Going with the Flow Distributed Computing for Systems Biology Using Taverna

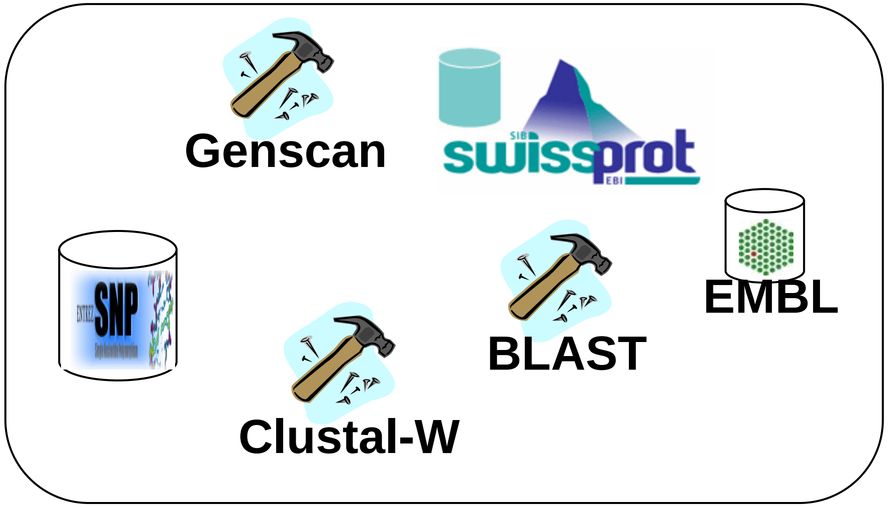
Prof Carole Goble

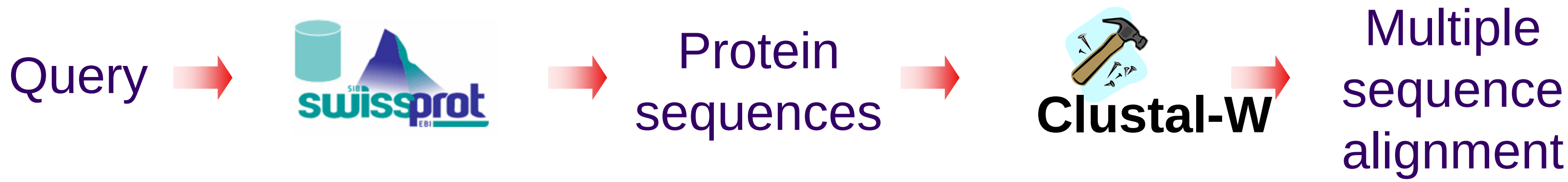
The University of Manchester, UK

http://www.mygrid.org.uk
http://www.omii.ac.uk

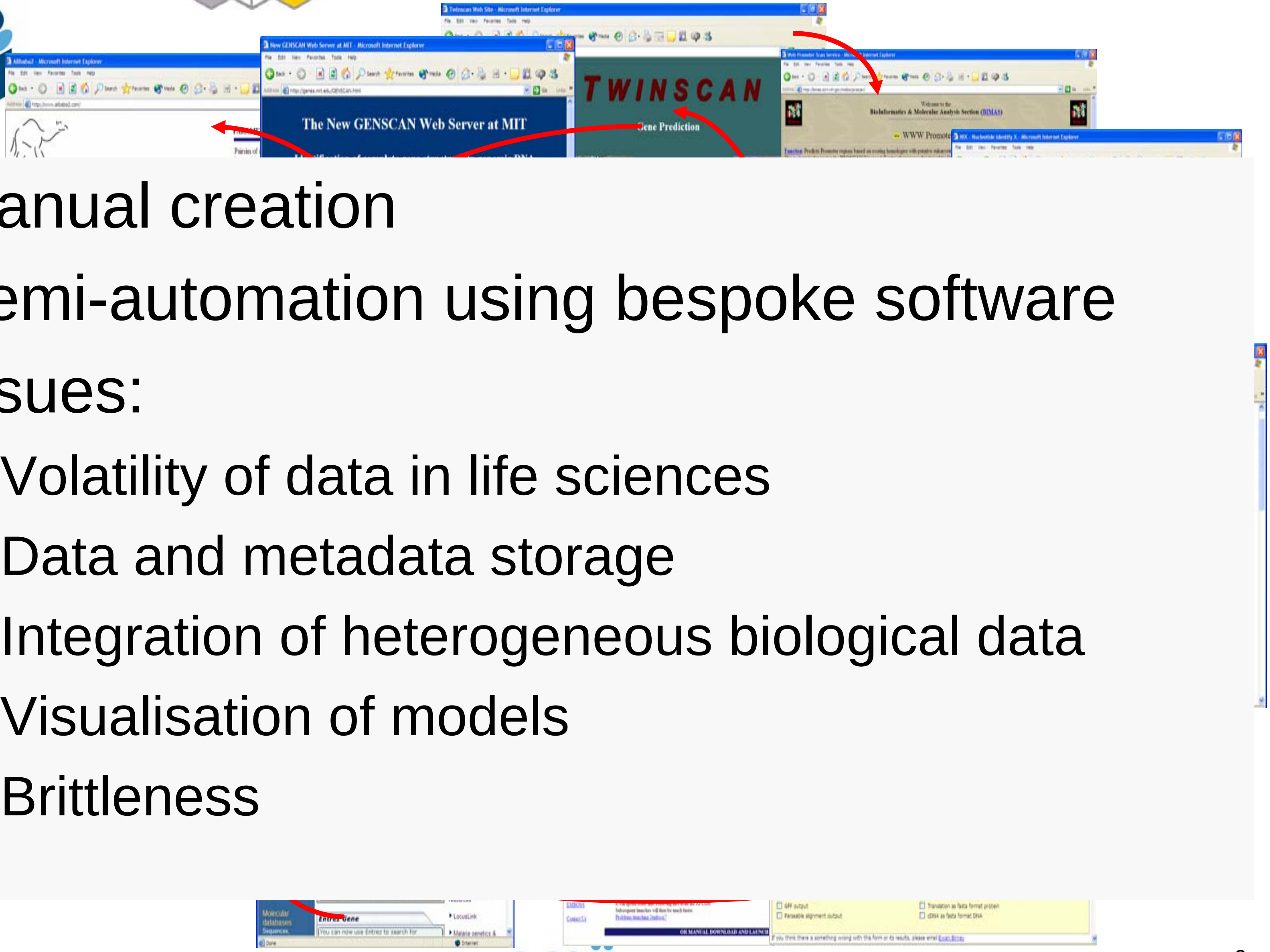
Data pipelines in bioinformatics

Resources /Services

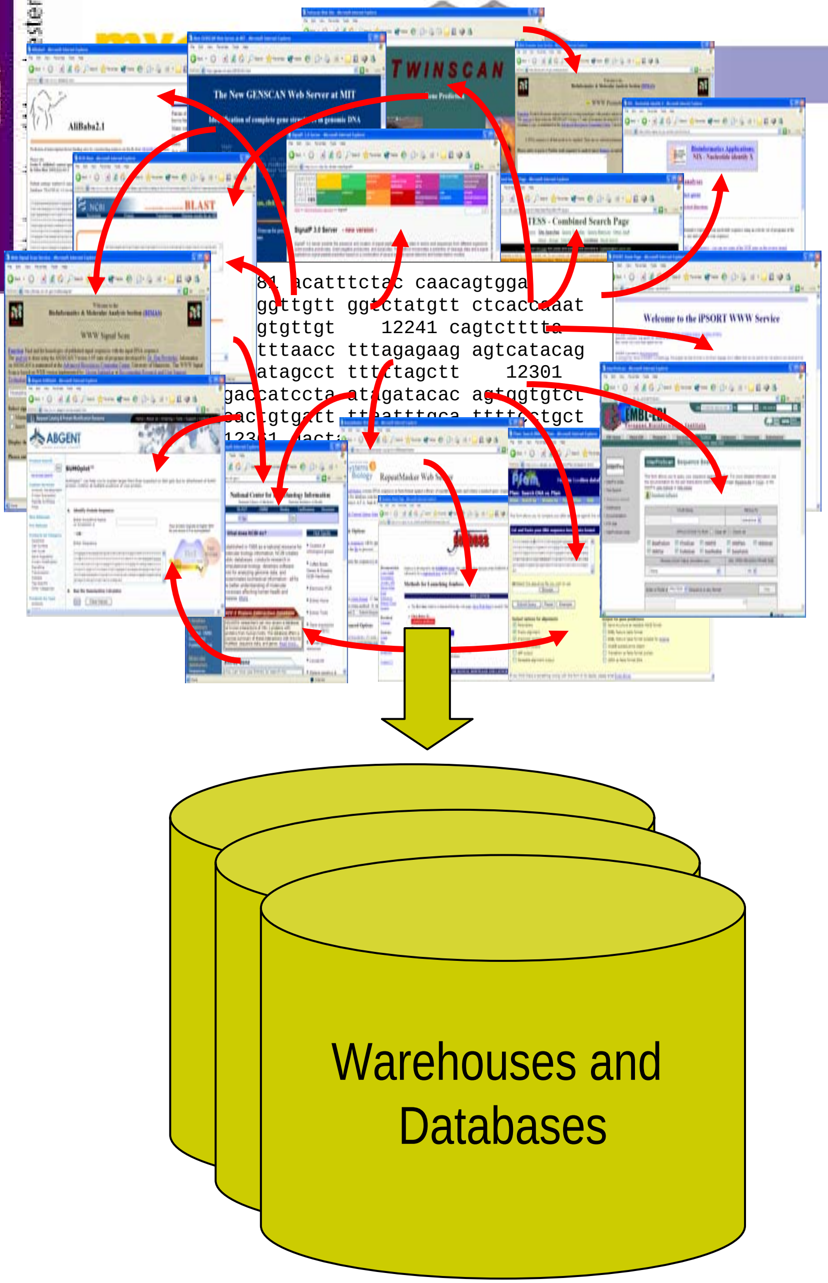




Example *in silico* experiment: Investigate the evolutionary relationships between proteins [Peter Li]



- Manual creation
- Semi-automation using bespoke software
- Issues:
  - Volatility of data in life sciences
  - Data and metadata storage
  - Integration of heterogeneous biological data
  - Visualisation of models
  - Brittleness

