Toward interoperable bioscience data

Susanna-Assunta Sansone[1,39], Philippe Rocca-Serra[1,39], Dawn Field[2], Eamonn Maguire[1], Chris Taylor[2,3], Oliver Hofmann[4], Hong Fang[5], Steffen Neumann[6], Weida Tong[7], Linda Amaral-Zettler[8], Kimberly Begley[4,9], Tim Booth[2], Lydie Bougueleret[10], Gully Burns[11], Brad Chapman[4], Tim Clark[12,13], Lee-Ann Coleman[14], Jay Copeland[15], Sudeshna Das[12,13], Antoine de Daruvar[16,17], Paula de Matos[3], Ian Dix[18], Scott Edmunds[19], Chris T Evelo[20,21], Mark J Forster[22], Pascale Gaudet[23,24], Jack Gilbert[25], Carole Goble[26], Julian L Griffin[27,28], Daniel Jacob[17,29], Jos Kleinjans[30], Lee Harland[31], Kenneth Haug[3], Henning Hermjakob[3], Shannan J Ho Sui[4], Alain Laederach[32], Shaoguang Liang[19], Stephen Marshall[33], Annette McGrath[34], Emily Merrill[13], Dorothy Reilly[33], Magali Roux[35,36], Caroline E Shamu[15], Catherine A Shang[37], Christoph Steinbeck[3], Anne Trefethen[1], Bryn Williams-Jones[31], Katherine Wolstencroft[26], Ioannis Xenarios[10,38] & Winston Hide[4]

To make full use of 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 describe the prerequisites for data commoning and present an established and growing ecosystem of solutions using the shared 'Investigation-Study-Assay' framework to support that vision.

To tackle complex scientific questions, experimental datasets from different sources often need to be harmonized in regard to structure, formatting and annotation so as to open their content to (integrative) analysis. Vast swathes of bioscience data remain locked in esoteric formats, are described using nonstandard terminology, lack sufficient contextual information or simply are never shared due to the perceived cost or futility of the exercise. This loss of value continues to engender standardization initiatives and drives the ongoing conversation about the encouragement of data sharing through appropriate reward mechanisms.

Minimum reporting guidelines, terminologies and formats (hereafter referred to generally as reporting standards) are increasingly used in the structuring and curation of datasets, enabling data sharing to varying degrees. However, the mountain of frameworks needed to support data sharing between communities inhibits the development of tools for data management, reuse and integration. Here we describe a way in which a group of data producers and consumers work within an invisible metadata framework that enables the coordinated use of reporting standards by service providers and circumvents many of the problems caused by data diversity. The same framework enables researchers, bioinformaticians and data managers to operate within an open data commons.

A full list of author affiliations appears at the end of the paper.

From reusable data to reproducible research

Shared, annotated research data and methods offer new discovery opportunities and prevent unnecessary repetition of work. Although funding agencies, journals and community initiatives encourage good data stewardship and sharing through the use of community reporting standards, data sharing remains challenging[1–3]. More significant coordination has occurred in the food and drug regulatory arena[4] and in commercial science, where investments in procedures and tools that integrate external sources with internal data now enhance decision-making processes[5].

Funding agency 'encouragement' has normally taken the form of top-down data sharing policies. Increasingly, however, funding agencies are also requiring specific data management, preservation and sharing plans in grant applications and are monitoring adherence[6]. Such an approach requires researchers to follow or develop best practices collaboratively. These practices are also emerging organically through the provision of independent databases, tools and curators, driven by advocates of the sharing of both pre- and post-publication data[7,8]. To build an interoperable open data ecosystem will require leveraging all of these positive efforts and further increasing community buy-in.

Time to leap outside the box

Overall, most stakeholder groups accept the principles of data sharing, but in practice, achieving compliance is challenging, especially when new technologies or combinations of technologies are employed. The current wealth of domain-specific reporting standards provides proof of stakeholders' engagement with standardization and sharing, but the use of combinations of technologies presents challenges[9,10]. Descriptions of investigations of biological systems in which source material has been subject to several kinds of analyses (for example, genomic sequencing, protein-protein interaction assays and the measurement of metabolite concentrations) are particularly challenging to share as coherent units of research because of the diversity of reporting standards with which the parts must be formally represented. Equally, most repositories are designed for specific assay types, necessitating the fragmentation of complex datasets[11–15].

One way forward is to establish reciprocal data exchange between major repositories, but budgetary constraints limit such activities[15,16], and a crop of differing methodologies still imposes barriers[11,12].

