open middleware infrastructure institute uk

myGrid

MANCHESTER 1824
The University of Manchester

EPSRC

The Taverna Workflow Workbench and Tool Suite

Professor Carole Goble

The University of Manchester, UK, carole.goble@manchester.ac.uk

Open Middleware Infrastructure Institute UK http://www.omii.ac.uk

EGEE 2007, October 1-5, Budapest, Hungary.

Talk Roadmap

- Survey of Taverna, its philosophy, its background and its associated toolsets.
- Some success stories.
- More on the components and their status
  - The new Taverna2 enactor.
- Highlight some EGEE2 collaboration points.

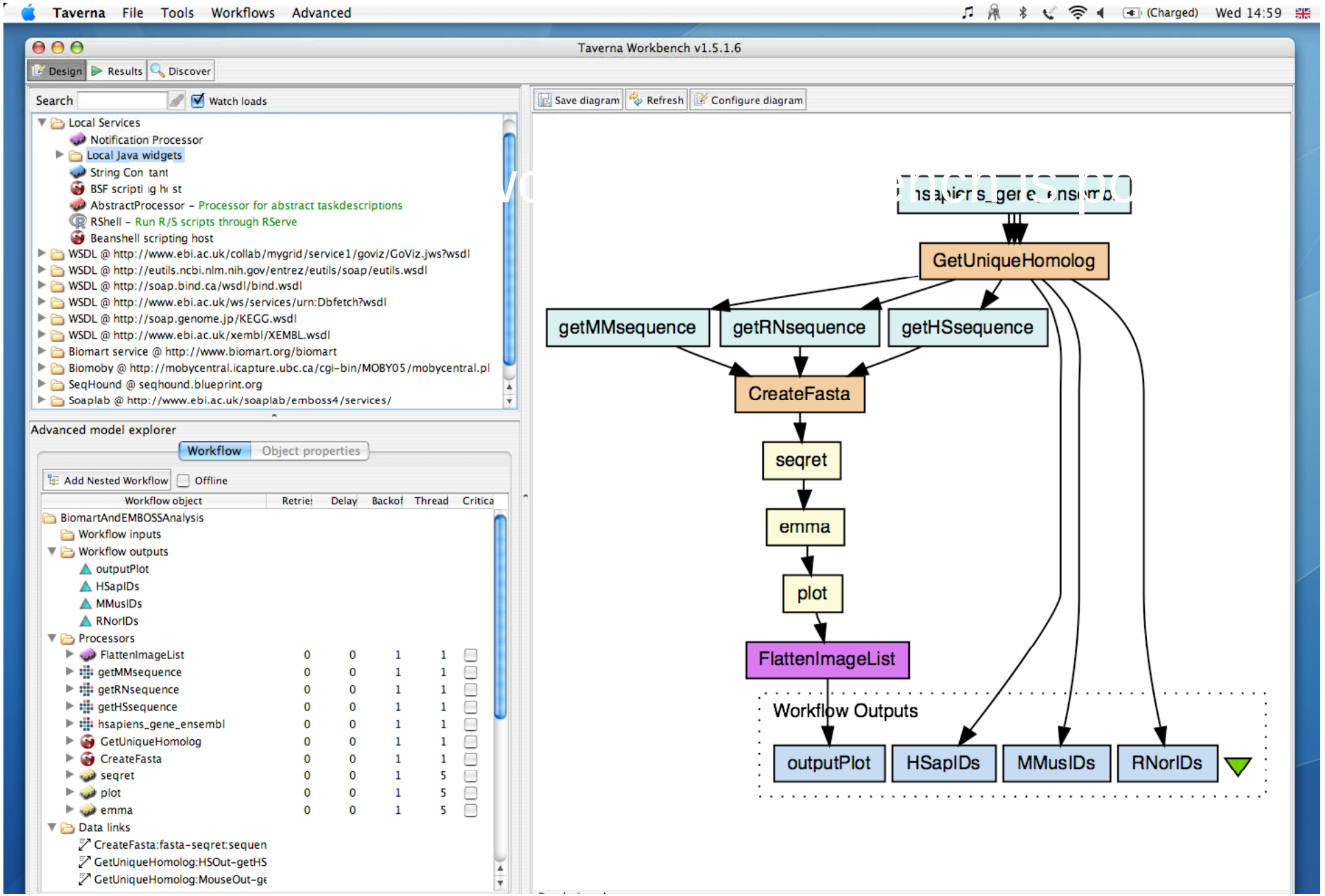
http://www.mygrid.org.uk

- A suite of Open Source middleware for Life Scientists that enables them to:
  - Create and undertake workflows
    - Scufl dataflow language in terms of user's abstraction
    - Extensible Taverna workflow enactor and workbench
    - Web browser toolkit and application toolkit
  - Share and reuse those workflows, and their results.
    - myExperiment social network, bazaar and portal
- UK e-Science pilot project since 2001.
- Open Middleware Infrastructure Institute UK
- Dedicated development team of 11 developers, 1 curator, 3 outreach workers, 1 project manager
- Team of PALs – embedded or tightly-coupled users and developers

Characteristics

- Aimed at the real needs of real bioinformaticians
  - Though application agnostic – no biology.
  - Workbench – expert bioinformaticians and WF developers
  - Web browser toolkit – bio users
  - Client application and service toolkits – developers
- Application Dataflow over Services
  - “Processors/Activities”: Web services, REST services, Beanshell scripts, R scripts, java apps, perl scripts...
  - Data type agnostic.
- “Come as you are” philosophy
  - Little obligation on 3rd party service providers.
  - Shim services handle inter-service mismatches.
- Open Architecture
  - Extensibility points throughout enactor, workbench and toolsets
- Shield users from service execution complexity yet enable them to control their own data flow.

Taverna Workflow Workbench



e-Services and Service Providers in the CLOUD

- Independent third party world-wide service providers of applications, tools and data sets. In the Cloud.
  - 850 databases, 166 web servers Nucleic Acids Research Jan 2006
- My local applications, tools and datasets. In the Enterprise. In the laboratory.
- Easily incorporate new service without coding. So even more services from the cloud and enterprise.

- 3500+ service operations
- All major providers
- Integration application for service providers like BioMOBY and BioMART

e-Scientists in the CLOUD

- Individual life scientists, in under-resourced labs, who use other people's applications, with little systems support.
  - Exploratory workflows
  - Developers (often) the users.
  - Consumers are providers.

- A distributed, disconnected community of scientists.
- Decoupled suppliers and consumers of services and workflows.
- Scientists in an enterprise and in large projects
- Scientists out of the enterprise, in small projects or sole traders.

200+ projects and sites, ~1000 individual users.
Users throughout UK, USA, Europe, and SE Asia

