Metabolomics (2015) 11:1587–1597
DOI 10.1007/s11306-015-0810-y

ORIGINAL ARTICLE

COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access

Reza M. Salek[1,15] · Steffen Neumann[2] · Daniel Schober[2] · Jan Hummel[3] · Kenny Billiau[3] · Joachim Kopka[3] · Elon Correa[10] · Theo Reijmers[12] · Antonio Rosato[4] · Leonardo Tenori[4,19] · Paola Turano[4] · Silvia Marin[5] · Catherine Deborde[6] · Daniel Jacob[6] · Dominique Rolin[6] · Benjamin Dartigues[7] · Pablo Conesa[1] · Kenneth Haug[1] · Philippe Rocca-Serra[8] · Steve O'Hagan[10] · Jie Hao[16] · Michael van Vliet[12] · Marko Sysi-Aho[13] · Christian Ludwig[17] · Jildau Bouwman[11] · Marta Cascante[5] · Timothy Ebbels[16] · Julian L. Griffin[14,15] · Annick Moing[6] · Macha Nikolski[9] · Matej Oresic[13] · Susanna-Assunta Sansone[8] · Mark R. Viant[18] · Royston Goodacre[10] · Ulrich L. Günther[17] · Thomas Hankemeier[12] · Claudio Luchinat[4] · Dirk Walther[3] · Christoph Steinbeck[1]

Received: 26 September 2014 / Accepted: 14 May 2015 / Published online: 26 May 2015

Abstract Metabolomics has become a crucial phenotyping technique in a range of research fields including medicine, the life sciences, biotechnology and the environmental sciences. This necessitates the transfer of experimental information between research groups, as well as potentially to publishers and funders. After the initial efforts of the metabolomics standards initiative, minimum reporting standards were proposed which included the concepts for metabolomics databases. Built by the community, standards and infrastructure for metabolomics are still needed to allow

✉ Christoph Steinbeck
steinbeck@ebi.ac.uk

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

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

[3] Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany

[4] Magnetic Resonance Center (CERM), University of Florence, 50019 Sesto Fiorentino, FI, Italy

[5] Department of Biochemistry and Molecular Biology, Faculty of Biology, IBUB, Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Spain

[6] INRA, Univ. Bordeaux, UMR1332 Fruit Biology and Pathology, Metabolome Facility of Bordeaux - MetaboHUB, Functional Genomics Center, IBVM, Centre INRA Bordeaux, 71 av Edouard Bourlaux, 33140 Villenave d'Ornon, France

[7] Centre of bioinformatics of Bordeaux (CBiB), University of Bordeaux, 33000 Bordeaux, France

[8] University of Oxford e-Research Centre, 7 Keble Road, Oxford OX1 3QG, UK

[9] University of Bordeaux, CBiB/LaBRI, 33000 Bordeaux, France

[10] School of Chemistry & Manchester Institute of Biotechnology, University of Manchester, 131 Princess St., Manchester M1 7DN, UK

[11] Microbiology & Systems Biology, TNO, Zeist, The Netherlands

[12] Division of Analytical Biosciences, Leiden Academic Center for Drug Research, Leiden University, Leiden, The Netherlands

[13] Zora Biosciences OY, 02150 Espoo, Finland

[14] Medical Research Council Human Nutrition Research, Fulbour Road, Cambridge CB1 9NL, UK

[15] Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK

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

[17] School of Cancer Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

[18] School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

[19] FiorGen Foundation, 50019 Sesto Fiorentino, FI, Italy

storage, exchange, comparison and re-utilization of metabolomics data. The Framework Programme 7 EU Initiative 'coordination of standards in metabolomics' (COSMOS) is developing a robust data infrastructure and exchange standards for metabolomics data and metadata. This is to support workflows for a broad range of metabolomics applications within the European metabolomics community and the wider metabolomics and biomedical communities' participation. Here we announce our concepts and efforts asking for re-engagement of the metabolomics community, academics and industry, journal publishers, software and hardware vendors, as well as those interested in standardisation worldwide (addressing missing metabolomics ontologies, complex-metadata capturing and XML based open source data exchange format), to join and work towards updating and implementing metabolomics standards.

Keywords Metabolomics · Metabonomics · Data standards · Data exchange · e-Infrastructure · Coordination and data sharing community

1 Introduction

Metabolomics (Bundy et al. 2009; Clayton et al. 2006; Eckhart et al. 2012; Holmes et al. 2008)[1] and fluxomics (metabolic flux analysis, Zamboni, Nicola et al. "13C-based metabolic flux analysis." *Nature protocols* 4.6 (2009): 878–892) measurements mark the end point closest to the phenotype of organisms, reflecting changes in organisms influenced by external parameters such as nutritional, environmental or toxicological interactions. In this context, due to its dynamic nature, metabolomics is of considerable value for examples in personalised medicine, especially as it captures rapid responses close to the phenotype and in concert with the genome, transcriptome and epigenome (van der Greef et al. 2006, Nicholson et al. 2011). For such methods to succeed in a personalised medicine context, robust traceable standardisation is essential, covering storage and exchange of metabolomics and fluxomics data. Moreover, new applications that link *metabolomics and biobanks* are emerging: metabolomics may be used as an efficient tool to monitor the quality of stored samples and to establish the optimal standard operating procedures (SOPs) for the pre-analytical handling of bio-specimens (Bernini et al. 2011). Metabolomics is rapidly becoming an essential tool in the *screening of food products,* which is highly regulated and follows standard guidelines. Furthermore it is being investigated as a potentially transformative technology for the *screening of chemical safety*, not only for traditional industrial and domestic chemicals but also for the safety assessments of engineered nanomaterials as well as novel compounds generated through synthetic biology.

Considering the diversity and breadth of metabolomics applications, not forgetting complexity and diversity of the analytical technologies in use, there is a clearly identified need for standardisation that evolves with the technologies and is sufficiently inclusive to cover all metabolomics applications.

