nature biotechnology

PERSPECTIVE

The minimum information about a proteomics experiment (MIAPE)

Chris F Taylor[1,2], Norman W Paton[1,3], Kathryn S Lilley[1,4], Pierre-Alain Binz[1,5,6], Randall K Julian Jr[1,7], Andrew R Jones[1,3], Weimin Zhu[1,2], Rolf Apweiler[1,2], Ruedi Aebersold[1,8], Eric W Deutsch[1,9], Michael J Dunn[10], Albert J R Heck[11], Alexander Leitner[12], Marcus Macht[13], Matthias Mann[14], Lennart Martens[1,2], Thomas A Neubert[15], Scott D Patterson[16], Peipei Ping[17], Sean L Seymour[1,18], Puneet Souda[19], Akira Tsugita[20], Joel Vandekerckhove[21], Thomas M Vondriska[22], Julian P Whitelegge[19], Marc R Wilkins[23], Ioannnis Xenarios[24], John R Yates III[25] & Henning Hermjakob[1,2]

Both the generation and the analysis of proteomics data are now widespread, and high-throughput approaches are commonplace. Protocols continue to increase in complexity as methods and technologies evolve and diversify. To encourage the standardized collection, integration, storage and dissemination of proteomics data, the Human Proteome Organization's Proteomics Standards Initiative develops guidance modules for reporting the use of techniques such as gel electrophoresis and mass spectrometry. This paper describes the processes and principles underpinning the development of these modules; discusses the ramifications for various interest groups such as experimentalists, funders, publishers and the private sector; addresses the issue of overlap with other reporting guidelines; and highlights the criticality of appropriate tools and resources in enabling 'MIAPE-compliant' reporting.

[1]The HUPO Proteomics Standards Initiative. [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [3]School of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK. [4]Cambridge Centre for Proteomics, University of Cambridge, Cambridge, Cambridgeshire CB2 1QW, UK. [5]Swiss Institute of Bioinformatics, Geneva, Switzerland. [6]GeneBio SA, Geneva, Switzerland. [7]Indigo BioSystems, Indianapolis, Indiana, USA. [8]Swiss Federal Institute of Technology and University of Zurich, Zurich, Switzerland. [9]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA. [10]Proteome Research Centre, Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. [11]Utrecht University and Netherlands Proteomics Centre, Sorbonnelaan 16, 3584 CA Utrecht, The Netherlands. [12]Department of Analytical Chemistry and Food Chemistry, University of Vienna, 1090 Vienna, Austria. [13]Bruker Daltonik GmbH, Bremen, Germany. [14]Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, D-82152 Martinsried, Germany. [15]Skirball Institute of Biomolecular Medicine and Department of Pharmacology, New York University School of Medicine, New York, New York 10016, USA. [16]Molecular Sciences, Amgen Inc., Thousand Oaks, California, USA. [17]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, California 90075, USA. [18]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404, USA. [19]The Pasarow Mass Spectrometry Laboratory, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, and The Molecular Biology Institute, and The Brain Research Institute, University of California, Los Angeles, California 90095, USA. [20]Proteomics Research Laboratory, Tokyo Rikakikai Co., Tsukuba, Japan. [21]Department of Biochemistry, Ghent University and Department of Medical Protein Research, Flanders Institute for Biotechnology, B-9000 Ghent, Belgium. [22]Department of Medicine (Cardiology) and Department of Anesthesiology, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA. [23]School of Biotechnology and Biomolecular Sciences, University of New South Wales, New South Wales 2052, Australia. [24]Serono Pharmaceutical Research Institute, 14, Chemin des Aulx, 1228 Plan-les-Ouates, Geneva, Switzerland. [25]The Scripps Research Institute, 10550 North Torrey Pines Road, SR11, La Jolla, California 92037, USA. Correspondence should be addressed to C.F.T. chris.taylor@ebi.ac.uk.

Published online 8 August 2007; doi:10.1038/1329

The burgeoning of public repositories of experimentally derived genomic[1,2] and transcriptomic[3–5] data and the concomitant increases in protein sequence[6] and integrated[7,8] databases are well documented, providing to the scientific community a rich set of resources that has enabled more rapid advances in the understanding of gene function than would otherwise have been possible. Practitioners of proteomics—the direct study of sets of proteins occurring together in particular parts or states of biological entities (that is, proteomes)—are now beginning to share data through communal resources[9–14]. This raises challenging issues: for example, should datasets contain a level of description beyond what is found in the 'average' published paper; how can experimentalists be supported in collecting and transmitting such information; and what might the benefits of more systematic reporting be, for individuals and the wider community?