Researchers acting as data consumers also face challenges when the component parts of an investigation are scattered across databases. Fragmented datasets can only be reassembled by those equipped to navigate the various reporting guidelines, terminologies and formats involved[17]. Cross-cutting, topic-specific reference datasets have been assembled, but predominantly by large initiatives (such as Sage Commons) and programs (such as ENCODE or the US National Institutes of Health–National Institute of Allergy and Infectious Diseases' Bioinformatics Resource Centers (BRCs)). These limitations fuel the indifference researchers feel about investing significant effort to share their data[18].

As the main facilitators of data sharing, major public repositories are evolving to support the structure and detail increasingly present in complex, multipart datasets (such as the US National Center for Biotechnology Information's BioSample system). By importing data from external files under their own schemata, databases provide badly needed integration. The speed of this evolution is dependent on access to highly skilled biocurators able to generate and validate complex annotations, increasing the pressure on data producers to quality check data before submission[19].

ISA commons: a part of the data-commoning revolution

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 'data bonanza'. Here we refer to efforts toward such positive solutions as 'data commoning'. **Box 1** presents an exemplar ecosystem of data curation and sharing solutions from groups working together to create a cross-domain data sharing vision of the future. These collaborative groups 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, systems biology, transcriptomics and toxicogenomics, but also communities working to characterize nucleic acid structures and to build a library of cellular signatures. This emerging commons depends on its participants' use of the metadata categories 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement). This so-called ISA framework is the backbone upon which the discovery, exchange and informed integration of data sets articulate with one another.

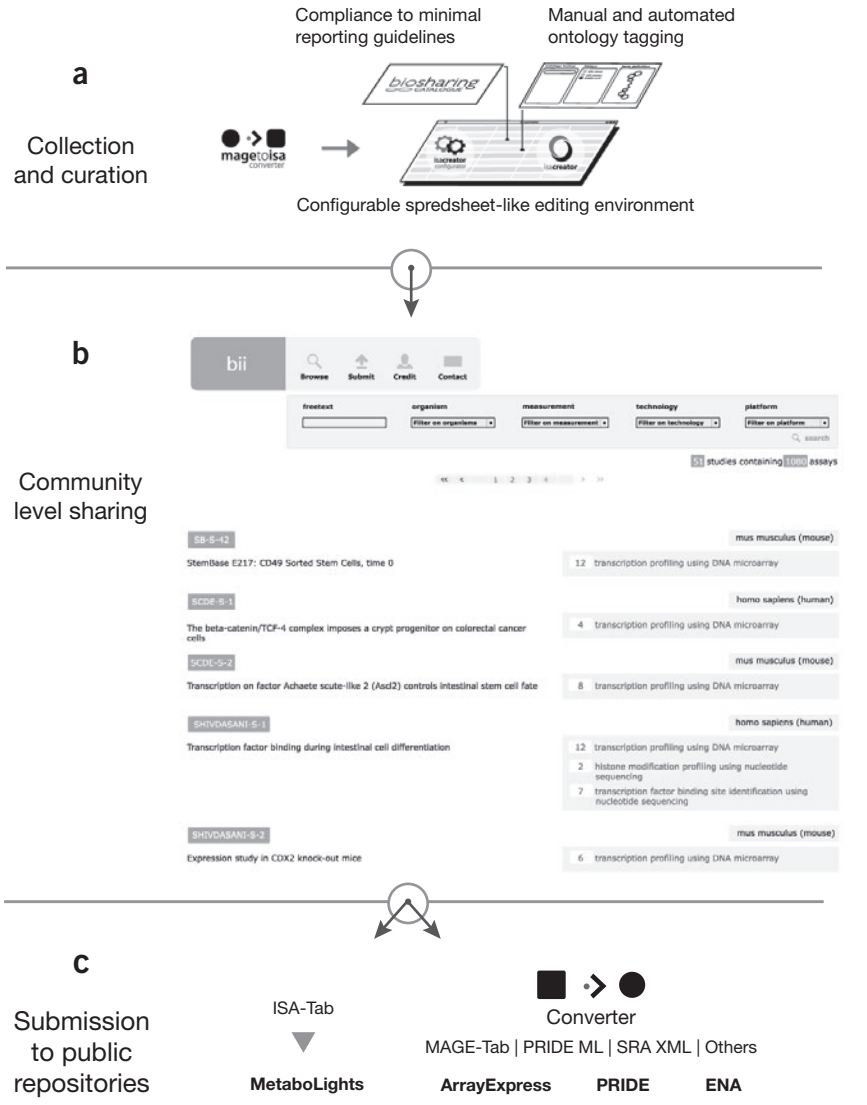

Figure 1 The ISA framework in action in the stem cell–based system of the Harvard Stem Cell Institute (HSCI). The data management workflow of the HSCI's Stem Cell Discovery Engine (SCDE) system, powered by the ISA framework. (**a**) Curators use the ISAconfigurator and ISAcreator software modules to consistently curate a variety of internally generated stem cell-based genomics profiles according to community-developed minimum information guidelines and terminologies; published transcriptomics-based studies are also collected via the MAGEtoISA module, then curated and enriched for consistency. (**b**) Consistently represented investigations are loaded in the BioInvestigation Index (BII) component that stores and serves the (public and private) data sets to the HSCI and wider community. (**c**) Upon publication, investigations are directly submitted to those public repositories using ISA-Tab format, or converted to/from other supported formats via the ISAconverter.