2 What has been achieved so far in metabolomics standards

The momentum for metabolomics standards started in 2004–2005 with initiatives such as the standard metabolic reporting structure initiative or SMRS (Lindon et al. 2005) and the Architecture for Metabolomics consortium or Armet (Jenkins et al. 2004); these were mainly focused on an aspect of metabolomics standards, for example nuclear magnetic resonance (NMR) based metabonomics or plant-based metabolomics. There were several other initiatives at the time, however all efforts eventually resulted in the formation of the metabolomics standards initiative (MSI) in 2005 (Castle et al. 2006; Fiehn et al. 2006). This was focused on community-agreed minimum reporting standards and providing initial efforts on the descriptions of the experimental metadata describing a metabolomics study. This culminated in a series of manuscripts published in 2007 that considered all the components undertaken in metabolomics experiments (Sansone et al. 2007; Fiehn et al. 2007; Hardy and Taylor 2007) summarized in (Goodacre 2014). One major outcome was the formation of five different working groups (WG) to consider each aspect of the metabolomic pipeline; biological context metadata WG, chemical analysis WG, data processing WG, ontology WG and exchange format WG, with the task of collecting relevant metabolomics standards and a forum for discussion (Goodacre et al. 2007; Morrison et al. 2007; Rubtsov et al. 2007; Sumner et al. 2007; Werf et al. 2007). However, there have been limited practical applications for such descriptions, with some exceptions (Ludwig et al. 2012; Bais et al. 2010; Ferry-Dumazet et al. 2011; Griffin et al. 2011; Scholz and Fiehn 2007), in part owing to a lack of tools to facilitate implementation or a widely used database to enforce such standards. Most projects or databases focused on one particular technology or limited to a particular species or type of analytical technique. In order to implement agreed and acceptable guidelines on reporting identified metabolites, an application platform such as database i.e. a metabolomics repository in addition to a journal publication is required. 2012 saw the release of MetaboLights (http://www.ebi.ac.uk/metabolights), the

[1] "Metabolome | Metabolomics Definition | InTechOpen." 2012. 20 Aug. 2013 http://www.intechopen.com/books/metabolomics.

first general purpose database in metabolomics, developed and maintained by the European Bioinformatics Institute (EMBL-EBI), one of the largest open access data providers in the world (Haug et al. 2013; Salek et al. 2013a). MetaboLights combines small molecule 'reference' layer with information about individual metabolites, their chemistry, spectrometry and biological roles with a study archive, where primary data and metadata from metabolomics studies are ontologically tagged and stored. Such depositions receive a stable identifier for each study, which can be quoted in related publications and can be used to access the data long term. Making metabolomics data publicly accessible allows it to justify researchers' findings in a peer-reviewed publication, increases the possibility of wider collaborations within the metabolomics community and ultimately gives a study higher visibility and increased citation (Nature Genetics 2009). MetaboLights adheres to MSI standards and uses the Investigation/Study/Assay (ISA) tab-delimited format (Rocca-Serra et al. 2010), which makes it interoperable with a large number of other ongoing projects dealing with biological study data, including other 'omics datasets. Specific scientific fields have developed their own systems biology solutions (e.g. dbNP developed by NuGO, covering the Nutritional Phenotype), and the metabolomics data of such sites should be made exchangeable with other metabolomics databases (for instance by implementing an export to ISA-Tab format for the study metadata as well as exporting the results).

3 COSMOS: the way forward in standards

The FP7 EU Initiative 'coordination of standards in metabolomics' (COSMOS) brings together leading vendors, researchers and bioinformaticians of the European metabolomics community, members of the MSI, members of the international Metabolomics Society, along with other stakeholders worldwide. One of the COSMOS initiative goals is to develop a robust data infrastructure for metabolomics data and metadata representation and exchange in order to support workflows for a broad range of metabolomics applications (Salek et al. 2013b; Steinbeck et al. 2012). The potential of metabolomics cannot be harvested without major standardisation of formats and terminologies, therefore we leverage on and extend earlier efforts initiated by the MSI and currently operating under the Metabolomics Society, in part via that society's dedicated Data Standards task group. As is the case for other high-throughput "-omics" disciplines, metabolomics is seeing a paradigm shift from hypothesis-driven to data-driven science (Cox and Mann 2011; Goodacre et al. 2004). As a result, metabolomics data are constantly growing with a plethora of analysis tools. Cross-site data comparison remains a challenging task due to the different access modalities for the different local repositories. Hence, currently the generated data often ends up in data silos or worse as data dumps or 'data-graveyards'. This situation constitutes a need for the establishment of open data standards and accessible repositories that allow researchers to store, exchange and compare metabolomics data with pertinent metadata information, and thus communicate on a scientific level without getting stuck in vendor specific data formats. As different scientific fields continue to develop their own specific solutions, due to specific analytical solutions or meta-data requirements, COSMOS will invite these fields to adhere to the general metabolomics standards and export to the metabolomics solutions developed by COSMOS. To compare data between labs, the data needs to be stored in a way that allows concise objective interpretation and reproducibility, i.e. the type, origin and treatment of samples and corresponding spectra needs to be described in an unambiguous manner using a common communication channel. Here, controlled vocabularies (CVs) and ontologies can be used to standardise the terminology used to represent scientific facts, e.g. tissue or fluid description, sample storage, preparation and analysis conditions. Another benefit of CVs are their knowledge representation capabilities, i.e. their taxonomic backbone that can be exploited to gather more subsumptive (a more general/abstract) or a more excluding, search specific attributes.

To work out commonly agreed-upon metabolomics data standards, the COSMOS initiative coordinates with metabolomics and bioinformatics experts to work on open data exchange formats (syntax) and data semantics that maximize interoperability with other omics standards (Nature Genetics 2012). This is achieved among other solutions, by using (i) the general-purpose Investigation/Study/Assay tabular format or ISA-Tab (Rocca-Serra et al. 2010) for the experimental information and (ii) adapting the XML-based formats for the instrument-derived "raw" data types by the proteomics standards initiative (PSI) (Orchard et al. 2003b; Orchard et al. 2003a), e.g. *mzML* (Martens et al. 2011). Data completeness can then be verified using validator software enforcing minimum information recommendations such as the MSI Core Information for Metabolomics Reporting (CIMR; http://biosharing.org/bsg-000175) (Sumner et al. 2007). The standardisation efforts in COSMOS for nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) data, with potential to encompass new and alternative technologies as they are developed, and supporting tools, will form the basis for funders and publishers to recommend data deposition. The submitted dataset, in repositories such as MetaboLights (Haug et al. 2013) or the Netherlands Metabolomics Centre (NMC) Data Support Platform (DSP) could then be used to justify findings in a publication. Unlike in other -omics

