Taverna and ^{my}Grid

Open Workflow for Life Sciences

Tom Oinn

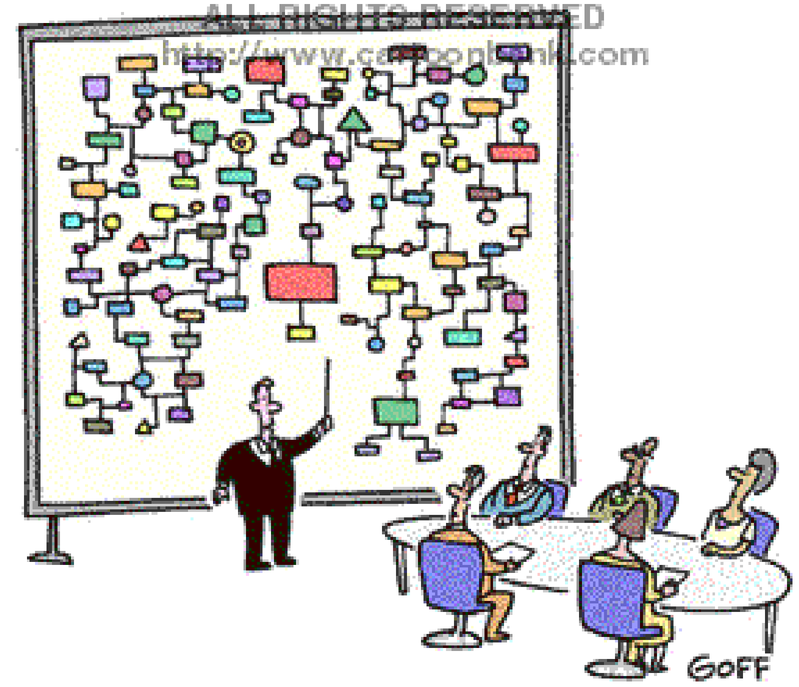
tmo@ebi.ac.uk

EPSRC

Engineering and Physical Sciences
Research Council

What, who, why?

- Taverna – a workflow development and enactment environment
- Who – part of myGrid, an EPSRC funded UK eScience Pilot project coordinated by Carole Goble at Manchester University
- Why – because bioinformatics is hard enough without turning users into web spiders ☺



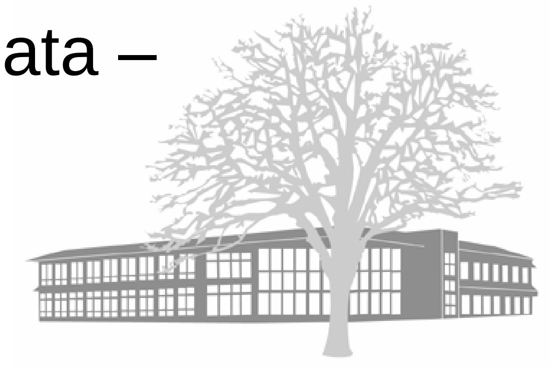
"And that's why we need a computer."

Old approach

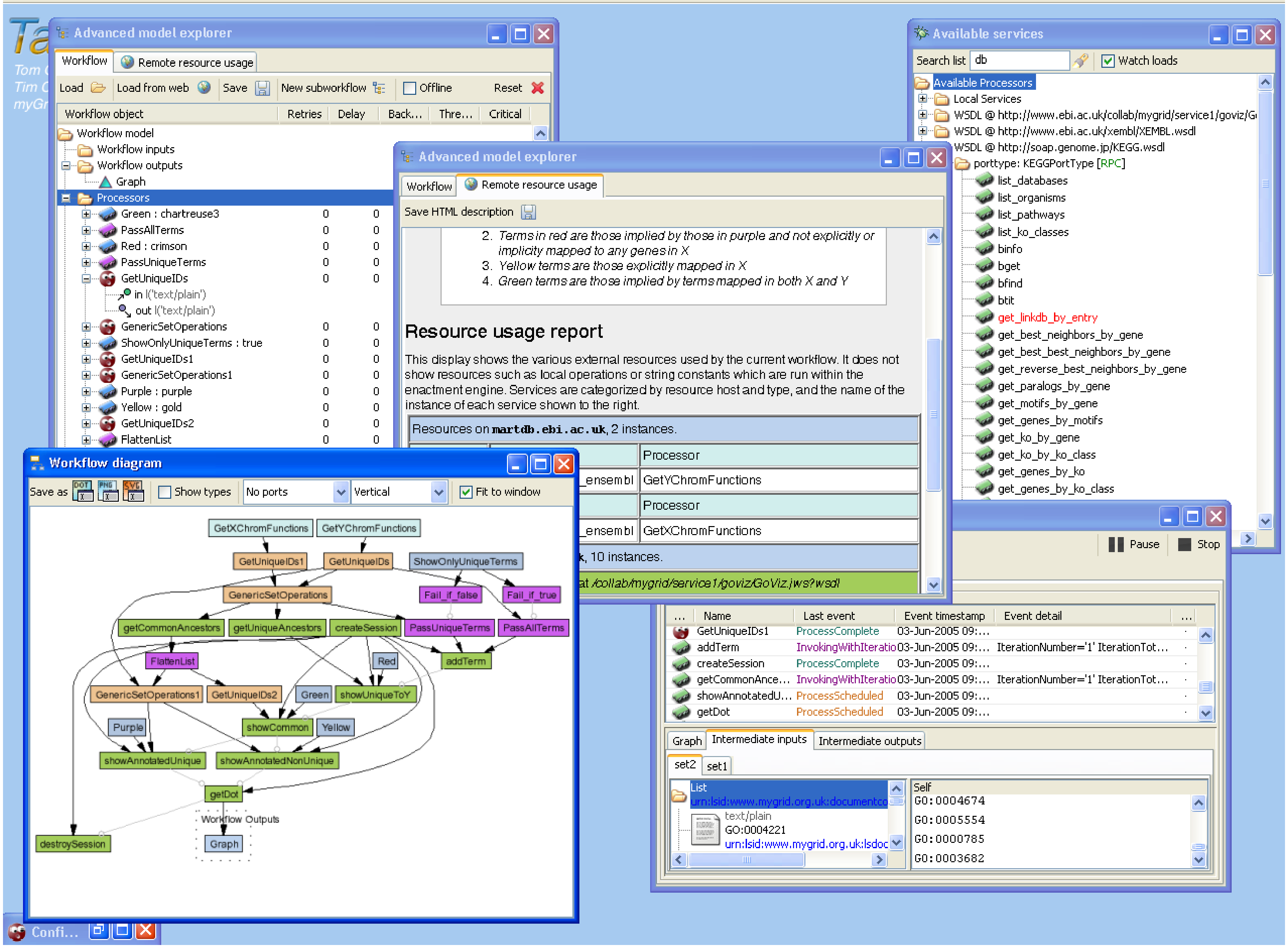
- Cut and paste, cgi, shell scripting, ftp, excel
- Time intensive
- Manual process, fails to scale sensibly
- Hard to document and reproduce
  - Good scientific discipline hard to maintain
- Boring, waste of highly trained scientists

