COMPUTIS ML - a standardised data format for imaging mass spectrometry

Thorsten Schramm, Alfons Hester, Ivo Klinkert, Ron M. A. Heeren, Markus Stoeckli, Jean-Pierre Both, A. Brunelle, Bernhard Spengler, Andreas Römpp

COMPUTIS ML - a standardised data format for imaging mass spectrometry

Thorsten Schramm[1], Alfons Hester[1], Ivo Klinkert[2], Ron M. A. Heeren[2], Markus Stöckli[3], Jean-Pierre Both[4], Alain Brunelle[5], Bernhard Spengler[1], Andreas Römpp[1]

[1]Institute of Inorganic and Analytical Chemistry, Justus Liebig University (JLU), D-35392 Giessen, Germany.
[2]FOM Institute for Atomic and Molecular Physics (AMOLF), 1009 DB Amsterdam, The Netherlands
[3]Novartis Pharma AG, Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland
[4]Commissariat à l'Énergie Atomique (CEA), Salclay, France
[5]Centre National de la Recherche Scientifique (CNRS), 91190 Gif sur Yvette, France

Introduction

Problem:

- EU sponsored project COMPUTIS: development of new technologies for molecular imaging mass spectrometry including comparison of different instruments and processing techniques
- every institute uses its own software and thus its own data format
- data formats are defined depending on the nature of the data e.g. continuous or event based data
- developing software to convert data into the data format of every participant is very time consuming

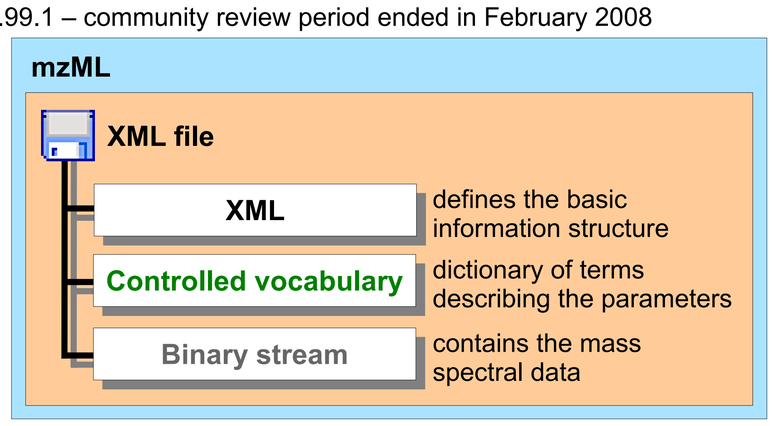
Solution:

- definition of a common data format every participant can read/write with his own software
- mzML is a good basis to solve this problem
- COMPUTIS ML: an extended version of mzML to include additional parameters for imaging mass spectrometry

mzML

History:

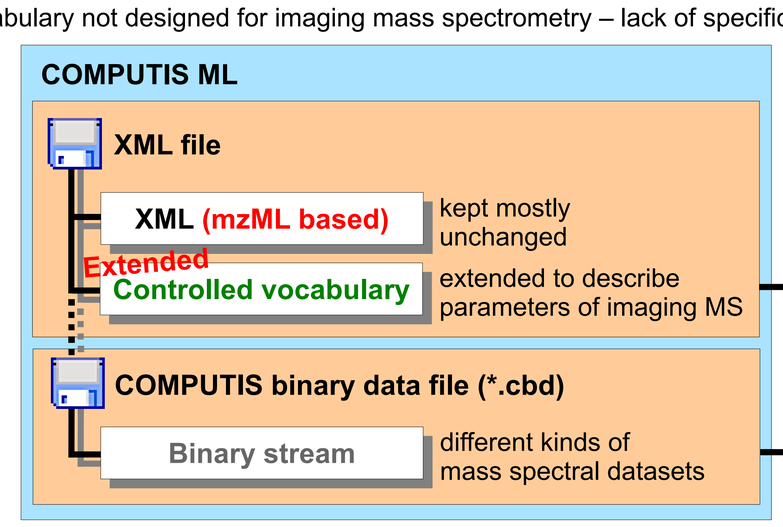
- mzML is the successor of mzData and mzXML – both XML based but not compatible
- developed in cooperation of HUPO-PSI (Proteome Standard Initiative) and ISB (Institute of Systems Biology, Seattle)
- latest version 0.99.1 – community review period ended in February 2008



COMPUTIS ML

Why a new format?

- very large imaging datasets (several gigabytes) → long processing times
- controlled vocabulary not designed for imaging mass spectrometry – lack of specific parameters



Features of COMPUTIS ML:

- XML structure of mzML – kept mostly unchanged
- extended controlled vocabulary – includes additional information needed to describe parameters of imaging experiments
- external binary file (COMPUTIS binary data, *.cbd) for mass spectral data → faster access and processing times

Advantages:

- easy conversion into mzML files
- offers possibility to use any software able to read/write mzML data files
  - commercial and non-commercial software e.g. visualisation, (pre)processing tools etc.
  - protein databases will offer a mzML upload API for easy communication

XML part of a COMPUTIS ML file:

```
...
<binaryFile path="C:\Data\Peptides\Sample1104.cbd"/>          link to external binary file
...
<spectrum id="S19" scanNumber="19" msLevel="1">
 <cvParam cvLabel="MS" accession="MS:1000580" name="MSn spectrum" value=""/>
 <spectrumDescription>
  <cvParam cvLabel="MS" accession="MS:1000127" name="centroid mass spectrum" value=""/>
  <cvParam cvLabel="MS" accession="MS:1000528" name="lowest m/z value" value="400.39"/>
  <scan  instrumentRef="LCQ Deca">                            Controlled vocabulary
...
    <selectionWindowList count="1">
...
    </selectionWindowList>                                    <XML structure/>
  </scan>
 </spectrumDescription>
 <binaryDataArray dataProcessingRef="Xcalibur Processing">
...
  <binary>1258</binary>                                       offset of spectrum in binary file
 </binaryDataArray>
...
```

Extended Controlled vocabulary

What is a controlled vocabulary (cv)?

