nature biotechnology

PERSPECTIVE

A proposed framework for the description of plant metabolomics experiments and their results

Helen Jenkins[1], Nigel Hardy[1], Manfred Beckmann[2], John Draper[2], Aileen R Smith[2], Janet Taylor[1,21], Oliver Fiehn[3], Royston Goodacre[4], Raoul J Bino[5,6], Robert Hall[5], Joachim Kopka[3], Geoffrey A Lane[7], B Markus Lange[8], Jang R Liu[9], Pedro Mendes[10], Basil J Nikolau[11], Stephen G Oliver[12], Norman W Paton[13], Sue Rhee[14], Ute Roessner-Tunali[15], Kazuki Saito[16], Jørn Smedsgaard[17], Lloyd W Sumner[18], Trevor Wang[19], Sean Walsh[19], Eve Syrkin Wurtele[20] & Douglas B Kell[4]

The study of the metabolite complement of biological samples, known as metabolomics, is creating large amounts of data, and support for handling these data sets is required to facilitate meaningful analyses that will answer biological questions. We present a data model for plant metabolomics known as ArMet (architecture for metabolomics). It encompasses the entire experimental time line from experiment definition and description of biological source material, through sample growth and preparation to the results of chemical analysis. Such formal data descriptions, which specify the full experimental context, enable principled comparison of data sets, allow proper interpretation of experimental results, permit the repetition of experiments and provide a basis for the design of systems for data storage and transmission. The current design and example implementations are freely available (http://www.armet.org/). We seek to advance discussion and community adoption of a standard for metabolomics, which would promote principled collection, storage and transmission of experiment data.

Functional genomic research is generating large amounts of data. These must be transmitted, stored safely with adequate curation and made available in convenient and supportive ways for statistical analyses and data mining. To do this, well-designed data standards[1] are required.

The DNA microarray community has developed MIAME[2] (minimum information about a microarray experiment) as a definition of what should be recorded for a transcriptome experiment. MAGE[3] (microarray gene expression) is an associated data exchange format. Relational database (RDB) implementations of MIAME/MAGE are in operation[4,5]. The structure has been widely accepted and MIAME compliance is now required by many journals[6–8].

PEDRo[9] (proteomics experiment data repository) is a UML[10] (unified modeling language) model, with both RDB and XML (extensible markup language; http://www.w3.org/TR/xml11/) implementations, which seeks to meet the requirements of the proteomics field. PEDRo was adopted by the Human Proteome Organization's (HUPO) Proteomics Standards Initiative (PSI)[11] as the basis for their own object model (PSI-OM[12]; http://psidev.sourceforge.net/gps/).

'Metabolomics' seeks to estimate the complement of metabolites present in a sample (for a more detailed explanation, see refs. 13–24). Several factors, including the dynamic nature of metabolites and the range and complexity of estimation techniques, mean that metabolomics (in common with transcriptomics and proteomics) requires significant data handling support[18,25,26]. Although reference databases of chemical information (e.g., refs. 27–29) and ontologies (e.g., ref. 30 and http://www.plantontology.org/) exist to support metabolomics, a metabolomics-specific standard equivalent to MIAME, PEDRo or PSI-OM, for describing metabolite complements in their experimental context is not currently available[18]. Such a description is needed to facilitate proper interpretation of experiment results, data set transfer, laboratory interoperability, meaningful comparison of data sets and replication of experiments.

