What is it, and how did it get there?

(or, making data make sense…)

Chris Taylor, European Bioinformatics Institute

Metabolomics Society, http://www.metabolomicssociety.org/

HUPO Proteomics Standards Initiative, http://psidev.sf.net/

Roles of Data Standards Bodies

Many data formats, many databases, varying requirements

- *Examples; MAGE (transcriptomics), SMRS/FDA requirements, MCP Guidelines, ASTM/ANSI (AnIML [animl.sf.net]), PSI projects*
- *Role of data standards bodies..?*
  - *data exchange/transport formats*
  - *minimum reporting requirements*
  - *controlled vocabularies / ontologies (≡ dictionary)*
- *Not to…*
  - *specify database structure*
    …different uses, warehousing
  - *recommend experimental protocols (aaaaargh!!!)*
- *Inclusivity is the key*
  - ***more important** than the engineering*
  - *active outreach, globally (so that silence = assent)*

Metabolomics Soc. — Background

Founded in 2004 with several objectives:

- *Promote growth and development of the field internationally*
- *Provide the opportunity for collaboration and association*
  - *workers in that science and in related sciences*
  - *academia, government and industry*
- *Provide opportunities for presentation of research achievements and creation of workshops*
- *Promote the publication of meritorious research in the field.*

Met. Soc. — Working groups

Oversight committee plus the following:

A. Biological sample context

Extensions: in vivo / mammalian biology, plant biology, in vitro / cell culture biology, environmental analysis

B. Chemical analysis

Lab techniques (mass spec, CE, etc.)

C. Data analysis

Bioinformatics, statistics

D. Ontology

Scope = all of the above (→ FuGO)

E. Data Exchange

Scope = all of the above (NMR format , FuGE, reuse of PSI formats)

A, B and C will all generate guidelines and SOPs…

HUPO — Background

Founded in 2001 with several objectives:

- *Consolidate national and regional proteome organisations*
- *Assist in the coordination of public proteome initiatives*
- *Engage in scientific and educational activities*

Tissue proteome projects and other initiatives:

- *Plasma, Liver, Brain, Glyco; PSI and Antibody initiative*

Congresses (in addition to many smaller meetings):

- *Paris, Nov. 2002; Montreal, Oct. 2003; Beijing, October 2004*

HUPO PSI — Workgroups

MI: Molecular Interactions

- First project to mature — Molecular Interaction Format (MIF)
  - Hermjakob et al., Nature Biotech (2004) **22**(2): 177-83
- Several PPI databases exist; BIND, DIP, MINT, Hybrigenics, IntAct
  - data provided in many different formats, not synchronised
- The standard (MIF) defines a minimal data model
  - allows scientists to provide core data, with back references
  - simplifies synchronisation (**http://imex.sourceforge.net/**)
  - viewable with Cytoscape™ (http://www.cytoscape.org/)

PTM: Post-translational modifications

- Combining ResID, UniMod (and DeltaMass) into one resource

GPS: General Proteomics Standards

- Overall view of proteomics workflows, inter-omics collaboration

MS: Mass Spectrometry

- Formats, standards and vocabulary for mass spectrometry

GPS & MS: Rationale

The generation and analysis of proteome data are widespread

- *High-throughput approaches are commonplace (if troublesome)*
- *Techniques continue to increase in complexity (DiGE, iTRAQ, SPR)*

Publicly available proteomics data is rather limited

- *Sample extraction and preparation usually undocumented*
- *Analytical methods employed in deriving conclusions absent*
- *Still no widely used databases of (for example) mass spectral data*

Standard integrated representations of methods (metadata) and data from proteomics experiments are required

- *Will facilitate handling, exchange and dissemination of data*
  - *development of effective search/analysis tools*
  - *derivation of maximum value from data sets*

GPS & MS: The Scope of our Remit

Sample generation

Origin of sample

project-context, (hypothesis), organism, environment, preparation, paper citations