At the heart of the ISA framework is the extensible, hierarchical 'ISA-Tab' file format[20] that can be used alone or as a template for a variety of spreadsheet-based formats for data sharing[21]. ISA-Tab was developed by mapping a number of public repositories' submission formats onto one structure for representing experimental metadata, leveraging common elements while keeping data files external in their native or community-specific formats. ISA-Tab offers the chance for both project-specific and public repositories to adopt a common file format for representing experimental metadata, increasing the flow of richly described investigations into the public domain.

The modular ISA software suite, which implements the ISA-Tab format, acts to

BOX 1 EXAMPLES OF THE GROWING ECOSYSTEM OF ISA COMMONS PARTICIPANTS

To better understand the utility of the ISA framework, we present here a series of brief case studies in which one or more of its elements have been embedded in open-source systems that facilitate standards-compliant collection, curation, management, distribution and reuse of data within a community. Other emerging systems include MeRy-B and the Biomedical Information Research Network (BIRN) BioScholar Knowledge Management system, the Harvard Medical School Library of Integrated Network–based Cellular Signatures (LINCS) effort and ArrayTrack at the Center for Bioinformatics of the US Food and Drug Administration (FDA), along with internal systems at the Leibniz Institute of Plant Biochemistry, the Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites (MIRADA LTERs), the International Census of Marine Microbes (ICoMM), the Environmental Microbiology activities at the Argonne National Laboratory, the Bioplatforms Australia consortium and the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia. Furthermore, ISA-Tab is used to facilitate the sharing of chemical and enzymatic structure-probing data in the Single Nucleotide Resolution Nucleic Acid Structure Mapping (SNRNASM) annotation guidelines. An instance of selected ISA software components is also being integrated as part of an extended workflow for a microarray gene expression resource at The Novartis Institutes for BioMedical Research (NIBR) to facilitate research aimed at drug discovery and development.

GigaScience. Now the world's largest sequencing center, BGI (formerly known as the Beijing Genomics Institute) is centrally involved in many large international sequencing projects. To speed the review, publication and sharing of large-scale data sets, BGI has launched GigaScience, a combined database and journal using BGI's cloud computing and server infrastructure. GigaScience will use the ISA Infrastructure to capture many kinds of study and assay metadata along with relationships between data set components. Through implementation of DataCite's Digital Object Identifiers (DOIs), data sets will be fully trackable and citable, supporting the awarding of credit to data producers.

HSCI Blood Genomics Repository. The Harvard Stem Cell Institute (HSCI) Blood Genomics Repository holds hematopoietic (blood) stem cell data from HSCI Blood program researchers studying the molecular and cellular characteristics and pathways involved in hematopoietic stem cell self-renewal. The repository comprises heavily curated data from gene expression, epigenetic modification and transcription factor–binding studies using various technologies and platforms, and it is made available in the form of ISA-compatible files.

HSCI Stem Cell Discovery Engine. The Stem Cell Discovery Engine (SCDE) is a manually curated public resource with a focus on cancer, powered by the ISA software suite and hosted by the HSCI. SCDE handles the submission, integration, visualization and dissemination of high-throughput studies and provides linked molecular analysis through Galaxy to experimental metadata. Data sets selected for inclusion are annotated using public resources and then expertly curated to ensure accuracy, consistency, compliance with relevant reporting requirements and appropriate use of terminologies.

MetaboLights. The MetaboLights resource will include the first public cross-species, cross-application database at the European Bioinformatics Institute (EBI) accepting metabolite structures and other data from metabolomic experiments. A curated reference layer with spectroscopic, chemical and biological information about metabolites will be developed to enhance submitted data. The project uses the ISA infrastructure and will publish customized templates for capturing study information, and assays using nuclear magnetic resonance and mass spectrometry, using common terminologies.

NERC EnvBase. The UK Natural Environmental Research Council's (NERC) Environmental Bioinformatics Centre (NEBC) collects and catalogs data sets from environmental and functional genomics investigations by the NERC research community and their international collaborators. Using the ISA infrastructure, the NEBC's data catalog, EnvBase, has recently been expanded to hold and serve investigations curated to meet community-developed standards requirements—in particular, standards developed and maintained by Genomic Standards Consortium (GSC) relevant to metagenomic investigations. The collection of experimental metadata at source is facilitated by the deployment of the editor component on a Bio-Linux platform.