domains, as for example ArrayExpress (Parkinson et al. 2009) and Gene Expression Omnibus GEO (Barrett et al. 2013) for transcriptomics, the previous lack of such open, centralised and persistent data deposition repositories in the metabolomics field has been criticised by journal editors who face the tedious task of having to judge whether conclusions based upon megavariate data are sound and justified. Here open-access repositories using our standards, data curation and capture tools, such as the ISA software suite (Rocca-Serra et al. 2010), as well as others, will facilitate curation and storage of the metadata at the source, and streamline submission to MetaboLights. A growing number of data publication journals, e.g. BioMedCentral's GigaScience[2] and Nature Publishing Group's *Scientific Data*[3] now support the ISA format for supplementary experimental data and as a means to capture metadata descriptions. ISA-Tab format is currently in use and supported by public data repositories such EMBL-EBI Metabolights (accounting for about 200 datasets, 90 of which are currently publicly available), but also several major european toxicogenomics projects (Carcinogenomics,[4] DiXa[5] and InnoMed PredTox and ToxBank[6]). These projects fully exploit the capability of the ISA-Tab format to support an array of assay type allowing to recording multi-omics assays. Furthermore, ISA developers have a range of tools for converting from various sources (ArrayExpress,[7] SRA[8]) into ISA-Tab format.

Ultimately, we hope that COSMOS will help experts in NMR spectroscopy and MS-based metabolomics to communicate their results in a more objective comprehensive, persistent and efficient way, and spanning and integrating multiple domains such as medical, environmental, plant and food sciences.

Although funding by the European Community is limited to a number of European expert scientists, COSMOS links to major initiatives world-wide. For example with the US National Institutes of Health (NIH) Common Funds Metabolomics Initiative (http://commonfund.nih.gov/Metabolomics/) which has awarded funding for six Regional Comprehensive Metabolomics Research Cores (RCMRC), and a Data Repository and Coordination Centre (DRCC), to act as a North American hub for metabolomics related research.[9] COSMOS also reaches and establishes links with other related e-infrastructures initiative such as the new European-wide ELIXIR project,[10] Biobanking and Biomolecular Resources Research Infrastructure[11] (BBMRI) via BioMedbridges[12] consortium, Human Metabolome Database (HMDB) in Canada (Wishart et al. 2013), Platform for RIKEN Metabolomics (PRIMe) in Japan (Sakurai et al. 2013) and Beijing Genomics Institute (BGI)[13] in China.

3.1 Metabolomics data exchange standards

The COSMOS work on standardisation aims to build on the foundational work by PSI and MSI and further develop and contribute to data exchange formats, ranging from raw data in MS and NMR, the reporting of metabolite quantification and metabolite identification, to the experimental metadata. We aim to extend the open standards for MS data exchange initiated by PSI, such as mzML (Martens et al. 2011), mzIdentML (Jones et al. 2012) and mzQuantML (Walzer et al. 2013) to meet the requirement of metabolomics experiments for reporting MS experiments. One example are GC–MS based metabolomics experiments, where data are often available in either a closed vendor format or as netCDF, where the latter provides only very few metadata acquisition parameters and fails to capture advanced MS experiments such as GC × GC–MS or tandem-MS with GC–MS instruments. This requirement led us to augment the PSI-MS controlled vocabulary with GC specific terms and concepts, which have already been included in the current PSI-MS ontology. To avoid a chicken-and-egg problem, we have collected raw data examples "in the wild"; and checked which GC–MS vendor formats can be converted to mzML. Currently, file formats by Agilent Technologies, Bruker Biosciences Corporation, Waters Corporation and Thermo Fisher Scientific are companies that their file format is readily supported by the Proteowizard Open Source converter (Kessner et al. 2008; Chambers et al. 2012). Other companies, such as LECO Corporation and Bruker Biosciences in addition have software to export their file format to mzML. On the consumer side, we surveyed which mzML parsing libraries are available for the community. Parsers for mzML exist for the languages C++, Java, R and Python,

[2] "GigaScience." 2011. 20 Aug. 2013 http://www.gigasciencejournal.com/.

[3] "Scientific Data - Nature." 2013. 20 Aug. 2013 http://www.nature.com/scientificdata/.

[4] Vinken "The carcinoGENOMICS project: critical selection of model …" 2008. http://www.ncbi.nlm.nih.gov/pubmed/18514569.

[5] "diXa." 2012. 12 Mar. 2015 http://www.dixa-fp7.eu/.

[6] "ToxBank | Data warehouse for toxicity data management." 2010. 12 Mar. 2015 http://toxbank.net/.

[7] "ArrayExpress < EMBL-EBI - European Bioinformatics Institute." 2008. 12 Mar. 2015 https://www.ebi.ac.uk/arrayexpress/.

[8] "Home - SRA - NCBI." 2009. 12 Mar. 2015 <http://www.ncbi.nlm.nih.gov/sra>.

[9] "NIH announces new program in metabolomics." 2012. 20 Aug. 2013 http://www.nih.gov/news/health/sep2012/od-19.htm.

[10] "home | ELIXIR." 2007. 20 Aug. 2013 http://www.elixir-europe.org/.

[11] "BBMRI: Home." 2008. 20 Aug. 2013 http://www.bbmri.eu/.

[12] "BioMedBridges: Home." 2012. 20 Aug. 2013 http://www.biomedbridges.eu/.

[13] "BGI." 2009. 20 Aug. 2013 http://www.genomics.cn/.

which should cover the majority of the current software developments in the metabolomics community, and we shall of course supplement to these as necessary. Therefore, mzML can be a strong suggestion or even a requirement for data deposition in public repositories. Formats that capture metabolite identification and reporting of quantification results also need to be adapted for MS metabolomics experiments, and require real-world tests, software support and community adoption. We hope to make COSMOS a platform for community engagement with adaptation and development of these formats to suit the metabolomics community needs, and as mentioned above we regularly consult with the Metabolomics Society. In addition, COSMOS developing the missing XML exchange formats for NMR spectroscopy such as nmrML, nmrIdentML, nmrQuantML and nmrTab needed by databases and open source software such as NMRLab/MetaboLab (Ludwig and Gunther 2011) Bayesian AuTomated Metabolite Analyser for NMR spectra (BATMAN) (Hao et al. 2012) and rNMR (Lewis et al. 2009). These developments take place on www.nmrml.org and https://github.com/nmrML/ and include the XML schema, controlled vocabulary (Schober et al. 2014), example files, and reader, writer, conversion and validation software. As part of this approach we have begun and will continue to interact with the wider community to ensure wide adoption and call for implementations of the standards during the design phase, which helps to catch design errors before the standard is published. With semantic web technology in mind, these standards will pave the way for metabolomics data to be part of the world of linked (and open) data (Murray-Rust 2008). Preliminary work in currently underway, leveraging work by the ISAteam in the field of linked data to offer MetaboLights metadata content as linked data (González-Beltrán et al. 2014a).

