FuGE: A framework for developing standards for functional genomics

Andrew Jones

School of Computer Science,

University of Manchester

MGED Society

Metabomeeting 2.0 10/01/2006

Proteomics Standards Initiative

Overview

- Challenge of building data standards
- Introduction to FuGE
- Current status
- Experience with formats developed using FuGE
  - Example: Sample Processing markup language

Data standards for functional genomics

Major challenge developing standards:

- Technology still evolving

- Heterogeneous data formats from instruments

- "Important" info about starting sample is almost unlimited

- Lots of metadata required to validate results

BUT:

- Most of these problems are shared by microarrays, proteomics, metabolomics etc.
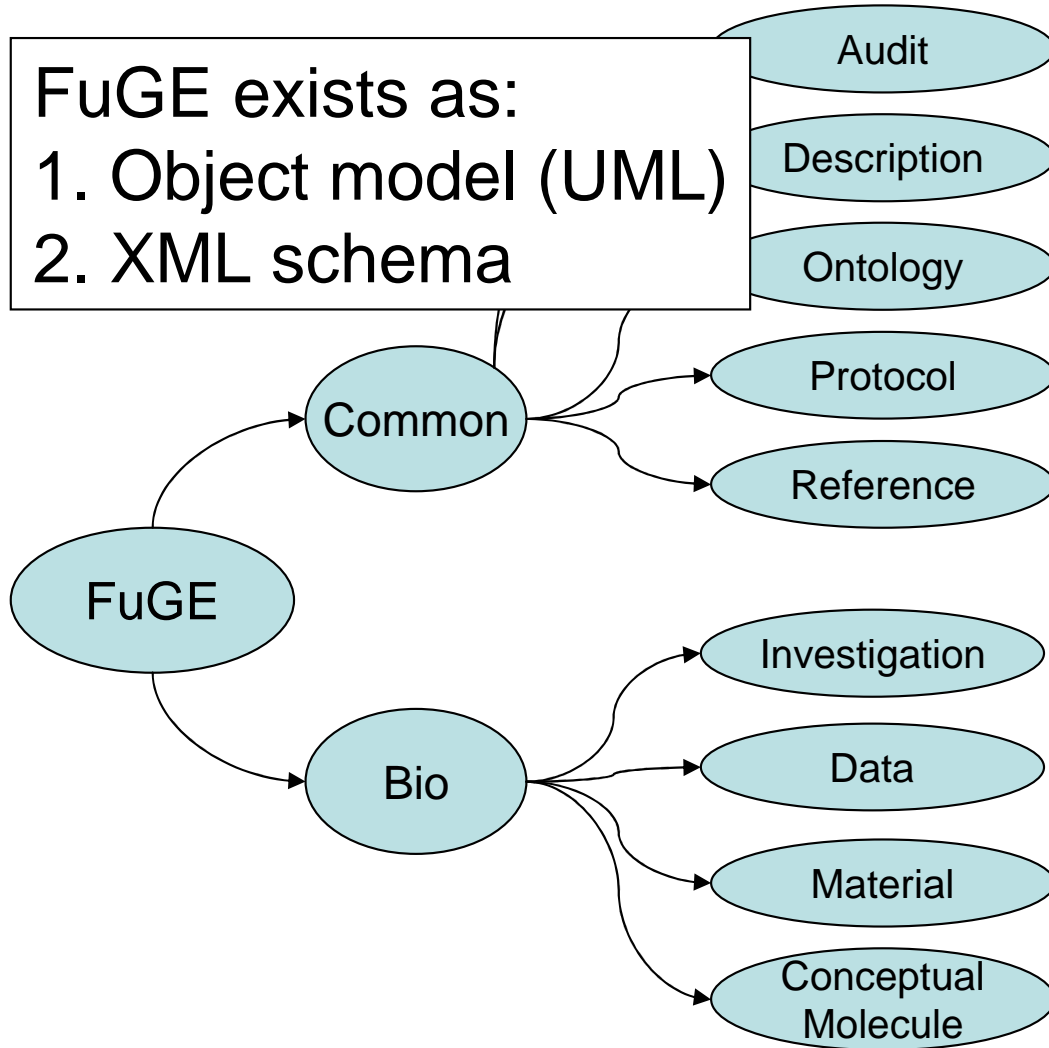
FuGE