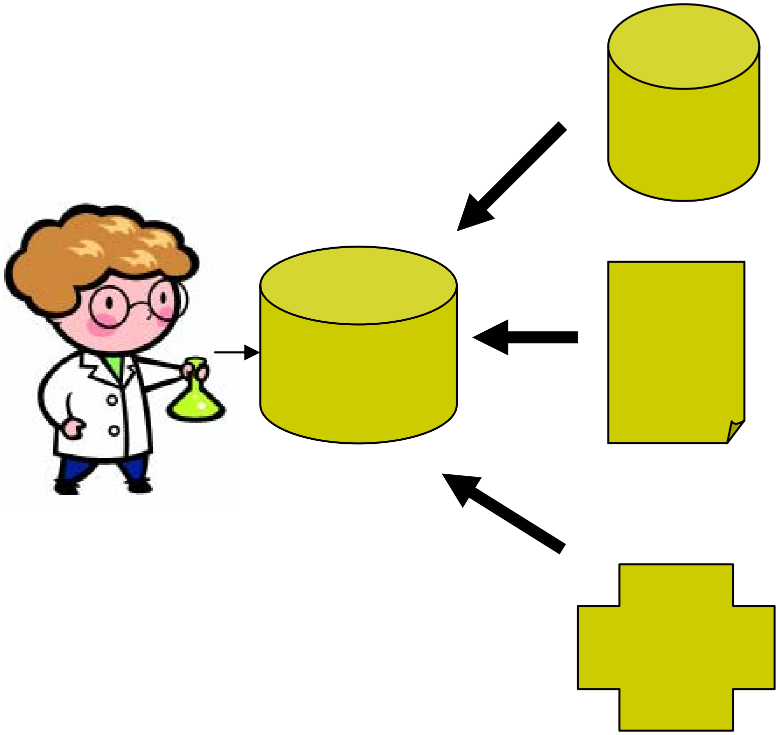

TWINSCAN
AliBaba2.1
The New GENSCAN Web Server at MIT
Identification of complete gene structures in genomic DNA
BLAST
SignalP 3.0 Server
TESS - Combined Search Page
Bioinformatics Applications
Welcome to the iPSORT WWW Service
WWW Signal Scan
ABGENT
National Center for Biotechnology Information
RepeatMasker Web Server
Pfam
EMBL-EBI
81 acatttctac caacagtgga
ggttgtt ggtctatgtt ctcaccaaat
gtgttgt 12241 cagtctttta
tttaacc tttagagaag agtcatacag
atagcct ttttagctt 12301
gaccatccta atagatacac agtggtgtct
cactgtgatt ttaatttgca ttttcctgct
12361 act

Warehouses and Databases

© omii

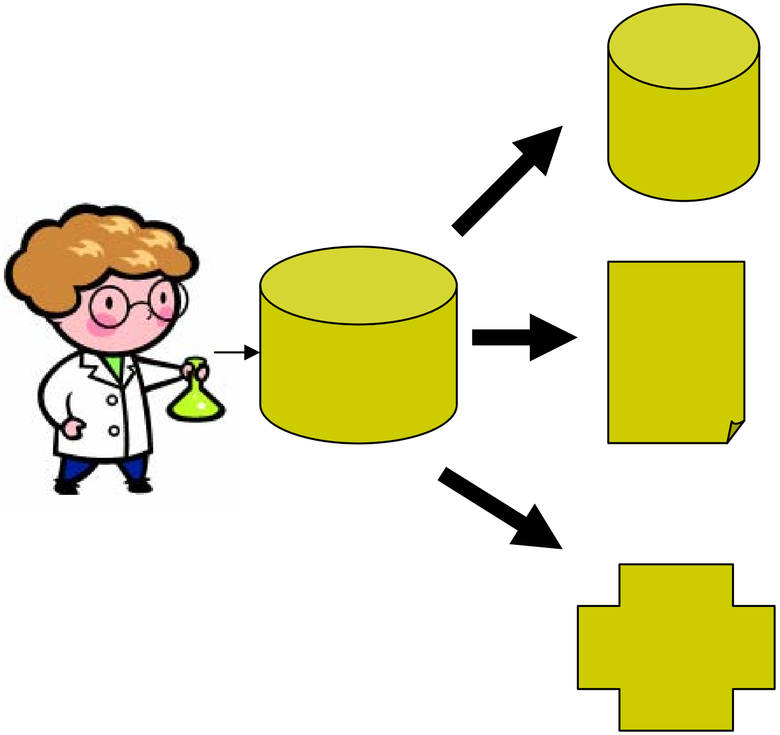
Data Warehouse

- Copy the data sets
- Combine them into a pre-determined model before query
- Query that model
- Clean data
- Refresh, Fixed
- High cost, Front loaded
- You can only use what has been set up for you.



Distributed Database Integration

- Marshal the data sets
- Combine it into a pre-determined model when you query
- Always fresh
- Map from model to databases dynamically
- More flexible but still depends on model
- High cost
- You can only use what has been set up

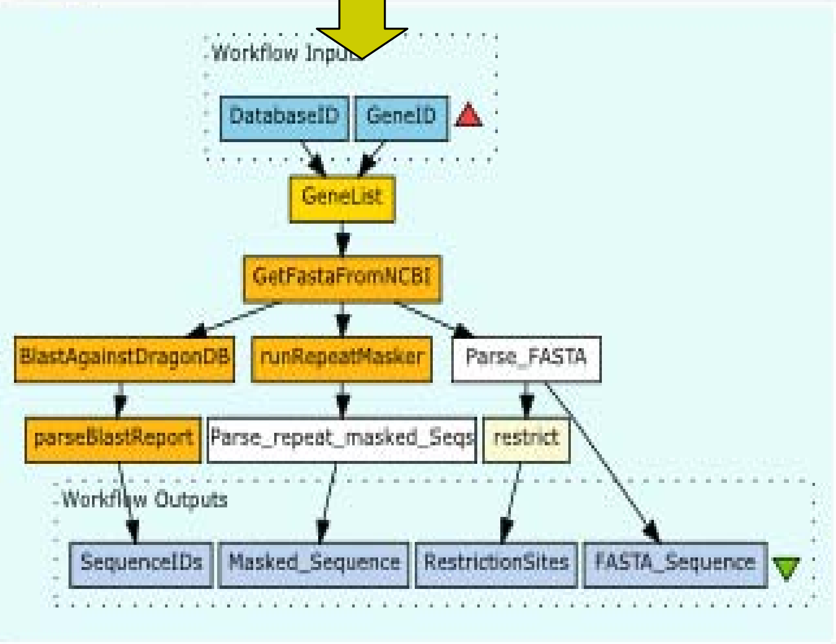


© omii

TWINSCAN
The New GENSCAN Web Server at MIT
AliBaba2.1
BLAST
ABGENT
Welcome to the iPSORT WWW Service
EMBL-EBI
RepeatMasker Web Server
Pfam
81 acatttctac caacagtgga
ggttgtt ggtctatgtt ctcaccaaat
gtgttgt 12241 cagtctttta
tttaacc tttagagaag agtcatacag
atagcct tttctagctt 12301
gacatcccta atagatacac agtggtgtct
cactgtgatt ttaatttgca ttttctgct
12361 act

Protocol

Create a gene list in Excel
Go to NCBI
Retrieve FASTA for each gene
DragonDB Blast each sequence
Copy/paste IDs into a spreadsheet
Run Repeat Masker on each sequence
copy/paste masked sequence into Excel
Run MacVector cut each seq with EcoRI



[Mark Wilkinson, 2006 BioMOBY]

It would be good if you could systematically automate…

Make data sets / resources / tools / codes / models accessible to a computer.

And cope when they change

And run them where they are hosted ….

It would be good if you could systematically automate…

….Link together resources

Automate the protocol so I don’t have to do it every time I need to repeat the search or re-run the analysis.

And do it accurately and systematically every time without mistakes. And not get bored and sloppy. And be comprehensive too….

© omii

It would be good if you could systematically automate…

…Rerun it over and over and over and over again. Automatically. And keep the log of what actually happened. Automatically.

Manage the results of the protocol. Not just the data results, but the evidence for the results, the source of the data, the log of what you did and why. Helpful when you publish!....

It would be good if you could systematically automate…

…Record this protocol, share it with colleagues.

Fiddle with it.

Remember what it was 2 weeks later.

Adapt a colleague’s or expert’s to suit you

Give your protocol to a colleague…

It would be good if you could systematically automate…

And do it in my lab without having to have a lot of systems administrators and developers building databases for me.

Or writing Perl.

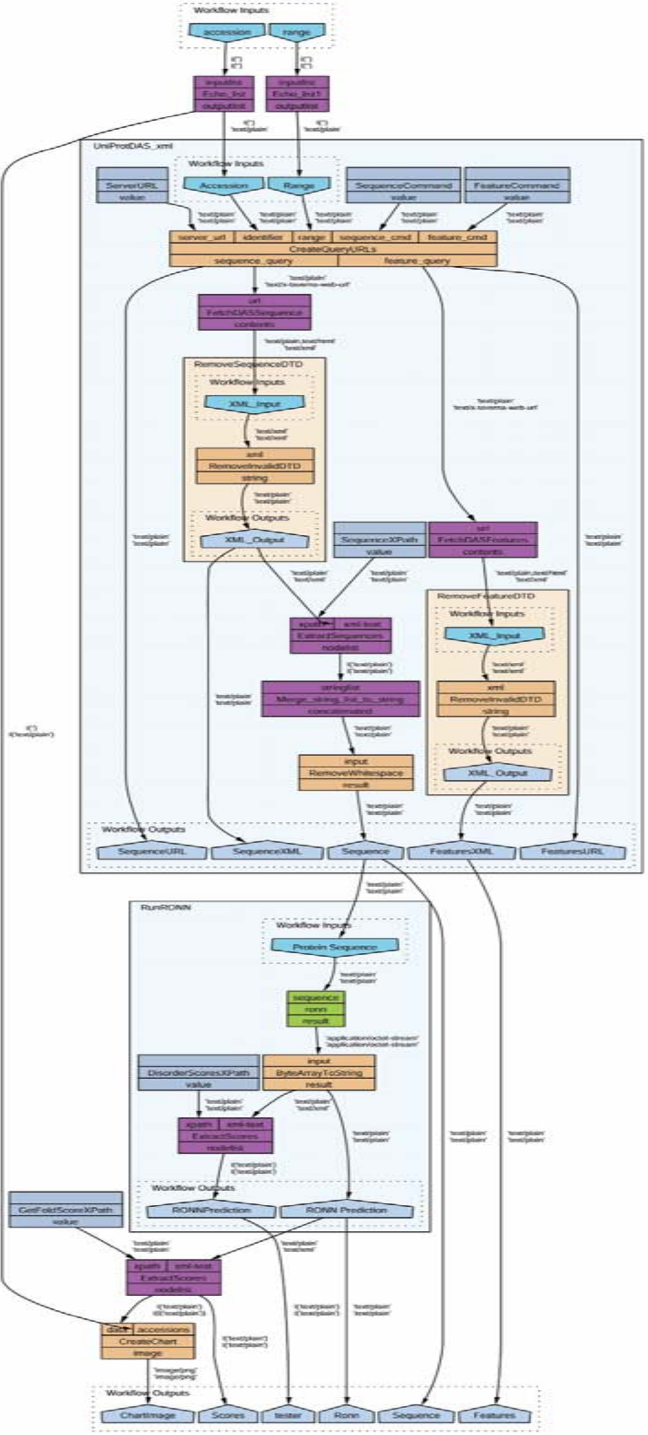
And it runs on my crappy laptop.