3.2 Metabolomics databases and repositories

The power of broad, system-wide -omics relies on the potential to interrogate datasets from new perspectives. Researchers not involved in the original data generation process may reuse data differently from the original purpose that motivated the data collection. Unleashing this potential also in the heterogeneous metabolomics landscape requires the availability of metabolite level (ideally quantified) and profile data along with adequate metadata. Therefore, the COSMOS consortium is committed to develop the MetaboLights database further as a centralised data exchange and storage platform. MetaboLights will serve as a common publication hub and make it possible to connect different resources while keeping the data interoperable (e.g. connect to data in other resources, such as the NMC-DSP (van Ommen et al. 2010)). In the metabolomics field, a large number of custom and often technology focused, substance-class, or species-centric databases exist and are continuously developed for example; the Golm Metabolome Database or GMD (Hummel et al. 2010) LipidMaps (Fahy et al. 2009), PlantMetabolomics.org (Bais et al. 2010) and MeRy-B (Ferry-Dumazet et al. 2011). Defining a sensible balance of centralised versus decentralised information storage can be resolved by developing and applying standards and exchange formats. Also, buy-in from users as well as publishers will have to be achieved. The journal *Metabolomics*, which is the official journal of the Metabolomics Society, has since 2010 encouraged authors to ensure their papers are as MSI compliant as possible (Goodacre 2010) and is committed to supporting the COSMOS consortium in its endeavours for metabolomics standardisation. This ethos is also being adopted by other journals including *Metabolite, EMBO* and others to join.

Integrative analysis of datasets is essential in order to achieve better understanding of phenotypes. Moreover, interfacing with dedicated databases utilising metadata annotation tools will engage and enable a broad user base to export data from their local systems into ISA-Tab formatted data sets, and subsequently to easily import or submit to *MetaboLights*. MetaboLights and the ISA team have been working on implementing principled curation guidelines, ensuring consistency in the reporting of experimental designs. As a machine readable XML dialect, the schema based XEML (Hannemann et al. 2009) provides means to store experimental design and metadata describing the actual experiment, together with links to one or more independent databases hosting the actual experimental results as well as export the results in ISA-Tab format. The XEML-Lab sources and binaries for different operating systems can be accessed and downloaded from https://github.com/cbib/XEML-Lab.

In addition to the obvious breadth of experimental conditions, the diversity of laboratory specific SOPs, even within the most commonly employed measurement techniques such as NMR and MS, renders the joint interpretation of data produced by different labs difficult. Hence, standardised and machine-readable metadata describing all aspect of experimental conditions are an essential prerequisite to allow a quick, objective and hence meaningful selection of experiments suitable for comparison. In a single-user environment, experimental metadata annotation can be efficiently handled using the ISAcreator, part of the ISA software suite. In addition, the COSMOS consortium also aims to develop standards to connect to existing specialised databases such as the GMD, MeRy-B, the NMC-DSP/dbNP, as well as other similar resources using alternative metadata annotation tools such as XEML (Hannemann et al. 2009) or the automatic processing pipelines within Bioconductor packages (Gonzalez-Beltran et at. 2014b) and Bioportal powered ontology (Maguire et al. 2013).

3.3 Data deposition workflow

Making raw data available to the interested research community has clear benefits to the transparency and trustworthiness of those scientific studies. Scientists might choose a variety of resources for their data deposition, depending on their preferred technology. A comprehensive workflow should protect data proprietary interests, security (data will not be made publicly available until the associated publication has a bona fide DOI or the authors request immediate data release), and confidentiality as required. To ensure proper reporting of metabolomics data and metadata (Salek et al. 2013c), COSMOS will set clear procedures for data submission and deposition, as well as metabolomics results reporting considering publishing requirements. These will be in line with the existing MSI guidelines. These guidelines are currently being carefully discussed, elaborated and agreed by all COSMOS partners. COSMOS is also taking every opportunity to engage fully with stakeholders and potential collaborators on planning, discussion and implementation of the guidelines for data deposition workflow. Careful planning of the data deposition flow, its control policies and actions, will ensure that the utility of the system is maximized and quality-controlled for use inside and outside Europe. A proposed model for the data deposition workflow, drafted from discussions within COSMOS, is shown in Fig. 1. The workflow definition will prioritize simplicity, usability, annotation quality, the plurality of metabolomics resources and databases, to ensure connectivity between similar studies and to provide rapid matching results to the end users. We envisage that in the future additional purpose-built databases will be created that can potentially be integrated into the proposed workflow. This will include MetaboLights and the NIH funded Metabolomics Workbench.[14] The first phase of the data deposition cycle is temporary and all data and associated information are kept private. Once the study has been officially published and the depositor agrees to share the data (making it open access), the COSMOS "metabolomeXchange" system will automatically announce and broadcast availability of such studies to the broad research community (Fig. 2). In addition to the minimum required metadata (e.g. accession, title, abstract, publication date, URL and submitters name), the COSMOS "metabolomeXchange" system allows datasets to be annotated using additional metadata information. This would enable the metabolomics community (both metabolomics researchers and databases) to query efficiently and readily identify interesting and reusable metabolomics data sets.

The default for most journals will be to promote open access of the data as soon as the study is published, while the accepted academic standard is to allow others to access the work. All parties involved will benefit from sharing raw data, processed data, metadata, statistical methods and source codes. By increasing the visibility of their work, depositors are likely to boost citations. Publishers and journals will expose their publications to a greater number of potential readers and enhance the overall impact of the work. In addition, through COSMOS the research community will gain free access to a vast amount of well documented and easy to access scientific information.

3.4 Coordination with BioMedBridges and biomedical ESFRI infrastructures

COSMOS aims at building a network of close interactions with the European biomedical infrastructures. A particular interest is in the infrastructures for which metabolomics is most relevant such as BBMRI, ELIXIR, EU-Openscreen,[15] EuroBioimaging[16] and EURODISH,[17] all of which are participating in the EC-funded project BioMedBridges.[18] Our main objective is to obtain indicators, useful to prioritize various COSMOS activities, in order to obtain effectively responses needed in large-scale EU biomedical infrastructures. BioMedBridges and COSMOS are reasonably complementary as they address different levels of information—COSMOS is more focused on the experimental side whereas BioMedBridges tackles the higher complexity of human disease, investigated by a plethora of different technologies, of which one is metabolomics. Another important collaboration is COSMOS with BBMRI. Human biobanks are structured resources that store: (a) human biological materials and/or information generated from the analysis of the same and (b) extensive associated information. An emerging aspect would be the usage of metabolomics as an efficient tool to monitor pre-analytical sample variations as metabolites are known to be prone to degradation phenomena, possibly more so than other clinical biomarkers (e.g., DNA, RNA, proteins). Another application of metabolomic profiling in relation to biobanking is in assessment of quality and usage history for samples and their collection, handling, traceability and storage. Here COSMOS may play a key role in the development of relevant required standardised formats.

