imzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data.

Thorsten Schramm, Alfons Hester, Ivo Klinkert, Jean-Pierre Both, Ron M A Heeren, Alain Brunelle, Olivier Laprévote, Nicolas Desbenoit, Marie-France Robbe, Markus Stoeckli, et al.

Elsevier Editorial System(tm) for Journal of Proteomics
Manuscript Draft

Manuscript Number: JPROT-D-12-00290R1

Title: imzML - a common data format for the flexible exchange and processing of mass spectrometry imaging data

Article Type: SI: Mass Spectrometry Imaging

Section/Category: Technical Note

Keywords: mass spectrometry imaging
data format
data processing
standardization

Corresponding Author: Dr Andreas Roempp, PhD

Corresponding Author's Institution: Justus Liebig University

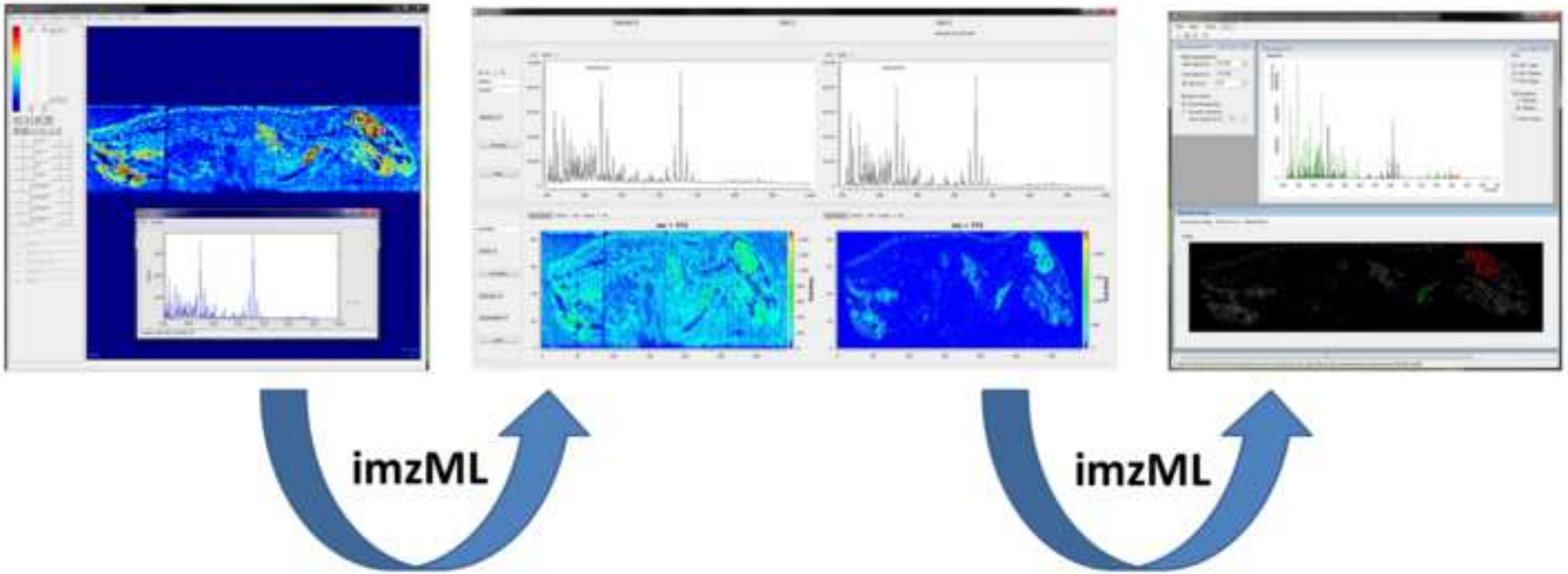
First Author: Thorsten Schramm

Order of Authors: Thorsten Schramm; Alfons Hester; Ivo Klinkert; Jean-Pierre Both; Ron M Heeren; Alain Brunelle; Olivier Laprevote; Marie-France Robbe; Markus Stoeckli; Bernhard Spengler; Andreas Roempp, PhD