And be …

- Un-biased and Unambiguous in my science
- Systematic
- Efficient
- Scalable
- Flexible
- Customisable
- Transparent in my scientific method

The Two W's

- Web Services
  - Technology and standard for exposing code / database with an API that can be consumed by a third party remotely.
  - Describes how to interact with it.
- Workflows
  - General technique for describing and enacting a process
  - Describes *what* you want to do, not *how* you want to do it

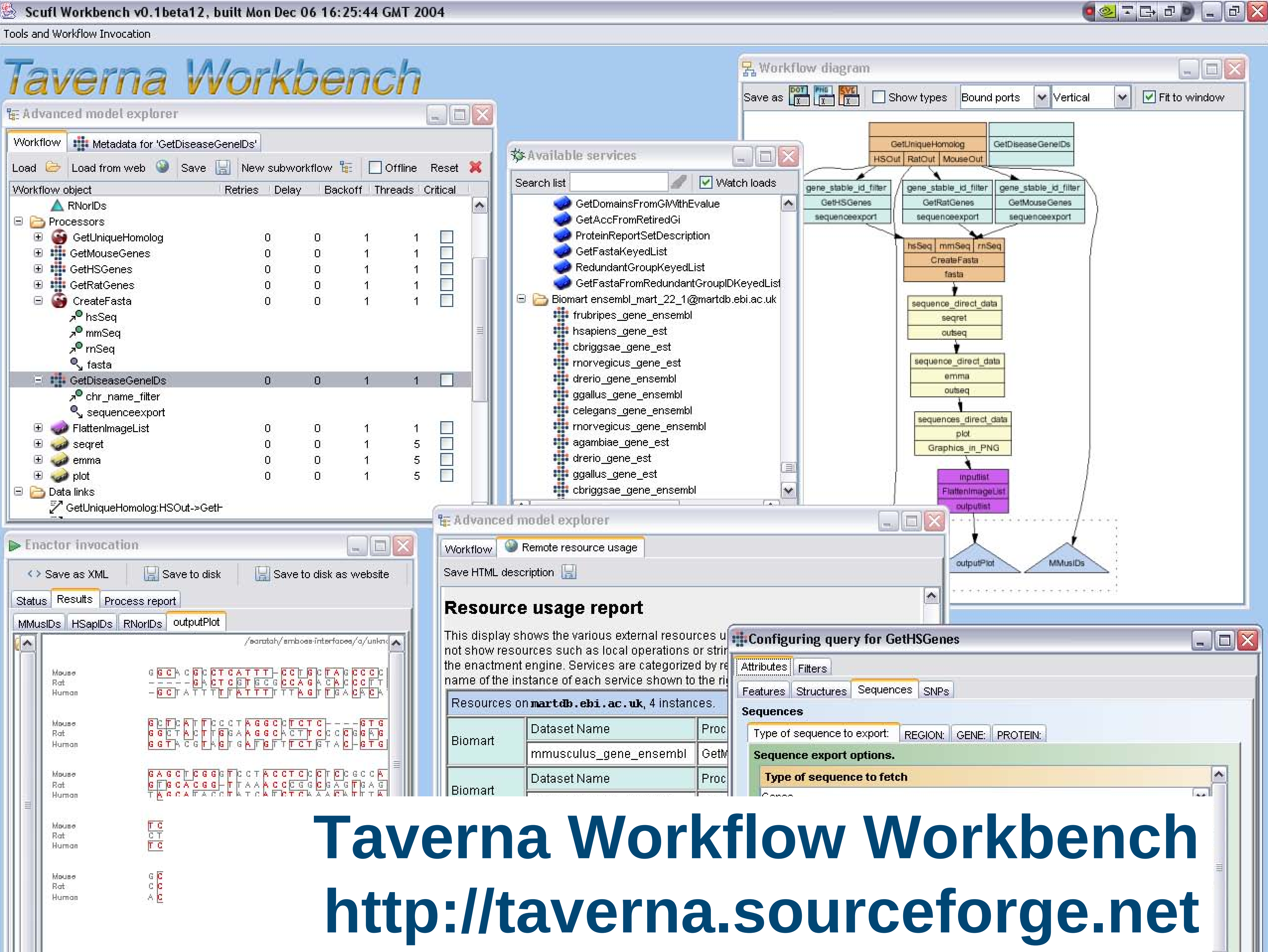


Workflows

- Workflow language specifies how bioinformatics processes fit together.
- High level workflow diagram separated from any lower level coding – you don’t have to be a coder to build workflows.
- Workflow is a kind of script or protocol that you configure when you run it.
- Easier to explain, share, relocate, reuse and repurpose.

myGrid

- myGrid http://www.mygrid.org.uk
- UK e-Science pilot project since 2001
- Build middleware for Life Scientists that enables them to undertake *in silico* experiments and share those experiments and their results.
- Individual scientists, in under-resourced labs, who use other people’s applications.
- Open source.
- Workflows.
- Data flows. Ad hoc & exploratory



Scufl Workbench v0.1beta12, built Mon Dec 06 16:25:44 GMT 2004

Tools and Workflow Invocation

Taverna Workbench

Advanced model explorer

Workflow | Metadata for 'GetDiseaseGeneIDs'

Load | Load from web | Save | New subworkflow | □ Offline | Reset

Workflow object	Retries	Delay	Backoff	Threads	Critical
RNorIDs					
⊟ Processors					
⊞ GetUniqueHomolog	0	0	1	1	□
⊞ GetMouseGenes	0	0	1	1	□
⊞ GetHSGenes	0	0	1	1	□
⊞ GetRatGenes	0	0	1	1	□
⊟ CreateFasta	0	0	1	1	□
hsSeq					
mmSeq					
rnSeq					
fasta					
⊟ GetDiseaseGeneIDs	0	0	1	1	□
chr_name_filter					
sequenceexport					
⊞ FlattenImageList	0	0	1	1	□
⊞ seqret	0	0	1	5	□
⊞ emma	0	0	1	5	□
⊞ plot	0	0	1	5	□
⊟ Data links					
GetUniqueHomolog:HSOut->GetH					

Available services

Search list | ☑ Watch loads

- GetDomainsFromGiWithEvalue
- GetAccFromRetiredGi
- ProteinReportSetDescription
- GetFastaKeyedList
- RedundantGroupKeyedList
- GetFastaFromRedundantGroupIDKeyedList
- ⊟ Biomart ensembl_mart_22_1@martdb.ebi.ac.uk
  - frubripes_gene_ensembl
  - hsapiens_gene_est
  - cbriggsae_gene_est
  - rnorvegicus_gene_est
  - drerio_gene_ensembl
  - ggallus_gene_ensembl
  - celegans_gene_ensembl
  - rnorvegicus_gene_ensembl
  - agambiae_gene_est
  - drerio_gene_est
  - ggallus_gene_est
  - cbriggsae_gene_ensembl

Workflow diagram

Save as DOT PNG SVG | □ Show types | Bound ports | Vertical | ☑ Fit to window

Enactor invocation

Save as XML | Save to disk | Save to disk as website

Status | Results | Process report

MMusIDs | HSapIDs | RNorIDs | outputPlot

/saratah/emboss-interfaces/a/unkno

```
Mouse    GGCACGCCTCATTT-CCTGCTAGCCCC
Rat      ------GACTCGTGCGCCAGACACCCTT
Human    -GCTATTTTTATTTTTTAGTTGACACA

Mouse    GCTCATTCCCTAGGCCTCTC----GTG
Rat      GGCTACTTGGAAGGCACTTCCCCGGAG
Human    GGTACGTAGTGATGTTTCTGTAC-GTG

Mouse    GAGCTCGGGTCCTACCTCCCTCCGCCA
Rat      GTGCACGG-TTAAACCCGGCGAGTGAG
Human    TAGCATACCTATCATCTCAAACATTTA

Mouse    TC
Rat      CT
Human    TC

Mouse    GC
Rat      CC
Human    AC
```

Advanced model explorer

Workflow | Remote resource usage

Save HTML description

Resource usage report

This display shows the various external resources u
not show resources such as local operations or strir
the enactment engine. Services are categorized by re
name of the instance of each service shown to the ri

Resources on **martdb.ebi.ac.uk**, 4 instances.

Biomart	Dataset Name	Proc
	mmusculus_gene_ensembl	GetM
Biomart	Dataset Name	Proc

Configuring query for GetHSGenes

Attributes | Filters

Features | Structures | Sequences | SNPs

Sequences

Type of sequence to export: | REGION: | GENE: | PROTEIN:

Sequence export options.

Type of sequence to fetch

Gene

Taverna Workflow Workbench

http://taverna.sourceforge.net

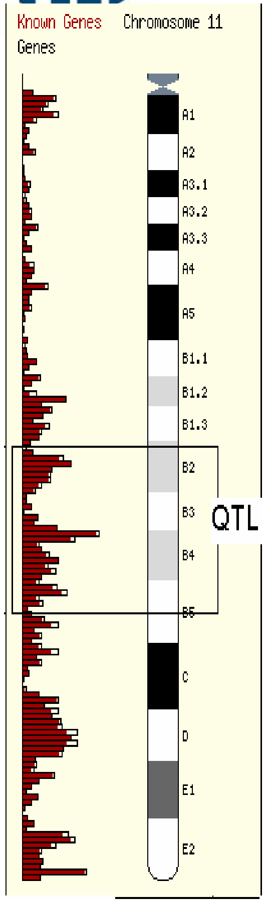
The Wellcome Trust Funded
Host-Pathogen Project

Trypanosomiasis in Africa

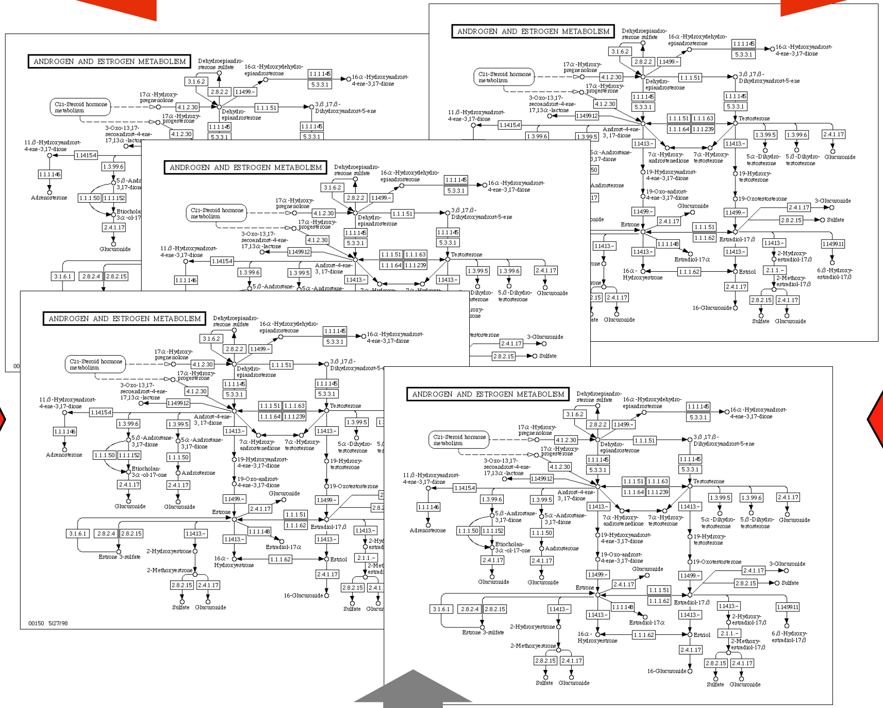
http://www.genomics.liv.ac.uk/tryps/trypsindex.html

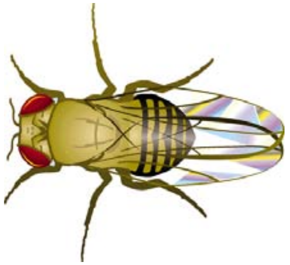
Genotype

Phenotype

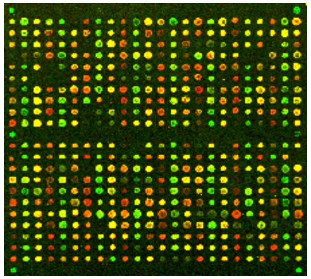


200


ANDROGEN AND ESTROGEN METABOLISM
ANDROGEN AND ESTROGEN METABOLISM
ANDROGEN AND ESTROGEN METABOLISM
ANDROGEN AND ESTROGEN METABOLISM
ANDROGEN AND ESTROGEN METABOLISM



?



