Encoding MS2 Clustering Results in mzTab

# Introduction

The aim of this addition to the mzTab format specification is to incorporate the results of MS2 clustering tools directly in mzTab. As a core prerequisite this proposal explicitly aims to provide a method that keeps the format backward compatible to the current release 1.0.

This addition is being developed to support the following use cases:

1. Store information about the clusters formed by a clustering process
2. Store (clustering specific) information about the spectra that are part of a cluster
3. Support the direct connection of identification and quantification data stored in mzTab with the clustering results

# Format Specification

This format specification is intended as an addition to the already existing format specification of mzTab version 1.0.

Two new table based sections will be added to encode the clustering information in mzTab: 1) The **cluster** table holds information about the formed clusters and 2) the **clustered spectrum** table links the spectra with the stored clusters.

The existing **line prefixes** will be extended as follows:

* CLH for the cluster table header
* CLS for rows of the cluster table
* CSH for the clustered spectrum table header
* CSS for the rows of the clustered spectrum table

The clusters’ representative spectra are stored in an external file. Even though this file is not the result of a specific ms run, the already existing method to reference external spectra in mzTab will not be changed. Therefore, the corresponding peak list files must be listed as “ms\_runs” in the metadata section of the mzTab file. To highlight that this ms\_run is representing clustering results a new field is added to the **metadata section**:

### ms\_run[1-n]-clustering\_algorithm

|  |  |
| --- | --- |
| Description: | This metadata section field holds the name of the used clustering algorithm. It is additionally used to highlight that a certain referenced ms data file is holding clustering results. |
| Type: | String |
| Mandatory: |  |
| Example: | MTD ms\_run[1]-clustering\_algorithm ms-cluster |

The PSM and the peptide section are implicitly linked to the cluster section by referring to the same spectrum. Based on this newly added field, the parsing software can easily determine that this ms data file is holding clustering results and that additional information about the actual clustering results may be present in the file. More importantly, this method is fully backward compatible to the mzTab format specification 1.0.

Additional metadata about the used clustering algorithm can be encoded using the already existing “software[1-n]” field in combination with the “software[1-n]-setting[1-n]” field. Since clustering algorithms are still evolving any more specific fields are in my opinion not really useful.

## Cluster Section

The cluster section is table based. The following list of columns is a first suggestion.

### CLUSTER\_ID

|  |  |
| --- | --- |
| Description: | A unique identifier for a cluster within the mzTab file. |
| Type: | String |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID …  CLS CL\_1 … |

### precursor\_mz

|  |  |
| --- | --- |
| Description: | The cluster’s consensus spectrum’s precursor m/z. Depending on the used algorithm this value will either be the average precursor m/z of all contained spectra or the precursor m/z of the best representative spectrum. |
| Type: | Double |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID precursor\_mz …  CLS CL\_1 350.00 … |

### precursor\_intensity

|  |  |
| --- | --- |
| Description: | The cluster’s consensus spectrum’s precursor intensity. Depending on the used algorithm this value will either be the average precursor intensity of all contained spectra or the precursor intensity of the best representative spectrum. |
| Type: | Double |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID precursor\_mz precursor\_intensity …  CLS CL\_1 350.00 1000123 … |

### charge

|  |  |
| --- | --- |
| Description: | The cluster’s consensus spectrum’s charge. Depending on the used algorithm this value will either be the most common charge of the contained spectra or the charge of the best representative spectrum. |
| Type: | Integer |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID precursor\_mz charge …  CLS CL\_1 350.00 2 … |

### spectrum\_ref

|  |  |
| --- | --- |
| Description: | Reference to the spectrum representing the cluster. The reference must be in the format ms\_run[1-n]:{SPECTRA\_REF} where SPECTRA\_REF MUST follow the format defined in the mzTab format specification section 5.2.  Depending on the algorithm this will be either a unified consensus spectrum or the best representative spectrum. |
| Type: | String |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID … spectrum\_ref …  CLS CL\_1 … ms\_run[1]:index=5 … |

### size

|  |  |
| --- | --- |
| Description: | The number of spectra within the cluster. |
| Type: | Integer |
| Mandatory: |  |
| Example: | CLH CLUSTER\_ID … size …  CLS CL\_1 … 10 … |

### Additional Information to Add (Points for discussion)

* Add reliability information about the accuracy of the clustering process?

## Clustered Spectrum Section

The clustered spectrum section is a table based section. It is not very elegant to have such a section only representing the spectrum’s metadata. Unfortunately, I don’t really see another way to reference the spectra within a cluster in any other sensible way.

### spectrum\_ref

|  |  |
| --- | --- |
| Description: | Reference to the actual spectrum |
| Type: | String |
| Mandatory: |  |
| Example: | CSH spectrum\_ref …  CSS ms\_run[1]:index=5 … |
|  |  |

### cluster\_id

|  |  |
| --- | --- |
| Description: | Reference to the cluster the spectrum is part of. In case the spectrum is found in multiple clusters, multiple rows MUST be added for the same spectrum using the same spectrum\_ref. |
| Type: | String |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id …  CSS ms\_run[1]:index=5 CL\_1 … |

### cluster\_id

|  |  |
| --- | --- |
| Description: | Reference to the cluster the spectrum is part of. In case the spectrum is found in multiple clusters, multiple rows MUST be added for the same spectrum using the same spectrum\_ref. |
| Type: | String |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id …  CSS ms\_run[1]:index=5 CL\_1 … |

### precursor\_mz

|  |  |
| --- | --- |
| Description: | The spectrum’s precursor’s m/z. |
| Type: | Double |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id precursor\_mz …  CSS ms\_run[1]:index=5 CL\_1 350.00 … |

### precursor\_intensity

|  |  |
| --- | --- |
| Description: | The spectrum’s precursor’s intensity. |
| Type: | Double |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id … precursor\_intensity …  CSS ms\_run[1]:index=5 CL\_1 … 1000123 … |

### charge

|  |  |
| --- | --- |
| Description: | The spectrum’s precursor’s charge. |
| Type: | Integer |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id … charge …  CSS ms\_run[1]:index=5 CL\_1 … 2 … |

### similarity

|  |  |
| --- | --- |
| Description: | The similarity of the spectrum to the cluster’s consensus spectrum / best representative depending on the used algorithm (this metric might not be applicable to many algorithms). |
| Type: | Double |
| Mandatory: |  |
| Example: | CSH spectrum\_ref cluster\_id … similarity …  CSS ms\_run[1]:index=5 CL\_1 … 0.85 … |