The twenty minute guide to mzTab

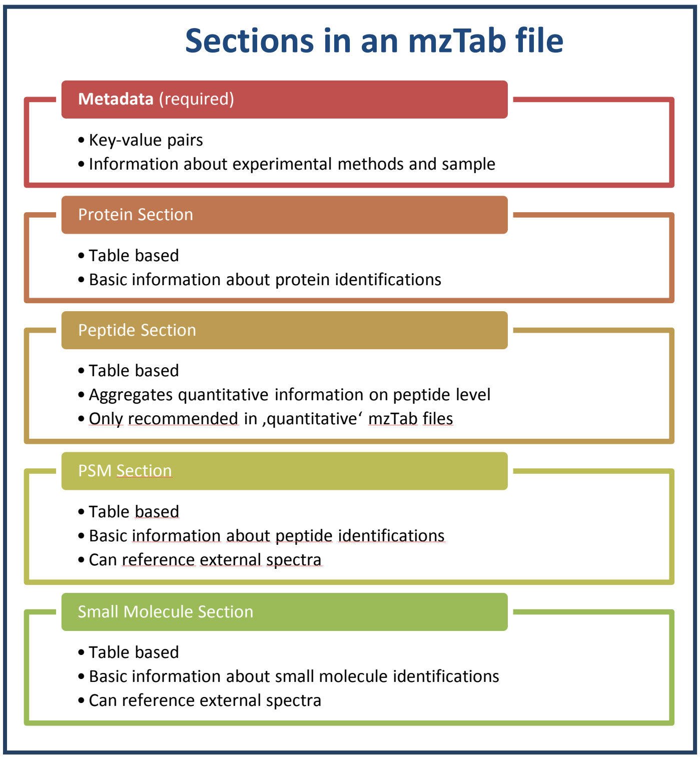
Johannes Griss & Juan Antonio Vizcaíno, EMBL-EBI, juan@ebi.ac.uk, June 2014

Introduction

The purpose of this guide is to give a quick introduction on how to use mzTab efficiently. It is targeted at both, developers and end-users alike. This guide is not intended to give a complete and detailed overview of mzTab but should only be a quick and easy to understand introduction. The complete format specification as well as example files can be found at http://mztab.googlecode.com.

Basic structure

mzTab files can have five sections: The metadata section, the protein section, the peptide section, the peptide-spectrum match (PSM) section and the small molecule section (see Figure 1). All of these sections, apart from the metadata section are optional and may not be present in every file.



Sections in an mzTab file

Metadata (required)

- Key-value pairs
- Information about experimental methods and sample

Protein Section

- Table based
- Basic information about protein identifications

Peptide Section

- Table based
- Aggregates quantitative information on peptide level
- Only recommended in ‚quantitative' mzTab files

PSM Section

- Table based
- Basic information about peptide identifications
- Can reference external spectra

Small Molecule Section

- Table based
- Basic information about small molecule identifications
- Can reference external spectra

Figure 1: Basic structure of an mzTab file.

All lines in an mzTab file start with a three letter code to identify the information held by the line:

MTD	for metadata
PRH	for the protein table header line (the column labels)
PRT	for rows of the protein table
PEH	for the peptide table header line (the column labels)
PEP	for rows of the peptide table
PSH	for the PSM table header line (the column labels)
PSM	for rows in the PSM table
SMH	for small molecule table header line (the column labels)
SML	for rows of the small molecule table

COM for comment lines

The header lines of the table based sections (protein, peptide, PSM, small molecule) must be at the top of these sections and must only occur once in the file (since every section must only occur once).

For developers:

mzTab is a tab separated file format. The three letter codes must be separated by a tab from the next field. Also, field names and values in the metadata section are separated by tabs as are the columns in the table based sections.