Our approach

- Capture the scientific method as a formal process model
- Allow users to construct such models from libraries of available components in a graphical editing environment with semantic support
- Publish process definitions as scientific methods, enact and automatically scale to large data sets, multiple runs
- Automatically collect enactment metadata – workflow provenance.



Taverna Scufl Workbench v1.2
http://taverna.sf.net
Design
Enactment
Analysis
Collaboration
SoapLab
moby
iT
EBI
myGrid
Blueprint
djm,jf,mg,pl,ktg,mp,ms,tmo,mf,ek,pa



Workflow object	Retries	Delay	Back…	Thre…	Critical
Workflow model					
Workflow inputs					
Workflow outputs					
Graph					
Processors					
Green : chartreuse3	0	0			
PassAllTerms	0	0			
Red : crimson	0	0			
PassUniqueTerms	0	0			
GetUniqueIDs	0	0			
in l('text/plain')					
out l('text/plain')					
GenericSetOperations	0	0			
ShowOnlyUniqueTerms : true	0	0			
GetUniqueIDs1	0	0			
GenericSetOperations1	0	0			
Purple : purple	0	0			
Yellow : gold	0	0			
GetUniqueIDs2	0	0			
FlattenList	0	0			

Search list: db

Available Processors
Local Services
WSDL @ http://www.ebi.ac.uk/collab/mygrid/service1/goviz/G
WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
WSDL @ http://soap.genome.jp/KEGG.wsdl
porttype: KEGGPortType [RPC]
list_databases
list_organisms
list_pathways
list_ko_classes
binfo
bget
bfind
btit
get_linkdb_by_entry
get_best_neighbors_by_gene
get_best_best_neighbors_by_gene
get_reverse_best_neighbors_by_gene
get_paralogs_by_gene
get_motifs_by_gene
get_genes_by_motifs
get_ko_by_gene
get_ko_by_ko_class
get_genes_by_ko
get_genes_by_ko_class

Save HTML description

2. *Terms in red are those implied by those in purple and not explicitly or implicity mapped to any genes in X*
3. *Yellow terms are those explicitly mapped in X*
4. *Green terms are those implied by terms mapped in both X and Y*

Resource usage report

This display shows the various external resources used by the current workflow. It does not show resources such as local operations or string constants which are run within the enactment engine. Services are categorized by resource host and type, and the name of the instance of each service shown to the right.

Resources on **martdb.ebi.ac.uk**, 2 instances.

	Processor
_ensembl	GetYChromFunctions
	Processor
_ensembl	GetXChromFunctions

k, 10 instances.

at /collab/mygrid/service1/goviz/GoViz.jws?wsdl

No ports | Vertical | ☑ Fit to window

GetXChromFunctions
GetYChromFunctions
GetUniqueIDs1
GetUniqueIDs
ShowOnlyUniqueTerms
GenericSetOperations
Fail_if_false
Fail_if_true
getCommonAncestors
getUniqueAncestors
createSession
PassUniqueTerms
PassAllTerms
FlattenList
Red
addTerm
GenericSetOperations1
GetUniqueIDs2
Green
showUniqueToY
Purple
showCommon
Yellow
showAnnotatedUnique
showAnnotatedNonUnique
getDot
Workflow Outputs
destroySession
Graph

…	Name	Last event	Event timestamp	Event detail
	GetUniqueIDs1	ProcessComplete	03-Jun-2005 09:…	
	addTerm	InvokingWithIteratio	03-Jun-2005 09:…	IterationNumber='1' IterationTot…
	createSession	ProcessComplete	03-Jun-2005 09:…	
	getCommonAnce…	InvokingWithIteratio	03-Jun-2005 09:…	IterationNumber='1' IterationTot…
	showAnnotatedU…	ProcessScheduled	03-Jun-2005 09:…	
	getDot	ProcessScheduled	03-Jun-2005 09:…	

