The HUPO Proteomics Standards Initiative (HUPO-PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. The PSI Workshop at the Tenth Annual HUPO World Congress in Geneva was opened by Pierre-Alain Binz (SIB) who welcomed delegates and briefly summarised the work of the initiative over the last 10 years, including the release of the Minimum Information About a Proteomics Experiment (MIAPE) standards, exchange formats and the accompanying controlled vocabularies (CVs) for several aspects of a mass spectrometry worktrack and for molecular interactions.

Sandra Orchard (EMBL-EBI) described the work of the molecular interactions worktrack over the last few months. The established Minimum Information required for a Molecular Interaction experiment standards document has now been supplemented by Minimum Information About a Protein Affinity Reagent, describing the interactions made by protein affinity reagents and by Minimum Information About a Bioactive Entity, a small molecule-target binding standard. All of these data types can be contained within the existing PSI-MI XML and tab-delineated exchange formats, and described by terms in the existing PSI-MI CV. The shared usage of a common format has enabled the creation of the Proteomics Standards Initiative Common QUery interface (PSICQUIC) web service. At the time of the HUPO congress, users could already access over 30 million interaction evidences spread over 20 sites with a single query and a reference implementation has enabled such a query system to be built into several existing websites. The PSICQUIC project site (http://psicquic.googlecode.com/) offers open-source client libraries and code examples, and all participating resources are listed in the PSICQUIC registry (http://code.google.com/p/psicquic/wiki/Registry). A distributed interaction confidence scoring architecture (PSIscore) has been developed to take advantage of this service, enabling individual research groups to create specific scoring approaches and users to select one or more of these to subsequently score their own data against. The International Molecular Exchange consortium (IMEx) also uses PSICQUIC to make the IMEx set of high-quality, non-redundant protein–protein interaction records available to both the IMEx website and those of individual member databases though a tagging process (http://www.imexconsortium.org). An ever-increasing corpus of consistently annotated interaction records is now available from this site.

Eric Deutsch (ISB, Seattle) then detailed the progress made by the Mass Spectrometry Standards Working Group. mzML, a stable format created to encode mass spectrometer output data, has now been implemented by an increasing number of vendors. Programming libraries exist in a number of languages, including jmzML, a Java API and OpenMS, an open-source C11 library for mass spectrometry, which provides classes for reading and writing mzML, that can easily be integrated in other software tools. imzML, developed to store MS imaging data, has been aligned with mzML, but differs in the storage of mass spectral data in a binary file. Talks are on-going with the metabolomics community with regard to adoption of these standards for small molecule usage. A digital signature mechanism is currently being investigated as a possible addition to the index wrapper schema.

The efforts of the workgroup are currently concentrating on the development of TraML, a new standardized format for encoding transition lists and associated metadata. A Java API, jTraML, is available (http://code.google.com/p/jtraml/wiki/TraML) and a two-way conversion tool, between TraML documents and vendor specific files, facilitates the adoption process. All of the mass spectrometry data formats (mzML, mzIdentML, mzQuantML and TraML) share a single-controlled vocabulary of terms for annotation (PSI-MS) and this is now in the process of being re-ordered and expanded. Finally, MIAPE-MS, the mass spectrometry minimum information standard, has been updated in line with advances in the field since its original publication, and following discussions with a panel of proteomics journal editors at the Toronto HUPO Congress in 2009. However, synchronization with MIAPE-MSI and MIAPE-Quant (see below) has to be completed before publication.

Martin Eisenacher (MPC, Ruhr-Universit.at Bochum) went on to speak about the formats being developed to capture the output of spectrum identification results. mzIdentML has been updated from version 1.0 to 1.1, which is expected to be the stable version. It enables a description of the search with enough detail to enable the user to reproduce the search (software, sequence database and parameters utilized) and also a report of the final results, i.e. the peptides and proteins identified together with their scores. The format can encompass the results of an MS, MS/MS or MSn, spectral library searches, the results of decoy searches and multiple search results combined into a single protein list. A semantic validator for mzIdentML – checking the correct usage of controlled vocabulary terms – is also available (see http://www.psidev.info/index.php?q5 node/304). The standard is now being widely implemented with exporters available for most major search engines (see http://www.psidev.info/index.php?q5node/408 for a full list) although not all of them are updated to version 1.1 yet.