[14] "Metabolomics Workbench." 2013. 20 Aug. 2013 <http://www.metabolomicsworkbench.org/>.

[15] "WELCOME TO EU-OPENSCREEN." 2009. 20 Aug. 2013 http://www.eu-openscreen.de/.

[16] "Euro-BioImaging: Research Infrastructure for Imaging Technologies …" 2009. 20 Aug. 2013 http://www.eurobioimaging.eu/.

[17] "EuroDISH: Home." 2013. 18 Sep. 2013 http://www.eurodish.eu/.

[18] "BioMedBridges: Home." 2012. 20 Aug. 2013 http://www.biomedbridges.eu/.

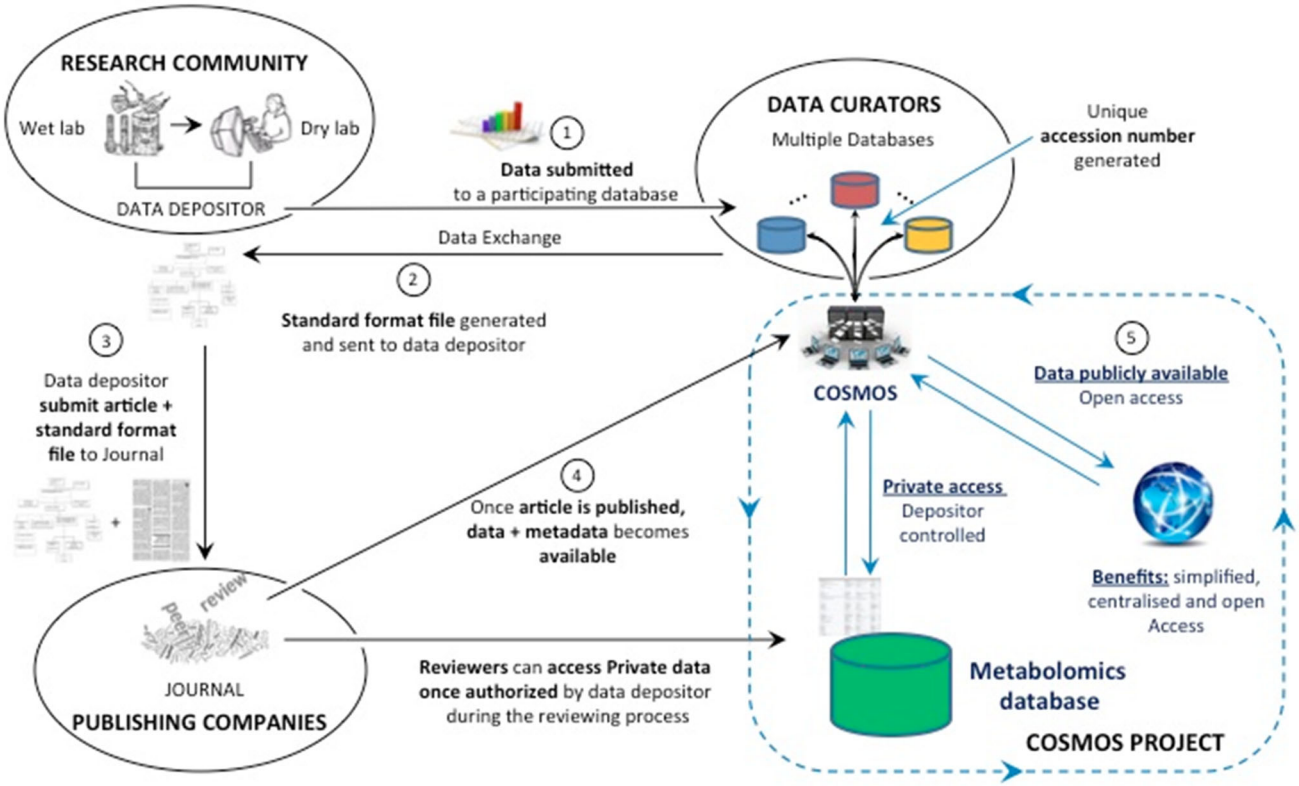

Fig. 1 Initial model for the COSMOS data deposition workflow system. The data deposition cycle is initiated when a Submitter (who has generated or owns the study material) submits his/her metabolomics study to a specific associated database (*1*). Once the data submission has been completed, fulfilling the requirement of the associated repository submission guideline, a unique COSMOS accession number will be generated. The COSMOS engine will then properly annotate, format and store the minimum agreed metadata according to proposed reporting standards suggested by COSMOS partners (*2*, *3*). COSMOS will bring together publishers and other metabolomics repositories to come to final agreement on a data workflow specifying minimum metadata exchange, associated raw data, source code and any additional information that can be shared (*4*, *5*). Open access to the system will ensure that any interested party can benefit from the standardized resources

Formal interactions between COSMOS and BBMRI are being established to coordinate efforts: first, the coordinator of BBMRI, has been nominated in the Advisory Board of COSMOS; second, and also within the frame of BBMRI (which specifically aims at integrating existing biomolecular resources, technologies, standards and know-how), the University of Florence Magnetic Resonance Center (CERM/CIRMMP) partner of COSMOS is establishing a European multi-site Expert Center on metabolomics.

3.5 Outreach and dissemination

We will maintain a close link between the COSMOS consortium and the wider metabolomics and biomedical communities, as well as other related scientific fields. We shall continue to use metabolomics workshops, meetings and conferences to interact with the wider communities, metabolomics or others, to get all parties involved with the initiative. In addition, we will use publications and social media, news articles and blogs to raise awareness of metabolomics standards and the services provided by the COSMOS consortium. This will be from data submission to providing support on different views on how metabolomics data may be reported. There is a great need for increasingly stringent requirements for data quality that is publicly available to the whole metabolomics community. Through the existing framework of the Metabolomics Society, we will ensure broad community input into the services and standards for metabolomics data representation developed by COSMOS consortium, and it may come as no surprise that several of the authors have served on one or more of the Boards of the Metabolomics Society.

