## 1D NMR FID & Spectrum Raw Data Representation

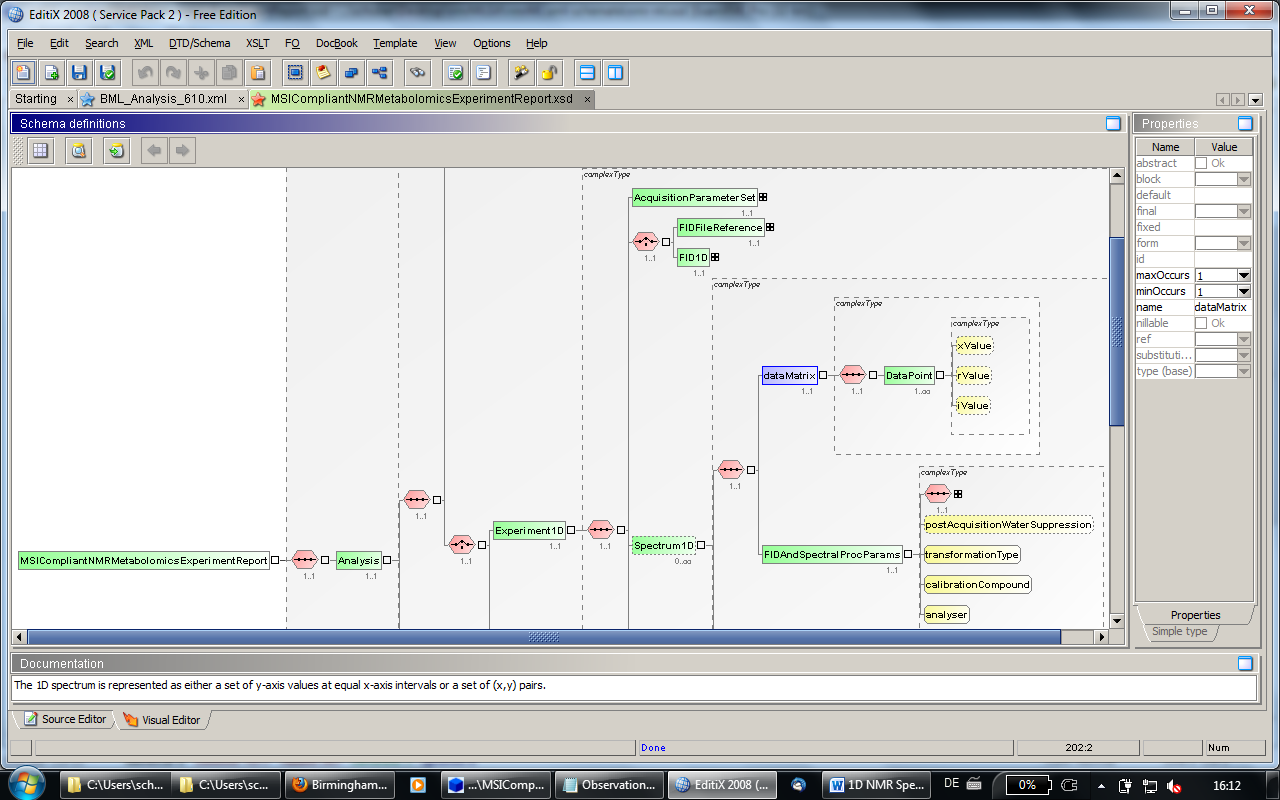
## in the Rubtsov, Wishard and PSI mzML XSDs

Here we try to investigate and compare modules and Substructures in present nmr ml representations. It is also a test on what tools can be helpful to compare xml and xsd structures of our different nmrML lineages and visualize those design differences. Here, I have compared the way the nmr raw data (FID and 1D Spectrum) is represented in the two nmrML predecessors (the Rubtsov and Wishard MLs) and how (mass-) spectral data is captured in PSI mzML. For each storage format, I show the corresponding xsd sections in a graphical way, the xsd code section and a snippet of an xml example for concrete illustration.

