**mzTab: exchange format for proteomics and metabolomics results**

Status of This Document

This document presents the final specification of the mzTab data format developed by members of the Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) Proteomics Informatics (PI) Working Group, in collaboration with the Metabolomics Standards initiative (MSI) and COSMOS (COordination of Standards in MetabOlomicS) organizations. Distribution is unlimited.

Version of This Document

The current version of this document is: version 1.1.0-draft, Aug 2017.

# 

# Abstract

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) and Metabolomics Standards initiative (MSI) define community standards for data representation in proteomics/metabolomics to facilitate data comparison, exchange and verification. In this context, the two organizations are working together on a shared standard for downstream results, following mass spectrometry (MS) analysis. This document defines a tab delimited text file format to report proteomics and metabolomics results.

**Contents**

[Abstract 1](#_Toc489862664)

[1. Introduction 4](#_Toc489862665)

[1.1 Background 4](#_Toc489862666)

[1.2 Document Structure 5](#_Toc489862667)

[2. Use Cases for mzTab 5](#_Toc489862668)

[3. Notational Conventions 6](#_Toc489862669)

[4. Relationship to Other Specifications 6](#_Toc489862670)

[4.1 The PSI Mass Spectrometry Controlled Vocabulary (CV) 6](#_Toc489862671)

[5. Resolved Design and scope issues 7](#_Toc489862672)

[5.1 Handling updates to the controlled vocabulary 8](#_Toc489862673)

[5.2 Use of identifiers for input spectra to a search 8](#_Toc489862674)

[5.3 Recommendations for reporting replicates within experimental designs 9](#_Toc489862675)

[5.4 mzTab types ‘Identification’ and ‘Quantification’ 11](#_Toc489862676)

[5.5 mzTab modes ‘Summary’ and ‘Complete’ 11](#_Toc489862677)

[5.6 Recommendations for reporting protein inference 16](#_Toc489862678)

[5.7 Recommendations for reporting quantification results 16](#_Toc489862679)

[5.8 Reporting modifications and amino acid substitutions 17](#_Toc489862680)

[5.9 Encoding missing values, zeroes, nulls, infinity and calculation errors 19](#_Toc489862681)

[5.10 Number of peptides reported 19](#_Toc489862682)

[5.11 Reliability score 20](#_Toc489862683)

[5.12 Comments on Specific Use Cases 20](#_Toc489862684)

[5.13 Other supporting materials 22](#_Toc489862685)

[6. Format specification 23](#_Toc489862686)

[6.1 Sections 24](#_Toc489862687)

[6.2 Metadata Section 24](#_Toc489862688)

[MTD uri[1] http://www.ebi.ac.uk/pride/url/to/experiment 29](#_Toc489862689)

[1mandatory if protein section is present 32](#_Toc489862690)

[1mandatory if quantification is reported on assays 36](#_Toc489862691)

[2not recommended for identification only files 36](#_Toc489862692)

[String 39](#_Toc489862693)

[**Mandatory** 39](#_Toc489862694)

[**Example:** 39](#_Toc489862695)

[MTD colunit-protein molecular\_mass=[UO, UO:0000222, kilodalton,] 39](#_Toc489862696)

[String 39](#_Toc489862697)

[**Example:** 39](#_Toc489862698)

[MTD colunit-peptide retention\_time=[UO,UO:0000031, minute,] 39](#_Toc489862699)

[String 40](#_Toc489862700)

[**Example:** 40](#_Toc489862701)

[MTD colunit-psm retention\_time=[UO,UO:0000031, minute,] 40](#_Toc489862702)

[String 40](#_Toc489862703)

[**Example:** 40](#_Toc489862704)

[String 40](#_Toc489862705)

[**Example:** 40](#_Toc489862706)

[String 40](#_Toc489862707)

[**Example:** 40](#_Toc489862708)

[6.3 Protein Section 40](#_Toc489862709)

[6.4 Peptide Section 47](#_Toc489862710)

[1 47](#_Toc489862711)

[1 47](#_Toc489862712)

[1Not recommended in identification only files 47](#_Toc489862713)

[1 47](#_Toc489862714)

[1 47](#_Toc489862715)

[1 47](#_Toc489862716)

[1 47](#_Toc489862717)

[Identification 47](#_Toc489862718)

[1 47](#_Toc489862719)

[1 47](#_Toc489862720)

[1 48](#_Toc489862721)

[1 48](#_Toc489862722)

[1 48](#_Toc489862723)

[1 48](#_Toc489862724)

[1 48](#_Toc489862725)

[1 48](#_Toc489862726)

[1 48](#_Toc489862727)

[1 48](#_Toc489862728)

[1 49](#_Toc489862729)

[1 49](#_Toc489862730)

[1 49](#_Toc489862731)

[1 49](#_Toc489862732)

[1 50](#_Toc489862733)

[1 50](#_Toc489862734)

[1 50](#_Toc489862735)

[1 50](#_Toc489862736)

[1 50](#_Toc489862737)

[1 50](#_Toc489862738)

[1 50](#_Toc489862739)

[1 50](#_Toc489862740)

[1 51](#_Toc489862741)

[1 51](#_Toc489862742)

[1 51](#_Toc489862743)

[1 51](#_Toc489862744)

[2Mandatory only if MS2 based quantification is used 51](#_Toc489862745)

[1 51](#_Toc489862746)

[1 51](#_Toc489862747)

[1 51](#_Toc489862748)

[1 51](#_Toc489862749)

[1 51](#_Toc489862750)

[1 51](#_Toc489862751)

[1 52](#_Toc489862752)

[1 52](#_Toc489862753)

[1 52](#_Toc489862754)

[1 52](#_Toc489862755)

[6.5 PSM Section 52](#_Toc489862756)

[6.6 Small Molecule Section 62](#_Toc489862757)

[6.7 Small Molecule Feature (SMF) Section 72](#_Toc489862758)

[6.8 Small Molecule Evidence (SME) Section 82](#_Toc489862759)

[7. Non-supported use cases 86](#_Toc489862760)

[8. Conclusions 86](#_Toc489862761)

[9. Authors 86](#_Toc489862762)

[10. Contributors 87](#_Toc489862763)

[11. References 87](#_Toc489862764)

[12. Intellectual Property Statement 88](#_Toc489862765)

[TradeMark Section 88](#_Toc489862766)

[Copyright Notice 88](#_Toc489862767)

# Introduction

## 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):

1. *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.
2. *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.
3. *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.
4. *Act as an output format of (web-) services* that report MS-based results and thus can produce standardized result pages.
5. *Allow the combination of MS-based proteomics and metabolomics experimental results* within a single file.
6. *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.

