The open source ISA software suite and its international user community: knowledge management of experimental data

Alejandra González-Beltrán✉, Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone
[1]University of Oxford, Oxford e-Research Centre, Oxford, United Kingdom

Motivation and Objectives

Both in academia and industry, data generation is currently in the order of petabytes in the biomedical domain. The availability of this massive amount of data brings with it many challenges, especially when considering data sharing and integration aiming at a later re-use. In this context, the adoption of standard formats, minimum information guidelines and terminologies/ontologies for the rich annotation of experimental data is crucial. Annotation is a time-consuming task that must be supported by software tools, which should also enable querying, linking, integrating, reasoning and analysing the data as well as the information about it.

The Investigation/Study/Assay (ISA) infrastructure (Rocca-Serra et al 2010) aims at facilitating this rich description of heterogeneous experimental data and supporting the different steps of the data management workflow. The infrastructure revolves around a general-purpose file format (ISA-Tab) and includes an open source software suite supporting compliance with community standards and dealing with the harmonization of the experimental metadata. The ultimate goal is to allow for the gradual progression from unstructured, usually non-digital metadata kept in lab notebooks to structure data that can be interpreted by machines (see Figure 1). The success of the ISA infrastructure is evidenced by the growing ISA Commons community (Sansone et al 2012), which encompasses increasingly diverse domains varying from metabolomics, (meta)genomics, proteomics, system biology to environmental health, environmental genomics and stem cell discovery (Ho Sui et al 2012).

We will present the components of the ISA infrastructure, the rationale behind them and their evolution. In particular, we will introduce our efforts to expand the infrastructure into three important directions: collaboration in a cloud environment, support for analysis with R, and the semantic web world. We will show use cases to exemplify the usage of the ISA infrastructure.

Methods

The ISA infrastructure software suite is the first one to support both experimentalists and curators in the description of multi-assay experiments (Rocca-Serra et al 2010). Studies using high-throughput (post)genomic technologies may involve multiple assays. For example, a system biology study in yeast (Castrillo et al 2007) includes transcription, metabolite and protein profiling using DNA microarray, NMR spectroscopy and mass spec-

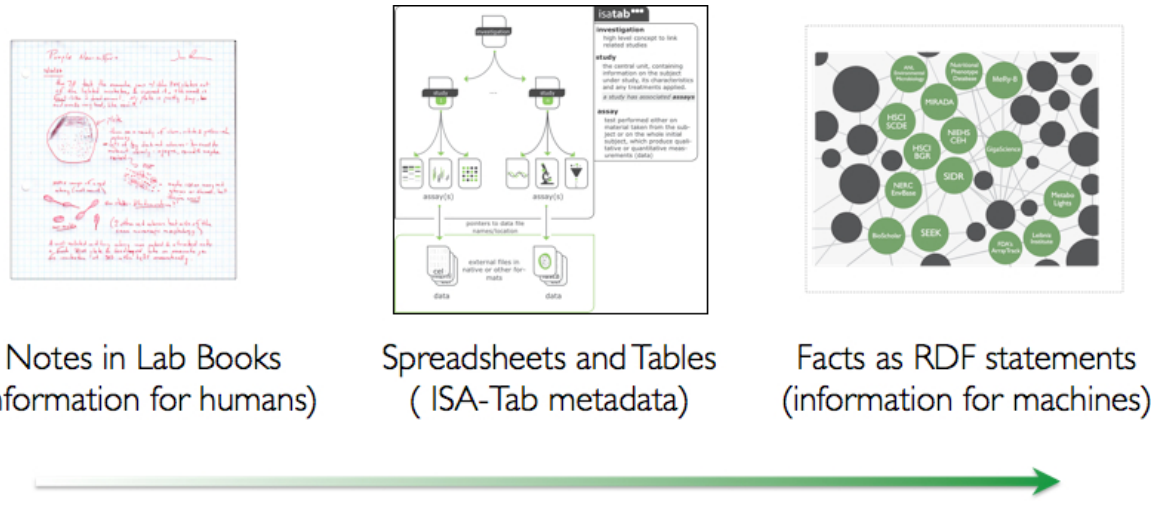

Figure 1. Experimental information with increasing level of structure.

troscopy, respectively.(BII-S-1: http://www.ebi.ac.uk/bioinvindex/study.seam?studyId=BII-S-1, (last accessed on 23rd July 2012).

The ISA-Tab format was designed to be domain-agnostic and includes the description of the experiments' contextual information such as the samples characteristics, the technology and measurement types, the instruments' parameters used, among other things. The availability of this metadata is crucial for the reproducibility of the experiments and posterior data re-use", i.e. add 'the' in 'for the reproducibility.

The ISA tools are *open source* (http://isatools.org/, GitHub: https://github.com/ISA-tools, (last accessed on 23rd July 2012) and follow a modular architecture, with standalone components providing functionality for each step of the data management workflow: *ISAcreator* offers spreadsheet-based data acquisition and curation relying on BioPortal services (http://bioportal.bioontology.org); *ISAconfigurato*r enables conformance to reporting standards (data formats, minimum information checklists, terminologies/ontologies); the *Bioinvestigation Index* provides for storage in a searchable repository with configurable access; *ISAconverter* facilitates reformatting for a growing number of acceptable formats; the tools enable submission of data to public repositories, validation and visualization.