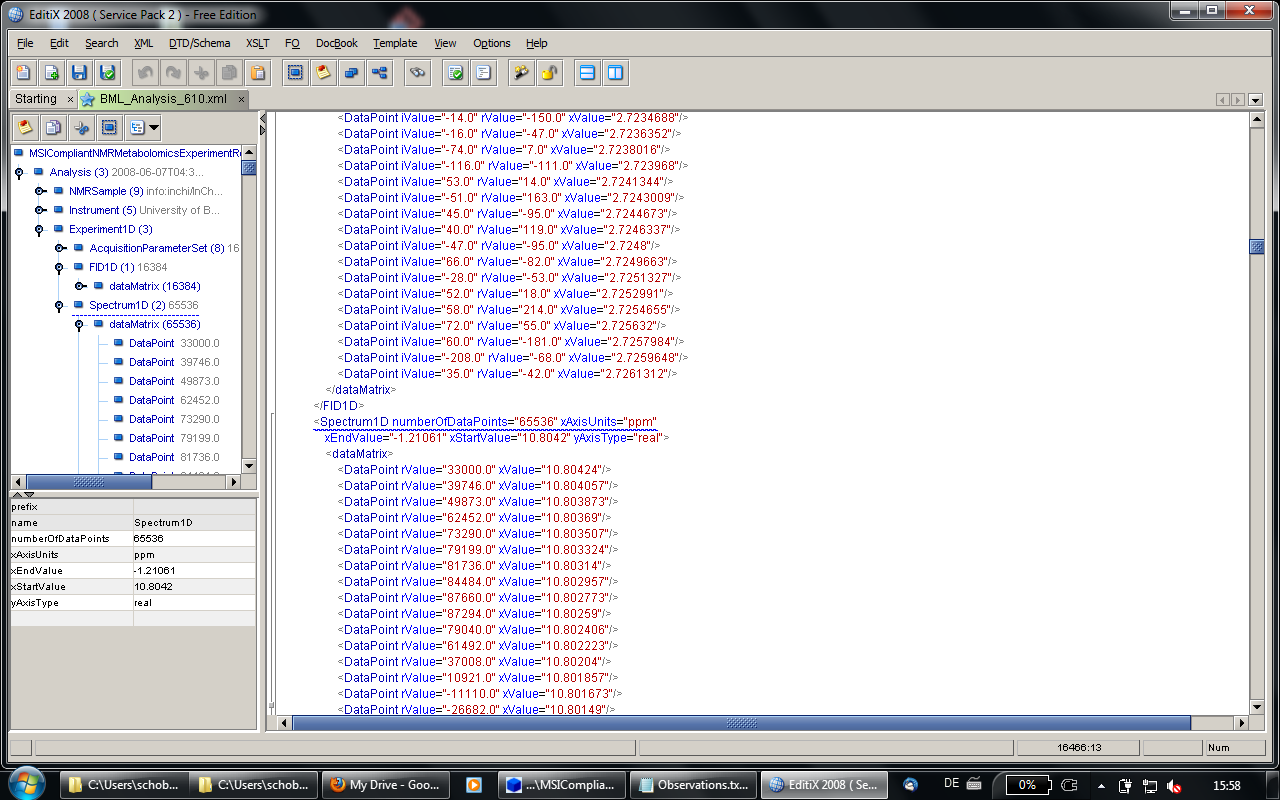
We can now discuss the pros and cons of a particular design decision, i.e. the representation of the numeric raw data for FIDs and spectra.

Birmingham NMR:

Xsd:



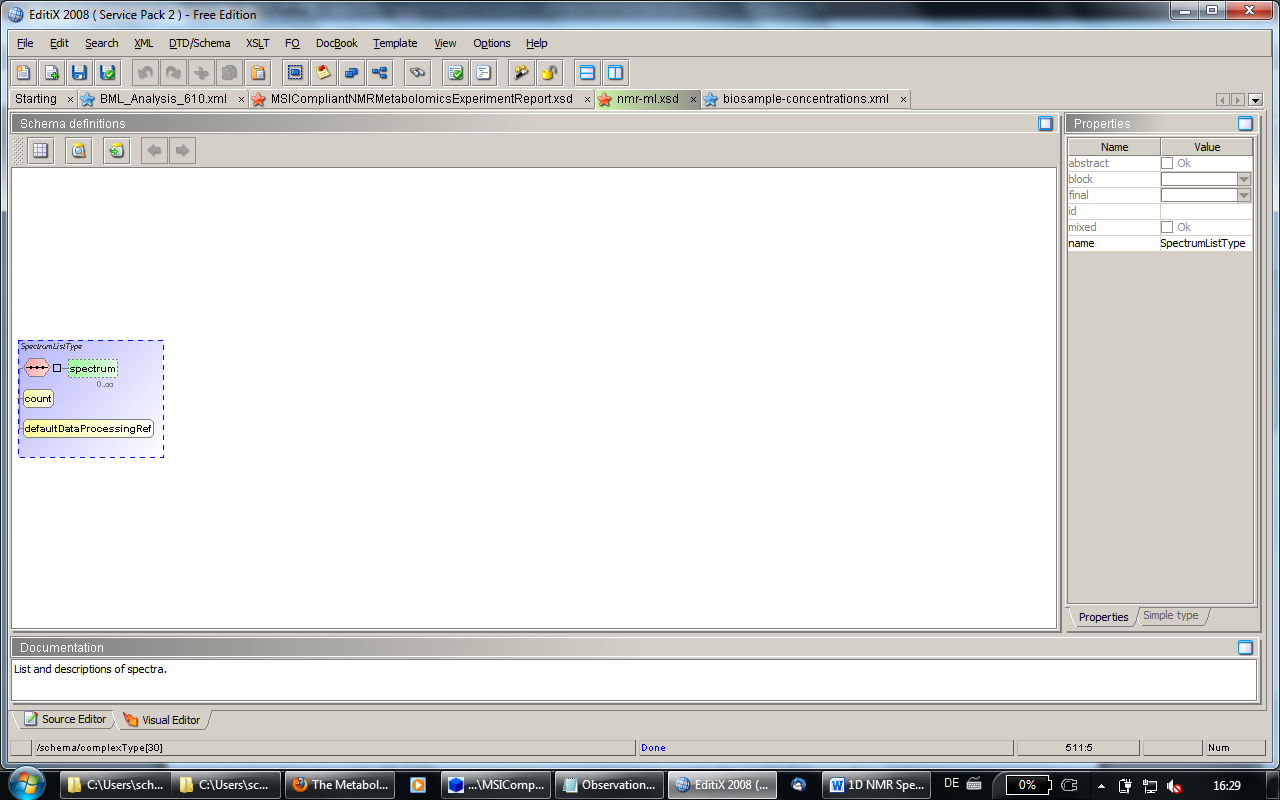
This is how JCamp does the raw data capturing. The advantage is that it is directly human readable. The disadvantage is that it is not compressed and takes a lot of space in the file, which makes it unhandy and bulky. As an alternative option, the data could be stored as an external file locally   
or on a server. There's a field fidFileRef that is supposed to hold a reference to that  file in the form of Uniform Resource Identifier.