FuGE = Functional Genomics Experiment (Object Model / Markup Language)

- Model of the common components in different types of FG experiments
- Shared base for different data formats
- Goals:
  - Improved integration of different data types
  - Simplify development of data standards
  - Single framework for describing laboratory workflows for functional genomics (e.g. for linking unrelated data formats)

FuGE structure

FuGE exists as:
1. Object model (UML)
2. XML schema

Audit

Description

Ontology

Protocol

Reference

Common

FuGE

Bio

Investigation

Data

Material

Conceptual Molecule

Common:
•Auditing and Security settings
•Referencing external resources
•Protocols
•Standard object identification system (LSID)

Bio:
•Investigation structure (and experimental variables)
•Data
•Materials (organisms, solutions, compounds)
•Theoretical molecules e.g. sequences

http://fuge.sourceforge.net/
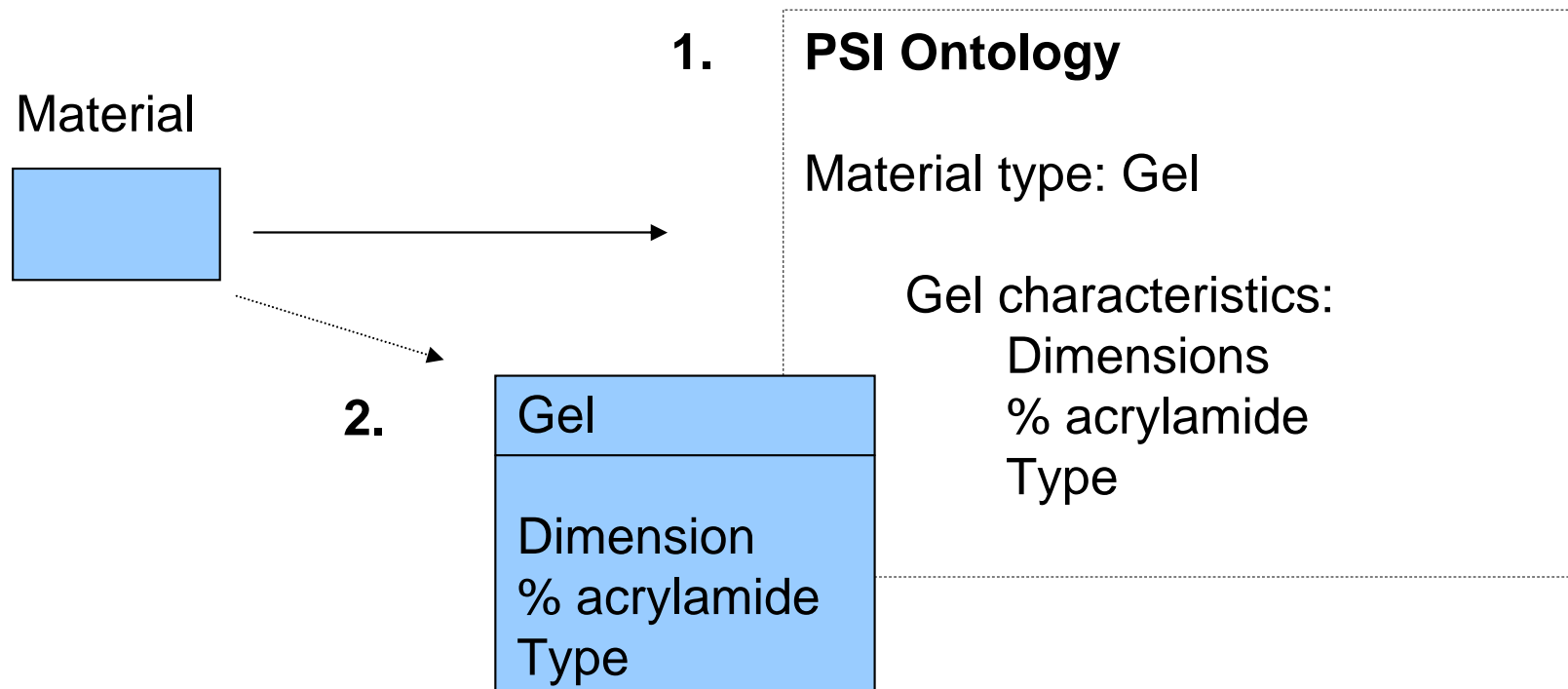
Capturing lab workflows

