**PSI Recommendation**

Eric W. Deutsch, Institute for Systems Biology

Hans Vissers, Waters Corporation

Hannes Röst, University of Toronto

Matt Chambers, Consultant

Andrew Dowsey, University of Bristol

Pierre-Alain Binz, Centre Hospitalier Universitaire Vaudois

**Recommendation for encoding data independent acquisition and ion mobility data in mzML 1.1**

Status of this document

DRAFT Version 0.9.5

To be versioned as 1.0 upon approval.

# 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. This document presents a consensus community practice description on how to encode the data from various data independent analysis (DIA) workflows, including scanning quadrupole data independent acquisition (SONAR, Waters Corporation) and ion mobility assisted data independent acquisition workflows (HDMSE, Waters Corporation, and diaPASEF, Bruker Daltonik) in the PSI standard format mzML 1.1. Although a schema change from mzML 1.1 to 1.2 was considered, it was determined that the data could be encoded with only the addition of controlled vocabulary terms and the additional guidance provided in this document, and no schema change was advisable.

Contents

[Abstract 1](#_heading=h.gjdgxs)

[1.](#_heading=h.30j0zll) Introduction 2

[1.1](#_heading=h.1fob9te) Description of the need 2

[1.2](#_heading=h.tyjcwt) Requirements 2

[1.3](#_heading=h.3dy6vkm) Issues to be addressed 2

[2.](#_heading=h.1t3h5sf) Notational Conventions 2

[3.](#_heading=h.2s8eyo1) The Format Implementation 3

[3.1](#_heading=h.17dp8vu) The documentation 3

[3.2](#_heading=h.3rdcrjn) Relationship to other specifications 3

[3.3](#_heading=h.26in1rg) Adjustments for raw DIA data such as SONAR, HDMSE, or diaPASEF data 3

[3.4](#_heading=h.lnxbz9) Adjustments for deconvoluted DIA data such as SONAR or HDMSE data 4

[3.5](#_heading=h.35nkun2) Adjustments for differentiating between low and high energy scans 5

[3.6](#_heading=h.1ksv4uv) Adjustments for front-end ion mobility filtering 5

[3.7](#_heading=h.44sinio) Parallel subsampled data arrays 6

[3.8](#_heading=h.2jxsxqh) Additional compression types 6

[4.](#_heading=h.z337ya) Authors Information 7

[5.](#_heading=h.3j2qqm3) Contributors 7

[6.](#_heading=h.1y810tw) Intellectual Property Statement 7

[7.](#_heading=h.4i7ojhp) Copyright Notice 8

[8.](#_heading=h.2bn6wsx) Glossary 8

[9.](#_heading=h.qsh70q) References 8

# Introduction

## Description of the need

The mzML format (1) is an open, comprehensive XML-based format for encoding mass spectrometer output data. It is the only such format approved by the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI). The format has remained stable and unchanged since the mzML 1.1 release in June 2009, although the addition of controlled vocabulary (CV) terms to the PSI-MS CV (2) have enabled it to encode data from new workflows since its release.

Recently developed workflows by Waters Corporation include SONAR (scanning quadrupole data independent acquisition) (3) and HDMSE (ion mobility assisted data independent acquisition) (4),and Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF) (5). It has become important to clarify how data from these workflows SHOULD be encoded in the mzML 1.1 format to minimize differences in implementation amongst software packages that can properly handle these workflows. There are two primary forms of the data from these two workflows: the original high dimensional data and the reduced-dimensionality deconvoluted data. Although the primary driver of these requirements has been from Waters Corporation and Bruker Daltonik, it is desirable to describe these new technologies in somewhat generic terms that can apply to similar technologies from other vendors.

## Requirements

The main requirements to be fulfilled are:

* Provide a clear single recommended way to encode and interpret the raw output data from SONAR, HDMSE, diaPASEF and similar data-independent acquisition (DIA) workflows in mzML 1.1.
* Provide a clear single recommended way to encode and interpret the deconvoluted data from SONAR, HDMSE, and similar DIA workflows in mzML 1.1.
* Encode the aforementioned information without changing the XML schema.

