PSI final recommendation

PSI Proteomics Informatics Working Group
Angel Pizarro, Center for Bioinformatics, University of Pennsylvania
David Creasy, MatrixScience
Phil Jones, European Bioinformatics Institute
Andreas Bertsch, Eberhard Karls University Tübingen
Jenny Siepen, University of Manchester
Martin Eisenacher, Medizinisches Proteom-Center, Ruhr-Universität Bochum
Andrew Jones, University of Liverpool

August 2009

mzIdentML: exchange format for peptides and proteins identified from mass spectra

Status of This Document

This document presents a final specification for the mzIdentML data format developed by the HUPO Proteomics Standards Initiative. Distribution is unlimited.

Version of This Document
The current version of this document is: version 1.0.0 August 2009.

Abstract

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. The Proteomics Informatics Working Group is developing standards for describing the results of identification and quantitation processes for proteins, peptides and protein modifications from mass spectrometry. This document defines an XML schema that can be used to describe the outputs of proteomics search engines.

Contents

1. Introduction

1.1 Background

This document addresses the systematic description of polypeptide identification and characterisation based upon mass spectrometry. A large number of different proteomics search engines are available that produce output in a variety of different formats. It is intended that mzIdentML will provide a common format for the export of identification results from any search engine. The format was originally developed under the name AnalysisXML as a format for several types of computational analyses performed over mass spectra in the proteomics context. It has been decided to split development into two formats: mzIdentML for peptide and protein identification (described here) and mzQuantML (to be described in a future specification document).

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

T1. *The discovery of relevant results,* so that, for example, data sets in a database that use a particular technique or combination of techniques can be identified and studied by experimentalists during experiment design or data analysis.
T2. *The sharing of best practice*, so that, for example, analyses that have been particularly successful at identifying a certain group of peptides/proteins can be interpreted by consumers of the data.
T3. *The evaluation of results*, so that, for example, sufficient information is provided about how a particular analysis was performed to allow the results to be critically evaluated.
T4. *The sharing of data sets,* so that, for example, public repositories can import or export data, or multi-site projects can share results to support integrated analysis.
T5. *The creation of a format for input to analysis software*, for example, allowing software to be designed that provides a meta-score over the output from several search engines.

The description of the analysis of proteomics mass spectra requires that models describe: (i) the identity and configuration of software used to perform the analysis and the protocol used to apply this software to the analysis; (ii) the identity of molecules; and (iii) the way in which these relate to other techniques to form a proteomics workflow. Most of this document is concerned with (i) and (ii) – the identification of the key features of different techniques that are required to support the tasks T1 to T5 above. Models of type (iii) are created by developments in the context of the Functional Genomics Experimental Object Model (FuGE), which defines model components of relevance to a wide range of experimental techniques. Several components from FuGE are re-used in the development of mzIdentML.

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 when the specification is stable.

1.2 Document Structure

The remainder of this document is structured as follows. Section 2 lists use cases for which mzIdentML is created 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 for the XML schema which is generated automatically and several parts of the schema are documented in more detail in Section 6.1. Conclusions are presented in Section 8.

2. Use Cases for mzIdentML

The following use cases have driven the development of the mzIdentML data model and XML schema, and are used to define the scope of the format in version 1.

1. It should be possible to create a tool that loads an mzIdentML document and enables users to examine results from an MS, MS-MS or MSn. (For MSn searches, the assumption is that matches will be of a similar format to those from MS-MS searches and there will be no attempt to model combining, say MS4 matches with the corresponding MS3 and MS-MS results). There should be sufficient information for the tool to generate output reports that conform to the requirements made by journals for publication and that conform to the relevant MIAPE guidelines. For example:
   · For a Peptide Mass Fingerprint (PMF) search, it should be possible to display the spectrum and show the matches of the peaks to the relevant peptides, but only if the spectrum is available.
   · For an MS-MS search, it should be possible to locate which spectrum matched to which peptide in the original file.
2. There should be sufficient information stored in the instance document to enable a user to run the same search on the same or another search engine. This means that all search parameters should be described in sufficient detail and that sufficient information is available to determine which database (if any) the data were searched against. The peak lists data (if any) do not need to be included in the instance document, but do need to be suitably referenced.
3. A PMF search and an MS-MS search of the same sample can be saved in the same instance document as long as the result is one combined protein list.
4. It should be possible to save the results of searching a decoy database in the same instance document as the results from the target database. It should then be possible to write a viewer application that enables a user to investigate the effect of changing, for example, a threshold value on the false discovery rate. This would only be possible if results that are generally considered lower quality from the search are also saved in the mzIdentML document (rather than just top matches) and if the results from the decoy search are also saved. It would only be possible to do this at the peptide level for an MS-MS search, because changing thresholds would normally have some effect on the protein grouping algorithm.
5. It should be possible to save manual or automated annotation of proteins/peptides in an instance document. A third party tool could be used to save annotations and validations of identified proteins/peptides to an existing instance document.
6. It should be possible to save the results from a search of a metabolically labelled sample. For example, with a 14N/15N experiment, two separate sets of amino acid masses are used, and it must be possible to tell which masses were used for each peptide result.
7. For a search of multiple peak lists, it should be possible to identify the spectrum that obtained a match to a particular peptide or protein reported by the search engine. For example, in an LC-MS-MS run, it should be possible to refer back to the spectrum in the peak list file that was searched and from there, if the information is available, to be able to determine the retention time of the spectrum. For an mzML file, the unique 'id' of the spectrum should be available. For other peak list formats, some other unique identifier should be stored where possible. There is no requirement to store other redundant information in the mzIdentML file that will be available in the peak list data.
8. It should be possible to search a file to retrieve all molecules that have a specified modification.

9. It should be possible to store the results of a search of spectra against other spectra - i.e. a spectral library search.
10. It should be possible to store the results of a top down search i.e. analysis of complete proteins.
11. Support should be provided for storing fragmentation data so that for example viewers could display which ions in the input data match predicted ion fragment masses.
12. There should be support for storing the results of searches of peptides against nucleic acid databases, including the information about which translation frame the matches were found in.
13. It should be possible to combine the results from multiple search engines into one mzIdentML document. For example, the peptide identification results from two different search engines could be combined using a third tool to give one set of protein results.

There will be limited support for the following use cases:

1. *De novo*. *De novo* peptide sequencing results will be supported to the extent that it will be possible to enumerate and record all possible matches found by a *de novo* technique, however, we anticipate that this will produce extremely large files. In later versions of mzIdentML, solutions will be investigated for defining a standard way of reporting ambiguous combinations of residues and we invite proposals in this area.

The following use cases will not be supported in version 1 of mzIdentML:

1. It should be possible to store relative and absolute quantitation information at the peptide and protein level using all the popular techniques [to be developed in a separate format called mzQuantML].
2. Support for LC-MS biomarker discovery.
3. Support for complex workflows where multiple data processing algorithms are combined in a pipeline; i.e. only “final” results are represented in mzIdentML v1, no intermediate results.
4. Support for tag searches, in which short sequences defined by a *de novo* process are used to pre-filter a sequence database prior to a complete search.

3. Concepts and Terminology

This document assumes familiarity with XML Schema notation (www.w3.org/XML/Schema). 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 [RFC2119].

4. Relationship to Other Specifications

The specification described in this document is not being 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. *MIAPE MSI* (http://www.psidev.info/miape/msi/). The Minimum Information About a Proteomics Experiment: Mass spectrometry Informatics document defines a checklist of information that should be reported about such a study. It is expected that mzIdentML will be used to support MIAPE:MSI compliant submissions to public repositories (as demonstrated in mzIdentML_MIAPE.doc).
2. *FuGE* (http://fuge.sourceforge.net). FuGE is a data model in UML, and an associated XML rendering, that represents various high-level concepts that are characteristic of functional genomics, such as investigations and protocols. FuGE has been developed by representatives of several standards bodies, with a view to making the representation of functional genomic data sets more consistent, and as such more easily shared and compared. The FuGE specifications are available from [Jones 07].

3. *mzML* (http://www.psidev.info/index.php?q=node/80). mzML is the PSI standard for capturing mass spectra / peak lists resulting from mass spectrometry in proteomics. It is RECOMMENDED that mzIdentML should be used in conjunction with mzML, although it will be possible to use mzIdentML with other formats of mass spectra. This document does not assume familiarity with mzML.

4.1 Important concepts from FuGE

mzIdentML makes use of several components from FuGE to allow the format to be more easily integrated with other FuGE-based formats. However, FuGE is a large, flexible specification that can cover a variety of concepts not required for mzIdentML. As such, it was decided to remove a number of elements from the FuGE XSD to make the format as simple as possible to implement. The components of the FuGE model used by mzIdentML are described briefly here. Furthermore, some minor changes have been made to FuGE, such that it uses the same conventions as mzML, since close compatibility with mzML was deemed a high priority. In this context, the altered FuGE schema has been renamed “FuGE-light”.

In the mzIdentML schema the following elements from FuGE have been extended or used without extension:

- <*Identifiable*> - Elements in mzIdentML that are referenced elsewhere in the file are subclasses of FuGE <Identifiable>, which gives the element a mandatory attribute to store a unique identifier, and an optional attribute to store a human readable name. A change is made to FuGE in that the attribute has been changed from “identifier” to “id” to match mzML.
- <*Protocol*> – Instances of <Protocol> represent a description of, for example, standard operating procedures or data processing instructions. In mzIdentML, extensions have been created to model analyses and the associated sets of parameters used in a data analysis routine (for peptide and protein identification).
- <*ProtocolApplication*> represents the running of a <Protocol>, mapping the input and output data sets, and thus allowing different processes to be tied together through a chain of inputs and outputs. In mzIdentML version 1.0, there is minimal flexibility for tying together complex identification workflows, however <ProtocolApplication> has been incorporated into the schema so that in future versions, such flexibility may be supported.
- <*Material*> from FuGE represents materials or samples and is extended in mzIdentML to capture descriptions of the starting sample(s) that has been analysed. Controlled vocabulary or ontology terms can be attached to represent the properties of the sample.
- <*Software*> – the FuGE <Software> class is extended in mzIdentML to provide additional details about the specific software used for proteome informatics.
- *Controlled vocabulary terms* - An alteration has been made in the FuGE schema, to use the mzML mechanism for referencing controlled vocabulary terms, rather than the more complex ontology term specification in the standard FuGE schema.

4.2 The PSI Mass Spectrometry Controlled Vocabulary (CV)

The PSI-MS controlled vocabulary is intended to provide terms for annotation of mzML and mzIdentML files. The CV has been generated by 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 by have a “has_units” key in the CV itself

(relationship: has_units: UO:0000221 ! dalton). The details of which terms are allowed or required in a given schema section is reported in the mapping file (Section 4.3).

As recommended by the PSI CV guidelines, psi-ms.obo should be dynamically maintained via the psidev-ms-vocab@lists.sourceforge.net mailing list that allows any user to request new terms in agreement with the community involved. Once a consensus is reached among the community the new terms are added within few 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.sourceforge.net/mod/data/PSI-MOD.obo
- Unimod modifications database - http://www.unimod.org/obo/unimod.obo

4.3 Validation of controlled vocabulary terms

The correct usage of controlled vocabulary terms within mzIdentML is governed by the use of a mapping file which defines each XML location (XPath) where a <cvParam> instance can be used, and the allowed terms from the PSI-MS, or other, controlled vocabularies. The mapping file is read and interpreted by validation software, checking that the data annotation is consistent. The mapping file needs to be checked and updated when the structure of CV is changed, and in some instances when new terms are added to the CV. The draft specifications for the mapping file can be found here: http://www.psidev.info/files/validator/PSI-Mapping.doc. XML paths are associated with CV terms along with a requirement level (MAY, SHOULD or MUST) defining what should be reported by validation software if one of the mapped terms is not provided in an instance document. Example validation software based on the mapping file has been implemented as part of OpenMS: www.psidev.info/validator, which has been used to perform syntactic and semantic validation of the example files listed in Section 5.4.

5. Resolved Design and scope issues

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

5.1.1 Quantitation

There is a clear requirement for a standard data format that supports quantitative data from studies of peptide and proteins. During the development process, several attempts were made to model the range of analysis

procedures currently used in proteome studies that produce quantitative data under the name AnalysisXML. The variability in the different techniques employed (e.g. labelled, label-free) and the continual evolution of new techniques resulted in considerable delays in getting a version 1.0 of AnalysisXML produced. It was decided at the 2008 PSI meeting in Toledo that the best course of action would be to get a stable format released without support for quantitative data, rather than facing further delays. At the 2009 PSI meeting in Turku, several quantification use cases were examined and it was demonstrated that a format for quantification would be simpler to develop independently of the format for identification. It was thus decided to split the development of AnalysisXML into two formats: mzIdentML and mzQuantML. It is expected that mzQuantML will follow a broadly similar structure as the upper level hierarchy of mzIdentML. An instance of mzQuantML will reference back to <SpectrumIdentificationItem> and <ProteinDetectionHypothesis> within an mzIdentML file for peptide and protein identifications respectively. This design is relatively intuitive since software typically performs identification and quantification in independent processes.

5.1.2 Use of FuGE-light schema

FuGE is a data model designed to be extended to create technology-specific data formats. The advantages of using FuGE as a basis for development are that several different formats would share the same core structure, simplifying the integration of data produced using different technologies. Two problems were faced that prevented FuGE being used in its native mode: (i) FuGE development is centred on Unified Modeling Language (and mapped automatically to XML Schema) and (ii) the FuGE specification is capable of representing a wide range of use cases, and as such, is a large complex specification. First, it was decided that the Proteome Informatics WG found working directly with XML Schema resulted in faster development cycles. Second, few developers had a working knowledge of FuGE, and it was therefore decided that by cutting down FuGE to the essential classes would speed development, and would help implementers of the format. Third, it was decided that close compatibility with mzML was a high priority, and as such several attributes were renamed to match those in mzML, as detailed in Section 4.1. The advantages of using FuGE-light in its present form are that extension points have been defined, for example, allowing quantitative models to be added in later versions, and the extension points are standardised across all FuGE implementing systems and formats.

5.1.3 Handling updates to the controlled vocabulary

There is a difficult issue with respect to how software should encode CV terms, such that changes to core can be accommodated. This issue is discussed at length in the mzML specification document [Deutsch08], and mzIdentML follows the same convention. In brief, when a new term is required, the file producers must contact the CV working group and request the new term. It is anticipated that problems may arise if a consumer of the file encounters a new CV term and they are not working from the latest version of the CV file. It has been decided that rather than aim for a workaround to this issue, it can be expected that data file consumers must ensure that the OBO file is up-to-date.

5.1.4 Use of identifiers for input spectra to a search

A <SpectrumIdentificationResult> is linked to the source spectrum (in an external file) from which the identifications are made by way of a reference in the spectrumID attribute and via the <SpectraData> 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. A CV term MUST be imported into the <SpectraData> element to demonstrate which system for identifying input spectra is being used in the spectrumID attribute of <SpectrumIdentificationResult>. *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 conversion of 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 conversion of 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.

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 mzIdentML, the spectrumID attribute should be constructed following the data type specification in Table 1. As an example, to reference the third spectrum in an mgf (Mascot Generic Format) file:

```
<SpectrumIdentificationResult id="Res1" spectrumID="index=3" SpectraData_ref="InputSpectra1">
```

...

```
<SpectraData location="local/mgf/merge.mgf" id="SD_1" >
  <fileFormat>
    <cvParam accession="MS:1001062" name="Mascot MGF file" cvRef="PSI-MS" />
  </fileFormat>
  <spectrumIDFormat>
    <cvParam accession="MS:1000774" name="multiple peak list nativeID format" cvRef="PSI-MS" />
  </spectrumIDFormat>
</SpectraData>
```

Spectra represented in mzML (in the <Spectrum> element) have a unique identifier within the “id” attribute, formulated as above depending on the source of the file. If the source file is mzML, <SpectrumIdentificationResult> MUST reference the value in “id” attribute to reference the spectrum that was searched.

5.1.5 Recommendations for reporting multiple spectrum identifications and protein hypotheses

There has been discussion of including a recommendation in this specification for what should be reported to allow statistical processing of results. For example, it has been noted that without peptide identifications reported for all (or most) spectra, it is difficult to perform comparative statistical analysis without a reference point. As discussed in Section 7.4, mzIdentML allows multiple peptide and protein identifications to be included with a flag for those identifications that the file producer deems to have passed a threshold. This structure MAY be used to provide sufficient information to allow further statistical processing to be carried out but it has been decided that recommendations about the level of detail to report are handled as part of the MIAPE MSI document.

5.1.6 Exclusion of information relating to mass spectral data

It has been decided that the peak list that was searched should remain external to the format, for example referenced as an mzML file. Similarly other data items that may be used during a search, but can be retrieved from the source spectra file are not duplicated in mzIdentML, such as retention time.

5.2 Open Issues

None at present, any issues identified during the document process will appear here.

5.3 Comments on Specific Use Cases

Many special use cases for mzIdentML were considered during its development. Each of these use cases has a corresponding example file that exercises the relevant part of the schema and provides a reference implementation example (see supporting documentation). Authors of software that create mzIdentML are encouraged to examine the examples that accompany this format release before implementing the writer. Further, such authors are encouraged to use the validator before releasing any new writer code and working with the PSI PI Working Group to resolve any issues. In the subsections below, we comment on a few of the notable use cases that were considered.

5.3.1 Multiple database search engines

Proteomics groups now commonly analyze MS data using multiple search engines and combine results to improve the number of peptide and protein identifications that can be made. The output of such approaches can be represented in mzIdentML as follows (see Section 6 for documentation of the model elements). Each database search SHOULD be represented by an instance of <SpectrumIdentification> (application of the protocol) which references the <SpectrumIdentificationProtocol> and the output data: an instance of <SpectrumIdentifcationList>. As such, if three database search engines are used, there SHOULD be three instances each of <SpectrumIdentification>, <SpectrumIdentificationProtocol> and <SpectrumIdentifcationList>. Results are then combined into a list of proteins by a separate process, represented as one instance of <ProteinDetection> (application of the protocol), which references one instance of <ProteinDetectionProtocol> and references (as input) the three instances of <SpectrumIdentificationList>. The output of <ProteinDetection> is one instance of <ProteinDetectionList>. If a secondary scoring scheme is used to weigh evidence for peptide-spectrum matches according to the search engines that have identified them, any consensus or composite scores should be assigned to each <SpectrumIdentificationItem> within parallel lists.

It was decided that more complex arrangements of workflows cannot be represented in mzIdentML version 1, such as different protein lists produced by each search engine, then combined by an additional process, since it becomes difficult to define which are “final” and which are “intermediate” results for data consumers and implementers of databases. Such workflows may be incorporated into later versions of the format.

5.3.2 Spectral library searches

An alternative to sequence database searches for identifying peptides from MS data is to search a pre-compiled library of peptide-spectrum matches. These spectral library searches are supported in mzIdentML. The recommended encoding is similar to sequence database search results, the main difference being that rather than protein sequences represented in the <DBSequence> element, the peptide sequence for each library entry is stored here instead. Additional information about the peptide-spectrum match, such as observed modifications and consensus scores, can be stored as CV terms within each <DBSequence> entry.

