



Middleware for *in silico* Biology

Professor Carole Goble
University of Manchester
<http://www.mygrid.org.uk>

GGF Summer School 24th July 2004, Italy





Vision: Collaboratory

A collaboratory is

...a center without walls, in which the nation's researchers can perform their research without regard to geographical location, interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries

William Wulf, 1989
U.S. National Science Foundation



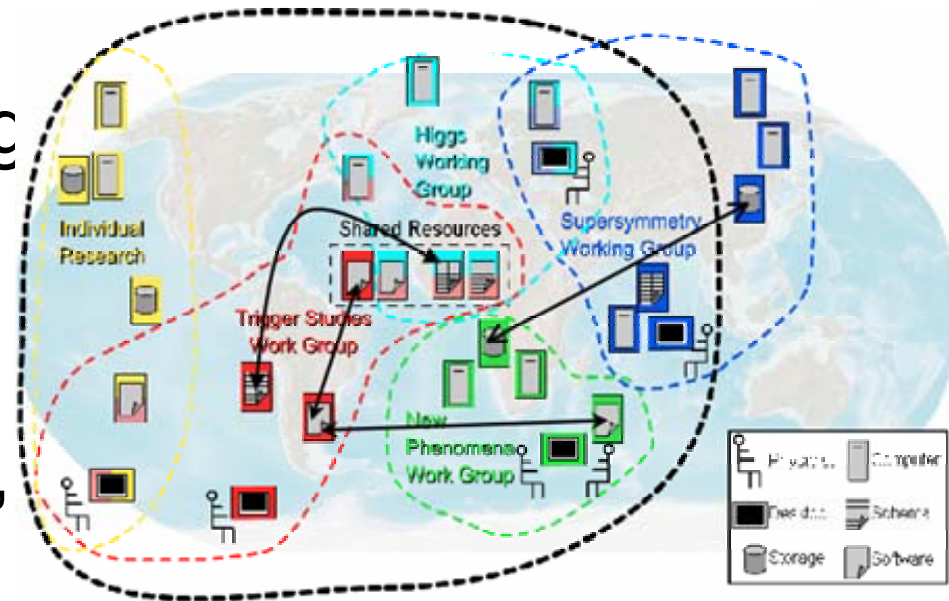
Vision: The Grid

Grid computing has emerged as an important new field, distinguished from conventional distributed computing by its focus on large-scale resource sharing, innovative applications, and, in some cases, high-performance orientation...we [define] the "Grid problem"...as flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources - what we refer to as **virtual organizations**

From "The Anatomy of the Grid: Enabling Scalable Virtual Organizations" by Foster, Kesselman and Tuecke

Knowledge workers, fluid communities

- Capturing, generating, gathering, integrating, sharing, processing, analysing, weeding, cleaning, correlating, archiving, retiring knowledge
- Much of it not theirs & not of their creation
- Much of it destined for others



- Know-how as important as know-what
- Know-why, when, where, who as important

Roadmap

- Part 1
 - Application context
- Part 2
 - Architecture
 - Information and Workflows
 - Semantics and provenance
- Part 3
 - Wrap up



myGrid is an EPSRC funded UK eScience Program Pilot Project



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



Application Testbeds



Grave's Disease

- Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle
- Autoimmune disease of the thyroid
- Discover all you can about a gene: Affymetrix microarray analysis, Gene annotation
- Services from Japan, Hong Kong, various sites in UK

Williams-Beuren Syndrome

- Hannah Tipney, May Tassabehji, Andy Brass, St Mary's Hospital, Manchester, UK
- Microdeletion of 155 Mbases on Chromosome 7
- Characterise an unknown gene: Gene alerting service, gene and protein annotation
- Services from USA, Japan, various sites in UK

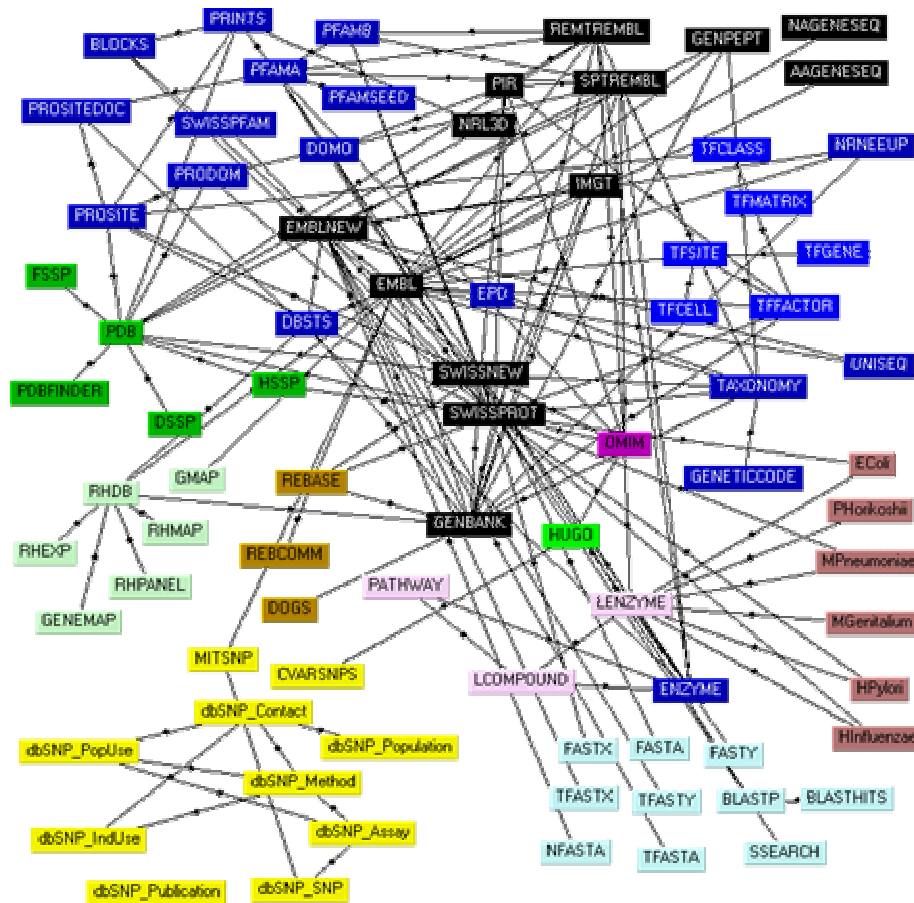


Trypanosomiasis in cattle

- Steve Kemp, University of Liverpool, UK
- Annotation pipelines and Gene expression analysis Services from USA, Japan, various sites in UK

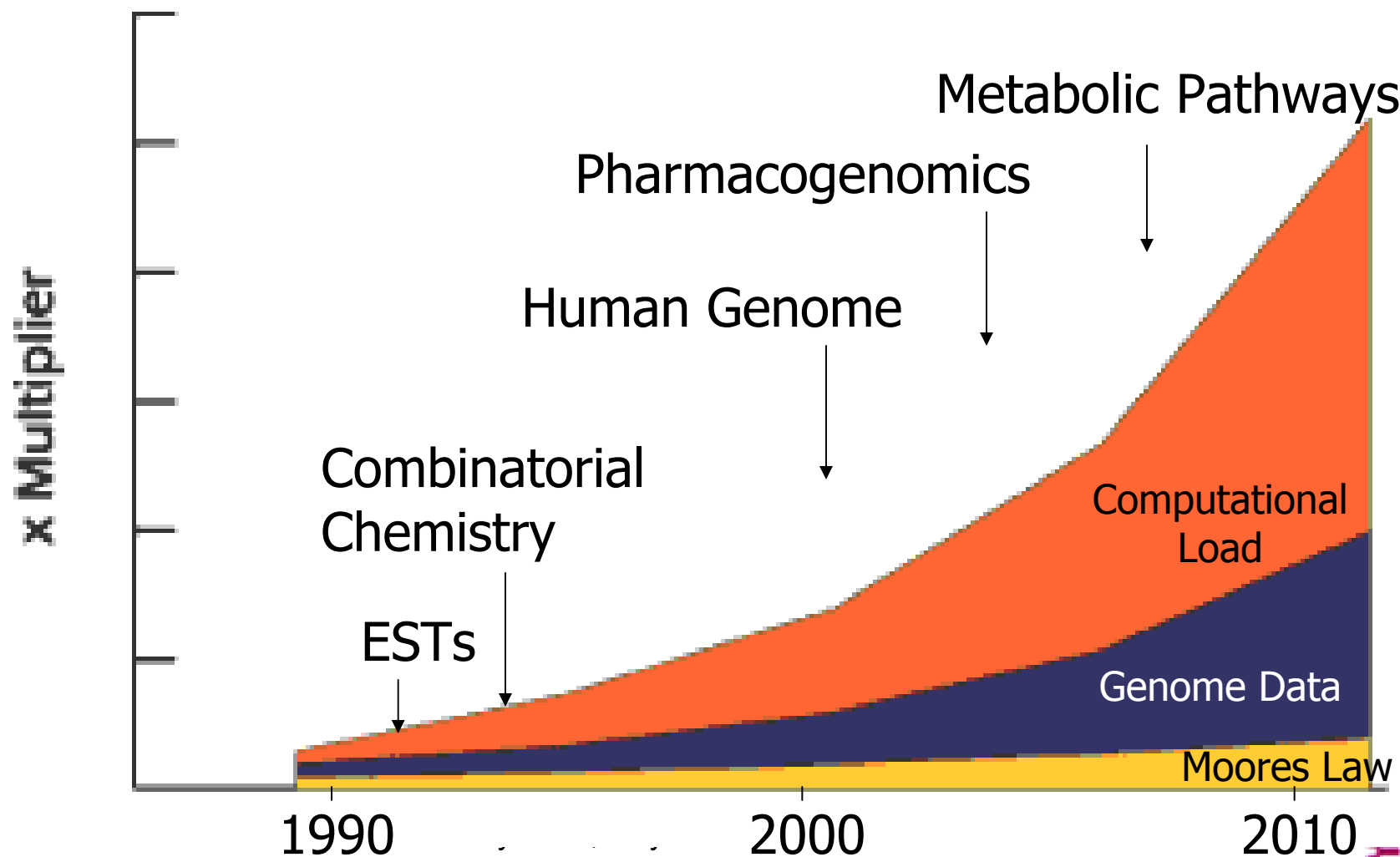
GC

Life Sciences: knowledge generation



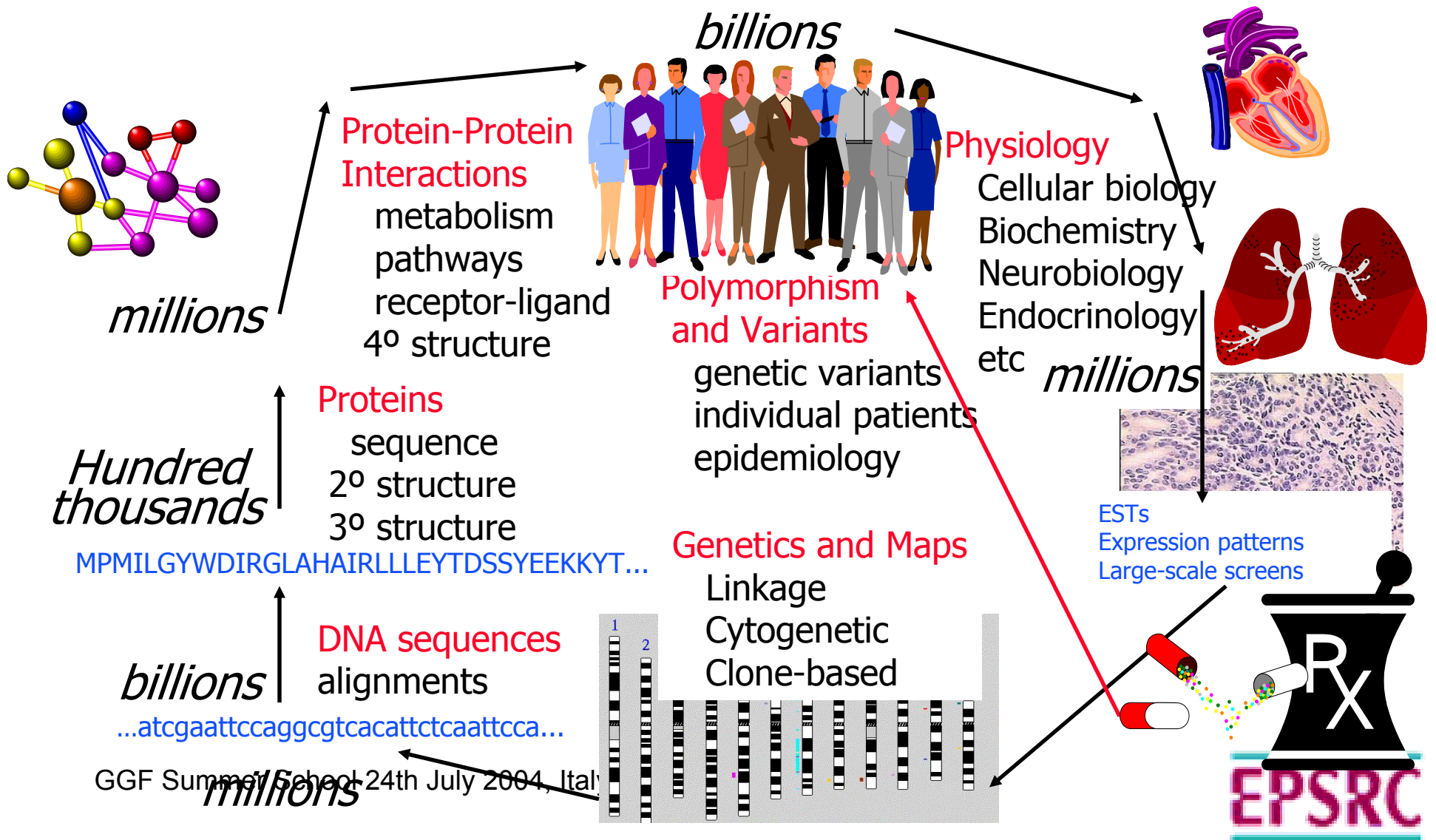
- Informational Science
- Large Scale
- Distributed
- No one organisation owns it all
- Integrating across scales, models, types, communities
- Small groups drawing on pooled resources

Data deluge, processing bottleneck



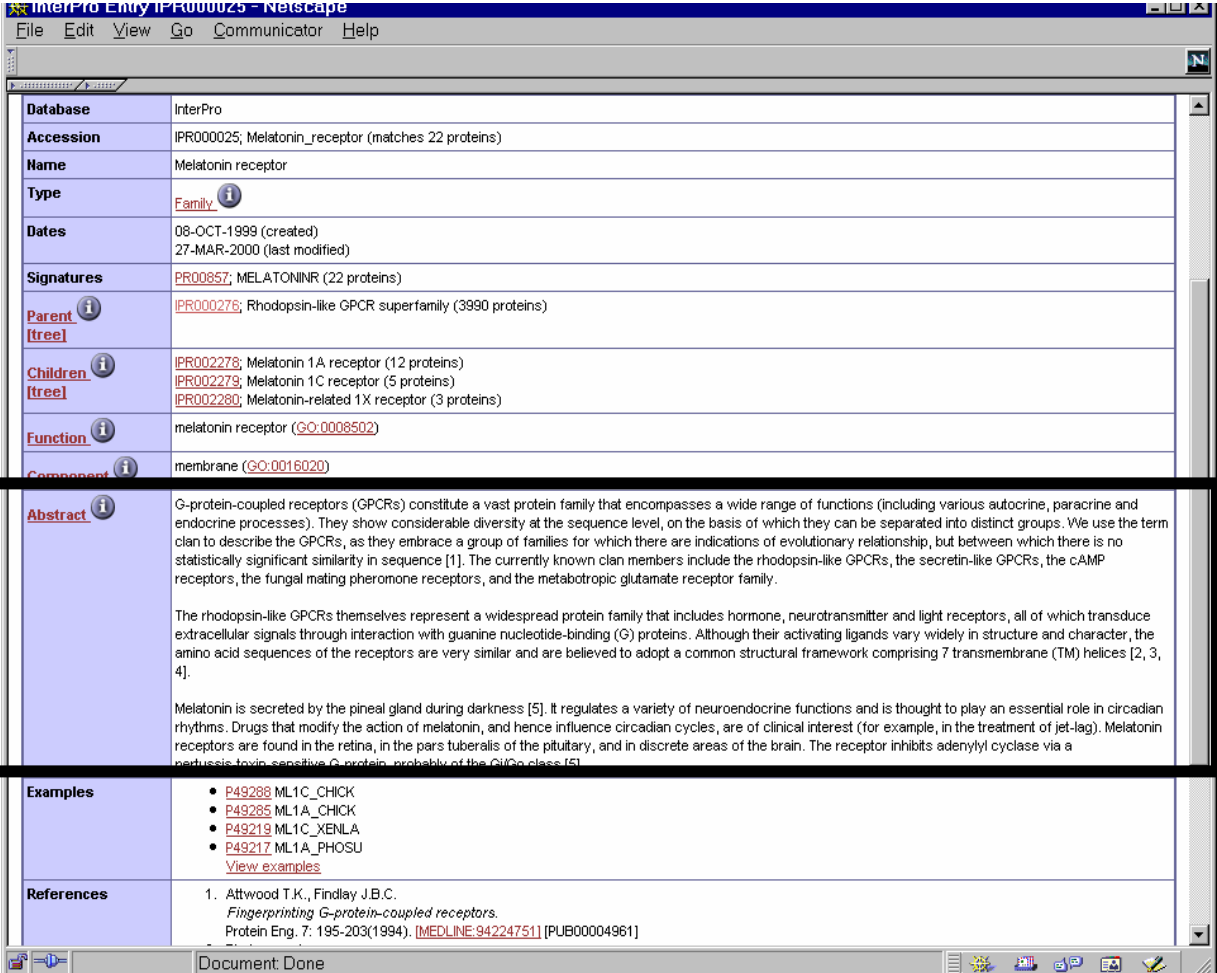


Union of lots of small experiments



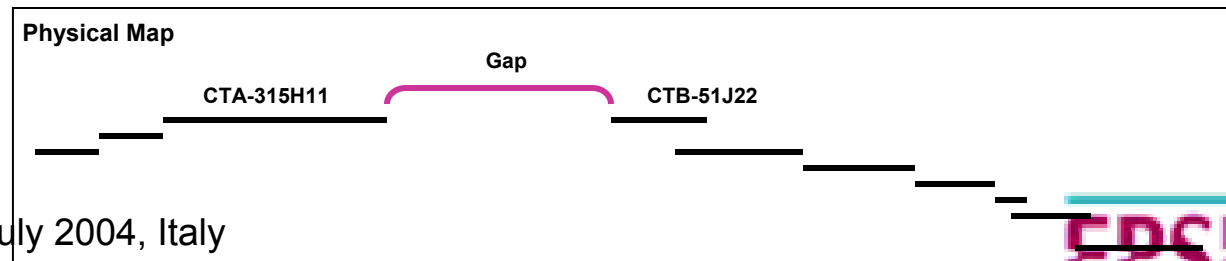
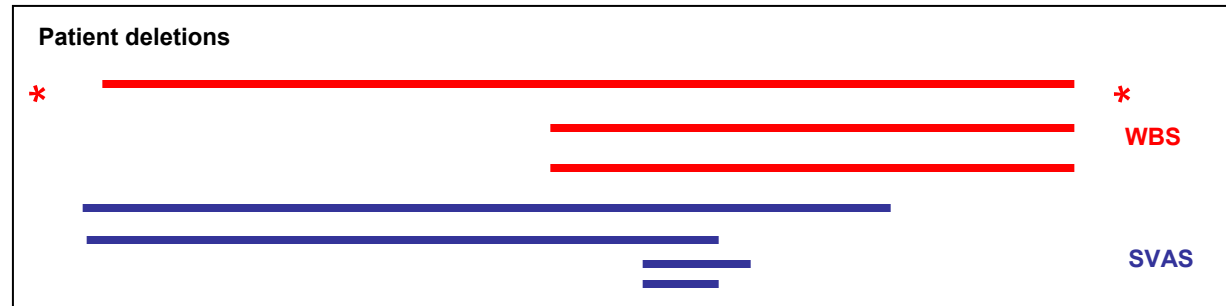
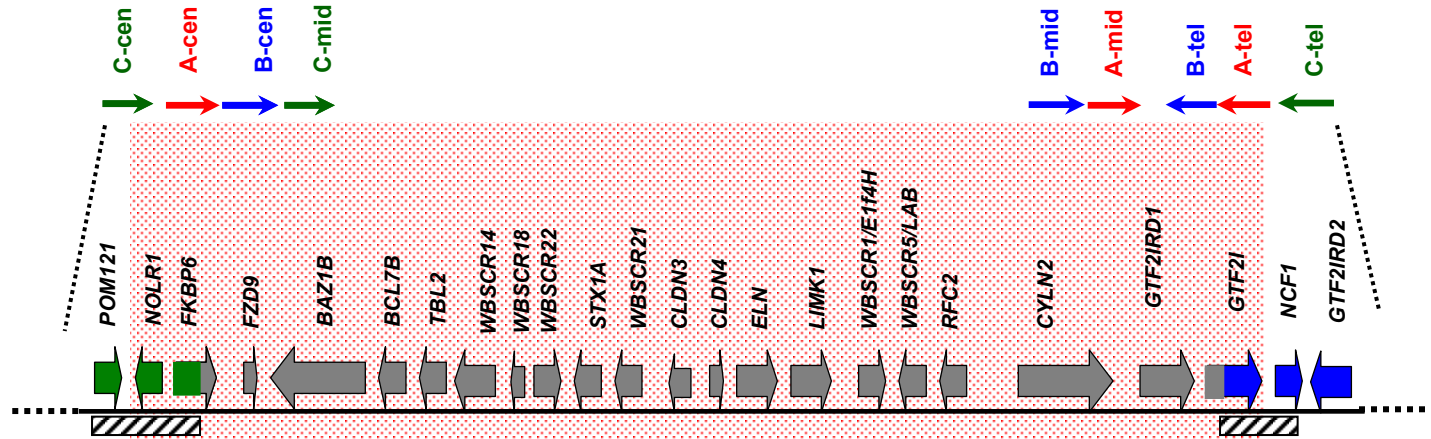
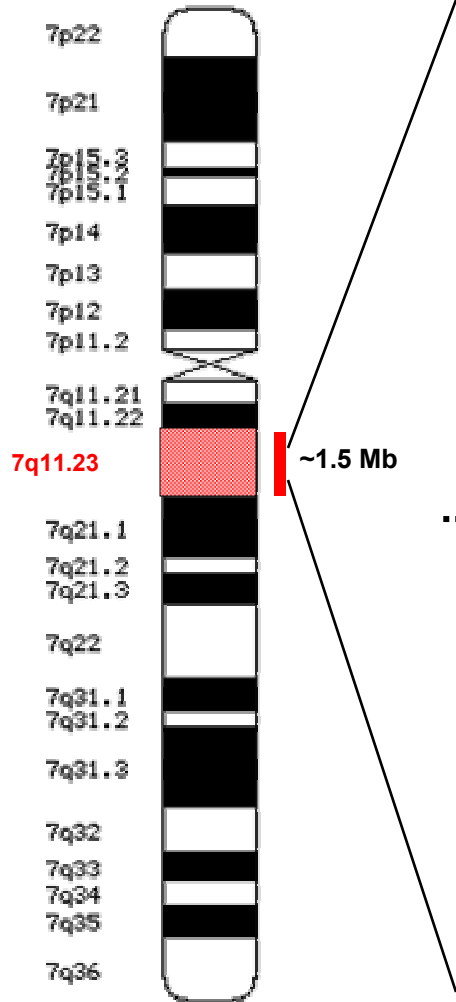
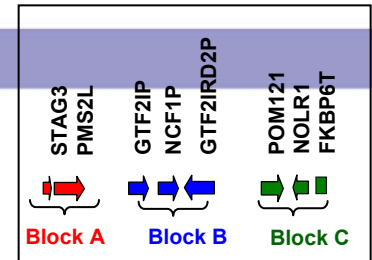
What data do I get?

- Descriptive as well as numeric
- Literature
- Analogy/ knowledge-based



Database	InterPro
Accession	IPR000025; Melatonin_receptor (matches 22 proteins)
Name	Melatonin receptor
Type	Family i
Dates	08-OCT-1999 (created) 27-MAR-2000 (last modified)
Signatures	PR00857 ; MELATONINR (22 proteins)
Parent i [tree]	IPR000276 ; Rhodopsin-like GPCR superfamily (3990 proteins)
Children i [tree]	IPR002276 ; Melatonin 1A receptor (12 proteins) IPR002279 ; Melatonin 1C receptor (5 proteins) IPR002280 ; Melatonin-related 1X receptor (3 proteins)
Function i	melatonin receptor (GO:0008502)
Component i	membrane (GO:0016020)
Abstract i	<p>G-protein-coupled receptors (GPCRs) constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the term clan to describe the GPCRs, as they embrace a group of families for which there are indications of evolutionary relationship, but between which there is no statistically significant similarity in sequence [1]. The currently known clan members include the rhodopsin-like GPCRs, the secretin-like GPCRs, the cAMP receptors, the fungal mating pheromone receptors, and the metabotropic glutamate receptor family.</p> <p>The rhodopsin-like GPCRs themselves represent a widespread protein family that includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework comprising 7 transmembrane (TM) helices [2, 3, 4].</p> <p>Melatonin is secreted by the pineal gland during darkness [5]. It regulates a variety of neuroendocrine functions and is thought to play an essential role in circadian rhythms. Drugs that modify the action of melatonin, and hence influence circadian cycles, are of clinical interest (for example, in the treatment of jet-lag). Melatonin receptors are found in the retina, in the pars tuberalis of the pituitary, and in discrete areas of the brain. The receptor inhibits adenylyl cyclase via a pertussis toxin sensitive G-protein, probably of the Gi/Go class [5].</p>
Examples	<ul style="list-style-type: none"> • P49288 ML1C_CHICK • P49285 ML1A_CHICK • P49219 ML1C_XENLA • P49217 ML1A_PHOSU View examples
References	<ol style="list-style-type: none"> 1. Attwood T.K., Findlay J.B.C. <i>Fingerprinting G-protein-coupled receptors.</i> Protein Eng. 7: 195-203(1994). [MEDLINE:94224751] [PUB00004961]

The bottleneck is not computation Its integration



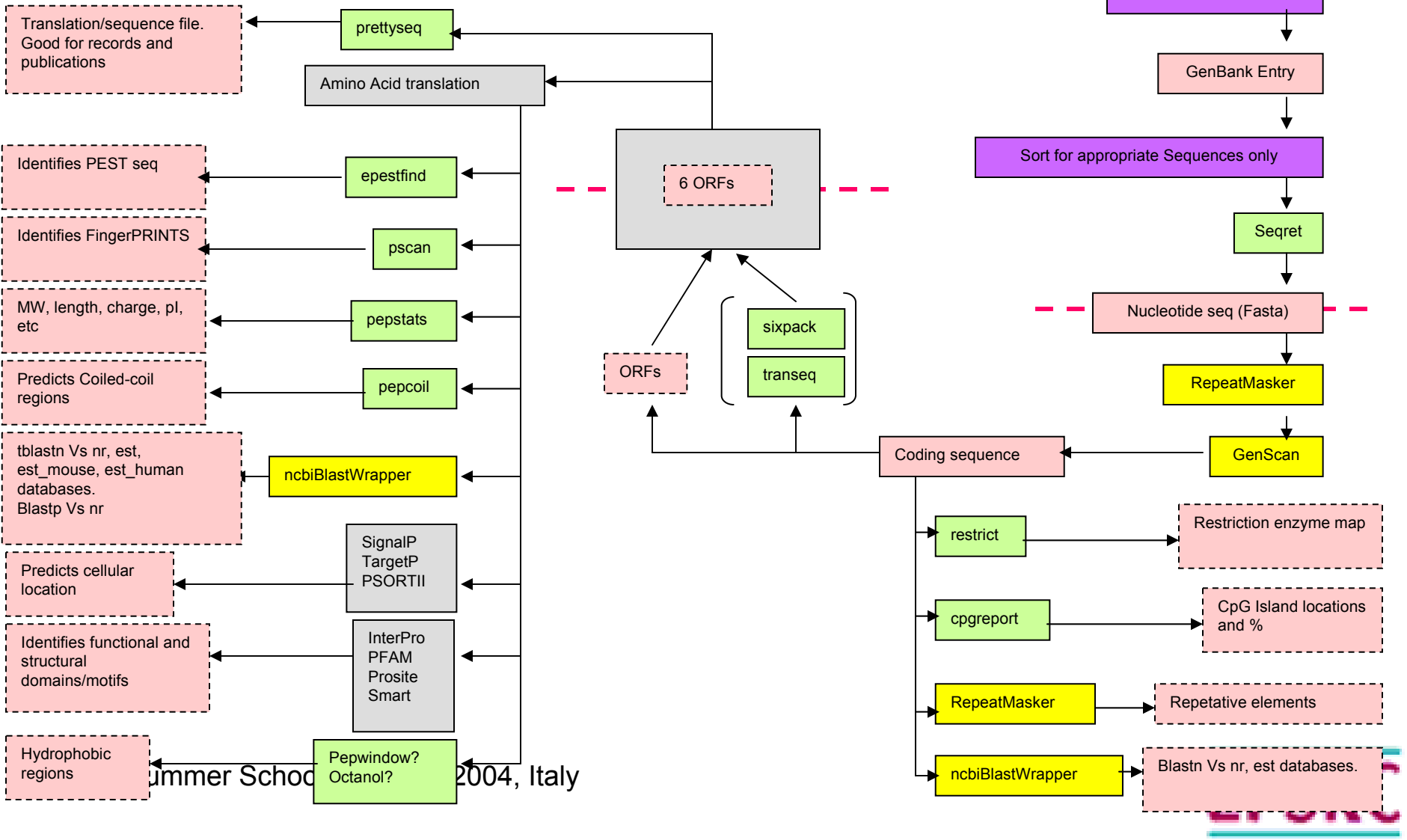
Chr 7 ~155 Mb

GGF Summer School 24th July 2004, Italy

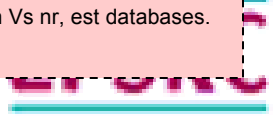


- Pink: Outputs/inputs of a service
- Purple: Taylor-made services
- Green: Emboss soaplab services
- Yellow: Manchester soaplab services
- Grey: Unknowns

Interoperability



Summer School 2004, Italy



The problem

Two major steps:

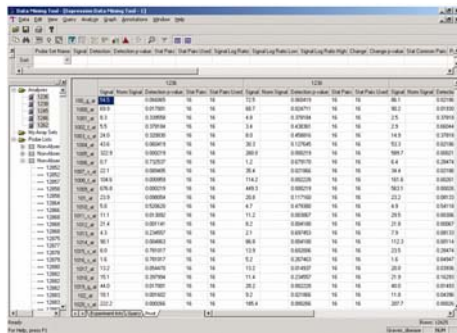
- Extend into the gap: Similarity searches; RepeatMasker, BLAST
- Characterise the new sequence: NIX, Interpro, etc...

