Practical ten minutes guide for requesting new CV terms in the PSI-MS CV

Gerhard Mayer[1], Juan Antonio Vizcaíno[2], Daniel Schober[3]
[1] Medizinisches Proteom-Center, Ruhr-Universität Bochum, Universitätsstraße 150, D-44801 Bochum, Germany
[2] EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[3] Department of Stress- and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany

Abbreviations:

CV	Controlled Vocabulary
HUPO	Human Proteome Organization
MS	Mass Spectrometry
OBO	Open Biomedical Ontologies
OLS	Ontology Lookup Service
PSI	Proteomics Standards Initiative

Scope of this document:

This document is intended as an introductory guide for people who want to convert their proteomics identification and / or quantification results into the standard formats mzIdentML [1], mzQuantML [2] and/or mzTab (mztab.googlecode.com) of the HUPO-PSI organization. It is written especially for people who are total novices in using Controlled Vocabulary (CV) terms [3] for semantic annotation of data in the HUPO-PSI formats. Often they do not know if the terms they need are already present in the CV, or if they should request new CV terms for their needs. For such cases the following can be used as a starting point for those people who want to check if they need new terms and about how to request them. This document serves as a guideline for explaining the CV term request procedure via a practical example use case scenario.

1. Hypothetical problem description:

Assume you want to check if you need to request new Controlled Vocabulary (CV) terms for the search engine named ‘MassMatrix’ (MM).

Example MS/MS search results from MM:
http://www.massmatrix.net/massmatrix/mm-results/32846/RPMI_RE8_01_246.html

Peptide terms:

Index	scan#	charge	score	pp	pp_2	pp_{tag}	m/z	MW(obs)
MW	delta	Miss	Unique	sequence + modifications [start:end], e.g.				
1153	3659	+2	12	10.9	14.4	2.7	614.8204	1228.6336
1228.628	0.0056	0	x	VEIIANDQGNR				

So you have the following data describing a peptide search result:
index (arbitrary search engine created index)
scan# (MS scan #)
charge (predicted charge)
score (discriminant score based on heuristic model)
pp (probabilistic score based on number of matched peaks)
pp2 (probabilistic score based on ion intensity distribution of matched peaks)
pptag (probabilistic score based on cosecutiveness of matched peaks)
m/z (observed peptide m/z)
MW(obs) (observed zero charge peptide MW)
MW (zero charge theoretical peptide MW)
delta (MW difference between theo and observed)
miss (missed cleavages)
Unique (unique sequence in search space)
sequence + modifications (peptide spectral match + modifications and their locations)

Protein terms:

Protein Mass:	72288.436 (monoisotopic) 72332.412(average)
Protein Score:	408
Protein pp:	13872.0
Fasta Line:	sp|P11021|GRP78_HUMAN 78 kDa glucose-regulated protein OS=Homo sapiens GN=HSPA5 PE=1 SV=2
Sequence Coverage:	27%
Sequence Tag Coverage:	12%

Protein Mass
Protein Score (heuristic discriminant protein score)
Protein pp (probabilistic discriminant protein score)
Sequence coverage (based on peptide match)
Sequence tag coverage (based on amino acids bracketed by product ions)

MM search machine input parameters:

version:	MassMatrix 2.4.2, Feb 22 2012
Tandem MS/MS data file:	RPMI_RE8_01_246.mgf
Database:	Uniprot_Human_Complete_with_Isoforms_9_23_2011.fasta
Decoy sequences:	reversed
Digestion:	Trypsin(no P rule)
Fragmentation:	CID
Ion Mode:	Positive
Non-monoisotopic ions:	yes
Modifications:	CAMC: Iodoacetamide derivative (Carbamidomethyl) of C
Fixed Modifications:	none
Maximum # Missed Cleavages:	3
Maximum Length of Peptides:	40
Minimum Length of Peptides:	6
Peptide Mass Tolerance:	±20.00 ppm
Fragment Mass Tolerance:	±0.10 Da(CID)
Mass:	monoisotopic
Minimum Score of Output:	10 (CID)
Minimum pp Value of Output:	5.0
Minimum pp2Value of Output:	5.0
Minimum PP_{tag} of output:	1.3
Minimum CLpp of Output:	0.0
Minimum CLpp2 of Output:	0.0
Minimum protein score:	5.0
Max # PTM per peptide:	2
Maximum # of matches/Spec :	1
Maximum # of combs/peptide:	1
Cross linkage search:	Disabled
Total # of MS/MS spectra:	5704
Protein sequences checked:	144780
Peptide sequences checked:	7702451
Peptides checked:	1.275947e+07
R^2 of LR model for t_R vs H:	N/A or failed
MS/MS tag quantitation:	disabled
Wall clock time:	0hr 2min 18sec
Date and time:	Thu Feb 28 13:10:05 2013

2. How to know which CV terms you should use / you need for converting your search results into mzIdentML [1]?

Now assume that you want to convert the search results into a PSI standard format, like for instance mzIdentML. You first have to determine which CV terms are already available in psi-ms.obo [4]. Only if you need terms that are not covered yet by appropriate existing obo CVs, you need to request these terms anew.