The role of experimental metadata

There are many different subsets of the 'total' (that is, all parts and all states) proteome of an organism, just as there are many related patterns of gene transcription, each distinguished both by cell type and by condition. Furthermore, protein identifications obtained by mass spectrometry, for example, are dependent on the separation technologies employed, the particular mass spectrometer, and the protein identification tool and database with which identities were assigned. To understand an analysis, perform comparisons between datasets, or derive statistics from their aggregation, it is crucial to understand both the biological and the methodological contexts. Inadequate description can allow inappropriate experimental design and random or systematic errors to go undetected. Conversely, confidence in data and data analysis can be increased by, for example, reporting the performance of appropriate calibration runs[15] or using power analyses to support the particular study design[16].

Proteomics data should therefore ideally be accompanied by contextualizing 'metadata' (essentially 'data about the data'), making explicit both where samples came from and how analyses were performed. To

that end, the Proteomics Standards Initiative (PSI; http://www.psidev.info)[17,18] develops guidance documents specifying the data and metadata that should be collected from various proteomics workflows (**Box 1**), known collectively as the "minimum information about a proteomics experiment" (MIAPE) guidelines.

There are several precedents in the biomedical sciences for such reporting prescriptions. The "minimum information about a microarray experiment" (MIAME) guidelines[19], which deal specifically with transcriptomics data, has become an accepted community standard; the original paper had been cited in >1,100 published papers (source: Google Scholar), many of which describe MIAME-compliant software development. Several major journals now require that papers reporting transcriptomics experiments are accompanied by the MIAME-prescribed set of metadata, either as supplementary information supplied to the journal or in the form of a database submission, as a prerequisite for publication[20–22]. MIAME demonstrably facilitates the reuse of data from 'compliant' repositories such as ArrayExpress[4] in new research[23,24], although there are still issues with respect to the willingness[25] or ability[26,27] of authors to submit fully compliant datasets. Other examples of established reporting requirements include CONSORT[28] (randomized controlled trials), QUORUM[29] (meta-analyses of randomized, controlled trials), STARD[30] (assessment of the accuracy of diagnostic tests), REMARK[31] (tumor-marker prognostic studies) and the Clinical Data Interchange Standards Consortium's SEND[32] (non-clinical toxicology). Studies of some of these standards[33,34] have found that they increase the quality of reports without incurring a disproportionate cost for users. Other sets of reporting requirements are under development across the biosciences. For example, some fifteen ongoing projects (including MIAPE) are now listed on the website of a central registry of reporting guidelines, the "minimal information for biological and biomedical investigations" (MIBBI)[35] (http://mibbi.sf.net)—a clear demonstration of the appetite in the bioscience community for the regularization of reporting.

MIAPE: principles, process and product

Proteomics workflows frequently consist of sample collection and processing, separation by liquid chromatography or some form of electrophoresis, examination of separands by mass spectrometry, identity assignment, and absolute or relative quantification of proteins through bioinformatic analysis of the mass spectra generated[36,37]. However, there are many technologies available, each enabling the analysis of different (although frequently overlapping) 'subproteomes' differentiated by mass, surface charge or localization in time or space. These various technologies, many of which are still evolving, produce a dauntingly diverse array of data. Additionally, the future holds the promise of many new technologies: improved prefractionation and depletion techniques, evolving array- and bead-based technologies, and new types of mass spectrometer components. The volume and complexity of (meta)data potentially available from most proteomics experiments makes the issue of what to keep and what to discard an important one. To guide specific decisions on the data and metadata that should be required by each MIAPE module, we employ two very general criteria:

1. Sufficiency. The MIAPE guidelines should require sufficient information about a dataset and its experimental context to allow a reader to understand and critically evaluate the interpretation and conclusions, and to support their experimental corroboration.
2. Practicability. Achieving compliance with MIAPE should not be so burdensome as to prohibit its widespread use.

The resulting guidelines, about which some general information is provided in **Box 2**, are reminiscent of the schema underlying a laboratory information management system (LIMS) in that they require not only data but also metadata that are relevant to the discovery and interpretation of the results.

The development of MIAPE documents

This document exists to make explicit the scope, purpose and manner of use of the modular MIAPE guidelines that accompany it, and to lay out the principles underlying module production. This document should be stable, as the principles described herein are technique independent and should therefore remain valid for the foreseeable future. The associated modules will be more labile; they will both evolve over time and increase in number, to track changes in experimental techniques. We recognize, however, that it is just as desirable to have stable reporting guidelines as it is to have stable data formats. There will, therefore, be a period of at least one year between successive versions of any one module.