- Numerous web-based services (i.e. BLAST, RepeatMasker)
- Cutting and pasting between screens
- Large number of steps
- Frequently repeated – info now rapidly added to public databases
- Don't always get results
- Time consuming
- Huge amount of interrelated data is produced – handled in lab book and files saved to local hard drive
- Mundane
- Much knowledge remains undocumented
- Bioinformatician does the analysis

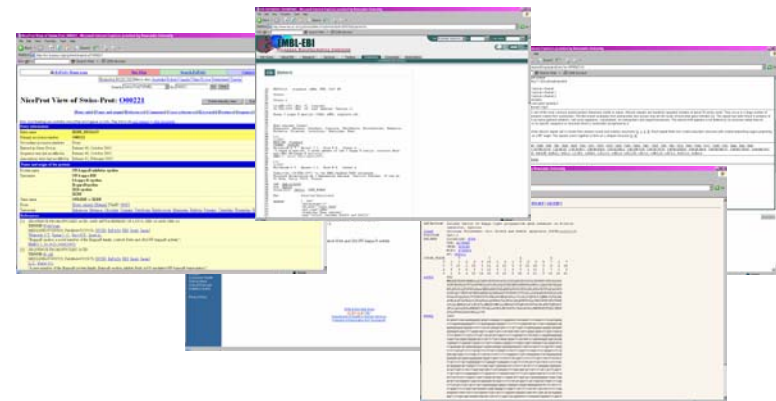
Classical Approach to the Bioinformatics

Data Analysis - Microarray

Import microarray data to Affymetrix data Mining Tool, Run Analyses and select



Study Annotations for many different Genes



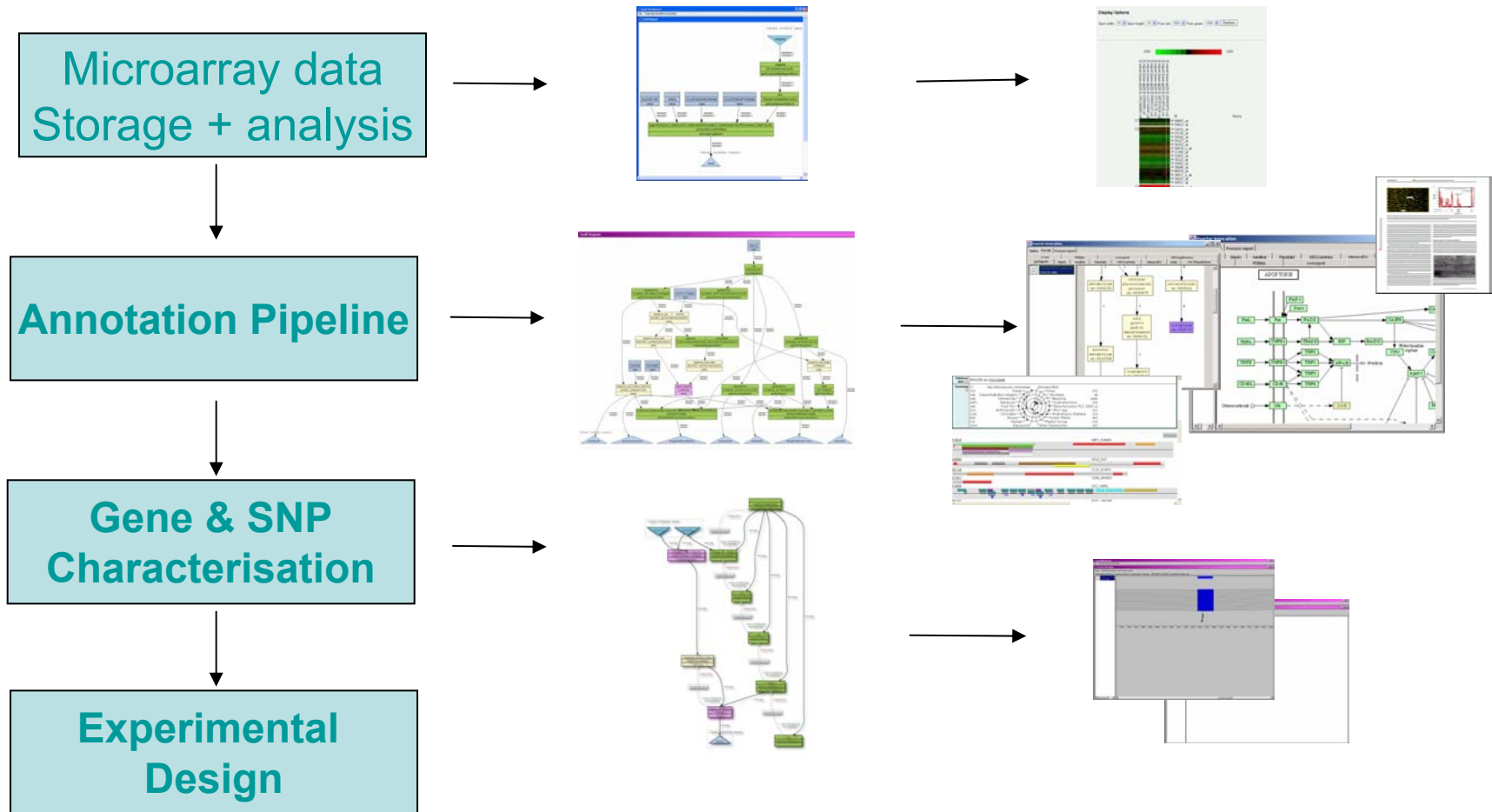
Select Gene and Visually examine SNPs lying within



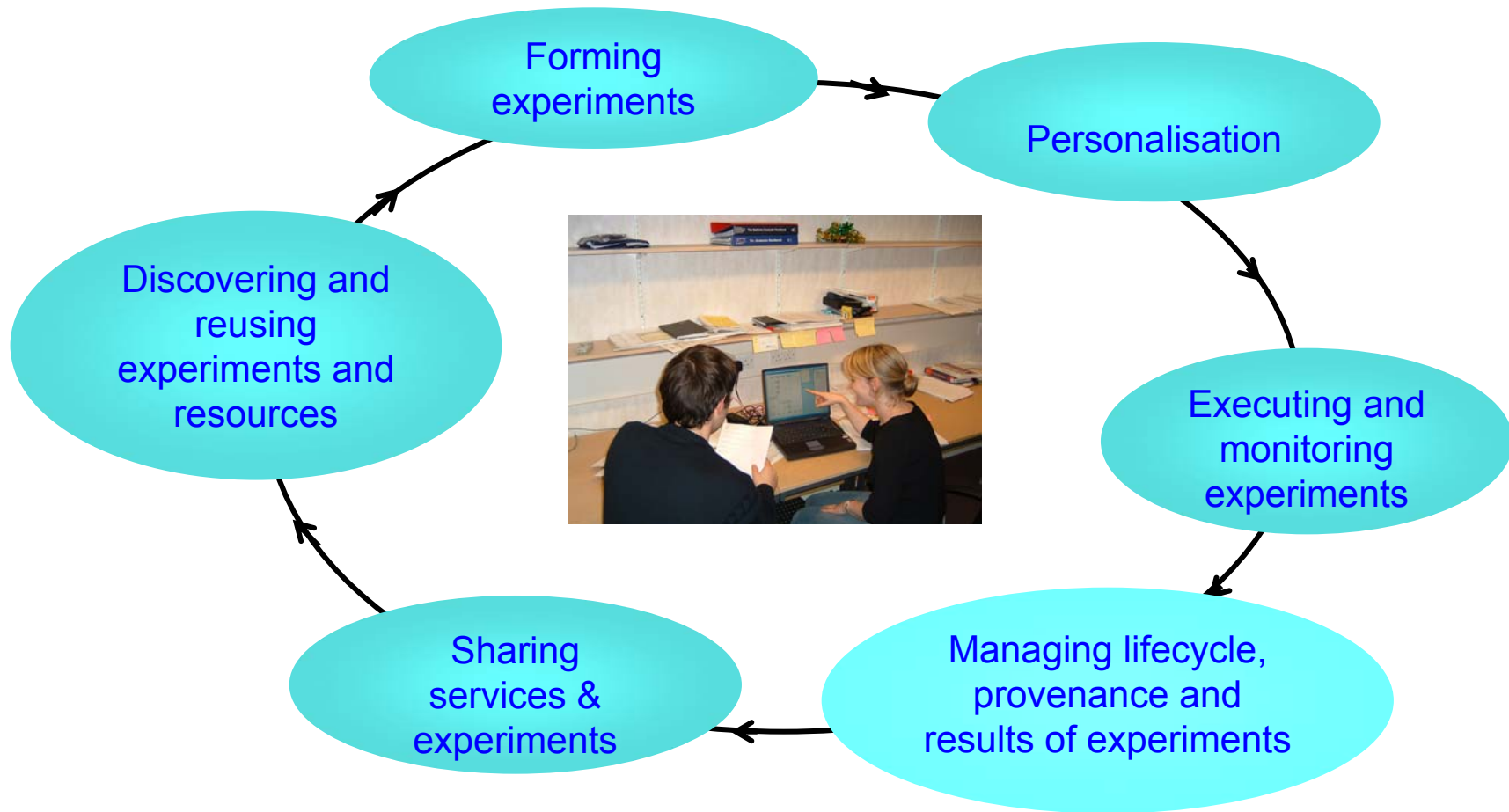
Experiment Design to test Hypotheses Find restriction sites and design primers by eye for genotyping experiments



The Graves' Disease Scenario



Experiment life cycle



The Grid is a technology; the scientist wants a solution.

Scientists...

- ...Experiment
 - Can workflow be used as an experimental method?
 - How many times has this experiment been run?
- ...Analyze
 - How do we manage the results to draw conclusions from them?
- ...Collaborate
 - Can we share workflows, results, metadata etc?
- ...Publish
 - Can we link to these workflows and results from our papers?
- ...Review
 - Can I find, comprehend and review your work?
 - How was that result derived?

Scufl Workbench

File Tools and Workflow Invocation

Run Workflow

Input Document

```
<?xml version="1.0" encoding="UTF-8"
<b:dataThingMap xmlns:b="http://org.
```

Input Document

Run Workflow

h beta8

Glover,
of the

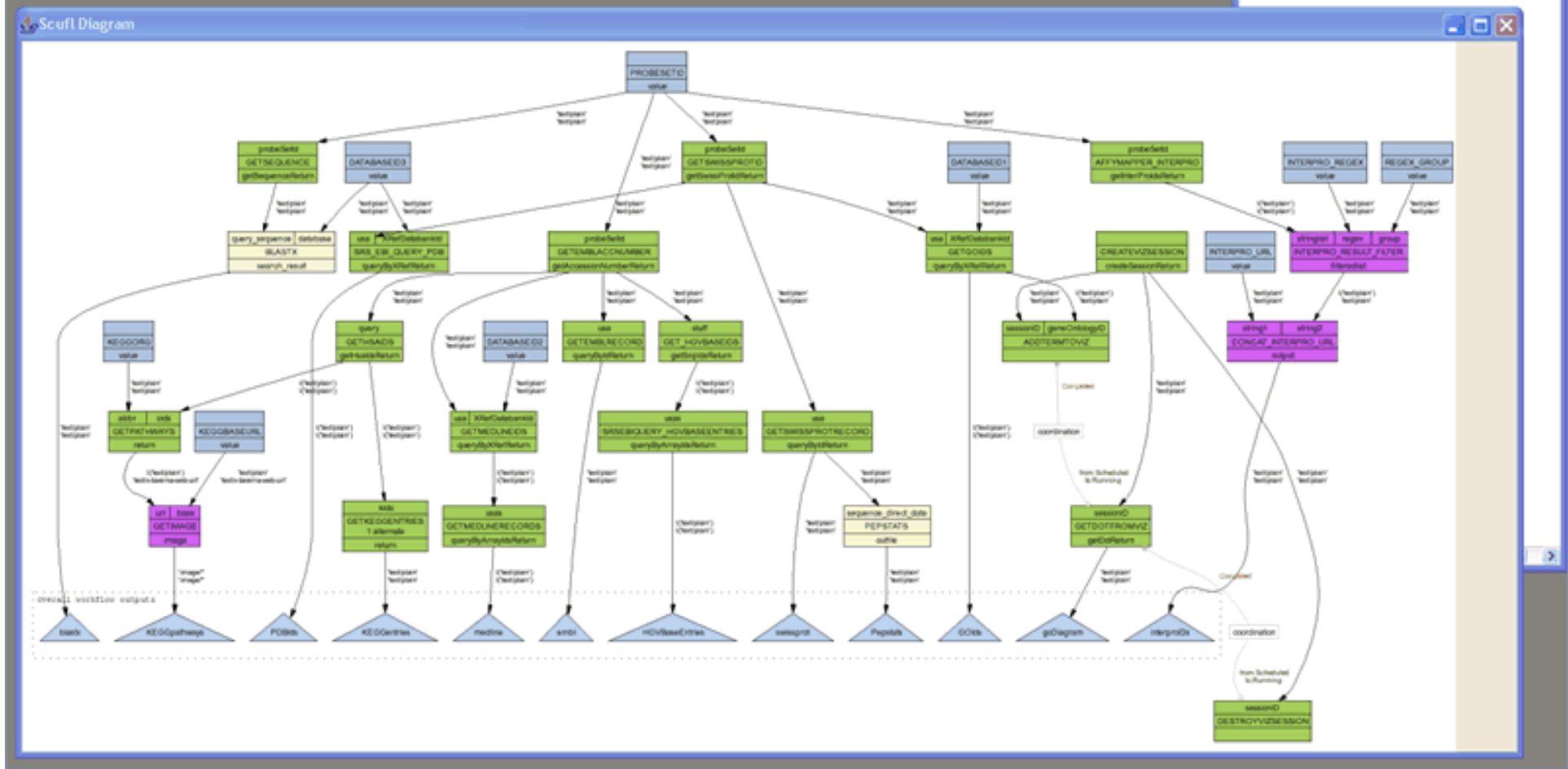
Scufl Model Explorer

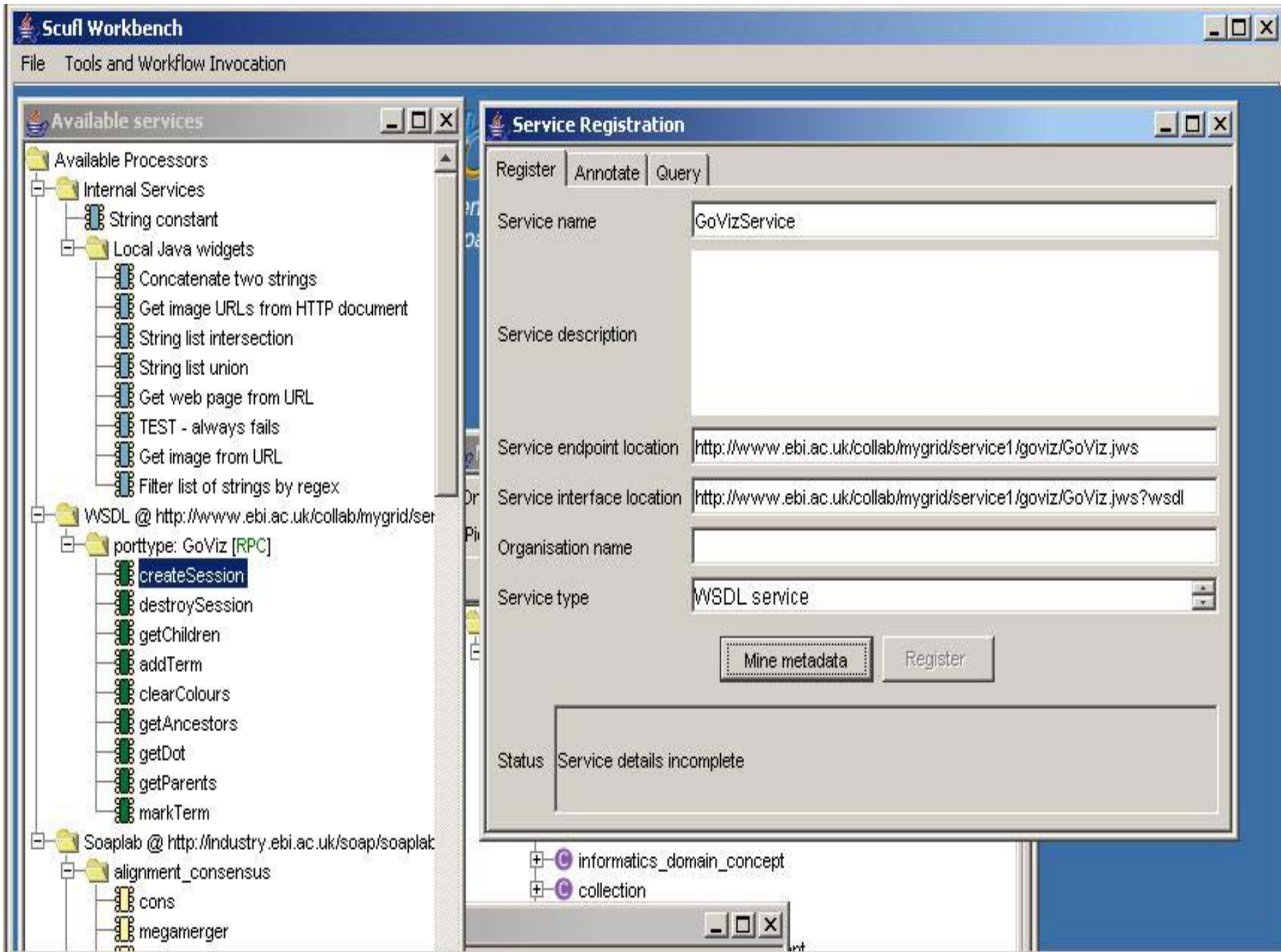
Workflow object	Retries	Delay	Backoff
Workflow model			
Workflow inputs			
goDiagram			
embl			
medline			
swissprot			
blastc			
interproDs			
PCIDs			
GOids			
KEGGentries			

Available services

Available Processors

- Internal Services
- Local Java widgets
 - Filter list of strings by regex
 - String list intersection
 - Get web page from URL
 - String list union
 - Concatenate two strings
 - Filter list of strings extracting match to a reg
 - TEST - always fails
 - Get image from URL
 - Get image URLs from HTTP document
 - String constant







http://archer2.cs.nott.ac.uk:8080/WorkflowPortal/portal.jsessionid=E42278568720E152



Home Bookmarks Download Marketplace



Welcome to myGrid!

Welcome test test

My Pages: test page

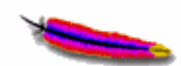
Customize: [HTML](#) [WML](#)
[Edit account: test](#)
[Logout](#)

workflow
Workflow status
[Run workflow](#)

Submission date	Status	Execution time(h..h:mm:ss)	Links
06/07/2004 10:47:55	COMPLETE	0:00:05	delete provenance report results
06/07/2004 15:08:23	COMPLETE	0:00:03	delete provenance report results
07/07/2004 11:19:05	COMPLETE	0:00:05	delete provenance report results
13/07/2004 13:49:27	COMPLETE	0:00:16	delete provenance report results
13/07/2004 13:51:16	COMPLETE	0:00:03	delete provenance report results

Apache Jetspeed Portal - Version 1.4
© Apache Software Foundation 1999-2003

[Support and Additional Information](#)



Taverna Workbench

Advanced model explorer

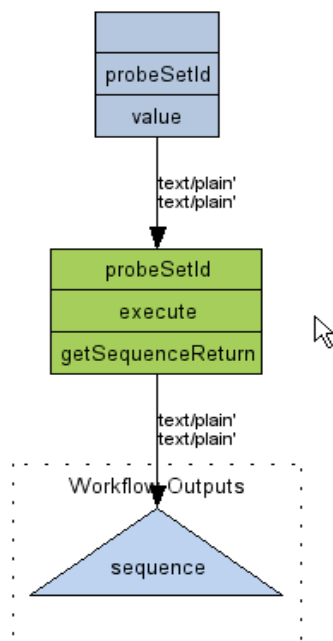
Workflow | Object properties

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

Workflow object	Retries	Delay	Backoff
Workflow model			
Workflow inputs			
Workflow outputs			
sequence			
Processors			
probeSetId : 1001_at	0	0	1
execute	0	0	1
probeSetId			
getSequenceReturn			
Data links			
probeSetId.value->execute.probeSetId			
execute.getSequenceReturn->sequen			

Workflow diagram

Save as | Show types | Bound ports | Fit to window



TavernaFetaGUI

Query | Result

AffymetrixMapper ...

BlastService

Service name: AffymetrixMapperService

Service description: Makes calls to methods in the Mapper class to retrieve various mappings from Affymetrix probe set Ids. The source of the mappings originate from the Affymetrix's NetAffix web site. Apparently, the quality of the mappings from probe set Ids to other databases is somewhat dependent on the intital mapping to Unigene. There might not be a mapping to a Unigene Id (or other databases for that matter) for every probe set Id; in these cases, it might be useful to BLAST the probe set sequence against EMBL to look for highly similar sequences and see if they are associated with a Unigene cluster.

Service endpoint location: <http://mygrid.ncl.ac.uk/axis/services/affymapper>

Service interface location: <http://mygrid.ncl.ac.uk/axis/services/affymapper?wsdl>

Organisation name: Newcastle University

Service type: WSDL service

Operation Name: getSequence

Enactor invocation
🔍 📄 🗑

Status Results Results as XML Provenance Text Provenance Tree

tmapPlot prophetOutput outputPlot

List

- application/octet-stream,image/png
- application/octet-stream,image/png

Save to file

CPS2-DRDME	AAHEKAMRREQA KKMNVKSLRSEED-CDKSAEGKLAKVALLTISLWFMAWT	287
CPS2-DRDPS	AAHEKAMRREQA KKMNVKSLRSEED-CDKSAENKLAKVALLTISLWFMAWT	287
CPS2-LIMPO	AHEKAMRREQA KKMNVKSLRANEDQKQSAECRLAKVAMMTVSLWFMAWT	286
CPS2-HEMSA	FAHEKAMRREQA KKMNVSLRSNEA-DAAQAEIRIAKVALYNVSLWFLCWT	285
CPS2-SCHGR	REHEKAMRREQA KKMNVKSLQSNADTEAQAECIRIAKVALTFFLFLCWT	283
CPS2-PATYE	VCKD-----BRKNGIRAKRYTTRPFIQDTEQRYFFLSFLMMAKFMVAWT	254
CPS2-DRDME	PYLVIICYFGLFKIDG-LTPLTTIWGATFAKTSAYYNPIVYGISHPKYRIY	346
CPS2-DRDPS	PYLVIICYFGLFKIDG-LTPLTTIWGATFAKTSAYYNPIVYGISHPNDRLY	346
CPS2-LIMPO	PYLIIAWAGVFSSTRLTPLATIWSYFAKNSCYNPIVYGISHPRYKAA	338
CPS2-HEMSA	PYALISCKGVMGDTSGITPLVSTLPALEAKSCSCYNPFVYAI SHPKYRLA	345
CPS2-SCHGR	PYAVVAMISAFBNRAALTLPLSTMPAVTAKVSCDPWVYAINHPFRRE	343
CPS2-PATYE	PYAIMSKLALISSFNV--ENSFAALPLFLAKASCAYNPFIIYAFTRKNERDT	302
CPS2-DRDME	LKERCPMVCVFNTDEPKPDAPASDTETTFSEADSKA-----	361
CPS2-DRDPS	LKERCPMVCVCGTTDEPKPDAPPSTETTFSEADSKD-----	361
CPS2-LIMPO	LYQRFPSLACSSSESGSDVKSASATMTMEEKPKSPEA-----	376
CPS2-HEMSA	ITRHLPWFCYHETEETKSNDDSSQSNSTVAQDKA-----	377
CPS2-SCHGR	VQRMKWLHLGEDARSSKSDTSSSATDRIVGNYSKASA-----	360
CPS2-PATYE	VVEIMAPWTTTRRYGVSTLPWPQVTTYPRRRTSKVMTTDIEFPDDNIFIVN	352
CPS2-DRDME	-----	361
CPS2-DRDPS	-----	361
CPS2-LIMPO	-----	376
CPS2-HEMSA	-----	377
CPS2-SCHGR	-----	360
CPS2-PATYE	SSVNGPTVKREKIYQRNPI NVRLGIKIEPRDSRAATENTFTA DFSVI	369

Scufl Workbench

File Tools and Workflow Invocation

Run Workflow

Input Document	Accession	Size	Species	Chromosome
bare_seq_in	7717376	44.1	Homo sapiens	chromosome 21
old_result	5629923	44.1	Homo sapiens	12q22 BAC RPCI
species	34367431	44.1	Homo sapiens	mRNA
chromosome	16304790	44.1	Human	chromosome 14
	34533695	44.1	Homo sapiens	cDNA F
	20377057	44.1	Homo sapiens	chromo
	17977487	44.1	Homo sapiens	BAC cl
	17048246	44.1	Homo sapiens	chromo
	5757554	44.1	Homo sapiens	PAC clone RP3-
	4176355	44.1	Homo sapiens	chromosome 4 c
	4191263	44.1	Human	DNA sequence from clo
	14485328	44.1	Human	DNA sequence
	2029100	44.1	Homo sapiens	chromosome 21q
	2828772	44.1	Homo sapiens	chromosome 17,
	9798442	42.1	Human	DNA sequence from clo

Scufl Model Explorer

Workflow object

Workflow inputs

- bare_seq_in
- old_result
- species
- chromosome

Workflow outputs

- simple
- comparison
- genbank
- fasta_out
- missed

Available services

Available Processors

- Local Java widgets
- Filter list of strings by regex
- String list intersection
- Get web page from URL
- String list union
- Concatenate two strings
- Filter list of strings extracting match to a reg
- TEST - always fails
- Get image from URL
- Get image URLs from HTTP document
- String constant

```

    graph TD
      bare_seq_in --> masked[REPEATMASKER]
      masked --> ncbi_blast[NCBI BLAST]
      ncbi_blast --> simplifier[SIMPLIFIER]
      simplifier --> comparer[COMPARER]
      old_result --> comparer
      species --> retriever[RETRIEVE]
      chromosome --> retriever
      comparer --> simple[simple]
      comparer --> comparison[comparison]
      retriever --> genbank[genbank]
      retriever --> fasta_out[fasta_out]
      retriever --> missed[missed]
  
```

Enactor invocation

Status

Processor stati

Type	Name	Last event	Event timestamp	Event detail
	simplifier	ProcessComplete	31-Mar-2004 16:30:53	
	comparer	ProcessComplete	31-Mar-2004 16:30:54	
	ncbiblast	ProcessComplete	31-Mar-2004 16:30:49	
	repeatmasker	ProcessComplete	31-Mar-2004 16:30:47	
	retrieve	Invoking	31-Mar-2004 16:30:54	

Intermediate inputs: masked

Intermediate outputs: >UnnamedSeq1

```

AAAGCTTTTCTGGCACTGTTTCCTTCTCTGATAACCAGAGAAGGAAAAG
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GGCAAGCCCTGCTCCTCCGGGCTTCACTCTGCACACTGTAACTGGG
GTTAAATGGGCTCACTGGACTGTTGAGCGGAGCTGGAGGAGGCTCGGA
AGCAACATGGGTGGTCACTTCTCTGATTCAGGGAGAAAACACACAAGAGG
  
```

Enactor invocation

Save all results

Status | Provenance Tree | Process report

Processor stati

Type	Name	Last event	Event timestamp	Event detail
	comparer	ProcessScheduled	09-Jul-2004 23:24:04	
	simplifier	ServiceError	09-Jul-2004 23:24:57	Message='Output 'outp...
	repeatmasker	ProcessComplete	09-Jul-2004 23:24:18	
	ncbiblast	ProcessComplete	09-Jul-2004 23:24:53	
	ebi_blast_ncbi	Invoking	09-Jul-2004 23:24:18	

Intermediate inputs | Intermediate outputs

query_sequence

test plan
Click to view...

```
>UnnamedSeq1
AAGCTTTTCTGGCACTGTTTCCTTCTTCCTGATAACCAGAGAAGGAAAAG
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGACCTT
GGCAAGCCTGTCTCCTCCGGGCTTCACTCTGCACACCTGTAACCTGGG
GTAAATGGGCTCACCTGGACTGTTGAGCGGAGCTGGGAGGAGGTCTGGA
```

Enactor invocation

Status | Results | Results as XML | Provenance Text | Provenance

- workflowReport workflowID="FlowID.org.embl.ebi.escience"
 - processorList
 - processor name="string1"
 - processor name="string2"
 - processor name="failingthing"
 - ProcessComplete TimeStamp="13-Feb-2004 13:56:25"
 - Invoking TimeStamp="13-Feb-2004 13:56:25"
 - AlternateProcessScheduled TimeStamp="13-Feb-2004 13:56:25"
 - s:local maxretries="0", retrybackoff="0.0", retrydelay="0", xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha"
 - org.embl.ebi.escience.scuflworkers.java.StringConcat
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:25"
 - WaitingToRetry MaxRetries="2", RetryNumber="2", TimeDelay="2000", TimeStamp="13-Feb-2004 13:56:23"
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:23"
 - WaitingToRetry MaxRetries="2", RetryNumber="1", TimeDelay="1000", TimeStamp="13-Feb-2004 13:56:22"
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:22"
 - Invoking TimeStamp="13-Feb-2004 13:56:22"
 - ProcessScheduled TimeStamp="13-Feb-2004 13:56:22"
 - s:local maxretries="2", retrybackoff="2.0", retrydelay="1000", xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha"
 - org.embl.ebi.escience.scuflworkers.java.TestAlwaysFailingProcessor

urn:lsid:pdb.org:PDB:2ACE - Microsoft Internet Explorer

Address: [lsid:um:lsid:pdb.org:PDB:2ACE](urn:lsid:um:lsid:pdb.org:PDB:2ACE)

IBM LSID Launchpad CONFIGURE HELP I3C


<urn:lsid:pdb.org:pdb:2ace>
Formatted content is available for this resource, and is stored as

- <urn:lsid:pdb.org:pdb:2ace-pdb>
Type: urn:lsid:i3c.org:types:content
Format: urn:lsid:i3c.org:formats:pdb
This resource represents content in PDB
It can be viewed using RasWin.
Launch... Save As...
- <urn:lsid:pdb.org:pdb:2ace-mmCIF>
Type: urn:lsid:i3c.org:types:content
Format: urn:lsid:i3c.org:formats:mmCIF
This resource represents content in mmCIF
No application is registered to view this
Launch... Save As...
- <urn:lsid:pdb.org:pdb:2ace-fasta>
Type: urn:lsid:i3c.org:types:content
Format: urn:lsid:i3c.org:formats:fasta
This resource represents content in FASTA
No application is registered to view this
Launch... Save As...
- <urn:lsid:pdb.org:pdb:2ace-jpg>
Type: urn:lsid:i3c.org:types:content
Format: urn:lsid:i3c.org:formats:jpg
This resource represents content in JPEG image format.
It can be viewed using Internet Explorer.
Launch... Save As...

2ACE X-RAY DIFFRACTION
File Edit Display Colours Options Settings Export Help

Local intranet

Schema.ad Schema.ad urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12



urn:lsid:ncbi.nlm

Commands

- File away
- Rename
- View in QMol

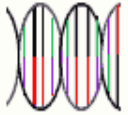
Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Available views

- Browse view
- Calendar



urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12

Pubmed

None specified; click here to add Edit ▾

urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:pred...

taxon:9606 Edit ▾

Sequence Summary

Name: None specified; click here to add ▾

urn:lsid:ncbi.r AC009070 Edit ▾

Extracted resources

Extracted resources

Commands

- Add existing items
- Add new item
- Clear collection/list
- Create checkbox aspect
- File away
- Rename

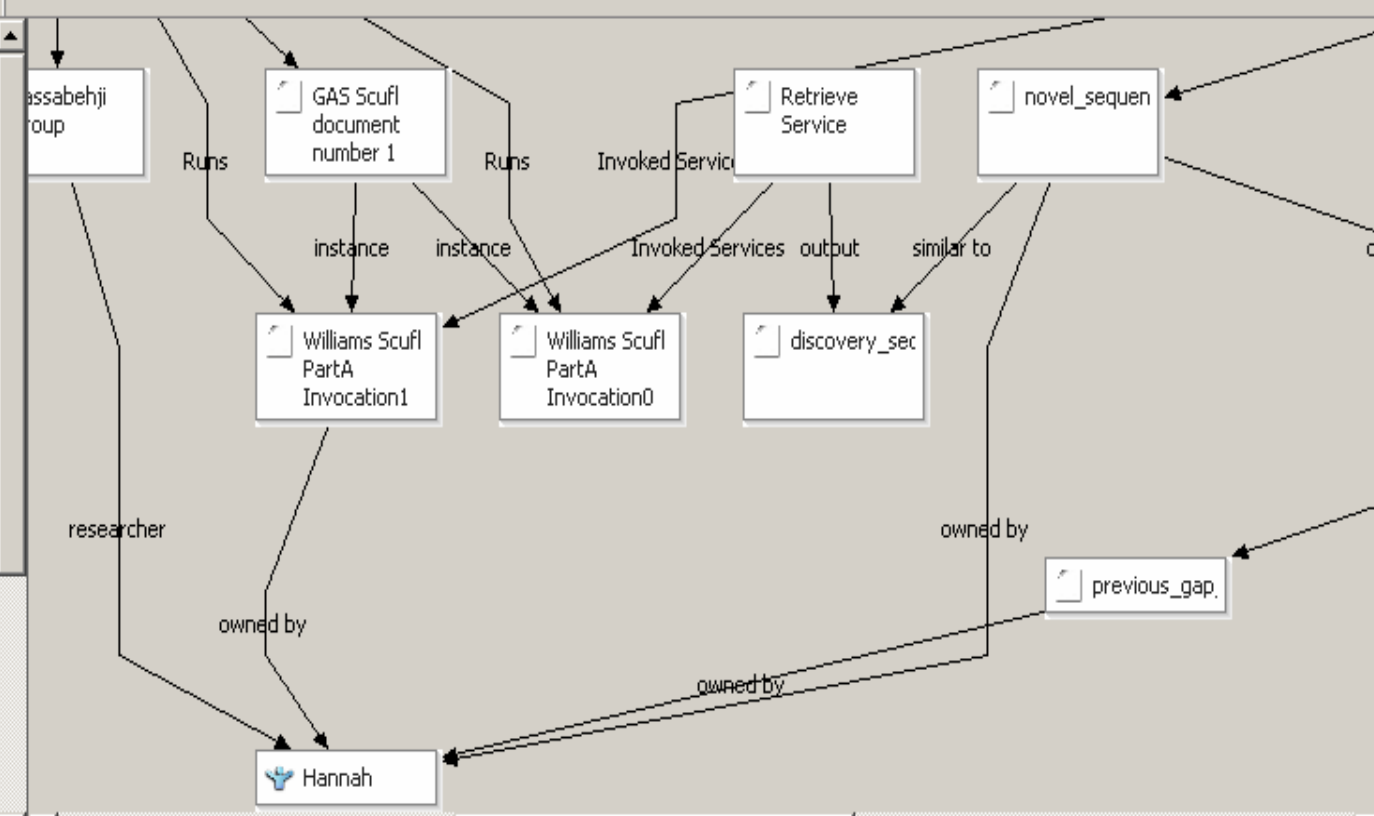
Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Available views

- Browse view
- Calendar



```

graph TD
    subgraph "assabehji group"
        A[assabehji group]
    end
    subgraph "GAS ScufI"
        B[GAS ScufI document number 1]
    end
    subgraph "Retrieve Service"
        C[Retrieve Service]
    end
    subgraph "novel_sequen"
        D[novel_sequen]
    end
    subgraph "Williams ScufI PartA"
        E[Williams ScufI PartA Invocation1]
        F[Williams ScufI PartA Invocation0]
    end
    subgraph "discovery_sec"
        G[discovery_sec]
    end
    subgraph "previous_gap"
        H[previous_gap]
    end
    subgraph "Hannah"
        I[Hannah]
    end

    A -- Runs --> B
    A -- Runs --> C
    B -- instance --> E
    C -- Invoked Service --> E
    C -- Invoked Services --> F
    C -- output --> G
    D -- similar to --> G
    E -- owned by --> I
    F -- owned by --> I
    G -- owned by --> I
    H -- owned by --> I
    
```

Choose an arr...