Modelling an experimental design in mzTab

mzTab supports the reporting of technical/biological replicates within experimental designs using an adaptation of the system originally developed for mzQuantML. This is made up of four components:

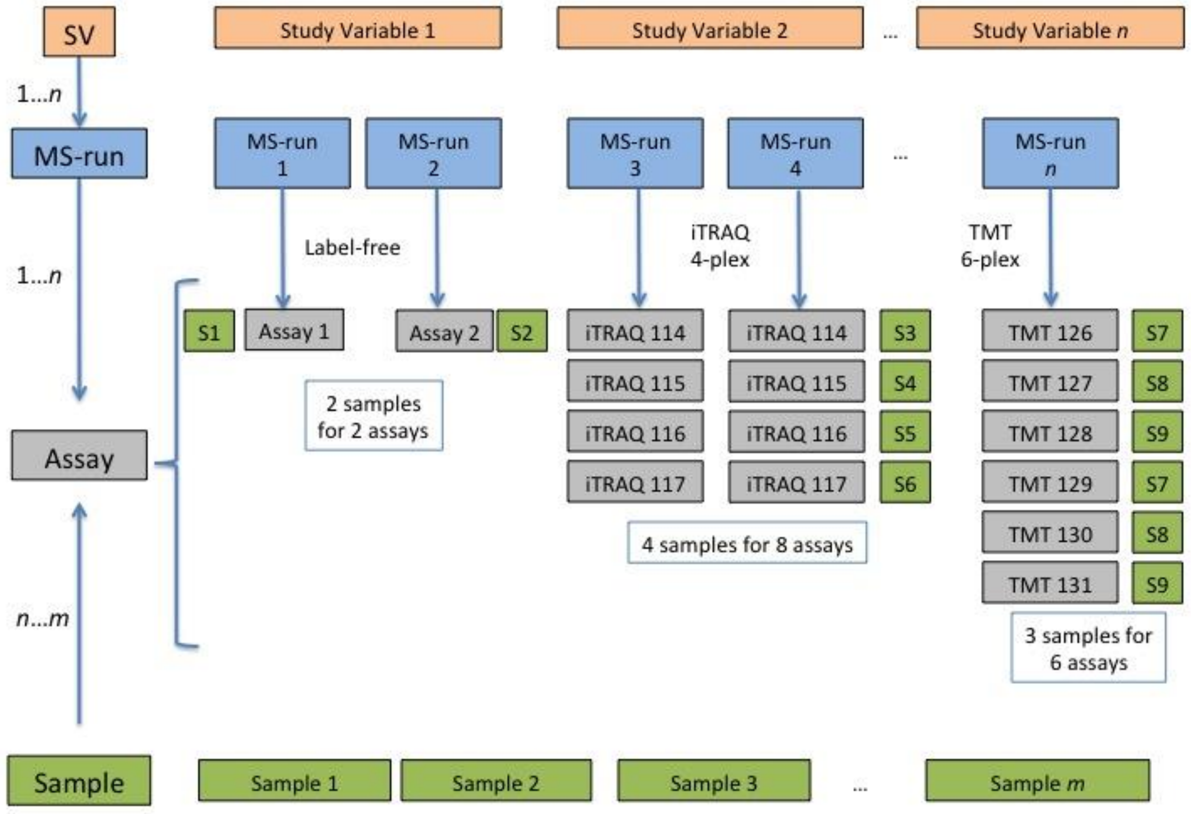


Figure 2: Diagram summarizing the relation between Study Variables (SVs), MS runs, assays and samples.

- Study variable: Study variables represent the core final results of the study (ie. 'inflammatory response' vs. 'control'). Often, these will have been derived from averaging the results of a group of replicate measurements (assays). In files where such assays are reported, study variables reference and thereby group assays. The same concept has been defined 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:** Any quantitative measurement about the sample (in this case through MS) is reported as an assay. In label-free MS analysis one assay is usually mapped to one MS run. If multiplexed techniques are used multiple assays are mapped to one MS run (e.g. iTRAQ). In these cases additional information about the used tag (as a property of the assay) can be reported in the metadata section.

- **Sample:** A sample represents any analysed biological material to which descriptors of species such as cell/tissue type can be applied. In all mzTab files, these can be reported in the metadata section as "sample[1-n]-description". Samples are not mandatory in mzTab, since many software packages cannot determine what type of sample was analysed (e.g. whether biological or technical replication was performed).