Initial versions of all modules are generated through the PSI, drawing on community expertise gathered at PSI and other meetings, through the discussion lists, and through direct interactions with experts in the relevant techniques. These 'candidate' documents then enter the PSI document process, where they are subject to public comment and (anonymous) review, including tests by experimentalists in the lab of the feasibility of collecting the required information, and are finally submitted for publication in a peer-reviewed journal. Throughout their development and after their release, modules can be discussed in a number of ways: (i) on e-mail discussion lists (details on the website, http://www.psidev.info); (ii) on the scheduled open teleconferences held by the working groups; and (iii) by attending one of the free annual meetings. Conflicts arising in relation to extant or emerging modules should be resolved through the discussion lists and teleconferences where possible. If consensus cannot be achieved informally then issues will be resolved at the next annual meeting, by vote if necessary.

The MIAPE modules

Each MIAPE module relates to a particular technology or group of technologies. The first modules that will be deployed are briefly described in **Box 1**. The most recent versions of all modules are available from the MIAPE home page on the PSI website (http://www.psidev.info/miape). Where a particular (newly developed or legacy) technique is not yet covered by MIAPE, authors should attempt to match existing modules in terms of depth of coverage; ideally the expansion of an existing module or the development of a new module for that technique should then be raised through the available discussion lists (as described at http://www.psidev.info).

'Shared' modules

The number of ongoing reporting requirements projects raises a concern about overlap between independently generated standards. For example, were MIAPE and MIAME to offer differing prescriptions for describing study organisms, the subsequent integration of proteome and transcriptome datasets would be unnecessarily complicated. We understood this early in the development process; thus, the PSI has focused on techniques that are more or less specific to proteomics rather than very general areas, such as the description of the biological material under study (which has relevance for many areas of bioscience). However, there are two complicating factors. First, proteomics, metabolomics and the like are not completely discrete entities; for example, mass spectrometry is as much a tool of metabolomics as it is of proteomics. Second, different kinds of study often require different levels of detail; for example, with respect to the husbandry of animal study subjects, dietary information is key for metabolomics but relatively unimportant for genomic sequencing. Reporting requirements for all technologies, protocols or entities that have relevance for many kinds of bioscience should

Box 1 The MIAPE modules

The various MIAPE modules are briefly described here, along with their status at the time of writing (in brackets). Recent versions of all modules that have progressed significantly are available from the MIAPE home page (http://www.psidev.info/miape).

- Study design and sample generation [to be developed collaboratively through MIBBI[35]]
  Experimental motivation and design; factors of interest; origin and preprocessing of biological material; numbers of replicates; relationship to other studies; miscellaneous administrative detail.
- Separations and sample handling [to be developed collaboratively through MIBBI[35]]
  The use of various techniques (excepting column chromatography, gel electrophoresis and capillary electrophoresis) to fractionate, deplete or otherwise manipulate a sample before analysis; also, sample storage and transport.
- Column chromatography [in the PSI document process]
  The use of columns, of all scales and flow rates.
- Capillary electrophoresis [drafting]
  The performance of any of the wide range of capillary electrophoresis protocols.
- Mass spectrometry [manuscript submitted for publication]
  The use of a mass spectrometer; the generation of peak lists from raw data; quantification based on the use of an isotopic or chemical label (the application of that label, though, is a form of 'sample handling', and is therefore captured elsewhere).
- Informatics for mass spectrometry [manuscript submitted for publication]
  The use of processing engines to analyze mass spectrometry data (both spectra and ion chromatograms). This includes search engines that assign peptides, proteins or biological class membership to spectra; the matching of assigned peptides, proteins or *de novo* sequences against a named database; quantification and the use of quality control measures.
- Gel electrophoresis [manuscript submitted for publication]
  The use of gel-based electrophoretic separation techniques, single- or multidimensional, native or denaturing; various visualization techniques, including 'electroblotting'; image acquisition.
- Gel image informatics [drafting]
  The processing, analysis and interrelation of gel images (to identify spots, measure relative intensities, or warp images to align them).
- Protein and peptide arrays [exploratory discussions]
  Array type, design and construction; experimental protocol; data collection and initial analysis.
- Statistical analysis of data [to be developed collaboratively through MIBBI[35]]
  Applicable to qualitative, quantitative and comparative studies: the use of generic data transformation algorithms (for example, normalization); the calculation of descriptive statistics, such as confidence intervals; methods used to sum, average, cluster or otherwise compare datasets.
- Molecular interaction experiments [published in this issue[46], p. 894]
  The use of any of a range of techniques to determine a set of interacting molecules, within the context of a particular experiment. This includes such techniques as yeast two-hybrid and tandem affinity purification assays. This checklist is published separately under the title MIMIx ("minimum information about a molecular interaction experiment"), but is a MIAPE module.