- Bioinformatic arrows
- Provenance Graph
- Show arrows based on the ontology
- Show arrows based on the schema

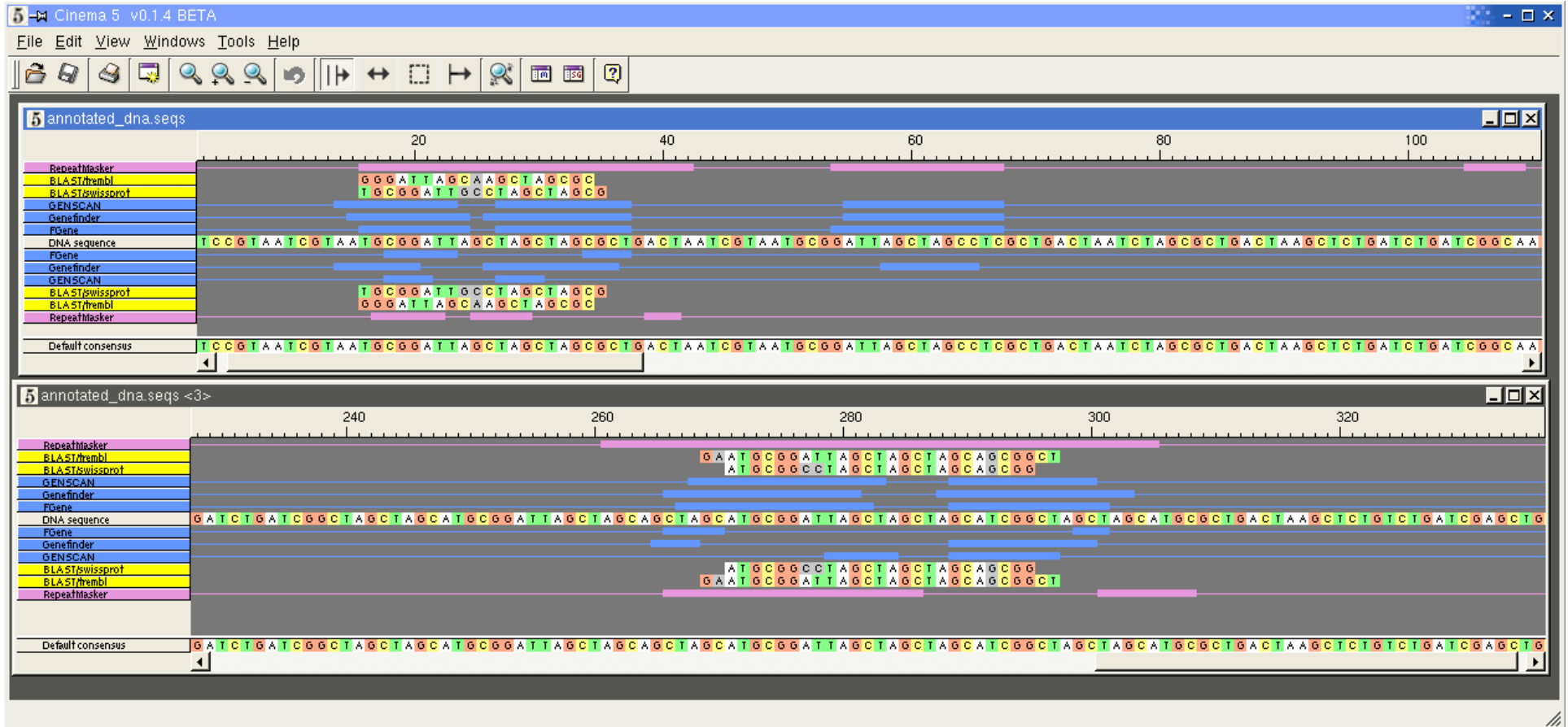
Add ▾

Available arrows for Provenance Graph

- ↪ Designs
- ↪ Invoked Services
- 11 more it...

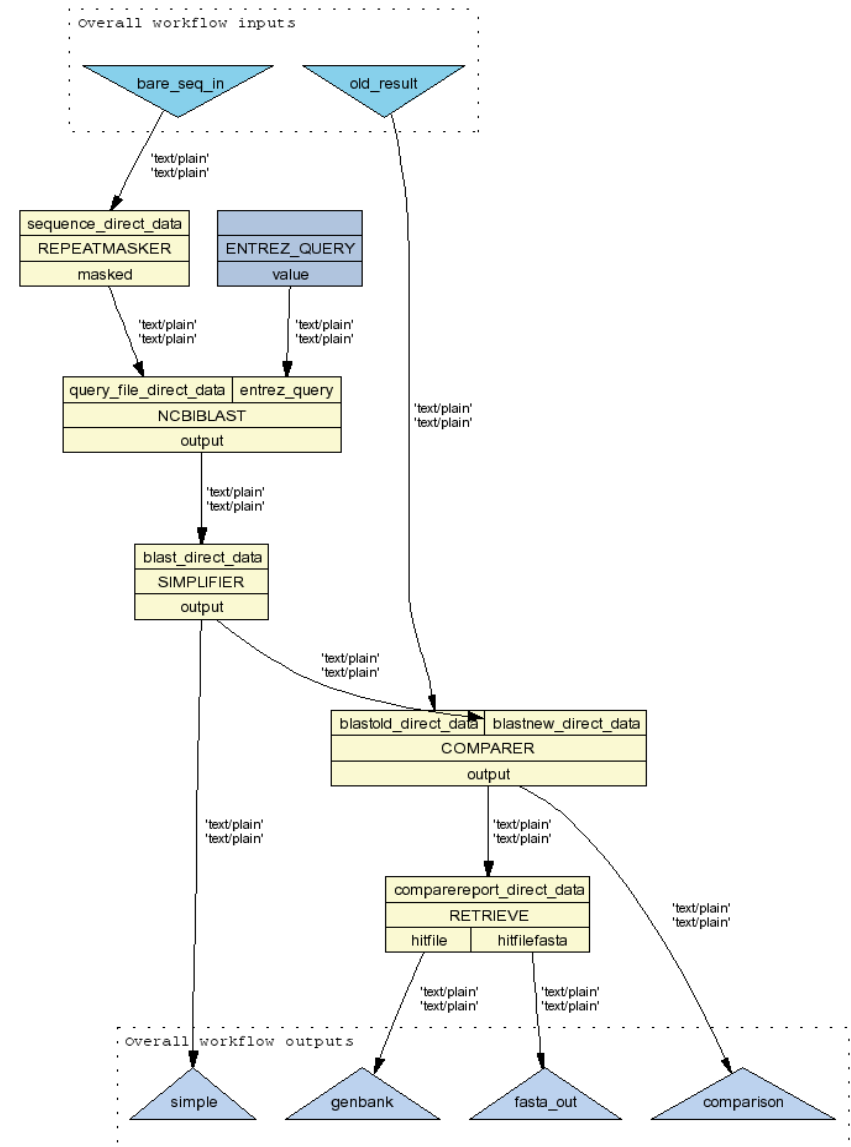


Results displayed using Cinema



WBS Life Cycle

- Wrap services as web services
- Register them
- Build a workflow using the services
- Evolve the workflow
- Run it over and over again in case data has changed
- Record results & provenance
- Inspect and compare results & provenance
- Set up event notification to fire the workflow
- Set up a portal to run the workflow
- Publish the workflow template in a registry to share with the world



Delivering results

William-Beuren Syndrome

- Cuts down the time taken to perform one pipeline from 2 weeks to 2 hours
- Much more systematic collection and analysis. More regularly undertaken. Less boring. Less prone to mistakes.
- Once notification installed won't even have to initiate it.
- Possible lead already found – but I can't tell you.
- Benchmark: first run though of two iterations of workflows
 - Reduced gap by 267 693 bp at its centromeric end
 - Correctly located all seven known genes in this region
 - Identified 33 of the 36 known exons residing in this location

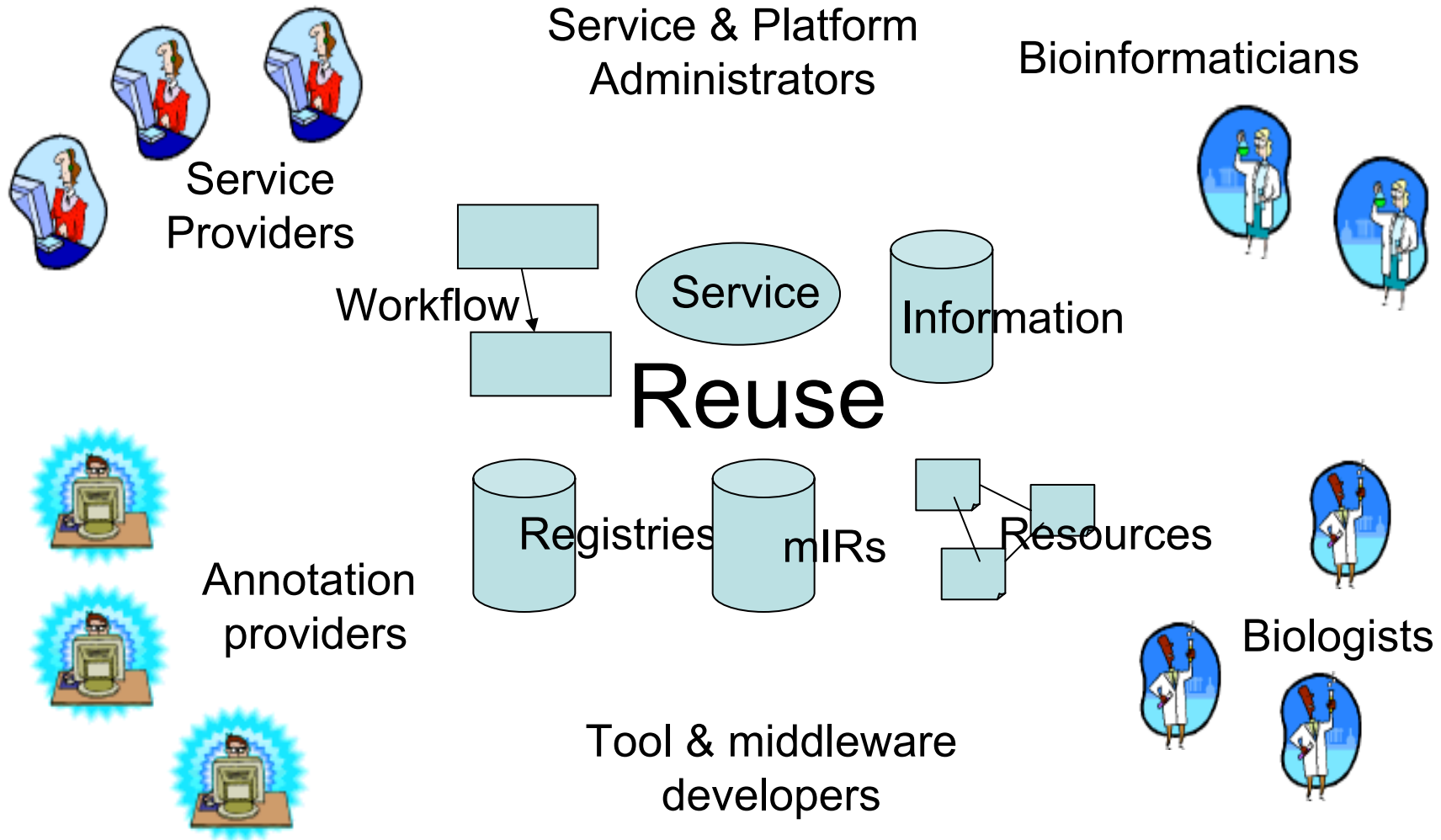
Delivering results

- Easy to get started with Taverna
- Sharing happens
 - IPR issues, and suspicions still abound
- Network effect necessary and happens

- Managed the transition from generic middleware development to practical day to day useful services.
- Architecture is solid.
- SOA – good idea

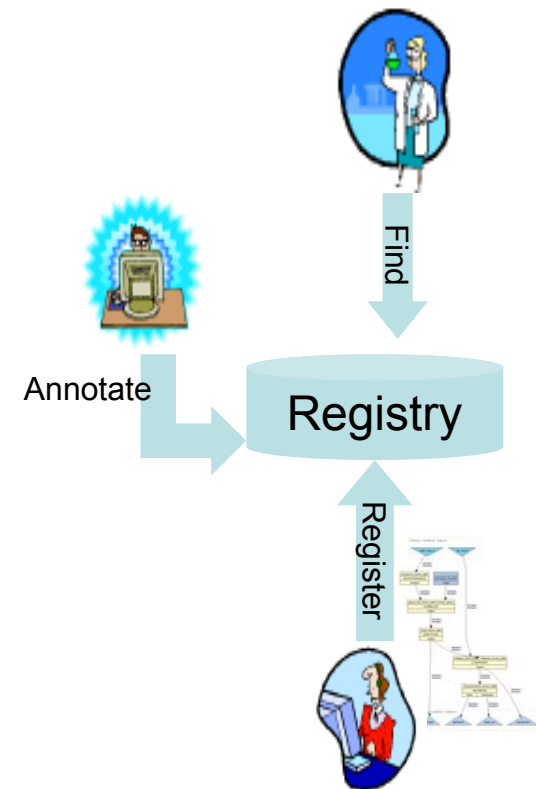


Virtual organisations



Collaborative e-Science

- High level services for e-Science experimental management;
 - Provenance
 - Event notification
 - Personalisation
- Sharing knowledge and sharing components
 - Scientific discovery is personal & global.
 - Federated third party registries for workflows and services
 - Workflow and service discovery for reuse and repurposing





Roadmap

- Part 1
 - Application context
- Part 2
 - Architecture
 - Information and Workflows
 - Semantics and provenance
- Part 3
 - Wrap up



Key Characteristics

- Data Intensive, Up stream analysis
- Pipelines - experiments as workflows (chiefly)
- Adhoc exploratory investigative workflows for individuals from no particular a priori community
- **Openness – the services are not ours.**
- Low activation energy, incremental take on
- Foundations for sharing knowledge and sharing experimental objects
- Multiple stakeholders
- Collection of components for assembly

Openness

- Openness
 - open source
 - open world of services
 - open extensible technology
 - open to wider eScience context
 - open to user feedback
 - open to third party metadata



Putting the user first

User-driven end to end scenarios essential

Whole solution that fits with them

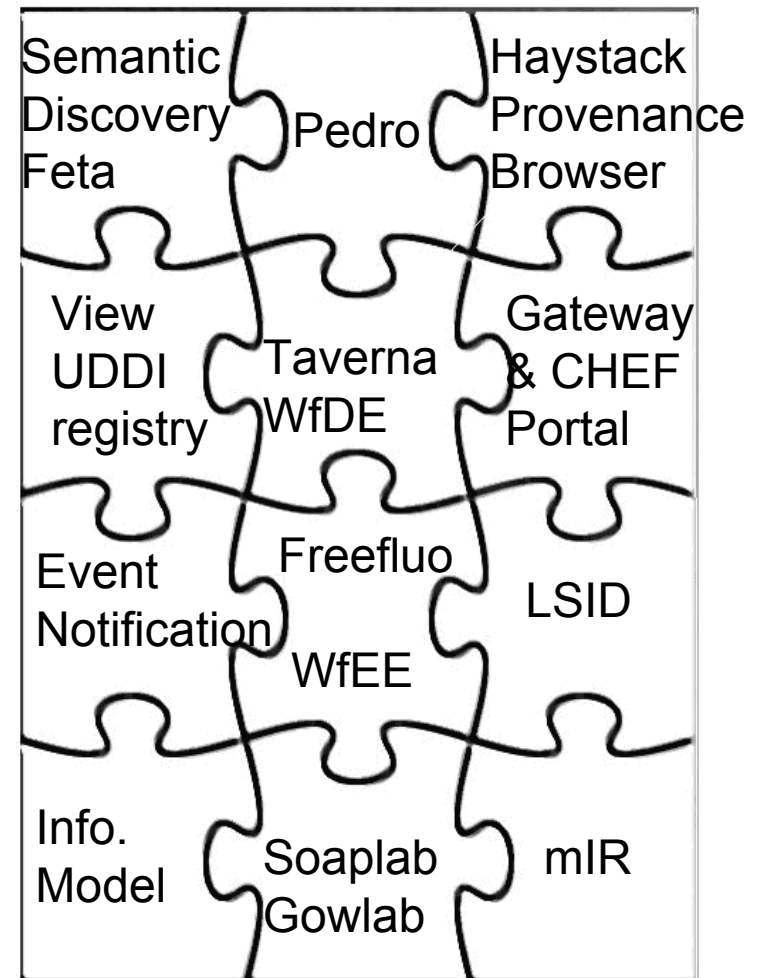
Users vs Machines (vs Interesting computer science)

- Mismatch for information needs
 - Scufi instead of BPEL/WSFL
 - Layers of Provenance
 - Service/workflow descriptions for PEOPLE not just machines
 - Bury complexity, increasingly simplify
- Bioinformaticans **HARDLY EVER** want to have their services automatically selected
 - Except SHIMs, Replicas, User specified equivalences

Service providers and developers are users too!

In a nutshell

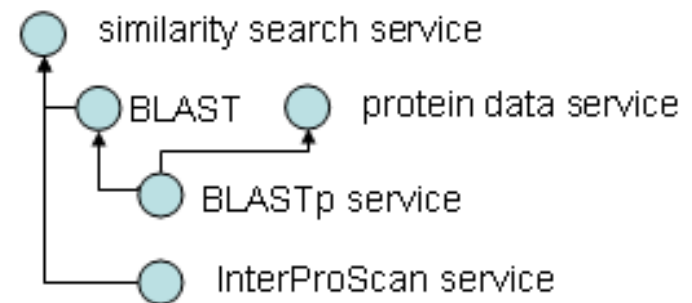
- Bioinformatics toolkit
- Open (Web) Services
 - myGrid components and external domain services
 - Publication, discovery, interoperation, composition, decommissioning of myGrid services
 - No control or influence over domain service providers
- Metadata Driven
 - LSIDs, Common information model, Ontologies, Semantic Web technologies
- Open extensible architecture
 - Assemble your own components
 - Designed to work together
 - Loosely coupled



- Standards based
- (Web) Service Oriented Architecture
 - Publication, discovery, interoperation, composition, decommissioning of myGrid services
 - Web services communication fabric
 - XML document types
 - LSIDs for identifying resources
- Implemented in Java using Axis and Tomcat
 - WS-I -> OGSA / WSRF

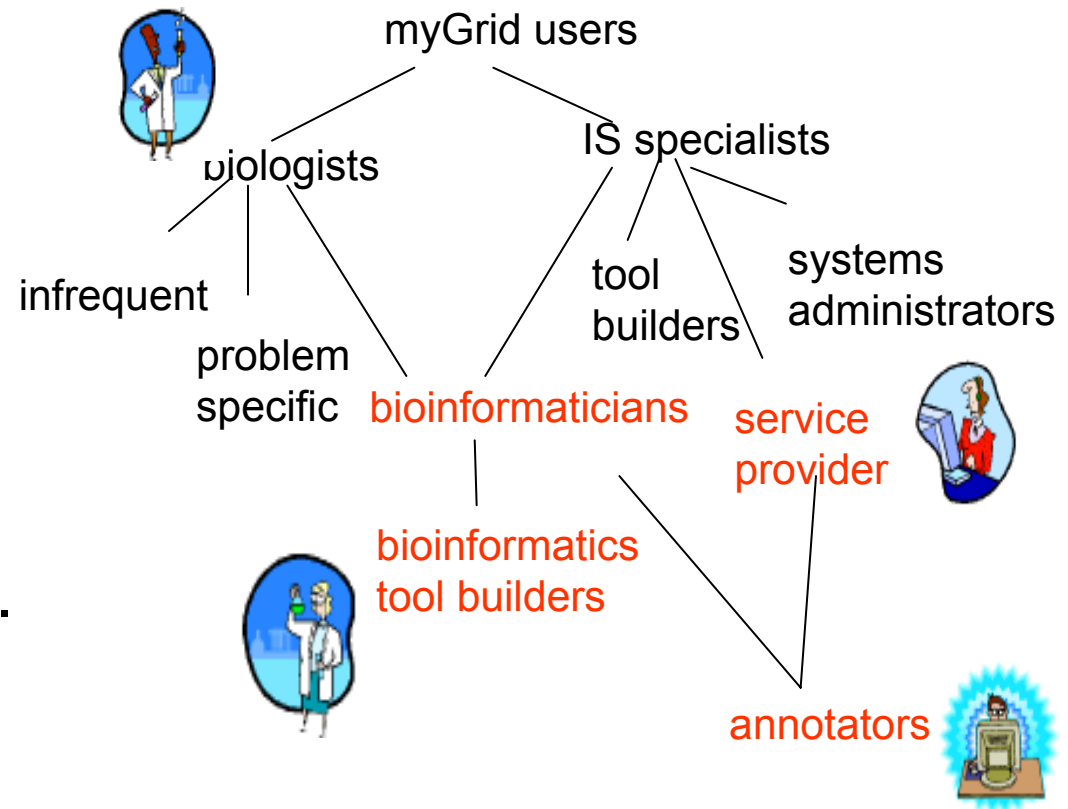
- Metadata driven
 - RDF-coded metadata
 - OWL-coded ontologies
 - Common information model

Extract of service classification

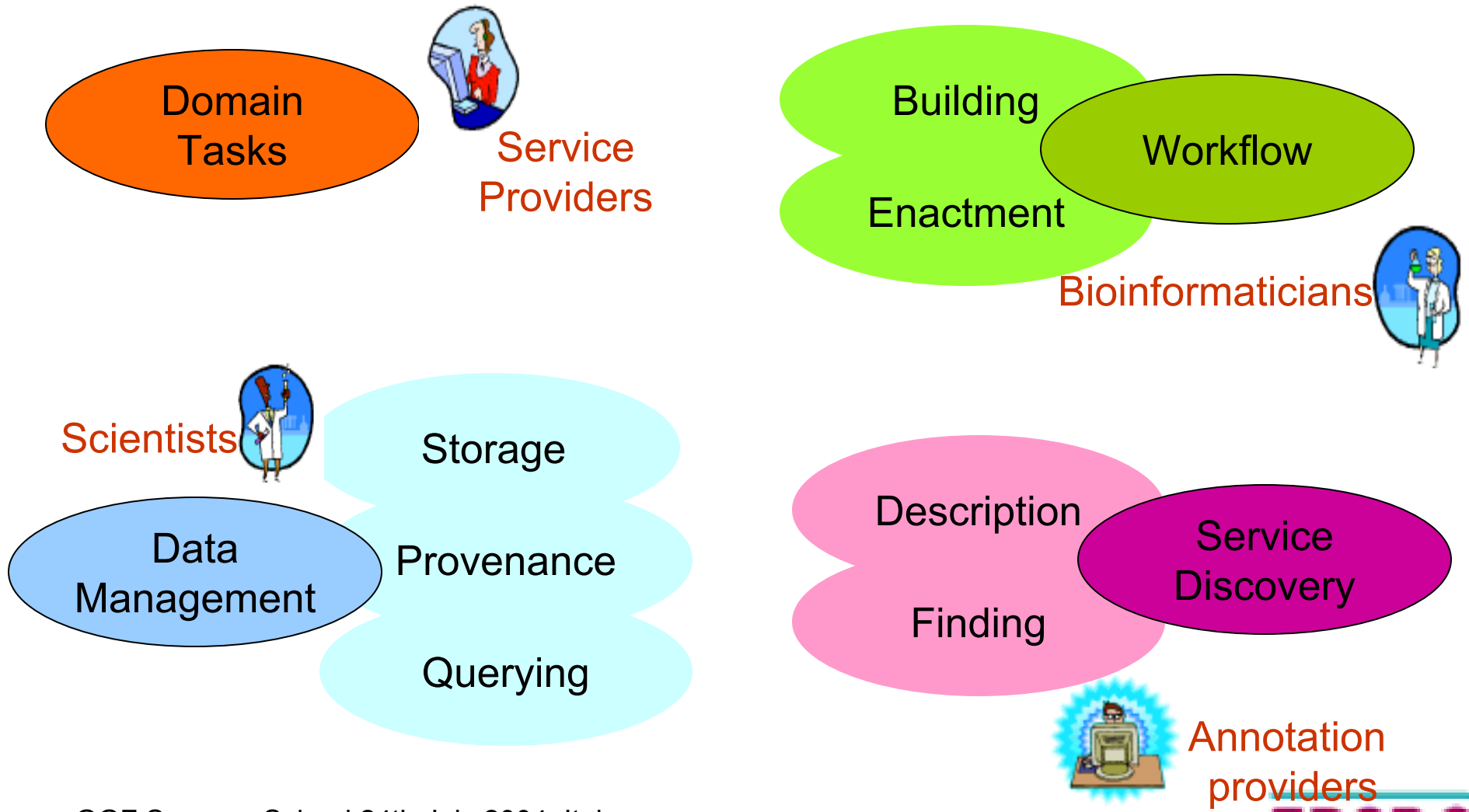


Stakeholders

- Middleware for
- Tool Developers
- Bioinformaticians
- Service Providers
- Biologists are indirectly supported by the portals and apps these develop.

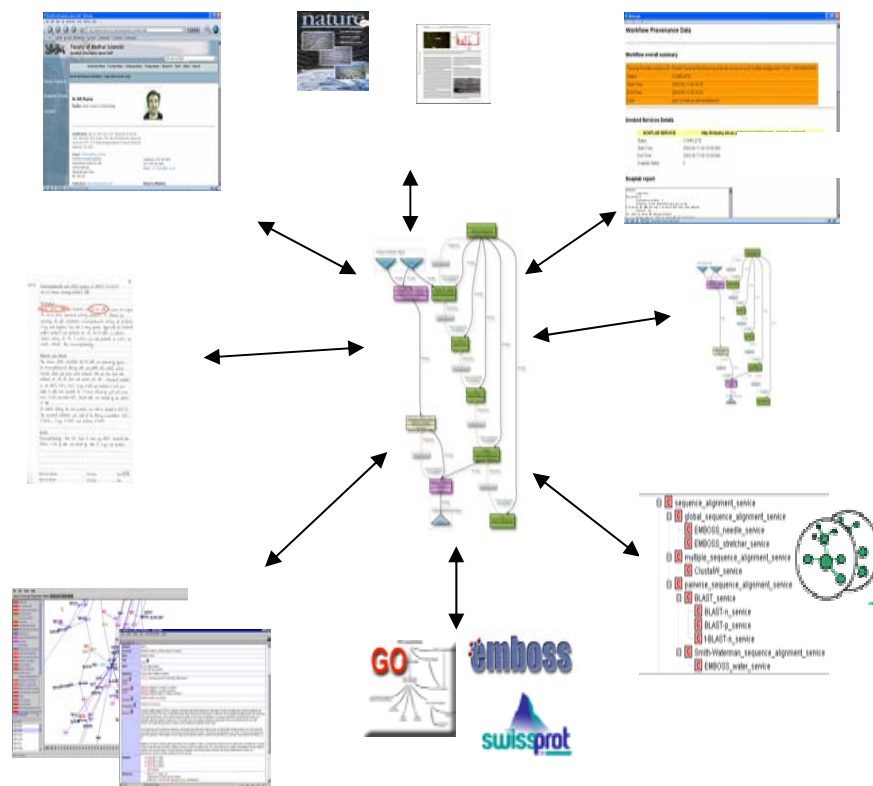


Collections of Tasks

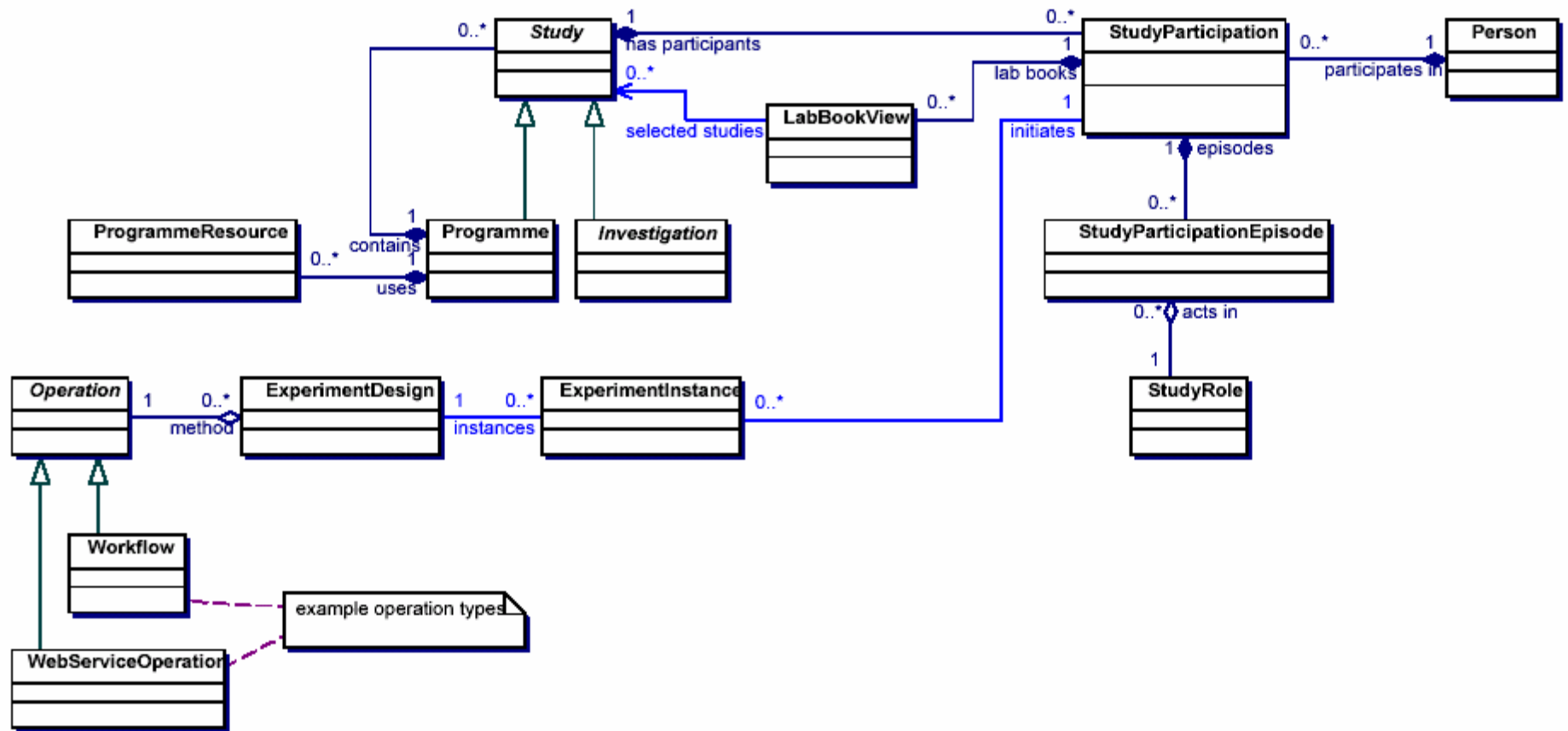


Investigation = set of experiments + metadata

- *Experimental design components*
- *Experimental instances* that are records of enacted experiments
- *Experimental glue* that groups and links design and instance components
- Life Science IDs, URIs, RDF

