Data Integration in myGrid with Taverna

Duncan Hull
University of Manchester, UK
(on behalf of the myGrid team)

International Workshop on the Interoperability
of Biological Information Resources
(IBIR 2006), Tokyo, Japan

17th March 2006

Outline

- The myGrid project and Taverna, its flagship product
  - Funding: e-Science and OMII
- Taverna version 1.0: a product of the myGrid **development track**
  - Motivating scenario: Data Integration in the Life Sciences:
    - Supporting the *in silico* experimentation life cycle
    - User requirements
  - Demonstration: What you can do now
  - Architecture
  - Lessons learnt from development that feed the research
- myGrid **research track** which feeds into next version
  - Semantic Web Services and workflow repositories
  - Provenance
  - The Grid
- Taverna version 2: What we learnt from version 1, how to make it better

Introduction

- UK e-Science Pilot Project
  - Phase 1: 2001 – 2005
  - £3.5 million
- OMII-UK
  - Phase 2: 2005 - To 2009
  - £2 million

Particular thanks to the other members of the Taverna project, http://taverna.sf.net

Funding: e-Science and OMII

"e-Science is about global collaboration in key areas of science and the next generation of [computing] infrastructure that will enable it."

Sir John Taylor
Director Office of Science and Technology, UK

e.g. e-Science analagous to e-Business. Not only for Life Sciences, but also Physical Sciences myGrid aimed to support the e-scientist

- OMII Open Middleware Infrastructure: omii.ac.uk
  - Aims to be "…the source for reliable, interoperable and open-source Grid middleware, ensuring the success of Grid-enabled e-Science".

myGrid

Problem: e-Science

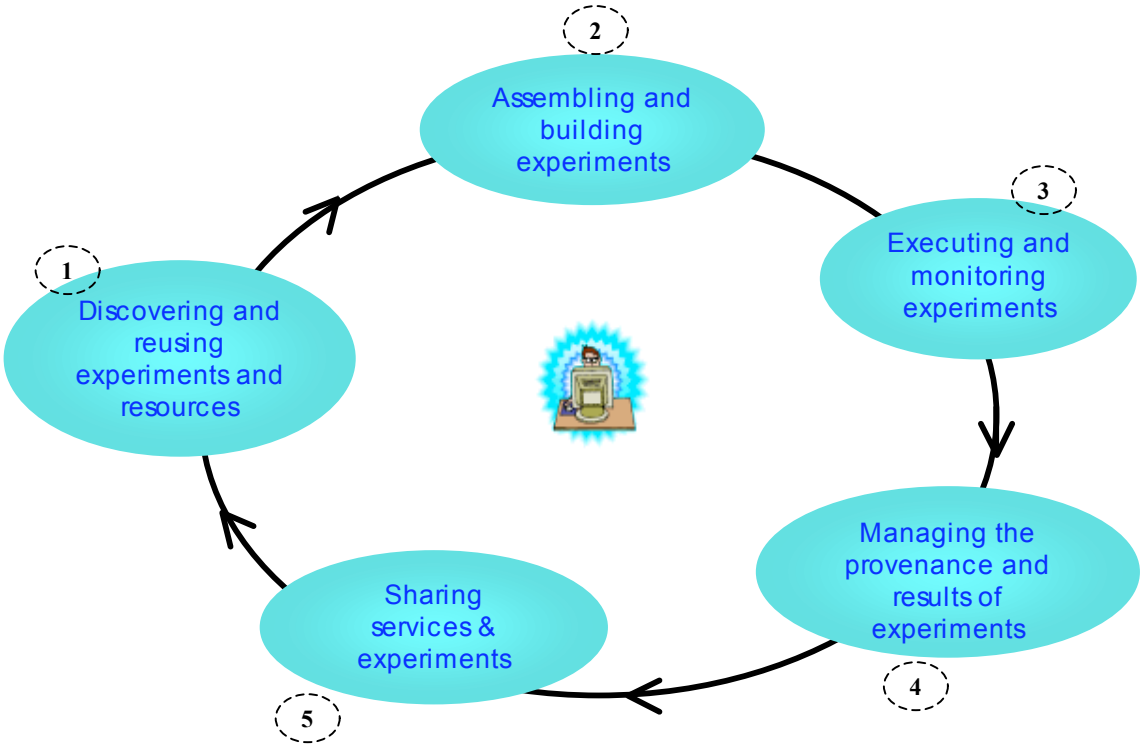
- Life Sciences, especially molecular biology, has terabytes of heterogeneous, autonomous data and tools on the Web that need to integrating in order to understand DNA, genes, genomes, proteins, biological pathways etc
- 858 public databases
  - MY Galperin. The molecular biology database collection: 2006 update. *Nucleic Acids Research*, 34(Database issue):3-5, Jan 2006.
- 150+ public web servers
  - JA Fox, SL Butland, S McMillan, G Campbell, and BF Ouellette. The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Research*, 33(Web Server issue):3-4, Jul 2005.

Problem…continued

- Between 2,000 and 3,000 public services (e.g. sequence analysis programs like BLAST that use Web Service standards like WSDL and SOAP)

- All these databases, servers and services allow us to perform many different sorts of computations on DNA, RNA and Proteins
  - Genome annotation
  - Systems biology
  - Phylogenetics, evolution
  - Microarray analysis
- (e-)Scientists need combine all these resources in their experiments, *in silico*, e.g. on the web

myGrid *In silico* Experiment Lifecycle



-Service-oriented middleware and tools that formalize and support the lifecycle:

- -Service/Experiment Discovery — Feta
- -Service Selection
- -Service Composition
- -Service Execution & Execution Reporting
- -Result Display

Taverna & Freefluo

- -Result Storage and Management — Provenance

Using Workflows is one way to make these experiments structured, shareable, repeatable and verifiable.

