

Towards a standard representation of metabolomics experiments and their results: ArMet

Nigel Hardy¹, Helen Fuell¹, Manfred Beckmann², John Draper², Oliver Fiehn³, Royston Goodacre⁴, Birgit Linkohr³, Aileen Smith².

¹University of Wales, Aberystwyth Department of Computer Science
³Max Planck Institute of Molecular Plant Physiology

²University of Wales, Aberystwyth Institute of Biological Science
⁴Department of Chemistry, University of Manchester Institute of Science and Technology

Objectives

We are developing a standard representation of the data associated with metabolomics research. Such a representation can be used to develop data-handling systems to support a range of activities including:

- Distributed collection of data across labs
- Long term storage
- Statistical analysis and data-mining of the results in the context of the experiments in which they were collected
- Re-analysis of data in new contexts

Additional expected benefits

- Uniform recording of experiments
- Logical completeness and internal consistency of data sets
- Reliable and verified exchange of data sets
- Principled comparison of data from a range of techniques
- Support for design of new experiments
- Promotion of standard operating procedures

Design Issues

Data Re-use

Metabolomics is expensive and frequently requires collaborations between different sites and labs. Once data are collected, it is important to make the fullest use of them by re-analysis in a variety of contexts and in association with other data sets.

A range of uses

A database must support the detailed QA and statistical analysis requirements of well defined experiments carried out in a single lab or well integrated community. On the other hand we believe it is important to provide the opportunity to mine data across many experiments, technologies, species, labs etc. in an environment where the extent of **comparability** can be properly assessed.

Textual data

Most of the data will be numerical but some will be textual and this leaves room for ambiguity and confusion. We have adopted **controlled vocabularies** for data of this sort. This promotes consistency and supports effective interaction with other types of database.

Noisy data

Growth conditions, harvesting and quenching techniques, sample preparation and analytical machine operation are all sources of variation which must be controlled. The ways in which they are controlled will affect dataset comparability.