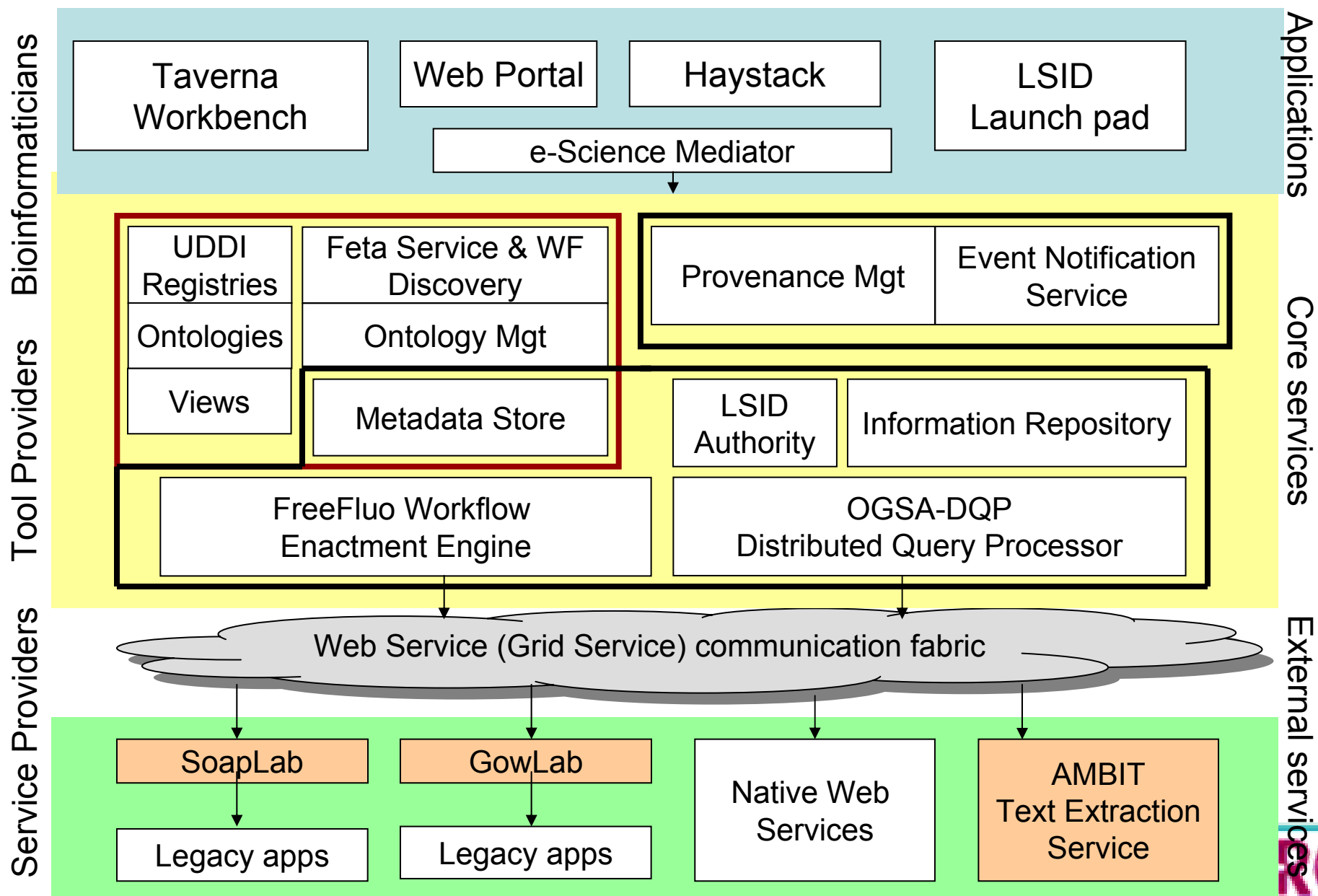


Experimental entities

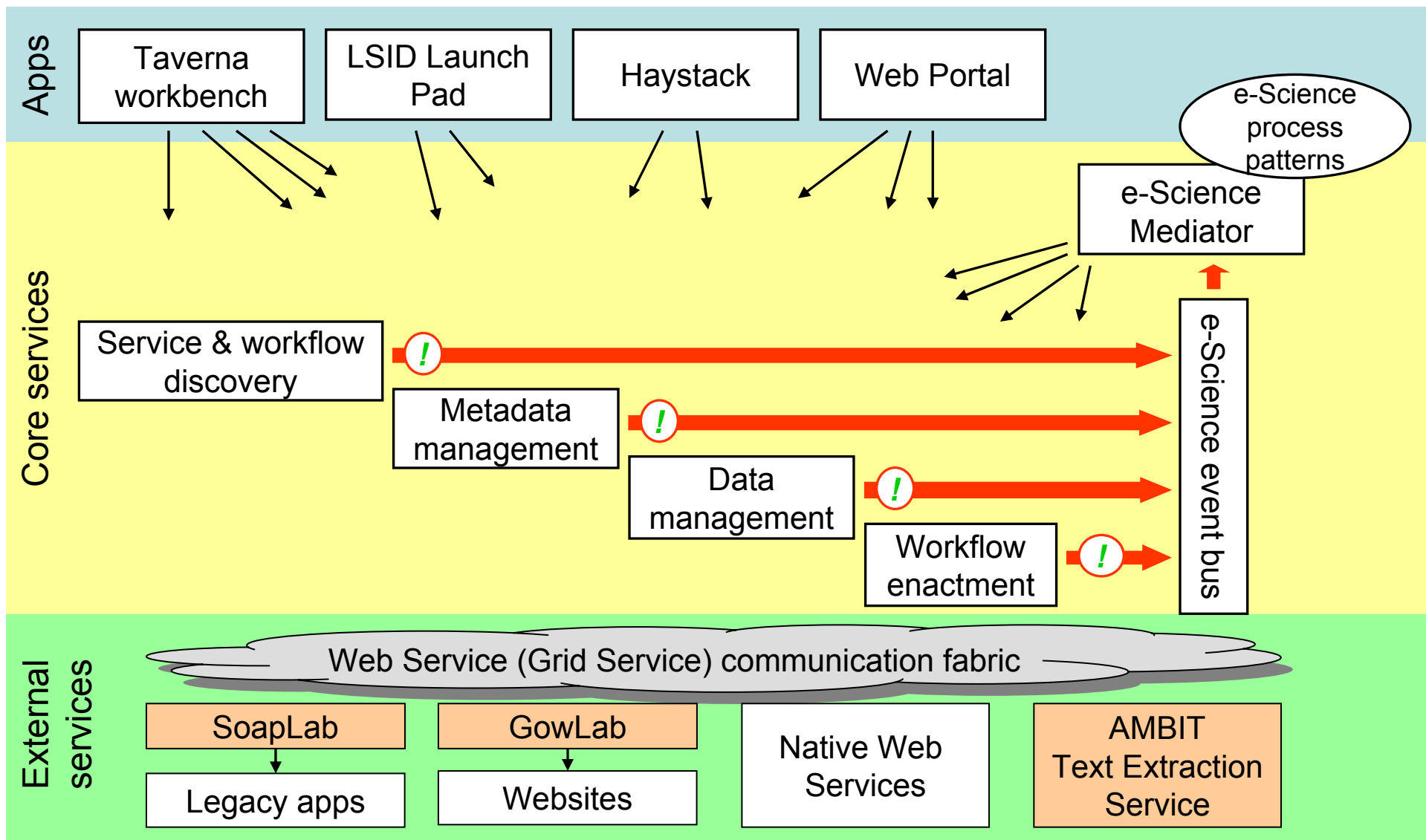




myGrid Service Stack

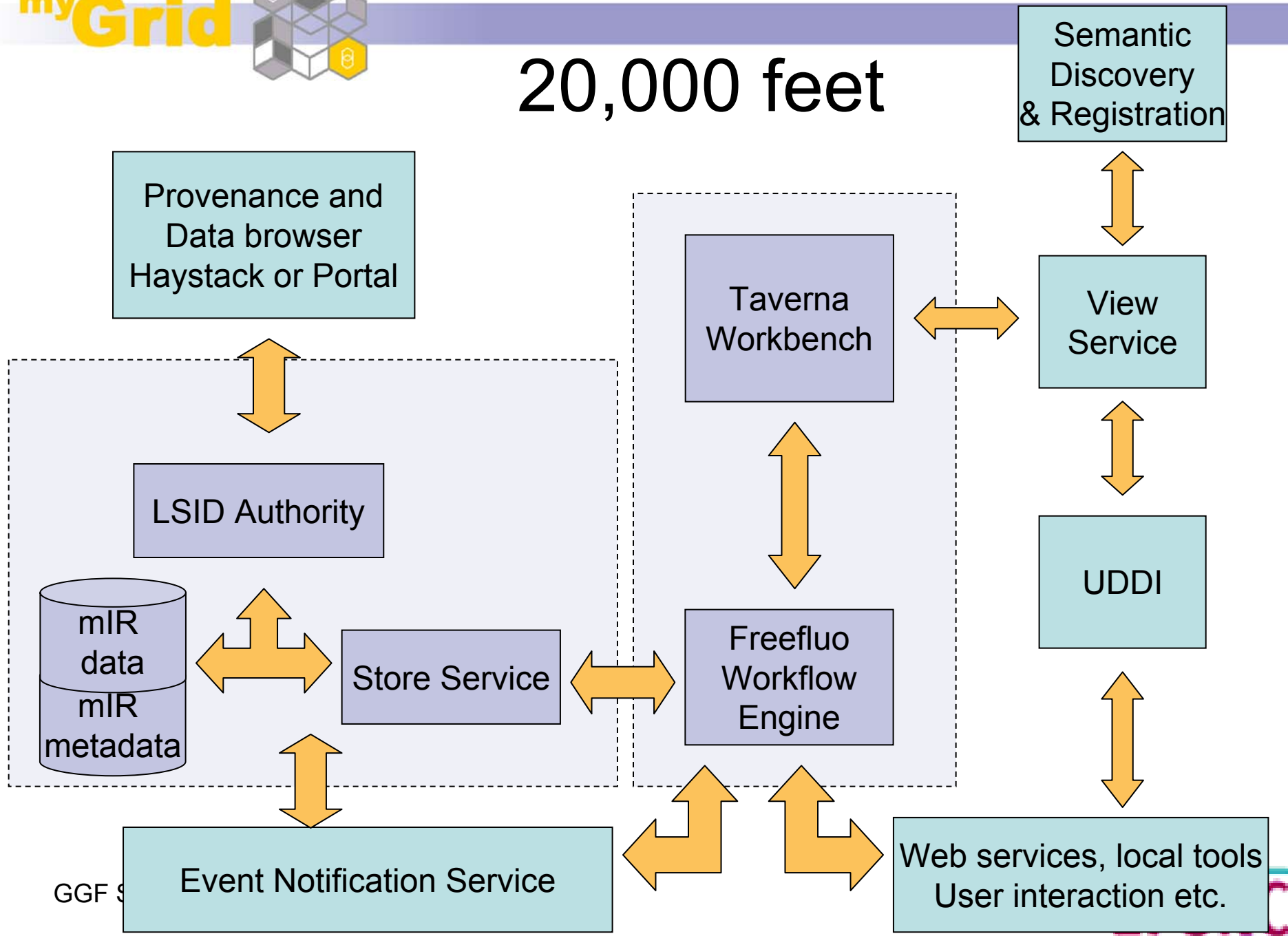


Service stack





20,000 feet



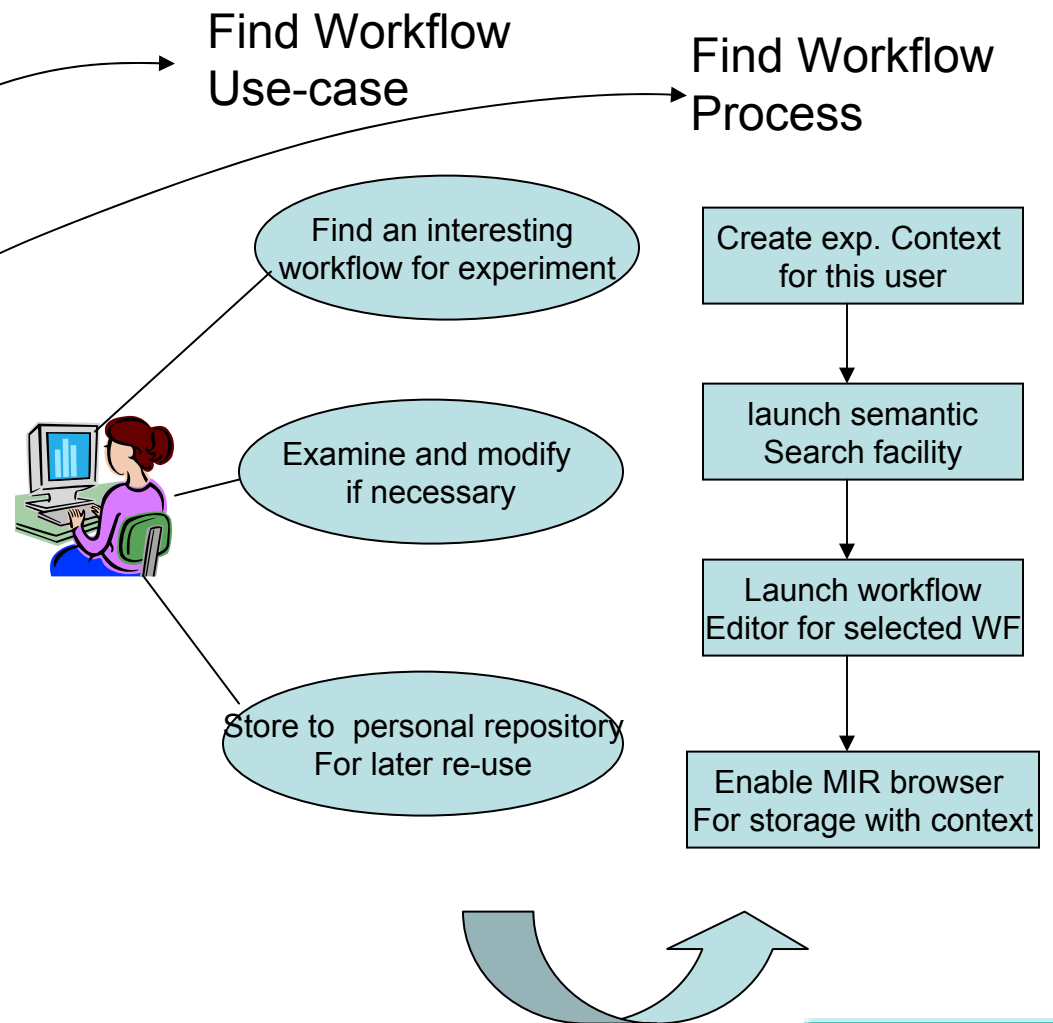


e-Science Mediator

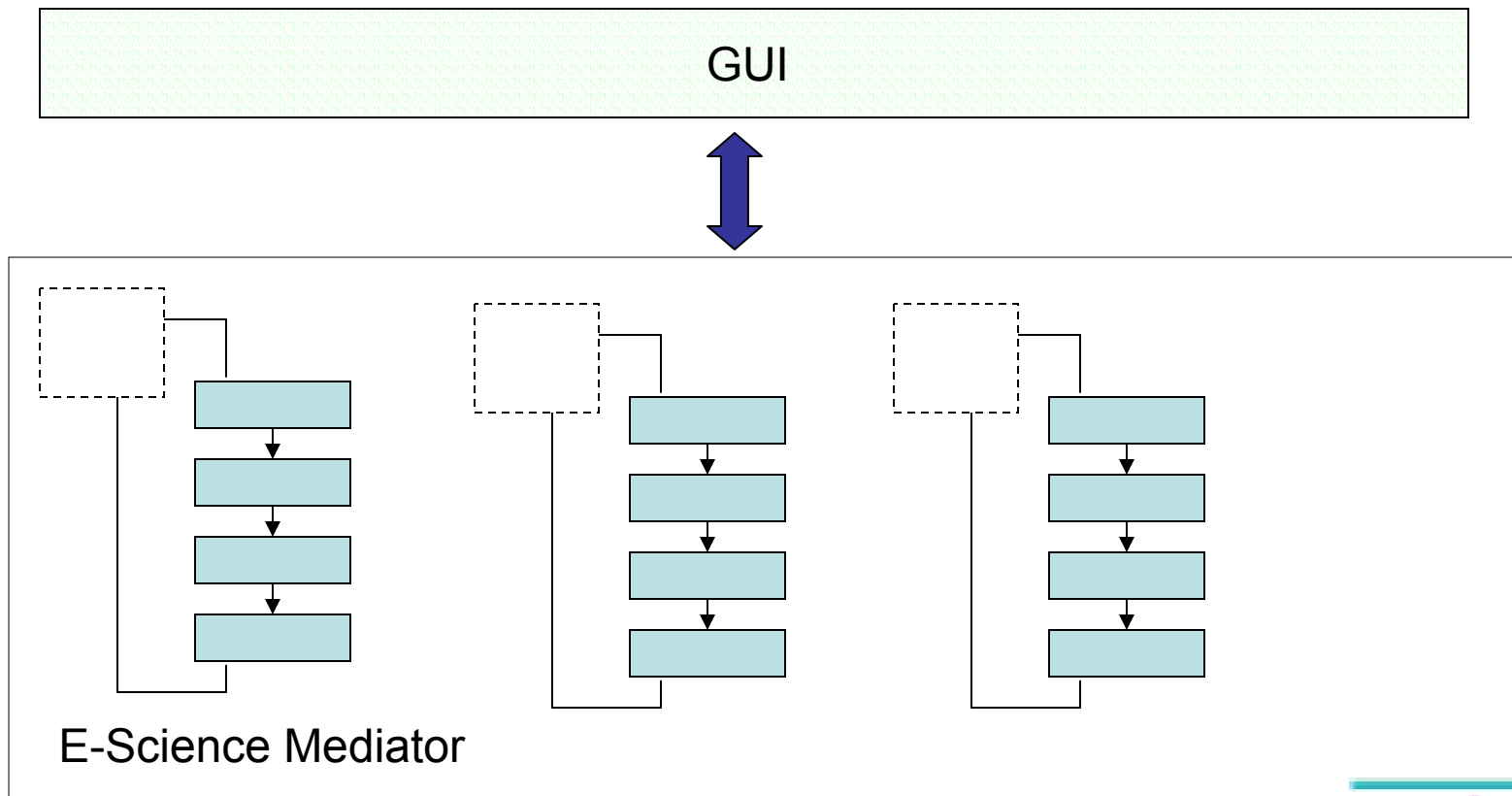
1. Application oriented: directly supports the e-Scientist by:
 - providing pre-configured e-Science processes templates (i.e. system-level workflows)
 - helping in capturing and maintaining context information (via the information model) that is relevant to the interpretation and sharing of the results of the e-science experiments.
 - Facilitating personalisation and collaboration
2. Middleware oriented: contributes to the synergy between myGrid services by:
 - Acting as a sink for e-Science events initiated by myGrid components
 - Interpreting the intercepted events and triggering interactions with other related components entailed by the semantics of those events
 - Compensating for possible impedance mismatches with other services both in terms of data types and interaction protocols

Supporting the e-scientist

- Recurring use-cases can be captured
- Then corresponding process templates can be authored
- e-science mediator makes processes available to the user

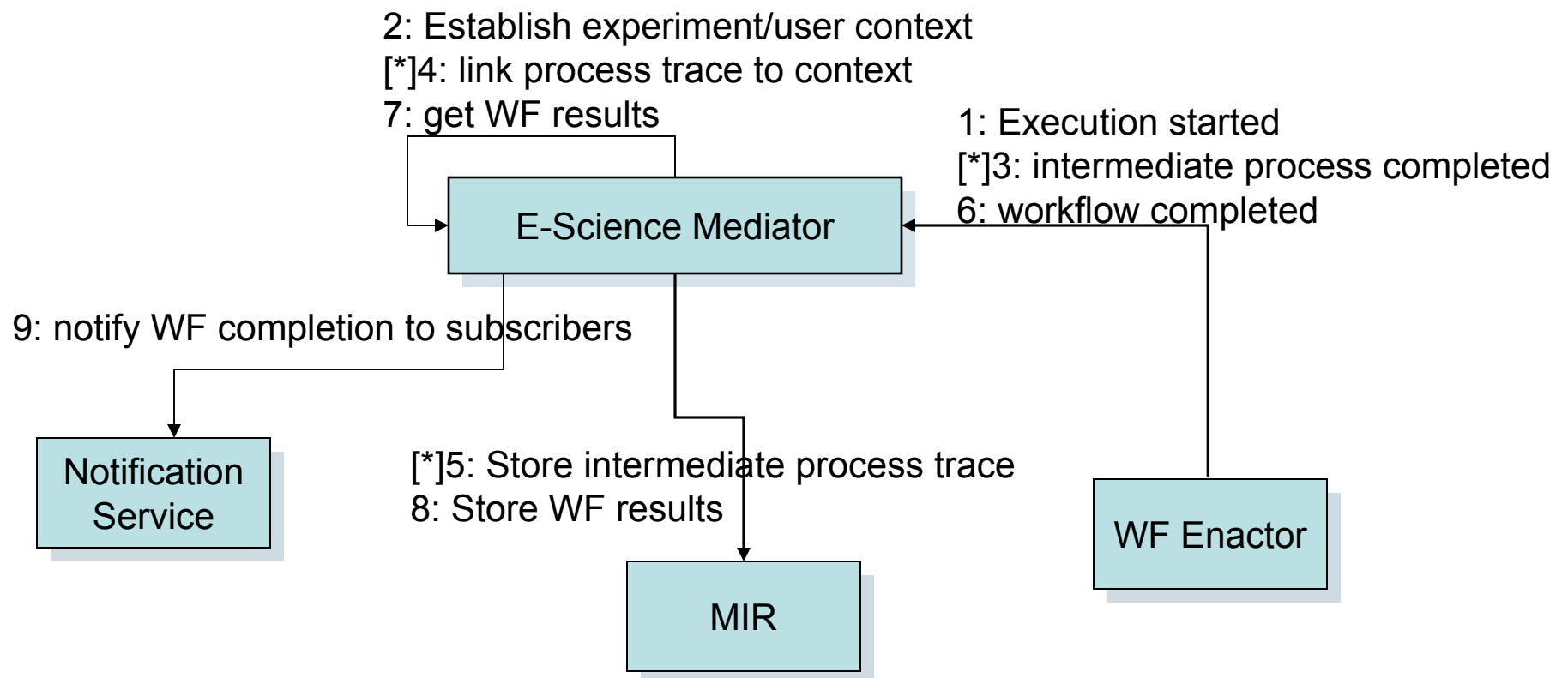


- E-Science process templates maintained by the mediator can derive the GUI generation and interaction with the user

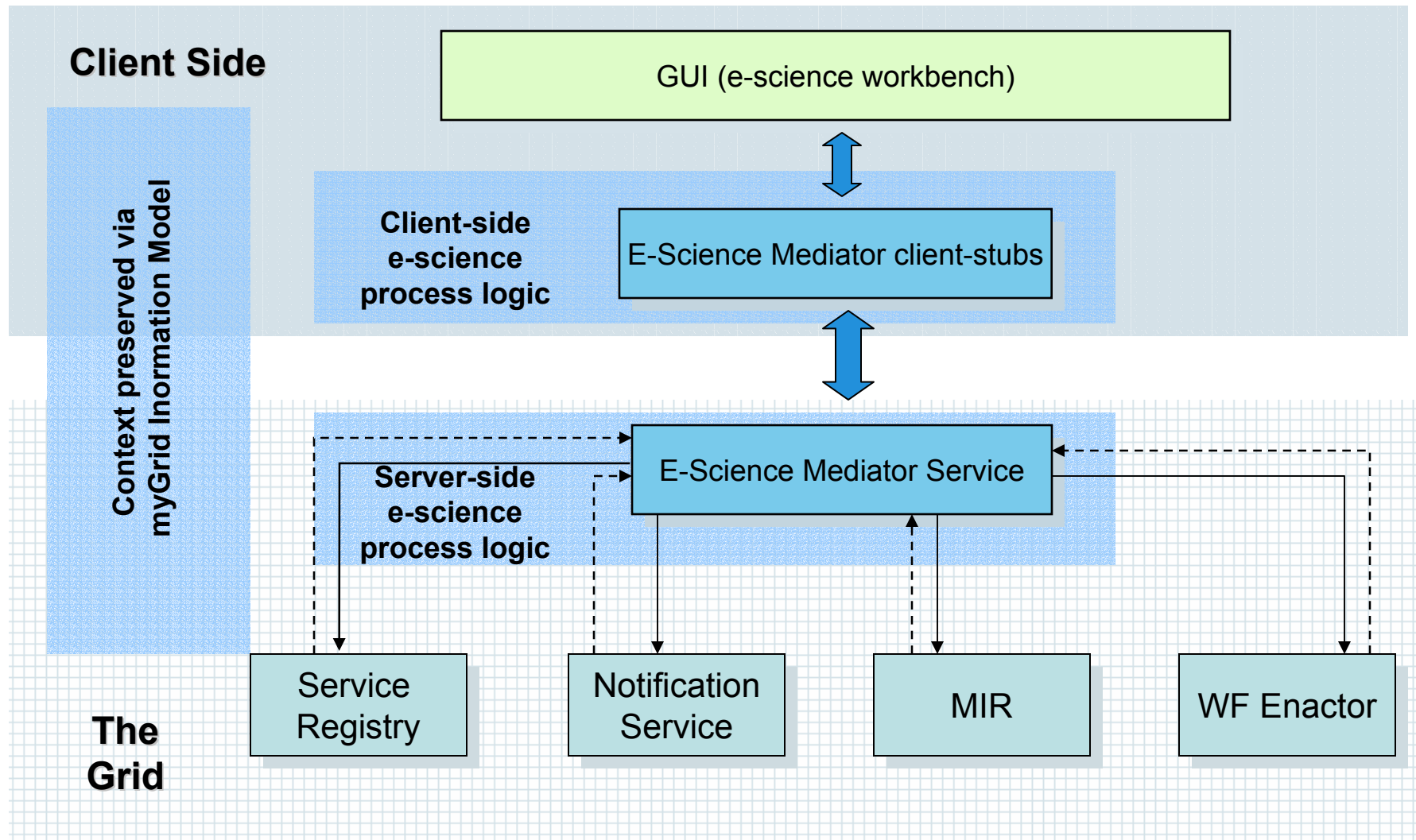


Mediating between services

Example: mediation during a workflow execution

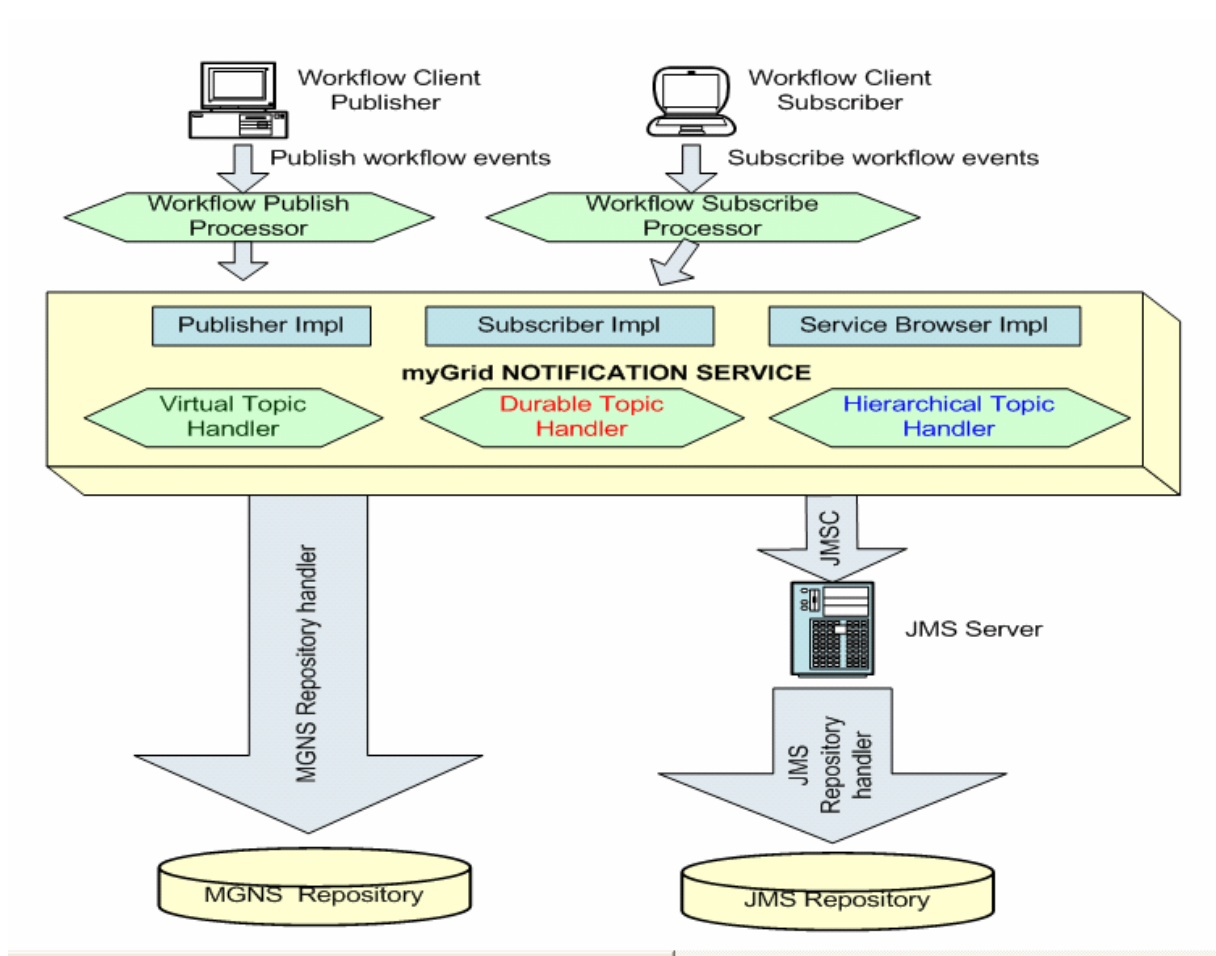


Simplified Architecture



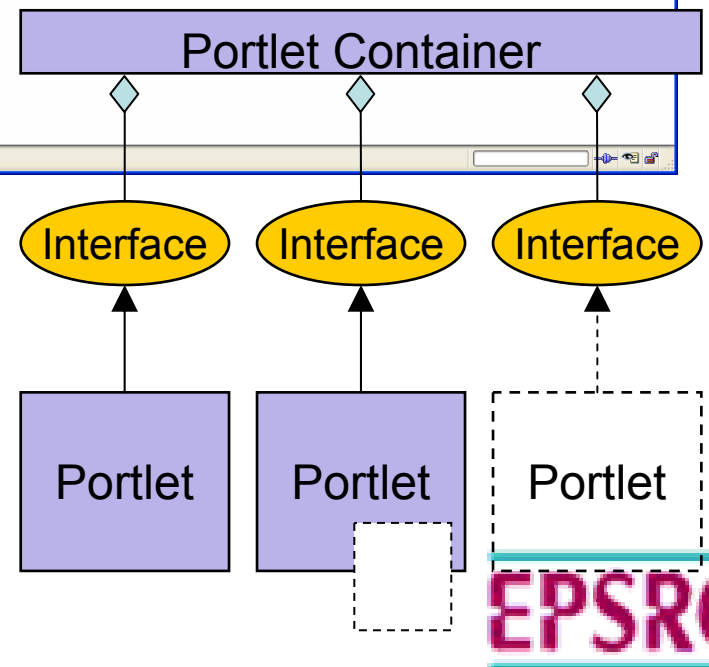
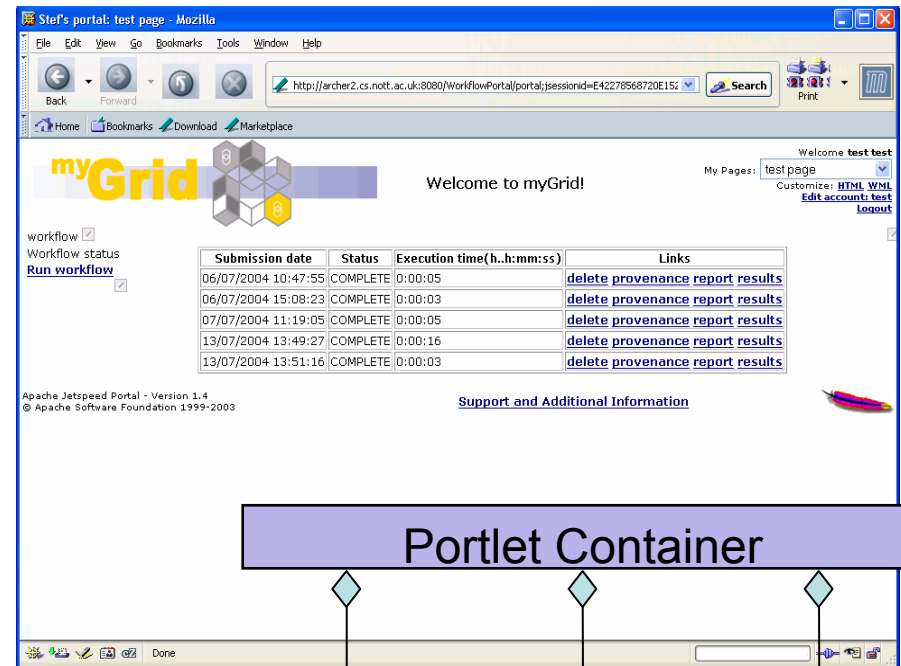
Event notification Service

- Publish/subscribe model
 - Topic based (cf. JMS topics, CORBA channels)
 - Hierarchic topics
 - Persistent event storage
 - Subscription leases
 - Federation for scalability & reliability
 - Event filtering

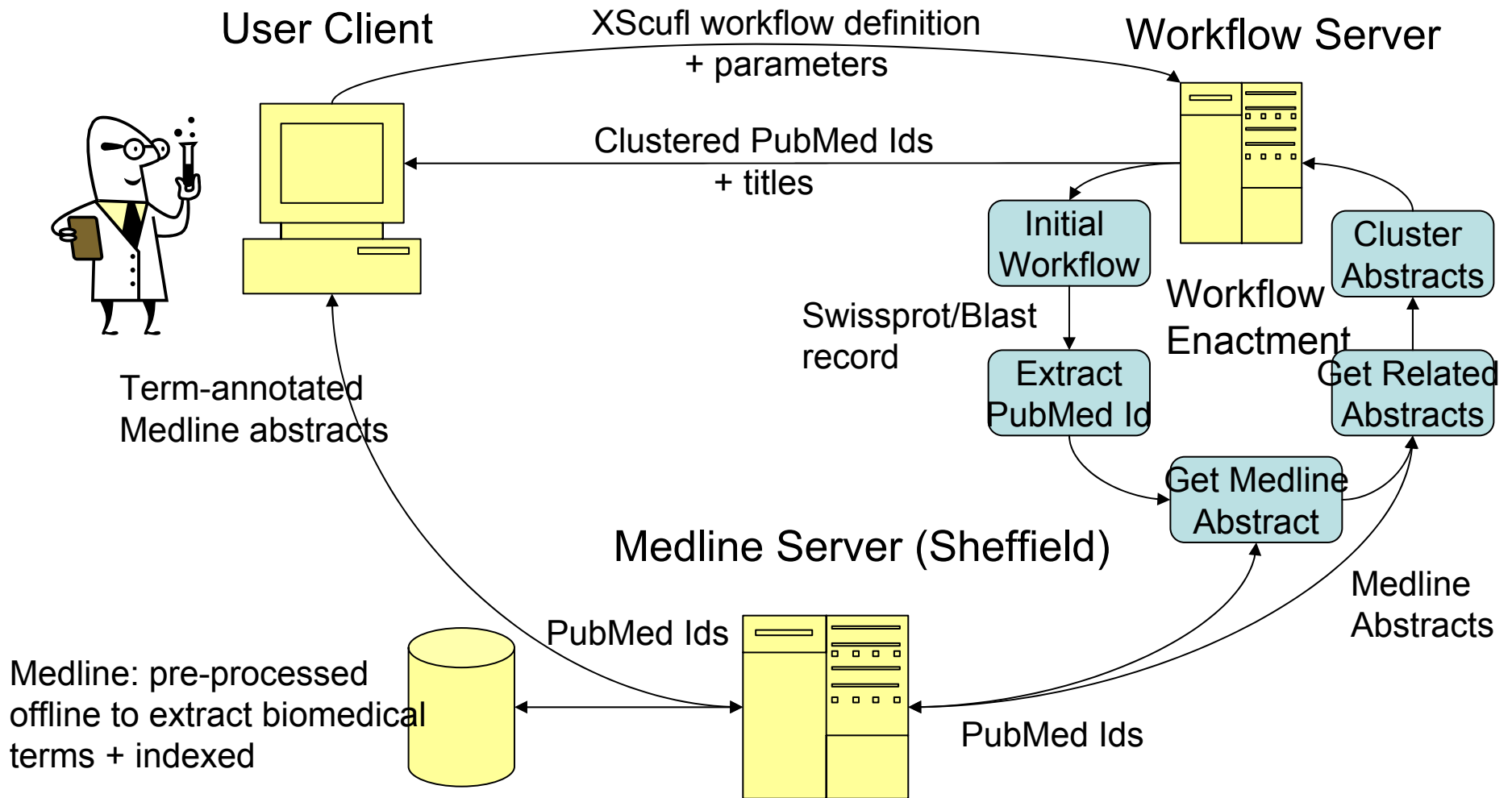


Portal toolkit for bioinformaticians

- Target application
 - Williams-Beuren Syndrome
 - Fixed set of workflows
- Extra myGrid portlets
 - Configurable
 - Workflow enactment
 - Workflow scheduling
 - Completion notification
 - Results browsing
- Based on CHEF & Jetspeed-1
 - Portlets for team collaboration