4 Concluding remarks

COSMOS will implement harmonised and compatible data deposition strategies and contribute to annotation workflows, providing data producers involved in metabolomics experiments with a single point of submission, while allowing other data entry points through facilitation of interoperability. The data deposition and exchange workflow

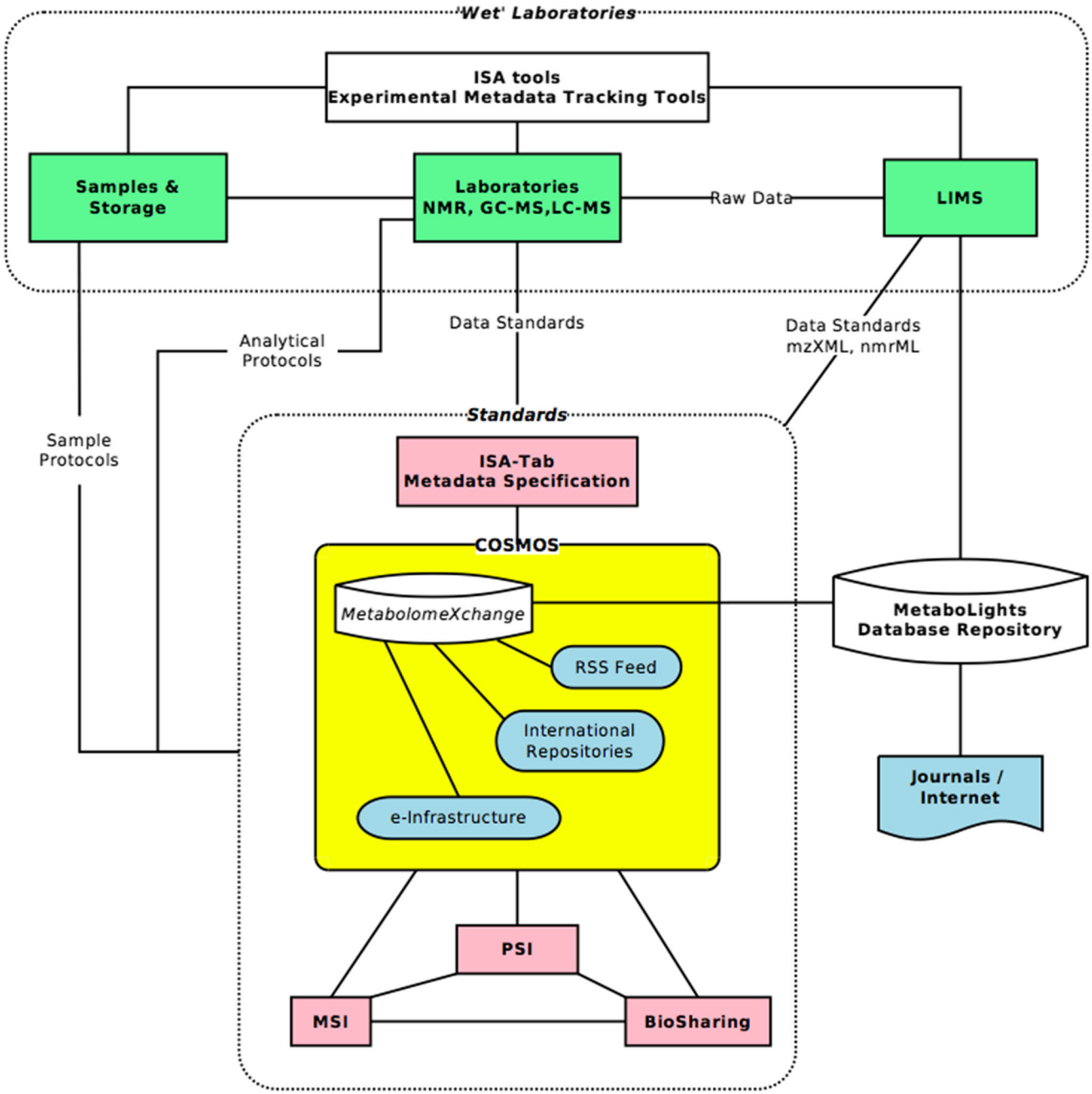

Fig. 2 COSMOS initiative workflow. Overview of COSMOS initiative role in metabolomics standards, databases, data exchange and dissemination of metabolomics experiments. *Green* metabolomics labs experimental workflow from lab based data generation, metadata collection to interaction with LIMS systems. *Pink* standardisation initiative and minimum information reporting agreement involved or used within the COSMOS project. *Blue* Dissemination and role of journal and link to other e-infrastructures. *White* databases and tools used to capture experimental data and metadata (Color figure online)

in the COSMOS consortium will be formally defined, agreed, and documented in relation with MetaboLights and all partnering databases in Europe and worldwide that would like to participate, and we welcome discussions from other data providers and software houses. A guideline for submitted data will be generated and COSMOS will strive to make sure that all metabolomics data submitted to partner databases are exchangeable with this standard. Since the adoption of minimal standards for metabolomics by the relevant journals is a major goal of this coordination action, we will consult with journal publishers and ensure data annotation quality and consistency, according to the required standard level set by each journal. For example, collaboration with data journals such as *GigaScience* and

Nature's *Scientific Data* will be streamlined, given these journals already use the ISA-Tab format for data submission. We are also working closely with *Metabolomics,* which is the official journal of the Metabolomics Society.

The COSMOS consortium ultimately develops the standards and infrastructure for—and with—the metabolomics and fluxomics community. For the most efficient interaction we have already, and will continue, to organise stakeholder meetings as satellite events to major metabolomics meetings, individual staff exchange between partners, as well as larger workshops. These efforts will directly enable the implementation of COSMOS important deliverable—that of a robust data infrastructure and mechanisms for standards metabolomics data representation and data/meta-data exchange that will enrich metabolomics science.

Acknowledgments The authors would like to acknowledge the contribution of Michael Wilson and David Wishart in NMR standardisation work carried out within COSMOS. COSMOS advisory board; Drs. David Wishart, Oliver Fiehn, Masanori Arita, Yulan Wang, Richard Beger, Rima Kaddurah-Daouk and Kurt Zatloukal, Finally European Commission framework seven for funding COSMOS grant EC312941.

