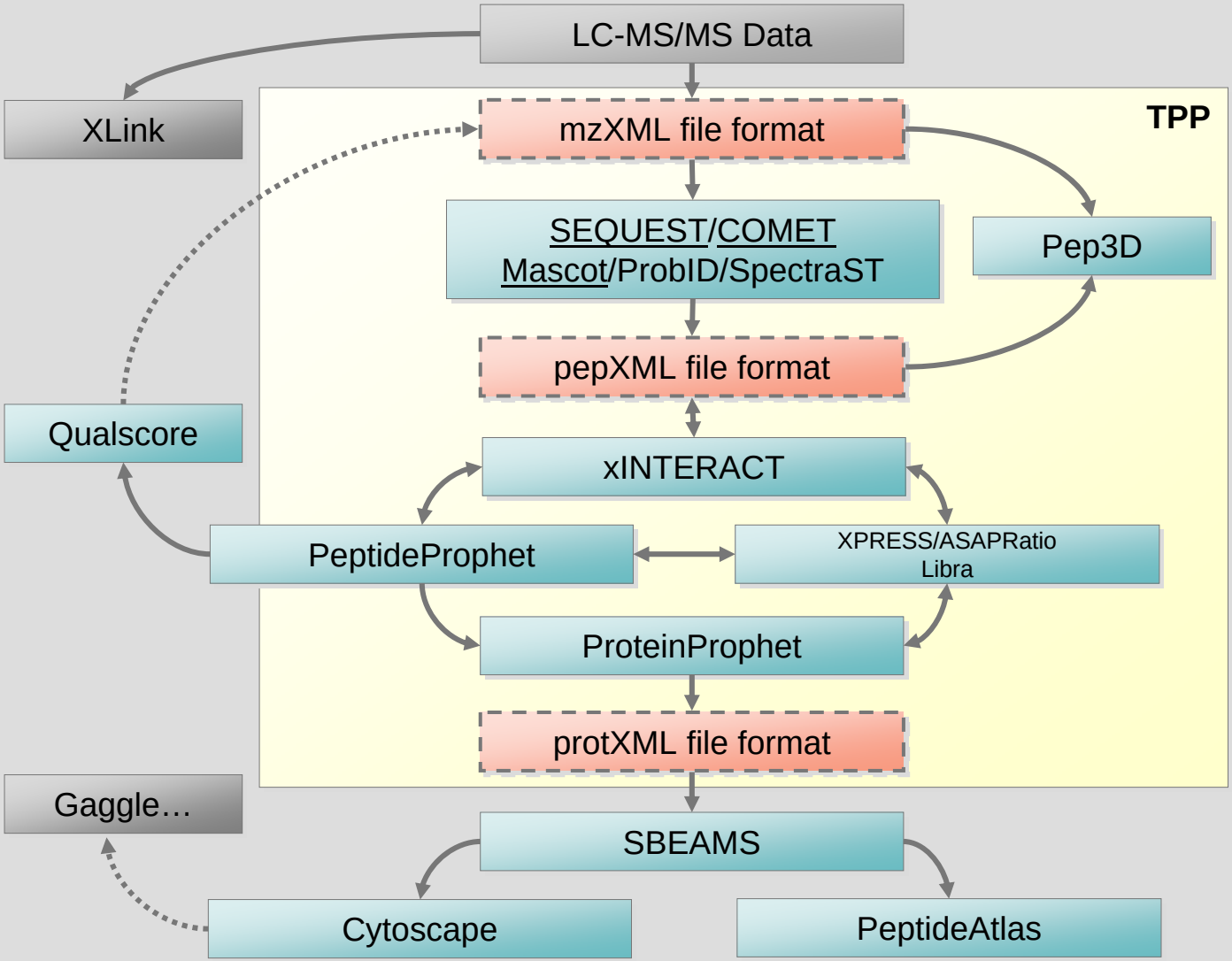
Formats: mzXML, PepXML, ProtXML

David Shteynberg

NHLBI PROTEOMICS

INSTITUTE FOR Systems Biology

Formats

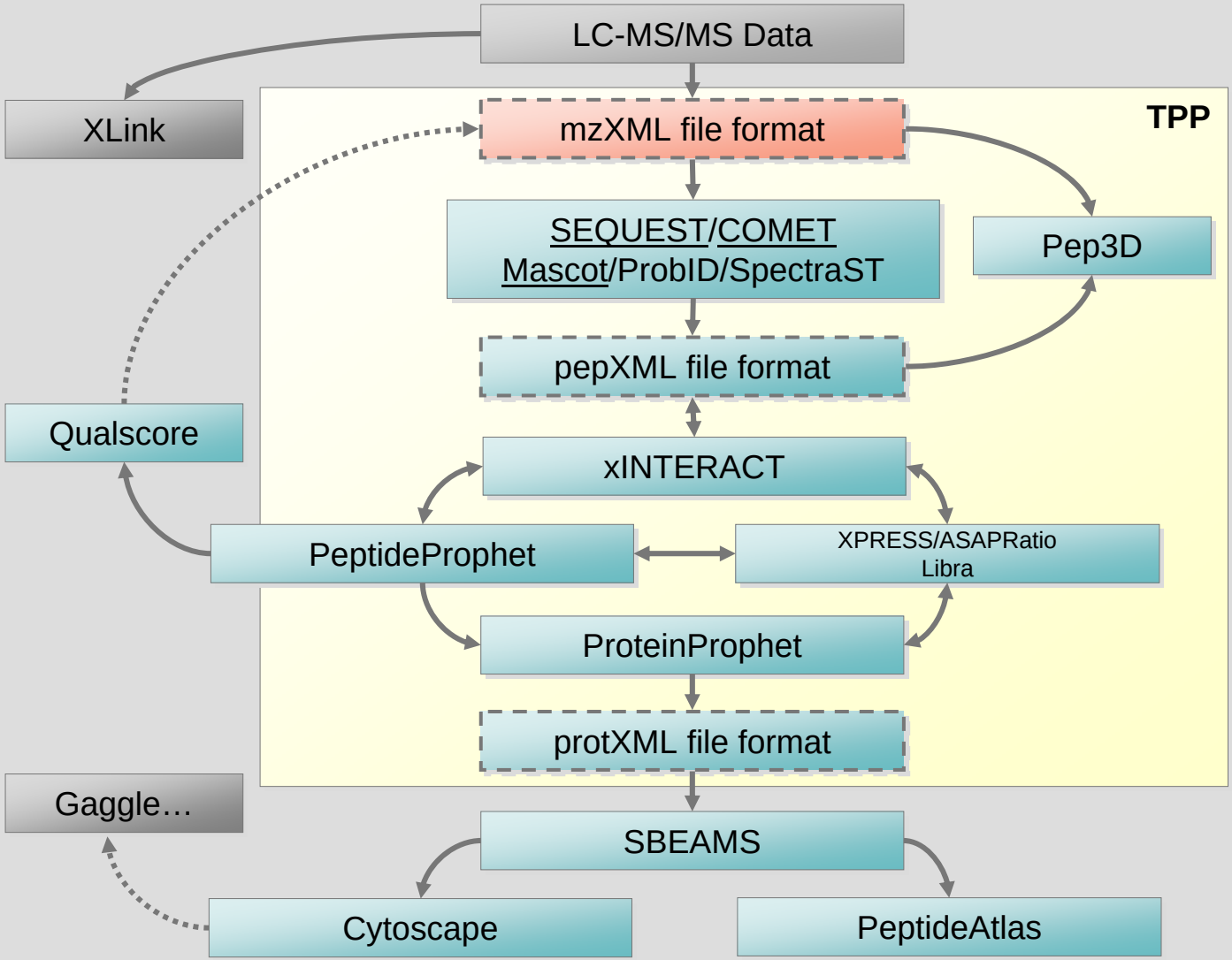

LC-MS/MS Data
XLink
TPP
mzXML file format
SEQUEST/COMET
Mascot/ProbID/SpectraST
Pep3D
pepXML file format
Qualscore
xINTERACT
PeptideProphet
XPRESS/ASAPRatio
Libra
ProteinProphet
protXML file format
Gaggle…
SBEAMS
Cytoscape
PeptideAtlas