Text Services



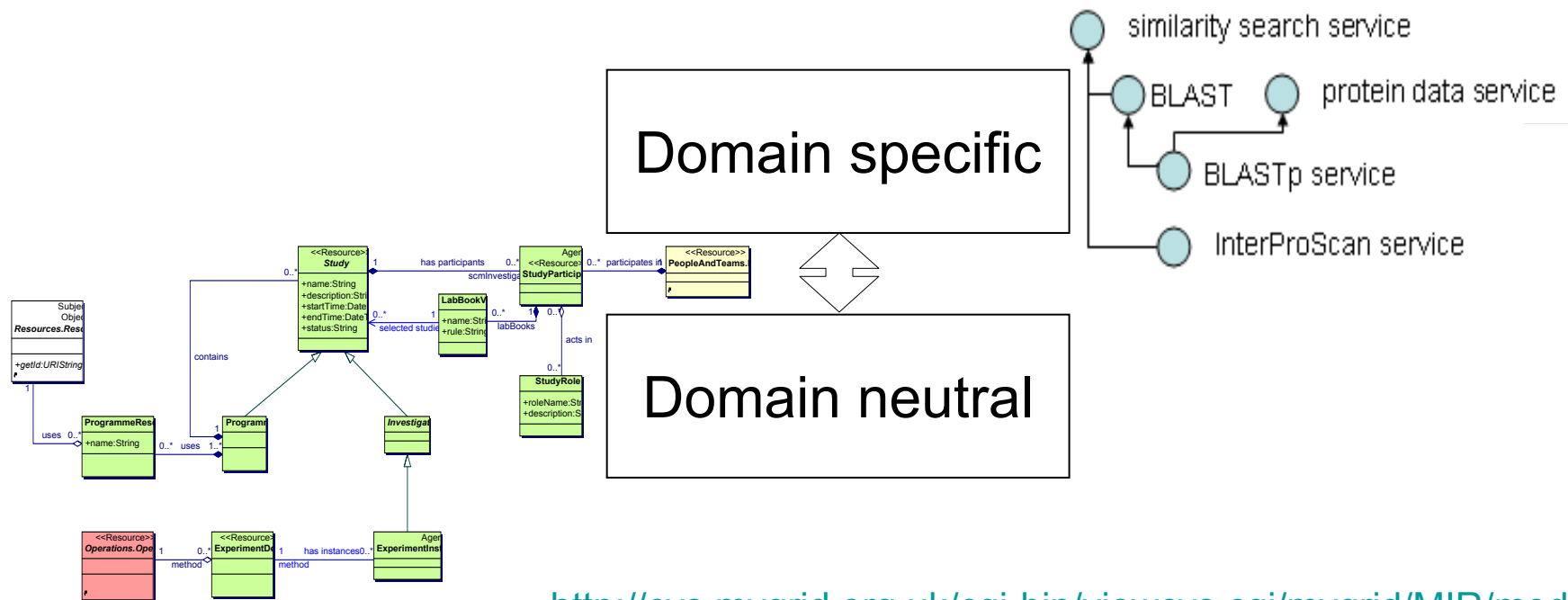
Roadmap

- Part 1
 - Application context
- Part 2
 - Architecture
 - Information and Workflows
 - Semantics and provenance
- Part 3
 - Wrap up



Information Model v2

myGrid components form a loosely coupled system
 An Information Model for e-Science experiments
 Based on CCLRC scientific metadata model
 XML messages between services conform to the IMv2



<http://cvs.mygrid.org.uk/cgi-bin/viewcvs.cgi/mygrid/MIR/model/>

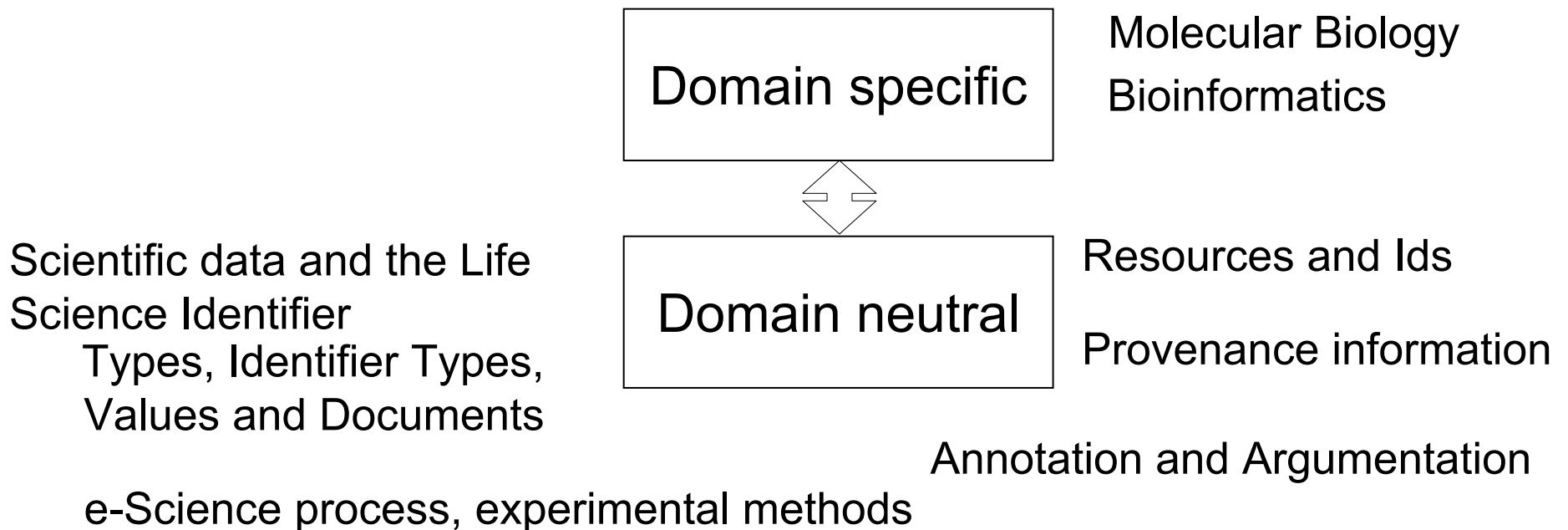
Nick Sharman, Nedim Alpdemir, Justin Ferris, Mark Greenwood, Peter Li, Chris Wroe, *The myGrid Information Model*, Proc UK e-Science 2nd All Hands Meeting, Nottingham, UK 1-3 Sept 2004.

GGF Summer School 24th July 20



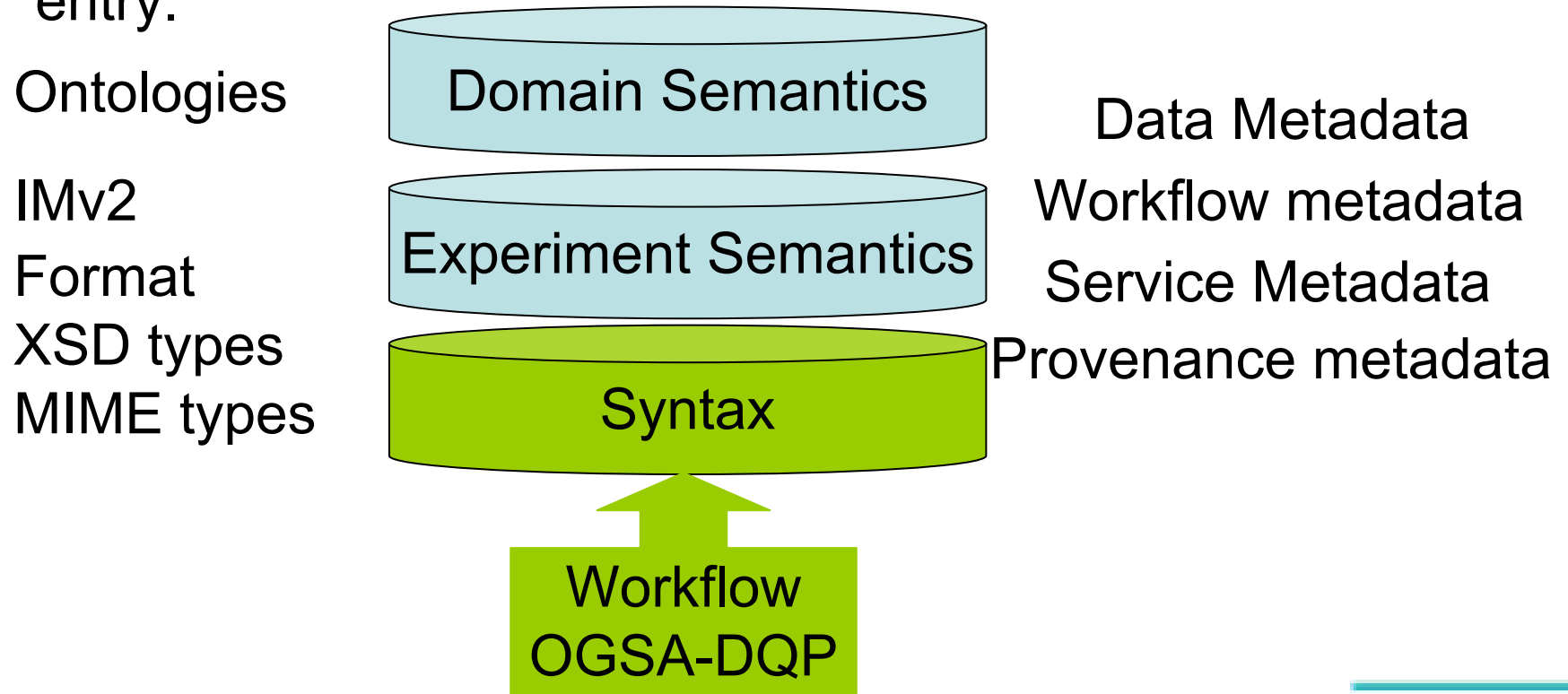
Information Model v2

myGrid components form a loosely coupled system
An Information Model for e-Science experiments
Based on CCLRC scientific metadata model
XML messages between services conform to the IMv2

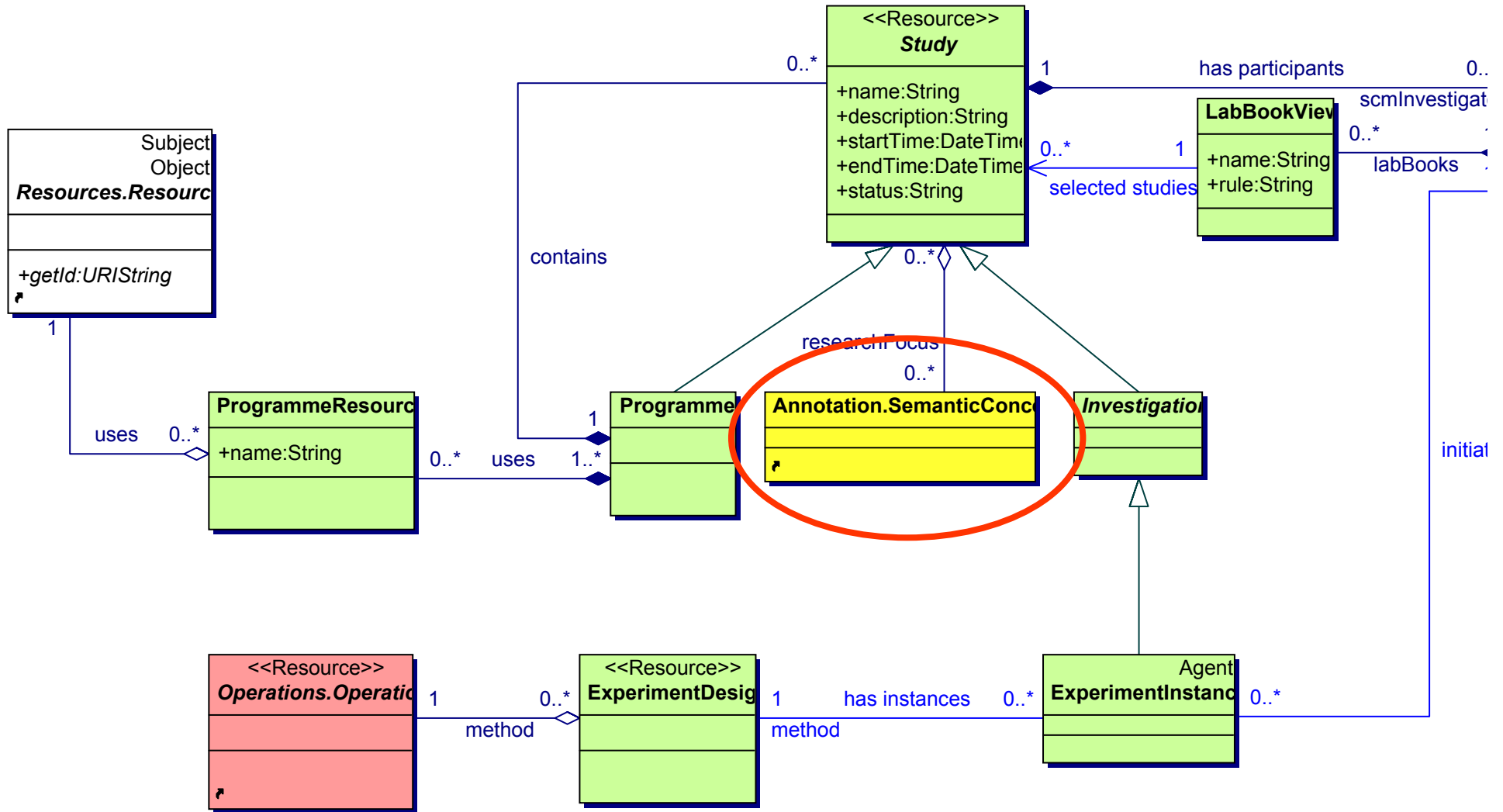


Layered Semantics

- Domain Semantics layered on top of domain neutral but scientific data model
- Reducing the activation energy, lowering barriers of entry.

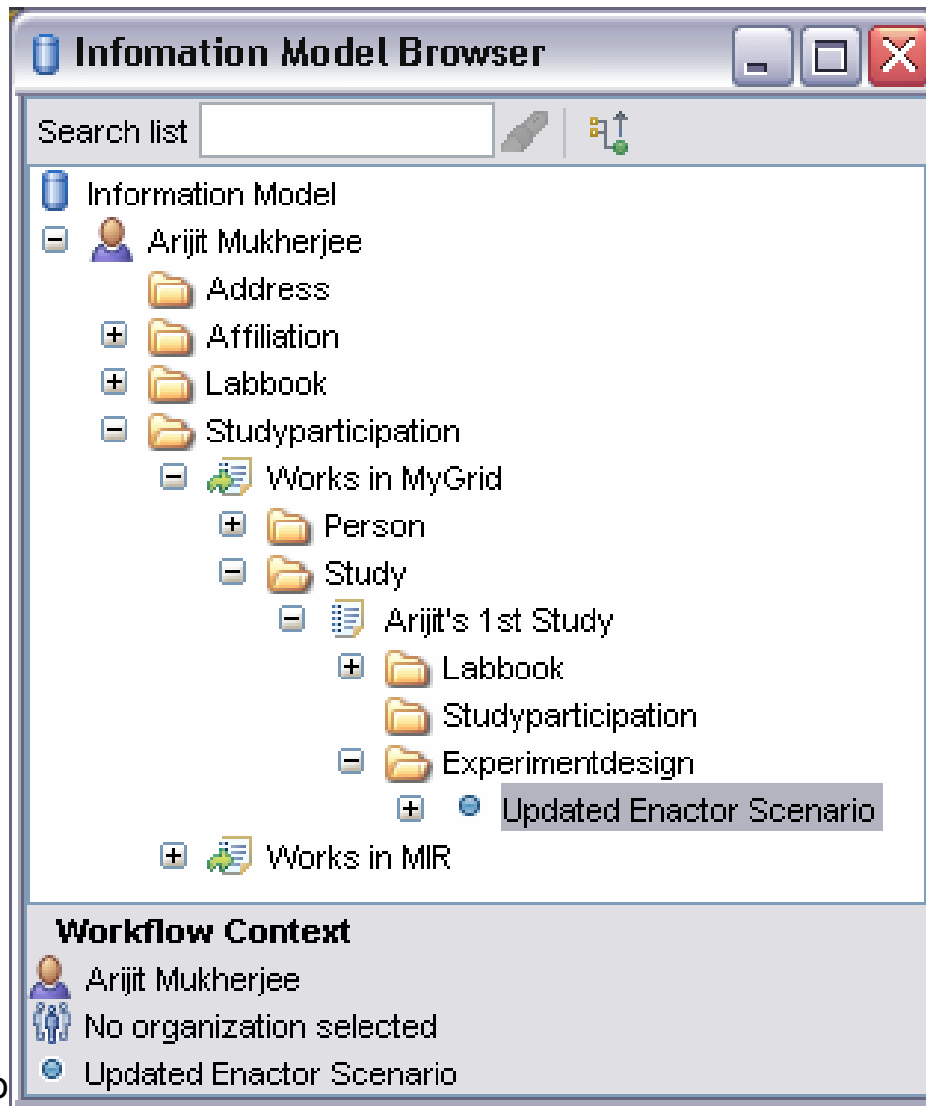


Experimental entities





View over the MIR



Life Science IDs

- Each database on the web has:
 - Different policies for assigning and maintaining identifiers, dealing with versioning etc.
 - Different mechanism for retrieving an item given an ID.
- Life Science IDs designed to harmonise the retrieval of data.
- Emerging standard for bioinformatics
 - I3C, OMG Life Sciences Group, W3C
- Defines:

URI for life science resources

T. Clark, S. Martin & T. Liefeld: *Globally distributed object identification for biological knowledge bases*, Briefings in Bioinformatics Vol 5 No 1 pp 59-70, March 2004

What is an LSID?

urn:lsid:AuthorityID:NamespaceID:ObjectID:[
RevisionID]

urn:lsid:ncbi.nlm.nih.gov:GenBank:T48601:2

urn:lsid:ebi.ac.uk:SWISS-
PROT.accession:P34355:3

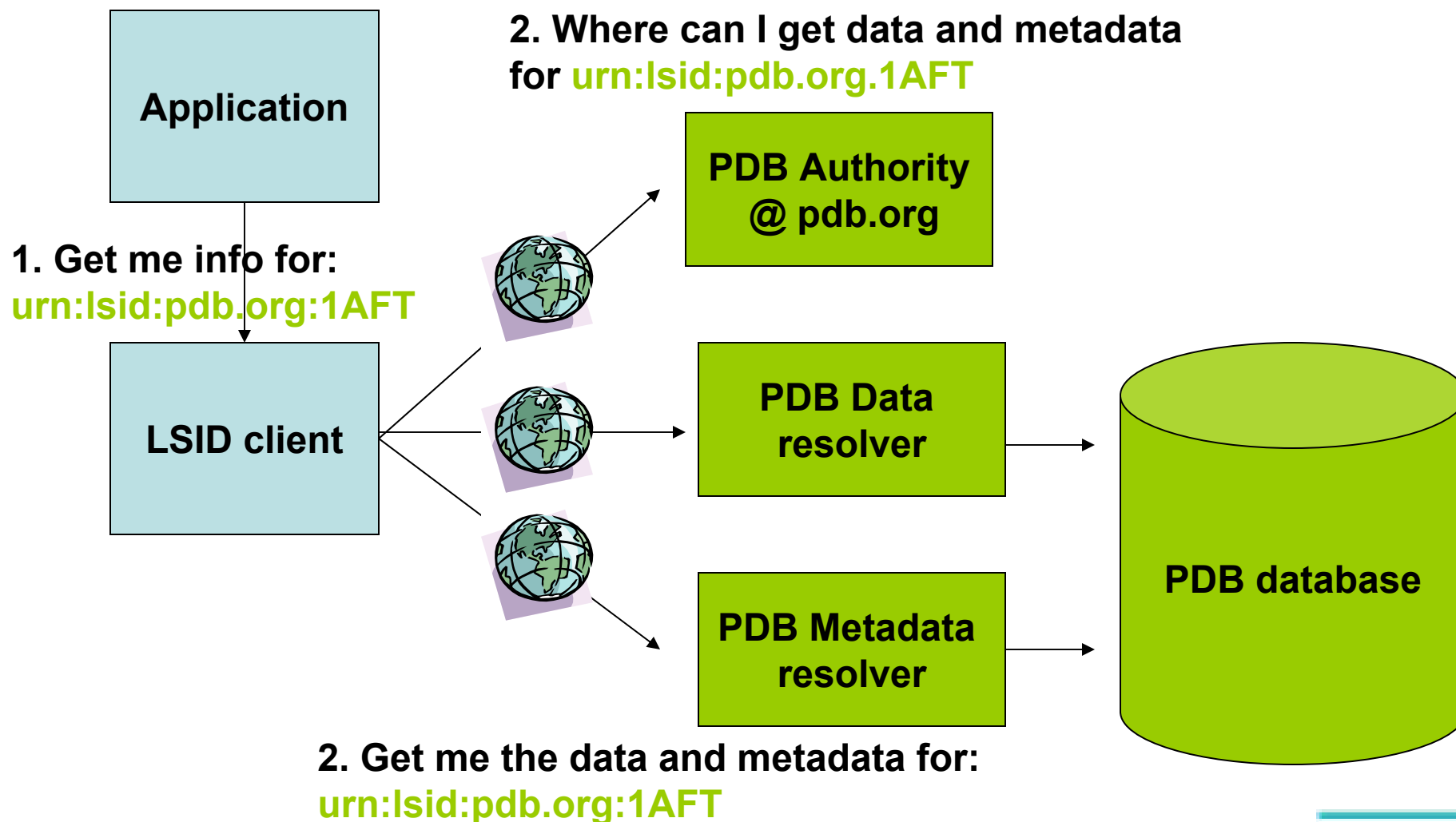
urn:lsid:rcsb.org:PDB:1D4X:22

- **LSID Designator:** A mandatory preface that notes that the item being identified is a life science-specific resource
- **Authority Identifier:** An Internet domain owned by the organization that assigns an LSID to a resource

LSID Properties

- Unique authority for each identifier
- Multiple resolution services, supporting:
 - Data retrieval – data immutable: data returned for a given LSID must always be the same
 - caches
 - Metadata retrieval – mutable and resolver-specific
 - annotation services. More on this in Part 4
- Resolution discovery service
 - Implemented over DNS/DDNS (Optional)
- Authority commitment: must **always** maintain an authority at e.g. pdb.org that can point to data and metadata resolvers.

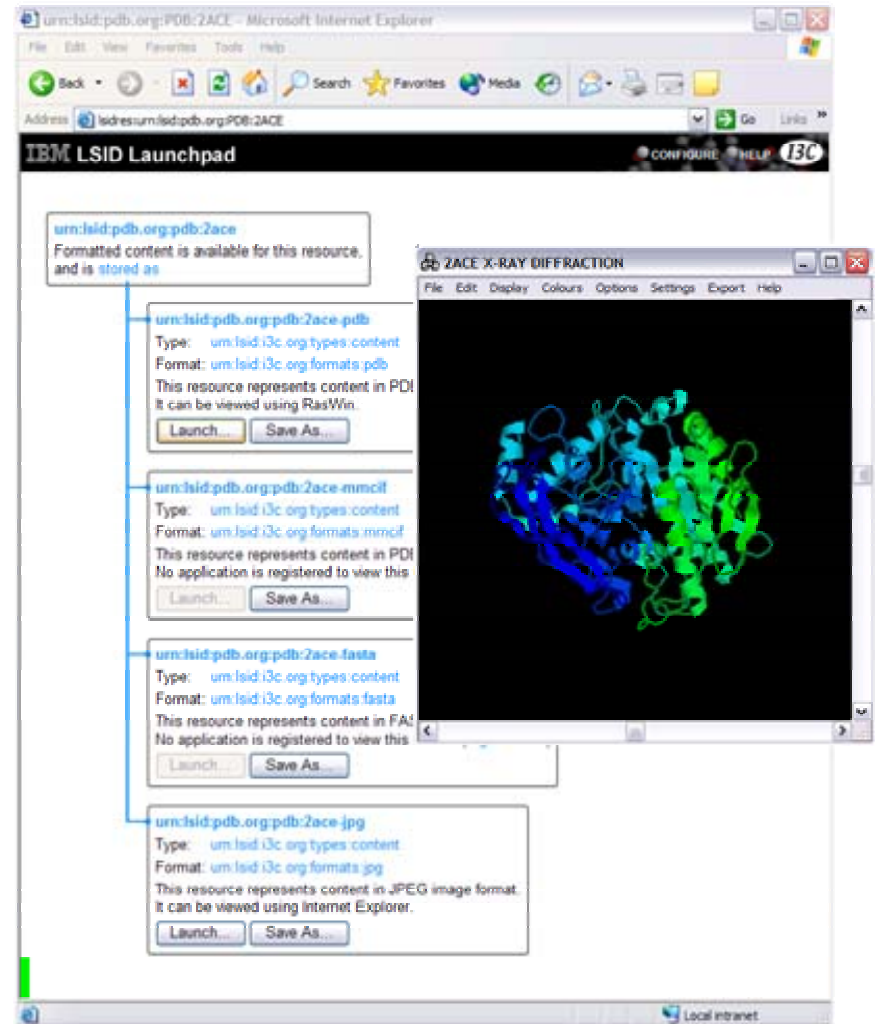
How is data retrieved?



LSID Components

- IBM built client and server implementations in Perl, Java, C++
- Straightforward to wrap an existing database as a source of data or metadata
- Client simple to use
- LSID Launchpad adds LSID

<http://www-124.ibm.com/developerworks/oss/lsid/>
Internet Explorer





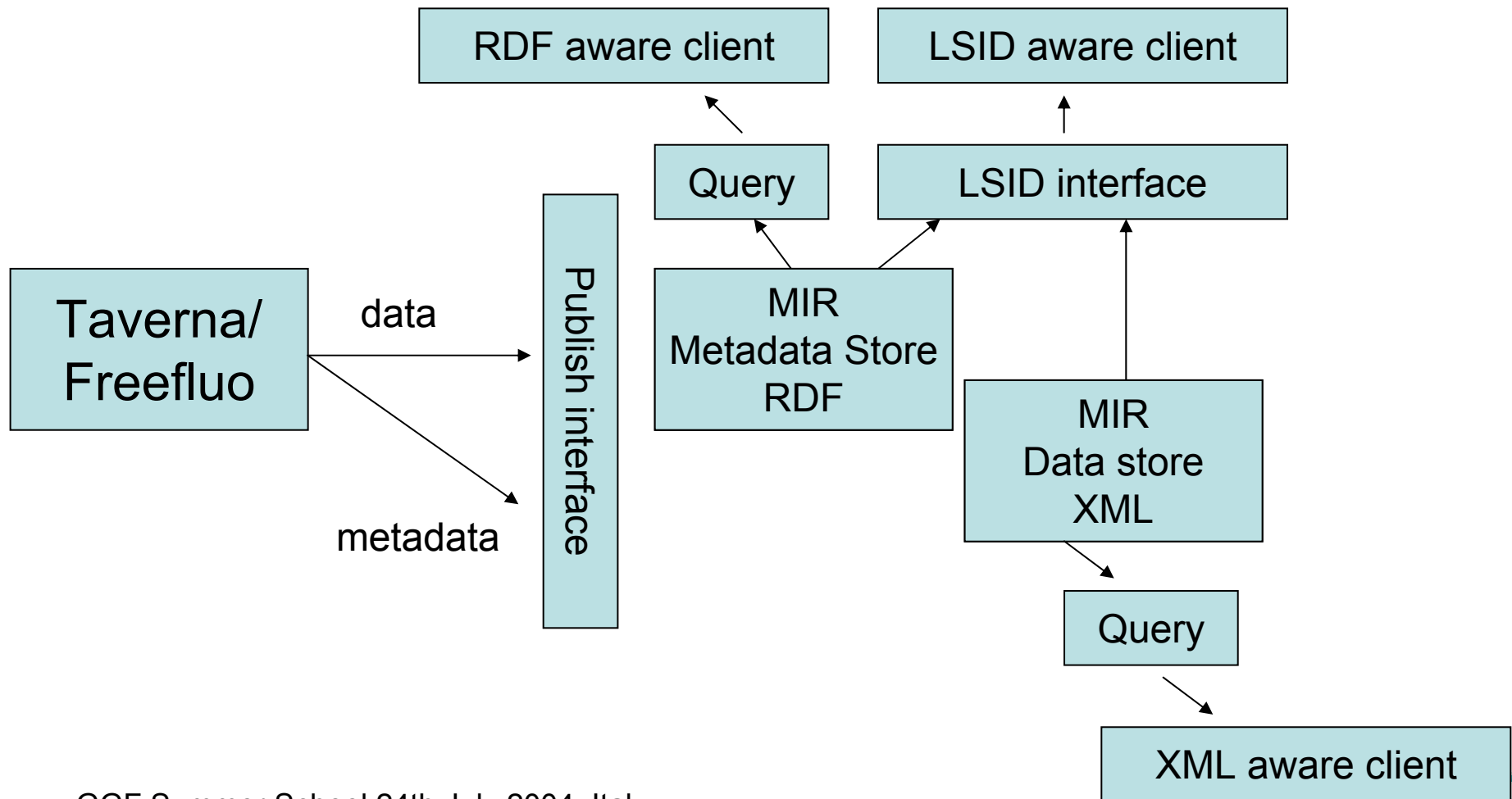
Use within myGrid

- Needed an identifier for our own experimental resources
 - workflows, experiments, new data results etc
- All and everything identified with LSIDs
- LSID saves us having to invent our own conventions and code.
- Can pass references to data around and be reassured the other party will know how to resolve that reference
- Resolution services:
 - Data: myGrid Information Repository (MIR)
 - Metadata: myGrid Metadata Store (RDF-based)