Genes captured in microarray experiment and present in QTL region

Microarray + QTL

Phenotypic response investigated using microarray in form of expressed genes or evidence provided through QTL mapping

[Andy Brass, Steve Kemp, Paul Fisher, 2006]

Key:

A – Retrieve genes in QTL region

B – Annotate genes with external database Ids

C – Cross-reference Ids with KEGG gene ids

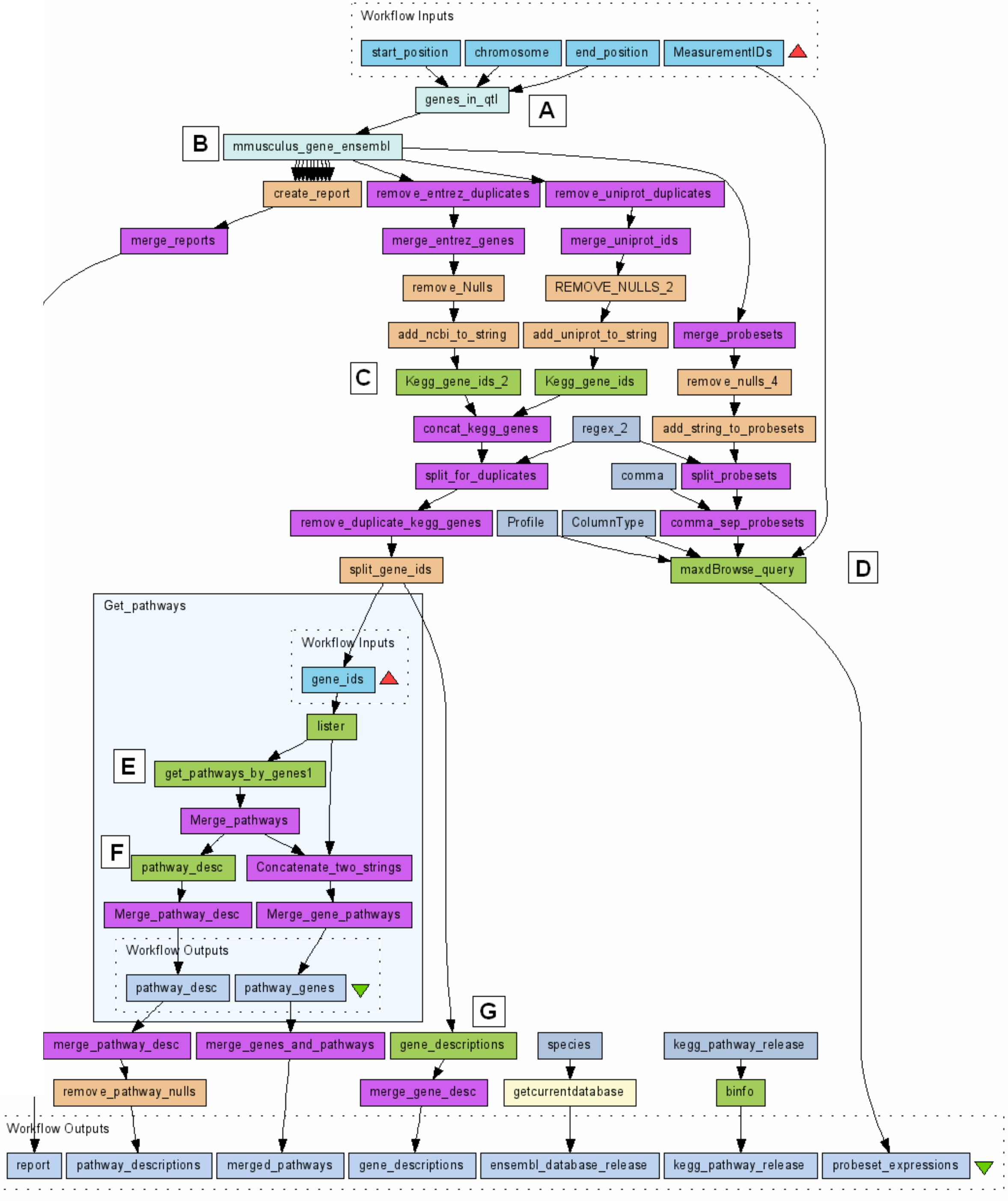
D – Retrieve microarray data from MaxD database

E – For each KEGG gene get the pathways it’s involved in

F – For each pathway get a description of what it does

G – For each KEGG gene get a description of what it does

[Andy Brass, Steve Kemp, Paul Fisher, 2006]



Result

- Captured the pathways returned by QTL and Microarray workflows over the MaxD microarray database
- Identified a pathway for which its correlating gene (Daxx) is believed to play a role in trypanosomiasis resistance.
- Manually analysis on the microarray and QTL data had failed to identify this gene as a candidate.

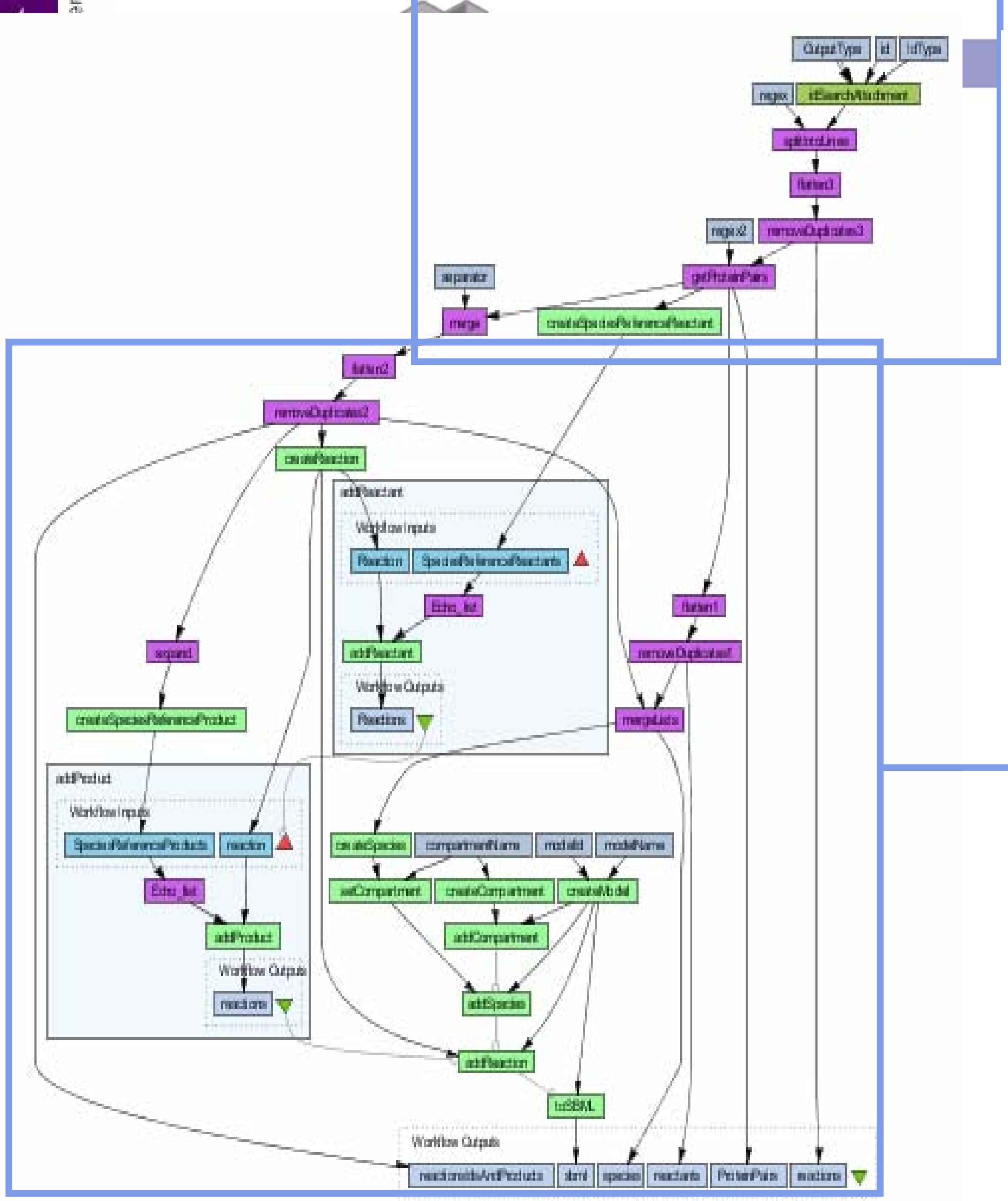
[Andy Brass, Steve Kemp, Paul Fisher, 2006]

Trichuris muris (mouse whipworm) infection

- Identified the biological pathways involved in sex dependence in the mouse model, previously believed to be involved in the ability of mice to expel the parasite.
- Manual experimentation: Two year study of candidate genes, processes unidentified
- Workflows: trypanosomiasis cattle experiment, was reused without change.
- Analysis of the data by a biologist found the processes in a couple of days.

[Joanne Pennock, Paul Fisher, 2006]

