PSI Recommendation
PSI Proteomics Informatics Workgroup
Version 1.0, release candidate 5

Johannes Griss, European Bioinformatics Institute
Timo Sachsenberg, University of Tübingen
Mathias Walzer, University of Tübingen
Oliver Kohlbacher, University of Tübingen
Andrew R. Jones, University of Liverpool
Henning Hermjakob, European Bioinformatics Institute
Juan Antonio Vizcaíno, European Bioinformatics Institute

June 1 2012
Updated: December 11, 2013

mzTab: exchange format for proteomics and metabolomics results

Status of This Document

This document presents a draft specification for the mzTab data format developed by members of the Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) Proteomics Informatics (PI) Working Group. Distribution is unlimited.

Version of This Document
The current version of this document is: version 1.0, release candidate 5, 11 December 2013.

Abstract

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. The Proteomics Informatics Working Group is developing standards for describing the results of identification and quantification processes for proteins, peptides, small molecules and protein modifications from mass spectrometry. This document defines a tab delimited text file format to report proteomics and metabolomics results.

Contents

1. Introduction

1.1 Background

This document addresses the systematic description of peptide, protein, and small molecule identification and quantification data retrieved from mass spectrometry (MS)-based experiments. A large number of software tools are available that analyze MS data and produce a variety of different output data formats. The HUPO Proteomics Standards Initiative (PSI) has developed several vendor-neutral data formats to overcome this heterogeneity of data formats for MS data. Currently, the PSI promotes the usage of three file formats to report an experiment’s data: mzML to store the pure MS data (i.e. the spectra and chromatograms), mzIdentML to store (poly)peptide identifications and potentially inferred protein identifications, and mzQuantML to store quantitative data associated with these results. All three of these formats are XML-based and require sophisticated software to access the stored data.

While full, detailed representation of MS data including provenance is essential for researchers in the field, many downstream analysis use cases are only concerned with the *results* of the experiment in an easily accessible format. In addition, there is a trend for performing more integrated experimental workflows involving both proteomics and metabolomics data. Thus, the current lack of standardization in the field of metabolomics was taken into account in the development of the format presented here, and structures were developed that can report protein, peptide, and small molecule MS based data.

mzTab is intended as a lightweight supplement to the already existing standard file formats, providing a summary, similar to the supplementary table of results of a scientific publication.

mzTab files can contain protein, peptide, and small molecule identifications together with basic quantitative information. mzTab is not intended to store an experiment's complete data / evidence but only its final reported results. This format is also intended to provide local LIMS systems as well as MS proteomics repositories a simple way to share and combine basic information.

mzTab has been developed with a view to support the following general tasks (more specific use cases are provided in Section 2):

T1. *Facilitate the sharing of final experimental results,* especially with researchers outside the field of proteomics that i) lack specialized software to parse the existing PSI's XML-based standard file formats, and ii) are only interested in the final reported results and not in all the details related to the data processing due to the inherent complexity of MS proteomics data. Furthermore, this should encourage the development of small innovative tools without the requirement of parsing huge XML files, which might be outside the scope of many bioinformaticians.

T2. *Export of results to external software,* that is not able to parse proteomics/metabolomics specific data formats but can handle simple tab-delimited file formats. As a guideline the file format is designed to be viewable by programs such as Microsoft Excel® and Open Office Spreadsheet.

T3. *Contain the results of an experiment in a single file*, and thus not require linking two files to retrieve identification and quantification results to again simplify the processing of the data.

T4. *Act as an output format of (web-) services* that report MS-based results and thus can produce standardized result pages.

T5. *Allow the combination of MS-based proteomics and metabolomics experimental results* within a single file.

T6. *Be able to link to the external experimental evidence* (i.e. the mass spectra in different formats), following the same approach used in mzIdentML and mzQuantML.

This document presents a specification, not a tutorial. As such, the presentation of technical details is deliberately direct. The role of the text is to describe the model and justify design decisions made. The document does not discuss how the models should be used in practice, consider tool support for data capture or storage, or provide comprehensive examples of the models in use. It is anticipated that tutorial material will be developed independently of this specification.

1.2 Document Structure

The remainder of this document is structured as follows. Section 2 lists use cases mzTab is designed to support. Section 3 describes the terminology used. Section 4 describes how the specification presented in Section 6 relates to other specifications, both those that it extends and those that it is intended to complement. Section 5 discusses the reasoning behind several design decisions taken. Section 6 contains the documentation of the file. Section 7 lists use cases that are currently not supported. Conclusions are presented in Section 8.

2. Use Cases for mzTab

The following cases of usage have driven the development of the mzTab data model, and are used to define the scope of the format in version 1.0.

1. mzTab files should be simple enough to make proteomics/metabolomics results accessible to people outside the respective fields. This should facilitate the sharing of data beyond the borders of the fields and make it accessible to non-experts.
2. mzTab files should contain sufficient information to provide an electronic summary of all findings in a proteomics/metabolomics study to permit its use as a standard documentation format for 'supplementary material' sections of publications in proteomics and metabolomics. It should thus be able to replace PDF tables as a way of reporting peptides and proteins and make published identification and quantification information more accessible.
3. mzTab files should enable reporting at different levels of detail: ranging from a simple summary of the final results to a detailed reporting including the experimental design. In practise, when different samples and assays (including replicates) are reported in a single mzTab file, this file can be generated in two ways: 'Summary' mode, and 'Complete' mode. In 'Summary' full results per assay/replicate need not be included, only the final data for the experimental conditions analysed must be present. In 'Complete' mode, all the results per assay/replicate need to be detailed.
4. It should be possible to open mzTab files with "standard" software such as Microsoft Excel® or Open Office Spreadsheet. This should furthermore improve the usability of the format to people outside the fields of proteomics/metabolomics.
5. It should be possible to export proteomics data from, for example, mzIdentML/ mzQuantML files into mzTab to then load this data into, for example, statistical tools such as those provided through the R programming language. With the current formats, complex conversion software would be needed to make proteomics results available to such environments.
6. mzTab files should make MS derived results easily accessible to scripting languages allowing bioinformaticians to develop software without the overhead of developing sophisticated parsing code. Since mzTab files will be comparatively small, the data from multiple experiments can be processed at once without requiring special resource management techniques.
7. It should be possible to contain the complete final results of an MS-based proteomics/metabolomics experiment in a single file. This should furthermore reduce the complexity of sharing and processing an experiment's final results. mzTab files should be able to store quantitative values for protein, peptide, and small molecule identifications. Furthermore, mzTab files should contain basic protein inference information and modification position ambiguity information. Additionally, mzTab files should be able to report merged results from multiple search engines.
8. It should be useful as an output format by web-services that can then be readily accessed by tools supporting mzTab.
9. As mzTab files only contain an experiment's core results, all entries should link back to their source. Furthermore, it should be possible to directly link a given peptide / small molecule identification to its source spectrum in an external MS data file. The same referencing system as in mzIdentML/mzQuantML should be used.

3. Notational Conventions

The key words "MUST," "MUST NOT," "REQUIRED," "SHALL," "SHALL NOT," "SHOULD," "SHOULD NOT," "RECOMMENDED," "MAY," and "OPTIONAL" are to be interpreted as described in RFC-2119 (Bradner 1997).

4. Relationship to Other Specifications

The specification described in this document has not been developed in isolation; indeed, it is designed to be complementary to, and thus used in conjunction with, several existing and emerging models. Related specifications include the following:

1. *mzML* (http://www.psidev.info/mzml). mzML is the PSI standard for capturing mass spectra / peak lists resulting from mass spectrometry in proteomics (Martens, L., *et al.* 2011). mzTab files MAY be used in conjunction with mzML, although it will be possible to use mzTab with other formats of mass spectra. This document does not assume familiarity with mzML.
2. *mzIdentML* (http://www.psidev.info/mzidentml). mzIdentML is the PSI standard for capturing of peptide and protein identification data (Jones, A. R., *et al.* 2012). mzTab files MAY reference mzIdentML files that then contain the detailed evidence of the reported identifications.
3. *mzQuantML* (http://www.psidev.info/mzquantml). mzQuantML is the PSI standard for capturing quantitative proteomics data from mass spectrometry (Walzer, M. *et al.* 2013). mzTab files that report quantitative data MAY reference mzQuantML files for detailed evidence of the reported values.

4.1 The PSI Mass Spectrometry Controlled Vocabulary (CV)

The PSI-MS controlled vocabulary is intended to provide terms for annotation of mzML, mzIdentML, and mzQuantML files. The CV has been generated with a collection of terms from software vendors and academic groups working in the area of mass spectrometry and proteome informatics. Some terms describe attributes that must be coupled with a numerical value attribute in the CvParam element (e.g. MS:1001191 “p-value”) and optionally a unit for that value (e.g. MS:1001117, “theoretical mass”, units = “dalton”). The terms that require a value are denoted by having a “datatype” key-value pair in the CV itself: MS:1001172 "mascot:expectation value" value-type:xsd:double. Terms that need to be qualified with units are denoted with a “has_units” key in the CV itself (relationship: has_units: UO:0000221 ! dalton).

As recommended by the PSI CV guidelines, psi-ms.obo should be dynamically maintained via the psidev-ms-vocab@lists.sourceforge.net mailing list that allows any user to request new terms in agreement with the community involved. Once a consensus is reached among the community the new terms are added within a few business days. If there is no obvious consensus, the CV coordinators committee should vote and make a decision. A new psi-ms.obo should then be released by updating the file on the CVS server without changing the name of the file (this would alter the propagation of the file to the OBO website and to other ontology services that rely on file stable URI). For this reason an internal version number with two decimals (x.y.z) should be increased:

- x should be increased when a first level term is renamed, added, deleted or rearranged in the structure. Such rearrangement will be rare and is very likely to have repercussion on the mapping.
- y should be increased when any other term except the first level one is altered.
- z should be increased when there is no term addition or deletion but just editing on the definitions or other minor changes.

The following ontologies or controlled vocabularies specified below may also be suitable or required in certain instances:

- Unit Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=unit)
- ChEBI (http://www.ebi.ac.uk/chebi/)
- OBI (Ontology of Biological Investigations - http://obi.sourceforge.net/)
- PSI Protein modifications workgroup - http://psidev.cvs.sourceforge.net/psidev/psi/mod/data/PSI-MOD.obo
- Unimod modifications database - http://www.unimod.org/obo/unimod.obo
- PRIDE Controlled Vocabulary (http://ebi-pride.googlecode.com/svn/trunk/pride-core/schema/pride_cv.obo)
- NEWT UniProt Taxonomy Database (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=NEWT)
- BRENDA tissue/ enzyme source (http://www.brenda-enzymes.info/ontology/tissue/tree/update/update_files/BrendaTissueOBO).
- Cell Type ontology (http://obo.cvs.sourceforge.net/obo/obo/ontology/anatomy/cell_type/cell.obo).

5. Resolved Design and scope issues

There were several issues regarding the design of the format that were not clear cut, and a design choice was made that was not completely agreeable to everyone. So that these issues do not keep coming up, we document the issues here and why the decision that is implemented was made.

5.1 Handling updates to the controlled vocabulary

There is a difficult issue with respect to how software should encode CV terms, such that changes to the core can be accommodated. This issue is discussed at length in the mzML specification document (Martens, L *et al.* 2011), and mzTab follows the same convention. In brief, when a new term is required, the file producers must contact the CV working group (via the mailing list psidev-ms-vocab@lists.sourceforge.net) and request the new term. It is anticipated that problems may arise if a consumer of the file encounters a new CV term and they are not working from the latest version of the CV file. It has been decided that rather than aim for a workaround to this issue, it can be expected that data file consumers must ensure that the OBO file is up-to-date.

5.2 Use of identifiers for input spectra to a search

PSMs and small molecules MUST be linked to an identifier of the source spectrum (in an external file) from which the identifications are made by way of a reference in the spectra_ref attribute and via the ms_run element which stores the URL of the file in the location attribute.

It is advantageous if there is a consistent system for identifying spectra in different file formats. The following table is implemented in the PSI-MS CV for providing consistent identifiers for different spectrum file formats. This is the exact same approach followed in mzIdentML and mzQuantML. *Note, this table shows examples from the CV but will be extended. The CV holds the definite specification for legal encodings of spectrumID values.*

ID	Term	Data type	Comment

MS:1000768	Thermo nativeID format	controllerType=xsd:nonNegativeInteger
		controllerNumber=xsd:positiveInteger
		scan=xsd:positiveInteger.
			controller=0 is usually the mass spectrometer

MS:1000769	Waters nativeID format	function=xsd:positiveInteger
		process=xsd:nonNegativeInteger
		scan=xsd:nonNegativeInteger	

MS:1000770	WIFF nativeID format	sample=xsd:nonNegativeInteger
		period=xsd:nonNegativeInteger
		cycle=xsd:nonNegativeInteger
		experiment=xsd:nonNegativeInteger	

MS:1000771	Bruker/Agilent YEP nativeID format	scan=xsd:nonNegativeInteger	
MS:1000772	Bruker BAF nativeID format	scan=xsd:nonNegativeInteger	
MS:1000773	Bruker FID nativeID format	file=xsd:IDREF	The nativeID must be the same as the source file ID
MS:1000774	multiple peak list nativeID format	index=xsd:nonNegativeInteger	Used for referencing peak list files with multiple spectra, i.e. MGF, PKL, merged DTA files. Index is the spectrum number in the file, starting from 0.
MS:1000775	single peak list nativeID format	file=xsd:IDREF	The nativeID must be the same as the source file ID. Used for referencing peak list files with one spectrum per file, typically in a folder of PKL or DTAs, where each sourceFileRef is different
MS:1000776	scan number only nativeID format	scan=xsd:nonNegativeInteger	Used for conversion from mzXML, or a DTA folder where native scan numbers can be derived.
MS:1000777	spectrum identifier nativeID format	spectrum=xsd:nonNegativeInteger	Used for conversion from mzData. The spectrum id attribute is referenced.
MS:1001530	mzML unique identifier	xsd:string	Used for referencing mzML. The value of the spectrum id attribute is referenced directly.

Table 1 Controlled vocabulary terms and rules implemented in the PSI-MS CV for formulating the “nativeID” to identify spectra in different file formats.

In mzTab, the spectra_ref attribute should be constructed following the data type specification in Table 1. As an example, to reference the third spectrum (index = 2) in an MGF (Mascot Generic Format) file:

```
MTD   ms_run[1]-format     [MS, MS:1001062, Mascot MGF file, ]
MTD   ms_run[1]-id_format  [MS, MS:1000774, multiple peak list nativeID format, ]
```

…

```
PSH   sequence   ...   spectra_ref               ...
PSM   NILNELFQR  ...   ms_run[1]:index=2        ...
```

Example: Reference the spectrum with identifier “scan=11665” in an mzML file.

```
MTD   ms_run[1]-format    [MS, MS:1000584, mzML file, ]
MTD   ms_run[1]-id_format  [MS, MS:1001530, mzML unique identifier, ]
```

...

```
PSH   sequence   ...   spectra_ref            ...
PSM   NILNELFQR  ...   ms_run[1]:scan=11665      ...
```

5.3 Recommendations for reporting replicates within experimental designs

Modeling the correct reporting of technical/biological replicates within experimental designs is supported in mzTab using an adaptation of the system originally developed for mzQuantML comprising four components described below (Figure 1). These components have various cross-references and MUST be used in different types of mzTab files, as described in Section 5.4:

- Study variable – The variables about which the final results of a study are reported, which may have been derived following averaging across a group of replicate measurements (assays). In files where assays are reported, study variables have references to assays. The same concept has been defined by others as “experimental factor”.
- MS run – An MS run is effectively one run (or set of runs on pre-fractionated samples) on an MS instrument, and is referenced from assay in different contexts.
- Assay – The application of a measurement about the sample (in this case through MS) – producing values about small molecules, peptides or proteins. One assay is typically mapped to one MS run in the case of label-free MS analysis or multiple assays are mapped to one MS run for multiplexed techniques, along with a description of the label or tag applied.
- Sample – a biological material that has been analysed, to which descriptors of species, cell/tissue type etc. can be attached. In all of types of mzTab file, these MAY be reported in the metadata section as sample[1-n]-description. Samples are NOT MANDATORY in mzTab, since many software packages cannot determine what type of sample was analysed (e.g. whether biological or technical replication was performed).