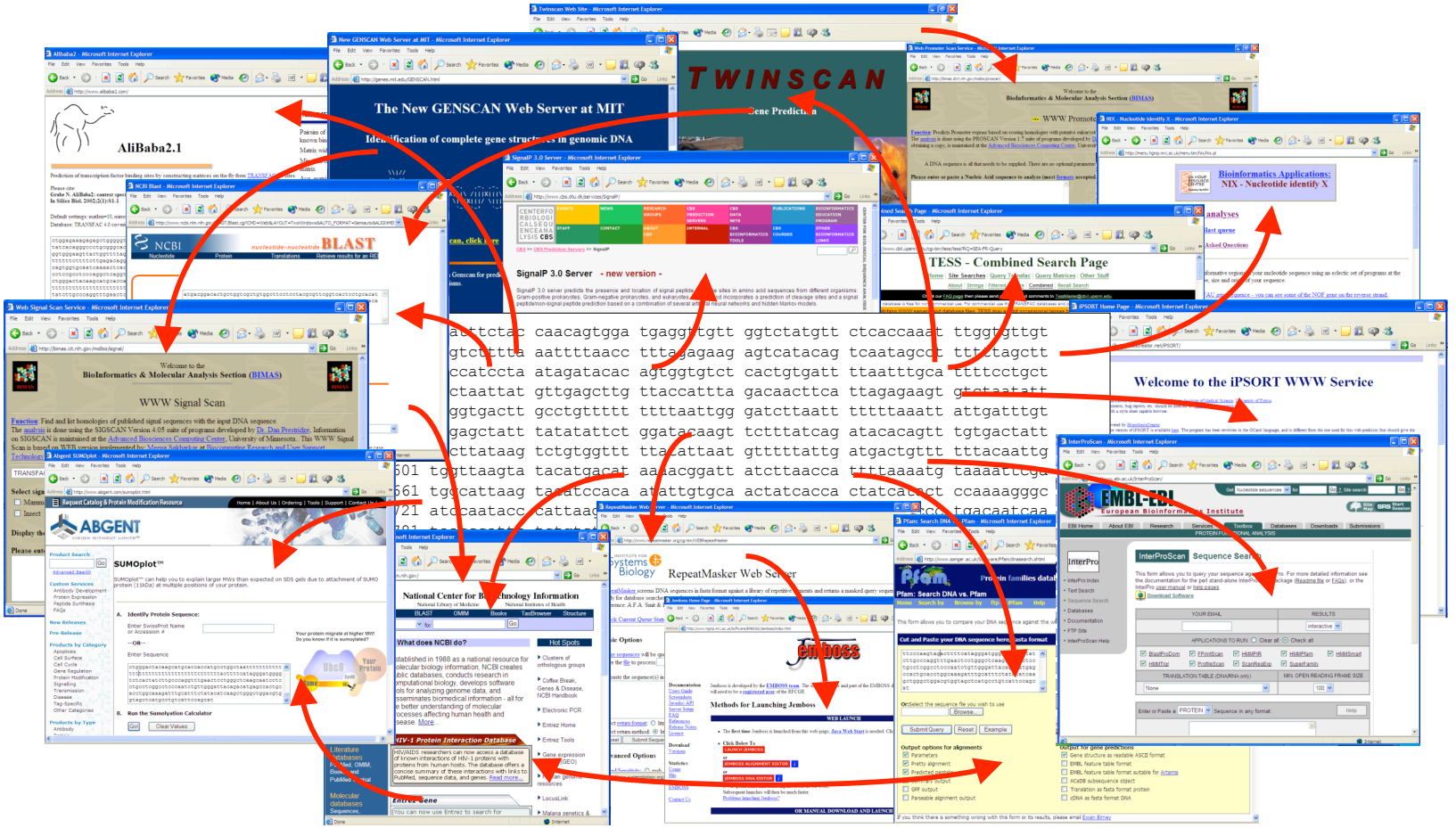
Example: Case study

1. Identify new sequences to close a gap in a highly repetitive region of human chromosome 7, implicated in WBS
2. Characterise the new sequence (DNA and protein)
   - Comparative/speculative reasoning, (making predictions based on previously made **similar** observations)
   - Repetitive application of standard bioinformatics techniques using Web based forms
3. GenBank, BLAST, RepeatMasker, InterProScan etc standard tools and databases (GenBank)

See: Robert D. Stevens, Hannah J. Tipney, Chris Wroe, Tom Oinn, Martin Senger, Phillip W Lord, Carole A. Goble, Andy Brass, and May Tassabehji. Exploring Williams-Beuren Syndrome Using myGrid. *Bioinformatics*, 20:i303-i310, 2004.

myGrid

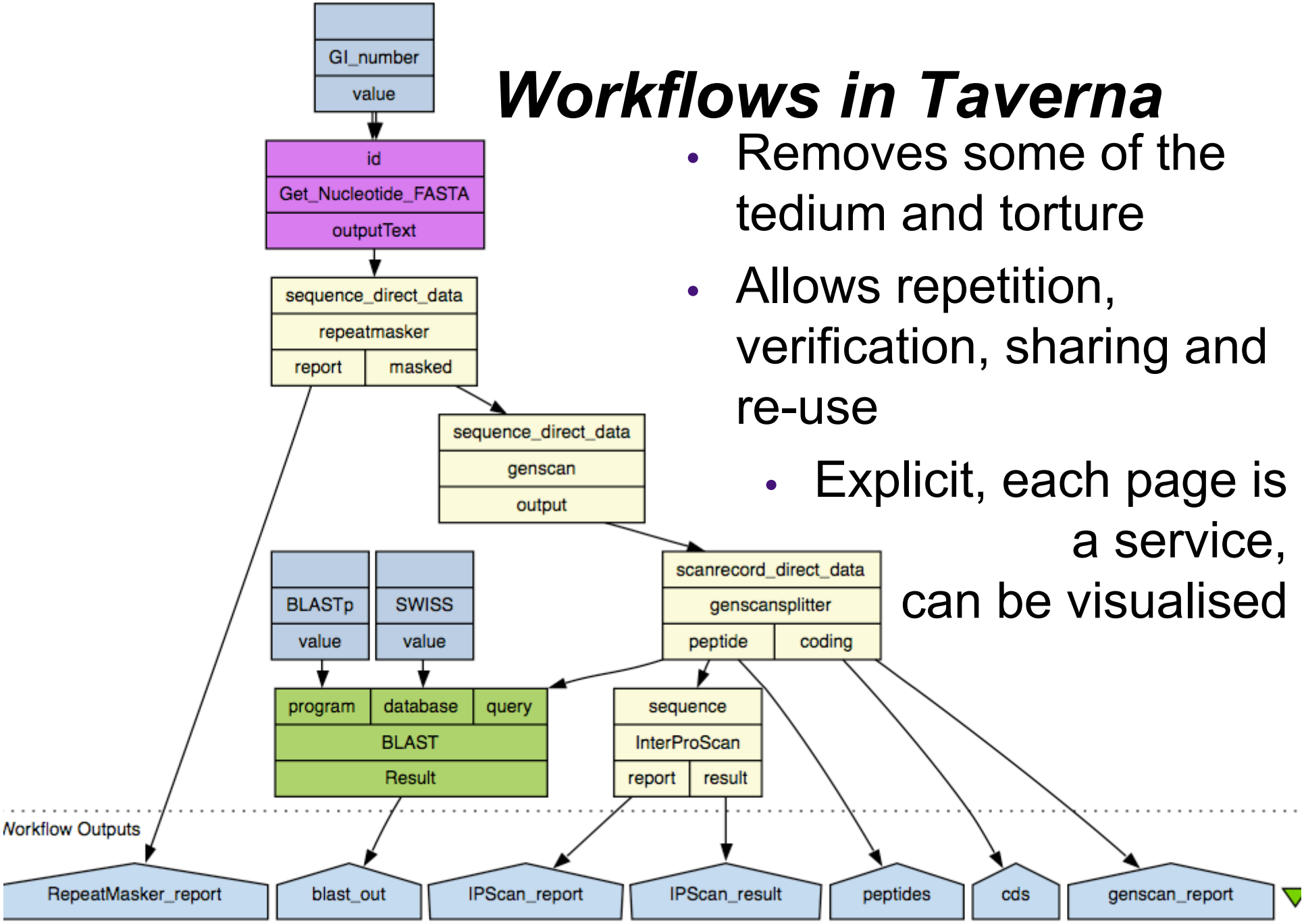
“Cut-and-paste”



Can’t repeat, share, modify or verify these experiments.
Not a robust solution.

Workflows in Taverna

- Removes some of the tedium and torture
- Allows repetition, verification, sharing and re-use
  - Explicit, each page is a service, can be visualised