## Issues to be addressed

The main issues to be addressed by the recommendation are:

* The only issue to be addressed is the potential confusion or ambiguity on how the aforementioned data types are encoded in mzML 1.1.

# Notational Conventions

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 (6).

Essentially, “MUST” denotes REQUIRED, “SHOULD” denotes recommended, “MAY” denotes optional.

# The Format Implementation

## The documentation

All information for this best practice recommendation is provided in this document. Potential updates to this document or other follow-on recommendations for other workflows are provided at the mzML documentation page at the HUPO-PSI website (<http://www.psidev.info/mzML>).

## Relationship to other specifications

This recommendation is a direct addendum to the main mzML 1.1 specification, which can be found at <http://www.psidev.info/mzML>.

## Adjustments for raw DIA data such as SONAR, HDMSE, or diaPASEF data

When raw (i.e. not deconvolved) data from SONAR, HDMSE, diaPASEF or similar workflows are encoded in mzML, the file content header SHOULD contain one or more of the following CV term sets, depending on the appropriate workflow:

|  |  |  |
| --- | --- | --- |
| Workflow | Description | CV term set |
| MSE  AIF  p2cid | data independent acquisition from dissociation of full mass range | - data independent acquisition  - dissociation of full mass range |
| SONAR | data independent acquisition from dissociation of scanning quadrupole across mass range | - data independent acquisition  - dissociation of scanning quadrupole across a specified mass range |
| SWATH-MS  FT-ARM  HRM  PAcIFIC | data independent acquisition from dissociation of sequential mass ranges | - data independent acquisition  - dissociation of sequential mass ranges |
| diaPASEF | data independent acquisition from dissociation of sequential mass ranges separated by ion mobility | - data independent acquisition  - dissociation of sequential mass ranges  - ion mobility separation |
| HDMSE  IMS-AIF | data independent acquisition from dissociation of full mass range separated by ion mobility | - data independent acquisition  - dissociation of full mass range  - ion mobility separation |

Raw MS1 and MS2 scans are encoded according to the original mzML 1.1 specification.

However, for ion mobility data, instead of encoding each ion mobility frame to a different <spectrum>, which leads to millions of spectra per run, it is recommended that each sweep of ion mobility is recorded as a single <spectrum> with an extra ion mobility dimension. In addition to the “m/z array” and “intensity array” (and potentially other arrays) that make up each spectrum, an additional parallel array SHOULD be provided with the most appropriate child term of “ion mobility array” (MS:1002893), depending on the units and degree of processing. The ambiguous parent term MS:1002893 itself MUST NOT be used in a data file.

There was some concern that in the original mzML 1.1 format there should never be two identical m/z values in an “m/z array”. But this was not explicitly stipulated in the mzML 1.1 specification, and with the addition of ion mobility arrays, repetition in the “m/z array” will be common (having differing ion mobility array values). This means that all mzML parsers SHOULD be checked/updated to handle mzML data with non-increasing (identical or decreasing) m/z values in the "m/z array". All parsers SHOULD also be updated to warn or halt gracefully if they are not prepared to handle the kinds of data described herein.

The set of binary arrays SHOULD be sorted in order of ascending m/z, with the ion mobility value (any type) in ascending order as the second sorting criterion if it is present. It is expected that this order will compress most efficiently. However, this sort order is not enforced by the mzML validator. Software SHOULD be prepared to handle other orders.

Example HDMSE and diaPASEF mzML data are available at:

<http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD008362>

<https://github.com/OpenMS/OpenMS/blob/develop/src/tests/topp/OpenSwathWorkflow_1_input.mzML>

For SONAR or other scanning quadrupole data, the lower bound and upper bound of the scanning quadrupole isolation window MUST be provided using two additional data arrays next to the m/z array and intensity arrays as follows:

<cvParam cvRef="MS" accession="MS:1003157" name="scanning quadrupole position lower bound m/z array" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />

and

<cvParam cvRef="MS" accession="MS:1003158" name="scanning quadrupole position upper bound m/z array" unitAccession="MS:1000040" unitName="m/z" unitCvRef="MS" />

