

# Creation of an open standard file format for the representation of MS data

P. Pedrioli<sup>1</sup>, J. Eng<sup>1</sup>, R. Hubley<sup>1</sup>, B. Pratt<sup>2</sup>, E. Nilsson<sup>2</sup>, A. Taylor<sup>3</sup>, R. Aebersold<sup>1</sup>  
<sup>1</sup>Institute for Systems Biology, Seattle, WA; <sup>2</sup>Insilicos, Seattle, WA; <sup>3</sup>Amgen, Seattle, WA

## Overview

To address the difficulties presented by the introduction of a new mass spectrometer into a pre-existing data analysis framework, we developed an XML based common file format for MS data. The adoption of an open standard will provide programmers with an easy way to access this kind of information, thus facilitating development and distribution of software in this field. Additionally, the use of an architecture and operating system independent representation will ease the exchange of datasets between collaborators and ultimately allow for the creation of public data repositories. Although this format was created to address our specific needs in quantitative proteomics experiments, due to the expandable nature of XML, we believe that it will be easily amenable to other types of analysis.

### A diverse set of Mass Spectrometers allows for more flexibility! However ...

- Different MS manufacturers store **data in different formats**. These formats are incompatible with each other and often times proprietary.
- **This limits:**
- **Data analysis**
  - Software from the same company that sells the instrument needs to be used to analyse or even just to look at the data
  - The sources of these programs are not released and as a result one **cannot quickly modify** them to fit a particular experimental setup
  - **Comparing results** generated on different machines becomes **difficult**
- **Exchange** of RAW datasets and creation of public repositories
- **Software development** and distribution
- Ease of **integration** of a new instrument in a pre-existing data analysis pipeline

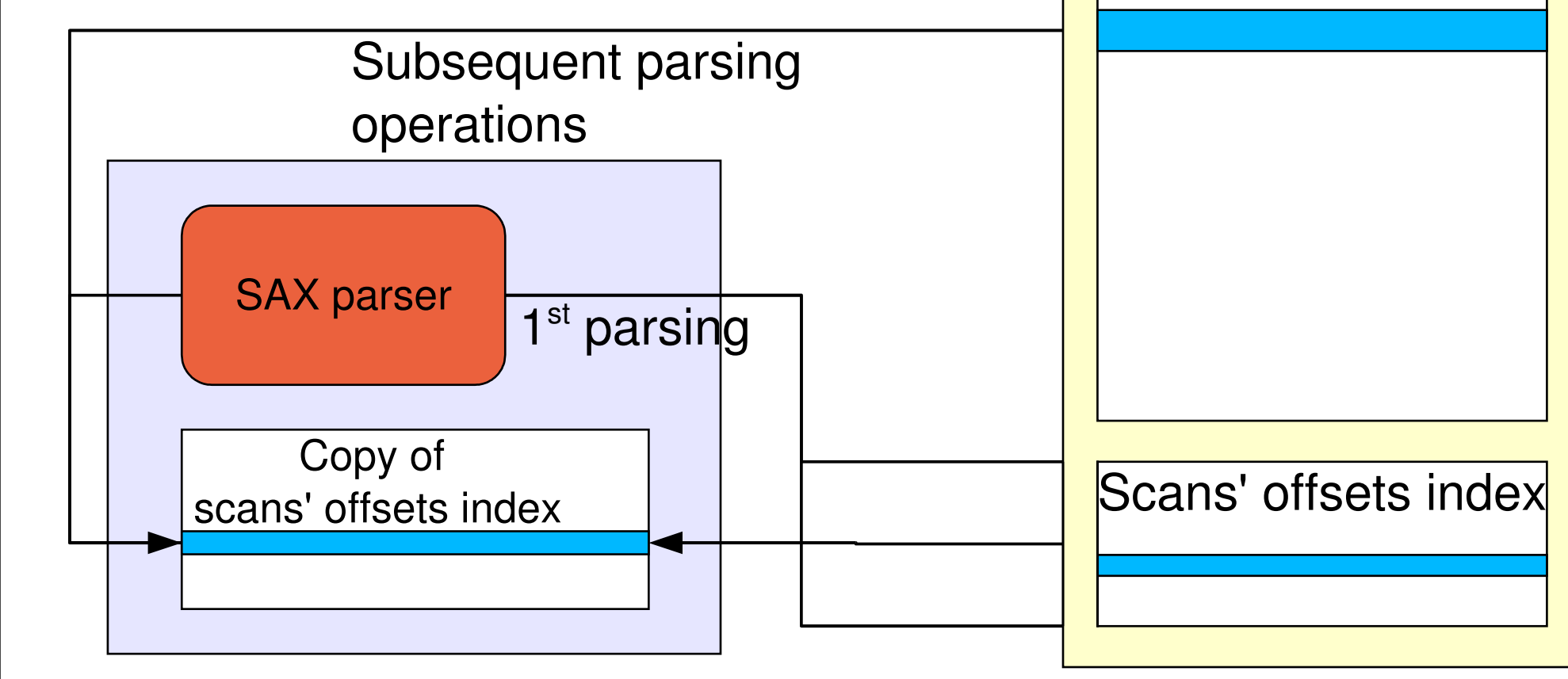
### In the ideal world ...