When integrating new terms (e.g. for peptide scores, protein scores and input parameters) needed for the CV-based description of the new search engine, we have to ensure that no redundant terms are created in addition to already existing descriptors. We want to make sure that terms with the same meaning are not defined again and again for each search engine. Hence one should first check, for which of the scores resp. input parameters there are already fitting terms defined in the existing CVs, resp. if the terms really have to be created anew. This can be quite laborious, especially for those not well acquainted with the PSI-MS ontology. We therefore suggest, first to get familiar with the structure of the PSI-MS ontology, e.g. by consulting the publication [3]. Then, one must check for each candidate term, if it is already present in the CV, resp. if it is needed, or not. Here the Ontology Lookup Service (OLS) web site and the OBO-Edit tool can help for browsing the ontology content. In the following we outline the general term request scenario and exemplify it for some of the scores and input parameters mentioned above.

For instance for the **peptide term** '`charge (predicted charge)`' one must use the already existing term 'MS:1000041 (charge state)' and state the charge in its value slot.

Other examples for such matches would be:

'm/z (observed peptide m/z)'	-->	MS:1000040 (m/z)
Sequence Coverage	-->	MS:1001093 (sequence coverage)

Note also that for properties, for which we have an attribute in the corresponding XML schema (.xsd file), e.g. mzIdentML, we should not define a new CV term, but use the suitable attribute instead. For example, for your two peptide terms 'MW(obs) (observed zero charge peptide MW)' and 'MW (zero charge theoretical peptide MW)' one must use the two attributes 'experimentalMassToCharge' resp. 'calculatedMassToCharge' of the 'SpectrumIdentificationItem' element of mzIdentML, e.g.

```
<SpectrumIdentificationItem id="SEQ_spec1_pep1" peptide_ref="prot1_pep1" chargeState="1"
calculatedMassToCharge="1507.6950" experimentalMassToCharge="1507.696" passThreshold="true"
rank="1">
```

For 'delta (MW difference between theory and observed)' one can use the term 'MS:1000904 (product ion m/z delta)' if one takes the charge into consideration.

For the **input parameter terms** there are also some terms already contained in the CV, e.g. for specifying the 'Database' one can use the terms '`MS:1001012 (database source)`' and '`MS:1001013 (database name)`', whereas for the database version one must use the version attribute of the 'SearchDatabase' XML element in mzIdentML instead of the obsoleted term '`MS:1001016 (database version)`', e.g.

```
<SearchDatabase id="ipi.HUMAN_decoy"
location="file://C://DBServer/ipi.HUMAN/3.15/ipi.HUMAN_decoy.fasta" version="3.15"
releaseDate="2006-02-22T09:30:47Z" numDatabaseSequences="58099">
```

In general one should avoid using obsoleted terms or terms from the purgatory branch, which contains terms we are planning to obsolete soon.

Other matches for the search input parameters would be:

- For specifying 'Fragment mass tolerance' resp. 'Peptide mass tolerance' one must use the terms 'MS:1001412 (search tolerance plus value)' and 'MS:1001413 (search tolerance minus value)'.
- For 'Ion Mode' one can use the 'MS:1000465 (scan polarity)' term.
- The 'Fragmentation' matches to the 'MS:1000008 (ionization type)' term.
- For the 'Digestion' input parameter one can use the 'MS:1001251 (Trypsin)' or another cleavage agent name.
- For specifying the date and time one can use the term 'MS:1000747 (completion time)'.
- For specifying the 'Decoy Sequences' one must can the terms 'MS:1001194 (quality estimation with decoy database)', 'MS:1001195 (decoy DB type reverse)' and 'MS:1001196 (decoy DB type randomized)'.
- For specifying the 'Modifications' one must use the CV terms from either the PSI-MOD or Unimod ontologies [4].