NIEHS Center for Environmental Health. The National Institute of Environmental Health Sciences' Center for Environmental Health at Harvard works to preserve a diverse array of data from environmental research, population-, patient- and laboratory-based studies, and published data sets imported from other databases. The ISA infrastructure serves as the base for this institutional repository and will also serve as a 'resource locator', allowing new investigators to quickly identify collaborators and available preliminary data from historical studies, reducing redundancy.

Nutritional Phenotype Database. The Nutritional Phenotype Database (dbNP) facilitates the sharing of large-scale laboratory clinical intervention and observation studies relating to food intake between Dutch research groups and with international consortia. Their harmonization of study description, following the ISA approach, allows cross-experiment comparisons and facilitates the querying of data at the biological outcome level (for example, by pathway).

SEEK. The SEEK is a web-based registry and repository for systems biology data, models and experiments. Originally developed for SysMO, a pan-European consortium studying dynamic molecular processes in microorganisms, it has since been adopted to handle data sets from other large systems biology projects. The SEEK 'experimental contexts' follow the ISA approach for conversion to other formats.

SIDR. The Standards-based Infrastructure with Distributed Resources (SIDR) works to collect, preserve and disseminate genomics and functional genomics data sets from a variety of French National Centre for Scientific Research's groups. The various experiment types are structured following the ISA approach, identified with DOIs, and also provided in several formats. Part of a broader approach, SIDR aims to address complex issues in systems biology and is being customized for the translational medicine domain.

(i) regularize local collection and management of experimental metadata, (ii) reduce the adoption barrier for using community minimum reporting guidelines and terminologies through customizable configuration, (iii) facilitate consistent curation at source and (iv) support direct submission to a growing number of public repositories, both in ISA-Tab format (such as MetaboLights and the other systems shown in **Box 1**) and through conversion to other supported formats[12–14]. An example of the ISA framework in action is illustrated by the Harvard Stem Cell Institute (HCSI)'s Stem Cell Discovery Engine (SCDE)[22] and shown in **Figure 1**.

Without community-level harmonization and interoperability, many community projects risk becoming data silos, aggravating the problem. Using the shared, metadata-focused ISA framework, it is now possible to aggregate investigations in community 'staging posts', merge them in various combinations, perform meta-analyses and more straightforwardly submit to public repositories. Furthermore, simplifying the integration of bioscience data can only speed systems biology research[23] and improve the ability of the R&D community to utilize shared data[24].

The growing number of communities using the ISA framework adds credibility to this metadata-focused data sharing vision. Taking this a step further, **Figure 2** shows how these communities' systems—a mix of public and internal tools that use ISA software components or, minimally, the ISA-Tab format—will progressively interrelate to build the 'ISA commons'. Activities are already underway under the auspices of the World Wide Web Consortium (W3C) Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG)'s Scientific Discourse task force to generate serialized ISA-Tab metadata in compliance with the recommendations of the international Linked Data community[25]. Semantic integration of bioscience data with the wider corpus of human knowledge then becomes more straightforward.

BioSharing: standard cooperating procedures

It is widely acknowledged that unlocking shared data promises to accelerate discovery, but this process requires new models for the way we collaborate[1–3,5,6,17,18,26]. But reporting standards often have different levels of maturity, and inevitably, duplication of effort. Communication between standards initiatives is pivotal to ensure that a common or at least complementary set of standards exists and is widely used by the academic and commercial sectors to maximize the utility of shared data. Building on the effort of the Minimum Information for Biological and Biomedical Investigations (MIBBI) portal[10], the BioSharing initiative works to strengthen collaborations between researchers, funders, industry and journals and to discourage redundant (if unintentional) competition between standards-generating groups[27]. The BioSharing catalog maps the landscape of standards and the systems implementing them, and it also works to build graphs of complementarities in scope and functionality. In time and after consultation, a set of criteria for assessing the usability and popularity of standards will be implemented to maximize their adoption and use to assist the

