

# The First RSBI (ISA-TAB) Workshop: “Can a Simple Format Work for Complex Studies?”

Susanna-Assunta Sansone,<sup>1</sup> Philippe Rocca-Serra,<sup>1</sup> Marco Brandizi,<sup>1</sup> Alvis Brazma,<sup>1</sup> Dawn Field,<sup>2</sup> Jennifer Fostel,<sup>3</sup> Andrew G. Garrow,<sup>4</sup> Jack Gilbert,<sup>5</sup> Federico Goodsaid,<sup>6</sup> Nigel Hardy,<sup>7</sup> Phil Jones,<sup>1</sup> Allyson Lister,<sup>8</sup> Michael Miller,<sup>9</sup> Norman Morrison,<sup>2,10</sup> Tim Rayner,<sup>1</sup> Nataliya Sklyar,<sup>1</sup> Chris Taylor,<sup>1</sup> Weida Tong,<sup>11</sup> Guy Warner,<sup>4</sup> Stefan Wiemann,<sup>12</sup> and Members of the RSBI Working Group\*

## Abstract

This article summarizes the motivation for, and the proceedings of, the first ISA-TAB workshop held December 6–8, 2007, at the EBI, Cambridge, UK. This exploratory workshop, organized by members of the Microarray Gene Expression Data (MGED) Society’s Reporting Structure for Biological Investigations (RSBI) working group, brought together a group of developers of a range of collaborative systems to discuss the use of a common format to address the pressing need of reporting and communicating data and metadata from biological, biomedical, and environmental studies employing combinations of genomics, transcriptomics, proteomics, and metabolomics technologies along with more conventional methodologies. The expertise of the participants comprised database development, data management, and hands-on experience in the development of data communication standards. The workshop’s outcomes are set to help formalize the proposed Investigation, Study, Assay (ISA)-TAB tab-delimited format for representing and communicating experimental metadata. This article is part of the special issue of OMICS on the activities of the Genomics Standards Consortium (GSC).

## The OMICS Standards Scenario

THE MARRIAGE OF CONVENTIONAL METHODS with (meta)genomics, transcriptomics, proteomics, and metabolomics technologies (hereafter referred as “omics”) has created not only opportunities, but also substantial new informatics challenges. For example, consider the reporting of a complex multiomic study looking at the effect on a number of subjects of a compound inducing liver damage by characterizing the metabolic profile of their urine (by mass spec-

troscopy), measuring protein and gene expression in the liver (by mass spectrometry and DNA microarrays, respectively), and conducting conventional histopathological analysis. To coordinate the reporting of such a heterogeneous study requires new approaches for communicating the complex metadata (i.e., sample characteristics, study design, and execution) required to correctly interpret the final results.

Many groups are rising to this challenge, and standards for describing, formatting, submitting, and exchanging both data and metadata from such complex studies are being de-

\*See the RSBI and ISA-TAB Websites for a complete list of contributors. Web URL: <http://www.mged.org/Workgroups/rsbi> and <http://isatab.sf.net>

<sup>1</sup>EMBL-EBI The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

<sup>2</sup>NERC Centre for Ecology and Hydrology, Mansfield Rd, Oxford, United Kingdom.

<sup>3</sup>NIH, National Institute of Environmental Health Science (NIEHS), Research Triangle Park, North Carolina.

<sup>4</sup>Unilever, Safety and Environmental Assurance Centre, Colworth Park, Sharnbrook, Bedford, United Kingdom.

<sup>5</sup>Plymouth Marine Laboratory, Prospect Place, Plymouth, United Kingdom.

<sup>6</sup>Office of Clinical Pharmacology, Office of Translational Science, Center for Drug Evaluation and Research (CDER) U.S. Food and Drug Administration (FDA), Silver Spring, Maryland.

<sup>7</sup>Department of Computer Science, Aberystwyth University, Ceredigion, Wales, United Kingdom.

<sup>8</sup>CISBAN & School of Computing Science, Newcastle University, Newcastle upon Tyne, United Kingdom.

<sup>9</sup>Rosetta Biosoftware, Seattle, Washington.