## 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.

# 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.

# 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](#_ENREF_1)).

# 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.

## 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](mailto: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>).

# 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.

## 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](mailto: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.

## 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 ...

## Recommendations for reporting replicates within experimental designs

Modeling the correct reporting of technical/biological replicates within experimental designs is supported in mzTab as shown in 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 analyzed, 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 analyzed (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 analyzed by MS.
* Technical replicates are where same samples are analyzed 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.

## Recommendations for reporting quantification results

Quantitative technologies generally result in some kind of abundance measurement of the identified analyte. Furthermore, several of the available techniques, furthermore, allow/require multiple samples to be multiplexed and analyzed in a single MS run – for example in label-based techniques, such as SILAC/N15 where quantification occurs on MS1 data or in tag-based techniques, such as iTRAQ/TMT where quantification occurs in MS2 data.

One measurement of a small molecule, peptide or protein is mapped to the concept of assay in multiplexed techniques and label-free approaches 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/N15) 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.

Ratios can be generated by the file consumers based on the abundance values of the relevant conditions.

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

## Reporting derivatization approaches

## *Some text needed in here about how to encode derivatization results*

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

In the table-based sections there MUST NOT be any empty cells. In case a given property is not available “null” MUST be used, but this is only allowed for cells in which isNullable= “true”. 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 “NaN” (for Not a Number). In some cases, there is ambiguity with respect to these cases: e.g. if there are alignment issues and it is unclear whether a molecule has been quantified with zero abundance or

## Encoding numerical data with a standard encoding

[Insert some text in here about standard numerical encoding, e.g. US default style “x.x”, i.e. using a period for decimal separation and no commas to separate thousands.]

## 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.

## Support for positive and negative modes or pre-fractionation

[Ideal encoding is to put these into separate mzTab files]

## Referencing evidence for small molecule identifications

[Insert text in here to explain about how to encode evidence where multiple features are used to determine molecule identification; this should be optional column on SML row]

## 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.

### 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

### Referencing external resources

[Text in here?]

## Other supporting materials

[Insert references to example files]

# 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
  + SMH for small molecule table header line (the column labels)
  + SML for rows of the small molecule table
  + SMF for rows of the small molecule feature table
  + SME for rows of the small molecule evidence table
  + HEADERS?
  + 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,]   
  [,,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)

## 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).

## Metadata Section

The metadata section provides additional information about the dataset(s) reported in the mzTab file. All fields in the metadata section are optional apart from those noted as mandatory. 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.

**Core Metadata**

### mzTab-version

|  |  |
| --- | --- |
| **Description:** | The version of the mzTab file. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD mzTab-version 1.1.0 |

### mzTab-ID

|  |  |
| --- | --- |
| **Description:** | The ID of the mzTab file. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD mzTab-ID PRIDE\_1234 |

### title

|  |  |
| --- | --- |
| **Description:** | The file’s human readable title. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD title My first test experiment |

### description

|  |  |
| --- | --- |
| **Description:** | The file’s human readable description. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD description An experiment investigating the effects of Il-6. |

### sample\_processing[1-n]

|  |  |
| --- | --- |
| **Description:** | 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 “|”. |
| **Type:** | Parameter List |
| **Mandatory** | False |
| **Example:** | MTD sample\_processing[1] [SEP, SEP:00173, SDS PAGE,] MTD sample\_processing[2] [SEP, SEP:00142, enzyme digestion,]|[MS, …  MS:1001251, Trypsin, ] |

### instrument[1-n]-name

|  |  |
| --- | --- |
| **Description:** | The name of the instrument used in the experiment. Multiple instruments are numbered 1..n. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | 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] |

### instrument[1-n]-source

|  |  |
| --- | --- |
| **Description:** | The instrument's source used in the experiment. Multiple instruments are numbered 1..n. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD instrument[1]-source [MS, MS:1000073, ESI,] … MTD instrument[2]-source [MS, MS:1000598, ETD,] |

### instrument[1-n]-analyzer[1-n]

|  |  |
| --- | --- |
| **Description:** | The instrument’s analyzer type used in the experiment. Multiple instruments are enumerated 1..n. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD instrument[1]-analyzer[1] [MS, MS:1000291, linear ion trap,] … MTD instrument[2]-analyzer[1] [MS, MS:1000484, orbitrap,] |

### instrument[1-n]-detector

|  |  |
| --- | --- |
| **Description:** | The instrument's detector type used in the experiment. Multiple instruments are numbered 1..n. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD instrument[1]-detector [MS, MS:1000253, electron multiplier,] … MTD instrument[2]-detector [MS, MS:1000348, focal plane collector,] |

### software[1-n]

|  |  |
| --- | --- |
| **Description:** | 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. |
| **Type:** | Parameter |
| **Mandatory** | True |
| **Example:** | MTD software[1] [MS, MS:1001207, Mascot, 2.3] MTD software[2] [MS, MS:1001561, Scaffold, 1.0] |

### software[1-n]-setting[1-n]

|  |  |
| --- | --- |
| **Description:** | 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. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD software[1]-setting Fragment tolerance = 0.1 Da  MTD software[2]-setting Parent tolerance = 0.5 Da |

### publication[1-n]

|  |  |
| --- | --- |
| **Description:** | 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 “|”. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD publication[1] pubmed:21063943|doi:10.1007/978-1-60761-987-1\_6 MTD publication[2] pubmed:20615486|doi:10.1016/j.jprot.2010.06.008 |

### contact[1-n]-name

|  |  |
| --- | --- |
| **Description:** | 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). |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD contact[1]-name James D. Watson … MTD contact[2]-name Francis Crick |

### contact[1-n]-affiliation

|  |  |
| --- | --- |
| **Description:** | The contact’s affiliation. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD contact[1]-affiliation Cambridge University, UK MTD contact[2]-affiliation Cambridge University, UK |

### contact[1-n]-email

|  |  |
| --- | --- |
| **Description:** | The contact’s e-mail address. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD contact[1]-email watson@cam.ac.uk … MTD contact[2]-email crick@cam.ac.uk |