The need for metabolomics data standards is evident. Currently, the Metabolomics Society (a group founded from academia, government

[1]Department of Computer Science and [2]Institute of Biological Sciences, University of Wales, Penglais, Aberystwyth, Ceredigion, Wales, UK, SY23 3DB. [3]Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany. [4]School of Chemistry, The University of Manchester, PO Box 88, Manchester, UK, M60 1QD. [5]Plant Research International, POB 16, 6700 AA Wageningen, The Netherlands. [6]Centre for BioSystems Genomics, POB 98, 6700 AB Wageningen. [7]AgResearch Grasslands Research Centre, Private Bag 11008, Palmerston North, New Zealand. [8]Institute of Biological Chemistry and Center for Integrated Biotechnology, Washington State University, Pullman, Washington 99164-6340, USA. [9]Plant Cell Biotechnology Laboratory of KRIBB and NRL of Eugentech Inc., PO Box 115, Yuseong, Daejeon, 305-600 Korea. [10]Virginia Bioinformatics Institute, Virginia Tech, 1880 Pratt Drive, Blacksburg, Virginia 24061, USA. [11]Center for Designer Crops, W.M. Keck Metabolomics Research Laboratory, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, USA. [12]School of Biological Sciences and [13]Department of Computer Science, University of Manchester, Oxford Road, Manchester, UK, M13 9PL. [14]The Arabidopsis Information Resource, Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA. [15]Australian Centre for Plant Functional Genomics, School of Botany, University of Melbourne, 3010 Victoria, Australia. [16]Graduate School of Pharmaceutical Sciences, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba 263-8522, Japan. [17]Technical University of Denmark, BioCentrum-DTU, Søltofts Plads Building 221, Kgs Lyngby, 2800 Denmark. [18]The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401, USA. [19]John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH. [20]Department of Genetics, Development and Cell Biology, 441 Bessey Hall, Iowa State University, Ames, Iowa 50011, USA. [21]Present address: IGER, Plas Gogerddan, Aberystwyth, Ceredigion, Wales, UK, SY23 3EB. Correspondence should be addressed to N.W.H. (e-mail: *nwh@aber.ac.uk*).

Published online 6 December; doi:10.1038/nbt1041

Box 1 A brief UML tutorial

A UML class is a model element that represents a concept from the domain being modeled. Classes can be viewed as templates, for example, a 'Genotype' class provides a template for describing plant genotypes; a system may then use multiple instances of this class to describe different genotypes. Classes are depicted using segmented boxes (**Fig. 2**). Individual data items (attributes), used to describe a concept, are listed inside the box. The class 'Environment' contains attributes 'growthLocation,' 'dateTime' and 'protocolName' (**Fig. 2a**).

A UML package is a collection of related model elements. Packages are depicted using the notation in **Figure 1**. For example, 'BiologicalSource' contains 'Genotype' and 'Source.' Dashed arrows between packages indicate dependencies between model elements. For example, there is a model element in package 'BiologicalSource' that is dependent on an element in 'Growth.' This implies that if data in the element in 'Growth' were altered, data in the dependent element in 'BiologicalSource' may also require alteration.

Lines between classes represent associations between them. Annotations to these associations represent multiplicity and participation constraints. The annotation '1..*' beside 'Explant' (**Fig. 2b**) indicates that each instance of a 'Sample' should be associated with at least one and possibly many (*) instances of 'Explant.' Examples of other constraints are '0..1' (a single, optional association), '1..1' (a single, required association), '0..*' (zero or many associations) and '1..20' (a minimum of 1 to a maximum of 20 associations). Associations may be annotated with association classes, for example 'BulkExplant' (**Fig. 2d**), to describe an association more fully.

ArMet employs two special types of association: containment and specialization. A containment association exists between 'Treatment' and 'Environment' (**Fig. 2a**). The class at the diamond end of the association is the container, that is, a 'Treatment' contains (comprises) one (or more) 'Environment' class(es).