<sup>10</sup>NERC Bioinformatics Center (NEBC), Centre for Ecology and Hydrology, Oxford, United Kingdom, and the University of Manchester, School of Computer Science, Manchester, United Kingdom.

<sup>11</sup>FDA’s National Center for Toxicological Research (NCTR), Center for Toxicoinformatics, Jefferson, Arkansas.

<sup>12</sup>German Cancer Research Center, Heidelberg, Germany.

veloped. Currently, several standards initiatives occupy strategic positions in the international scenario, largely falling into two groups identifiable by the needs of their respective user communities. One group of initiatives is driven by regulatory frameworks, and most significantly, focuses on the Voluntary eXploratory Data Submissions (VXDS) and electronic data submission programs of the Food and Drug Administration (FDA) (Frueh, 2006; Tong et al., 2007; U.S. HHS/FDA Guidance for Industry: Pharmacogenomic data submissions, 2005). These initiatives include long-standing efforts in the clinical and nonclinical domains (<http://www.cdisc.org/standards/index.html>) alongside more recent activities in the pharmacogenomics area, bringing the added complexity of omics technologies to biomedical studies (Shabo, 2006). A second group of initiatives addressing particular technologies or defined domains of application have emerged from the academic community, and in many cases benefit from the support of commercial organizations. These initiatives are focused on supporting tool interoperability and data exchange among public and proprietary systems, by developing common minimal requirements, formats, and terminologies (Ball and Brazma, 2006; Deutsch et al., 2008; Field and Sansone, 2006; Field et al., this issue; Le Novère et al., 2005; Orchard and Hermjakob, 2007; Sansone et al., 2007; Wiemann et al., community consultation).

Although we currently lack a global initiative ready to bring these counterparts under one umbrella, among the academic community several synergistic activities have begun that aim to foster the harmonization and consolidation of the three *kinds* of standards being developed (checklists, syntax, and semantics). More than 20 groups are now participating in the Minimum Information for Biomedical or Biological Investigations (MIBBI) project; set to be a one-stop shop for those exploring the range of extant checklists and to foster collaborative, integrative development (Taylor et al., 2008). Several groups participate in the Functional Genomics (FuGE) project to develop a single generic data model that will underpin a variety of XML-based formats by providing a single common framework (Jones et al., 2007). Over 60 groups participate in the Open Biological Ontology Foundry (Smith et al., 2007, <http://www.obofoundry.org>), with the objective of developing interoperable ontologies; and approx 20 communities are contributing to the creation of the Ontology for Biomedical Investigation (OBI), which will support the description of experimental metadata in a stan-

dardized manner across a variety of biological and medical domains. Managing this process of consensus-building from start to finish takes time, resources, and expertise; the time available to invest in finding commonalities and building synergies among projects is limited due to lack of resources. Lacking formal funding, developers participate on a voluntary basis, because the lack of standardization is an unacceptable state of affairs for omics researchers, and is repeatedly proving to be a significant bottleneck in the collection, sharing, and integration of data. The massively collaborative nature of these undertakings mandates frequent face-to-face workshops to create the necessary conditions for the building of consensus.

### Rationale and Overview of the Workshop

This workshop is part of a series on omics data standards funded by a Biotechnology and Biological Sciences Research Council (BBSRC) award to the European Bioinformatics Institute (EBI). These workshops are designed to (1) advance the coordinated development of MIBBI, FuGE and OBI, (2) identify stable subsets of those projects outputs that can be implemented and tested, and (3) discuss interim solutions to tackle today's need for describing, formatting, submitting, and exchanging both data and metadata while these synergistic projects remain works in progress. The integrative elements binding these themed workshops together are largely informed by the opinions and requirements of the Reporting Structure for Biological Investigations (RSBI, <http://www.mged.org/Workgroups/rsbi>) working group (Sansone et al., 2006). RSBI was established in 2004 under the Microarray and Gene Expression Data (MGED) Society umbrella (Ball and Brazma, 2006). RSBI has been conceived of as a "single point of focus" for groups already independently developing standards-supported databases and tools for biological, biomedical, and environmental studies employing omics technologies and more conventional methodologies.