5.4 Other supporting materials

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

All example files can be downloaded manually from:
http://code.google.com/p/psi-pi/source/browse/trunk/examples/

a) Mascot_MSMS_example.mzid - a simple example of four MS/MS spectra searched against a protein database with Mascot

b) Mascot_N15_example.mzid - an example of a search using two sets of residue masses, ^{14}N and ^{15}N with Mascot.

c) Mascot_NA_example.mzid - an example of a search against an EST database with Mascot.

d) Mascot_top_down_example.mzid - a single MS/MS spectra from an intact protein, searched with Mascot.

e) MPC_example.mzid – an example of PSMs from different search engines, assembled into proteins using a third-party algorithm; false-discovery estimation using decoy database.

f) omssa_example_full.mzid - cut down example MS-MS search results including decoy matches from OMSSA.

g) PMF_example.mzid - example Peptide Mass Fingerprint search with Mascot.

h) Sequest_example.mzid - a simple example derived from a “.out” file produced by SEQUEST.

i) spectraST.mzid - example of a spectral library search using SpectraST.

j) xtandem_example_full.mzid - example MS-MS search results including decoy matches from X!Tandem.

k) Mascot_mzml_example.mzid – an example search of a spectrum file in mzML format.

6. Model in XML Schema

The following documentation is automatically generated from the XML Schema.

6.1 Element <mzIdentML>

Definition: The upper-most hierarchy level of mzIdentML with sub-containers for example describing software, protocols and search results (spectrum identifications or protein detection results).

Type: PSI-PI.Main.mzIdentMLType

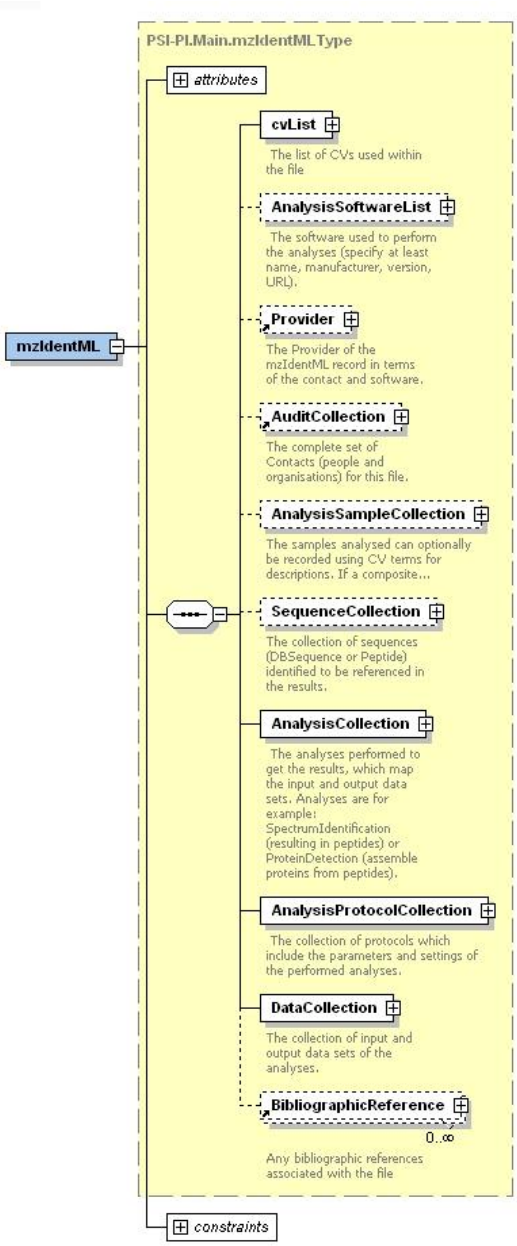
Attributes:

Attribute Name	Data Type	Use	Definition
creationDate	xsd:dateTime	optional	The date on which the file was produced.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
version	psi-pi:versionRegex	required	The version of the schema this instance document refers to, in the format x.y.z. Changes to z should not affect prevent instance documents from validating.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvList	1	1	The list of CVs used within the file
AnalysisSoftwareList	0	1	The software used to perform the analyses (specify at least name, manufacturer, version, URL).
Provider	0	1	The Provider of the mzIdentML record in terms of the contact and software.
AuditCollection	0	1	The complete set of Contacts (people and organisations) for this file.
AnalysisSampleCollection	0	1	The samples analysed can optionally be recorded using CV terms for descriptions. If a composite...
SequenceCollection	0	1	The collection of sequences (DBSequence or Peptide) identified to be referenced in the results.
AnalysisCollection	1	1	The analyses performed to get the results, which map the input and output data sets. Analyses are for example: SpectrumIdentification (resulting in peptides) or ProteinDetection (assemble proteins from peptides).
AnalysisProtocolCollection	1	1	The collection of protocols which include the parameters and settings of the performed analyses.
DataCollection	1	1	The collection of input and output data sets of the analyses.
BibliographicReference	0	unbounded	Any bibliographic references associated with the file

Graphical Context:



Example Context:

```
<mzIdentML id="MPC_use_case" creationDate="2008-11-28T13:56:00"
xmlns="http://psidev.info/psi/pi/mzIdentML/1.0" xmlns:pf="http://psidev.info/fuge-light/1.0" xmlns:PSI-
MS="http://psidev.info/psi/pi/mzIdentML/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://psidev.info/psi/pi/mzIdentML/1.0 ../schema/mzIdentML_working.xsd"
version="0.9.9">
  <cvList>
    <cv id="PSI-MS" fullName="Proteomics Standards Initiative Mass Spectrometry Vocabularies"
URI="http://www.psidev.info/PSI-MS" version="2.0.0"/>
    <cv id="BTO" fullName="BRENDA tissue 7 enzyme source" URI="http://www.brenda-enzymes.info/"
version="12/2007"/>
    <cv id="UNIMOD" fullName="UNIMOD CV for modifications" URI="http://www.unimod.org/obo/unimod.obo"/>
    <cv id="UO" fullName="Unit Ontology"
URI="http://obo.cvs.sourceforge.net/*checkout*/obo/obo/ontology/phenotype/unit.obo"/>
  ...
</mzIdentML>
```

6.2 Element <cvList>

Definition: The list of CVs used within the file

Type: cvListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cv	1	unbounded	A source controlled vocabulary from which cvParams will be obtained.

Example Context:

```
  <cvList>
    <cv id="PSI-MS" fullName="Proteomics Standards Initiative Mass Spectrometry Vocabularies"
URI="http://www.psidev.info/PSI-MS" version="2.0.0"/>
    <cv id="BTO" fullName="BRENDA tissue 7 enzyme source" URI="http://www.brenda-enzymes.info/"
version="12/2007"/>
    <cv id="UNIMOD" fullName="UNIMOD CV for modifications"
URI="http://www.unimod.org/obo/unimod.obo"/>
    <cv id="UO" fullName="Unit Ontology"
URI="http://obo.cvs.sourceforge.net/*checkout*/obo/obo/ontology/phenotype/unit.obo"/>
  </cvList>
```

6.3 Element <AnalysisSoftwareList>

Definition: The software used to perform the analyses (specify at least name, manufacturer, version, URL).

Type: AnalysisSoftwareListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
AnalysisSoftware	1	unbounded	The software used for performing the analyses.

Example Context:

```
  <AnalysisSoftwareList>
    <AnalysisSoftware id="SEQUEST_SW" name="ThermoFisher TurboSequest" version="PVM Slave v.27 (rev.
12)" URI="http://www.thermo.com/com/cda/product/detail/1,,16483,00.html">
      <ContactRole Contact_ref="THERMO">
        <role>
          <cvParam accession="MS:1001267" name="software vendor" cvRef="PSI-MS"/>
        </role>
      </ContactRole>
  ...
</AnalysisSoftwareList>
```

6.4 Element <Provider>

Definition: The Provider of the mzIdentML record in terms of the contact and software.

Attributes:

Attribute Name	Data Type	Use	Definition
Software_ref	xsd:string	optional	The Software that produced the document instance.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
ContactRole	0	1	The Contact that provided the document instance.

Example Context:

```
<Provider id="PROVIDER">
  <ContactRole Contact_ref="PERSON_DOC_OWNER">
    <role>
      <cvParam accession="MS:1001271" name="researcher" cvRef="PSI-MS" />
    </role>
  </ContactRole>
</Provider>
```

6.5 Element <AuditCollection>

Definition: The complete set of Contacts (people and organisations) for this file.

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Organization	0	unbounded	The complete set of Contacts.
Person	0	unbounded	The complete set of Contacts.

Example Context:

```
<AuditCollection>
  <Organization id="BRUKER" name="Bruker Daltonics GmbH" address="Bremen, Germany"/>
  <Organization id="MATRIXSCIENCE" name="MatrixScience" address="UK"/>
  <Organization id="THERMO" name="Thermo Corp." address="USA"/>
  <Person id="MPCMEYER" name="Prof. Dr. Helmut E. Meyer" address="Universitaetsstr. 150, D-44795
Bochum, Germany" email="helmut.e.meyer@rub.de">
    <affiliations Organization_ref="MPCINSTITUTE"/>
  </Person>
 ...
</AuditCollection>
```

6.6 Element <AnalysisSampleCollection>

Definition: The samples analysed can optionally be recorded using CV terms for descriptions. If a composite...

Type: AnalysisSampleCollectionType

Attributes: None

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Sample	1	unbounded	A description of the sample analysed by mass spectrometry using CVParams or UserParams. If a composite sample has been analysed, a parent sample should be defined, which references subsamples.

Example Context:

```
<AnalysisSampleCollection>
  <Sample id="sample1">
    <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="9606"/>
    <cvParam accession="BTO:0000255" name="brain cell line" cvRef="BTO"/>
  </Sample>
</AnalysisSampleCollection>
```

6.7 Element <SequenceCollection>

Definition: The collection of sequences (DBSequence or Peptide) identified to be referenced in the results.

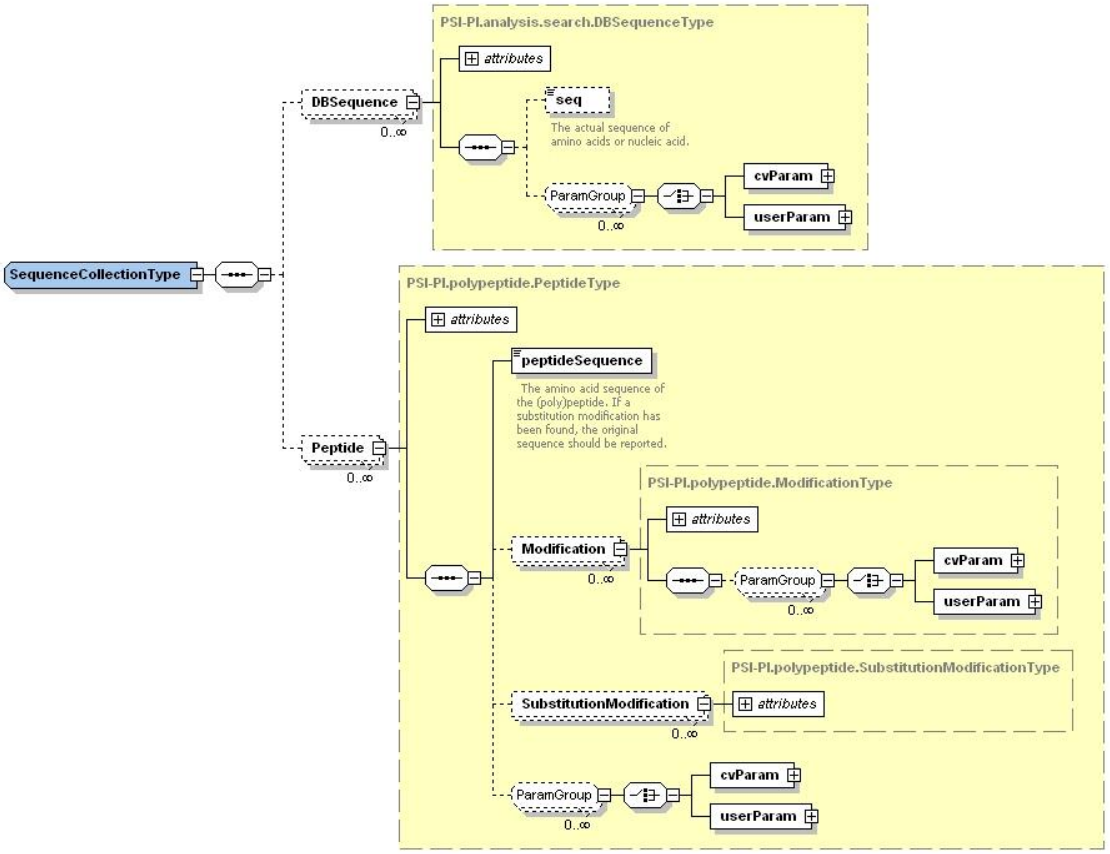
Type: SequenceCollectionType

Attributes: None

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
DBSequence	0	unbounded	A database sequence from the specified SearchDatabase (nucleic acid or amino acid). If the sequence is nucleic acid, the source nucleic acid sequence should be given in the seq attribute rather than a translated sequence.
Peptide	0	unbounded	One (poly)peptide (a sequence with modifications).

Graphical Context:



Example Context:

```
  <SequenceCollection>
    <DBSequence id="prot1_IPI" accession="IPI00013808.1" SearchDatabase_ref="ipi.HUMAN_decoy">
      <seq>MVDYH...GESDL</seq>
      <cvParam accession="MS:1001088" name="protein description" cvRef="PSI-MS"
value="IPI:IPI00013808.1|SWISS-PROT:O43707|TREMBL:Q96BG6|ENSEMBL:ENSP00000252699|REFSEQ:NP_004915|H-
INV:HIT000032172|VEGA:OTTHUMP00000076071;OTTHUMP00000174445 Tax_Id=9606 Gene_Symbol=ACTN4 Alpha-
actinin-4"/>
    </DBSequence>
    <DBSequence id="prot2_IPI" accession="IPI00554648.1" SearchDatabase_ref="ipi.HUMAN_decoy">
      <seq>SIRVTQK...VLPK</seq>
  ...
</SequenceCollection>
```

6.8 Element <AnalysisCollection>

Definition: The analyses performed to get the results, which map the input and output data sets. Analyses are for example: SpectrumIdentification (resulting in peptides) or ProteinDetection (assemble proteins from peptides).

Type: AnalysisCollectionType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SpectrumIdentification	1	unbounded	An Analysis which tries to identify peptides in input spectra, referencing the database searched, the input spectra, the output results and the protocol that is run.
ProteinDetection	0	1	An Analysis which assembles a set of peptides (e.g. from a spectra search analysis) to proteins.

Example Context:

```
  <AnalysisCollection>
    <SpectrumIdentification id="SEQUEST_analysis" SpectrumIdentificationProtocol_ref="SEQUEST_proto"
SpectrumIdentificationList_ref="SEQUEST_results" activityDate="2007-05-12T13:00:00">
      <InputSpectra SpectraData_ref="LCMALDI_spectra"/>
      <SearchDatabase SearchDatabase_ref="ipi.HUMAN_decoy"/>
    </SpectrumIdentification>
    <SpectrumIdentification id="Mascot_analysis" SpectrumIdentificationProtocol_ref="Mascot_proto"
SpectrumIdentificationList_ref="Mascot_results" activityDate="2007-05-12T14:00:00">
      <InputSpectra SpectraData_ref="LCMALDI_spectra"/>
```

```
  ...
</AnalysisCollection>
```

6.9 Element <AnalysisProtocolCollection>

Definition: The collection of protocols which include the parameters and settings of the performed analyses.

Type: AnalysisProtocolCollectionType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SpectrumIdentificationProtocol	1	unbounded	The parameters and settings of a SpectrumIdentification analysis.
ProteinDetectionProtocol	0	1	The parameters and settings of a ProteinDetection process.

Example Context:

```
<AnalysisProtocolCollection>
  <SpectrumIdentificationProtocol id="SIP" AnalysisSoftware_ref="AS_mascot_server">
    <SearchType>
      <cvParam accession="MS:1001081" name="pmf search" cvRef="PSI-MS" value=""/>
    </SearchType>
    <AdditionalSearchParams>
      <userParam name="Mascot User Comment" value="Figure 8. MALDI-TOF spectrum of an in-gel tryptic
digest of a protein isolated from a thermophilic bacterium"/>
  ...
</AnalysisProtocolCollection>
```

6.10 Element <DataCollection>

Definition: The collection of input and output data sets of the analyses.

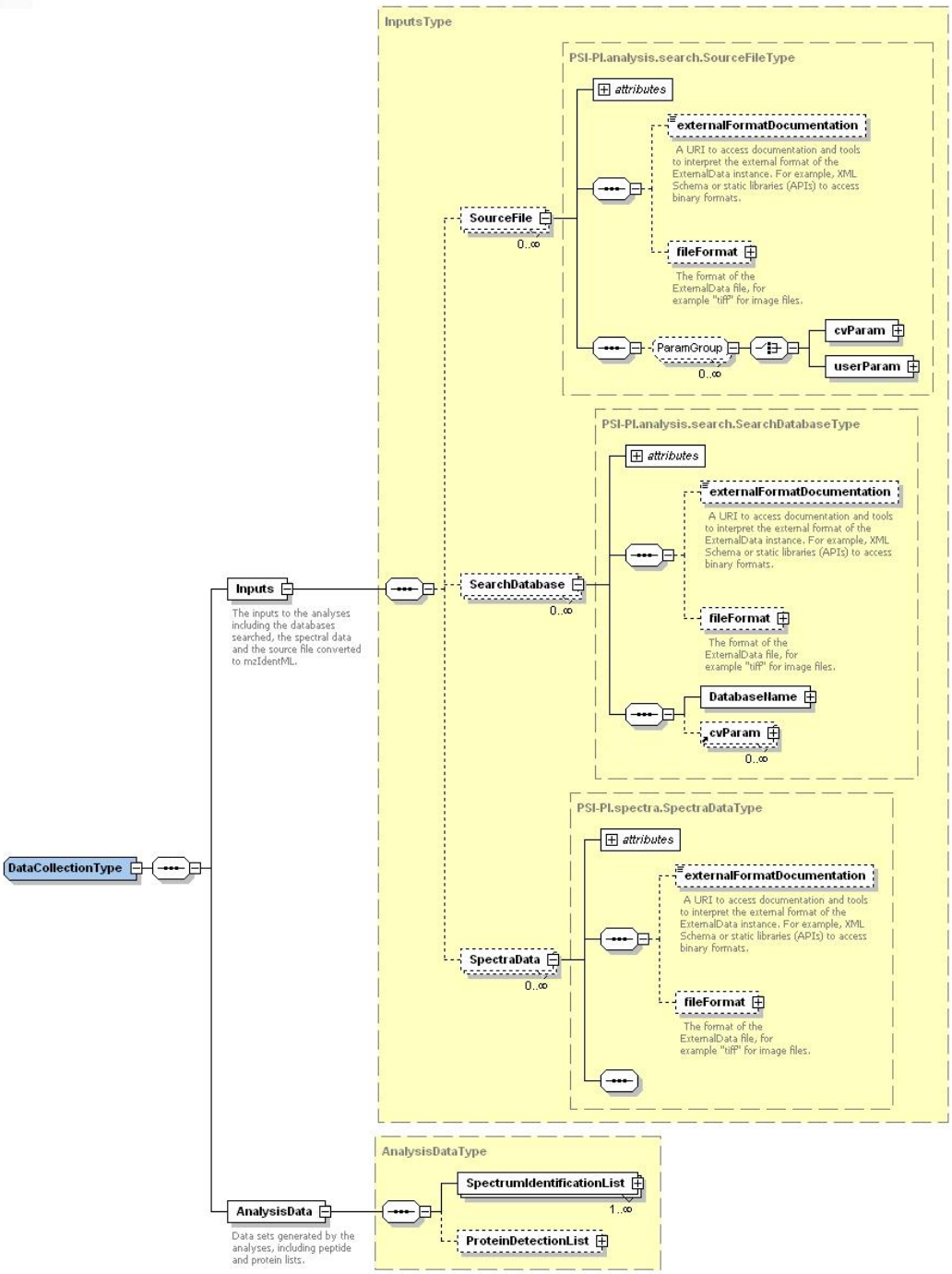
Type: DataCollectionType

Attributes: None

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Inputs	1	1	The inputs to the analyses including the databases searched, the spectral data and the source file converted to mzIdentML.
AnalysisData	1	1	Data sets generated by the analyses, including peptide and protein lists.