therefore be developed in common between the relevant standards bodies (or by way of representative collaborations if no official standards body exists). In many cases, a 'tiered' solution should be sought (for example, for genomic sequencing, identify the source of the organism only; for proteomics or metabolomics, also give feeding schedule; and so forth). To address all of these concerns, the PSI has become an active participant in the MIBBI project[35], which aims to anticipate or remedy such overlaps between sets of requirements.

Cui bono?

We now address the question most famously asked by the Roman orator Cicero: "For whose benefit?" The cost of acceding to requests for richer annotation of proteomics data falls almost completely on the shoulders of the experimentalists generating the data, so what is the justification for their increased expenditure of time and resources?

Several groups stand to benefit from the acceptance of MIAPE by the proteomics community: users of 'MIAPE-compliant' public repositories ('data consumers'), public-sector data producers, and the private sector. Researchers acting as data consumers (whether in the public or private sector) stand to benefit because:

- Data sets generated by specific techniques can be easily identified and retrieved (or excluded).
- Data can be used for purposes different from those for which the data were generated.
- Data and analyses can be assessed in the light of the methods deployed.
- Protocols associated with high-quality data can be more easily retrieved.
- Sufficient information will be available to enable parallel or orthogonal studies to confirm or refute a given result.

However, experimentalists are often under severe time, budget and productivity constraints. They must be assured that they too will reap direct benefits, not just the kudos of enhancing the publicly available corpus of biological data. For public-sector data producers the direct benefits accruing to the routine capture of MIAPE-specified (meta)data include:

- Straightforward, regularized promotion of new protocols and best practice to others.
- Obviation of the need to repeatedly construct sets of appropriate contextualizing information for a project:
  - Facilitates the sharing of data with collaborators.
  - Avoids the risk of loss of information through staff turnover.
  - Enables time-efficient handover of projects from one researcher to another.
- Support for the assessment of results that may have been generated months or even years ago (perhaps in answer to referees' comments during the process of publishing work).
- The performance of better-informed comparisons of datasets,