Clear definitions of biological and technical replicates are difficult to provide as these are somewhat dependent upon the biological domain. However, we use the following general definitions in mzTab.

- Biological replicates are where different samples have been analysed by MS.
- Technical replicates are where same samples are analysed multiple times by (LC)-MS.

Note: there is deliberately no attempt to define the boundary of the term “sample”.

If sample level information is provided optimally, it is expected that *n* biological replicates can be mapped to sample[1-n]; *m* technical replicate measurements of sample 1 SHOULD be mapped to assay[1-m] referencing sample[1] (for example). However, an open challenge remains since analysis software is often not aware of whether replicates (multiple MS runs) are originally biological or technical in nature. As such, the default behavior for mzTab exporters from quantitative software is to exclude sample level information and report quantitative data for assay[1-n] and/or study_variable[1-n] depending on whether it is a ‘Complete’ or ‘Summary’ file. Additional annotation software would typically be required to add the sample-level information, as provided (often manually) by the user.

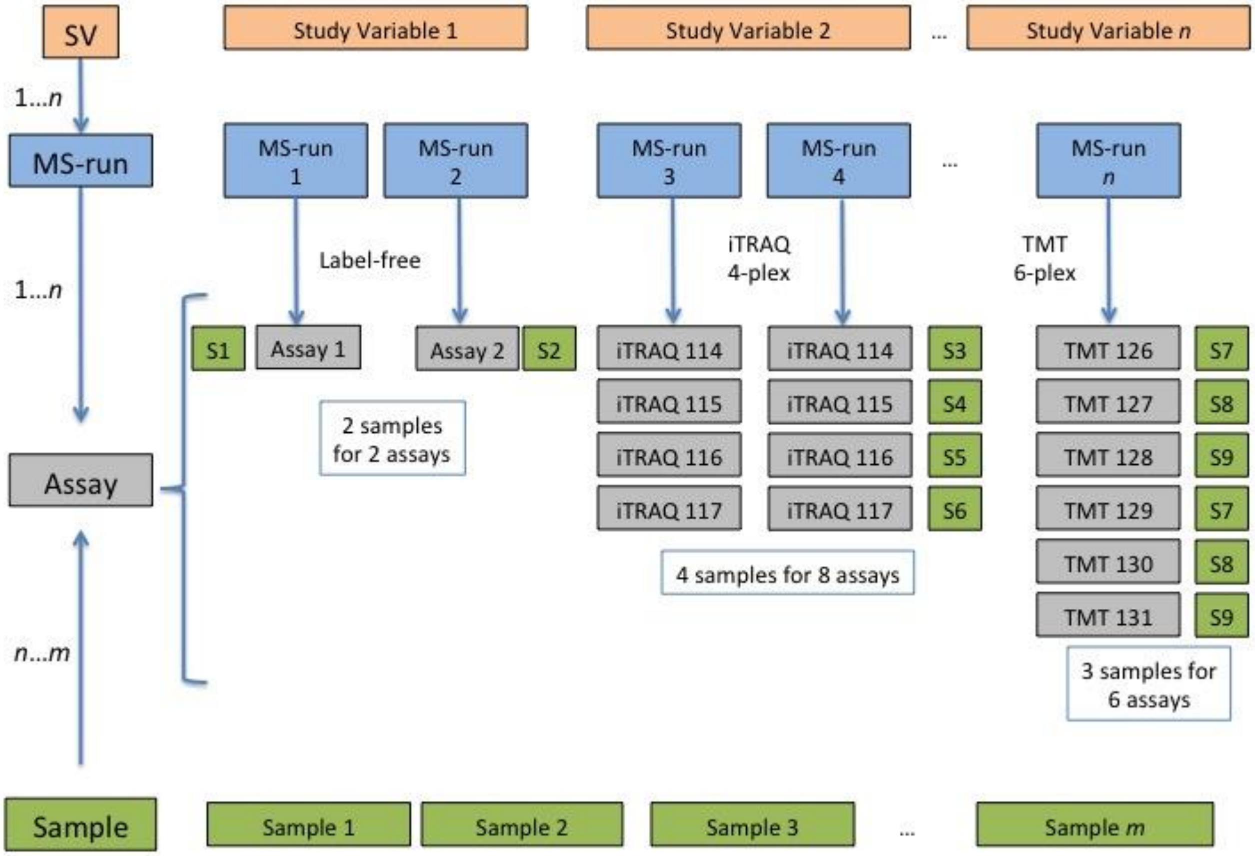

Figure 1. Diagram summarizing the relation between Study Variables (SVs), MS runs, assays and samples.

5.4 mzTab types ‘Identification’ and ‘Quantification’

There are two types of mzTab files which MUST be specified using the mandatory metadata field ‘mzTab-type’ (‘Identification’ or ‘Quantification’). ‘Identification’ MUST be used to report raw peptide, protein and small molecule identifications. The type ‘Quantification’ MUST be used for quantification results (which optionally might contain identification results about the quantified protein/peptide or small molecules). ‘Quantification’ files MUST always report quantification data on the level of study variables and MAY report quantification data on the level of assays. In contrast, ‘Identification’ files MAY contain neither study variables nor assays but only report identifications on the level of MS runs. Of course, ‘Identification’ files SHOULD include information about study variables and assays if this information is available.

Providing metadata on samples is not mandatory in both mzTab types as most software for quantification and identification can't readily export this information.

5.5 mzTab modes 'Summary' and 'Complete'

There are two modes of reporting data in mzTab files: as 'Identification' and 'Quantification' type results. The type MUST be specified by the mandatory metadata field 'mzTab-mode' ('Summary' and 'Complete'). The 'Summary' mode is used to report final results (e.g. quantification data at the level of study variables). The 'Complete' mode is used if all quantification data is provided (e.g. quantification on the assay level and on the study variable level).

The MANDATORY fields in the Metadata Section 'mzTab-mode' and 'mzTab-type' MUST therefore be present to indicate which type of file it is. In general, "null" values SHOULD not be given within any column of a "Complete" file if the information is available. Tables 2-6 indicate which metadata or columns are mandatory for a specific mzTab-mode ('Summary' and 'Complete') and mzTab-type ('Identification' and 'Quantification') in the different sections.

In general, "null" values SHOULD not be used within any column of a "Complete" file if the information is available. This is the nomenclature used in these tables:

S … required in summary file	*s* … optional in summary file
C … required in complete file	*c* … optional in complete file
SV … study variable	

Metadata Section

Field Name	Identification	Quantification
mzTab-version	**SC**	**SC**
mzTab-mode	**SC**	**SC**
mzTab-type	**SC**	**SC**
description	**SC**	**SC**
ms_run[1-n]-location	**SC**	**SC**
fixed_mod[1-n]	**SC** (if PSM section present)	**SC** (if PSM section present)
variable_mod[1-n]	**SC** (if PSM section present)	**SC** (if PSM section present)
protein-quantification-unit		**SC** (if protein section present)
peptide-quantification-unit		**SC** (if peptide section present)
smallmolecule- quantification -unit		**SC** (if small molecule section present)
study_variable[1-n]-description		**SC**
software[1-n]	_s_**C**	_s_**C**
quantification_method		_s_**C**
assay[1-n]-ms_run_ref	*sc* (required if assays reported)	_s_**C** (required if assays reported)
assay[1-n]-quantification_reagent		_s_**C**
study_variable[1-n]-assay_refs		_s_**C**
quantification_method		_s_**C**
mzTab-ID	*sc*	*sc*
title	*sc*	*sc*
sample_processing[1-n]	*sc*	*sc*
instrument[1-n]-name	*sc*	*sc*
instrument[1-n]-source	*sc*	*sc*
instrument[1-n]-analyzer	*sc*	*sc*

instrument[1-n]-detector	*sc*	*sc*
software[1-n]-setting	*sc*	*sc*
false_discovery_rate	*sc*	*sc*
publication[1-n]	*sc*	*sc*
contact-name[1-n]	*sc*	*sc*
contact-affiliation[1-n]	*sc*	*sc*
contact-email[1-n]	*sc*	*sc*
uri[1-n]	*sc*	*sc*
fixed_mod[1-n]-site	*sc*	*sc*
fixed_mod[1-n]-position	*sc*	*c*
variable_mod[1-n]-site	*sc*	*sc*
variable_mod[1-n]-position	*sc*	*sc*
ms_run[1-n]-format	*sc*	*sc*
ms_run[1-n]-id_format	*sc*	*sc*
ms_run[1-n]-fragmentation_method	*sc*	*sc*
custom[1-n]	*sc*	*sc*
sample[1-n]-species[1-n]	*sc*	*sc*
sample[1-n]-tissue[1-n]	*sc*	*sc*
sample[1-n]-cell_type[1-n]	*sc*	*sc*
sample[1-n]-disease[1-n]	*sc*	*sc*
sample[1-n]-description	*sc*	*sc*
sample[1-n]-custom[1-n]	*sc*	*sc*
assay[1-n]-sample_refs	*sc*	*sc*
study_variable[1-n]-description	*sc* (required if SV reported)	*sc* (required if SV reported)
study_variable[1-n]-sample_refs	*sc*	*sc*
study_variable[1-n]-assay_refs	*sc*	sC
assay[1-n]-quantification_mod[1-n]		*sc*
assay[1-n]-quantification_mod[1-n]-position		*sc*
assay[1-n]-quantification_mod[1-n]-site		*sc*
assay[1-n]-sample_refs		*sc*
cv[1-n]-label	*sc*	*sc*
cv[1-n]-full_name	*sc*	*sc*
cv[1-n]-version	*sc*	*sc*
cv[1-n]-url	*sc*	*sc*
colunit_protein	*sc*	*sc*
colunit_peptide	*sc*	*sc*
colunit_psm	*sc*	*sc*
colunit_small_molecule	*sc*	*sc*
mzTab-ID	*sc*	*sc*

Table 2. Mandatory and optional metadata in the Metadata section

Protein Section

Field Name	Identification	Quantification
accession	**SC**	**SC**
description	**SC**	**SC**
taxid	**SC**	**SC**
species	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
search_engine	**SC**	**SC**
best_search_engine_score	**SC**	**SC**
ambiguity_members	**SC**	**SC**
modifications	**SC**	**SC**

protein_coverage	*s***C**	*s***C**
protein_abundance_study_variable[1-n]		**SC**
protein_abundance_stdev_study_variable[1-n]		**SC**
protein_abundance_std_error_study_variable[1-n]		**SC**
search_engine_score_ms_run[1-n]	*s***C**	*s***C**
num_psms_ms_run[1-n]	*s***C**	sc
num_peptides_distinct_ms_run[1-n]	*s***C**	sc
num_peptide_unique_ms_run[1-n]	*s***C**	sc
protein_abundance_assay[1-n]		*s***C**
opt_global_*	*sc*	*sc*
go_terms	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*
num_psms_ms_run[1-n]		*sc*

Table 3. Mandatory and optional columns in the Protein section

Peptide Section (not recommended in 'Identification' files)

Field Name	Identification	Quantification
sequence		**SC**
accession		**SC**
unique		**SC**
database		**SC**
database_version		**SC**
search_engine		**SC**
best_search_engine_score		**SC**
modifications		**SC**
retention_time		**SC**
retention_time_window		**SC**
charge		**SC**
mass_to_charge		**SC**
peptide_abundance_study_variable[1-n]		**SC**
peptide_abundance_stdev_study_variable[1-n]		**SC**
peptide_abundance_std_error_study_variable[1-n]		**SC**
search_engine_score_ms_run[1-n]		*s***C**
peptide_abundance_assay[1-n]		*s***C**
spectra_ref		*s***C** (if MS2 based quantification is used)
opt_global_*		*sc*
reliability		*sc*
uri		*sc*

Table 4. Mandatory and optional columns in the Peptide section

PSM Section

Field Name	Identification	Quantification
sequence	**SC**	**SC**
PSM_ID	**SC**	**SC**
accession	**SC**	**SC**
unique	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
search_engine	**SC**	**SC**
search_engine_score	**SC**	**SC**
modifications	**SC**	**SC**

spectra_ref	**SC**	**SC**
retention_time	**SC**	**SC**
charge	**SC**	**SC**
exp_mass_to_charge	**SC**	**SC**
calc_mass_to_charge	**SC**	**SC**
pre	**SC**	**SC**
post	**SC**	**SC**
start	**SC**	**SC**
end	**SC**	**SC**
opt_global_*	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*

Table 5. Mandatory and optional columns in the PSM section

Small Molecule Section

Field Name	Identification	Quantification
identifier	**SC**	**SC**
chemical_formula	**SC**	**SC**
smiles	**SC**	**SC**
inchi_key	**SC**	**SC**
description	**SC**	**SC**
exp_mass_to_charge	**SC**	**SC**
calc_mass_to_charge	**SC**	**SC**
charge	**SC**	**SC**
retention time	**SC**	**SC**
taxid	**SC**	**SC**
species	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
spectra_ref	**SC**	**SC**
search_engine	**SC**	**SC**
best_search_engine_score	**SC**	**SC**
modifications	**SC**	**SC**
smallmolecule_abundance_assay[1-n]		**SC** (if assays reported)
smallmolecule_abundance_study_variable[1-n]		**SC** (if study vars. reported)
smallmolecule_stdev_study_variable[1-n]		**SC** (if study vars. reported)
smallmolecule_std_error_study_variable[1-n]		**SC** (if study vars. reported)
search_engine_score_ms_run[1-n]		s**C**
opt_global_*	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*

Table 6. Mandatory and optional columns in the Small Molecule section

5.6 Recommendations for reporting protein inference

There are multiple approaches to how protein inference can be reported. mzTab is designed to only hold experimental results, which in proteomics experiments can be very complex. At the same time, for downstream statistical analysis there is a need to simplify this problem. It is not possible to model detailed protein inference data without a significant level of complexity at the file format level. Therefore, it was decided to have only limited support for protein inference/grouping reporting in mzTab files. Protein entries in mzTab files contain the field ambiguity_members. The protein accessions listed in this field should identify proteins that were also identified through the same set of peptides or spectra, or proteins supported by a largely overlapping set of evidence, and could also be a viable candidate for the “true” identification of the entity reported. It is RECOMMENDED that “subset proteins” that are unlikely to have been identified SHOULD NOT be reported here. The mapping of a single peptide-spectrum match (PSM) to multiple accessions is supported through the reporting of the same PSM on multiple rows of the PSM section, as exemplified below.

```
COM Example of how protein inference is reported. Other sections and several columns are omitted.
...
PRH   accession  …   ambiguity_members               …
PRT   P14602     …   Q340U4, P16627  …
...
PSH            sequence    PSM_ID    accession       unique  …
PSM            DWYPAHSR    4         P14602          0       …
PSM            DWYPAHSR    4         Q340U4          0       …
PSM            DWYPAHSR    4         P16627          0       …
```

5.7 Recommendations for reporting quantification results

Quantitative technologies generally result in some kind of abundance measurement of the identified analyte. Several of the available techniques, furthermore, allow/require multiple similar samples to be multiplexed and analyzed in a single MS run – for example in label-based techniques, such as SILAC/N^{15} where quantification occurs on MS^1 data or in tag-based techniques, such as iTRAQ/TMT where quantification occurs in MS^2 data.

One measurement of a small molecule, peptide or protein is mapped to the concept of assay for both multiplexed techniques and label-free techniques in Complete files. Each assay MUST have a reference to the quantification reagent/label used (“unlabelled” in the label-free case and the “light” channel in SILAC/N^{15}) and each assay MUST have a reference to the ms_run[1_n] from which it originated. As such, in multiplexed techniques where *n* reagents are used within one analysis, assay[1-n] MUST reference the same ms_run.