Background on XML

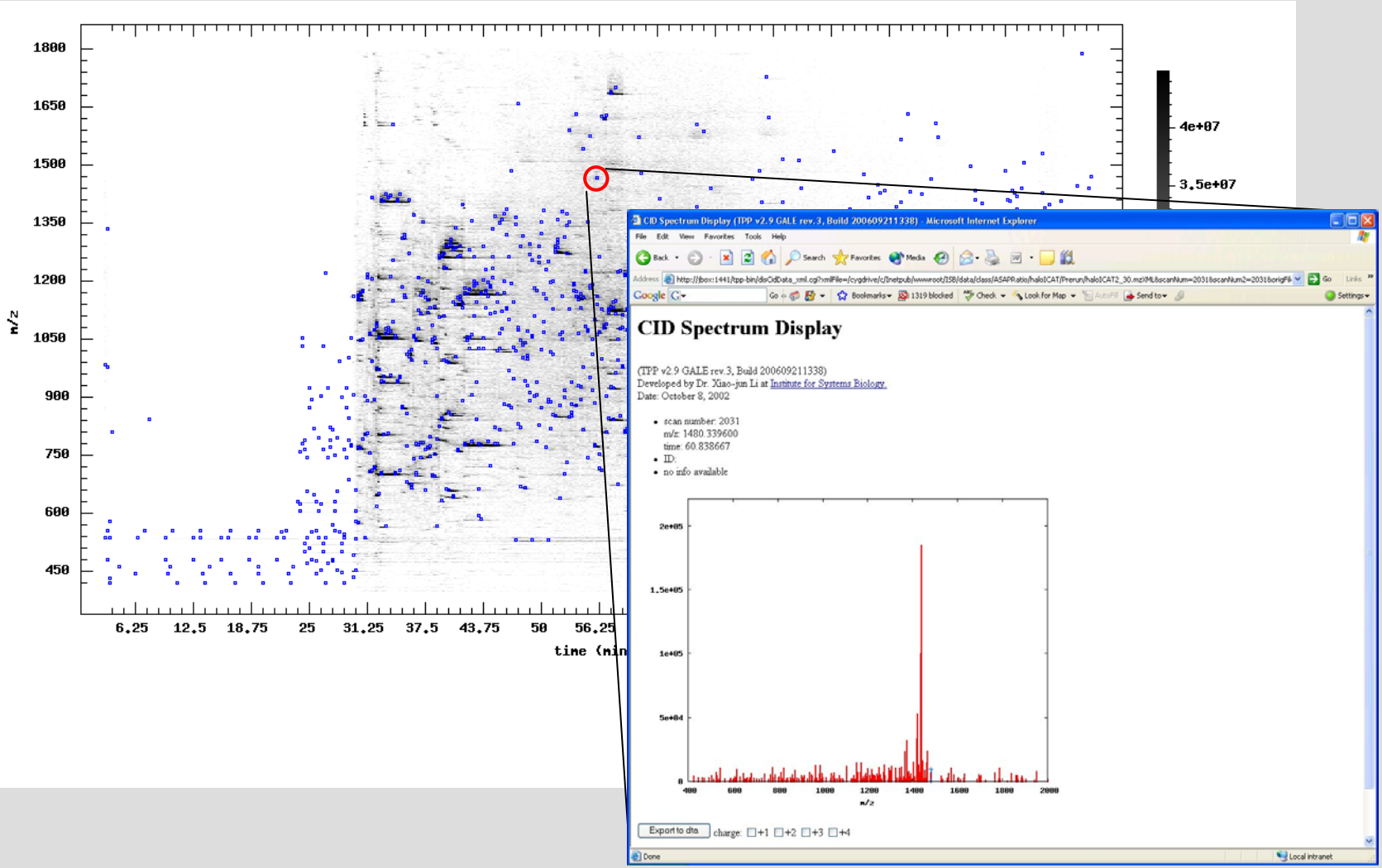
- XML (**eXtensible** Markup Language) allows the creation of **self-describing** formats.
- XML consists of **text** organized **within tags**.
  - Human readable
- The actual data is represented in **XML instance documents**.
- The structure of an instance document is defined by the **XML Schema**.

```
<book  type="SciFi"
       location="A-13">
       <author>Douglas Adams</author>
       <title>The Hitchhiker's Guide to the Galaxy</title>
</book>
```

mzXML


LC-MS/MS Data
XLink
TPP
mzXML file format
SEQUEST/COMET
Mascot/ProbID/SpectraST
Pep3D
pepXML file format
Qualscore
xINTERACT
PeptideProphet
XPRESS/ASAPRatio
Libra
ProteinProphet
protXML file format
Gaggle…
SBEAMS
Cytoscape
PeptideAtlas

LC-MS/MS Raw Data



Pedrioli *et al*., Nature Biotechnology, 22, 1459-66, 2004

PERSPECTIVE

nature biotechnology

A common open representation of mass spectrometry data and its application to proteomics research

Patrick G A Pedrioli[1], Jimmy K Eng[1], Robert Hubley[1], Mathijs Vogelzang[1], Eric W Deutsch[1], Brian Raught[1], Brian Pratt[2], Erik Nilsson[2], Ruth H Angeletti[3], Rolf Apweiler[4], Kei Cheung[5], Catherine E Costello[6], Henning Hermjakob[4], Sequin Huang[6], Randall K Julian, Jr[7], Eugene Kapp[8], Mark E McComb[6], Stephen G Oliver[9], Gilbert Omenn[10], Norman W Paton[11], Richard Simpson[8], Richard Smith[12], Chris F Taylor[4], Weimin Zhu[4] & Ruedi Aebersold[1]

A broad range of mass spectrometers are used in mass spectrometry (MS)-based proteomics research. Each type of instrument possesses a unique design, data system and performance specifications, resulting in strengths and weaknesses for different types of experiments. Unfortunately, the native binary data formats produced by each type of mass spectrometer also differ and are usually proprietary. The diverse, nontransparent nature of the data structure complicates the integration of new instruments into preexisting infrastructure, impedes the analysis, exchange, comparison and publication

components creates a wide range of MS instruments, each one with particular strengths and weaknesses for certain types of proteomics experiments. It is thus common to find more than one type of mass spectrometer in a typical proteomics laboratory. After acquisition, an MS instrument's output (from now on referred to as native data) is subjected to a number of analytical steps, commonly supported by proprietary programs supplied by the instrument manufacturer. Thus, in a laboratory environment that operates different types of mass spectrometers and multiple data analysis programs, it becomes extremely challenging to make meaningful comparisons of results

ublishing Group http://www.nature.com/naturebiotechnology

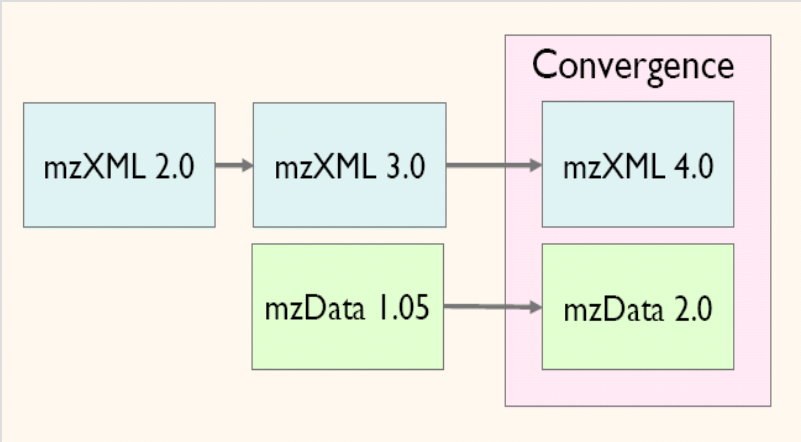
Motivation for mzXML

Data is in varied formats -- difficult to:

- Apply Common Data Analysis Pipeline
- Share data with others
- Share tools with others

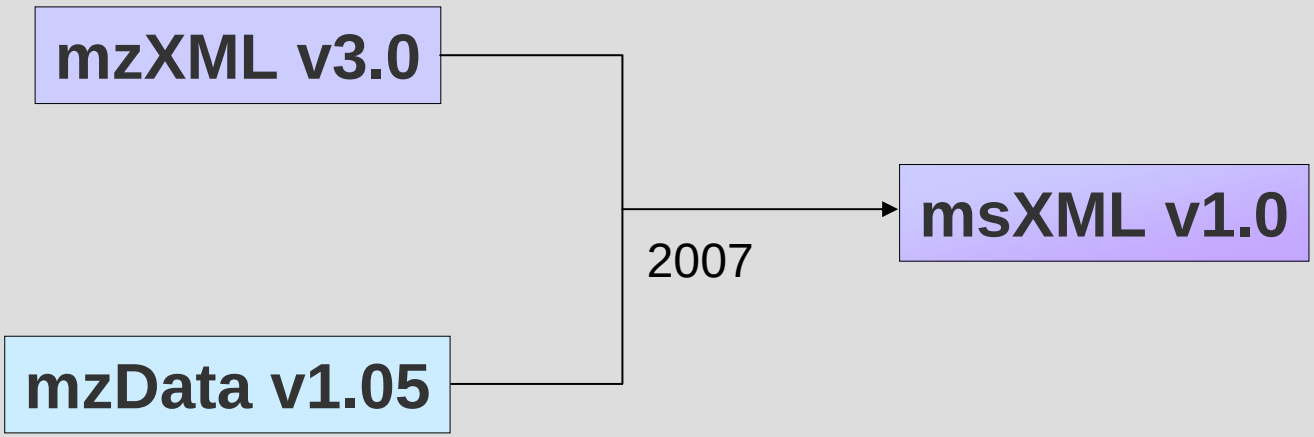
Need for Open Standards

- Different MS manufacturers store **data in different proprietary formats**
- **Open Standards enable exchange** of data and software tools
- Several existing open formats

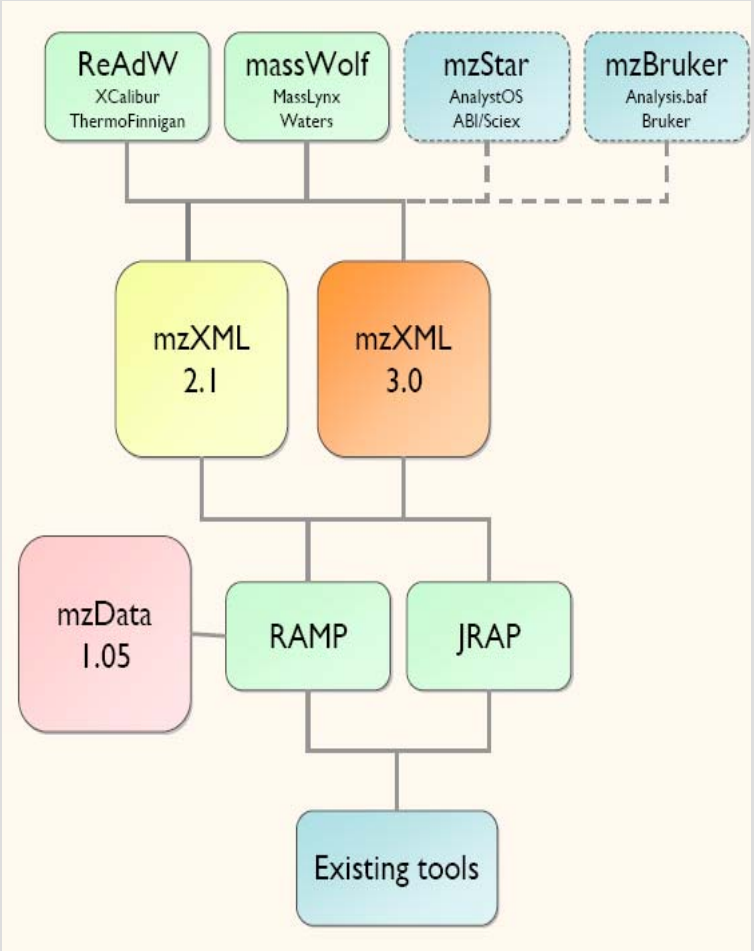


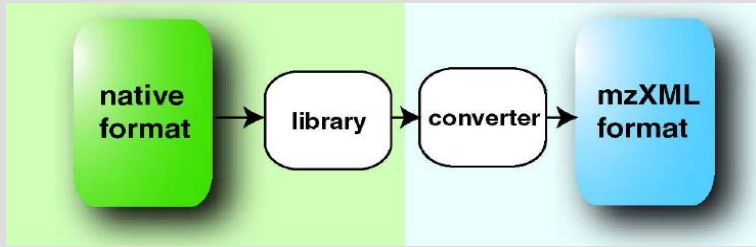
HUPO PSI – msXML v1.0

Convergence of open proteomics formats



Reading And Writing mzXML

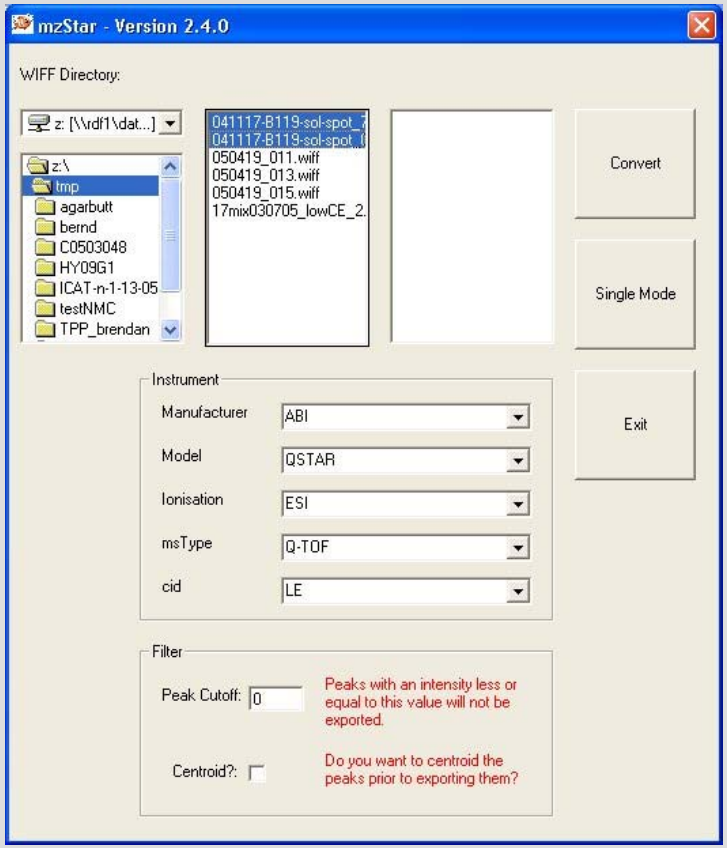




- **Converters**
  - **Input**: Instrument Format
  - **Output**: mzXML
- **Parsers**
  - Libraries for decoding mzXML/mzData encoded spectra

