MCP Papers in Press. Published on May 17, 2017 as Manuscript M117.068429

The mzIdentML data standard version 1.2, supporting advances in proteome informatics

Juan Antonio Vizcaíno[1], Gerhard Mayer[2], Simon Perkins[3], Harald Barsnes[4,5,6], Marc Vaudel[4,6,7], Yasset Perez-Riverol[1], Tobias Ternent[1], Julian Uszkoreit[2], Martin Eisenacher[2], Lutz Fischer[8], Juri Rappsilber[8,9], Eugen Netz[10], Mathias Walzer[11], Oliver Kohlbacher[10,11,12,13], Alexander Leitner[14], Robert J. Chalkley[15], Fawaz Ghali[3], Salvador Martínez-Bartolomé[16], Eric W. Deutsch[17] and Andrew R. Jones[3,*]

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

[2] Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, D-44801 Bochum, Germany.

[3] Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK.

[4] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway.

[5] Computational Biology Unit, Department of Informatics, University of Bergen, Norway.

[6] KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway.

[7] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway.

[8] Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom.

[9] Chair of Bioanalytics, Institute of Biotechnology Technische Universität Berlin, 13355 Berlin, Germany.

[10] Biomolecular Interactions group, Max Planck Institute for Developmental Biology, Tübingen D-72076, Germany.

[11] Center for Bioinformatics, University of Tübingen, 72076 Tübingen, Germany.

[12] Dept. of Computer Science, University of Tübingen, Germany.

[13] Quantitative Biology Center, University of Tübingen, Germany.

[14] Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Auguste-Piccard-Hof 1, 8093 Zurich, Switzerland.

[15]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, 94143, USA.

[16] Department of Chemical Physiology, The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, CA, 92037, USA.

[17] Institute for Systems Biology, Seattle, WA, 98109, USA.

* Corresponding author. andrew.jones@liv.ac.uk.

Postal address: Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK.

Running title: mzIdentML data standard version 1.2

Abstract

The first stable version of the Proteomics Standards Initiative mzIdentML open data standard (version 1.1) was published in 2012 – capturing the outputs of peptide and protein identification software. In the intervening years, the standard has become well supported in both commercial and open software, as well as a submission and download format for public repositories. Here we report a new release of mzIdentML (version 1.2) that is required to keep pace with emerging practice in proteome informatics. New features have been added to support: (i) scores associated with localization of modifications on peptides; (ii) statistics performed at the level of peptides; (iii) identification of cross-linked peptides; and (iv) support for proteogenomics approaches. In addition, there is now improved support for the encoding of *de novo* sequencing of peptides, spectral library searches and protein inference. As a key point, the underlying XML schema has only undergone very minor modifications to simplify as much as possible the transition from version 1.1 to version 1.2 for implementers, but there have been several notable updates to the format specification, implementation guidelines, controlled vocabularies and validation software. mzIdentML 1.2 can be described as backwards compatible, in that reading software designed for mzIdentML 1.1 should function in most cases without adaptation. We anticipate that these developments will provide a continued stable base for software teams working to implement the standard. All the related documentation is accessible at http://www.psidev.info/mzidentml.

Introduction

The Proteomics Standards Initiative (PSI) has taken the role of developing standard file formats for different aspects of mass spectrometry (MS) based analysis (for a review see (1)). These include the mzML format, which is able to store raw MS data suitable for quantitation processes, as well as processed peak lists for searching (2). Downstream of mzML, several formats serve different use cases. To capture the results of identification approaches, such as the use of proteomics search engines, the mzIdentML format was released as a stable version (version 1.1) in the last quarter of 2011, and published in 2012 (3). A separate format for quantitative results was built, called mzQuantML – initially defined for large scale discovery approaches (4) and updated to also support targeted approaches, such as selected reaction monitoring (5). More recently, in 2014, a lighter tab-delimited format called mzTab was also developed by the PSI to capture identification and quantification results (6).

mzML, mzIdentML and mzQuantML are represented in the Extensible Markup Language (XML). The three formats are defined by an XML Schema Definition (XSD) file, which have all been fixed on a particular version for some time: mzML (version 1.1 since 2009), mzIdentML (version 1.1 since 2011) and mzQuantML (version 1.0 since 2013). The PSI has also built a large controlled vocabulary (CV), the PSI-MS CV (7), containing more than 2,700 terms and definitions in a structured hierarchy (version 4.0.12, May 2017). To ensure that terms are used correctly within the formats, an additional mechanism has been built on top of XSD validation, called PSI semantic validation (8). This mechanism is encoded as a *CV mapping file* listing all places in each format where CV terms are allowed, and specifying a branch or branches of the PSI-MS CV where terms can be sourced, called a *semantic rule*. Each semantic rule is accompanied by a requirement level – MAY, SHOULD or MUST (formal keywords), triggering particular behaviour in custom validation software (information, warning, and error, respectively). To create a valid mzIdentML file thus requires it to be syntactically correct (passing XSD validation) and semantically correct (passing PSI semantic validation i.e. intended CV terms are present and correct), and these features have been implemented in validation software (9).

Since its initial stable release, the adoption of mzIdentML 1.1 has increased enormously (Table 1). Most notably, several popular proteomics software now export mzIdentML natively, and this list is growing regularly. This includes analysis tools such as MS-GF+ (10), Mascot (*Matrix Science,* from version 2.4), ProteinPilot (*SCIEX,* from version 5.0), ProteinLynx Global Server (*Waters Corp.*, from version 3.0.3 onwards), PEAKS (11), Scaffold (12), Byonic (*Protein Metrics*), MyriMatch (13), PeptideShaker (14), Crux (15), OpenMS (16), mzID in Bioconductor (17) and PIA (Protein Inference Algorithms) (18). It is also planned that ProteomeDiscoverer (*Thermo Fisher Scientific*) will export the format natively in its next version, to be released later in 2017. It is used as an import format to quantitation software, such as Progenesis QI for proteomics (*Waters Corp.*) e.g. to import results from Byonic. In addition, it is becoming increasingly common that mzIdentML becomes the final output of analysis pipelines, e.g. ProteoAnnotator (19). For software not natively implementing the format, some converters are available, e.g. ProCon (20) and ProteoWizard (21). Visualisation software for mzIdentML files is also now available, most notably the open source PRIDE Inspector tool (22), which was updated in 2016 to fully support mzIdentML, and the ProteoIDViewer (9). Some of these tools are reusing open source libraries tailored to the format such as jmzIdentML (23), mzid Library (9) and the ms-data-core-api (24). Finally, it is important to highlight that mzIdentML is now

fully supported as a data submission format, via the 'complete' submission mechanism (enabling full search capabilities and visualisation of the data), to the PRIDE, MassIVE and jPOST data repositories, as part of the ProteomeXchange Consortium (25).

An important consideration for data standards is the balance between stability and innovation. A standard that is updated at regular intervals causes problems for the developer community – including those writing exporters from their own software, as well as those wishing to write parsers for data produced by others. However, it is also important for standards organisations to update formats periodically as the requirements of the field evolve. In this article, we are reporting an update from mzIdentML version 1.1 to version 1.2 to cope with several features that have been requested by software teams or by the wider proteomics community, and that were not specified previously. New features have generally been implemented by adding new CV terms to the format, and updating the way in which terms are used in a valid mzIdentML file as opposed to making changes to the XML schema itself. However, several minor updates have been made to the mzIdentML XSD file to fix bugs or important omissions (for example concerning whether elements are mandatory or optional), which will overall improve the ease of development around the format. The background to the major new features and improvements is summarized in the following sections. The changes made to mzIdentML 1.2 can be described as backwards compatible, in that reading software designed for mzIdentML 1.1 should function in the vast majority of cases without adaptation. It is understood that standards should remain stable for significant periods of time to ensure ease of adoption by the developer community, and as such, we have attempted to add new features in a manner that will make it as straightforward as possible for existing adopters of the format.