Sample processing, gels and gel informatics

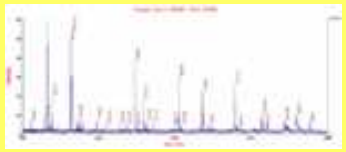
Gels (GelML+), columns, 'chips', others

images, gel type and ranges, band/spot coordinates, quantitation

stationary and mobile phases, flow rate, temperature, fraction details

Mass Spectrometry (mzData)

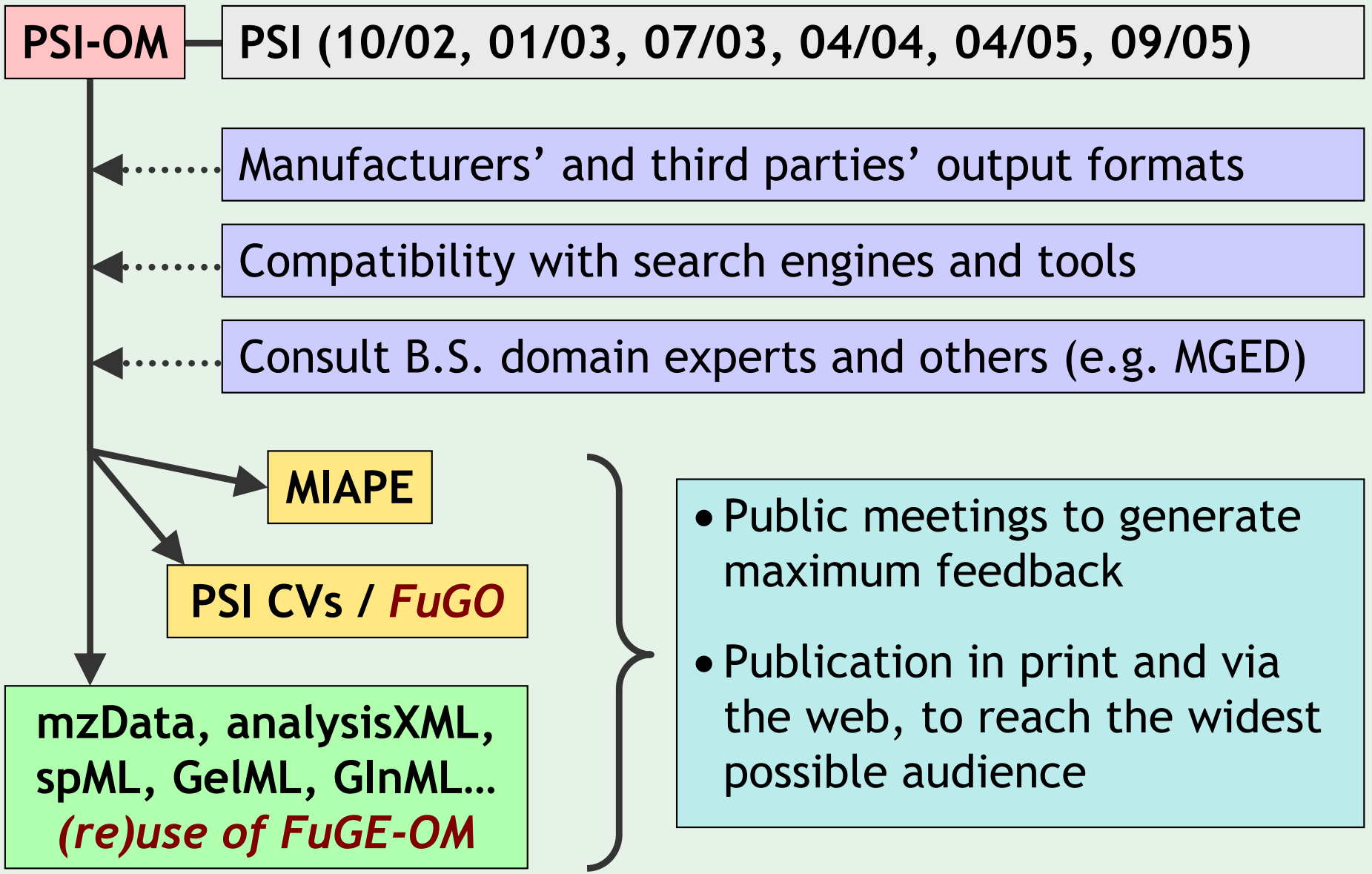
machine type, ion source, voltages



Mass Spec Informatics (analysisXML)

peak lists, database name + version, partial sequence, search parameters, search hits, accession numbers, quantitation