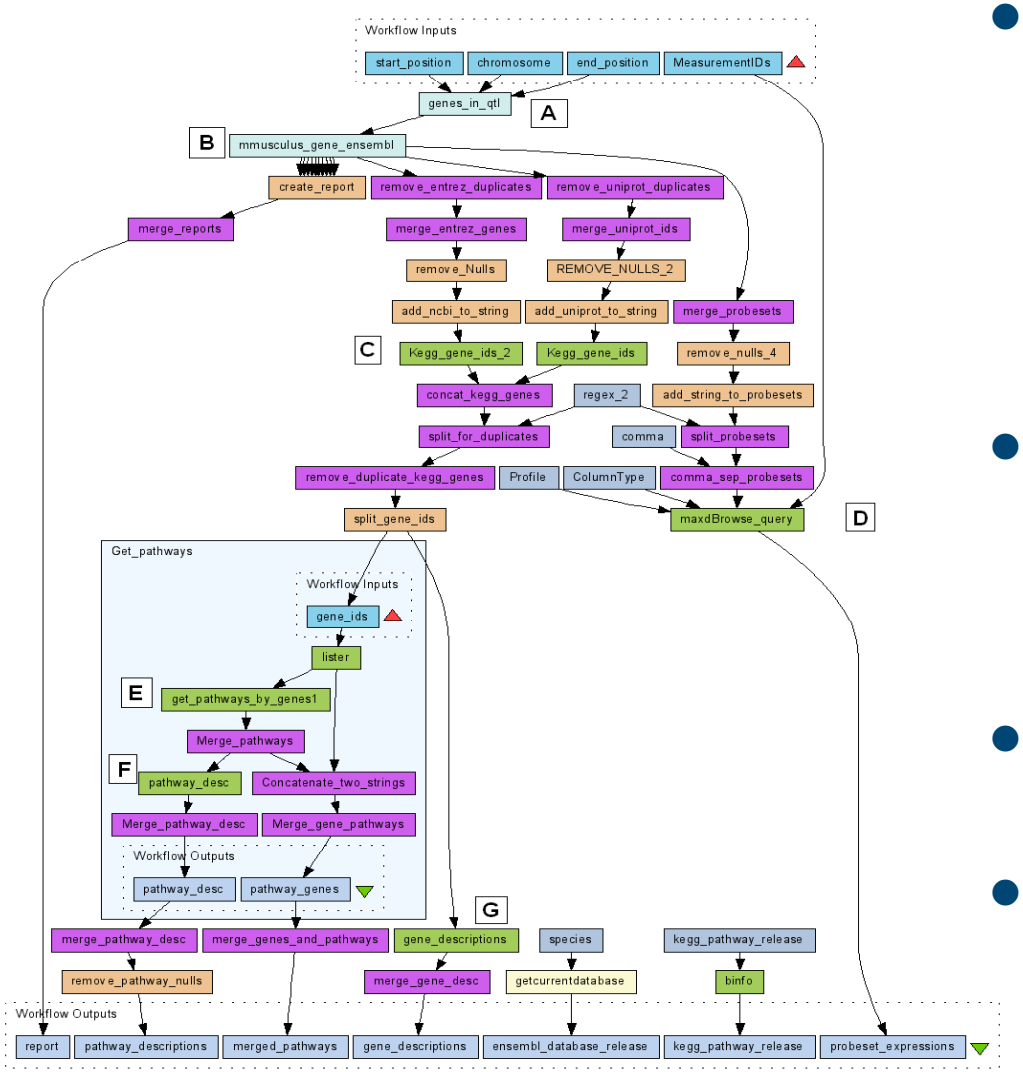
Example

The Wellcome Trust Host-Pathogen Project

Trypanosomiasis in African Cattle

Multi-disciplinary, Multi-site, Many people

http://www.genomics.liv.ac.uk/tryps/trypsindex.html



- Identified a metabolic pathway for which its correlating gene (Daxx) is believed to play a role in trypanosomiasis resistance.
- Manual analysis on the microarray and QTL data failed to identify this gene as a candidate.
- Repetitive, unbiased analysis.
- Generated a hypothesis to be tested in the laboratory

Published online 20 August 2007

Nucleic Acids Research, 2007, Vol. 35, No. 16 **5625–5633**
doi:10.1093/nar/gkm623

A systematic strategy for large-scale analysis of genotype–phenotype correlations: identification of candidate genes involved in African trypanosomiasis

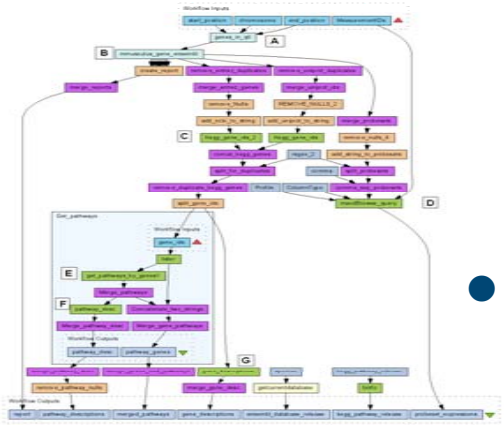
Paul Fisher[1,*], Cornelia Hedeler[1], Katherine Wolstencroft[1], Helen Hulme[1], Harry Noyes[2], Stephen Kemp[2], Robert Stevens[1] and Andrew Brass[1,3]

[1]School of Computer Science, Kilburn Building, University of Manchester, Oxford Road, Manchester, M13 9PL, [2]School of Biological Sciences, Biosciences Building, University of Liverpool, Crown Street, Liverpool, L69 7ZB and [3]Faculty of Life Science, Michael Smith Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

Received June 20, 2007; Revised and Accepted July 30, 2007

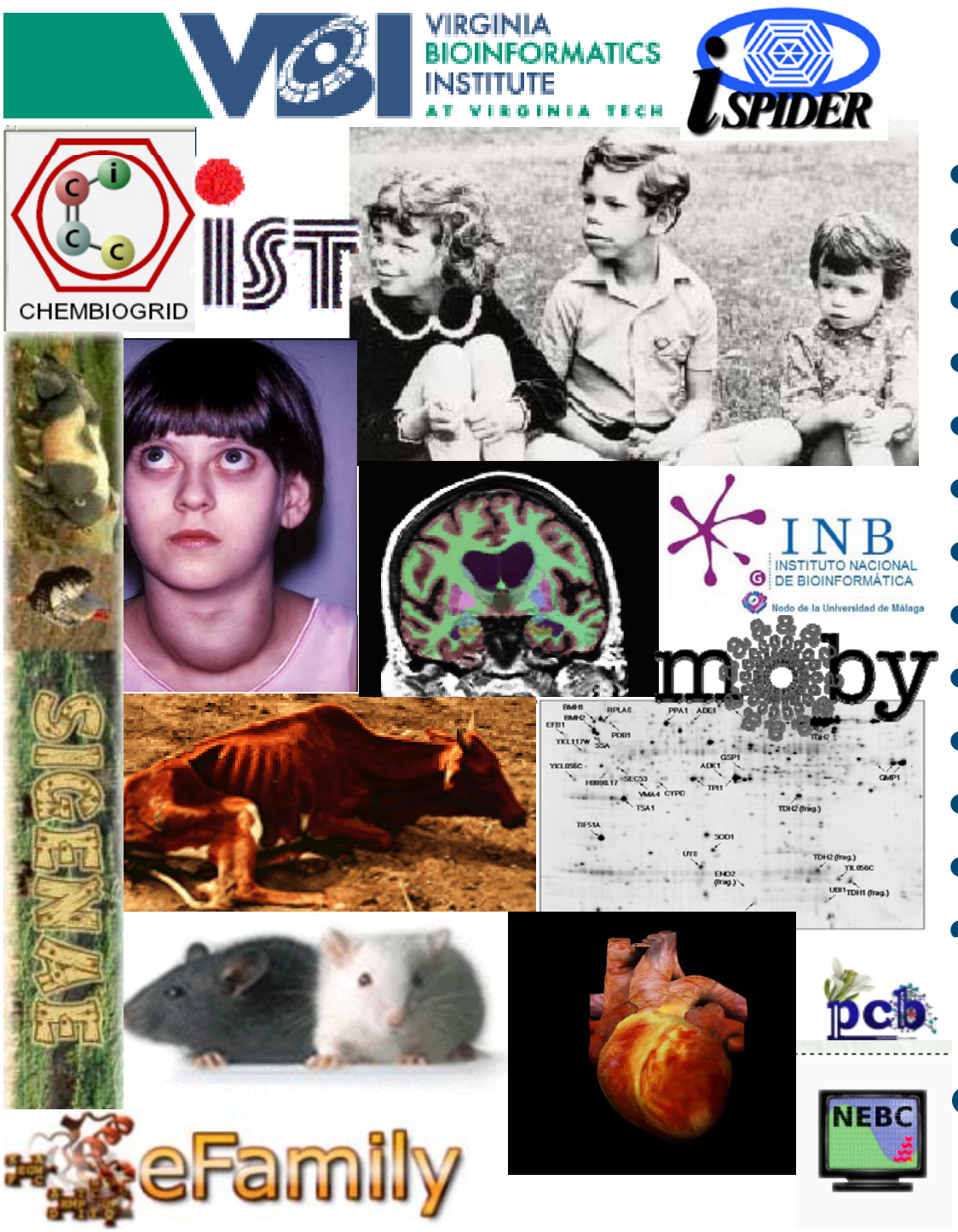
- High cholesterol in African cattle identified as a *protective* factor against death from trypanosomiasis

- Is high cholesterol a protective factor in humans undergoing extreme inflammation in intensive care?



Trichuris muris (mouse whipworm) infection

- Trypanosomiasis cattle workflow reused without change to identify the biological pathways involved in sex dependence in the mouse model, previously believed to be involved in the ability of mice to expel the whipworm parasite.
- Previously a manual two year study of candidate genes had failed to do this.
- Accelerated productivity
  - Workflow market.
  - Workflow pattern books.
  - Discovery factory.