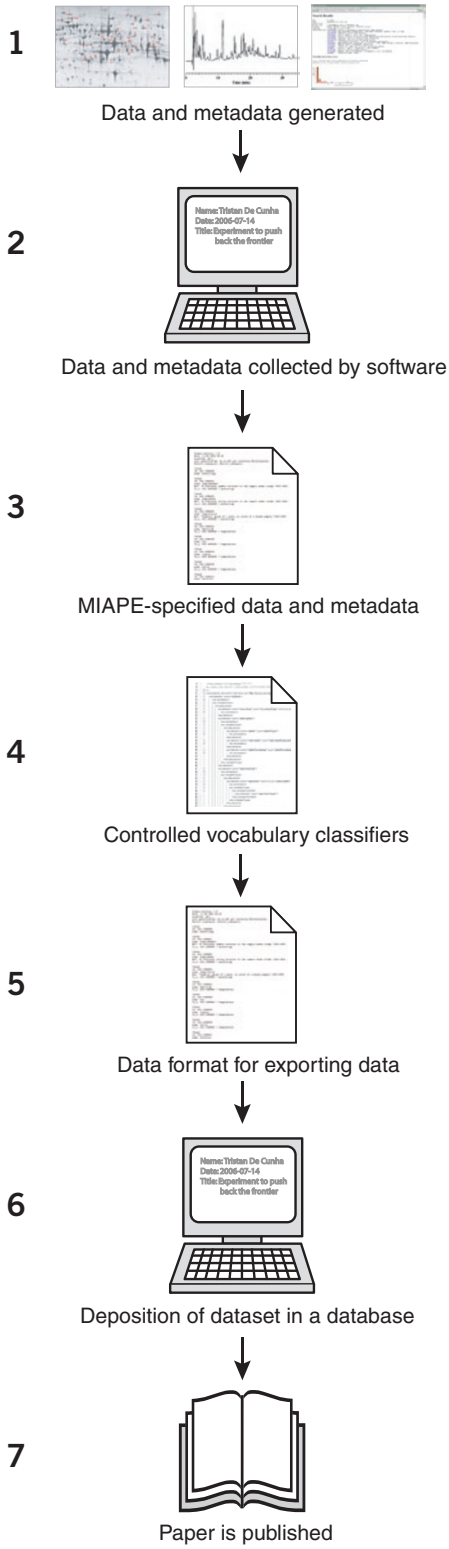

Figure 1 An example of MIAPE-compliant data management. (1) Data and metadata are generated by an experiment; (2) some form of software collects the data and metadata, either by importing from computer-controlled instruments or from manual data entry; (3) MIAPE specifies the data and metadata to be requested by the software tool; (4) a controlled vocabulary supplies classifiers via the software; (5) the software uses a data format specification when exporting a MIAPE-compliant dataset; (6) the dataset is stored in a MIAPE-compliant database and assigned an accession number; (7) a paper, including the appropriate accession number, is published in a journal.

increasing the likelihood of discovering the factors (both controlled and uncontrolled) that might differentiate them. Importantly, this includes the discovery of sources of systematic or random error by correlating data with metadata features such as the equipment used or the date or operator.

- The ability to aggregate proteomics datasets in an informed manner, and to combine proteomics data with data from other kinds of study (for example, correlating changes in mRNA and protein expression levels in response to a stimulus).

All of the above cases are affected by decisions on what information to capture and keep from an experiment and what to ignore or dispose of; MIAPE is relevant to such decisions, because it incorporates the views of many experimentalists considering diverse issues.

Private sector interest in MIAPE has slightly different drivers. The pharmaceutical industry, for example, is accustomed to archiving all the data and metadata generated in the course of an experiment under the US Food and Drug Administration's 21 CFR Part 11 requirements. So arguments based on the benefits of capturing particular kinds of (meta)data do not apply. For that sector the issue is mainly one of efficiency: capturing a reduced set of metadata in a rigorous way (in addition to archiving the full set) facilitates more efficient retrieval, reanalysis and integration of data. The validity of this argument is supported by the abundance of sophisticated software designed to distill all-encompassing datasets into useful summary reports.

Support from funding agencies and journals

At present, projected benefits for data consumers may not directly justify the use of data producers' time, largely because of the mechanisms by which publicly funded researchers are assessed—usually through publications. However, this situation is changing. Funders around the world see that potential value is being squandered through failures to maximize the utility of datasets, which are frequently expensive to produce. Many already have data sharing policies that direct those receiving funds to make their data publicly available. Some go further; for example, the UK Biotechnology and Biological Sciences Research Council (BBSRC) has now finalized a policy statement[38] that not only requires plans to be established for the provision of access to datasets that were generated in the course of BBSRC-funded work (as many other funders, such as the US National Institutes of Health and the National Science Foundation, also do), but also require adherence to agreed community standards, where they exist. This stated aim of having data made publicly available in a manner compliant with community standards will enable researchers to seek support for time spent annotating and depositing their data.

Journals already support and in some cases drive initiatives to establish reporting guidelines, the aim being to ensure that reviewers and readers alike have access to sufficient information to make informed judgments about the content of papers. As stated above, several journals[20–22] already require that authors of transcriptomics papers satisfy the MIAME[19] requirements. Other journals, such as *Molecular and Cellular Proteomics*[39,40] and more recently *Proteomics*[41], have themselves driven processes to generate appropriate guidelines. However, both those journals' requirements incorporate guidance on appropriate experimental processes, whereas MIAPE simply requires the provision of sufficient information to allow quality to be independently assessed (for example, as a part of the review process). It has always been a matter of policy that the PSI should neither attempt to produce standard operating procedures specifying how particular techniques should be performed nor attempt to establish quality assessment benchmarks. We do not believe it is the job of this body to dictate to the proteomics community how it should perform experiments or analyses.

Box 2 Frequently asked questions about MIAPE

Below we clarify common questions about using and contributing to the MIAPE guidelines.

What is MIAPE and what is it for?

- MIAPE is a formal list of the items of information that should be provided when describing particular analytical techniques employed in a proteomics experiment, the data generated and any analyses performed.
- Primarily, MIAPE is a guide for people submitting datasets to databases; this will usually be in support of a publication. MIAPE is not a requirement to share data, merely a prescription for so doing.

How does MIAPE differ from journals' own 'guidelines to authors'?

- Compared with standard author guidance, MIAPE is much more specific. It explicitly lists every piece of information that should be provided, leaving nothing open to interpretation.
- In some cases, author guidance addresses the issue of quality by setting specific thresholds and operating procedures.
  - — MIAPE does not address quality in any form; such judgments are the province of reviewers (who will be better equipped to form such judgments if provided with a MIAPE-compliant dataset).
  - — Neither does MIAPE recommend any particular protocol. Techniques are addressed solely on the basis of frequency of use, and the guidelines for reporting any one technique are sufficiently flexible that they avoid prescribing the manner of use.

