The ISA Commons – Towards interoperable data in bioscience

Susanna-Assunta Sansone, Philippe Rocca-Serra, Eamonn Maguire on behalf of the ISA Commons’ community.
University of Oxford, Oxford e-Research Center, Oxford, UK
Correspondence: sa.sansone@gmail.com

ABSTRACT
To make full use of ‘omics research data, the bioscience community needs to adopt technologies and reward mechanisms that support interoperability and promote the growth of an open ‘data commoning’ culture. Here we present the prerequisites for data commoning and describe an established and growing ecosystem of solutions using shared concepts to support that vision. The ISA commons is a growing exemplar ecosystem of data curation and sharing solutions built on a common metadata tracking framework, providing tools and resources to create and manage large, heterogeneous data sets in a coherent manner (Sansone*, Rocca-Serra* et al., Nature Genetics, in press).

To tackle complex scientific questions, experimental data sets from different ‘omics sources often need to be harmonized with respect to their structure, formatting and annotation to open their content to (integrative) analysis. Vast swathes of bioscience data remain inaccessible; locked in esoteric formats, described using non-standard terminology, lacking sufficient contextual information or never actually shared due to the perceived cost or futility of the exercise. This clear loss of value continues to engender community standardization initiatives and drives the ongoing conversation about the encouragement of data sharing through appropriate reward mechanisms. However, the mountain of technical frameworks needed to support data sharing between communities continues to inhibit the development of general tools for data management and reuse. New solutions are required that deliver economies of scale in data capture and inherently support data integration, rendering the process of data capture and annotation scalable in the face of the current ‘omics data bonanza.

We present how a family of data producers, consumers and service providers enable the (invisible) use of the Investigation/Study/Assay (ISA) framework (Rocca-Serra et al, Bioinformatics, 2010; www.isa-tools.org) to harmonize metadata descriptions of disparate datasets, enabling data commoning through invisible compliance with the community standards described in the BioSharing catalogue (Field*, Sansone* et al. Science, 2009; www.biosharing.org). These collaborative groups include the BGI’s GigaScience journal and associated database just announced to the world, the EMBL-EBI’s new public metabolomics system and the Harvard’s Stem Cell discovery engine. These collaborative groups – along many others - are, in essence, on the
path to building a data commons, serving an increasingly diverse set of domains including environmental health, environmental genomics, metabolomics, (meta)genomics, proteomics, stem cell discovery, system biology, transcriptomics, toxicogenomics, but also communities working to characterize nucleic acid structures and to build a library of cellular signatures.

Our work is very timely and pivotal nowadays, as funders and journals require that researchers share their data, and encourage the enrichment and standardization of experimental context or metadata (i.e., sample characteristics, technology/measurement types and instrument parameters) to make datasets reliable and reusable, particularly if used to underpin future investigations. This situation, therefore, demands better curation at the source by data generators or community curators, using (open source) software with automated content validation that empower use of community-developed standards but hides it from users. Our job is to minimize the effort required, contributing to the realization of the data sharing vision and empower scientists to take data management and sharing into their own hands.

While efforts as the ISA Commons hold promise, it is a drop in the ocean compared to the daunting challenges ahead; for example, the integration of clinical and biological data in translational medicine; and the establishment of mechanisms to support credit for data sharing, which would benefit data producers for making their data accessible. Nonetheless, the vision of data sharing through a ‘commons’ is entirely technologically possible; but communities need agree the largely-organizational changes required. The continued collaborative development and uptake of standard frameworks, and the emergence of compliant tools and interoperable data sets, illustrates the potential of the horizontal, synergistic approach that is data commoning. Such horizontal integration transcends individual life science domains and assay- or technology-focused communities.