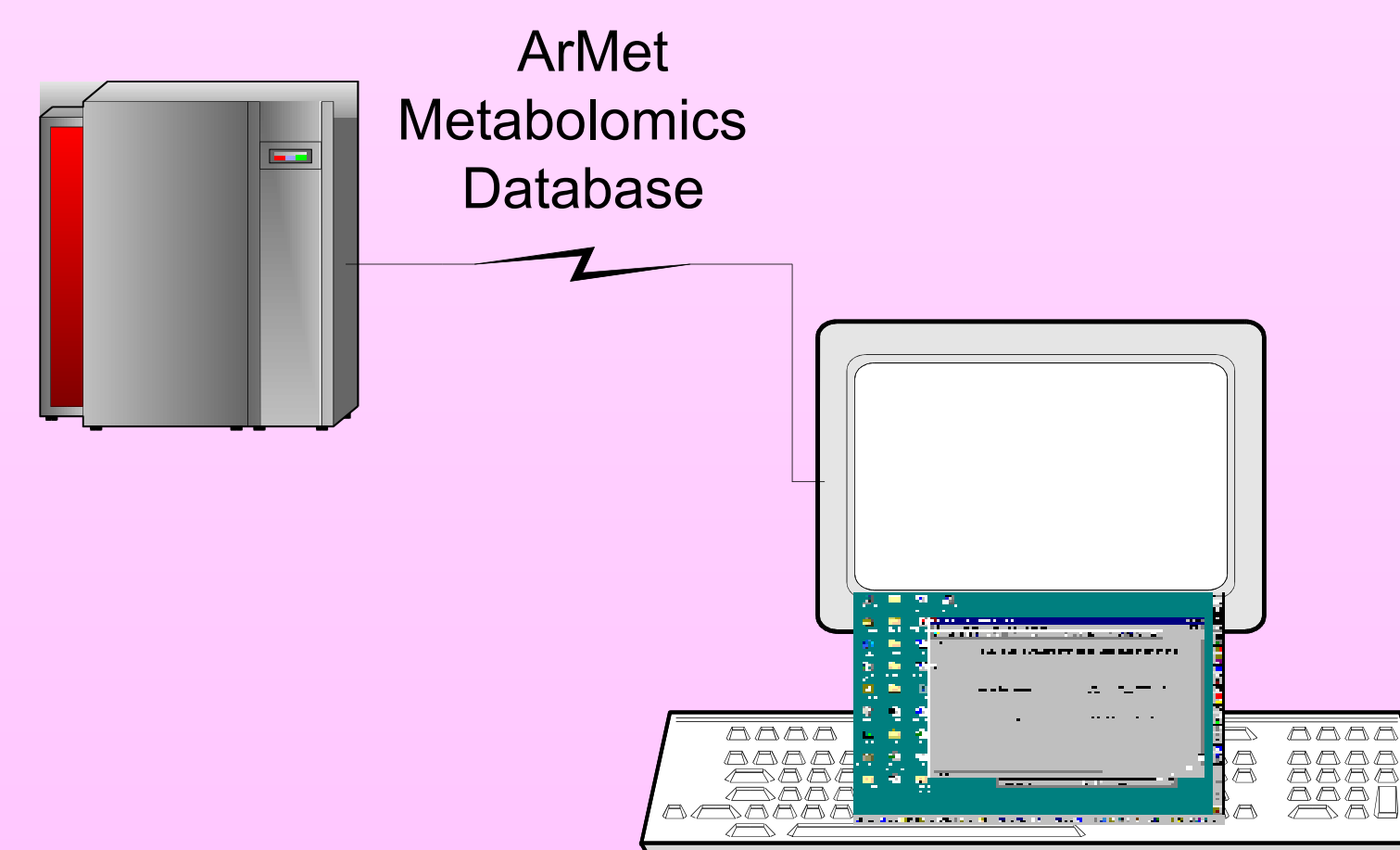
A range of analytical techniques

Various GC-MS methods, LC-MS, FT-IR, NMR etc. are used, each with different sample preparation requirements, operating parameters and producing results of different types which estimate the components of a metabolome with different biases. One way of characterising this range is under the headings of fingerprinting, targeted analysis, metabolite profiling and metabolomics¹.

Some conclusions on these issues

- Metadata - data about the data, providing context - represents a large part of the system. This allows dataset comparability to be assessed.
- Controlled vocabularies are used to ensure internal comparability of data sets and correct inter-connection with other databases.
- The full system will be complex.
- Different users will wish to have very different styles of metabolomics database.

¹ Oliver Fiehn, Metabolomics – the link between genotypes and phenotypes. Plant Molecular Biology 48:155-171, 2002.



The Architecture (ArMet)

A set of nine **packages** that describe the different phases of the metabolomics experimental timeline were identified using the detailed and specific design of the prototype:

- **Admin:** Experiment management data
- **Biological source:** Genotype, provenance and ID information for source material
- **Growth:** Environment descriptions and protocols for developing source material
- **Collection:** Sample gathering protocols (harvesting)
- **Sample handling:** Sample bulking/division description and storage protocols
- **Sample preparation:** Protocols for preparing samples for analysis
- **Analysis specific sample preparation:** Instrument specific preparation protocols
- **Instrumental analysis:** Description of the analysis of the metabolite content of samples
- **Metabolome Estimate:** Results sets with associated data processing protocols

Different projects/laboratories will have differing requirements for detailed meta-data to describe their experiments. To support this the project specific detail in the prototype design was abstracted to produce a core set of data items that provide a generic description for each package that is applicable to a wider range of experiments. **Sub-packages** may then be created that extend this core data to customise it for particular experimental environments. The core data, therefore, provides the lowest common denominator for the comparison of datasets, whilst experiments that are described using the same sub-packages may be compared in more detail.

This package hierarchy enables **incremental development** of ArMet. Sub-packages may be designed for new analytical technologies and experimental techniques as they develop, without affecting either the core or any existing sub-packages that are already in use. Development of ArMet involves the development of sub-packages and the management of sub-package specifications.

The prototype represents the first set of sub-packages, customised for our current project.

The Prototype

We carried out a **requirements analysis**, mainly involving staff on a particular plant metabolomics project. The set of techniques covered was therefore biased, though provision for alternative approaches was given consideration. Specifically the prototype design handles i) Arabidopsis thaliana grown in the greenhouse and phytotrons and field-grown potatoes ii) harvest and storage of those materials, iii) preparation for GC-MS analysis iv) GC-MS analysis v) preparation and storage of peak lists. (See the poster by Fuell et al.)