Abstract: The application of MS imaging is rapidly growing with a constantly increasing number of different instrumental systems and software tools. The data format imzML was developed to allow the flexible and efficient exchange of MS imaging data between different instruments and data analysis software. imzML data is divided in two files which are linked by a universally unique identifier (UUID). Experimental details are stored in an XML file which is based on the HUPO-PSI format mzML. Information is provided in the form of a 'controlled vocabulary' (CV) in order to unequivocally describe the parameters and to avoid redundancy in nomenclature. Mass spectral data are stored in a binary file in order to allow efficient storage. imzML is supported by a growing number of software tools. Users are no longer limited to proprietary software, but are able to use the processing software best suited for a specific question or application. MS imaging data from different instruments can be converted to imzML and displayed with identical parameters in one software package for easier comparison. All technical details necessary to implement imzML and additional background information is available at www.imzml.org.

*Graphical Abstract (for review)

Click here to download high resolution image



***Highlights (for review)**

Highlights

- Common format for exchange of mass spectrometry imaging data
- Comprehensive and efficient description of MS imaging experiments
- Converters and data analysis tools are available
- Flexible structure allows adaption to new methods and instruments

imzML – a common data format for the flexible exchange and processing of mass spectrometry imaging data

Thorsten Schramm[1], Alfons Hester[1], Ivo Klinkert[2], Jean-Pierre Both[3], Ron M. A. Heeren[2], Alain Brunelle[4], Olivier Laprévote[5], Nicolas Desbenoit[1], Marie-France Robbe[3], Markus Stoeckli[6], Bernhard Spengler[1] and Andreas Römpp[1*]

[1]Justus Liebig University, Giessen, Germany; [2]FOM Institute AMOLF, Amsterdam, Netherlands; [3]CEA-LIST, Saclay, France; [4]Centre de recherche de Gif, Institut de Chimie des Substances Naturelles, CNRS, Gif-sur-Yvette Cedex, France; [5]Chimie et Toxicologie Analytique et Cellulaire, Université Paris Descartes, Paris, France; [6]Novartis Institutes for BioMedical Research, Basel, Switzerland

* Corresponding author

Dr. Andreas Roempp
Institute of Inorganic and Analytical Chemistry
- Analytical Chemistry -
Justus Liebig University Giessen
Schubertstrasse 60, Build. 16
D-35392 Giessen, Germany
phone: +49-641-99 34802
fax: +49-641-99 34809
email: Andreas.Roempp@anorg.Chemie.uni-giessen.de

Abstract

The application of mass spectrometry imaging (MS imaging) is rapidly growing with a constantly increasing number of different instrumental systems and software tools. The data format imzML was developed to allow the flexible and efficient exchange of MS imaging data between different instruments and data analysis software. imzML data is divided in two files which are linked by a universally unique identifier (UUID). Experimental details are stored in an XML file which is based on the HUPO-PSI format mzML. Information is provided in the form of a 'controlled vocabulary' (CV) in order to unequivocally describe the parameters and to avoid redundancy in nomenclature. Mass spectral data are stored in a binary file in order to allow efficient storage. imzML is supported by a growing number of software tools. Users will be no longer limited to proprietary software, but are able to use the processing software best suited for a specific question or application. MS imaging data from different instruments can be converted to imzML and displayed with identical parameters in one software package for easier comparison. All technical details necessary to implement imzML and additional background information is available at www.imzml.org.

Introduction

In mass spectrometry imaging (MS Imaging) a sample of interest is scanned and the resulting ion signals are reconstructed into a pixel image. Several overviews of this method have been published recently [1, 2].The application of MS imaging is rapidly growing with a constantly increasing number of different instrumental systems and software tools. This results in a need of exchangeability of MS imaging data between different instruments and data analysis software. As for many analytical methods, data processing has become an essential part of the workflow for MS imaging. MS Imaging data consist of several thousand spectra which frequently result in data files of several gigabytes. Mass spectra of one experiment are usually acquired with identical measurement parameters. Existing standards such as the DICOM standard [3] for in-vivo imaging data or the mzML standard [4] established by HUPO-PSI [5] are not suitable to adequately represent an MS imaging experiment. Therefore a common data format dedicated to mass spectrometry imaging data was developed within the EU project COMPUTIS (www.computis.org). The data format is in part based on the HUPO-PSI format mzML (see details below) and is thus called imzML for 'imaging mzML'.

This technical report provides an overview of the most important properties and features of the imzML format. Additional information and all technical details necessary to implement imzML is provided on the website **www.imzml.org** and in a recently published book chapter [6].