If the data exporter wishes to report only final results for ‘Summary’ files (i.e. following averaging over replicates), then these MUST be reported as quantitative values in the columns associated with the study_variable[1-n] (e.g. protein_abundance_study_variable[1]). mzTab allows the reporting of abundance, standard deviation, and standard error for any study_variable. The unit of values in the abundance column MUST be specified in the metadata section of the mzTab file. The reported values SHOULD represent the final result of the performed data analysis. The exact meaning of the values will thus depend on the used analysis pipeline and quantitation method and is not expected to be comparable across multiple mzTab files.

See coding examples for SILAC, iTRAQ and label free approaches from the relevant example files (listed in Section 5.13).

5.8 Reporting modifications and amino acid substitutions

Modifications are defined in the meta-data section and reported in the modification columns of the protein, peptide or PSM section.

Defining modifications in the meta-data section:

The meta values “fixed_modification[1-n]” and “variable_modification[1-n]” describe all search modifications used to identify peptides and proteins of the mzTab file (e.g. carbamidomethylation, oxidation, labels/tags). This is the minimal information that MUST be provided for Complete Identification or Quantification files.

In addition, for each assay the optional meta-data assay[1-n]-quantification_mod* MAY be specified that allows to define details of modifications associated with the quantification reagent (e.g. SILAC label).

Reporting of modifications in columns of the protein, peptide and PSM sections:

Fixed modifications or modifications specified as quantification_modification in the metadata Section SHOULD NOT be reported in protein (PRT) and peptide rows (PEP). In contrast, all variable modifications plus fixed modifications like those induced by the quantification reagents MUST be reported in peptide spectrum match rows (PSM).

Modifications or substitutions are modelled using a specific modification object with the following format:

{position}{Parameter}-{Modification or Substitution identifier}|{neutral loss}

The number of modification (or substitution) objects MUST correspond to the number of identified modifications (or substitutions) on a given peptide or PSM. It is also expected that modifications SHOULD be reported for proteins using the same format. However, it is recognised that some export software may not be able to do this. If software cannot determine protein-level modifications, “null” MUST be used. If the software has determined that there are no modifications to a given protein “0” MUST be used.

{position} is mandatory. However, if it is not known (e.g. MS1 Peptide Mass Fingerprinting), ‘null’ must be used Terminal modifications in proteins and peptides MUST be reported with the position set to 0 (N-terminal) or the amino acid length +1 (C-terminal) respectively. N-terminal modifications that are specifically on one amino acid MUST still be reported at the position 0. This object allows modifications to be assigned to ambiguous locations, but only at the PSM and Peptide level. Ambiguity of modification position MUST NOT be reported at the Protein level. In that case, the modification element can be left empty. Ambiguous positions can be reported by separating the {position} and (optional) {cvParam} by an ‘|’ from the next position. Thereby, it is possible to report reliabilities / scores / probabilities etc. for every potential location.

```
Here only the modification field is given:

3-MOD:00412, 8-MOD:00412            TESTPEPTIDES with two known phosphorylation sites
3|4-MOD:00412, 8-MOD:00412          First phosphorylation site can be either on S or T
3|4|8-MOD:00412, 3|4|8-MOD:00412    Three possible positions for two phosphorylation sites
```

{Parameter} is optional. It MAY be used to report a numerical value e.g. a probability score associated with the modification or location.

```
Reporting the first two possible sites for the phosphorylation with given probability score
Here only the modification field is given:

3[MS,MS:1001876, modification probability, 0.8]|4[MS,MS:1001876, modification probability, 0.2]
MOD:00412, 8-MOD:00412
```

This option is not allowed though:

```
(3|4)[MS,MS:1001876, modification probability, 0.8]|7[MS,MS:1001876, modification probability, 0.2]-
MOD:00412
```

{Modification or Substitution identifier} for proteins and peptides modifications SHOULD be reported using either UNIMOD or PSI-MOD accessions. As these two ontologies are not applicable to small molecules, so-called CHEMMODs can also be defined. Two types of CHEMMODs are allowed: specifying a chemical formula or specifying a given *m/z* delta. Additionally, it is possible to report substitutions of amino acids using SUBST:{amino acid}. In these cases, the “sequence” column MUST contain the original, unaltered sequence. The list of allowed Modification or Substitution identifiers therefore is:

```
CHEMMOD:+NH4
CHEMMOD:-18.0913
UNIMOD:18
MOD:00815
SUBST:{amino acid}
```

CHEMMODs SHOULD NOT be used for protein/peptide modifications if the respective entry is present in either the PSI-MOD or the UNIMOD ontology. Furthermore, mass deltas SHOULD NOT be reported if the given delta can be expressed through a known and unambiguous chemical formula.

All (identified) variable modifications as well as fixed modifications MUST be reported for every identification.

{neutral loss} is optional. Neutral losses are reported as cvParams. They are reported in the same way that modification objects are (as separate, comma-separated objects in the modification column). The position for a neutral loss MAY be reported.

```
PEH  sequence           … modifications                                                              …
COM  Phosphorylation with a neutral loss:
PEP  EISILACEIR         … 3-UNIMOD:21,3-[MS, MS:1001524, fragment neutral loss, 63.998285],7-UNIMOD:4
…
COM  Neutral loss without an associated modification:
PEP  EISILACEIR         … [MS, MS:1001524, fragment neutral loss, 63.998285],7-UNIMOD:4              …
```

5.9 Encoding missing values, zeroes, nulls, infinity and calculation errors

In the table-based sections (protein, peptide, and small molecule) there MUST NOT be any empty cells. In case a given property is not available “null” MUST be used. This is, for example, the case when a URI is not available for a given protein (*i.e.* the table cell MUST NOT be empty but “null” has to be reported). If ratios are included and the denominator is zero, the “INF” value MUST be used. If the result leads to calculation errors (for example 0/0), this MUST be reported as “not a number” (“NaN”). In some cases, there is ambiguity with respect to these cases: e.g. in spectral counting if no peptide spectrum matches are observed for a given protein, it is open for debate as to whether its abundance is zero or missing (“null”).

5.10 Number of peptides reported

There are columns allowed in the protein section to report the number of peptides supporting a given protein identification, which are MANDATORY for Complete Identification files.

- num_psms_ms_run[1_n]
  - The count of the total significant PSMs that can be mapped to the reported protein
- num_peptides_distinct_ms_run[1_n]
  - The count of the number of different peptide sequences that have been identified above the significance threshold. Different modifications or charge states of the same peptide are not counted.
- num_peptides_unique_ms_run[1_n]
  - The number of peptides that can be mapped uniquely to the protein reported. If ambiguity members have been reported, the count MUST be derived from the number of peptides that can be uniquely mapped to the group of accessions, since the assumption is that these accessions are supported by the same evidence.

The idea of these three columns is to give the researcher a quick overview of how well a given protein identification is supported by peptide identifications for a given ms_run reported. The num_psms column also provides the opportunity for reporting pseudo-quantitative (label-free) values from approaches in which no explicit quantification has been performed.

5.11 Reliability score

All protein, peptide, psm and small molecule identifications reported in an mzTab file MAY be assigned a reliability score (column “reliability” in all tables). This reliability only applies to the identification reliability but not to modification position and or quantification reliabilities. The idea is to provide a way for researchers and/or MS proteomics or metabolomics repositories to score the reported identifications based on their own criteria. This score is completely resource-dependent and MUST NOT be interpreted as a comparable score between mzTab files generated from different resources. The criteria used to generate this score SHOULD be documented by the data providers. If this information is not provided by the producers of mzTab files, “null“ MUST be provided as the value for each of the protein, peptide or small molecule identification.

The reliability value, if provided, MUST be an integer between 1-3 in all but the *small molecule* section (see below) and SHOULD be interpreted as follows:

1: high reliability
2: medium reliability
3: poor reliability

For metabolomics (*small molecule* section), according to current MSI agreement, it should be reported as an integer between 1-4 and should be interpreted as follows:

1: identified metabolites
2: putatively annotated compounds
3: putatively characterized compound classes
4: unknown compounds

The idea behind this score was to mimic the general concept of “resource based trust”. For example, if one resource reports identifications with a given reliability this would be interpreted differently as an identification reported from another resource – depending on who is responsible for the given resource and how it is built. If resources now report their reliabilities using this metric and document how this metric is generated, a user can base his own interpretation of the results based on his trust in the resource. Furthermore, approaches to make various search engine scores comparable have failed so far. To prevent the notion that the reported scores represent comparable probabilities this very abstract metric was chosen. Resources MUST explicitly specify how these reliability scores are calculated and what metric they represent.

5.12 Comments on Specific Use Cases

Many special use cases for mzTab were considered during its development. Each of these use cases has a corresponding example file that exercises the relevant part of the format and provides a reference implementation example (see supporting documentation). Authors of software that create mzTab are encouraged to examine the examples that accompany this format release before implementing the writer.

5.12.1 Multiple database search engines

Proteomics groups now commonly analyze MS data using multiple search engines and combine results to improve the number of peptide and protein identifications that can be made. The output of such approaches can be represented in mzTab as follows: mzTab files SHOULD only contain the “final” protein list generated by any such workflow. Any protein, peptide, and small molecule can be associated with any number of search engines as well as multiple search engine scores. Thus, it is possible to report which element was identified by which search engine together with the resulting scores.

5.12.2 Adding optional columns

Additional columns MAY be added to the end of rows in all the table-based sections (protein, peptide, PSM and small molecule). These columns represent information not included by default in the currently defined fields and differ from the specification of optionality with regards to columns that MUST be present in Summary or Complete files (Tables 2 and 3).

These column headers MUST start with the prefix “opt_” followed by the identifier of the object they reference: assay, study variable, MS run or “global” (if the value relates to all replicates). Column names MUST only contain the following characters: ‘A’-‘Z’, ‘a’-‘z’, ‘0’-‘9’, ‘_’, ‘-’, ‘[’, ‘]’, and ‘:’. CV parameter accessions MAY be used for optional columns following the format: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘_’.

The information stored within an optional column is completely up to the resource that generates the file. It MUST not be assumed that optional columns having the same name in different mzTab files contain the same type of information. CV parameter accessions MAY be used as optional column names according to the following convention: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘_’.

```
COM   Example showing how emPAI values are reported in an additional column from MS run 1 using
COM   MS CV parameter “emPAI value” (MS:1001905)
…
PRH   accession  …   opt_ms_run[1]_cv_MS:1001905_emPAI_value
PRT   P12345     …   0.658
```

5.12.3 Referencing external resources (i.e. mzIdentML or mzQuantML files)

In mzTab all identifications MAY reference external resources that contain detailed evidence for the identification. This link is stored in the “uri” column of the respective table. This field MUST NOT be used to reference an external MS data file. MS data files should be referenced using the method described in Section 5.2.

Where these URIs point to depends on the resource that generated the mzTab file. If, for example, PeptideAtlas was exporting data in the mzTab format the URI would be expected to point to the identification’s entry within the respective PeptideAtlas build. mzTab files originating from an mzIdentML file MAY reference the mzIdentML file using the URI column. In case quantitative values are reported coming from an mzQuantML file, the mzQuantML file SHOULD be referenced as it contains the reference to the underlying mzIdentML file.

5.12.4 Reporting sequence ambiguity

In MS based proteomics approaches, some amino acids cannot be unambiguously identified. To report such ambiguous amino acid identifications, the following symbols SHOULD be used:

Asparagine or aspartic acid	B
Glutamine or glutamic acid	Z
Leucine or Isoleucine	J
Unspecified or unknown amino acid	X

5.12.5 Reporting decoy peptide identifications

To report the results of a target-decoy search, decoy identifications MAY be labeled using the optional column “opt_global_cv_MS:1002217_decoy_peptide”. The value of this column MUST be a Boolean (1/0).

5.13 Other supporting materials

The following example instance documents are available and between them cover all the use cases supported.

All example files can be downloaded from:
http://code.google.com/p/mztab/wiki/ExampleFiles

a) SILAC_CQI.mzTab - (hand crafted) Report of a minimal "Complete Quantification report" SILAC experiment, quantification on 2 study variables (control/treatment), 3+3 assays (replicates) reported, identifications reported.
b) iTRAQ_CQI.mzTab - (hand crafted) Report of a minimal "Complete Quantification report" iTRAQ experiment, quantification on 4 study variables (t=0, t=1, t=2, t=3), 4*3 assays (3 replicate experiments) reported, identifications reported.
c) labelfree_CQI.mzTab – (hand crafted) Report of a minimal "Complete Quantification report" label free experiment, quantification on 2 study variables (control/treatment), 3+3 assays (replicates) reported, identifications reported.
d) protein_SQ.mzTab – (hand crafted) Report of a minimal "Summary Quantification report" experiment, quantification on 2 study variables (control/treatment), no assays (replicates) reported, no identifications reported.
e) protein_CQ.mzTab – (hand crafted) Report of a minimal "Complete Quantification report" SILAC experiment, quantification on 2 study variables (control/treatment), 3+3 assays (replicates) reported, no identifications reported.

f) peptide_SQ.mzTab – (hand crafted) Report of a minimal "Summary Quantification report" experiment, quantification on 2 study variables (control/treatment), no assays (replicates) reported, no identifications reported.
g) PSM_SQ.mzTab – (hand crafted) Report of a minimal "Summary Identification report" with PSMs only.
h) protein_and_PSM_SI.mzTab – Report of a "Summary Identification report" with protein identification and PSMs
i) PRIDE_Exp_Complete_Ac_16649.xml-mztab.txt - file generated using the mztab-exporter (converted PRIDE experiment accession 16649) containing iTRAQ data.
j) lipidomics-HFD-LD-study-TG.mzTab – File generated by the LipidDataAnalyzer (LDA) mzTab export for small molecules. Report of a "Complete Quanification report" lipidomics experiment for the lipid class TG. Quantification on 3 study variables (HFD/FED/FAS), 6+6+6 assays (biological replicates) reported, identifications reported.
k) lipidomics-HFD-LD-study-PL-DG-SM.mzTab – File generated by the LDA mzTab export for small molecules. Report of a "Complete Quanification report" lipidomics experiment for the lipid classes SM, PE, PC, LPC, DG, PS. Quantification on 3 study variables (HFD/FED/FAS), 6+6+6 assays (biological replicates) reported, identifications reported.
l) MaxQuant_SILAC.mztab – MaxQuant example generated by the MaxQuant mzTab exporter. Two B-cell lymphoma cell lines. File Quantification of a subset of two B-cell lymphoma cell lines using the Super SILAC approach measured as single shots in three replicates.
m) Cytidine.mzTab – File generated manually. It describes the identification of cytidine.

6. Format specification

This section describes the structure of an mzTab file.

- **Field separator**
  The column delimiter is the Unicode Horizontal Tab character (Unicode codepoint 0009).
- **File encoding**
  The UTF-8 encoding of the Unicode character set is the preferred encoding for mzTab files. However, parsers should be able to recognize commonly used encodings.
- **Case sensitivity**
  All column labels and field names are case-sensitive.
- **Line prefix**
  Every line in an mzTab file MUST start with a three letter code identifying the type of line delimited by a Tab character. The three letter codes are as follows:
  - MTD for metadata
  - PRH for the protein table header line (the column labels)
  - PRT for rows of the protein table
  - PEH for the peptide table header line (the column labels)
  - PEP for rows of the peptide table
  - PSH for the PSM table header (the column labels)
  - PSM for rows of the PSM table
  - SMH for small molecule table header line (the column labels)
  - SML for rows of the small molecule table
  - COM for comment lines
- **Header lines**
  Each table based section (protein, peptide, PSM and small molecule) MUST start with

the corresponding header line. These header lines MUST only occur once in the document since each section also MUST only occur once.