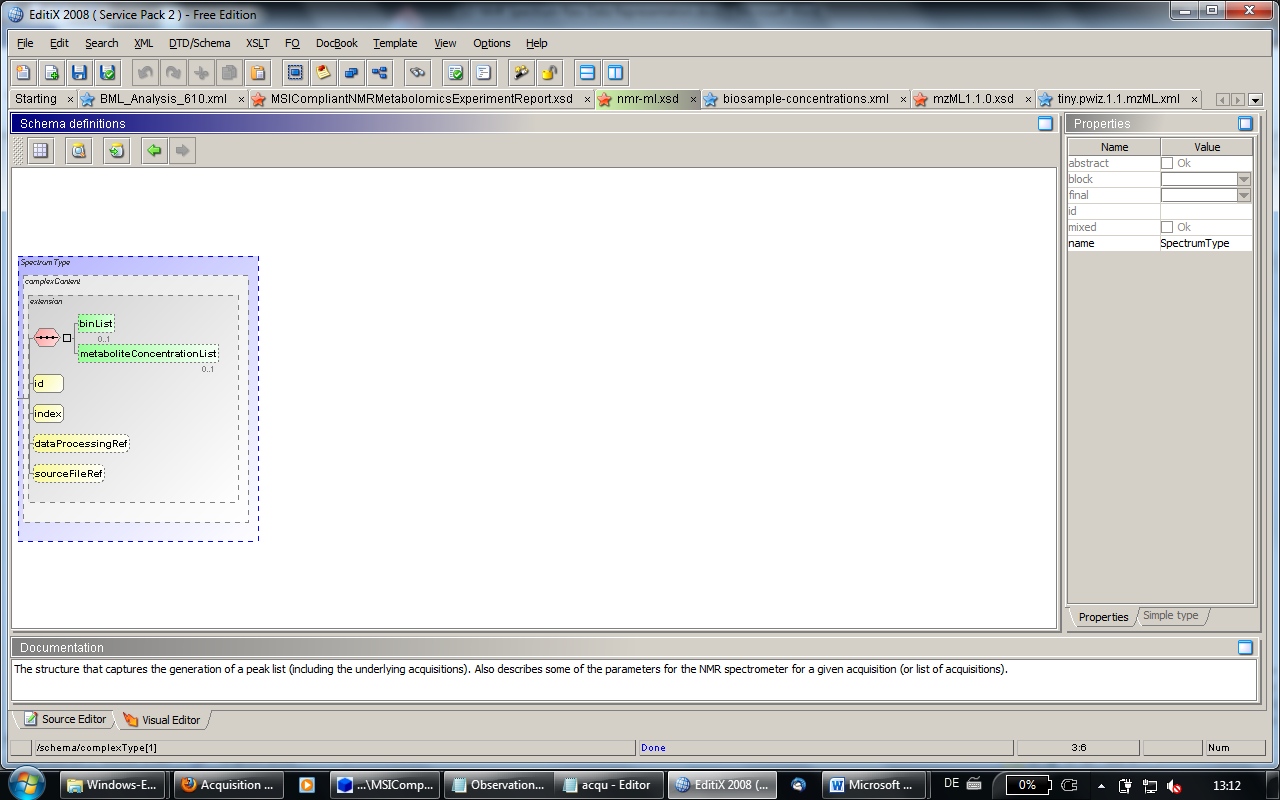
xml example:  


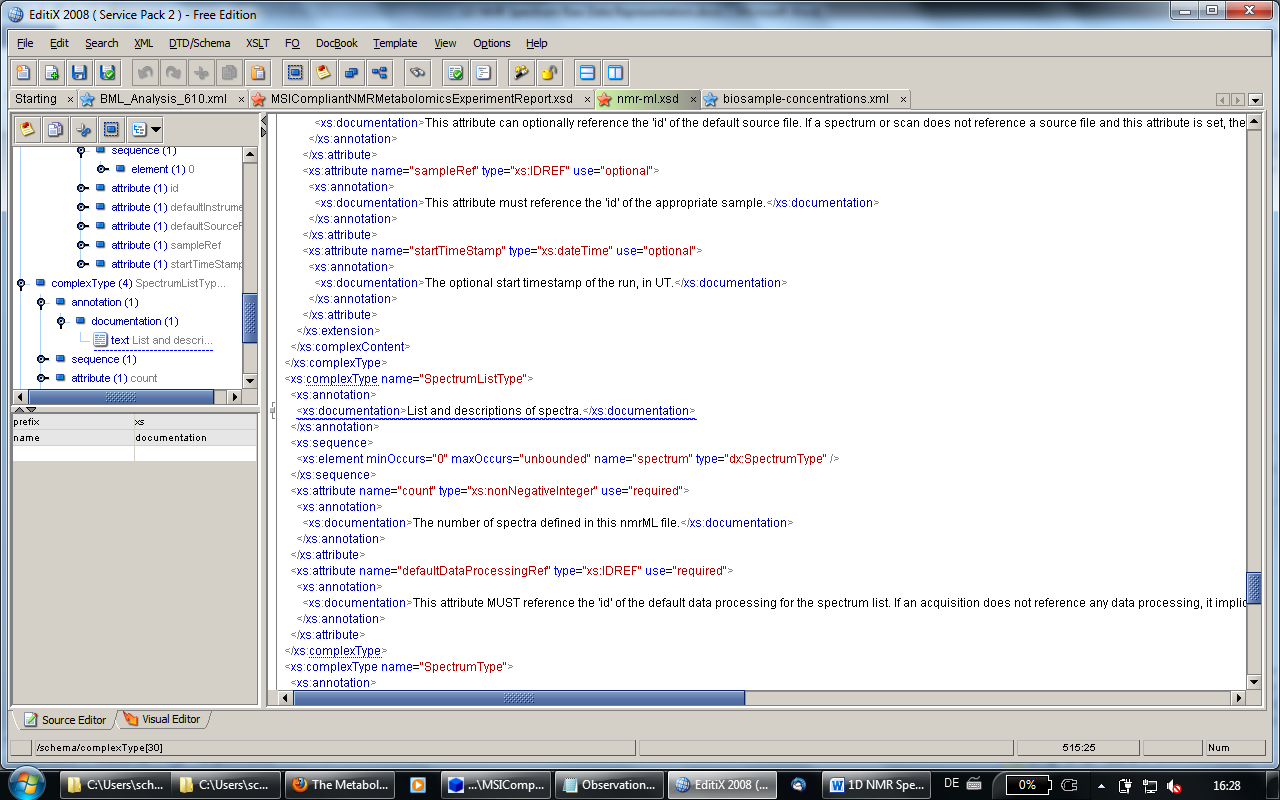
This large pile of numeric raw data should be captured as binary in our new version of the nmrML.