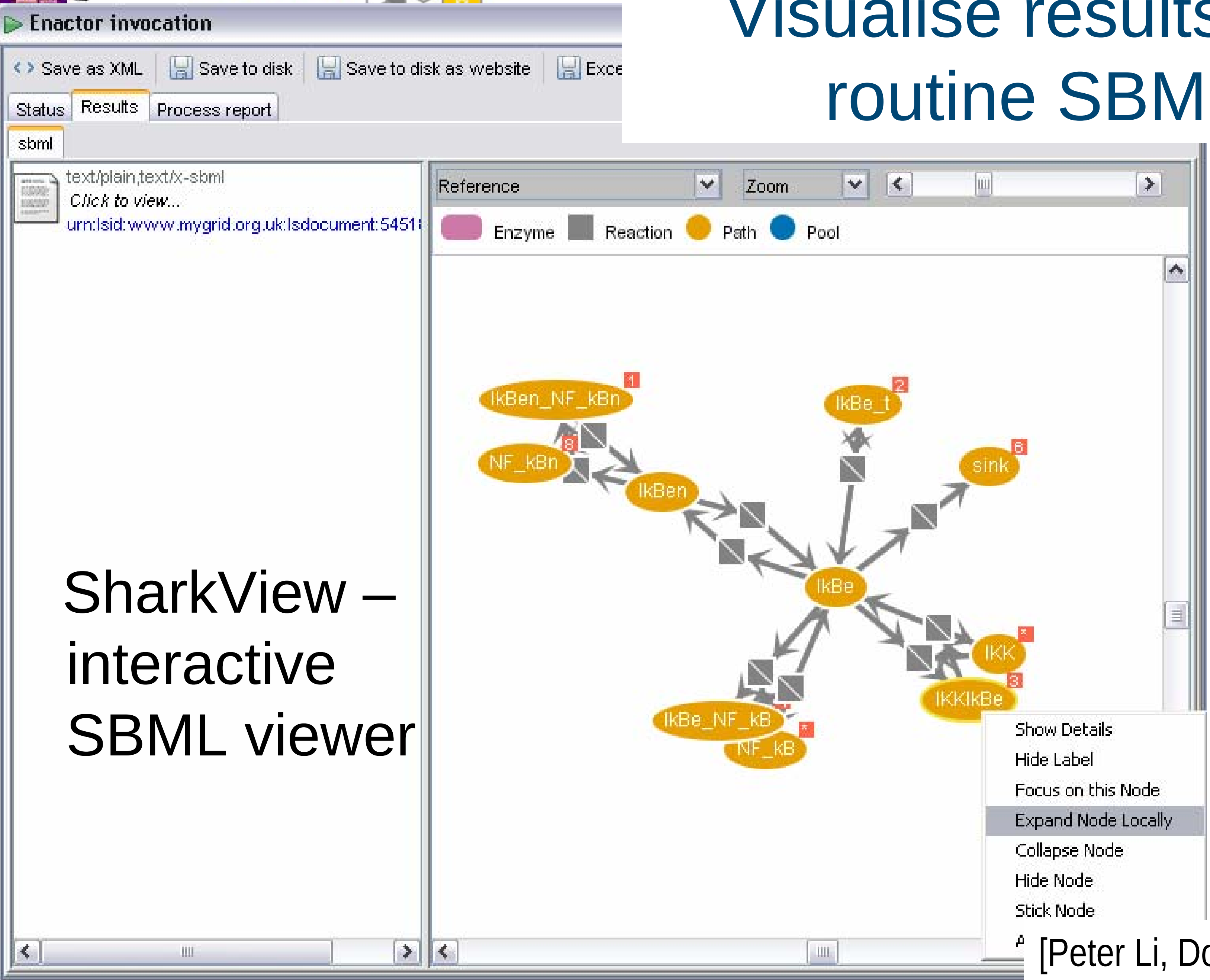
Pull Public Databases + inHouse Data = Model



Core SBML model construction workflow

[Peter Li, Doug Kell, 2006]

Visualise results using routine SBML tools



text/plain,text/x-sbml

Click to view...

urn:lsid:www.mygrid.org.uk:lsdocument:5451

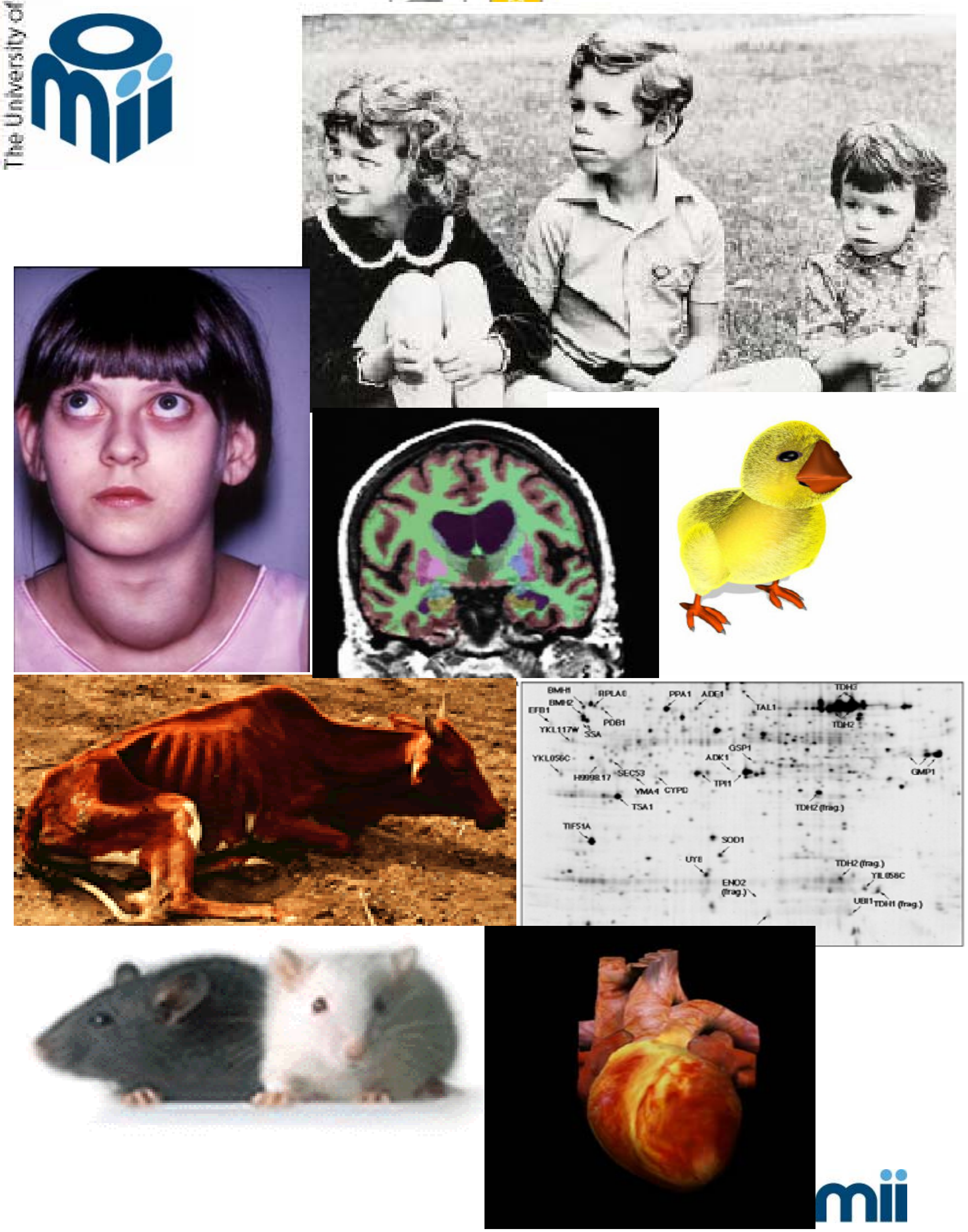
SharkView – interactive SBML viewer

[Peter Li, Doug Kell, 2006]

Model construction: Post-Taverna

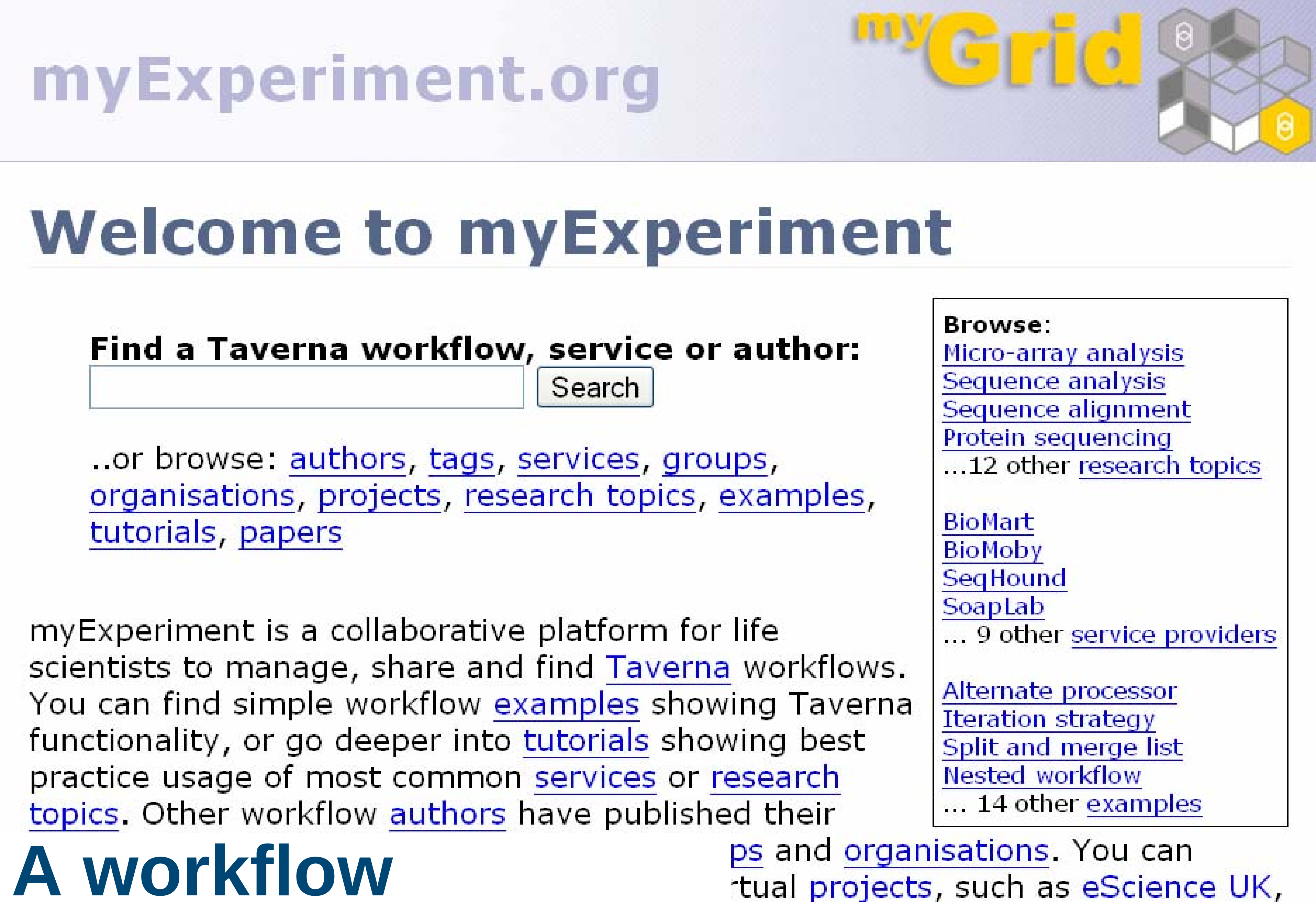
- Captures the scientific process of model construction as workflows
- Workflows enacted ‘on demand’ to construct most up-to-date models using the latest data
- Models are pushed into a data model of choice
- Provide various ways of visualising models

Multi-disp.