The goals of this workshop were twofold: first, to create an "exchange network test bed," of RSBI groups and other collaborators that have expressed an interest in leveraging on MIBBI, FuGE, and OBI; second, to evaluate a straw man proposal addressing the pressing need for a format with which to communicate study data and metadata while a complete set of interoperable FuGE-based modules remains a work in progress. The participants in this first workshop,

TABLE 1. THE NODES OF THE "EXCHANGE NETWORK TEST BED" THAT WERE REPRESENTED AT THE FIRST ISA-TAB WORKSHOP

<i>System</i>	<i>URL</i>
ArrayExpress	<a href="http://www.ebi.ac.uk/arrayexpress">http://www.ebi.ac.uk/arrayexpress</a>
ArrayTrack	<a href="http://www.fda.gov/nctr/science/centers/toxicoinformatics">http://www.fda.gov/nctr/science/centers/toxicoinformatics</a>
BioInvIndex	<a href="http://www.ebi.ac.uk/net-project">http://www.ebi.ac.uk/net-project</a>
CEBS	<a href="http://cebs.niehs.nih.gov">http://cebs.niehs.nih.gov</a>
Coral	<a href="http://www.unilever.co.uk">http://www.unilever.co.uk</a>
GSC Genome Catalogue	<a href="http://gensc.org/">http://gensc.org/</a>
Omixed	<a href="http://www.omixed.org">http://www.omixed.org</a>
PRIDE	<a href="http://www.ebi.ac.uk/pride">http://www.ebi.ac.uk/pride</a>
Rosetta Resolver <sup>®</sup>	<a href="http://www.rosettatabio.com/products/resolver">http://www.rosettatabio.com/products/resolver</a>
Simbioms	<a href="http://www.simbioms.org">http://www.simbioms.org</a>
SyMBA	<a href="http://symba.sf.net">http://symba.sf.net</a>

all of whom are authors of this article, included representatives of public and proprietary repositories, and public and commercial software developers (listed in Table 1), working variously in academic, industrial and governmental groups contributing to one or more standardization initiatives or to the development of submission policies requiring the use of reporting standards for omics-based data. In addition to MIBBI, OBI, FuGE, and MGED, the Proteomics Standards Initiative (PSI; Orchard and Hermjakob, 2007), the Metabolomics Standards Initiative (MSI; Sansone et al., 2007), and the Minimum Information About a Cellular Assay group (MIACA; Wiemann et al., community consultation) (Table 2) were represented at the meeting.

The workshop began with a welcome from the organizer, **Susanna-Assunta Sansone (EBI)**, who set the context for the event; shortly followed by presentations of the MIBBI, FuGE, and OBI projects, providing information on current status and planned activities. For the second session, the meeting participants provided brief overviews of their systems; to be of most use, all had been asked to highlight where, in their view, common standards could (potentially) bring benefits. **Weida Tong (NCTR-FDA)**, who cochaired the workshop, gave an overview of the Study Data Tabulation Model (STDM, <http://www.cdisc.org/standards/index.html>) and that format's requirements and challenges in the context of VXDS submissions to the FDA. A few themes emerged from the presentations; the most important being the need for a simple format to submit or exchange studies employing omics technologies along with more conventional methodologies, while a complete set of interoperable XML modules, such as FuGE-based community formats, are still under development. The Investigation/Study/Assay (ISA) tab-delimited (TAB) format was then presented by **Philippe Rocca-Serra (EBI)** as a straw man proposal in the third session, which was followed by open discussion.

Overall, this first exploratory workshop produced general consensus around the ISA-TAB proposal, and in addition, a clear work plan to refine and test it further. The next sections provide a brief overview of the ISA-TAB proposal and present the heterogeneous group of nodes participating in the "exchange network test bed" (Table 1).