Evolution of HUPO PSI Standards



mzData now established…

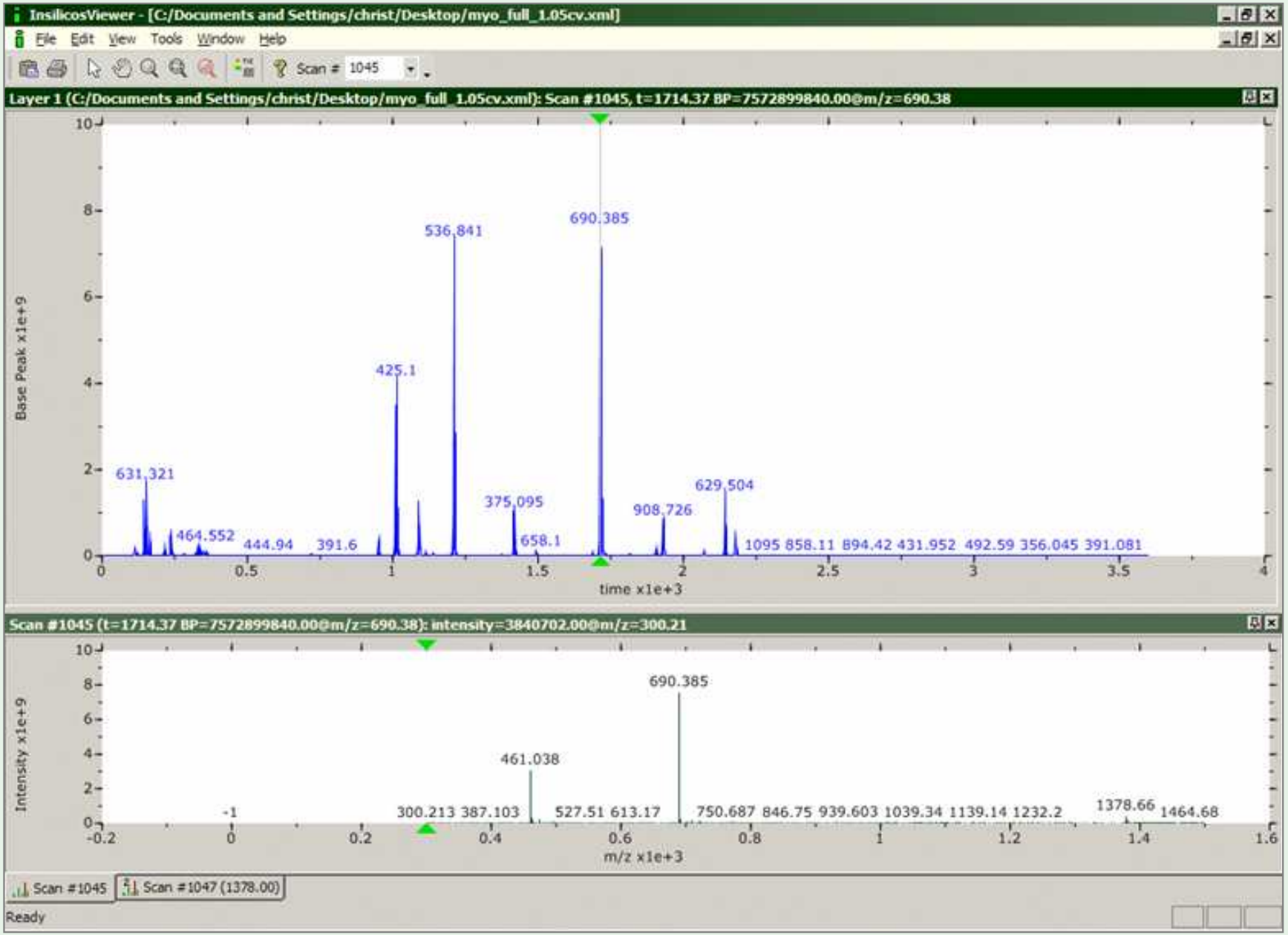
Mass spec vendors (in bundled software)