Compliance with ethical requirements This article does not contain any studies with human or animal subjects.

Conflict of interest The authors have no conflicts of interest to declare.

References

Bais, P., Moon, S. M., He, K., Leitao, R., Dreher, K., Walk, T., et al. (2010). PlantMetabolomics.org: a web portal for plant metabolomics experiments. [Research Support, U.S. Gov't, Non-P.H.S.]. *Plant Physiology, 152*(4), 1807–1816. doi:10.1104/pp.109.151027.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets: update. [Research Support, N.I.H., Intramural]. *Nucleic Acids Research, 41*, D991–D995. doi:10.1093/nar/gks1193.

Bernini, P., Bertini, I., Luchinat, C., Nincheri, P., Staderini, S., & Turano, P. (2011). Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. [Research Support, Non-U.S. Gov't]. *Journal of Biomolecular NMR, 49*(3–4), 231–243. doi:10.1007/s10858-011-9489-1.

Bundy, J., Davey, M., & Viant, M. (2009). Environmental metabolomics: A critical review and future perspectives. *Metabolomics, 5*(1), 3–21. doi:10.1007/s11306-008-0152-0.

Castle, A. L., Fiehn, O., Kaddurah-Daouk, R., & Lindon, J. C. (2006). Metabolomics standards workshop and the development of international standards for reporting metabolomics experimental results. [Consensus Development Conference, NIH]. *Briefings in Bioinformatics, 7*(2), 159–165. doi:10.1093/bib/bbl008.

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology, 30*(10), 918–920. doi:10.1038/nbt.2377.

Clayton, T. A., Lindon, J. C., Cloarec, O., Antti, H., Charuel, C., Hanton, G., et al. (2006). Pharmaco-metabonomic phenotyping and personalized drug treatment. [Research Support, Non-U.S. Gov't]. *Nature, 440*(7087), 1073–1077. doi:10.1038/nature04648.

Cox, J., & Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. [Research Support, Non-U.S. Gov't Review]. *Annual Review of Biochemistry, 80*, 273–299. doi:10.1146/annurev-biochem-061308-093216.

Eckhart, A. D., Beebe, K., & Milburn, M. (2012). Metabolomics as a key integrator for "omic" advancement of personalized medicine and future therapies. [Review]. *Clinical and Translational Science, 5*(3), 285–288. doi:10.1111/j.1752-8062.2011.00388.x.

Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R., Shimizu, T., et al. (2009). Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of Lipid Research, 50*, S9–S14. doi:10.1194/jlr.R800095-JLR200.

Ferry-Dumazet, H., Gil, L., Deborde, C., Moing, A., Bernillon, S., Rolin, D., et al. (2011). MeRy-B: A web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. [Research Support, Non-U.S. Gov't]. *BMC Plant Biology, 11*, 104. doi:10.1186/1471-2229-11-104.

Fiehn, O., Kristal, B., van Ommen, B., Sumner, L. W., Sansone, S. A., Taylor, C., et al. (2006). Establishing reporting standards for metabolomic and metabonomic studies: A call for participation. *OMICS, 10*(2), 158–163. doi:10.1089/omi.2006.10.158.

Fiehn, O., Robertson, D., Griffin, J., Werf, M., Nikolau, B., Morrison, N., et al. (2007). The metabolomics standards initiative (MSI). *Metabolomics, 3*(3), 175–178. doi:10.1007/s11306-007-0070-6.

González-Beltran, A., Neumann, S., & Maguire, E. (2014). The risa R/bioconductor package: Integrative data analysis from experimental metadata and back again. *BMC Bioinformatics, 15*(1), S11. doi:10.1186/1471-2105-15-S1-S11.

González-Beltrán, A., et al. (2014). LinkedISA: Semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics, 15*(14), S4.

Goodacre, R. (2010). An overflow of… what else but metabolism! *Metabolomics, 6*(1), 1–2. doi:10.1007/s11306-010-0201-3.

Goodacre, R. (2014). Water, water, every where, but rarely any drop to drink. *Metabolomics, 10*(1), 5–7.

Goodacre, R., Broadhurst, D., Smilde, A., Kristal, B., Baker, J. D., Beger, R., et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics, 3*(3), 231–241. doi:10.1007/s11306-007-0081-3.

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. [Research Support, Non-U.S. Gov't Review]. *Trends in Biotechnology, 22*(5), 245–252. doi:10.1016/j.tibtech.2004.03.007.

Griffin, J. L., Atherton, H. J., Steinbeck, C., & Salek, R. M. (2011). A Metadata description of the data in "A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human.". *BMC Research Notes, 4*, 272. doi:10.1186/1756-0500-4-272.

Hannemann, J., Poorter, H., Usadel, B., Blasing, O. E., Finck, A., Tardieu, F., et al. (2009). Xeml Lab: A tool that supports the

design of experiments at a graphical interface and generates computer-readable metadata files, which capture information about genotypes, growth conditions, environmental perturbations and sampling strategy. *Plant, Cell and Environment, 32*(9), 1185–1200. doi:10.1111/j.1365-3040.2009.01964.x.

Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. (2012). BATMAN: An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics, 28*(15), 2088–2090. doi:10.1093/bioinformatics/bts308.

Hardy, N., & Taylor, C. (2007). A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics, 3*(3), 243–248. doi:10.1007/s11306-007-0071-5.

Haug, K., Salek, R. M., Conesa, P., Hastings, J., deMatos, P., Rijnbeek, M., et al. (2013). MetaboLights: An open-access general-purpose repository for metabolomics studies and associated meta-data. [Research Support, Non-U.S. Gov't]. *Nucleic Acids Research, 41*, D781–D786. doi:10.1093/nar/gks1004.

Holmes, E., Wilson, I. D., & Nicholson, J. K. (2008). Metabolic phenotyping in health and disease. *Cell, 134*(5), 714–717. doi:10.1016/j.cell.2008.08.026.

Hummel, J., Strehmel, N., Selbig, J., Walther, D., & Kopka, J. (2010). Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics, 6*(2), 322–333. doi:10.1007/s11306-010-0198-7.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A. R., Taylor, J., et al. (2004). A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology, 22*(12), 1601–1606. doi:10.1038/nbt1041.

Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., et al. (2012). The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular and Cellular Proteomics: MCP, 11*(7), M111 014381. doi:10.1074/mcp.M111.014381.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534–2536. doi:10.1093/bioinformatics/btn323.