- Systems biology
- Proteomics
- Gene/protein annotation
- Microarray data analysis
- Medical image analysis
- Heart simulations
- High throughput screening
- Phenotypical studies
- Phylogeny
- Plants,
- Mouse, Human
- Astronomy
- Music

- **Taverna 1.6.1**

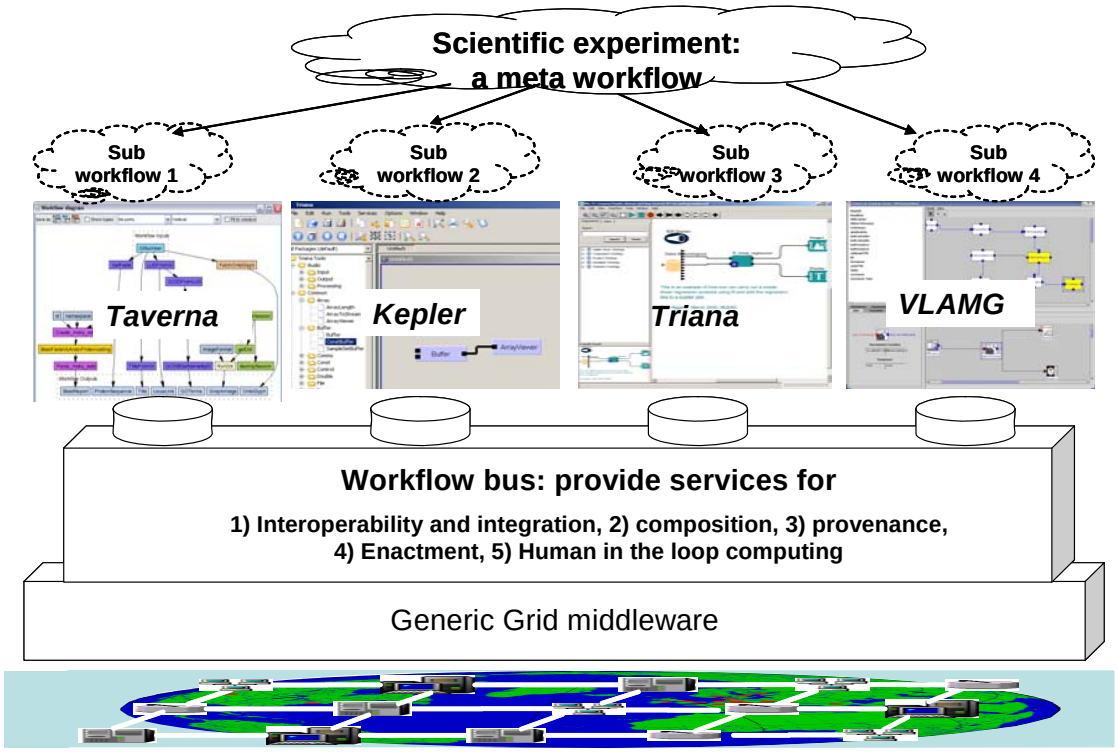
Colleagues

- The identification of a *mutation associated with the autoimmune disorder Graves' Disease* in the I kappa B-epsilon gene
- The study of Addison's Disease, an autoimmune disorder.
- Identification and classification of *proteins secreted by the anthrax bacterium*, Bacillus anthracis.
- Studies in the *metagenomics of freshwater microorganisms*
  - Pearce, Wipat, Newcastle, UK
- The *first complete and accurate map of the region of chromosome 7 involved in Williams-Beuren Syndrome* (WBS)
  - Tassabehji, St Mary's Hospital, Manchester, UK
- *Automatic target selection for protein structure and function studies*
  - EBI-Sanger Centre, UK
- Automatic reconstruction of *genome-scale yeast metabolic pathways* from distributed data in the life sciences to create and manipulate Systems Biology Markup Models.
  - Manchester Centre for Integrative Systems Biology, UK
- Data Warehouse pipelines
  - Proteomics - ISPIDER project, UK
  - Model organisms – eFungi Warehouse, UK

Strangers (and now Colleagues)

- Adopted by the National BioInformatics Centre (Netherlands), Genome Canada, BioLinux distribution
- National Cancer Research Centre, Genoa
- EU Projects: EMBRACE, Sealife, SIMDAT, CHASE
- Spike sorting raw EEG data for Neuro images, (CARMEN Project)
- Biotext mining workflows for the Adaptive Information Disclosure Application Toolbox, (VLe programme, The Netherlands)
- Chemoinformatics (Unilever Centre Cambridge, Indiana University)
- Crop research (Leibniz Institute of Plant Genetics and Crop Plant Research and a bunch of other plant people)
- Routine data integration (TaWeka (Edinburgh), DGEMap (Edinburgh), University of Georgia, everywhere)
- Genome annotation pipelines (Bergen Center for Computational Science, everywhere)
- Parameter sweeps across heart simulations (Oxford). Image processing (Manchester)
- Commercially: Drug design simulations (Chimatica); Data integration (BioTeam); Genome annotation pipelines (KooPrime)
- Teaching – widely.
- Music (University of Bath), Astronomy (AstroGrid), Multimedia (CHASE)

Tools and Platform developers



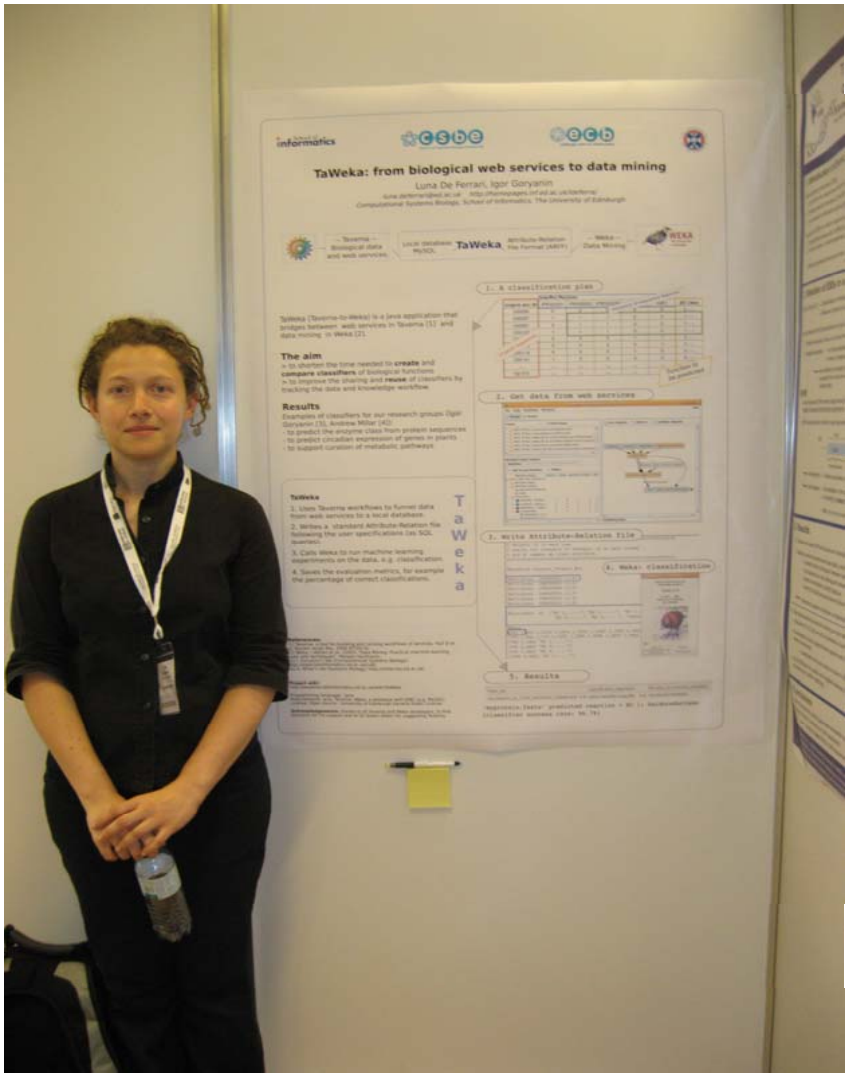
Z. Zhao et al., "Workflow bus for e-Science", in IEEE e-Science 2006, Amsterdam

vl·e virtual laboratory for e·science

- VL-e workflow bus
- VBI Pathport Toolbus
- EBI Distributed Annotation Service
- Bergen interaction service
- BioMOBY
- BioMART
- Utopia tool suite for interactive visualisation and data analysis
- BioLinux Distribution
- KOOPlatform

Current talks:

- Pegasus /DAGMan
- caGRID
- VisTrails



Bioinformatics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS

Institution: Institute for Development Policy and management, University of Manchester Sign In as Personal Su

Oxford Journals > Life Sciences > Bioinformatics > Volume 22, Number 10 > Pp. 1260-1261

XQTav: an XQuery processor for Taverna environment

Jacek Sroka [1,*], Grzegorz Kaczor [1], Jerzy Tyszkiewicz [1] and Andrzej M. Kierzek [2,*]

[1] Warsaw University Warsaw, Poland
[2] SBMS, University of Surrey Guildford, GU2 7XH, UK

*To whom correspondence should be addressed.

on of bioinformatic
ning popularity fast,
yet robust graphical

open middleware infrastructure institute uk

myGrid

MANCHESTER 1824
The University of Manchester

EPSRC

38,673 total sourceforge downloads (30 Sept 2007).