- Thermo (*completed*)
- Bruker Daltonic (*completed*)
- Kratos / Shimadzu (*ongoing*)
- AB / MDS Sciex (*ongoing*)
- Agilent (*ongoing*)
- Waters (*ongoing*)

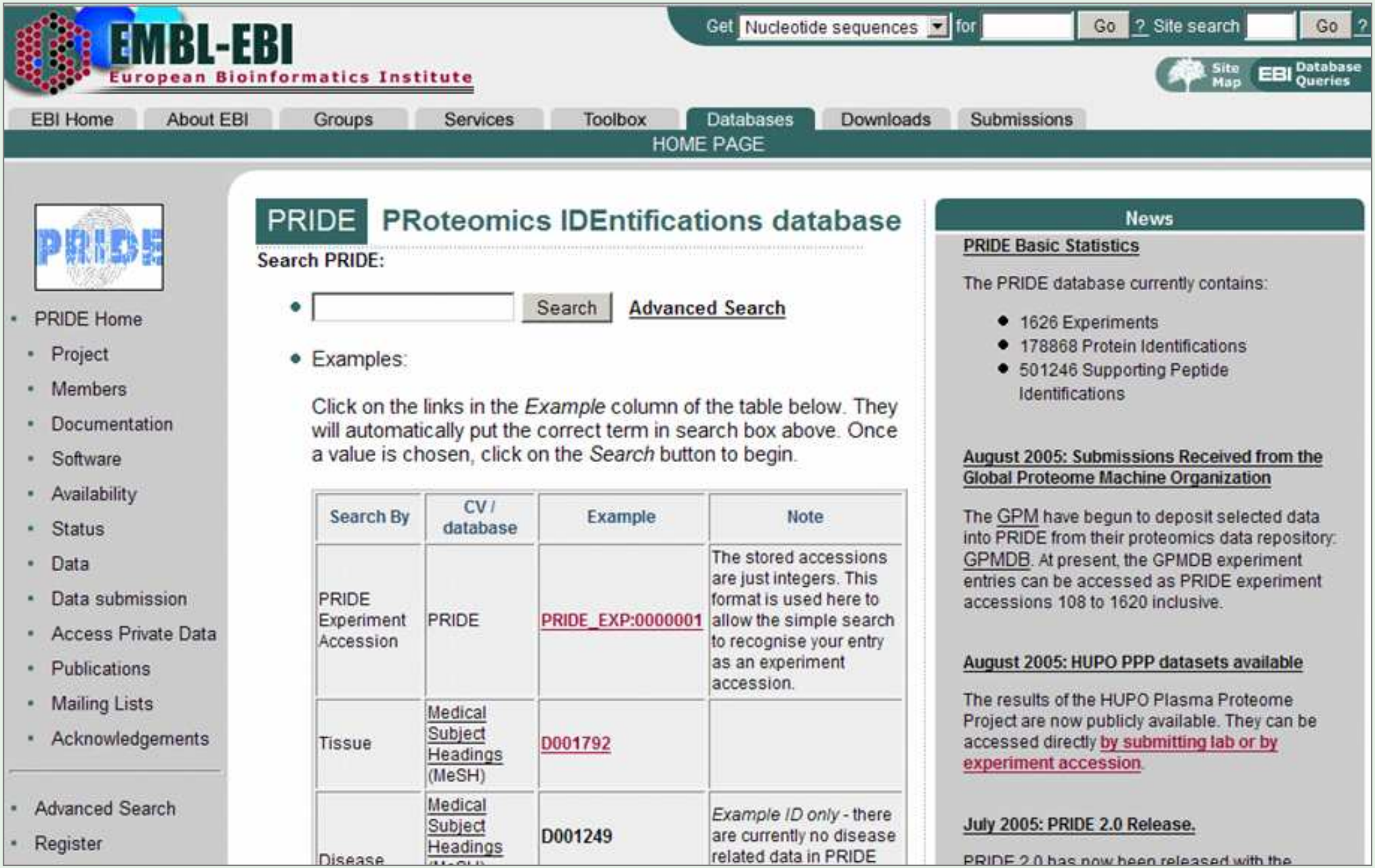
Software companies / organisations

- GeneBio [Phenyx] (*completed*)
- Matrix Science [Mascot] (*completed*)
- Swegene Bioinformatics [Proteios] (*completed*)
- The GPM Organization [X!Tandem] (*completed*)
- Swiss Institute for Bioinformatics [Aldente] (*ongoing*)
- Proteome Systems Ltd. (*ongoing*)
- Insilicos [viewer] *(completed)*
- European Bioinformatics Institute [PRIDE] *(completed)*

Insilicos Viewer



http://www.ebi.ac.uk/pride