### ISA-TAB in a Nutshell

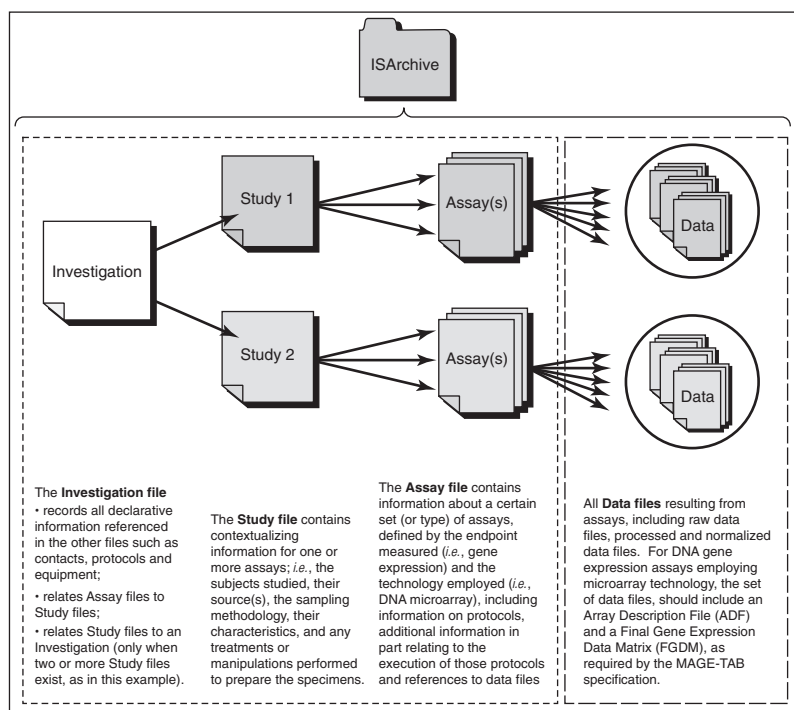
Investigation, Study, and Assay are the three key entities (Sansone et al., 2006) around which the ISA-TAB framework is built; these assist in structuring metadata and describing the relationship of samples to data (Fig. 1). The ISA-TAB pro-

posal builds on the successful uptake of the MicroArray Gene Expression (MAGE) TAB format, which supports the management, exchange and submission of microarray-based experiment data and metadata (Rayner et al., 2006). MAGE-TAB was designed for use by laboratories with little or no bioinformatics support, rendering them unable to deal with the complexity of MAGE Markup Language (ML) formatted files (Spellman et al., 2002), which often are also exceptionally large—too large to be easily read by most people, and often too large to be read by most software programs (Maier et al., 2008). ISA-TAB can be viewed as an extended version of the MAGE-TAB paradigm, sharing its motivation for the use of tab-delimited text files; that is, that they can be created programmatically or by using spreadsheet software such as Microsoft Excel, where they can easily be viewed and edited by researchers. Like MAGE-TAB before it, ISA-TAB is simply a format with which to communicate information. Neither minimum requirements nor the use of controlled terminologies are within the scope of this proposal. Therefore, the decision on how to regulate its use (e.g., by enforcing MIBBI minimum requirements, or mandating the use of OBO Foundry terminologies) is solely a matter for those who will implement the format in their system. The ISA-TAB actually employs MAGE-TAB syntax to ensure backward compatibility with existing MAGE-TAB files, to facilitate the future adoption of one common format. However, ISA-TAB has a number of additional features making it a more general framework that can capture the complexity of studies employing a combination of technologies. For example, where omics-based technologies are being used in clinical or non-clinical studies, ISA-TAB can complement existing biomedical formats such as the SDTM by formally capturing information about the interrelationship of the various parts.

It is important to maintain an alignment between the concepts in ISA-TAB and (some of) the objects in the FuGE model, partly as that model is integral to the development of MAGE-ML v2. The ISA-TAB format could be seen as competing with XML-based formats, whether existing or under development, such as the FuGE-ML. However, ISA-TAB addresses an immediate need, whereas a complete set of FuGE-based modules or other interoperable XML is still some way off. Once such formats do become available, ISA-TAB can continue to serve those with little or no bioinformatics support, as well as finding utility as a user-friendly presentation layer for XML-based formats (via an XSL transformation); that is, in the manner of the HTML rendering of MAGE-ML documents (<http://www.ebi.ac.uk/~rocca/MAGE-XSLT/HTML%20rendering%20of%20MAGE.htm>).