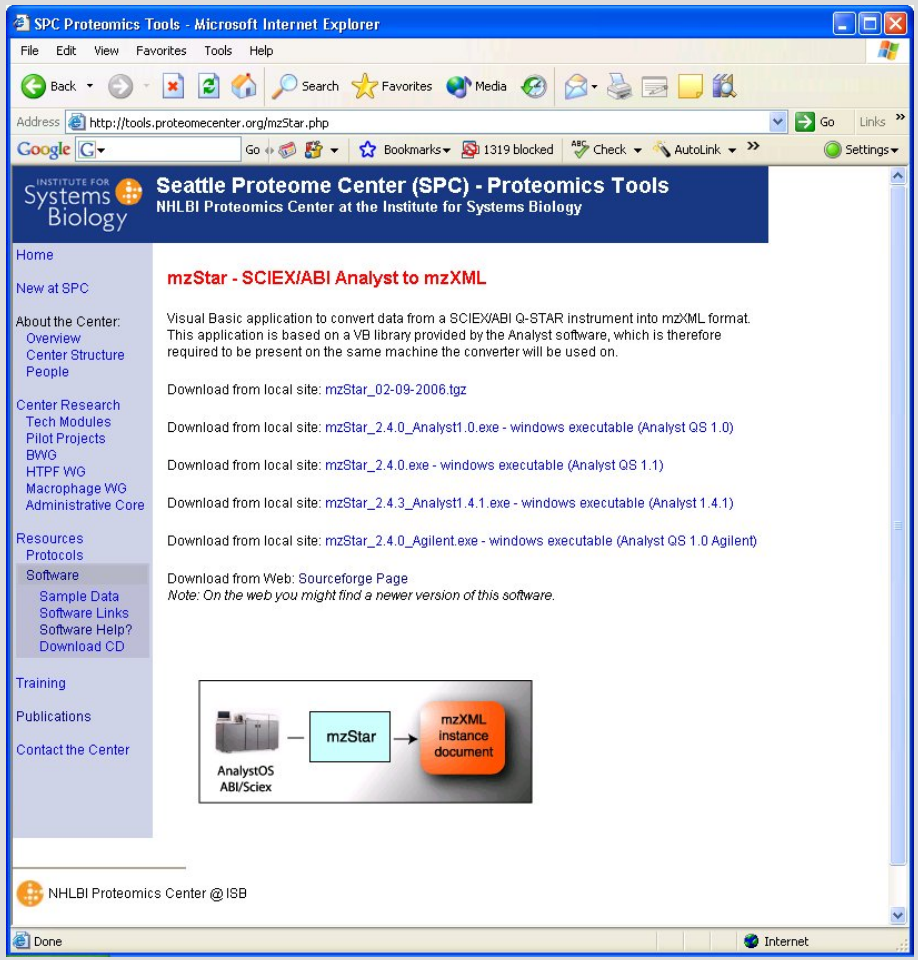
mzStar – ABI/Sciex

- Stand-alone Windows application for converting .wiff files
- Requires correct version of AnalystQS
- Features
  - Batch mode
  - Peak filtering
  - Centroiding



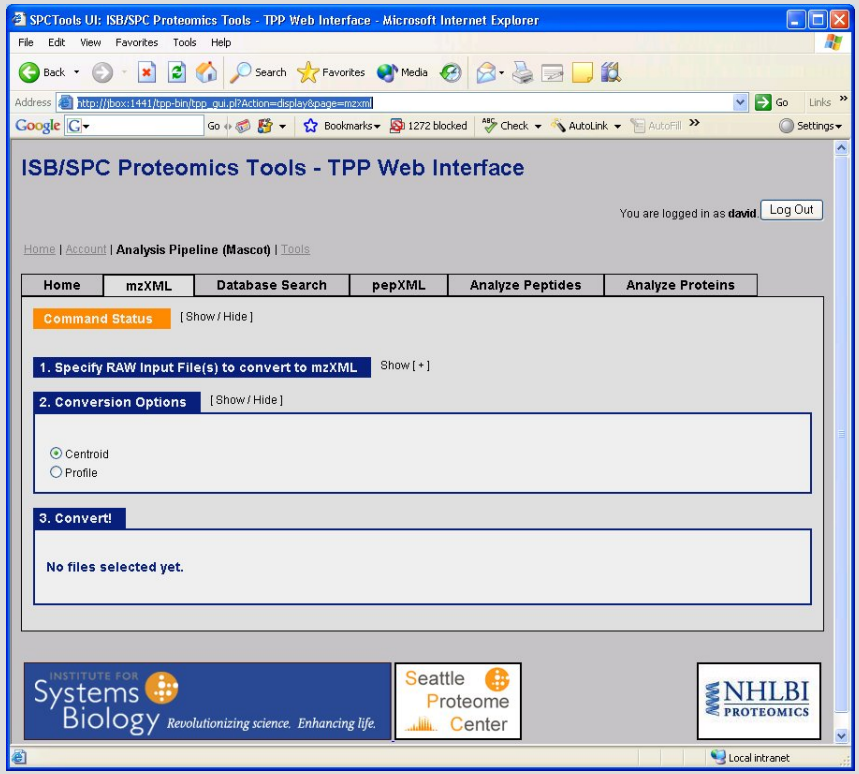
http://tools.proteomecenter.org/mzStar.php

http://tools.proteomecenter.org/mzStar.php



ReAdW – ThermoFinnigan

- Converts .RAW files generated by LTQ/LCQ
- Requires XCalibur
- Integrated with Petunia interface
  - Batch mode
- Centroiding option