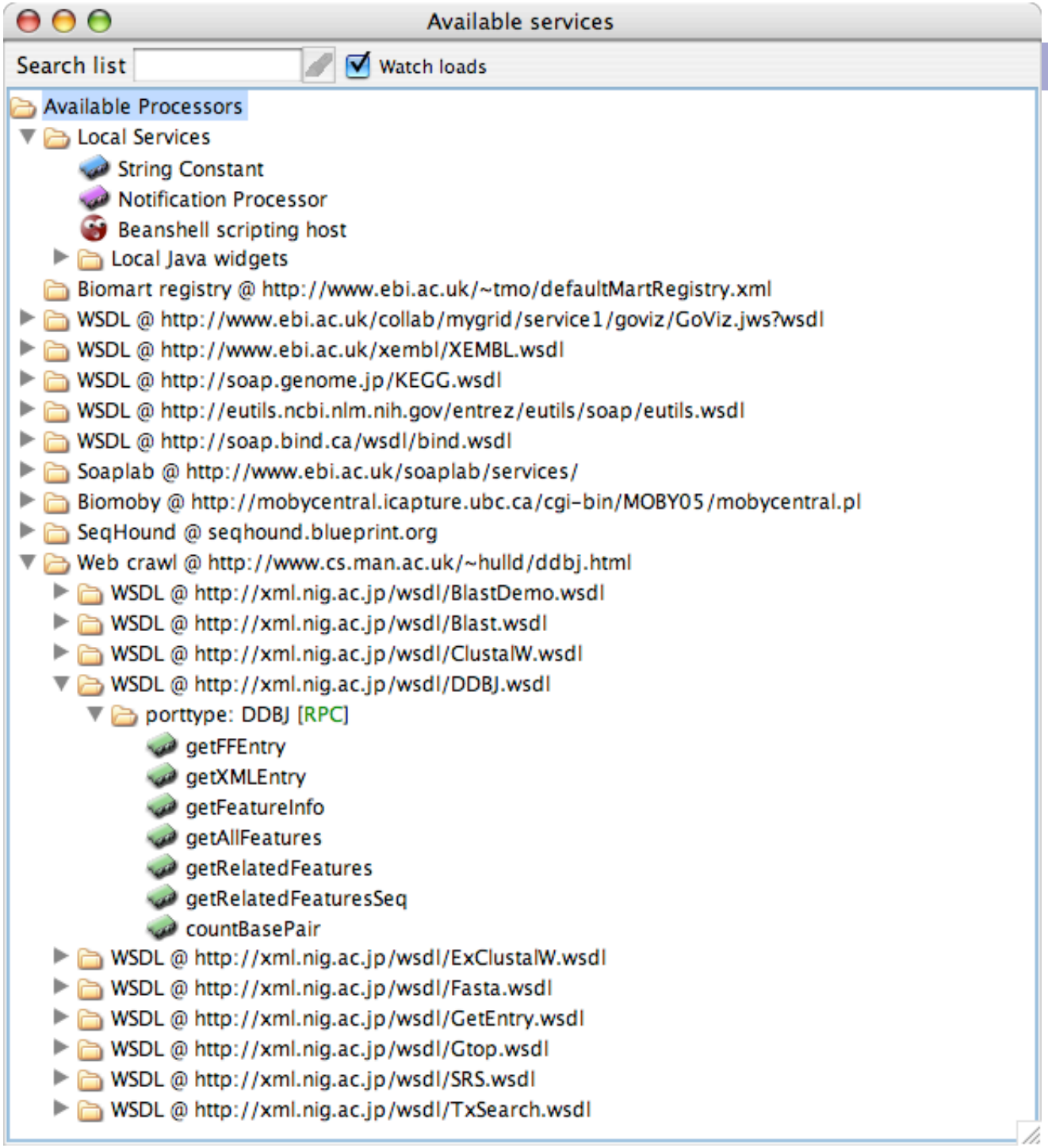


What is the Taverna Workbench?

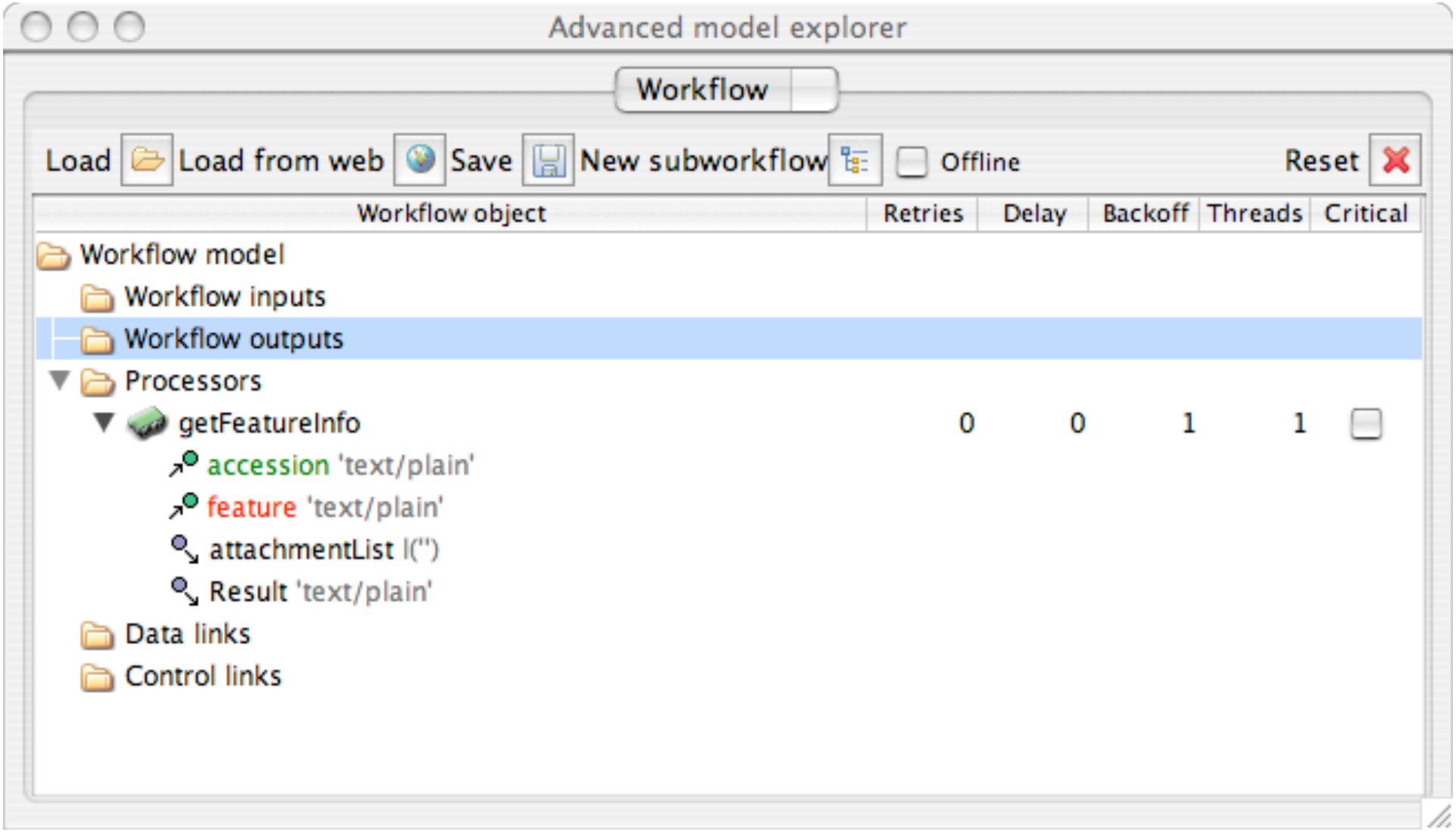
- A “super client” that
  - Allows scientists to graphically construct complex processes in the form of workflows expressed in the Simplified ConceptUal workFlow Language (Scufl)
    - A Scufl workflow?
      - Set of processors that make up a process
      - Definitions about how data moves between these processors (data links)
      - Simple conditional branching using control flow (co-ordination links)
      - Specification of what needs to be done but not how to do it
      - Interacts with the enactment engine (FreeFluo) to execute the workflow
  - Insulates scientist from complexity of invoking web services