1. Sample
e.g....

2. Separation protocol
e.g....

3. Data acquisition
e.g....

4. Data transformation
e.g. database search creates results
(new Data objects)
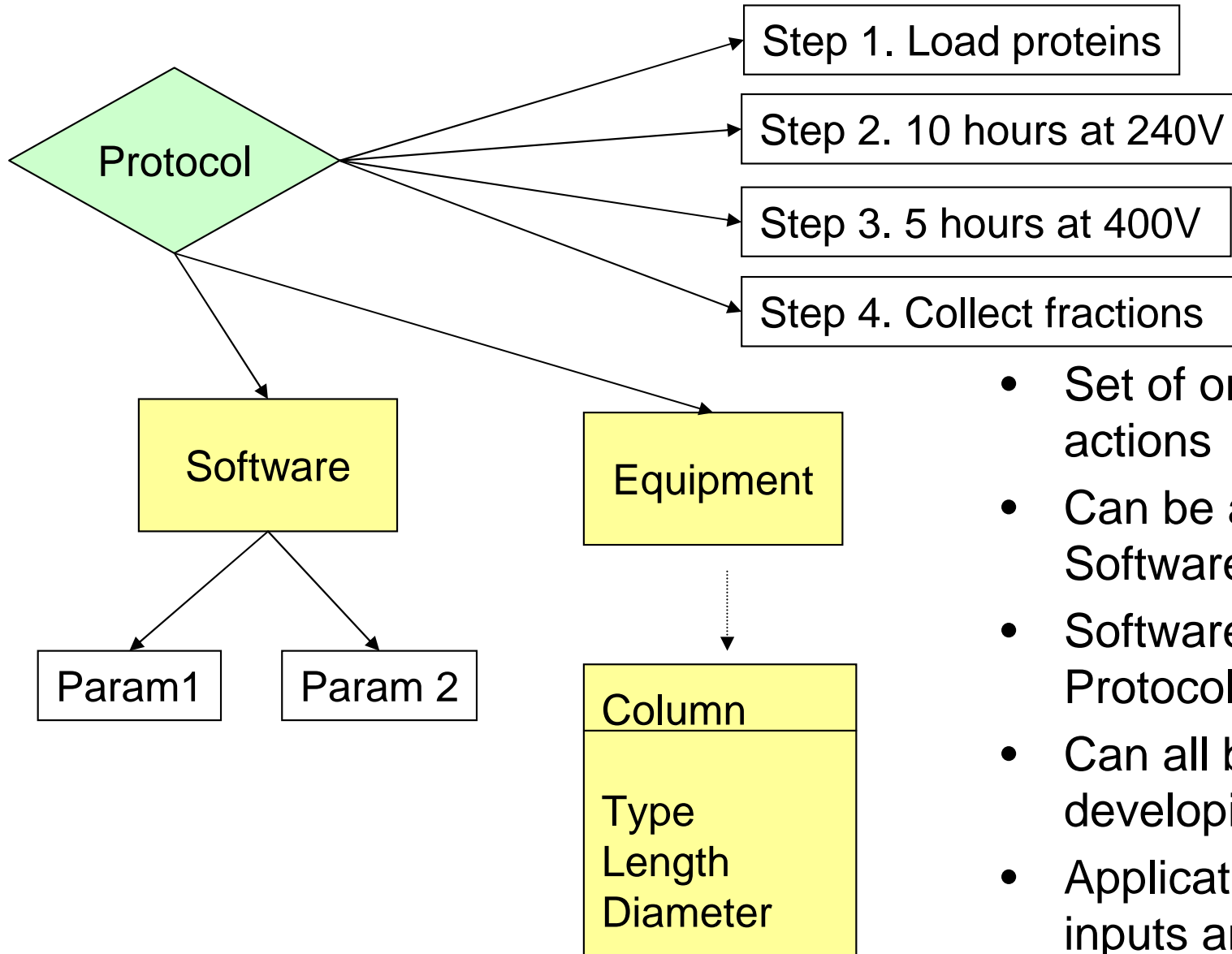
Material