## Adjustments for deconvoluted DIA data such as SONAR or HDMSE data

We define deconvoluted data as average properties of an analyte post peak-detection, weighted charge state reduction, and/or adduct aggregation. When deconvoluted data from SONAR, HDMSE, or similar workflows are encoded in mzML, the file Content header MUST contain one of the following CV term sets, depending on the appropriate workflow:

|  |  |  |
| --- | --- | --- |
| Workflow | Description | CV term set |
| MS^e  AIF  p2cid | data independent acquisition from dissociation of full mass range | - data independent acquisition  - dissociation of full mass range  - deconvoluted data |
| SONAR | data independent acquisition from dissociation of scanning quadrupole across mass range | - data independent acquisition  - dissociation of scanning quadrupole across a specified mass range  - deconvoluted data |
| SWATH-MS  FT-ARM  HRM  PAcIFIC | data independent acquisition from dissociation of sequential mass ranges | - data independent acquisition  - dissociation of sequential mass ranges  - deconvoluted data |
| diaPASEF | data independent acquisition from dissociation of sequential mass ranges separated by ion mobility | - data independent acquisition  - dissociation of sequential mass ranges  - ion mobility separation  - deconvoluted data |
| HDMSE  IMS-AIF | data independent acquisition from dissociation of full mass range separated by ion mobility | - data independent acquisition  - dissociation of full mass range  - ion mobility separation  - deconvoluted data |

Furthermore, the <selectedIonList> element MUST contain a list of all of the measured analyte sibling ions (i.e. different charge states of the same molecule) that are associated with the fragment ion spectrum. This is in contrast to mzML from other workflows where the entries in the <selectedIonlist> are one or more unrelated analyte ions that happened to be included in the isolation window. For example:

<selectedIonList count="2">

<selectedIon>

<cvParam cvRef="MS" accession="MS:1000744" name="selected ion m/z" value="666.679163727272" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>

<cvParam cvRef="MS" accession="MS:1000041" name="charge state" value="2"/>

<cvParam cvRef="MS" accession="MS:1000042" name="peak intensity" value="0.87e05" unitCvRef="MS" unitAccession="MS:1000131" unitName="number of detector counts"/>

</selectedIon>

<selectedIon>

<cvParam cvRef="MS" accession="MS:1000744" name="selected ion m/z" value="333.679163727272" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>

<cvParam cvRef="MS" accession="MS:1000041" name="charge state" value="3"/>

<cvParam cvRef="MS" accession="MS:1000042" name="peak intensity" value="0.40e05" unitCvRef="MS" unitAccession="MS:1000131" unitName="number of detector counts"/>

</selectedIon>

</selectedIonList>

Here, both selected ions are 2+ and 3+ siblings of the same analyte from which the deconvoluted fragment ion spectrum is created. Software reading this information MAY display a single analyte where the charge-reduced precursor mass (M+H)+ and total ion intensity have been computed from the above information, with an option to see the individual ions.

Moreover, for any DIA acquisition of a full mass range (such as MSE or HDMSE data), the <isolationWindow> element SHOULD contain the full isolation window information with an added term to denote that the isolation is a full range, such as:

<cvParam cvRef="MS" accession="MS:1000827" name="isolation window target m/z" value="50.0" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>

<cvParam cvRef="MS" accession="MS:1000828" name="isolation window lower offset" value="0" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>

<cvParam cvRef="MS" accession="MS:1000829" name="isolation window upper offset" value="1950" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>

<cvParam cvRef="MS" accession="MS:1003159" name="isolation window full range"/>

Note that the <isolationWindow> element is required by the current schema and this requirement is maintained.

## Adjustments for differentiating between low and high energy scans

There has been substantial confusion in the community about the best way to differentiate between the low energy and high energy scans from Waters Corporation instruments for MSE, HDMSE, and SONAR acquisitions. Several possible strategies were proposed, but the selected implementation requires no schema change and is as follows.

