Bio-GraphIIn: a graph-based, integrative and semantically-enabled repository for life science experimental data

Alejandra Gonzalez-Beltran✉, **Eamonn Maguire, Pavlos Georgiou, Susanna-Assunta Sansone, Philippe Rocca-Serra**

University of Oxford, United Kingdom

Received 9 July 2013; **Accepted** 10 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

We present the design and architecture of Bio-GraphIIn or "Biological Graph Investigation Index", an integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental metadata.

Motivation and Objectives

Biological and biomedical experiments often rely on a multiplicity of methods to monitor distinct biological signals from a given sample. This is the case, for example, in multi-omic experiments, where samples are studied using several post-genomic techniques (e.g. proteomics, transcriptomics). This variety of methods produce heterogeneous data, whose analysis results should be considered in an integrated manner to provide new insights at the systems biology level. Interpretation of results, as well as potential reuse of data, demands access to the provenance of data and sample information from the overall metadata payload.

Major primary databases, such as those maintained by the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), support technology-centric formats and, often, single data types. This state of affairs hampers realising truly integrative work. NCBI and EBI maintain centralised BioSample databases (Barrett *et al.*, 2012; Gostev *et al.*, 2012) to link data back to the originating samples. Yet, owing to the current one-way cross reference, technology-specific databases remain insular.

A second observation is that most of the existing databases support retrospective submissions. In other words, metadata and associated data are deposited in one go, as a bundle when experiments have been completed, and submission systems seldom allow for incremental deposition. Thus, when errors (from spelling mistakes to more serious issues) in the metadata arise, edits to existing repositories are not particularly straightforward, often requiring deleting the submission and re-submitting an entire dataset.

In this work, we present the design and architecture of Bio-GraphIIn (pronounced "biographene"), or "Biological Graph Investigation Index", an integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental metadata that:

1. relies on the ISA-TAB format for the description of experiments and their provenance, as it is multi-purpose and supports multiple data types (Rocca-Serra *et al.*, 2010) ; this makes Bio-GraphIIn an integrative repository;
2. exploits ISA2OWL project (Gonzalez-Beltran *et al.*, 2012) to provide a semantically explicit representation of ISA-TAB formatted experimental information, expressed in a graph model provided by semantic web technologies such as the Resource Description Framework (RDF);
3. supports semantically-rich and traversal queries within and across experiments; for example, involving elements from their design, such as retrieve all investigations whose treatment groups sizes are greater than four, or experiments corresponding to control animals, or retrieving all the data associated with certain samples;
4. enhances user experience by taking advantage of experimental design information to improve experimental metadata in more meaningful ways, such as the study groups defined as combination of factor values or as dynamic groupings;
5. supports Create, Read, Update, Delete (CRUD) operations in a web-interface, allowing for the creation of metadata from the experimental planning phase up to the data analysis. It also offers third party curation/correction possibilities that often lack.

This work improves over existing components from the Investigation/Study/Assay (ISA) Infrastructure (Rocca-Serra *et al.*, 2010), described briefly in the following paragraphs. The ISA infrastructure is a metadata tracking framework that was designed to deal with multi-omic experiments and it is based on three pillars:

1. the multi-purpose ISA-TAB format for the description of the experimental design, factors, what is being measured, the characteristics of the samples, the technology used, the assays, and so on;
2. a software suite allowing for the curation, creation, conversion to other formats such as those supported by public data repositories and RDF/OWL, links to analysis platforms and publication to data journals (Rocca-Serra *et al.*, 2010; Maguire *et al.*, 2013; Gonzalez-Beltran *et al.*, 2013);
3. an international and active user community grouped in the ISA commons (Sansone *et al.*, 2012).

The ISA infrastructure has implemented components for data persistence, namely: the Bio-Investigation Index[1] (BII) web-application, database and database manager tool. These components have been successfully used in systems such as the Stem Cell Discovery Engine (Ho Sui *et al.*, 2012). Contrary to other omics data repositories, as the BII is based on ISA, it supports multiple data types, without the need for a federated infrastructure where each data type is stored in a different endpoint. Similar to other data repositories, BII supports browsing of the stored experiments, comprehensive free text search, filtering according to organism, measurement, technology and platform, and programmatic access. However, BII is 'read only' and does not exploit any semantic features, nor does it allow 'slicing and dicing' across datasets. Bio-GraphIIn is the new generation of the BII and is designed to extend BII's functionality as per the five points above, addressing requests obtained from our user community.