- **Dates**
  Dates and times MUST be supplied in the ISO 8601 format (“YYYY-MM-DD”, “YYYY-MM-DDTHH:MMZ” respectively).
- **Decimal separator**
  In mzTab files the dot (“.”) MUST be used as decimal separator. Thousand separators MUST NOT be used in mzTab files.
- **Comment lines and empty lines**
  Comment lines can be placed anywhere in an mzTab file. These lines must start with the three-letter code COM and are ignored by most parsers. Empty lines can also occur anywhere in an mzTab file and are ignored.
- **Params**
  mzTab makes use of CV parameters. As mzTab is expected to be used in several experimental environments where parameters might not yet be available for the generated scores etc. all parameters can either report CV parameters or user parameters that only contain a name and a value.
  Parameters are always reported as [CV label, accession, name, value]. Any field that is not available MUST be left empty.

```
[MS, MS:1001207, Mascot,]
[MS, MS:1001171, Mascot:score, 40.21]
[,,A user parameter, The value]
```

In case, the name of the param contains commas, quotes MUST be added to avoid problems with the parsing: [label, accession, “first part of the param name , second part of the name”, value].

```
[MOD, MOD:00648, “N,O-diacetylated L-serine”,]
```

- **Sample IDs**
  To be able to supply metadata specific to each sample, ids in the format sample[1-n] are used.

```
MTD    sample[1]-species[1]    [NEWT, 9606, Homo sapiens (Human), ]
```

- **Assay IDs**
  To be able to supply metadata specific to each assay, ids in the format assay[1-n] are used.

```
MTD    assay[1]-quantification_reagent      [MS,MS:1002038,unlabeled sample,]
```

- **Study variable IDs**
  To be able to supply metadata specific to each study variable (grouping of assays), ids in the format study_variable[1-n] are used.

```
MTD    study_variable[1]-description Group B (spike-in 0.74 fmol/uL)
```

6.1 Sections

mzTab files can contain five different sections. The MANDATORY metadata section is made up of key-value pairs. The other four sections are OPTIONAL: protein, peptide, PSM and small molecule section are table-based.

Every section in an mzTab file MUST only occur once if present. If the PSM, Peptide and Protein Sections are present, the information MUST be consistent between these sections. Field names with indices in square brackets MUST be numbered sequentially and non-decreasing (starting at the first value indicated in the bracket; single integer steps).

6.2 Metadata Section

The metadata section can provide additional information about the dataset(s) reported in the mzTab file. All fields in the metadata section are optional apart from five exceptions:

- “mzTab-version” MUST always be reported.
- “mzTab-mode” MUST always be reported. Two modes are possible: ‘Summary’ and ‘Complete’.
- “mzTab-type” MUST always be reported. Two types are possible: ‘Quantification’ or ‘Identification’. Any analyses generating both quantification and identification results MUST be flagged as ‘Quantification’.
- “description” MUST always be reported.
- “ms_run-location[1-n]” MUST always be reported.

In addition, various other metadata parameters are REQUIRED for different file types, as defined above and in Tables 2-6.

The fields in the metadata section should be reported in order of the various fields listed here. The field’s name and value MUST be separated by a tab character:

```
MTD    publication  [PRIDE, PRIDE:00000029, PubMed, 12345]
```

In the following list of fields any term encapsulated by {} is meant as a variable which MUST be replaced accordingly.

6.2.1 mzTab-version

<table>
<tr><td><b>Description:</b></td><td>The version of the mzTab file.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   mzTab-version   1.0</code></td></tr>
</table>

6.2.2 mzTab-mode

<table>
<tr><td><b>Description:</b></td><td>The results included in an mzTab file can be reported in 2 ways: ‘Complete’ (when results for each assay/replicate are included) and ‘Summary’, when only the most representative results are reported.</td></tr>
<tr><td><b>Type:</b></td><td>Enum</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   mzTab-mode   Complete</code>
<code>MTD   mzTab-mode   Summary</code></td></tr>
</table>

6.2.3 mzTab-type

<table>
<tr><td><b>Description:</b></td><td>The results included in an mzTab file MUST be flagged as ‘Identification’ or ‘Quantification’ - the latter encompassing approaches that are quantification only or quantification and identification.</td></tr>
</table>

<table>
<tr><td><b>Type:</b></td><td>Enum</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD   mzTab-type   Quantification
MTD   mzTab-type   Identification</td></tr>
</table>

6.2.4 mzTab-ID

<table>
<tr><td><b>Description:</b></td><td>The ID of the mzTab file.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD   mzTab-ID PRIDE_1234</td></tr>
</table>

6.2.5 title

<table>
<tr><td><b>Description:</b></td><td>The file’s human readable title.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD   title   My first test experiment</td></tr>
</table>

6.2.6 description

<table>
<tr><td><b>Description:</b></td><td>The file’s human readable description.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD   description   An experiment investigating the effects of Il-6.</td></tr>
</table>

6.2.7 sample_processing[1-n]

<table>
<tr><td><b>Description:</b></td><td>A list of parameters describing a sample processing step. The order of the data_processing items should reflect the order these processing steps were performed in. If multiple parameters are given for a step these MUST be separated by a “|”.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD sample_processing[1] [SEP, SEP:00173, SDS PAGE,]
MTD sample_processing[2] [SEP, SEP:00142, enzyme digestion,]|[MS, …
MS:1001251, Trypsin,]</td></tr>
</table>

6.2.8 instrument[1-n]-name

<table>
<tr><td><b>Description:</b></td><td>The name of the instrument used in the experiment. Multiple instruments are numbered 1..n.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>MTD instrument[1]-name [MS, MS:1000449, LTQ Orbitrap,]
…
MTD instrument[2]-name [MS, MS:1000031, Instrument model, name of the instrument not included in the CV]</td></tr>
</table>

6.2.9 instrument[1-n]-source

<table>
<tr><td><b>Description:</b></td><td>The instrument's source used in the experiment. Multiple instruments are numbered 1..n.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD instrument[1]-source [MS, MS:1000073, ESI,]</code>
…
<code>MTD instrument[2]-source [MS, MS:1000598, ETD,]</code></td></tr>
</table>

6.2.10 instrument[1-n]-analyzer

<table>
<tr><td><b>Description:</b></td><td>The instrument’s analyzer type used in the experiment. Multiple instruments are enumerated 1..n.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD instrument[1]-analyzer [MS, MS:1000291, linear ion trap,]</code>
…
<code>MTD instrument[2]-analyzer [MS, MS:1000484, orbitrap,]</code></td></tr>
</table>

6.2.11 instrument[1-n]-detector

<table>
<tr><td><b>Description:</b></td><td>The instrument's detector type used in the experiment. Multiple instruments are numbered 1..n.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD instrument[1]-detector [MS, MS:1000253, electron multiplier,]</code>
…
<code>MTD instrument[2]-detector [MS, MS:1000348, focal plane collector,]</code></td></tr>
</table>

6.2.12 software[1-n]

<table>
<tr><td><b>Description:</b></td><td>Software used to analyze the data and obtain the reported results. The parameter’s value SHOULD contain the software’s version. The order (numbering) should reflect the order in which the tools were used.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD software[1] [MS, MS:1001207, Mascot, 2.3]</code>
<code>MTD software[2] [MS, MS:1001561, Scaffold, 1.0]</code></td></tr>
</table>

6.2.13 software[1-n]-setting[1-n]

<table>
<tr><td><b>Description:</b></td><td>A software setting used. This field MAY occur multiple times for a single software. The value of this field is deliberately set as a String, since there currently do not exist cvParams for every possible setting.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD software[1]-setting Fragment tolerance = 0.1 Da</code>
<code>MTD software[2]-setting Parent tolerance = 0.5 Da</code></td></tr>
</table>

6.2.14 false_discovery_rate

<table>
<tr><td><b>Description:</b></td><td>The file’s false discovery rate(s) reported at the PSM, peptide, and/or protein</td></tr>
</table>

<table>
<tr><td></td><td>level. False Localization Rate (FLD) for the reporting of modifications can also be reported here. Multiple parameters MUST be separated by “|”.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD false_discovery_rate [MS, MS:1001364, pep:global FDR, 0.01]|…</code>
<code>[MS, MS:1001214, prot:global FDR, 0.08]</code></td></tr>
</table>

6.2.15 publication[1-n]

<table>
<tr><td><b>Description:</b></td><td>A publication associated with this file. Several publications can be given by indicating the number in the square brackets after “publication”. PubMed ids must be prefixed by “pubmed:”, DOIs by “doi:”. Multiple identifiers MUST be separated by “|”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD publication[1] pubmed:21063943|doi:10.1007/978-1-60761-987-1_6</code>
<code>MTD publication[2] pubmed:20615486|doi:10.1016/j.jprot.2010.06.008</code></td></tr>
</table>

6.2.16 contact[1-n]-name

<table>
<tr><td><b>Description:</b></td><td>The contact's name. Several contacts can be given by indicating the number in the square brackets after "contact". A contact has to be supplied in the format [first name] [initials] [last name] (see example).</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD contact[1]-name James D. Watson</code>
<code>…</code>
<code>MTD contact[2]-name Francis Crick</code></td></tr>
</table>

6.2.17 contact[1-n]-affiliation

<table>
<tr><td><b>Description:</b></td><td>The contact’s affiliation.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD contact[1]-affiliation Cambridge University, UK</code>
<code>MTD contact[2]-affiliation Cambridge University, UK</code></td></tr>
</table>

6.2.18 contact[1-n]-email

<table>
<tr><td><b>Description:</b></td><td>The contact’s e-mail address.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD contact[1]-email watson@cam.ac.uk</code>
<code>…</code>
<code>MTD contact[2]-email crick@cam.ac.uk</code></td></tr>
</table>

6.2.19 uri[1-n]

<table>
<tr><td><b>Description:</b></td><td>A URI pointing to the file's source data (e.g., a PRIDE experiment or a PeptideAtlas build).</td></tr>
<tr><td><b>Type:</b></td><td>URI</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr></table></td></tr>
</table>

Identification		

Example:

```
MTD  uri[1]  http://www.ebi.ac.uk/pride/url/to/experiment
MTD  uri[2]  http://proteomecentral.proteomexchange.org/cgi/GetDataset
```

6.2.20 fixed_mod[1-n]

Description: A parameter describing a fixed modifications searched for. Multiple fixed modifications are numbered 1..n.

Type: Parameter

Mandatory

	Summary	Complete
Quantification	(✓)[1]	✓
Identification	(✓)[1]	✓

[1] mandatory if PSM section is present

Example:

```
MTD  fixed_mod[1]  [UNIMOD, UNIMOD:4, Carbamidomethyl, ]
MTD  fixed_mod[2]  [UNIMOD, UNIMOD:35, Oxidation, ]
```

6.2.21 fixed_mod[1-n]-site

Description: A string describing a fixed modifications site. Following the unimod convention, modification site is a residue (e.g. “M”), terminus (“N-term” or “C-term”) or both (e.g. “N-term Q” or “C-term K”).

Type: String

Mandatory

	Summary	Complete
Quantification		
Identification		

Example:

```
MTD  fixed_mod[1]  [UNIMOD, UNIMOD:35, Oxidation, ]
MTD  fixed_mod[1]-site  M
…
MTD  fixed_mod[2] [UNIMOD, UNIMOD:1, Acetyl, ]
MTD  fixed_mod[2]-site  N-term
…
MTD  fixed_mod[3]  [UNIMOD, UNIMOD:2, Amidated, ]
MTD  fixed_mod[3]-site  C-term
```

6.2.22 fixed_mod[1-n]-position

Description: A string describing the term specifity of a fixed modification. Following the unimod convention, term specifity is denoted by the strings “Anywhere”, “Any N-term”, “Any C-term”, “Protein N-term”, “Protein C-term”.

Type: String

Mandatory

	Summary	Complete
Quantification		
Identification		

Example:

```
MTD  fixed_mod[1]  [UNIMOD, UNIMOD:35, Oxidation, ]
MTD  fixed_mod[1]-site  M
…
MTD  fixed_mod[2] [UNIMOD, UNIMOD:1, Acetyl, ]
MTD  fixed_mod[2]-site  N-term
MTD  fixed_mod[2]-position  Protein N-term
…
MTD  fixed_mod[3]  [UNIMOD, UNIMOD:2, Amidated, ]
MTD  fixed_mod[3]-site  C-term
MTD  fixed_mod[3]-position  Protein C-term
```

6.2.23 variable_mod[1-n]

Description: A parameter describing a variable modifications searched for. Multiple variable modifications are numbered 1.. n.

Type: Parameter

Mandatory

	Summary	Complete
Quantification	(✓)[1]	✓
Identification	(✓)[1]	✓

[1] mandatory if PSM section is present

Example:

```
MTD  variable_mod[1]  [UNIMOD, UNIMOD:21, Phospho, ]
MTD  variable_mod[1]  [UNIMOD, UNIMOD:35, Oxidation, ]
```

6.2.24 variable_mod[1-n]-site

Description:	A string describing a variable modifications site. Following the unimod convention, modification site is a residue (e.g. "M"), terminus ("N-term" or "C-term") or both (e.g. "N-term Q" or "C-term K").
Type:	String

Mandatory

	Summary	Complete
Quantification		
Identification		

Example:

```
MTD  variable_mod[1]  [UNIMOD, UNIMOD:35, Oxidation, ]
MTD  variable_mod[1]-site  M
…
MTD  variable_mod[2] [UNIMOD, UNIMOD:1, Acetyl, ]
MTD  variable_mod[2]-site  N-term
…
MTD  variable_mod[3]  [UNIMOD, UNIMOD:2, Amidated, ]
MTD  variable_mod[3]-site  C-term
```

6.2.25 variable_mod[1-n]-position

Description:	A string describing the term specifity of a variable modification. Following the unimod convention, term specifity is denoted by the strings "Anywhere", "Any N-term", "Any C-term", "Protein N-term", "Protein C-term".
Type:	String

Mandatory

	Summary	Complete
Quantification		
Identification		

Example:

```
MTD  variable_mod[1]  [UNIMOD, UNIMOD:35, Oxidation, ]
MTD  variable_mod[1]-site  M
…
MTD  variable_mod[2] [UNIMOD, UNIMOD:1, Acetyl, ]
MTD  variable_mod[2]-site  N-term
MTD  variable_mod[2]-position  Protein N-term
…
MTD  variable_mod[3]  [UNIMOD, UNIMOD:2, Amidated, ]
MTD  variable_mod[3]-site  C-term
MTD  variable_mod[3]-position  Protein C-term
```

6.2.26 quantification_method

Description:	The quantification method used in the experiment reported in the file.
Type:	Parameter

Mandatory

	Summary	Complete
Quantification		✓
Identification		

Example:

```
MTD  quantification_method  [MS, MS:1001837, iTRAQ quantitation analysis, ]
```

6.2.27 protein-quantification_unit

Description:	Defines what type of units is reported in the protein quantification fields.
Type:	Parameter

Mandatory

	Summary	Complete
Quantification	(✓)[1]	(✓)[1]
Identification		

[1] mandatory if protein section is present

Example:

```
MTD  protein-quantification_unit  [PRIDE, PRIDE:0000395, Ratio, ]
```

6.2.28 peptide-quantification_unit

Description:	Defines what type of units is reported in the peptide quantification fields.
Type:	Parameter

<table>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if peptide section is present</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD peptide-quantification_unit [PRIDE, PRIDE:0000395, Ratio,]</code></td></tr>
</table>

6.2.29 small_molecule-quantification_unit

<table>
<tr><td><b>Description:</b></td><td>Defines what type of units is reported in the small molecule quantification fields.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if small molecule section is present</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD small_molecule-quantification_unit [PRIDE, PRIDE:0000395, Ratio,]</code></td></tr>
</table>