See below an example corresponding to one SILAC experiment:

```
COM  Report of a minimal "Complete Quantification report" SILAC experiment, quantification on
COM  2 study variables (control/treatment), 3+3 assays (replicates) reported, no
COM  identifications reported.
COM  Internally 3 replicates/assays have been used to obtain quantification values, stdev and
COM  stderror
MTD  mzTab-version                          1.0.0
MTD  mzTab-mode                             Complete
MTD  mzTab-type                             Quantification
MTD  description                            mzTab example file for reporting a summary…
MTD  ms_run[1]-location                     file://C:\path\to\my\file1.mzML
MTD  ms_run[2]-location                     file://C:\path\to\my\file2.mzML
MTD  ms_run[3]-location                     file://C:\path\to\my\file3.mzML
MTD  ms_run[4]-location                     file://C:\path\to\my\file4.mzML
MTD  protein-quantification_unit            [PRIDE, PRIDE:0000393, Relative quantification
                                                                                       unit,]
MTD  software[1]                             [MS, MS:1001583, MaxQuant,]
MTD  psm_search_engine_score[1]             [MS, MS:1001979, MaxQuant:PTM score,]
MTD  fixed_mod[1]                           [UNIMOD, UNIMOD:4, Carbamidomethyl, ]
MTD  fixed_mod[2]                           [UNIMOD, UNIMOD:188, Label:13C(6), ]
MTD  variable_mod[1]                        [UNIMOD, UNIMOD:35, Oxidation, ]
MTD  quantification_method                  [MS, MS:1001835, SILAC, ]
MTD  assay[1]-quantification_reagent        [PRIDE, PRIDE:0000326, SILAC light, ]
MTD  assay[2]-quantification_reagent        [PRIDE, PRIDE:0000325, SILAC heavy, ]
MTD  assay[3]-quantification_reagent        [PRIDE, PRIDE:0000326, SILAC light, ]
MTD  assay[4]-quantification_reagent        [PRIDE, PRIDE:0000325, SILAC heavy, ]
MTD  assay[1]-ms_run_ref                    ms_run[1]
MTD  assay[2]-ms_run_ref                    ms_run[1]
MTD  assay[3]-ms_run_ref                    ms_run[2]
MTD  assay[4]-ms_run_ref                    ms_run[2]
MTD  study_variable[1]-assay_refs           assay[1],assay[3]
MTD  study_variable[2]-assay_refs           assay[2],assay[4]
MTD  study_variable[1]-description          heat shock response of control
MTD  study_variable[2]-description          heat shock response of treatment
```

Metadata section in mzTab

The metadata section in mzTab files contains information about the units and consists of key - value pairs separated by a tab. A complete list of available fields can be found in the specification document.

```
COM  Example of the metadata section for an identification file.
MTD    mzTab-version               1.0 rc5
MTD    mzTab-mode                  Complete
MTD    mzTab-type                  Identification
MTD    mzTab-ID                    PRIDE assay metadata example
MTD    title                       COFRADIC N-terminal proteome of unstimulated human blood…
MTD    instrument[1]-name          [PRIDE, PRIDE:0000131, Instrument model, Micromass Q-TOF I]
MTD    instrument[1]-source        [MS, MS:1000008, Ionization Type, ESI]
MTD    instrument[1]-analyzer      [1][MS, MS:1000010, Analyzer Type, Quadrupole-TOF]
MTD    instrument[1]-detector      [MS, MS:1000026, Detector Type, MultiChannelPlate]
MTD    software[1]                 [MS, MS:1001456, analysis software, MassLynx v3.5]
MTD    protein_search_engine_score[1][MS, MS:1002367, probability for proteins,]
MTD    publication[1]              pubmed:16038019|pubmed:12665801|pubmed:16518876
MTD    contact[1]-name             Kristian Flikka
MTD    contact[1]-affiliation      Computational Biology Unit, University of Bergen
MTD    contact[1]-email            flikka@ii.uib.no
MTD    ms_run[1]-format            [MS, MS:1000564, PSI mzData file, ]
MTD    ms_run[1]-location  …
       ftp://ftp.ebi.ac.uk/pub/databases/pride/PRIDE_Exp_Complete_Ac_1643.xml
```

```
MTD    ms_run[1]-id_format         [MS, MS:1000777, spectrum identifier nativeID format, ]
MTD    sample[1]-species[1]         [NEWT, 9606, Homo sapiens (Human), ]
MTD    sample[1]-cell_type[1]       [CL, CL:0000233, platelet, ]
MTD    sample[1]-custom[1]          [MeSH, D001792, blood_platelets, ]
MTD    assay[1]-sample_ref          sample[1]
MTD    assay[1]-ms_run_ref          ms_run[1]
```