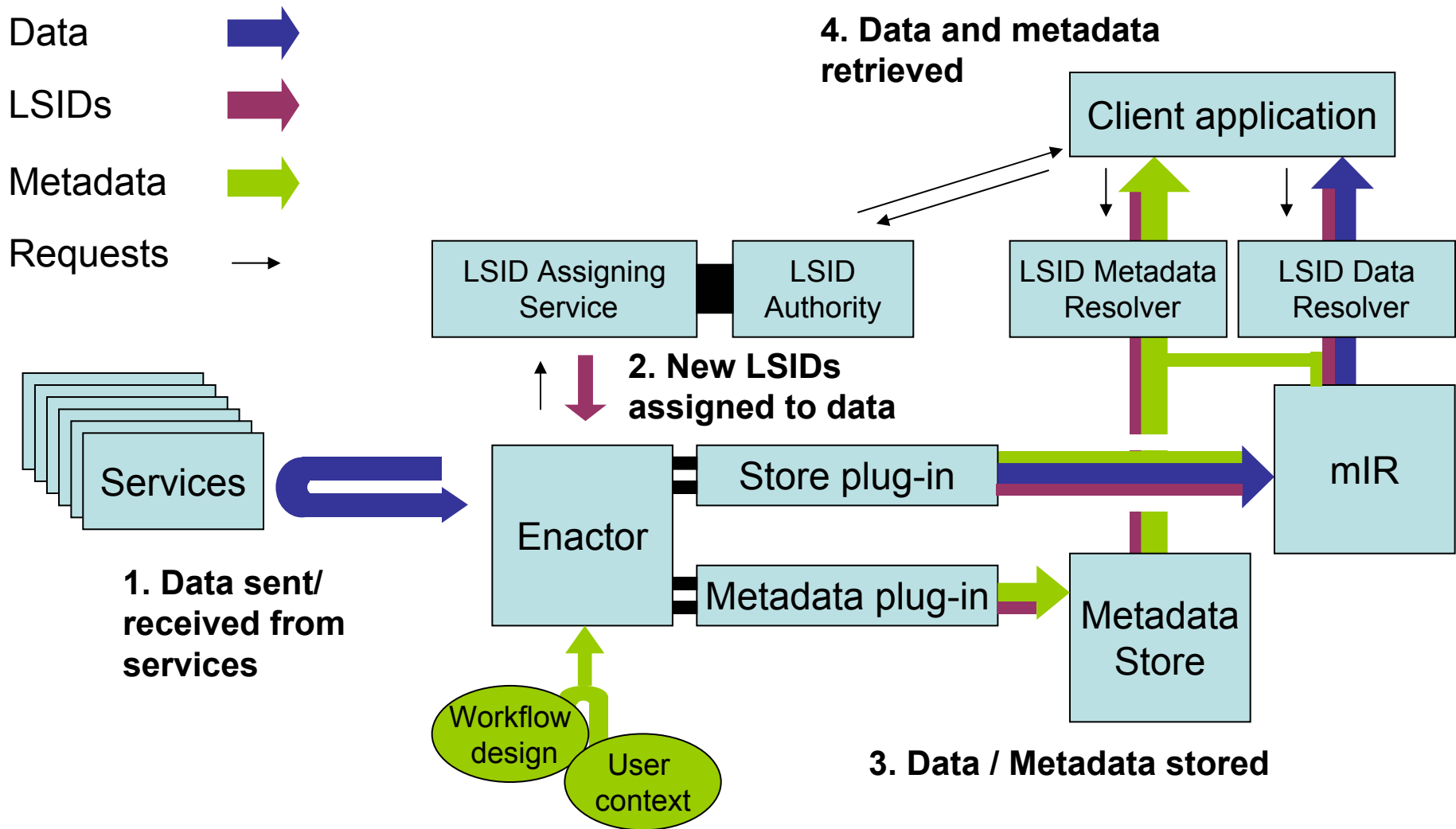
GGF Summer School 24th July 2004, Italy



Information Access

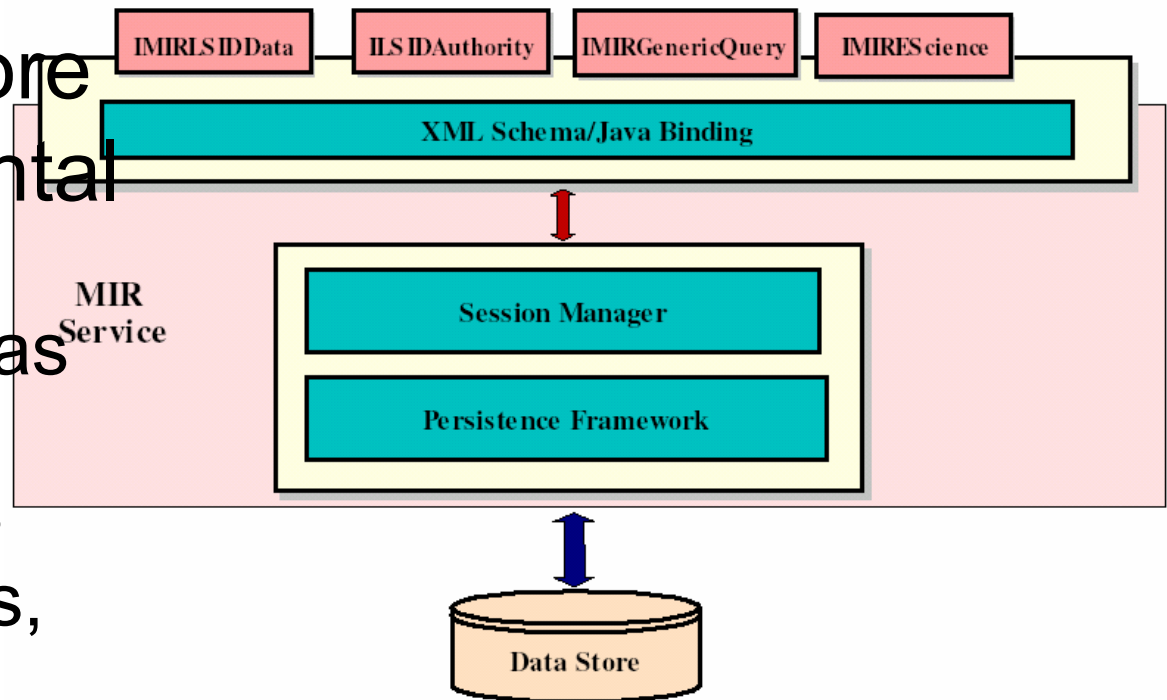


LSID Assignment



Information Storage

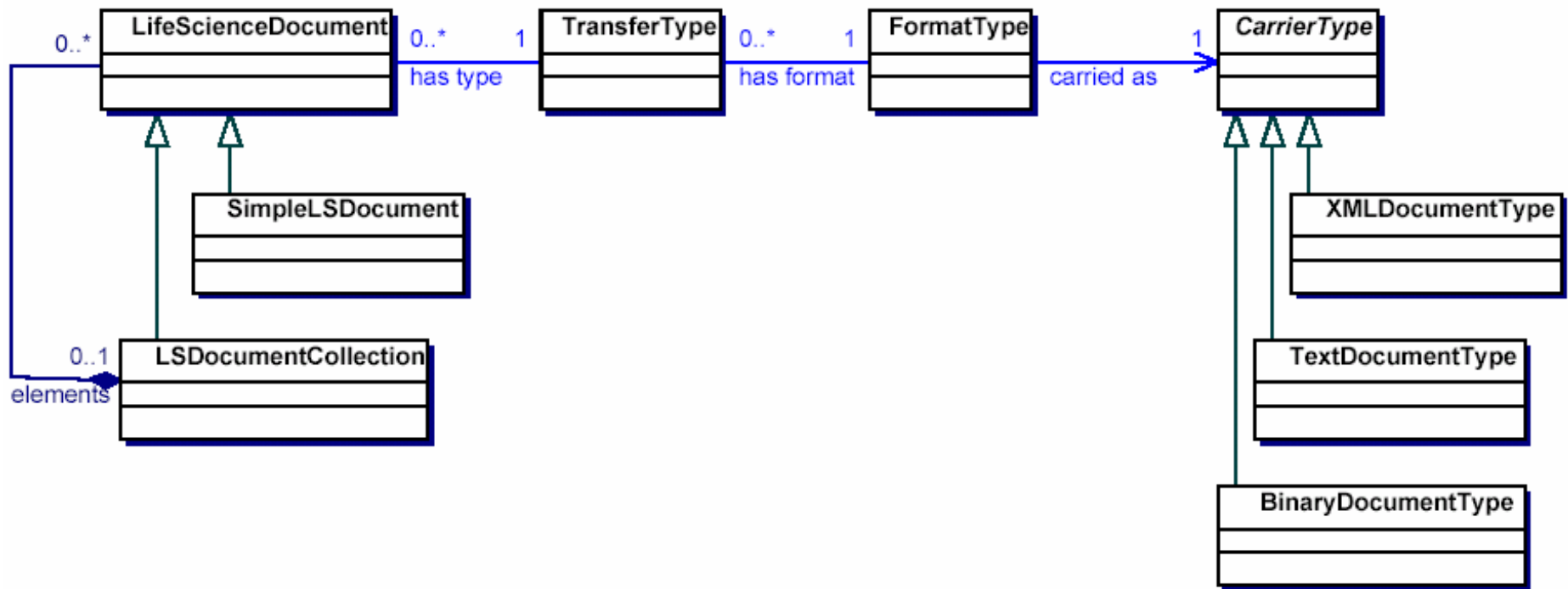
- The MIR data store
- Stores experimental components
 - Workflow specs as XML Scufi docs
 - Data, XML notes
 - Types: XML docs, Relational
- Every entry has Dublin Core provenance attributes
- Every entry can have (multiple) ontology expressions
- Multiple IDs



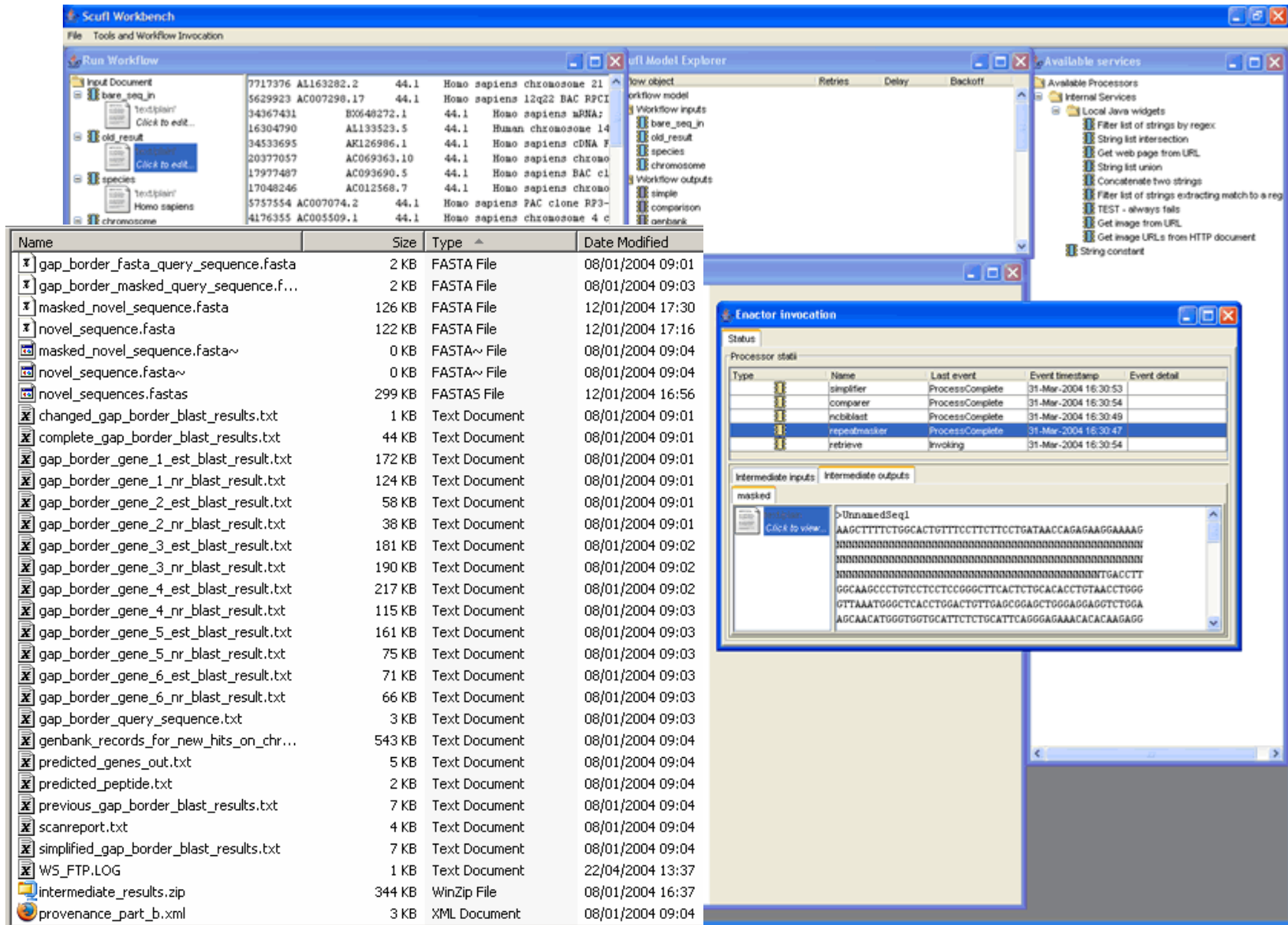
Metamodel for Types

- Necessary to identify the type and format of each datum of interest so that it can (only) be input to type-compatible viewers, services and workflows.

Can't fix this - working in an open world



Intermediate Results



The screenshot displays the Scuff Workbench interface. At the top, there's a menu bar with 'File', 'Tools and Workflow Invocation'. Below it, several panes are visible:

- Run Workflow:** A tree view showing workflow steps like 'Input Document', 'bare_seq_in', 'old_result', 'species', and 'Homo sapiens chromosome'.
- Workflow Model Explorer:** A diagram showing workflow objects, inputs, and outputs.
- Available services:** A list of services including 'Local Java widgets' and 'Internal Services'.
- Enactor Invocation:** A detailed view of a processor's status and intermediate outputs.

The 'Enactor Invocation' window shows the following processor status table:

Type	Name	Last event	Event timestamp	Event detail
	simplifier	ProcessComplete	31-Mar-2004 16:30:53	
	comparer	ProcessComplete	31-Mar-2004 16:30:54	
	ncbiblast	ProcessComplete	31-Mar-2004 16:30:49	
	repeatsmasker	ProcessComplete	31-Mar-2004 16:30:47	
	retrieve	Invoking	31-Mar-2004 16:30:54	

Below the table, the 'Intermediate outputs' section shows a 'masked' sequence:

```
>UnnamedSeq1
AAGCTTTTCTGGCACTGTTCTCTTCTCTGATAACCCAGAGAAGGAAAAG
#####
#####
#####
GGCAAGCCCTGTCTCTCTCCGGGCTTCACTCTGCACACCTGTAACCTGGG
GTTAAATGGGCTCACTTGGACTGTTGAGCGGAGCTGGAGGAGGCTTGGG
AGCAACATGGGTGGTGCATTTCTCTGATTCAGGGAGAAACACACAAGAGG
```

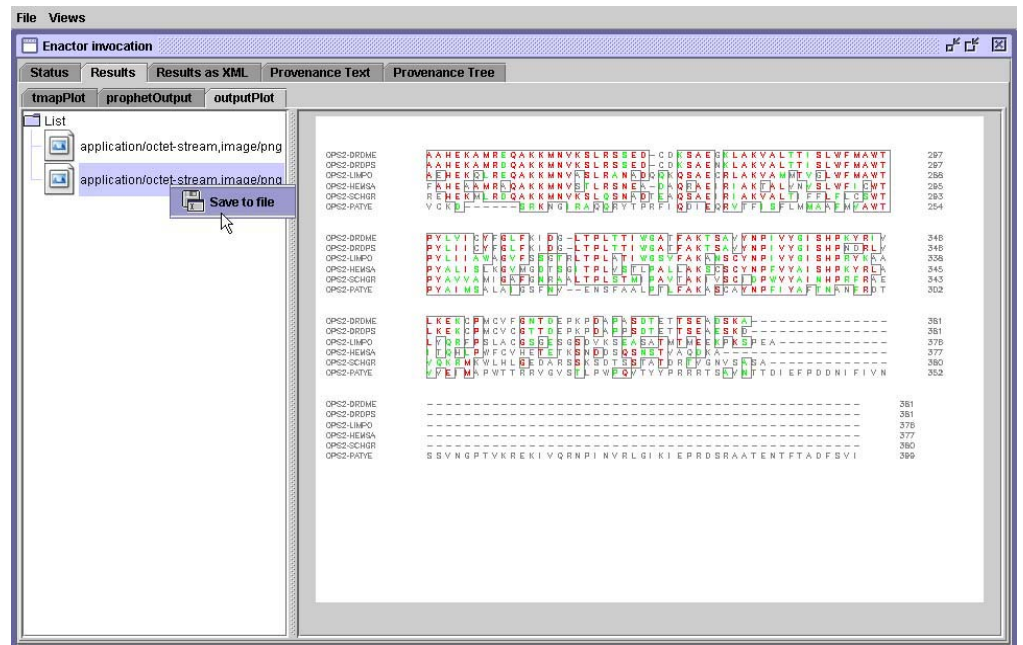
The main window also displays a file explorer with the following table:

Name	Size	Type	Date Modified
gap_border_fasta_query_sequence.fasta	2 KB	FASTA File	08/01/2004 09:01
gap_border_masked_query_sequence.f...	2 KB	FASTA File	08/01/2004 09:03
masked_novel_sequence.fasta	126 KB	FASTA File	12/01/2004 17:30
novel_sequence.fasta	122 KB	FASTA File	12/01/2004 17:16
masked_novel_sequence.fasta~	0 KB	FASTA~ File	08/01/2004 09:04
novel_sequence.fasta~	0 KB	FASTA~ File	08/01/2004 09:04
novel_sequences.fastas	299 KB	FASTAS File	12/01/2004 16:56
changed_gap_border_blast_results.txt	1 KB	Text Document	08/01/2004 09:01
complete_gap_border_blast_results.txt	44 KB	Text Document	08/01/2004 09:01
gap_border_gene_1_est_blast_result.txt	172 KB	Text Document	08/01/2004 09:01
gap_border_gene_1_nr_blast_result.txt	124 KB	Text Document	08/01/2004 09:01
gap_border_gene_2_est_blast_result.txt	58 KB	Text Document	08/01/2004 09:01
gap_border_gene_2_nr_blast_result.txt	38 KB	Text Document	08/01/2004 09:01
gap_border_gene_3_est_blast_result.txt	181 KB	Text Document	08/01/2004 09:02
gap_border_gene_3_nr_blast_result.txt	190 KB	Text Document	08/01/2004 09:02
gap_border_gene_4_est_blast_result.txt	217 KB	Text Document	08/01/2004 09:02
gap_border_gene_4_nr_blast_result.txt	115 KB	Text Document	08/01/2004 09:03
gap_border_gene_5_est_blast_result.txt	161 KB	Text Document	08/01/2004 09:03
gap_border_gene_5_nr_blast_result.txt	75 KB	Text Document	08/01/2004 09:03
gap_border_gene_6_est_blast_result.txt	71 KB	Text Document	08/01/2004 09:03
gap_border_gene_6_nr_blast_result.txt	66 KB	Text Document	08/01/2004 09:03
gap_border_query_sequence.txt	3 KB	Text Document	08/01/2004 09:03
genbank_records_for_new_hits_on_chr...	543 KB	Text Document	08/01/2004 09:04
predicted_genes_out.txt	5 KB	Text Document	08/01/2004 09:04
predicted_peptide.txt	2 KB	Text Document	08/01/2004 09:04
previous_gap_border_blast_results.txt	7 KB	Text Document	08/01/2004 09:04
scanreport.txt	4 KB	Text Document	08/01/2004 09:04
simplified_gap_border_blast_results.txt	7 KB	Text Document	08/01/2004 09:04
WS_FTP.LOG	1 KB	Text Document	22/04/2004 13:37
intermediate_results.zip	344 KB	WinZip File	08/01/2004 16:37
provenance_part_b.xml	3 KB	XML Document	08/01/2004 09:04

Results Management

- Taverna/Freefluo WfEE agnostic about the data flowing through it.
- As objects progress through tagged with terms from ontologies, free text descriptions and MIME types, and which may contain arbitrary collection structures.
- Using the metadata hints we can locate and launch a suitable view.

GGF Summer School 24th July 2004, Italy



Advanced model explorer

Workflow | Object properties |

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

Workflow object	Retries	Delay	Backoff
-----------------	---------	-------	---------

Workflow model

- Workflow inputs
 - fasta_out
 - HGMP_ID
 - HGMP_PVW
 - description
 - email
- Workflow outputs
 - Blastn_est_out
 - peptide
 - CDS
 - GS_Report
 - ListerOut
 - Blastn_nr_out
 - RM_masked
 - RM_out
 - Rtbl
 - PrettyORFs
 - All_ORFs
 - CpG_Islands
 - CpG_Composition
 - PrettyTranslation
 - Restriction Map

Run Workflow

Load Inputs | New Input | New List | Remove

Input Document

- fasta_out
- HGMP_ID
- HGMP_PVW
- description
- email

Load | Load from URL

```
tgccacgtctggagtgcaggggtgtgatctctcctgggtcaagtgtctcctgcctcagttacagatgtgcaccaccacacctgctaatttgggggttctgcatgttggccagcgtggtctctccccatctctgaaaggagaacaacttgcctcctgtcaccagctccatggtccaagaactctgttcccccttccaccacccccagccccagtttagccaatctccactgtcctcttatgtcttaggaaacaaagaaaaataagccagagactggtctactggggaccttccatagacaagcc
```

Run Workflow

Enactor invocation

Save all results

Status | Results | Provenance Tree | Process report

Tess bigAppletUrl	TW out peptide	PrettyTranslation	COPYRIGHT	PrettyORFs
Blastn nr mouse_out	Tess tableOut	Rtbl	Tess tableUrl	CpG Composition
peptide	RM out	Tess poissonUrl		
NIX Results	Tess modelUrl	All ORFs	comparer cds	CpG Islands
Tess smalAppletUrl	CDS	Tess annotatedUrl		
comparer gene	RM masked	6ORF Translation	signalscan_out	Blastn_est mouse_out
TW_out_cds	Tess_annotatedOut	Restriction_Map	Blastn_est_out	Promotorscan_out

List

Sequences producing significant alignments:

```
gi|14458769|gb|BI051239.1|BI051239 CM4-GN0367-110101-664-e02 GN0...
gi|26430541|dbj|BY249029.1|BY249029 BY249029 RIKEN full-length e...
gi|26414151|dbj|BY233041.1|BY233041 BY233041 RIKEN full-length e...
gi|26411480|dbj|BY230370.1|BY230370 BY230370 RIKEN full-length e...
gi|26407984|dbj|BY226875.1|BY226875 BY226875 RIKEN full-length e...
gi|26368441|dbj|BY192371.1|BY192371 BY192371 RIKEN full-length e...
gi|21856503|gb|BQ717606.1|BQ717606 AGENCOURT_8305092 Lupski_symp...
gi|16527352|gb|BM012998.1|BM012998 603637780F1 NIH_MGC_47 Homo s...
gi|11058209|gb|BF180067.1|BF180067 601806426F1 NCI_CGAP_Mam5 Mus...
gi|8893024|dbj|BB224412.1|BB224412 BB224412 RIKEN full-length en...
gi|6224086|gb|AW155217.1|AW155217 mgie0002H03f Rice blast infect...
gi|3054106|gb|AA914714.1|AA914714 vz03e08.rl Soares_mammary_glan...
```

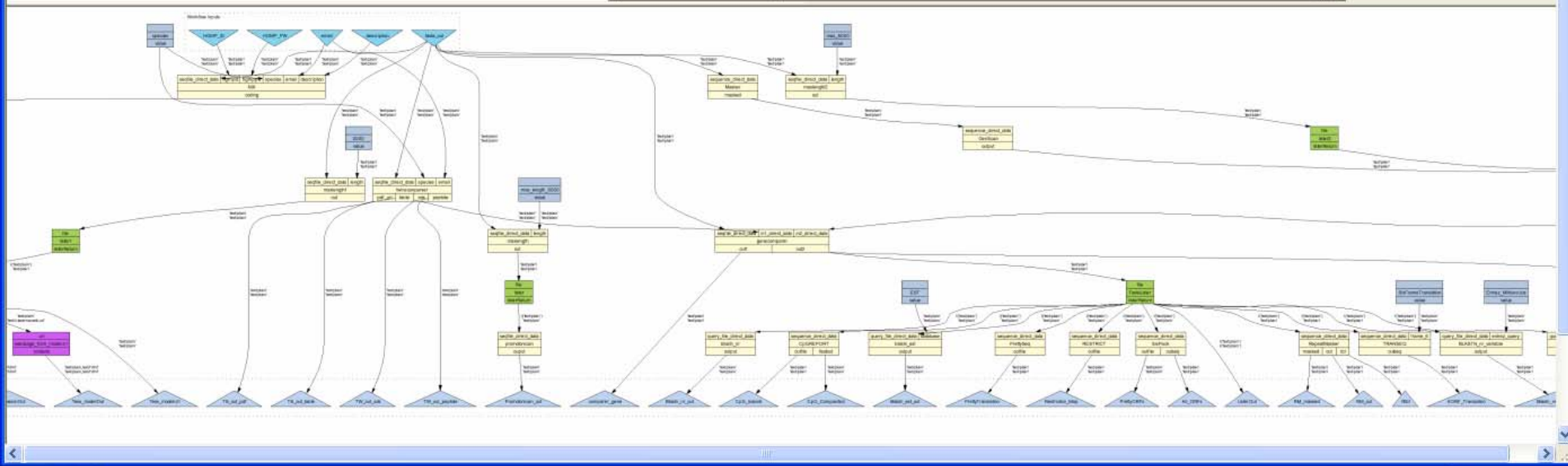
ALIGNMENTS

Available services

ch list

available Processors

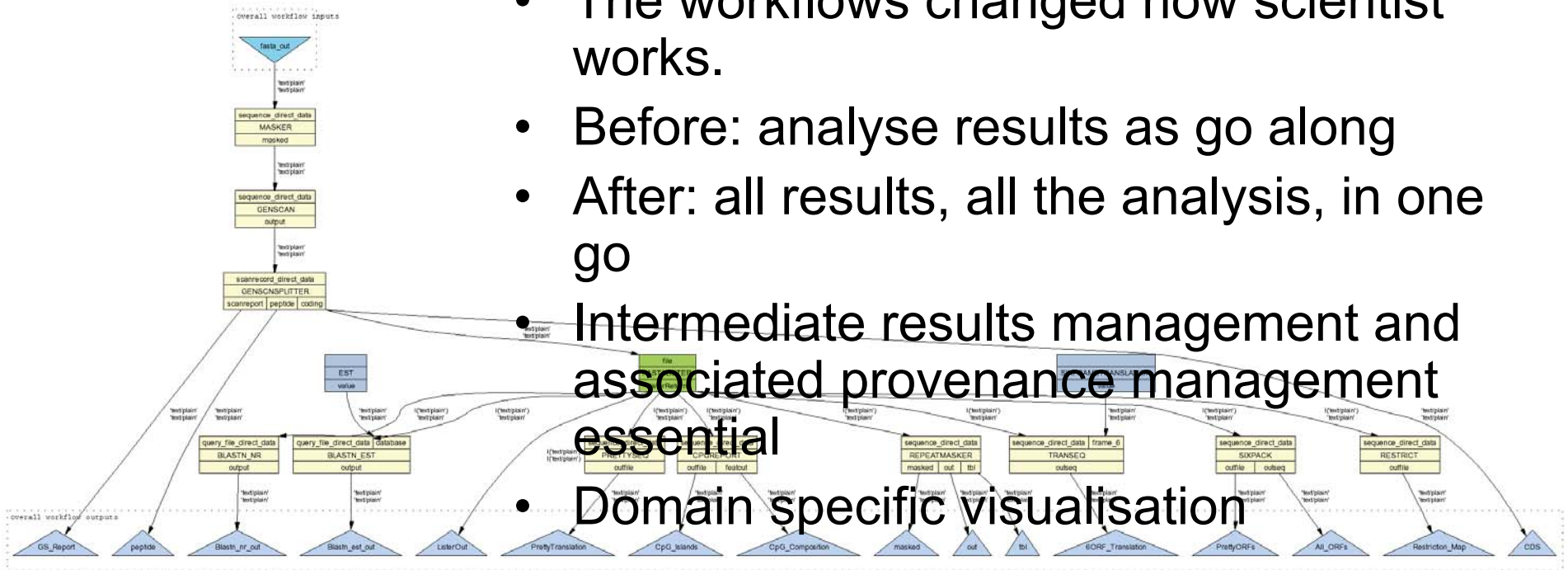
- Local Services
 - Soaplab @ http://industry.ebi.ac.uk/
 - Biomoby @ http://mobycentral.cbr.n
 - WSDL @ http://phoebus.cs.man.ac.
 - porttype: FastaLister [RPC]
 - lister
- Soaplab @ http://phoebus.cs.man.a
- adaptor
- database
- documentation
- parsers
- seq_analysis
 - blastcomparer
 - blasthelp
 - blastsimplifier
 - genbankretrieve
 - genscan
 - genscansplitter
 - ncbiblastwrapper
 - repeatmasker
 - genscanreportparser
 - psortiiwrapper
 - ipsortiiwrapper
 - fasta to numbered



Results Amplification


One input

- Automated annotation workflows produce lots of heterogeneous data
- The workflows changed how scientist works.
- Before: analyse results as go along
- After: all results, all the analysis, in one go
- Intermediate results management and associated provenance management essential
- Domain specific visualisation



Many outputs

Schema.ad Schema.ad urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12

 urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12

Commands

- File away
- Rename
- View in QMol

Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Pubmed

None specified; click here to add Edit ▾


Sequence Summary

Name: None specified; click here to add ▾

urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:pred... Edit ▾

taxon:9606 Edit ▾

Extracted resources

 **Extracted resources**

Commands

- Add existing items
- Add new item
- Clear collection/list
- Create checkbox aspect
- File away
- Rename

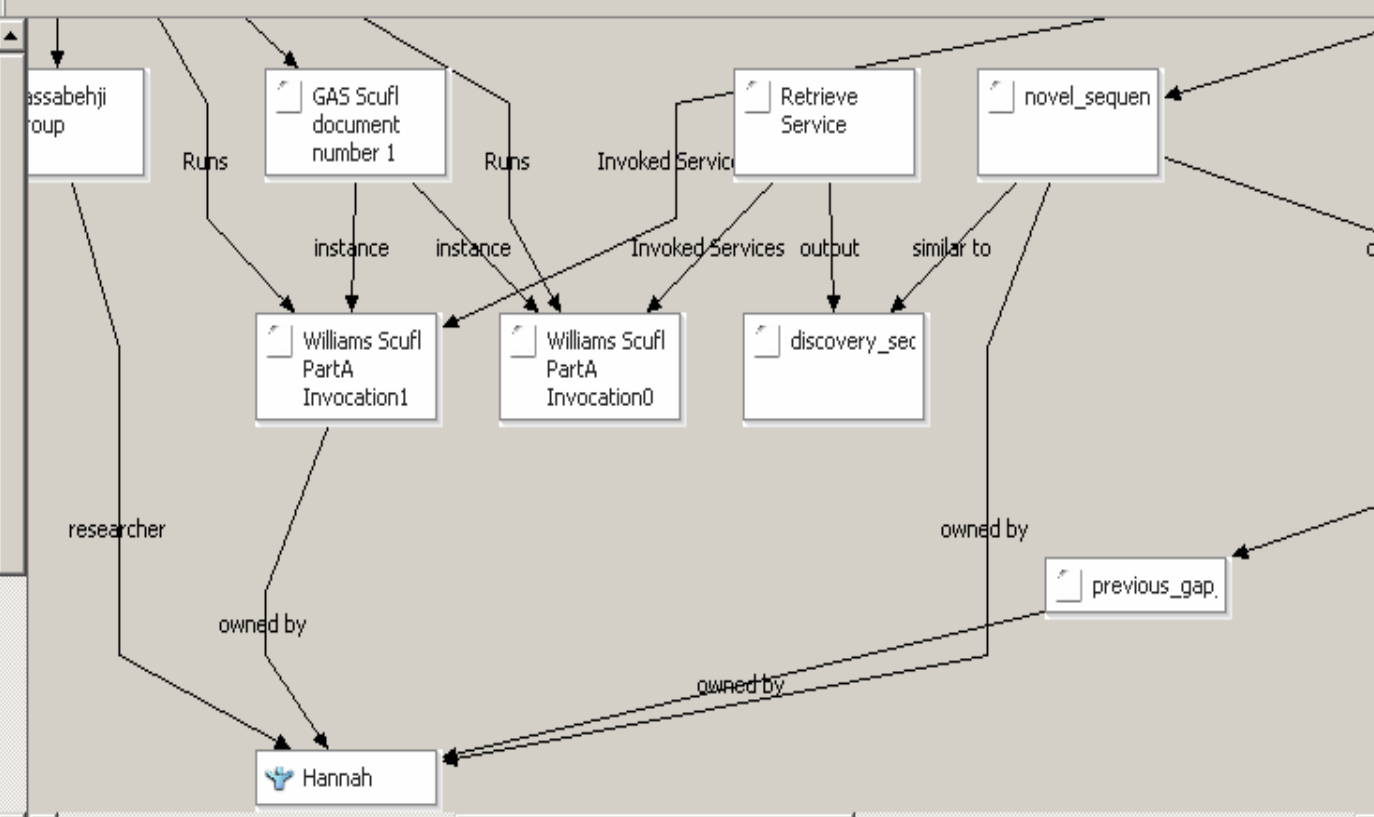
Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Available views

- Browse view
- Calendar



```

graph TD
    subgraph "assabehji group"
        A[assabehji group]
    end
    subgraph "GAS ScufI"
        B[GAS ScufI document number 1]
    end
    subgraph "Retrieve Service"
        C[Retrieve Service]
    end
    subgraph "novel_sequen"
        D[novel_sequen]
    end
    subgraph "Williams ScufI PartA"
        E[Williams ScufI PartA Invocation1]
        F[Williams ScufI PartA Invocation0]
    end
    subgraph "discovery_sec"
        G[discovery_sec]
    end
    subgraph "previous_gap"
        H[previous_gap]
    end
    subgraph "Hannah"
        I[Hannah]
    end

    A -- Runs --> B
    A -- Runs --> E
    B -- instance --> E
    B -- instance --> F
    C -- Invoked Service --> E
    C -- Invoked Service --> F
    C -- output --> G
    D -- similar to --> G
    E -- owned by --> I
    F -- owned by --> I
    G -- owned by --> I
    H -- owned by --> I
  
```

Choose an arr...

- Bioinformatic arrows
- Provenance Graph
- Show arrows based on the ontology
- Show arrows based on the schema

Add ▾

Available arrows for Provenance Graph

- Designs
- Invoked Services
- 11 more it...

Domain Services

- Native WSDL Web services
 - DDBJ, NCBI BLAST, PathPort
- BioMOBY Web services
 - Single function stereotype
- Wrapped legacy services
 - Stateful interaction stereotype
 - One button wrapping
 - SoapLab for command-line tools

For each application
CreateJob
Run
WaitFor
GetResults
Destroy

GGF – Summer School 24th July 2014, slide 11
– SoapLab for screen scraped web pages

Domain Services

- Lots of them ~ 300
- Open world: we don't own them
- Many produce **text** not numbers
- Many are unique, single site
- Need lots of genuine **redundant replica** services
- **Unreliable and unstable**
 - Research level software
 - Reliant on other peoples servers
- Services in the wild rare - **significant time to wrap** applications as web services (licensing, installation,

Domain Services in WBS

- Repeatmasker
- NCBI_BLAST
- Modified BLAST
- GenScan
- PSORTII
- iPSORT
- TargetP
- Various EMBOSS services
- InterProScan
- BLAST2
- NIX
- TESS
- TWINSCAN
- Alibaba?