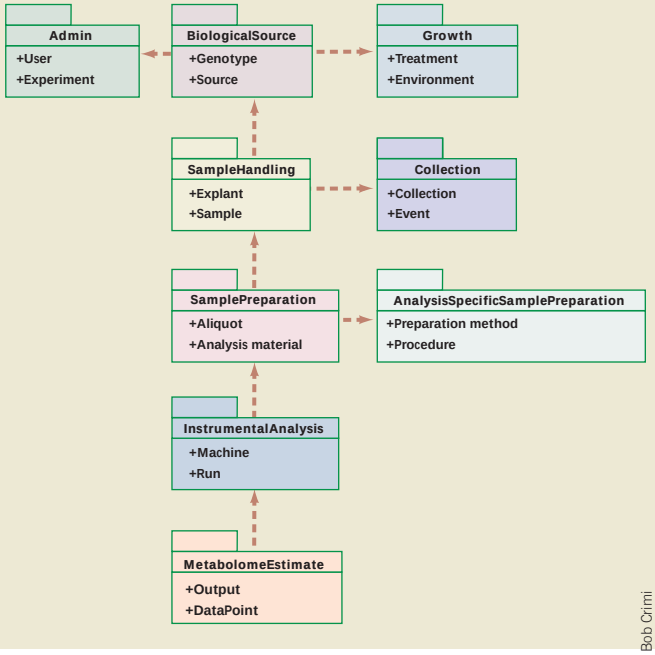
Specialization describes class inheritance. 'ControlledGrowth' and 'FieldGrowth' are specializations (subclasses) of the more general 'Environment' class (the superclass) (**Fig. 2c**). A subclass inherits the attributes of its superclass and may include additional attributes to describe the specialization. **Figures 2c,d** extend the core data definitions (superclasses) in **Figures 2a,b**. To indicate this relationship the names of the classes that they extend are included.

Specializations may be constrained to describe more fully the relationship between superclass and subclasses. All of the specializations used in this paper are 'mandatory-or,' meaning that each instance of the superclass must be associated with precisely one instance of a subclass.

and industry, which was announced in March 2004) aims to address this issue (http://metabolomicssociety.org/). A standard metabonomics reporting structure (SMRS) is also being developed by a group headed by Imperial College, London (J. Lindon, personal communication). 'Metabonomics' is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"[31] These groups have yet to publish. The need for data standards for plant metabolomics is also discussed regularly at congresses organized by the International Committee for Plant Metabolomics (http://www.metabolomics.nl/) and recently, Bino *et al*[32]. have suggested a checklist of the information necessary to provide context for metabolomics data that is to be published, which they have called MIAMET (minimum information on a metabolomics experiment). MIAMET represents a first positive step in the direction of a standard, but does not provide the complete formal data description of the specific required data items necessary for the development of supportive data handling systems.

In this article, we present ArMet as a contribution to the establishment of standards in the field. ArMet is a complete formal data description for plant metabolomics that supports not only data sets but also the necessary metadata (data about the data) to provide experimental context. The structure provides a framework that may be used to organize the community-driven process of development and enhancement to produce a widely accepted standard.

To encourage discussion and community adoption of a data standard for metabolomics, the ArMet data model is freely available (see **Supplementary Notes** 'ArMet design' online). We welcome feedback and will coordinate future development and dissemination (see http://www.armet.org/).



Bob Crimi

Figure 1 The nine ArMet components and their interdependencies describing the key concepts of a metabolomics experiment modeled using standard UML packages. For explanation of the notation see **Box 1**. Each component contains core data items (modeled as attributes of the classes listed within the packages), which serve as a minimal description for each component.

ArMet

To capture the required metadata, ArMet encompasses the entire experiment time line and organizes it into nine subunits termed 'com-

Table 1 The components of ArMet

Component	Data supported by component
'Admin'	Experiment description and contact details
'BiologicalSource'	Genotype, provenance and identification data for items of biological material
'Growth'	Description of the environments in which the biological material developed
'Collection'	Procedures followed for gathering samples from items of biological material
'SampleHandling'	Handling and storage procedures following collection
'SamplePreparation'	Protocols for preparing samples for presentation to analytical instruments
'AnalysisSpecificSamplePreparation'	Protocols specific to particular analytical technologies
'InstrumentalAnalysis'	Process description of the chemical analysis of samples, including descriptions of analytical instruments and their operational parameters, quality control protocols and references to archive copies of raw results
'MetabolomeEstimate'	The output from the analytical instruments after it has been processed from raw data to produce a metabolome description and metadata about the processing

ponents.' We have used the widely adopted modeling standard UML to represent ArMet. ArMet's components are modeled as the UML packages (see **Box 1**) shown in **Figure 1**. UML classes are used to model the key concepts for each component and are listed within the packages. **Table 1** provides a brief description of the data that each component is designed to support.