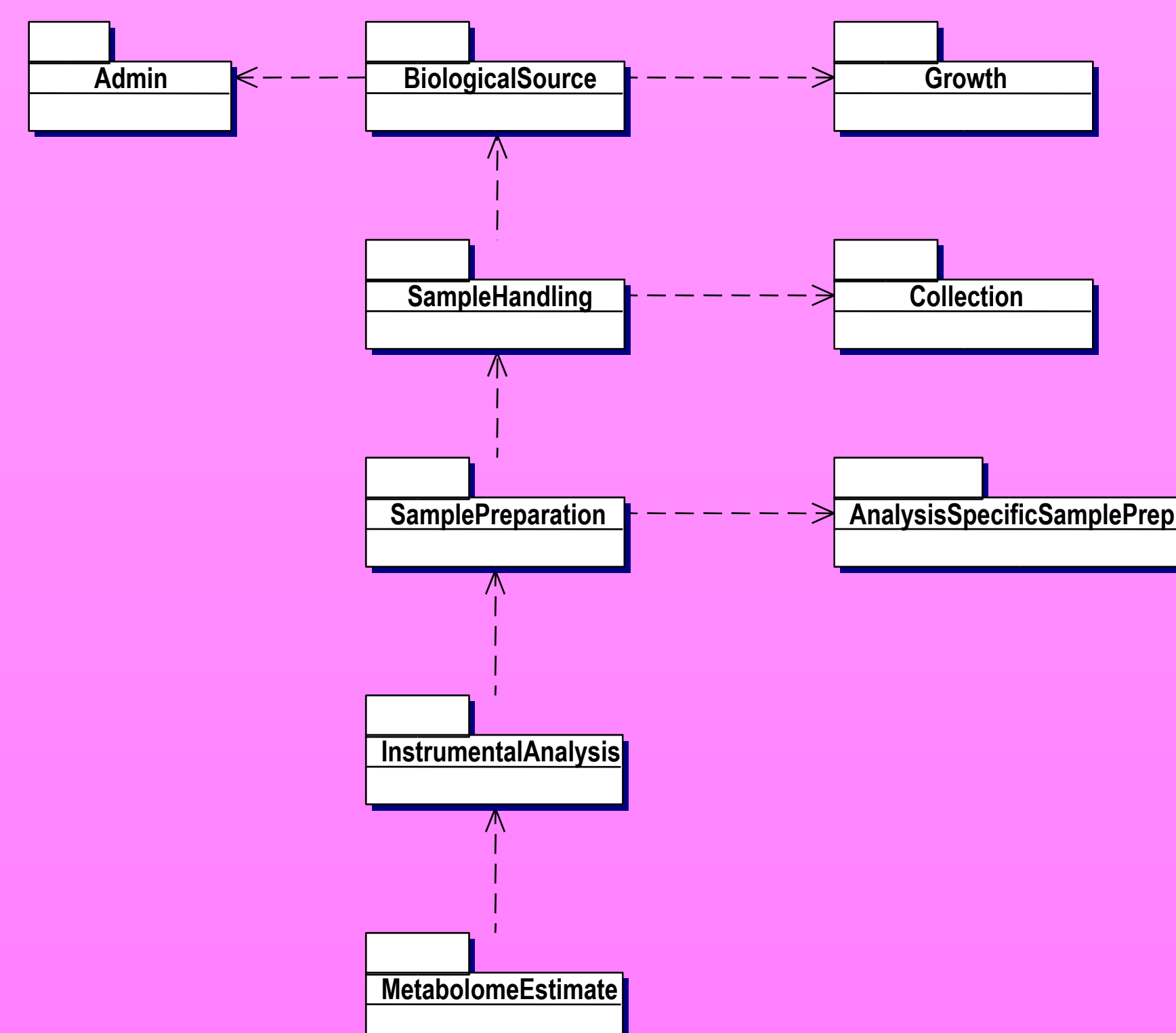
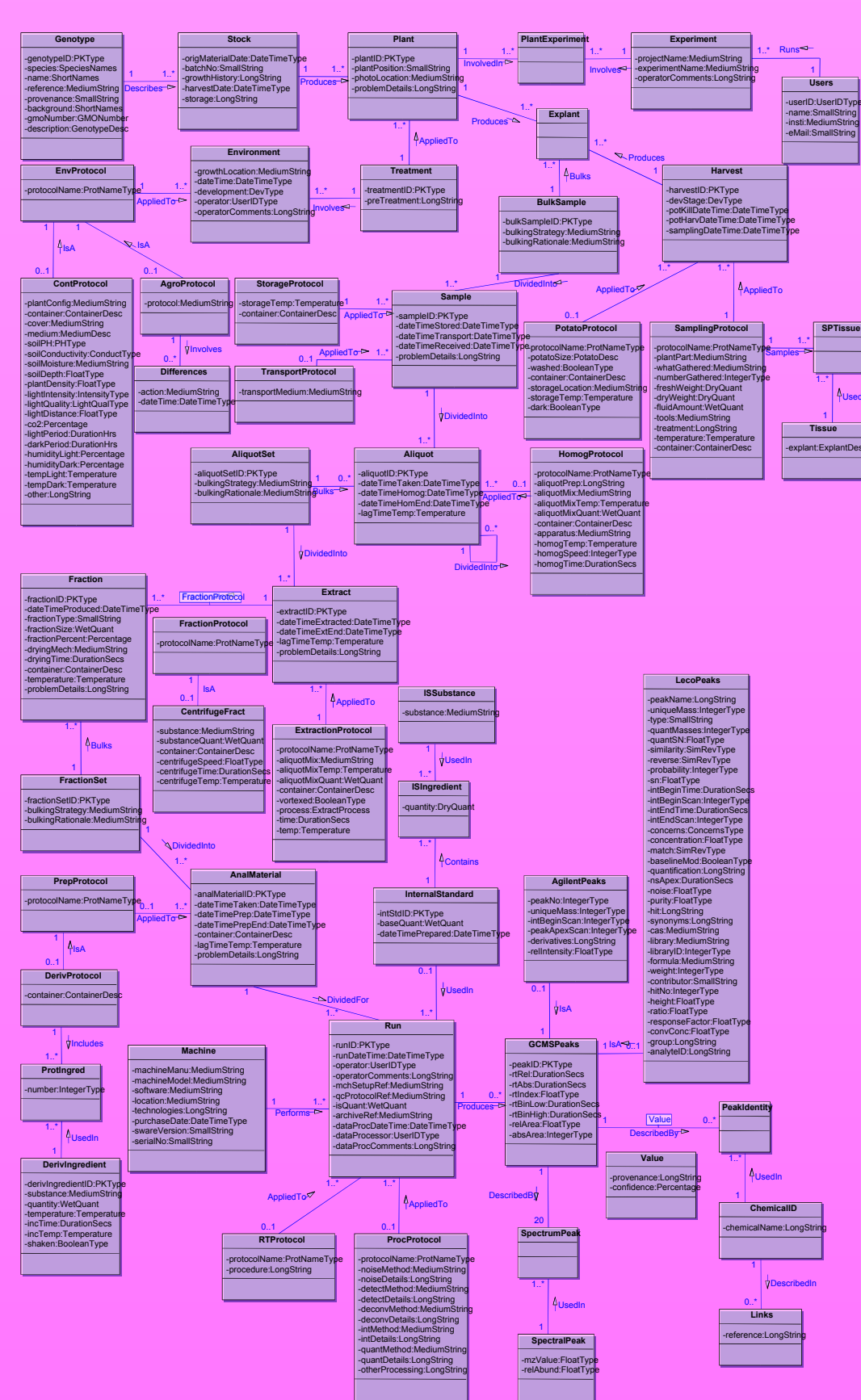
The results of this analysis were represented in the Universal Modelling Language (UML)¹, a widely used modern standard which incorporates many computing design techniques. A UML model permits and encourages unambiguous specifications, supports dissemination of these specifications and makes it much easier to produce multiple consistent implementations and to track change and enhancement. As with MAGE² and PEDRo³ the UML model is the definition of prototype. Databases, transmission formats (including XML) and front ends for data collection and retrieval are all possible implementations of it.

An implementation of the prototype was built using Oracle with a Microsoft Office front end. These choices were based on management considerations and no aspect of the prototype is dependent on proprietary features of either system.

¹UML <http://www.omg.org/uml/>

²Spellman et al. "Design and implementation of microarray gene expression markup language (MAGE-ML)". Genome Biology 3(9):

³Taylor et al. "A systematic approach to modeling, capturing and disseminating proteomics experimental data". Nat. Biotech. 21:247-254. March 2003



An Architecture for Metabolomics

ArMet