There would be a **common file format for all MS** and it would be:

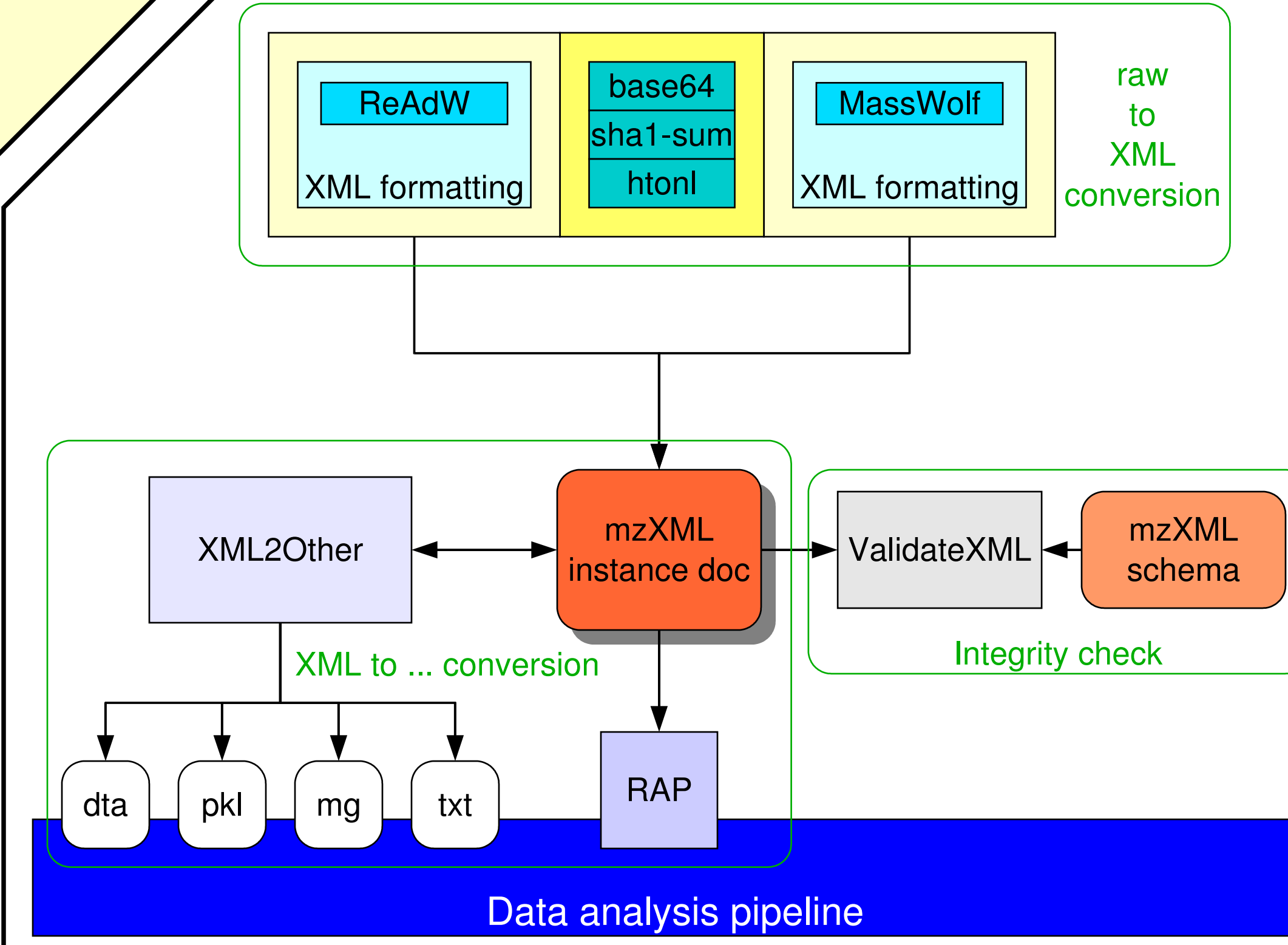
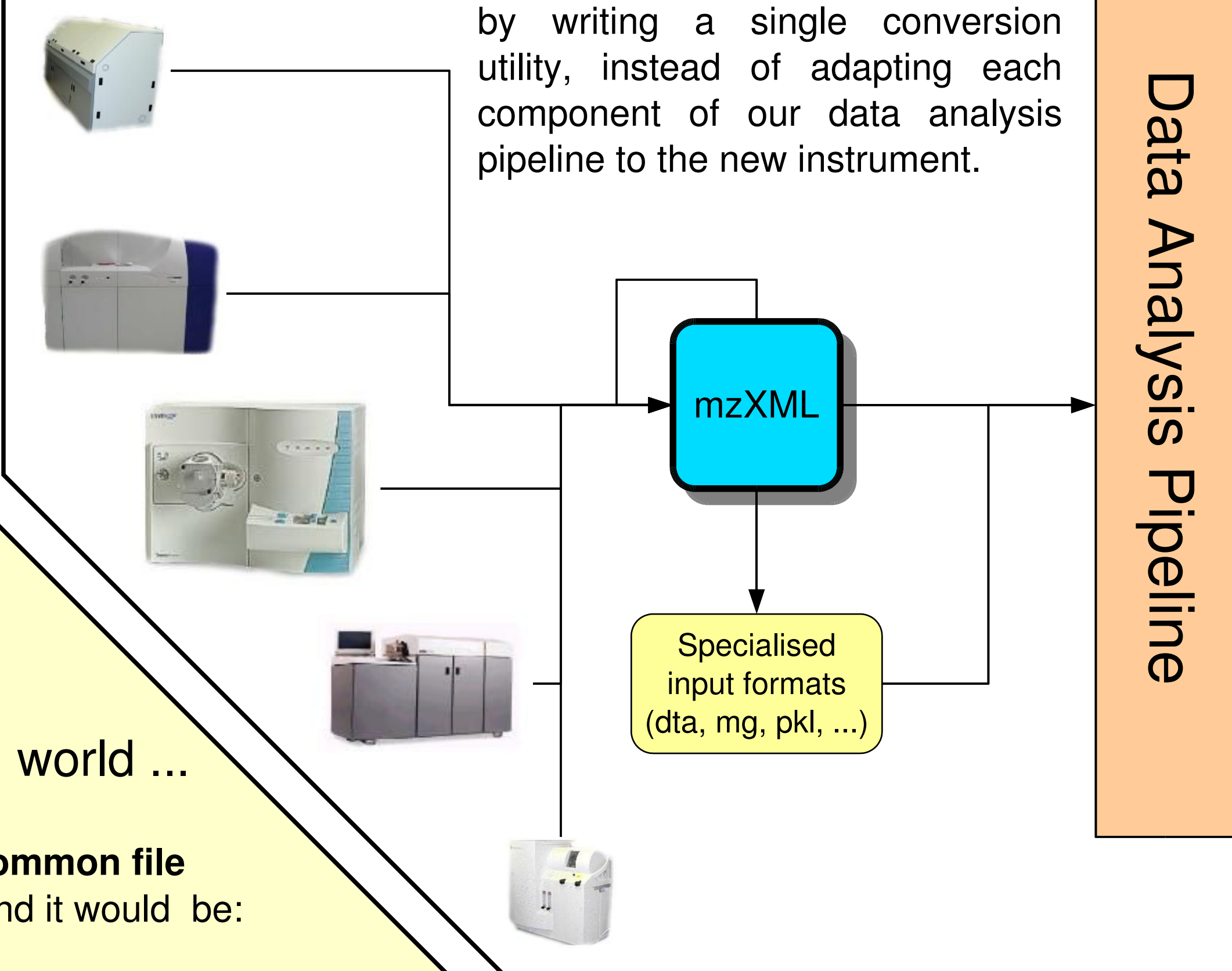
- **Compact**
  - **Precise**
  - **Fast**
  - **Machine independent**
  - **Easy to expand**
  - **Public domain**
- Binary  
XML

Some of our MS generate ~1GB of compressed binary data per hour. It is extremely inconvenient to represent that much information in a clear text format while retaining the necessary precision. The use of **Base64 encoding** (the same encoding used to attach binary files to emails and used in GAML by Thermo LabSystems) can **help minimize this size increase**.