Data
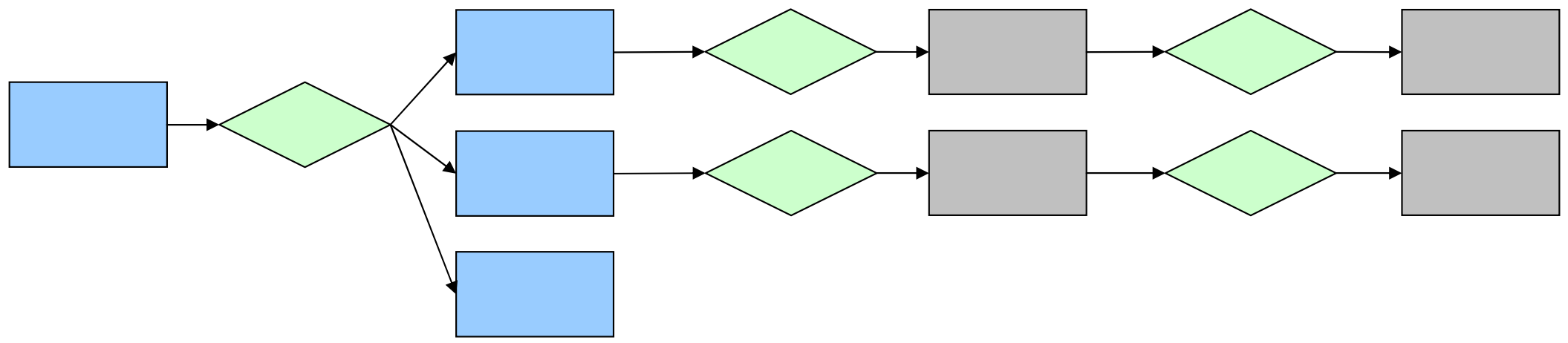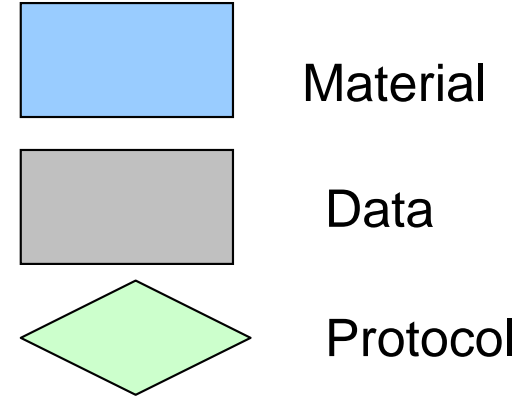
Protocol

What are Materials and Protocols?

FuGE provides 2 options:
1. Describe Material using external ontologies (controlled vocabularies)
2. Extend Material model with required attributes

Protocols

Protocol → Step 1. Load proteins

Protocol → Step 2. 10 hours at 240V

Protocol → Step 3. 5 hours at 400V

Protocol → Step 4. Collect fractions

Protocol → Software → Param1, Param 2

Protocol → Equipment → Column (Type, Length, Diameter)

- Set of ordered simple (atomic) actions
- Can be associated with Software and Equipment
- Software, Equipment & Protocols all have parameters
- Can all be extended for developing own format
- Application of protocol can map inputs and outputs

Protocols

Input material

Column
ProtocolApplication

Output material

Output data

Sample

Fraction

Chromatogram

omic)