Methodology

MS imaging data in imzML is divided into two separate files (Figure 1). This structure, consisting of a small file (text or XML) and a larger (binary) file, has been shown to be efficient in previous data formats for MS imaging, for example in BioMap [7] and internal data formats at FOM Institute and Justus Liebig University. All metadata (e.g. instrumental parameters, sample details) are stored in an XML file. Mass spectral data is stored in a binary file to ensure efficient storage. Corresponding XML and binary files contain a universally unique identifier (UUID) [8] in order to link them unequivocally and to prevent loss of information.

The **XML file** (*.imzml) contains all relevant information about the MS imaging experiment. In order to stay as close as possible to existing formats this metadata file is based on the mass

spectrometry data standard mzML [4] developed by HUPO-PSI [9]. Discussions with HUPO-PSI resulted in the strategy to maintain imzML as a separate format, but keep the XML structure consistent with mzML. Information is provided in the form of a 'controlled vocabulary' (CV) which is stored in an open biomedical ontology [10]. This approach is used in order to unequivocally describe the parameters and to avoid redundancy in nomenclature. The mzML controlled vocabulary [11] has been supplemented by parameters which are necessary for a comprehensive description of MS imaging experiments. These additional imzML CV parameters (including x/y position, scan direction/pattern, pixel size) are stored in the imagingMS.obo file [12].

The **imaging binary data file** (*.ibd) contains the mass spectral data of the MS imaging measurement. In order to ensure efficient storage, two different binary modes are defined: continuous and processed. 'Continuous' means that an intensity value is stored for each m/z bin even if there is no measurable signal (resulting in an intensity of zero). This data structure is used for many MALDI-TOF mass spectrometers. As a result the m/z axis is identical for all mass spectra of one image. Therefore it is sufficient to store the m/z array only once in the binary file. This structure can reduce the file size significantly (up to a factor of 2). On the other hand mass spectra are often processed before they are stored e.g. for noise-reduction, peak-picking, deisotoping. This results in discontinuous and non-constant m/z arrays. In this case the m/z array has to be stored for each spectrum separately. This data structure was termed 'Processed'.

The two files (XML and binary) are connected by offset values in the XML file that indicate the position of the corresponding data in the binary file. This allows for fast reading/access of large data sets. The combination of non-redundant metadata representation and binary data leads to efficient storage. The resulting datasets are comparable in size to the proprietary data and the separate metadata file allows flexible handling of large datasets. This is important since MS imaging experiments typically consist of several thousand spectra, and increased file size due to format conversion could lead to data sets which are very difficult to handle.

imzML has been extensively discussed with academic and industrial users at various occasions in recent years. The current version 1.1.0 of imzML was released at the International Mass Spectrometry Conference (IMSC) in Bremen, Germany in September 2009. The flexible structure based on XML and the controlled vocabulary allows for compatibility with new instruments and methods in the future. Technical documentation

including XML scheme, mapping file, controlled vocabulary (obo file) and example files are provided on www.imzml.org.

Implementation and Discussion

A number of software tools already support imzML. This allows for choosing from a (growing) number of options to display and process MS imaging data. Users will be no longer limited to proprietary software, but are able to use the processing software best suited for a specific question or application. Multiple software tools can also be combined in one workflow using imzML as demonstrated in Figure 2. In this example a whole body rat section was coated with α-Cyano-4-hydroxycinnamic acid (CHCA) and analysed with a QStar mass spectrometer (AB Sciex). Mass spectra were acquired for 344 x 87 pixels (500 µm pixel size) in the mass range m/z = 300-1000 u. Data was saved in the Analyze 7.5 data format and opened in 'BioMap' for initial inspection (Figure 2A). Subsequently the dataset was converted to imzML using the 'Toimzml' converter developed by the CEA-LIST institute. The imzML file was opened in 'SpectViewer' (developed by CEA-LIST) and a wavelet denoising procedure was applied in order to improve image quality (Figure 2B). This particular processing function is not available in the other software tools. The processed data was saved in a new imzML file which was subjected to further analysis.