6.2.30 ms_run[1-n]-format

<table>
<tr><td><b>Description:</b></td><td>A parameter specifying the data format of the external MS data file.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD ms_run[1]-format [MS, MS:1000584, mzML file,]</code>
…
<code>MTD ms_run[2]-format [MS, MS:1001062, Mascot MGF file,]</code></td></tr>
</table>

6.2.31 ms_run[1-n]-location

<table>
<tr><td><b>Description:</b></td><td>Location of the external data file. If the actual location of the MS run is unknown, a “null” MUST be used as a place holder value.</td></tr>
<tr><td><b>Type:</b></td><td>URL</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td></td><td></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD ms_run_location[1] file://C:\path\to\my\file</code>
…
<code>MTD ms_run_location[2] ftp://ftp.ebi.ac.uk/path/to/file</code></td></tr>
</table>

6.2.32 ms_run[1-n]-id_format

<table>
<tr><td><b>Description:</b></td><td>Parameter specifying the id format used in the external data file.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD ms_run[1]-id_format [MS, MS:1000530, mzML unique identifier,]</code>
…
<code>MTD ms_run[2]-id_format [MS, MS:1000774, multiple peak list …</code>
<code>nativeID format,]</code></td></tr>
</table>

6.2.33 ms_run[1-n]-fragmentation_method

<table>
<tr><td><b>Description:</b></td><td>A list of “|” separated parameters describing all the types of fragmentation used in a given ms run.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD ms_run[1]-fragmentation_method [MS, MS:1000133, CID,]</code>
…
<code>MTD ms_run[2]-fragmentation_method [MS, MS:1000422, HCD …,]</code></td></tr>
</table>

6.2.34 custom[1-n]

<table>
<tr><td><b>Description:</b></td><td>Any additional parameters describing the analysis reported.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD custom[1] [,,MS operator, Florian]</code></td></tr>
</table>

6.2.35 sample[1-n]-species[1-n]

<table>
<tr><td><b>Description:</b></td><td>The respective species of the samples analysed.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>COM Experiment where all samples consisted of the same two species</code>
<code>MTD sample[1]-species[1] [NEWT, 9606, Homo sapiens (Human),]</code>
<code>MTD sample[2]-species[1] [NEWT, 12059, Rhinovirus,]</code>

<code>COM Experiment where different two samples from different species (combinations)</code>
<code>COM were analysed as biological replicates.</code>

<code>MTD sample[1]-species[1] [NEWT, 9606, Homo sapiens (Human),]</code>
<code>MTD sample[1]-species[2] [NEWT, 573824, Human rhinovirus 1,]</code>
<code>MTD sample[2]-species[1] [NEWT, 9606, Homo sapiens (Human),]</code>
<code>MTD sample[2]-species[2] [NEWT, 12130, Human rhinovirus 2,]</code></td></tr>
</table>

6.2.36 sample[1-n]-tissue[1-n]

<table>
<tr><td><b>Description:</b></td><td>The respective tissue(s) of the sample.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD sample[1]-tissue[1] [BTO, BTO:0000759, liver,]</code></td></tr>
</table>

6.2.37 sample[1-n]-cell_type[1-n]

<table>
<tr><td><b>Description:</b></td><td>The respective cell type(s) of the sample.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD sample[1]-cell_type[1] [CL, CL:0000182, hepatocyte,]</code></td></tr>
</table>

6.2.38 sample[1-n]-disease[1-n]

<table>
<tr><td><b>Description:</b></td><td>The respective disease(s) of the sample.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD sample[1]-disease[1] [DOID, DOID:684, hepatocellular carcinoma,]</code>
<code>MTD sample[1]-disease[2] [DOID, DOID:9451, alcoholic fatty liver,]</code></td></tr>
</table>

6.2.39 sample[1-n]-description

<table>
<tr><td><b>Description:</b></td><td>A human readable description of the sample.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD sample[1]-description Hepatocellular carcinoma samples.</code></td></tr>
</table>

<table>
<tr><td></td><td><code>MTD sample[2]-description Healthy control samples.</code></td></tr>
</table>

6.2.40 sample[1-n]-custom[1-n]

<table>
<tr><td><b>Description:</b></td><td>Parameters describing the sample’s additional properties.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td>
<table>
<tr><th></th><th>Summary</th><th>Complete</th></tr>
<tr><td>Quantification</td><td></td><td></td></tr>
<tr><td>Identification</td><td></td><td></td></tr>
</table>
</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD sample[1]-custom[1]   [,,Extraction date, 2011-12-21]</code>
<code>MTD sample[1]-custom[2]   [,,Extraction reason, liver biopsy]</code></td></tr>
</table>

6.2.41 assay[1-n]-quantification_reagent

<table>
<tr><td><b>Description:</b></td><td>The reagent used to label the sample in the assay. For label-free analyses the “unlabeled sample” CV term SHOULD be used. For the “light” channel in label-based experiments the appropriate CV term specifying the labelling channel should be used.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td>
<table>
<tr><th></th><th>Summary</th><th>Complete</th></tr>
<tr><td>Quantification</td><td>(✓)[1]</td><td>✓</td></tr>
<tr><td>Identification</td><td>[2]</td><td>[2]</td></tr>
</table>
[1] mandatory if quantification is reported on assays
[2] not recommended for identification only files
</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD assay[1]-quantification_reagent [PRIDE,PRIDE:0000114,iTRAQ reagent,114]</code>
<code>MTD assay[2]-quantification_reagent [PRIDE,PRIDE:0000115,iTRAQ reagent,115]</code>

<code>OR</code>

<code>MTD   assay[1]-quantification_reagent   [MS,MS:1002038,unlabeled sample,]</code>

<code>OR</code>

<code>MTD   assay[1]-quantification_reagent   [PRIDE, PRIDE:0000326, SILAC light]</code>
<code>MTD   assay[2]-quantification_reagent   [PRIDE, PRIDE:0000325, SILAC heavy]</code></td></tr>
</table>

6.2.42 assay[1-n]-quantification_mod[1-n]

<table>
<tr><td><b>Description:</b></td><td>A parameter describing a modification associated with a quantification_reagent. Multiple modifications are numbered 1..n.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter</td></tr>
<tr><td><b>Mandatory</b></td><td>
<table>
<tr><th></th><th>Summary</th><th>Complete</th></tr>
<tr><td>Quantification</td><td></td><td></td></tr>
<tr><td>Identification</td><td>[1]</td><td>[1]</td></tr>
</table>
[1] not recommended for identification only files
</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD assay[2]-quantification_mod[1] [UNIMOD, UNIMOD:188, Label:13C(6),]</code></td></tr>
</table>

6.2.43 assay[1-n]-quantification_mod[1-n]-site

<table>
<tr><td><b>Description:</b></td><td>A string describing the modifications site. Following the unimod convention, modification site is a residue (e.g. “M”), terminus (“N-term” or “C-term”) or both (e.g. “N-term Q” or “C-term K”).</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td>
<table>
<tr><th></th><th>Summary</th><th>Complete</th></tr>
<tr><td>Quantification</td><td></td><td></td></tr>
<tr><td>Identification</td><td>[1]</td><td>[1]</td></tr>
</table>
[1] not recommended for identification only files
</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD assay[2]-quantification_mod[1] [UNIMOD, UNIMOD:188, Label:13C(6),]</code>
<code>MTD assay[2]-quantification_mod[2] [UNIMOD, UNIMOD:188, Label:13C(6),]</code>
<code>MTD assay[2]-quantification_mod[1]-site   R</code>
<code>MTD assay[2]-quantification_mod[2]-site   K</code></td></tr>
</table>

6.2.44 assay[1-n]-quantification_mod[1-n]-position

<table>
<tr><td><b>Description:</b></td><td>A string describing the term specifity of the modification. Following the unimod convention, term specifity is denoted by the strings “Anywhere”, “Any N-term”, “Any C-term”, “Protein N-term”, “Protein C-term”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] not recommended for identification only files</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD assay[2]-quantification_mod[1] [UNIMOD, UNIMOD:188, Label:13C(6),]</code>
<code>MTD assay[2]-quantification_mod[2] [UNIMOD, UNIMOD:188, Label:13C(6),]</code>
<code>MTD assay[2]-quantification_mod[1]-site   R</code>
<code>MTD assay[2]-quantification_mod[2]-site   K</code>
<code>MTD assay[2]-quantification_mod[1]-position   Anywhere</code>
<code>MTD assay[2]-quantification_mod[2]-position   Anywhere</code></td></tr>
</table>

6.2.45 assay[1-n]-sample_ref

<table>
<tr><td><b>Description:</b></td><td>An association from a given assay to the sample analysed.</td></tr>
<tr><td><b>Type:</b></td><td>{SAMPLE_ID}</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD assay[1]-sample_ref sample[1]</code>
<code>MTD assay[2]-sample_ref sample[2]</code></td></tr>
</table>

6.2.46 assay[1-n]-ms_run_ref

<table>
<tr><td><b>Description:</b></td><td>An association from a given assay to the source MS run.</td></tr>
<tr><td><b>Type:</b></td><td>{MS_RUN_ID}</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>✓</td></tr><tr><td>Identification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr></table>[1] mandatory if assays are reported</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   assay[1]-ms_run_ref   ms_run[1]</code></td></tr>
</table>

6.2.47 study_variable[1-n]-assay_refs

<table>
<tr><td><b>Description:</b></td><td>Comma-separated references to the IDs of assays grouped in the study variable.</td></tr>
<tr><td><b>Type:</b></td><td>{ASSAY_ID}, ...</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>✓</td></tr><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if both assays and study variables are reported</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   study_variable[1]-assay_refs assay[1], assay[2], assay[3]</code></td></tr>
</table>

6.2.48 study_variable[1-n]-sample_refs

<table>
<tr><td><b>Description:</b></td><td>Comma-separated references to the samples that were analysed in the study variable.</td></tr>
<tr><td><b>Type:</b></td><td>{SAMPLE_ID}, ... {SAMPLE_ID}</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   study_variable[1]-sample_refs sample[1]</code></td></tr>
</table>

6.2.49 study_variable[1-n]-description

Description:	A textual description of the study variable.
Type:	String

<table>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>✓</td></tr><tr><td>Identification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr></table>[1] mandatory of study variables reported</td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   study_variable[1]-description Group B (spike-in 0.74 fmol/uL)</code></td></tr>
</table>

6.2.50 cv[1-n]-label

<table>
<tr><td><b>Description:</b></td><td>A string describing the labels of the controlled vocabularies/ontologies used in the mzTab file</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD cv[1]-label MS</code>
…</td></tr>
</table>

6.2.51 cv[1-n]-full_name

<table>
<tr><td><b>Description:</b></td><td>A string describing the full names of the controlled vocabularies/ontologies used in the mzTab file</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD cv[1]-full_name MS</code>
…</td></tr>
</table>

6.2.52 cv[1-n]-version

<table>
<tr><td><b>Description:</b></td><td>A string describing the version of the controlled vocabularies/ontologies used in the mzTab file</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD cv[1]-version 3.54.0</code>
…</td></tr>
</table>

6.2.53 cv[1-n]-url

<table>
<tr><td><b>Description:</b></td><td>A string containing the URLs of the controlled vocabularies/ontologies used in the mzTab file</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD cv[1]-url http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo</code>
…</td></tr>
</table>

6.2.54 colunit-protein

<table>
<tr><td><b>Description:</b></td><td>Defines the unit for the data reported in a column of the protein section. The format of the value has to be {column name}={Parameter defining the unit}</td></tr>
</table>

<table>
<tr><td></td><td>This field MUST NOT be used to define a unit for quantification columns. The unit used for protein quantification values MUST be set in <em>protein-quantification_unit</em>.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD</code></td></tr>
</table>

6.2.55 colunit-peptide

<table>
<tr><td><b>Description:</b></td><td>Defines the used unit for a column in the peptide section. The format of the value has to be {column name}={Parameter defining the unit}
This field MUST NOT be used to define a unit for quantification columns. The unit used for peptide quantification values MUST be set in peptide-quantification_unit.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   colunit-peptide retention_time=[UO,UO:0000031, minute,]</code></td></tr>
</table>

6.2.56 colunit-psm

<table>
<tr><td><b>Description:</b></td><td>Defines the used unit for a column in the PSM section. The format of the value has to be {column name}={Parameter defining the unit}
This field MUST NOT be used to define a unit for quantification columns. The unit used for peptide quantification values MUST be set in peptide-quantification_unit.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   colunit-psm retention_time=[UO,UO:0000031, minute,]</code></td></tr>
</table>

6.2.57 colunit-small_molecule

<table>
<tr><td><b>Description:</b></td><td>Defines the used unit for a column in the small molecule section. The format of the value has to be {column name}={Parameter defining the unit}
This field MUST NOT be used to define a unit for quantification columns. The unit used for small molecule quantification values MUST be set in small_molecule-quantification_unit.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>MTD   colunit-small_molecule retention_time=[UO,UO:0000031, minute,]</code></td></tr>
</table>

6.3 Protein Section

The protein section is table-based. The protein section MUST always come after the metadata section. All table columns MUST be tab-separated. There MUST NOT be any empty cells. Missing values MUST be reported using “null”. Most columns are mandatory. The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

6.3.1 accession

<table>
<tr><td><b>Description:</b></td><td>The accession of the protein in the source database. A protein accession MUST be unique within one mzTab file. If different quantification values are required for the same underlying accession, for example if differentially modified forms of a protein have been quantified, a the suffix [1-n] SHOULD be appended to the accession e.g. P12345[1], P12345[2].</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession …
PRT P12345 …
PRT P12346 …</td></tr>
</table>

6.3.2 description

<table>
<tr><td><b>Description:</b></td><td>The protein’s name and or description line.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession description …
PRT P12345 Aspartate aminotransferase, mitochondrial …
PRT P12346 Serotransferrin …</td></tr>
</table>

6.3.3 taxid

<table>
<tr><td><b>Description:</b></td><td>The NCBI/NEWT taxonomy id for the species the protein was identified in.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession … taxid …
PRT P12345 … 10116 …
PRT P12346 … 10116 …</td></tr>
</table>

6.3.4 species

<table>
<tr><td><b>Description:</b></td><td>The human readable species the protein was identified in - this SHOULD be the NCBI entry’s name.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession … taxid species …
PRT P12345 … 10116 Rattus norvegicus (Rat) …
PRT P12346 … 10116 Rattus norvegicus (Rat) …</td></tr>
</table>

6.3.5 database

<table>
<tr><td><b>Description:</b></td><td>The protein database used for the search (could theoretically come from a different species). Wherever possible the Miriam (http://www.ebi.ac.uk/miriam) assigned name SHOULD be used.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession … taxid species database …
PRT P12345 … 10116 Rattus norvegicus (Rat) UniProtKB …
PRT P12346 … 10116 Rattus norvegicus (Rat) UniProtKB …</td></tr>
</table>

6.3.6 database_version

<table>
<tr><td><b>Description:</b></td><td>The protein database's version – in case there is no version available (custom build) the creation / download (e.g., for NCBI nr) date SHOULD be given. Additionally, the number of entries in the database MAY be reported in round brackets after the version in the format: {version} ({#entries} entries), for example “2011-11 (1234 entries)”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PRH accession … taxid species   database   database_version …</code>
<code>PRT P12345   … 10116 Rattus norvegicus (Rat)   UniProtKB 2011_11   …</code>
<code>PRT P12346   … 10116 Rattus norvegicus (Rat)   UniProtKB 2011_11   …</code></td></tr>
</table>

6.3.7 search_engine

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine(s) used to identify this protein. Search engines MUST be supplied as parameters.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>COM In this example the first protein was identified by Mascot and Sequest while</code>
<code>COM the second protein was only identified by Mascot.</code>
<code>PRH accession … search_engine   …</code>
<code>PRT P12345   … [MS,MS:1001207,Mascot,]|[MS,MS:1001208,Sequest,]   …</code>
<code>PRT P12346   … [MS,MS:1001207,Mascot,]   …</code></td></tr>
</table>