The number of required columns in the protein table depends on the type of mzTab file (‘Identification’ / ‘Quantification’) and the used mode (‘Complete’ / ‘Summary’):

Metadata Section

Field Name	Identification	Quantification
mzTab-version	**SC**	**SC**
mzTab-mode	**SC**	**SC**
mzTab-type	**SC**	**SC**
description	**SC**	**SC**
ms_run[1-n]-location	**SC**	**SC**
protein_search_engine_score[1-n]	**SC** (if protein section present)	**SC** (if protein section present)
peptide_search_engine_score[1-n]	**SC** (if peptide section present)	**SC** (if peptide section present)
psm_search_engine_score[1-n]	**SC** (if PSM section present)	**SC** (if PSM section present)
smallmolecule_search_engine_score[1-n]	**SC** (if small molecule section present)	**SC** (if small molecule section present)
fixed_mod[1-n]	**SC**	**SC**
variable_mod[1-n]	**SC**	**SC**
protein-quantification-unit	(not used)	**SC** (if protein section present)
peptide-quantification-unit	(not used)	**SC** (if peptide section present)
smallmolecule-quantification-unit	(not used)	**SC** (if small molecule section present)
study_variable[1-n]-description	(not used)	**SC**
software[1-n]	*s***C**	*s***C**
quantification_method	(not used)	*s***C**
assay[1-n]-ms_run_ref	*sc* (required if assays reported)	*s***C** (required if assays reported)
assay[1-n]-quantification_reagent	(not used)	*s***C**
mzTab-ID	*sc*	*sc*
title	*sc*	*sc*
sample_processing[1-n]	*sc*	*sc*
instrument[1-n]-name	*sc*	*sc*
instrument[1-n]-source	*sc*	*sc*
instrument[1-n]-analyzer[1-n]	*sc*	*sc*
instrument[1-n]-detector	*sc*	*sc*
software[1-n]-setting[1-n]	*sc*	*sc*
false_discovery_rate	*sc*	*sc*
publication[1-n]	*sc*	*sc*
contact[1-n]-name	*sc*	*sc*
contact[1-n]-affiliation	*sc*	*sc*
contact[1-n]-email	*sc*	*sc*
uri[1-n]	*sc*	*sc*
fixed_mod[1-n]-site	*sc*	*sc*
fixed_mod[1-n]-position	*sc*	*sc*
variable_mod[1-n]-site	*sc*	*sc*
variable_mod[1-n]-position	*sc*	*sc*
ms_run[1-n]-format	*sc*	*sc*
ms_run[1-n]-id_format	*sc* (required if ms_run[1-n]-format reported)	*sc* (required if ms_run[1-n]-format reported)
ms_run[1-n]-fragmentation_method	*sc*	*sc*

ms_run[1-n]-hash	*sc*	*sc*
ms_run[1-n]-hash_method	*sc* (required if ms_run[1-n]-hash reported)	*sc* (required if ms_run[1-n]-hash reported)
custom[1-n]	*sc*	*sc*
sample[1-n]-species[1-n]	*sc*	*sc*
sample[1-n]-tissue[1-n]	*sc*	*sc*
sample[1-n]-cell_type[1-n]	*sc*	*sc*
sample[1-n]-disease[1-n]	*sc*	*sc*
sample[1-n]-description	*sc*	*sc*
sample[1-n]-custom[1-n]	*sc*	*sc*
study_variable[1-n]-description	*sc* (required if SV reported)	*sc* (required if SV reported)
study_variable[1-n]-sample_refs	*sc*	*sc*
study_variable[1-n]-assay_ref	*sc*	*s***C**
assay[1-n]-quantification_mod[1-n]	(not used)	*sc*
assay[1-n]-quantification_mod[1-n]-position	(not used)	*sc*
assay[1-n]-quantification_mod[1-n]-site	(not used)	*sc*
assay[1-n]-sample_refs	(not used)	*sc*
cv[1-n]-label	*sc*	*sc*
cv[1-n]-full_name	*sc*	*sc*
cv[1-n]-version	*sc*	*sc*
cv[1-n]-url	*sc*	*sc*
colunit_protein	*sc*	*sc*
colunit_peptide	*sc*	*sc*
colunit_psm	*sc*	*sc*
colunit_small_molecule	*sc*	*sc*