The experiment data (data about an experiment) are specified by way of a core set of data items for each component. These core items (modeled as attributes of the classes within the packages) are relevant to all plant metabolomics experiments and serve as a minimal description for each component. **Figure 2a**,**b** depicts two components as examples (see **Supplementary Notes** 'ArMet design' online for the full model). These core data provide a basis for cross-laboratory data exchange and data mining. However, as the examples show, the number of experiment details contained in the core is currently quite limited.

Metabolomics is a young discipline and the methodologies employed are varied and rapidly evolving. The specification of a more detailed core would be presumptuous at this stage. The component-based approach adopted for ArMet provides a basis for the definition of extensions to the core data to support the detailed requirements of the range of methodologies employed by different projects, experiments and laboratories. The core data have been designed to provide an abstraction of these detailed requirements by modeling only key concepts for each experiment area and relying on references to externally documented protocols. Therefore, although sparse, it is useful, coherent and logically complete.

Several extensions (termed 'subcomponents') to the core data will be required for each component to describe, in detail, the range of experiment methodologies. As examples, **Figure 2c** shows a subcomponent of the 'Growth' component to support a project that involves pre-treatment of seeds before sowing in either a field or controlled growth environment, whereas **Figure 2d** shows a subcomponent of the 'SampleHandling' component to support an experiment in which four 'Explants' (pieces of plant material taken during a harvest) are bulked and then divided to produce two 'Samples,' which are then transported and stored before preparation for analysis.

Any application-specific data model that can demonstrate support for the core data model, and any system that can export the same, may be viewed as a subcomponent of ArMet. For specific subcomponent design and implementation this support could be effected through the use of class inheritance or any other implementation-specific paradigm.

As with the core definition, we ask developers to make their subcomponents freely available for reuse by other parties. Thereby, any data set that conforms to ArMet will be comparable with any other conforming data set at the core level and will be further compatible, at a greater level of detail, with other data sets created using the same experimental methods and described using a readily available subcomponent. This approach is designed to yield the following benefits:

- movement towards community standardization of experiment methodologies and their description;
- the ability to assess and make a decision about the comparability of a data set against preexisting ones;

