

Data, Metadata and Reporting Standards for Metabolomics

Nigel Hardy

Dept. Computer Science

Aberystwyth University

Metabolomics

- “all low molecular weight molecules”
- biochemical complement of
 - cell, tissue, organ, organism, colony, medium
- a branch of “functional genomics”
 - Genomes
 - Transcriptomics
 - Proteomics
 - Metabolomics
 - Greatest complexity?
- Contributing to “Systems biology”

The shape of the problem

- Diversity of chemical species
 - Plant kingdom
 - 100,000? 200,000? 250,000?
 - One organ – 1,000-2,000
 - Animals
 - xenobiotics
 - co-mensals
- Spatial distribution
 - sub-cellular compartmentalization
- Organisms are not isolated
- Dynamic range of concentrations
 - many order of magnitude
- Speed of turnover
 - $\ll 1s$
 - intermediate compounds not detected

Defined by Technology?

- Detection
 - Mass spectrometry (**MS**)
 - “molecular” ions or fragmentation patterns
 - Nuclear magnetic resonance spectrometry (**NMR**)
 - Infra-red absorption (**IR**)

- Prior separation (Optional)

- gas chromatography (**GC**)
- liquid chromatography (**LC**)
- capillary electrophoresis (**CE**)

Many combinations

GC-MS

LC-MS

LC-NMR

...

Applications

- Plants, animals, microbes, medical, environmental
- Discrimination
 - No chemical explanation required
 - Food adulteration and quality
 - Pre-clinical drug trials
 - ? medical diagnosis
- Scientific investigation of metabolism
 - Pathways and fluxes
- Systems biology
 - Integration with data from transcriptomics, proteomics ...

Fingerprinting and “metabolomics” and “true metabolomics”

- Fingerprint
 - No chemical identity
- Targetted metabolomics
 - Estimate pre-selected metabolites
 - Estimating abundance
- True metabolomics
 - Detect and estimate metabolites
 - Detecting chemical species and their abundance

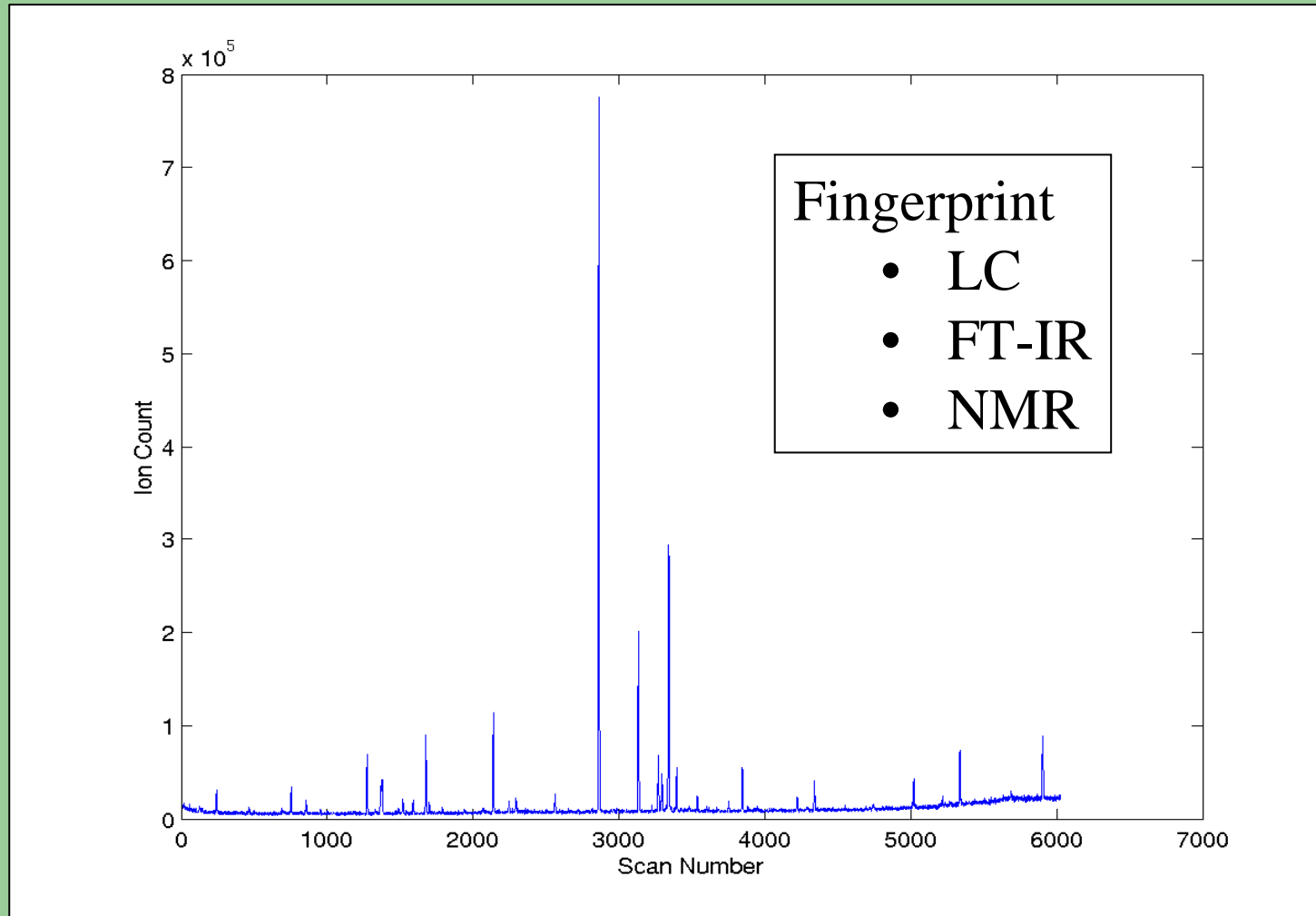
c.f. “Footprint”

What does “no value” mean?