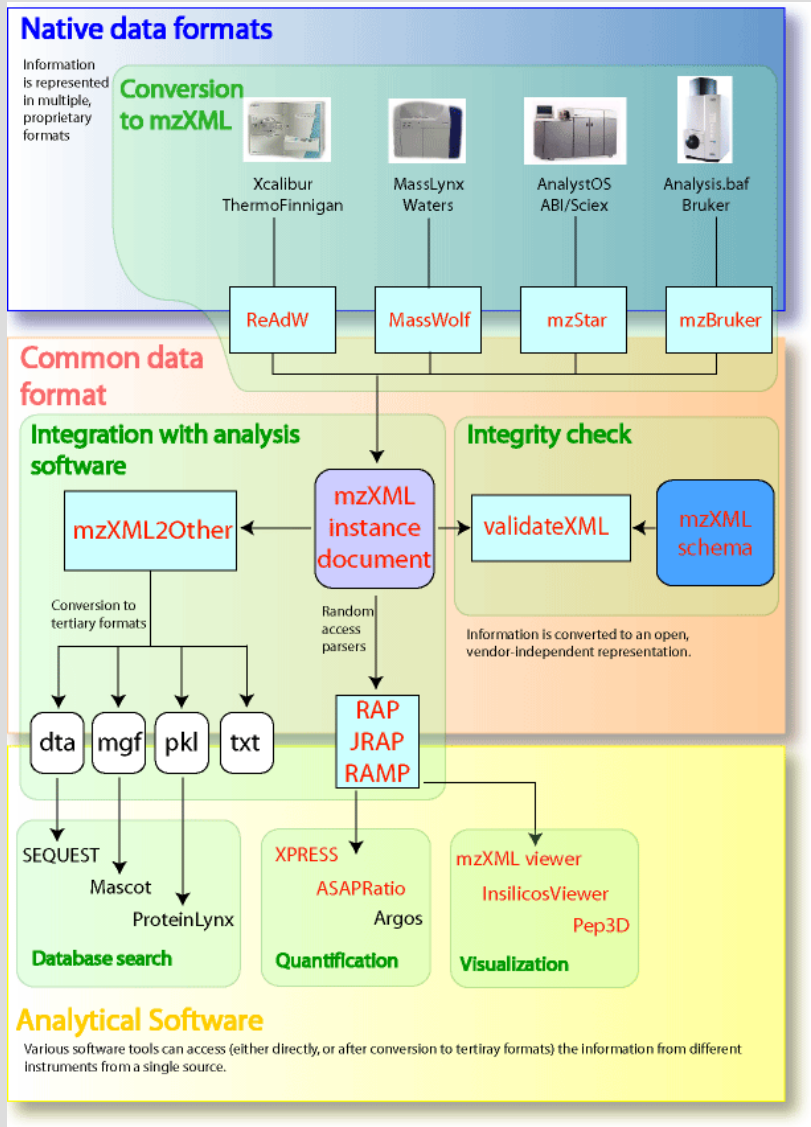


Other Converters

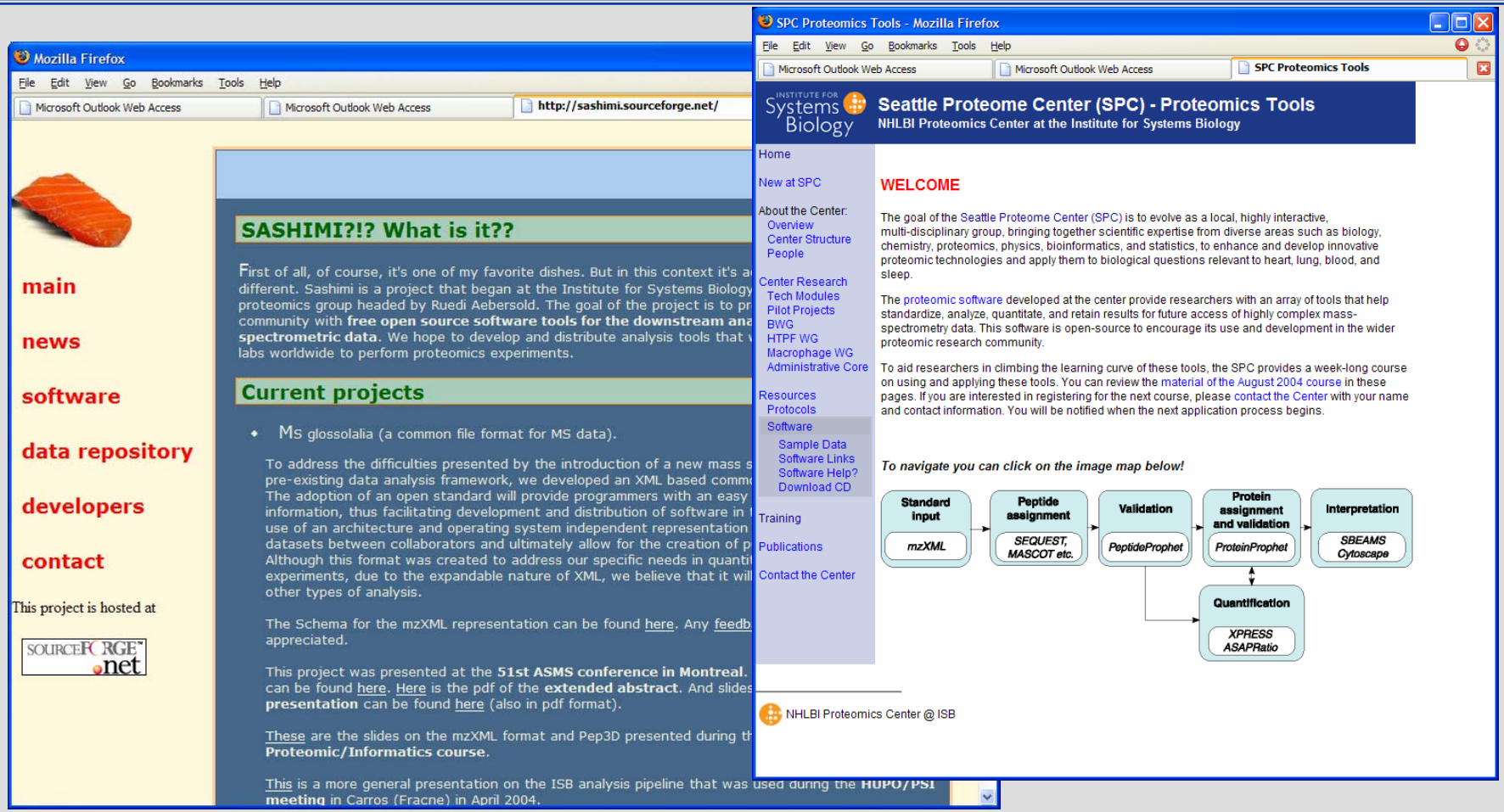
- http://tools.proteomecenter.org/MassWolf.php
  - Waters Micromass MassLynx
- http://tools.proteomecenter.org/mzBruker.php
  - Converts analysis.baf files
- Converter for Agilent .yep format also exists

mzXML Consumers

- Converters of mzXML into other formats
  - e.g DB Search Engine Inputs
  - mzXML2Other
    - Output: mgf, dta …
- Database Search Engines
  - e.g. X!Tandem, Phenyx
- MS Data Visualization Tools
  - e.g. Pep3D, InsilicosViewer