Work is currently focusing on the development of mzQuantML, the data exchange format for quantitative data. The format is required to encompass final abundance values (relative or absolute) for peptides, proteins and protein groups, quantification of post-translational modifications, abundance values at the level of a single run and relative values of groups of runs, relationships between features either on different regions of the same spectrum or on different spectra that report on the same peptide or small molecule, and details about pre-fractionation sufficient to describe the combination of multiple input data files. mzQuantML 1.0.0 supports MS1 label-free intensity, MS1 label-based (e.g. SILAC and metabolic labelling such as 15N, MS2 tag-based (e.g. iTRAQ/TMT) and MS2 spectral counting. Additionally, it is expected to be suitable also for quantification by selected reaction monitoring (SRM), absolute quantification based on averaging the intensities of features, small molecule quantification (in metabolomics), MS2 intensity-based approaches and MS2 label-based approaches but further development and testing is required by users of these techniques. More documentation is available in the mzQuantML Google code project at http://code.google.com/p/mzquantml/. The standard was recently submitted to the PSI document process – a standardized workflow for community and public review before release of a standard. More implementations are being encouraged prior to publication.

MIAPE guidelines for quantitative proteomics, MIAPE-Quant, are now being defined with aim of standardising the reporting of such data and this work was described by Salvador Martı´nez-Bartolome´ (ProteoRed). Existing quantitative elements have been removed from both MIAPE-MS and MIAPE-MSI, and both these documents are being updated in parallel. A range of techniques, including isotope-coded protein labeling, iTRAQ, SILAC, label-free methods such as spectral counting, and SRM/MRM have been included in the document. Example documents for each technique have been written. The document is currently being circulated to interested groups for comment.

mzIdentML and mzQuantML are established XML-based formats, but their inherent complexity requires sophisticated bioinformatics expertise. Juan Antonio Vizcaíno (EMBL-EBI) spoke about the development of mzTab, a lightweight, tab-delimited file format for proteomics experiments (http://code.google. com/p/mztab/). This format is designed to provide a simple summary of results, allowing the reporting of experimental results such as peptide/protein IDs, quantitative data and associated metadata, with one file having the capability to contain multiple experiments. The aim of the format is to present the results of a proteomics experiment in a computationally accessible overview, not to provide the detailed evidence for these results, or allow recreating the process which led to them. Both of these functions are established through links to more detailed representations in other formats, such as mzIdentML and mzQuantML. mzTab has also been designed to report MS metabolomics results.

Juan Pablo Albar (ProteoRed) detailed a 2D DIGE case study on the effects of salbutamol on the rat muscle proteome as an example of best practices in proteomics data sharing. The gel-based protein identification data set was deposited in the PRoteomics IDEntifications (PRIDE) database (http://www.ebi.ac.uk/pride), using a new software tool, the PRIDESpotMapper (http://proteo.cnb.csic.es/pridespotmapper/), which works in conjunction with the PRIDE Converter application. Additionally, the ProteoRed MIAPE generator tool was used to create and share a complete and compliant set of MIAPE reports for this experiment. This means that the dataset is now publicly available with sufficient metadata to allow other groups to verify the findings, perform data mining or integrate this with in-house data.

Finally, Henning Hermjakob (EMBL-EBI) concluded with a description of the ProteomeXchange consortium (http://www.proteomexchange.org), established to provide a single point of submission of MS proteomics data to the main existing proteomics repositories, and encourage data exchange between them. Initial submission of raw, processed and metadata files will, at the moment, only be via the PRIDE database. Typically, PRIDE will deposit the raw instrument output files into Tranche or into a similar resource that can store them (generally called raw files archive), and will get the links back into PRIDE. Submission of search engine output files to the ProteomeXchange consortium is optional. These files will be stored in the raw files archive too. By default, all the data will remain private during the review process.

Secondary data resources (such as PeptideAtlas [www.peptideatlas.org/] and UniProtKB [http://www.uniprot.org]) will only have access to the data once it has been made publicly available at which point a ProteomeXchange message (XML file) would be generated by PRIDE and distributed via a RSS feed. This mechanism will be used to announce to the other members of the ProteomeXchange consortium (and the proteomics community) that a new ProteomeXchange project is now available. The system will be tested in the first months of 2012 and it is expected to be operational by summer 2012.

Henning concluded by thanking delegates for attending the session and keeping up a lively discussion during the question time following each talk. The next HUPO-PSI meeting will be in San Diego, USA on March 12–14. Further details will be made available at the PSI website (www.psidev.info), and registration will be free, as in previous years. All interested parties are invited to attend.