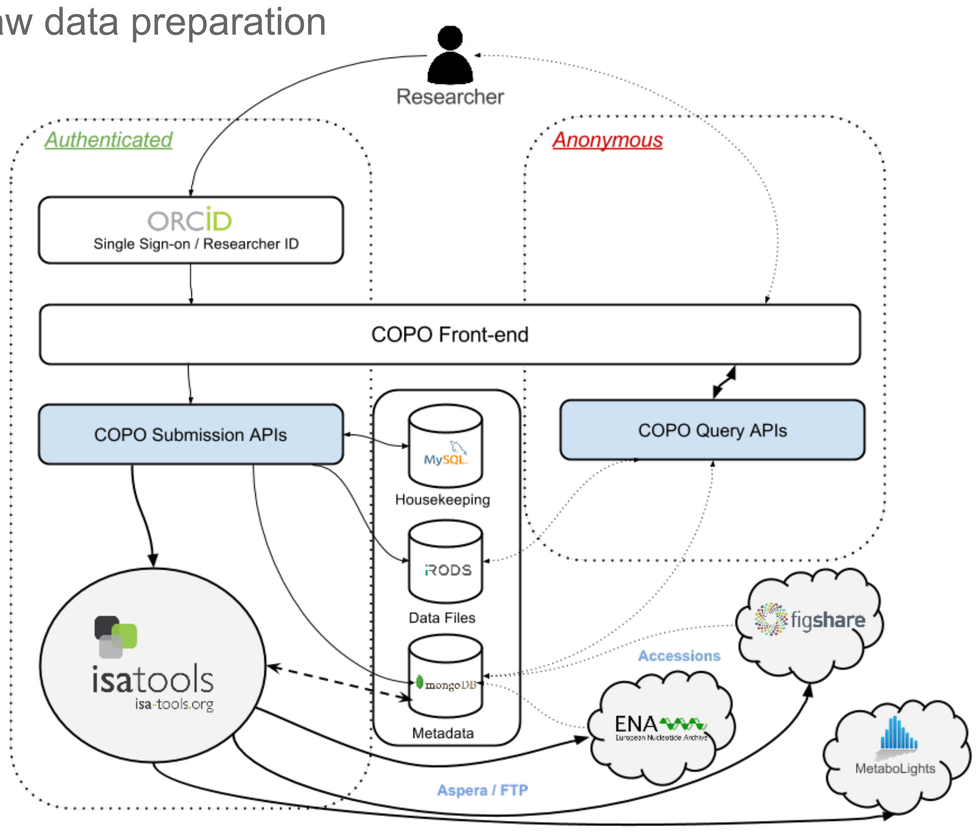
Metadata submission formats to public repositories are heterogeneous, often requiring manual authoring of complex markup documents, taking scientists out of their fields of expertise. Modern -omics methods are producing more data than can be easily stored, let alone downloaded, making cloud-based analysis software highly desirable.

We present **Collaborative Open Plant Omics (COPO)**, a brokering service between plant scientists and public repositories, which enables aggregation and publication of research outputs as well as providing easy access to existing data services comprising disparate sources of information via web interfaces and Application Programming Interfaces (APIs).

COPO streamlines the process of data deposition to public repositories by hiding much of the complexity of metadata capture and data management from the end-user. The 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. PDFs, raw data, contextual supplementary information) relating to a body of work are stored in COPO collections and represented by common standards, which are publicly searchable. Bundles of multiple data objects themselves can then be deposited directly into public repositories through COPO interfaces.

Core Features

- **User Interfaces**
  - Web-based tools enable consolidated access to a range of data repositories
  - Facilitate interactions between end-users and software suites for metadata annotation and raw data preparation
- **Data Deposition, Querying and Publication**
  - **APIs** facilitate **deposition of data and metadata** to public repositories (e.g. European Nucleotide Archive, Figshare)
  - **APIs** allow **querying of metadata** and **access to research artefacts** deposited in public archives
  - APIs to facilitate submission of data and metadata to **publication platforms** such as Scientific Data and F1000



Platform in Development

- Framework being built on Python, Django, MongoDB, JSON-LD, ISATools, jQuery and Bootstrap technologies
- **Single sign-on (SSO) leverages ORCID** to provide researcher profile data
- Anonymous users able to search the COPO index for research artefacts
- Deposition functionality available to authenticated users
- Initial **EMBL-EBI repository deposition support** (European Nucleotide Archive (ENA), MetaboLights) provided by Aspera-powered data transfer and ISA API integration.
- **Figshare deposition** of secondary research artefacts (PDFs, images, figures, supplementary data, etc) supported
- The complexity of deposition services is hidden from end users, who simply fill out clean, intuitive web forms

Metadata Management

- The **Investigation/Study/Assay (ISA) formats** and ISA software suite enable experimental metadata attribution, management and conversion between metadata formats
- Scientific metadata is captured in the **ISA model,** comprising information about investigators, objectives/hypotheses, publications, subjects, experimental design, experimental workflow and assays
- ISA-based metadata is represented as **ISA-JSON**, and integrated within broader **COPO-JSON** metadata that encompasses infrastructural information
- Both JSON implementations conform to **JSON-LD linked data schemas**
- JSON metadata fragments are stored in a **MongoDB** document-based database
- ISA converters allow **traversal between representations** of the same metadata, e.g. **ISATab** to/from **ISA-JSON** and public repositories formats

Future Work

- A **large network of linked metadata** means semantic meaning can be attached to research artefacts
- **Semantic inferences** can be made over artefacts providing a richer search experience than text based search alone
- Link to cloud-based analysis services such as **Galaxy** and **iPlant**
- Provision of **visualisation for network graphs** of linked metadata, allowing discovery of useful datasets, investigations and protocols
- **Support for more repositories** and open publishing platforms such as GigaScience, F1000, Scientific Data, Dryad

TGAC The Genome Analysis Centre™ · University of Oxford · The University of Warwick · EMBL-EBI · BBSRC bioscience for the future