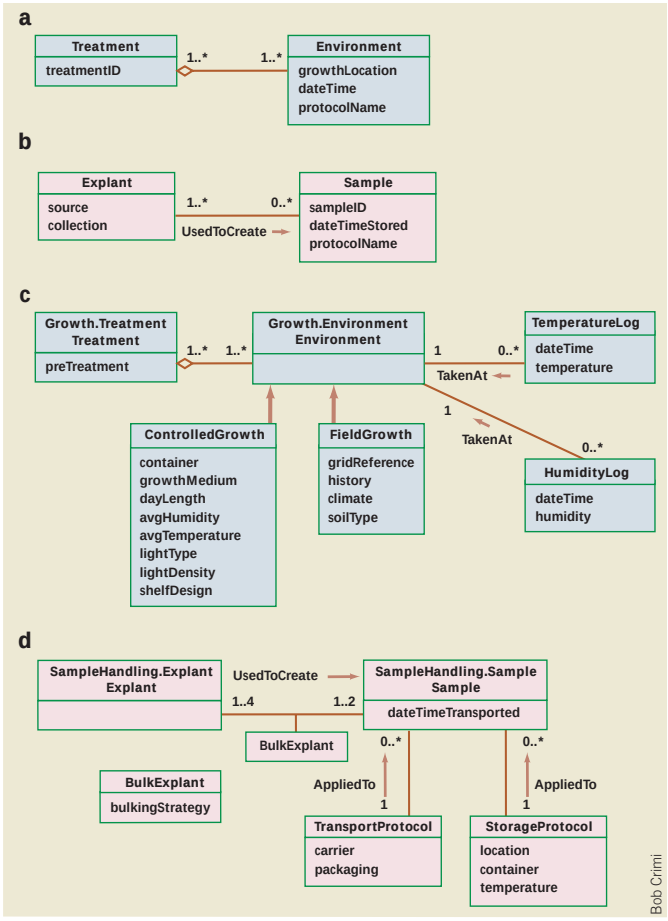


Figure 2 ArMet core components and detailed subcomponents. (**a**) The core definition of the 'Growth' component. (**b**) The core definition of the 'SampleHandling' component. (**c**) An example subcomponent of the 'Growth' component. (**d**) An example subcomponent of the 'SampleHandling' component. For explanation of the notation, see **Box 1**.

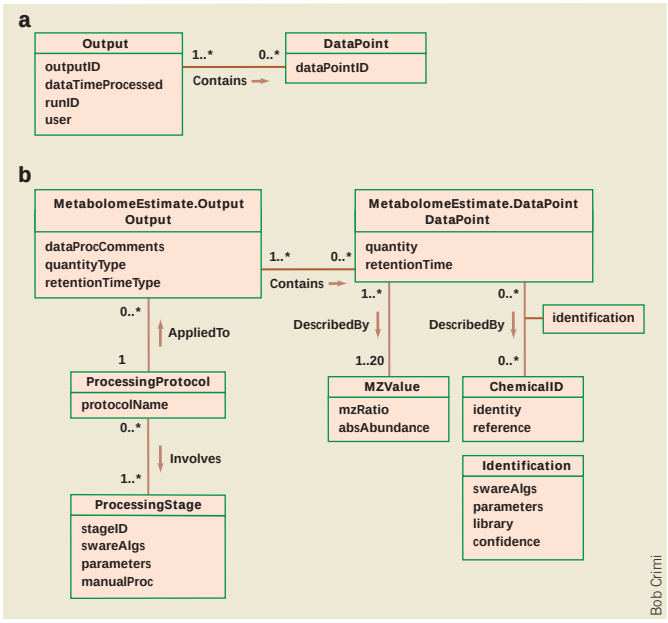


Figure 3 The 'MetabolomeEstimate' component. (**a**) The core definition. (**b**) An example subcomponent to support metabolomics data, as produced by GC-MS. For explanation of the notation, see **Box 1**.

• a basis for future enhancement of the core data to include detail that is common to all of the subcomponents of a component.

A parallel activity to the development of subcomponents is the development and enhancement of publicly available controlled vocabularies to provide control over the values of attributes. Managed, community-led development of such vocabularies will lead to global use, which will in turn result in enhanced data set comparability.

ArMet requires the use of controlled vocabularies but does not specify which vocabularies should be used. Instead, the attributes that are controlled by vocabularies are designed to support not just data values, but also authorities that describe their provenance, for example, a recognized ontology (e.g., ref. 30 and http://www.plantontology.org/) or the local list of terms for a laboratory (e.g., **Supplementary Notes** 'Controlled vocabulary examples' online). Thus, the authorities also imply the status of a particular value.

Of the nine ArMet components, the first eight define the metadata required to place metabolomics data sets in context. The ninth component, 'MetabolomeEstimate,' is designed to support the data sets themselves as metabolome descriptions produced by processing the raw output from analytical instruments. Such descriptions support detailed analysis and comparison of experiment results. The structure of these descriptions is not only dependent on the type of question being asked and, therefore, the analytical approach that is being taken, but also on the nature of the metabolome descriptions produced from the output of the many analytical technologies used in metabolomics. For ArMet, we have adopted the characterization of four different analytical approaches (targeted analysis, metabolite profiling, metabolomics and fingerprinting) described by Fiehn[14]. Subcomponents of the 'MetabolomeEstimate' component will therefore be required to support these four descriptions.