### uri[1-n]

|  |  |
| --- | --- |
| **Description:** | A URI pointing to the file's source data (e.g., a PRIDE experiment, PeptideAtlas build or MetaboLights records). |
| **Type:** | URI |
| **Mandatory** | False |
| **Example:** | MTD uri[1] http://www.ebi.ac.uk/pride/url/to/experiment  MTD uri[2] http://proteomecentral.proteomexchange.org/cgi/GetDataset |

### quantification\_method

|  |  |
| --- | --- |
| **Description:** | The quantification method used in the experiment reported in the file. |
| **Type:** | Parameter |
| **Mandatory** | True |
| **Example:** | MTD quantification\_method [MS, MS:1001837, iTRAQ quantitation analysis, ]  MTD quantification\_method [MS, MS:1001838, SRM quantitation analysis, ] |

### assay[1-n]

|  |  |
| --- | --- |
| **Description:** | A name for each assay, to serve as a list of the assays that MUST be reported in the following tables. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD assay[1] first assay  MTD assay[2] second assay |

### assay[1-n]-custom[1-n]

|  |  |
| --- | --- |
| **Description:** | Additional parameters or values for a given assay. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD assay[1]-custom[1] [MS, MS:100XXXX, TO\_COMPLETE, ] |

### assay[1-n]-external\_uri

|  |  |
| --- | --- |
| **Description:** | A reference to further information about the assay, for example via a reference to an object within an ISA-TAB file. |
| **Type:** | URI |
| **Mandatory** | False |
| **Example:** | MTD assay[1]-external\_uri [Example URI to insert] |

### study\_variable[1-n]

|  |  |
| --- | --- |
| **Description:** | A name for each study variable (experimental condition or factor), to serve as a list of the study variables that MUST be reported in the following tables. For software that does not capture study variables, a single study variable MUST be reported, linking to all assays. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD study\_variable[1] [,,control,]  MTD study\_variable[2] [,,1 minute,]  MTD study\_variable[3] [,,5 minute,] |

### assay[1-n]-sample\_ref

|  |  |
| --- | --- |
| **Description:** | An association from a given assay to the sample analysed. |
| **Type:** | {SAMPLE\_ID} |
| **Mandatory** | False |
| **Example:** | MTD assay[1]-sample\_ref sample[1] MTD assay[2]-sample\_ref sample[2] |

### assay[1-n]-ms\_run\_ref

|  |  |
| --- | --- |
| **Description:** | An association from a given assay to the source MS run. All assays MUST reference exactly one ms\_run unless a workflow with pre-fractionation is being encoded, in which case each assay MUST reference *n* ms\_runs where *n* fractions have been collected.  Multiple assays SHOULD reference the same ms\_run to capture multiplexed experimental designs. |
| **Type:** | {MS\_RUN\_ID} |
| **Mandatory** | True |
| **Example:** | MTD assay[1]-ms\_run\_ref ms\_run[1] |

### study\_variable[1-n]-assay\_refs

|  |  |
| --- | --- |
| **Description:** | Comma-separated references to the IDs of assays grouped in the study variable. |
| **Type:** | {ASSAY\_ID}, ... |
| **Mandatory** | True |
| **Example:** | MTD study\_variable[1]-assay\_refs assay[1], assay[2], assay[3] |

### study\_variable\_function[1-n]

|  |  |
| --- | --- |
| **Description:** | The function used to calculate the study variable quantification value if it is reported and the operation used is not arithmetic mean (default) e.g. “geometric mean”, “median”. Multiple terms can be provided if for example, imputation approaches are to be reported. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD small\_molecule-quantification\_unit [PRIDE, PRIDE:0000395, Ratio, ] |

### study\_variable[1-n]-description

|  |  |
| --- | --- |
| **Description:** | A textual description of the study variable. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD study\_variable[1]-description Group B (spike-in 0.74 fmol/uL) |

### study\_variable[1-n]-factors

|  |  |
| --- | --- |
| **Description:** | Additional parameters or factors, separated by bars, that are known about study variables allowing the capture of more complex, such as nested designs. |
| **Type:** | Param List |
| **Mandatory** | False |
| **Example:** | MTD study\_variable[1]-factors [EXAMPLE HEREe.g. param1 = geneKO; param2 = drug treatment] |

### ms\_run[1-n]-location

|  |  |
| --- | --- |
| **Description:** | Location of the external data file e.g. raw files on which analysis has been performed. If the actual location of the MS run is unknown, a “null” MUST be used as a place holder value, since the [1-n] cardinality is referenced elsewhere. If pre-fractionation has been performed, then the second field should be used for the locations of the grouped fractions files. |
| **Type:** | URL |
| **Mandatory** | True |
| **Example:** | MTD ms\_run\_location[1] file://C:\path\to\my\file … MTD ms\_run\_location[2] <ftp://ftp.ebi.ac.uk/path/to/file>  OR  MTD ms\_run\_location[1]-[1] <ftp://ftp.ebi.ac.uk/path/to/file_fraction1>  MTD ms\_run\_location[1]-[2] ftp://ftp.ebi.ac.uk/path/to/file\_fraction2 |

### ms\_run[1-n]-format

|  |  |
| --- | --- |
| **Description:** | A parameter specifying the data format of the external MS data file. If ms\_run[1-n]-format is present, ms\_run[1-n]-id\_format SHOULD also be present. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD ms\_run[1]-format [MS, MS:1000584, mzML file, ]  MTD ms\_run[1]-id\_format [MS, MS:1000530, mzML unique identifier, ] … MTD ms\_run[2]-format [MS, MS:1001062, Mascot MGF file, ]  MTD ms\_run[2]-id\_format [MS, MS:1000774, multiple peak list nativeID format, ] |

### ms\_run[1-n]-id\_format

|  |  |
| --- | --- |
| **Description:** | Parameter specifying the id format used in the external data file. If ms\_run[1-n]-id\_format is present, ms\_run[1-n]-format SHOULD also be present. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD ms\_run[1]-format [MS, MS:1000584, mzML file, ]  MTD ms\_run[1]-id\_format [MS, MS:1000530, mzML unique identifier, ] … MTD ms\_run[2]-format [MS, MS:1001062, Mascot MGF file, ]  MTD ms\_run[2]-id\_format [MS, MS:1000774, multiple peak list nativeID format, ] |

### ms\_run[1-n]-fragmentation\_method[1-n]

|  |  |
| --- | --- |
| **Description:** | The type(s) of fragmentation used in a given ms run. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD ms\_run[1]-fragmentation\_method[1] [MS, MS:1000133, CID, ] … MTD ms\_run[1]-fragmentation\_method[2] [MS, MS:1000422, HCD …, ] |