myGrid

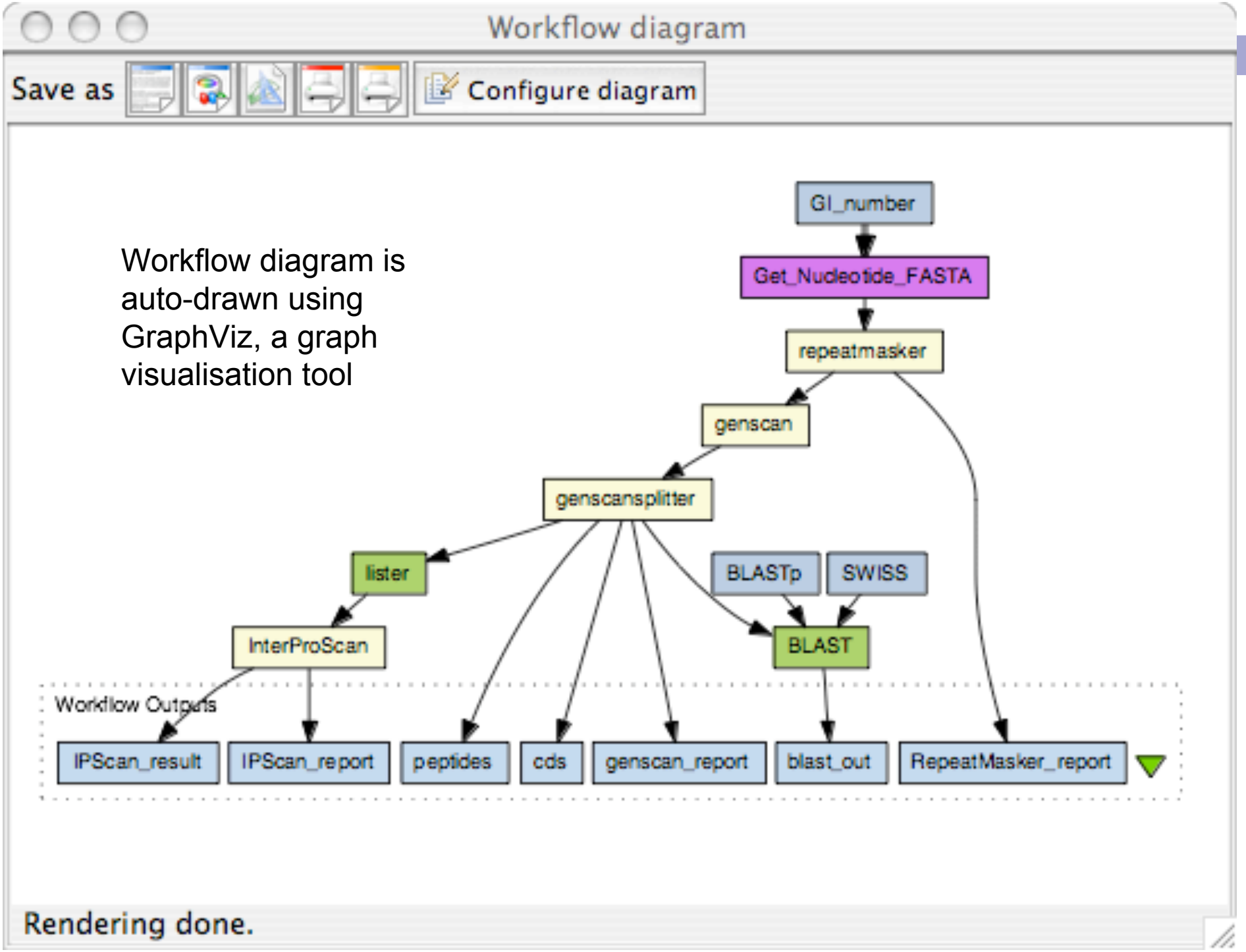
Demonstration of Taverna 1.3



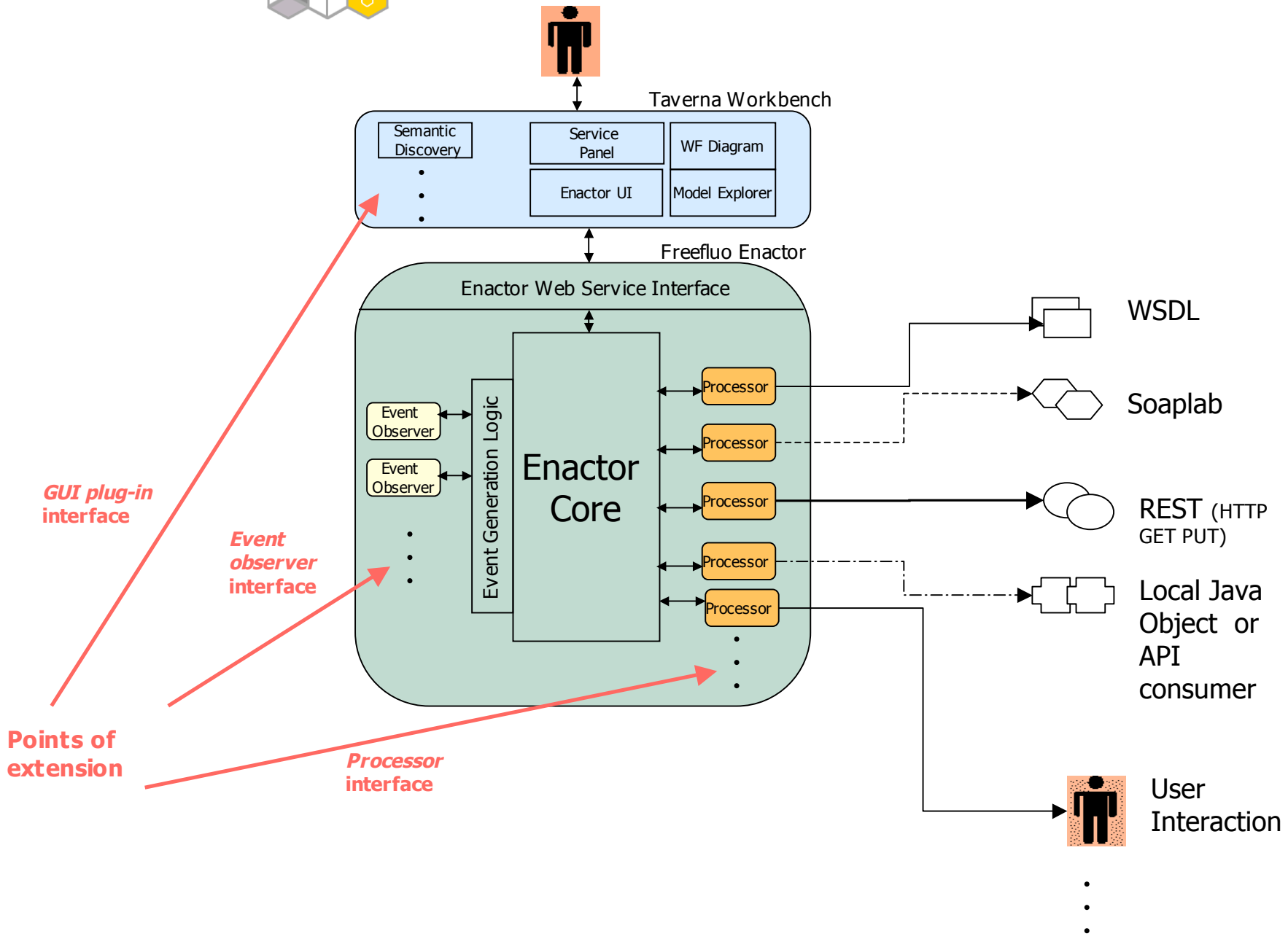
- Different types of processor: (each with its own invocation mechanism) e.g.
- Local java widgets
- Beanshell
- SOAPlab
- BioMOBY
- BioMART
- Can add abitary WSDL



- Shows inputs and outputs with names and types
- Can connect up inputs to outputs or add control co-ordination


Workflow diagram
Save as
Configure diagram
Workflow diagram is auto-drawn using GraphViz, a graph visualisation tool
GI_number
Get_Nucleotide_FASTA
repeatmasker
genscan
genscansplitter
lister
BLASTp
SWISS
InterProScan
BLAST
Workflow Outputs
IPScan_result
IPScan_report
peptides
cds
genscan_report
blast_out
RepeatMasker_report
Rendering done.

