Management of Proteomics Data:

2D Gel Electrophoresis and Other Methods

Philip Andrews

National Resource for Proteomics & Pathway Mapping

Michigan Proteome Consortium
University of Michigan

Outline of Presentation

- Introduction and Point of View.
- Role of Standards in Proteomics
- Management, Annotation, Distribution of Proteomics Data.
- Role of Open Source in Proteomics

Challenges in Proteomics.

- The gaps between Basic Research, Technical Development, and Commercialization narrow dramatically in rapidly-evolving fields.
- Commitment to a single technology may be fatal in a rapidly evolving field.
- Computational and IT infrastructure is critical and limiting for Proteomics.

Proteome Informatics and the Rate of Progress in Proteomics

- Availability of appropriate software.
- Effective information management.
- Lack of basic standards.

Challenges in Proteome Informatics

- **Proteome technologies evolve rapidly.**
  - Software always lags behind hardware.
  - Software always lags behind applications.

- **Current database structures are inadequate** – missing data, data quality, complex interactions, changing interactions, new data types, pedigree, etc.

Challenges in Proteome Informatics II

- Ability to extract knowledge from large, complex biological datasets is still evolving.
- Mechanisms for annotation of genome databases need to be improved.
- Planning large-scale experiments must be automated.

Data Challenges in Proteomics

- High-throughput technologies generate large amounts of data.
- Data are heterogeneous.
- Data relationships are complex.
- There are minimal standards.

Data Complexity: **Proteome Mapping Data**

- 2D Gel Images
- MS spectra
- MS/MS spectra
- LC MS
- 2D LC/ MS/MS
- Tune files
- Sample data
- Data analysis parameters

Proteomics Is Both Research and Production.

- Standards are more easily applied to production.
- Standards should conform to technology, not the other way around.

Why Do We Need Proteome File Standards?

- Standardize reporting.
- Data pipelines require batch export of files.
- Allow more facile development of open source software for proteomics.
- Data longevity.

XML Files for Proteomics

- Pros
  - Structured
  - Easily readable
  - Translatable into other formats
  - Amenable to open source development
- Cons
  - Inefficient
  - Complexity can be problem

Development of Proteomics XML Standards

- ISB (www.isb.org)
  - mzXML
- EBI/HUPO (www.hupo.org, www.pedro.org)
  - Pedro
  - mzDATA, MIAPE
- www.gaml.org

XML Summary

- A near-term solution.
- Useful for data exchange.
- Multiple formats will be necessary.
- Formats will change.
- Mechanism for timely updates will be necessary.
- We will support all major XML formats.

But Other Formats Could Be Better

- Computationally, restructuring files to better fit your data structure can lead to increases in efficiency.
  - It is important that the file format be publicly available.
  - Parsers and translators should also be available.
  - Batch conversions should be supported.

- **The standard should be openess in data formats.**

Information Management in Proteomics

- Low level data management (e.g., LIMS).
- Curation tools (goal is to automate).
- Higher level information management (e.g., metadata).
- Aggregation and Integration systems.

PRIME

Proteome Research Information Management Environment

General Goals:

- Integrated, simple, flexible system to acquire, manage, and mine Proteomics data (code generation).
- Useable by distributed groups.
- Support all data types and standards.
- Secure.