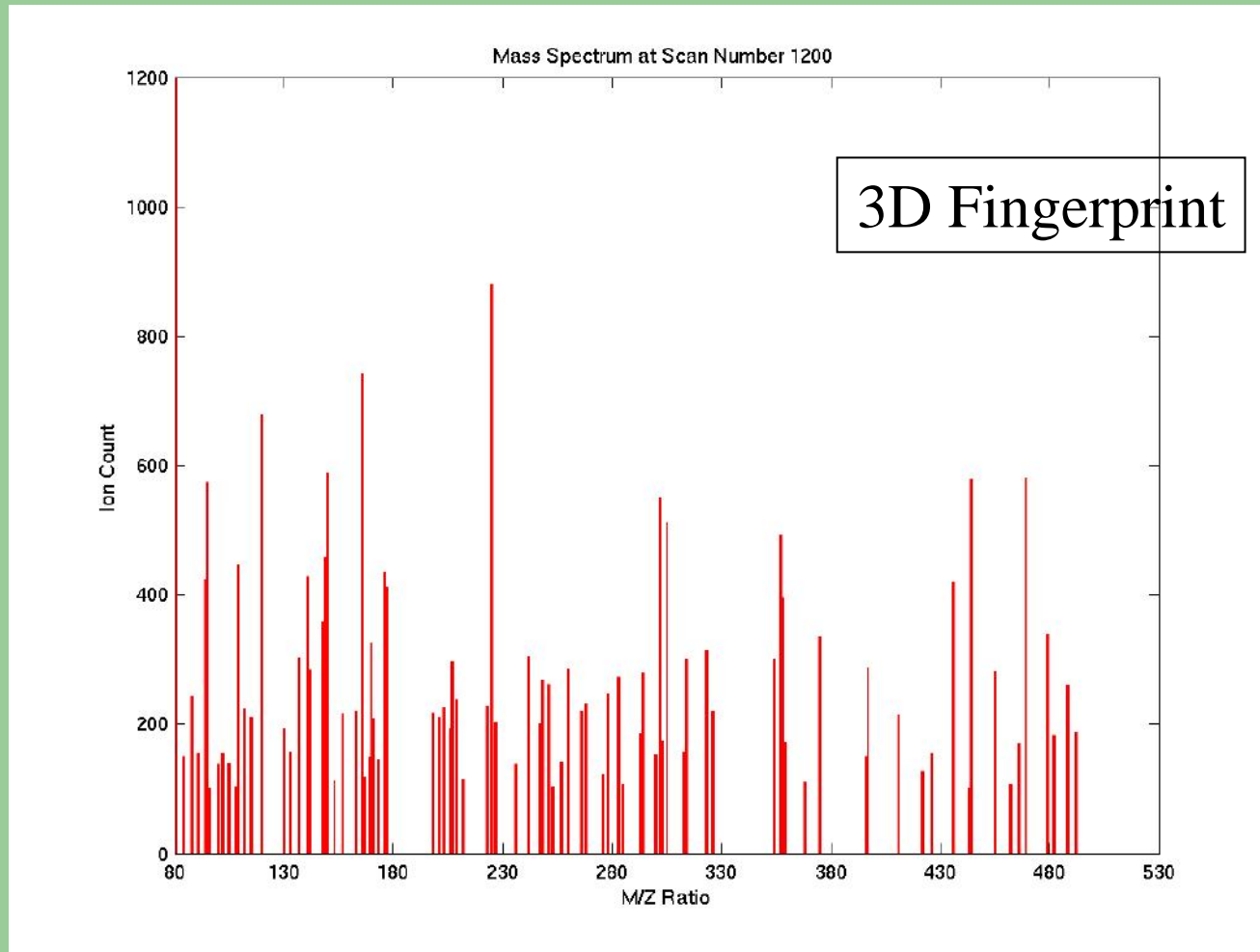
GC-MS data

an example of metabolomic results

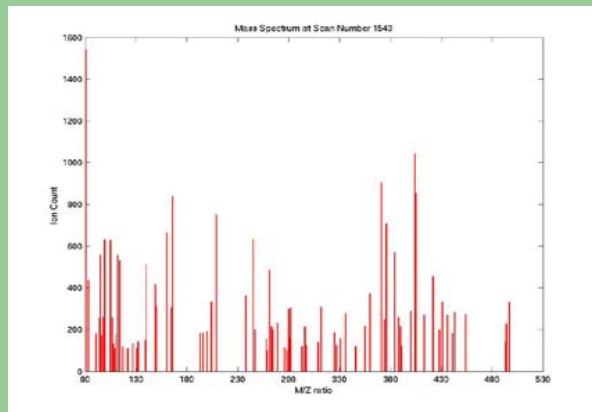
Total Ion Count



Mass Spectrum

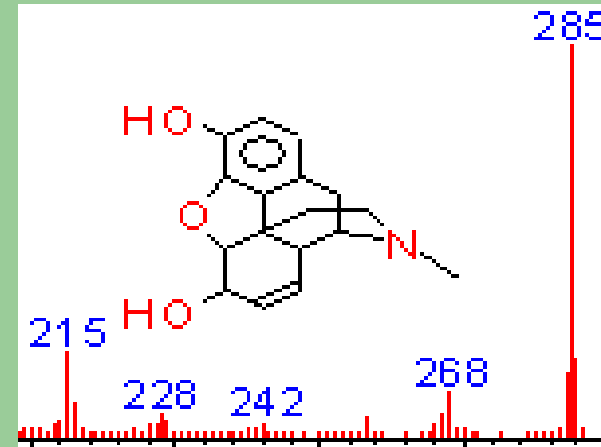


Pre-processing



Mass Spectrum

Peak Labelling



Chemical Identity

Also for

- LC-MS
- NMR
- ...

The data are noisy

- Metabolomics only just works
 - Nature (even in a lab) is highly diverse
 - “Natural variation”
 - The systems are highly dynamic
 - hard to take a “snapshot”
 - The laboratory preparation procedures are complex
 - Operator variance
 - The analytical equipment is complex
- Controlling variance
 - Experimental design
 - Statistical control
 - Rigorous standards in the lab
 - process quality control
- Require traceability
- Quality control and quality assurance necessary
 - c.f. industrial automation

Common experience of data quality

- Many “results” are artifacts
 - Lab technician
 - Time of day of harvesting
 - Day of the week
 - Phase of the moon
 - ...
- Metabolomics typically “high throughput”
 - many workers
 - many locations & machines
 - environmental change
 - machine drift
 - ...
- “meta-data” are crucial

Multivariate data analysis and data mining

- Principle Components Analysis
- Discriminant Function Analysis
- Partial Least Squares
- Cluster Analysis
- Minimum Spanning trees
- Random Forrest
- Neural networks
- Genetic algorithms
- Genetic Programming
- Models for:
 - Discrimination
 - Identification
 - Detecting differences
- Distinguishing the signal