Taverna - Architecture



An Open World

- Open source (LGPL)
- Open domain services and resources
- Open community
- Open application
  - Nothing specific to biology, although oriented to it
- Open model and open data
  - No prescribed typing or domain data model
  - A layered information model
- Open architecture
  - Service Oriented Architecture
  - Loosely coupled, Web services based

open source™

open middleware infrastructure institute

myGrid

Open environment

NCBI

National Center for Biotechnology Information (USA)

DDBJ DNA Data Bank of Japan

Tokyo, Japan

EMBL-EBI European Bioinformatics Institute

Cambridge, UK

biomart

SoapLab

wellcome trust sanger institute

KEGG Kyoto Encyclopedia of Genes and Genomes

MOUNT SINAI HOSPITAL

PDBj XP Protein Data Bank Japan Experimetal Version

PathPort The Pathogen Portal Web Project

emboss

moby

SeqVISTA

LION

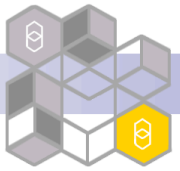
SRS

SeqHound

caBIO

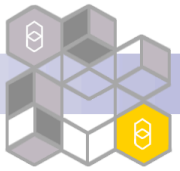
Taverna 1.3 Support

- Taverna has a user community, (developer and user mailing lists) “taverna-hackers”, “taverna-users”
- ~1500 installations, 14,000 downloads, part of bio-linux http://envgen.nox.ac.uk/biolinux.html
- Has a user manual
- Is written in Java, so can be used on Windows, Mac and Linux (90% of the binary downloads are the windows version)
- Has User days, demos at conferences e.g. Intelligent Systems in Molecular Biology (ISMB 2004-2006) and in Manchester
- All accessible from http://taverna.sf.net
- Publications…see one-page sheet that accompanies this talk
  - Also we have submitted an updated description of Taverna to the 2006 Nucleic Acids Web Server issue which we hope will be published in July



Strengths & weaknesses

- Not enforcing a common type system
  - Objects passed around are largely opaque to the middleware hence provides application interoperation rather than application integration
  - PRO: can quickly add new services, arbitrary WSDL files, more services than BioMOBY
    CON: joining services is difficult, requires shims, less metadata than BioMOBY
- Service oriented architecture
  - PRO: Don't have to install tools and databases locally, access them over the web
    CON:Services can be unreliable and poorly described with licensing issues



Lessons learnt

- Services can be difficult to find because they are poorly described (more later)
- Inevitably, services don't fit together neatly
- Many "shim" services needed, to align inputs and outputs in a pipeline. Close integration in truly open environments is (and always was) a hard problem
- Web Services stack is difficult to debug, Taverna builds on third-party toolkits like Axis, WSDL4J, WSIF which often provide poor error reporting
- Sharing workflows, users are cautious, IPR, privacy, security, advantage to competitors?
- We really need a proper registry! Flexibility of not having one has its advantages…

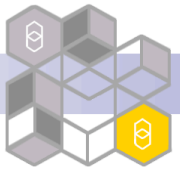
Lessons learnt part 2

- One of the most difficult problems isn't really gathering and co-ordinating services, but gathering and co-ordinating *results, e.g. provenance*
- Getting the abstraction level right, Xscufl worklfows seem to be the appropriate abstraction for many bioinformaticians
- We need more services, more replicas of services (for failover), better reliability, stability
- Visualisation is a (unforseen) key benefit, graph drawing using GraphViz

Research track: three areas

1. Semantic Web Services and workflows
   - Reasoning over metadata
   - Workflow repository http://workflows.mygrid.org.uk
2. Provenance
   - The who, why, what, when and where of an experiment
   - LSIDs
3. The Grid

These two rely on metadata in RDF and OWL

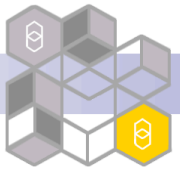


Semantic Web Services?

- Annotate services with ontology terms using the Web Ontology Language (OWL) and RDF
  "Enables automating interoperation, integration, discovery"

see Sheila McIlraith, T. Son, and H. Zeng. Semantic Web Services. IEEE *Intelligent Systems*, pages 46-53, March-April 2001.

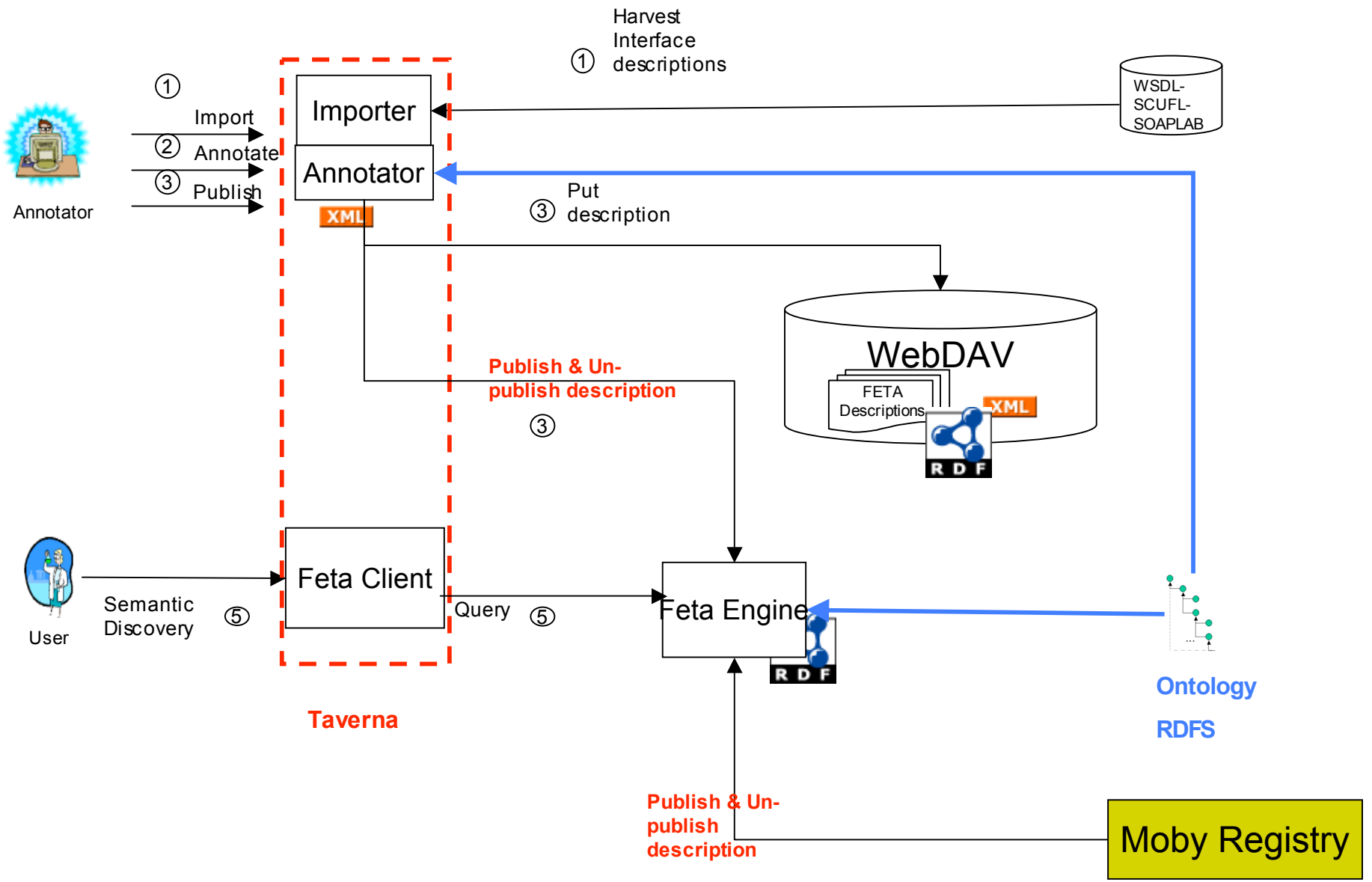
Use reasoners to annotate and classify services and retrieve them "semantic discovery"