Wishard NMR:

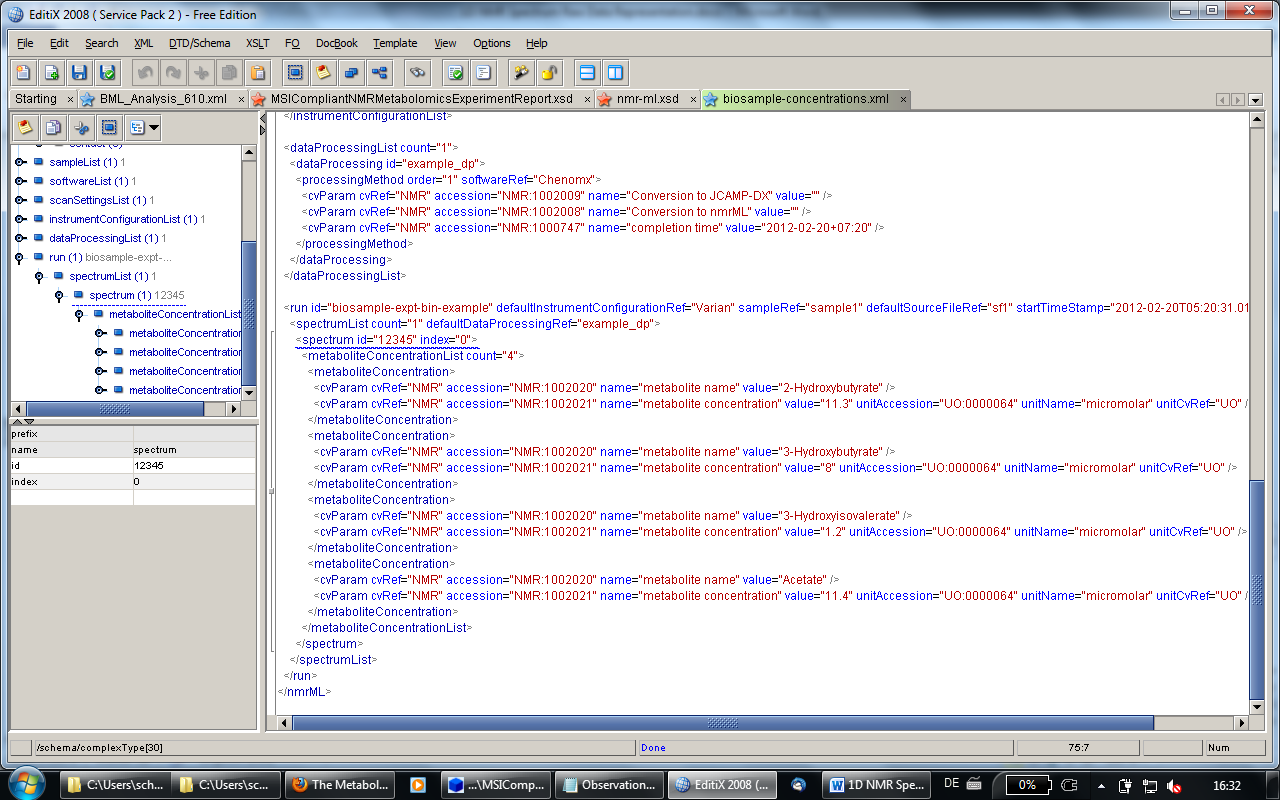
Xsd:



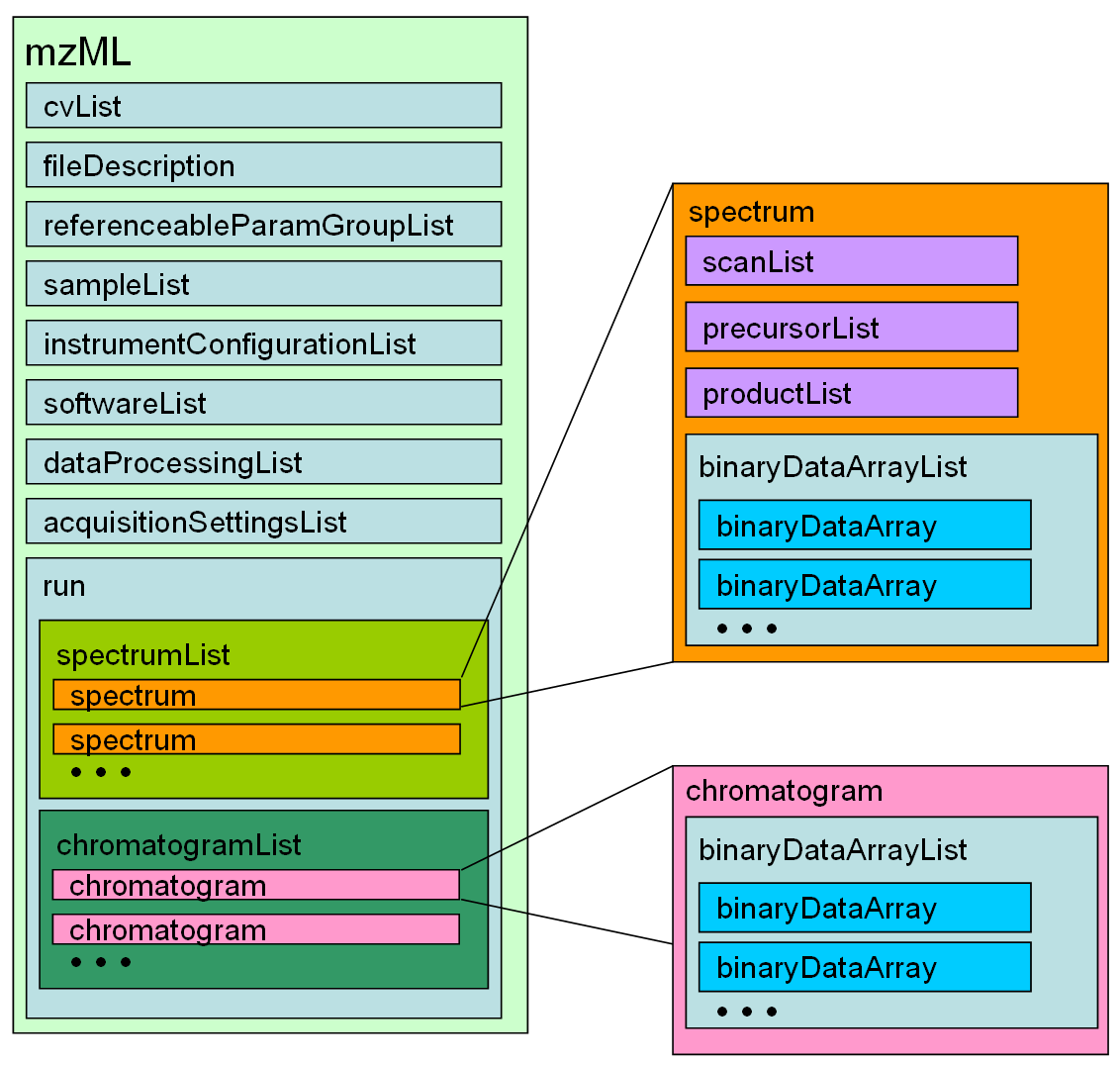


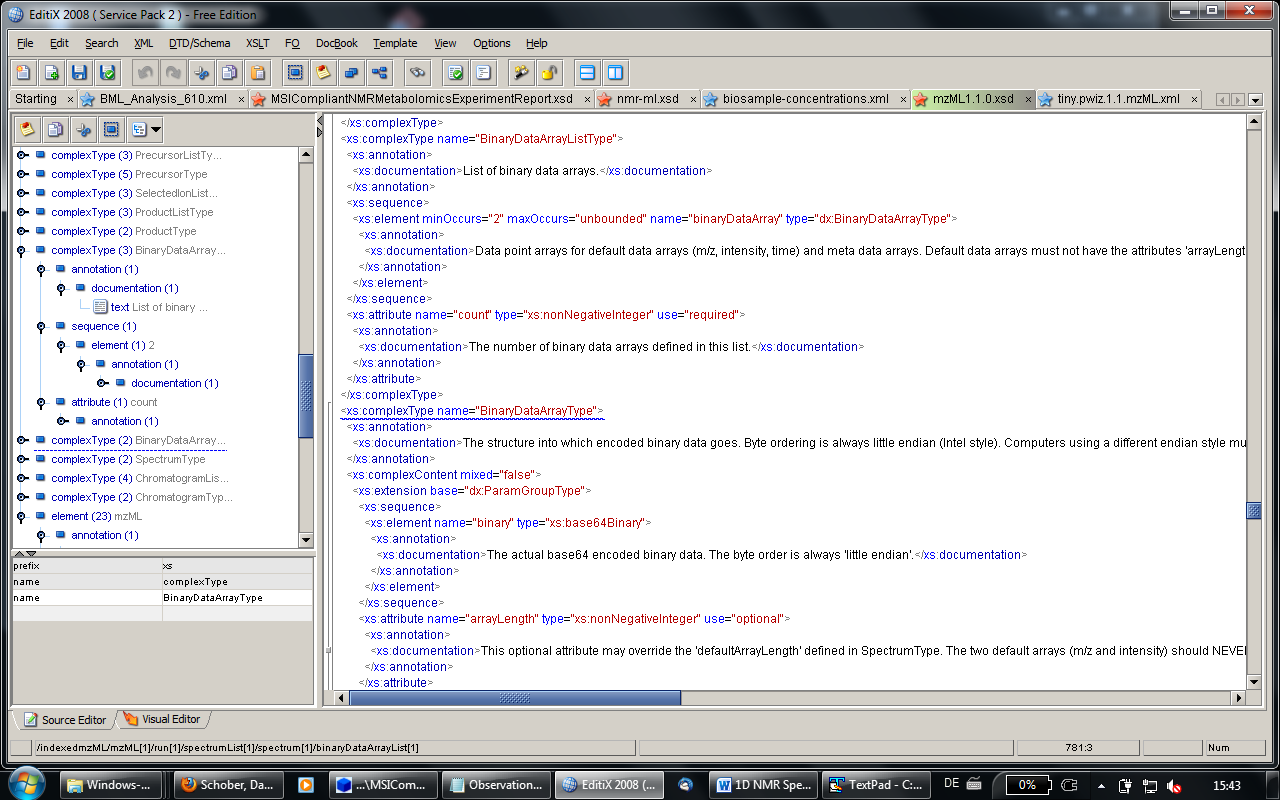


Xml example:

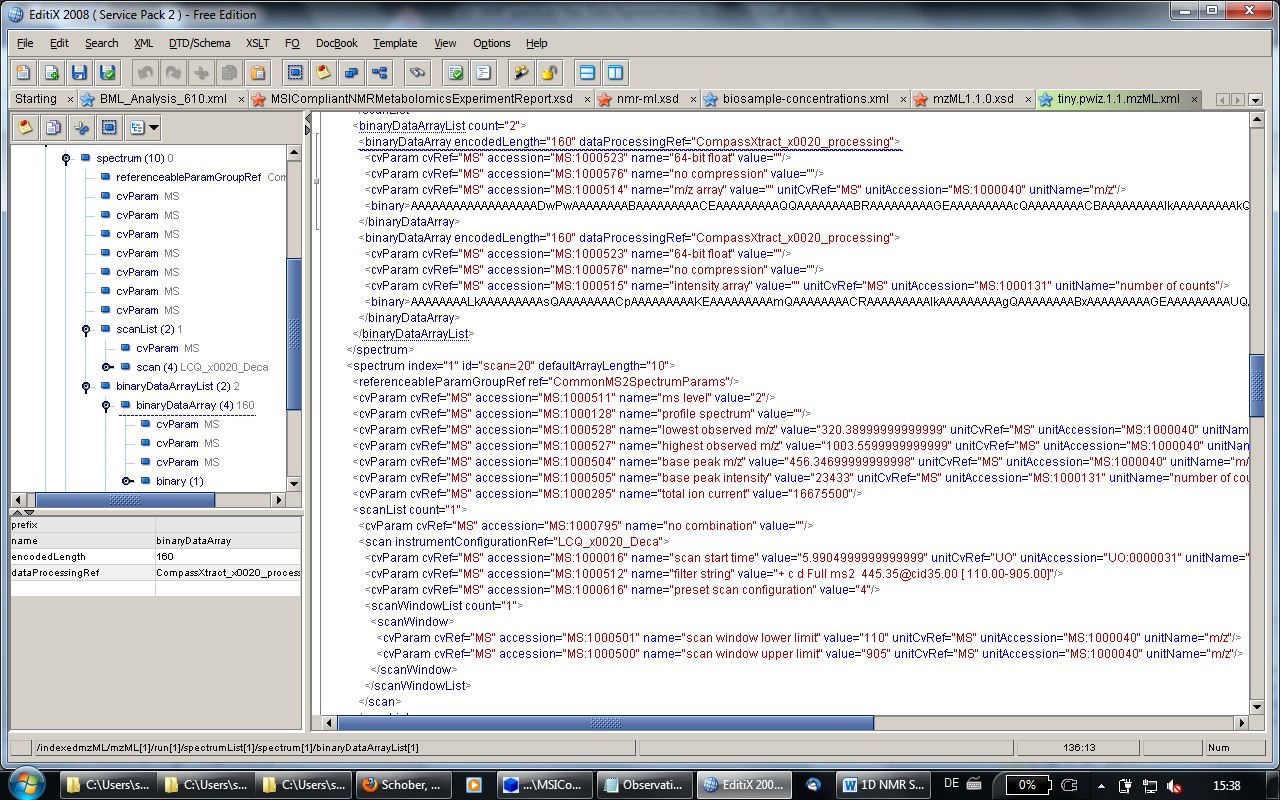


Here is what PSI did in its mzML schema:





In the xml example they have the raw data in as a 64bit encoded. This is much more compact yet still reproduceable.



## Element <binaryDataArrayList>

|  |  |
| --- | --- |
| **Definition:** | List of binary data arrays. |
| **Type:** | dx:BinaryDataArrayListType |
| **Attributes:** | |  |  |  |  | | --- | --- | --- | --- | | **Attribute Name** | **Data Type** | **Use** | **Definition** | | count | xs:nonNegativeInteger | required | The number of binary data arrays defined in this list. | |
| **Subelements:** | |  |  |  |  | | --- | --- | --- | --- | | **Subelement Name** | **min** | **max** | **Definition** | | [binaryDataArray](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#binaryDataArray) | 2 | unlim | Data point arrays for default data arrays (m/z, intensity, time) and meta data arrays. Default data arrays MUST not have the attributes 'arrayLength' and 'dataProcessingRef'. | |
| **Graphical Context:** | mzML_figure_018_binaryDataArrayList |
| **Example Context:** | <binaryDataArrayList count="2">  <binaryDataArray encodedLength="160" dataProcessingRef="XcaliburProcessing">  <cvParam cvRef="MS" accession="MS:1000523" name="64-bit float" value=""/>  <cvParam cvRef="MS" accession="MS:1000576" name="no compression" value=""/>  <cvParam cvRef="MS" accession="MS:1000514" name="m/z array" value="" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <binary>AAAAAAAAAAAAAAAAAADwPwAAAAAAAABAAAAAAAAACEAAAAAAAA...</binary>  </binaryDataArray>  ...  </binaryDataArrayList> |