Upcoming formats and the use of FuGE

***GelML** (alpha release, spring 2006)*

- *Preparation and running of 1D and 2D gels*
- *Image capture, image description (resolution etc.)*

***GInML** (alpha release, spring 2006)*

- *Analysis of gel images*
  - *spot finding, image warping, spot volumes, ensembles*

***spML** (alpha release, spring 2006)*

- *Columns, centrifugation, CE, miscellaneous IEF methods*
- *Tagging, digesting*

***AnalysisXML** (beta release, spring 2006)*

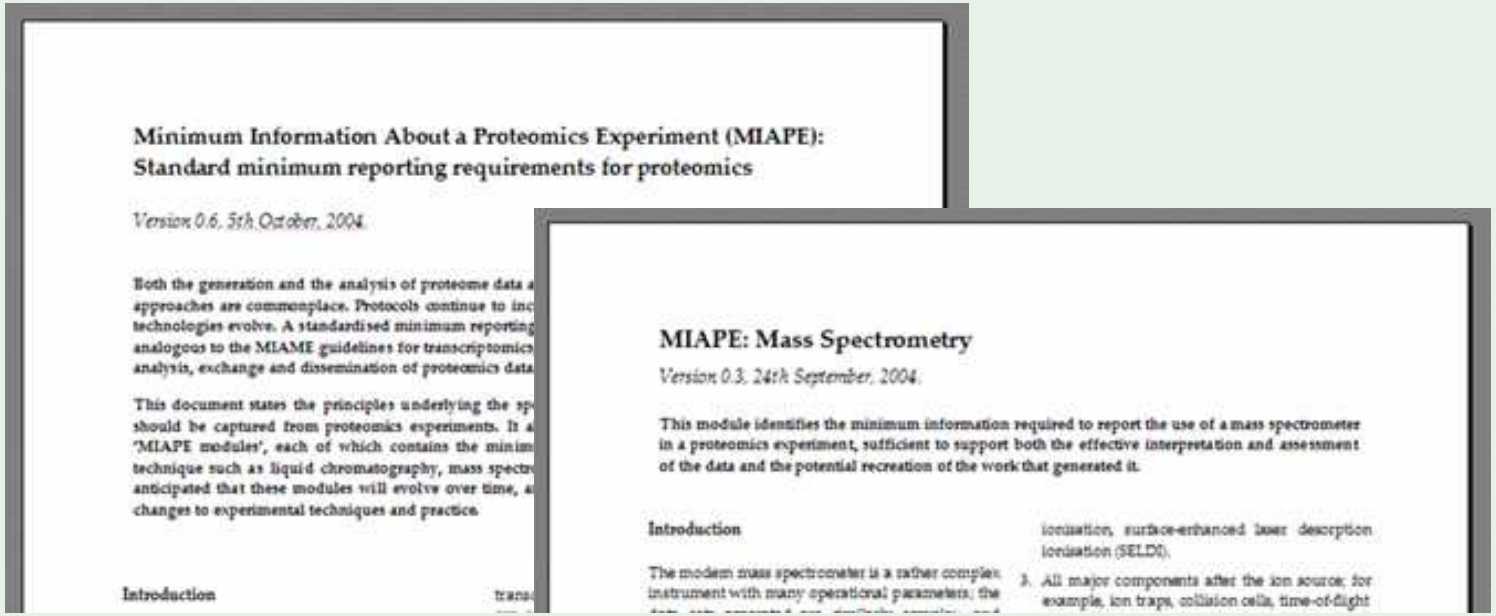
- *What flew down my mass spec?*
  - *protein/peptide ID (Mascot, Sequest), lipids, metabolites*