Controlled vocabulary schemes mandate the uses of predefined, authorised terms that have been preselected by the designer of the controlled vocabulary as opposed to natural language vocabularies where there is no restriction on the vocabulary that can be used. [...] controlled vocabularies reduce ambiguity inherent in normal human languages where the same concept can be given different names and ensure consistency. (www.wikipedia.org/wiki/Controlled_vocabulary)

How to use a controlled vocabulary?

- declaration of the used controlled vocabulary (including the URI where the definition file can be found) in the XML part of the COMPUTIS ML file

```
<cvList count="1">
   <cv cvLabel="MS" fullName="Proteomics Standards Initiative Mass Spectrometry Ontology" version="2.0.2"
      URI="http://psidev.sourceforge.net/ms/xml/mzdata/psi-ms.2.0.2.obo"/>
</cvList>
```

- *.obo is a Open Biomedical Ontology – an example of a controlled vocabulary

Example of a controlled vocabulary parameter description:

```
<cvParam cvLabel="MS"                  short tag as defined in the <cvList> in the COMPUTIS ML file
         accession="MS:1000528"        accession number of the parameter
         name="lowest m/z value"       actual name of the parameter
         value="400.39"                value of the parameter, may be absent, if not applicable
/>
```

Modifications of the controlled vocabulary?

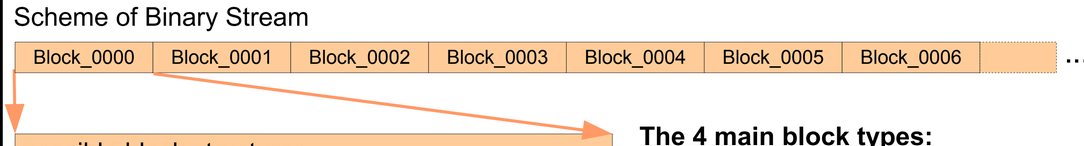
- controlled vocabulary is extended to meet the demands of imaging mass spectrometry
- new entries to describe the special parameters of spectra in images e.g.:
  - image dimensions in µm and pixels
  - scan patterns and positions of the sample stage and/or ion beam etc.
- COMPUTIS is in contact with PSI to make these changes part of mzML

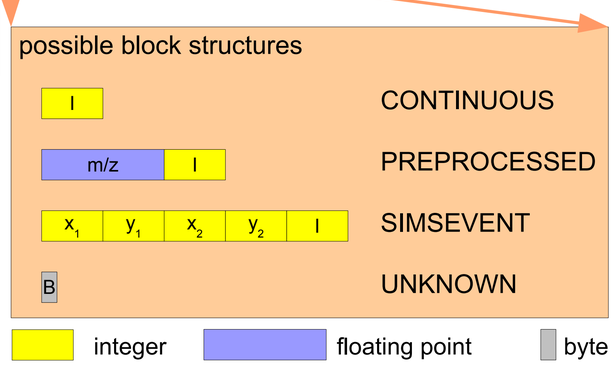
Binary stream

Binary file and COMPUTIS ML file:

- a binary file is a stream of bytes with no textual information, let us call it binary stream
- all spectra are stored in one binary file so this binary stream can be seen as a sequence of spectra
- each single information of a spectrum ist stored in blocks of equal length and equal inner structure
- therefore a binary stream is a sequence of equally structured blocks of equal length
- the binary stream has to be interpreted (block length and inner structure of each block)
- a binary file always belongs to one unique XML file (e.g. inner block structure is stored in the XML file)
- fast access to some (specific) spectra is performed by a table with spectrum offsets (in the XML file)

Scheme of Binary Stream





The 4 main block types:

- COMPUTIS partners agreed on 3 main data types :
  - CONTINUOUS : sequence of intensities (= I)
  - PREPROCESSED : seq. of [m/z, I]-pairs
  - SIMSEVENT : seq of (x_1,y_1,x_2,y_2,I)-quintuples
  - UNKNOWN : sequence of bytes (every file can be seen as a sequence of bytes)

How many differently structured binary blocks can be defined?

- integer values may be stored as byte (1 byte), word (2 bytes) or longword (4 bytes)
- floating point values may be stored as single or double (cf. IEEE 754)
- integers stored as word or longword might be stored in Big or Little Endian
- total amount of different block types is 1140 (5 continuous, 10 preprocessed and 1125 simsevent)

Outlook

Next steps:

- version 1.0 of the mzML data format by HUPO-PSI and ISB expected soon
- final adjustment of the COMPUTIS ML data format to adapt the changes made between versions 0.99.1 and 1.0 of mzML
- finalise the COMPUTIS controlled vocabulary for imaging purposes
- contribute the changes to the PSI
  - official extension of mzML to meet the demands of imaging purposes
  - official extension of the controlled vocabulary

Acknowledgements

This work was supported by the EU, COMPUTIS Project No. 518194