The 'Datacube Explorer' developed by AMOLF allows for fast browsing of MS images and can thus be used to visually search for signals that show interesting spatial features. The denoised imzML data set was displayed in Datacube Explorer and spectra from two regions of interest (ROI) were extracted (Figure 2C). The data analysis tool 'Mirion' developed by JLU can be used to semiautomatically generate MS images based on parameters such as pixel coverage or incremental mass. It can also be used to open multiple files as shown in Figure 2D (original and denoised imzML data set), data from these files can be combined for example by RGB overlay (not shown). The combination of processing steps within the four programmes is only possible due to the standardised and open data format imzML. A detailed description of the different software packages would be beyond the scope of this technical note. The purpose of the example shown in Figure 2 is to demonstrate the possibility to open and process imzML files in different MS imaging software packages. Details about functionalities of SpectViewer, Datacube Explorer and Mirion will be provided in separate publications which are currently in preparation.

The demonstrated workflow is only one example for the use of imzML for analysis of MSI data with different tools. Another application is the comparison of measurements from different instruments. Options for data processing (e.g. binning, normalization) and displaying (e.g. interpolation options, color schemes) vary strongly between different MS imaging software tools and can make a direct comparison of data difficult. This problem can be avoided, if all data sets are converted to imzML and the data is displayed with identical parameters in one software package. Examples of imzML data sets from different instrument platforms are shown in Figure 3. All images are displayed in the same software package (two masses in green and red, no interpolation). Data from the most commonly used matrix-assisted desorption/ionization (MALDI) mass spectrometers (Thermo, Bruker, AB Sciex and Waters) are shown in Figures 3A-D. A data set originating from secondary ion mass spectrometry (SIMS) is shown in Figure 3E. The majority of MS imaging applications are based on MALDI and SIMS, but other ionization techniques are increasingly used. Data acquisition in desorption electrospray ionization (DESI) mode is based on one data file per line scan and conversion to imzML therefore differs from most MALDI experiments where usually a single file is acquired per image. The 'RAW to imzML' converter developed by Justus Liebig University was modified to convert multiple line scans into a unified imzML file. An example of a DESI imzML file generated with this converter is shown in Figure 3F. This conversion tool has also been used in previous studies in combination with the Datacube Explorer for analyzing and displaying MS imaging data[13, 14]. More information on sample details, image properties and conversion method are provided in the 'imzML gallery' on www.imzml.org. This page also includes additional examples of MS images generated with imzML.

The examples discussed above illustrate the functionalities and possibilities of the imzML data format. A number of additional software tools are currently adopted for usage with imzML. This includes 'MSImageView' the successor of BioMap which is probably the most widely used independent software to analyze MS imaging data. The first two commercial software packages supporting imzML have been released recently [15, 16]. The concept of imzML is also actively supported by major vendors of instrumentation for MS imaging including Waters Corporation (Manchester, United Kingdom), Thermo Fisher Scientific (Bremen, Germany) and Bruker Daltonik (Bremen, Germany). Their data formats are integrated in the imzML workflow through export options in proprietary software suites (Waters) or by external converters being developed by the MS imaging community with technical support from the companies (Thermo, Bruker). An updated list of available tools

and workflows can be found on the imzML website. imzML was chosen as the central platform for exchange of data in the European COST network 'Mass Spectrometry Imaging: New Tools for Healthcare Research' which includes more than 30 MS imaging groups [17]. The future development of imzML-based software will also be coordinated in this framework. These activities ensure a wide distribution and active development of imzML as the data standard for mass spectrometry imaging data.

Conclusions

The data format imzML enables efficient and flexible exchange of MS imaging data. It allows for a flexible selection of data analysis tools. imzML is increasingly used in the MS imaging community. A number of software tools are available and many more are currently being adapted to imzML. A common data format is not a mere technical detail, but has significant impact on practical work: fully searchable mass spectrometry imaging data can be shared with collaborators in biological or clinical labs without being restricted to proprietary vendor software. imzML is thus an important step towards more efficient collaboration as well as more flexible and transparent data processing in mass spectrometry imaging. These are key requirements for MS imaging to finally be established as a routine method.

Acknowledgements

First and foremost we would like to thank all the members of the HUPO Proteomics Standards Initiative (PSI) for developing the data format mzML which has been used as a model for imzML. We are especially thankful to Lennart Martens, Eric Deutsch, Randall Julian and Martin Eisenacher for many helpful discussions in the last years. We would like to thank Sabine Guenther, Corinna Henkel, Brendan Prideaux, Emmanuelle Claude, Zoltan Takats for providing MS imaging data sets. This work was supported by the European Union (Contract LSHG-CT-2005-518194 COMPUTIS). JLU acknowledges support by the Hessian Ministry of Science and Art (LOEWE focus Ambiprobe).