The top-to-bottom reading nature of the SAX parser is not compatible with applications (such as our quantitation software) requiring non-sequential access to the scans. We obviated this problem by indexing the positions of each scan in a given XML file. At parsing time the **index** can be used to adjust the input stream to a scan-specific offset, thereby obtaining **random access to the data**.

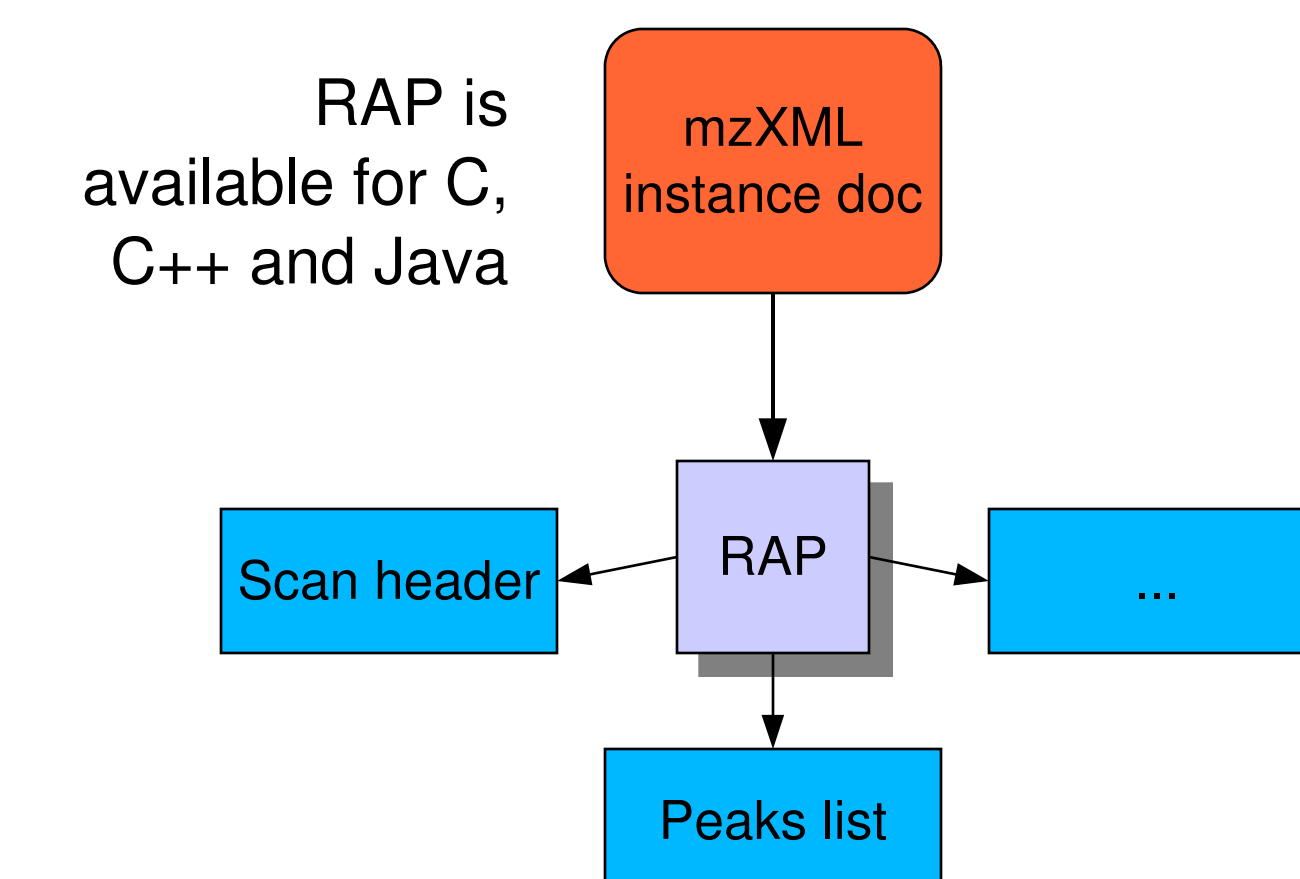


We developed an intermediary file format based on the eXtensible Markup Language (XML). XML has many desirable properties: it is based on a public domain standard maintained by a neutral organization (the World Wide Web Consortium) and it is supported on a very broad range of machines. Having a common representation allows us to easily integrate a new instrument in the laboratory. This is achieved by writing a single conversion utility, instead of adapting each component of our data analysis pipeline to the new instrument.