FuGE (http://fuge.sf.net)

- *Act as the 'glue' for all these modular formats*
- *Provides superclasses to develop all of the above from*

Minimum Reporting Requirement

- ***MIAPE (http://psidev.sf.net/gps/#miape)***
  - *Minimum Information About a Proteomics Experiment*
  - *Requirement to be enforced by journals, repositories, funders*
  - *Technology-specific modules associated with a parent document*
    - *Users should assemble the relevant modules into a bespoke reporting requirement for their particular workflow*



Minimum Information About a Proteomics Experiment (MIAPE):
Standard minimum reporting requirements for proteomics

Version 0.6, 5th October, 2004.

Both the generation and the analysis of proteome data a
approaches are commonplace. Protocols continue to inc
technologies evolve. A standardised minimum reporting
analogous to the MIAME guidelines for transcriptomics
analysis, exchange and dissemination of proteomics data

This document states the principles underlying the sp
should be captured from proteomics experiments. It a
'MIAPE modules', each of which contains the minim
technique such as liquid chromatography, mass spectr
anticipated that these modules will evolve over time, a
changes to experimental techniques and practice.

Introduction

MIAPE: Mass Spectrometry

Version 0.3, 24th September, 2004.

This module identifies the minimum information required to report the use of a mass spectrometer in a proteomics experiment, sufficient to support both the effective interpretation and assessment of the data and the potential recreation of the work that generated it.

Introduction

The modern mass spectrometer is a rather complex instrument with many operational parameters; the

ionisation, surface-enhanced laser desorption ionisation (SELDI).

3. All major components after the ion source; for example, ion traps, collision cells, time-of-flight

Minimum Reporting Requirement

The MIAPE 'parent document'— underlying principles:

1. Sufficiency

- *Unambiguous description of the experimental context*
- *Allow understanding of the results and their interpretation*
- *Sufficient to permit a critical evaluation of same*
- *In principle allow recreation of the work*

2. Practicability

- *Achieving MIAPE compliance should not be so burdensome as to prohibit the widespread use of the guidelines*

The MIAPE 'modules' — technology-specific guidelines

- *One module per technology; MS, MSI, GE, (GI, SP), (SO & PS..?)*
- *Each module (to be) validated by an ad hoc expert committee*
- *Users assemble modules into a bespoke reporting requirement*