Methods

Graphical user interface (GUI) and database backend - requirements elicitation.

In order to gather the requirements for the Bio-GraphIIn system, we analysed existing data repositories and their functionalities. We also contacted a number of biologists, some within the ISA commons and others not familiar with the ISA infrastructure, and performed semi-structured interviews. The main outcomes of this requirement analysis phase are the basis for the new functionality. This includes the support for CRUD operations, the GUI views at each level of the ISA hierarchy, the ability to retrieve raw and derived data files from samples that satisfy certain conditions on their characteristics across multiple experiments, which depicts not only explicit metadata from ISA-TAB but also elements from the experimental design, such as study groups. In terms of the interface design, biologists have once again stated their preference for spreadsheet-like interfaces, so Bio-GraphIIn relies on a tabular format for ISA-TAB creation.

Service-oriented architecture and implementation details

Figure 1 depicts the modular software architecture of the Bio-GraphIIn system. The web application is based on the Django Web Framework[2] and it relies on two RESTful web services when creating or uploading datasets: one for validation of the ISA-TAB files (wrapping the ISAvalidator code) and another one for conversion to RDF (wrapping the ISA2OWL conversion code). In addition, the system persists the graph representation of the ISA-TAB datasets into a graph database, and relies on SPARQL queries through a Storage And Inference Layer (SAIL), using the TinkerPop open-source stack[3] to interact with the Graphical User Interface.

Experimental metadata representation using semantic web and graph technologies.

The ISA-TAB format lends itself very well for a graph representation, as it describes the experimental workflow: material entities and data files can be represented as graph nodes, whose transformations are described by processes, specified in experimental protocols. The ISA2OWL project (Gonzalez-Beltran *et al.*, 2012) has developed a semantic representation of the ISA-TAB syntax, where the relationships between the highly interconnected ISA elements is made explicit and tagged with ontology terms. These include, for example, the relationships among material and data nodes and their related ISA processes. This