Graphical Context:



Example Context:

```
  <DataCollection>
    <Inputs>
      <SourceFile id="SF1" location="proteinscape://www.medizinisches-proteom-
center.de/PSServer/Project/Sample/Separation_1D_LC/Fraction_X/SpectraData/Results1">
        <fileFormat>
          <cvParam accession="MS:1001275" name="ProteinScape SearchEvent" cvRef="PSI-MS"/>
        </fileFormat>
      </SourceFile>
  ...
</DataCollection>
```

6.11 Element <BibliographicReference>

Definition: Any bibliographic references associated with the file

Attributes: none

Subelements: none

Example Context:

```
<BibliographicReference authors="David N. Perkins, Darryl J. C. Pappin, David M. Creasy, John S.
Cottrell" editor="" id="10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2"
name="Probability-based protein identification by searching sequence databases using mass spectrometry
data" issue="18" pages="3551-3567" publication="Electrophoresis" volume="20" year="1999"
publisher="Wiley VCH" title="Probability-based protein identification by searching sequence databases
using mass spectrometry data"/>
```

6.12 Element <cv>

Definition: A source controlled vocabulary from which cvParams will be obtained.

Attributes: none

Subelements: none

Example Context:

```
<cv id="PSI-MS" fullName="Proteomics Standards Initiative Mass Spectrometry Vocabularies"
URI="http://www.psidev.info/PSI-MS" version="2.0.0"></cv>
```

6.13 Element <AnalysisSoftware>

Definition: The software used for performing the analyses.

Type: PSI-PI.analysis.search.AnalysisSoftwareType

Attributes:

Attribute Name	Data Type	Use	Definition
URI	xsd:anyURI	optional	URI of the analysis software e.g. manufacturer's website
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
version	xsd:string	optional	The version of Software used.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
ContactRole	0	1	The Contact that provided the document instance.
SoftwareName	1	1	The name of the analysis software package, sourced from a CV if available.
Customizations	0	1	Any customizations to the software, such as alternative scoring mechanisms implemented, should be documented here as free text.

Example Context:

```
    <AnalysisSoftware id="SEQUEST_SW" name="ThermoFisher TurboSequest" version="PVM Slave v.27 (rev.
12)" URI="http://www.thermo.com/com/cda/product/detail/1,,16483,00.html">
      <ContactRole Contact_ref="THERMO">
        <role>
          <cvParam accession="MS:1001267" name="software vendor" cvRef="PSI-MS"/>
        </role>
      </ContactRole>
      <SoftwareName>
  ...
</AnalysisSoftware>
```

6.14 Element <Sample>

Definition: A description of the sample analysed by mass spectrometry using CVParams or UserParams. If a composite sample has been analysed, a parent sample should be defined, which references subsamples.

Type: SampleType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
ContactRole	0	unbounded	The Contact that provided the document instance.
subSample	0	unbounded	References to the individual component samples within a mixed parent sample.

Example Context:

```
<Sample id="sample1">
  <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="9606"/>
  <cvParam accession="BTO:0000255" name="brain cell line" cvRef="BTO"/>
</Sample>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisSampleCollection/Sample
SHOULD supply a *child* term of MS:1001089 (molecule taxonomy) one or more times
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001467 (taxonomy: NCBI TaxID)
  e.g.: MS:1001468 (taxonomy: common name)
  e.g.: MS:1001469 (taxonomy: scientific name)
  e.g.: MS:1001470 (taxonomy: Swiss-Prot ID)
SHOULD supply a *child* term of BTO:0000000 (brenda source tissue ontology) one or more times
```

6.15 Element <DBSequence>

Definition: A database sequence from the specified SearchDatabase (nucleic acid or amino acid). If the sequence is nucleic acid, the source nucleic acid sequence should be given in the seq attribute rather than a translated sequence.

Type: PSI-PI.analysis.search.DBSequenceType

Attributes:

Attribute Name	Data Type	Use	Definition
SearchDatabase_ref	xsd:string	required	The source database of this sequence.
accession	xsd:string	required	The unique accession of this sequence.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
length	xsd:int	optional	The length of the sequence as a number of bases or residues.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
seq	0	1	The actual sequence of amino acids or nucleic acid.
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<DBSequence id="prot5_IPI" accession="IPI00398776.3" SearchDatabase_ref="ipi.HUMAN_decoy">
  <seq>MKIVP...GPESAVA</seq>
  <cvParam accession="MS:1001088" name="protein description" cvRef="PSI-MS"
value=">IPI:IPI00398776.3|TREMBL:Q6S379;Q96IE3|REFSEQ:NP_958783 Tax_Id=9606 plectin 1 isoform 7"/>
</DBSequence>
```

cvParam Mapping Rules:

```
Path /mzIdentML/SequenceCollection/DBSequence
MAY supply a *child* term of MS:1001342 (database sequence details) one or more times
  e.g.: MS:1001088 (protein description)
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001343 (NA sequence)
  e.g.: MS:1001344 (AA sequence)
  e.g.: MS:1001467 (taxonomy: NCBI TaxID)
  e.g.: MS:1001468 (taxonomy: common name)
  e.g.: MS:1001469 (taxonomy: scientific name)
  e.g.: MS:1001470 (taxonomy: Swiss-Prot ID)
MAY supply a *child* term of MS:1001089 (molecule taxonomy) one or more times
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001467 (taxonomy: NCBI TaxID)
  e.g.: MS:1001468 (taxonomy: common name)
  e.g.: MS:1001469 (taxonomy: scientific name)
  e.g.: MS:1001470 (taxonomy: Swiss-Prot ID)
```

6.16 Element <Peptide>

Definition: One (poly)peptide (a sequence with modifications).

Type: PSI-PI.polypeptide.PeptideType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
peptideSequence	1	1	The amino acid sequence of the (poly)peptide. If a substitution modification has been found, the original sequence should be reported.
Modification	0	unbounded	A molecule modification specification. If n modifications have been found on a peptide, there should be n instances of Modification. If multiple modifications are provided as cvParams, it is assumed that the modification is ambiguous i.e. one modification or another. If no CVParams are provided it is assumed that the delta has not been matched to a known modification. A neutral loss should be defined as an additional CVParam within Modification. If more complex information should be given about neutral losses (such as presence/absence on particular product ions), this can additionally be encoded within the FragmentationArray.
SubstitutionModification	0	unbounded	A modification where one residue is substituted by another (amino acid change).
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<Peptide id="peptide_176_4">
  <peptideSequence>EMMYKIAAMQSVDIDPATVK</peptideSequence>
  <Modification location="2" residues="M" monoisotopicMassDelta="15.994919">
    <cvParam accession="UNIMOD:35" name="Oxidation" cvRef="UNIMOD" />
    <cvParam accession="MS:1001524" name="fragment neutral loss" cvRef="PSI-MS" value="63.998285"
unitAccession="UO:0000221" unitName="dalton" unitCvRef="UO"/>
  </Modification>
  <Modification location="3" residues="M" monoisotopicMassDelta="15.994919">
  ...
</Peptide>
```

cvParam Mapping Rules:

```
Path /mzIdentML/SequenceCollection/Peptide
MAY supply a *child* term of MS:1001355 (peptide descriptions) one or more times
```

6.17 Element <SpectrumIdentification>

Definition: An Analysis which tries to identify peptides in input spectra, referencing the database searched, the input spectra, the output results and the protocol that is run.

Type: PSI-PI.analysis.search.SpectrumIdentificationType

Attributes:

Attribute Name	Data Type	Use	Definition
SpectrumIdentificationList_ref	xsd:string	required	A reference to the SpectrumIdentificationList produced by this analysis in the DataCollection section.
SpectrumIdentificationProtocol_ref	xsd:string	required	A reference to the search protocol used for this SpectrumIdentification.
activityDate	xsd:dateTime	optional	When the protocol was applied.

id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
InputSpectra	1	unbounded	One of the spectra data sets used (can be several).
SearchDatabase	1	unbounded	One of the search databases used (can be several).

Example Context:

```
    <SpectrumIdentification id="SEQUEST_analysis" SpectrumIdentificationProtocol_ref="SEQUEST_proto"
SpectrumIdentificationList_ref="SEQUEST_results" activityDate="2007-05-12T13:00:00">
      <InputSpectra SpectraData_ref="LCMALDI_spectra"/>
      <SearchDatabase SearchDatabase_ref="ipi.HUMAN_decoy"/>
    </SpectrumIdentification>
```

6.18 Element <ProteinDetection>

Definition: An Analysis which assembles a set of peptides (e.g. from a spectra search analysis) to proteins.

Type: PSI-PI.analysis.process.ProteinDetectionType

Attributes:

Attribute Name	Data Type	Use	Definition
ProteinDetectionList_ref	xsd:string	required	A reference to the ProteinDetectionList in the DataCollection section.
ProteinDetectionProtocol_ref	xsd:string	required	A reference to the detection protocol used for this ProteinDetection.
activityDate	xsd:dateTime	optional	When the protocol was applied.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
InputSpectrumIdentifications	1	unbounded	The lists of spectrum identifications that are input to the protein detection process.

Example Context:

```
    <ProteinDetection id="ProteinExtractor_analysis"
ProteinDetectionProtocol_ref="ProteinExtractor_proto"
ProteinDetectionList_ref="ProteinExtractor_results" activityDate="2007-05-12T15:30:00">
      <InputSpectrumIdentifications SpectrumIdentificationList_ref="SEQUEST_results"/>
      <InputSpectrumIdentifications SpectrumIdentificationList_ref="Mascot_results"/>
    </ProteinDetection>
```

6.19 Element <SpectrumIdentificationProtocol>

Definition: The parameters and settings of a SpectrumIdentification analysis.

Type: PSI-PI.analysis.search.SpectrumIdentificationProtocolType

Attributes:

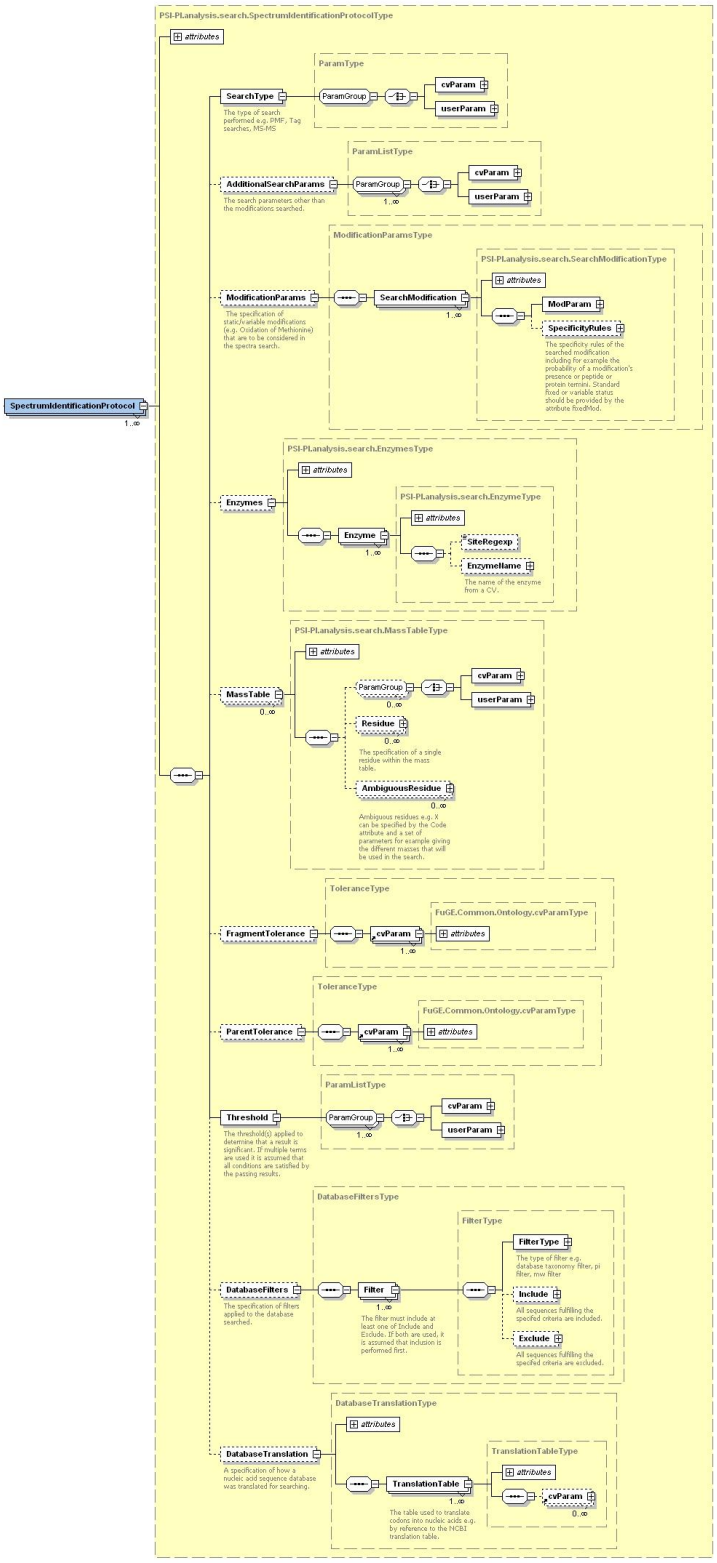
Attribute Name	Data Type	Use	Definition
AnalysisSoftware_ref	xsd:string	required	The search algorithm used, given as a reference to the SoftwareCollection section.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition

SearchType	1	1	The type of search performed e.g. PMF, Tag searches, MS-MS
AdditionalSearchParams	0	1	The search parameters other than the modifications searched.
ModificationParams	0	1	The specification of static/variable modifications (e.g. Oxidation of Methionine) that are to be considered in the spectra search.
Enzymes	0	1	The list of enzymes used in experiment
MassTable	0	unbounded	The masses of residues used in the search.
FragmentTolerance	0	1	The tolerance of the search given as a plus and minus value with units.
ParentTolerance	0	1	The tolerance of the search given as a plus and minus value with units.
Threshold	1	1	The threshold(s) applied to determine that a result is significant. If multiple terms are used it is assumed that all conditions are satisfied by the passing results.
DatabaseFilters	0	1	The specification of filters applied to the database searched.
DatabaseTranslation	0	1	A specification of how a nucleic acid sequence database was translated for searching.

Graphical Context:



Example Context:

```
<SpectrumIdentificationProtocol id="SIP" AnalysisSoftware_ref="AS_mascot_server">
  <SearchType>
    <cvParam accession="MS:1001081" name="pmf search" cvRef="PSI-MS" value=""/>
  </SearchType>
  <AdditionalSearchParams>
    <userParam name="Mascot User Comment" value="Figure 8. MALDI-TOF spectrum of an in-gel tryptic
digest of a protein isolated from a thermophilic bacterium"/>
    <cvParam accession="MS:1001211" name="parent mass type mono"    cvRef="PSI-MS"/>
  ...
</SpectrumIdentificationProtocol>
```

6.20 Element <ProteinDetectionProtocol>

Definition: The parameters and settings of a ProteinDetection process.

Type: PSI-PI.analysis.process.ProteinDetectionProtocolType

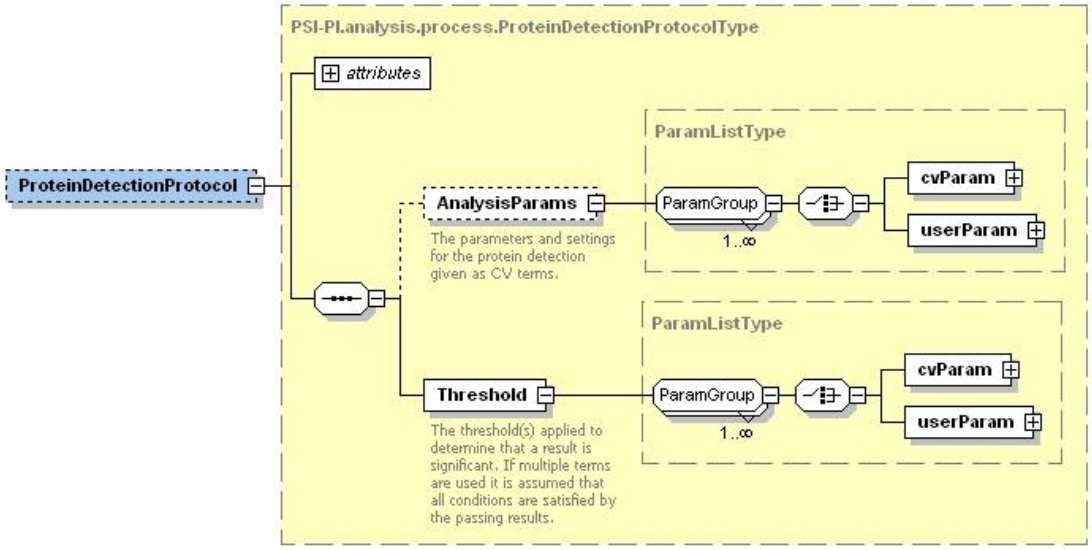
Attributes:

Attribute Name	Data Type	Use	Definition
AnalysisSoftware_ref	xsd:string	required	The protein detection software used, given as a reference to the SoftwareCollection section.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
AnalysisParams	0	1	The parameters and settings for the protein detection given as CV terms.
Threshold	1	1	The threshold(s) applied to determine that a result is significant. If multiple terms are used it is assumed that all conditions are satisfied by the passing results.

Graphical Context:



Example Context:

```
<ProteinDetectionProtocol id="PDP_MascotParser_1"  AnalysisSoftware_ref="AS_mascot_parser">
  <AnalysisParams>
    <cvParam accession="MS:1001316" name="mascot:SigThreshold" cvRef="PSI-MS" value="0.05"/>
    <cvParam accession="MS:1001317" name="mascot:MaxProteinHits" cvRef="PSI-MS" value="Auto"/>
    <cvParam accession="MS:1001318" name="mascot:ProteinScoringMethod" cvRef="PSI-MS"
value="Standard"/>
    <cvParam accession="MS:1001319" name="mascot:MinMSMSThreshold" cvRef="PSI-MS" value="0"/>
    <cvParam accession="MS:1001320" name="mascot:ShowHomologousProteinsWithSamePeptides" cvRef="PSI-MS"
value="1"/>
  ...
</ProteinDetectionProtocol>
```

6.21 Element <Inputs>

Definition: The inputs to the analyses including the databases searched, the spectral data and the source file converted to mzIdentML.

Type: InputsType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition

SourceFile	0	unbounded	A file from which this mzIdentML instance was created.
SearchDatabase	0	unbounded	One of the search databases used (can be several).
SpectraData	0	unbounded	A data set containing spectra data (consisting of one or more spectra).