TABLE 2. THE STANDARDIZATION INITIATIVES THAT WERE REPRESENTED AT THE FIRST ISA-TAB WORKSHOP

<i>Initiative</i>	<i>URL</i>	<i>Domain</i>
GSC	<a href="http://gensc.org">http://gensc.org</a>	Genomics
MGED	<a href="http://www.mged.org">http://www.mged.org</a>	Transcriptomics
MIACA	<a href="http://miaca.sourceforge.net">http://miaca.sourceforge.net</a>	Cellular assay
MSI	<a href="http://msi-workgroups.sourceforge.net">http://msi-workgroups.sourceforge.net</a>	Metabolomics
PSI	<a href="http://www.psidev.info">http://www.psidev.info</a>	Proteomics
FuGE	<a href="http://fuge.sf.net">http://fuge.sf.net</a>	Generic data model
MIBBI	<a href="http://mibbi.sf.net">http://mibbi.sf.net</a>	Modular checklists
OBI	<a href="http://obi.sf.net">http://obi.sf.net</a>	Common terminology



**FIG. 1.** An overview of the ISA-TAB structure is shown here; for submission or transfer, files can be packaged into an ISArchive. Detailed description of each file is available in the ISA-TAB v0.2 specification (<http://isatab.sourceforge.net>).

For the development of this format, openness and due process must apply. To ensure that all groups have access to information, a public Web site has been created at SourceForge (<http://isatab.sourceforge.net>). It contains the latest ISA-TAB v0.2 specification, an alignment with MAGE-TAB, the list of participants and their systems and a link to a mailing list to which interested parties can subscribe. Several example ISA-TAB “instance” files are being created from published biological, biomedical, and environmental studies, to be posted on the Web site. For example, the first was created by members of the Genomics Standards Consortium (GSC) (Field et al., 2008, this issue) using a dataset published by Gilbert et al. (2008) originating from a joint metagenomics and meta-transcriptomics study looking at the effect of ocean acidification on phytoplankton and bacterioplankton.

### The Exchange Network Test Bed

The initial motivation for creating the ISA-TAB straw man proposal was to meet the needs of the **BioInvestigation Index (BioInvIndex)** system at EBI (Table 1). BioInvIndex aims to create a common structured representation of the metadata and the sample-data relationship for biological, biomedical, and environmental studies employing omics-based technologies along with more conventional methodologies. The ISA-TAB format is being developed to assist users make combined submissions to EBI public archives; for example, to ArrayExpress, PRIDE and a metabolomics repository to be developed in the near future. It has been clear from the outset that the ISA-TAB framework could also serve as a common crossplatform format, thereby greatly benefiting other collaborators; to pipeline omics-based experimental data into EBI public repositories, to enable their users to import data from EBI repositories into their local systems, or

simply to exchange data among themselves. Therefore, in a collaborative spirit the development of the ISA-TAB has been opened up and shared with a wider community.

The section below provides a brief overview of the collaborative systems whose developers constitute the nodes of this “exchange network test bed” for ISA-TAB format.

**ArrayExpress** is the EBI public resource for microarray and transcriptomics data (Parkinson et al., 2007); it supports the Minimum Information About a Microarray Experiment standard (MIAME; Brazma et al., 2001). The system consists of two parts: the ArrayExpress repository, which is one of the databases recommended by the MGED Society for archiving publication-related microarray data, and the warehouse of gene expression profiles, which uses the data in the repository. ArrayExpress supports all MGED standards and recommendations; it accepts and distributes data in the MAGE-TAB format. ArrayExpress will ultimately use the ISA-TAB format, it being a generalization of MAGE-TAB.

**ArrayTrack** is a publicly available FDA genomic tool that has been used for the FDA review of genomic data submissions (Tong et al., 2004, 2007). It provides an integrated solution for managing, analyzing, and interpreting microarray gene expression data. ArrayTrack is MIAME supportive for storing both microarray data and experiment parameters associated with a pharmacogenomics or toxicogenomics study. However, recently the FDA has encountered multiomic data sets, submitted by industry as part of the VXDS program. The ISA-TAB crossplatform standard can be used to manage multiomic data in ArrayTrack, and additionally, provides a means to communicate multiomic data between ArrayTrack and other software platforms.