~1,500 per month.

Ranked 210 sourceforge activity (06 June 07).

Stop

http://carlin.myexperiment.org/profile/show/6

mozilla.org | mozillaZine | mozdev.org | myBibSonomy | http://www.aloud.c... | postBookmark | postPublication | post to del.icio.us | my del.icio.us

Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

myexperiment

Inbox | Logged in as: Carole Goble

Logout

My | People | Projects | Workflows | About

Workflows | Search

Duncan Hull

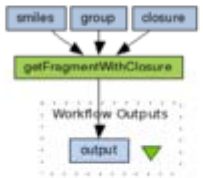
I'm one of Dougs Thugs and a user advocate for myExperiment.

Actions

Send Message

Remove friend

Workflows



Sets

Coming soon

Contacts

June Finch.

Giovanni Dall'Olio

Roy Goodacre

Use of the myExperiment beta is subject to the terms and conditions and the understanding that bugs may still exist and you may lose data. Report a bug

myexperiment

Inbox | Logged in as: Carole Goble

Logout

My | People | Projects | Workflows | About

Upload a new workflow | Bulk Upload (ZIP) | Download

Workflows | Search

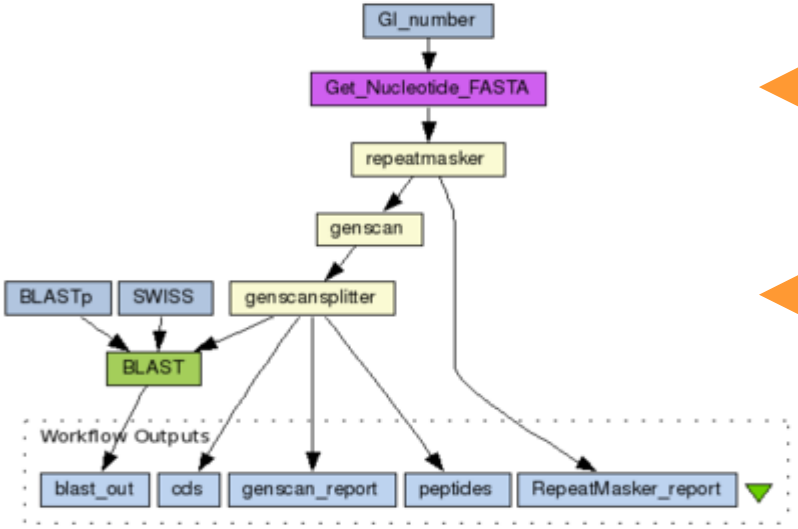
Bookmark this workflow

Collaborative, Social Bookmarking

Genome annotation pipeline demonstrator workflow for Nucleic Acids Research

Part of a workflow by Hannah Tipney, adapted by Duncan Hull using GenScan, RepeatMasker and BLAST. http://dx.doi.org/10.1093/nar/gkl320

Uploaded on 20 July, 2007 by Duncan Hull



Content Sharing

Application Execution

Social Recommendations

80% (1 rating)

Download this workflow

Tags

- BLAST
- genscan
- repeatmasker
- genome annotation
- nar

Collaborative, Social Tagging

Comments

http://kosh.ecs.soton.ac.uk:3000/jobs/new/8

Google

myexperiment

Inbox | Logged in as: Don Cruickshank

Logout

My | People | Projects | Workflows | About

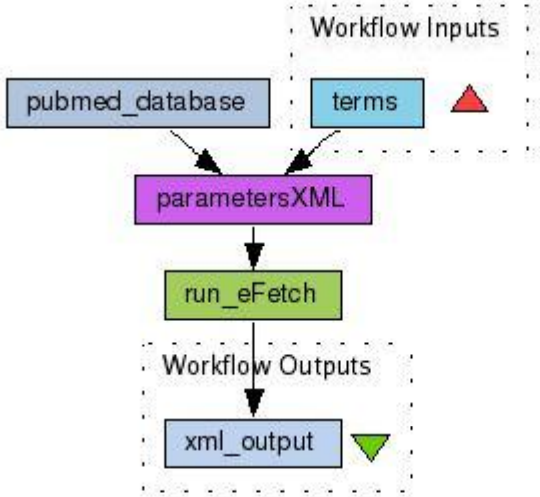
List jobs

Workflows | Search

Create job for workflow "Perform a text based search through PubMed"

This workflow takes in a number of search terms in which to perform a search over the PubMed literature database. These search terms may be input as if entered in the web based version of PubMed. The output from this workflow is a list of PubMed identifiers in xml based format

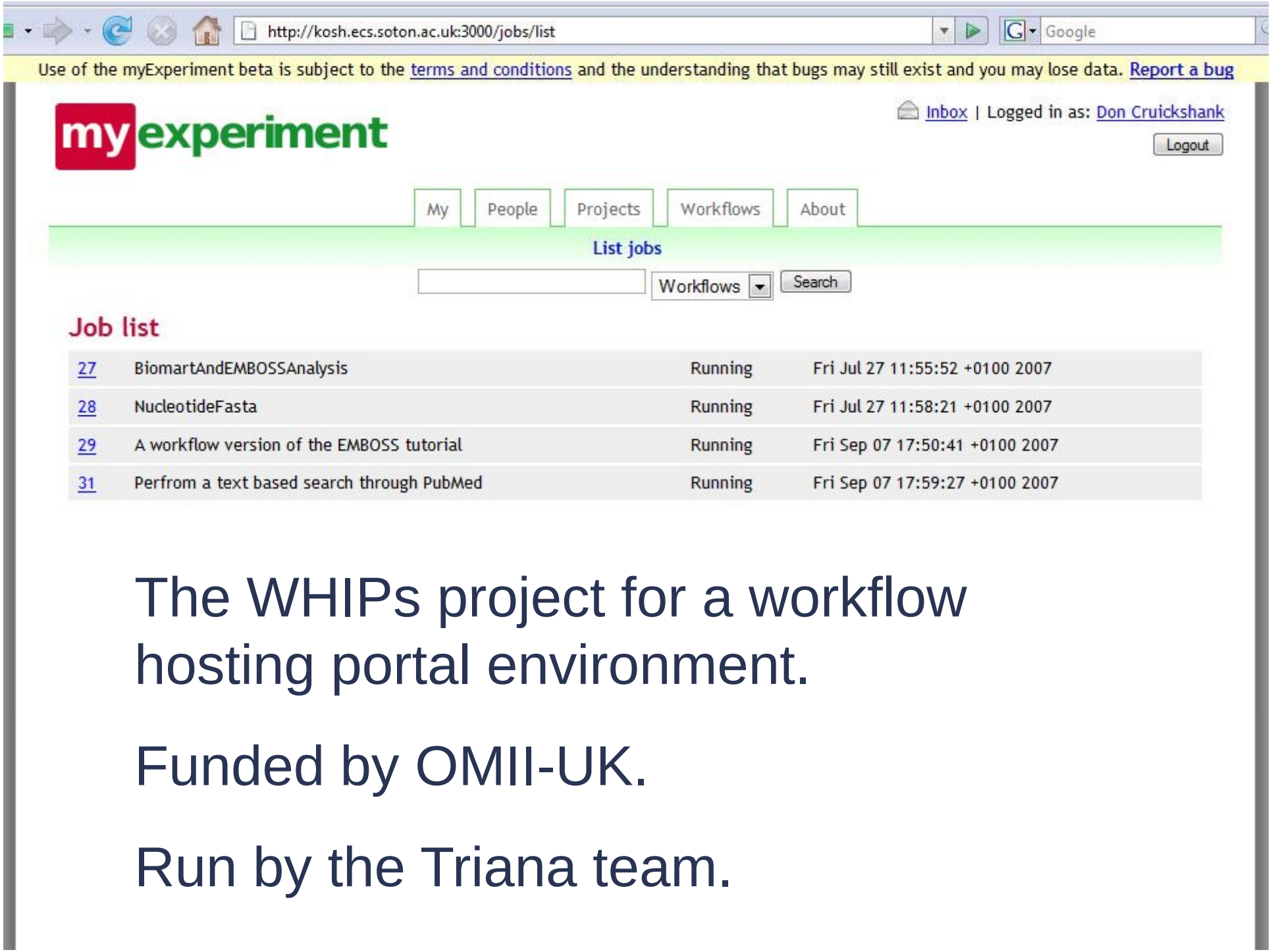
Uploaded on 07 September, 2007 by Don Cruickshank