As an example, **Figure 3** depicts a subcomponent to support metabolomics (q.v.) data sets produced from gas chromatography mass spectrometry (GC-MS) output. Following the principle of public availability, the current version of its design is freely available (see **Supplementary Notes** 'ArMet subcomponent design: GC-MS' online).

To produce a quantified list of detected compounds from raw GC-MS output noise removal, detection of chromatographic peaks, peak deconvolution and peak quantification are performed, details of which must be provided as metadata. In **Figure 3b,** this is modeled as a named 'ProcessingProtocol' that comprises a number of 'ProcessingStages'. Complete data sets, with their processing, are described by the extensions to the core 'Output' attributes, whereas the detected compounds they contain are modeled by the extensions to the core 'DataPoint' attributes and the 'ChemicalID,' 'MZValue' and 'Identification' classes. The process of chemical identification involves mass spectral comparison of experimental GC-MS spectra with those of authentic standards contained in mass spectral libraries. Stein *et al*[33]. evaluated three algorithms and two distance measures used to perform mass spectral comparison and reported a best accuracy of 74.9% correct identifications. A follow-up study by McLafferty *et al*[34]. achieved a best accuracy of 77%. These levels of accuracy show that the identifications that result from automatic mass spectral comparison are tentative at best. **Figure 3b**, therefore, supports multiple candidate chemical identities for each peak, each annotated with metadata describing the process of spectral comparison. Note that the 'ChemicalID' class contains an 'identify' 'reference' pair. The 'identity' is a chemical name or other identifier, such as an International Union of Pure and Applied Chemistry (IUPAC, Research Triangle Park, NC, USA)–National Institute of Standards and Technology (NIST, Bethesda, MD, US) Chemical Identifier (INChI) (http://www.iupac.org/projects/2000/2000-025-1-800.html), or a chemical ontology class in the case of partially identified components, whereas the 'reference' provides a link to an entry in an external library of structure information for the chemical or chemical ontology class (e.g., refs. 27–29 and http://www.ebi.ac.uk/chebi/).

As a compound in a metabolomics data set may be associated with zero or more chemical identities, this information cannot be used as its unique label. Therefore, this subcomponent includes quantity, retention and spectral data for each peak as modeled by the 'DataPoint' and 'MZValue' classes. A recent study[34] has looked at the number of mass spectral peaks required for successful comparison of spectra in the context of library identifications, finding that 15 peaks were 87% as effective as 150 peaks, whereas 18 peaks were 97% as effective. On this basis, and to yield the benefit of reduced data storage requirements in database implementations, the subcomponent supports a representative mass spectrum with a maximum of 20 peaks for each compound.

In practice, different personnel working in different contexts will provide and retrieve data. A benefit of the component-based approach is that it lends itself to the development of customizations to support particular functions, that is, selections of components implemented as independent systems. Screenshots from an example customization may be found in the **Supplementary Notes** 'ArMet example customization' online.

ArMet and other standards

ArMet and MIAMET can be viewed as complementary proposals. Where MIAMET provides a checklist of information that should be described in metabolomics publications, ArMet provides a formal data definition to support automatic data set comparison and mining and the development of systems for data storage and exchange. Collaborative future development will further ensure that a MIAMET compliant experiment description would imply ArMet compliance and vice versa.

MIAME, PEDRo, PSI-OM and ArMet have been developed independently to accommodate the specific requirements of their fields. Metabolomics analyses will be performed in conjunction with those on

the transcriptome and proteome in integrative experiments and therefore there is a urgent need to create a common standard for experiment descriptions. However, it is imperative that this common description contains the metadata necessary to fully evaluate the different 'omic' analyses that are performed on samples from a single experiment. A recent initiative in this area has created SysBio-OM[35], a model that integrates MAGE-OM and PEDRo and provides support for description of metabolomics experiments in toxicogenomics.