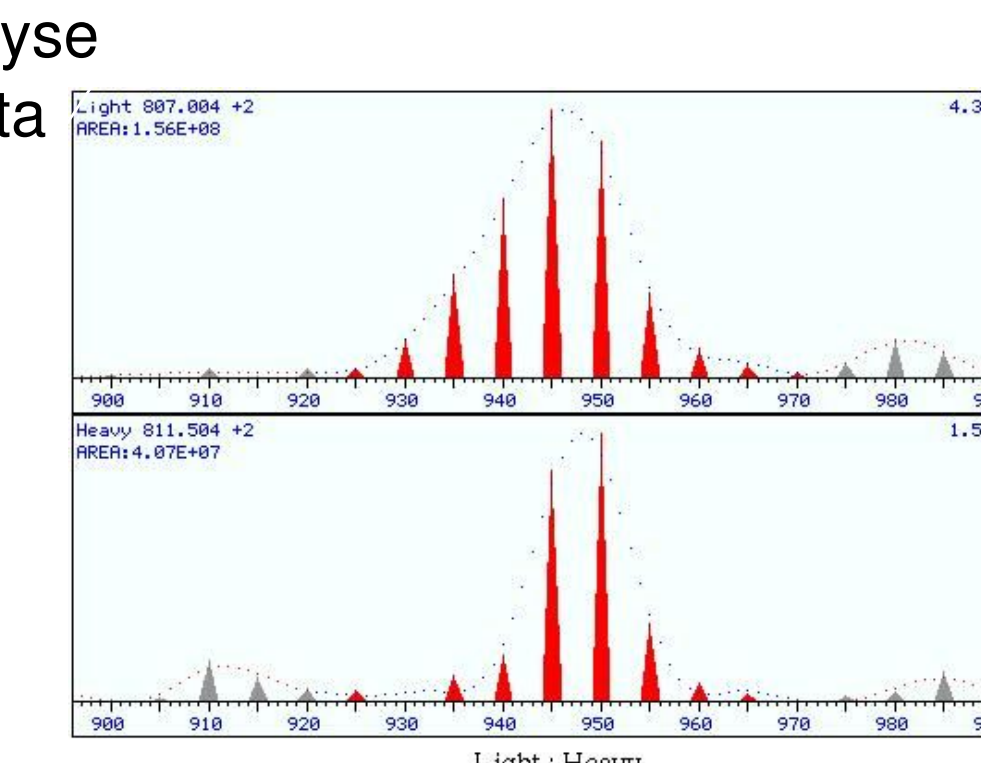
### RandomAccessParser

Provide an easy way to **retrieve information from an mzXML instance document**. Make use of the index and offers multiple functions for specific and fast data access.