I hope these hints can give you some idea on how to proceed in the usage of psi-ms.obo for mzIdentML. If one needs new specific terms to annotate / describe a search engine, then post the required terms together with a short definition and example data on our psidev-vocab mailing list. Make sure that for every requested term you provide a clear and meaningful description in their def tags (def: "... ." [PSI:MS]). Note also, that for each term belonging to a score one should also provide an ordering information, i.e. either

```
    has_order: MS:1002108 ! higher score better
or
    has_order: MS:1002109 ! lower score better
```

Another rule to follow is that one should avoid acronyms and abbreviations in term names, see the naming conventions of the OBO Foundry (http://www.obofoundry.org/wiki/index.php/Naming).

Among others the following terms are **peptide / protein scores** *specific* for the new search engine:
pp (probabilistic score based on number of matched peaks)
pp2 (probabilistic score based on ion intensity distribution of matched peaks)
pptag (probabilistic score based on cosecutiveness of matched peaks)

For these one can request to include them into the PSI-MS CV, e.g. as

```
    [Term]
    id: MS:10023xy
    name: MassMatrix : Peptide Probility
    def: "Probabilistic score based on number of matched peaks." [PSI:PI]
    xref: value-type:xsd\:double "The allowed value-type for this CV term."
    is_a: MS:1001143 ! search engine specific score for PSMs
    is_a: MS:1001153 ! search engine specific score
    has_order: MS:100210x ! ... score better resp.
```

Same for some of the **input parameters**, e.g.
Minimum CLpp of Output
Minimum CLpp2 of Output
... and maybe some other terms

One can request to include them into the PSI-MS CV as something like

```
    [Term]
    id: MS:10023zz
    name: MassMatrix input parameter
    def: "Search engine input parameters specific toMassMatrix." [PSI:PI]
    is_a: MS:1001302 ! search engine specific input parameter

    [Term]
    id: MS:10023xy
    name: MassMatrix:Minimum CLpp
    def: "... (Provide a meaningful definition here)." [PSI:MS]
    xref: value-type:xsd\:float "The allowed value-type for this CV term."
    is_a: MS:10023zz ! MassMatrix input parameter
```

... and so on.

So one should always first check all the scores / input parameters in detail: whether there is already an existing term in psi-ms.obo or an appropriate attribute in mzIdentML, or if a new term should be included in the PSI-MS CV. In the latter case collect all the terms you need and send them for discussion /review to the psidev-vocab mailing list.

References:

[1] A.R. Jones, M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S.J. Hubbard, J.N. Selley, B.C. Searle, J. Shofstahl, S.L. Seymour, R. Julian, P.A. Binz, E.W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J.A. Vizcaíno, M. Chambers, A. Pizarro, D. Creasy, **The mzIdentML data standard for mass spectrometry-based proteomics results**, Mol Cell Proteomics, 11 (2012) M111 014381.

[2] M. Walzer, D. Qi, G. Mayer, J. Uszkoreit, M. Eisenacher, T. Sachsenberg, E.W. Deutsch, F. Reisinger, J.A.Vizcaíno, J. A. Medina-Aunon, J.P. Albar, O. Kohlbacher, A.R. Jones, **The mzQuantML data standard for quantitative studies in proteomics**, Molecular & Cellular Proteomics, 2013 Apr 18. [Epub ahead of print].

[3] Mayer G., Jones A.R., Binz P.-A., Deutsch E.W., Orchard S., Montecchi-Palazzi L., Vizcaíno J.A., Hermjakob H., Ovelleiro D., Julian R., Stephan C., Meyer H.E., Eisenacher M. **Controlled Vocabularies and**

Ontologies in Proteomics: Overview, Principles and Practice, Biochim. Biophys. Acta (2013), doi:10.1016/j.bbapap.2013.02.017 [PMID: 23429179].

[4] Mayer G., Montecchi-Palazzi L., Ovelleiro D., Jones A.R., Binz P.-A., Deutsch E.W., Chambers M., Kallhardt M., Levander F., Shofstahl J., Orchard S., Vizcaíno J.A., Hermjakob H., Stephan C., Meyer H.E., Eisenacher M. **The HUPO Proteomics Standards Initiative – Mass Spectrometry Controlled Vocabulary**, Database (2013), doi:10.1093/database/bat009 [PMID: 23482073].

Web links:

- Controlled Vocabularies
- mzIdentML standard format for reporting proteomics identification results
- mzQuantML standard format for reporting proteomics quantification results
- Naming conventions of the OBO Foundry
- OBO-Edit Ontology editor
- OLS – Ontology Lookup Service of the EBI for browsing the ontologies
- psidev-vocab mailing list
- psi-ms.obo file