Metabolomics and principled computer-based data handling

Why it must happen

Practical Reasons

1. Laboratory management
2. Automation
3. Multi-site research

Scientific Advancement Reasons

4. Re-analysis at a later date
5. Re-use of data in new “experiments”

Legal and regulatory reasons

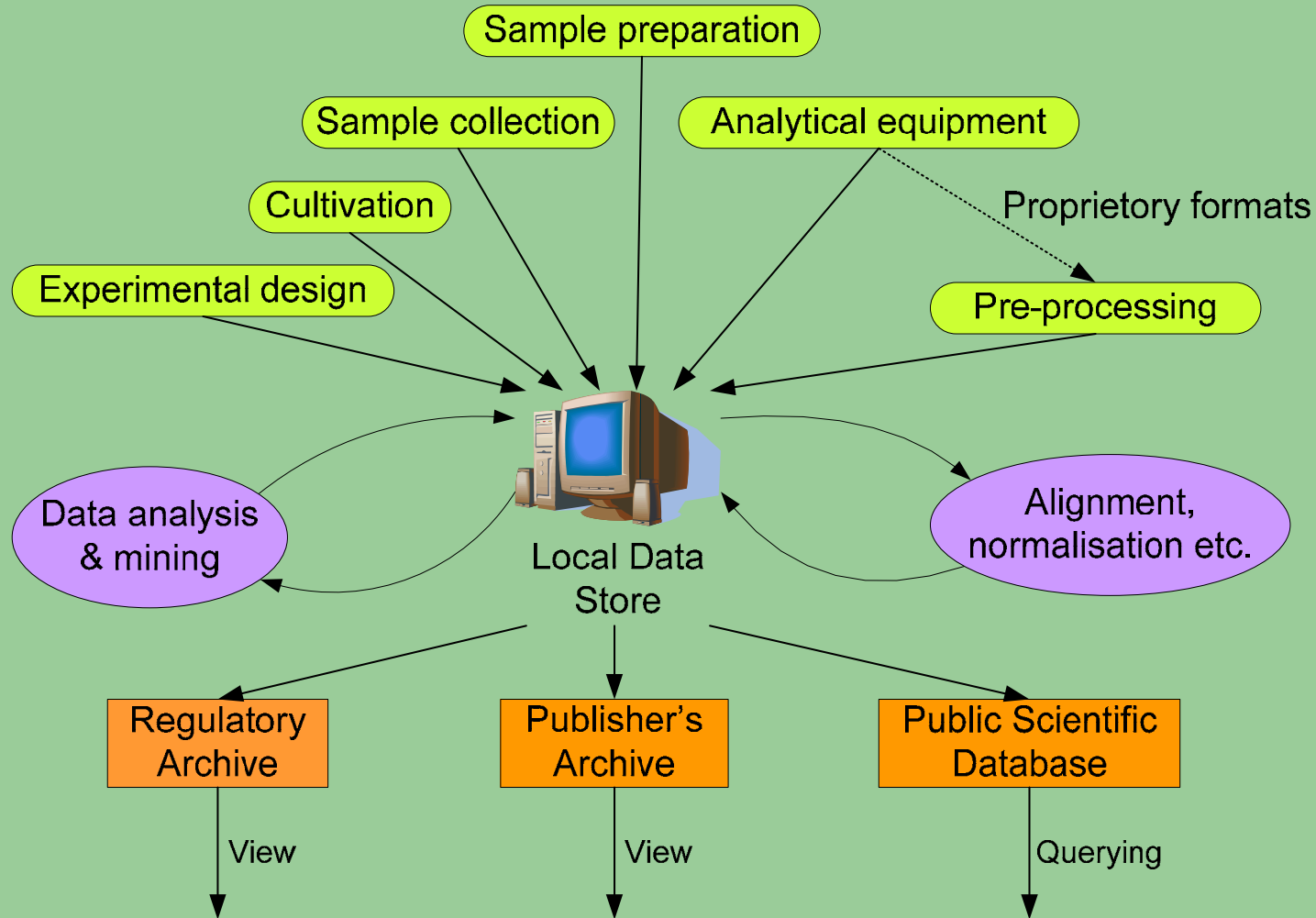
6. Food, Drugs Environmental

National Research Council (US)

- “An author's obligation is not only to release data and materials to enable others to verify or replicate published findings (as journals already implicitly or explicitly require) but also to provide them in a form on which other scientists can build with further research.”
- “If publicly accessible repositories for data have been agreed on by a community of researchers and are in general use, the relevant data should be deposited in one of these repositories by the time of publication.”
- “... these repositories help define consistent policies of data format and content”
- Specialist communities must develop technical standards

Committee on Responsibilities of Authorship in the Biological Sciences of the National Research Council (US, 2003)

Metabolomics data handling



Standards for which data?

- Mendes (2002) looks at it in terms of storage
- Databases for:
 - ... storing metabolite profiles, raw data and metadata (= LIMS)
 - ... storing metabolite profiles for a single species (portal)
 - ... combining metabolite profiles for many species and conditions (c.f. ArrayExpress)
 - (he perhaps missed formally lodged evidence – regulatory/publication)
 - ... cataloguing all known metabolites in each species
 - ... containing reference biochemical information.

ArMet work at Aberystwyth

ArMet

nature
biotechnology

w.nature.com/naturebiotechnology

A proposed framework for the description of plant metabolomics experiments and their results

Helen Jenkins¹, Nigel Hardy¹, Manfred Beckmann², John Draper², Aileen R Smith², Janet Taylor^{1,21}, Oliver Fiehn³, Royston Goodacre⁴, Raoul J Bino^{5,6}, Robert Hall⁵, Joachim Kopka³, Geoffrey A Lane⁷, B Markus Lange⁸, Jang R Liu⁹, Pedro Mendes¹⁰, Basil J Nikolau¹¹, Stephen G Oliver¹², Norman W Paton¹³, Sue Rhee¹⁴, Ute Roessner-Tunali¹⁵, Kazuki Saito¹⁶, Jørn Smedsgaard¹⁷, Lloyd W Sumner¹⁸, Trevor Wang¹⁹, Sean Walsh¹⁹, Eve Syrkin Wurtele²⁰ & Douglas B Kell⁴

(Jenkins, Hardy et al. 2004)