Modification localization. Even if the identification of a peptide carrying one or more post-translational modifications (PTMs) or chemical modifications can be confirmed with high certainty, the exact residue on which the modification resides may be ambiguous, particularly for modifications that are known to occur on multiple residues, such as phosphorylation on S, T or Y. A variety of approaches have been developed that give scores or statistical measures to modification localization, as reviewed in (26). There is a growing interest in annotating proteomes with previously "confirmed" PTM sites on proteins, and thus if MS-derived data is to be used for this purpose, it is imperative that the evidence trail is adequately reported in the standard, which could not previously be achieved in a clear way.

Cross-linking. Cross-linking MS has become a standard tool in structural biology investigations of multi-protein complexes (27) and can lead to detailed models of proteins (28). The principle of the technique involves the use of a chemical reagent to cross-link residues close in physical space in the three-dimensional structure of a protein or protein complex. Cross-linked peptides are identified by database searching and can reveal which residues are in close proximity in the folded proteins. For instance, such information can be integrated with other sources in structural biology to compute structures of biomolecular systems (29). Upwards of 30 specifically designed search engines and associated statistical techniques are available for identifying cross-linked peptides. The data types resulting from such software require considerable changes to the standard reporting guidelines as used in the rest of the proteomics field.

Peptide-level statistics. In large-scale (discovery) proteomics MS/MS approaches, there has been much work over the last two decades to improve statistical approaches, to ensure that results from different analytical pipelines are more comparable and to accurately estimate the false discovery rate (FDR). The approach as first described was applied to peptide-spectrum matches (PSMs), and the field adopted a consensus 1% FDR threshold. However, high intensity peptides eluting over a retention time higher than the dynamic exclusion of the instrument generate high-similarity fragment spectra which, when correctly identified, result in redundant PSMs, whereas false positives are more evenly distributed across peptides. As a result, if a fixed threshold (e.g. 1%) is used for false discovery at the PSM-level, it is likely that the actual level of false discovery at the peptide-level is somewhat higher. This phenomenon is important in a variety of cases, for example the identification of phosphopeptides (or peptides containing other PTMs) or in the annotation of genomes, where it is important to control FDR at the level of the individual peptide sequence. Consequently, new structures have been added to mzIdentML 1.2 to support the grouping of PSMs into peptide units, and the reporting of scores for peptides (as well or instead of PSM scores).

Proteogenomics. In these approaches (30), searches are performed against databases that are generated using genomic and/or transcriptomic sequence information, from which novel peptides and sequence variants can be identified. One of the key concepts required is the mapping of peptides back to gene models and chromosomes, for example demonstrating evidence where peptides map across splice junctions. To ensure that a consistent export is possible from mzIdentML to formats designed specifically for genome visualisation or annotation, e.g. the BED or SAM/BAM formats (31), and their recently developed proteomics counterpart PSI formats proBed (see http://www.psidev.info/probed) and proBAM (http://www.psidev.info/probam), in mzIdentML 1.2, a consistent encoding of the chromosomal mappings for peptides has been developed.

Protein grouping. The protein inference problem has been widely discussed in the literature (32, 33). Most identification pipelines now report grouped protein identifications where ambiguity cannot be resolved e.g. proteins have been identified from the same set of peptides. For each group, one or more *leading* or *representative* proteins can be reported. In mzIdentML 1.1, a two level hierarchy was defined for capturing evidence at the level of the *group*, and the level of individual database *accession numbers*. However, a higher-level concept emerged in some approaches of a protein *cluster* or *family* (34), inside which groups of different proteins shared some peptides in common, but also had independent evidence. The mzIdentML 1.1 specifications left the developers of export software to choose how to use these structures, and the result is that inconsistencies arose around the encoding of clusters/families (which were not explicitly mentioned in the format specification), as well as the definition of group leading proteins. A PSI working group investigated the issue at length, taking on board a wide range of opinions, and examining all popular approaches in software. From the working group a new specification emerged, described previously (35), and now included in this stable release of mzIdentML 1.2.

Experimental Procedures

The development of mzIdentML 1.2 started in 2012 and it has been an open process *via* conference calls, discussions at the PSI annual meetings and smaller workshops. The specifications have been submitted to the PSI document process (36) for review, during which time external reviewers can

provide feedback on the specifications and they are available for public comments, enabling broad input on the specifications. The model is accompanied by CV terms and definitions as part of the PSI-MS CV, also used in other PSI data formats and actively maintained by the PSI MS and PI (Proteomics Informatics) working groups, as well as a newly developed CV for cross-linking reagents and modifications called XLMOD-CV. The complete mzIdentML 1.2 specification document, the new XLMOD-CV, example files and additional documentation can be found at http://www.psidev.info/mzidentml.

mzIdentML overview

Here we briefly describe the structure of mzIdentML files, as a basis for demonstrating the mechanism used to add the new features. As mentioned above, this overall file structure is nearly identical in mzIdentML 1.2 and mzIdentML 1.1. We have attempted to encode the new use cases without changing the core model of the format, to simplify adoption by the developer community. The core model is summarised in Figure 1. The core data type in proteomics identification approaches (by MS/MS) is the PSM. The majority of search engines output one or more ranked explanations (peptide sequences) that match each collected MS/MS fragmentation spectrum, associated with scores or statistical values. In mzIdentML, such data is recorded in a section of the file, called the *<SpectrumIdentificationList>*, which contains a set of elements called *<SpectrumIdentificationResult>*, each storing all reported identifications from a single spectrum. One *<SpectrumIdentificationResult>* has an attribute enabling reading software to identify the spectrum that was searched (in an external file), then lists an ordered set of *<SpectrumIdentificationItem>* elements, each one being a single PSM. The key attributes of the *<SpectrumIdentificationItem>* are the calculated and experimental *m/z* values, the rank, a reference to the peptide that has been identified, and an additional set of scores or statistics, represented as CV terms (list of *<cvParam>* elements). An example PSM represented in mzIdentML (either in version 1.1 or 1.2, the representation does not change) is given in Figure 1A. The *<peptide_ref>* element contains a reference to a separate element in the file containing the *<Peptide>* object, such that if multiple PSMs identify the same peptide, the peptide details are only recorded once in the file to save space (Figure 1B).

The *<SpectrumIdentificationItem>* also has one or more *<PeptideEvidenceRef>* elements (Figure 1C), which reference to a second external object, capturing the protein sequences in which the peptide can be found (assuming a digestion with the given enzyme rules). The *<PeptideEvidence>* element also refers to the *<Peptide>* object and has a second external reference to *<DBSequence>,* which captures a protein sequence entry in the database that was searched (Figure 1D).

Protein and grouped protein results are held in a separate part of the file, called the *<ProteinDetectionList>*. A set of proteins with shared evidence are reported under *<ProteinAmbiguityGroup>*, and the evidence for a single protein accession number being identified is captured under *<ProteinDetectionHypothesis>*, which references the set of PSMs (*<SpectrumIdentificationItem>* elements) on which it is based (Figure 1E). For a complete description of the mzIdentML specification, see the original publication (3) and the PSI website (http://www.psidev.info/mzidentml).