Example Context:

```
    <Inputs>
      <SourceFile id="SF1" location="proteinscape://www.medizinisches-proteom-
center.de/PSServer/Project/Sample/Separation_1D_LC/Fraction_X/SpectraData/Results1">
        <fileFormat>
          <cvParam accession="MS:1001275" name="ProteinScape SearchEvent" cvRef="PSI-MS"/>
        </fileFormat>
      </SourceFile>
      <SearchDatabase id="ipi.HUMAN_decoy" location="uri://www.medizinisches-proteom-
center.de/ipi.HUMAN_decoy/3.15" version="3.15" releaseDate="22 February, 2006"
numDatabaseSequences="58099">
  ...
</Inputs>
```

6.22 Element <AnalysisData>

Definition: Data sets generated by the analyses, including peptide and protein lists.

Type: AnalysisDataType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SpectrumIdentificationList	1	unbounded	Represents the set of all search results from SpectrumIdentification.
ProteinDetectionList	0	1	The protein list resulting from a protein detection process.

Example Context:

```
<AnalysisData>
  <SpectrumIdentificationList id="SIL_1" numSequencesSearched="257964">
    <SpectrumIdentificationResult id="SIR_1" spectrumID="1" SpectraData_ref="SD_1">
      <SpectrumIdentificationItem id="SII_1_1"  calculatedMassToCharge="1107.534897" chargeState="1"
experimentalMassToCharge="1108.53" Peptide_ref="peptide_1_1" rank="0" passThreshold="true">
        <PeptideEvidence id="PE_1_1_UVRB_THET8" start="542" end="550" pre="R" post="V"
missedCleavages="0" isDecoy="false" DBSequence_Ref="DBSeq_UVRB_THET8" />
      </SpectrumIdentificationItem>
      <SpectrumIdentificationItem id="SII_1_8"  calculatedMassToCharge="1107.617584" chargeState="1"
experimentalMassToCharge="1108.53" Peptide_ref="peptide_1_8" rank="0" passThreshold="true">
  ...
</AnalysisData>
```

6.23 Element <ContactRole>

Definition: The Contact that provided the document instance.

Type: FuGE.Common.Audit.ContactRoleType

Attributes:

Attribute Name	Data Type	Use	Definition
Contact_ref	xsd:string	required	When a ContactRole is used, it specifies which Contact the role is associated with.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
role	1	1	The roles (lab equipment sales, contractor, etc.) the Contact fills.

Example Context:

```
      <ContactRole Contact_ref="ORG_MSL">
        <role>
          <cvParam accession="MS:1001267" name="software vendor" cvRef="PSI-MS"/>
        </role>
      </ContactRole>
```

6.24 Element <SoftwareName>

Definition: The name of the analysis software package, sourced from a CV if available.

Type: ParamType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<SoftwareName>
  <cvParam accession="MS:1000734" name="ProteinScape" cvRef="PSI-MS"/>
</SoftwareName>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisSoftwareList/AnalysisSoftware/SoftwareName
MUST supply a *child* term of MS:1001456 (analysis software) one or more times
  e.g.: MS:1000532 (Xcalibur)
  e.g.: MS:1000533 (Bioworks)
  e.g.: MS:1000534 (MassLynx)
  e.g.: MS:1000535 (FlexAnalysis)
  e.g.: MS:1000536 (Data Explorer)
  e.g.: MS:1000537 (4700 Explorer)
  e.g.: MS:1000539 (Voyager Biospectrometry Workstation System)
  e.g.: MS:1000551 (Analyst)
  e.g.: MS:1000600 (Proteios)
  e.g.: MS:1000601 (ProteinLynx Global Server)
  et al.
```

6.25 Element <Customizations>

Definition: Any customizations to the software, such as alternative scoring mechanisms implemented, should be documented here as free text.

Type: xsd:string

Attributes: none

Subelements: none

Example Context:

```
<Customizations>
  No customisations
</Customizations>
```

6.26 Element <subSample>

Definition: References to the individual component samples within a mixed parent sample.

Type: subSampleType

Attributes:

Attribute Name	Data Type	Use	Definition
Sample_ref	xsd:string	required	Reference to the individual component samples within a mixed parent sample.

Subelements: none

Example Context:

```
<subSample Sample_ref="Sample_light"/>
```

6.27 Element <seq>

Definition: The actual sequence of amino acids or nucleic acid.

Type: sequence

Attributes: none

Subelements: none

Example Context:

```
<seq>MKIVPDER...SAVA</seq>
```

6.28 Element <peptideSequence>

Definition: The amino acid sequence of the (poly)peptide. If a substitution modification has been found, the original sequence should be reported.

Type: sequence

Attributes: none

Subelements: none

Example

```
<peptideSequence>GLSDGEWQQG</peptideSequence>
```

Context:

6.29 Element <Modification>

Definition: A molecule modification specification. If n modifications have been found on a peptide, there should be n instances of Modification. If multiple modifications are provided as cvParams, it is assumed that the modification is ambiguous i.e. one modification or another. If no CVParams are provided it is assumed that the delta has not been matched to a known modification. A neutral loss should be defined as an additional CVParam within Modification. If more complex information should be given about neutral losses (such as presence/absence on particular product ions), this can additionally be encoded within the FragmentationArray.

Type: PSI-PI.polypeptide.ModificationType

Attributes:

Attribute Name	Data Type	Use	Definition
avgMassDelta	xsd:double	optional	Atomic mass delta considering the natural distribution of isotopes in Daltons.
location	xsd:int	optional	Location of the modification within the peptide - position in peptide sequence, counted from the N-terminus residue, starting at position 1. Specific modifications to the N-terminus should be given the location 0. Modification to the C-terminus should be given as peptide length + 1.
monoisotopicMassDelta	xsd:double	optional	Atomic mass delta when assuming only the most common isotope of elements in Daltons.
residues	listOfChars	optional	Specification of the residue (amino acid) on which the modification occurs. If multiple values are given, it is assumed that the exact residue modified is unknown i.e. the modification is to ONE of the residues listed. Multiple residues would usually only be specified for PMF data.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<Modification location="10" residues="M" monoisotopicMassDelta="15.994919">
  <cvParam accession="UNIMOD:35" name="Oxidation" cvRef="UNIMOD" />
  <cvParam accession="MS:1001524" name="fragment neutral loss" cvRef="PSI-MS" value="63.998285"
unitAccession="UO:0000221" unitName="dalton" unitCvRef="UO"/>
</Modification>
```

cvParam Mapping Rules:

```
Path /mzIdentML/SequenceCollection/Peptide/Modification
MUST supply a *child* term of UNIMOD:0 (UNIMOD root) one or more times OR
MUST supply a *child* term of MS:1001471 (peptide modification details) one or more times
  e.g.: MS:1001460 (unknown modification)
  e.g.: MS:1001524 (fragment neutral loss)
  e.g.: MS:1001525 (precursor neutral loss )
MUST supply a *child* term of MOD:00000 (protein modification) one or more times
```

6.30 Element <SubstitutionModification>

Definition: A modification where one residue is substituted by another (amino acid change).

Type: PSI-PI.polypeptide.SubstitutionModificationType

Attributes:

Attribute Name	Data Type	Use	Definition
avgMassDelta	xsd:double	optional	Atomic mass delta considering the natural distribution of isotopes in Daltons. This should only be reported if the original amino acid is known i.e. it is not "X"

location	xsd:int	optional	Location of the modification within the peptide - position in peptide sequence, counted from the N-terminus residue, starting at position 1. Specific modifications to the N-terminus should be given the location 0. Modification to the C-terminus should be given as peptide length + 1.
monoisotopicMassDelta	xsd:double	optional	Atomic mass delta when assuming only the most common isotope of elements in Daltons. This should only be reported if the original amino acid is known i.e. it is not "X"
originalResidue	xsd:restriction base="xsd:string"	required	The original residue before replacement.
replacementResidue	xsd:restriction base="xsd:string"	required	The residue that replaced the originalResidue.

Subelements: none

Example Context:

```
<SubstitutionModification location="7" originalResidue="X" replacementResidue="N"/>
```

6.31 Element <InputSpectra>

Definition: One of the spectra data sets used (can be several).

Type: InputSpectraType

Attributes:

Attribute Name	Data Type	Use	Definition
SpectraData_ref	xsd:string	optional	A reference to the SpectraData element which locates the input spectra to an external file.

Subelements: none

Example Context:

```
<InputSpectra SpectraData_ref="LCMALDI_spectra"/>
```

6.32 Element <SearchDatabase>

Definition: One of the search databases used (can be several).

Type: SearchDatabaseType

Attributes:

Attribute Name	Data Type	Use	Definition
SearchDatabase_ref	xsd:string	optional	A reference to the database searched.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
location	xsd:anyURI	required	The location of the data file.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
numDatabaseSequences	xsd:long	optional	The total number of sequences in the database.
numResidues	xsd:long	optional	The number of residues in the database.
releaseDate	xsd:string	optional	The release date of the database.
version	xsd:string	optional	The version of the database.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
externalFormatDocumentation	0	1	A URI to access documentation and tools to interpret the external format of the ExternalData instance. For example, XML Schema or static libraries (APIs) to access binary formats.
fileFormat	0	1	The format of the ExternalData file, for example "tiff" for image files.

DatabaseName	1	1	The database name may be given as a cvParam if it maps exactly to one of the release databases listed in the CV, otherwise a userParam should be used.
cvParam	0	unbounded	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
      <SearchDatabase id="ipi.HUMAN_decoy" location="uri://www.medizinisches-proteom-
center.de/ipi.HUMAN_decoy/3.15" version="3.15" releaseDate="22 February, 2006"
numDatabaseSequences="58099">
        <DatabaseName>
          <userParam name="MPC ipi.HUMAN_decoy"/>
        </DatabaseName>
        <cvParam accession="MS:1001300" name="decoy DB from IPI_human" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001197" name="DB composition target+decoy" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001452" name="decoy DB type shuffle" cvRef="PSI-MS"/>
  ...
</SearchDatabase>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/Inputs/SearchDatabase
MAY supply a *child* term of MS:1000561 (data file checksum type) one or more times
  e.g.: MS:1000568 (MD5)
  e.g.: MS:1000569 (SHA-1)
MAY supply a *child* term of MS:1001011 (search database details) one or more times
  e.g.: MS:1001014 (database local file path)
  e.g.: MS:1001015 (database original uri)
  e.g.: MS:1001016 (database version)
  e.g.: MS:1001017 (database release date)
  e.g.: MS:1001020 (DB filter taxonomy)
  e.g.: MS:1001021 (DB filter on accession numbers)
  e.g.: MS:1001022 (DB MW filter)
  e.g.: MS:1001023 (DB PI filter)
  e.g.: MS:1001024 (translation frame)
  e.g.: MS:1001025 (translation table)
  et al.
```

6.33 Element <InputSpectrumIdentifications>

Definition: The lists of spectrum identifications that are input to the protein detection process.

Type: InputSpectrumIdentificationsType

Attributes:

Attribute Name	Data Type	Use	Definition
SpectrumIdentificationList_ref	xsd:string	required	A reference to the list of spectrum identifications that were input to the process.

Subelements: none

Example Context:

```
      <InputSpectrumIdentifications SpectrumIdentificationList_ref="SEQUEST_results"/>
```

6.34 Element <SearchType>

Definition: The type of search performed e.g. PMF, Tag searches, MS-MS

Type: ParamType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<SearchType>
  <cvParam accession="MS:1001083" name="ms-ms search" cvRef="PSI-MS" value=""/>
</SearchType>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/SearchType
MUST supply a *child* term of MS:1001080 (search type) one or more times
  e.g.: MS:1001010 (de novo search)
  e.g.: MS:1001031 (spectral library search)
  e.g.: MS:1001081 (pmf search)
  e.g.: MS:1001082 (tag search)
  e.g.: MS:1001083 (ms-ms search)
```

6.35 Element <AdditionalSearchParams>

Definition: The search parameters other than the modifications searched.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
      <AdditionalSearchParams>
        <cvParam accession="MS:1001211" name="parent mass type mono" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001256" name="fragment mass type mono" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001259" name="param: immonium ion" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001108" name="param: a ion" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001146" name="param: a ion-NH3" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001148" name="param: a ion-H2O" cvRef="PSI-MS"/>
  ...
</AdditionalSearchParams>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/AdditionalSearchParams
MAY supply a *child* term of MS:1001302 (search engine specific input parameter) one or more times
  e.g.: MS:1001005 (sequest:CleavesAt)
  e.g.: MS:1001007 (sequest:OutputLines)
  e.g.: MS:1001009 (sequest:DescriptionLines)
  e.g.: MS:1001026 (sequest:NormalizeXCorrValues)
  e.g.: MS:1001028 (sequest:SequenceHeaderFilter)
  e.g.: MS:1001032 (sequest:SequencePartialFilter)
  e.g.: MS:1001037 (sequest:ShowFragmentIons)
  e.g.: MS:1001038 (sequest:Consensus)
  e.g.: MS:1001042 (sequest:LimitTo)
  e.g.: MS:1001046 (sequest:sort_by_dCn)
  et al.
MAY supply a *child* term of MS:1001066 (ions series considered in search) one or more times
  e.g.: MS:1001108 (param: a ion)
  e.g.: MS:1001118 (param: b ion)
  e.g.: MS:1001119 (param: c ion)
  e.g.: MS:1001146 (param: a ion-NH3)
  e.g.: MS:1001148 (param: a ion-H2O)
  e.g.: MS:1001149 (param: b ion-NH3)
  e.g.: MS:1001150 (param: b ion-H2O)
  e.g.: MS:1001151 (param: y ion-NH3)
  e.g.: MS:1001152 (param: y ion-H2O)
  e.g.: MS:1001257 (param: v ion)
  et al.
MAY supply a *child* term of MS:1001210 (mass type settings) one or more times
  e.g.: MS:1001211 (parent mass type mono)
  e.g.: MS:1001212 (parent mass type average)
  e.g.: MS:1001255 (fragment mass type average)
  e.g.: MS:1001256 (fragment mass type mono)
```

6.36 Element <ModificationParams>

Definition: The specification of static/variable modifications (e.g. Oxidation of Methionine) that are to be considered in the spectra search.

Type: ModificationParamsType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SearchModification	1	unbounded	Specification of a search modification as parameter for a spectra search. Contains the name of the modification, the mass, the specificity and whether it is a static modification.

Example Context:

```
<ModificationParams>
  <SearchModification fixedMod="false" >
    <ModParam massDelta="57.021469" residues="C">
      <cvParam accession="UNIMOD:4" name="Carbamidomethyl" cvRef="UNIMOD"/>
    </ModParam>
  </SearchModification>
  <SearchModification fixedMod="false" >
  ...
```

```
</ModificationParams>
```

6.37 Element <Enzymes>

Definition: The list of enzymes used in experiment

Type: PSI-PI.analysis.search.EnzymesType

Attributes:

Attribute Name	Data Type	Use	Definition
independent	xsd:boolean	optional	If there are multiple enzymes specified, this attribute is set to true if cleavage with different enzymes is performed independently

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Enzyme	1	unbounded	The details of an individual cleavage enzyme should be provided by giving a regular expression or a CV term if a "standard" enzyme cleavage has been performed.

Example Context:

```
<Enzymes independent="0">
  <Enzyme id="ENZ_0" CTermGain="OH" NTermGain="H" missedCleavages="1" semiSpecific="0">
    <SiteRegexp><![CDATA[(?<=M)]]></SiteRegexp>
    <EnzymeName>
      <cvParam accession="MS:1001307" name="CNBr" cvRef="PSI-MS" />
    </EnzymeName>
  </Enzyme>
  ...
</Enzymes>
```

6.38 Element <MassTable>

Definition: The masses of residues used in the search.

Type: PSI-PI.analysis.search.MassTableType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
msLevel	listOfIntegers	required	The MS spectrum that the MassTable refers to e.g. "1" for MS1 "2" for MS2 or "1 2" for MS1 or MS2
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
Residue	0	unbounded	The specification of a single residue within the mass table.
AmbiguousResidue	0	unbounded	Ambiguous residues e.g. X can be specified by the Code attribute and a set of parameters for example giving the different masses that will be used in the search.

Example Context:

```
<MassTable id="MT_light" msLevel="1 2">
  <Residue Code="A" Mass="71.037113805"/>
  <Residue Code="C" Mass="103.009184505"/>
  <Residue Code="D" Mass="115.026943065"/>
  <Residue Code="E" Mass="129.042593135"/>
  <Residue Code="F" Mass="147.068413945"/>
  <Residue Code="G" Mass="57.021463735"/>
  ...
</MassTable>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/MassTable
MAY supply a *child* term of MS:1001354 (mass table options) one or more times
  e.g.: MS:1001346 (AAIndex mass table)
```

6.39 Element <FragmentTolerance>

Definition: The tolerance of the search given as a plus and minus value with units.

Type: ToleranceType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	The tolerance of the search given as a plus and minus value with units.

Example Context:

```
      <FragmentTolerance>
        <cvParam accession="MS:1001412" name="search tolerance plus value" cvRef="PSI-MS" value="0.9"
unitAccession="UO:0000221" unitName="dalton" unitCvRef="UO"/>
        <cvParam accession="MS:1001413" name="search tolerance minus value" cvRef="PSI-MS"
value="0.9" unitAccession="UO:0000221" unitName="dalton" unitCvRef="UO"/>
      </FragmentTolerance>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/FragmentTolerance
MUST supply term MS:1001412 (search tolerance plus value) only once
MUST supply term MS:1001413 (search tolerance minus value) only once
```

6.40 Element <ParentTolerance>

Definition: The tolerance of the search given as a plus and minus value with units.

Type: ToleranceType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	The tolerance of the search given as a plus and minus value with units.

Example Context:

```
      <ParentTolerance>
        <cvParam accession="MS:1001412" name="search tolerance plus value" cvRef="PSI-MS"
value="75.0" unitAccession="UO:0000169" unitName="parts per million" unitCvRef="UO"/>
        <cvParam accession="MS:1001413" name="search tolerance minus value" cvRef="PSI-MS"
value="75.0" unitAccession="UO:0000169" unitName="parts per million" unitCvRef="UO"/>
      </ParentTolerance>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/ParentTolerance
MUST supply term MS:1001412 (search tolerance plus value) only once
MUST supply term MS:1001413 (search tolerance minus value) only once
```

6.41 Element <Threshold>

Definition: The threshold(s) applied to determine that a result is significant. If multiple terms are used it is assumed that all conditions are satisfied by the passing results.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<Threshold>
  <cvParam accession="MS:1001316" name="mascot:SigThreshold"
cvRef="PSI-MS" value="0.05"/>
</Threshold>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/Threshold
MUST supply a *child* term of MS:1001302 (search engine specific input parameter) only once
  e.g.: MS:1001005 (sequest:CleavesAt)
  e.g.: MS:1001007 (sequest:OutputLines)
  e.g.: MS:1001009 (sequest:DescriptionLines)
  e.g.: MS:1001026 (sequest:NormalizeXCorrValues)
  e.g.: MS:1001028 (sequest:SequenceHeaderFilter)
  e.g.: MS:1001032 (sequest:SequencePartialFilter)
  e.g.: MS:1001037 (sequest:ShowFragmentIons)
  e.g.: MS:1001038 (sequest:Consensus)
  e.g.: MS:1001042 (sequest:LimitTo)
```

```
    e.g.: MS:1001046 (sequest:sort_by_dCn)
    et al.
  MUST supply a *child* term of MS:1001153 (search engine specific score) only once
    e.g.: MS:1001154 (sequest:probability)
    e.g.: MS:1001155 (sequest:xcorr)
    e.g.: MS:1001156 (sequest:deltacn)
    e.g.: MS:1001157 (sequest:sp)
    e.g.: MS:1001158 (sequest:Uniq)
    e.g.: MS:1001159 (sequest:expectation value)
    e.g.: MS:1001160 (sequest:sf)
    e.g.: MS:1001161 (sequest:matched ions)
    e.g.: MS:1001162 (sequest:total ions)
    e.g.: MS:1001163 (sequest:consensus score)
    et al.
  MUST supply term MS:1001494 (no threshold) only once
  MUST supply term MS:1001448 (pep:FDR threshold) only once
  Path /mzIdentML/AnalysisProtocolCollection/ProteinDetectionProtocol/Threshold
  MUST supply a *child* term of MS:1001302 (search engine specific input parameter) only once
    e.g.: MS:1001005 (sequest:CleavesAt)
    e.g.: MS:1001007 (sequest:OutputLines)
    e.g.: MS:1001009 (sequest:DescriptionLines)
    e.g.: MS:1001026 (sequest:NormalizeXCorrValues)
    e.g.: MS:1001028 (sequest:SequenceHeaderFilter)
    e.g.: MS:1001032 (sequest:SequencePartialFilter)
    e.g.: MS:1001037 (sequest:ShowFragmentIons)
    e.g.: MS:1001038 (sequest:Consensus)
    e.g.: MS:1001042 (sequest:LimitTo)
    e.g.: MS:1001046 (sequest:sort_by_dCn)
    et al.
  MUST supply a *child* term of MS:1001153 (search engine specific score) only once
    e.g.: MS:1001154 (sequest:probability)
    e.g.: MS:1001155 (sequest:xcorr)
    e.g.: MS:1001156 (sequest:deltacn)
    e.g.: MS:1001157 (sequest:sp)
    e.g.: MS:1001158 (sequest:Uniq)
    e.g.: MS:1001159 (sequest:expectation value)
    e.g.: MS:1001160 (sequest:sf)
    e.g.: MS:1001161 (sequest:matched ions)
    e.g.: MS:1001162 (sequest:total ions)
    e.g.: MS:1001163 (sequest:consensus score)
    et al.
  MUST supply term MS:1001447 (prot:FDR threshold) only once
  MUST supply term MS:1001494 (no threshold) only once
```

6.42 Element <DatabaseFilters>

Definition: The specification of filters applied to the database searched.