### ms\_run[1-n]-hash

|  |  |
| --- | --- |
| **Description:** | Hash value of the corresponding external MS data file defined in ms\_run[1-n]-location. If ms\_run[1-n]-hash is present, ms\_run[1-n]-hash\_method SHOULD also be present. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD ms\_run[1]-hash\_method [MS, MS: MS:1000569, SHA-1, ] MTD ms\_run[1]-hash de9f2c7fd25e1b3afad3e85a0bd17d9b100db4b3 |

### ms\_run[1-n]-hash\_method

|  |  |
| --- | --- |
| **Description:** | A parameter specifying the hash methods used to generate the String in ms\_run[1-n]-hash. Specifics of the hash method used MAY follow the definitions of the mzML format. If ms\_run[1-n]-hash is present, ms\_run[1-n]-hash\_method SHOULD also be present. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD ms\_run[1]-hash\_method [MS, MS: MS:1000569, SHA-1, ] MTD ms\_run[1]-hash de9f2c7fd25e1b3afad3e85a0bd17d9b100db4b3 |

### custom[1-n]

|  |  |
| --- | --- |
| **Description:** | Any additional parameters describing the analysis reported. |
| **Type:** | Parameter |
| **Mandatory** | |  |  |  | | --- | --- | --- | |  | Summary | Complete | | Quantification |  |  | | Identification |  |  | | Metabolomics |  |  | |
| **Example:** | MTD custom[1] [,,MS operator, Florian] |

### sample[1-n]

|  |  |
| --- | --- |
| **Description:** | A name for each biological sample, to serve as a list of the samples to be referenced elsewhere in the file. Samples are not mandatory in mzTab files, since the biological origin of analysed samples may often not be known to quantification software. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-custom[1] [,,Extraction date, 2011-12-21] MTD sample[1]-custom[2] [,,Extraction reason, liver biopsy] |

### sample[1-n]-species[1-n]

|  |  |
| --- | --- |
| **Description:** | The respective species of the samples analysed. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | COM Experiment where all samples consisted of the same two species MTD sample[1]-species[1] [NEWT, 9606, Homo sapiens (Human), ] MTD sample[2]-species[1] [NEWT, 12059, Rhinovirus, ]  COM Experiment where different two samples from different species (combinations) COM were analysed as biological replicates.  MTD sample[1]-species[1] [NEWT, 9606, Homo sapiens (Human), ] MTD sample[1]-species[2] [NEWT, 573824, Human rhinovirus 1, ] MTD sample[2]-species[1] [NEWT, 9606, Homo sapiens (Human), ] MTD sample[2]-species[2] [NEWT, 12130, Human rhinovirus 2, ] |

### sample[1-n]-tissue[1-n]

|  |  |
| --- | --- |
| **Description:** | The respective tissue(s) of the sample. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-tissue[1] [BTO, BTO:0000759, liver, ] |

### sample[1-n]-cell\_type[1-n]

|  |  |
| --- | --- |
| **Description:** | The respective cell type(s) of the sample. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-cell\_type[1] [CL, CL:0000182, hepatocyte, ] |

### sample[1-n]-disease[1-n]

|  |  |
| --- | --- |
| **Description:** | The respective disease(s) of the sample. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-disease[1] [DOID, DOID:684, hepatocellular carcinoma, ] MTD sample[1]-disease[2] [DOID, DOID:9451, alcoholic fatty liver, ] |

### sample[1-n]-description

|  |  |
| --- | --- |
| **Description:** | A human readable description of the sample. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-description Hepatocellular carcinoma samples. MTD sample[2]-description Healthy control samples. |

### sample[1-n]-custom[1-n]

|  |  |
| --- | --- |
| **Description:** | Parameters describing the sample’s additional properties. |
| **Type:** | Parameter |
| **Mandatory** | False |
| **Example:** | MTD sample[1]-custom[1] [,,Extraction date, 2011-12-21] MTD sample[1]-custom[2] [,,Extraction reason, liver biopsy] |

### cv[1-n]-label

|  |  |
| --- | --- |
| **Description:** | A string describing the labels of the controlled vocabularies/ontologies used in the mzTab file |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD cv[1]-label MS  … |

### cv[1-n]-full\_name

|  |  |
| --- | --- |
| **Description:** | A string describing the full names of the controlled vocabularies/ontologies used in the mzTab file |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD cv[1]-full\_name PSI-MS controlled vocabulary  … |

### cv[1-n]-version

|  |  |
| --- | --- |
| **Description:** | A string describing the version of the controlled vocabularies/ontologies used in the mzTab file |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD cv[1]-version 3.54.0  … |

### cv[1-n]-url

|  |  |
| --- | --- |
| **Description:** | A string containing the URLs of the controlled vocabularies/ontologies used in the mzTab file |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD cv[1]-url <http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo>  … |

### database[1-n]

|  |  |
| --- | --- |
| **Description:** | The description of databases used. For cases, where a known database has not been used for identification, a userParam SHOULD be inserted to describe any identification performed or simply “no database”. |
| **Type:** | Param |
| **Mandatory** | True |
| **Example:** | MTD database[1] [MIRIAM,MIR:00100079 , “HMDB”, ]  MTD database[2] [, , “No database”, ]  MTD database[2] [MIRIAM,MIR:00000002 , “CHEBI”, ] |

### database[1-n]-prefix

|  |  |
| --- | --- |
| **Description:** | The prefix used in the “identifier” column of data tables. This MUST be used even for the “no database” case e.g. using prefix “nd”. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD database[1]-prefix hmdb  MTD database[2]-prefix nd |

### database[1-n]-version

|  |  |
| --- | --- |
| **Description:** | The database version is mandatory where identification has been performed. This may be a formal version number e.g. “1.4.1”, a date of access “27/10/2016” or “Unknown” if there is no suitable version that can be annotated. |
| **Type:** | String |
| **Mandatory** | True |
| **Example:** | MTD database[1]-version 3.6 |

### database[1-n]-url

|  |  |
| --- | --- |
| **Description:** | The URL to the database. |
| **Type:** | URL |
| **Mandatory** | True |
| **Example:** | database[1]-url http://www.hmdb.ca/ |

## Metabolomics Metadata

The metadata fields in this section MAY be reported in a metabolomics type file, but MUST NOT be reported in a proteomics file.