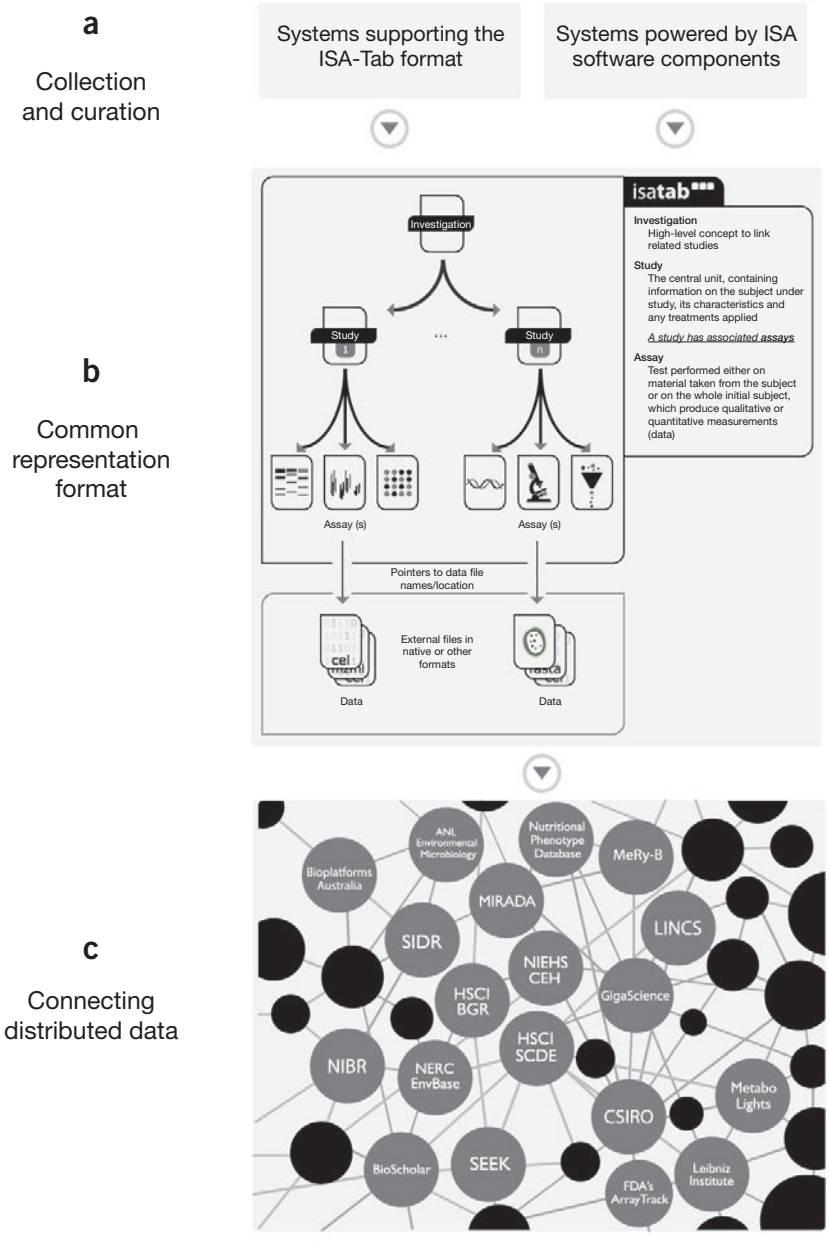

Figure 2 Building the 'ISA commons', a growing ecosystem of resources that work to provide a data commons. (**a**) Data sets of interest to each community are collected and curated. (**b**) Capture systems, either powered by the ISA software suite or supporting the hierarchical ISA-Tab structure, deliver a common representation of experimental content that transcends individual domains. (**c**) To achieve broader data integration, the next step is to explore the growing Linked Data universe. The European Innovative Medicines Initiative (IMI) Open PHACTS project, for example, will use semantic web approaches to make existing knowledge available for linking, querying and where possible, reasoning. This project will benefit greatly from study descriptions that draw on the ISA model to connect quantified information held in semantic triple stores to data from actual experiments performed. As a result, the project will connect public and private datasets to genomics resources, enabling the combination of existing and new experimental data.

virtuous data cycle—from generation to standardization through publication to subsequent sharing and reuse.

The research community requires solutions that accommodate the current 'wealth' of standards and resources, but hides it from users, thereby simplifying their efforts to meet (or ideally, exceed) applicable reporting requirements. Although ongoing activities hold promise, they are a drop in the ocean compared to the daunting challenges ahead: for example, the integration of clinical and biological data in translational medicine[28] and the establishment of mechanisms to support credit for data sharing, which would benefit data producers for making their data accessible (for example, refs. 29,30).

Nonetheless, the vision of data sharing through a 'commons' is entirely technologically possible; communities simply need agree on the largely organizational changes required. The continued collaborative development and uptake of standard frameworks, and the emergence of compliant tools and interoperable data sets such as we have described, 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.

A growing movement

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, and allowing users of (parts of) data sets to 'connect the metadata dots'. We are open to coordinating efforts with other data commons working on similar and related aspects of the same problem, who we invite to adopt and contribute to the further evolution of the ISA framework—the results of years of effort to agree to a basic *lingua franca* for the standards community.