Par   eters

Column

Type
Length
Diameter

- Can all be extended for developing own format
- **Application of protocol can map inputs and outputs**
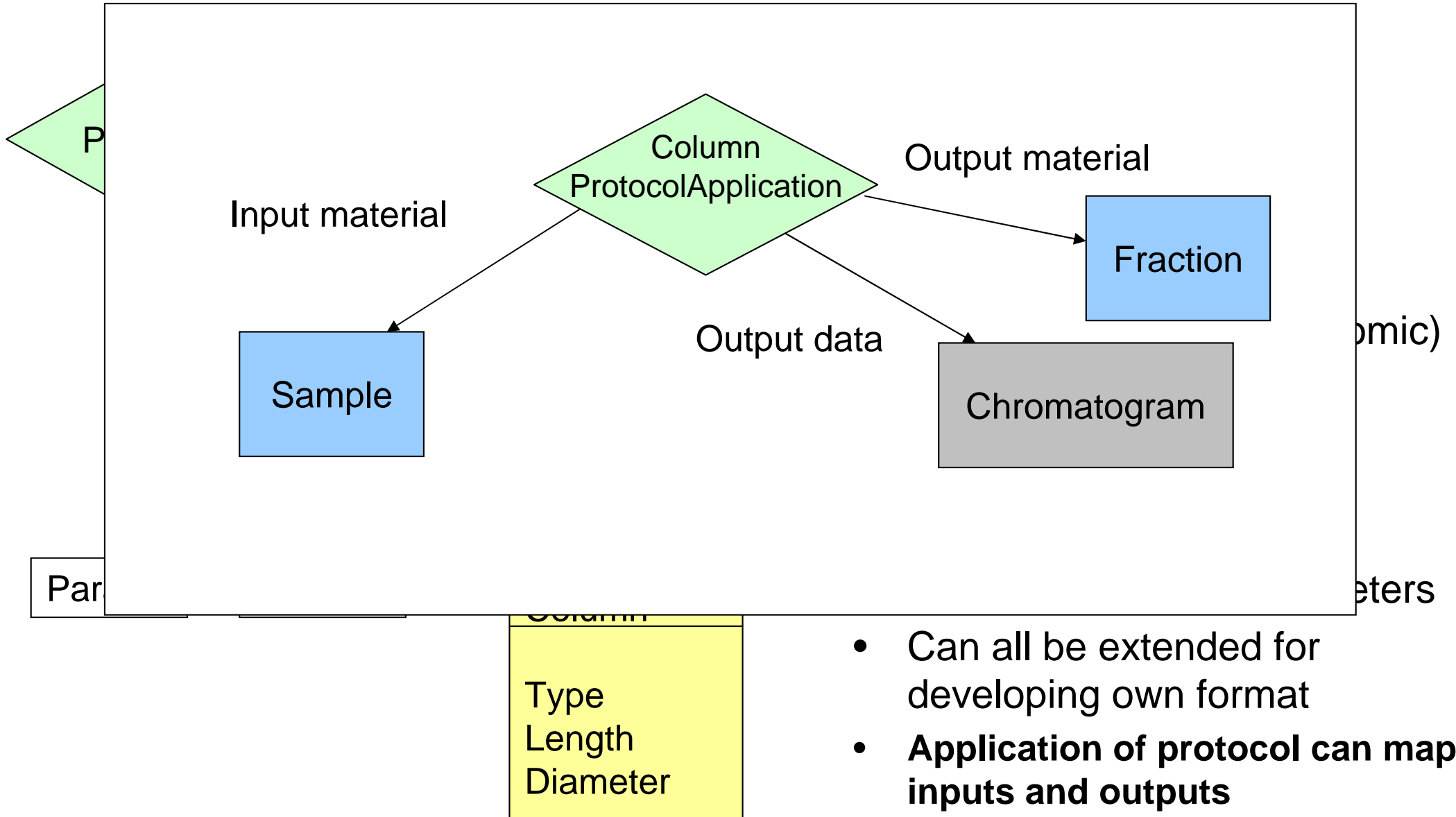
Status of FuGE

- Milestone 1 release - Sep 2005
- Milestone 2 release - Dec 2005
- FuGE version 1.0 - Spring 2006
  - Will include UML, XML Schema, documentation and software library

Formats using FuGE:

- MAGE-ML version 2 (MGED)
- GelML, spML (PSI)
- Planned for mzData v2, analysisXML (PSI)

spML (Sample Processing - ML)

- Model for sample processing and separation techniques
- Extends from FuGE
- Includes models for:
  - Capillary electrophoresis
  - Columns e.g. liquid chromatography
  - Centrifugation
  - Rotofors *(type of isoelectric focussing)*
  - General separation, splitting, combining etc.
  - Protocols for defining buffers / solutions etc.