**CEBS** is the Chemical Effects in Biological Systems, an integrated public repository for toxicologic and toxicogenomics data, including the study design and timeline, clini-

cal chemistry and histopathology findings, and microarray and proteomics data. CEBS contains modules for the study component of each investigation, and stores data in data-type specific modules. The system permits users to explore data by integrating across studies and across data modules, and then either to download or to analyze the data of interest within CEBS. To facilitate the publication, exchange and reuse of data the system is engaged in developing tools for depositors and annotation formats for study data and meta-data. The development of ISA-TAB is an important part of this effort.

**Coral** is an investigation-centric data management system used within Unilever's Safety and Environment Assurance Centre (SEAC). The system has to be able to capture descriptions of complex scientific investigations, allowing the linkage of a single sample to multiple analyses employing various assays. Key to Coral's success was that it can be readily configured for new assays (transcriptomic, proteomic, etc.) and sample types, making it robust to both technological developments and new experimental designs. When configuring Coral for new data types, it is imperative that it conforms to community standards, so that data can be exchanged with external systems. Therefore, established standards such as MIAME and the Minimum Information About a Proteomics Experiment (MIAPE; Taylor et al., 2007) are used (Brazma et al., 2001; Taylor et al., 2007), in addition to emerging standards in the metabolomics domain (Sansone et al., 2007). ISA-TAB will allow the system to better define the essential meta-information that it should capture from investigation and assay types not previously encountered, in order to ensure interoperability with other ISA-TAB compliant systems.

The GSC's **Genome Catalog** is a pilot online repository for holding case studies that have helped the development of its Minimum Information about a Genome Sequence (MIGS) specification (Field et al., 2008). This database is now also serving as a repository for collected information. Input of data is currently managed through a set of web forms, but this only serves the needs of genome and metagenome authors with one or two projects to register. For users with more, submission of data should be tackled programmatically; to that end an XML implementation of MIGS is being developed (Kottmann, 2008). For those users who do not want to use XML, or cannot easily do so, a spreadsheet based submission format would be ideal. Therefore, the GSC is contributing to the development of ISA-TAB as it greatly prefers to work with the international community in a multiomic context rather than build a GSC-specific solution.

**Omixed** is a software system designed to manage scientific research data using an intuitive, Web-based interface. Omixed is developed by the Natural Environmental Research Council's Environmental Bioinformatics Center (NEBC; <http://nebc.nox.ac.uk>) to serve the needs of a large user base of interconnected labs throughout the UK who are engaged in environmental research. This system allows the collection and management of multiomic data in a robust, secure, and collaborative manner. NERC strongly supports the use of international standards and promotes long-term data storage; therefore, it is a firm requirement that Omixed facilitate convenient, standardized data capture, support integrative data analyses, and enable the publication of data

sets to public repositories. The ISA-TAB format will allow Omixed users to submit and import studies to and from other compliant systems.

**PRIDE** is the Proteomics IDentifications Database, a repository of identifications of proteins, peptides, and protein modifications by mass spectrometry (Jones et al., 2008). The vast majority of data in PRIDE originate from direct submissions; however, the complexity of the XML format describing a complete PRIDE experiment places a significant burden on laboratories with limited bioinformatics support. The PRIDE Proteome Harvest Data Submission Spreadsheet (<http://www.ebi.ac.uk/pride/proteomeharvest/index.html>), which allows a PRIDE submission to be constructed in a tabular format, helps to resolve this problem, and has proven popular with PRIDE data submitters. The creation of a general solution such as ISA-TAB for the submission of proteomics data (under the wider umbrella of omics in general) will therefore benefit both the PRIDE project and the community that it supports.

**Rosetta Resolver**<sup>®</sup> Gene Expression Analysis System is a comprehensive, flexible gene expression data analysis, management, and storage environment developed by Rosetta Biosoftware. The system allows for automated data import from multiple array technologies, including those manufactured by Affymetrix<sup>®</sup>, Agilent<sup>®</sup>, and Illumina<sup>®</sup>. The Resolver system supports multiple data standards such as CDISC/SEND and MAGE-ML, and is MIAME-compliant. MAGE-ML allows the Resolver system to import and export data from multiple Resolver clients, as well as from external software applications. The Resolver system also supports FDA Code of Federal Regulations (21 CFR Part 11, 2000) compliance as validated by industry audits, and data exchange between internal or external collaborators (including with the FDA in a variety of formats). ISA-TAB has the potential to be a crossplatform standard that can be easily parsed and written by the Resolver system.