Input data

terms: nucleic translation

Submit job

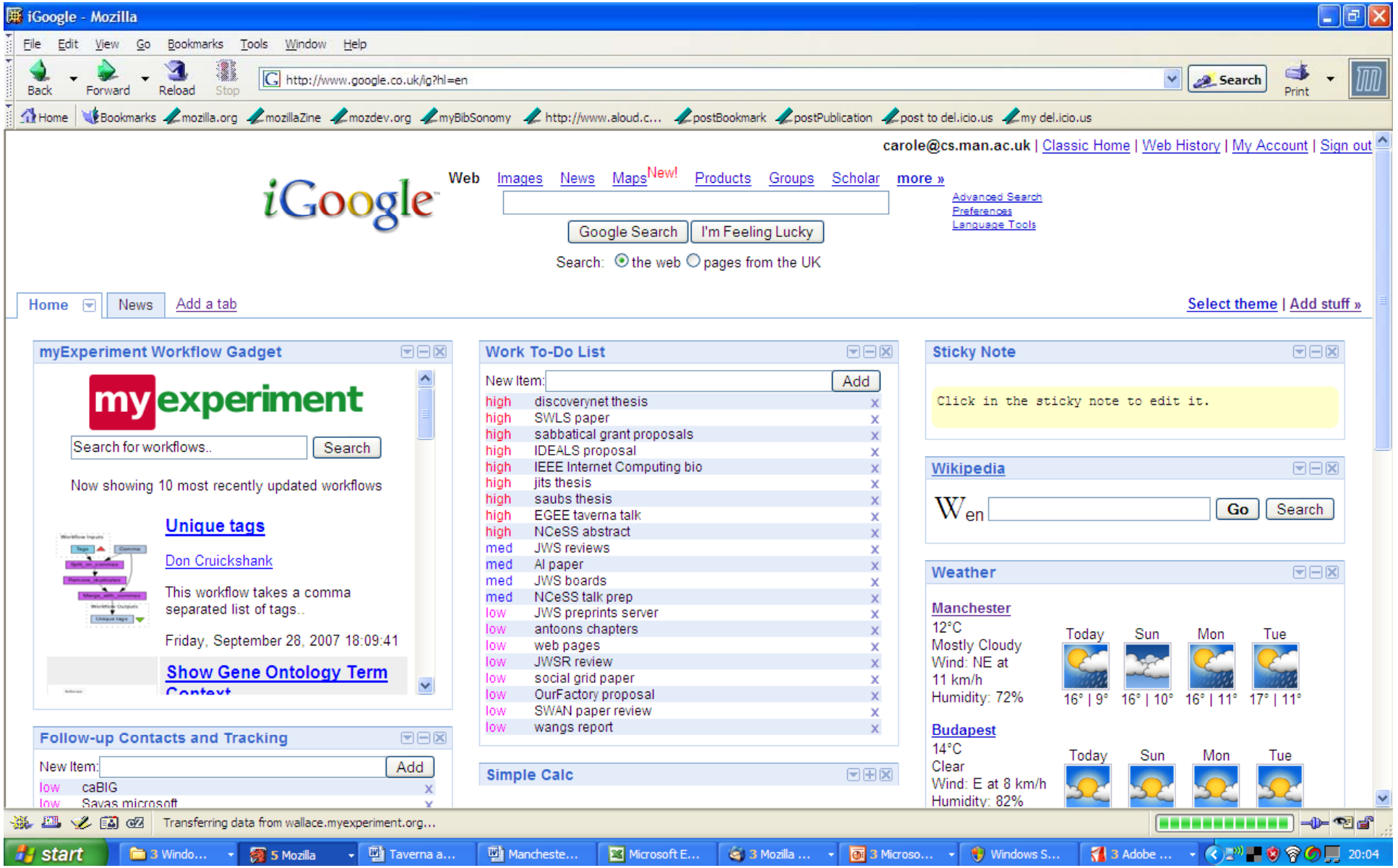


The WHIPs project for a workflow hosting portal environment.

Funded by OMII-UK.

Run by the Triana team.

Google Gadgets



More on the Components

Components

- Service Provider Components
  - Processor/Activity plug-ins, Java API consumer
  - Soaplab wrapper toolkit. **Soaplab2 shortly in beta.**
  - **BioNanny service monitoring service.**
- Management tools suite
  - RAVEN update management on the fly class loading, customised GUIs.
- Taverna as a service
  - Remote execution of Taverna outside workbench
  - **Taverna server**
  - **Improved client support for application and tool developers.**
  - **Workflows as content syndication for Web 2.0 browser mash ups**

Information and Cataloguing Services

- Feta Semantic Information service
  - Discovery over Service registry and Workflow repository.
  - OWL Ontology 710 classes. Full time curator.
  - WSDL (and WADL) Scavengers.
  - Smoother integration into Taverna 2
  - Absorbed into myExperiment for out-of-Taverna use.

- Semantic Curation Suite
  - Service provider curation and mass tagging
  - Expanded automated tagging

Data and Metadata

LogBook Provenance collection and access

- Ontology 35 classes. Life Science Identifier scheme
- Default BAKLAVA Store. LogBook Web Service
- **Complete overhaul for Taverna 2**

Data Management

- Default LogBook
- Your Store / File Management system / Database
- **More direct support in Taverna 2**
- **Better packaging with popular data solutions.**

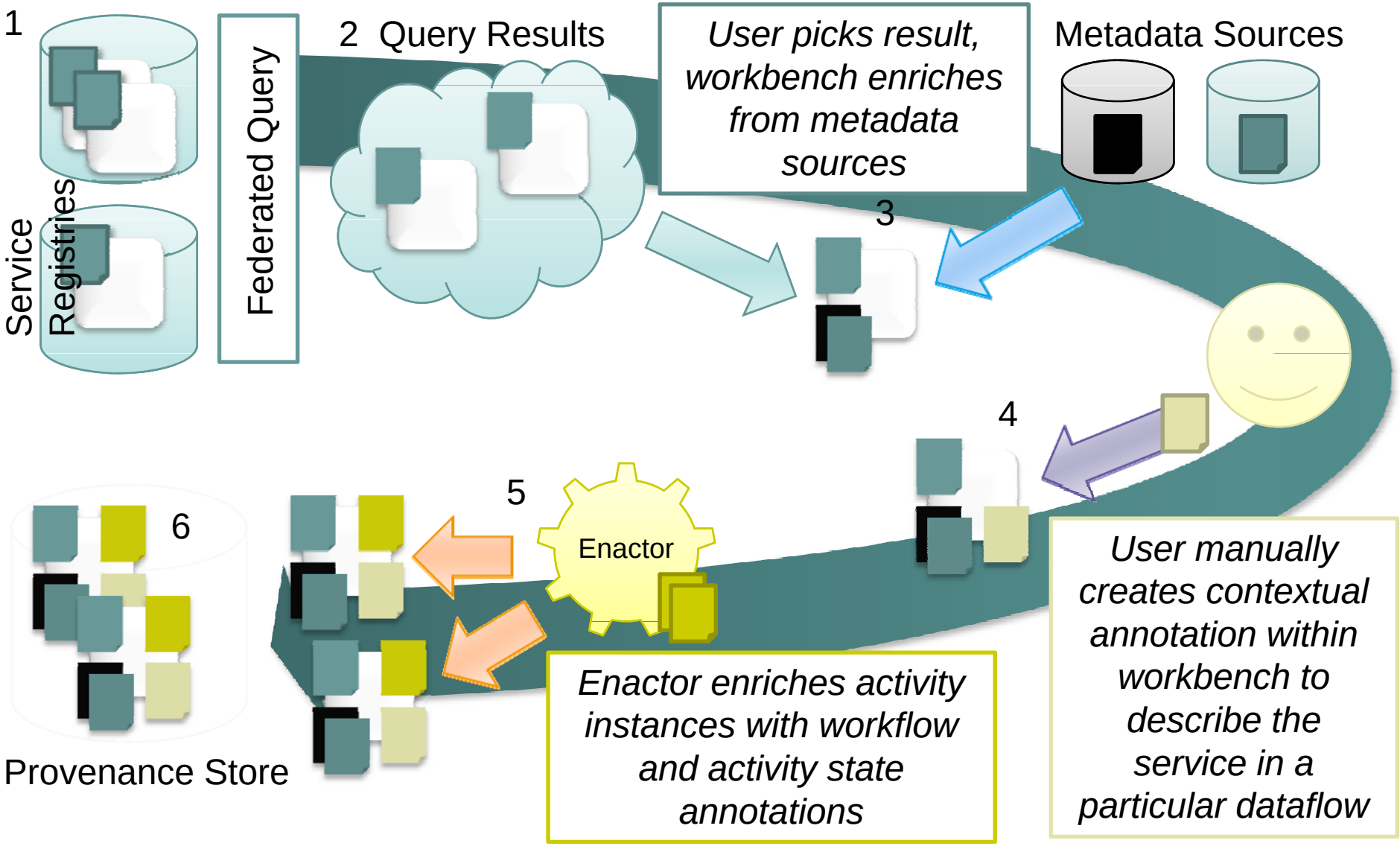
Extensible Annotation Objects

- A la S-OGSA / SOKU framework

Extensible Annotation Objects

- Workflow entities (activities, dataflows, edges, input and output ports etc.) are described by annotation objects
- Annotation objects are extensible
  - Mapping to ontology, free text, MIME types etc...
- Annotations on activities are transferred to descriptions of the inputs and outputs of those activities
- Both enactment (log, audit) and data (lineage, semantic network construction) provenance are captured by the enactor
- Provenance capture uses metadata from annotation objects
- S-OGSA / SOKU idea