MIAPE: Summary

The MIAPE minimum reporting requirement

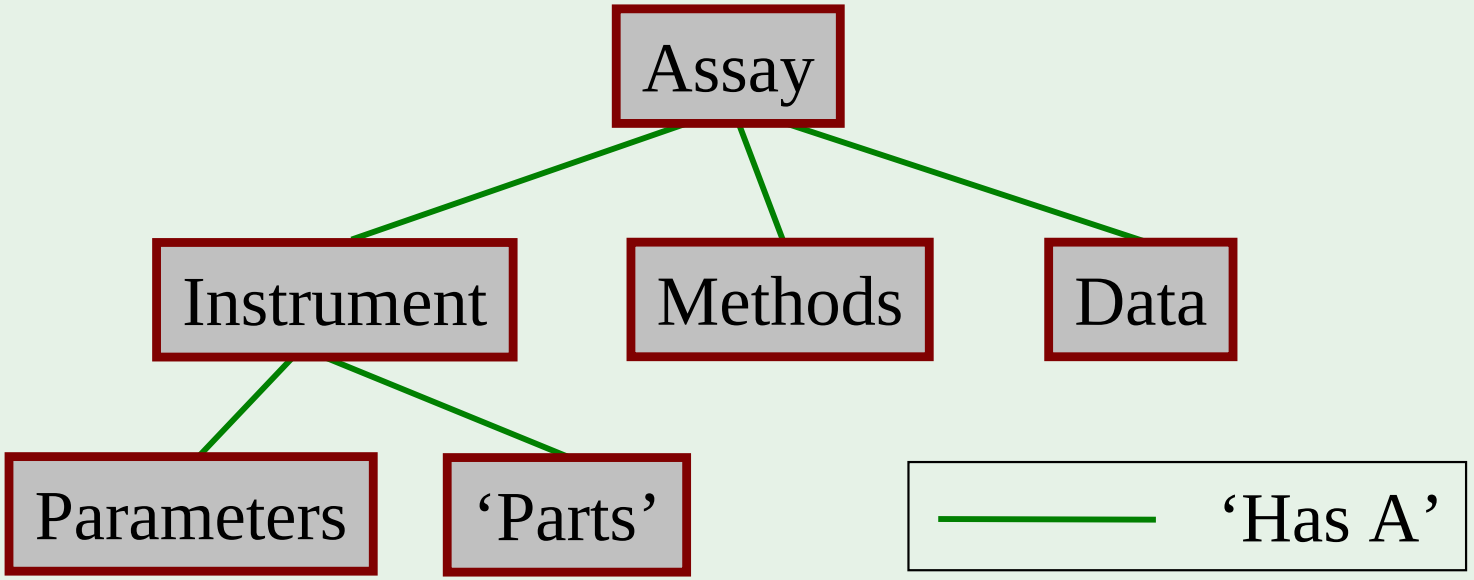
- Establish provenance and relevance of a data set
- Sophisticated search and analysis becomes possible
  - e.g. exclude data generated by a particular technique
- Extraction of maximum value from MIAPE-compliant data sets
- High-level module (project design) requires broad input
  - All biological and technological communities

Progress in gaining acceptance

- Positive response from UK-based funders and government
  - 'Recommend' for now, enforce in the future
- Interest expressed by the National Institutes of Health (USA)
- Strong support from the major journals in the field
  - MIAPE will be factored into the process of developing common 'quality' guidelines

The Functional Genomics Future

- ***Functional Genomics Ontology (FuGO)***
  - *Common descriptors, domain-specific extended sets*
    http://fugo.sourceforge.net/
- ***Functional Genomics Experiment (FuGE) [OM / ML]***
  - *provide superclasses (to anchor other models) and workflow 'glue'* **http://fuge.sourceforge.net/**

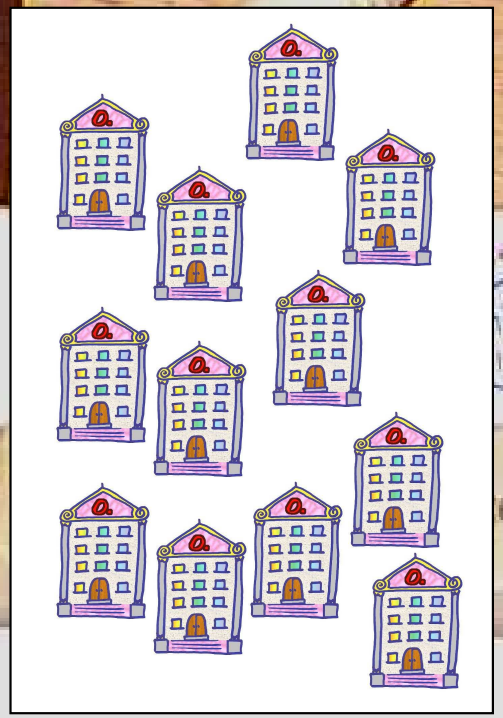


The Functional Genomics Future

- ***Core biological description should be shared***
  - *Study design and sample generation*
  - *Requires extensive community liaison*
  - *Heavy impact on all three resources (ontology, formats, reporting)*
- ***A 'grand collaboration' (data formats & ontology)***
  - ***Proteomics; PSI** (fairly well covered here…)*
  - ***Transcriptomics; Microarray Gene Expression Data Soc.***
    - *Academia, industry , journals, government*
    - *MIAME, MAGE, MGED Ontology; RSBI; NWG*
  - ***Metabol/nomics; Metabolomics Society***
    - *New body; academia, industry, journals, government*
  - ***Others**: Genome sequencing ('MIGS'), lipidomics, ++*

Variety — is it the spice of life?