Low-energy scans will be encoded as MS Level 1 scans. High-energy scans will be encoded as MS Level 2 scans with an isolationWindow that is set to be same width as the scanWindow, with a target m/z that is set as the lower bound of the window, as described in section 3.4 and the collision energy applied encoded as usual.

## Adjustments for front-end ion mobility filtering

Some instruments support filtering of ions entering a mass spectrometer by ion mobility. A common mechanism is the high-field asymmetric waveform ion mobility spectrometry (FAIMS) device, for which each scan may have a single compensation voltage value. If FAIMS information is available it MUST be encoded in mzML files by specifying the “FAIMS compensation voltage” CV term (MS:1001581) in the <scan> element. For example:

<scan instrumentConfigurationRef="IC1">

<cvParam cvRef="MS" accession="MS:1000016" value="6.437239112" name="scan start time" unitAccession="UO:0000031" unitName="minute" unitCvRef="UO" />

<cvParam cvRef="MS" accession="MS:1000927" value="118" name="ion injection time" unitAccession="UO:0000028" unitName="millisecond" unitCvRef="UO" />

<cvParam cvRef="MS" accession="MS:1001581" name="FAIMS compensation voltage" value="-54.0" unitAccession="UO:0000218" unitName="volt" unitCvRef="UO"/>

Note that some converters prior to 2021-04-01 placed the “FAIMS compensation voltage” term in the <spectrum> element. Readers MAY also consult the <spectrum> element to see if the FAIMS compensation voltage term is present there.

A similar technology is the SelexION device from SCIEX. For data acquired with the SolexION device, if the device voltage is available, the information MUST be encoded in mzML files by specifying the “solexION differential mobility” (???) CV term in the <scan> element. For example:

<scan instrumentConfigurationRef="IC1">

<cvParam cvRef="MS" accession="MS:1000016" value="12.39486" name="scan start time" unitAccession="UO:0000031" unitName="minute" unitCvRef="UO" />

<cvParam cvRef="MS" accession="MS:1000927" value="60" name="ion injection time" unitAccession="UO:0000028" unitName="millisecond" unitCvRef="UO" />

<cvParam cvRef="MS" accession="MS:1001581" name="SelexION differential mobility" value="-54.0" unitAccession="UO:0000218" unitName="volt" unitCvRef="UO"/>

<spectrum index="100" id="controllerType=0 controllerNumber=1 scan=101" defaultArrayLength="67">

<cvParam cvRef="MS" accession="MS:1000579" name="MS1 spectrum" value=""/>

<cvParam cvRef="MS" accession="MS:1000511" name="ms level" value="1"/>

<cvParam cvRef="MS" accession="???" name="SelexION differential mobility separation device"/>

<cvParam cvRef="MS" accession="???" name="differential mobility" value="20.0"/>

...

## Parallel subsampled data arrays

In a standard mzML <spectrum> <binaryDataArrayList>, each of the data arrays SHOULD have the same number of elements. However, there is a requested use case to allow a second parallel array that is sampled with fewer m/z data points. The requested use case is for storing subsampled (i.e. interpolated to fewer data points) noise information. In order to support such a use case, the following set of “sampled” data arrays are available in the CV for use here:

id: MS:1002743

name: sampled noise m/z array

def: "A data array of parallel, independent m/z values for a sampling of noise across a spectrum (typically much smaller than MS:1000514, the m/z array)." [PSI:MS]

id: MS:1002744

name: sampled noise intensity array

def: "A data array of intensity values for the amplitude of noise variation superposed on the baseline (MS:1002745) across a spectrum (for use with MS:0002743, sampled noise m/z array)." [PSI:MS]

id: MS:1002745

name: sampled noise baseline array

def: "A data array of baseline intensity values (the intensity in the absence of analytes) for a sampling of noise across a spectrum (for use with MS:0002743, sampled noise m/z array)." [PSI:MS]

## Additional compression types

At the original release of mzML 1.1, only “zlib compression” and “no compression” were supported compression types. There are now 11 different compression settings permitted:

(MS:1000576) no compression

(MS:1000574) zlib compression

(MS:1002312) MS-Numpress linear prediction compression

(MS:1002746) MS-Numpress linear prediction compression followed by zlib compression

(MS:1002313) MS-Numpress positive integer compression

(MS:1002747) MS-Numpress positive integer compression followed by zlib compression

(MS:1002314) MS-Numpress short logged float compression

(MS:1002748) MS-Numpress short logged float compression followed by zlib compression

(MS:1003088) truncation and zlib compression

(MS:1003089) truncation, delta prediction and zlib compression

(MS:1003090) truncation, linear prediction and zlib compression

along with 2 additional terms that can describe settable parameters:

(MS:1003091) binary data compression parameter

(MS:1003092) number of mantissa bits truncated

These compression types have been implemented in ProteoWizard for reading and writing, and other software tools may encounter files written with these compression techniques.

# Authors Information

Eric W. Deutsch

Institute for Systems Biology, Seattle WA, USA

edeutsch@systemsbiology.org

Hans Vissers

Waters Corporation

Stamford Avenue | Altrincham Road, Wilmslow, SK9 4AX, UK

hans\_vissers@waters.com

Matt Chambers, Bioinformatics Consultant

Stamford CT, USA

[matt.chambers42@gmail.com](mailto:matt.chambers42@gmail.com)

Hannes Röst

University of Toronto

160 College Street, Room 604  
Toronto, ON M5S 3E1, Canada

hannes.rost@utoronto.ca

Andrew Dowsey

University of Bristol

Oakfield Grove Bristol BS8 2BN, UK

andrew.dowsey@bristol.ac.uk

Pierre-Alain Binz

CHUV Centre Universitaire Hospitalier Vaudois,

CH-1011 Lausanne 14, Switzerland

pierre-alain.binz@chuv.ch

# Contributors

In addition to the authors, a number of additional contributions have been made during the preparation process. The contributors who actively participated to the recommendation documentation are:

Pierre-Alain Binz, Centre Hospitalier Universitaire Vaudois

Matt Chambers, Vanderbilt University

Joshua Klein, Boston University

# 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 which 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

Copyright (C) Proteomics Standards Initiative (2017). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

# Glossary

Not used.

# References

1. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011) mzML--a community standard for mass spectrometry data. *Mol. Cell. Proteomics MCP* 10, R110.000133

2. Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., Jones, A. R., Binz, P.-A., Deutsch, E. W., Chambers, M., Kallhardt, M., Levander, F., Shofstahl, J., Orchard, S., Vizcaíno, J. A., Hermjakob, H., Stephan, C., Meyer, H. E., Eisenacher, M., and HUPO-PSI Group (2013) The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database J. Biol. Databases Curation* 2013, bat009

3. Moseley, M. A., Hughes, C. J., Juvvadi, P. R., Soderblom, E. J., Lennon, S., Perkins, S. R., Thompson, J. W., Steinbach, W. J., Geromanos, S. J., Wildgoose, J., Langridge, J. I., Richardson, K., and Vissers, J. P. C. (2017) Scanning Quadrupole Data Independent Acquisition - Part A. Qualitative and Quantitative Characterization. *J. Proteome Res.*,

4. Rodriguez-Suarez, E., Hughes, C., Gethings, L., Giles, K., Wildgoose, J., Stapels, M., E. Fadgen, K., J. Geromanos, S., P.C. Vissers, J., Elortza, F., and I. Langridge, J. (2013) An Ion Mobility Assisted Data Independent LC-MS Strategy for the Analysis of Complex Biological Samples. *Curr. Anal. Chem.* 9, 199–211

5.. Meier F, Brunner AD,Frank M, Ha A, Voytik E, Kaspar-Schoenefeld S, Lubeck M,Raether O,Aebersold R, Collins BC, Röst HL, Mann M. (2019) Parallel accumulation – serial fragmentation combined with data-independent acquisition (diaPASEF): Bottom-up proteomics with near optimal ion usage. BioRxiv. https://doi.org/10.1101/656207

5. Bradner, S. (1997) *Key words for use in RFCs to Indicate Requirement Levels* (RFC Editor)