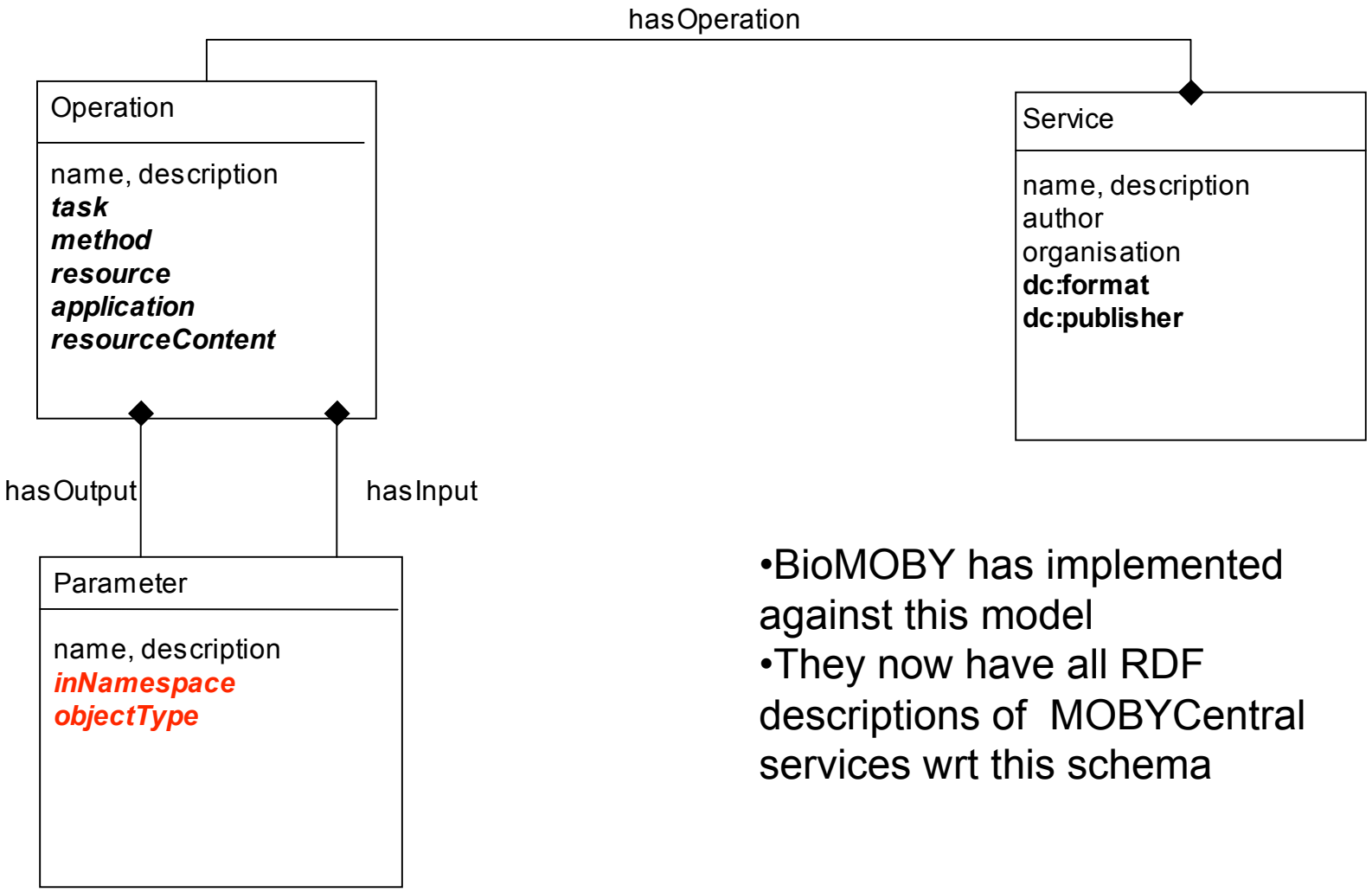
Semantic Web Services

- WSDL in the wild
  - Cryptic operation names “run”, “get”
  - Cryptic parameter names “in0”, “in1”, “out1”
  - Most data “typed” as `xsd:string`
  - ….But these hide complex legacy flat-file formats e.g. BLAST reports and Database records etc
  - Extensive use of XML schema (e.g. complex types) is rare but does happen e.g. NCBI e-utils WSDL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl
  - So we need to annotate WSDL somehow, two different mechanisms

Feta Engine 1.0



myGrid-BioMoby Service Model



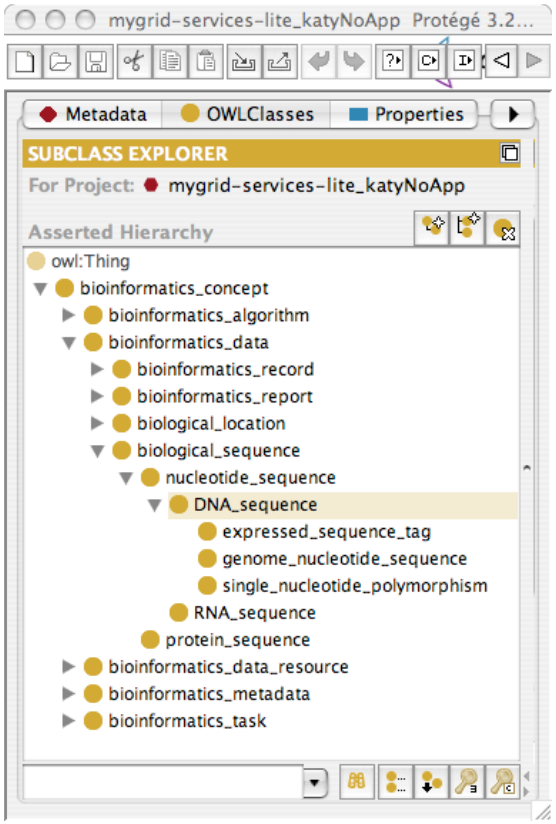
- BioMOBY has implemented against this model
- They now have all RDF descriptions of MOBYCentral services wrt this schema

BioMOBY and Taverna

- Shared model `objectType`

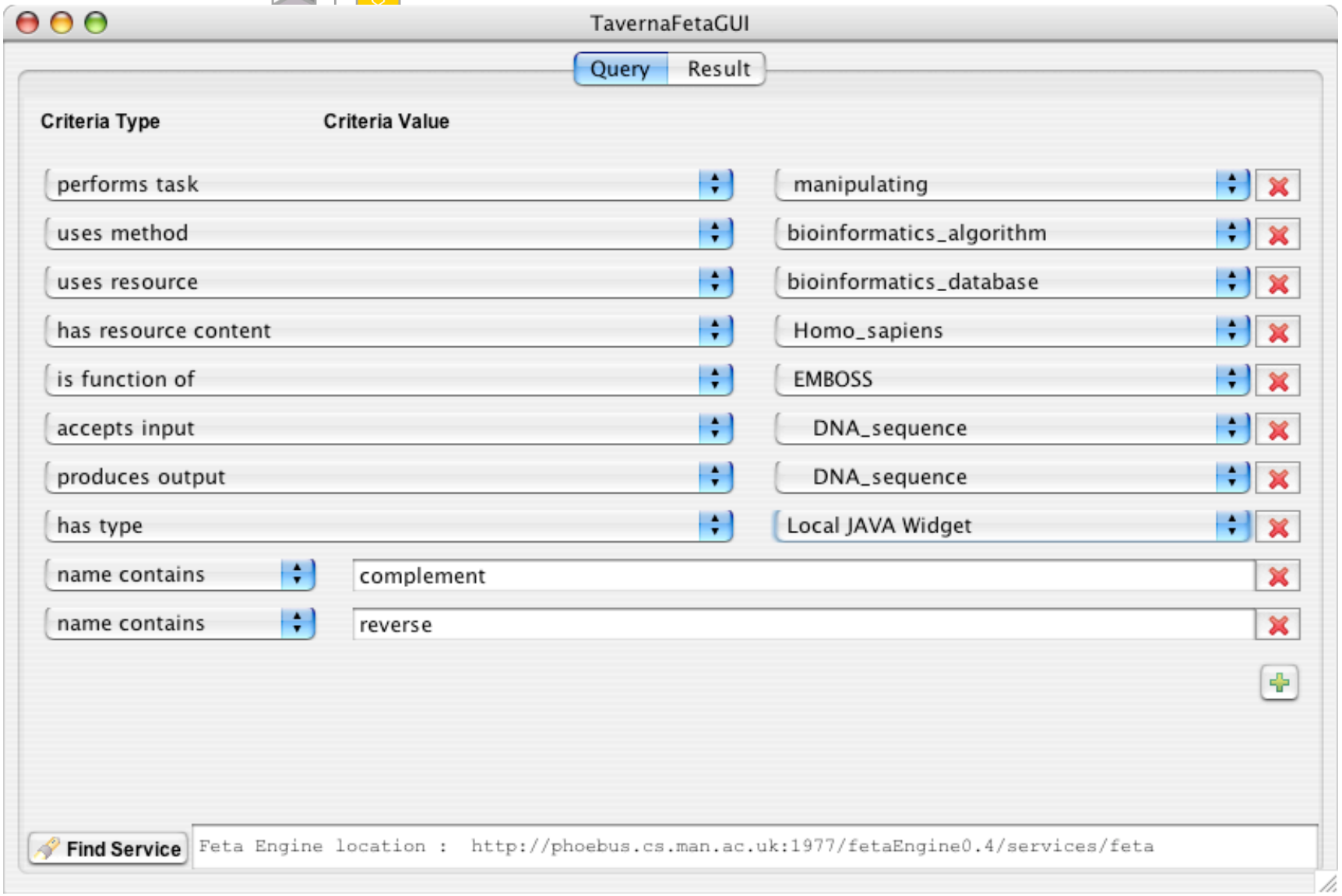
myGrid ontology (OWL) with more complex relations