Results and Discussion

The following section describes the implementation of new features in mzIdentML 1.2 only. Due to the complexity and diversity of analysis workflows that need to be represented in the mzIdentML 1.2 format, a "flag" was added in the top part of the file, which enables reading software to determine which, if any, new features have been added and need to be considered. This mandatory requirement is met by adding an additional CV term in the *<SpectrumIdentificationProtocol>* element depending on the type of workflow represented (Table 2).

Modification position scoring

First, to ensure that downstream software is aware that a file contains modification position scores, a CV term is added to the *<SpectrumIdentificationProtocol>* called "modification localization scoring" (MS:1002491), as shown in Figure 2. Once this term is detected in the file, the validation (and reading) software expect the following additional features to be present. First, some approaches apply a statistical threshold for accepting or rejecting that a modification position has been confidently identified, which can be reported in the *<Threshold>* element. The (re-usable) *<Peptide>* element has an attribute *via* which the residue and location of a modification can be recorded. To remain backwards compatible, we recommend that the software implementing modification scoring in mzIdentML should continue to use these attributes, populating it with the most likely modification position. A new CV term (mandatory when MS:1002491 is present in *<SpectrumIdentificationProtocol>*) must be added to every *<Modification>* element, called "modification index" (MS:1002504), where the value serves as a unique identifier (local only to the containing *<Peptide>*) to be referenced from *<SpectrumIdentificationItem>*. The modification scores (from any algorithm or scoring system) themselves can be added as CV terms with a value provided as a regular expression of four values in a defined order: *MOD_INDEX*, *SCORE*, *POSITION*, *PASS_THRESHOLD*. *MOD_INDEX* is a reference to the "modification index" identifier provided in the referenced *<Peptide>* - *<Modification>* element. This is required in case there are two or more different types of modification on the same peptide, which could otherwise not be distinguished by position alone. The MOD_INDEX thus ensures that the correct CV term for the modification being scored is referenced. *SCORE* is the score or statistical value for the given position. *POSITION* is the scored modification position with respect to the peptide sequence (where position=0 is used to indicate the N-terminus, and position=peptide length+1 is used to indicate the C-terminus). The *POSITION* can include the bar symbol '|', as a logical OR, if the score relates to multiple positions that cannot be distinguished. *PASS_THRESHOLD* holds a Boolean (true, false) value to indicate whether the modification position passes the threshold described above. If a reader of a file wishes to determine all the sites identified without ambiguity given the threshold written to the file, one could retrieve all those PSMs with modification scores having *PASS_THRESHOLD* equals true.

Where modification position scoring, and similarly peptide-level statistics (discussed below), have been performed by post-processing software rather than the initial search engine, any relevant parameters of the post-processing should be added under *<AdditionalSearchParams>,* and the software description under *<AnalysisSoftwareList>* (not shown).

MS/MS cross-linking approaches

Search engines that are able to identify cross-linked peptides report PSMs in a broadly similar manner to regular search engines. However, where an identification is made indicating that a

spectrum matches a cross-linked pair of peptides, there may be a score for the overall identification, as well as independent scores for the alpha (cross-link donor) and beta (cross-link acceptor) peptides. This arises since it is common for fragment products to be identified only from one or the other peptide chain, and thus a given result may include higher confidence in one peptide than in the other (37). To fulfil these requirements in mzIdentML 1.2, the following adaptations were made (Figure 3). First of all, the <*SpectrumIdentificationProtocol*> must contain the CV term "cross-link search" (MS:1002494) as shown in Figure 3B. Once this term is detected in the file, the validation and implementing software will expect the following features to be present. First, a mechanism has been added for relating two different <*Peptide*> elements together, using the CV terms "cross-link donor" and "cross-link acceptor" where an identical (and within-file unique) value indicates that they are grouped together (Figure 3C). The <*Modification*> element has an attribute called *monoisotopicMassDelta*, and by convention it is expected that the cross-link donor contains the complete mass delta introduced by the cross-linking reagent, and that the cross-link acceptor reports a mass shift delta of zero. As no current CV is designed for cross-linking modifications, to capture the modification masses, site specificity and common names for cross-linking reagents, we have created a new CV (XLMOD-CV) to which new terms can be added by request.

Second, a convention was also introduced within a given <*SpectrumIdentificationResult*>. There, a pair of cross-linked peptides are reported as two instances of <*SpectrumIdentificationItem*> linked together by sharing the same value for the rank attribute, and through having a shared local unique identifier as the value for the CV term "cross-link spectrum identification item" (MS:1002511), as shown in Figure 3D. If the search engine has produced a single score for the cross-linked pair, both <*SpectrumIdentificationItem*> elements must carry the identical score (same CV term name and value, as in Figure 3D), but the two chains may also have additional, independent scores if needed (not shown). Finally, mechanisms have also been developed that enable evidence derived from cross-linked peptides pairs containing differential stable isotope labels to be encoded, as well as protein interaction evidence (not shown, see the specification document for more details). This overall mechanism can be extended to report more than two peptides that are sequentially cross-linked. More complex scenarios are not supported in mzIdentML 1.2.

Peptide-level statistics

To encode peptide-level scores or statistics in mzIdentML, first, an additional CV term "peptide-level scoring" (MS:1002490) must be included in <*SpectrumIdentificationProtocol*> (Suppl. Figure 1). Second, there are various mechanisms by which a set of PSMs can be collapsed down to a peptide-level, depending on the purpose of the routine. In contexts such as genome annotation, an application may only require evidence for whether a given peptide sequence has been confidently identified regardless of its modification status, and thus different PSMs giving evidence for both modified and unmodified forms of the peptide could be grouped together. In other cases, such as providing evidence for particular PTMs, grouping of PSMs into peptides must differentiate between modification statuses. Three CV terms have been added to the PSI-MS CV: "group PSMs by sequence" (MS:1002496), "group PSMs by sequence with modifications" (MS:1002497) and "group PSMs by sequence with modifications and charge" (MS:1002498) to cover the most common scenarios. Further CV terms for other grouping mechanisms can be added on request. One of the main reasons for performing peptide-level analysis is to apply a threshold, such as 1% FDR, for selecting data for downstream analysis, which can now be added to the search protocol. In addition, as explained, a mechanism is then needed for capturing how different PSMs are grouped into a

single peptide. This is achieved by adding a CV term "peptide group ID" (MS:1002520) to every PSM (*<SpectrumIdentificationItem>*) in the file, whereby the associated value is a unique identifier shared between all PSMs in the same peptide group. In Suppl. Figure 1, the unique identifier used is the peptide sequence itself (since when grouping by sequence irrespective of the modification status, this value must be unique), although this could be any arbitrary value such as an integer code. Finally, the mzIdentML file must be able to record scores or statistical values at the peptide-level. This is achieved *via* adding CV terms with identical values to all PSMs within the same peptide-group. There are now branches within the PSI-MS CV providing different scores for PSMs and peptides from which suitable terms can be sourced. The use of peptide-level scoring and export to mzIdentML 1.2 has already been added to PeptideShaker (14) and ProteoAnnotator (19).