### derivatization\_agent[1-n]

|  |  |
| --- | --- |
| **Description:** | A description of derivatization agents applied to small molecules, using userParams or cvParams where possible. |
| **Type:** | Param |
| **Mandatory** | False |
| **Example:** | MTD derivatization\_agent[1] [, PUBCHEM:00XXX, idomethylation, ] |

### small\_molecule-quantification\_unit

|  |  |
| --- | --- |
| **Description:** | Defines what type of units is reported in the small molecule quantification fields. |
| **Type:** | Parameter |
| **Mandatory** | True |
| **Example:** | MTD small\_molecule-quantification\_unit [PSI-MS, MS:000XXXX, Progenesis QI Normalised Abundance, ] |

### small\_molecule\_feature-quantification\_unit

|  |  |
| --- | --- |
| **Description:** | Defines what type of units is reported in the small molecule feature quantification fields. |
| **Type:** | Parameter |
| **Mandatory** | True |
| **Example:** | MTD small\_molecule\_feature-quantification\_unit [PSI-MS, MS:000XXXX, Progenesis QI Normalised Abundance, ] |

### small\_molecule-identification\_reliability

|  |  |
| --- | --- |
| **Description:** | The system used for giving reliability codes to small molecule identifications MUST be specified if not using the default codes. |
| **Type:** | Param |
| **Mandatory** | False |
| **Example:** | MTD small\_molecule-quantification\_unit [PRIDE, PRIDE:0000395, Ratio, ] |

### smallmolecule\_id\_confidence\_measure[1-n]

|  |  |
| --- | --- |
| **Description:** | The type of small molecule confidence measures or scores MUST be reported as a CV parameter [1-n].  The order of the scores SHOULD reflect their importance for the identification and be used to determine the identification’s rank. |
| **Type:** | Param |
| **Mandatory** | True |
| **Example:** | MTD 6.2.17 smallmolecule\_id\_confidence\_measure[1] [, , LipidDataAnalyzer,] |

### colunit-small\_molecule

|  |  |
| --- | --- |
| **Description:** | 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. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD colunit-small\_molecule retention\_time=[UO,UO:0000031, minute,] |

### colunit-small\_molecule\_feature

|  |  |
| --- | --- |
| **Description:** | Defines the used unit for a column in the small molecule feature 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. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD colunit-small\_molecule retention\_time=[UO,UO:0000031, minute,] |

### colunit-small\_molecule\_evidence

|  |  |
| --- | --- |
| **Description:** | Defines the used unit for a column in the small molecule evidence section. The format of the value has to be {column name}={Parameter defining the unit}. |
| **Type:** | String |
| **Mandatory** | False |
| **Example:** | MTD colunit-small\_molecule retention\_time=[UO,UO:0000031, minute,] |

## Small Molecule Section

The small molecule section is table-based. The small molecule section MUST always come after the metadata section in a metabolomics type file. All table columns MUST be Tab separated. There MUST NOT be any empty cells; missing values MUST be reported using “null” for columns where Is Nullable = “True”.

Each row of the small molecule section is intended to report one final result to be communicated in terms of a molecule that has been quantified. In many cases, this may be the molecule of biological interest, although in some cases, the final result could be a derivatized form as appropriate – although it is desirable for the database identifier(s) to reference to the biological (non-derivatized) form. In general, different adduct forms would generally be reported in the Small Molecule Feature section.

The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

All columns are MANDATORY except for “opt\_” columns.

### SML\_ID

|  |  |
| --- | --- |
| **Description:** | A within file unique identifier for the small molecule. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SMH SML\_ID …  SML 1 …  SML 2 … |

### SMF\_ID\_REFS

|  |  |
| --- | --- |
| **Description:** | References to all the features on which quantitation has been based (SMF elements) via referencing SMF\_ID values. Multiple values SHOULD be provided as a “|” separated list. This MAY be null only if this is a Summary file. |
| **Type:** | {SMF\_ID} list |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID SMF\_ID\_REFS SML 1 2|3|11… |

### database\_identifier

|  |  |
| --- | --- |
| **Description:** | A list of “|” separated possible identifiers for the small molecule; multiple values MUST only be provided to indicate ambiguity in the identification of the molecule and not to demonstrate different identifier types for the same molecule. Alternative identifiers for the same molecule MAY be provided as optional columns.  The database identifier must be preceded by the resource description (prefix) followed by a colon, as specified in the Metadata section.  A null value MAY be provided if the identification is sufficiently ambiguous as to be meaningless for reporting or the small molecule has not been identified. |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID identifier …  SML 1 CID:00027395 …  SML 2 HMDB:HMDB12345 … |

### chemical\_formula

|  |  |
| --- | --- |
| **Description:** | A list of “|” separated potential chemical formulae of the reported compound. The number of values provided MUST match the number of entities reported under “database\_identifier”, even if this leads to redundant reporting of information (i.e. if ambiguity can be resolved in the chemical formula), and the validation software will throw an error if the number of “|” symbols does not match. “null” values between bars are allowed.  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.  **Example:** N-acetylglucosamine would be encoded by the string “C8H15NO6” |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … chemical\_formula … SML 1 … C17H20N4O2 … |

### smiles

|  |  |
| --- | --- |
| **Description:** | A list of “|” separated potential molecule structures in the simplified molecular-input line-entry system (SMILES) for the small molecule. The number of values provided MUST match the number of entities reported under “database\_identifier”, and the validation software will throw an error if the number of “|” symbols does not match. “null” values between bars are allowed. |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … chemical\_formula smiles … SML 1 … C17H20N4O2 C1=CC=C(C=C1)CCNC(=O)CCNNC(=O)C2=CC=NC=C2 … |

### inchi

|  |  |
| --- | --- |
| **Description:** | A list of “|” separated potential standard IUPAC International Chemical Identifier (InChI) Keys of the given substance.  The number of values provided MUST match the number of entities reported under “database\_identifier”, , even if this leads to redundant information being reported (i.e. if ambiguity can be resolved in the InChi), and the validation software will throw an error if the number of “|” symbols does not match. “null” values between bars are allowed. |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … chemical\_formula … inchi … SML 1 … C17H20N4O2 … QXBMEGUKVLFJAM-UHFFFAOYSA-N … |

### chemical\_name

|  |  |
| --- | --- |
| **Description:** | A list of “|” separated possible chemical/common names for the small molecule, or general description if a chemical name is unavailable. Multiple names are only to demonstrate ambiguity in the identification. The number of values provided MUST match the number of values provided in database\_identifier. A null value MAY be provided if the small molecule cannot be identified with reasonable confidence. |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … description … SML 1 … N-(2-phenylethyl)-3-[2-(pyridine-4-carbonyl)hydrazinyl]propanamide… |