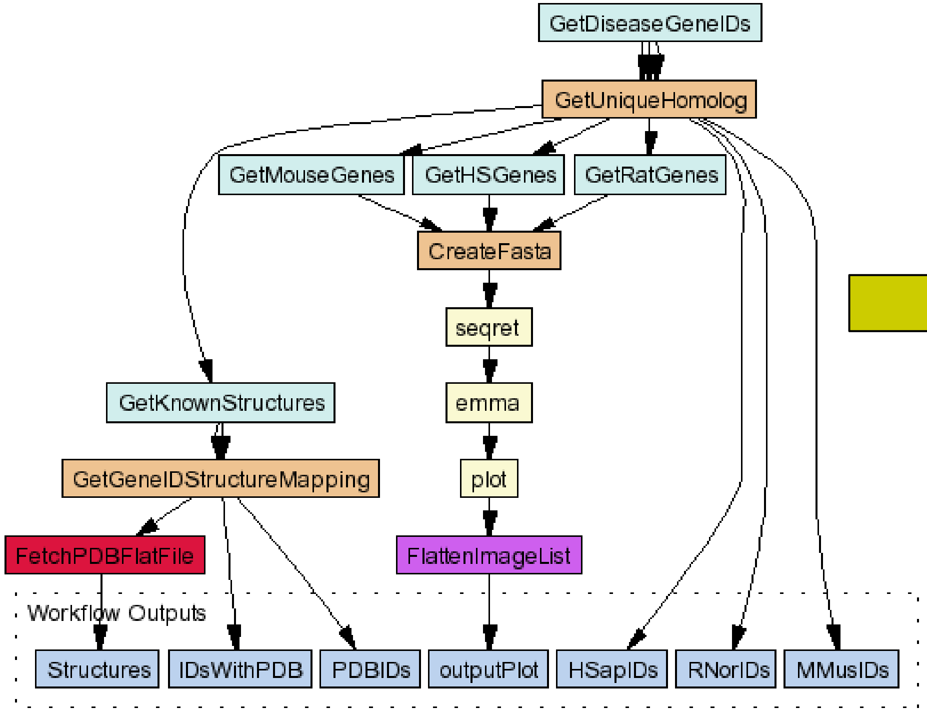
List
urn:lsid:www.mygrid.org.uk:documentco
text/plain
GO:0004221
urn:lsid:www.mygrid.org.uk:lsdoc

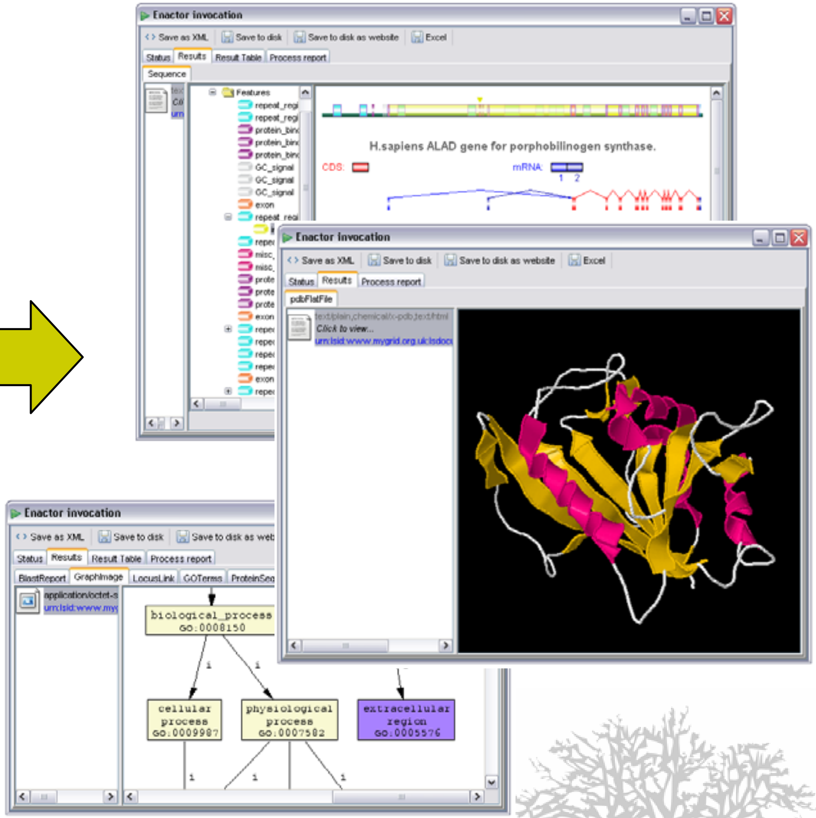
Self
GO:0004674
GO:0005554
GO:0000785
GO:0003682

What can we integrate?