Type: DatabaseFiltersType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Filter	1	unbounded	The filter MUST include at least one of Include and Exclude. If both are used, it is assumed that inclusion is performed first.

Example Context:

```
<DatabaseFilters>
  <Filter>
    <FilterType>
      <cvParam accession="MS:1001020" name="DB filter taxonomy" cvRef="PSI-MS" />
    </FilterType>
    <Include>
      <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="33208"/>
  ...
</DatabaseFilters>
```

6.43 Element <DatabaseTranslation>

Definition: A specification of how a nucleic acid sequence database was translated for searching.

Type: DatabaseTranslationType

Attributes:

Attribute Name	Data Type	Use	Definition
frames	listOfAllowedFrames	optional	The frames in which the nucleic acid sequence has been translated as a space separated list

Subelements:

Subelement	minOccurs	maxOccurs	Definition

Name			
TranslationTable	1	unbounded	The table used to translate codons into nucleic acids e.g. by reference to the NCBI translation table.

Example Context:

```
<DatabaseTranslation frames="1 2 3 -1 -2 -3">
  <TranslationTable id="TT_1" name="Standard">
    <cvParam accession="MS:1001025" name="translation table" cvRef="PSI-MS"
value="FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG" />
    <cvParam accession="MS:1001410" name="translation start codons" cvRef="PSI-MS" value="---M---------
------M---------------M---------------------------" />
    <cvParam accession="MS:1001423" name="translation table description" cvRef="PSI-MS"
value="http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG1" />
  </TranslationTable>
  <TranslationTable id="TT_2" name="Vertebrate Mitochondrial">
  ...
</DatabaseTranslation>
```

6.44 Element <AnalysisParams>

Definition: The parameters and settings for the protein detection given as CV terms.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<AnalysisParams>
  <cvParam accession="MS:1001316" name="mascot:SigThreshold" cvRef="PSI-MS" value="0.05"/>
  <cvParam accession="MS:1001317" name="mascot:MaxProteinHits" cvRef="PSI-MS" value="Auto"/>
  <cvParam accession="MS:1001318" name="mascot:ProteinScoringMethod" cvRef="PSI-MS"
value="Standard"/>
  <cvParam accession="MS:1001319" name="mascot:MinMSMSThreshold" cvRef="PSI-MS" value="0"/>
  <cvParam accession="MS:1001320" name="mascot:ShowHomologousProteinsWithSamePeptides"
cvRef="PSI-MS" value="1"/>
  <cvParam accession="MS:1001321" name="mascot:ShowHomologousProteinsWithSubsetOfPeptides"
cvRef="PSI-MS" value="1"/>
  ...
</AnalysisParams>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/ProteinDetectionProtocol/AnalysisParams
MAY supply a *child* term of MS:1001302 (search engine specific input parameter) one or more times
  e.g.: MS:1001005 (sequest:CleavesAt)
  e.g.: MS:1001007 (sequest:OutputLines)
  e.g.: MS:1001009 (sequest:DescriptionLines)
  e.g.: MS:1001026 (sequest:NormalizeXCorrValues)
  e.g.: MS:1001028 (sequest:SequenceHeaderFilter)
  e.g.: MS:1001032 (sequest:SequencePartialFilter)
  e.g.: MS:1001037 (sequest:ShowFragmentIons)
  e.g.: MS:1001038 (sequest:Consensus)
  e.g.: MS:1001042 (sequest:LimitTo)
  e.g.: MS:1001046 (sequest:sort_by_dCn)
  et al.
MAY supply a *child* term of MS:1001194 (quality estimation with decoy database) one or more times
```

6.45 Element <SourceFile>

Definition: A file from which this mzIdentML instance was created.

Type: PSI-PI.analysis.search.SourceFileType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
location	xsd:anyURI	required	The location of the data file.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
externalFormatDocumentation	0	1	A URI to access documentation and tools to interpret the external format of the ExternalData instance. For example, XML Schema or static libraries (APIs) to access binary formats.
fileFormat	0	1	The format of the ExternalData file, for example "tiff" for image files.
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
      <SourceFile id="SF1" location="proteinscape://www.medizinisches-proteom-
center.de/PSServer/Project/Sample/Separation_1D_LC/Fraction_X/SpectraData/Results1">
        <fileFormat>
          <cvParam accession="MS:1001275" name="ProteinScape SearchEvent" cvRef="PSI-MS"/>
        </fileFormat>
      </SourceFile>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/Inputs/SourceFile
MAY supply a *child* term of MS:1000561 (data file checksum type) one or more times
  e.g.: MS:1000568 (MD5)
  e.g.: MS:1000569 (SHA-1)
```

6.46 Element <SpectraData>

Definition: A data set containing spectra data (consisting of one or more spectra).

Type: PSI-PI.spectra.SpectraDataType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
location	xsd:anyURI	required	The location of the data file.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
externalFormatDocumentation	0	1	A URI to access documentation and tools to interpret the external format of the ExternalData instance. For example, XML Schema or static libraries (APIs) to access binary formats.
fileFormat	0	1	The format of the ExternalData file, for example "tiff" for image files.
spectrumIDFormat	1	1	The format of the spectrum identifier within the source file.

Example Context:

```
<SpectraData location="file:///C:/DOCUME~1/DAVIDC~1.MAT/LOCALS~1/Temp/Dis83.tmp" id="SD_1">
  <fileFormat>
    <cvParam accession="MS:1001062" name="Mascot MGF file" cvRef="PSI-MS" />
  </fileFormat>
  <spectrumIDFormat>
    <cvParam accession="MS:1001528" name="Mascot query number" cvRef="PSI-MS" />
  </spectrumIDFormat>
</SpectraData>
```

6.47 Element <SpectrumIdentificationList>

Definition: Represents the set of all search results from SpectrumIdentification.

Type: PSI-PI.analysis.search.SpectrumIdentificationListType

Attributes:

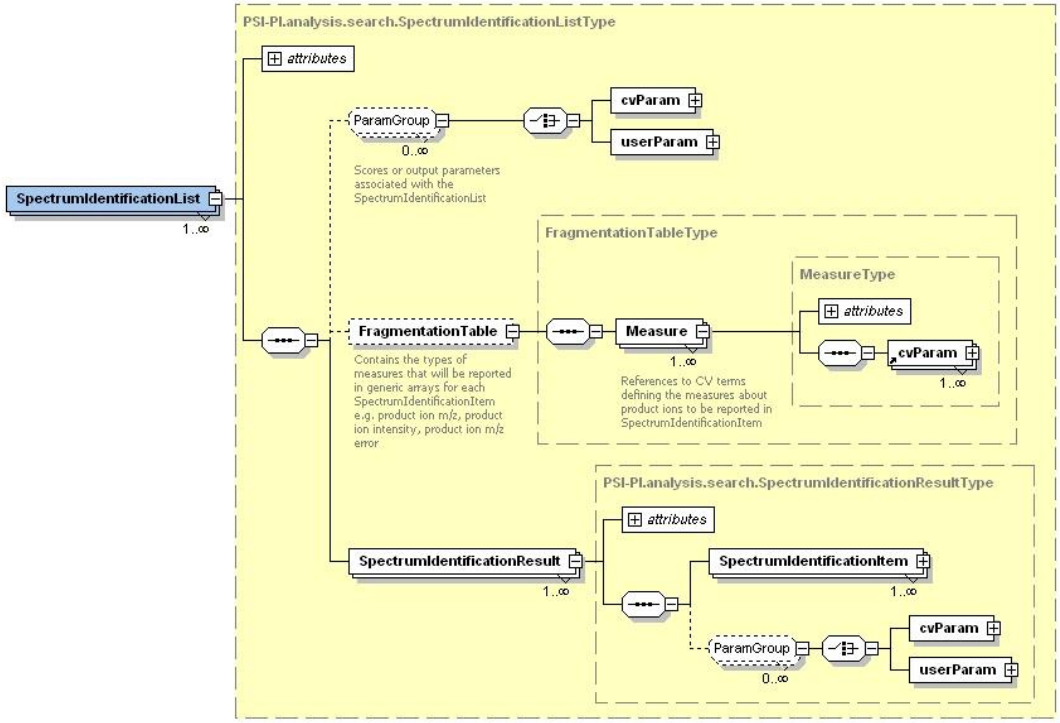
Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.

name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
numSequencesSearched	xsd:long	optional	This value should be provided unless a de novo search has been performed.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
FragmentationTable	0	1	Contains the types of measures that will be reported in generic arrays for each SpectrumIdentificationItem e.g. product ion m/z, product ion intensity, product ion m/z error
SpectrumIdentificationResult	1	unbounded	All identifications made from searching one spectrum. For PMF data, all peptide identifications will be listed underneath as SpectrumIdentificationItems. For MS/MS data, there will be ranked SpectrumIdentificationItems corresponding to possible different peptide IDs.

Graphical Context:



Example Context:

```
<SpectrumIdentificationList id="SIL_1" numSequencesSearched="257964">
  <SpectrumIdentificationResult id="SIR_1" spectrumID="1" SpectraData_ref="SD_1">
    <SpectrumIdentificationItem id="SII_1_1"  calculatedMassToCharge="1107.534897" chargeState="1"
experimentalMassToCharge="1108.53" Peptide_ref="peptide_1_1" rank="0" passThreshold="true">
      <PeptideEvidence id="PE_1_1_UVRB_THET8" start="542" end="550" pre="R" post="V"
missedCleavages="0" isDecoy="false" DBSequence_Ref="DBSeq_UVRB_THET8" />
    </SpectrumIdentificationItem>
    <SpectrumIdentificationItem id="SII_1_8"  calculatedMassToCharge="1107.617584" chargeState="1"
experimentalMassToCharge="1108.53" Peptide_ref="peptide_1_8" rank="0" passThreshold="true">
    </SpectrumIdentificationItem>
  ...
</SpectrumIdentificationList>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/SpectrumIdentificationList
MAY supply a *child* term of MS:1001184 (search statistics) one or more times
  e.g.: MS:1001035 (date / time search performed)
  e.g.: MS:1001036 (search time taken)
  e.g.: MS:1001177 (number of molecular hypothesis considered)
```

6.48 Element <ProteinDetectionList>

Definition: The protein list resulting from a protein detection process.

Type: PSI-PI.analysis.process.ProteinDetectionListType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
ProteinAmbiguityGroup	0	unbounded	A set of logically related results from a protein detection, for example to represent conflicting assignments of peptides to proteins.

Example Context:

```
<ProteinDetectionList id="PDL_1">
  <ProteinAmbiguityGroup id="PAG_hit_1" >
    <ProteinDetectionHypothesis id="PDH_MYG_EQUBU" DBSequence_ref="DBSeq_MYG_EQUBU"
passThreshold="true">
      <PeptideHypothesis  PeptideEvidence_Ref="PE_1_1_MYG_EQUBU" />
      <cvParam accession="MS:1001171" name="mascot:score" cvRef="PSI-MS" value="405.72" />
      <cvParam accession="MS:1001093" name="sequence coverage" cvRef="PSI-MS" value="99" />
      <cvParam accession="MS:1001097" name="distinct peptide sequences" cvRef="PSI-MS" value="1" />
  ...
</ProteinDetectionList>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/ProteinDetectionList
MAY supply a *child* term of MS:1001184 (search statistics) one or more times
  e.g.: MS:1001035 (date / time search performed)
  e.g.: MS:1001036 (search time taken)
  e.g.: MS:1001177 (number of molecular hypothesis considered)
```

6.49 Element <role>

Definition: The roles (lab equipment sales, contractor, etc.) the Contact fills.

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<role>
    <cvParam accession="MS:1001267" name="software vendor" cvRef="PSI-MS"/>
</role>
```

6.50 Element <SearchModification>

Definition: Specification of a search modification as parameter for a spectra search. Contains the name of the modification, the mass, the specificity and whether it is a static modification.

Type: PSI-PI.analysis.search.SearchModificationType

Attributes:

Attribute Name	Data Type	Use	Definition
fixedMod	xsd:boolean	required	True, if the modification is static (i.e. occurs always).

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
ModParam	1	1	The modification searched for, sourced from e.g. UniMod and the mass delta
SpecificityRules	0	1	The specificity rules of the searched modification including for

			example the probability of a modification's presence or peptide or protein termini. Standard fixed or variable status should be provided by the attribute fixedMod.

Example Context:

```
<SearchModification fixedMod="false" >
  <ModParam massDelta="127.063324" residues="">
    <cvParam accession="UNIMOD:29" name="SMA" cvRef="UNIMOD"/>
  </ModParam>
  <SpecificityRules>
    <cvParam accession="MS:1001189" cvRef="PSI-MS" name="modification specificity N-term"/>
  </SpecificityRules>
  ...
</SearchModification>
```

6.51 Element <Enzyme>

Definition: The details of an individual cleavage enzyme should be provided by giving a regular expression or a CV term if a "standard" enzyme cleavage has been performed.

Type: PSI-PI.analysis.search.EnzymeType

Attributes:

Attribute Name	Data Type	Use	Definition
CTermGain	xsd:restriction base="xsd:string"	optional	Element formula gained at CTerm.
NTermGain	xsd:restriction base="xsd:string"	optional	Element formula gained at NTerm.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
minDistance	xsd:restriction base="xsd:int"	optional	Minimal distance for another cleavage (minimum: 1).
missedCleavages	xsd:int	optional	The number of missed cleavage sites allowed by the search. The attribute MUST be provided if an enzyme has been used.
semiSpecific	xsd:boolean	optional	Set to true if the enzyme cleaves semi-specifically (i.e. one terminus MUST cleave according to the rules, the other can cleave at any residue), false if the enzyme cleavage is assumed to be specific to both termini (accepting for any missed cleavages).

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SiteRegexp	0	1	Regular expression for specifying the enzyme cleavage site.
EnzymeName	0	1	The name of the enzyme from a CV.

Example Context:

```
<Enzyme id="ENZ_1" CTermGain="OH" NTermGain="H" missedCleavages="1" semiSpecific="0">
  <SiteRegexp><![CDATA[(?<=[KR])(?!P)]]></SiteRegexp>
  <EnzymeName>
    <cvParam accession="MS:1001251" name="Trypsin" cvRef="PSI-MS" />
  </EnzymeName>
</Enzyme>
```

6.52 Element <Residue>

Definition: The specification of a single residue within the mass table.

Type: ResidueType

Attributes:

Attribute Name	Data Type	Use	Definition
Code	chars	required	The single letter code for the residue.
Mass	xsd:float	required	The residue mass in Daltons (not including any fixed modifications).

Subelements: none

Example

```
<Residue Code="C" Mass="103.009184505"/>
```

Context:

6.53 Element <AmbiguousResidue>

Definition: Ambiguous residues e.g. X can be specified by the Code attribute and a set of parameters for example giving the different masses that will be used in the search.

Type: AmbiguousResidueType

Attributes:

Attribute Name	Data Type	Use	Definition
Code	chars	required	The single letter code of the ambiguous residue e.g. X.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<AmbiguousResidue Code="X">
  <cvParam accession="MS:1001360" name="alternate single letter codes" cvRef="PSI-MS" value="A C D E F G H I K L M N O P Q R S T U V W Y"/>
</AmbiguousResidue>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/MassTable/AmbiguousResidue
MAY supply a *child* term of MS:1001359 (ambiguous residues) one or more times
  e.g.: MS:1001360 (alternate single letter codes)
  e.g.: MS:1001361 (alternate mass)
```

6.54 Element <Filter>

Definition: The filter MUST include at least one of Include and Exclude. If both are used, it is assumed that inclusion is performed first.

Type: FilterType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
FilterType	1	1	The type of filter e.g. database taxonomy filter, pi filter, mw filter
Include	0	1	All sequences fulfilling the specifed criteria are included.
Exclude	0	1	All sequences fulfilling the specifed criteria are excluded.

Example Context:

```
<Filter>
  <FilterType>
    <cvParam accession="MS:1001020" name="DB filter taxonomy" cvRef="PSI-MS" />
  </FilterType>
  <Include>
    <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="33208"/>
  </Include>
  ...
</Filter>
```

6.55 Element <TranslationTable>

Definition: The table used to translate codons into nucleic acids e.g. by reference to the NCBI translation table.