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

Proteins in mzTab

Protein identifications are reported in the protein section. The protein section is table based. The table header is identified by the prefix “PRH”, entries in the protein table are identified through “PRT”. The protein section must only be present once. Columns are separated by a tab.

```
COM  Example of the protein section. Other sections are omitted.
PRH  accession  description  taxid  species  database  database_version         …
PRT  P12345     mAspAT       9986   Rabbit   UniProtKB 2013_08                   …
PRT  P02042     Hemoglobin   9606   Human    UniProtKB 2013_08                   …
```

The number of required columns in the protein table depends on the type of mzTab file (‘Identification’ / ‘Quantification’) and the used mode (‘complete’ / ‘summary’):

Field Name	Identification	Quantification
accession	**SC**	**SC**
description	**SC**	**SC**
taxid	**SC**	**SC**
species	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
search_engine	**SC**	**SC**
best_search_engine_score[1-n]	**SC**	**SC**
ambiguity_members	**SC**	**SC**
modifications	**SC**	**SC**

protein_coverage	*s***C**	*s***C**
protein_abundance_study_variable[1-n]	(not used)	**SC**
protein_abundance_stdev_study_variable[1-n]	(not used)	**SC**
protein_abundance_std_error_study_variable[1-n]	(not used)	**SC**
search_engine_score[1-n]_ms_run[1-n]	*s***C**	*s***C**
num_psms_ms_run[1-n]	*s***C**	*sc*
num_peptides_distinct_ms_run[1-n]	*s***C**	*sc*
num_peptide_unique_ms_run[1-n]	*s***C**	*sc*
protein_abundance_assay[1-n]	(not used)	*s***C**
opt_{identifier}_*	*sc*	*sc*
go_terms	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*

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

Peptides in mzTab

The peptide section is similar to the PSM section but used to report quantitative results aggregated on the peptide level. It should therefore not be used in 'Identification' files. It is table based and columns are separated by a tab. The header of the peptide table is indicated by "PEH", and entries in the table by "PEP". The peptide section must also be present only once.

```
PEH   sequence       accession    unique   database    database_version   search_engine
PEP   KLVILEGELER    IPI00010779  0        UniProtKB   2013_08     [MS,MS:1001207,Mascot,]
PEP   KQAEDRCK       IPI00513698  0        UniProtKB   2013_08     [MS,MS:1001207,Mascot,]
PEP   LATALQK        IPI00218319  1        UniProtKB   2013_08     [MS,MS:1001207,Mascot,]
PEP   LATALQKLEEAEK  IPI00218319  1        UniProtKB   2013_08     [MS,MS:1001207,Mascot,]
```

The number of required columns depends on the mzTab file's type and mode:

Field Name	Identification	Quantification
sequence	(not used)	**SC**
accession	(not used)	**SC**
unique	(not used)	**SC**
database	(not used)	**SC**
database_version	(not used)	**SC**
search_engine	(not used)	**SC**
best_search_engine_score[1-n]	(not used)	**SC**
modifications	(not used)	**SC**
retention_time	(not used)	**SC**
retention_time_window	(not used)	**SC**
charge	(not used)	**SC**
mass_to_charge	(not used)	**SC**
peptide_abundance_study_variable[1-n]	(not used)	**SC**
peptide_abundance_stdev_study_variable[1-n]	(not used)	**SC**
peptide_abundance_std_error_study_variable[1-n]	(not used)	**SC**
search_engine_score[1-n]_ms_run[1-n]	(not used)	*s***C**
peptide_abundance_assay[1-n]	(not used)	*s***C**
spectra_ref	(not used)	*s***C** (if MS2 based quantification is used)
opt_{identifier}_*	(not used)	*sc*
reliability	(not used)	*sc*
uri	(not used)	*sc*

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

PSMs in mzTab

The PSM section is used to report peptide identifications on a per spectrum level and is the recommended way to report peptides in 'Identification' files. It is similar to the protein section and also table based with the columns separated by a tab. If a peptide can be assigned to multiple proteins, this PSM MUST be reported multiple times (see PSM_ID 4 in the example below). The PSM section must also be present only once.

```
COM Example of the PSM section. Other sections and several columns are omitted.
PSH        sequence                PSM_ID    accession   unique    database    database_version…
PSM        QTQTFTTYSDNQPGVL        1         P63017      1         UniProtKB   2013_08            …
PSM        AVVNGYSASDTVGAGFAQAK    2         Q8K0U4      1         UniProtKB   2013_08            …
PSM        ALLRLHQECEKLK           3         Q61699      1         UniProtKB   2013_08            …
PSM        DWYPAHSR                4         P14602      0         UniProtKB   2013_08            …
PSM        DWYPAHSR                4         Q340U4      0         UniProtKB   2013_08            …
PSM        DWYPAHSR                4         P16627      0         UniProtKB   2013_08            …
PSM        MNQSNASPTLDGLFR         5         P14602      1         UniProtKB   2013_08            …
```