- Web services defined by WSDL
  - *Pathport, BIND, Gene Ontology, DBFetch, FASTA, InterproScan, NCBI eUtils…*
- Complex analysis services conforming to Life Science Analysis Engine (LSAE) specification
  - *EMBOSS, Jess, any arbitrary legacy C, PERL or Shell script*
- BioMoby services (www.biomoby.org)
  - *PlaNeT, IRI, Spanish Bioinformatics Network, Genome Prairie…*
- Biomart Database Queries
  - *Ensembl, DbSNP, VEGA…*
- Local embedded scripts via Java, Perl, Python, Ruby etc.
- Seqhound Genomic data warehouse
  - *Genbank, LocusLink, GO*
- Styx Grid Service
  - *Environmental eScience, ocean temperature analysis etc*
- Arbitrary 3rd Party APIs i.e. BioJava, JUMBO, caBIG

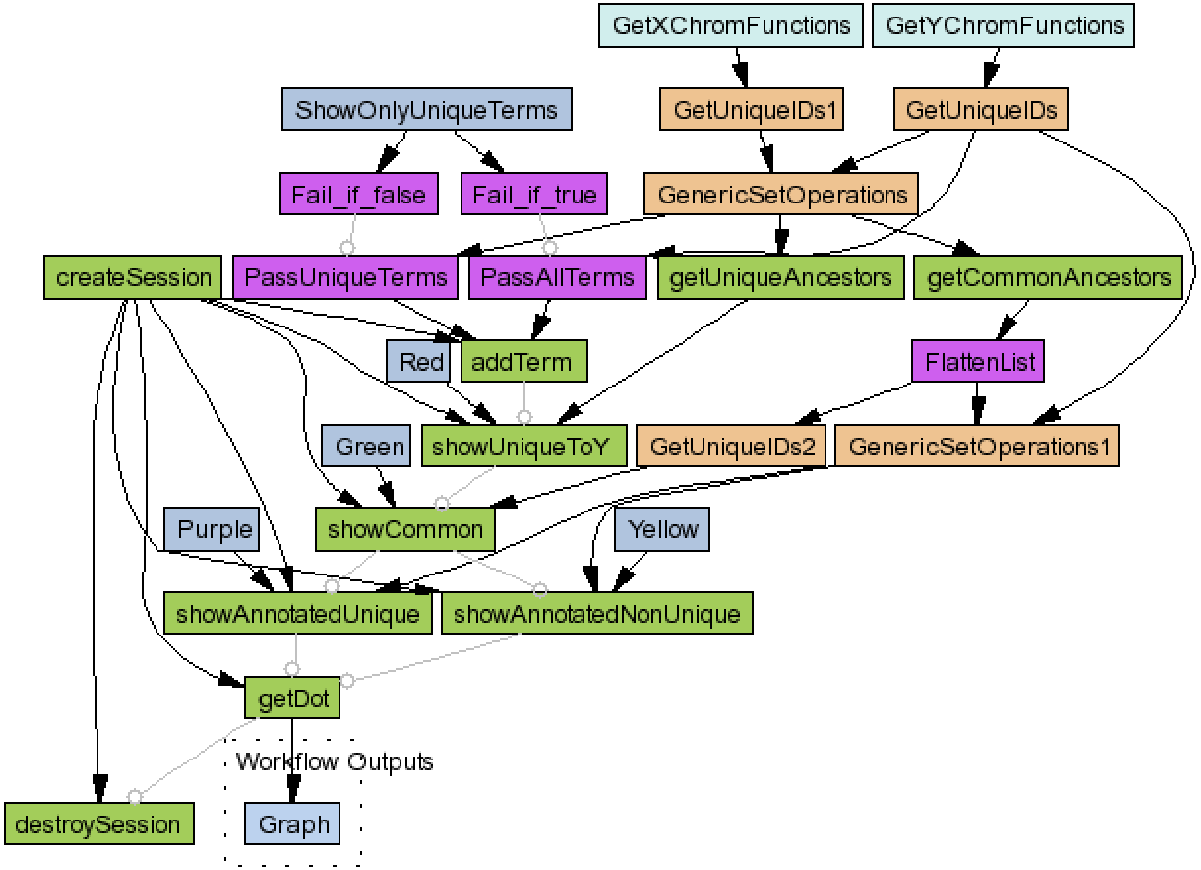
Comparative Genomics





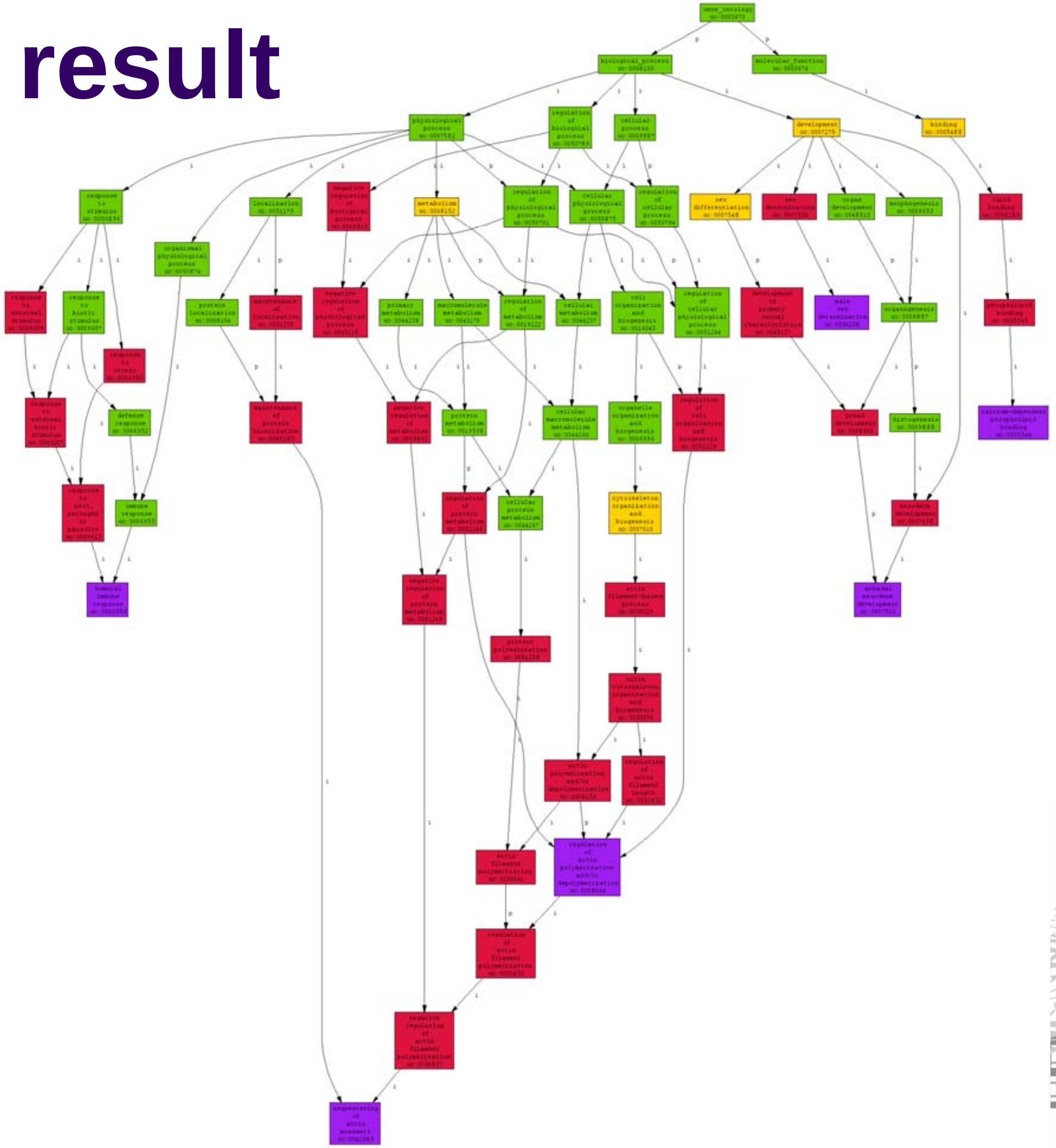
BiomartAndEMBOSSAnalysis.xml

Functional Analysis Workflow…



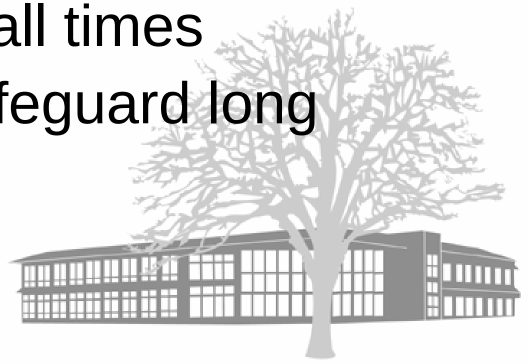
CompareXandYFunctions.xml

…and result

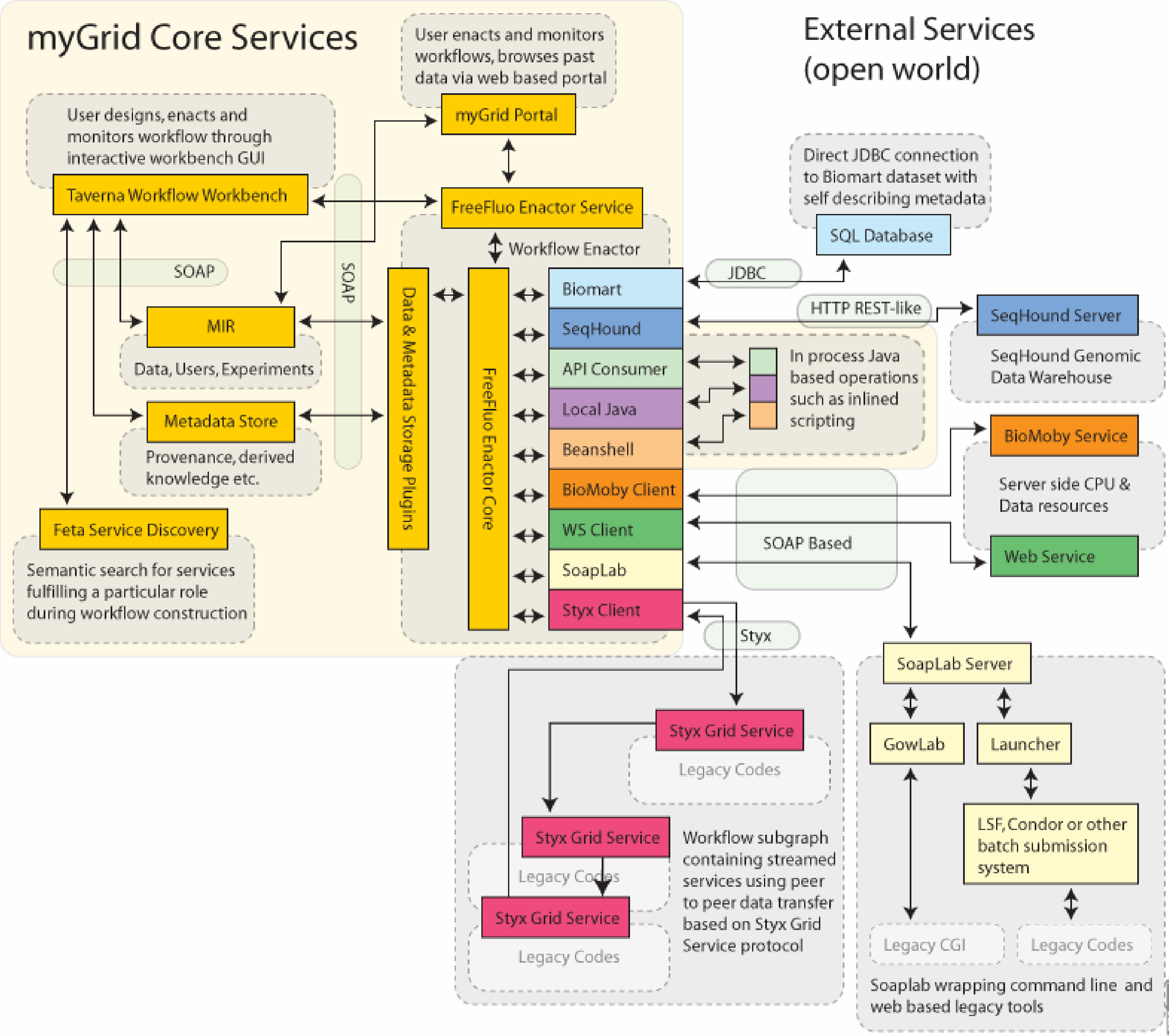

gene_ontology
biological_process
molecular_function
physiological process
regulation of biological process
cellular process
development
binding
response to stimulus
localization
negative regulation of biological process
metabolism
regulation of physiological process
cellular physiological process
regulation of cellular process
sex differentiation
sex determination