ArMet works in synergy with laboratory information management systems (LIMS) and other existing standards. In appropriate circumstances, ArMet could provide a design for customization of a LIMS to support the metabolomics process, which, therefore, would become an implementation vehicle for ArMet. Similarly existing standards for analytical data storage and transmission—for example, the IUPAC Working Party on Spectroscopic Data Standards standard JCAMP-DX[36], the American Society for Testing Materials' (Conshohocken, PA, USA) Analytical Data Interchange format (ANDI, http://pubs.acs.org/hotartcl/tcaw/98/may/stan.html), the climate data format standard (netCDF, http://www.unidata.ucar.edu/packages/netcdf/) and analytical information markup language (AnIML, http://animl.sourceforge.net/)—could be viewed as part of the definition of subcomponents for the 'InstrumentalAnalysis' component to satisfy the requirement for access to the analytical machine output.

Application, current status and future development

We present ArMet, a freely available data model for describing plant metabolomics experiments. An XMI (XML metadata interchange) version of the UML model is available (**Supplementary Notes** 'ArMet design XMI' online) together with example implementations in SQL (structured query language) and XML (**Supplementary Notes** 'ArMet core SQL' and 'ArMet core XML' online). Multiple implementation structures for use in different settings are derived from models, such as ArMet. Presentation to users can be varied and targeted. Web-based forms as front ends to remote database access (RDB) implementations may be appropriate for low-volume data input and simple querying (see **Supplementary Notes** 'ArMet on-line forms access' online). Open database connectivity (ODBC, http://msdn.microsoft.com/library/) and related standards support programmatic access to RDB implementations for producing specialized programs. Offline database implementations (typically with a well recognized 'look-and-feel') can support workplace use (see **Supplementary Notes** 'ArMet example customization' online). Microsoft Excel (Version 2003, Microsoft, Seattle, WA, USA) supports XML schema mapping, permitting its use as an ArMet compliant data entry tool, whereas the PEDRo data capture tool[9] is an example of a more specialized XML schema-driven application.

ArMet compliant databases and data handling systems are in use on two major projects involving a complete set of subcomponents to support experiments with *Arabidopsis thaliana* and *Solanum tuberosum*. Work is ongoing to design additional metadata subcomponents for alternative experimental methodologies and 'Fingerprinting' subcomponents for Fourier transform-infrared (FT-IR) spectrometry[37] and mass profiling by direct-injection electrospray ionization-mass spectrometry (ESI-MS)[38–40]. Also in progress is adaptation of ArMet for use in microbial metabolomics.