What else do I need to use MIAPE?

- Minimally, one needs a method of capturing information; this could be as simple as a pen and paper! Normally, though, one would require the following (**Fig. 1**):
  - — A data capture tool (Excel, for example, has been modified for this purpose).
  - — A source of controlled vocabulary terms, such as the lists produced by the PSI.
  - — A suitable data format, such as those produced by the PSI.
  - — A database in which to store the data file once completed.

What if no appropriate MIAPE module exists for my technique?

- Where guidance for reporting a particular technique does not exist, experimentalists should refer to the two general principles described here (sufficiency and practicability) in preparing their report. Ideally, they would also propose a new module to the PSI.

How can I have input into future versions of the MIAPE modules?

- Although the PSI has pledged to keep all modules stable for as long as possible (at least one year per version), updates and extensions will be required as new techniques are developed and existing ones evolve. The PSI maintains several channels through which anyone can contribute to the evolutionary process. These are:
  - — Mailing lists, details of which can be found at the website.
  - — Open teleconferences, announced on the mailing lists.
  - — Annual and other meetings, announced on the web site and through the mailing lists.

Ultimately, we anticipate that the publication of a proteomics experiment will involve (i) the submission of a 'classical' paper for the print or electronic version of the journal; (ii) the submission of optional supplementary data to the journal, providing further analysis; and (iii) the submission of a supporting dataset, annotated to the level specified by MIAPE, to a MIAPE-compliant database. This scenario has two benefits: first, the foundations of conclusions presented in papers would routinely be made available for inspection; second, journals would retain the right to determine the level of detail appropriate for the paper while still requiring compliance with MIAPE. For this situation to obtain, it is clear that MIAPE should endeavor to include all the (non-quality-related) components of guidelines such as those produced by *Molecular and Cellular Proteomics* and *Proteomics*. This is already the case for developed MIAPE modules, as analyses conducted by the PSI have established.

File formats, controlled vocabularies, tools and databases

The information requested by MIAPE modules could be captured and transmitted in any number of ways; a typical workflow is outlined in **Figure 1**. MIAPE is an implementation-independent description of information that should be transmitted along with the core data from an experiment. MIAPE is not a data format (such as an XML or a tab-delimited text file), nor is it a source of 'controlled vocabulary' terms (well characterized, consensus terms for use predominantly in data files). Furthermore, neither the MIAPE modules nor this document make the use of any particular file format or any other resource a condition of MIAPE compliance; this is important, as to require the use of a particular informatics framework could, for example, exclude those who would wish to comply with MIAPE in a purely commercial setting where a LIMS stores data using a proprietary format. However, the PSI also develops data formats and vocabularies (see http://www.psidev.info); a specific 'use case' for these is the transmission of MIAPE-specified information. Note also that although there is no requirement in MIAPE that data must be made publicly accessible, we endorse public access to data as being of benefit to science.

Neither journals nor funders will move to require MIAPE compliance until tools and databases are in place to ease data capture and support data sharing. The prevalence of computer technology in the lab is a boon here; instrument manufacturers and LIMS and analysis software vendors have shown continued interest in MIAPE (evidenced by the regular attendance of representatives of such companies at the PSI's public meetings). Support from such commercial vendors will greatly simplify the reporting process.

Tool support is important for the success of guidelines such as MIAPE. The policy initiatives of various funders and the position taken by many

journals has now created a market for commercial software that, first, supports the capture of the required data either directly from instruments or by acting as an 'electronic lab book' and, second, supports the export of MIAPE-compliant reports (that is, data files of some description containing everything required by the guidelines). Ideally such software will be able to export standard data formats, such as those developed by the PSI.

Of course the funds required to purchase commercial software may not be available to all; only once appropriate free tools become available will it be reasonable to expect that experimentalists comply fully with MIAPE and its kin. It is not unrealistic to expect such tools to appear. Substantial tool development can be achieved by the public sector, as shown by projects such as CPAS[42], and public funders look ever more favorably on projects that aim to develop appropriate tools to support data sharing. The PSI will provide, as a base solution, specially designed Microsoft Excel spreadsheets similar to the ProteomeHarvest tool that has proved so successful as a submission route for the PRIDE database[12], one per MIAPE module, to assist in capturing the MIAPE-specified set of data and metadata. These Excel spreadsheets are intended to be used only for low volumes of data, such as the data supporting a paper. For situations where legacy software or large data volumes make MIAPE compliance challenging, the PSI will encourage the development of appropriate tools. We anticipate that both free and commercial software will track standards as they are produced.