RDF model with ISA, HAS & HASA relations



http://biomoby.org/RESOURCES/MOBY-S/Objects

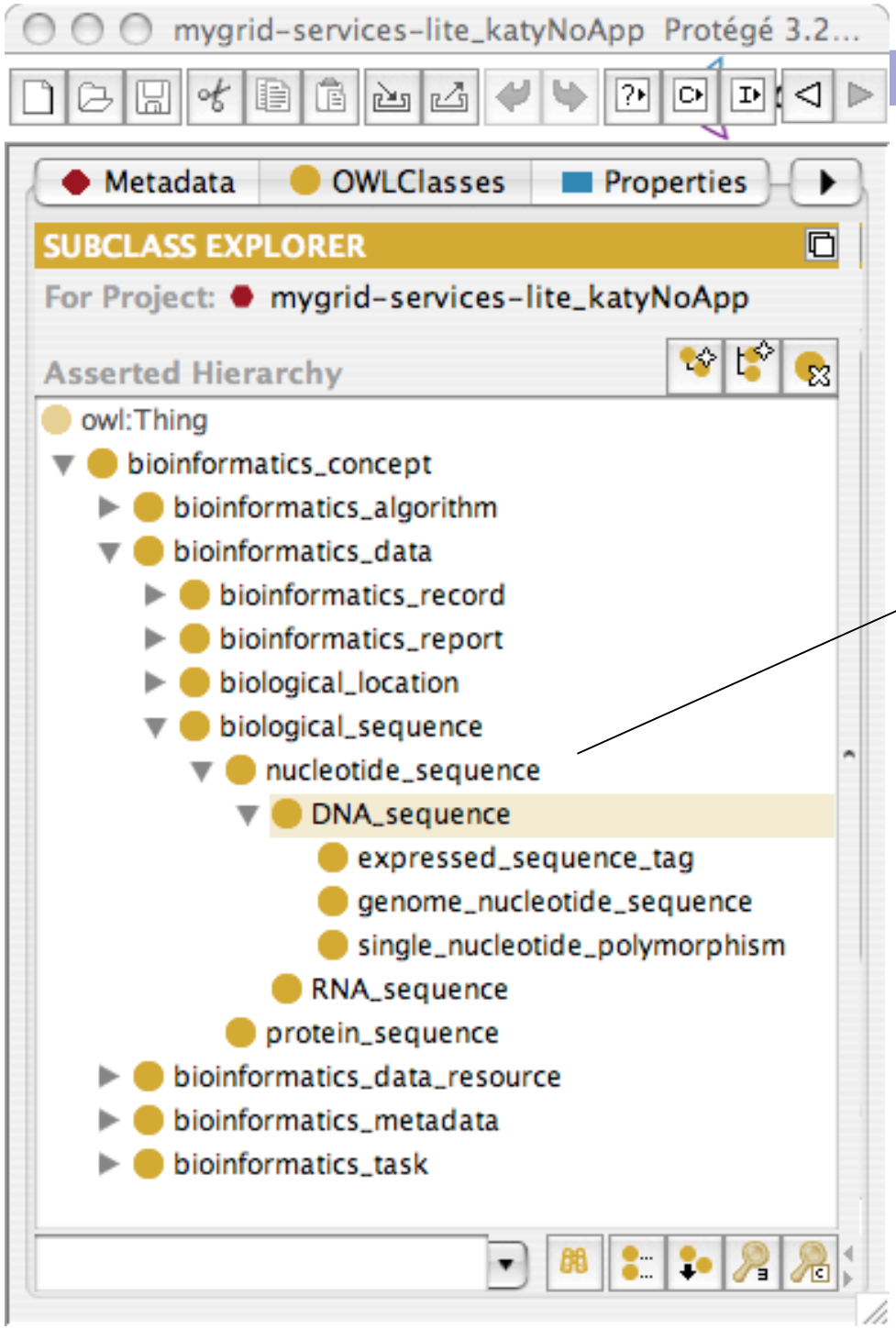
myGrid


TavernaFetaGUI
Query
Result
Criteria Type
Criteria Value
performs task
manipulating
uses method
bioinformatics_algorithm
uses resource
bioinformatics_database
has resource content
Homo_sapiens
is function of
EMBOSS
accepts input
DNA_sequence
produces output
DNA_sequence
has type
Local JAVA Widget
name contains
complement
name contains
reverse
Find Service
Feta Engine location : http://phoebus.cs.man.ac.uk:1977/fetaEngine0.4/services/feta

myGrid

Current Status: myGrid Ontology

- Aims to capture domain knowledge
- Similar to Gene Ontology, but used to annotate web services instead of Proteins
  - Provides the vocabulary
  - Modules for
    - Service Ontology,
    - Bioinformatics
    - Molecular Biology
- Two forms exist
  - OWL (using OWL-S)
  - RDF(S)



Useful for finding services

e.g. Find me a service that accepts /produces this input / ouput, or its subclasses / subclasses