We urge new communities interested in breaching the boundary of their own biodomain to join the growing ISA network and the BioSharing initiative, thereby contributing to the realization of this data-sharing vision: to empower ever more scientists to take data management and sharing into their own hands, using community standards while remaining blissfully unaware of the underlying complexities of the implementation of those standards.

Note: The views presented in this article do not necessarily reflect those of the US Food and Drug Administration.

URLs. BGI, http://en.genomics.cn/; BioLinux, http://nebc.nerc.ac.uk/tools/bio-linux; Bioplatforms Australia, http://bioplatforms.com.au/; CSIRO, http://www.bioinformatics.csiro.au; **BioSharing**, http://biosharing.org/; BIRN BioScholar Knowledge Management system, http://bmkeg.isi.edu/; DataCite's DOIs, http://www.datacite.org/; dbNP, http://www.dbnp.org/; ENCODE, http://encodeproject.org/ENCODE/dataStandards.html; Galaxy, http://galaxy.psu.edu/; GSC, http://gensc.org/; GigaScience, www.gigasciencejournal.com/; HSCI's SCDE, http://discovery.hsci.harvard.edu/; HSCI's Blood Genomics Repository, http://bloodprogram.hsci.harvard.edu/; ICoMM, http://icomm.mbl.edu/; IMI Open PHACTS, http://www.openphacts.org/; **ISA Commons**, http://www.isacommons.org/; **ISA software suite and ISA-Tab**, http://www.isa-tools.org/; Leibniz Institute of Plant Biochemistry, http://www.ipb-halle.de/en/research/stress-and-developmental-biology/research/bioinformatics-mass-spectrometry/research-projects; LINCS, http://lincs.hms.harvard.edu/; Linked Data, http://linkeddata.org/; MeRy-B, http://www.cbib.u-bordeaux2.fr/MERYB/index.php; http://listserver.ebi.ac.uk/mailman/listinfo/metabolights/; MIRADA LTERs, http://amarallab.mbl.edu/mirada/mirada.html/; NIEHS' Center for Environmental Health, http://www.hsph.harvard.edu/research/niehs; NCBI's BioSample, http://www.ncbi.nlm.nih.gov/biosample; NERC EnvBase, http://bii.nwl.ac.uk/; NIBR, http://www.nibr.com/; NIH-NIAID's BRCs (Bioinformatics Resource Centers), http://www.niaid.nih.gov/labsandresources/resources/brc; Sage Commons, http://sagebase.org/commons/; SEEK, http://www.sysmo-db.org/; SIDR, http://sidr-dr.inist.fr/; SNRNASM, http://snrnasm.bio.unc.edu/; SysMO, http://www.sysmo.net/; http://www.fda.gov/AboutFDA/CentersOffices/OC/OfficeofScientificandMedicalPrograms/NCTR/WhatWeDo/NCTRCentersofExcellence/ucm078990.htm/; W3C HCLSIG Scientific Discourse task force, http://www.w3.org/wiki/HCLSIG/SWANSIOC.

ACKNOWLEDGMENTS

S.-A.S. and P.R.-S. owe debts of gratitude to the many collaborators involved in the ISA Commons, and particularly to the EU CarcinoGENOMICS partners and developers who have contributed to the ISA framework and to the creation of the Commons over the years. We specifically acknowledge M. Brandizi and A. Santarsiero. The authors also acknowledge the following funding sources in particular: UK Biotechnology and Biological Sciences Research Council (BBSRC) BB/I000771/1 to S.-A.S. and A.T.; UK BBSRC BB/I025840/1 to S.-A.S.; UK BBSRC BB/I000917/1 to D.F.; EU CarcinoGENOMICS (PL037712) to J.K.; US National Institutes of Health (NIH) 1RC2CA148222-01 to W.H. and the HSCI; US MIRADA LTERS DEB-0717390 and Alfred P. Sloan Foundation (ICoMM) to L.A.-Z.; Swiss Federal Government through the Federal Office of Education and Science (FOES) to L.B. and I.X.; EU Innovative Medicines Initiative (IMI) Open PHACTS 115191 to C.T.E.; US Department of Energy (DOE) DE-AC02-06CH11357 and Arthur P. Sloan Foundation (2011-6-05) to J.G.; UK BBSRC SysMO-DB2 BB/I004637/1 and BBG0102181 to C.G.; UK BBSRC BB/I000933/1 to C.S. and J.L.G.; UK MRC UD99999906 to J.L.G.; US NIH R21 MH087336 (National Institute of Mental Health) and R00 GM079953 (National Institute of General Medical Science) to A.L.; NIH U54 HG006097 to J.C. and C.E.S.; Australian government through the National Collaborative Research Infrastructure Strategy (NCRIS); BIRN U24-RR025736 and BioScholar RO1-GM083871 to G.B. and the 2009 Super Science initiative to C.A.S.