## Element <binaryDataArray>

|  |  |
| --- | --- |
| **Definition:** | Data point arrays for default data arrays (m/z, intensity, time) and meta data arrays. Default data arrays MUST not have the attributes 'arrayLength' and 'dataProcessingRef'. |
| **Type:** | dx:BinaryDataArrayType |
| **Attributes:** | |  |  |  |  | | --- | --- | --- | --- | | **Attribute Name** | **Data Type** | **Use** | **Definition** | | arrayLength | xs:nonNegativeInteger | optional | This optional attribute may override the 'defaultArrayLength' defined in SpectrumType. The two default arrays (m/z and intensity) should NEVER use this override option, and should therefore adhere to the 'defaultArrayLength' defined in SpectrumType. Parsing software can thus safely choose to ignore arrays of lengths different from the one defined in the 'defaultArrayLength' SpectrumType element. | | dataProcessingRef | xs:IDREF | optional | This optional attribute may reference the 'id' attribute of the appropriate dataProcessing. | | encodedLength | xs:nonNegativeInteger | required | The encoded length of the binary data array. | |
| **Subelements:** | |  |  |  |  | | --- | --- | --- | --- | | **Subelement Name** | **min** | **max** | **Definition** | | [referenceableParamGroupRef](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#referenceableParamGroupRef) | 0 | unlim | A reference to a previously defined ParamGroup, which is a reusable container of one or more cvParams. | | [cvParam](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#cvParam) | 0 | unlim | This element holds additional data or annotation. Only controlled values are allowed here. | | [userParam](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#userParam) | 0 | unlim | Uncontrolled user parameters (essentially allowing free text). Before using these, one should verify whether there is an appropriate CV term available, and if so, use the CV term instead | | [binary](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#binary) | 1 | 1 | The actual base64 encoded binary data. The byte order is always 'little endian'. | |
| **Example Context:** | <binaryDataArray encodedLength="160" dataProcessingRef="XcaliburProcessing">  <cvParam cvRef="MS" accession="MS:1000523" name="64-bit float" value=""/>  <cvParam cvRef="MS" accession="MS:1000576" name="no compression" value=""/>  <cvParam cvRef="MS" accession="MS:1000515" name="intensity array" value="" unitCvRef="MS" unitAccession="MS:1000131" unitName="number of counts"/>  <binary>AAAAAAAALkAAAAAAAAAsQAAAAAAAACpAAAAAAAAAKEAAAAAAAA...</binary>  </binaryDataArray> |
| **cvParam Mapping Rules:** | Path mzML/run/chromatogramList/chromatogram/binaryDataArrayList/binaryDataArray  MUST supply a \*child\* term of MS:1000572 (binary data compression type) only once  e.g.: MS:1000574 (zlib compression)  e.g.: MS:1000576 (no compression)  MUST supply a \*child\* term of MS:1000513 (binary data array) only once  e.g.: MS:1000514 (m/z array)  e.g.: MS:1000515 (intensity array)  e.g.: MS:1000516 (charge array)  e.g.: MS:1000517 (signal to noise array)  e.g.: MS:1000595 (time array)  e.g.: MS:1000617 (wavelength array)  e.g.: MS:1000786 (non-standard data array)  e.g.: MS:1000820 (flow rate array)  e.g.: MS:1000821 (pressure array)  e.g.: MS:1000822 (temperature array)  MUST supply a \*child\* term of MS:1000518 (binary data type) only once  e.g.: MS:1000521 (32-bit float)  e.g.: MS:1000523 (64-bit float)  Path mzML/run/spectrumList/spectrum/binaryDataArrayList/binaryDataArray  MUST supply a \*child\* term of MS:1000572 (binary data compression type) only once  e.g.: MS:1000574 (zlib compression)  e.g.: MS:1000576 (no compression)  MUST supply a \*child\* term of MS:1000513 (binary data array) only once  e.g.: MS:1000514 (m/z array)  e.g.: MS:1000515 (intensity array)  e.g.: MS:1000516 (charge array)  e.g.: MS:1000517 (signal to noise array)  e.g.: MS:1000595 (time array)  e.g.: MS:1000617 (wavelength array)  e.g.: MS:1000786 (non-standard data array)  e.g.: MS:1000820 (flow rate array)  e.g.: MS:1000821 (pressure array)  e.g.: MS:1000822 (temperature array)  MUST supply a \*child\* term of MS:1000518 (binary data type) only once  e.g.: MS:1000521 (32-bit float)  e.g.: MS:1000523 (64-bit float) |
| **Example cvParams:** | <cvParam cvRef="MS" accession="MS:1000523" name="64-bit float" value=""/>  <cvParam cvRef="MS" accession="MS:1000576" name="no compression" value=""/>  <cvParam cvRef="MS" accession="MS:1000514" name="m/z array" value="" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000515" name="intensity array" value="" unitCvRef="MS" unitAccession="MS:1000131" unitName="number of counts"/>  <cvParam cvRef="MS" accession="MS:1000595" name="time array" value="" unitCvRef="UO" unitAccession="UO:0000010" unitName="second"/>  <cvParam cvRef="MS" accession="MS:1000521" name="32-bit float"/>  <cvParam cvRef="MS" accession="MS:1000574" name="zlib compression" value=""/> |
| **Notes and Constraints:** | The arrayLength attribute need only be specified if it is different from the defaultArrayLength specified in the <spectrum> element. |