Encoding proteogenomics approaches

Proteogenomics data requires storing the results of mapping peptide sequences identified back onto gene models and source chromosomes potentially coming from different genome builds. This is achieved in mzIdentML 1.2 as follows. First, an additional CV term "proteogenomics search" (MS:1002635) is included in *<SpectrumIdentificationProtocol>* (Suppl. Figure 2). CV terms have been created to enable the mapping of peptides back to specific positions on chromosomes, accounting for regions where it has been inferred that a peptide is mapped across an intron boundary to different exons. CV terms related to peptide sequences (e.g. peptide coordinates, number of exons, etc) must be included in *<PeptideEvidence>* elements, whereas CV terms related to the gene model/resulting protein (genome build, chromosome name and strand) must be included in *<DBSequence>* elements (representing the database protein sequence), as indicated in Suppl. Figure 2.

Other changes in mzIdentML 1.2

Various changes have also been made in mzIdentML 1.2 and in the accompanying implementation guidelines to better accommodate four additional common use cases: pre-fractionation of samples, approaches for *de novo* sequencing of peptides, spectral library searches and the use of multiple search engines in one combined analysis. We have also significantly improved the reporting of protein-level results, derived by protein inference, which was reported in detail here (35).

Pre-fractionation. A single mzIdentML file is intended to encompass the analysis of a single sample, either as a result of a single MS run, or as the end result of multiple MS runs from the same sample where pre-fractionation has occurred. However, to simplify the reading of mzIdentML files by software, in both mzIdentML 1.1 and continued in mzIdentML 1.2, there is a restriction that only a single list of proteins (one *<ProteinDetectionList>*) can be given in one file, although multiple *<SpectrumIdentificationList>* elements can be provided. We have amended the specification document to clarify the cases where one or many mzIdentML files are expected in cases of sample pre-fractionation. In brief, where protein inference is performed over *n* lists of PSMs (one per fraction) to produce a single protein list, this should be stored in a single mzIdentML file with *n* *<SpectrumIdentificationList>* elements and a single *<ProteinDetectionList>*. If protein inference happens independently on each fraction, then *n* mzIdentML files should be used, each containing a single *<SpectrumIdentificationList>* and one *<ProteinDetectionList>*. For a more full discussion, consult the mzIdentML 1.2. specification document available from the PSI website.

Multiple search engines. It has been widely reported that there are gains in sensitivity for peptide and protein identification through the use of multiple search engines (38-40). In mzIdentML 1.1, it was already described how such approaches could be encoded and exported from software, but the resulting scheme was difficult to implement for reading software. The challenge arises since an mzIdentML 1.1 file could contain the search engine results as reported by the original search engines, as well as list of re-scored PSMs, which were used for protein inference, and constitute the "final" results of the process. As such, in mzIdentML 1.2, we have specified that there can only be a single result for each spectrum searched (i.e. the spectrum identifier is unique within the file), thus enforcing that only "final" results after performing post-processing or combination can be validly reported.

De novo sequencing. There are several software packages that aim to derive complete or near complete peptide sequences directly from the spectrum without requiring a sequence database. The mzIdentML 1.1 specifications discussed that such approaches could theoretically be supported, but relevant examples files were not produced at the time due to an apparent lack of demand. It has since become evident that supporting *de novo* results was not straightforward, as there was a mandatory requirement for every PSM reported to record one to many relationships to protein sequences (see *<PeptideEvidenceRef>* on Figure 1A and *<PeptideEvidence>* on Figure 1C). In *de novo* approaches there is no need to relate a peptide sequence to a parent protein, and as such this cardinality has been relaxed to zero to many in mzIdentML 1.2, only when the export software includes the CV term as a "flag" in the *<SpectrumIdentificationProtocol>* "de novo search" (MS:1001010). In other cases, the validation software will then report an error if relationships to one or more proteins are not recorded for any PSMs.

Spectral library searching. mzIdentML 1.2 can also support searches against pre-annotated spectral libraries. The standard case for representing PSMs is modelled with scores or statistics on *<SpectrumIdentificationItem>* referencing to a *<Peptide>* element. For sequence database searches, *<Peptide>* stores the (sequence) database entry against which a spectrum has been matched. For spectral library searches, *<Peptide>* should store a representation of the spectral library entry, annotated with any metadata about the library entry (such as confidence scores or metrics for the entry itself), with or without a peptide sequence depending on what is contained within the library (i.e. matches against previously unidentified library entries can be supported). As for *de novo* sequencing, it is not mandatory either to provide links between peptides and database proteins *(<DBSequence>* elements) from which the peptide sequence could have been derived, since these associations may be unknown.

Guide on implementing new features in mzIdentML 1.2

In this article we describe several extensions to mzIdentML, and introduce version 1.2 from version 1.1, to support use cases that were not anticipated previously. However, aside from required minor changes to cardinality in several places (a few attributes changing to become optional or mandatory), the resulting XML schema for mzIdentML 1.2 is identical to mzIdentML 1.1. As such, we anticipate that for groups who have already implemented mzIdentML, only minor changes would be required to accommodate both mzIdentML 1.1 and 1.2 files. We expect that mzIdentML 1.1 files will remain in circulation for several years. The changes made to mzIdentML 1.2 can be described as

backwards compatible, in that reading software designed for mzIdentML 1.1 should function in the vast majority of cases without adaptation. However, for new implementers of mzIdentML for export from software, we strongly encourage developers to follow the mzIdentML 1.2 guidelines.

In the case of cross-linking, this is still a relatively specialised field, and thus it would not be expected for general reading software to be able to handle the extensions described beyond general reading of PSMs at this point in time. Peptide-level statistics and modification location scoring are becoming more prevalent in proteome informatics, and thus we strongly encourage development teams to support these features for both file reading and writing.

Conclusions

The mzIdentML standard for peptide and protein identification data has been stable for around five years, and has steadily grown in use to support data interchange between software tools, as well as a data repository submission format. Here we report updates to the standard to enhance its support and usability for unanticipated requirements when the standard was initially released. We have attempted to encode these use cases without adapting the core model of the format to simplify adoption by the developer community. The PSI remains a free and open consortium of interested parties, and we encourage critical feedback, suggestions and contributions via attendance at a PSI annual meeting, conference calls or our mailing lists (see http://www.psidev.info/).

Acknowledgements

A.R.J. gratefully acknowledges funding from BBSRC [BB/K020145/1, BB/L024128/1, BB/H024654/1, BB/L005239/1]. J.A.V., Y.P.R. and T.T acknowledge funding from BBSRC [BB/K01997X/1, BB/L024225/1], Wellcome Trust [grant number WT101477MA] and EMBL core funding. G.M. and J.U. are funded by the BMBF grant de.NBI - German Network for Bioinformatics Infrastructure (FKZ 031 A 534A). M.E. is funded by PURE, a project of North Rhine-Westphalia, a federal state of Germany. R.J.C. acknowledges funding from NIH NIGMS grant 8P41GM103481. L.F. and J.R. are supported by the Wellcome Trust [grant numbers 103139, 092076, 108504, 101477]. E.W.D. acknowledges funding from NIH NIGMS grant R01GM087221 and NIH NIBIB grant U54EB020406. H.B. is supported by the Bergen Research Foundation and the Research Council of Norway. M.W. and O.K. acknowledge funding from DFG (SFB685) and from the European Union FP7 programme [PRIME-XS, grant number 262067].