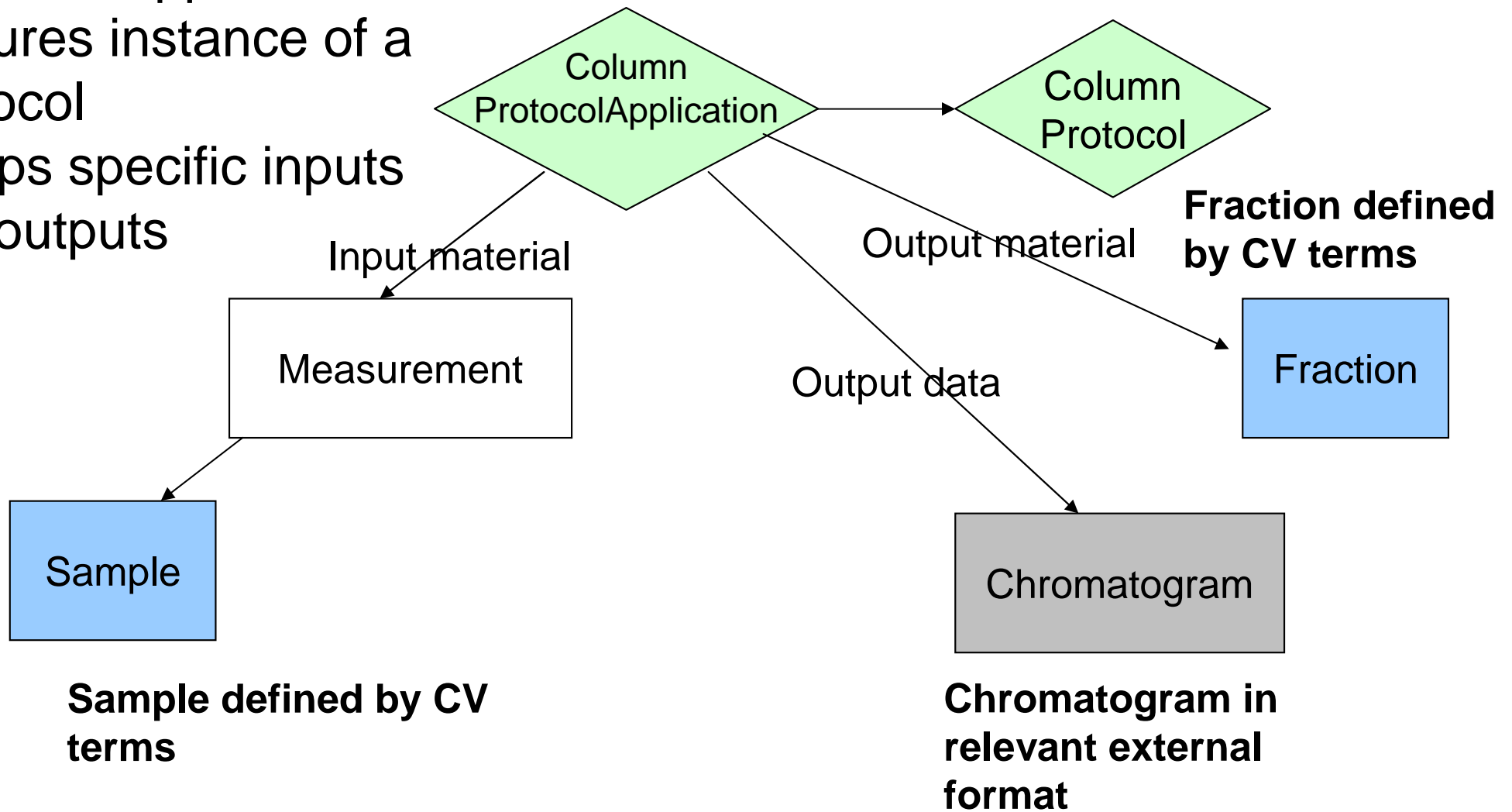http://psidev.sourceforge.net/gps/

spML – Column Model

Equipment

Column
___
Type
Length
Diameter
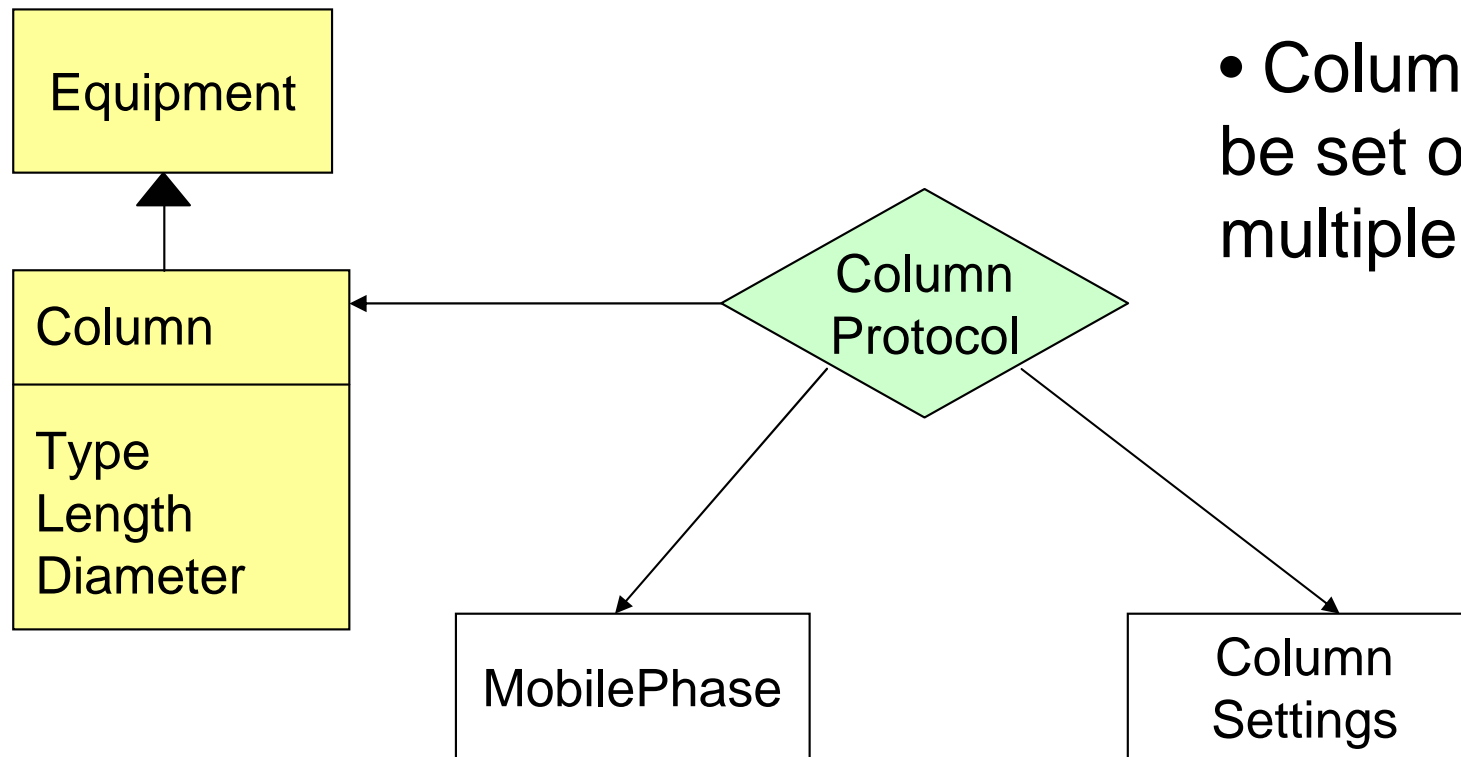
Column Protocol

- ColumnProtocol can be set once and used multiple times

MobilePhase

Column Settings

http://psidev.sourceforge.net/gps/

spML – Column Model

- ProtocolApplication captures instance of a Protocol
- Maps specific inputs and outputs

Column ProtocolApplication

Column Protocol

Fraction defined by CV terms

Input material

Output material

Measurement

Output data

Fraction

Sample

Chromatogram

Sample defined by CV terms

Chromatogram in relevant external format

http://psidev.sourceforge.net/gps/

spML (Sample Processing - ML)

- spML milestone 1 – Dec 2005
- Aim for milestone 2 March/April 2006
- Could be applicable for sample processing in metabolomics
  - Example: could include gas chromatography and other separations
  - Would like to encourage feedback and testing
- Can be integrated with other parts of a lab workflow using FuGE

http://psidev.sourceforge.net/gps/

Conclusions

- FuGE adopted by microarray and proteomics standards bodies
- Experience with building formats by extension
- Simplifies format development
  - Developers can focus on what to capture rather than how to build model
  - Framework will allow auto-generation of XML Schema & software library for all extensions
- Major benefits if all data formats for functional genomics share a common structure
  - Improved integration of data

Acknowledgements

- ## FuGE contributors
  - Angel Pizarro (UPenn), Michael Miller (Rosetta), Paul Spellman (Lawrence Berkley)
  - PSI
  - MGED
  - Fred Hutchinson Cancer Research Centre
  - Genologics

Email: ajones@cs.man.ac.uk

FuGE Web: http://fuge.sourceforge.net/

spML Web: http://psidev.sourceforge.net/gps/