

Embedding standards in metabolomics: the Metabolomics Society data standards task group

Reza M. Salek¹ · Masanori Arita^{2,3} · Saravanan Dayalan⁴ · Timothy Ebbels⁵ · Andrew R. Jones⁶ · Steffen Neumann⁷ · Philippe Rocca-Serra⁸ · Mark R. Viant⁹ · Juan-Antonio Vizcaíno¹

Published online: 3 June 2015

© Springer Science+Business Media New York 2015

Metabolomics has reached a maturity as a field. Yet, there remain some challenges to overcome, particularly in metabolite identification and reporting such results, as well as the need for continuous improvements in data standards and data sharing. For over a decade now, the metabolomics community, has built a dedicated international society with an affiliated journal, *Metabolomics*.¹ In 2007 several leaders within the community, established a set of standards and minimum reporting guidelines for experimental descriptors and data, known as the Metabolomics Standards Initiative (MSI), summarized by Goodacre et al.

(Goodacre 2013). Since 2012, resources and repositories have been established, notably EMBL-EBI MetaboLights (Haug et al. 2012) and the NIH funded Metabolomics Workbench,² where experimental data and metadata can be shared with the community, all of which is publicly accessible.

A central tenet of science is the reproducibility of results. However, this may be challenging to achieve in metabolomics, owing to the complex nature of the metabolome, the diversity of technologies and data analysis techniques used (Beisken et al. 2015). Despite these difficulties the principle of reproducibility must hold. Data sharing is not just simply making raw files available via a website link nor sharing the end results of data processing and analysis pipelines, usually in an excel spreadsheet. Key steps are required to achieve meaningful data sharing, ensuring that the results are reusable and the experimental results can be reproduced. Additionally, substantial curation efforts are often required to ensure optimal reporting, enriched metadata annotation within a study, but also to ensure consistency across studies, which may be achieved through checking compliance with annotation checklists such as the MSI guidelines (Salek et al. 2013). Ideally, data sharing should be shouldered by dedicated, institution backed repositories, thus guaranteeing continued support and long-term preservation.

Initial standardization efforts focused on study description and instrument generated metadata reporting. Experimental metadata associated to datasets can now be

✉ Reza M. Salek
reza.salek@ebi.ac.uk

¹ European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

² National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

³ RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

⁴ Metabolomics Australia, The University of Melbourne, Parkville, VIC 3010, Australia

⁵ Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, South Kensington, London SW7 2AZ, UK

⁶ Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK

⁷ Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany

⁸ Oxford e-Research Centre, University of Oxford, 7 Keble road, Oxford OX1 3QG, UK

⁹ School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

¹ “Metabolomics—Springer.” 2012. 21 May 2015 <http://link.springer.com/journal/11306>.

² “Metabolomics Workbench: NIH Common Fund...” 2013. 19 May 2015. <http://www.metabolomicsworkbench.org/nihmetabolomics/fundingopportunities.html>.

reported, relying on the ISA-Tab format to support a manuscript. The ISA-Tab format is a metadata standard that has gained a lot of momentum since first being released in 2008 (Rocca-Serra et al. 2010); it has now been adopted by Publishers (e.g. Nature Publishing Group, GigaScience) and vendors (e.g. Biocrates AG). Instrument vendors and software companies usually each create specific data formats that are dependent on a commercial package or tools to even be able to view the raw files. Fortunately, there are solutions, one being conversion of the proprietary data formats into open formats. The most popular and more developed are extensible markup language (XML) based and vendor independent data standards, such as mzML for mass spectrometry and nmrML for NMR raw data. The latter, nmrML, has recently been developed by the COSMOS consortium, The Metabolomics Innovation Centre in Canada, and other partners. The ‘COordination Of Standards In MetabOlogicS’ (COSMOS)³ is an European Framework 7 funded initiative that aims to develop a robust data infrastructure for metabolomics data and metadata representation. Another format that can potentially help with reporting metabolites identified is the “tab” separated mzTab file format (Griss et al. 2014). Originally developed by the Human Proteome Organization (HUPO)—Proteomics Standards Initiative (PSI) community for reporting proteomics experiments, it also includes support for small molecule (or metabolite) identification reporting. With the COSMOS initiative and other task groups within the Metabolomics Society, we can bring together leading vendors, researchers and bioinformaticians, members of the MSI, and international communities, such as HUPO-PSI⁴ to develop, support and adopt such open source data/metadata exchange formats and workflows. This is a continuous and prolonged effort, as technologies are constantly changing, new ones are introduced, reporting requirements change or are enhanced as the community evolves. Such endeavours require continuous

and renewed support, collaboration and information dissemination, working closely and engaging with developers, vendors and researchers. This networking would be one of the main aims of this task group, to act as a bridge or an official body to bring together and coordinate such effort. Further, the Data Standards Task Group will be a forum to foster interactions and collaboration with data producers, journal editors, reviewers and referees to increase and facilitate data review and the evaluation process to thus deliver a better ecosystem ensuring data discovery, data availability, data reuse and data citation. This last aspect is key to ensuring a proper accreditation of scientific output, giving equal weight to datasets, as is currently given to a manuscript.

References

- Beisken, S., Eiden, M., & Salek, R. (2015). Getting the right answers: understanding metabolomics challenges. *Expert Review of Molecular Diagnostics*, 15(1), 97–109.
- Goodacre, R. (2013). Water, water, every where, but rarely any drop to drink. *Metabolomics*, 10(1), 5–7.
- Griss, J., Jones, A., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., et al. (2014). The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular and Cellular Proteomics*, 13(10), 2765–2775.
- Haug, K., Salek, R., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2012). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1), D781–D786.
- Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., et al. (2010). ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18), 2354–2356.
- Salek, R., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendrakar, T., et al. (2013). The MetaboLights repository: curation challenges in metabolomics. *Database*. doi:10.1093/database/bat029.

³ “COSMOS—COordination of Standards in MetabOlogicS... 2012. 22 May. 2015 <http://www.cosmos-fp7.eu/>.

⁴ “HUPO Proteomics Standards Initiative: HUPO-PSI Working...“ 2007. 21 May. 2015 <http://www.psidev.info/>.