References

1. Deutsch, E. W., Albar, J. P., Binz, P. A., Eisenacher, M., Jones, A. R., Mayer, G., Omenn, G. S., Orchard, S., Vizcaino, J. A., and Hermjakob, H. (2015) Development of data representation standards by the human proteome organization proteomics standards initiative. *J Am Med Inform Assoc* 22, 495-506

2. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011) mzML—a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics* 10, R110.000133

3. Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S., Selley, J., Searle, B., Shofstahl, J., Seymour, S., Julian, R., Binz, P.-A., Deutsch, E. W., Hermjakob, H., Reisinger,

F., Griss, J., Vizcaino, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Molecular & Cellular Proteomics* 11, M111.014381

4. Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., Gonzalez-Galarza, F. F., Fan, J., Bessant, C., Deutsch, E. W., Reisinger, F., Vizcaino, J. A., Medina-Aunon, J. A., Albar, J. P., Kohlbacher, O., and Jones, A. R. (2013) The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Molecular & Cellular Proteomics*, mcp.O113.028506

5. Qi, D., Lawless, C., Teleman, J., Levander, F., Holman, S. W., Hubbard, S., and Jones, A. R. (2015) Representation of selected-reaction monitoring data in the mzQuantML data standard. *Proteomics* 15, 2592-2596

6. Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q. W., Del Toro, N., Perez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaino, J. A., and Hermjakob, H. (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & cellular proteomics : MCP* 13, 2765-2775

7. Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., Jones, A. R., Binz, P.-A., Deutsch, E. W., Chambers, M., Kallhardt, M., Levander, F., Shofstahl, J., Orchard, S., Antonio Vizcaíno, J., Hermjakob, H., Stephan, C., Meyer, H. E., and Eisenacher, M. (2013) The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. *Database* 2013, bat009

8. Montecchi-Palazzi, L., Kerrien, S., Reisinger, F., Aranda, B., Jones, A. R., Martens, L., and Hermjakob, H. (2009) The PSI semantic validator: A framework to check MIAPE compliance of proteomics data. *PROTEOMICS* 9, 5112-5119

9. Ghali, F., Krishna, R., Lukasse, P., Martínez-Bartolomé, S., Reisinger, F., Hermjakob, H., Vizcaíno, J. A., and Jones, A. R. (2013) Tools (Viewer, Library and Validator) that Facilitate Use of the Peptide and Protein Identification Standard Format, Termed mzIdentML. *Molecular & Cellular Proteomics* 12, 3026-3035

10. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 5, 5277

11. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012) PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 11, M111 010587

12. Searle, B. C. (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 10, 1265-1269

13. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6, 654-661

14. Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L., and Barsnes, H. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 33, 22-24

15. Park, C. Y., Klammer, A. A., Kall, L., MacCoss, M. J., and Noble, W. S. (2008) Rapid and accurate peptide identification from tandem mass spectra. *J Proteome Res* 7, 3022-3027

16. Rost, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., Wojnar, D., Wolski, W. E., Schilling, O., Choudhary, J. S., Malmstrom, L., Aebersold, R., Reinert, K., and Kohlbacher, O. (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Meth* 13, 741-748

17. Pedersen, T., Gatto, L., and Gibb, S. (2016) mzID: An mzIdentML parser for R. R package version 1.10.2. *http://bioconductor.org/packages/release/bioc/html/mzID.html*

18. Uszkoreit, J., Maerkens, A., Perez-Riverol, Y., Meyer, H. E., Marcus, K., Stephan, C., Kohlbacher, O., and Eisenacher, M. (2015) PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J Proteome Res* 14, 2988-2997
19. Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., and Jones, A. R. (2014) ProteoAnnotator--open source proteogenomics annotation software supporting PSI standards. *Proteomics* 14, 2731-2741
20. Mayer, G., Stephan, C., Meyer, H. E., Kohl, M., Marcus, K., and Eisenacher, M. (2015) ProCon - PROteomics CONversion tool. *Journal of proteomics* 129, 56-62
21. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotech* 30, 918-920
22. Perez-Riverol, Y., Xu, Q. W., Wang, R., Uszkoreit, J., Griss, J., Sanchez, A., Reisinger, F., Csordas, A., Ternent, T., Del-Toro, N., Dianes, J. A., Eisenacher, M., Hermjakob, H., and Vizcaino, J. A. (2016) PRIDE Inspector Toolsuite: Moving Toward a Universal Visualization Tool for Proteomics Data Standard Formats and Quality Assessment of ProteomeXchange Datasets. *Mol Cell Proteomics* 15, 305-317
23. Reisinger, F., Krishna, R., Ghali, F., Ríos, D., Hermjakob, H., Antonio Vizcaíno, J., and Jones, A. R. (2012) jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. *PROTEOMICS* 12, 790-794
24. Perez-Riverol, Y., Uszkoreit, J., Sanchez, A., Ternent, T., Del Toro, N., Hermjakob, H., Vizcaino, J. A., and Wang, R. (2015) ms-data-core-api: an open-source, metadata-oriented library for computational proteomics. *Bioinformatics (Oxford, England)* 31, 2903-2905
25. Ternent, T., Csordas, A., Qi, D., Gómez-Baena, G., Beynon, R. J., Jones, A. R., Hermjakob, H., and Vizcaíno, J. A. (2014) How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics* 14, 2233-2241
26. Chalkley, R. J., and Clauser, K. R. (2012) Modification Site Localization Scoring: Strategies and Performance. *Molecular & Cellular Proteomics* 11, 3-14
27. Leitner, A., Faini, M., Stengel, F., and Aebersold, R. (2016) Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in biochemical sciences* 41, 20-32
28. Belsom, A., Schneider, M., Fischer, L., Brock, O., and Rappsilber, J. (2016) Serum Albumin Domain Structures in Human Blood Serum by Mass Spectrometry and Computational Biology. *Molecular & cellular proteomics : MCP* 15, 1105-1116
29. Sali, A., Berman, H. M., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S. K., Markley, J., Nakamura, H., Adams, P., Bonvin, A. M., Chiu, W., Peraro, M. D., Di Maio, F., Ferrin, T. E., Grunewald, K., Gutmanas, A., Henderson, R., Hummer, G., Iwasaki, K., Johnson, G., Lawson, C. L., Meiler, J., Marti-Renom, M. A., Montelione, G. T., Nilges, M., Nussinov, R., Patwardhan, A., Rappsilber, J., Read, R. J., Saibil, H., Schroder, G. F., Schwieters, C. D., Seidel, C. A., Svergun, D., Topf, M., Ulrich, E. L., Velankar, S., and Westbrook, J. D. (2015) Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure (London, England : 1993)* 23, 1156-1167
30. Nesvizhskii, A. I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat Meth* 11, 1114-1125
31. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25, 2078-2079
32. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Molecular & Cellular Proteomics* 4, 1419-1440

33. Rappsilber, J., and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends in biochemical sciences* 27, 74-78
34. Koskinen, V. R., Emery, P. A., Creasy, D. M., and Cottrell, J. S. (2011) Hierarchical Clustering of Shotgun Proteomics Data. *Molecular & Cellular Proteomics* 10
35. Seymour, S. L., Farrah, T., Binz, P. A., Chalkley, R. J., Cottrell, J. S., Searle, B. C., Tabb, D. L., Vizcaino, J. A., Prieto, G., Uszkoreit, J., Eisenacher, M., Martinez-Bartolome, S., Ghali, F., and Jones, A. R. (2014) A standardized framing for reporting protein identifications in mzIdentML 1.2. *Proteomics* 14, 2389-2399
36. Vizcaíno, J. A., Martens, L., Hermjakob, H., Julian, R. K., and Paton, N. W. (2007) The PSI formal document process and its implementation on the PSI website. *PROTEOMICS* 7, 2355-2357
37. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, A. L., and Chalkley, R. J. (2014) Matching Cross-linked Peptide Spectra: Only as Good as the Worse Identification. *Molecular & Cellular Proteomics* 13, 420-434
38. Jones, A. R., Siepen, J. A., Hubbard, S. J., and Paton, N. W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *PROTEOMICS* 9, 1220-1229
39. Searle, B. C., Turner, M., and Nesvizhskii, A. I. (2008) Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies. *J. Proteome Res.* 7, 245-253
40. Shteynberg, D., Deutsch, E. W., Lam, H., Eng, J. K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R. L., Aebersold, R., and Nesvizhskii, A. I. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Molecular & cellular proteomics : MCP* 10, M111 007690