AUTHOR CONTRIBUTIONS

S.-A.S. and P.R.-S. designed and led the development of the ISA framework and the BioSharing catalogue. D.F. and S.-A.S. are the cofunders of the BioSharing initiative. E.M. is the lead engineer of the ISA framework and, with P.R.-S., of the BioSharing site. C.T. coordinates the MIBBI portal. W.H. conceived SCDE and the role of an ISA approach to integration and within its stem cell systems, W.H., O.H., B.C., S.J.H.S. and K.B. contributed to the development of the ISA framework and worked on the SCDE. W.T. and H.F. contributed to the development of the ISA framework and strategies to integrate it with the FDA's ArrayTrack tool. S.N. contributed to the development of the ISA framework and developed workflows to integrate it with lab equipment. L.A.-Z. worked toward the implementation of ISA for the MIRADA-LTERS and ICoMM data sets. T.B. developed the NERC Environmental Bioinformatics Center (NEBC) EnvBase catalogue. G.B. worked toward the implementation of ISA for the BIRN BioScholar Knowledge Management system. T.C. leads the W3C working subgroup on Scientific Discourse; S.D. led the development of the Harvard Stem Cell Institute (HSCI) Blood Genomics repository, and M.E. worked on the integration of ISA-Tab into the system. L.-A.C. assisted the ISA developers to make use of the DataCite Metadata Store to mint Digital Object Identifiers (DOIs). J.C. and C.E.S. worked toward the implementation of ISA for use with HMS LINCS data. A.d.D. and D.J. worked toward the implementation of ISA for the MeRy-B knowledgebase. S.E. and S.L. worked on the integration of the ISA framework into the *GigaScience* and BGI database infrastructure. C.T.E. worked toward the implementation of ISA in the dbNP database and provided links to the Open PHACTS project. J.G. worked toward the implementation of ISA at the Argonne National Laboratory. C.G. and K.W. worked on the implementation of ISA-Tab in the SEEK platform. J.K. led the CarcinoGENOMICS project under which the ISA framework was first funded and developed. K.H., P.d.M. and C.S. developed the MetaboLights, powered by the ISA framework. A.L. led the implementation of the ISA-Tab in the SNRNASM annotation guidelines. S.M. and D.R. worked toward the integration of selected ISA software components as part of an extended workflow at NIBR. M.R. headed the development of the SIDR repository and the implementation of the ISA-Tab format. A.M. worked toward the implementation of ISA at CSIRO. C.A.S. worked toward the implementation of ISA at Bioplatforms Australia.
A.T., B.W.-J., H.H., I.D., I.X., J.L.G., L.B., L.H., M.J.F. and P.G., along with all the other authors, have provided advice, suggestions and feedback to S.-A.S. and P.R.-S. during the design and development phase of the ISA framework. In particular, P.G. was also closely involved in the BioSharing effort, and L.H. and B.W.-J. were pivotal for the links to the Pistoia Alliance, industry groups and the IMI Open PHACTS project. All the authors have contributed to the preparation of the manuscript at all stages; in particular, E.M. developed the figures and S.-A.S., P.R.-S., D.F. and C.T. led the writing process.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Editorial *Nature* **461**, 145 (2009).
2. Editorial *Nat. Genet.* **42**, 1 (2010).
3. Editorial *Science* **331**, 692 (2011).
4. Hamburg, M.A. *Science* **331**, 987 (2011).
5. Barnes, M.R. *et al. Nat. Rev. Drug Discov.* **8**, 701–708 (2009).
6. Field, D. *et al. Science* **326**, 234–236 (2009).
7. Birney, E. *et al. Nature* **461**, 168–170 (2009).
8. Schofield, P.N. *et al. Nature* **461**, 171–173 (2009).
9. Smith, B. *et al. Nat. Biotechnol.* **25**, 1251–1255 (2007).
10. Taylor, C.F. *et al. Nat. Biotechnol.* **26**, 889 (2008).
11. Barrett, T. *et al. Nucleic Acids Res.* **37**, D885–D890 (2009).
12. Parkinson, H. *et al. Nucleic Acids Res.* **37**, D868–D872 (2009).
13. Vizcaíno, J.A. *et al. Nucleic Acids Res.* **38**, D736–D742 (2010).
14. Shumway, M. *et al. Nucleic Acids Res.* **38**, D870–D871 (2010).
15. Editorial *Genome Biol.* **12**, 402 (2011).
16. Mervis, J. *Science* **332**, 291 (2011).
17. Harland, L. *et al. Drug Discov. Today* **16**, 940–947 (2011).
18. Nelson, B. *Nature* **461**, 160–163 (2009).
19. Howe, D. *et al. Nature* **455**, 47–50 (2008).
20. Rocca-Serra, P. *et al. Bioinformatics* **26**, 2354–2356 (2010).
21. Rocca-Serra, P. *et al. Bioinformatics* **26**, 2354–2356 (2010).
22. Ho Sui, S.J. *et al. Nucleic Acids Res.* published online, doi:10.1093/nar/gkr1051 (24 November 2011).
23. Demir, E. *et al. Nat. Biotechnol.* **28**, 935–942 (2010).
24. Harland, L. & Forster, M. *Open Source Software in Life Science Research: Practical Solutions in the Pharmaceutical Industry and Beyond* (Biohealthcare Publishing, Oxford, 2012).
25. Chen, B. *et al. BMC Bioinformatics* **11**, 255 (2010).
26. Rocca-Serra, P. *et al. RNA* **17**, 1204–1212 (2011).
27. Editorial *Nat. Genet.* **43**, 501 (2011).
28. Sorani, M.D. *et al. Drug Discov. Today* **15**, 741–748 (2010).
29. Editorial *Nat. Biotechnol.* **27**, 579 (2009).
30. Thorisson, G.A. *Nat. Biotechnol.* **27**, 984–985 (2009).