Having collected the data and metadata from an experiment and encoded it in a PSI format using the appropriate controlled vocabulary, experimentalists will need a place to deposit those data. As stated above, there are now a number of broadly appropriate databases[9–14], although at present only PRIDE[12] can store the level of description MIAPE requires in a structured manner. MIAPE should no more require the use of any particular repository than it should require the use of any one format, although the PRIDE repository will endeavor to support PSI standards as they emerge.

Finally, we consider the issue of validating whether an allegedly MIAPE-compliant dataset actually contains all that MIAPE specifies. Were a journal to require MIAPE compliance, reviewers would benefit from assistance in checking that datasets submitted in support of publications contain the required information. Repositories can help in this; as a likely destination for datasets supporting unpublished work, they can put mechanisms in place to check for compliance upon submission of a dataset and approve that dataset on the reviewers' behalf. This kind of service has recently been offered[27,43] in the context of transcriptomics by the ArrayExpress[4] and GEO[5] repositories.

Conclusion

One of the main objectives of the MIAPE process is to increase the value derived by the scientific community from ongoing experimentation in proteomics, through community processes that support sharing, dissemination and reanalysis of datasets, and that assist in establishing and promoting best practice in specific technical areas. Guidelines such as these promote transparency in experimental reporting, enhance accessibility to data and support effective quality assessment, thereby increasing the general value of a body of work (and by extension the competitiveness of the originators of that work).

The MIAPE guidelines require a fairly rich description without being overly burdensome: much of the required data should be readily available in electronic form and therefore amenable to export, especially as vendors of instruments, analysis software and LIMS implement standards-compliant export facilities. Additionally, a substantial portion of the captured metadata will be common to many experiments, permitting some economies of scale through reuse. MIAPE therefore provides a sound base for developers of repositories and tools to work from, and a rational framework for journals and funders to consider enforcing.

Several benefits will arise from the widespread acceptance of MIAPE. Compliant datasets will contain sufficient information to quickly establish the provenance and relevance (to the researcher) of a dataset. Additionally, tools will be developed that afford easy access to, and analysis of, large numbers of such datasets. Tool development will be facilitated by standardized XML-based data transport formats and controlled vocabulary terms generated by the PSI. The MIAPE modules constitute a set of 'requirements documents' for such development. Details of all PSI standards development projects can be found on the project website (http://www.psidev.info).

In this age of genome- and proteome-scale experiments, the need to standardize the content of reports of biological experiments is evident if we are to extract full value from our activities[44,45]. It is our hope that this document and the modules accompanying it will begin to fulfill this need for proteomics researchers and for the proteomics community as a whole, increasing the value of both individual pieces of work and of the general, diverse corpus to which so many contribute.

ACKNOWLEDGMENTS

The people who have contributed to the evolution of this document are too numerous to name, but it is important to note the contribution made by the PSI's many participants, who have contributed through attendance at meetings, web-based discussion and other forms of communication. Many others have also commented, at conferences and elsewhere: academics, commercial scientists, vendors, funders, editors and others.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Cochrane, G., *et al.* EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.* **34** (Database issue) D10–D15 (2006).
2. Galperin, M.Y. The molecular biology database collection: 2006 update. *Nucleic Acids Res.* **34** (Database issue), D3–D5 (2006).
3. Ball, C.A., *et al.* The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* **33** (Database issue), D580–D582 (2005).
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. Barrett, T. & Edgar, R. Gene Expression Omnibus (GEO): microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.* **411**, 352–369 (2006).
6. Wu, C.H., *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34** (Database issue) D187–D191 (2006).
7 Birney, E., *et al.* Ensembl 2006. *Nucleic Acids Res.* **34** (Database issue), D556–D561 (2006).
8. Pruitt, K.D., Tatusova, T., Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33** (Database issue) D501–D504 (2005).
9. Appel, R.D. *et al.* Federated 2-DE database: a simple means of publishing 2-DE data. *Electrophoresis* **17**, 540–546 (1996).
10. Babnigg, G., Giometti, C.S. GELBANK: a database of annotated two-dimensional gel electrophoresis patterns of biological systems with completed genomes. *Nucleic Acids Res.* **32** (Database issue), D582–D585 (2004).
11. Garwood, K. *et al.* PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics [online]* **5**, 68 (2004).
12. Jones, P., *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34** (Database issue), D659–D663 (2006).
13. Prince, J.T. *et al.* The need for a public proteomics repository. *Nat. Biotechnol.* **22**, 471–472 (2004).
14. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
15. Hogan, J.M., Higdon, R. & Kolker, E. Experimental standards for high throughput proteomics. *OMICS* **10**, 152–157 (2006).
16. Lenth, R.V. Some practical guidelines for effective sample size determination. *Am. Stat.* **55**, 187–193 (2001).
17. Orchard, S., Hermjakob, H. & Apweiler, R. The proteomics standards initiative. *Proteomics* **3**, 1374–1376 (2003).