PRIME

Proteome Research Information Management Environment

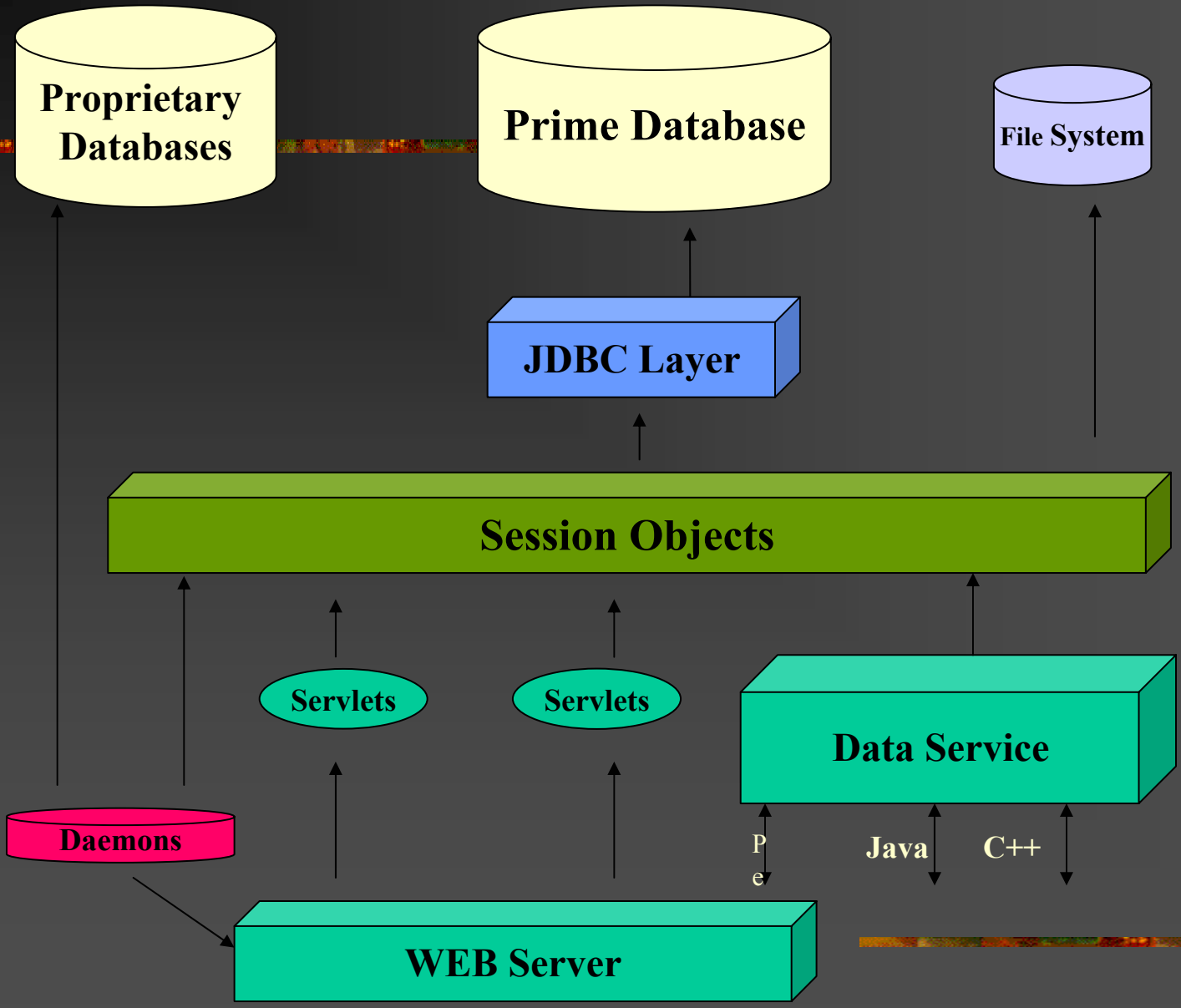
Specific Design Goals:

- Open architecture development.
- Multi-tier with HTML user interface.
- Distributed system.
- Scaleable system.
- Compatible with other databases.
- Flexibly accommodate data types.
- Low maintenance.
- Easily extensible.
- Developed using open standards.

System Components

- Laboratory data management system,
- Data viewers (2D Gel images, chromatograms, and mass spectra),
- Automated data collection from instruments,
- Automated protein database search engines (Mascot, Prot. Prospector, X!Tandem),
- Data discovery toolkits.