Progressive enrichment of Extensible Annotation Objects



More on the Taverna Enactor

Abstraction

Extensibility points

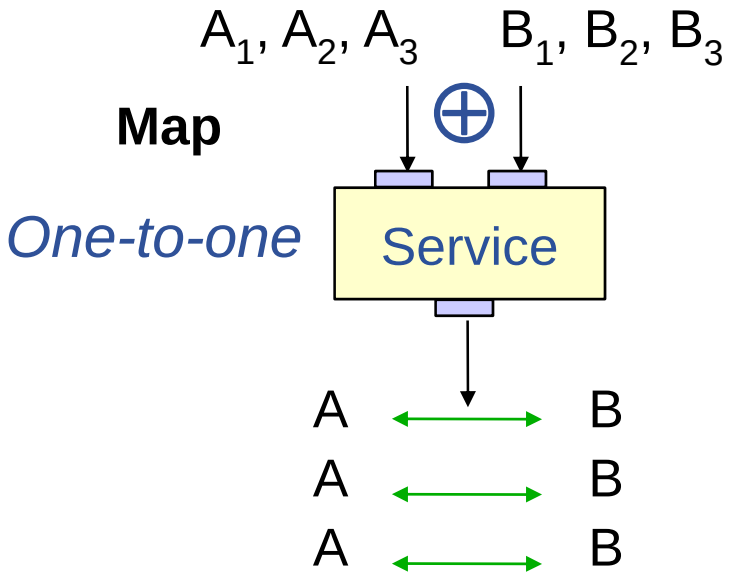
Taverna 2

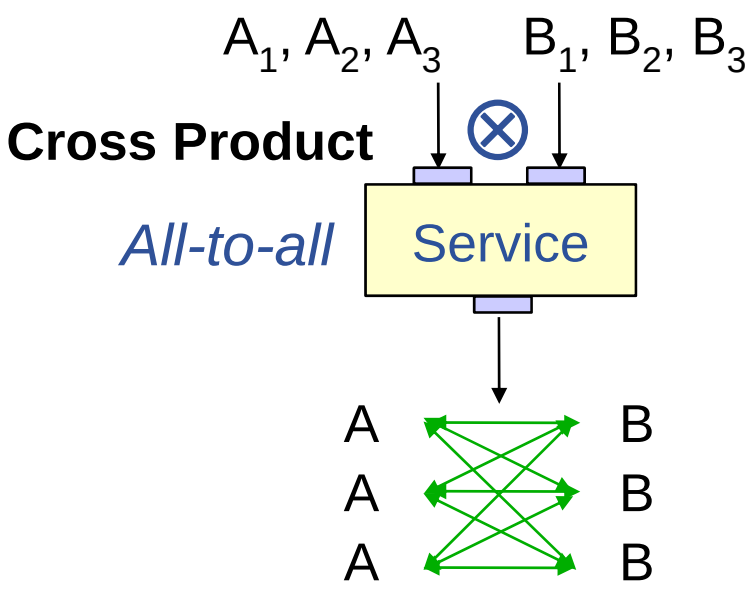
Philosophy

- Don’t mix science and mechanics
  - In a lab you don’t document getting the reagent from the storage cupboard. So don’t expose stuff like data staging or service invocation protocols. Or other plumbing.
  - Let plumbing problem be scientific or domain specific plumbing problems
  - Shield your users as much as possible.
  - Expose the scientific protocol
- Keep the barriers to adoption for your service provider low too.

What IS the enactor?

- The computational lambda calculus obtained by augmenting lambda calculus with a list operator monad
- Collection management. Implicit iteration. Nesting.
- Data composition patterns. Data manager.