Background to ArMet

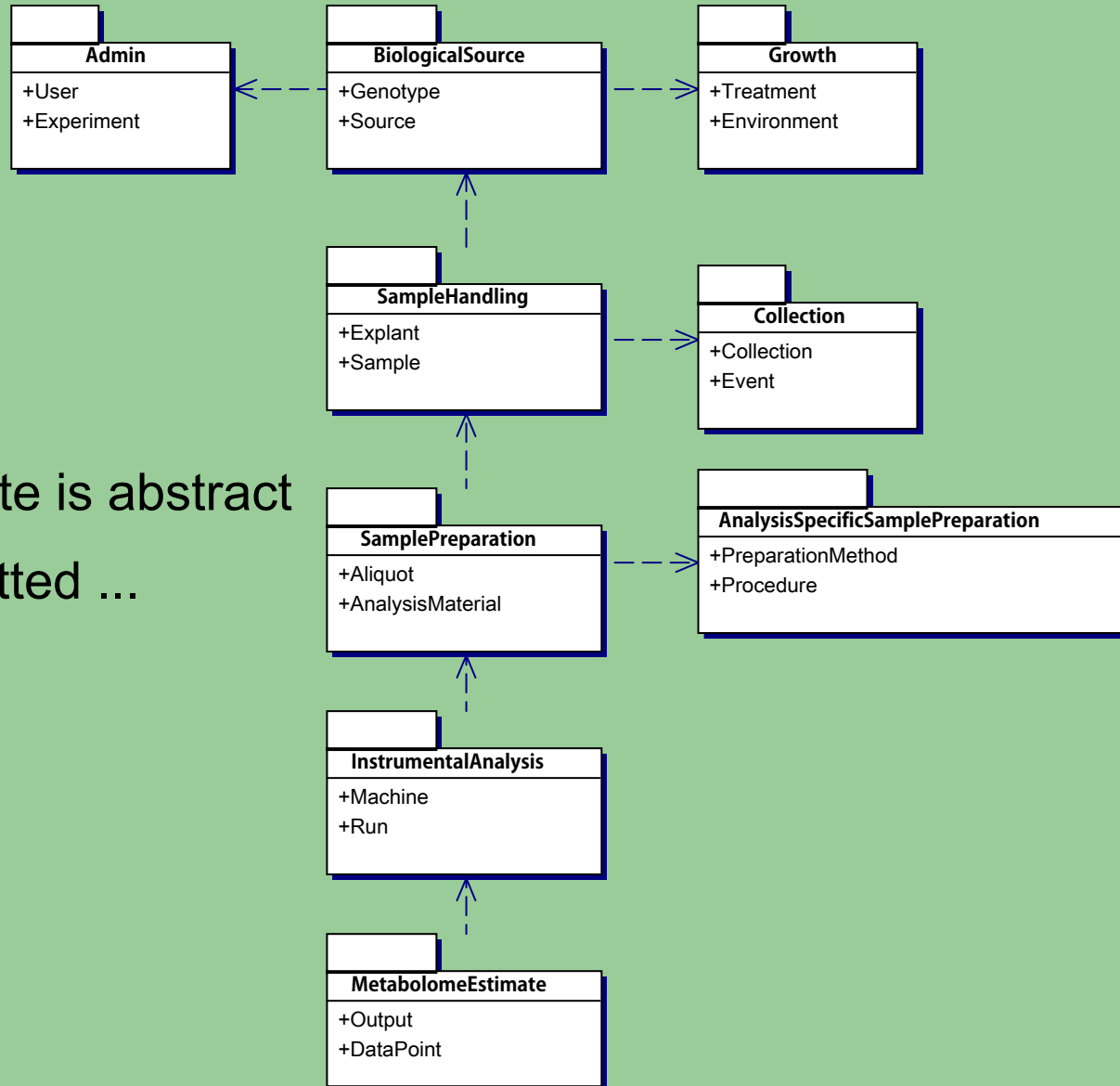
- United Kingdom Food Standards Agency
 - Department of Computer Science, Aberystwyth
 - Institute of Biological Sciences, Aberystwyth
 - John DRAPER
 - School of Chemistry, University of Manchester
 - Douglas KELL
 - Max Planck Institute for Molecular Plant Physiology, Golm, Germany
 - Oliver FIEHN
- Investigation of metabolomics as a platform technology for food safety assesment
- Published with community support

ArMet was ...

- The first published description of required data for metabolomics
- Biased? (Incomplete systems analysis)
 - Plants (GM potatoes)
 - Field and greenhouse
 - GC-MS
- A UML model
- Implementations
 - XML (transmission)
 - SQL (Storage and retrieval)
 - (Oracle)
- Extensible