1. Inaccessible



2. Multifarious

3. “Just right”

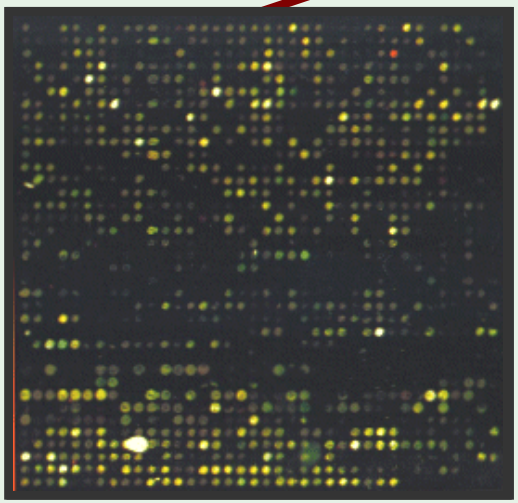
- *The ‘wealth’ of data formats is already a major problem*
- *For ontologies (including accessions), the problems could grow*
- *General reporting requirements are still a rare beast but…*

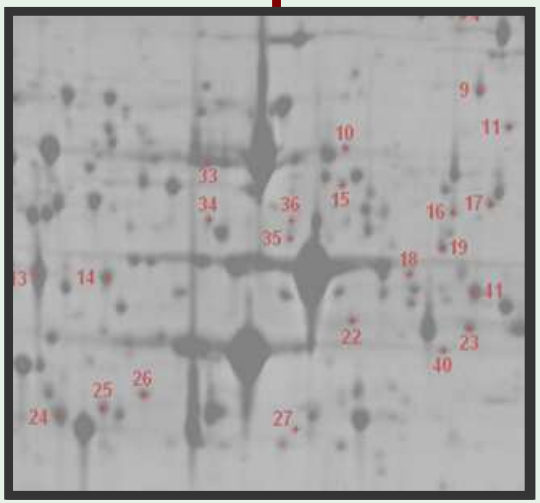
Synergy for (not so) Dummies™

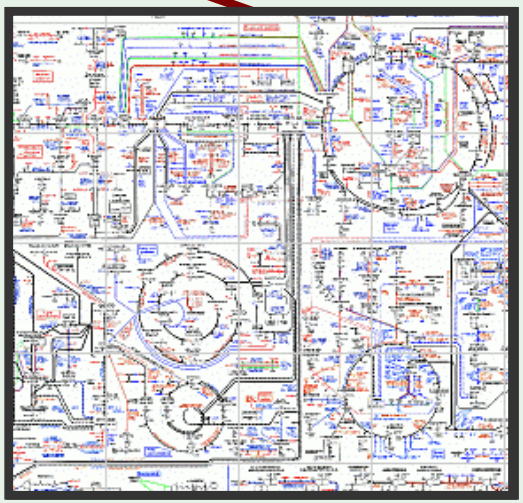
Diverse community-specific extensions

Generic Features (origin of biomaterial)

Generic Features (experimental design)







My other favourite keyboard…

International Speak Like a Pirate Day: September 19th

Issues for tomorrow's session

Room G4, Gibbs Building, King's College
13.30-15.00 tomorrow (10th)

Issues for discussion:

- ***Status of extant projects***
- ***Codevelopment of resources with proteomics, transcriptomics, etc.***
  - *Data capture formats (fuge.sf.net, mzData, spML, NMR format?)*
  - *Ontology terms (fugo.sf.net)*
  - *Reporting requirements (SMRS++, MIAPE, MIAME, MIGS, etc.)*
- ***Future meetings schedule***
  - *Next PSI meeting: 21-23 April, SF*
  - *MetSoc meeting: 24-28 June, Boston*