- ~20000 downloads
- Users in US, Singapore, UK, Europe, Australia,
- Systems biology
- Proteomics
- Gene/protein annotation
- Microarray data analysis
- Medical image analysis
- Heart simulations
- High throughput screening
- Phenotypical studies
- Plants, Mouse, Human
- Astronomy
- Dilbert Cartoons

A workflow marketplace



Welcome to myExperiment

Find a Taverna workflow, service or author:

..or browse: authors, tags, services, groups, organisations, projects, research topics, examples, tutorials, papers

myExperiment is a collaborative platform for life scientists to manage, share and find Taverna workflows. You can find simple workflow examples showing Taverna functionality, or go deeper into tutorials showing best practice usage of most common services or research topics. Other workflow authors have published their

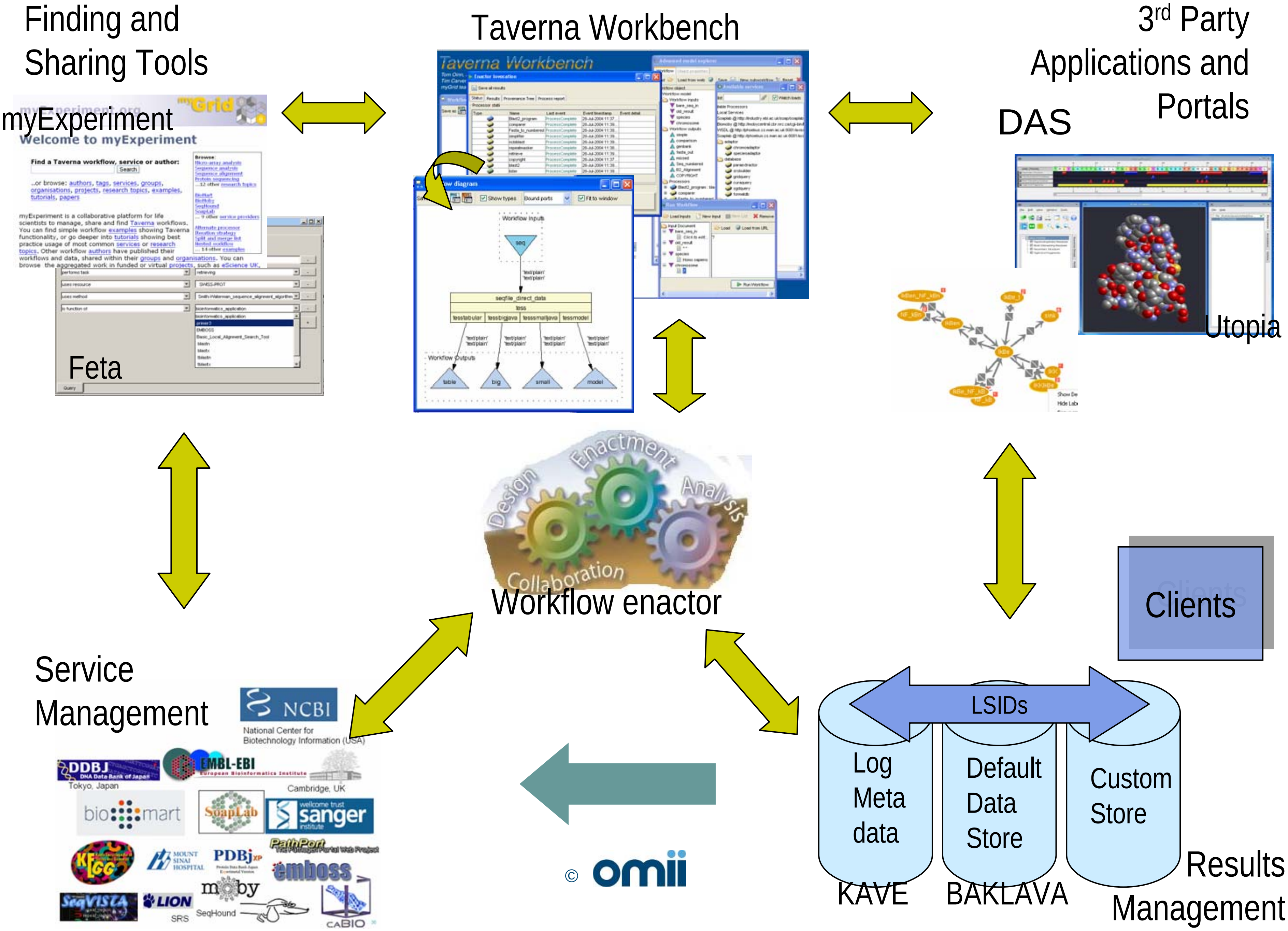
ps and organisations. You can

rtual projects, such as eScience UK,

Browse:
Micro-array analysis
Sequence analysis
Sequence alignment
Protein sequencing
...12 other research topics

BioMart
BioMoby
SeqHound
SoapLab
... 9 other service providers

Alternate processor
Iteration strategy
Split and merge list
Nested workflow
... 14 other examples



Finding and Sharing Tools

myExperiment

Welcome to myExperiment

Find a Taverna workflow, service or author:

..or browse: authors, tags, services, groups, organisations, projects, research topics, examples, tutorials, papers

myExperiment is a collaborative platform for life scientists to manage, share and find Taverna workflows. You can find simple workflow examples showing Taverna functionality, or go deeper into tutorials showing best practice usage of most common services or research topics. Other workflow authors have published their workflows and data, shared within their groups and organisations. You can browse the aggregated work in funded or virtual projects, such as eScience UK,

Feta

Taverna Workbench

3rd Party Applications and Portals

DAS

Utopia

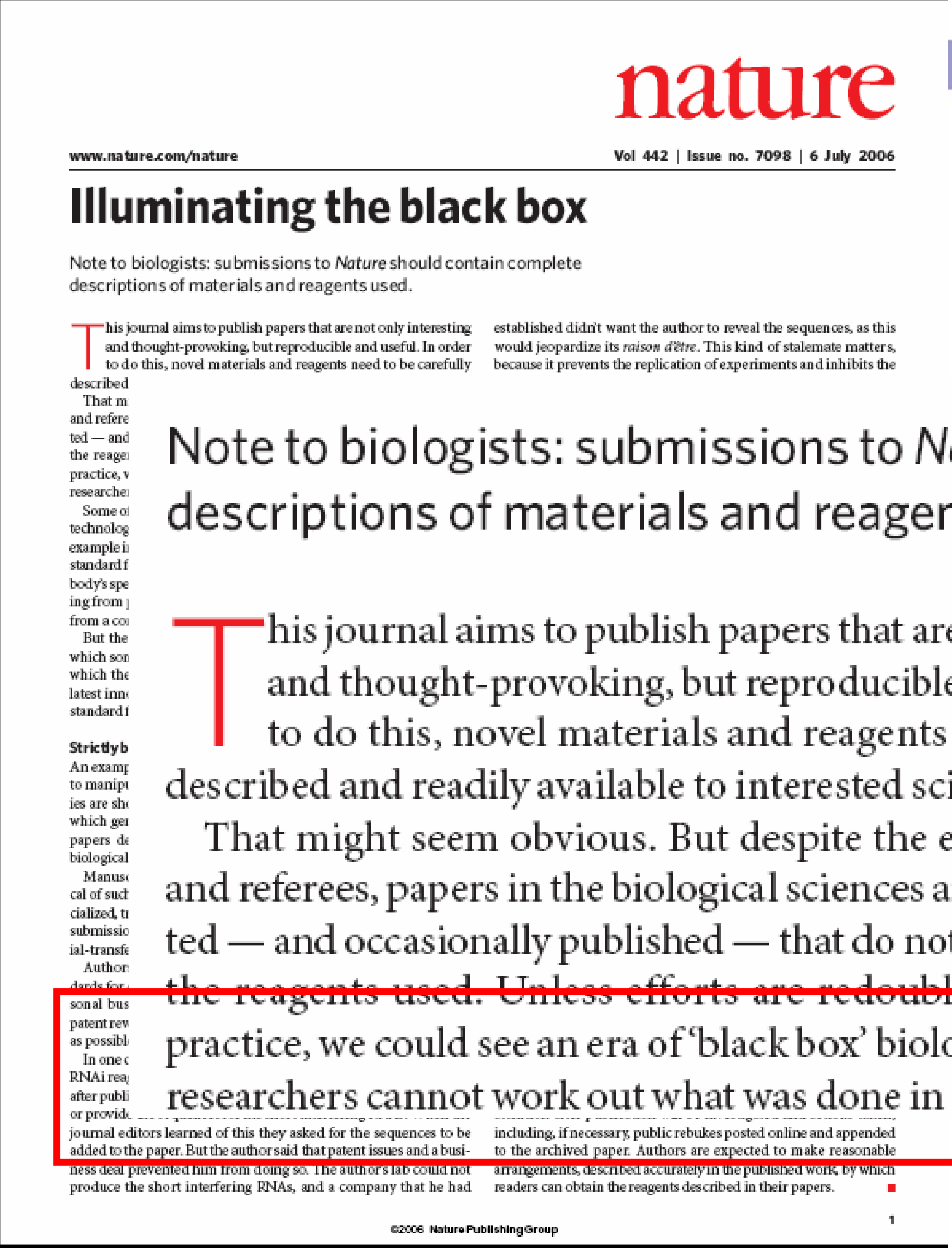
Workflow enactor

Clients

Service Management

Results Management

Transparency



nature

www.nature.com/nature Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully
described
That m
and refere
ted — and
the reage
practice, w
researche
Some o
technolog
example i
standard f
body's spe
ing from
from a co
But the
which son
which the
latest inn
standard f

Strictly b
An examp
to manip
ies are sh
which ge
papers d
biological
Manus
cal of suc
cialized, t
submissio
ial-transfe
Author
dards for
sonal bus
patent rev
as possibl
In one
RNAi rea
after publi
or provid
journal editors learned of this they asked for the sequences to be added to the paper. But the author said that patent issues and a business deal prevented him from doing so. The author's lab could not produce the short interfering RNAs, and a company that he had

established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the

including, if necessary, public rebukes posted online and appended to the archived paper. Authors are expected to make reasonable arrangements, described accurately in the published work, by which readers can obtain the reagents described in their papers. ■

1

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

That might seem obvious. But despite the efforts of our editors and referees, papers in the biological sciences are still being submitted — and occasionally published — that do not adequately describe the reagents used. Unless efforts are redoubled to eliminate this practice, we could see an era of 'black box' biology, in which outside researchers cannot work out what was done in an experiment.