organ development
morphogenesis
lipid binding
organismal physiological process
response to external stimulus
response to biotic stimulus
protein localization
maintenance of localization
negative regulation of physiological process
primary metabolism
macromolecule metabolism
regulation of metabolism
cellular metabolism
cell organization and biogenesis
regulation of cellular physiological process
development of primary sexual characteristics
male sex determination
organogenesis
phospholipid binding
response to stress
response to external biotic stimulus
defense response
maintenance of protein localization
negative regulation of metabolism
protein metabolism
cellular macromolecule metabolism
organelle organization and biogenesis
regulation of cell organization and biogenesis
gonad development
histogenesis
calcium-dependent phospholipid binding
response to pest, pathogen or parasite
immune response
regulation of protein metabolism
cellular protein metabolism
cytoskeleton organization and biogenesis
mesoderm development
humoral immune response
negative regulation of protein metabolism
actin filament-based process
gonadal mesoderm development
protein polymerization
actin cytoskeleton organization and biogenesis
actin polymerization and/or depolymerization
regulation of actin filament length
actin filament polymerization
regulation of actin polymerization and/or depolymerization
regulation of actin filament polymerization
negative regulation of actin filament polymerization
sequestering of actin monomers
i
p

Philosophy

- Open world approach for services
  - Do not require service providers to change
  - Maximize interoperability
  - Extend on demand
    - Minimalist functional core, declarative language, many plugin extension points
- Open development approach as well
  - LGPL License
  - Transparent, public development process
  - CVS, Mailing lists, website are all public at all times
  - Avoid institutional ‘ownership’ of code to safeguard long term future development



Taverna network architecture diagram



myGrid Core Services

User enacts and monitors workflows, browses past data via web based portal

myGrid Portal

User designs, enacts and monitors workflow through interactive workbench GUI

Taverna Workflow Workbench

FreeFluo Enactor Service

Workflow Enactor

SOAP

SOAP

MIR

Data, Users, Experiments

Metadata Store

Provenance, derived knowledge etc.

Feta Service Discovery

Semantic search for services fulfilling a particular role during workflow construction

Data & Metadata Storage Plugins

FreeFluo Enactor Core

Biomart

SeqHound

API Consumer

Local Java

Beanshell

BioMoby Client

WS Client

SoapLab

Styx Client

External Services (open world)