Abbreviations

CV	Controlled Vocabulary
FDR	False Discovery Rate
MS	Mass Spectrometry
PIA	Protein Inference Algorithms
PSI	Proteomics Standards Initiative
PSM	Peptide Spectrum Match
PTM	Post-Translational Modification
XML	Extensible Markup Language
XSD	XML Schema Definition

Tables

Tool	Type	Status / Description	URL	I/E	F/C
Byonic (Protein Metrics Inc.)	Search	Byonic search engine supports mzIdentML 1.1 as an output format	http://www.proteinmetrics.com/products/byonic/	E	C
Crux	Search	Supports mzIdentML 1.1 as an output format and reads mzIdentML 1.1 to generate spectral count data	http://crux.ms/	I & E	F
IDPicker	Grouping	Version 3.x implements mzIdentML 1.1 import	https://medschool.vanderbilt.edu/msrc-bioinformatics/software	I	F
IP2	Search & Quant	*Integrated Proteomics Pipeline* supports export of results into mzIdentML 1.1	http://www.integratedproteomics.com/	E	C
Iquant	Quant	Automated pipeline for quantification by using isobaric tags; identification results are imported *via* mzIdentML 1.1	https://sourceforge.net/projects/iquant/	I	F
jmzIdentML	IO	Java API for reading and writing mzIdentML 1.1	https://github.com/PRIDE-Utilities/jmzIdentML	I & E	F
jPOST	Database	identification result files can be uploaded in mzIdentML 1.1	http://jpostdb.org/	I	F
Mascot (Matrix Science)	Search & Quant	mzIdentML version 1.1 available in Mascot version 2.4+	http://www.matrixscience.com/	E	C
MassIVE	Database	identification files can be uploaded in mzIdentML 1.1	https://massive.ucsd.edu	I	F
ms-data-core-api	IO	Java API that supports reading of PSI standard and open formats e.g. mzML, mzIdentML, mzTab, mgf and others.	https://github.com/PRIDE-Utilities/ms-data-core-api	I	F
MS-GF+	Search	Full support for exporting identification results into mzIdentML 1.1	https://omics.pnl.gov/software/ms-gf	E	F
MyriM**atch**	Search	Identifications exported in mzIdentML 1.1	https://medschool.vanderbilt.edu/msrc-bioinformatics/software	E	F
mzID package	IO	R package available through Bioconductor supporting v 1.1	http://www.bioconductor.org/packages/release/bioc/html/mzID.html	I	F
mzidLibrary	Post-processing	Routines and viewer (stats, protein inference, CSV import/export, proteogenomics) supporting v1.1 and 1.2	https://github.com/PGB-LIV/mzidlib	I & E	F
OMSSA [mzidLib]	Search	Converter from OMSSA .omx files to v1.1 or 1.2 in mzidLibrary.	https://github.com/PGB-LIV/mzidlib	E	F
OpenMS	Pipeline	mzIdentML 1.1 fully supported in release 1.9 +	https://www.openms.de/	I & E	F
PAnalyzer	Grouping	Used for protein grouping; it imports and exports mzIdentML (v1.1 and 1.2)	https://github.com/akrogp/EhuBio/wiki/Panalyzer	I & E	F
PEAKS (Bioinformatics Solutions Inc.)	Search & Quant	Native export of mzIdentML version 1.1	http://www.bioinfor.com/	E	C
PeptideShaker	Post-processing	Java stand-alone tool for the analysis and post-processing of proteomics experiments; it support mzIdentML 1.1 & 1.2	http://compomics.github.io/projects/peptide-shaker.html	I & E	F

PGA	Proteogenomics	Software for creating RNA-Seq based databases; it supports v1.1 as an input format for post-processing.	http://www.bioconductor.org/packages/devel/bioc/html/PGA.html	I	F
PIA	Grouping	Toolbox for protein inference and identification analysis; it supports mzIdentML 1.1.	https://github.com/mpc-bioinformatics/pia	I & E	F
ProteinLynx Global Server	Search & Quant	Peptide/protein identification and quantification software; it supports export to mzIdentML in version 3.0.3+	www.waters.com/waters/en_GB/ProteinLynx-Global-SERVER-(PLGS)/nav.htm?cid=513821	E	C
PRIDE	Database	mzIdentML 1.1 fully supported as an import format as part of the "complete" dataset submission pipeline	https://www.ebi.ac.uk/pride/archive/	I	F
PRIDE Inspector	Visualisation	Java stand-alone tool that can be used to visualise mzIdentML 1.1 files, independently or together with the corresponding mass spectra files (available in any open formats e.g. mzML, mzXML, mgf, dta, pkl, and apl).	https://github.com/PRIDE-Toolsuite/pride-inspector	I	F
Progenesis QI for proteomics (Waters Corp.)	Quant	Label-free quantification software can read identifications from Byonic in mzIdentML 1.1	http://www.nonlinear.com/progenesis/qi-for-proteomics/	I	C
ProteinPilot	Search & Quant	ProteinPilot 5.0+ exports search results in mzIdentML version 1.2.	https://sciex.com/products/software/proteinpilot-software	E	C
ProteinScape (Bruker)	Search & Quant	It imports search engine results other than Mascot in mzIdentML 1.1	https://www.bruker.com/products/mass-spectrometry-and-separations/ms-software/proteinscape/overview.html	I	C
SEQUEST / Proteome Discoverer (Thermo) [m2Lite / ProCon]	Search & Quant	Conversion of msf files from Proteome Discoverer to mzIdentML 1.1 *via* m2Lite or ProCon (ProCon also supports ProteinScape and Comet conversions).	https://bitbucket.org/paiyetan/m2lite/downloads/ http://www.ruhr-uni-bochum.de/mpc/software/ProCon/index.html.en	E	F*
ProteoAnnotator	Proteo-genomics	Proteogenomics software that uses mzIdentML 1.1 as its internal file format	http://www.proteoannotator.org/	E	F
ProteoWizard	IO	pepXML converter available and support for reading/writing mzIdentML 1.1	http://proteowizard.sourceforge.net/	I & E	F
Scaffold	Search & quant	Scaffold 4.0+ supports reading and writing of mzIdentML 1.1	http://www.proteomesoftware.com/products/scaffold/	I & E	C
Skyline	Quant	SRM/MRM/PRM, DIA and targeted DDA software can import mzIdentML 1.1 for spectral library construction	https://skyline.ms	I	F
TagRecon	Variant ID	Identifications exported in mzIdentML 1.1	https://medschool.vanderbilt.edu/msrc-bioinformatics/software	E	F
Trans Proteomic Pipeline [ProteoWizard]	Pipeline	pepXML to mzIdentML 1.1 converter available from ProteoWizard	http://proteowizard.sourceforge.net/	I & E	F
X!Tandem	Search	Converter from X!Tandem XML files to mzIdentML 1.1 or	https://github.com/PGB-LIV/mzidlib	E	F

[mzidLib]	1.2 as part of the mzidLibrary.