### theoretical\_neutral\_mass

|  |  |
| --- | --- |
| **Description:** | The small molecule’s precursor’s calculated (theoretical) neutral mass.  The number of values provided MUST match the number of values provided in identifier. Multiple values should be separated by “|”.  This MAY be null for unidentified small molecules. |
| **Type:** | Double List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … calc\_neutral\_mass … SML 1 … 1234.5 … |

### exp\_mass\_to\_charge

|  |  |
| --- | --- |
| **Description:** | The mass to charge of the small molecule’s primary adduct form (e.g. mean m/z across assays), assumed by default to be the protonated (positive mode) or de-protonated (negative mode), otherwise the first reported adduct under the adduct ions column. For GC-MS approaches, this MAY be the m/z of the ion used for quantification. |
| **Type:** | Double |
| **Is Nullable:** | **FALSE** |
| **Example:** | SMH SML\_ID … exp\_mass\_to\_charge … SME 1 … 348.65 … |

### retention\_time

|  |  |
| --- | --- |
| **Description:** | The apex of the small molecule’s primary adduct form on the retention time axis in a Master or aggregate MS run. Retention time MUST be reported in seconds, except units MUST be reported in the Metadata Section (“XXXX”). Retention time values for individual MS runs (i.e. before alignment) MAY be reported as optional columns. Retention time SHOULD only be null in the case of direct infusion MS or other techniques where a retention time value is absent or unknown. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … retention\_time … SML 1 … 638 … |

### adduct\_ions

|  |  |
| --- | --- |
| **Description:** | A “|” separated list of adducts for this this molecule, following the general style in the 2013 IUPAC recommendations on [terms relating to MS](http://dx.doi.org/10.1351/PAC-REC-06-04-06) e.g. [M+H]+, [M+Na]+, [M+NH4]+, [M-H]-, [M+Cl]-. If the adduct classification is ambiguous with regards to identification evidence it MAY be null. |
| **Type:** | String List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … adduct ions … SML 1 … [M+H]1+ | [M+Na]1+ … |

### reliability

|  |  |
| --- | --- |
| **Description:** | The reliability of the given small molecule identification. By default, the following system is used.  This must be supplied by the resource and MUST be reported as an integer between 1-4 [PUT REFERENCE IN HERE]:  1: identified metabolites  2: putatively annotated compounds  3: putatively characterized compound classes  4: unknown compounds  -1: reliability not specified  These MAY be replaced using a suitable CV term in the Metadata section e.g. to use MSI recommendation levels.  If the export software package does not use export a value, then a value of 5 should be given as “-1”.  A String data type is set to allow for different systems to be specified in the metadata section. |
| **Type:** | String |
| **Is Nullable:** | **FALSE** |
| **Example:** | SMH identifier … reliability … SML 1 … 3 … |

### uri

|  |  |
| --- | --- |
| **Description:** | A URI pointing to the small molecule’s entry in a reference database (e.g., the small molecule’s HMDB or KEGG entry). In the case of ambiguity of identification, the number of values provided MUST match the number of values provided in identifier. |
| **Type:** | URI List |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … uri … SML 1 … example\_URL … |

### best\_search\_engine

|  |  |
| --- | --- |
| **Description:** | The search engine that identified this small molecule with highest confidence. |
| **Type:** | Parameter |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … best\_search\_engine … SML 1 … [MS, MS:1001477, SpectraST,] … |

### best\_smallmolecule\_id\_confidence\_measure[1\_n]

|  |  |
| --- | --- |
| **Description:** | The best confidence measure in identification (for this type of score) for the given small molecule across all assays. The type of score MUST be defined in the metadata section. If the small molecule was not identified by the specified search engine, “null” MUST be reported. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | MTD smallmolecule\_id\_confidence\_measure[1]\_ms\_run[1] [MS, MS:1001419, SpectraST:discriminant score F,]  …  SMH SML\_ID … smallmolecule\_id\_confidence[1]\_measure\_ms\_run[1] … SML 1 … 0.7 … |

### smallmolecule\_abundance\_assay[1-n]

|  |  |
| --- | --- |
| **Description:** | The small molecule’s abundance in every assay described in the metadata section MUST be reported. Null or zero values may be reported as appropriate. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … smallmolecule\_abundance\_assay[1] … SML 1 … 0.3 … |

### smallmolecule\_abundance\_study\_variable[1-n]

|  |  |
| --- | --- |
| **Description:** | The small molecule’s abundance in all the study variables described in the Metadata section, calculated using the method as described in the Metadata section (default = arithmetic mean across assays). Null or zero values may be reported as appropriate. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … smallmolecule\_abundance\_study\_variable[1] … SML 1 … 0.3 … |

### smallmolecule\_abundance\_coeffvar\_study\_variable [1-n]

|  |  |
| --- | --- |
| **Description:** | The co-efficient of variation of the small molecule’s abundance in the given study variable. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … smallmolecule\_abundance\_study\_variable[1] smallmolecule\_abundance\_ coeffvar\_study\_variable[1]… SML 1 … 0.3 0.04 … |

### opt\_{identifier}\_\*

|  |  |
| --- | --- |
| **Description:** | Additional columns can be added to the end of the small molecule 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\_{identifier}\_cv\_{accession}\_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘\_’. |
| **Type:** | Column |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … opt\_assay[1]\_my\_value opt\_global\_another\_value SML 1 … My value some other value |

**Example optional columns:**

* Species
* Taxid
* Retention time index values normalised to a given scale
* Identification scores specific to each assay
* Raw quantification values, assuming normalised values are provided in the standard assay quantification columns.

## Small Molecule Feature (SMF) Section

The small molecule feature section is table-based, representing individual MS regions (generally considered to be the elution profile for all isotopomers formed from a single charge state of a molecule), that have been measured/quantified. Different adducts or derivatives and different charge states of individual molecules should be reported as separate SMF rows.

The small molecule feature section MUST always come after the Small Molecule Table. All table columns MUST be Tab separated. There MUST NOT be any empty cells. Missing values MUST be reported using “null”.

The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

All columns are MANDATORY except for “opt\_” columns.