Prime Architecture

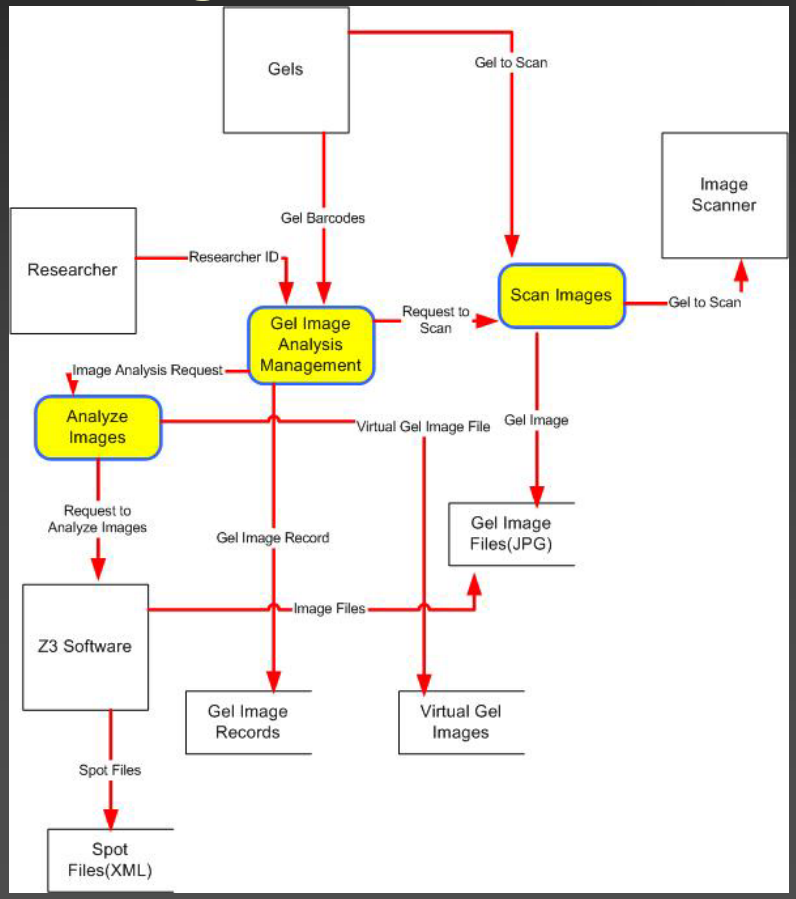


Work Flow Levels

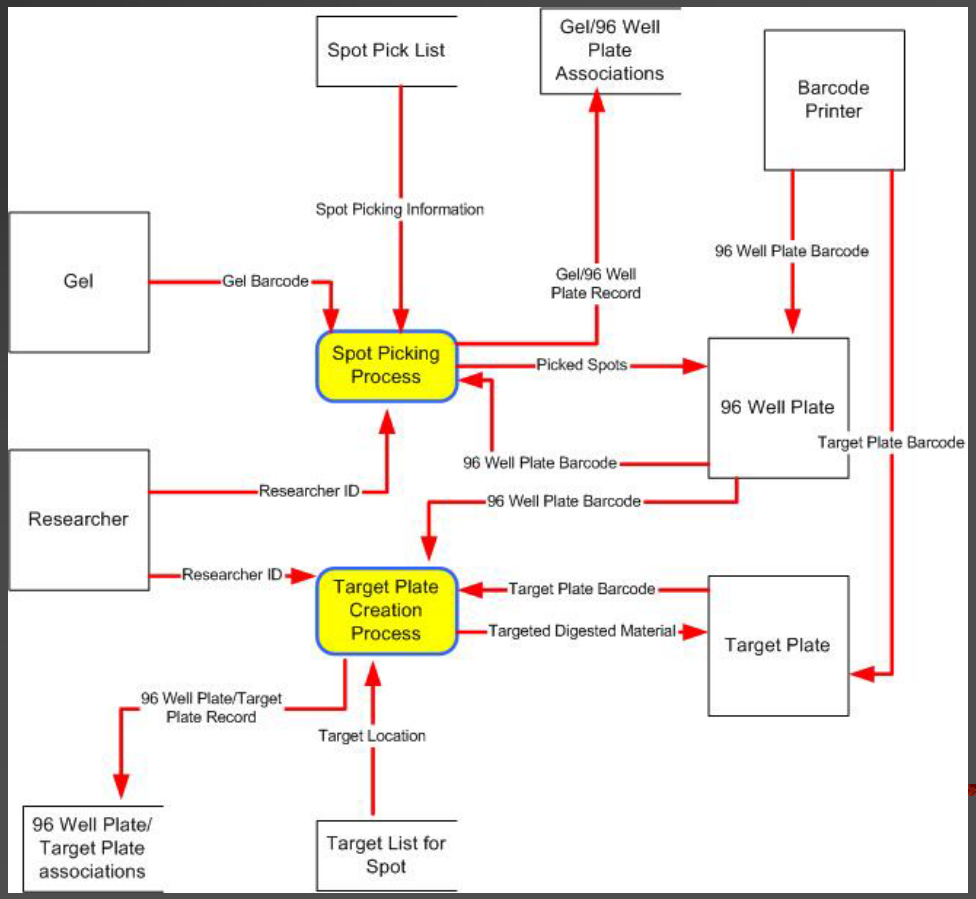
- Lab Level (Samples)
- Data Level (Processing)
- Administrative Level (Paper)