Type: TranslationTableType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<TranslationTable id="TT_4" name="Mold Mitochondrial; Protozoan Mitochondrial; Coelenterate
Mitochondrial; Mycoplasma; Spiroplasma">
  <cvParam accession="MS:1001025" name="translation table" cvRef="PSI-MS"
value="FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG" />
  <cvParam accession="MS:1001410" name="translation start codons" cvRef="PSI-MS" value="--MM----------
--M----------MMMM----------M----------" />
  <cvParam accession="MS:1001423" name="translation table description" cvRef="PSI-MS"
value="http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG4" />
</TranslationTable>
```

cvParam Mapping Rules:

```
Path
/mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/DatabaseTranslation/TranslationTable
MUST supply term MS:1001410 (translation start codons) only once
MUST supply term MS:1001025 (translation table) only once
MUST supply term MS:1001423 (translation table description) only once
```

6.56 Element <externalFormatDocumentation>

Definition: A URI to access documentation and tools to interpret the external format of the ExternalData instance. For example, XML Schema or static libraries (APIs) to access binary formats.

Type: xsd:anyURI

Attributes: none

Subelements: none

Example Context:

```
<externalFormatDocumentation>http://www.matrixscience.com/help/data_file_help.html</externalFormatDocumentation>
```

6.57 Element <fileFormat>

Definition: The format of the ExternalData file, for example "tiff" for image files.

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<fileFormat>
        <cvParam accession="MS:1001275" name="ProteinScape SearchEvent" cvRef="PSI-MS"/>
</fileFormat>
```

6.58 Element <FragmentationTable>

Definition: Contains the types of measures that will be reported in generic arrays for each SpectrumIdentificationItem e.g. product ion m/z, product ion intensity, product ion m/z error

Type: FragmentationTableType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
Measure	1	unbounded	References to CV terms defining the measures about product ions to be reported in SpectrumIdentificationItem

Example Context:

```
    <FragmentationTable>
        <Measure id="m_mz">
          <cvParam cvRef="PSI-MS" accession="MS:1001225" name="product ion m/z"/>
        </Measure>
        <Measure id="m_intensity">
          <cvParam cvRef="PSI-MS" accession="MS:1001226" name="product ion intensity"/>
        </Measure>
  ...
</FragmentationTable>
```

6.59 Element <SpectrumIdentificationResult>

Definition: All identifications made from searching one spectrum. For PMF data, all peptide identifications will be listed underneath as SpectrumIdentificationItems. For MS/MS data, there will be ranked SpectrumIdentificationItems corresponding to possible different peptide IDs.

Type: PSI-PI.analysis.search.SpectrumIdentificationResultType

Attributes:

Attribute Name	Data	Use	Definition

	Type		
SpectraData_ref	xsd:string	required	A reference to a spectra data set (e.g. a spectra file).
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
spectrumID	xsd:string	required	The locally unique id for the spectrum in the spectra data set specified by SpectraData_ref. External guidelines are provided on the use of consistent identifiers for spectra in different external formats.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
SpectrumIdentificationItem	1	unbounded	An identification of a single (poly)peptide, resulting from querying an input spectra, along with the set of confidence values for that identification. PeptideEvidence elements should be given for all mappings of the corresponding Peptide sequence within protein sequences.
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<SpectrumIdentificationResult id="SIR_1" spectrumID="query_1" SpectraData_ref="SD_1">
  <SpectrumIdentificationItem id="SII_1_1"  calculatedMassToCharge="670.86261" chargeState="2"
experimentalMassToCharge="671.9" Peptide_ref="peptide_1_1" rank="1" passThreshold="true">
    <PeptideEvidence id="PE_1_1_HSP70_ECHGR" start="161" end="172" pre="K" post="I" missedCleavages="0"
isDecoy="false" DBSequence_Ref="DBSeq_HSP70_ECHGR" />
    <PeptideEvidence id="PE_1_1_HSP70_ONCMY" start="160" end="171" pre="K" post="L" missedCleavages="0"
isDecoy="false" DBSequence_Ref="DBSeq_HSP70_ONCMY" />
    <PeptideEvidence id="PE_1_1_HSP7C_ICTPU" start="160" end="171" pre="K" post="I" missedCleavages="0"
isDecoy="false" DBSequence_Ref="DBSeq_HSP7C_ICTPU" />
    <PeptideEvidence id="PE_1_1_HSP7C_ORYLA" start="160" end="171" pre="K" post="I" missedCleavages="0"
isDecoy="false" DBSequence_Ref="DBSeq_HSP7C_ORYLA" />
    <PeptideEvidence id="PE_1_1_HSP7D_MANSE" start="160" end="171" pre="K" post="I" missedCleavages="0"
isDecoy="false" DBSequence_Ref="DBSeq_HSP7D_MANSE" />
  ...
</SpectrumIdentificationResult>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/SpectrumIdentificationList/SpectrumIdentificationResult
MAY supply a *child* term of MS:1001405 (spectrum identification result details) one or more times
  e.g.: MS:1000796 (spectrum title)
  e.g.: MS:1000797 (peak list scans)
  e.g.: MS:1000798 (peak list raw scans)
  e.g.: MS:1001030 (number of peptide seqs compared to each spectrum)
  e.g.: MS:1001035 (date / time search performed)
  e.g.: MS:1001036 (search time taken)
  e.g.: MS:1001088 (protein description)
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001093 (sequence coverage)
  e.g.: MS:1001097 (distinct peptide sequences)
  et al.
```

6.60 Element <ProteinAmbiguityGroup>

Definition: A set of logically related results from a protein detection, for example to represent conflicting assignments of peptides to proteins.

Type: PSI-PI.analysis.process.ProteinAmbiguityGroupType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition

ProteinDetectionHypothesis	1	unbounded	A single result of the ProteinDetection analysis (i.e. a protein).
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<ProteinAmbiguityGroup id="PAG_hit_1" >
  <ProteinDetectionHypothesis id="PDH_HSP7D_MANSE" DBSequence_ref="DBSeq_HSP7D_MANSE"
passThreshold="true">
    <PeptideHypothesis  PeptideEvidence_Ref="PE_1_1_HSP7D_MANSE" />
    <PeptideHypothesis  PeptideEvidence_Ref="PE_3_1_HSP7D_MANSE" />
    <cvParam accession="MS:1001171" name="mascot:score" cvRef="PSI-MS" value="104.854382332144" />
    <cvParam accession="MS:1001093" name="sequence coverage" cvRef="PSI-MS" value="4" />
    <cvParam accession="MS:1001097" name="distinct peptide sequences" cvRef="PSI-MS" value="2" />
  ...
</ProteinAmbiguityGroup>
```

6.61 Element <cvParam>

Definition: A single entry from an ontology or a controlled vocabulary.

Type: FuGE.Common.Ontology.cvParamType

Attributes:

Attribute Name	Data Type	Use	Definition
accession	xsd:string	required	The accession or ID number of this CV term in the source CV.
cvRef	xsd:string	required	A reference to the cv element from which this term originates.
name	xsd:string	required	The name of the parameter.
unitAccession	xsd:string	optional	An accession number identifying the unit within the OBO foundry Unit CV.
unitCvRef	xsd:string	optional	If a unit term is referenced, this attribute MUST refer to the CV 'id' attribute defined in the cvList in this file.
unitName	xsd:string	optional	The name of the unit.
value	xsd:string	optional	The user-entered value of the parameter.

Subelements: none

Example Context:

```
      <cvParam accession="MS:1001088" name="protein description" cvRef="PSI-MS"
value="">IPI:IPI00414676.5|SWISS-
PROT:P08238|TREMBL:Q5T9W7;Q6PK50;Q9H6X9|ENSEMBL:ENSP00000325875|REFSEQ:NP_031381|H-
INV:HIT000008644;HIT000032091;HIT000034201;HIT000035963;HIT000036733;HIT000049765;
HIT000057726|VEGA:OTTHUMP00000016517;OTTHUMP00000016518;OTTHUMP00000016519 Tax_Id=9606 Heat shock
protein HSP 90-beta"/>
```

6.62 Element <userParam>

Definition: A single user-defined parameter.

Type: FuGE.Common.Ontology.userParamType

Attributes:

Attribute Name	Data Type	Use	Definition
name	xsd:string	required	The name of the parameter.
unitAccession	xsd:string	optional	An accession number identifying the unit within the OBO foundry Unit CV.
unitCvRef	xsd:string	optional	If a unit term is referenced, this attribute MUST refer to the CV 'id' attribute defined in the cvList in this file.
unitName	xsd:string	optional	The name of the unit.
value	xsd:string	optional	The user-entered value of the parameter.

Subelements: none

Example Context:

```
<userParam name="Mascot User Comment" value="Example Mascot MS-MS search for PSI mzIdentML"/>
```

6.63 Element <ModParam>

Definition: The modification searched for, sourced from e.g. UniMod and the mass delta

Type: PSI-PI.polypeptide.ModParamType

Attributes:

Attribute Name	Data Type	Use	Definition
massDelta	xsd:float	required	The mass delta of the searched modification in Daltons
residues	listOfChars	required	The residue(s) searched with the specified modification

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<ModParam residues="M" massDelta="15.994914622">
  <cvParam accession="UNIMOD:35" name="Oxidation" cvRef="UNIMOD"/>
</ModParam>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/ModificationParams/
SearchModification/ModParam
MUST supply a *child* term of UNIMOD:0 (UNIMOD root) one or more times
MUST supply a *child* term of MS:1001471 (peptide modification details) one or more times
  e.g.: MS:1001460 (unknown modification)
  e.g.: MS:1001524 (fragment neutral loss)
  e.g.: MS:1001525 (precursor neutral loss )
MUST supply a *child* term of MOD:00000 (protein modification) one or more times
```

6.64 Element <SpecificityRules>

Definition: The specificity rules of the searched modification including for example the probability of a modification's presence or peptide or protein termini. Standard fixed or variable status should be provided by the attribute fixedMod.

Type: SpecificityRulesType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<SpecificityRules>
  <cvParam accession="MS:1001189" cvRef="PSI-MS" name="modification specificity N-term"/>
</SpecificityRules>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/ModificationParams/
SearchModification/SpecificityRules
MUST supply a *child* term of MS:1001056 (modification specificity rule) one or more times
  e.g.: MS:1001189 (modification specificity N-term)
  e.g.: MS:1001190 (modification specificity C-term)
```

6.65 Element <SiteRegexp>

Definition: Regular expression for specifying the enzyme cleavage site.

Type: PSI-PI.analysis.search.SiteRegexpType

Attributes: none

Subelements: none

Example Context:

```
<SiteRegexp><![CDATA[(?<=[KR])(?!P)]]></SiteRegexp>
```

6.66 Element <EnzymeName>

Definition: The name of the enzyme from a CV.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<EnzymeName>
  <cvParam accession="MS:1001251" name="Trypsin" cvRef="PSI-MS"/>
</EnzymeName>
```

cvParam Mapping Rules:

```
Path /mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/Enzymes/Enzyme/EnzymeName
MAY supply a *child* term of MS:1001045 (cleavage agent name) only once
  e.g.: MS:1001091 (NoEnzyme)
  e.g.: MS:1001251 (Trypsin)
  e.g.: MS:1001303 (Arg-C)
  e.g.: MS:1001304 (Asp-N)
  e.g.: MS:1001305 (Asp-N_ambic)
  e.g.: MS:1001306 (Chymotrypsin)
  e.g.: MS:1001307 (CNBr)
  e.g.: MS:1001308 (Formic_acid)
  e.g.: MS:1001309 (Lys-C)
  e.g.: MS:1001310 (Lys-C/P)
  et al.
```

6.67 Element <FilterType>

Definition: The type of filter e.g. database taxonomy filter, pi filter, mw filter

Type: ParamType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<FilterType>
  <cvParam accession="MS:1001020" name="DB filter taxonomy" cvRef="PSI-MS" />
</FilterType>
```

cvParam Mapping Rules:

```
Path
/mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/DatabaseFilters/Filter/FilterType
MUST supply a *child* term of MS:1001511 (Sequence database filter types) one or more times
  e.g.: MS:1001020 (DB filter taxonomy)
  e.g.: MS:1001021 (DB filter on accession numbers)
  e.g.: MS:1001022 (DB MW filter)
  e.g.: MS:1001023 (DB PI filter)
  e.g.: MS:1001027 (DB filter on sequence pattern)
```

6.68 Element <Include>

Definition: All sequences fulfilling the specifed criteria are included.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<Include>
  <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="33208"/>
</Include>
```

cvParam Mapping Rules:

```
Path
/mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/DatabaseFilters/Filter/Include
MAY supply a *child* term of MS:1001512 (Sequence database filters) one or more times
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001201 (DB MW filter maximum)
  e.g.: MS:1001202 (DB MW filter minimum)
  e.g.: MS:1001203 (DB PI filter maximum)
  e.g.: MS:1001204 (DB PI filter minimum)
  e.g.: MS:1001467 (taxonomy: NCBI TaxID)
  e.g.: MS:1001468 (taxonomy: common name)
  e.g.: MS:1001469 (taxonomy: scientific name)
  e.g.: MS:1001470 (taxonomy: Swiss-Prot ID)
  e.g.: MS:1001513 (DB sequence filter pattern)
  et al.
```

6.69 Element <Exclude>

Definition: All sequences fulfilling the specifed criteria are excluded.

Type: ParamListType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<Exclude>
  <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="45251"/>
</Exclude>
```

cvParam Mapping Rules:

```
Path
/mzIdentML/AnalysisProtocolCollection/SpectrumIdentificationProtocol/DatabaseFilters/Filter/Exclude
MAY supply a *child* term of MS:1001512 (Sequence database filters) one or more times
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001201 (DB MW filter maximum)
  e.g.: MS:1001202 (DB MW filter minimum)
  e.g.: MS:1001203 (DB PI filter maximum)
  e.g.: MS:1001204 (DB PI filter minimum)
  e.g.: MS:1001467 (taxonomy: NCBI TaxID)
  e.g.: MS:1001468 (taxonomy: common name)
  e.g.: MS:1001469 (taxonomy: scientific name)
  e.g.: MS:1001470 (taxonomy: Swiss-Prot ID)
  e.g.: MS:1001513 (DB sequence filter pattern)
  et al.
```

6.70 Element <Measure>

Definition: References to CV terms defining the measures about product ions to be reported in SpectrumIdentificationItem

Type: MeasureType

Attributes:

Attribute Name	Data Type	Use	Definition
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	A single entry from an ontology or a controlled vocabulary.

Example Context:

```
<Measure id="m_error">
  <cvParam cvRef="PSI-MS" accession="MS:1001227" name="product ion m/z error"
unitAccession="MS:1000040" unitName="m/z" unitCvRef="PSI-MS"/>
</Measure>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/SpectrumIdentificationList/
FragmentationTable/Measure
MUST supply a *child* term of  (???) one or more times
MUST supply term MS:1001226 (product ion intensity) only once
MUST supply term MS:1001225 (product ion m/z) only once
MUST supply term MS:1001227 (product ion m/z error) only once
```

6.71 Element <SpectrumIdentificationItem>

Definition: An identification of a single (poly)peptide, resulting from querying an input spectra, along with the set of confidence values for that identification. PeptideEvidence elements should be given for all mappings of the corresponding Peptide sequence within protein sequences.

Type: PSI-PI.analysis.search.SpectrumIdentificationItemType

Attributes:

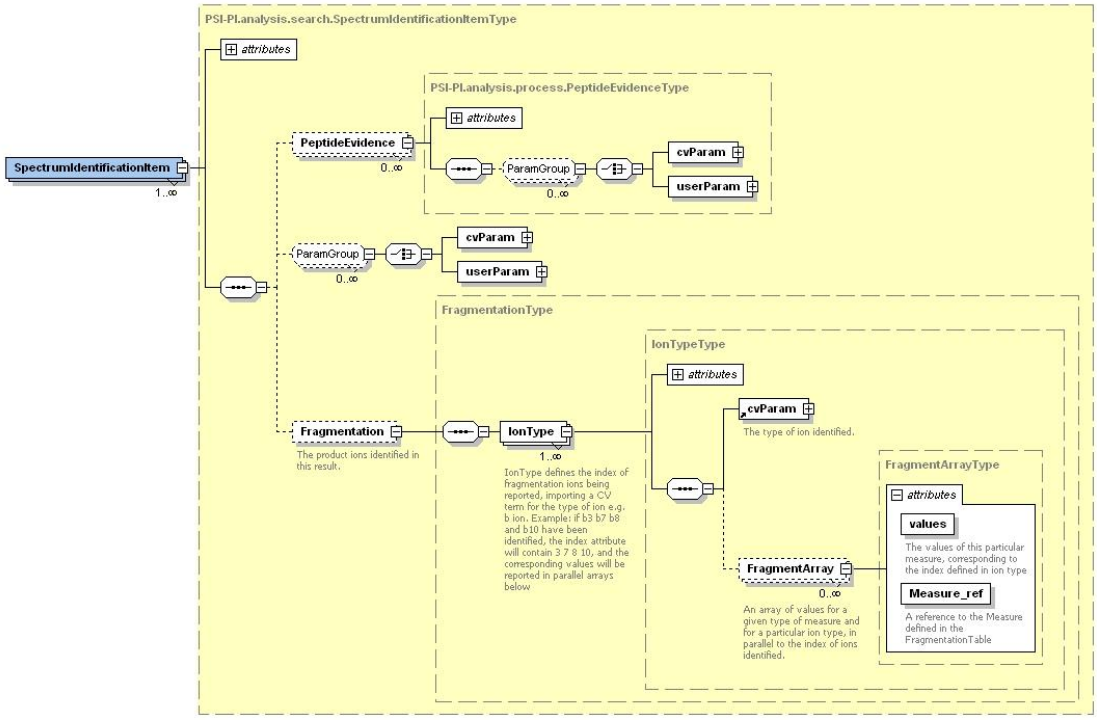
Attribute Name	Data Type	Use	Definition

MassTable_ref	xsd:string	optional	A reference should be given to the MassTable used to calculate the sequenceMass only if more than one MassTable has been given
Peptide_ref	xsd:string	optional	A reference to the identified (poly)peptide sequence in the Peptide element.
Sample_ref	xsd:string	optional	A reference should be provided to link the SpectrumIdentificationItem to a Sample if more than one sample has been described in the AnalysisSampleCollection.
calculatedMassToCharge	xsd:double	optional	The theoretical mass-to-charge value calculated for the peptide in Daltons / charge.
calculatedPI	xsd:float	optional	The calculated isoelectric point of the (poly)peptide, with relevant modifications included. Do not supply this value if the PI cannot be calcuated properly.
chargeState	xsd:int	required	The charge state of the identified peptide.
experimentalMassToCharge	xsd:double	required	The mass-to-charge value measured in the experiment in Daltons / charge.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
passThreshold	xsd:boolean	required	Set to true if the producers of the file has deemed that the identification has passed a given threshold or been validated as correct. If no such threshold has been set, value of true should be given for all results.
rank	xsd:int	required	For an MS/MS result set, this is the rank of the identification quality as scored by the search engine. 1 is the top rank. If multiple identifications have the same top score, they should all be assigned rank =1. For PMF data, the rank attribute may be meaningless and values of rank = 0 should be given.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
PeptideEvidence	0	unbounded	PeptideEvidence maps a spectrum identification to DBSequence in which such a peptide is located.
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
Fragmentation	0	1	The product ions identified in this result.

Graphical Context:



Example Context:

```
<SpectrumIdentificationItem id="SII_10_18"  calculatedMassToCharge="2040.993866" chargeState="1"
experimentalMassToCharge="2041.99" Peptide_ref="peptide_10_18" rank="0" passThreshold="true">
  <PeptideEvidence id="PE_10_18_VP3_BTV10" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV10" />
  <PeptideEvidence id="PE_10_18_VP3_BTV11" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV11" />
  <PeptideEvidence id="PE_10_18_VP3_BTV17" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV17" />
  <PeptideEvidence id="PE_10_18_VP3_BTV1S" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV1S" />
  <PeptideEvidence id="PE_10_18_VP3_BTV1A" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV1A" />
  <PeptideEvidence id="PE_10_18_VP3_BTV13" start="152" end="169" pre="R" post="N" missedCleavages="1"
isDecoy="false" DBSequence_Ref="DBSeq_VP3_BTV13" />
  ...
</SpectrumIdentificationItem>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/SpectrumIdentificationList/SpectrumIdentificationResult/
SpectrumIdentificationItem
MAY supply a *child* term of MS:1001405 (spectrum identification result details) one or more times
  e.g.: MS:1000796 (spectrum title)
  e.g.: MS:1000797 (peak list scans)
  e.g.: MS:1000798 (peak list raw scans)
  e.g.: MS:1001030 (number of peptide seqs compared to each spectrum)
  e.g.: MS:1001035 (date / time search performed)
  e.g.: MS:1001036 (search time taken)
  e.g.: MS:1001088 (protein description)
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001093 (sequence coverage)
  e.g.: MS:1001097 (distinct peptide sequences)
  et al.
```

6.72 Element <ProteinDetectionHypothesis>

Definition: A single result of the ProteinDetection analysis (i.e. a protein).