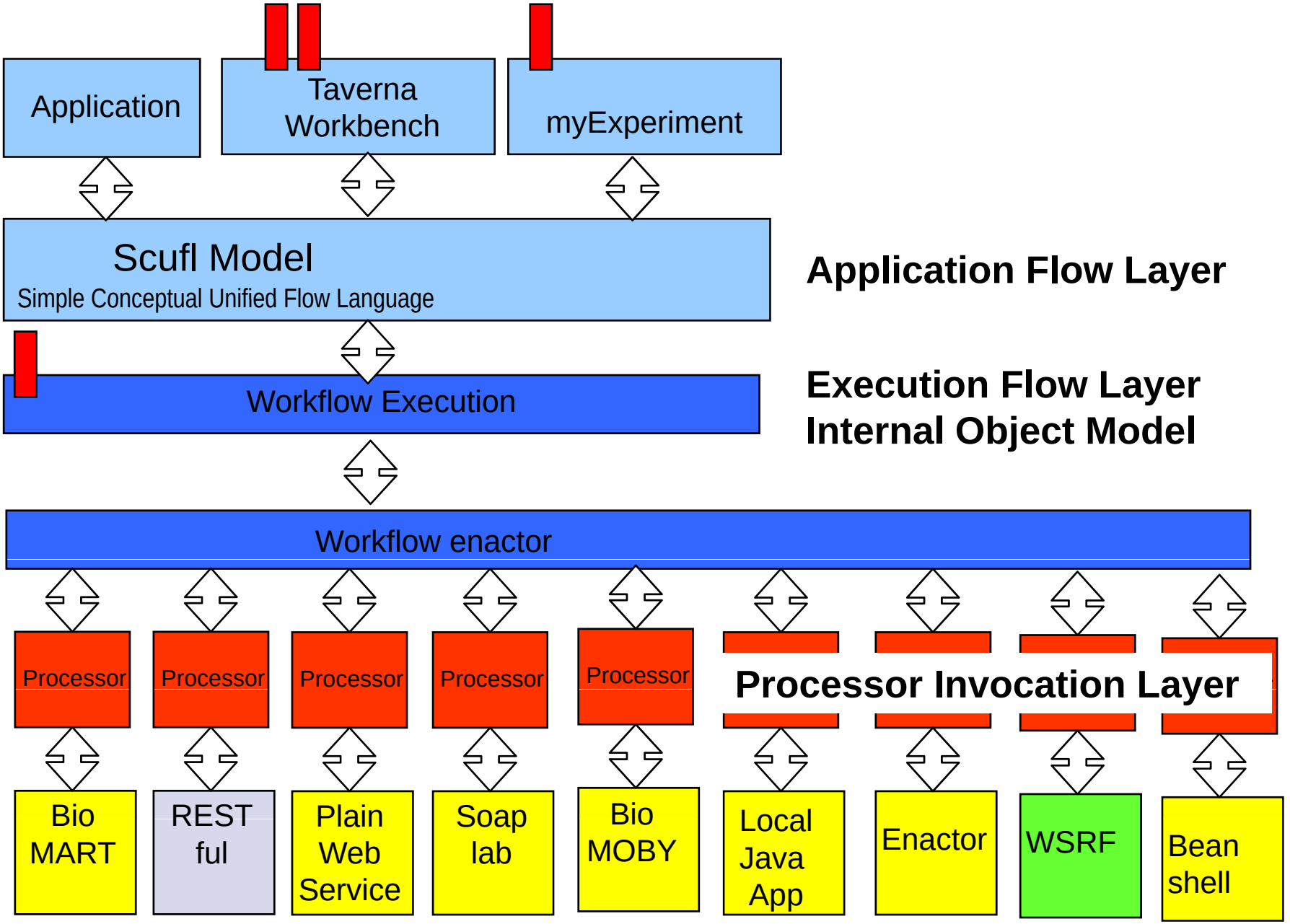


With thanks to Vangelis Floros,

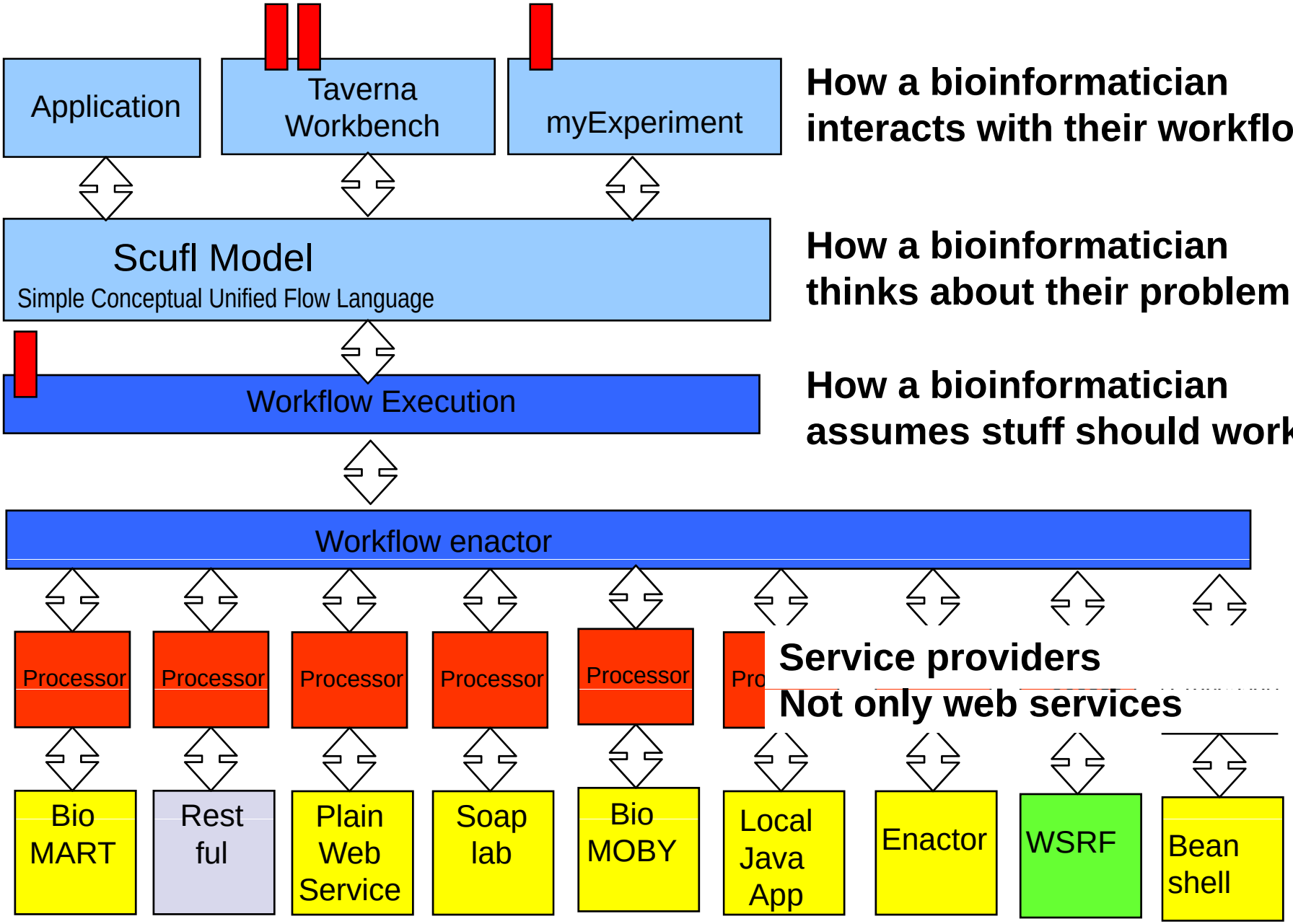
Oinn, T., et al Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice & Experience* **18, 1067-1100 (2006).**

Turi, Missier, De Roure, Goble and Oinn. *Taverna Workflows: Syntax and Semantics*. To appear in Proc 3rd IEEE Intl Conf e-Science and Grid Computing 2007