References

[1] Chughtai K, Heeren RMA. Mass Spectrometric Imaging for Biomedical Tissue Analysis. Chem Rev. 2010;110:3237-77.

[2] McDonnell LA, Heeren RMA. Imaging mass spectrometry. Mass Spectrometry Reviews. 2007;26:606-43.

[3] DICOM. Digital imaging and communications in medicine - DICOM.http://medical.nema.org.

[4] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz P-A, Deutsch EW. mzML - a Community Standard for Mass Spectrometry Data. Molecular & Cellular Proteomics. 2011;10.

[5] Hermjakob H. The HUPO proteomics standards initiative - Overcoming the fragmentation of proteomics data. Proteomics. 2006:34-8.

[6] Römpp A, Schramm T, Hester A, Klinkert I, Heeren R, Stöckli M, Both J-P, Brunelle A, Spengler B. Imaging mzML (imzML) – a common data format for imaging mass spectrometry. In: Hamacher M, Stephan C, Eisenacher M, editors. Data Mining in Proteomics. New York: Humana Press; 2010. p. 205-24.

[7] Biomap. http://maldi-msi.org/index.php?option=com_content&task=view&id=14&Itemid=39.

[8] Network_Working_Group. RFC 4122 - A Universally Unique IDentifier (UUID) URN Namespace.http://tools.ietf.org/html/rfc4122.

[9] HUPO-PSI. The Human Proteome Organization: Proteomics Standards Initiative.http://psidev.info.

[10] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S, Consortium OBI. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25:1251-5.

[11] HUPO-PSI. Mass Spectrometry Standards Working Group - controlled vocabulary.http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo.

[12] MS Imaging controlled vocabulary.http://www.maldi-msi.org/index.php?option=com_content&view=article&id=193&Itemid=66.

[13] Li B, Bjarnholt N, Hansen SH, Janfelt C. Characterization of barley leaf tissue using direct and indirect desorption electrospray ionization imaging mass spectrometry. J Mass Spectrom. 2011;46:1241-6.

[14] Thunig J, Hansen SH, Janfelt C. Analysis of Secondary Plant Metabolites by Indirect Desorption Electrospray Ionization Imaging Mass Spectrometry. Anal Chem. 2011;83:3256-9.

[15] PREMIERBiosoft. MALDIVision. http://www.premierbiosoft.com/maldi-tissue-imaging/index.html.

[16] Imabiotech. Quantinetix MALDI Imaging Software. http://www.imabiotech.com/Quantinetix-TM-Maldi-Imaging.html?lang=en.

[17] McDonnell LA, Heeren RMA, Andrén PE, Stoeckli M, Corthals GL. Going forward: Increasing the accessibility of imaging mass spectrometry. Journal of Proteomics.

Figure 1: Scheme of imzML structure (adapted from www.imzml.org).

Figure 2: Example of sequential processing of an MS imaging data set based on imzML. MS imaging data acquired on an AB Sciex QStar instrument was converted to imzML. The imzML file was opened and processed in different software packages for analysis of MS imaging data. A) BioMap: Selected ion image of m/z = 772 and averaged mass spectrum, B) SpectViewer: averaged mass spectrum and selected ion image of m/z = 772 for original data (left) and after application of wavelet denoising algorithm (right), C) Datacube Explorer: mass spectrum of full image and two regions of interest (ROI) which are indicated in selected ion image of m/z 772, D) Mirion: selected ion image of m/z = 772 from original (top) and denoised (bottom) data set.