Type: PSI-PI.analysis.process.ProteinDetectionHypothesisType

Attributes:

Attribute Name	Data Type	Use	Definition
DBSequence_ref	xsd:string	optional	A reference to the corresponding DBSequence entry. This is optional and redundant, because the PeptideEvidence elements referenced from here also map to the DBSequence.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.

passThreshold	xsd:boolean	required	Set to true if the producers of the file has deemed that the ProteinDetectionHypothesis has passed a given threshold or been validated as correct. If no such threshold has been set, value of true should be given for all results.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
PeptideHypothesis	1	unbounded	Peptide evidence on which this ProteinHypothesis is based by reference to a PeptideEvidence element in a SpectrumIdentificationItem.
cvParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	0	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
          <ProteinDetectionHypothesis id="id_prot3" passThreshold="false">
            <PeptideHypothesis PeptideEvidence_Ref="PE1_SEQ_spec15_pep1"/>
            <PeptideHypothesis PeptideEvidence_Ref="PE1_SEQ_spec20_pep1"/>
            <cvParam accession="MS:1001093" name="sequence coverage" cvRef="PSI-MS" value="0.59"/>
            <cvParam accession="MS:1001301" name="protein rank" cvRef="PSI-MS" value="3"/>
            <cvParam accession="MS:1001097" name="distinct peptide sequences" cvRef="PSI-MS" value="2"/>
            <cvParam accession="MS:1001250" name="local FDR" cvRef="PSI-MS" value="33.33"
unitAccession="UO:0000187" unitName="percent" unitCvRef="UO"/>
  ...
</ProteinDetectionHypothesis>
```

cvParam Mapping Rules:

```
Path
/mzIdentML/DataCollection/AnalysisData/ProteinDetectionList/ProteinAmbiguityGroup/ProteinDetectionHypothesis
MAY supply a *child* term of  (???) one or more times
MAY supply a *child* term of MS:1001153 (search engine specific score) one or more times
  e.g.: MS:1001154 (sequest:probability)
  e.g.: MS:1001155 (sequest:xcorr)
  e.g.: MS:1001156 (sequest:deltacn)
  e.g.: MS:1001157 (sequest:sp)
  e.g.: MS:1001158 (sequest:Uniq)
  e.g.: MS:1001159 (sequest:expectation value)
  e.g.: MS:1001160 (sequest:sf)
  e.g.: MS:1001161 (sequest:matched ions)
  e.g.: MS:1001162 (sequest:total ions)
  e.g.: MS:1001163 (sequest:consensus score)
  et al.
MAY supply a *child* term of MS:1001060 (quality estimation method details) one or more times
  e.g.: MS:1001058 (quality estimation by manual validation)
  e.g.: MS:1001194 (quality estimation with decoy database)
  e.g.: MS:1001447 (prot:FDR threshold)
  e.g.: MS:1001448 (pep:FDR threshold)
  e.g.: MS:1001454 (quality estimation with implicite decoy sequences)
  e.g.: MS:1001494 (no threshold)
MAY supply a *child* term of MS:1001085 (protein result details) one or more times
  e.g.: MS:1001088 (protein description)
  e.g.: MS:1001090 (taxonomy nomenclature)
  e.g.: MS:1001093 (sequence coverage)
  e.g.: MS:1001097 (distinct peptide sequences)
  e.g.: MS:1001098 (confident distinct peptide sequences)
  e.g.: MS:1001099 (confident peptide qualification)
  e.g.: MS:1001100 (confident peptide)
  e.g.: MS:1001101 (protein group/subset relationship)
  e.g.: MS:1001125 (manual validation)
  e.g.: MS:1001157 (sequest:sp)
  et al.
```

6.73 Element <PeptideEvidence>

Definition: PeptideEvidence maps a spectrum identification to DBSequence in which such a peptide is located.

Type: PSI-PI.analysis.process.PeptideEvidenceType

Attributes:

Attribute Name	Data Type	Use	Definition
DBSequence_Ref	xsd:string	required	A reference to the sequence from which this identification has been made.
TranslationTable_ref	xsd:string	optional	A reference to the translation table used if this is PeptideEvidence derived from nucleic acid sequence
end	xsd:int	optional	The index position of the last amino acid of the

			peptide inside the protein sequence, where the first amino acid of the protein sequence is position 1.
frame	psi-pi:allowed_frames	optional	The translation frame of this sequence if this is PeptideEvidence derived from nucleic acid sequence
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
isDecoy	xsd:boolean	optional	Set to true if the peptide is matched to a decoy sequence.
missedCleavages	xsd:int	optional	Number of missed cleavage sites (not required if no enzyme has been used).
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
post	xsd:restriction base="xsd:string"	optional	Post flanking residue. If the peptide is C-terminal, post="-" and not post="". If for any reason it is unknown (e.g. denovo), post="?" should be used.
pre	xsd:restriction base="xsd:string"	optional	Previous flanking residue. If the peptide is N-terminal, pre="-" and not pre="". If for any reason it is unknown (e.g. denovo), pre="?" should be used.
start	xsd:int	optional	Start position of the peptide inside the protein sequence, where the first amino acid of the protein sequence is position 1.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	unbounded	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<PeptideEvidence id="PE_4_1_gi|152812279" start="130" end="139" pre="R" post="V" missedCleavages="0"
TranslationTable_ref="TT_1" frame="2" isDecoy="false" DBSequence_Ref="DBSeq_gi|152812279" />
```

6.74 Element <Fragmentation>

Definition: The product ions identified in this result.

Type: FragmentationType

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
IonType	1	unbounded	IonType defines the index of fragmentation ions being reported, importing a CV term for the type of ion e.g. b ion. Example: if b3 b7 b8 and b10 have been identified, the index attribute will contain 3 7 8 10, and the corresponding values will be reported in parallel arrays below

Example Context:

```
<Fragmentation>
  <IonType index="5 6 7 11 12 15 17 21 22 23 24 26 27 30 31 33 34 35 38 39 40 41 42 43 44 45 46 47 48
50 51 52 53 54 55 56 59 61 62 65 69 73 76 78 83 86 96" charge="1">
    <cvParam cvRef="PSI-MS" accession="MS:1001231" name="frag: c ion"/>
    <FragmentArray values="447.216737 576.26154 762.38947 1230.608358 1344.651696 1686.8204 1913.9818
2342.1727 2413.208 2470.2302 2607.2849 2792.370851 2921.413169 3246.646488 3402.747973 3662.8968
3763.9465 3820.9689 4184.122757 4285.1695 4398.254701 4527.295803 4655.3939 4802.459719 4917.4869
5045.580864 5192.648576 5320.7443 5457.7927 5698.9794 5800.027 5929.0677 6000.106688 6129.151636
6260.190033 6388.2896 6675.395079 6903.506231 7031.607787 7353.773068 7766.0424 8108.2303 8391.422
8647.6118 9235.887 9549.050832 10610.644 " Measure_ref="m_mz"/>
    <FragmentArray values="4380000 854900 7506000 30210000 12170000 11670000 9618000 10810000 8991000
15620000 5489000 13310000 21520000 31540000 19790000 5384000 9486000 12940000 28610000 8175000 24870000
46070000 10500000 46930000 19710000 22930000 31270000 10140000 5954000 7813000 17080000 9987000
16830000 35420000 14530000 5249000 56680000 13460000 14780000 17410000 5084000 8978000 8186000 7804000
7299000 22380000 6606000" Measure_ref="m_intensity"/>
    <FragmentArray values="-0.0031 -0.0008 0.0478 -0.0030 -0.0026 -0.0030 -0.0050 -0.0048 -0.0067 -
```

```
0.0059 -0.0101 -0.0042 -0.0045 -0.0077 -0.0073 -0.0110 -0.0090 -0.0080 -0.0084 -0.0094 -0.0082 -0.0097
-0.0066 -0.0092 -0.0090 -0.0100 -0.0107 -0.0099 -0.0204 -0.0127 -0.0128 -0.0147 -0.0128 -0.0105 -0.0126
-0.0080 -0.0142 -0.0141 -0.0075 -0.0175 -0.0168 -0.0192 -0.0171 -0.0172 -0.0189 -0.0188 -0.0238"
Measure_ref="m_error"/>
  </IonType>
  ...
</Fragmentation>
```

6.75 Element <PeptideHypothesis>

Definition: Peptide evidence on which this ProteinHypothesis is based by reference to a PeptideEvidence element in a SpectrumIdentificationItem.

Type: PeptideHypothesisType

Attributes:

Attribute Name	Data Type	Use	Definition
PeptideEvidence_Ref	xsd:string	required	A reference to the PeptideEvidence element on which this hypothesis is based.

Subelements: none

Example Context:

```
<PeptideHypothesis PeptideEvidence_Ref="PE1_SEQ_spec10_pep1"/>
```

6.76 Element <IonType>

Definition: IonType defines the index of fragmentation ions being reported, importing a CV term for the type of ion e.g. b ion. Example: if b3 b7 b8 and b10 have been identified, the index attribute will contain 3 7 8 10, and the corresponding values will be reported in parallel arrays below

Type: IonTypeType

Attributes:

Attribute Name	Data Type	Use	Definition
charge	xsd:int	required	The charge of the identified fragmentation ions.
index	listOfIntegers	optional	The index of ions identified as integers, following standard notation for a-c, x-z e.g. if b3 b5 and b6 have been identified, the index would store "3 5 6". For internal ions, the index contains pairs defining the start and end point - see specification document for examples. For immonium ions, the index is the position of the identified ion within the peptide sequence - if the peptide contains the same amino acid in multiple positions that cannot be distinguished, all positions should be given.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	A single entry from an ontology or a controlled vocabulary.
FragmentArray	0	unbounded	An array of values for a given type of measure and for a particular ion type, in parallel to the index of ions identified.

Example Context:

```
<IonType index="5 6 7 11 12 15 17 21 22 23 24 26 27 30 31 33 34 35 38 39 40 41 42 43 44 45 46 47 48 50
51 52 53 54 55 56 59 61 62 65 69 73 76 78 83 86 96" charge="1">
  <cvParam cvRef="PSI-MS" accession="MS:1001231" name="frag: c ion"/>
  <FragmentArray values="447.216737 576.26154 762.38947 1230.608358 1344.651696 1686.8204 1913.9818
2342.1727 2413.208 2470.2302 2607.2849 2792.370851 2921.413169 3246.646488 3402.747973 3662.8968
3763.9465 3820.9689 4184.122757 4285.1695 4398.254701 4527.295803 4655.3939 4802.459719 4917.4869
5045.580864 5192.648576 5320.7443 5457.7927 5698.9794 5800.027 5929.0677 6000.106688 6129.151636
6260.190033 6388.2896 6675.395079 6903.506231 7031.607787 7353.773068 7766.0424 8108.2303 8391.422
8647.6118 9235.887 9549.050832 10610.644 " Measure_ref="m_mz"/>
  <FragmentArray values="4380000 854900 7506000 30210000 12170000 11670000 9618000 10810000 8991000
15620000 5489000 13310000 21520000 31540000 19790000 5384000 9486000 12940000 28610000 8175000 24870000
46070000 10500000 46930000 19710000 22930000 31270000 10140000 5954000 7813000 17080000 9987000
16830000 35420000 14530000 5249000 56680000 13460000 14780000 17410000 5084000 8978000 8186000 7804000
7299000 22380000 6606000" Measure_ref="m_intensity"/>
  <FragmentArray values="-0.0031 -0.0008 0.0478 -0.0030 -0.0026 -0.0030 -0.0050 -0.0048 -0.0067 -0.0059
-0.0101 -0.0042 -0.0045 -0.0077 -0.0073 -0.0110 -0.0090 -0.0080 -0.0084 -0.0094 -0.0082 -0.0097 -0.0066
-0.0092 -0.0090 -0.0100 -0.0107 -0.0099 -0.0204 -0.0127 -0.0128 -0.0147 -0.0128 -0.0105 -0.0126 -0.0080
-0.0142 -0.0141 -0.0075 -0.0175 -0.0168 -0.0192 -0.0171 -0.0172 -0.0189 -0.0188 -0.0238"
Measure_ref="m_error"/>
</IonType>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/AnalysisData/SpectrumIdentificationList/SpectrumIdentificationResult/
SpectrumIdentificationItem/Fragmentation/IonType
MAY supply a *child* term of MS:1001221 (fragmentation information) one or more times
```

```
e.g.: MS:1001220 (frag: y ion)
e.g.: MS:1001222 (frag: b ion - H2O)
e.g.: MS:1001223 (frag: y ion - H2O)
e.g.: MS:1001224 (frag: b ion)
e.g.: MS:1001225 (product ion m/z)
e.g.: MS:1001226 (product ion intensity)
e.g.: MS:1001227 (product ion m/z error)
e.g.: MS:1001228 (frag: x ion)
e.g.: MS:1001229 (frag: a ion)
e.g.: MS:1001230 (frag: z ion)
et al.
```

6.77 Element <FragmentArray>

Definition: An array of values for a given type of measure and for a particular ion type, in parallel to the index of ions identified.

Type: FragmentArrayType

Attributes:

Attribute Name	Data Type	Use	Definition
Measure_ref	xsd:string	required	A reference to the Measure defined in the FragmentationTable
values	listOfFloats	required	The values of this particular measure, corresponding to the index defined in ion type

Subelements: none

Example Context:

```
<FragmentArray values="447.216737 576.26154 762.38947 1230.608358 1344.651696 1686.8204 1913.9818
2342.1727 2413.208 2470.2302 2607.2849 2792.370851 2921.413169 3246.646488 3402.747973 3662.8968
3763.9465 3820.9689 4184.122757 4285.1695 4398.254701 4527.295803 4655.3939 4802.459719 4917.4869
5045.580864 5192.648576 5320.7443 5457.7927 5698.9794 5800.027 5929.0677 6000.106688 6129.151636
6260.190033 6388.2896 6675.395079 6903.506231 7031.607787 7353.773068 7766.0424 8108.2303 8391.422
8647.6118 9235.887 9549.050832 10610.644 " Measure_ref="m_mz"/>
```

6.78 Element <Organization>

Definition: Organizations are entities like companies, universities, government agencies for which the attributes are self describing.

Type: FuGE.Common.Audit.OrganizationType

Attributes:

Attribute Name	Data Type	Use	Definition
address	xsd:string	optional	The address of the Contact.
email	xsd:string	optional	The email address of the Contact.
fax	xsd:string	optional	The fax number of the Contact.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
phone	xsd:string	optional	The telephone number of the Contact including the suitable area codes.
tollFreePhone	xsd:string	optional	A toll free phone number for the Contact, including suitable area codes.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
parent	0	1	The containing organization (the university or business which a lab belongs to, etc.)

Example Context:

```
<Organization id="ORG_MSL" name="Matrix Science Limited" address="64 Baker Street, London W1U 7GB, UK"
email="support@matrixscience.com"  fax="+44 (0)20 7224 1344" phone="+44 (0)20 7486 1050" />
```

6.79 Element <Person>

Definition: A person for which the attributes are self describing.

Type: FuGE.Common.Audit.PersonType

Attributes:

Attribute Name	Data Type	Use	Definition
address	xsd:string	optional	The address of the Contact.
email	xsd:string	optional	The email address of the Contact.
fax	xsd:string	optional	The fax number of the Contact.
firstName	xsd:string	optional	The Person's first name.
id	xsd:string	required	An identifier is an unambiguous string that is unique within the scope (i.e. a document, a set of related documents, or a repository) of its use.
lastName	xsd:string	optional	The Person's last/family name.
midInitials	xsd:string	optional	The Person's middle initial.
name	xsd:string	optional	The potentially ambiguous common identifier, such as a human-readable name for the instance.
phone	xsd:string	optional	The telephone number of the Contact including the suitable area codes.
tollFreePhone	xsd:string	optional	A toll free phone number for the Contact, including suitable area codes.

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
affiliations	0	unbounded	The organization a person belongs to.

Example Context:

```
    <Person id="MPCMEYER" name="Prof. Dr. Helmut E. Meyer" address="Universitaetsstr. 150, D-44795
Bochum, Germany" email="helmut.e.meyer@rub.de">
      <affiliations Organization_ref="MPCINSTITUTE"/>
    </Person>
```

6.80 Element <parent>

Definition: The containing organization (the university or business which a lab belongs to, etc.)

Type:

Attributes:

Attribute Name	Data Type	Use	Definition
Organization_ref	xsd:string	required	Organizations are entities like companies, universities, government agencies for which the attributes are self describing.

Subelements: none

Example Context:

```
<parent Organization_ref="RUB"/>
```

6.81 Element <affiliations>

Definition: The organization a person belongs to.

Type:

Attributes:

Attribute Name	Data Type	Use	Definition
Organization_ref	xsd:string	required	Organizations are entities like companies, universities, government agencies for which the attributes are self describing.

Subelements: none

Example Context:

```
<affiliations Organization_ref="MPCINSTITUTE"/>
```

6.82 Element <DatabaseName>

Definition: The database name may be given as a cvParam if it maps exactly to one of the release databases listed in the CV, otherwise a userParam should be used.