**Simbioms** is the System for Information Management in Biomedical Studies, a lightweight open source software package for managing biomedical research data. The system, originally developed for a large international genetic epidemiology project, is Web based and can be run either on a central server for collaborative projects, or on a researcher's PC. It consists of two components: The Patient and Sample System for Information Management (PASSIM) (Vikсна et al., 2007), and Assay and Data Management System (AIMS). It is customized for a wide range of high throughput technologies and applications including genotyping. The system can import and export data in tab-delimited format and will be made fully compliant with ISA-TAB once the standard is finalized.

**SyMBA** is the Systems and Molecular Biology Data and Metadata Archive, comprising a versioned database, a Java toolkit, and a Web front-end, all based on the FuGE-OM. SyMBA archives, stores, and retrieves raw high-throughput data and metadata in a single database. Originally developed to meet the needs of an interdisciplinary systems biology center, it is now an open-source, multideveloper project available from SourceForge. Making use of community data standards such as FuGE, SyMBA reduces development time while simultaneously increasing compatibility with other data providers. A format such as ISA-TAB would enable

SyMBA users to submit their FuGE-formatted data and metadata to the EBI public databases until direct FuGE submission becomes available.

### Conclusions and Next Steps

This first exploratory workshop produced a general consensus on the ISA-TAB proposal and on the importance of having (1) a simple format that can easily be created, viewed, and edited by researchers with little or no bioinformatics support, and (2) a common user-friendly presentation layer for (interoperable) XML-based formats, once a complete set of such formats becomes available. The attendees agreed that follow-up meetings to coordinate developmental activity would be needed. Funding for future workshops has been secured by the BBSRC award to EBI with contributions from NERC's NEBC and the GSC. The next workshop is planned for June 16–18, 2008, at the EBI; on that occasion ISA-TAB v0.2 will be examined in the light of a series of real case examples being produced by the participating communities, and will be revised as necessary. The agenda will also include a discussion on, and examples of the kinds of implementations and tools (that need to be) developed to assist users create ISA-TAB files. Documentation will be finalized and a manuscript drafted; the intent being to publish it as an open access paper. Anyone interested in knowing more about or joining this effort is encouraged to subscribe to the mailing list.

### Acknowledgments

We thank the BBSRC for funding the workshop through an award to Susanna-Assunta Sansone (WODS BB/E025080/1). This author also gratefully acknowledges the European integrated project CarcinoGENOMICS (<http://www.carcinogenomics.eu>, PL 037712) for supporting the development of the BioInvIndex system and thanks Dawn Field and George Garrity for editing this special issue.

### Author Disclosure Statement

The authors declare that no competing financial interests exist.