The required columns depend on the mzTab file's type and mode:

Field Name	Identification	Quantification
sequence	**SC**	**SC**
PSM_ID	**SC**	**SC**
accession	**SC**	**SC**
unique	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
search_engine	**SC**	**SC**
search_engine_score[1-n]	**SC**	**SC**
modifications	**SC**	**SC**
spectra_ref	**SC**	**SC**
retention_time	**SC**	**SC**
charge	**SC**	**SC**
exp_mass_to_charge	**SC**	**SC**
calc_mass_to_charge	**SC**	**SC**
pre	**SC**	**SC**
post	**SC**	**SC**
start	**SC**	**SC**
end	**SC**	**SC**
opt_{identifier}_*	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*

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

Small Molecules in mzTab

The small molecule section is also a table based section (same rules apply). Small molecules are identified through an "identifier" in mzTab. This identifier can be any text that sensibly identifies the given small molecule in the given field of research. These identifiers should generally be entries in compound databases used in the respective field (for example, Human Metabolome Database entries, ChEBI identifiers, PubChem IDs or LIPID MAPS IDs). Apart from this identifier, small molecules can be assigned a chemical formula, SMILES and/or InChi identifier, a human readable description, a *m/z* value, a charge state, retention time(s), a species, source database and search engine including score. We are aware, that these fields are not applicable to all fields of metabolomics, but we believe that they represent a sensible selection.

```
COM  Example of the small molecule section. Other sections are omitted. 'smiles' and
COM  'inchi_key' are not complete.

identifier     chemical_formula smiles       inchi_key      description    exp_mass_to_charge

CHEBI:17562    C9H13N3O5       Nc1ccn([C@@H]  UHDGCWIWMR…   Cytidine       244.0928
                               2O[C@H](CO)…
```

The required columns depend on the mzTab file's type and mode:

Field Name	Identification	Quantification
identifier	**SC**	**SC**
chemical_formula	**SC**	**SC**
smiles	**SC**	**SC**
inchi_key	**SC**	**SC**
description	**SC**	**SC**
exp_mass_to_charge	**SC**	**SC**
calc_mass_to_charge	**SC**	**SC**
charge	**SC**	**SC**
retention time	**SC**	**SC**
taxid	**SC**	**SC**
species	**SC**	**SC**
database	**SC**	**SC**
database_version	**SC**	**SC**
spectra_ref	**SC**	**SC**
search_engine	**SC**	**SC**
best_search_engine_score[1-n]	**SC**	**SC**
modifications	**SC**	**SC**
smallmolecule_abundance_assay[1-n]	(not used)	**SC** (if assays reported)
smallmolecule_abundance_study_variable[1-n]	(not used)	**SC** (if study vars. reported)
smallmolecule_stdev_study_variable[1-n]	(not used)	**SC** (if study vars. reported)
smallmolecule_std_error_study_variable[1-n]	(not used)	**SC** (if study vars. reported)
search_engine_score[1-n]_ms_run[1-n]	(not used)	s**C**
opt_{identifier}_*	*sc*	*sc*
reliability	*sc*	*sc*
uri	*sc*	*sc*

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

Missing values

In the table-based sections (protein, peptide, and small molecule) there MUST NOT be any empty cells. In case a given property is not available "null" MUST be used.

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

Reliability score

All protein, peptide, psm and small molecule identifications reported in an mzTab file can be assigned a reliability score (optional column "reliability" in all tables). The idea is to provide a way for researcher and/or MS proteomics or metabolomics repositories or data producers to score the reported identifications based on their own criteria. This score is completely resource-dependent and must not be seen 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.

The reliability is reported as 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 is 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. If resources now report their reliabilities using this metric and document how their 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, for example 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.