1 https://github.com/ISA-tools/BioInvIndex

2 http://djangoproject.com/

3 http://www.tinkerpop.com/

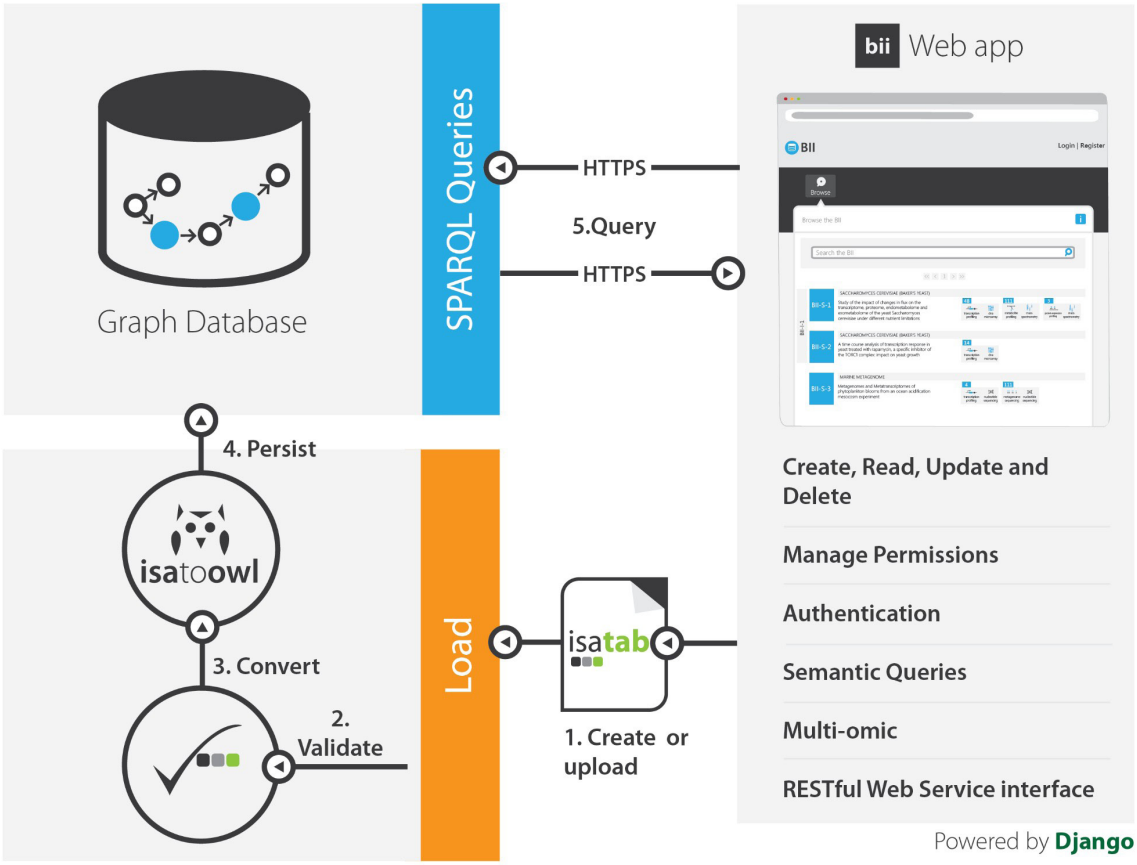

Figure 1. the software architecture of the Bio-GraphIIn system.

representation also allows to build links to external resources (e.g. publications, chemical compounds used in the description of experiments). The ISA2OWL component has been designed to support multiple semantic frameworks, which are specified through mapping files.

Graph databases, CRUD and query operations

For data persistence, Bio-GraphIIn relies on graph database technologies to exploit their ability to deal with highly interconnected data, their scalability and performance. In particular, their use was chosen due to the requirement to perform traversal queries such as those that relate samples to their associated data files. In order to be able to evaluate different existing technologies, the implementation relies on the TinkerPop Blueprints[4], a generic property graph model analogous to Java DataBase Connectivity (JDBC) but for graph databases. This implementation decision will allow us to evaluate different graph database implementations, including neo4j and RDF triple stores (e.g. Sesame).

As regards the CRUD operations, we expect update/delete operations to be more efficient with the underlying graph representation than storing the tabular representation (Brandizi *et al.*, 2012). For the query operations, we will also evaluate their performance for different underlying databases and present the preliminary results.

Results and Discussion

In this work, we presented the design and architecture of Bio-GraphIIn, a graph-based, integrative and semantically-enabled repository for heterogeneous biological and biomedical experimental data. Bio-GraphIIn is composed of a web application interface and a graph database back-end. It relies on a graph data model, as offered by semantic-web technologies such as RDF and OWL, to represent ISA-TAB datasets that describe biological and biomedical experiments relying on a variety of technologies.