Community engagement through case studies has been fundamental in the development of the infrastructure, and this community is grouped into the ISA commons (Sansone et al, 2012), ISA commons: http://isacommons.org/ last accessed on 23rd July 2012. The website attempts to list ISA users who have adopted or extended the format and/or tools for both public and private resources. An example resource using the ISA infrastructure is Metabolights (http://www.ebi.ac.uk/metabolights (last accessed on 27th September 2012) for metabolomics experiments.

The latest additions in the ISA software suite, described next, are: *OntoMaton, Risa and isa2owl.*

OntoMaton provides support for online collaborative data curation (Maguire et al 2012). It is implemented in Javascript using Google Apps Script API and offers functionality for searching bio-ontologies and for tagging free text with terms from ontologies. These functionalities rely on the NCBO BioPortal REST Services (http://www.bioontology.org/wiki/index.php/BioPortal_REST_services (last accessed on 23rd July 2012) and can be used for general semantic data annotation. Additionally, the ISAConfigurator 1.6 tool, which allows curators to create standards-compliant templates for ISA-Tab, has been extended to build templates to be included in the Google cloud environment and combined with OntoMaton. The OntoMaton Google Template can be found at: https://drive.google.com/templates?type=spreadsheets&q=ontomaton (last accessed on July 23rd 2012).

The Risa package, available in BioConductor 2.11 (http://www.bioconductor.org/packages/2.11/bioc/html/Risa.html (last accessed on 27th September 2012), offers methods for parsing ISA-Tab datasets and building R objects that can be used for analysis using domain specific packages. Risa also provides interfaces to some of these domain specific R packages, such as the xcms R package (Tautenhahn et al 2008) if there are mass spectrometry assays within the ISA-Tab data-set. Also, it is possible to augment the ISA-Tab metadata after analysis, and save the new files from R.

Last but not least, the isa2owl Java package follows our approach to expose ISA-Tab datasets to the Linked Data cloud. Our methodology relies on the definition of mapping files, aligning the ISA terminology with existing domain ontologies. A noteworthy mapping is that between ISA and the Ontology of Biomedical Investigations (OBI) (Brinkman et al 2010), which in turn is built in the Basic Formal Ontology (BFO) framework (Simon et al, 2006). Given such mapping, ISA-Tab datasets are parsed to populate the ontology, following Linked Data best practices such as the five star scheme whereby data is made available on the web in a structured non-proprietary format using URIs for identifying elements and linking to other data to provide context (Heath and Bizer 2011). This approach allows for semantic querying, discovery of links to other resources and reasoning over the ISA-Tab metadata.

Results and Discussion

The ISA infrastructure provides a comprehensive solution to the knowledge management challenges for experimental data in the biomedical domain. The modular architecture of the open source ISA tools enables users to adopt, and if necessary extend, one or more of the tools according to their specific needs. The underly-

ing ISA-Tab cross-domain format has proven to be generic enough and simple enough to be adopted by a large and growing community: the ISA commons.

The latest components of the ISA infrastructure offer support for collaborative semantic annotation in the cloud-computing environment of Google spreadsheets (OntoMaton), interface to popular data analysis packages (Risa), and a large number of new opportunities for querying, linking and reasoning about the data through transformation to RDF/OWL (isa2owl), making ISA-Tab datasets available in the linked data world. We are confident these tools will continue to facilitate important data management tasks in the context of massive amounts of data being generated in the biomedical domain.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council [grant BB/I025840/1 to SAS, BB/I000771/1, BB/I000917/1 to SAS]

References

1. Brinkman et al (2010), Modeling biomedical experimental processes with OBI. J Biomed Semantics 1(Suppl 1): S7. doi:10.1186/2041-1480-1-S1-S7
2. Castrillo et al (2007), Growth control of the eukaryote cell: a systems biology study in yeast, J. Biol. 6 (2):4. doi:10.1186/jbiol54
3. Heath and Bizer (2011), Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. doi:10.2200/S00334ED1V01Y2011WBE001
4. Ho Sui et al (2012), The Stem Cell Discovery Engine, Nucleic Acids Research. doi: 10.1093/nar/gkr1051
5. Maguire et al (2012). OntoMaton: bringing semantic annotation to Google spreadsheets for collaborative data management. Manuscript submitted.
6. Rocca-Serra et al (2010), ISA software suite. Bioinformatics, 26. doi:10.1093/bioinformatics/btq415
7. Sansone et al (2012), Toward interoperable bioscience data, Nature Genetics, 27. doi:10.1038/ng.1054
8. Simon et al (2006), Formal ontology for natural language processing and the integration of biomedical databases. Int J Med Inform. 2006 Mar-Apr;75(3-4):224-31. doi: 10.1016/j.ijmedinf.2005.07.015
9. Tautenhahn et al (2008), Highly sensitive feature detection for high resolution LC/MS BMC Bioinformatics, 9:504, doi:10.1186/1471-2105-9-504