## Element <spectrum>

|  |  |
| --- | --- |
| **Definition:** | The structure that captures the generation of a peak list (including the underlying acquisitions). Also describes some of the parameters for the mass spectrometer for a given acquisition (or list of acquisitions). |
| **Type:** | dx:SpectrumType |
| **Attributes:** | |  |  |  |  | | --- | --- | --- | --- | | **Attribute Name** | **Data Type** | **Use** | **Definition** | | dataProcessingRef | xs:IDREF | optional | This attribute can optionally reference the 'id' of the appropriate dataProcessing. | | defaultArrayLength | xs:int | required | Default length of binary data arrays contained in this element. | | id | xs:string (pattern: \S+=\S+( \S+=\S+)\*) | required | The native identifier for a spectrum. For unmerged native spectra or spectra from older open file formats, the format of the identifier is defined in the PSI-MS CV and referred to in the mzML header. External documents may use this identifier together with the mzML filename or accession to reference a particular spectrum. | | index | xs:nonNegativeInteger | required | The zero-based, consecutive index of the spectrum in the SpectrumList. | | sourceFileRef | xs:IDREF | optional | This attribute can optionally reference the 'id' of the appropriate sourceFile. | | spotID | xs:string | optional | The identifier for the spot from which this spectrum was derived, if a MALDI or similar run. | |
| **Subelements:** | |  |  |  |  | | --- | --- | --- | --- | | **Subelement Name** | **min** | **max** | **Definition** | | [referenceableParamGroupRef](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#referenceableParamGroupRef) | 0 | unlim | A reference to a previously defined ParamGroup, which is a reusable container of one or more cvParams. | | [cvParam](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#cvParam) | 0 | unlim | This element holds additional data or annotation. Only controlled values are allowed here. | | [userParam](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#userParam) | 0 | unlim | Uncontrolled user parameters (essentially allowing free text). Before using these, one should verify whether there is an appropriate CV term available, and if so, use the CV term instead | | [scanList](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#scanList) | 0 | 1 | List and descriptions of scans. | | [precursorList](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#precursorList) | 0 | 1 | List and descriptions of precursor isolations to the spectrum currently being described, ordered. | | [productList](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#productList) | 0 | 1 | List and descriptions of product isolations to the spectrum currently being described, ordered. | | [binaryDataArrayList](http://www.peptideatlas.org/tmp/mzML1.1.0_plain.html#binaryDataArrayList) | 0 | 1 | List of binary data arrays. | |
| **Graphical Context:** | mzML_figure_014_spectrum |
| **Example Context:** | <spectrum index="3" id="scan=22" spotID="A1,42x42,4242x4242" defaultArrayLength="15">  <referenceableParamGroupRef ref="CommonMS1SpectrumParams"/>  <cvParam cvRef="MS" accession="MS:1000511" name="ms level" value="1"/>  <cvParam cvRef="MS" accession="MS:1000127" name="centroid spectrum" value=""/>  <cvParam cvRef="MS" accession="MS:1000528" name="lowest observed m/z" value="142.38999999999999" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000527" name="highest observed m/z" value="942.55999999999995" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000504" name="base peak m/z" value="422.42000000000002" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  ...  </spectrum> |
| **cvParam Mapping Rules:** | Path mzML/run/spectrumList/spectrum  MAY supply a \*child\* term of MS:1000465 (scan polarity) only once  e.g.: MS:1000129 (negative scan)  e.g.: MS:1000130 (positive scan)  MUST supply a \*child\* term of MS:1000559 (spectrum type) only once  e.g.: MS:1000322 (charge inversion mass spectrum)  e.g.: MS:1000325 (constant neutral gain spectrum)  e.g.: MS:1000326 (constant neutral loss spectrum)  e.g.: MS:1000328 (e/2 mass spectrum)  e.g.: MS:1000341 (precursor ion spectrum)  e.g.: MS:1000581 (CRM spectrum)  e.g.: MS:1000582 (SIM spectrum)  e.g.: MS:1000583 (SRM spectrum)  e.g.: MS:1000789 (enhanced multiply charged spectrum)  e.g.: MS:1000790 (time-delayed fragmentation spectrum)  et al.  MUST supply term MS:1000525 (spectrum representation) or any of its children only once  e.g.: MS:1000127 (centroid spectrum)  e.g.: MS:1000128 (profile spectrum)  MAY supply a \*child\* term of MS:1000499 (spectrum attribute) one or more times  e.g.: MS:1000285 (total ion current)  e.g.: MS:1000497 (zoom scan)  e.g.: MS:1000504 (base peak m/z)  e.g.: MS:1000505 (base peak intensity)  e.g.: MS:1000511 (ms level)  e.g.: MS:1000527 (highest observed m/z)  e.g.: MS:1000528 (lowest observed m/z)  e.g.: MS:1000618 (highest observed wavelength)  e.g.: MS:1000619 (lowest observed wavelength)  e.g.: MS:1000796 (spectrum title)  et al. |
| **Example cvParams:** | <cvParam cvRef="MS" accession="MS:1000580" name="MSn spectrum"/>  <cvParam cvRef="MS" accession="MS:1000511" name="ms level" value="2"/>  <cvParam cvRef="MS" accession="MS:1000127" name="centroid spectrum"/>  <cvParam cvRef="MS" accession="MS:1000130" name="positive scan"/>  <cvParam cvRef="MS" accession="MS:1000285" name="total ion current" value="1.0289517E7"/>  <cvParam cvRef="MS" accession="MS:1000128" name="profile spectrum" value=""/>  <cvParam cvRef="MS" accession="MS:1000504" name="base peak m/z" value="810.415283203125" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000505" name="base peak intensity" value="1471973.875" unitCvRef="MS" unitAccession="MS:1000131" unitName="number of counts"/>  <cvParam cvRef="MS" accession="MS:1000528" name="lowest observed m/z" value="200.00018816645022" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000527" name="highest observed m/z" value="2000.0099466203771" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/>  <cvParam cvRef="MS" accession="MS:1000326" name="constant neutral loss spectrum"/> |
| **Notes and Constraints:** | id's MUST be unique within a file as constrained by a primary key. The format MUST follow the native ID guidelines for mzML If a scan yields no peaks, it should still be reported, but with a defaultArrayLength of 0 and no <binaryDataArrayList> element. |

--

Comparison in another tool (ExamXMLpro):