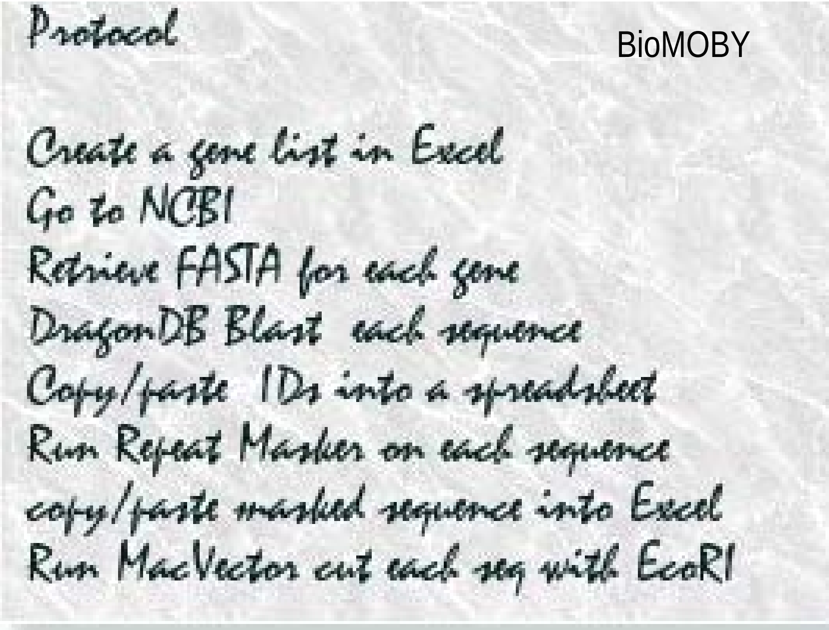
established didn't
would jeopardize
because it prevent
selection of appro
Some authors c
information or th
materials in quest
paper. This misses
ducing data that
consistent results

Provenance

- Who, What, Where, When, Why?, How?
- Context
- Interpretation
- Logging & Debugging
- Reproducibility and repeatability
- Evidence & Audit
- Non-repudiation
- Credit
- Credibility
- Accurate reuse and interpretation
- Smart re-running
- Cross experiment mining
- Just good scientific practice

© omii

Smart Tea



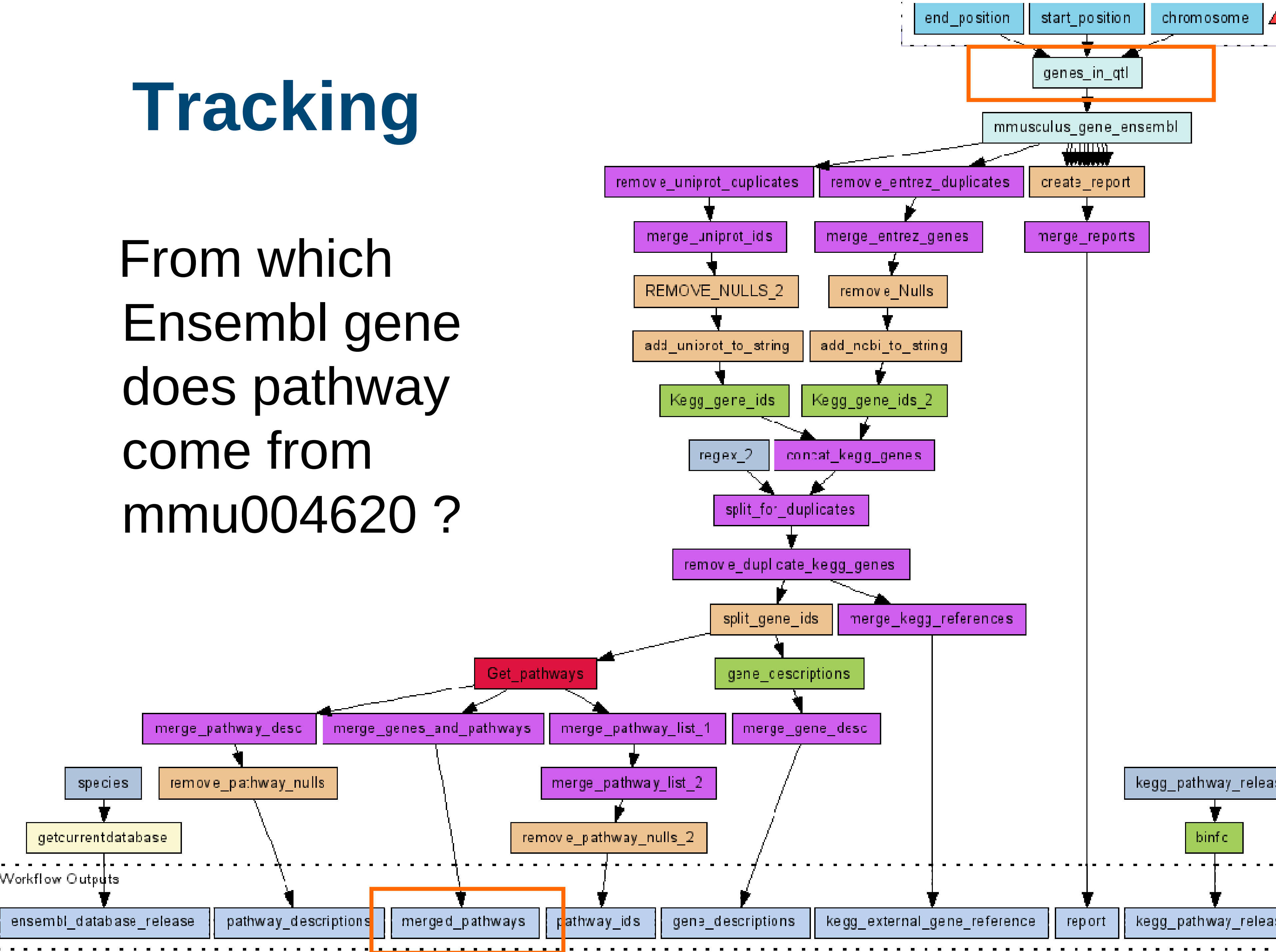
Protocol

BioMOBY

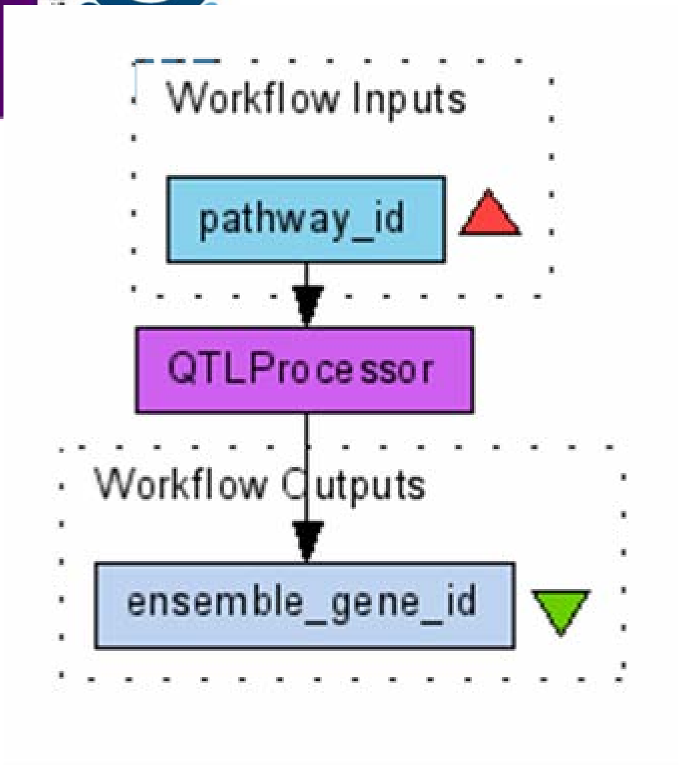
Create a gene list in Excel
Go to NCBI
Retrieve FASTA for each gene
DragonDB Blast each sequence
Copy/paste IDs into a spreadsheet
Run Repeat Masker on each sequence
copy/paste masked sequence into Excel
Run MacVector cut each seq with EcoRI

Tracking

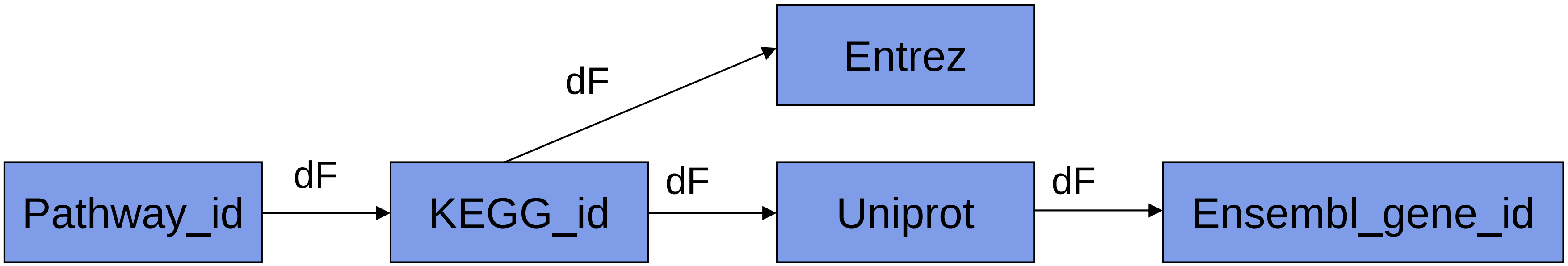
From which Ensembl gene does pathway come from mmu004620 ?



Workflows over Results

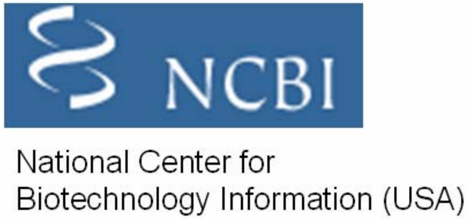


Automatically backtrack through the data provenance graph



An Open World

- Open domain services and resources.
- Taverna accesses 3000+ operation.
- Third party.
- All the major providers
  - NCBI, DDBJ, EBI …
- Enforce NO common data model.



- Quality Web Services considered desirable

.

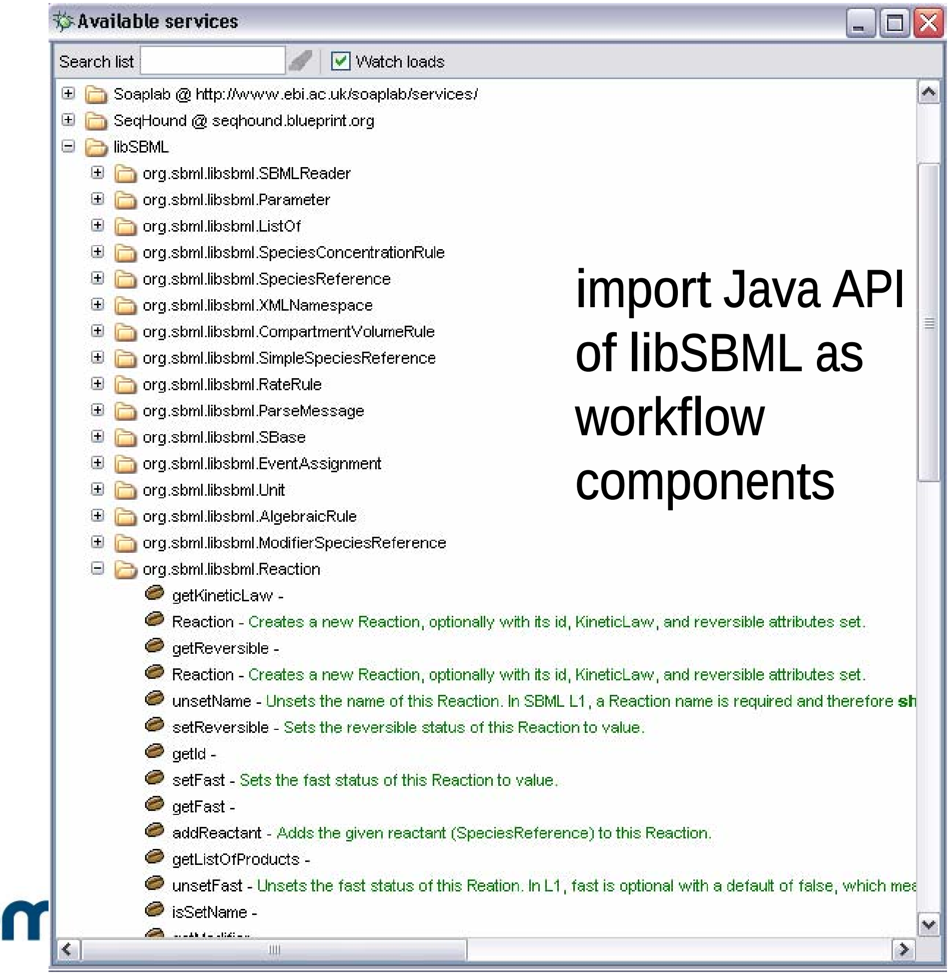
DDBJ
DNA Data Bank of Japan
Tokyo, Japan
EMBL-EBI
European Bioinformatics Institute
Cambridge, UK
bio mart
SoapLab
wellcome trust sanger institute
PathPort
The Pathogen Portal Web Project
KEGG
Kyoto Encyclopedia of Genes and Genomes
MOUNT SINAI HOSPITAL
PDBj XP
Protein Data Bank Japan Experimetal Version
emboss
moby
SeqVISTA
LION
SRS
SeqHound
CABIO