### SMF\_ID

|  |  |
| --- | --- |
| **Description:** | A within file unique identifier for the small molecule feature. |
| **Type:** | String |
| **Is Nullable:** | **FALSE** |
| **Example:** | SFH SMF\_ID …  SMF 1 …  SMF 2 … |

### SME\_ID\_REFS

|  |  |
| --- | --- |
| **Description:** | References to the identification evidence (SME elements) via referencing SME\_ID values. Multiple values MAY be provided as a “|” separated list to indicate ambiguity in the identification. For the case of a consensus approach where multiple adduct forms are used to infer the SML ID, different features should just reference the same SME\_ID value(s). |
| **Type:** | String (REF) list |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID SME\_ID\_REFS SMF 1 5|6|12… |

### SME\_ID\_REF\_Ambiguity\_code

|  |  |
| --- | --- |
| **Description:** | If multiple values are given under SME\_ID\_REFS, one of the following codes MUST be provided. 1=Ambiguous identification; 2=Only different evidence streams for the same molecule with no ambiguity; 3=Both ambiguous identification and multiple evidence streams. If there are no or one value under SME\_ID\_REFs, this MUST be reported as null. |
| **Type:** | Integer |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID SME\_ID\_REFS SME\_ID\_REF\_Ambiguity\_code SMF 1 5|6|12… 1 |

### charge

|  |  |
| --- | --- |
| **Description:** | The feature’s charge value. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SFH SMF\_ID … charge … SMF 1 … 1 … |

### adduct\_ion

|  |  |
| --- | --- |
| **Description:** | The assumed adduct classification of this molecule, following the general style in the 2013 IUPAC recommendations on terms relating to MS e.g. [M+H]+, [M+Na]+, [M+NH4]+, [M-H]-, [M+Cl]-. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID … adduct\_ion … SMF 1 … [M+H]1+ … |

### exp\_mass\_to\_charge

|  |  |
| --- | --- |
| **Description:** | The measured mass/charge value for the feature, by default assumed to be the mean across assays or a representative value. |
| **Type:** | Double |
| **Is Nullable:** | **FALSE** |
| **Example:** | SFH SMF\_ID … exp\_mass\_to\_charge … SML 1 … 1234.5 … |

### retention\_time

|  |  |
| --- | --- |
| **Description:** | The apex of the feature on the retention time axis, in a Master or aggregate MS run. Retention time MUST be reported in seconds. Retention time values for individual MS runs (i.e. before alignment) MAY be reported as optional columns. Retention time SHOULD only be null in the case of direct infusion MS or other techniques where a retention time value is absent or unknown. Relative retention time or retention time index values MAY be reported as optional columns, and could be considered for inclusion in future versions of mzTab as appropriate. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID … retention\_time … SMF 1 … 1345 … |

### retention\_time\_start

|  |  |
| --- | --- |
| **Description:** | The start time of the feature on the retention time axis, in a Master or aggregate MS run. Retention time MUST be reported in seconds. Retention time start and end SHOULD only be null in the case of direct infusion MS or other techniques where a retention time value is absent or unknown and MAY be reported in optional columns. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID … retention\_time\_start … SMF 1 … 1327 … |

### retention\_time\_end

|  |  |
| --- | --- |
| **Description:** | The end time of the feature on the retention time axis, in a Master or aggregate MS run. Retention time MUST be reported in seconds. Retention time start and end SHOULD only be null in the case of direct infusion MS or other techniques where a retention time value is absent or unknown and MAY be reported in optional columns.. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID … retention\_time\_start … SMF 1 … 1327 … |

### quant\_assay[1-n]

|  |  |
| --- | --- |
| **Description:** | The feature’s abundance in every assay described in the metadata section MUST be reported. Null or zero values may be reported as appropriate. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH SML\_ID … quant\_assay[1] … SMF 1 … 38648 … |

### opt\_{identifier}\_\*

|  |  |
| --- | --- |
| **Description:** | Additional columns can be added to the end of the small molecule feature 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\_{identifier}\_cv\_{accession}\_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘\_’. |
| **Type:** | Column |
| **Is Nullable:** | **TRUE** |
| **Example:** | SFH SMF\_ID … opt\_assay[1]\_my\_value opt\_global\_another\_value SMF 1 … My value some other value |

**Example optional columns:**

* (Apex) retention time values for each MS run pre-alignment
* Retention time index values normalised to a given scale
* Raw quantification values, assuming normalised values are provided in the standard assay quantification columns.
* Predicted retention time

## Small Molecule Evidence (SME) Section

The small molecule evidence section is table-based, representing evidence for identifications of small molecules/features, from database search or any other process used to give putative identifications to molecules.

The small molecule evidence section MUST always come after the Small Molecule Feature Table. All table columns MUST be Tab separated. There MUST NOT be any empty cells. Missing values MUST be reported using “null”.

The order of columns is not specified although for ease of human interpretation, it is RECOMMENDED to follow the order specified below.

All columns are MANDATORY except for “opt\_” columns.

### SME\_ID

|  |  |
| --- | --- |
| **Description:** | A within file unique identifier for the small molecule evidence result. |
| **Type:** | String |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID …  SME 1 … |

### evidence\_grouping\_ID

|  |  |
| --- | --- |
| **Description:** | A within file identifier for the data e.g. fragment spectrum, RT and m/z pair, isotope profile that was used for the identification process, to serve as a grouping mechanism, whereby multiple rows of data from the same data share the same ID. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID evidence\_grouping\_ID …  SME 1 1  SME 2 1  SME 3 1  (in this example three identifications were made from the same accurate mass search) |

### database\_identifier

|  |  |
| --- | --- |
| **Description:** | The putative identification for the small molecule sourced from an external database, using the same prefix specified in 1.1.1 small\_molecule-database[1-n]-prefix.  This could include additionally a chemical class or an identifier to a spectral library entity, even if its actual identity is unknown. |
| **Type:** | String |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID identifier …  SME 1 CID:00027395 …  SML 2 HMDB:HMDB12345 … |

### chemical\_formula

|  |  |
| --- | --- |
| **Description:** | The chemical formula of the identified compound e.g. in a database, assumed to match the calculated mass to charge (in some cases this will be the derivatized form, including adducts and protons) – the number of values provided MUST match the number of values provided in identifier.  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.  **Example:** N-acetylglucosamine would be encoded by the string “C8H15NO6” |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … chemical\_formula … SME 1 … C17H20N4O2 … |

### smiles

|  |  |
| --- | --- |
| **Description:** | The potential molecule’s structure in the simplified molecular-input line-entry system (SMILES) for the small molecule. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … chemical\_formula smiles … SML 1 … C17H20N4O2 C1=CC=C(C=C1)CCNC(=O)CCNNC(=O)C2=CC=NC=C2 … |

