# **Changes on the semantic validation from mzIdentML 1.1 to 1.2**

This document summarizes the changes performed on the rules to apply for the semantic validation of the mzIdentML that follow the new mzIdentML specification 1.2.

See new files at:

|  |  |  |
| --- | --- | --- |
| file | description | location |
| mzIdentML-mapping\_1.2.0.xml | cvMapping rules to apply | <https://psi-pi.googlecode.com/svn/trunk/validator/trunk/src/main/resources/mzIdentML-mapping_1.2.0.xml> |
| ObjectRules.1.2.0.xml | Objects rules to apply | <https://psi-pi.googlecode.com/svn/trunk/validator/trunk/src/main/resources/ObjectRules.1.2.0.xml> |
| ruleFilter\_semantic.xml | Conditional application of rules | <https://psi-pi.googlecode.com/svn/trunk/validator/trunk/src/main/resources/ruleFilter_semantic.xml> |

**1.** As in mzIdentML version 1.1, a single protein accession that has been cited by software (Figure 1A) is captured in mzIdentML in <ProteinDetectionHypothesis> (PDH).

a. A PDH MAY contain scores or statistical values produced by the export software, encoded as CV terms.

No modification required for MAY rule *ProteinAmbiguityGroupProteinDetectionHypothesis\_rule*

**2.** A “protein group” (Figure 1B), representing a “biological entity” for which the software claims independent evidence is present, MUST be mapped onto <ProteinAmbiguityGroup> (PAG).

a. A PAG MAY have additional scores produced by the export software, encoded as CV terms.

No modification required for MAY rule *ProteinAmbiguityGroup\_rule*

**3.** The reporting of protein identification thresholds is now mapped onto PAGs. There is no desire to change the core XML Schema Document (XSD) for mzIdentML and as such, a new CV term “protein group passes threshold” value= “xsd:boolean” MUST be present on every PAG (MS:1002415). If no thresholding has been done by the software, all protein groups MUST be annotated as “protein group passes threshold” value= “true”.

a. The attribute *passThreshold* = “true|false” remains present on PDH and MAY be used if software packages wish to report a two-level hierarchy of thresholds applied, however, it is not expected that consuming software will use this attribute to determine which proteins have been reported as identified.

"protein group passes threshold" in PAG: MUST rule *ProteinAmbiguityGroup\_must\_rule*

**4.** The <ProteinDetectionList> MUST contain the CV term “count of identified proteins” value=“xsd:integer” (MS:1002404). The value MUST be derived from the count of PAGs passing the threshold reported in the file and will be checked by validation software.

"count of identified proteins" in PDL: MUST rule *ProteinDetectionList\_must\_rule*

**5.** Few software packages report “protein clusters” at present (Figure 1C), but for those packages that wish to report clusters, a CV term “cluster identifier” value = “xsd:integer” SHOULD be used (MS:1002407). The integer identifier MUST be shared by all PAGs belonging to the same cluster.

Object rule: *ProteinDetectionListObjectRule*: Iterates over all PAGs and if one of them contains a "cluster identifier" CV, all the rest will also MUST have it. Check also that the value of the CV MUST be an integer.

An optional term “count of identified clusters” value = “xsd:integer” (MS:1002406) MAY be annotated on the <ProteinDetectionList>.

"count of identified clusters" in PDL: Added in MAY rule *ProteinDetectionList\_rule* (this rule already existed before. cvTermCombinationLogic changed to AND)

**6.** Every PDH MUST be annotated as either a “leading protein” (MS:1002401) or a “non-leading protein” (MS:1002402), as defined in Table 1, within a PAG. This recommendation thus makes it explicit for consuming software whether one or more proteins have stronger evidence than others in the group (see Table 2 for examples).

“leading protein” xor “non-leading protein” in PDH: MUST rule *ProteinAmbiguityGroupProteinDetectionHypothesis\_must\_xor\_rule*

**a.** An additional term, “group representative” (MS:1002403) MAY be used to annotate one PDH, which is also flagged as a “leading protein”, if the export software wishes to enforce that only one of potential several “leading proteins” will be interpreted by the consuming software as the representative of the group, for example acting as a tiebreaker.