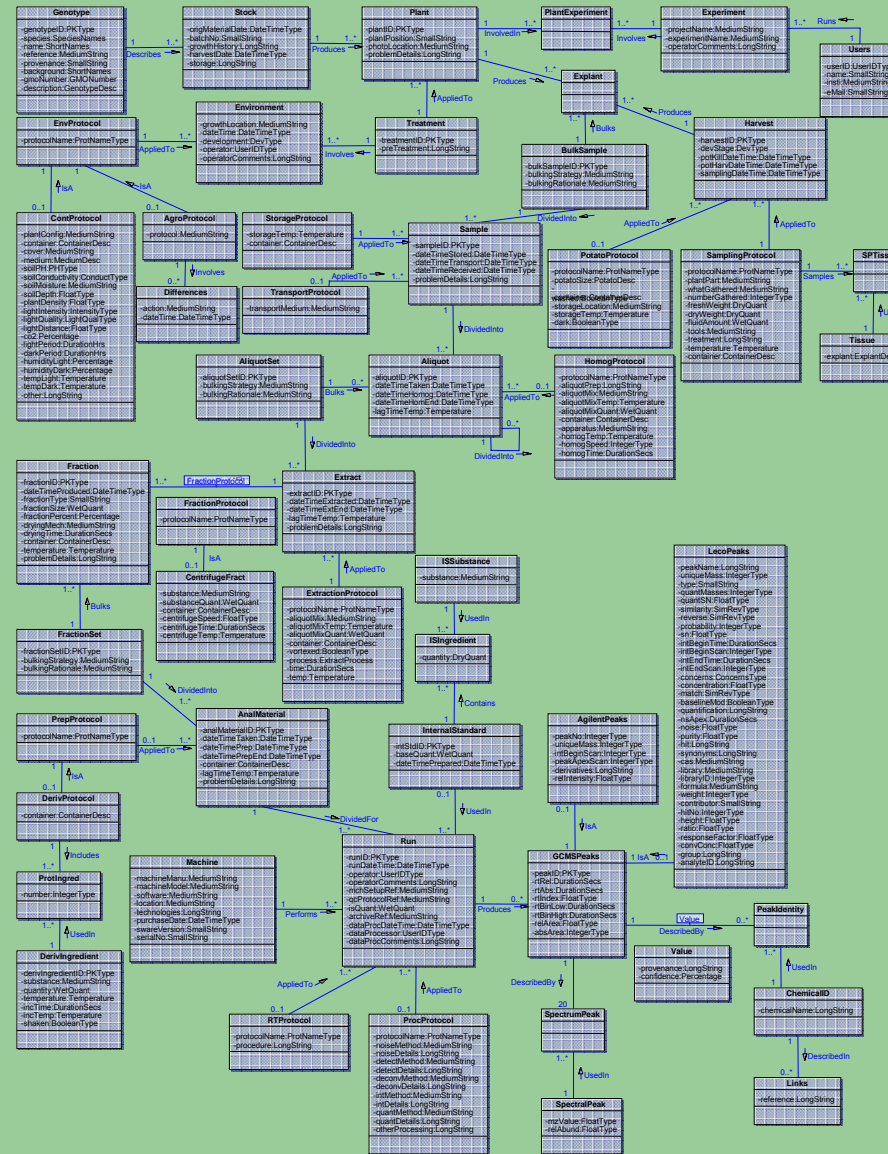
ArMet

- Metabolome Estimate is abstract
 - fingerprint, targeted ...
- None are final



Full UML

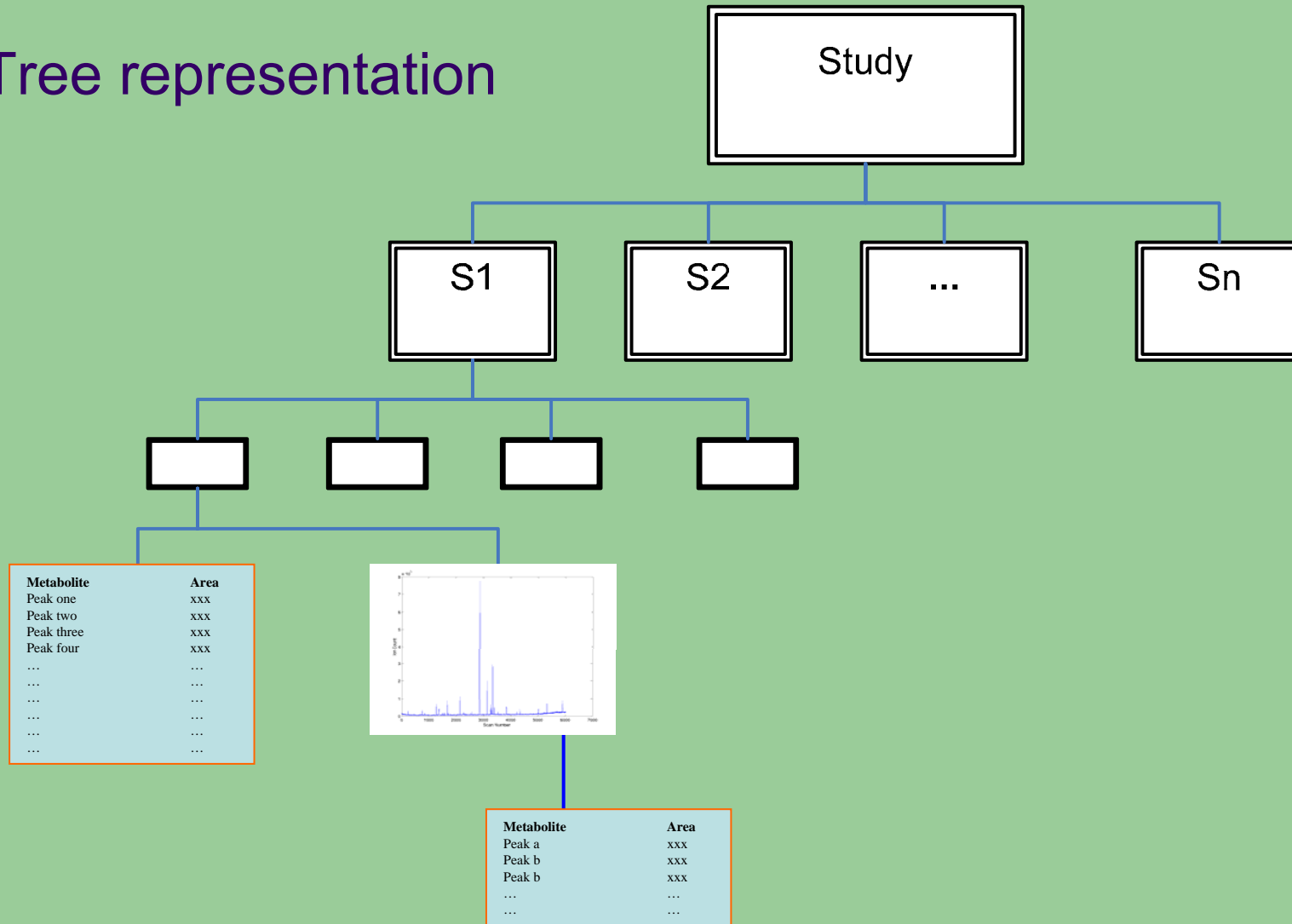
FSA G02006



ArMet is ...

- The basis of subsequent projects
 - HiMet
 - UK, reserach project
 - National Centre for Plant and Microbial Metabolomics
 - UK programme
 - META-PHOR
 - European Union project FOOD-CT-2006-036220
- Now in version 2
- Contributing to the MSI (q.v.)