Extensibility points for application developers



Extensibility points for application developers

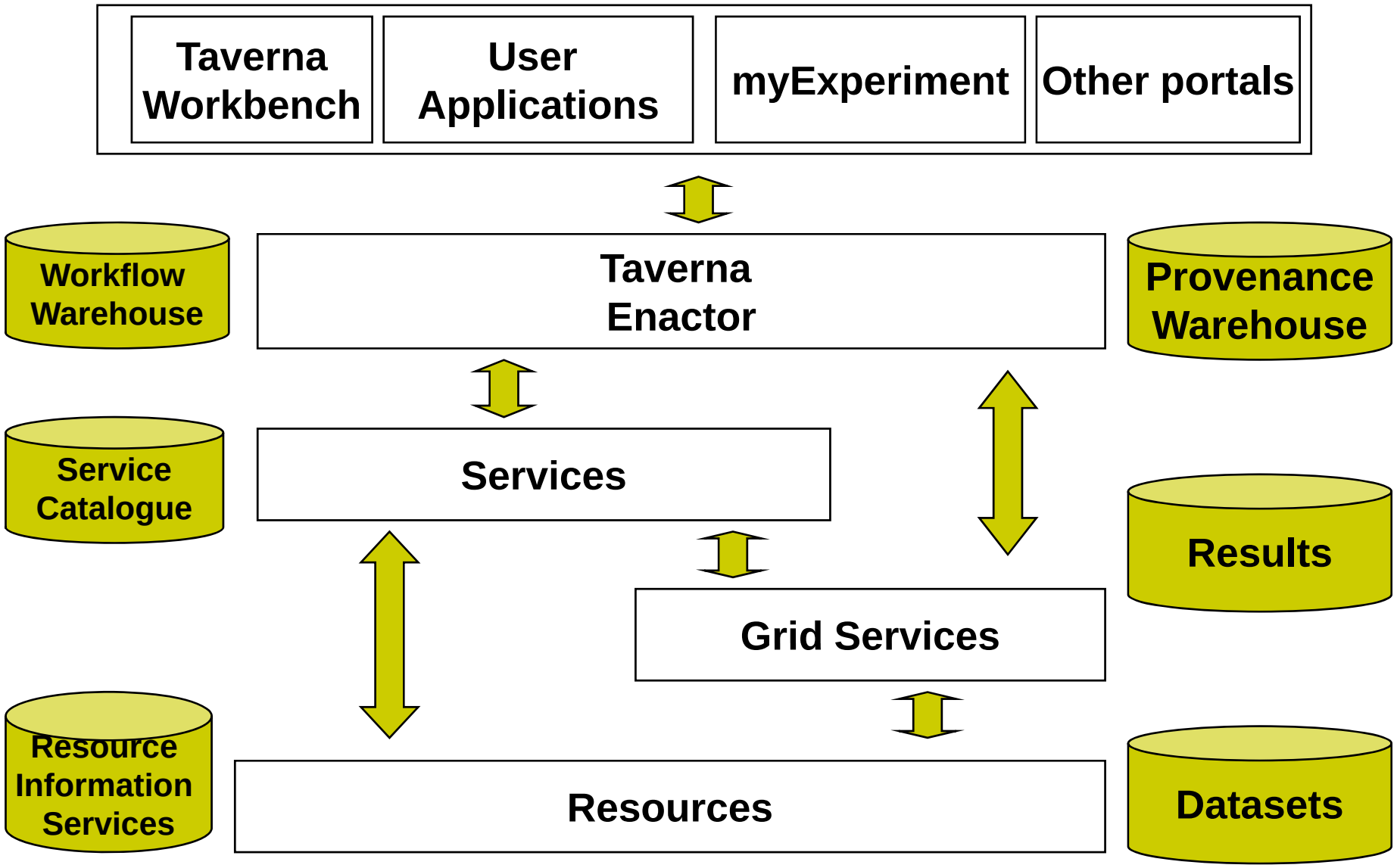


How a bioinformatician interacts with their workflow

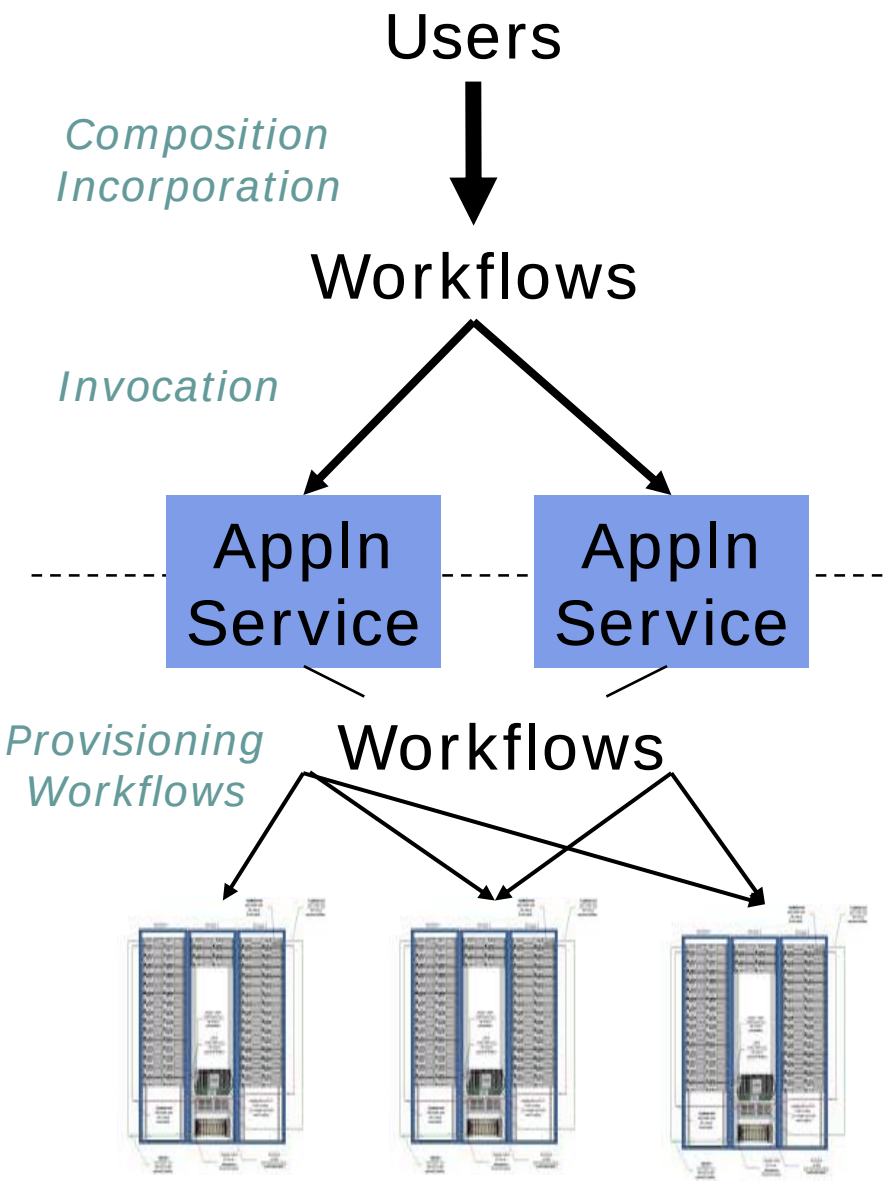
How a bioinformatician thinks about their problem

How a bioinformatician assumes stuff should work

Service providers
Not only web services


Taverna Workbench
User Applications
myExperiment
Other portals
Workflow Warehouse
Taverna Enactor
Provenance Warehouse
Service Catalogue
Services
Results
Grid Services
Resource Information Services
Resources
Datasets

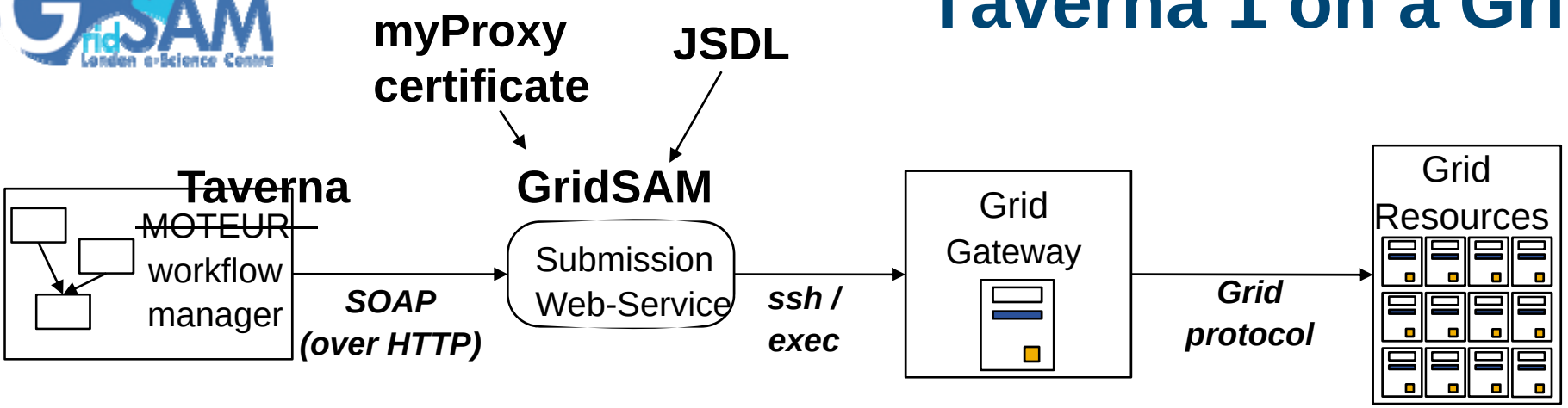
[Foster 2005]



- Service-oriented **applications**
  - Applications components of workflows
  - Compose applications into workflows
  - Incorporate workflows into applications

- Service-oriented **Grid Infrastructure**
  - Provision physical resources to support application workflows
  - Coordinate resources through workflows

Taverna 1 on a Grid



- Enacting services on a batch-oriented grid infrastructure
  - GridSAM submission web service
- From workflow manager to grid execution
  - Execution engine independent from grid middleware.
- Experiences
  - Chimatica, NWGrid, early pilots with EGEE1

With thanks to Vangelis Floros!

Taverna 2 in a nutshell beta January 2008

- Enactor rewrite with Extensibility points.
- Aim: Secure, long running workflows over large data sets.
- Enactor invocation extensions
  - Asynchronous processor and data streaming
  - Explicit monitoring and steering support
  - New dispatch layer better supports dynamic service binding and service invocation through a resource broker
  - Improved concurrency handling at the workflow level
- Data manager
  - Rewritten metadata annotation framework.
  - Implicit handling of data reference management and data movement - no explicit staging, minimised data shipping.
  - Support for heterogeneous data grid platforms.
- Security agent
  - Extension point to plug in your favourite AA scheme

Concurrency

- Four types
  - Parallelism in the service for job submissions over resources
  - Multiple parallel runs of the same workflow over the same data stream – spawning independent instances
  - Multiple parallel runs of the same workflow over different data sets or parameter sweeps – spawning independent enactors.
  - Multiple different workflows by different users through the same Taverna server – spawning independent enactors.

Linking Taverna 2 and EGEE 2?

- Linking with EGEE Information Services
  - To select services and resource intelligently.
  - To detect and response services (jobs) failures and running environment changes
- Cooperation with resource brokers
  - To support concurrent dynamic resource changes
- Take advantage of data services
  - EGEE's Data Catalog service, Replica service
  - Taverna's service tagging and annotation framework.
- Plug-in to the EGEE security infrastructure.

Linking Taverna 2 and EGEE 2?

- Dynamic VO
  - gLite needs to support dynamic VOs that only live during a workflow's lifetime.
- Good service/software discovery and running environment selection
  - To enable Taverna to make a workflow execution plan.
- Resource reservation
  - Workflow execution may need to guarantee (some) resources
- Resource failure support
  - Intelligent and automatic job re-submission on failure.
  - Automatic services migration/ running environment configuration to new computing resources for workflow execution.
- Monitoring tools and information services to provide meaningful, up-to-date information about service discovery, workflow execution, input/outputs etc.

Summary

- Taverna has a significant application user following because its application oriented.
  - Active, ongoing and expanding.
- myExperiment is the place we bring them, and other workflow users, together
- Taverna 2 gives us new opportunities to have better Grid integration.
- We could benefit from the gLite data management and security services.
- But it needs work and resources.

Acknowledgements

- Tom Oinn, Wei Xing, Katy Wolstencroft, David De Roure

- The myExperiment team.
- The myGrid team.
- The OMII-UK team.
- All our users.
- EPSRC and JISC for the funding.

For more stuff

- http://www.mygrid.org.uk
- http://taverna.sourceforge.net
- http://www.myexperiment.org
- http://www.omii.ac.uk