Direct JDBC connection to Biomart dataset with self describing metadata

SQL Database

JDBC

HTTP REST-like

SeqHound Server

SeqHound Genomic Data Warehouse

In process Java based operations such as inlined scripting

BioMoby Service

Server side CPU & Data resources

SOAP Based

Web Service

Styx

SoapLab Server

GowLab

Launcher

LSF, Condor or other batch submission system

Legacy CGI

Legacy Codes

Soaplab wrapping command line and web based legacy tools

Styx Grid Service

Legacy Codes

Styx Grid Service

Legacy Codes

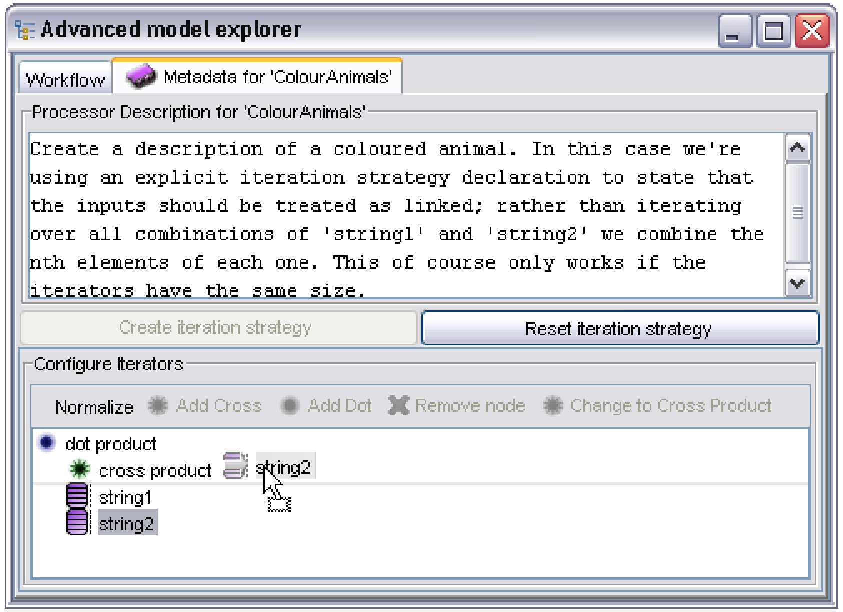
Styx Grid Service

Legacy Codes

Workflow subgraph containing streamed services using peer to peer data transfer based on Styx Grid Service protocol

Implicit Iteration

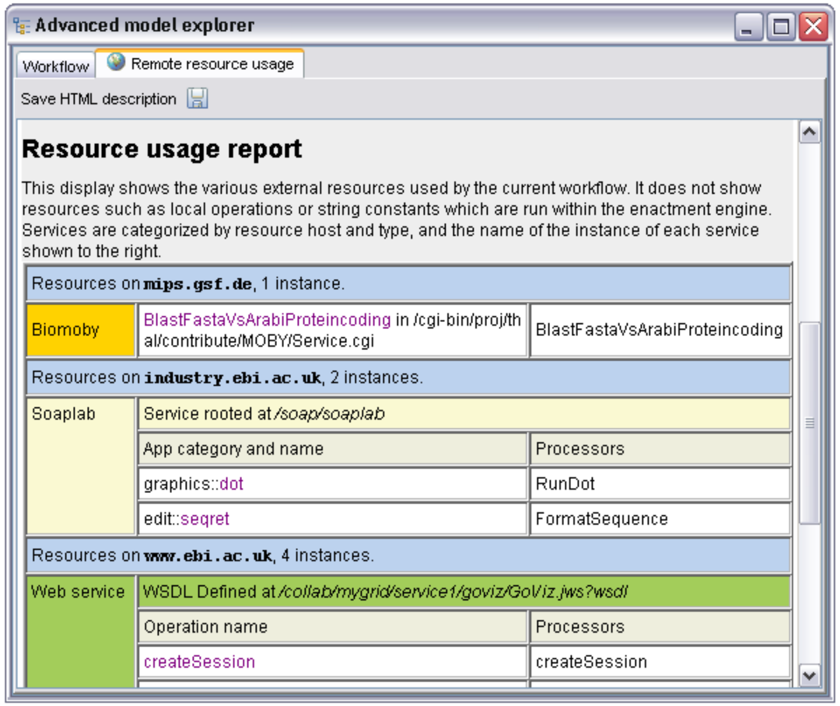
- Allows services to consume collections of items without service modification
- Equivalent to higher order map functions
- Graphical configuration
- Intuitively understood by our user community
- Scares computer scientists ☺



Create a description of a coloured animal. In this case we're using an explicit iteration strategy declaration to state that the inputs should be treated as linked; rather than iterating over all combinations of 'string1' and 'string2' we combine the nth elements of each one. This of course only works if the iterators have the same size.

Workflow summary views

- Diagram and HTML report of the structure of and resources used by the workflow
- Intended to be added to papers, websites etc.
- Can be used by portals, workflow repositories
- Supports reuse – very important!



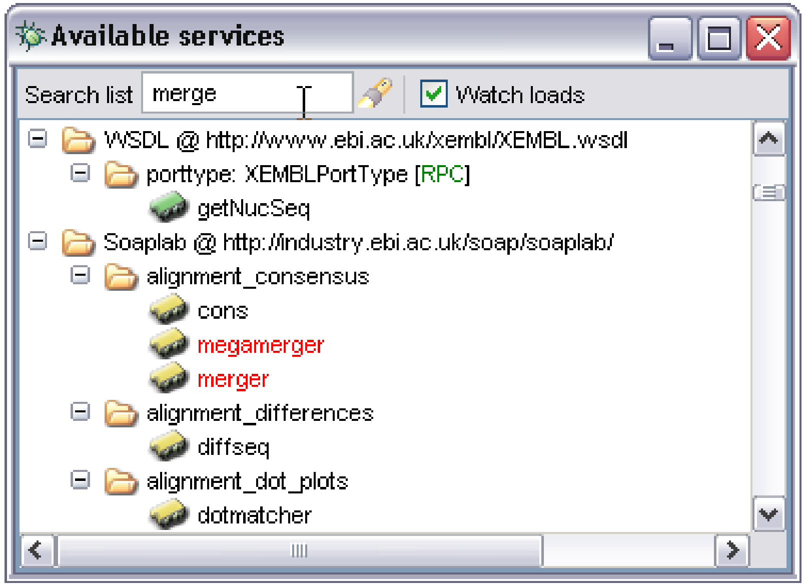
Resource usage report

This display shows the various external resources used by the current workflow. It does not show resources such as local operations or string constants which are run within the enactment engine. Services are categorized by resource host and type, and the name of the instance of each service shown to the right.