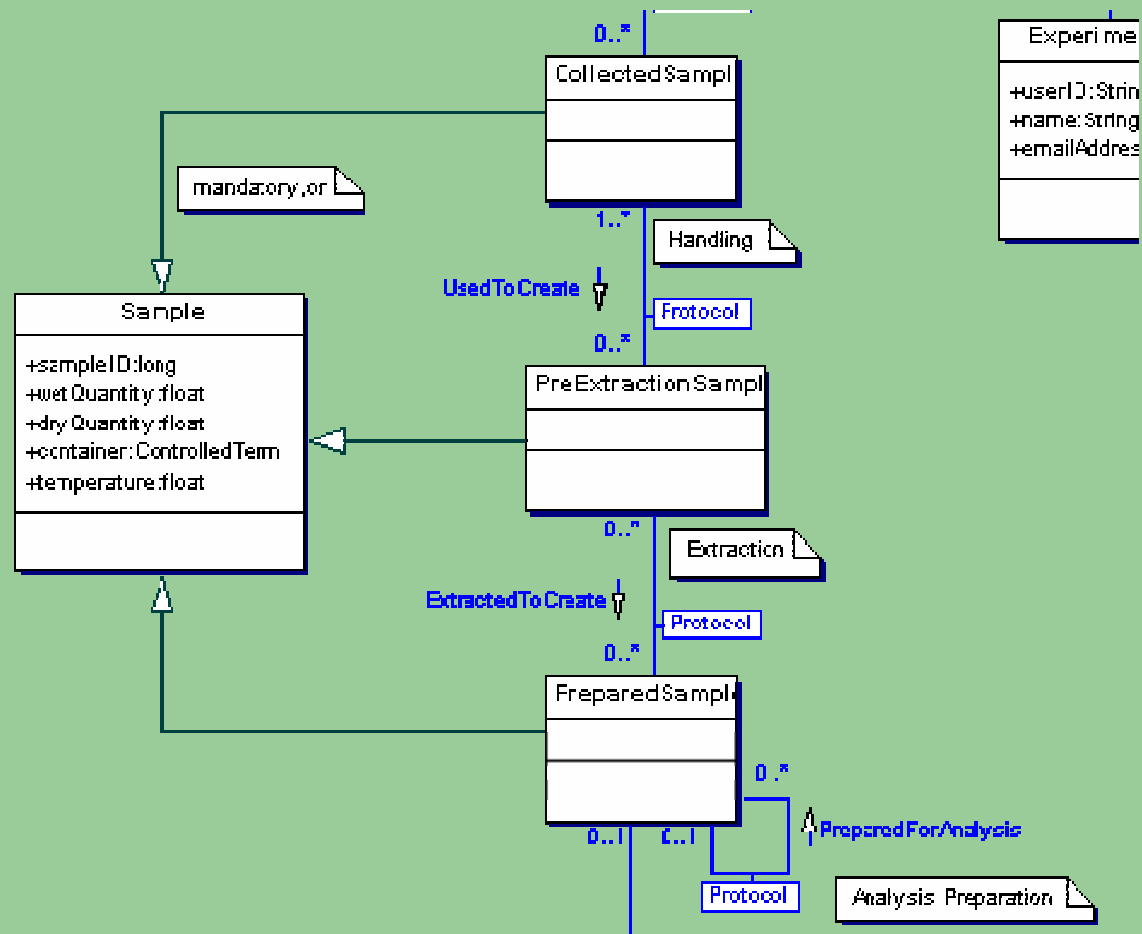
Tree representation



Universal relation

| Strain | Individual | Treatment | | Metabolite 1 | Metabolite 2 | Metabolite 3 | Metabolite 4 | Metabolite 5 | | Metabolite n |
|--------|------------|-----------|-----|--------------|--------------|--------------|--------------|--------------|-----|--------------|
| S1 | I1 | Hot | ... | 0.660 | 0.607 | 0.451 | 0.790 | 0.796 | ... | 0.927 |
| S1 | I2 | Cold | | 0.709 | 0.495 | 0.271 | 0.248 | 0.545 | | 0.953 |
| S1 | I3 | Hot | | 0.053 | 0.997 | 0.782 | 0.971 | 0.703 | | 0.453 |
| S1 | I4 | Cold | | 0.042 | 0.599 | 0.098 | 0.917 | 0.480 | | 0.990 |
| S2 | I5 | Hot | | 0.276 | 0.605 | 0.882 | 0.759 | 0.657 | | 0.104 |
| S2 | I6 | Cold | | 0.797 | 0.759 | 0.597 | 0.524 | 0.589 | | 0.835 |
| | | | | | ... | | | | | |
| Sn | In | Cold | | 0.379 | 0.522 | 0.520 | 0.552 | 0.265 | | 0.674 |

More complex? – ArMet 2



MeMo

- Spasić et al 2006
- ArMet based
- XML extensions
 - Protocols

Tool development based on ArMet

- Real data:
 - 1 data set (growth & harvest), 2 weeks, 3 iterations, failure!
- MS Access database (Jenkins, Johnson et al. 2005)
 - Database design was available
 - Front end customised to look like users' environment and procedures
- Complete datasets uploaded
 - <2 weeks of software engineer effort
 - “disposable”; cheap per-project/protocol

Microsoft Access - [SampleHandling : Form]

File Edit View Insert Format Records Tools Window Help

Tahoma 8 B I U

Harvest Shelves

To harvest your shelves first select the treatment that is taking place on them and create the collection that describes the harvest and then select the shelves below to identify the trays involved in that collection.

Select your Treatment: View Environment Protocol Details

To create your collection identifier click here: View Harvest Protocol Details

Collection ID: 1

1 2 3 4 5 6 7 8 9

To view a report of the trays harvested in a particular collection click here:

To view a report of the trays involved in a treatment, annotated with their collection IDs click here:

Form View

The screenshot shows a Microsoft Access application with two windows. The main window is titled 'SelectTrays : Form' and contains the following elements:

- Title Bar:** Microsoft Access - [SampleHandling : Form]
- Form Header:** Select Trays
- Text:** Select the half trays as you harvest them from shelf: 1
(The captions on the trays denote: number of plants harvested/number of plants unharvested)
- Table:**

| | | | | | |
|------|------|------|------|------|------|
| 3/9 | 0/12 | 0/12 | 4/8 | 0/12 | 0/12 |
| 0/12 | 0/12 | 7/5 | 0/12 | 0/12 | 2/10 |
| 0/12 | 0/12 | 5/7 | | | |
- Dialog Box:** HarvestNumber : Form. It contains the text 'How many plants have you taken from this tray today:' followed by a text box containing the number '6' and an 'OK' button.
- Notes Section:** A text box labeled 'Notes:' with instructions: 'Scroll through existing notes for this collection using the arrows or enter additional notes when the box is full'. Below the text box are left and right arrow buttons and a 'More Notes' button.
- Buttons:** 'Report' and 'Treatment Report' buttons are visible at the bottom of the form.
- Status Bar:** Form View

Development of new standards

Other initiatives - MIAMET

- Bino et al. 2004
- checklist of “minimum information” (reporting requirements)
- members of the *plant biology* community
 - significant emphasis of agronomic applications
 - significant emphasis on chromatography and mass spectrometry techniques
- “experiment” centred list
- some limited hints at data formatting, suggesting the use of NetCDF for raw MS data and JCAMP for NMR data

Other initiatives - SMRS

- Lindon et al 2005, Lindon 2005
- Bias to pre-clinical toxicology trials – Metabonomics
- Definitely a reporting standard proposal
- Noted need to cater for a variety of reporting circumstances:
 - “*the need for standards to facilitate communication between different fields of activity and to fulfil the needs of journal editors and regulatory agencies...*”
- noting that there are
 - “*fundamental differences between both the design and objectives of efforts focused on regulatory submission and those efforts focused on basic research*”
- Much less about domestic LIMS
- Reporting data analysis – models

Metabolomics Standards Initiative (MSI)

- NIH funded workshop, August 2005, Washington DC
 - <http://www.metabolomicssociety.org/nih.html>
- Decided to progress under the Metabolomics Society
 - <http://www.metabolomicssociety.org/>
- Metabolomics Standards Initiative
 - Operates via sourceforge
 - <http://msi-workgroups.sourceforge.net/>

Developing a metabolomics standard

- Reporting Standard (semantics)
- Data Modelling (logical)
 - Careful typing
- Ontologies
 - choose and/or define
- Data Format (syntax)

Reporting Standards

- Professional standards (for the biologist)
- Requirements specifications (for the software engineer)
- Transcriptomics:
 - MIAME (Brazma, Hingamp et al. 2001)
- Proteomics
 - MIAPE (Taylor, Hermjakob et al. 2006)
- Biochemical models
 - MIRIAM (Novere, Finney et al. 2005)
- MIBBI
 - <http://mibbi.sourceforge.net/>

MSI Reporting Standards

| Working Group | | Working Group chairs | |
|-----------------------------|----------------------------|--|--|
| Biological context metadata | | Don Robertson | |
| | mammalian/ <i>in vivo</i> | Jules Griffin | |
| | microbial/ <i>in vitro</i> | Mariët van der Werf | |
| | Plant biology | Basil Nikolau | |
| | Environmental | Dawn Field | |
| Chemical analysis | | Lloyd Sumner, Teresa Fan | |
| Data processing | | Roy Goodacre | |
| Ontology | | Susanna-Assunta Sansone, Ricardo Pietrobon | |
| Exchange format | | Nigel Hardy, Chris Taylor | |

MSI Reporting standards published September 2007

1. Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments
2. Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments
3. Minimum reporting standards for plant biology context information in metabolomic studies
4. Standard reporting requirements for biological samples in metabolomics experiments: environmental context
5. Proposed minimum reporting standards for chemical analysis
6. Proposed minimum reporting standards for data analysis in metabolomics

Metabolomics 3(3)

Data models

- Transcriptomics
 - MAGE - (Spellman, Miller et al. 2002)
 - FuGE
- Proteomics
 - PSI – FuGE - (Jones, Miller et al. 2007)
- Metabolomics
 - ArMet
 - MSI
 - FuGE?