**b.** If the export software does not explicitly flag one protein as the “group representative”, it is assumed that if consuming software requires a single accession to represent the group, an arbitrary choice will be made (among “leading proteins” only if these exist).

Object rule: *ProteinAmbiguityGroupObjectRule*:

- only one PDH in a PAG can be labeled as "group representative" (MS:1002403), and if found, the PDH must be labeled as a "leading protein" (MS:1002401), otherwise, a level-ERROR message will be reported.

- if all the PDHs labeled as "leading protein" in a PAG do not contain "group representative", a level-INFO message will be reported.

- if there is not any PDH in a PAG with "group representative" term, reports a level-INFO message

**7.** Any PDHs MAY be annotated with terms present in the CV for spectrum/sequence same-set, spectrum/sequence subset, spectrum/sequence subsumable, marginally distinguished and so on (Table 1).

**a.** A PDH MAY be annotated with more than one of these terms if appropriate to describe the complex set relationships that exist within a group.

**b.** Developers of software packages MAY propose additional terms for describing group membership of PDHs, which will be incorporated into the CV.

**c.** The associated value for these CV terms MAY be used to annotate which PDH(s) are the super/same-set of the annotated PDH.

**d.** There is no expectation that consuming software should be aware of these terms, but they may be useful in internal pipeline or visualization software packages that are specifically designed to work with this terminology set.

“protein group or subset relationship” in PDH: MAY rule *ProteinAmbiguityGroupProteinDetectionHypothesis\_rule\_may\_and*

**8.** Some PDHs could be mapped to more than one PAG, for example where proteins are multiply subsumed. To capture these cases, multiple PDHs in different PAGs MAY reference the same <DBSequence>.

No rule to be added

Additional validations added:

* ”spectrum identification list result details” in SIL: SHOULD rule *SpectrumIdentificationList\_PSMListType\_rule*
* Support for de novo data: Since PeptideEvidenceRef element on SII now is 0..n (before was 1..n):
  + - new cv rule “*DenovoSearchType\_rule*” that check if the term "de novo search" is present on APC/SIP/searchType. The result of this rule will be used for the application or not of these other following rules
  + - new object rule “*SpectrumIdentificationItemPeptideEvdenceRefObjectRule*” that reports an error if PeptideEvidenceRef element is not present on SII.
  + - new object rule “*SpectrumIdentificationItemNullPeptideEvdenceRefObjectRule*” that reports an error if PeptideEvidenceRef element is present on SII.
  + - to create a condition (until now, that mechanism was only used for MIAPE validation) that says: "if that cv rule fails, that is, it is not a de novo search, discard the errors from the *SpectrumIdentificationItemNullPeptideEvdenceRefObjectRule*. If that cv rules doesn’t fail, that is, it is a de novo search, discard the errors from the *SpectrumIdentificationItemPeptideEvdenceRefObjectRule*.
    - New file *ruleFilter\_semantic.xml* containing one rule:
      * <ruleConditions>
      * <ruleCondition valid="true" id="DenovoSearchType\_rule">
      * <ruleToSkip id="SpectrumIdentificationItemPeptideEvidenceRefObjectRule"/>
      * </ruleCondition>
      * <ruleCondition valid="false" id="DenovoSearchType\_rule">
      * <ruleToSkip id="SpectrumIdentificationItemNullPeptideEvidenceRefObjectRule"/>
      * <ruleToSkip id="DenovoSearchType\_rule"/> <!-- discard its error messages -->
      * </ruleCondition>
      * </ruleConditions>

Some minor modification over mzid1.1 validation rules:

* Allow just one child term of ”search type” in “APC/SIP/searchType”. So, isRepeatable changed to false in SearchType\_rule