Table 1. A summary of current software available for processing mzIdentML 1.1+ files by May 2017. Tool = Tool name, followed by (Vendor) [Converter if non-native support]. Type = “Search” (Search engine), “Quant” (Quantification software), “IO” (file input / output), “Pipeline” (processing pipeline), “Grouping” (protein grouping), “Post-Processing” (post-processing routines), “Proteogenomics” (proteogenomics software), “Variant ID” (variant identification software), “Visualisation” (visualisation tool) – in all cases referring to the named tool. URL = The web address of the tool itself or the conversion utility, if mzIdentML is not natively supported. I/E = IMPORT / EXPORT functionality. F / C = Free / Commercial, F* = the converter is free but the software is not. Additional abbreviations not indicated in the main text: DDA (Data Dependent Acquisition), DIA (Data Independent Acquisition), MRM (Multiple Reaction Monitoring), PRM (Parallel Reaction Monitoring).

CV term name	Accession number	Comments / Purpose
peptide-level scoring	MS:1002490	Statistics have been performed on non-redundant peptide identifications.
modification localization scoring	MS:1002491	Scoring has been performed on the sites of peptide modification.
consensus scoring	MS:1002492	Multiple search engines have been used for peptide identification.
sample pre-fractionation	MS:1002493	The file contains the results of merged pre-fractionation analyses.
cross-link search	MS:1002494	The search engine has analysed cross-linked (and regular) peptides, using the new encoding described here.
de novo search	MS:1001010	*De novo* sequencing of peptides has been performed, meaning that 0..*n* relationships from peptides to proteins are allowed (rather than 1..*n*).
proteogenomics search	MS:1002635	Peptides have been mapped back to genome level coordinates, stored in the file.
spectral library search	MS:1001031	The identifications have been made by searching against a spectral library. 0..*n* peptides to proteins are allowed (rather than 1..n elsewhere) for cases where peptide to protein relationships are unknown in the library, or where a library entry has been identified with no known peptide sequence.
no special processing	MS:1002495	Used to indicate that none of the above features have been included in the file.

Table 2. New CV terms in the PSI-MS CV that are now mandatory within the element <*SpectrumIdentificationProtocol*>, enabling the new features in mzIdentML 1.2 to be differentiated and recognised automatically by processing software. In the file, 1…*n* of the terms MUST be present.

Figures

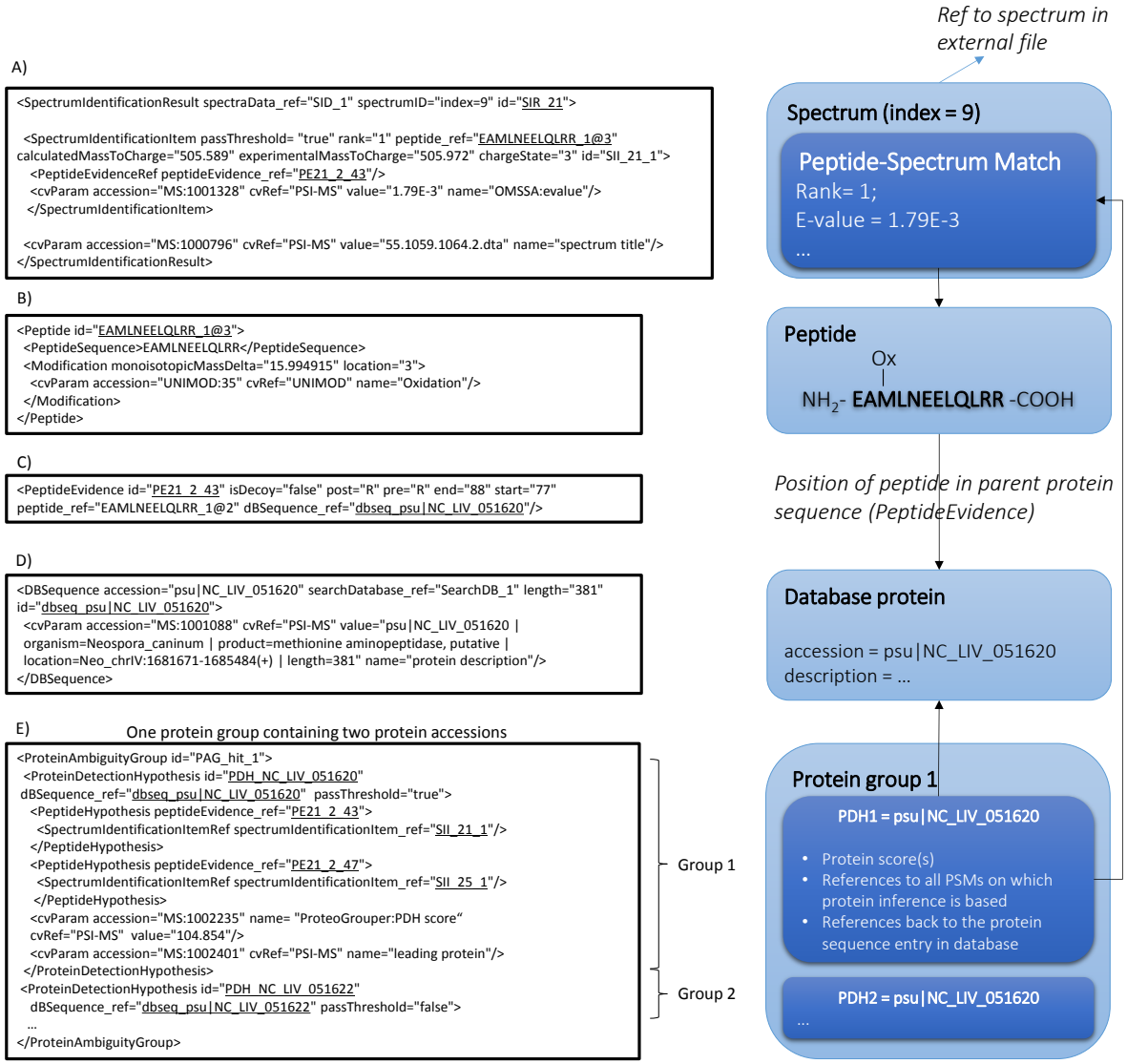

Figure 1. A) A single PSM is represented within *<SpectrumIdentificationItem>* including scores as values associated with standard terms sourced from the PSI-MS controlled vocabulary. Unique identifiers and references to other objects in the file are underlined. B) The peptide identified is stored elsewhere in the file, within the *<Peptide>* element, which can be referenced by an unlimited number of PSMs. C) All the proteins within which a peptide sequence can be located (given the enzyme specificity as defined) are linked *via* the *<PeptideEvidence>* element. D) Database proteins are represented in *<DBSequence>*. E) An identified group of proteins is stored within *<ProteinAmbiguityGroup>* ("Protein group" on the right panel) and *<ProteinDetectionHypothesis>* (PDH) - evidence at the level of a single database accession.