Resources on **mips.gsf.de**, 1 instance.		
Biomoby	BlastFastaVsArabiProteincoding in /cgi-bin/proj/thal/contribute/MOBY/Service.cgi	BlastFastaVsArabiProteincoding
Resources on **industry.ebi.ac.uk**, 2 instances.		
Soaplab	Service rooted at */soap/soaplab*	
	App category and name	Processors
	graphics::dot	RunDot
	edit::seqret	FormatSequence
Resources on **www.ebi.ac.uk**, 4 instances.		
Web service	WSDL Defined at */collab/mygrid/service1/goviz/GoViz.jws?wsdl*	
	Operation name	Processors
	createSession	createSession

Semantic and Naïve Search

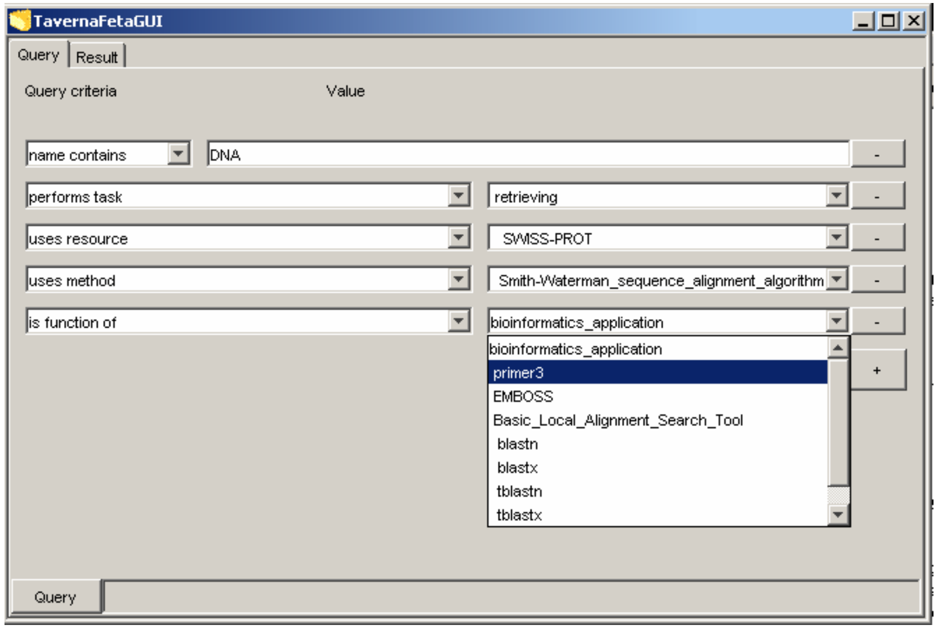
- Find services by name or…
- …by function, input types, resources



Available services

Search list: merge ☑ Watch loads

- WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
  - porttype: XEMBLPortType [RPC]
    - getNucSeq
- Soaplab @ http://industry.ebi.ac.uk/soap/soaplab/
  - alignment_consensus
    - cons
    - megamerger
    - merger
  - alignment_differences
    - diffseq
  - alignment_dot_plots
    - dotmatcher



TavernaFetaGUI

Query criteria	Value
name contains	DNA
performs task	retrieving
uses resource	SWISS-PROT
uses method	Smith-Waterman_sequence_alignment_algorithm
is function of	bioinformatics_application

bioinformatics_application
primer3
EMBOSS
Basic_Local_Alignment_Search_Tool
blastn
blastx
tblastn
tblastx

Successful?

- Over 1200 downloads of the workbench software for release 1.0
  - Averaging 10-15 downloads / day for release 1.1
  - Slightly scary 220 downloads in three days for 1.2 ☺
- Over 100 active mailing list participants
- Over 1300 available services
- Used across the world in widely differing projects, mostly but not all in bioinformatics (some cheminformatics)
- Active external developer community!

Taverna User Support

- Taverna has a self supporting user community
- Access help from other users and from the project developers via our mailing lists
- All accessible from http://taverna.sf.net
- We have a user manual! Please use it ☺

Where next?

- Funding
  - Core myGrid project has completed (3 years)
  - Follow-on platform grant for core team until 2008
  - Associated consumer / helper projects
    - Comparagrid, EMBRACE, iSpider…
- Will be used to…
  - Enhance the scalability of the workflow core
  - Investigate new interfaces (Dalec, Data driven workbench…)

Schedule

- 1.3 Release in September
  - Final version 1 release
- Moving to 2.0 with new workflow core by end 2005

Acknowledgements

myGrid is an EPSRC funded UK eScience Program Pilot Project

AstraZeneca

gsk GlaxoSmithKline

MERCK

EPISTEMICS LTD knowledge is our business

network inference

IBM

Sun microsystems

SOURCEFORGE.net

EMBL-EBI European Bioinformatics Institute

THE UNIVERSITY of MANCHESTER

THE UNIVERSITY OF NOTTINGHAM

University of Southampton

EPSRC Engineering and Physical Sciences Research Council

geneticxchange

UNIVERSITY OF NEWCASTLE UPON TYNE

iT innovation

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