4 http://www.tinkerpop.com/

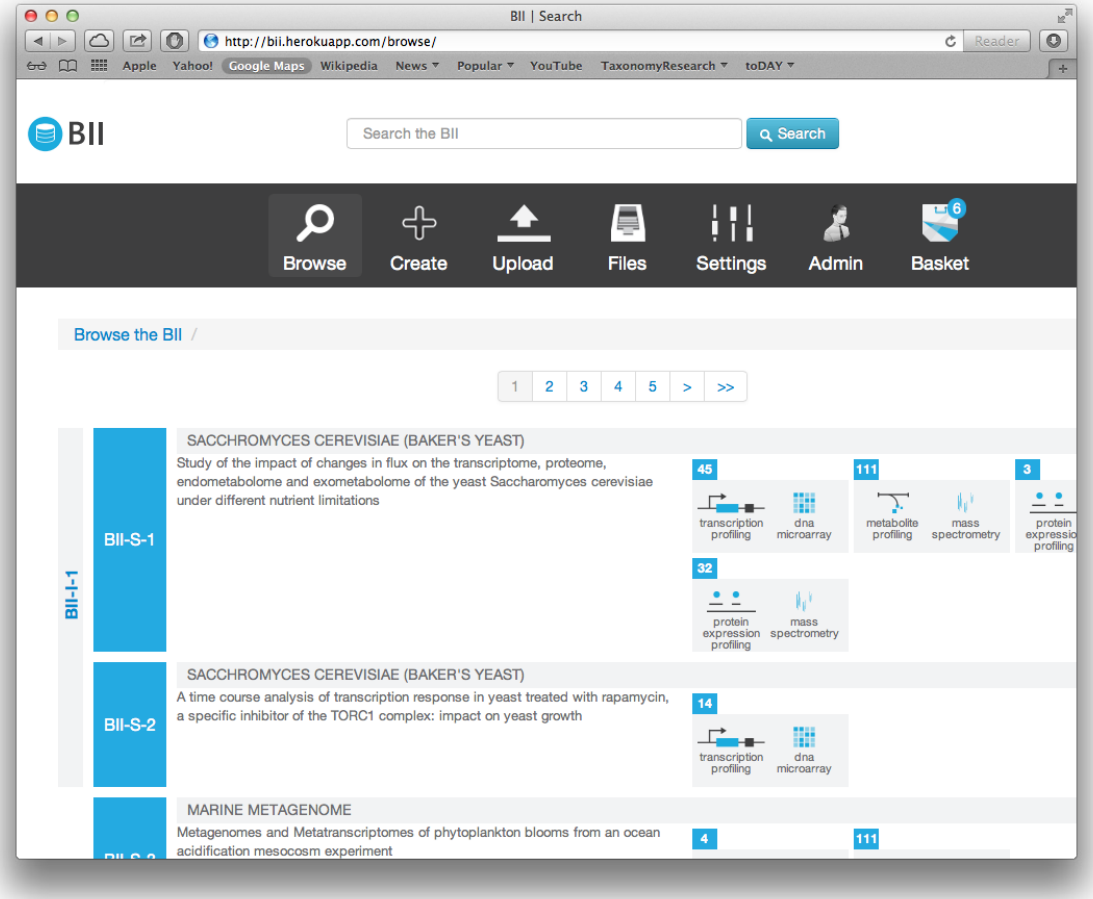

Figure 2. a screenshot of the Bio-GraphIIn system.

As ISA-TAB describes the experimental workflows, from the characteristics and preparation of the samples up to the data analysis performed and summary of results, Bio-GraphIIn supports the tracking of data provenance. Given the graph representation, traversal queries from samples to associate metadata are easily implemented. The latest instantiation of the Bio-GraphIIn database[5] is available on-line (see Figure 2 for a screenshot).

During the presentation, we will show the application using concrete multi-omic datasets and the operations that can be performed with them, and show our preliminary results on the performance analysis for uploading ISA-TAB datasets and for the CRUD operations implemented.

As future work, we will add a versioning feature. We will also investigate ways to associate the Bio-GraphIIn system with resources such as the Refinery Platform[6], a Django-based system for the integration of visualization and analysis of large-scale biological data based on the ISA-TAB format.

Acknowledgements

AGB, EM, SAS and PRS would like to thank their funding support to BBSRC BB/I000771/1, BB/I025840/1 and BB/J020265/1, EU COSMOS EC312941 and the University of Oxford e-Research Centre.

References

Barrett T, *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acid Research* **40**(Database issue), D57–D63. doi: 10.1093/nar/gkr1163

5 http://bii.oerc.ox.ac.uk

6 http://refinery-platform.org/

Brandizi M, *et al.* (2012) graph2tab, a library to convert experimental workflow graphs into tabular formats. *Bioinformatics* **28**(12), 1665-1667.doi:10.1093/bioinformatics/bts258

Gonzalez-Beltran A, *et al.* (2012) The open source ISA software suite and its international user community: knowledge management of experimental data. *EMBnet.journal* **18**, Suppl B, 35-37.

Gonzalez-Beltran A, *et al.* (2013) The Risa R/Bioconductor package: integrative data analysis from experimental metadata and back again. *BMC Bionformatics* In Press.

Gostev M, *et al.* (2012) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acid Research* **40**(Database issue), D64–D70. doi: 10.1093/nar/gkr937.

Ho Sui, *et al.* (2012) The Stem Cell Discovery Engine: an integrated repository and analysis system for cancer stem cell comparisons. *Nucleic Acid Research* **40**(Database issue), D984-991. doi: 10.1093/nar/gkr1051.

Maguire et al (2013) OntoMaton: a BioPortal powered ontology widget for Google Spreadsheets. *Bioinformatics* **29**(4), 525-527. doi: 10.1093/bioinformatics/bts718

Rocca-Serra P, *et al.* (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**(18), 2354-2356. doi: 10.1093/bioinformatics/btq415

Sansone S-A, *et al.* (2012) Towards interoperable bioscience data. *Nature Genetics* **44**, 121–126. doi:10.1038/ng.1054