Type:

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.
userParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<DatabaseName>
  <userParam name="ipi.HUMAN_decoy"/>
</DatabaseName>
```

cvParam Mapping Rules:

```
Path /mzIdentML/DataCollection/Inputs/SearchDatabase/DatabaseName
MAY supply a *child* term of MS:101013 (database name) one or more times
  e.g.: MS:101084 (database nr)
  e.g.: MS:101104 (database SwissProt)
  e.g.: MS:101142 (database IPI_human)
  e.g.: MS:101178 (database EST)
  e.g.: MS:101285 (database IPI_mouse)
  e.g.: MS:101286 (database IPI_rat)
  e.g.: MS:101287 (database IPI_zebrafish)
  e.g.: MS:101288 (database IPI_chicken)
  e.g.: MS:101289 (database IPI_cow)
  e.g.: MS:101290 (database IPI_arabidopsis)
```

6.83 Element <spectrumIDFormat>

Definition: The format of the spectrum identifier within the source file.

Type:

Attributes: none

Subelements:

Subelement Name	minOccurs	maxOccurs	Definition
cvParam	1	1	Abstract entity allowing either cvParam or userParam to be referenced in other schemas.

Example Context:

```
<spectrumIDFormat>
  <cvParam accession="MS:1001528" name="Mascot query number" cvRef="PSI-MS" />
</spectrumIDFormat>
```

cvParam Mapping Rules:

```
Path / mzIdentML/ DataCollection/Inputs/SpectraData/spectrumIDFormat
MUST supply a *child* term of MS:1000767 (native spectrum identifier format) only once OR
MUST supply a *child* term of MS:1001529 (spectra data details) only once
 e.g.: MS:1000768 ! Thermo nativeID format
 e.g.: MS:1000769 ! Waters nativeID format
 e.g.: MS:1000770 ! WIFF nativeID format
 e.g.: MS:1000771 ! Bruker/Agilent YEP nativeID format
 e.g.: MS:1000772 ! Bruker BAF nativeID format
 e.g.: MS:1000773 ! Bruker FID nativeID format
 e.g.: MS:1000774 ! multiple peak list nativeID format
 e.g.: MS:1000775 ! single peak list nativeID format
 e.g.: MS:1000776 ! scan number only nativeID format
 e.g.: MS:1000777 ! spectrum identifier nativeID format
 e.g.: MS:1000823 ! Bruker U2 nativeID format
 e.g.: MS:1000824 ! no nativeID format
 e.g.: MS:1001480 ! AB SCIEX TOF/TOF nativeID format
 e.g.: MS:1001508 ! Agilent MassHunter nativeID format
 e.g.: MS:1001526 ! spectrum from database nativeID format
 e.g.: MS:1001528 ! Mascot query number
 e.g.: MS:1001531 ! spectrum from ProteinScape database nativeID format
 e.g.: MS:1001532 ! spectrum from database nativeID format
```

7. Specific Comments on schema

In this section, several points of documentation are elaborated beyond the core specification in Section 6.

7.1 File extension

The file extension for mzIdentML files SHOULD be ".mzid".

7.2 Referencing elements within the document

A number of elements within the schema have an attribute which is used to reference an element elsewhere in the file using the unique identifier of the referenced element. These attributes are named following the convention: “[elementName]_ref”. The uniqueness of the value in the “id” attribute of elements is validated using xsd:key, and the integrity of the reference is validated using xsd:keyref, defined within the schema.

7.3 Searches against nucleotide sequences

Searches of Nucleic acid databases - The “seq” attribute on <DBSequence> SHOULD contain the nucleic acid sequence if a nucleic acid database was searched (rather than up to six translated sequences). <Peptide> represents the identified amino acid sequence (including modifications) and, as such, the <peptideSequence> elements SHOULD store the translated amino acid sequences. <PeptideEvidence> contains the DBSequence_Ref together with the translation frame and a TranslationTable_Ref attribute (see below). The Peptide_Ref is done in <SpectrumIdentificationItem> as in the case for an amino acid database. If a protein dectection is performed, there are <PeptideHypothesis> elements referencing <PeptideEvidence> elements from <SpectrumIdentificationItem> sections. For clarification, see the example instance document for a nucleic acid search (Section 5.4).

In the <SpectrumIdentificationProtocol>, <TranslationTable> is used to specify how nucleic acid sequences are translated into amino acid sequences as follows:

```
<DatabaseTranslation frames="1 2 3 -1 -2 -3">
  <TranslationTable id="TT_1" name="Standard">
    <cvParam accession="MS:1001025" name="translation table" cvRef="PSI-MS"
value="FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG" />
    <cvParam accession="MS:1001410" name="translation start codons" cvRef="PSI-MS" value="---M---------
------M---------------M----------------------------" />
    <cvParam accession="MS:1001423" name="translation table description" cvRef="PSI-MS"
value="http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG1" />
  </TranslationTable>
  <TranslationTable id="TT_2" name="Vertebrate Mitochondrial">
    <cvParam accession="MS:1001025" name="translation table" cvRef="PSI-MS"
value="FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG" />
    <cvParam accession="MS:1001410" name="translation start codons" cvRef="PSI-MS" value="-------------
------------------MMMM---------------M------------" />
    <cvParam accession="MS:1001423" name="translation table description" cvRef="PSI-MS"
value="http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes#SG2" />
  </TranslationTable>
```

The attribute “frames” specifies which frames are considered and one or more translation tables can be specified using CV parameters. The translation table is defined here:
http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/SEQFEAT.HTML#_Genetic_Codes:

"The genetic codes themselves are arrays of 64 amino acid codes. The index to the position in the array of the amino acid is derived from the codon by the following method:

index = (base1 16) + (base2 4) + (base3 1)
where T=0, C=1, A=2, G=3”

The same encoding technique is used to specify start codons. Alphabet names are prefixed with "s" (e.g. sncbieaa) to indicate start codon arrays. Each cell of a start codon array contains either the gap code ("-" for ncbieaa) or an amino acid code if it is valid to use the codon as a start codon. Currently all starts are set to code for methionine, since it has never been convincingly demonstrated that a protein can start with any other amino

acid. However, if other amino acids are shown to be used as starts, this structure can easily accommodate that information.

For each peptide, the frame and translation table should be specified in the PeptideEvidence:

```
<PeptideEvidence id=”1” TranslationTable_ref="TT_1" frame="1" />
```

7.4 Reporting peptide and protein identifications passing a significance threshold

The elements <SpectrumIdentificationItem> and <ProteinDetectionHypothesis> have a mandatory Boolean attribute passThreshold that allows a file producer to indicate that an identification has passed a given threshold or that it has been manually validated. Depending on the intended purpose of the file, the file producer MAY wish to report a number of identifications that fall below the given significance threshold, for example to allow global statistical analyses to be performed which are not possible if only identifications passing the threshold are reported. Thresholds for peptide-spectrum matches or for protein identification should be encoded as instances of <cvParam> within <SpectrumDetectionProtocol> or <ProteinDetectionProtocol> for example as follows. If the file producer does not want to indicate that a threshold has been set, all identifications MUST have passThreshold = “true” and the “no threshold” CV term should be given within the protocols.

```
<SpectrumIdentificationProtocol id="SIP" AnalysisSoftware_ref="AS_mascot_server">
        …
    <Threshold>
      <cvParam accession="MS:1001316" name="mascot:SigThreshold" cvRef="PSI-MS" value="0.05"/>
    </Threshold>


<ProteinDetectionProtocol id="PDP_MascotParser_1"  AnalysisSoftware_ref="AS_mascot_parser">
        …
  <Threshold>
    <cvParam accession="MS:1001316" name="mascot:SigThreshold" cvRef="PSI-MS" value="0.05"/>
  </Threshold>
```

7.5 Using decoy databases to set different thresholds of false discovery rate

mzIdentML supports the reporting of searches against decoy databases, constructed and searched using many of the currently known methods. A <SpectrumIdentificationItem> can be marked as matching a decoy peptide using the isDecoy attribute of the <PeptideEvidence> element, thus allowing the false discovery rate to be calculated across an entire file. The DBSequence_Ref references the decoy protein record.

Implementers of the format SHOULD report the peptide identifications that pass the threshold they wish to communicate to a consumer of the data. For example, a threshold could be set by p-value, false discovery rate, by a native search engine score (or a more complex system documented with CV terms in <Threshold>), and those peptides reported (passing the threshold) are used to determine which proteins have been detected. It is not guaranteed that a consumer of an mzIdentML file will be able to calculate other results, or global false discovery rates, using different thresholds from the reported information, although in some circumstances they may be able to, for example, if a user reports the complete output of a search against a target and decoy search.

```
<SearchDatabase location="/localdirectory/18.E_coli_K12_edit.fasta" id="K12_nosignal" name="K12"
numDatabaseSequences="9376" releaseDate="01-2008-08-2008" version="1.0" >
        <fileFormat>
                <cvParam accession="MS:1001348" name="FASTA format" cvRef="PSI-MS"/>
        </fileFormat>
        <DatabaseName>
                <userParam name="18.E_coli_K12_edit.fasta" />
        </DatabaseName>
        <cvParam accession="MS:1001197" name="DB composition target+decoy" cvRef="PSI-MS"/>
        <cvParam accession="MS:1001283" name="decoy DB accession regexp" value="REV_" cvRef="PSI-MS"/>
```

```
        <cvParam accession="MS:1001195" name="decoy DB type reverse" cvRef="PSI-MS"/>
</SearchDatabase>


<SpectrumIdentificationItem id="item_ref_16_1" calculatedMassToCharge="2096.47" chargeState="3"
experimentalMassToCharge="2096.21" Peptide_ref="peptide_353" passThreshold="0" rank="1">
        <PeptideEvidence id="PE_16_1_REV_SS2_P38097" start="753" end="771"
DBSequence_Ref="REV_SS2_P38097" isDecoy="true" />
                <cvParam accession="MS:1001328" name="OMSSA:evalue" cvRef="PSI-MS"  value="1.33646" />
                <cvParam accession="MS:1001329" name="OMSSA:pvalue" cvRef="PSI-MS"  value="0.00073351"
/>

…

</SpectrumIdentificationItem>
```

7.6 Database Filter

The format can specify that a sequence database has been filtered, for example based on PI, protein mass, taxonomy or even a set of accession numbers for a second pass search. For example all animals except mice would be encoded as (NCBI:33208 is metazoa, NCBI:10090 is *Mus musculus*):

```
<DatabaseFilters>
  <Filter>
    <FilterType>
      <cvParam accession="MS:1001020" name="DB filter taxonomy" cvRef="PSI-MS" />
    </FilterType>
   <Include>
      <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="33208"/>
   </Include>
   <Exclude>
     <cvParam accession="MS:1001467" name="taxonomy: NCBI TaxID" cvRef="PSI-MS" value="10090"/>
   </Exclude>
   </Filter>
</DatabaseFilters>
```

7.7 Types of parameters and values

There are several types for parameters that are used in the schema:

<ParamListType>: A list (i.e. unbounded number) of <ParamGroup>.
<ParamGroup>: A choice between <cvParam> or <userParam>.

<ParamType>: A single reference to <ParamGroup>, which allows a choice between either <cvParam> or <userParam> at the specified point in the schema.
<cvParamType>: A single entry from an ontology or a controlled vocabulary. Attributes: accession, cvRef, name, value, unitAccession, unitName, unitCvRef.
<userParamType>: A single user-defined parameter. Attributes: name, value, unitAccession, unitName, unitCvRef.

7.8 Reporting fragmentation ions

mzIdentML employs an array type structure to support the reporting of ion types identified in an MS/MS analysis, coupled with CV parameters to retain flexibility in the types of ion that can be reported. A brief example is given here to explain how these structures should be used where y11, y8 and y7 have been identified with charge = 2+. First, the types of measures to be reported are given in the <FragmentationTable> using <cvParam> instances. Second, each <SpectrumIdentificationItem> contains an index of values (11, 8 and 7 for each y ion) and parallel arrays that reference back to each <Measure> defined in the <FragmentationTable>. In the example, the y8 ion has a product ion m/z = 436.4, product ion intensity = 11 and product ion m/z error = 0.1284 (the second position in the index of each array).

```
<FragmentationTable>
  <Measure id="m_mz">
    <cvParam cvRef="PSI-MS" accession="MS:1001225" name="product ion m/z"/>
  </Measure>
  <Measure id="m_intensity">
    <cvParam cvRef="PSI-MS" accession="MS:1001226" name="product ion intensity"/>
  </Measure>
  <Measure id="m_error">
    <cvParam cvRef="PSI-MS" accession="MS:1001227" name="product ion m/z error"
unitAccession="MS:1000040" unitName="m/z" unitCvRef="PSI-MS"/>
  </Measure>
</FragmentationTable>
…

<IonType index="11 8 7" charge="2">
  <cvParam cvRef="PSI-MS" accession="MS:1001220" name="frag: y ion"/>
  <FragmentArray values="551.3 436.4 380.1 " Measure_ref="m_mz"/>
  <FragmentArray values="800 11 46" Measure_ref="m_intensity"/>
  <FragmentArray values="0.4752 0.1284 0.3704" Measure_ref="m_error"/>
</IonType>
```

7.8.1 Internal fragments and immonium ions

mzIdentML supports the reporting of internal fragment ions, of which an immonium ion is a special case comprising a single side chain (http://www.matrixscience.com/help/fragmentation_help.html). For internal and immonium ions, the index is used in two different ways. Internal fragments are reported using the index structure to identify the start and end of the ion within the sequence. The example shows how the index performs this different role, as it identifies pairs of internal ions: ya2-5, ya3-7, ya3-8, ya4-8, ya5-8, ya5-11, ya8-11.

```
<IonType index="2 5 3 7 3 8 4 8 5 8 5 11 8 11" charge="1">
  <cvParam cvRef="PSI-MS" accession="MS:1001366" name="frag: internal ya ion"/>
  <FragmentArray values="315.2 388.1 501.4 444.1 342.8 669.901495 412.4 " Measure_ref="m_mz"/>
  <FragmentArray values="44 63 10430 75 48 6420 31" Measure_ref="m_intensity"/>
  <FragmentArray values="-0.0027 -0.1191 0.0969 -0.1817 -0.4340 0.4721 0.1082" Measure_ref="m_error"/>
</IonType>
```

For immonium ions, the index is the position of the identified ion within the peptide sequence. If the peptide contains the same amino acid in multiple positions that cannot be distinguished, all positions should be given. Example, where immonium ions have been found matching T and G in the following peptide sequence FGGEENTY (positions 2 or 3, and position 7):

```
<IonType cvRef="PSI-PI" accession=" MS:1001239" name="frag: immonium ion" index="2 3 7" charge="1">
  <FragmentArray values="288.2 286.1 387.2 371.127841 " Measure_ref="m_mz"/>
  <FragmentArray values="2137 83 656 1663" Measure_ref="m_intensity"/>
  <FragmentArray values="0.0260 -0.1125 -0.0602 -0.1011" Measure_ref="m_error"/>
</IonType>
```

7.9 Enzyme definition

The <SpectrumIdentificationProtocol> SHOULD contain a specification of which enzyme (if any) was applied in the search. The element <Enzyme> has optional sub-elements for specifying the <EnzymeName> using a CV term and the cleavage site, using a regular expression. Regular expressions should be encoded following the notation of Perl Compatible Regular Expressions (PCRE regex, http://www.pcre.org, matching the syntax and semantics of Perl version 5). The PSI-MS CV contains terms for the most common enzymes with pre-defined regular expressions (Table 2). If the enzyme used is present in the PSI-MS CV, the term MUST be encoded under <EnzymeName> unless the rule given in the CV does not match that used by the software or if the enzyme used is not present in the CV, in which case the regular expression used MUST be given in the element <SiteRegexp>. If the <EnzymeName> element is used, the regular expression MAY also be provided additionally. For a no enzyme search, (i.e. one where there may be a cleavage at any residue), the cvTerm

MS:1001091 'NoEnzyme' MUST be specified, and the missedCleavages and semiSpecific attributes SHOULD NOT be specified. If two or more enzymes are used, multiple <Enzyme> elements SHOULD be provided rather than trying to build a regular expression covering all cleavage sites. If the software uses a name for an enzyme other than the one specified in the CV, a userParam MAY also be given.

The following guidelines SHOULD be followed when generating regular expressions in an instance document for enzymes not present in the CV: 1) use the PCRE supplied negation syntax for look-ahead and look-behind assertions and 2) use the most compact representation possible for a regex. The start of a match specifies the cleavage point. For example the enzyme Trypsin, which cleaves following a K or R residue unless the next residue is P, has the regular expression:

```
(?<=[KR])(?!P)
```

The ?<= is a "zero-width positive look-behind assertion", and [] means one of this character set. So, this rule is to look behind for a K or R. ?! is a zero-width positive look-ahead assertion, and ?!P means any character that is not P. An example of an "N-term" enzyme is Asp-N which cleaves before D or B. This can be described using the PCRE:

```
(?=[BD])
```

The ?= is a "zero-width positive look-ahead assertion."

A simple 3 line perl program can be written to test a regular expression:

```
$protein = "ABCDKPEFGHIJKLMNOPQRSTUVWXYZ";
@peptides = split(/(?<=[KR])( ?!P)/, $protein);
print join "\n", @peptides;
```

The program returns:

```
ABCDKPEFGHIJK
LMNOPQR
STUVWXYZ
```

Enzyme Name	Regular expression
Trypsin	(?<=[KR])(?!P)
Arg-C	(?<=R)(?!P)
Asp-N	(?=[BD])
Asp-N_ambic	(?=[DE])
Chymotrypsin	(?<=[FYWL])(?!P)
CNBr	(?<=M)
Formic_acid	((?<=D))|((?=D))
Lys-C	(?<=K)(?!P)
Lys-C/P	(?<=K)
PepsinA	(?<=[FL])
TrypChymo	(?<=[FYWLKR])(?!P)
Trypsin/P	(?<=[KR])
V8-DE	(?<=[BDEZ])(?!P)
V8-E	(?<=[EZ])(?!P)

Table 2 Common enzymes and the cleavage site specified as regular expressions as represented in the PSI-MS CV.

8. Conclusions

This document contains the specifications for using the mzIdentML format to represent results from peptide and protein identification pipelines, in the context of a proteomics investigation. This specification, in conjunction with the XML Schema, mapping file and CV constitute a proposal for a standard from the Proteomics Standards Initiative. These artefacts are currently undergoing the PSI document process standardization process, which will result in a standard officially sanctioned by PSI.

9. Authors and Contributors

Angel Pizarro
ITMAT Bioinformatics Facility
Biological Research Building
University of Pennsyvania
Philadelphia, PA 19104-6160
USA
angel@mail.med.upenn.edu

David Creasy
Matrix Science,
64 Baker Street
London W1U 7GB
UK
dcreasy@matrixscience.com

Philip Jones
EMBL-EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD
UK
pjones@ebi.ac.uk

Andreas Bertsch
Eberhard Karls University Tübingen
Sand 14
D-72076 Tübingen
Germany
bertsch@informatik.uni-tuebingen.de

Jenny Siepen
Faculty of Life Sciences
University of Manchester
M13 9PT
UK
jennifer.siepen@manchester.ac.uk

Martin Eisenacher
Medizinisches Proteom-Center (MPC)
Ruhr-Universität Bochum
Universitätsstr. 150
D-44801 Bochum
Germany
martin.eisenacher@ruhr-uni-bochum.de

Andrew Jones
Department of Pre-clinical Veterinary Science
Faculty of Veterinary Science
University of Liverpool
UK
Andrew.jones@liv.ac.uk

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

- Eric Deutsch, Institute for Systems Biology
- Matt Chambers, Vanderbilt University
- Simon Hubbard, University of Manchester
- Julian Selley, University of Manchester
- Zsuzsanna Bencsath-Makkai, Biomedical Engineering, McGill University
- Sean Seymour, Applied Biosystems
- Randy Julian, IndigoBio
- Pierre-Alain Binz, GeneBio Geneva
- Alex Masselot, GeneBio Geneva
- Juan Antonio Vizcaíno, European Bioinformatics Institute
- Lennart Martens, European Bioinformatics Institute
- Henning Hermjakob, European Bioinformatics Institute
- Luisa Montecchi, European Bioinformatics Institute
- Florian Reisinger, European Bioinformatics Institute
- Richard Cote, European Bioinformatics Institute
- Marc Sturm, Eberhard Karls University, Tübingen
- Jim Shofstahl, Thermo Fisher
- David Horn, Agilent
- Jimmy Eng, Fred Hutchinson Cancer Research
- Brian Searle, Proteome Software
- Phillip Young, Waters
- Michael Kohl, Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany
- Christian Stephan, Medizinisches Proteom-Center, Ruhr-Universität Bochum, Germany
- Eugene Kapp, Ludwig Institute for Cancer Research
- Michael Coleman, Stowers Institute.

10. References

- [RFC2119] Bradner, S. (1997). "Key words for use in RFCs to Indicate Requirement Levels, Internet Engineering Task Force, RFC 2119, http://www.ietf.org/rfc/rfc2119.txt.
- [Jones 07] Jones AR, Miller M, Spellman P and Pizarro A. Specification documentation for the Functional Genomics Experiment (FuGE) model: user guide. Version 1 (final): http://fuge.sourceforge.net/dev/V1Final/FuGE-v1-SpecDoc.doc.
- [Deutsch08] Deutsch EW, Martens L, Montecchi-Palazzi L, Binz P-A, Kessner D, Souda P mzML: Mass Spectrometry Markup Language, http://www.psidev.info/index.php?q=node/303.

11. 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).

Copyright Notice