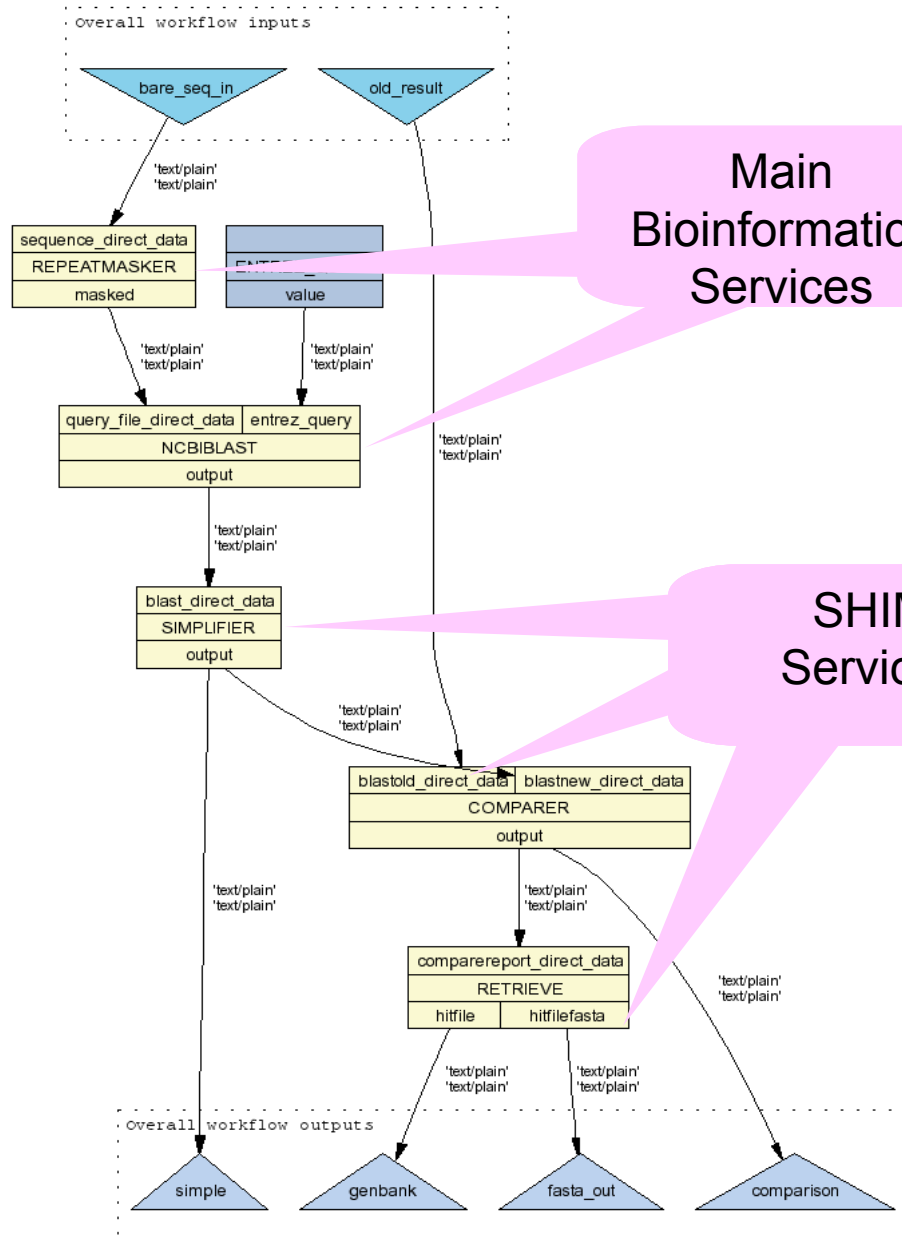
Can you guess what it is yet?

```
<?xml version='1.0' encoding='UTF-8'?>
<definitions name='Blast'>
  <message name='execute0In'>
    <part name='accession' type='xsd:string' />
  </message>
  <message name='execute0Out'>
    <part name='Result' type='xsd:string' />
  </message>
  <portType name='Blast'>
    <operation name='execute' parameterOrder='accession'>
      <documentation>Execute Blast</documentation>
      <input name='execute0In' message='tns:execute0In' />
      <output name='execute0Out' message='tns:execute0Out' />
    </operation>
  </portType>
  <binding name='Blast' type='tns:Blast'>
    <soap:binding style='rpc'
      transport='http://schemas.xmlsoap.org/soap/http' />
    <operation name='execute'>
      <soap:operation soapAction='execute' style='rpc' />
      <input name='execute0In'>
        <soap:body use='encoded' />
      </input>
      <output name='execute0Out'>
        <soap:body use='encoded' />
      </output>
    </operation>
  </binding>
</definitions>
```

GGF Summer School 24th July 2004, Italy



SHIM Services



- Explicitly capturing the process
- Unrecorded 'steps' which aren't realised until attempting to build something
- Services that enable domain services to fit together

Workflow development and enactment

- Freefluo workflow enactment engine
 - Processor & event observer plugin support
- Taverna development and execution environment
 - Workbench, workflow editor, tool plug-in support
- <http://taverna.sourceforge.net>
- Simple conceptual unified flow language (XScufl) wraps up units of

The screenshot displays the Scufi Workbench interface, which is used for workflow development and execution. The main window is titled "Run Workflow" and shows an "Input Document" with XML content. The interface includes several panels: "Run Workflow" (top left), "Scufi Model Editor" (top center), "Available services" (top right), and "Scufi Diagram" (bottom left). The "Available services" panel lists various tools like "Filter list of strings by regex", "String list intersection", "Get web page from URL", etc. The "Scufi Diagram" panel shows a workflow graph with nodes and connections. The "Input Document" panel shows XML content:

```
<?xml version="1.0" encoding="UTF-8"
<b:dataThingMap xmlns:b="http://org.
```

Taverna Workbench

Tom Oinn, Matthew Pocock, Justin Ferris, Darren Marvin, Kevin Glover, Tim Carver, Mark Greenwood, Peter Li, Anil Wipat and the rest of the myGrid team.

tree structure explorer

The 'Advanced model explorer' window displays a hierarchical tree structure of workflow objects. The root is 'Workflow model', which branches into 'Workflow inputs' and 'Workflow outputs'. Under 'Workflow inputs', there are several file-related objects like 'file1_id', 'file1_namespace', 'file1_value', 'file1_restrictout', 'file1_name', and 'file0_images'. Under 'Workflow outputs', there are 'Processors' such as 'file0_id : cho', 'file0_namespace : DragonDB:Allele', 'file0_Create_moby_data', 'file0_Parse_moby_data', 'file1_namespace : AGI_LocusCode', 'file1_id : At2g17950', 'file0_Decode_base64_to_byte', 'file1_Create_moby_data', 'file1_Parse_moby_data1', 'file1_Parse_moby_data', and 'file1_backtranseq'.

service palette shows a range of operations which can be used in the composition of a workflow

The 'Available services' window shows a search list and a 'Watch loads' checkbox. Below, a tree structure lists various services, including 'plantspGetRNA', 'www.pathbase.net', 'getPBfromGO', and 'arabidopsis.info'. Under 'arabidopsis.info', there are numerous services like 'get_ATH_insert_by_EMBL', 'getNASC_codebyKeyword', 'getArabidopsisImage_by_Keyword', 'getNASC_codebyAGI_locus', 'getPhenotype_by_NASC', 'getArabidopsisImage_by_NASC', 'getPhenotype_by_Keyword', 'get_insert_names_by_AGI', and 'get_ATH_insert_by_NASC_code'.

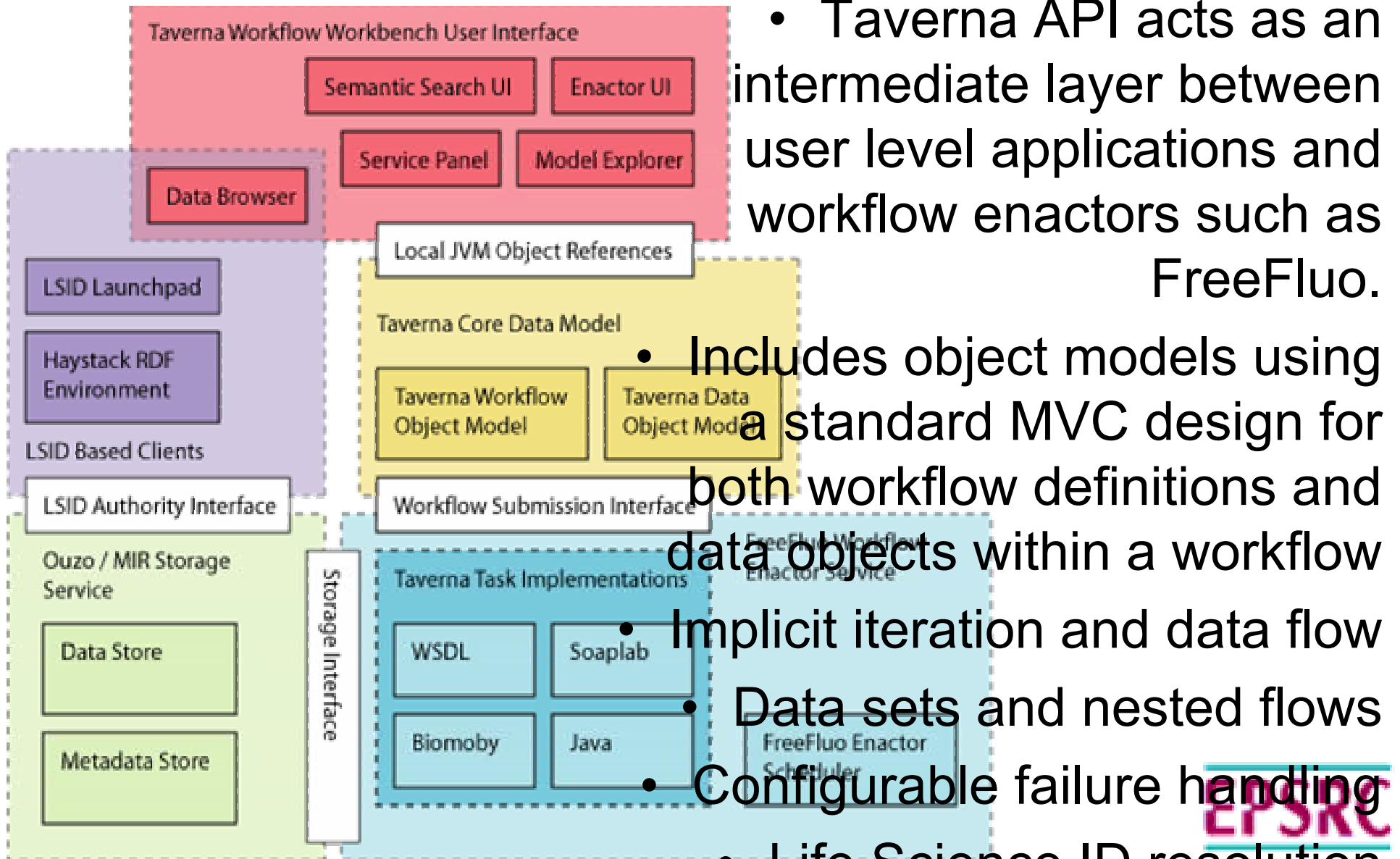
Results in enactor invocation window

The 'Enactor invocation' window shows a list of files under the 'file1_namespace' tab. The list contains several 'application/octet-stream,image/jpeg' entries. A context menu is open over one of the entries, showing options like 'Save to file', 'Viewers', 'Image', and 'Table'. The 'Image' option is selected, and a large image of a pink and white flower is displayed in the main area. A small white card with the text '51 CHO' is visible in the bottom left corner of the image.

graphical diagram

The 'Workflow diagram' window shows a graphical flowchart of the workflow. It starts with 'file0_namespace' and 'file0_id' boxes, which lead to a 'file0_Create_moby_data' box. This is followed by a 'file0_Parse_moby_data' box, then a 'file0_Decode_base64_to_byte' box. The final output is 'file0_images'. The diagram also shows the relationships between the workflow objects and the enactor invocation window.

Workflow environment

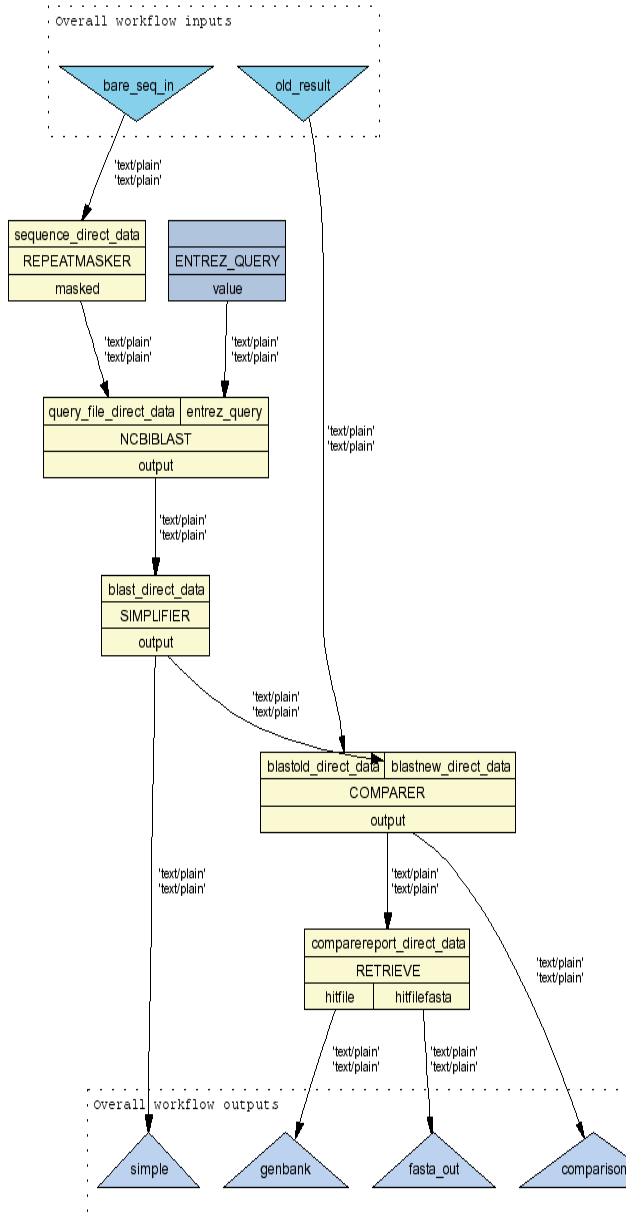


- Taverna API acts as an intermediate layer between user level applications and workflow enactors such as FreeFluo.

- Includes object models using a standard MVC design for both workflow definitions and data objects within a workflow
- Implicit iteration and data flow
- Data sets and nested flows
- Configurable failure handling



Scufl-Taverna-FreeFluo



- SCUFL - Simple Conceptual Unified Flow Language
- Started with WSFL ☹️ ... SCUFL provides a much higher level view on workflows, and therefore simpler and more user-focused.
- **Simple** – relies upon an inherently connected environment to reduce the

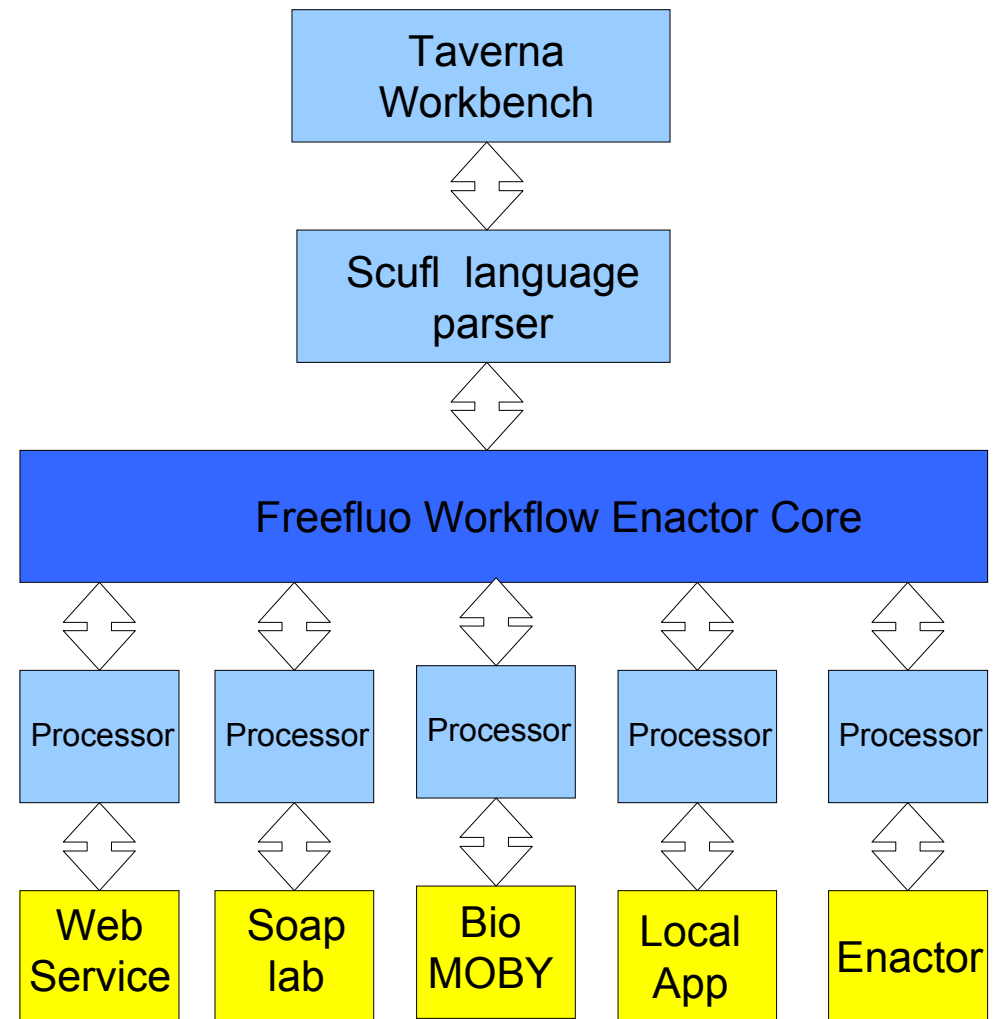
- Conceptual – one Processor in a SCUFL workflow maps as far as is possible to one conceptual operation as viewed by a non expert user

- Wrap up stateful service interactions into custom Processor

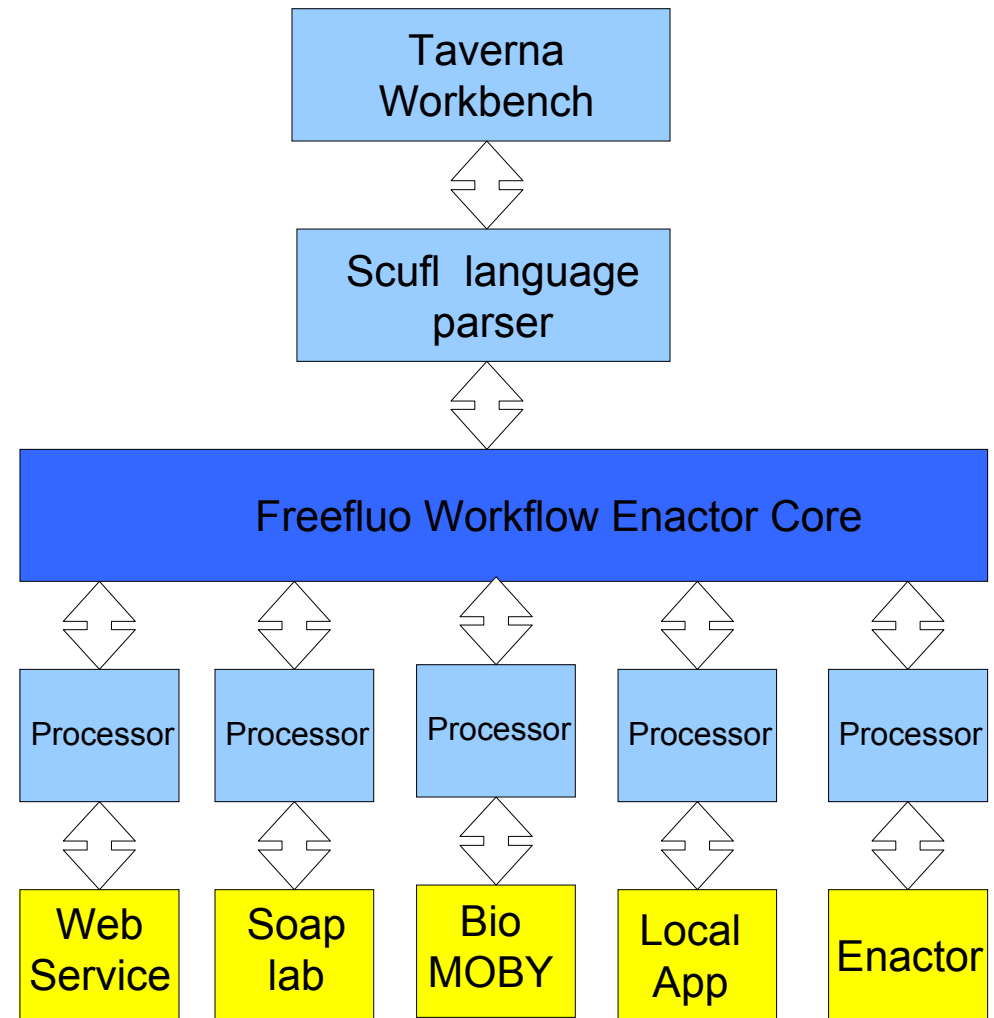
implementations

GGF Summer School 24th July 2004, Italy

- Lowers the barrier

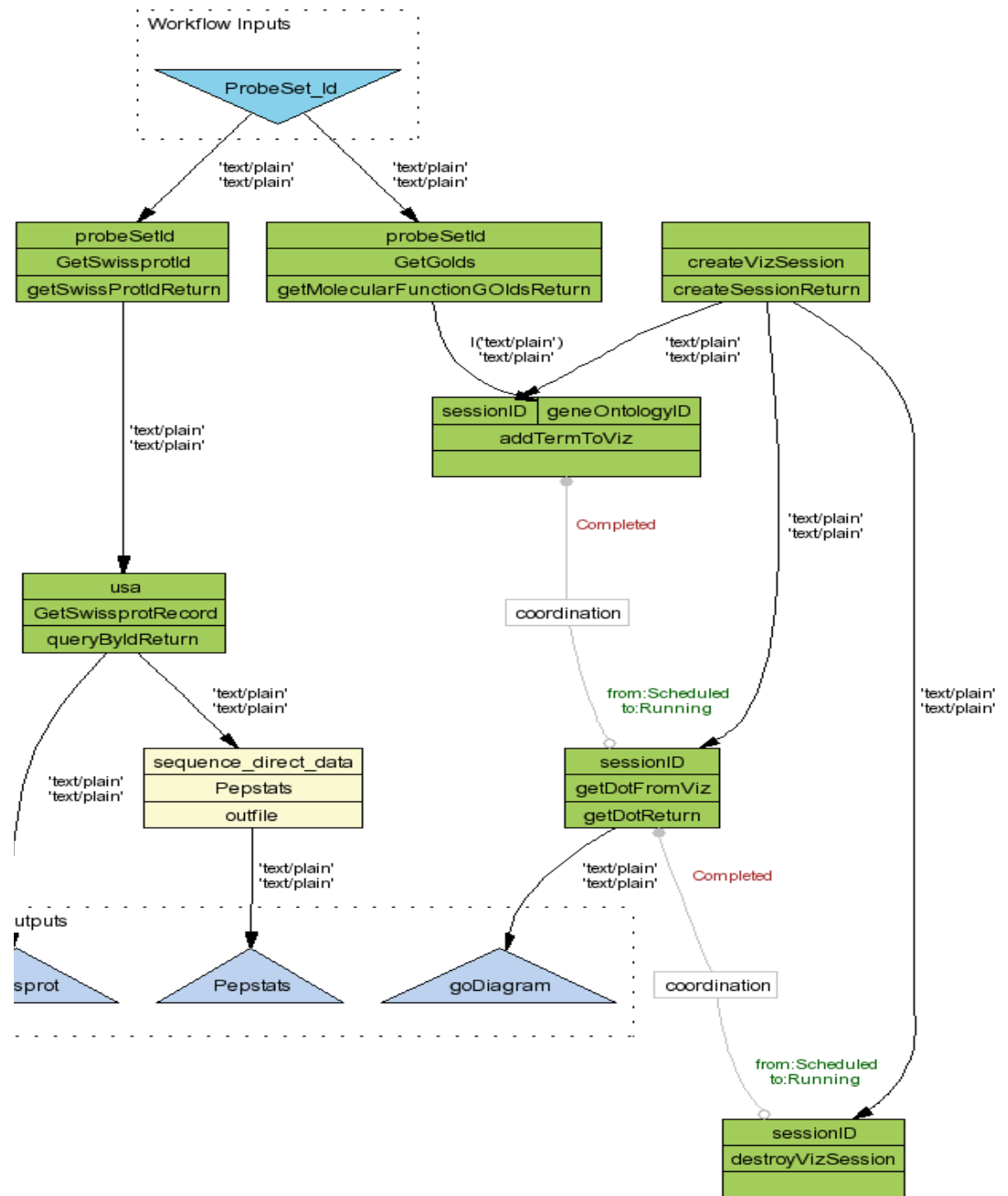


- Unified Flow Language – SCUFL does not dictate how the workflow is to be enacted, it is inherently declarative in intent.
- Can potentially be translated to other workflow languages.
- **Can be arbitrarily abstract any given**

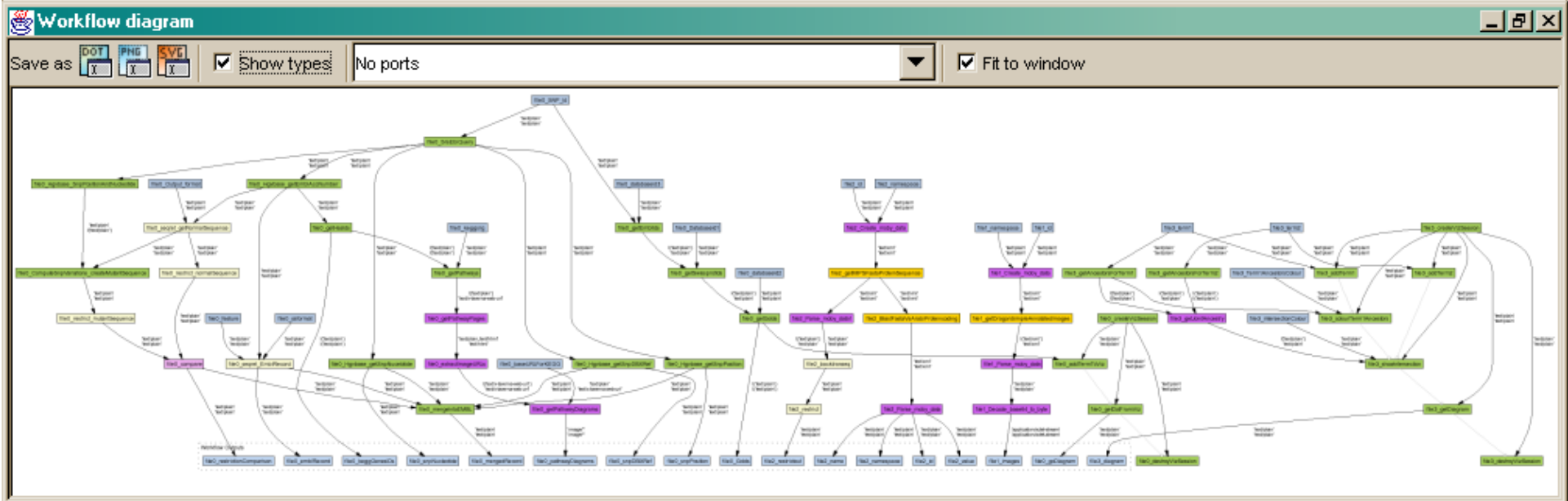




- One input, three outputs and eight processors.
- All the processors are labeled top to bottom with input ports, processor name and output ports.
- All the processors here are standard WSDL-described standard web services except for

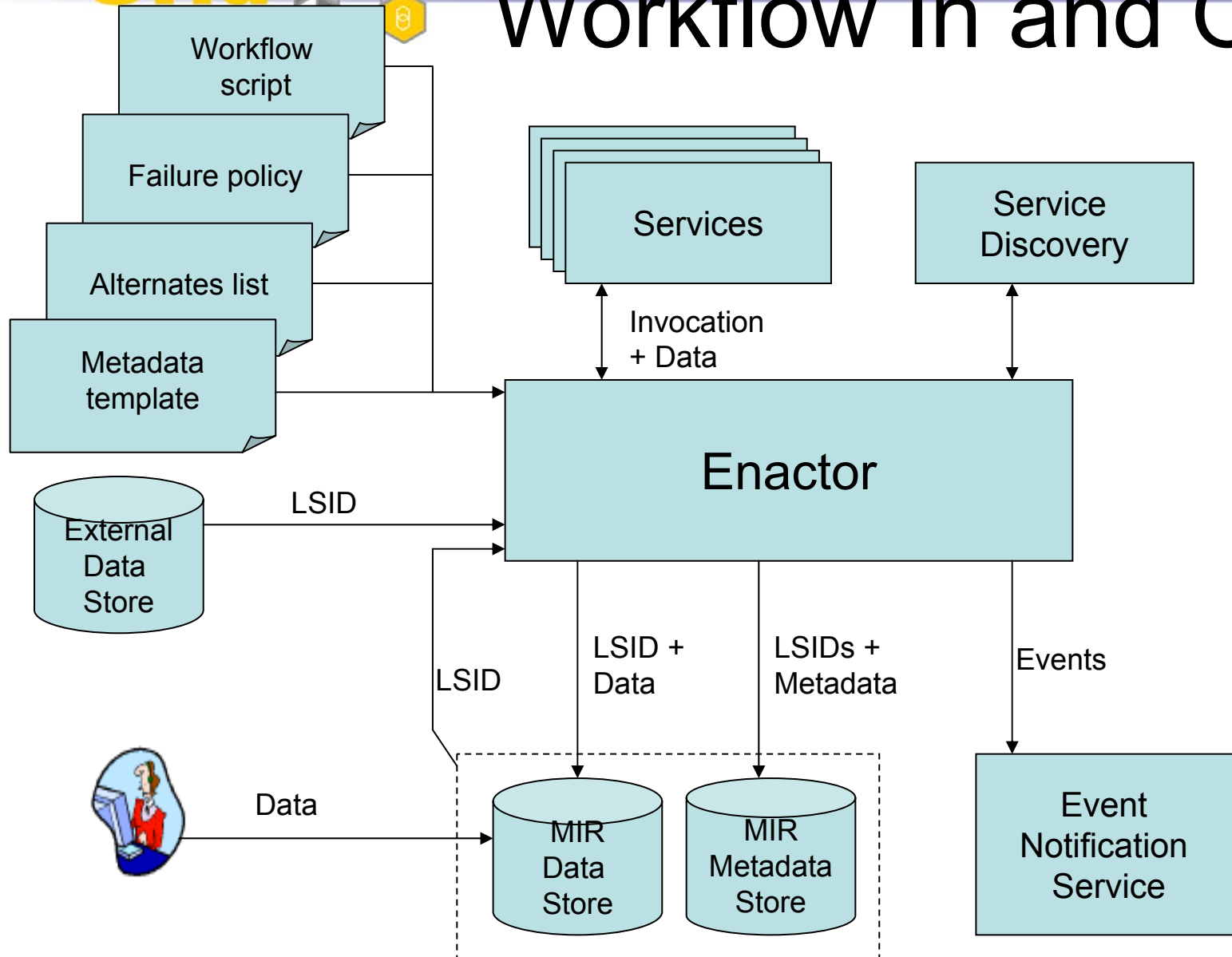


Tools and Workflow Invocation



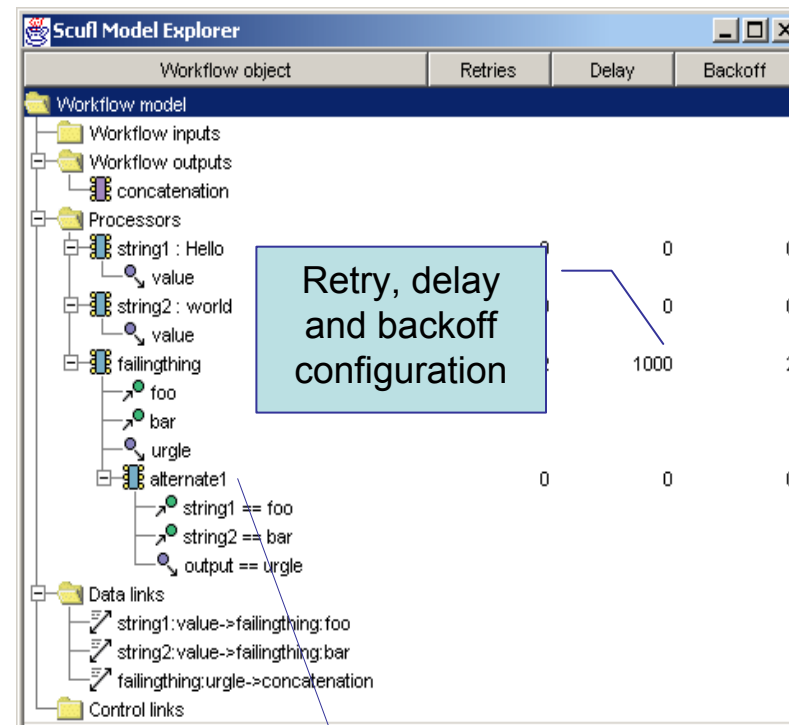


Workflow In and Outs



Fault tolerance

- Failure of workflow engine
 - P2P architecture
 - XML serialisation
 - Checkpointing
- Failure of services or network
 - User defined retry policy
 - Alternate replicas
 - Alternate list