[1]Oxford e-Research Centre, University of Oxford, Oxford, UK. [2]Natural Environment Research Council, Environmental Bioinformatics Centre, Wallingford Centre for Ecology and Hydrology (CEH), Oxford, UK. [3]European Molecular Biology Laboratory (EMBL) Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK. [4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [5]ICF International, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [6]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany. [7]Center for Bioinformatics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [8]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, International Census of Marine Microbes, Marine Biological Laboratory, Woods Hole, Massachusetts, USA. [9]Ontario Institute for Cancer Research, Informatics and Bio-computing, Toronto, Ontario, Canada. [10]Swiss Institute of Bioinformatics, Swiss-Prot Group, Geneva, Switzerland. [11]Information Sciences Institute, University of Southern California, Marina del Rey, California, USA. [12]Department of Neurology, Harvard Medical School, Boston, Massachusetts, USA. [13]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA. [14]The British Library, London, UK. [15]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [16]Laboratoire Bordelais de Recherche en Informatique (LaBRI), Université de Bordeaux, Centre National de la Recherche Scientifique (CNRS) Unité Mixte de Recherche (UMR) 5800, Talence Cedex, France. [17]Université de Bordeaux, Centre de Bioinformatique de Bordeaux (CBiB), Génomique Fonctionnelle Bordeaux, Bordeaux, France. [18]Knowledge Engineering & Information Science, Discovery Information, AstraZeneca plc, Macclesfield, UK. [19]GigaScience, BGI Shenzhen, Yantian, China. [20]Department of Bioinformatics BiGCaT, Maastricht University, Maastricht, The Netherlands. [21]NBIC Faculty, The Netherlands Bioinformatics Centre, Nijmegen, The Netherlands. [22]Syngenta RDIS, Jealott's Hill, Bracknell, UK. [23]Center for Genetic Medicine, Northwestern University, Chicago, Illinois, USA. [24]Swiss Institute of Bioinformatics, Computational Analysis and Laboratory Investigation of Proteins of Human Origin (CALIPHO), CMU 1, Geneva, Switzerland. [25]Argonne National Laboratory, Argonne, Illinois, USA. [26]School of Computer Science, University of Manchester, Manchester, UK. [27]Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK. [28]Elsie Widdowson Laboratory, Medical Research Council (MRC) Human Nutrition Research, Cambridge, UK. [29]Fruit Biology and Pathology Centre, Bordeaux, INRA, UMR 1332, Villenave d'Ornon, France. [30]Department of Toxicogenomics, Netherlands Toxicogenomics Centre, p/a Maastricht University, Maastricht, The Netherlands. [31]ConnectedDiscovery, London, UK. [32]Department of Biology, University of North Carolina, Chapel Hill, North Carolina, USA. [33]Developmental and Molecular Pathways, Quantitative Biology Unit, The Novartis Institutes for BioMedical Research, Cambridge, Massachusetts, USA. [34]CSIRO Mathematics, Informatics and Statistics, Canberra, Australia. [35]CNRS UPS76, Institute for Scientific and Technological Information, Vandoeuvre-lès-Nancy, France. [36]University of Pierre and Marie Curie CNRS UMR 7606, Paris, France. [37]Bioplatforms Australia, Macquarie University, Sydney, Australia. [38]Swiss Institute of Bioinformatics, Vital-IT, Lausanne, Switzerland. [39]These authors contributed equally to this work. Correspondence should be addressed to S.-A.S.
(e-mail: susanna-assunta.sansone@oerc.ox.ac.uk).