Quantitative Data

There are multiple quantification techniques available for MS-based experiments that often result in slightly different types of data. mzTab was not designed to capture any of these specific differences. The goal for mzTab was to provide a generic view on quantitative MS-based identification data that is applicable to as many different quantitation methods as possible. The method used in mzTab to model quantitative data is similar to the one used in mzQuantML and relies on "assays" and "study variables". "Assays" are used to report the actual measured values (*ie.* tag intensities) while "study variables" correspond to the final results from the study. A description of these items can be found above in "Modelling an experimental design in mzTab". Extensive example files on how to report different types of quantitation techniques can be found at https://code.google.com/p/mztab/wiki/ExampleFiles. See below an example corresponding to one SILAC experiment:

```
COM   Report of a minimal "Complete Quantification report" SILAC experiment, quantification on
COM   2 study variables (control/treatment), 3+3 assays (replicates) reported, no
COM   identifications reported. Internally 3 replicates/assays have been used to obtain
COM   quantification values, stdev and stderror
MTD   mzTab-version                       1.0.0
MTD   mzTab-mode                          Complete
MTD   mzTab-type                          Quantification
MTD   description                         mzTab example file for reporting a summary report
MTD   ms_run[1]-location                  file://C:\path\to\my\file1.mzML
MTD   ms_run[2]-location                  file://C:\path\to\my\file2.mzML
MTD   ms_run[3]-location                  file://C:\path\to\my\file3.mzML
MTD   ms_run[4]-location                  file://C:\path\to\my\file4.mzML
MTD   protein-quantification_unit         [PRIDE, PRIDE:0000393, Relative quantification unit,]
MTD   software[1]                         [MS, MS:1001583, MaxQuant,]
MTD   psm_search_engine_score[1]             [MS, MS:1001979, MaxQuant:PTM score,]
MTD   fixed_mod[1]                        [UNIMOD, UNIMOD:4, Carbamidomethyl, ]
MTD   fixed_mod[2]                        [UNIMOD, UNIMOD:188, Label:13C(6), ]
MTD   variable_mod[1]                     [UNIMOD, UNIMOD:35, Oxidation, ]
MTD   quantification_method               [MS, MS:1001835, SILAC, ]
MTD   assay[1]-quantification_reagent     [PRIDE, PRIDE:0000326, SILAC light, ]
MTD   assay[2]-quantification_reagent     [PRIDE, PRIDE:0000325, SILAC heavy, ]
MTD   assay[3]-quantification_reagent     [PRIDE, PRIDE:0000326, SILAC light, ]
MTD   assay[4]-quantification_reagent     [PRIDE, PRIDE:0000325, SILAC heavy, ]
MTD   assay[1]-ms_run_ref                 ms_run[1]
MTD   assay[2]-ms_run_ref                 ms_run[1]
MTD   assay[3]-ms_run_ref                 ms_run[2]
MTD   assay[4]-ms_run_ref                 ms_run[2]
MTD   study_variable[1]-assay_refs        assay[1],assay[3]
MTD   study_variable[2]-assay_refs        assay[2],assay[4]
MTD   study_variable[1]-description       heat shock response of control
MTD   study_variable[2]-description       heat shock response of treatment
```

Protein Inference

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

```
      COM   In the following example only one peptide was identified that can be attributed to
      COM   multiple proteins. The choice which one to pick as primary accession depends on
      COM   the resource generating the mzTab file.
      ...
      PRH   accession  …   ambiguity_members        …
      PRT   P14602     …   Q340U4, P16627           …
      ...
      PSH      sequence    PSM_ID    accession      unique …
      PSM      DWYPAHSR    4         P14602         0      …
      PSM      DWYPAHSR    4         Q340U4         0      …
      PSM      DWYPAHSR    4         P16627         0      …
```

Advanced topics

There are several other features in mzTab that could not be introduced here. Detailed information about these features can be found in the specification document such as:

- Reporting post-translational modifications (PTMs) including modification position ambiguity.
- Reporting results from multiple search engines.
- Referencing external spectra.
- Referencing external resources such as mzIdentML or mzQuantML files.
- Adding optional columns.
- Specifying a column's unit.

An up-to-date list of example files can be found at http://code.google.com/p/mztab/wiki/ExampleFiles. The specification document can be found at http://code.google.com/p/mztab/.