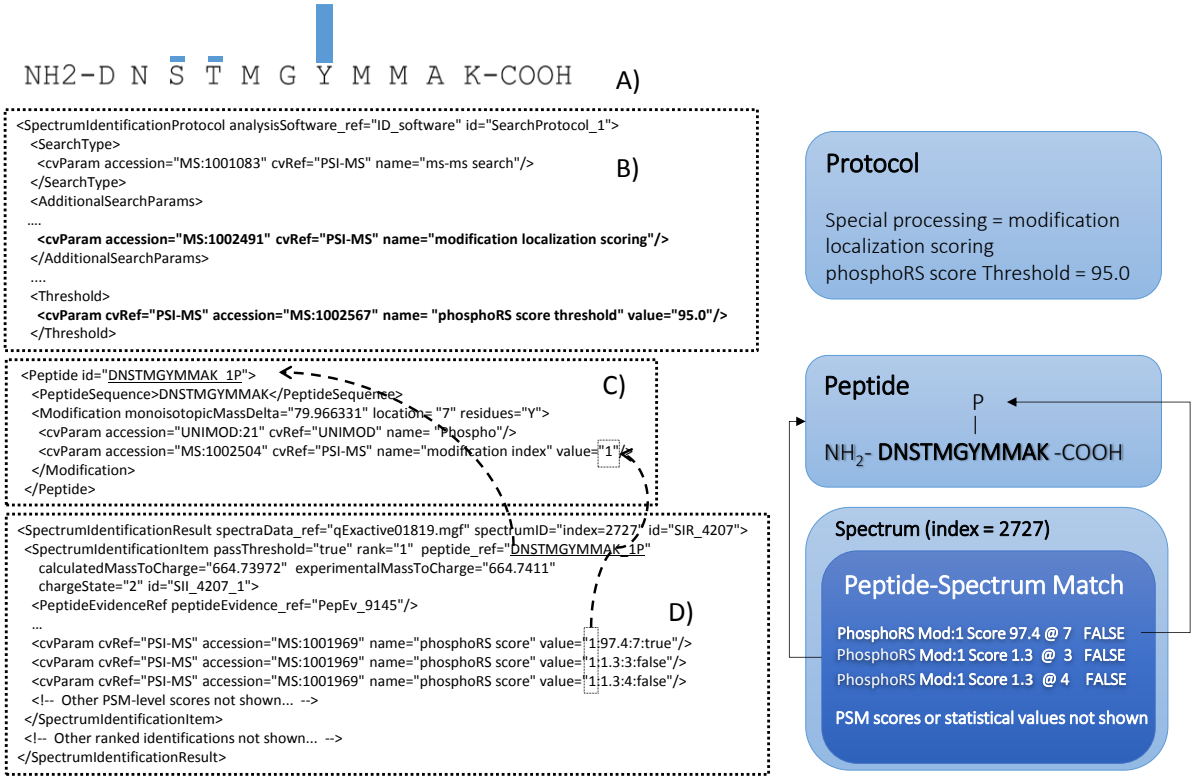

Figure 2. A) Graphical representation of the strength of evidence associated with one phosphorylation event on the peptide DNSTMGYMMAK. B) If modification re-scoring has been performed, the protocol must be flagged with the specific CV term and a threshold can be specified as to whether a given modification position has been confidently identified. C) The peptide and modification are represented in the re-usable *<Peptide>* and *<Modification>* elements. D) Modification localisation scores are included within *<SpectrumIdentificationItem>* following a given syntax: *MOD_INDEX:SCORE:POSITION:PASS_THRESHOLD*, where *MOD_INDEX* is the value referenced, allowing different modification types within a given *<Peptide>* element to be referenced, and POSITION is the position along the peptide chain (zero = N-terminus; peptide-length +1 = C-terminus).

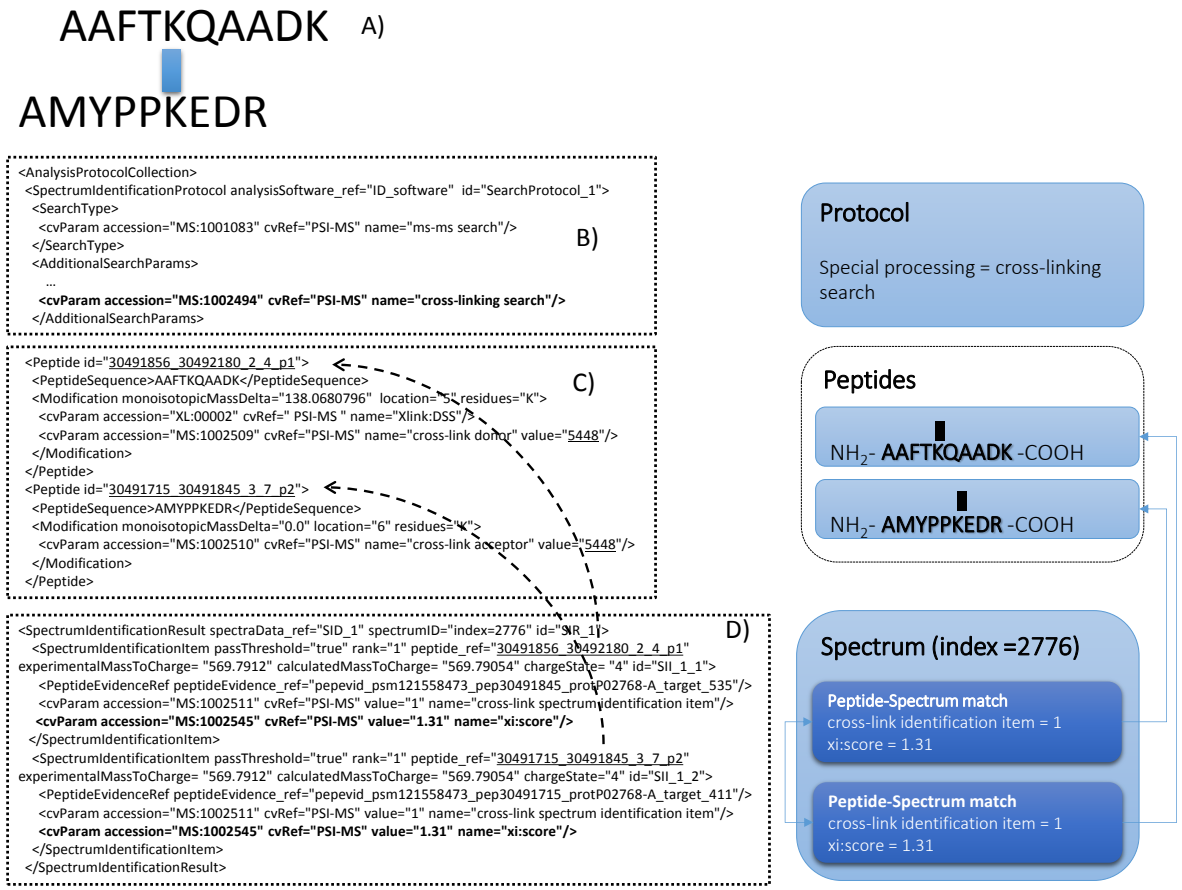

Figure 3. A) A graphical representation of the cross-linked peptide pair identified in the example. B) A specific CV term is added to the header of the file to indicate that this is a cross-linking search result set. C) The two peptide chains identified from a given spectrum are presented in a pair of *<Peptide>* and *<Modification>* elements linked via a shared, unique value in the *<cvParam>* element. The longer peptide is flagged as the cross-link donor (carrying the mass of the cross-linking reagent) and the other peptide is flagged as the cross-link acceptor with a zero mass on the *<Modification>*. D) The evidence for individual identifications is captured via two *<SpectrumIdentificationItem>* elements, which may share the same score (*<cvParam>*) for the paired identification, but may also store different, individual scores for each chain identified if appropriate (not shown).