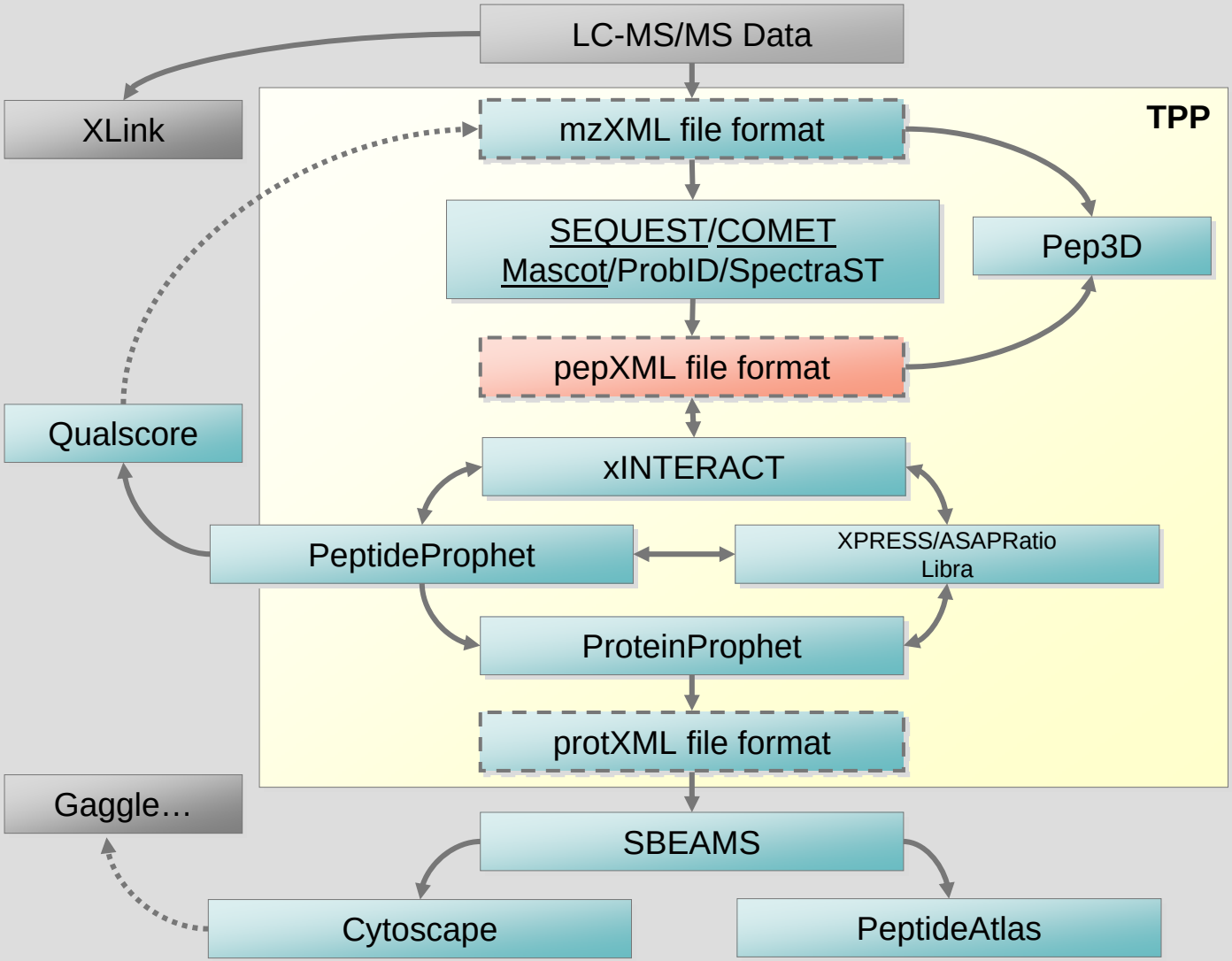
Project Homepage



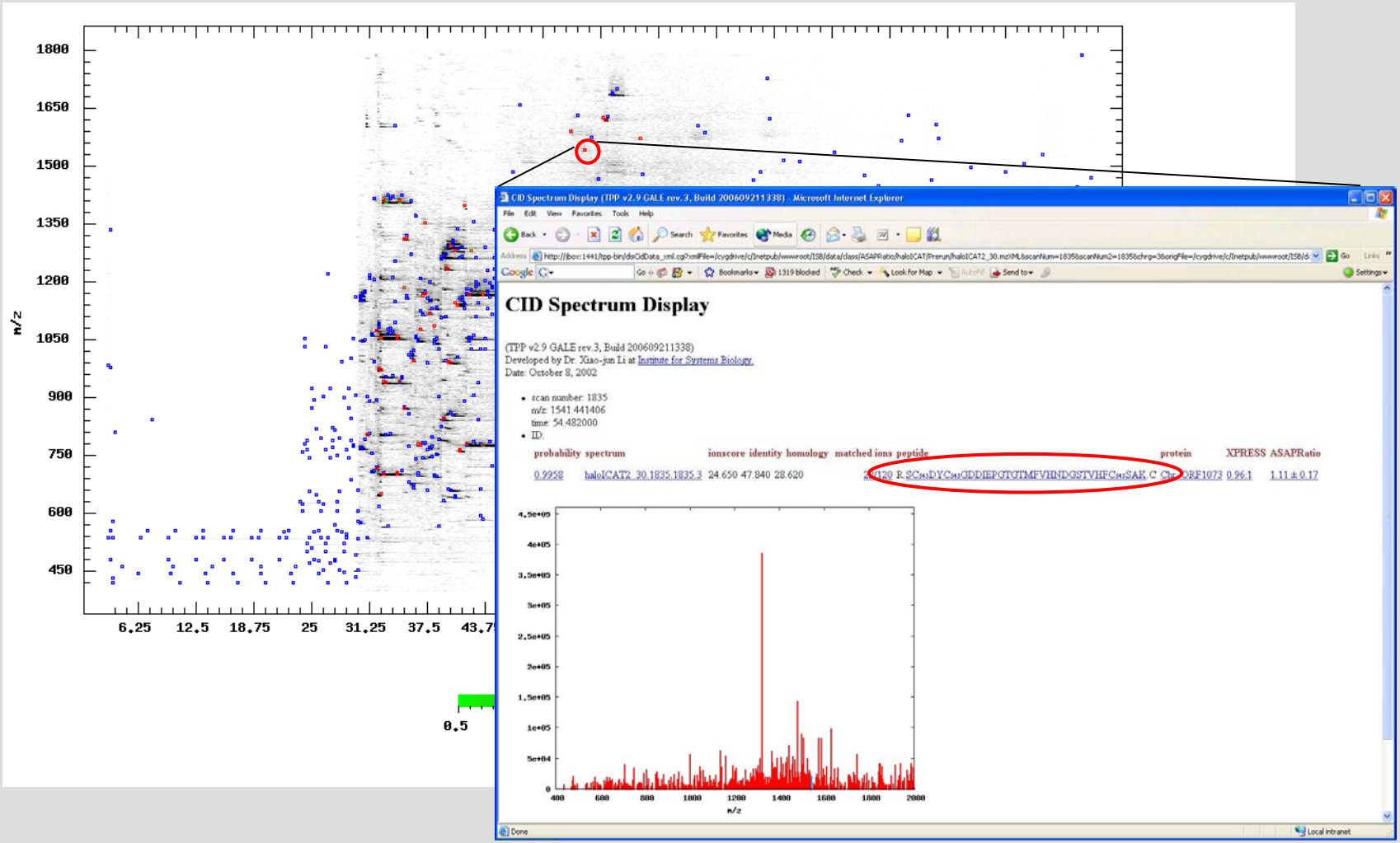
http://sashimi.sourceforge.net

http://tools.proteomecenter.org

pepXML


LC-MS/MS Data
XLink
TPP
mzXML file format
SEQUEST/COMET
Mascot/ProbID/SpectraST
Pep3D
pepXML file format
Qualscore
xINTERACT
PeptideProphet
XPRESS/ASAPRatio
Libra
ProteinProphet
protXML file format
Gaggle…
SBEAMS
Cytoscape
PeptideAtlas