Lab Level Work Flow

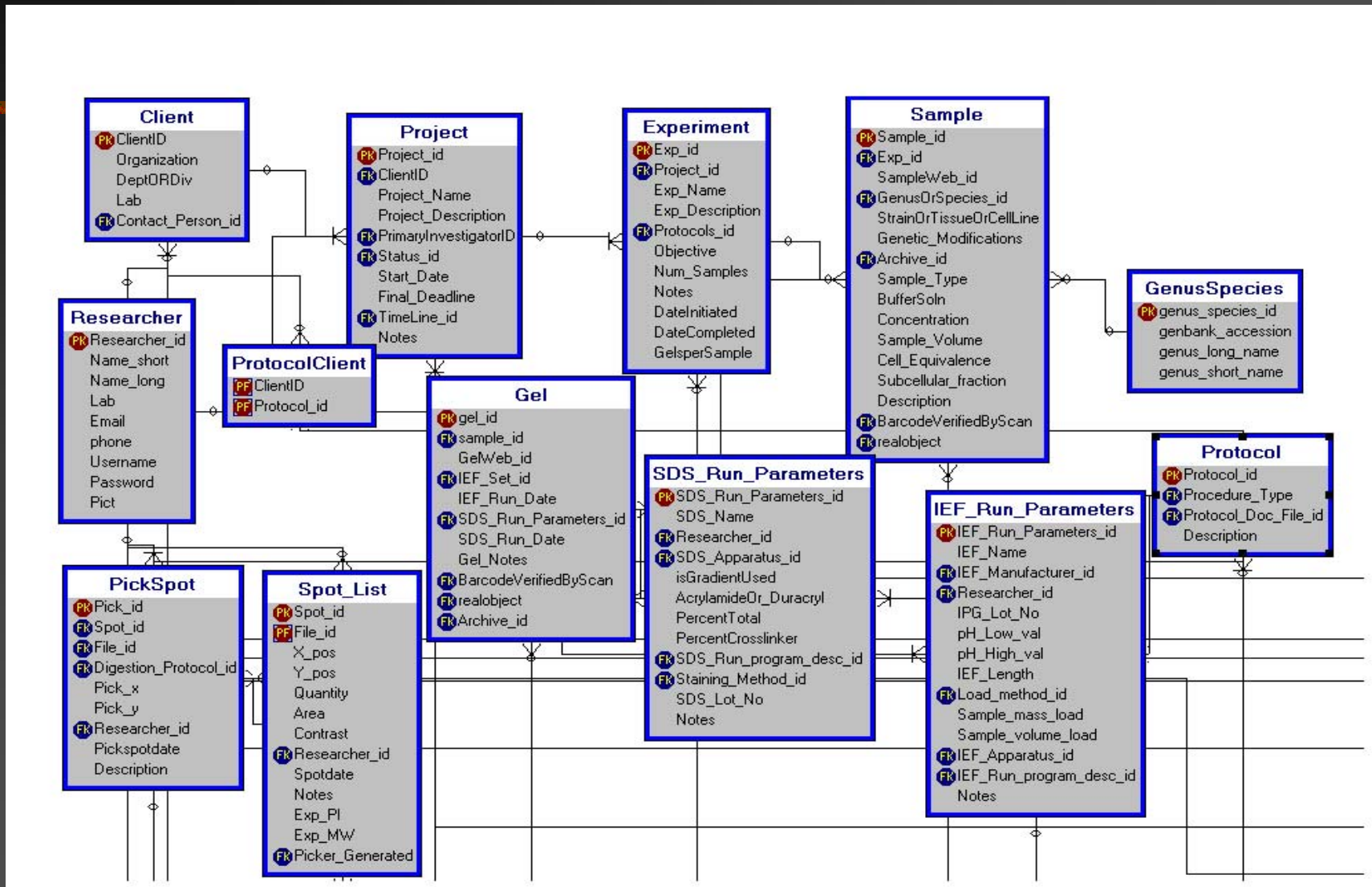
- Image Analysis Top Segment



- Robotics Top Segment



PRIME Table Structure Segment



PRIME Summary Stats

- 220 tables.
- ~2,200 Java source files
- ~ 10^6 lines of code.

- prime.proteome.med.umich.edu

Uses of PRIME

- Documentation and management.
- Curation.
- Collaboration tool.
- Provide data access for reviewers.
- Host public access to data.

Distribution of Proteomics Data: The Cathedral vs the Bazaar Revisited.

- Centralized system.
  - Must deal with many of the same issues as standards development.
  - May be better suited for metadata.
- Distributed system.
  - Distributes costs of maintenance.
  - Puts ‘ownership’ in hands of interested parties.

What Are Issues for Distributed Systems?

- Ownership of data.
- Persistence.
- Maintenance of context.
- Quality Control.
- Security.
- Cost (long-term maintenance).

Challenges in Proteome Informatics

- **Proteome technologies evolve rapidly.**
  - Software always lags behind hardware.
  - Software always lags behind applications.
- **Instrument development is negatively impacted by software development (cost and time).**

The (partial) Solution: Open Source Efforts in Proteomics

- Progress in Proteomics will be faster if a robust open source community is developed.
- Open source efforts allow the community to respond to new technologies rapidly.
- Open source allows each individual in the community to respond to their own needs.
- Cost of development is shared.
- Open Source is compatible with commercial proprietary software.