FuGE

- <http://fuge.sourceforge.net/>
- Functional Genomics Experiment model

“FuGE is a model of the shared components in different functional genomics domains”

Nature Biotechnology **25**(10):1127-1133

FuGE cont.

- FuGE is a UML model and a generated XML Schema
- FuGE facilitates the development of data standards in functional genomics in two ways:
 1. FuGE provides a model of common components in functional genomics investigations, such as materials, data, protocols, equipment and software. These models can be extended to develop modular data formats with consistent structure.
 2. FuGE provides a framework for capturing complete laboratory workflows, enabling the integration of pre-existing data formats. In this context, FuGE allows the capture of additional metadata that gives formats a context within the complete workflow.

Data values

Types

- models provide structure
- require data types
- many attributes (fields) are numeric
 - constraint on range and precision is important
- many attributes are not numeric
 - descriptive
 - species, variety, organ, cell type
 - equipment and procedures
 - metabolites
- controlled vocabularies or ontologies

What do words mean?

- natural language words are no good enough
 - may have more than one meaning
 - two words may mean the same thing
- **Controlled vocabularies**
 - to biologist:
 - list of terms with definitions
 - to the software engineer
 - enumerated types

Ontologies

- entities
- meaning
- relationships
 - has-a
 - is-a
- handles synonyms
- permit more complex comparisons
- *student is-a human*
- *woman is-a human*
- data items “woman” and “student” can have a more subtle relationship than \neq
- may biological and biomedical ontologies are under construction

Ontology example

- Which part of a plant was analysed?
- The Plant Ontology (PO) (Bruskiewich, Coe et al. 2002)
 - plant growth and development stages and plant structure
- An *inflorescence* is defined (as part of the *shoot*)
 - the potential parts of an *inflorescence* are described
 - includes *flower*
 - *ear* and *tassel* as specialist terms
 - for *Zea mays* (maize, indian corn)

Chemical Identity

Identification of chemical species

- Metabolites have “common” names
 - Vitamin C, Ascorbic acid
 - Clearly wrong
- Fully systematic IUPAC name
 - (+ Retained names)
 - A structure has no unique IUPAC name
 - May not be recognised
- Online catalogs and databases
 - Controversial
 - Contain mistakes
 - Lack specificity



Chemical species - Online Database IDs and URIs

<http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI%3A17634>

http://www.genome.ad.jp/dbget-bin/www_bget?cpd:C00031

<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=3333>

<http://webbook.nist.gov/cgi/cbook.cgi?ID=50-99-7>

- MIRIAM proposal (Novère et al 2005) provides an excellent example of the use of these

Chemical species - InChI

- From **IUPAC** - <http://www.iupac.org/inchi/>
 - *Objective:*
The objective of the IUPAC Chemical Identifier Project is to establish a unique label, the IUPAC Chemical Identifier, which would be a non-proprietary identifier for chemical substances that could be used in printed and electronic data sources thus enabling easier linking of diverse data compilations.
- fundamentally an algorithm
- InChI=1/C6H12O6/c7-1-3(9)5(11)6(12)4(10)2-8/h1,3-6,8-12H,2H2/t3-,4+,5+,6+/m0/s1
- Isomers and isotopes handled
 - or explicitly not specified

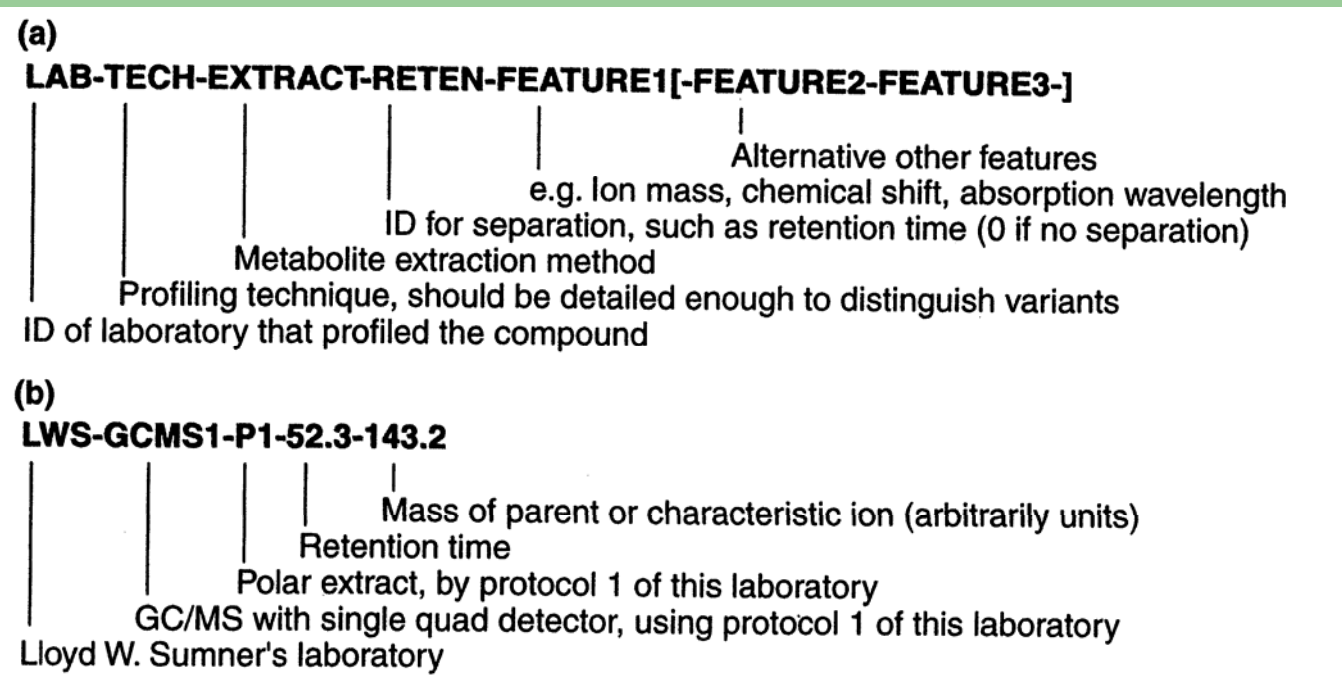
Chemical species: “Unknowns”

- Many peaks are not identified
 - Repeated found
 - “Known unknowns”
 - Significant
- “Labelled” locally by the lab
- Labelled by other labs
- They are characterised by signals from the analytical devices

- GC-MS, LC-MS ...