If you don’t provide a Web Service Interface…

- SoapLab

http://www.ebi.ac.uk/soaplab/

- Java API Consumer

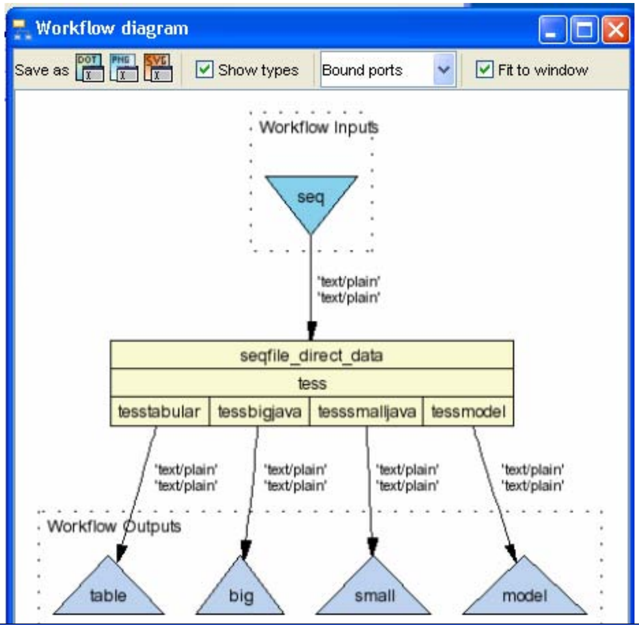


Soaplab @ http://www.ebi.ac.uk/soaplab/services/
SeqHound @ seqhound.blueprint.org
libSBML
org.sbml.libsbml.SBMLReader
org.sbml.libsbml.Parameter
org.sbml.libsbml.ListOf
org.sbml.libsbml.SpeciesConcentrationRule
org.sbml.libsbml.SpeciesReference
org.sbml.libsbml.XMLNamespace
org.sbml.libsbml.CompartmentVolumeRule
org.sbml.libsbml.SimpleSpeciesReference
org.sbml.libsbml.RateRule
org.sbml.libsbml.ParseMessage
org.sbml.libsbml.SBase
org.sbml.libsbml.EventAssignment
org.sbml.libsbml.Unit
org.sbml.libsbml.AlgebraicRule
org.sbml.libsbml.ModifierSpeciesReference
org.sbml.libsbml.Reaction
getKineticLaw -
Reaction - Creates a new Reaction, optionally with its id, KineticLaw, and reversible attributes set.
getReversible -
Reaction - Creates a new Reaction, optionally with its id, KineticLaw, and reversible attributes set.
unsetName - Unsets the name of this Reaction. In SBML L1, a Reaction name is required and therefore sh
setReversible - Sets the reversible status of this Reaction to value.
getId -
setFast - Sets the fast status of this Reaction to value.
getFast -
addReactant - Adds the given reactant (SpeciesReference) to this Reaction.
getListOfProducts -
unsetFast - Unsets the fast status of this Reation. In L1, fast is optional with a default of false, which mea
isSetName -

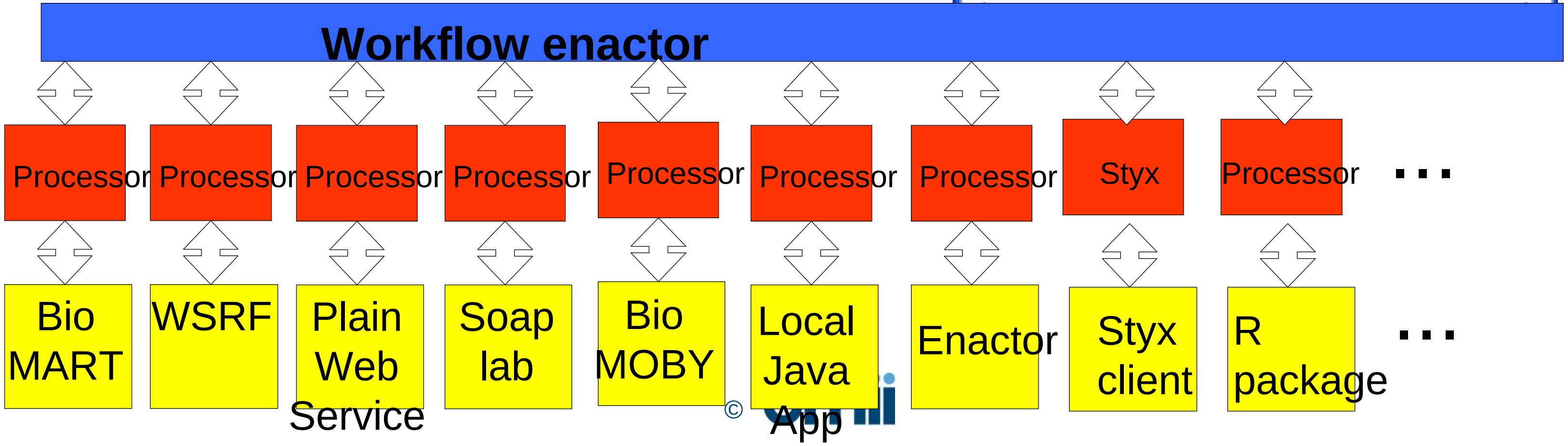
import Java API of libSBML as workflow components

Shield the Scientist

Bury the complexity

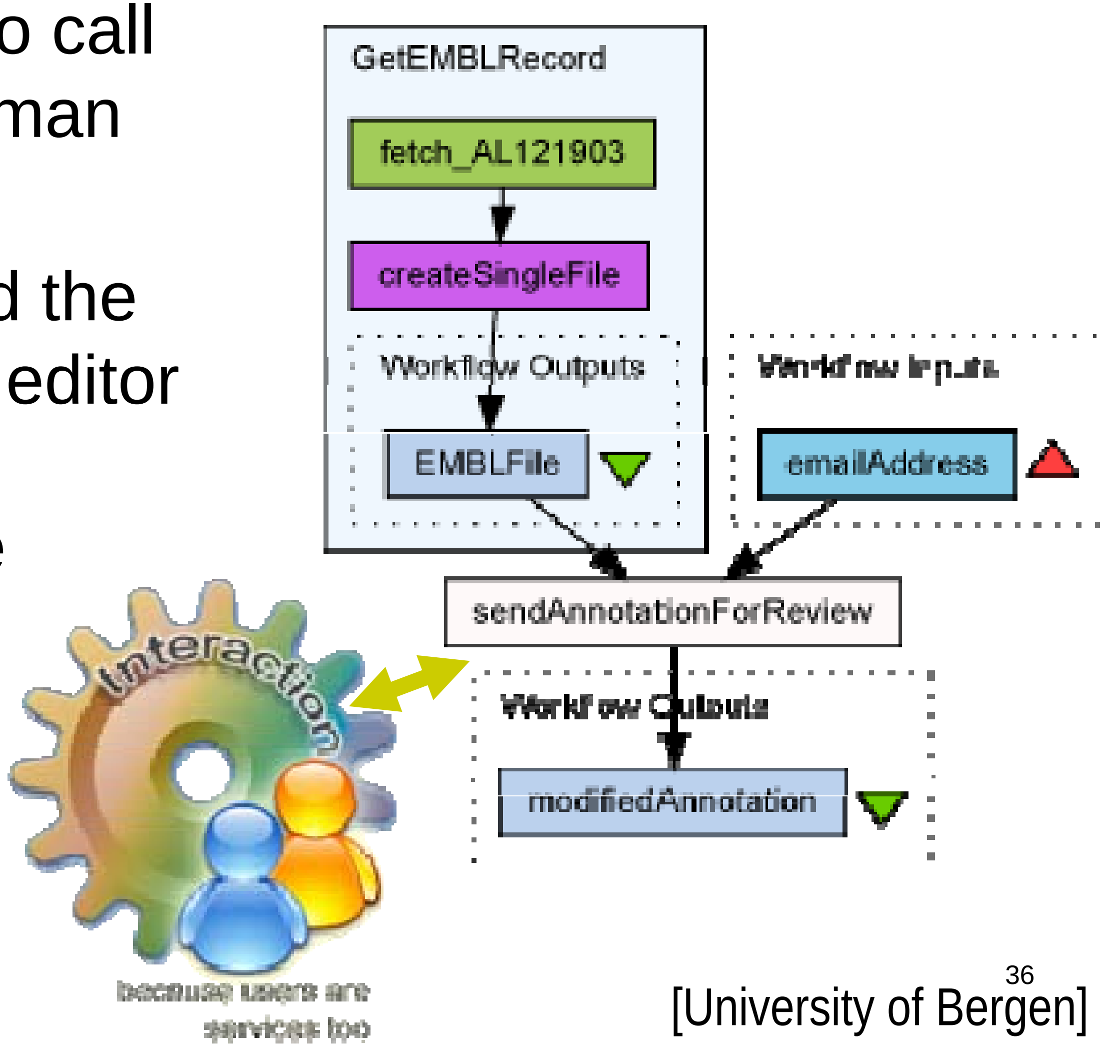

Workflow diagram

Save as DOT PNG SVG ☑ Show types Bound ports ☑ Fit to window



User Interaction

- Allows a workflow to call out to an expert human user
- E.g. Used to embed the Artemis annotation editor within an otherwise automated genome annotation pipeline



[University of Bergen]

No miracles here.

- Building good workflows
  - Pattern books
  - Best practice
  - Workflow packs
- Data integration
  - Still have to think about building models of results
- Services
  - Properly computer-accessible (Web) services
  - Maintenance

Changes to Scientific Practice

- Systematic and comprehensive automation.
  - Eliminated user bias and premature filtering of datasets and results leading to single sided, expert-driven hypotheses
- Dry people hypothesise, wet people validate.
  - “make sense of this data” -> “does this make sense?”
- Workflow factories.
  - Different dataset, different result
- Workflow market.
- Accurate provenance.

Conclusions

Distributed computing

1. Web Services

   Make your data or your code accessible to be a component in a …

2. Workflow

   For flexible, transparent and systematic encoding of protocols for linking services/processes up

Taverna http://taverna.sourceforge.net

myGrid http://www.mygrid.org.uk

OMII-UK http://www.omii.ac.uk

Acknowledgements

- Phase1 myGrid researchers, Phase2 OMII-UK, myGrid Research Team
- Tom Oinn (EBI), Martin Senger, Katy Wolstencroft
- Peter Li, Paul Fisher, Andy Brass, Robert Stevens, Mark Wilkinson
- EPSRC, Wellcome Foundation

Katy Wolsencroft

Tom Oinn

open middleware
infrastructure institute uk
www.omii.ac.uk