Figure 3: Examples of data sets from different instrument platforms that were converted to imzML. All images are displayed in the same software tool (Mirion) with identical settings (showing an overlay of two masses in red and green, no interpolation). Ionization type and mass spectrometer type are indicated for each measurement. A) mouse kidney, MALDI, LTQ Orbitrap (Thermo), data provided by Sabine Guenther (Justus Liebig University, Giessen, Germany), B) mouse brain (left half), MALDI, ultrafleXtreme (Bruker), data provided by Corinna Henkel (Bochum University, Germany), C) whole body rat section (head), MALDI, QStar (ABSciex), same data as shown in Figure 2, D) whole bod yrat section, MALDI, Synapt G2 (Waters), data provided by Emmanuelle Claude (Waters, Manchester, UK), E) rabbit eye, SIMS, TOF-SIMS IV (IonTOF), data provided by Alain Brunelle (CNRS, Gif-sur-Yvette, France), F) tumor tissue, DESI, Exactive (Thermo), data provided by Zoltan Takats (Imperial College, London, UK). More details on these data sets can be found in the 'imzML gallery' on the www.imzml.org.

Figure 1

Click here to download high resolution image

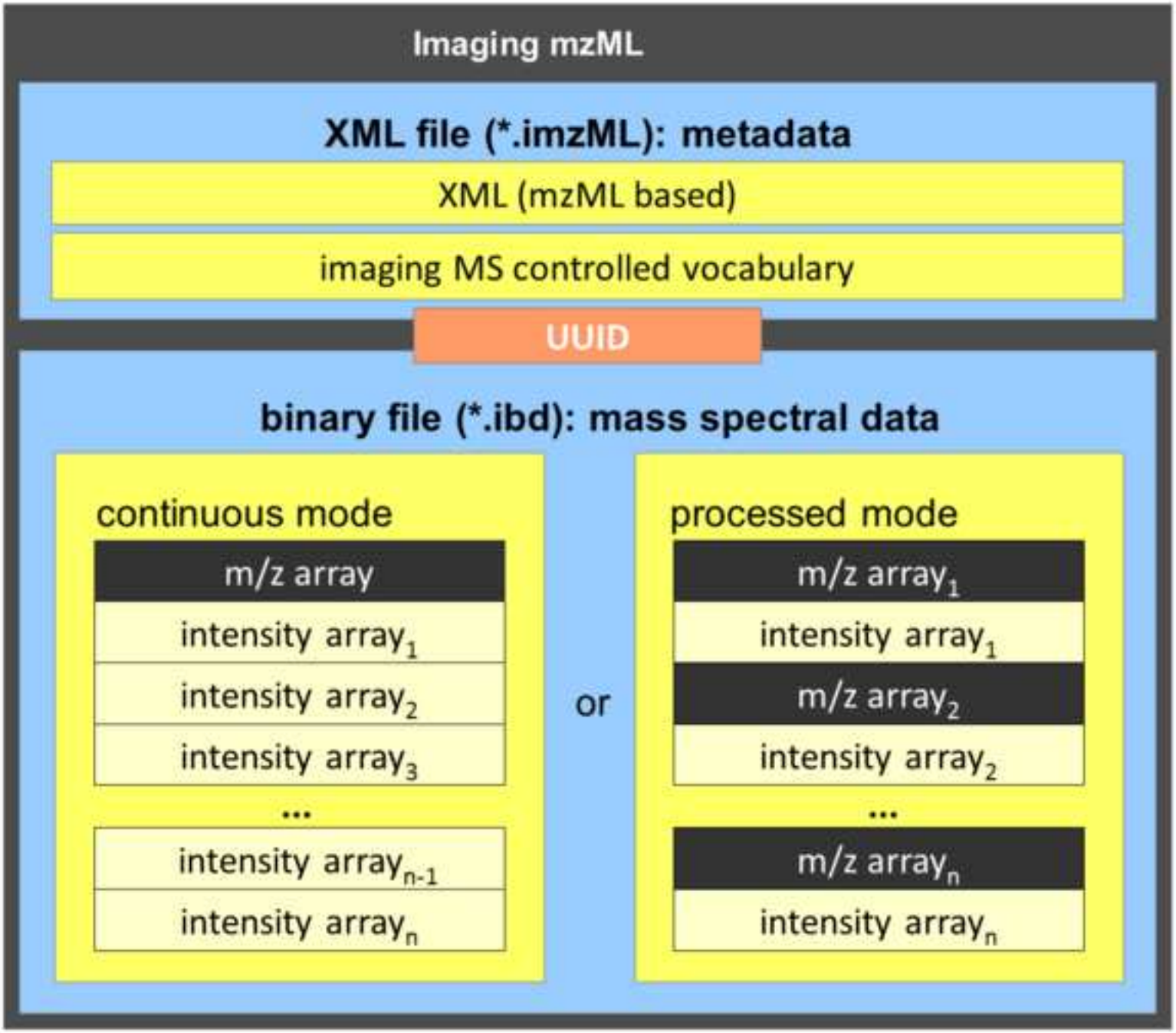


Figure 2

Click here to download high resolution image

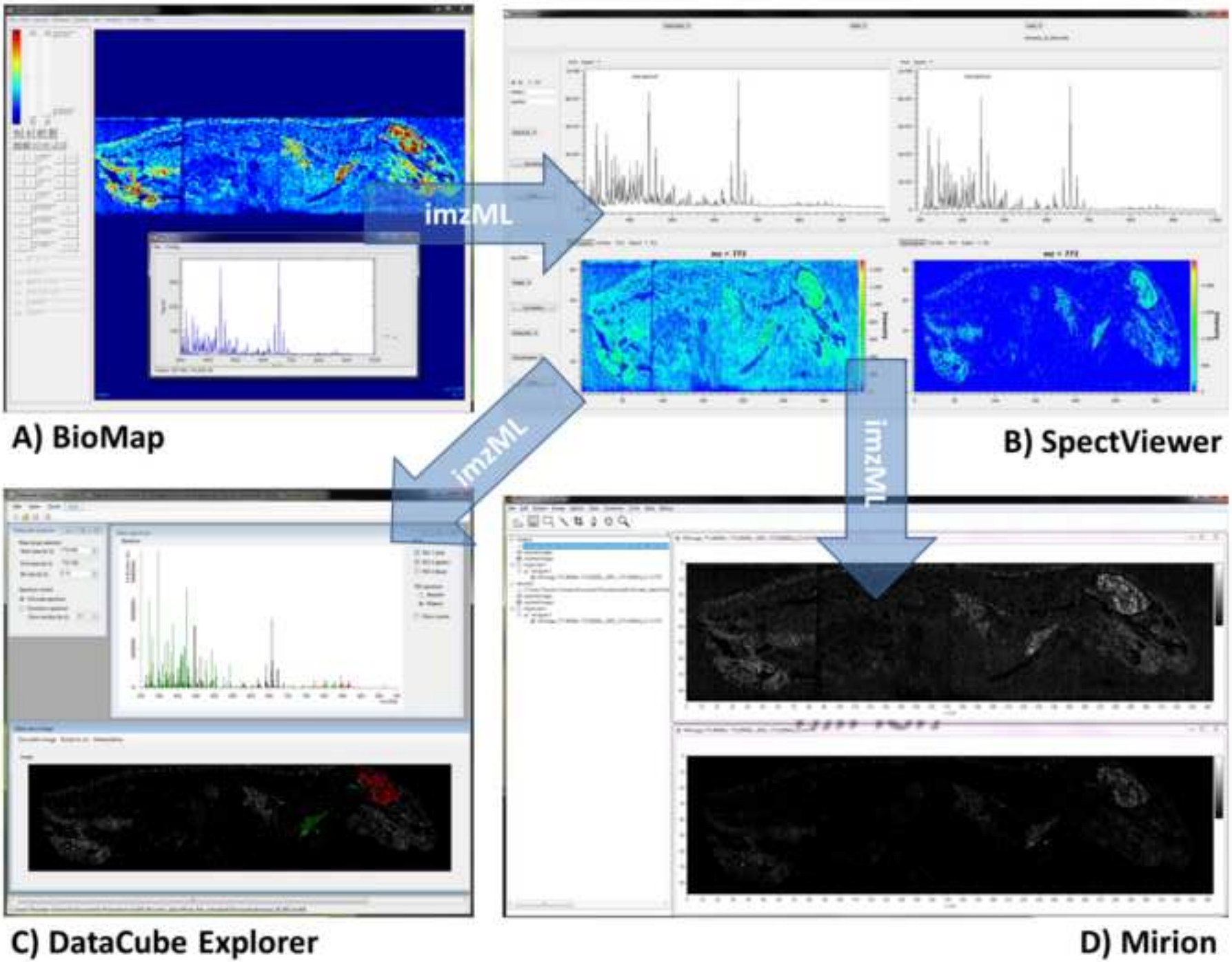


Figure 3

Click here to download high resolution image