LC-MS/MS Searched Data

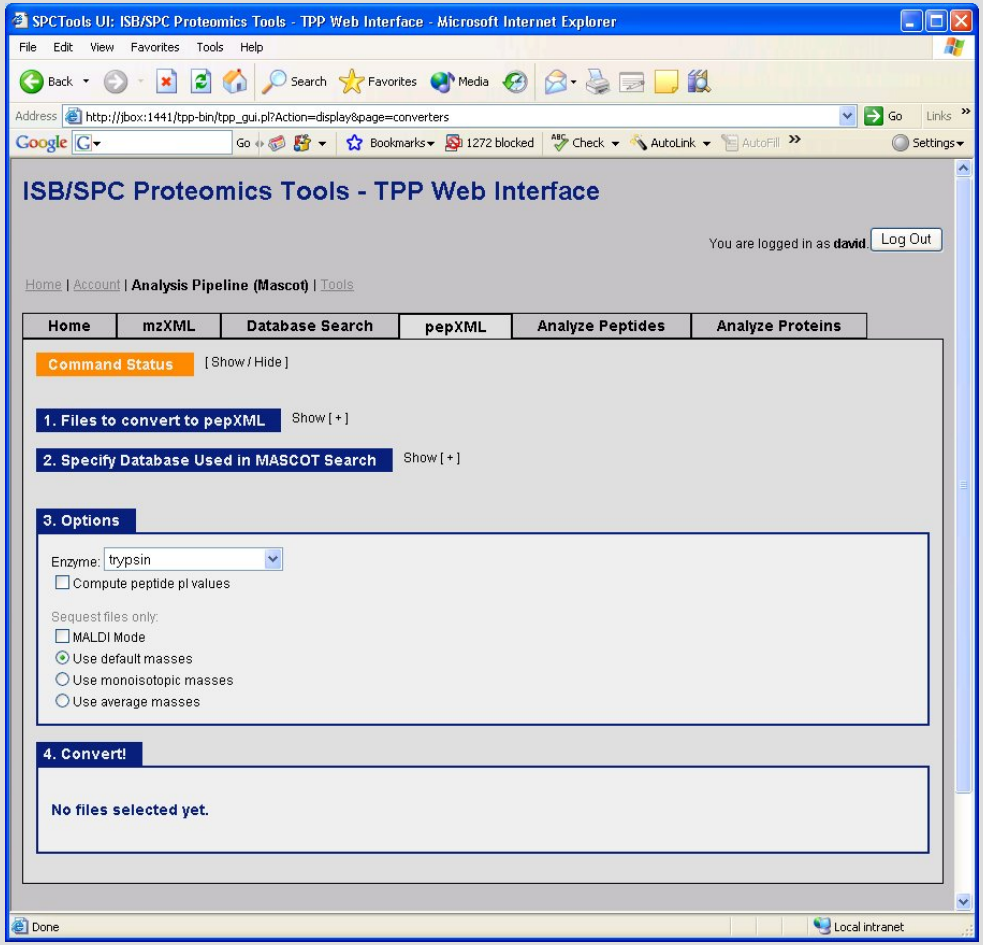


pepXML Features

- Standard format for representing searched data
- Stores information about peptide assignments to CID spectra
  - Mascot, Sequest, X!Tandem, ProbID, SpectraST
- Stores **PeptideProphet** validation results
- Stores results of Peptide Quantitation
  - XPress, ASAPRatio, Libra
- Stores DB Search results for one or more MS Runs
  - Search parameters stored separately for each run
- References one or more mzXML files containing the spectra searched

pepXML Converters

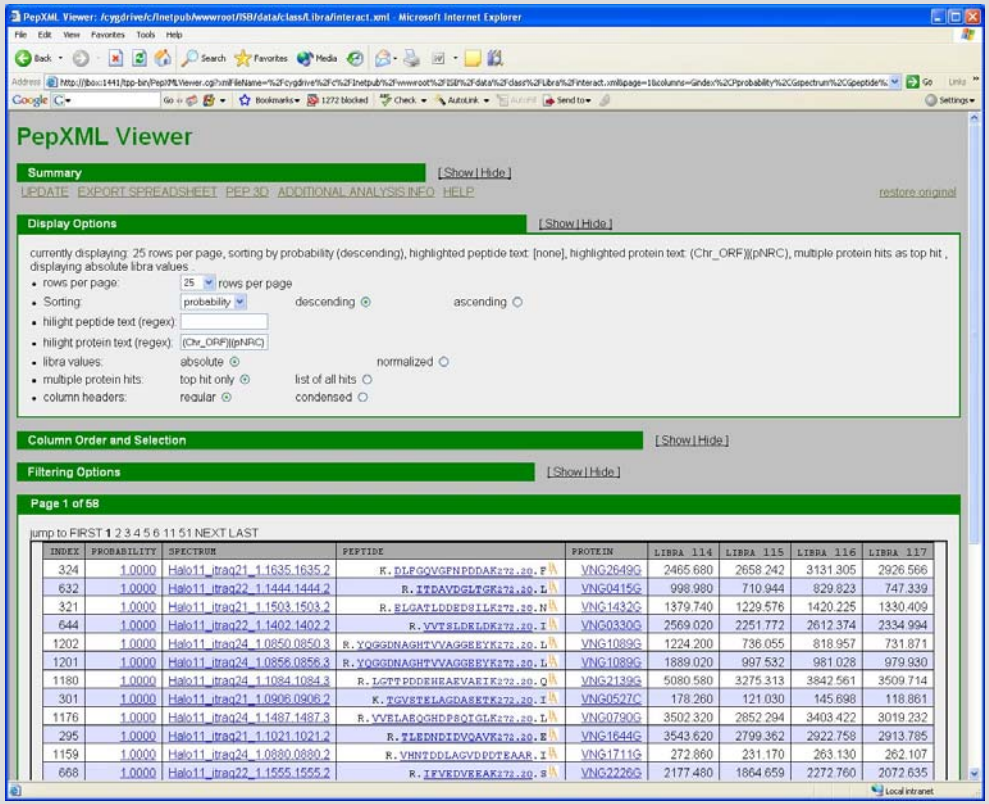
- Mascot2XML
  - .dat file input
  - Petunia pepXML tab
- Sequest2XML
  - .html file input
    - Output by runsearch
  - Peptunia pepXML tab
- Out2XML
  - Converts a directory of Sequest .out files