**9.** If the term **modification localization scoring** (MS:1002491) is present under

<SearchType> in SpectrumIdentificationProtocol, then the following term MUST be

present under SpectrumIdentificationItem (SII):

- modification rescoring:false localization rate (MS: 1002507) term itself with repetitions

**10.** If the term **peptide-level scoring** (MS:1002490) is present under <SearchType> in

SpectrumIdentificationProtocol, then the following triplet of terms MUST be

present under SpectrumIdentificationItem (SII):

- peptide group ID (MS: 1002520) term itself

- peptide-level statistical threshold (MS:1002484) childs thereof

- peptide passes threshold (MS:1002500) term itself

**11.** In **SIIModLocalizationScoringRule ("MS:1002507": modification rescoring:false localization rate)**: checks if in case of **modification localization scoring** the value for the cvParam 'modification rescoring:false localization rate' under the SpectrumIdentificationItem must have the format:

**MOD\_ORDER:SCORE:POSITION:PASS\_THRESHOLD** at with

**MOD\_ORDER** = <Modification> element order in the referenced <Peptide> object

**SCORE** = Score or statistical measure associated with the modification position

**POSITION** = Position of the modification on the peptide (N-terminus = 0, C-terminus = peptide length + 1).

If the score pertains to grouped positions, different positions MUST be separated by ‘|’

**PASS\_THRESHOLD** = true | false with regards to the threshold specified in Feature A. If no Threshold has

been specified, this MUST always be true.

e.g. '1:0.03:2|3:true' or '1:0.97:8|9:false' according to the regular expression

^\\d+:[0|1]{1}\\.\\d+:\\d+[|]{1}[\\d]+:(true|false){1}$

**12.** In **SILCheckForFinalPSMListRule ("MS:1002439": final PSM list) together with SIRUniqueSpectrumIDSpectrumRefCombinationRule**: Uniqueness check for combination of spectrumID and spectrumRef within "final PSM lists"

**13.** Suppresses repeated validation errors if they appear more than a few times (exact number is read from jSpinner; default value. 3)

**14.** Checks in **AdditionalSearchParamsRule** if one of the ‘special processing’ term is present in the AdditionalSearchParams:

* "MS:1001010": // de novo search
* "MS:1002490": // peptide-level scoring
* "MS:1002491": // modification localization scoring
* "MS:1002492": // consensus scoring
* "MS:1002493": // sample pre-fractionation
* "MS:1002494": // cross-linking search
* "MS:1002495": // no special processing

**15.** Other new CV mapping rules for mzIdentML 1.2:

SpectrumIdentificationList\_PSMListType\_rule MS:1002438 (spectrum identification list result details)

SHOULD be there.

SpectrumIdentificationList\_PSMListType\_rule MS:1001184 (search statistics) and

MS:1002406 (count of identified clusters) MAY be there

ProteinDetectionList\_must\_rule MS:1002404 (count of identified proteins) MUST be there

ProteinAmbiguityGroupProteinDetectionHypothesis\_rule\_may\_and

MS:1001101 (protein group or subset relationship) MAY be

there

ProteinAmbiguityGroupProteinDetectionHypothesis\_must\_xor\_rule

Either MS:1002401 (leading protein) or

MS:1002402 (non-leading protein) MUST be there

ProteinAmbiguityGroup\_must\_rule MS:1002415 (protein group passes threshold) MUST be there

PeptideLevelStatsSpectrumIdentificationItem\_must\_rule

The three CV terms MS:1002520 (peptide group ID),

MS:1002484 (peptide-level statistical threshold) and

MS:1002500 (peptide passes threshold) MUST be there

(DenovoSearchType\_rule MS:1001010 (de novo search) MAY be there) \*

PeptideLevelStatsSearchType\_rule MS:1002490 (peptide-level scoring) MAY be there

ModLocalizationSearchType\_rule MS:1002491 (modification localization scoring) MAY be there

ModLocalizationSpectrumIdentificationItem\_must\_rule

MS:1002507 (modification rescoring:false localization rate)

must be there