(Bino, Hall et al. 2004)

- Proposal incorporating:
 - Lab scope
 - Technology basis



ArMet GC-MS mechanism

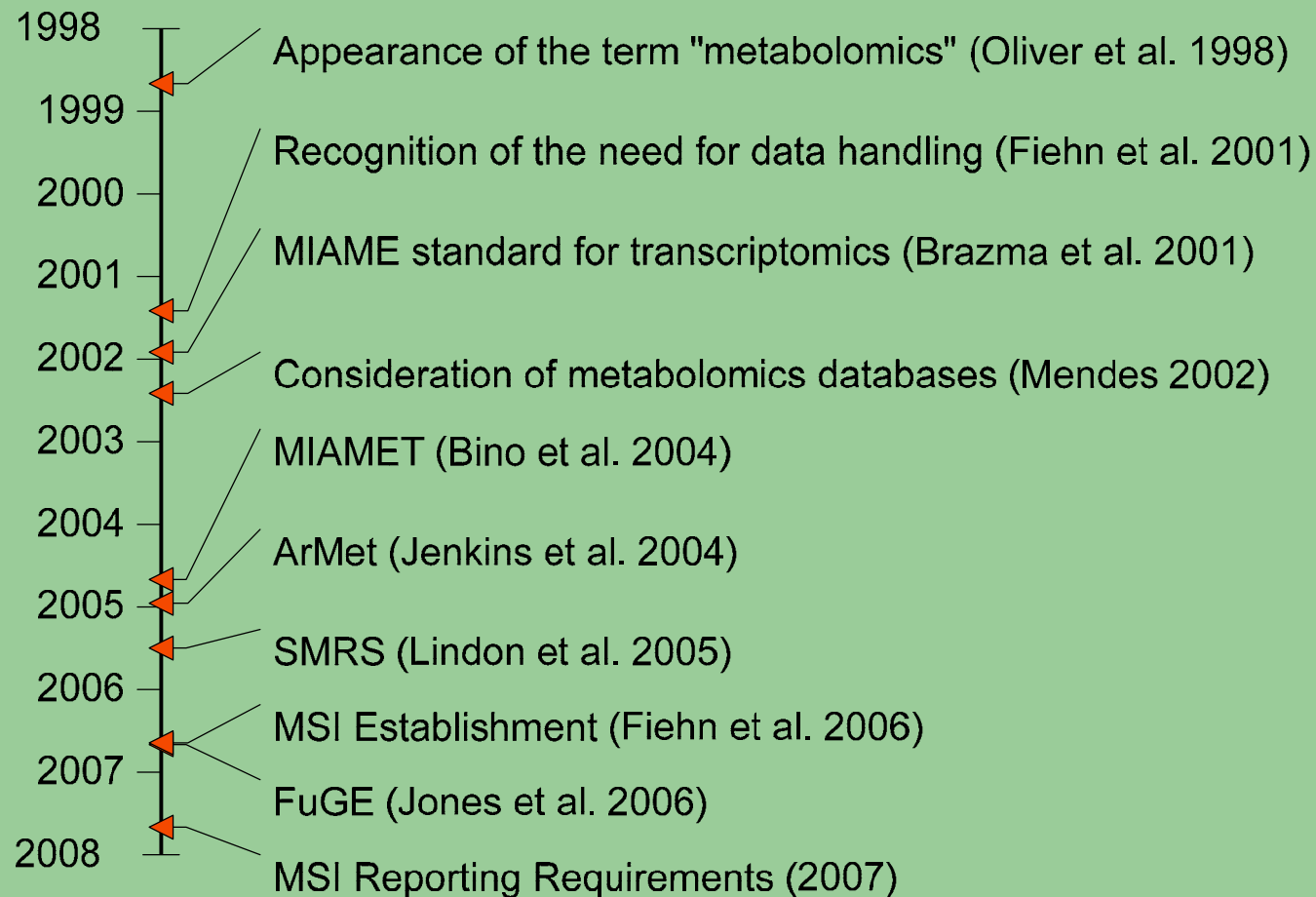
- Specific
 - Machine settings
 - 20 most abundant m/z values and intensities

URIs

- A subdivided namespace
 - Unique IDs can be created
 - Can retrieve data using it
 - Some argue an http url is appropriate
 - Permits subsequent identification
 - Avoids the issue of specific characterisation
- LSIDs
 - <http://lsid.sourceforge.net/>
 - Have a resolution mechanism
 - Can have associated (on-line) data
 - lsid:ncbi.nlm.nih.gov:pubmed:12571434
 - lsid:ncbi.nlm.nih.gov:GenBank:T48601:2
 - The lab provides an on-line name authority

Conclusions

Metabolomics data handling timeline



We are trying to:

- Increase the cost/benefit payoff of metabolomics research by
 - specifying comprehensive data sets
 - maximising the commonly appreciated meaning of data
 - supporting data repositories
 - and making the whole process as easy as possible

Acknowledgements

- Helen Jenkins, Computer Science, Aberystwyth
- Draper group, Institute of Biological Sciences, Aberystwyth
- Fiehn group (Golm, Germany) UC Davies, USE
- Kell group, Manchester, UK
- MSI working group members

- Funding
 - UK Food Standards Agency
 - UK BBSRC
 - EU Framework VI METAPHOR: FOOD-CT-2006-036220
 - Aberystwyth University

References

- Bino, R. J., R. D. Hall, et al. (2004). "Potential of metabolomics as a functional genomics tool." Trends In Plant Science **9**(9): 418-425.
- Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data." Nature Genetics **29**(4): 365-371.
- Bruskiwich, R., E. H. Coe, et al. (2002). "The Plant Ontology (TM) Consortium and plant ontologies." Comparative And Functional Genomics **3**(2): 137-142.
- Jenkins, H., N. Hardy, et al. (2004). "A proposed framework for the description of plant metabolomics experiments and their results." Nature Biotechnology **22**(12): 1601-1606. <http://dx.doi.org/10.1038/nbt1041>
- Jenkins, H., H. Johnson, et al. (2005). "Toward Supportive Data Collection Tools for Plant Metabolomics." Plant Physiology **138**(1): 67-77. <http://www.plantphysiol.org/cgi/content/abstract/138/1/67>
- Lindon JC, et al (2005) Summary recommendations for standardization and reporting of metabolic analyses. Nature Biotechnology **23**:833-838
- Lindon, J. C. (2005). *Standardisation of Reporting Methods for Metabolic Analyses: A Draft Policy Document from the Standard Metabolic Reporting Structures (SMRS) Group*. 2005. http://www.smrsgroup.org/documents/SMRS_policy_draft_v2.3.pdf
- Mendes, P. (2002). "Emerging bioinformatics for the metabolome." Briefings in Bioinformatics **3**(2): 134-145.
- Novere, N. L., A. Finney, et al. (2005). "Minimum information requested in the annotation of biochemical models (MIRIAM)." **23**(12): 1509. <http://dx.doi.org/10.1038/nbt1156>
- Spasić et al (2006) "MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics". BMC Bioinformatics **7**:281
- Spellman, P. T., M. Miller, et al. (2002). "Design and implementation of microarray gene expression markup language (MAGE-ML)." Genome Biology **3**(9): research0046.1-0046.9.
- Taylor, C. F., H. Hermjakob, et al. (2006). "The Work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI)." OMICS: A Journal of Integrative Biology **10**(2): 145-151.
- Jones A. R., M. Miller, et al. (2007) The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nature Biotechnology **25**(10):1127-1133