6.3.8 best_search_engine_score

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of the best search engine score(s) for the given protein across all replicates reported. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PRH accession … best_search_engine_score_ms_run[1]</code>
<code>…</code>
<code>PRT P12345   … [MS,MS:1001171,Mascot score,50]|[MS,MS:1001155,Sequest:xcorr,2] …</code>
<code>PRT P12346   … [MS,MS:1001171,Mascot score,47.2]   …</code></td></tr>
</table>

6.3.9 search_engine_score_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine score(s) for the given protein. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PRH accession … search_engine_score_ms_run[1]</code>
<code>…</code>
<code>PRT P12345   … [MS,MS:1001171,Mascot score,50]|[MS,MS:1001155,Sequest:xcorr,2] …</code>
<code>PRT P12346   … [MS,MS:1001171,Mascot score,47.2]   …</code></td></tr>
</table>

6.3.10 reliability

<table>
<tr><td><b>Description:</b></td><td>The reliability of the given protein identification. This must be supplied by the resource and has to be one of the following values:
1: high reliability
2: medium reliability
3: poor reliability</td></tr>
</table>

<table>
<tr><td></td><td>Important: An identification's reliability is resource-dependent.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRH accession … reliability …
PRT P12345 … 3 …
PRT P12346 … 1 …</td></tr>
</table>

6.3.11 num_psms_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>The count of the total significant PSMs that can be mapped to the reported protein.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>COM P12345 is identified through ABCM, ABCM+Oxidation, CDE, CDE
…
PRH accession … num_psms_ms_run[1] …
PRT P12345 … 4 …</td></tr>
</table>

6.3.12 num_peptides_distinct_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>The count of the number of different peptide sequences that have been identified above the significance threshold. Different modifications or charge states of the same peptide are not counted.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>COM P12345 is identified through ABCM, ABCM+Oxidation, CDE, CDE
…
PRH accession … num_peptides_distinct_ms_run[1] …
PRT P12345 … 3 …</td></tr>
</table>

6.3.13 num_peptides_unique_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>The number of peptides that can be mapped uniquely to the protein reported. If ambiguity members have been reported, the count MUST be derived from the number of peptides that can be uniquely mapped to the group of accessions, since the assumption is that these accessions are supported by the same evidence.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>COM P12345 is identified through ABCM, ABCM+Oxidation, CDE, CDE
COM ABCM is only from P12345, CDE from P12345 and P12346
…
PRH accession … num_peptides_unique_ms_run[1] …
PRT P12345 … 2 …</td></tr>
</table>

6.3.14 ambiguity_members

<table>
<tr><td><b>Description:</b></td><td>A comma-delimited list of protein accessions. This field should be set in the representative protein of the ambiguity group (the protein identified through the accession in the first column). The accessions listed in this field should identify proteins that could also be identified through these peptides (e.g. “same-set proteins”) but were not chosen by the researcher or resource, often for arbitrary reasons. It is NOT RECOMMENDED to report subset proteins as ambiguity_members, since the proteins reported here, together with the</td></tr>
</table>

<table>
<tr><td></td><td>representative protein are taken to be a group that cannot be separated based on the peptide evidence.</td></tr>
<tr><td><b>Type:</b></td><td>String List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>COM P12345, P12347, and P12348 can all be identified through the same peptides</code>
<code>…</code>
<code>PRH accession …   ambiguity_members …</code>
<code>PRT P12345   …   P12347,P12348   …</code></td></tr>
</table>

6.3.15 modifications

<table>
<tr><td><b>Description:</b></td><td>In contrast to the PSM section, fixed modifications or modifications caused by the quantification reagent (i.e. the SILAC/iTRAQ label) SHOULD NOT be reported in this column.
Column entries are a comma delimited list of modifications found in the given protein. Modifications have to be reported in the following format:
{position in protein}{Parameter}-{Modification or Substitution identifier}|{neutral loss}
Modification location scores cannot be supplied at the Protein level. Furthermore, in case a position is unknown no position information MAY be supplied.
Terminal modifications MUST be reported at position 0 or protein size + 1 respectively.
Valid modification identifiers are either PSI-MOD or UNIMOD accession (including the “MOD:” / “UNIMOD:” prefix) or CHEMMODS. CHEMMODS have the format CHEMMOD:+/-{chemical formula or <i>m/z</i> delta}. Valid CHEMMODS are for example “CHEMMOD:+NH4” or “CHEMMOD:-10.1098”. CHEMMODs MUST NOT be used if the modification can be reported using a PSI-MOD or UNIMOD accession. Mass deltas MUST NOT be used for CHEMMODs if the delta can be expressed through a known chemical formula.
Neutral losses MAY be reported as cvParams. If a neutral loss is not associated with an existing modification it is reported as separated comma-separated entry. Otherwise, the neutral loss MUST be reported after the modification it is associated with and separated by a ‘|’ from the modification.
Additionally, it is possible to report substitutions of amino acids using SUBST:{amino acid}.

If different modifications are identified from different ms_runs, a superset of the identified modifications SHOULD be reported here. Detailed modification mapping to individual ms_runs is provided through the PSM table.

If protein level modifications are not reported, a “null” MUST be used. If protein level modifications are reported but not present on a given protein, a “0” MUST be reported.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>COM Protein P12345 TESTPEPTIDES with 2 phosphorylation sites: TEpSTPEPpTIDES</code>

<code>COM Common use cases without score:</code></td></tr>
</table>

```
COM  Example 1: Both locations have been determined
PRH  accession   …  modifications                               …
PRT  P12345      …  3-MOD:00412,8-MOD:00412                     …

COM  Example 2: Like Ex. 1, but first site localization is ambiguous (S or T)
PRH  accession   …  modifications                               …
PRT  P12345      …  3|4-MOD:00412,8-MOD:00412                   …

COM  Example 3: Protein only known to contain two phosphor sites in the range 3 to 8
PRH  accession   …  modifications                               …
PRT  P12345      …  3|4|8-MOD:00412, 3|4|8-MOD:00412            …

COM  Example 4: No position information or only accurate mass available
PRH  accession   …  modifications                               …
PRT  P12345      …  CHEMMOD:+159.93                   …

COM  Common use cases with probability scores:
COM  Example 5: MOD:00412 with associated probabilities at position 3 and 4
COM             and a probability of 0.3 at position 8
PRH  accession   …  modifications                               …
PRT  P12345      …  3[MS,MS:1001876, modification probability, 0.8]|4[MS,MS:1001876,
modification probability, 0.2]-MOD:00412,8[MS,MS:1001876, modification probability,
0.3]-MOD:00412        …

COM Reporting substitutions
COM Example 6: Substitution of amino acid at position 3 with R (Original sequence is
reported in sequence column)
PRH  accession   …  modifications
PRT  P12345      …  3-SUBST:R
```

6.3.16 uri

<table>
<tr><td><b>Description:</b></td><td>A URI pointing to the protein's source entry in the unit it was identified in (e.g., the PRIDE database or a local database / file identifier).</td></tr>
<tr><td><b>Type:</b></td><td>URI</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … uri …
PRH P12345 … http://www.ebi.ac.uk/pride/url/to/P12345 …</td></tr>
</table>

6.3.17 go_terms

<table>
<tr><td><b>Description:</b></td><td>A ’|’-delimited list of GO accessions for this protein.</td></tr>
<tr><td><b>Type:</b></td><td>String List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … go_terms …
PRH P12345 … GO:0006457|GO:0005759|GO:0005886|GO:0004069 …</td></tr>
</table>

6.3.18 protein_coverage

<table>
<tr><td><b>Description:</b></td><td>A value between 0 and 1 defining the protein coverage.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td></td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … protein_coverage …
PRH P12345 … 0.4 …</td></tr>
</table>

6.3.19 protein_abundance_assay[1-n]

<table>
<tr><td><b>Description:</b></td><td>The protein's abundance as measured in the given assay through whatever technique was employed.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>✓</td></tr></table></td></tr>
</table>

<table>
<tr><td></td><td><table><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if quantification data is provided for assays</td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … protein_abundance_assay[1] … protein_abundance_assay[2] …
PRH P12345 … 0.4 … 0.2 …</td></tr>
</table>

6.3.20 protein_abundance_study_variable[1-n]

<table>
<tr><td><b>Description:</b></td><td>The protein's abundance as measured in the given Study Variable, for example mean or median of quantitative values reported in Assays.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … protein_abundance_study_variable[1] … protein_abundance_study_variable[2] …
PRH P12345 … 0.4 … 0.2 …</td></tr>
</table>

6.3.21 protein_abundance_stdev_study_variable[1-n]

<table>
<tr><td><b>Description:</b></td><td>The standard deviation of the protein's abundance. If a protein's abundance is given for a certain study variable, the corresponding standard deviation column MUST also be present (in case the value is not available "null" should be used).</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if protein_abundance_study_variable reported</td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … protein_abundance_stdev_study_variable[1] …
PRH P12345 … 0.4 …</td></tr>
</table>

6.3.22 protein_abundance_std_error_study_variable [1-n]

<table>
<tr><td><b>Description:</b></td><td>The standard error of the protein's abundance. If a protein's abundance is given for a certain study variable, the corresponding standard error column MUST also be present (in case the value is not available "null" should be used).</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>(✓)[1]</td><td>(✓)[1]</td></tr><tr><td>Identification</td><td></td><td></td></tr></table>[1] mandatory if protein_abundance_study_variable reported</td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … protein_abundance_study_variable[1] … protein_abundance_std_error_study_variable[1] …
PRH P12345 … 0.4 … 0.03 …</td></tr>
</table>

6.3.23 opt_global_*

<table>
<tr><td><b>Description:</b></td><td>Additional columns can be added to the end of the protein table. These column headers MUST start with the prefix "opt_" followed by the identifier of the object they reference: assay, study variable, MS run or "global" (if the value relates to all replicates). Column names MUST only contain the following characters: 'A'-'Z', 'a'-'z', '0'-'9', '_', '-', '[', ']', and ':'. CV parameter accessions MAY be used for optional columns following the format: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter's name MUST be replaced by '_'.</td></tr>
<tr><td><b>Type:</b></td><td>Column</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PRT accession … opt_assay[1]_my_value opt_global_another_value</td></tr>
</table>

	PRH P12345 … My value about assay[1] some other value that is across reps

6.4 Peptide Section

The peptide section is table based. The peptide section must always come after the metadata section and or protein section if these are present in the file. All table columns MUST be tab separated. There MUST NOT be any empty cells. Missing values MUST be reported using “null”. Most columns are mandatory. The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

6.4.1 sequence

<table>
<tr><td><b>Description:</b></td><td>The peptide's sequence</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence …
PEP KVPQVSTPTLVEVSR …
PEP EIEILACEIR …</td></tr>
</table>

6.4.2 accession

<table>
<tr><td><b>Description:</b></td><td>The protein's accession the peptide is associated with. In case no protein section is present in the file or the peptide was not assigned to a protein the field should be filled with “null”. If the peptide can be assigned to more than one protein, multiple rows SHOULD be provided for each peptide to protein mapping.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence accession …
PEP KVPQVSTPTLVEVSR P02768 …</td></tr>
</table>

6.4.3 unique

<table>
<tr><td><b>Description:</b></td><td>Indicates whether the peptide is unique for this protein in respect to the searched database.</td></tr>
<tr><td><b>Type:</b></td><td>Boolean (0/1)</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence accession unique …
PEP KVPQVSTPTLVEVSR P02768 0 …
PEP VFDEFKPLVEEPQNLIK P02768 1 …</td></tr>
</table>

6.4.4 database

<table>
<tr><td><b>Description:</b></td><td>The protein database used for the search (could theoretically come from a different species) and the peptide sequence comes from.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence accession unique database …
PEP KVPQVSTPTLVEVSR P02768 0 UniProtKB …</td></tr>
</table>

<table>
<tr><td></td><td><table><tr><td>PEP</td><td>VFDEFKPLVEEPQNLIK</td><td>P02768</td><td>1</td><td>UniProtKB</td><td>…</td></tr></table></td></tr>
</table>

6.4.5 database_version

<table>
<tr><td><b>Description:</b></td><td>The protein database's version – in case there is no version available (custom build) the creation / download (e.g., for NCBI nr) date should be given. Additionally, the number of entries in the database MAY be reported in round brackets after the version in the format: {version} ({#entries} entries), for example “2011-11 (1234 entries)”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td><table><tr><td>PEH</td><td>sequence</td><td>accession</td><td>unique</td><td>database</td><td>database_version</td><td>…</td></tr><tr><td>PEP</td><td>KVPQVSTPTLVEVSR</td><td>P02768</td><td>0</td><td>UniProtKB</td><td>2011_11</td><td>…</td></tr><tr><td>PEP</td><td>VFDEFKPLVEEPQNLIK</td><td>P02768</td><td>1</td><td>UniProtKB</td><td>2011_11</td><td>…</td></tr></table></td></tr>
</table>

6.4.6 search_engine

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine(s) used to identify this peptide. Search engines must be supplied as parameters.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td><table><tr><td>PEH</td><td>sequence</td><td>…</td><td>search_engine</td><td>…</td></tr><tr><td>PEP</td><td>KVPQVSTPTLVEVSR</td><td>…</td><td>[MS,MS:1001207,Mascot,]|[MS,MS:1001208,Sequest,]</td><td>…</td></tr><tr><td>PEP</td><td>VFDEFKPLVEEPQNLIK</td><td>…</td><td>[MS,MS:1001207,Mascot,]</td><td>…</td></tr></table></td></tr>
</table>

6.4.7 best_search_engine_score

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of best search engine score(s) for the given peptide across all replicates. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td><table><tr><td>PEH</td><td>sequence</td><td>…</td><td>best_search_engine_score</td><td>…</td></tr><tr><td>PEP</td><td>KVPQVSTPTLVEVSR</td><td>…</td><td>[MS,MS:1001155,Sequest:xcorr,2]</td><td>…</td></tr><tr><td>PEP</td><td>VFDEFKPLVEEPQNLIK</td><td>…</td><td>[MS,MS:1001171,Mascot score,47.2]</td><td>…</td></tr></table></td></tr>
</table>

6.4.8 search_engine_score_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine score(s) for the given peptide from a given MS run. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td><table><tr><td>PEH</td><td>sequence</td><td>…</td><td>search_engine_score_ms_run[1]</td><td>…</td></tr><tr><td>PEP</td><td>KVPQVSTPTLVEVSR</td><td>…</td><td>[MS,MS:1001155,Sequest:xcorr,2]</td><td>…</td></tr><tr><td>PEP</td><td>VFDEFKPLVEEPQNLIK</td><td>…</td><td>[MS,MS:1001171,Mascot score,47.2]</td><td>…</td></tr></table></td></tr>
</table>

6.4.9 reliability

<table>
<tr><td><b>Description:</b></td><td>The reliability of the given peptide identification. This must be supplied by the</td></tr>
</table>

<table>
<tr><td></td><td>resource and has to be one of the following values:
1: high reliability
2: medium reliability
3: poor reliability

Important: An identification's reliability is resource dependent.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1]Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … reliability …
PEP KVPQVSTPTLVEVSR … 3 …
PEP VFDEFKPLVEEPQNLIK … 1 …</td></tr>
</table>

6.4.10 modifications

<table>
<tr><td><b>Description:</b></td><td>The peptide's modifications or substitutions. To further distinguish peptide terminal modifications, these SHOULD be reported at position 0 or <i>peptide size</i> + 1 respectively. For detailed information see the modifications section in the protein table. If substitutions are reported, the “sequence” column MUST contain the original, unaltered sequence. Note that in contrast to the PSM section, fixed modifications or modifications caused by the quantification reagent i.e. the SILAC labels/tags SHOULD NOT be reported. It is thus also expected that modification reliability scores will typically be reported at the PSM-level only.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1]Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … modifications …
PEP KVPQVSTPTLVEVSR … 10-MOD:00412 …
PEP VFDEFKPLVEEPQNLIK … NULL …</td></tr>
</table>