We are not aware of any existing system or data definition that is being used as an agreed and freely available standard by the plant metabolomics community. We intend that ArMet will provide input to the community discussion leading to an agreed upon standard. We welcome feedback on the current design and suggestions for new subcomponents, coordination of which will be carried out via our website (http://www.armet.org/). As a first step, we plan to hold a workshop to discuss data standards for metabolomics in general and ArMet development in particular (see the website for further information and contact details).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the United Kingdom Food Standards Agency (under the G02006 project), the United Kingdom Biotechnology and Biological Sciences Research Council (particularly under the HiMet project) and the United Kingdom Engineering and Physical Sciences Research Council for support of their work in metabolomics. We would also like to thank personnel at the United Kingdom Institute of Grassland and Environmental Research and delegates at the 1st, 2nd and 3rd International Plant Metabolomics Conferences for many useful discussions.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Quackenbush, J. Data standards for 'omic' science. *Nat. Biotechnol.* **22**, 613–614 (2004).
2. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
3. Spellman, P.T. *et al.* Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3**, research0046.0041–0046.0049 (2002).
4. Brazma, A. *et al.* ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
5. Killion, P.J., Sherlock, G. & Iyer, V.R. The Longhorn Array Database (LAD): An open-source, MIAME compliant implementation of the Stanford Microarray database (SMD). *BMC Bioinformatics* **4**, 32 (2003). http://biomedcentral.com/1471-2105/4
6. Editorial. Microarray standards at last. *Nature* **419**, 323 (2002).
7. Glueck, S.B. & Dzau, V.J. Our new requirement for MIAME standards. *Physiol. Genomics* **13**, 1–2 (2003).
8. Oliver, S. On the MIAME standards and central repositories, of microarray. *Comp. Funct. Genomics* **4**, 1 (2003).
9. Taylor, C.F. *et al.* A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* **21**, 247–254 (2003).
10. Booch, G., Rumbaugh, J. & Jacobson, I. *The Unified Modeling Language User Guide* (Addison-Wesley, Reading, MA, 1999).
11. Orchard, S., Hermjakob, H. & Apweiler, R. The proteomics standards initiative. *Proteomics* **3**, 1374–1376 (2003).
12. Orchard, S. *et al.* Common interchange of standards for proteomics data: public availability of tools and schema. *Proteomics* **4**, 490–491 (2004).
13. Fiehn, O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics* **2**, 155–168 (2001).
14. Fiehn, O. Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
15. Fiehn, O. & Weckwerth, W. Deciphering metabolic networks. *Eur. J. Biochem.* **270**, 579–588 (2003).
16. Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. & Kell, D.B. Metabolomics by numbers:acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252 (2004).
17. Harrigan, G.G. & Goodacre, R. (eds.) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (Kluwer Academic Publishers, Boston, 2003).
18. Mendes, P. Emerging bioinformatics for the metabolome. *Brief. Bioinform.* **3**, 134–145 (2002).
19. Oliver, S.G., Winson, M.K., Kell, D.B. & Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**, 373–378 (1998).
20. Raamsdonk, L.M. *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* **19**, 45–50 (2001).
21. Roessner, U. *et al.* Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29 (2001).
22. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23**, 131–142 (2000).
23. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836 (2003).
24. Weckwerth, W. Metabolomics in systems biology. *Annu. Rev. Plant Biol.* **54**, 669–689 (2003).
25. Hall, R. *et al.* Plant metabolomics: the missing link in functional genomics strategies (Meeting report). *Plant Cell* **14**, 1437–1440 (2002).
26. van der Greef, J., van der Heiiden, R. & Verheij, E.R. The role of mass spectrometry in system biology: data processing and identification strategies in metabolomics. in *Advances in Mass Spectrometry*, vol. 16. (eds. Ashcroft, A.E., Brenton, G. & Monaghan, J.J.) 145–165 (Elsevier, Amsterdam, 2004).
27. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
28. Krieger, C.J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).

29. Mueller, L.A., Zhang, P. & Rhee, S.Y. AraCyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiol.* **132**, 453–460 (2003).
30. Harris, M.A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32 Special Issue**, D258–D261 (2004).
31. Nicholson, J.K., Lindon, J.C. & Homes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**, 1181–1189 (1999).
32. Bino, R.J. *et al.* Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9**, 418–425 (2004).
33. Stein, S.E. & Scott, D.R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
34. McLafferty, F.W., Zhang, M-Y., Stauffer, D.B. & Loh, S.Y. Comparison of algorithms and databases for matching unknown mass spectra. *J. Am. Soc. Mass Spectrom.* **9**, 92–95 (1998).
35. Xirasagar, S. *et al.* CEBS object model for systems biology data, SysBio-OM. *Bioinformatics* **20**, 2004–2015 (2004).
36. Lampen, P. *et al.* An extension to the JCAMP-DX standard file format, JCAMP-DX V.5.01 (IUPAC Recommendations 1999). *Pure Appl. Chem.* **71**, 1549–1556 (1999).
37. Griffiths, P.R. & de Haseth, J.A. *Fourier Transform Infrared Spectrometry*, vol. 83 (John Wiley & Sons, Chichester, England, 1986).
38. Allen, J.K. *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696 (2003).
39. Smedsgaard, J. *Terverticillate penicillia* studied by direct electrospray mass spectrometric profiling of crude extracts: II. Database and identification. *Biochemical Systematics and Ecology* **25**, 65–71 (1997).
40. Smedsgaard, J. & Frisvad, J.C. Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts. *J Microbiol. Methods* **25**, 5–17 (1996).