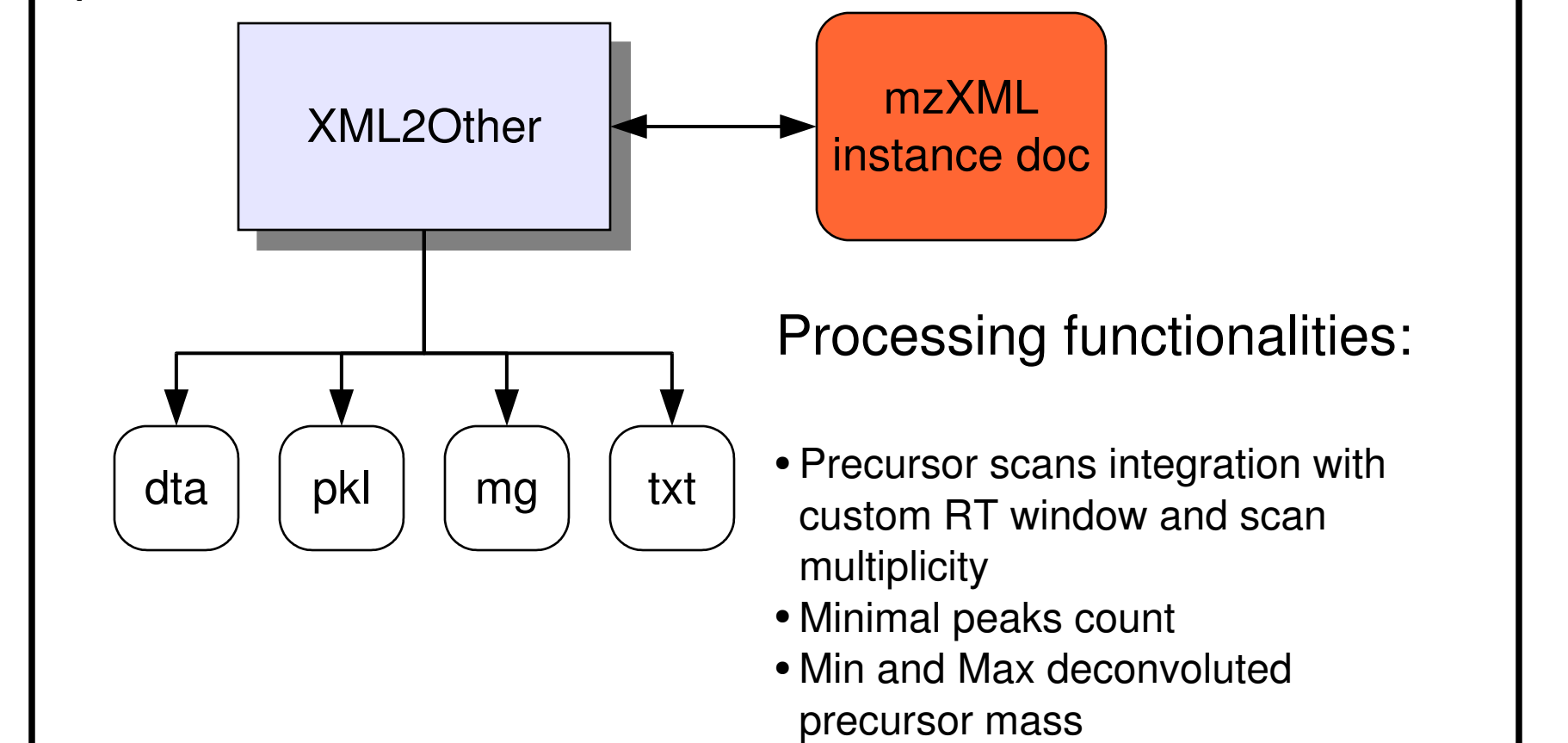
### XPRESS and RAP

XPRESS is an open source tool developed by J. Eng for relative **quantitation of isotopically labeled peptides**. With the use of RAP it has been successfully ported to analyse quantitative data from ICAT experiments represented in mzXML format.