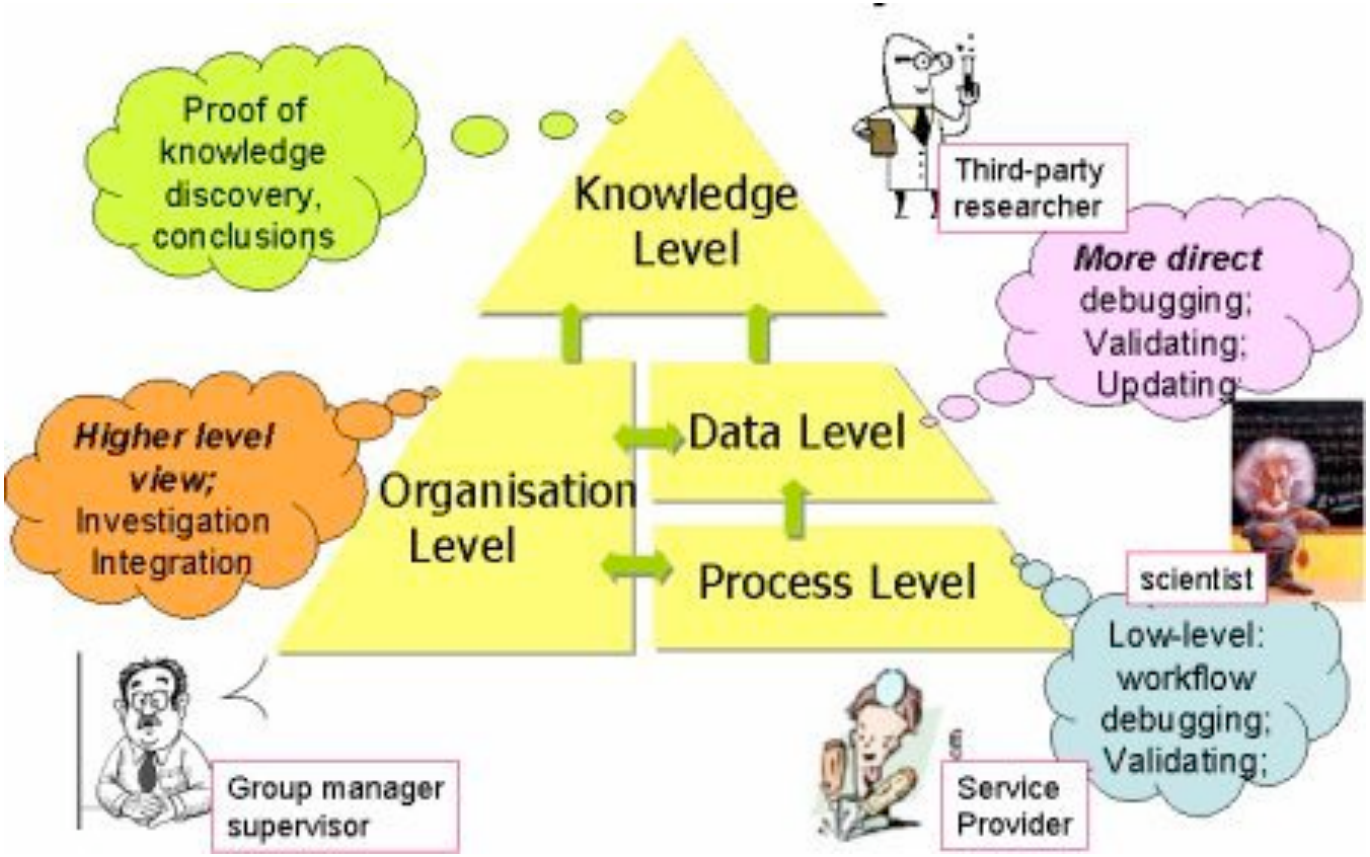
Semantic Workflows?

- Annotate services with ontology terms using the Web Ontology Language (OWL) and RDF
- See http://workflows.mygrid.org.uk and Antoon Goderis, University of Manchester
- Currently does syntactic graph matching on workflows, difficult getting a large number of workflows together

Provenance

- Generated using event-listeners, stored in database based on an RDF model, relies on uniquely identifying objects (workflows, people, genes etc) using LSIDs

Jun Zhao,
University of
Manchester



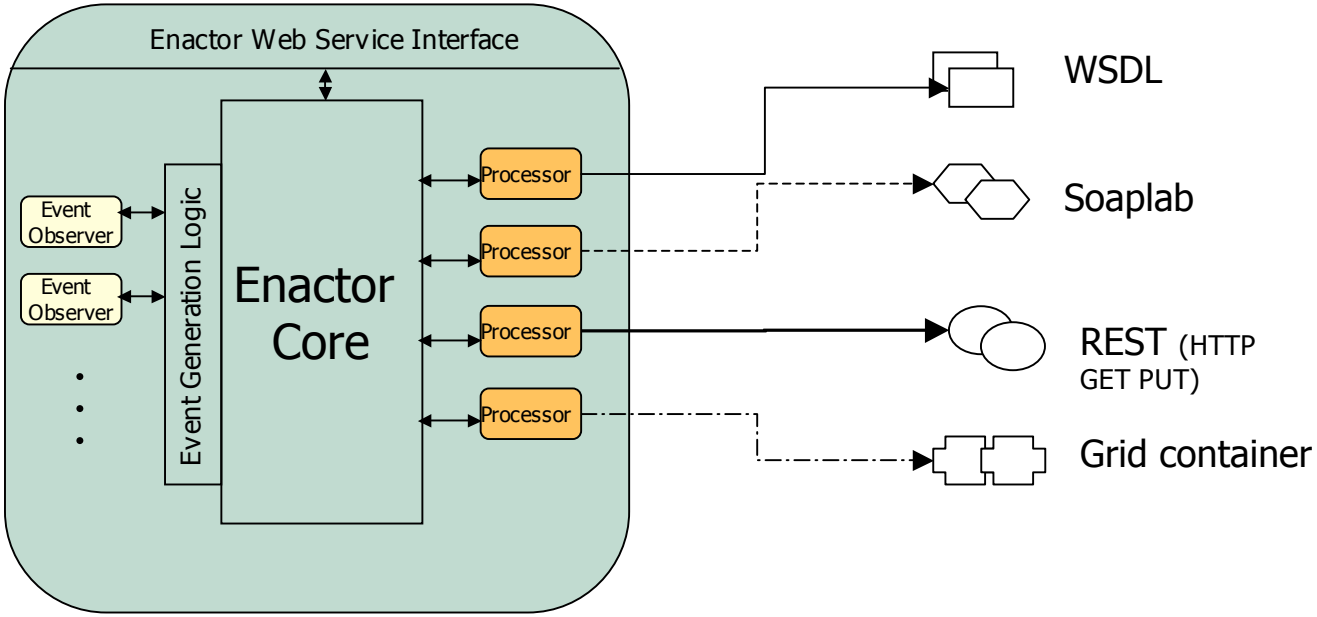
LSIDs

- Life Sciences Identifiers (LSIDs) are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources including genes, people, worklfow-runs etc
- Taverna 1.x uses these extensively for its Provenance, results gathering and management
- The most appropriate model for provenance is not really known, Jun is currently evaluating her model

Taverna 2.0: Scheduled 2007

- “hardening” Taverna
- Revised version of enactor, freefluo
- High-throughput workflows
- Long-running workflows
  - Especially using Grid: job submission services, monitoring services and large-scale data management services
- Semantics integrated more tightly, rather than an add-on

Adding Grid services



InterProScan and BLAST first Grid services

Conclusions

- Taverna, is already a useful tool for bioinformaticians, although there are some issues using version 1.x
- It provides an alternative a significant improvement on cut-and-paste experiments
- Taverna 2 will address the issues with Taverna 1, we'd like to make it more accessible to molecular biologists as well…

Acknowledgements

Pioneers

Hannah Tipney, May Tassabehji, Medical Genetics team at St Marys Hospital, Manchester, UK; **Simon Pearce, Claire Jennings**, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK; **Doug Kell, Peter Li,** Integrative Biology Centre, University of Manchester, UK; **Andy Brass, Paul Fisher**, Bio-Health Informatics Group, UoM, UK, **Simon Hubbard,** Faculty of Life Sciences, University of Manchester, UK

Core Research and Development

Nedim Alpdemir, Pinar Alper, Khalid Belhajjame, Tim Carver, Rich Cawley, Justin Ferris, Matthew Gamble, Kevin Glover, Mark Greenwood, Ananth Krishna, Peter Li, Phillip Lord, Darren Marvin, Simon Miles, Arijit Mukherjee, Tom Oinn, Stuart Owen, Juri Papay, Savas Parastatidis, Matthew Pocock, Stefan Rennick-Egglestone, Ian Roberts, Martin Senger, Nick Sharman, Stian Soiland, Victor Tan, Daniele Turi, David Withers, Katy Wolstencroft and Chris Wroe

Postgraduates

Tracy Craddock, Keith Flanagan, Antoon Goderis, Alastair Hampshire, Duncan Hull, Martin Szomszor, Jun Zhao

Investigators

Matthew Addis, Andy Brass, Alvaro Fernandes, Rob Gaizauskas, Carole Goble, Chris Greenhalgh, Luc Moreau, Norman Paton, Peter Rice, Alan Robinson, Robert Stevens, Paul Watson, Anil Wipat,

Industrial and major project collaborators

Dennis Quan, Sean Martin, Michael Niemi (IBM), Mark Wilkinson (BioMOBY)

Sponsors

EPSRC, Wellcome Trust, OMII