Lewis, I. A., Schommer, S. C., & Markley, J. L. (2009). rNMR: Open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance Chemistry, 47*(1), S123–S126. doi:10.1002/mrc.2526.

Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C., Craig, A., Pearce, J. T., et al. (2005). Summary recommendations for standardization and reporting of metabolic analyses. (Guideline). *Nature Biotechnology, 23*(7), 833–838. doi:10.1038/nbt0705-833.

Ludwig, C., Easton, J., Lodi, A., Tiziani, S., Manzoor, S., Southam, A., et al. (2012). Birmingham metabolite library: A publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics, 8*(1), 8–18. doi:10.1007/s11306-011-0347-7.

Ludwig, C., & Gunther, U. L. (2011). MetaboLab: Advanced NMR data processing and analysis for metabolomics. *BMC Bioinformatics, 12*, 366. doi:10.1186/1471-2105-12-366.

Maguire, E., Gonzalez-Beltran, A., Whetzel, P. L., Sansone, S. A., & Rocca-Serra, P. (2013). OntoMaton: A bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics, 29*(4), 525–527. doi:10.1093/bioinformatics/bts718.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., et al. (2011). mzML: A community standard for mass spectrometry data. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Molecular and Cellular Proteomics, 10*(1), R110 000133. doi:10.1074/mcp.R110.000133.

Morrison, N., Bearden, D., Bundy, J., Collette, T., Currie, F., Davey, M., et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: environmental context. *Metabolomics, 3*(3), 203–210. doi:10.1007/s11306-007-0067-1.

Murray-Rust, P. (2008). Open data in science. *Nature Precedings,*. doi:10.1016/j.serrev.2008.01.001.

Nature Genetics. (2009). Data producers deserve citation credit. (Editorial). *Nature Genetics, 41*(10), 1045. doi:10.1038/ng1009-1045.

Nature Genetics. (2012). It's not about the data. *Nature Genetics, 44*(2), 111. doi:10.1038/ng.1099.

Nicholson, J. K., Wilson, I. D., & Lindon, J. C. (2011). Pharmacometabonomics as an effector for personalized medicine. *Pharmacogenomics, 12*(1), 103–111.

Orchard, S., Hermjakob, H., & Apweiler, R. (2003a). The proteomics standards initiative. [Research Support, Non-U.S. Gov't]. *Proteomics, 3*(7), 1374–1376. doi:10.1002/pmic.200300496.

Orchard, S., Kersey, P., Hermjakob, H., & Apweiler, R. (2003b). The HUPO proteomics standards initiative meeting: Towards common standards for exchanging proteomics data. *Comparative and Functional Genomics, 4*(1), 16–19. doi:10.1002/cfg.232.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., et al. (2009). ArrayExpress update: From an archive of functional genomics experiments to the atlas of gene expression. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Nucleic Acids Research, 37*(Database issue), D868–D872.

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. [Research Support, Non-U.S. Gov't]. *Bioinformatics, 26*(18), 2354–2356. doi:10.1093/bioinformatics/btq415.

Rubtsov, D., Jenkins, H., Ludwig, C., Easton, J., Viant, M., Günther, U., et al. (2007). Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics, 3*(3), 223–229. doi:10.1007/s11306-006-0040-4.

Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K., et al. (2013). PRIMe update: Innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant and Cell Physiology, 54*(2), e5. doi:10.1093/pcp/pcs184.

Salek, R. M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., et al. (2013a). The MetaboLights repository: Curation challenges in metabolomics. [Research Support, Non-U.S. Gov't]. *Database,*. doi:10.1093/database/bat029.

Salek, R. M., Haug, K., & Steinbeck, C. (2013b). Dissemination of metabolomics results: Role of MetaboLights and COSMOS. (Editorial). *GigaScience, 2*(1), 8. doi:10.1186/2047-217X-2-8.

Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R., & Dunn, W. B. (2013c). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience, 2*(1), 13. doi:10.1186/2047-217X-2-13.

Sansone, S. A., Fan, T., Goodacre, R., Griffin, J. L., Hardy, N. W., Kaddurah-Daouk, R., et al. (2007). The metabolomics standards initiative. [Letter]. *Nature Biotechnology, 25*(8), 846–848. doi:10.1038/nbt0807-846b.

Schober, D., Wilson, M., Jacob., D., Moing, A., Mayer, G., Eisenacher, M., Salek, R.M., Neumann, S. (2014). Ontology usage in omics standards initiatives: Pros and cons of enriching XML data formats with controlled vocabulary terms. *Ontologies and Data in Life Sciences (ODLS2014)*, Freiburg im Breisgau, October 7–8, *IMISE-REPORT* Nr.1/2014, p. 36, from http://www.onto-med.de/obml/ws2014/odls2014report.pdf.

Scholz, M., & Fiehn, O. (2007). SetupX: A public study design database for metabolomic projects. [Research Support, N.I.H., Extramural]. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp 169–180.

Steinbeck, C., Conesa, P., Haug, K., Mahendraker, T., Williams, M., Maguire, E., et al. (2012). MetaboLights: Towards a new

COSMOS of metabolomics data management. *Metabolomics, 8*(5), 757–760. doi:10.1007/s11306-012-0462-0.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., et al. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics, 3*(3), 211–221. doi:10.1007/s11306-007-0082-2.

van der Greef, J., Hankemeier, T., & Mc Burney, R. N. (2006). Metabolomics-based systems biology and personalized medicine: Moving towards n = 1 clinical trials?. *Pharmacogenomics, 7*(7), 1087–1094.

van Ommen, B., Bouwman, J., Dragsted, L. O., Drevon, C. A., Elliott, R., de Groot, P., et al. (2010). Challenges of molecular nutrition research 6: The nutritional phenotype database to store, share and evaluate nutritional systems biology studies. *Genes and Nutrition, 5*(3), 189–203. doi:10.1007/s12263-010-0167-9.

Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., et al. (2013). The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Molecular and Cellular Proteomics : MCP, 12*(8), 2332–2340. doi:10.1074/mcp.O113.028506.

Werf, M. T., Takors, R., Smedsgaard, J. R., Nielsen, J., Ferenci, T., Portais, J., et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Microbial and in vitro biology experiments. *Metabolomics, 3*(3), 189–194. doi:10.1007/s11306-007-0080-4.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0: The human metabolome database in 2013. [Research Support, Non-U.S. Gov't]. *Nucleic Acids Research, 41*, D801–D807. doi:10.1093/nar/gks1065.