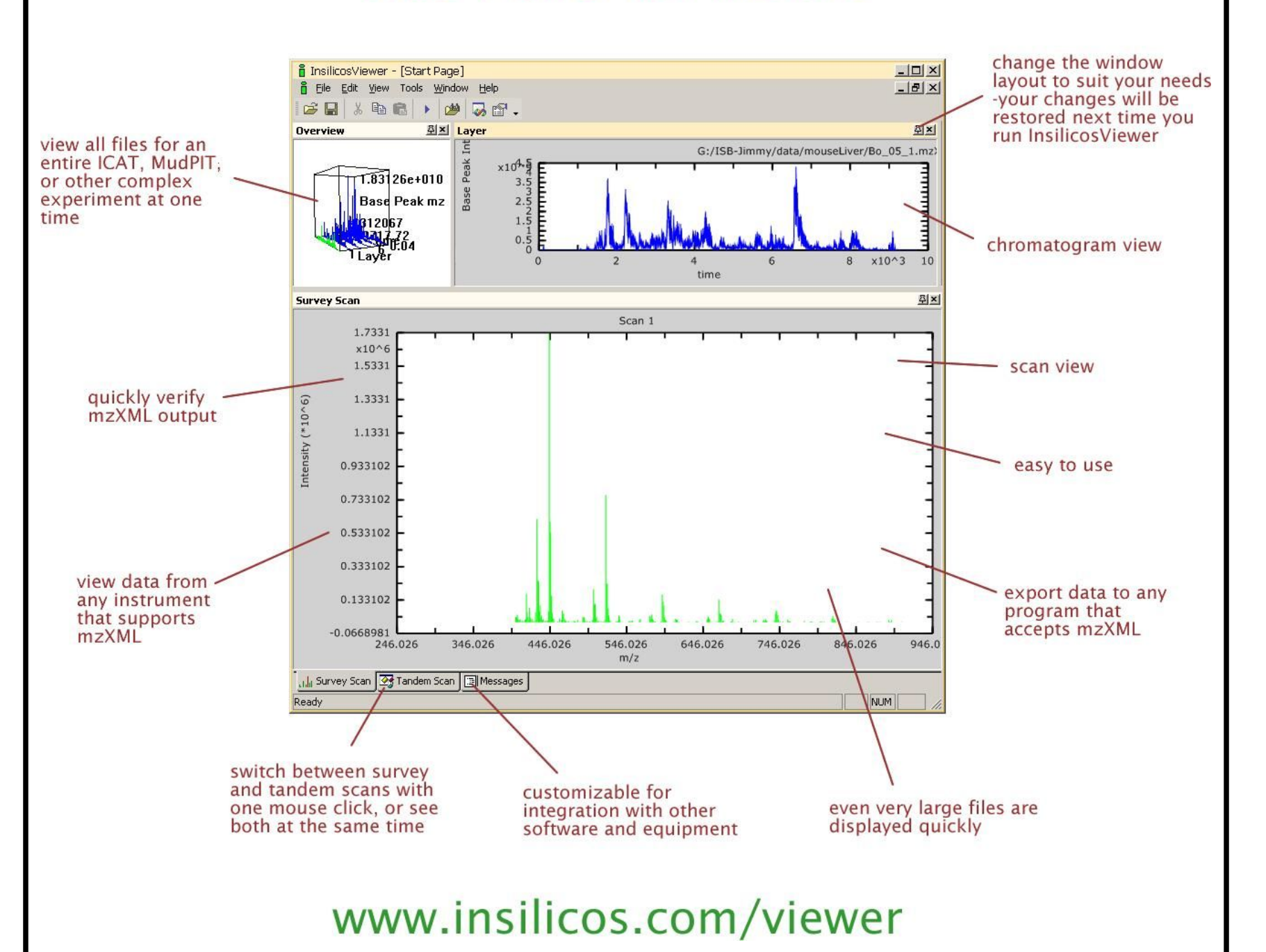


### MzXML2Other

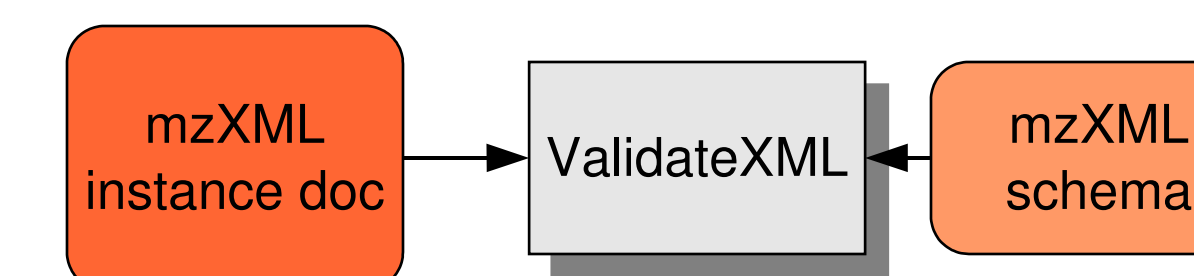
Convert mzXML into the formats supported by the more commonly used database search engines including SEQUEST [9] dta, Mascot generic, Micromass pkl and pure text formats.



### Free Viewer for mzXML



### ValidateXML



**Validates** a given instance document **against** the publicly available **schema** and checks the authenticity of its **checksum** in one single step. This is particularly useful to confirm that the data have not been corrupted during transfer from one machine to another.

### Conclusion & future directions

An XML based common file format for MS data was developed. We offer free open source utilities providing basic functionalities to work on datasets stored in this format, as well as software performing higher level manipulations such as relative quantitation of isotopically labeled peptides and data visualization. A collection of representative datasets from different instruments in raw and mzXML format can be accessed via our homepage. This collection will be updated in the future as RAW converters for more instruments will be developed. A new version of the Schema with support for MALDI instruments (the current one only supports ESI) is also being studied.

The XML schema for mzXML instance documents, as well as the sources for most of the programs presented on this poster can be obtained at:  
<http://sashimi.sourceforge.net/>