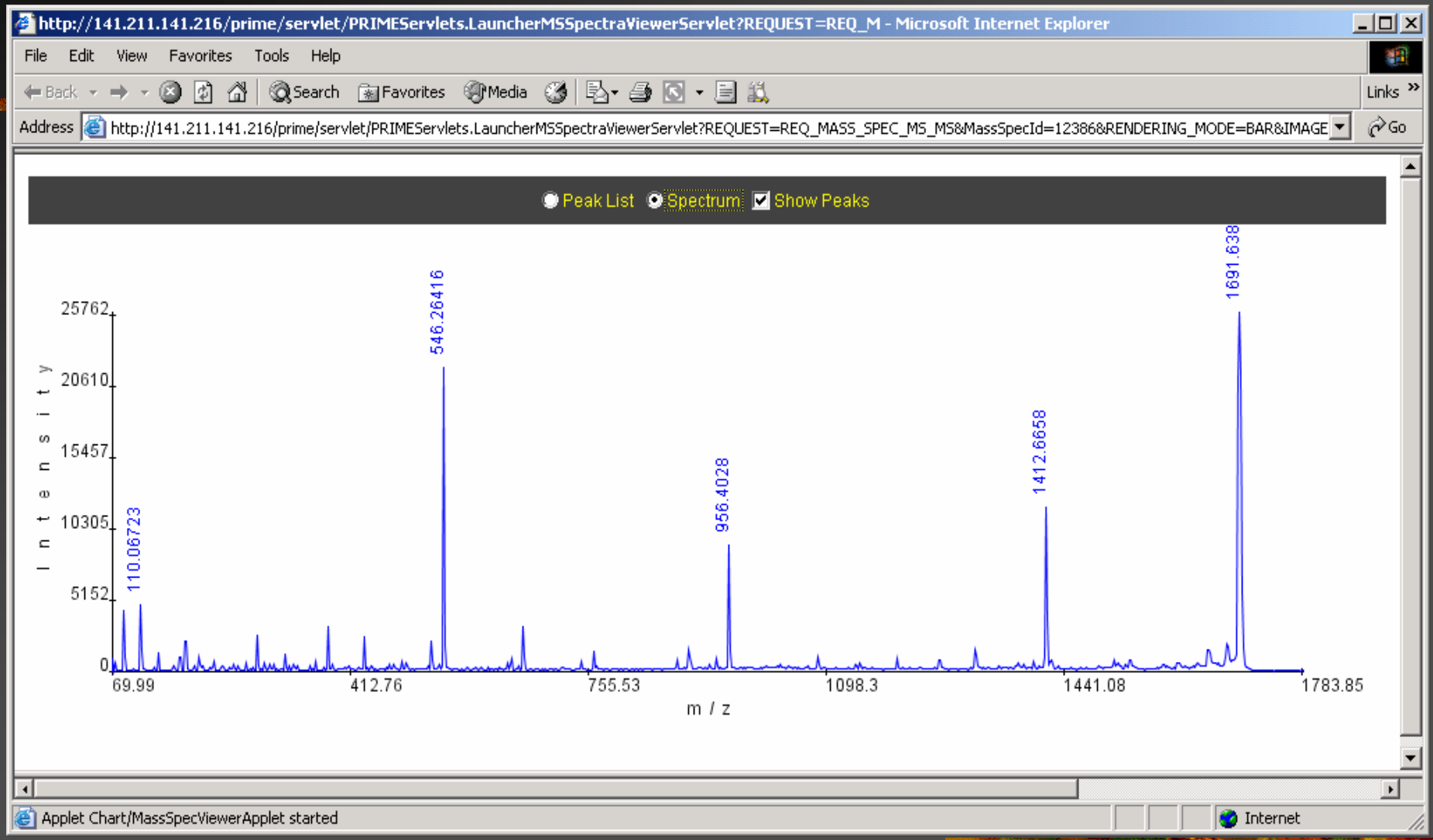
Open Source Websites

- www.proteomecommons.org
- www.thegpm.org
- http://www.systemsbiology.org/
- www.jasondunsmore.org/projects
- www.bioexchange.com/tools/
- http://bioinformatics.icmb.utexas.edu/OPD/
- http://www.peptideatlas.org/

www.proteomecommons.org

- Versioning system for organization and archives.
- Full source code and documentation downloadable.
- Spectra used for development and testing downloadable.
- Digital signatures used for security.
- Allows mirroring and bittorrent so users may host their own projects.
- Supports metainformation attached to projects.
- Code-in-progress accessible.

Spectrum Viewer Module in PRIME



Open Source Spectrum Viewer

- Uses WebStart
- Displays peak lists or spectra.
- Allows usual data manipulations.
- Generates peak lists.
- Allows spectrum annotation.
- Exports publishable-quality images.

Selected Datasets Distributed on Proteome Commons

- Development datasets
- ‘Gold Standard’ datasets
  - ~50 eukaryotic proteins
  - 400 human proteins
- Hosting/mirroring other datasets

Acknowledgements

Proteomics

- Mary Hurley (MPC)
- Eric Olsen (NRPP)
- Gary Rymar (NRPP)
- John Strahler (NRPP, MPC)
- Donna Veine (NRPP)
- Angela Walker (NRPP, MPC)
- Xuequn Xhu (NRPP)

NRPP

- Russ Finley (WSU)
- Brett Phinney (MSU)
- Trey Ideker (UCSD)
- Curt Wilkerson (MSU)

Acknowledgements

Proteome Informatics

- Thomas Blackwell (UM)
- Jayson Falkner (UM)
- Russ Finley (WSU)
- Catherine Grasso(UM)
- Trey Ideker (UCSD)
- George Michailidis (UM)
- Panagiotis Papoulias
- David States (UM)
- Peter Ulintz (UM)
- Curt Wilkerson (MSU)

Software Development

- David Lentz
- Narayani Anand
- Hsueling Chang
- Panagiotis Papoulias

Websites

- www.proteomecommons.org
- www.proteomeconsortium.org
- www.proteome.med.umich.edu
- www.proteomecenter.med.umich.edu