6.4.11 retention_time

<table>
<tr><td><b>Description:</b></td><td>A ‘|’-separated list of time points. Semantics may vary on how retention times are reported. For quantification approaches, different exporters MAY wish to export the retention times of all spectra used for quantification (e.g. in MS^2 approaches) or the centre point of the feature quantified for MS^1 approaches. It is assumed that the reported value(s) are for a given “master” peptide from one assay only (and the unlabeled peptide in label-based approaches). If the exporter wishes to export values for all assays, this can be done using optional columns. Retention time MUST be reported in seconds. Otherwise, units MUST be reported in the Metadata Section (“colunit-peptide”).</td></tr>
<tr><td><b>Type:</b></td><td>Double List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1]Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … retention_time …
PEP KVPQVSTPTLVEVSR … 10.2 …
PEP VFDEFKPLVEEPQNLIK … 15.8 …</td></tr>
</table>

6.4.12 retention_time_window

Description:	Start and end of the retention time window separated by a single ‘|’. Semantics

<table>
<tr><td></td><td>may vary but its primary intention is to report feature boundaries of eluting peptides (along with feature centroids in the retention_time column). It is assumed that the reported interval is for a given “master” peptide from one assay only (and the unlabeled peptide in label-based approaches). If the exporter wishes to export values for all assays, this can be done using optional columns. Retention time windows MUST be reported in seconds. Otherwise, units MUST be reported in the Metadata Section (“colunit-peptide”).</td></tr>
<tr><td><b>Type:</b></td><td>Double List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … retention_time_window …
PEP KVPQVSTPTLVEVSR … 1123.2|1145.3 …</td></tr>
</table>

6.4.13

6.4.14 charge

<table>
<tr><td><b>Description:</b></td><td>The charge assigned by the search engine/software. In case multiple charge states for the same peptide are observed these should be reported as distinct entries in the peptide table. In case the charge is unknown “null” MUST be used.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … charge …
PEP KVPQVSTPTLVEVSR … 2 …
PEP VFDEFKPLVEEPQNLIK … 3 …</td></tr>
</table>

6.4.15 mass_to_charge

<table>
<tr><td><b>Description:</b></td><td>The precursor’s experimental mass to charge (<i>m/z</i>). It is assumed that the reported value is for a given “master” peptide from one assay only (and the unlabeled peptide in label-based approaches). If the exporter wishes to export values for all assays, this can be done using optional columns.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … mass_to_charge …
PEP KVPQVSTPTLVEVSR … 1234.4 …
PEP VFDEFKPLVEEPQNLIK … 123.4 …</td></tr>
</table>

6.4.16 uri

<table>
<tr><td><b>Description:</b></td><td>A URI pointing to the peptide's entry in the experiment it was identified in (e.g., the peptide’s PRIDE entry).</td></tr>
<tr><td><b>Type:</b></td><td>URI</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td>PEH sequence … uri …
PEP KVPQVSTPTLVEVSR … http://www.ebi.ac.uk/pride/link/to/peptide …
PEP VFDEFKPLVEEPQNLIK … http://www.ebi.ac.uk/pride/link/to/peptide …</td></tr>
</table>

6.4.17 spectra_ref

<table>
<tr><td><b>Description:</b></td><td>Reference to spectra in a spectrum file. It is expected that spectra_ref SHOULD only be used for MS^2-based quantification approaches, in which retention time values cannot identify the spectra used for quantitation. The reference must be in the format <code>ms_run[1-n]:{SPECTRA_REF}</code> where SPECTRA_REF MUST follow the format defined in 5.2. Multiple spectra MUST be referenced using a “|” delimited list.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td>(✓)[2]</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files
[2] Mandatory only if MS2 based quantification is used</td></tr>
<tr><td><b>Example:</b></td><td><code>PEH sequence … spectra_ref …</code>
<code>PEP KVPQVSTPTLVEVSR … ms_run[1]:index=5 …</code>
<code>PEP VFDEFKPLVEEPQNLIK … ms_run[2]:index=7|ms_run[2]:index=9 …</code></td></tr>
</table>

6.4.18 peptide_abundance_assay[1-n]

<table>
<tr><td><b>Description:</b></td><td>The peptide’s abundance in the given assay.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files
[2] If quantification data is reported on assays level</td></tr>
<tr><td><b>Example:</b></td><td><code>PEH sequence … peptide_abundance_assay[1] peptide_abundance_assay[2]…</code>
<code>PEP KVPQVSTPTLVEVSR … 0.4 0.5</code></td></tr>
</table>

6.4.19 peptide_abundance_study_variable[1-n]

<table>
<tr><td><b>Description:</b></td><td>The peptide’s abundance in the given study variable, for example calculated as an average of assay values.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files
[2] mandatory if study variables are reported</td></tr>
<tr><td><b>Example:</b></td><td><code>PEH sequence … peptide_abundance_study_variable[1] …</code>
<code>PEP KVPQVSTPTLVEVSR … 0.4 …</code></td></tr>
</table>

6.4.20 peptide_abundance_stdev_study_variable[1-n]

<table>
<tr><td><b>Description:</b></td><td>The standard deviation of the peptide’s abundance for a given study variable.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>(✓)[2]</td><td>(✓)[2]</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1] Not recommended in identification only files
[2] mandatory if peptide_abundance_study_variable reported</td></tr>
<tr><td><b>Example:</b></td><td><code>PEH sequence … peptide_abundance_study_variable [1]</code>
<code>peptide_abundance_stdev_study_variable[1] …</code>
<code>PEP KVPQVSTPTLVEVSR … 0.4 0.2 …</code></td></tr>
</table>

6.4.21 peptide_abundance_std_error_study_variable[1-n]

<table>
<tr><td><b>Description:</b></td><td>The standard error of the peptide’s abundance for a given study variable.</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>(✓)[2]</td><td>(✓)[2]</td></tr></table></td></tr>
</table>

Identification	[1]	[1]

[1] Not recommended in identification only files
[2] mandatory if peptide_abundance_study_variable reported

Example:

```
PEH  sequence        …  peptide_abundance_study_variable[1] …
peptide_abundance_std_error_study_variable[1] …
PEP  KVPQVSTPTLVEVSR …  0.4                          … 0.2                                 …
```

6.4.22 opt_global_*

Description: Additional columns can be added to the end of the peptide table. These column headers MUST start with the prefix "opt_" followed by the identifier of the object they reference: assay, study variable, MS run or "global" (if the value relates to all replicates). Column names MUST only contain the following characters: 'A'-'Z', 'a'-'z', '0'-'9', '_', '-', '[', ']', and ':'. CV parameter accessions MAY be used for optional columns following the format: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter's name MUST be replaced by '_'.

Type: Column

Mandatory

	Summary	Complete
Quantification		
Identification	[1]	[1]

[1] Not recommended in identification only files

Example:

```
PRT  accession   …  opt_assay[1]_my_value  opt_global_another_value
PRH  P12345      …  My value about assay[1]      some other value that is across reps
```

6.5 PSM Section

The PSM section is table-based. The PSM section MUST always come after the metadata section, peptide section and or protein section if they are present in the file. All table columns MUST be tab separated. Missing values MUST be reported using "null". Most columns are mandatory. The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

6.5.1 sequence

Description: The peptide's sequence corresponding to the PSM

Type: String

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification	✓	✓

Example:

```
PSH  sequence                          …
PSM  KVPQVSTPTLVEVSR                   …
PSM  EIEILACEIR …
```

6.5.2 PSM_ID

Description: A unique identifier for a PSM within the file. If a PSM can be matched to multiple proteins, the same PSM should be represented on multiple rows with different accessions and the same PSM_ID.

Type: Integer

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification	✓	✓

Example:

```
PSH  sequence         PSM_ID  accession…
PSM  KVPQVSTPTLVEVSR  1     P02768     …
PSM  PEPTIDR  2     P04267     …
PSM  PEPTIDR  2     P04268     …
```

6.5.3 accession

Description: The protein's accession the corresponding peptide sequence (coming from the

<table>
<tr><td></td><td>PSM) is associated with. In case no protein section is present in the file or the peptide was not assigned to a protein the field should be filled with “null”. If the PSM can be assigned to more than one protein, the same PSM should be represented on multiple rows with the same unique identifier.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence accession …
PSM KVPQVSTPTLVEVSR P02768 …</td></tr>
</table>

6.5.4 unique

<table>
<tr><td><b>Description:</b></td><td>Indicates whether the peptide sequence (coming from the PSM) is unique for this protein in respect to the searched database.</td></tr>
<tr><td><b>Type:</b></td><td>Boolean (0/1)</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence accession unique …
PSM KVPQVSTPTLVEVSR P02768 0 …
PSM VFDEFKPLVEEPQNLIK P02768 1 …</td></tr>
</table>

6.5.5 database

<table>
<tr><td><b>Description:</b></td><td>The protein database used for the search (could theoretically come from a different species) and the peptide sequence comes from.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence accession unique database …
PSM KVPQVSTPTLVEVSR P02768 0 UniProtKB …
PSM VFDEFKPLVEEPQNLIK P02768 1 UniProtKB …</td></tr>
</table>

6.5.6 database_version

<table>
<tr><td><b>Description:</b></td><td>The protein database's version – in case there is no version available (custom build) the creation / download (e.g., for NCBI nr) date should be given. Additionally, the number of entries in the database MAY be reported in round brackets after the version in the format: {version} ({#entries} entries), for example “2011-11 (1234 entries)”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence accession unique database database_version …
PSM KVPQVSTPTLVEVSR P02768 0 UniProtKB 2011_11 …
PSM VFDEFKPLVEEPQNLIK P02768 1 UniProtKB 2011_11 …</td></tr>
</table>

6.5.7 search_engine

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine(s) used to create the PSM. Search engines must be supplied as parameters.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … search_engine …
PSM KVPQVSTPTLVEVSR … [MS,MS:1001207,Mascot,]|[MS,MS:1001208,Sequest,] …
PSM VFDEFKPLVEEPQNLIK … [MS,MS:1001207,Mascot,] …</td></tr>
</table>

6.5.8 search_engine_score

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine score(s) for the given PSM.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … search_engine_score …
PSM KVPQVSTPTLVEVSR … [MS,MS:1001155,Sequest:xcorr,2] …
PSM VFDEFKPLVEEPQNLIK … [MS,MS:1001171,Mascot score,47.2] …</td></tr>
</table>

6.5.9 reliability

<table>
<tr><td><b>Description:</b></td><td>The reliability of the given PSM. This must be supplied by the resource and has to be one of the following values:
1: high reliability
2: medium reliability
3: poor reliability

Important: An identification's reliability is resource dependent.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … reliability …
PSM KVPQVSTPTLVEVSR … 3 …
PSM VFDEFKPLVEEPQNLIK … 1 …</td></tr>
</table>

6.5.10 modifications

<table>
<tr><td><b>Description:</b></td><td>The peptide's (coming from the PSM) modifications or substitutions. To further distinguish peptide terminal modifications, these SHOULD be reported at position 0 or <i>peptide size</i> + 1 respectively. For detailed information see the modifications section in the protein table. If substitutions are reported, the “sequence” column MUST contain the original, unaltered sequence.
Note that in contrast to the PRT and PEP section all modifications (variable and fixed modifications, including those induced by quantification reagents) MUST BE reported in the PSM section.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … modifications …
PSM KVPQVSTPTLVEVSR … 10[MS,MS:100xxxx,Probability Score Y,0.8]-MOD:00412 …
PSM VFDEFKPLVEEPQNLIK … NULL …</td></tr>
</table>

6.5.11 retention_time

<table>
<tr><td><b>Description:</b></td><td>The retention time of the spectrum. A ‘|’-separated list of multiple time points is allowed in case multiple spectra were combined by the search engine to make the PSM. It MUST be reported in seconds. Otherwise, the units MUST be reported in the Metadata Section (‘columnit_psm’).</td></tr>
<tr><td><b>Type:</b></td><td>Double List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … retention_time …
PSM KVPQVSTPTLVEVSR … 10.2 …
PSM VFDEFKPLVEEPQNLIK … 15.8 …</td></tr>
</table>

6.5.12 charge

<table>
<tr><td><b>Description:</b></td><td>The charge assigned by the search engine/software.</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PSH sequence … charge …</code>
<code>PSM KVPQVSTPTLVEVSR … 2 …</code>
<code>PSM VFDEFKPLVEEPQNLIK … 3 …</code></td></tr>
</table>

6.5.13 exp_mass_to_charge

<table>
<tr><td><b>Description:</b></td><td>The PSM’s experimental mass to charge (<i>m/z</i>).</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PSH sequence … mass_to_charge …</code>
<code>PSM KVPQVSTPTLVEVSR … 1234.4 …</code>
<code>PSM VFDEFKPLVEEPQNLIK … 123.4 …</code></td></tr>
</table>

6.5.14 calc_mass_to_charge

<table>
<tr><td><b>Description:</b></td><td>The PSM’s calculated (theoretical) mass to charge (<i>m/z</i>).</td></tr>
<tr><td><b>Type:</b></td><td>Double</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PSH sequence … mass_to_charge …</code>
<code>PSM KVPQVSTPTLVEVSR … 1234.4 …</code>
<code>PSM VFDEFKPLVEEPQNLIK … 123.4 …</code></td></tr>
</table>

6.5.15 uri

<table>
<tr><td><b>Description:</b></td><td>A URI pointing to the PSM's entry in the experiment it was identified in (e.g., the peptide’s PRIDE entry).</td></tr>
<tr><td><b>Type:</b></td><td>URI</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PSH sequence … uri …</code>
<code>PSM KVPQVSTPTLVEVSR … http://www.ebi.ac.uk/pride/link/to/peptide …</code>
<code>PSM VFDEFKPLVEEPQNLIK … http://www.ebi.ac.uk/pride/link/to/peptide …</code></td></tr>
</table>

6.5.16 spectra_ref

<table>
<tr><td><b>Description:</b></td><td>Reference to a spectrum in a spectrum file. The reference must be in the format <code>ms_run[1-n]:{SPECTRA_REF}</code> where SPECTRA_REF MUST follow the format defined in 5.2. Multiple spectra MUST be referenced using a “|” delimited list for the (rare) cases in which search engines have combined multiple spectra to make identifications.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>PSH sequence … spectra_ref …</code>
<code>PSM KVPQVSTPTLVEVSR … ms_run[1]:index=5 …</code>
<code>PSM VFDEFKPLVEEPQNLIK … ms_run[2]:index=7|ms_run[2]:index=9 …</code></td></tr>
</table>

6.5.17 pre

<table>
<tr><td><b>Description:</b></td><td>Amino acid preceding the peptide (coming from the PSM) in the protein sequence. If unknown “null” MUST be used, if the peptide is N-terminal “-“</td></tr>
</table>

<table>
<tr><td></td><td>MUST be used.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … pre post …
PSM KVPQVSTPTLVEVSR … K D …
PSM VFDEFKPLVEEPQNLIK … R L …</td></tr>
</table>

6.5.18 post

<table>
<tr><td><b>Description:</b></td><td>Amino acid following the peptide (coming from the PSM) in the protein sequence. If unknown “null” MUST be used, if the peptide is C-terminal “-“ MUST be used.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … pre post …
PSM KVPQVSTPTLVEVSR … K D …
PSM VFDEFKPLVEEPQNLIK … R L …</td></tr>
</table>

6.5.19 start

<table>
<tr><td><b>Description:</b></td><td>The start position of the peptide (coming from the PSM) within the protein, counting 1 as the N-terminus of the protein.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … start end …
PSM KVPQVSTPTLVEVSR … 45 57 …
PSM VFDEFKPLVEEPQNLIK … 34 46 …</td></tr>
</table>

6.5.20 end

<table>
<tr><td><b>Description:</b></td><td>The end position of the peptide (coming from the PSM) within the protein, counting 1 as the N-terminus of the protein.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … start end …
PSM KVPQVSTPTLVEVSR … 45 57 …
PSM VFDEFKPLVEEPQNLIK … 34 46 …</td></tr>
</table>

6.5.21 opt_global_*

<table>
<tr><td><b>Description:</b></td><td>Additional columns can be added to the end of the PSM table. These column headers MUST start with the prefix “opt_” followed by the identifier of the object they reference: assay, study variable, MS run or “global” (if the value relates to all replicates). Column names MUST only contain the following characters: ‘A’-‘Z’, ‘a’-‘z’, ‘0’-‘9’, ‘_’, ‘-’, ‘[’, ‘]’, and ‘:’. CV parameter accessions MAY be used for optional columns following the format: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘_’.</td></tr>
<tr><td><b>Type:</b></td><td>Column</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><th></th><th>Summary</th><th>Complete</th></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td>PSH sequence … opt_assay[1]_my_value opt_global_another_value
PSM PEPTIDER … My value about assay[1] some other value that is across reps</td></tr>
</table>

6.6 Small Molecule Section

The small molecule section is table-based. The small molecule section MUST always come after the metadata section, peptide section and or protein section if they are present in the file. All table columns MUST be Tab separated. There MUST NOT be any empty cells. Missing values MUST be reported using “null”. Most columns are mandatory. The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

6.6.1 identifier

<table>
<tr><td><b>Description:</b></td><td>A list of “|” separated possible identifiers for these small molecules.
The database identifier must be preceded by the resource description followed by a colon (in case this is not already part of the identifier format).</td></tr>
<tr><td><b>Type:</b></td><td>String List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier   …</code>
<code>SML CID:00027395   …</code>
<code>SML HMDB:HMDB12345 …</code></td></tr>
</table>

6.6.2 chemical_formula

<table>
<tr><td><b>Description:</b></td><td>The chemical formula of the identified compound.
This should be specified in Hill notation (EA Hill 1900), i.e. elements in the order C, H and then alphabetically all other elements. Counts of one may be omitted. Elements should be capitalized properly to avoid confusion (e.g., “CO” vs. “Co”). The chemical formula reported should refer to the neutral form. Charge state is reported by the charge field. This permits the comparison of positive and negative mode results.