### References

- Ball, C.A., and Brazma, A. (2006). MGED standards: work in progress. *OMICS* **10**, 138–144.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* **29**, 365–371.
- Deutsch, E.W., Ball, C.A., Berman, J.J., et al. (2006). Minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). *Nat Biotechnol* **26**, 305–312.
- Field, D., Garrity, G.M., Gray, T., et al. (2008a). The “Minimum Information about a Genome Sequence” (MIGS) specification. *Nat Biotechnol* (in press).
- Field, D., Glockner, F.O., Garrity, et al. (2008b). Meeting report: The 4th Genomic Standards Consortium (GSC) workshop *OMICS* (this issue).
- Field, D., and Sansone, S.A. (2006). A special issue on Omic data standards. *OMICS* **10**, 84–93.
- Frueh, F.W. (2006). Impact of microarray data quality on genomic data submissions to the FDA. *Nat Biotechnol* **24**, 1105–1107.
- Gilbert, J.A., Edwards, R., Li, W., et al. (2008). Sequencing complex marine microbial metatranscriptomes with pyrosequencing technology. *Nat Methods* (in review).
- Jones, A.R., Miller, M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., et al. (2007). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotechnol* **25**, 1127–1133.
- Jones, P., Côté, R.G., Cho, S.Y., Martens, L., Quinn, A.F., Thorneycroft, D., et al. (2008). PRIDE: new developments and new datasets. *Nucleic Acids Res* **36**, D878–D883.
- Kottmann, R., Gray, T., Murphy, et al. (2008). A standard MIGS/MIMS compliant XML schema: towards the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* (this issue).
- Le Novère, N., Finney, A., Hucka, M., Bhalla, U.S., Campagna, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**, 1509–1515.
- Maier, D., Wymore, F., Sherlock, G., and Ball, C.A. (2008). The XBabelPhish MAGE-ML and XML translator. *BMC Bioinformatics* **18**, 28.
- Orchard, S., and Hermjakob, H. (2007). The HUPO proteomics standards initiative easing communication and minimizing data loss in a changing world. *Brief Bioinform* (in press).
- Parkinson, H., Kapushesky, M., Shojatalab, M., Abeyganawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**, D747–D750.
- Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, F., et al. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB *BMC Bioinformatics* **6**, 489.
- Sansone, S.A., Fan, T., Goodacre, R., Griffin, J.L., Hardy, N.W., Kaddurah-Daouk, R., et al. (2007). The metabolomics standards initiative. *Nat Biotechnol* **25**, 846–848.
- Sansone, S.A., Rocca-Serra, P., Tong, W., Fostel, J., Morrison, N., Jones, A.R., et al. (2006). A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. *OMICS* **10**, 164–171.
- Shabo, A. (2006). Clinical genomics data standards for pharmacogenetics and pharmacogenomics. *Pharmacogenomics* **7**, 247–253.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bag, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* **25**, 1251–1255.
- Spellman, P.T., Miller M., Stewart J., et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046.
- Taylor, C.F., Field, D., Sansone, S.A., et al. (2008). Promoting coherent minimum reporting requirements for biological and biomedical investigations: the MIBBI Project. *Nat Biotechnol* (in press).
- Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K., Jr., Jones, A.R., et al. (2007). The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **25**, 887–893.
- Tong, W., Harris, S.C., Fang, H., et al. (2007). An integrated bioinformatics infrastructure essential for advancing pharmacogenomics and personalized medicine in the context of the FDA's critical path initiative. *Drug Discov Today* **4**, 3–8.
- Tong, W., Harris, S., Cao, X., Fang, H., Shi, L., Fuscoe, J., et al. (2004). Development of public toxicogenomics software for microarray data management and analysis. *Mutat Res* **549**, 241–253.

- US FDA Code of Federal Regulations (21 CFR Part 11). Electronic Records; Electronic Signatures, 2000, [http://www.fda.gov/ora/compliance\\_ref/part11](http://www.fda.gov/ora/compliance_ref/part11).
- US HHS/FDA Guidance for Industry: Pharmacogenomic data submissions. (2005). <http://www.fda.gov/OHRMS/DOCKETS/98fr/2003d-0497-gdl0002.pdf>
- Viksna, J., Celms, E., Opmanis, M., Podnicks, K., Rucevskis, P., Zarins, A., et al. (2007). PASSIM—an open source software system for managing information in biomedical studies. *BMC Bioinformatics* **8**, 52.
- Waters, M., Stasiewicz, S., Merrick, B.A. et al. (2008). CEBS—chemical effects in biological systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res* **36**, D892–D900. Epub 2007, Oct. 25.
- Wiemann, S., Mehrle, A., Hahne, F., et al. MIACA—minimum information about a cellular assay, and the cellular assay object model. *Nat Biotechnol* community consultation: <http://www.nature.com/nbt/consult/index.html>

Address reprint requests to:

*Susanna-Assunta Sansone*

EMBL-EBI

*The European Bioinformatics Institute Wellcome Trust*

*Genome Campus*

*Hinxton, Cambridge, CB10 1 SD, United Kingdom.*

*E-mail: sansone@ebi.ac.uk*

or

*Philippe Rocca-Serra*

EMBL-EBI

*The European Bioinformatics Institute Wellcome Trust*

*Genome Campus*

*Hinxton, Cambridge, CB10 1 SD, United Kingdom.*

*E-mail: rocca@ebi.ac.uk*