### inchi

|  |  |
| --- | --- |
| **Description:** | A standard IUPAC International Chemical Identifier (InChI) for the given substance. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … chemical\_formula … inchi … SML 1 … C17H20N4O2 … QXBMEGUKVLFJAM-UHFFFAOYSA-N … |

### chemical\_name

|  |  |
| --- | --- |
| **Description:** | The small molecule’s chemical/common name, or general description if a chemical name is unavailable. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … chemical\_name … SML 1 … N-(2-phenylethyl)-3-[2-(pyridine-4-carbonyl)hydrazinyl]propanamide… |

### Modifications

|  |  |
| --- | --- |
| **Description:** | The small molecule’s modifications or derivatives from the evidence, reported using suitable userParam or cvParams as appropriate. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SMH identifier … modifications … SML CID:00027395 … [,, “methylated”,] … |

### adduct\_ion

|  |  |
| --- | --- |
| **Description:** | The assumed adduct classification of this molecule, following the general style in the 2013 IUPAC recommendations on terms relating to MS e.g. [M+H]+, [M+Na]+, [M+NH4]+, [M-H]-, [M+Cl]-. If the adduct classification is ambiguous with regards to identification evidence it MAY be null. |
| **Type:** | String |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … adduct\_ion … SME 1 … [M+H]1+ … |

### exp\_mass\_to\_charge

|  |  |
| --- | --- |
| **Description:** | The measured mass/charge value for the precursor ion. If multiple adduct forms have been combined into a single identification event/search, then a single value e.g. for the protonated form, SHOULD be reported here. |
| **Type:** | Double |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … exp\_mass\_to\_charge … SME 1 … 1234.5 … |

### charge

|  |  |
| --- | --- |
| **Description:** | The feature’s charge value. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … charge … SME 1 … 1 … |

### calc\_mass\_to\_charge

|  |  |
| --- | --- |
| **Description:** | The calculated (theoretical) mass/charge value for the small molecule or the database mass/charge value (for a spectral library match). |
| **Type:** | Double |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … calc\_mass\_to\_charge … SME 1 … 1234.71 … |

### calc\_neutral\_mass

|  |  |
| --- | --- |
| **Description:** | The calculated (theoretical) neutral mass value for the small molecule or the calculated neutral mass of the matched spectral library entry. |
| **Type:** | Double |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … calc\_neutral\_mass … SME 1 … 1233.71 … |

### uri

|  |  |
| --- | --- |
| **Description:** | A URI pointing to the small molecule’s entry in a database (e.g., the small molecule’s HMDB, Chebi or KEGG entry). |
| **Type:** | URI |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … uri … SME 1 … http://www.hmdb.ca/metabolites/HMDB00054 |

### spectra\_ref

|  |  |
| --- | --- |
| **Description:** | Reference to a spectrum in a spectrum file, for example a fragmentation spectrum has been used to support the identification. If a separate spectrum file has been used for fragmentation spectrum, this MUST be reported in the meta-data section as ms\_run[1]-fragmentation\_spectrum, otherwise the references are assumed to refer to the raw files reported under ms\_run[1]-location. 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 MUST be referenced using a “|” delimited list for the (rare) cases in which search engines have combined multiple spectra to make identifications. [Needs a note on how to reference chromatograms]  If a fragmentation spectrum has not been used, the value should indicate the ms\_run to which is identification is mapped e.g. “ms\_run[1]”. |
| **Type:** | String List |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … spectra\_ref … SME 1 … ms\_run[1]:index=5 … |

### identification\_method

|  |  |
| --- | --- |
| **Description:** | The database look-up, search engine or process that was used to identify this small molecule e.g. the name of software, database or manual curation etc. |
| **Type:** | Parameter |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … search\_engine … SME 1 … [MS, MS:1001477, SpectraST,] … |

### ms\_level

|  |  |
| --- | --- |
| **Description:** | The highest MS level used to inform identification e.g. MS1 (accurate mass only) = “1” or from an MS2 fragmentation spectrum = 2. For MS3 or other approaches, the appropriate integer MAY be used. For data independent approaches where fragmentation data is used, a value of 2 should be inserted. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … ms\_level … SME 1 … 2 … |

### smallmolecule\_id\_confidence\_measure[1-n]

|  |  |
| --- | --- |
| **Description:** | Any statistical value or score for the identification. The metadata section reports the type of score used, as smallmolecule\_id\_confidence\_measure[1-n] of type Param. |
| **Type:** | Double |
| **Is Nullable:** | **TRUE** |
| **Example:** | MTD smallmolecule\_id\_confidence\_measure[1] [MS, MS:1001419, SpectraST:discriminant score F,]  …  SEH SME\_ID … smallmolecule\_id\_confidence\_measure[1] … SME 1 … 0.7 … |

### rank

|  |  |
| --- | --- |
| **Description:** | The rank of this identification from this search engine as increasing integers from 1 (best ranked identification). Ties (equal score) are represented by using the same rank – defaults to 1 if there is no ranking system used. If no ranking system has been used, then a value of 0 SHOULD be used. |
| **Type:** | Integer |
| **Is Nullable:** | **FALSE** |
| **Example:** | SEH SME\_ID … rank … SME 1 … 1 … |

### opt\_{identifier}\_\*

|  |  |
| --- | --- |
| **Description:** | Additional columns can be added to the end of the small molecule evidence 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\_{identifier}\_cv\_{accession}\_{parameter name}. Spaces within the parameter’s name MUST be replaced by ‘\_’. |
| **Type:** | Column |
| **Is Nullable:** | **TRUE** |
| **Example:** | SEH SME\_ID … opt\_assay[1]\_my\_value opt\_global\_another\_value SML 1 … My value some other value |

**Example optional columns:**

* Additional statistical measures or annotations about evidence, such as decoy identifications.

# 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.

# 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.

# 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](mailto:juan@ebi.ac.uk)), Juan Antonio Vizcaíno ([juan@ebi.ac.uk](mailto:juan@ebi.ac.uk))

# 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.
* Fawaz Ghali, University of Liverpool, 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.
* Yasset Pérez-Riverol, European Bioinformatics Institute, Cambridge, United Kingdom.
* 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.
* Noemi del Toro, 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.

# 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.

# 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

Copyright (C) Proteomics Standards Initiative (2017). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

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."