<b>Example:</b> N-acetylglucosamine would be encoded by the string “C8H15NO6”</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier   chemical_formula …</code>
<code>SML CID:00027395   C17H20N4O2   …</code></td></tr>
</table>

6.6.3 smiles

<table>
<tr><td><b>Description:</b></td><td>The molecules structure in the simplified molecular-input line-entry system (SMILES). If there are more than one SMILES for a given small molecule, use the “|” separator.</td></tr>
<tr><td><b>Type:</b></td><td>String List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier   … chemical_formula smiles   …</code>
<code>SML CID:00027395   … C17H20N4O2   C1=CC=C(C=C1)CCNC(=O)CCNNC(=O)C2=CC=NC=C2 …</code></td></tr>
</table>

6.6.4 inchi_key

<table>
<tr><td><b>Description:</b></td><td>The standard IUPAC International Chemical Identifier (InChI) Key of the given substance. If there are more than one InChI identifier for a given small molecule, use the “|” separator.</td></tr>
<tr><td><b>Type:</b></td><td>String List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr></table></td></tr>
</table>

	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH identifier … chemical_formula … inchi_key … SML CID:00027395 … C17H20N4O2 … QXBMEGUKVLFJAM-UHFFFAOYSA-N …		

6.6.5 description

Description:	The small molecule's description / name.		
Type:	String		
Mandatory		Summary	Complete
	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH identifier … description … SML CID:00027395 … N-(2-phenylethyl)-3-[2-(pyridine-4-carbonyl)hydrazinyl]propanamide…		

6.6.6 exp_mass_to_charge

Description:	The small molecule's experimental mass to charge (*m/z*).		
Type:	Double		
Mandatory		Summary	Complete
	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH sequence … mass_to_charge … SMM CID:00027395 … 1234.4 …		

6.6.7 calc_mass_to_charge

Description:	The small molecule's precursor's calculated (theoretical) mass to charge ratio.		
Type:	Double		
Mandatory		Summary	Complete
	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH identifier … mass_to_charge … SML CID:00027395 … 1234.5 …		

6.6.8 charge

Description:	The charge assigned by the search engine/software.		
Type:	Integer		
Mandatory		Summary	Complete
	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH identifier … charge … SML CID:00027395 … 2 …		

6.6.9 retention_time

Description:	A '|'-separated list of time points. Semantics may vary. This time should refer to the small molecule's retention time if determined or the mid point between the first and last spectrum identifying the small molecule. It MUST be reported in seconds. Otherwise, the corresponding unit MUST be specified in the Metadata Section ('columnit_smallmolecule').		
Type:	Double List		
Mandatory		Summary	Complete
	Quantification	✓	✓
	Identification	✓	✓
Example:	SMH identifier … retention_time … SML CID:00027395 … 10.2|11.5 …		

6.6.10 taxid

Description:	The taxonomy id coming from the NEWT taxonomy for the species (if applicable).

<table>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … taxid …</code>
<code>SML CID:00027395 … null …</code></td></tr>
</table>

6.6.11 species

<table>
<tr><td><b>Description:</b></td><td>The species as a human readable string (if applicable).</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … species …</code>
<code>SML CID:00027395 … null …</code></td></tr>
</table>

6.6.12 database

<table>
<tr><td><b>Description:</b></td><td>Generally references the used spectral library (if applicable).</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … database …</code>
<code>SML CID:00027395 … name of used database …</code></td></tr>
</table>

6.6.13 database_version

<table>
<tr><td><b>Description:</b></td><td>Either the version of the used database if available or otherwise the date of creation.
Additionally, the number of entries in the database MAY be reported in round brackets after the version in the format: {version} ({#entries} entries), for example “2011-11 (1234 entries)”.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … database_version …</code>
<code>SML CID:00027395 … 2011-12-22 …</code></td></tr>
</table>

6.6.14 reliability

<table>
<tr><td><b>Description:</b></td><td>The reliability of the given small molecule identification. This must be supplied by the resource and MUST be reported as an integer between 1-4:
1: identified metabolites
2: putatively annotated compounds
3: putatively characterized compound classes
4: unknown compounds</td></tr>
<tr><td><b>Type:</b></td><td>Integer</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … reliability …</code>
<code>SML CID:00027395 … 3 …</code></td></tr>
</table>

6.6.15 uri

<table>
<tr><td><b>Description:</b></td><td>A URI pointing to the small molecule’s entry in the experiment it was identified in (e.g., the small molecule’s PRIDE entry).</td></tr>
<tr><td><b>Type:</b></td><td>URI</td></tr>
</table>

<table>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … uri …</code>
<code>SML CID:00027395 … http://www.ebi.ac.uk/pride/link/to/identification …</code></td></tr>
</table>

6.6.16 spectra_ref

<table>
<tr><td><b>Description:</b></td><td>Reference to a spectrum in a spectrum file. The reference must be in the format ms_run[1-n]:{SPECTRA_REF} where spectra_ref MUST follow the format defined in 5.2. Multiple spectra can be referenced using a “|” delimited list.</td></tr>
<tr><td><b>Type:</b></td><td>String</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … spectra_ref …</code>
<code>SML CID:00027395 … ms_run[1]:index=1002 …</code></td></tr>
</table>

6.6.17 search_engine

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine(s) used to identify this small molecule. Search engines must be supplied as parameters.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … search_engine …</code>
<code>SML CID:00027395 … [MS, MS:1001477, SpectraST,] …</code></td></tr>
</table>

6.6.18 best_search_engine_score

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of best search engine score(s) across replicates for the given small molecule. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td>✓</td><td>✓</td></tr><tr><td>Identification</td><td>✓</td><td>✓</td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … search_engine_score …</code>
<code>SML CID:00027395 … [MS, MS:1001419, SpectraST:discriminant score F, 0.7] …</code></td></tr>
</table>

6.6.19 search_engine_score_ms_run[1-n]

<table>
<tr><td><b>Description:</b></td><td>A “|” delimited list of search engine score(s) in each MS run for the given small molecule. Scores SHOULD be reported using CV parameters whenever possible.</td></tr>
<tr><td><b>Type:</b></td><td>Parameter List</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td>✓</td></tr><tr><td>Identification</td><td>[1]</td><td>[1]</td></tr></table>[1]Not recommended in identification only files</td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier … search_engine_score …</code>
<code>SML CID:00027395 … [MS, MS:1001419, SpectraST:discriminant score F, 0.7] …</code></td></tr>
</table>

6.6.20 modifications

<table>
<tr><td><b>Description:</b></td><td>The small molecule’s modifications or adducts. The position of the modification must be given relative to the small molecule’s beginning. The exact semantics of this position depends on the type of small molecule identified. In case the position information is unknown or not applicable it should not be supplied. For detailed information see protein table.</td></tr>
</table>

Type: String

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification	✓	✓

Example:

```
COM  example where an ammonium loss is found and the position is not
COM  applicable in the given small molecule

SMH  identifier    …  modifications   …
SML  CID:00027395  …  CHEMMOD:-NH4    …

COM  reporting adducts: sodiated glycine
SMH  …  formula  …  charge  …  modifications
SML  …  C2H5NO2  …       1  …  CHEMMOD:+Na-H
```

6.6.21 smallmolecule_abundance_assay[1-n]

Description: The small molecule's abundance in the given assays.

Type: Double

Mandatory

	Summary	Complete
Quantification	(✓)[1]	(✓)[1]
Identification		

[1] mandatory if assays are reported

Example:

```
SMH  identifier    …  smallmolecule_abundance_assay[1] …
SML  CID:00027395  …  0.3                              …
```

6.6.22 smallmolecule_abundance_study_variable[1-n]

Description: The small molecule's abundance in the given study variables.

Type: Double

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification		

Example:

```
SMH  identifier    …  smallmolecule_abundance_study_variable[1] …
SML  CID:00027395  …  0.3                              …
```

6.6.23 smallmolecule_abundance_stdev_study_variable [1-n]

Description: The standard deviation of the small molecule's abundance in the given study variable.

Type: Double

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification		

[1] In case the abundance for a respective study variable is given the standard deviation column MUST also be present (in case the value is not available "null" MUST be used).

Example:

```
SMH  identifier   … smallmolecule_abundance_study_variable[1]
smallmolecule_abundance_stdev_study_variable[1]…
SML  CID:00027395… 0.3                              0.04                    …
```

6.6.24 smallmolecule_abundance_std_error_study_variable[1-n]

Description: The standard error of the small molecule's abundance in the given study variable.

Type: Double

Mandatory

	Summary	Complete
Quantification	✓	✓
Identification		

[1] In case the abundance for a respective study variable is given the standard error column MUST also be present (in case the value is not available "null" MUST be used).

Example:

```
SMH  identifier    …  smallmolecule_abundance_std_error_study_variable[1] …
SML  CID:00027395  …  0.04                                    …
```

6.6.25 opt_global_*

Description: Additional columns can be added to the end of the small molecule table.

<table>
<tr><td></td><td>These column headers MUST start with the prefix “opt_” followed by the identifier of the object they reference: assay, study variable, MS run or “global” (if the value relates to all replicates). Column names MUST only contain the following characters: ‘A’-‘Z’, ‘a’-‘z’, ‘0’-‘9’, ‘_’, ‘-’, ‘[’, ‘]’, and ‘:’. CV parameter accessions MAY be used for optional columns following the format: opt_{OBJECT_ID}_cv_{accession}_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘_’.</td></tr>
<tr><td><b>Type:</b></td><td>Column</td></tr>
<tr><td><b>Mandatory</b></td><td><table><tr><td></td><td>Summary</td><td>Complete</td></tr><tr><td>Quantification</td><td></td><td></td></tr><tr><td>Identification</td><td></td><td></td></tr></table></td></tr>
<tr><td><b>Example:</b></td><td><code>SMH identifier   … opt_assay[1]_my_value opt_global_another_value</code>
<code>SML CID:00027395 … My value   some other value</code></td></tr>
</table>

7. Non-supported use cases

There are a number of use cases that were discussed during the development process and it was decided that they are not explicitly supported in mzTab version 1.0. They may be implemented in future versions of the standard.

- Sequence Tag approaches.
- Grouped modification position scoring systems.

8. Conclusions

This document contains the specifications for using the mzTab format to represent results from peptide, small molecule and protein identification pipelines, in the context of a proteomics investigation. This specification constitutes a proposal for a standard from the Proteomics Standards Initiative. These artefacts are currently undergoing the PSI document process, which will result in a standard officially sanctioned by PSI.

9. Authors

Johannes Griss, European Bioinformatics Institute, United Kingdom
Timo Sachsenberg, University of Tübingen, Germany
Mathias Walzer, Center for Bioinformatics, University of Tübingen, Germany
Oliver Kohlbacher, Center for Bioinformatics, University of Tübingen, Germany
Andrew R. Jones, University of Liverpool, United Kingdom
Henning Hermjakob, European Bioinformatics Institute, United Kingdom
Juan Antonio Vizcaíno, European Bioinformatics Institute, United Kingdom

Correspondence – Henning Hermjakob (hhe@ebi.ac.uk), Juan Antonio Vizcaíno (juan@ebi.ac.uk)

10. Contributors

In addition to the authors, the following people contributed to the model development, gave feedback or tested mzTab:

- Nuno Bandeira, Center for Computational Mass Spectrometry, University of California, San Diego, CA, USA.
- Robert J. Chalkley, Department of Pharmaceutical Chemistry, University of California San Francisco, CA, USA.
- Jürgen Cox, Max Planck Institute of Biochemistry, Martinsried, Germany.
- Martin Eisenacher, Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany.
- Jun Fan, Queen Mary University of London, United Kingdom.
- Laurent Gatto, University of Cambridge, Cambridge, United Kingdom.
- Jürgen Hartler, Graz University of Technology, Graz, Austria.
- Gerhard Mayer, Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany.
- Nadin Neuhauser, Max Planck Institute of Biochemistry, Martinsried, Germany.
- Steffen Neumann, Leibniz Institute of Plant Biochemistry, Halle, Germany.
- Florian Reisinger, European Bioinformatics Institute, Cambridge, United Kingdom.
- Reza M. Salek, European Bioinformatics Institute, Cambridge, United Kingdom.
- Christoph Steinbeck, European Bioinformatics Institute, Cambridge, United Kingdom.
- Gerhard Thallinger, Graz University of Technology, Graz, Austria.
- Ioannis Xenarios, Swiss Institute of Bioinformatics, Geneva, Switzerland.
- Qing-Wei Xu, European Bioinformatics Institute, United Kingdom.

11. References

- Bradner, S. (1997). Key words for use in RFCs to Indicate Requirement Levels, Internet Engineering Task Force. RFC 2119.
- Martens, L., et al. (2011). "mzML--a community standard for mass spectrometry data." Mol Cell Proteomics 10(1): R110 000133.
- Jones, A. R., et al. (2012). "The mzIdentML data standard for mass spectrometry-based proteomics results." Mol Cell Proteomics doi:10.1074/mcp.M111.014381
- EA Hill (1900). “ON A SYSTEM OF INDEXING CHEMICAL LITERATURE; ADOPTED BY THE CLASSIFICATION DIVISION OF THE U. S. PATENT OFFICE.” J. Am. Chem. Soc. 22 (8): 478–494. doi:10.1021/ja02046a005
- Walzer at al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics (2013) Mol Cell Proteomics doi: 10.1074 *mcp.O113.028506.*

12. Intellectual Property Statement

The PSI takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the PSI Chair.

The PSI invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to practice this recommendation. Please address the information to the PSI Chair (see contacts information at PSI website).

TradeMark Section

Microsoft Excel®

Copyright Notice

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."