18. Hermjakob, H. The HUPO Proteomics Standards Initiative — overcoming the fragmentation of proteomics data. *Proteomics* **6** (suppl. 2), 34–38 (2006).
19. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
20. Anonymous. Microarray standards at last. *Nature* **419**, 323 (2002).
21. Ball, C.A. *et al.* A guide to microarray experiments—an open letter to the scientific journals. *Lancet* **360**, 1019 (2002).
22. Information for authors. Cell Online <http://www.cell.com/misc/page?page=authors> (2006).
23. Sreenu, V.B., Kumar, P., Nagaraju, J. & Nagarajaram, H.A. Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity. *BMC Genomics [online]* **7**, 78 (2006).
24. Albers, C.J., Jansen, R.C., Kok, J., Kuipers, O.P. & van Hijum, S.A. SIMAGE: simulation of DNA-microarray gene expression data. *BMC Bioinformatics [online]* **7**, 205 (2006).
25. Larsson, O. & Sandberg, R. Lack of correct data format and comparability limits future integrative microarray research. *Nat. Biotechnol.* **24**, 1322–1323 (2006).
26. Burgoon, L.D. The need for standards, not guidelines, in biological data reporting and sharing. *Nat. Biotechnol.* **24**, 1369–1373 (2006).
27. Edgar, R. & Barrett, T. NCBI GEO standards and services for microarray data. *Nat. Biotechnol.* **24**, 1471–1472 (2006).
28. Moher, D., Schulz, K.F. & Altman, D.G. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**, 1191–1194 (2001).
29. Moher, D. *et al.* Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* **354**, 1896–1900 (1999).
30. Bossuyt, P.M. *et al.* Towards complete and, accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Br. Med. J.* **326**, 41–44 (2003).
31. McShane, L.M. *et al.* REporting recommendations for tumour MARKer prognostic studies (REMARK). *Eur. J. Cancer* **41**, 1690–1696 (2005).
32. CDISC SEND Team. Standard for exchange of nonclinical data (SEND). Implementation guide for animal toxicology studies. Version 2.3. CDISC Standards <http://www.cdisc.org/models/send/v2.3> (2005).
33. Plint, A.C. *et al.* Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med. J. Aust.* **185**, 263–267 (2006).
34. Smidt, N. *et al.* The quality of diagnostic accuracy studies since the STARD statement—has it improved? *Neurology* **67**, 792–797 (2006).
35. Taylor, C.F. *et al.* Promoting coherent minimum reporting requirements for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* (in the press).
36. Wilkins, M.R., Williams, K.L., Appel, R.D. & Hochstrasser, D.F. (eds.) *Proteome Research: New Frontiers in Functional Genomics* (Springer, Berlin, 1997).
37. Pennington, S.R. & Dunn, M.J. (eds.) *Proteomics: From Protein Sequence to Function.* (BIOS, Oxford, 2001).
38. BBSRC's data sharing policy. BBSRC: Biotechnology and Biological Sciences Research Council <http://www.bbsrc.ac.uk/support/guidelines/datasharing/context.html> (2006).
39. Publication guidelines for the analysis and documentation of peptide and protein identifications. Molecular & Cellular Proteomics <http://www.mcponline.org/misc/ParisReport_Final.shtml> (2007).
40. Carr, S. *et al.* The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics* **3**, 531–533 (2004).
41. Wilkins, M.R. *et al.* Guidelines for the next 10 years of proteomics. *Proteomics* **6**, 4–8 (2006).
42. Rauch, A. *et al.* Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **5**, 112–121 (2006).
43. Brazma, A. & Parkinson, H. ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat. Biotechnol.* **24**, 1321–1322 (2006).
44. Noble, W.S. Data hoarding is harming proteomics. *Nat. Biotechnol.* **22**, 1209 (2004).
45. Quackenbush, J. Standardizing the standards. *Mol. Syst. Biol.* **2**, 2006.0010 (2006).
46. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).