pepXML Writers/Readers

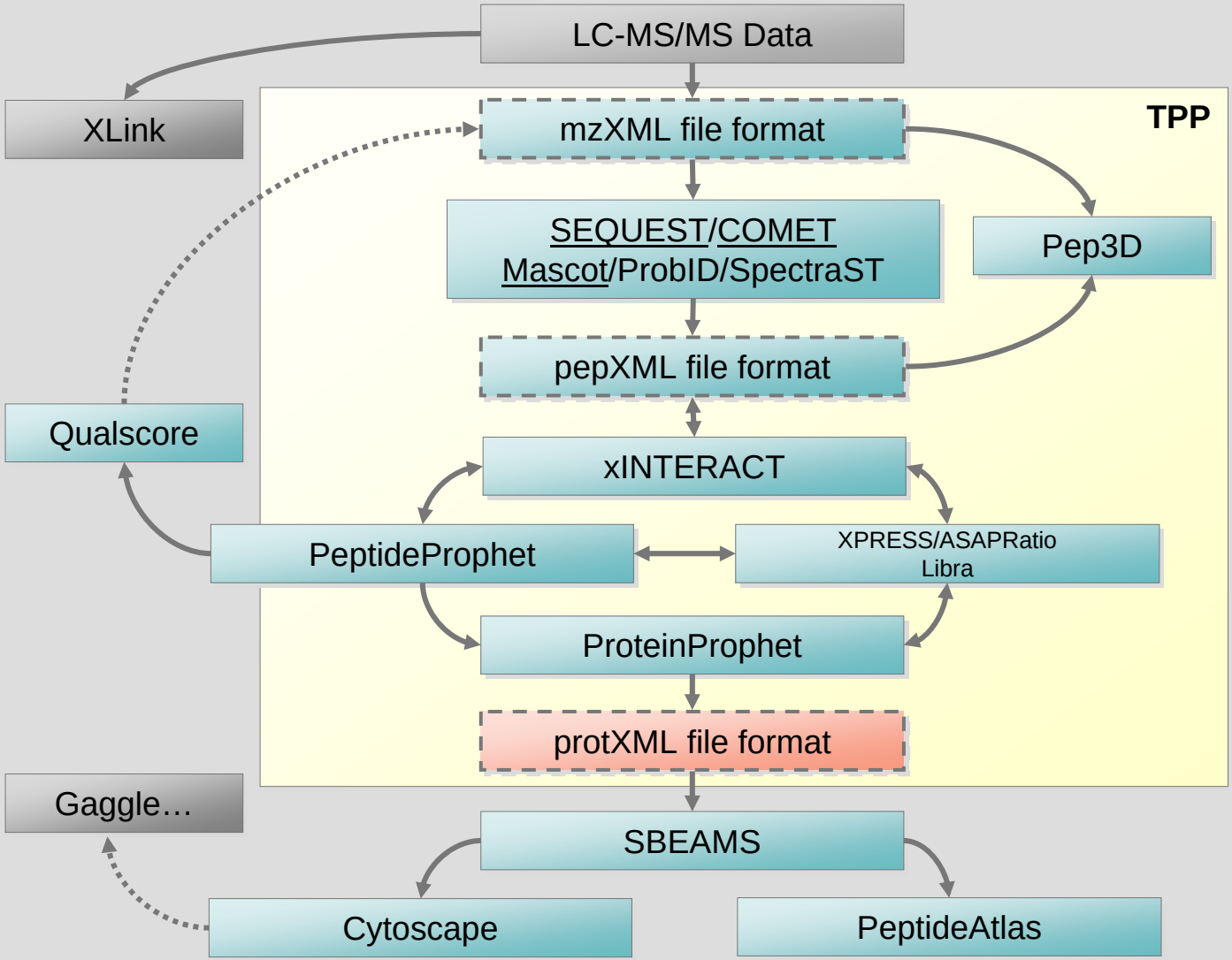
- **Writers**
  - **Converters**: Mascot2XML, Sequest2XML, Out2XML, Tandem2XML
  - **Search Tools**: SpectraST
  - **Validation Tools**: PeptideProphet
  - **Peptide Quantitation Tools**: Xpress, ASAPRatio, Libra
- **Readers**
  - **Visualization Tools**: Pep3D, PepXMLViewer
  - **Validation Tools**: PeptideProphet, ProteinProphet
  - **Protein Quantitation Tools**: Xpress, ASAPRatio, Libra
  - **Spectrum Filtering Tools**: Qualscore

pepXMLViewer.cgi



- Web application for parsing, sorting, filtering, displaying, and exporting results encoded in a pepXML document

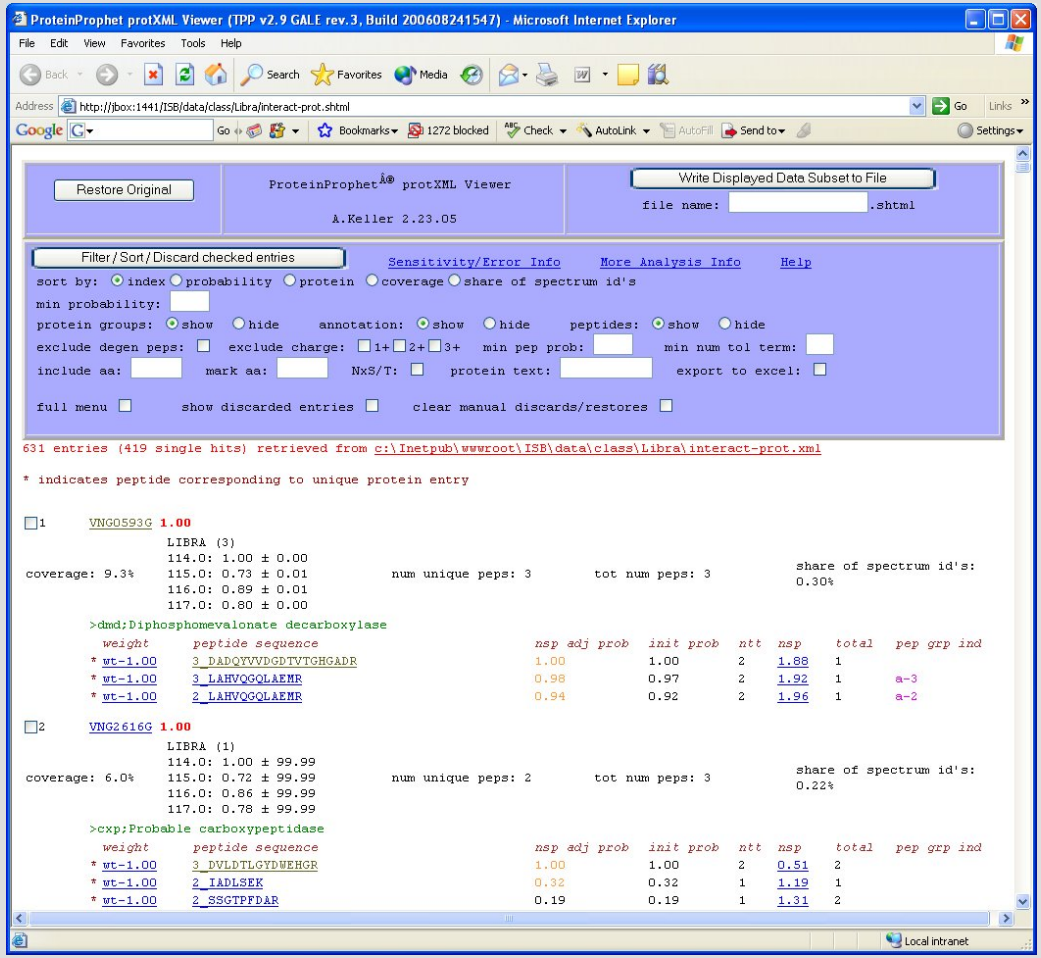
protXML


LC-MS/MS Data
XLink
TPP
mzXML file format
SEQUEST/COMET
Mascot/ProbID/SpectraST
Pep3D
pepXML file format
Qualscore
xINTERACT
PeptideProphet
XPRESS/ASAPRatio
Libra
ProteinProphet
protXML file format
Gaggle…
SBEAMS
Cytoscape
PeptideAtlas

protXML Features

- Open format for representing proteins identified by LC-MS/MS
- Stores ProteinProphet validation results
  - Inference of proteins based on peptide assignments to CID spectra
- Stores results of Protein Quantitation
  - XPress, ASAPRatio, Libra
- References one or more pepXML files
- Ingested by database storage and archiving systems
  - e.g. SBEAMS, CPAS

pepXML Writers/Readers

- **Writers**
  - **Protein Inference tools**: ProteinProphet
  - **Protein Quantitation tools**: Xpress, ASAPRatio, Libra
- **Readers**
  - **Visualization tools**: ProteinProphet Results Viewer (XSLT Translation)

protXML Display in TPP