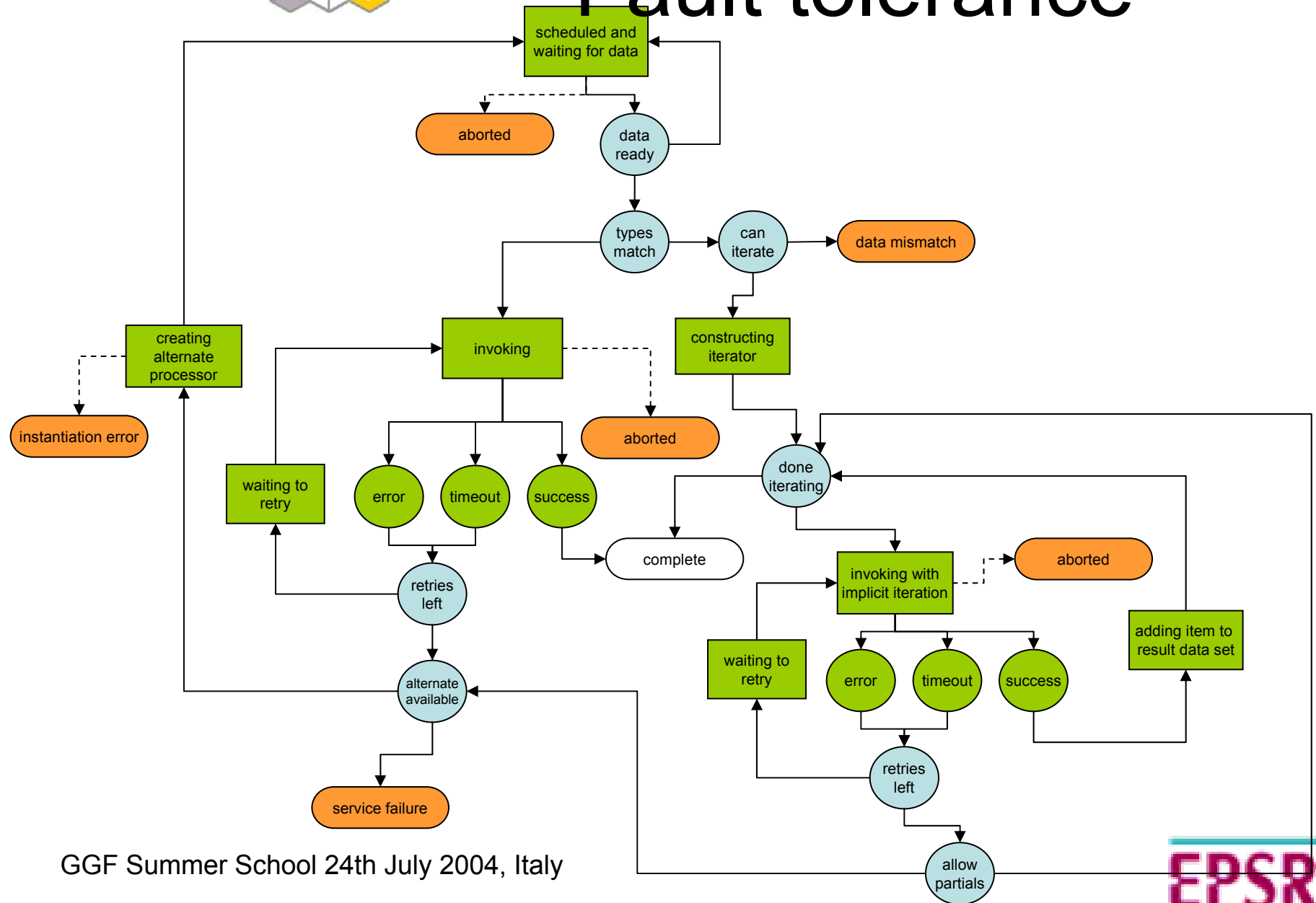


Retry, delay and backoff configuration

Alternate Processor



Fault tolerance



Status reporting

Enactor invocation

Save all results

Status | Provenance Tree | Process report

Processor status

Type	Name	Last event	Event timestamp	Event detail
	comparer	ProcessScheduled	09-Jul-2004 23:24:04	
	simplifier	ServiceError	09-Jul-2004 23:24:57	Message='Output 'outp...
	repeatmasker	ProcessComplete	09-Jul-2004 23:24:18	
	ncbiblast	ProcessComplete	09-Jul-2004 23:24:53	
	ebi_blast_ncbi	Invoking	09-Jul-2004 23:24:18	

Intermediate inputs | Intermediate outputs

query_sequence

```
>UnnamedSeq1
AAGCTTTTCTGGCACTGTTTCCTTCTTCTCGATAACCAGAGAAGGAAAAG
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
GGCAAGCCCTGTCTCCTCCGGGCTTCACTCTGCACACCTGTAACCTGGG
GTTAAATGGGCTCACCTGGACTGTTGAGCGGAGCTGGGAGGAGGTCTGGA
```

Enactor invocation

Status | Results | Results as XML | Provenance Text | Provenance

- workflowReport workflowID="FlowID.org.embl.ebi.escience"
 - processorList
 - processor name="string1"
 - processor name="string2"
 - processor name="failingthing"
 - ProcessComplete TimeStamp="13-Feb-2004 13:56:22"
 - Invoking TimeStamp="13-Feb-2004 13:56:25"
 - AlternateProcessScheduled TimeStamp="13-Feb-2004 13:56:22"
 - s:local maxretries="0", retrybackoff="0.0", retrydelay="0", xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha"
 - org.embl.ebi.escience.scuflworkers.java.StringConcat
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:25"
 - WaitingToRetry MaxRetries="2", RetryNumber="2", TimeDelay="2000", TimeStamp="13-Feb-2004 13:56:23"
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:23"
 - WaitingToRetry MaxRetries="2", RetryNumber="1", TimeDelay="1000", TimeStamp="13-Feb-2004 13:56:22"
 - ServiceError Message="This processor always fails!", TimeStamp="13-Feb-2004 13:56:22"
 - Invoking TimeStamp="13-Feb-2004 13:56:22"
 - ProcessScheduled TimeStamp="13-Feb-2004 13:56:22"
 - s:local maxretries="2", retrybackoff="2.0", retrydelay="1000", xmlns:s="http://org.embl.ebi.escience/xscufl/0.1alpha"
 - org.embl.ebi.escience.scuflworkers.java.TestAlwaysFailingProcessor

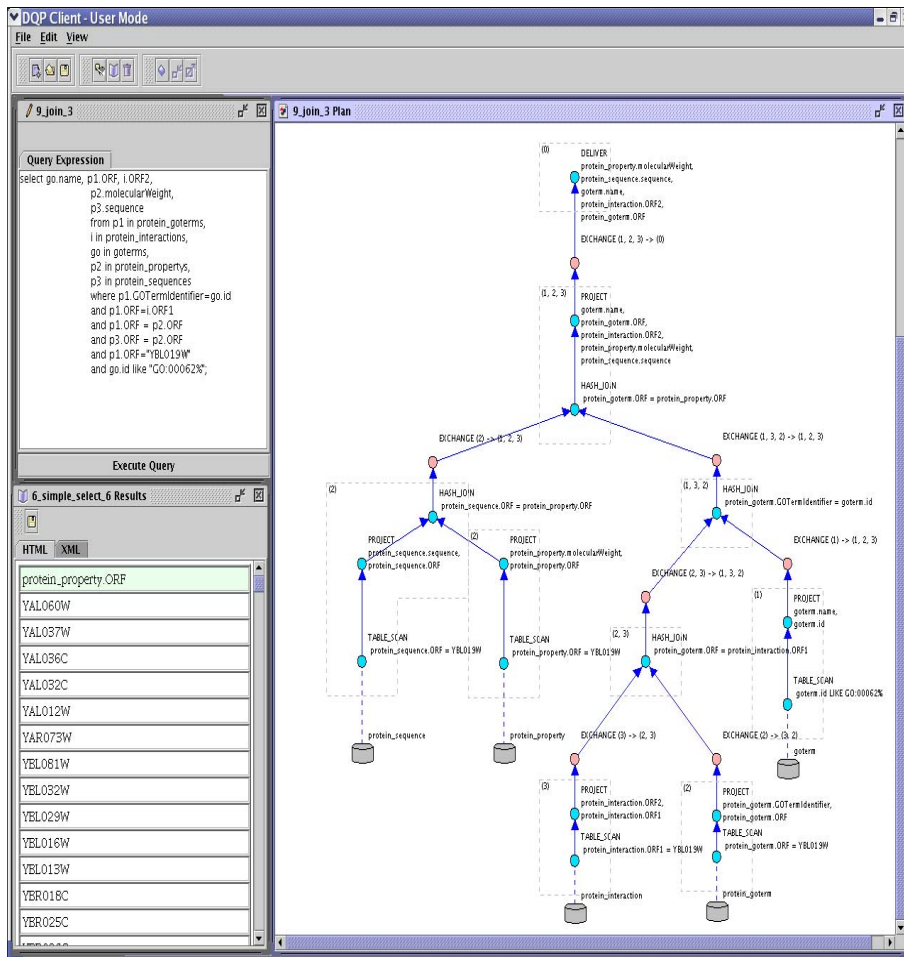
Whither BPEL?

- Focus: scripting simple request/response services vs. choreographing business processes
- Complexity: Scuf is simple enough for bioinformaticians to develop workflows
- Generality: Extensible *processor* support vs. Web Services only
- Provenance generation

What needs to be done

- Free-standing web service
- Long-running workflows
 - Computationally-intensive services
 - Access to a reliable high performance BLAST service that reflects NCBI Blast – NCBioGrid?
- Scalability
 - Large documents – data staging
- Debugging environment – services / workflows are brittle.
- Interactivity
 - Version 1 had user proxy as an actor
 - The Original Process split into 3 steps:
 - Identification of candidate overlapping nucleotide sequences
 - Characterisation of nucleotide sequence

OGSA-DQP



The screenshot shows the DQP Client interface with the following components:

- Query Expression:**

```
select go.name, p1.ORF, i.ORF2,
p2.molecularWeight,
p3.sequence
from p1 in protein_goterm,
i in protein_interactions,
go in goterm,
p2 in protein_properties,
p3 in protein_sequences
where p1.COTermIdentifier=go.id
and p1.ORF=i.ORF1
and p1.ORF = p2.ORF
and p3.ORF = p2.ORF
and p1.ORF="YBL019W"
and go.id like "CO.00062%";
```
- Execute Query:** A button to execute the query.
- 6. simple_select.6 Results:** A table of results with columns for HTML and XML. The first row is highlighted:

protein_property.ORF
YAL060W
YAL037W
YAL036C
YAL032C
YAL012W
YAR073W
YEL081W
YEL092W
YEL029W
YEL016W
YEL013W
YER018C
YER025C
- 9_join.3 Plan:** A complex query plan diagram showing the execution flow. It includes nodes for DELIVER, PROJECT, HASH_JOIN, EXCHANGE, and TABLE_SCAN, connected by arrows indicating data flow. The plan is annotated with various ORF values and identifiers.

- Used in Grave's Disease
- Uses OGSA-DAI data access services to access individual data resources.
- A single query to access and join data from more than one OGSA-DAI wrapped data resource.
- Supports orchestration of computational as

<http://www.ogsa-dai.org.uk/dqp>



Roadmap

- Part 1
 - Application context
- Part 2
 - Architecture
 - Information and Workflows
 - Semantics and provenance
- Part 3
 - Wrap up



Finding and selecting services

Activation energy gradient

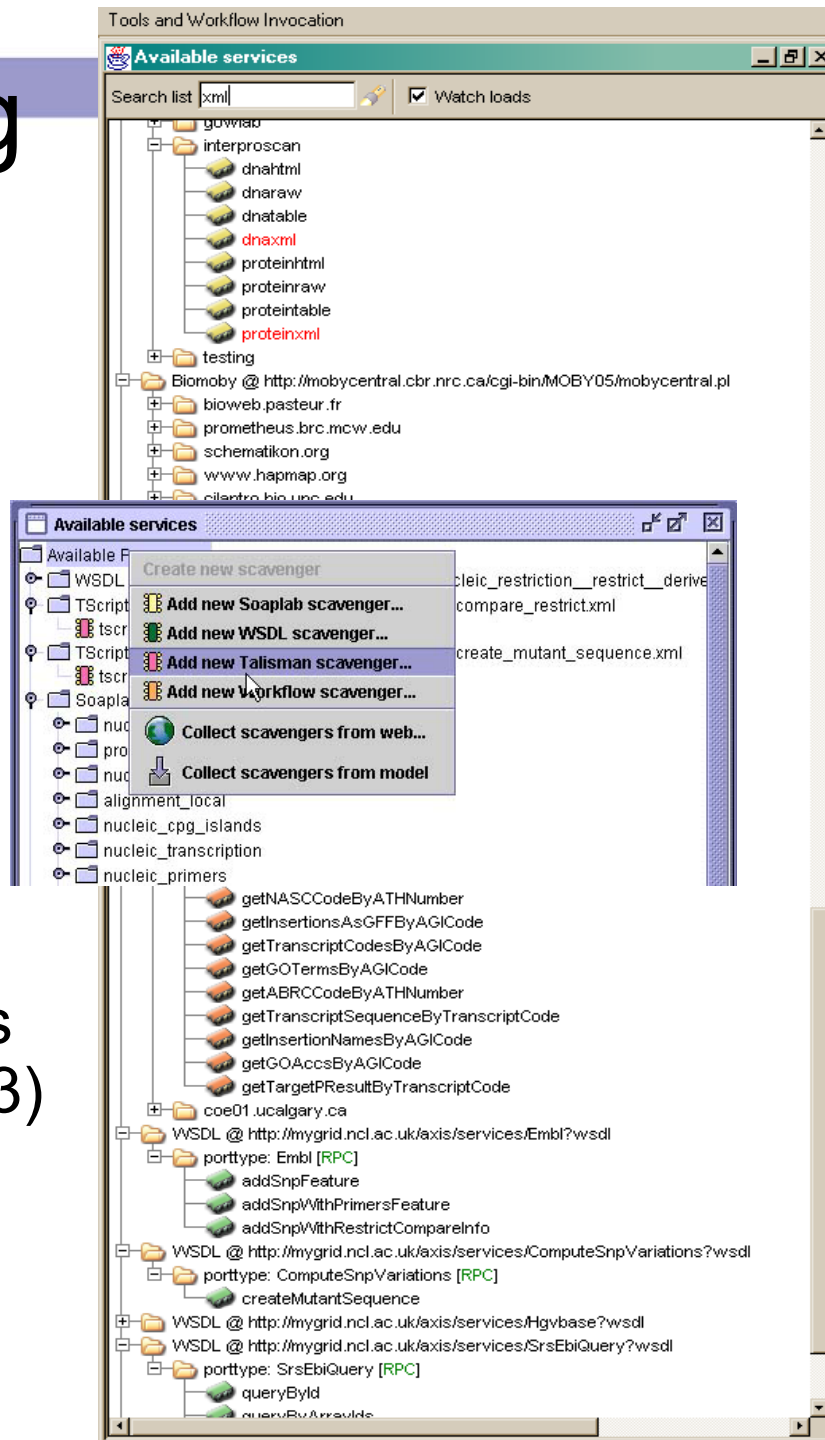
Unregistered services

- Scavenging
- URLs and Soaplab endpoints
 - Introspection

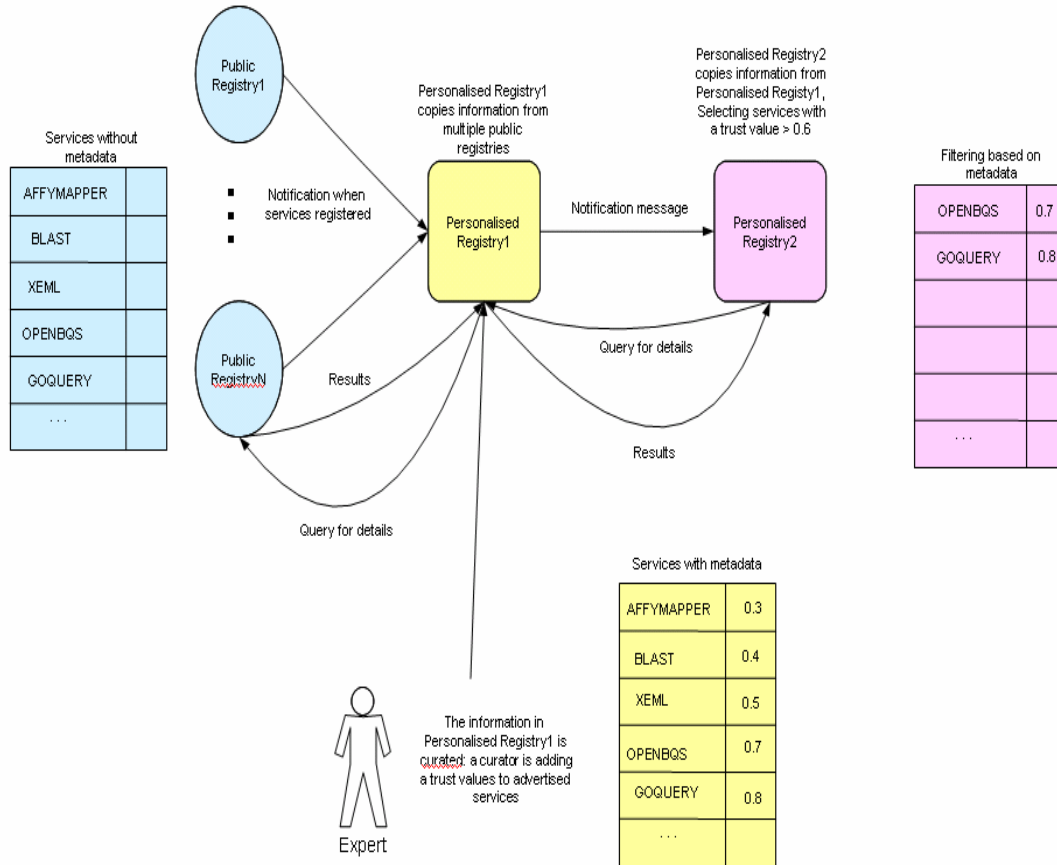
Registered services

- Word-based searching
- Semantic annotation for later discovery and (re)use by friends and strangers in your VO (Part 3)

Drag and drop services onto Taverna workbench

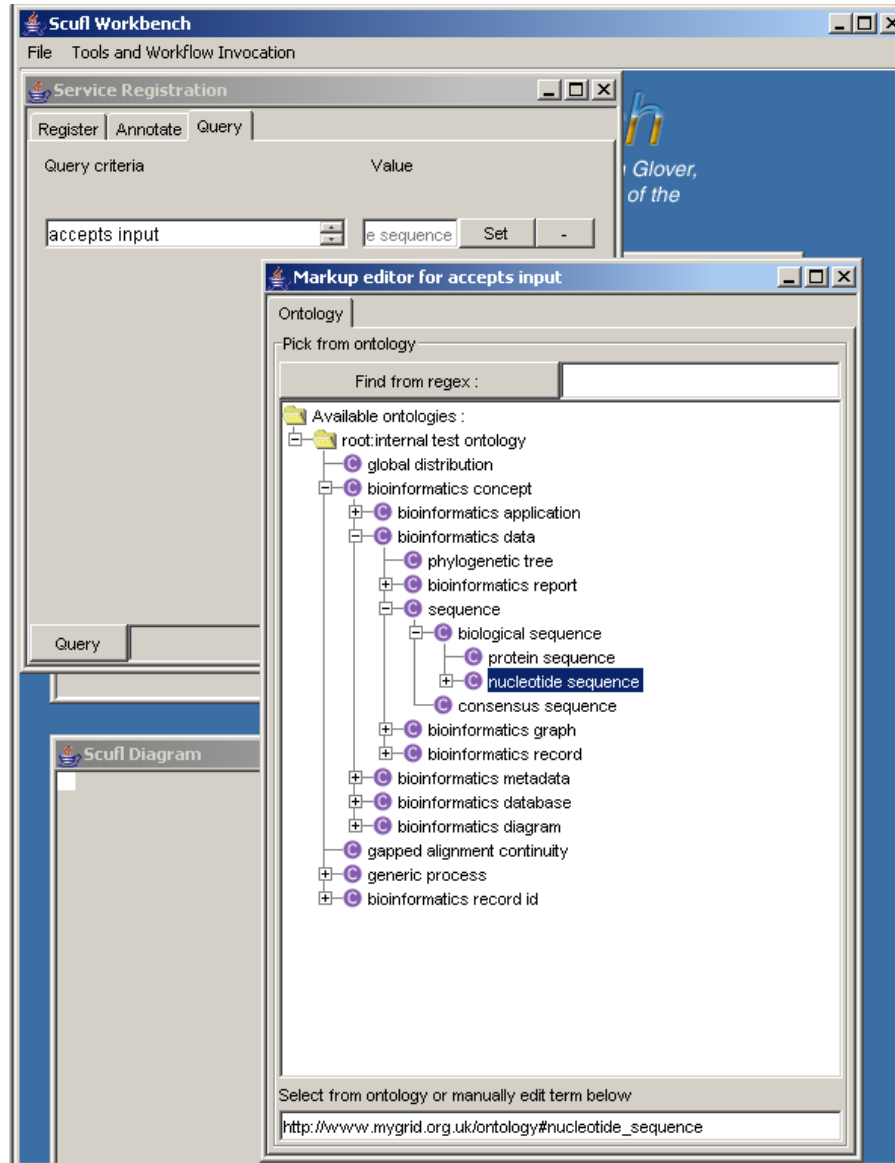


Registry View Service



- Registry
- Third party registries
- Third party services
- Third party annotation (RDF)
- Views over federated registries
- UDDI interfaces extended with RDF
- Federated views
 - Updated via Notification Service
 - Personalized based on Annotation
- Authorisation and IPR

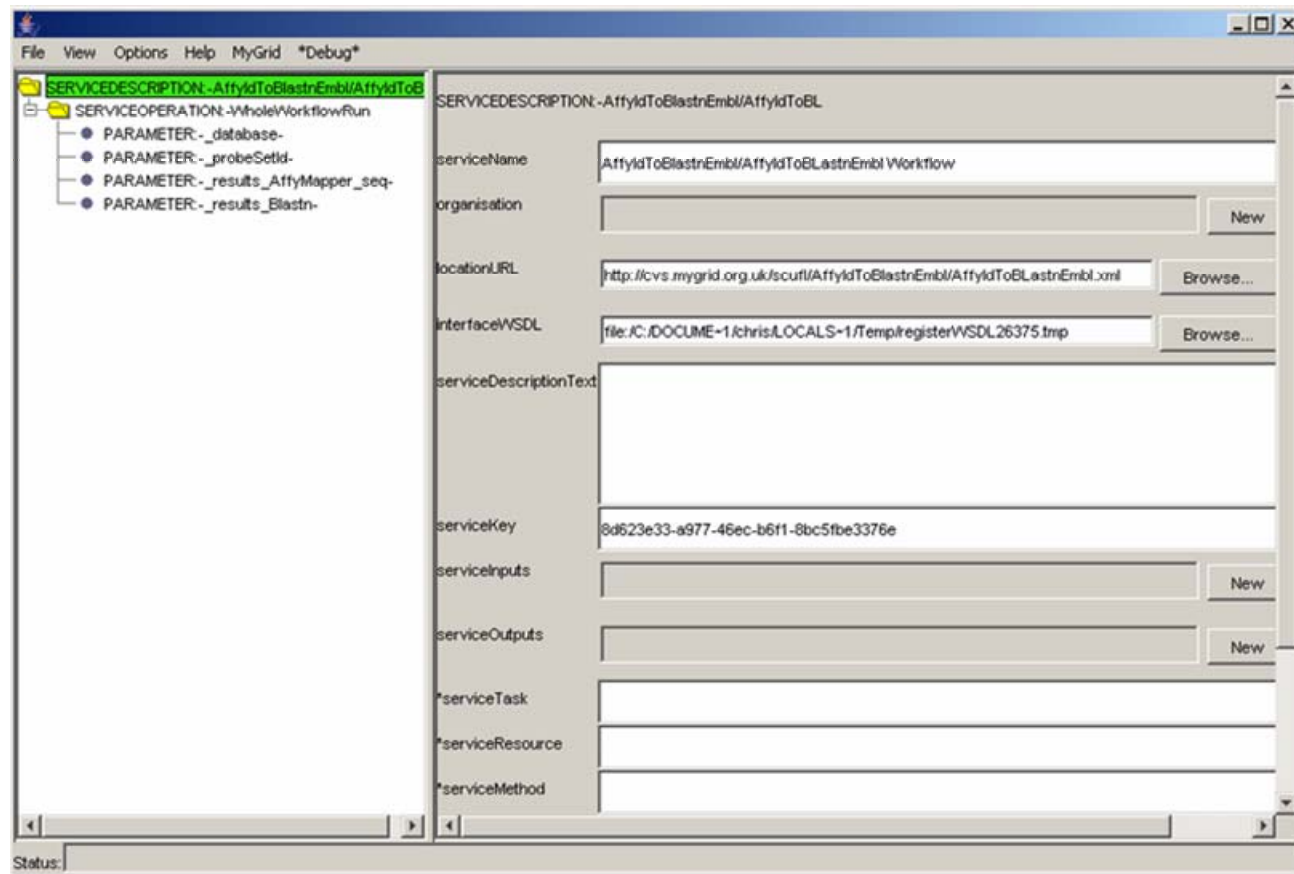
Semantic discovery



- User chooses services
- A common ontology is used to annotate and query any myGrid object including services.
- Discover workflows and services described in the registry via Taverna.
- Look for all workflows that accept an input of semantic type nucleotide sequence
- Aim to have semantic discovery over public view on the Web.



Workflow and service annotation



- Adding structured metadata to a workflow registration to enable others to discover and reuse it more effectively. E.g. what semantic type of input does it accept.

GGF Summer School 24th July 2004, Italy

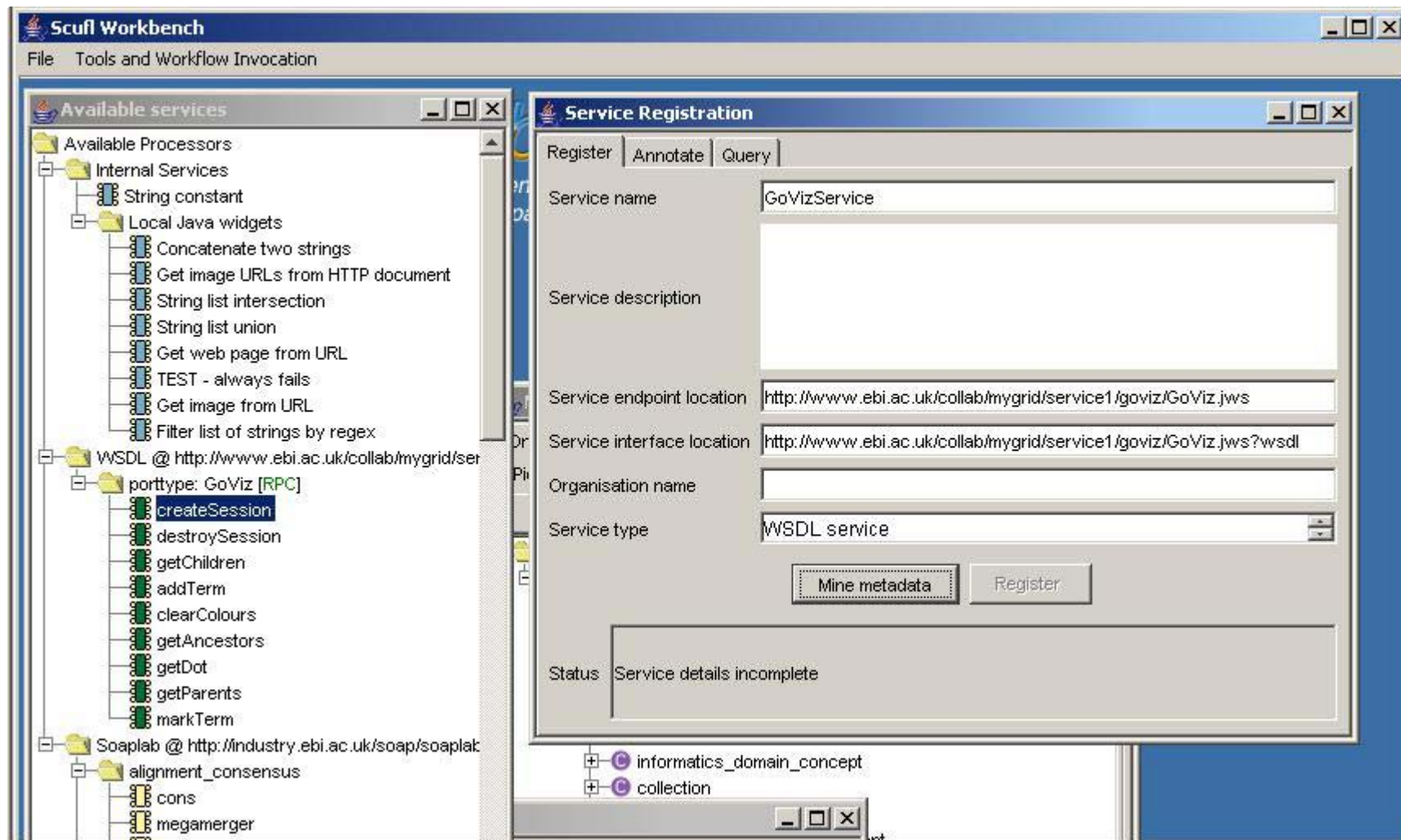


Can you guess what it is yet?

```
<?xml version='1.0' encoding='UTF-8'?>
<definitions name='Blast' >
  <message name='execute0In' >
    <part name='accession' type='xsd:string' />
  </message>
  <message name='execute0Out' >
    <part name='Result' type='xsd:string' />
  </message>
  <portType name='Blast' >
    <operation name='execute' parameterOrder='accession' >
      <documentation>Execute Blast</documentation>
      <input name='execute0In' message='tns:execute0In' />
      <output name='execute0Out' message='tns:execute0Out' />
    </operation>
  </portType>
  <binding name='Blast' type='tns:Blast' >
    <soap:binding style='rpc'
      transport='http://schemas.xmlsoap.org/soap/http' />
    <operation name='execute' >
      <soap:operation soapAction='execute' style='rpc' />
      <input name='execute0In' >
        <soap:body use='encoded' />
      </input>
      <output name='execute0Out' >
        <soap:body use='encoded' />
      </output>
    </operation>
  </binding>
</definitions>
```

Service Registration

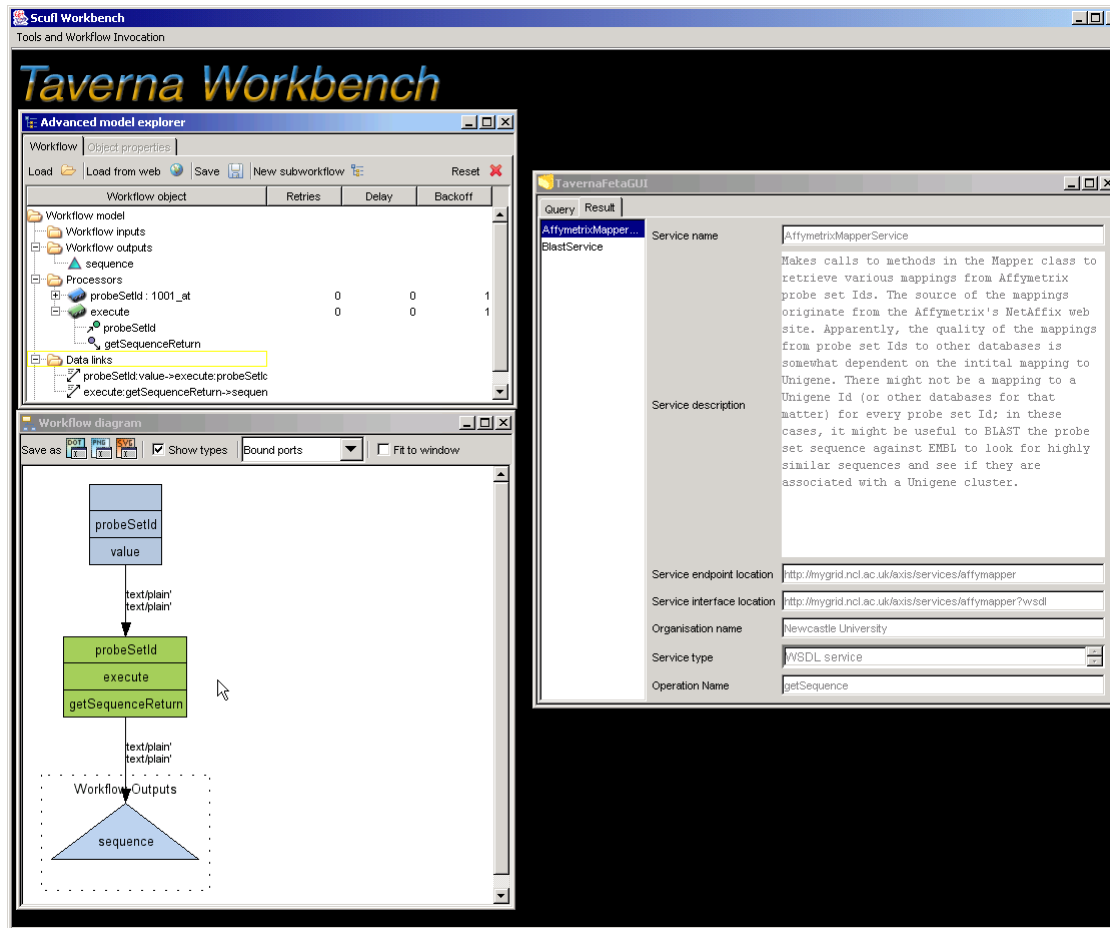
<http://pedro.man.ac.uk>



The screenshot shows the Scuff Workbench application with two main windows:

- Available services:** A tree view on the left showing a hierarchy of services. Under 'WSDL @ http://www.ebi.ac.uk/collab/mygrid/ser', the 'porttype: GoViz [RPC]' is expanded, and the 'createSession' operation is selected.
- Service Registration:** A dialog box on the right with the following fields:
 - Register | Annotate | Query (tabs)
 - Service name: GoVizService
 - Service description: (empty text area)
 - Service endpoint location: <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jws>
 - Service interface location: <http://www.ebi.ac.uk/collab/mygrid/service1/goviz/GoViz.jws?wsdl>
 - Organisation name: (empty text field)
 - Service type: WSDL service (dropdown menu)
 - Buttons: Mine metadata, Register
 - Status: Service details incomplete

Semantic Discovery



The screenshot displays the Taverna Workbench interface. On the left, the 'Advanced model explorer' shows a workflow tree with components like 'Workflow model', 'Workflow inputs', 'Workflow outputs', 'Processors', and 'Data links'. Below it, the 'Workflow diagram' shows a flow from 'probeSetId' to 'execute' and 'getSequenceReturn', which then connects to 'Workflow Outputs' and a 'sequence' node. On the right, the 'TavernaMetaGUI' window shows details for the 'AffymetrixMapperService', including its description, endpoint location, and organization name.

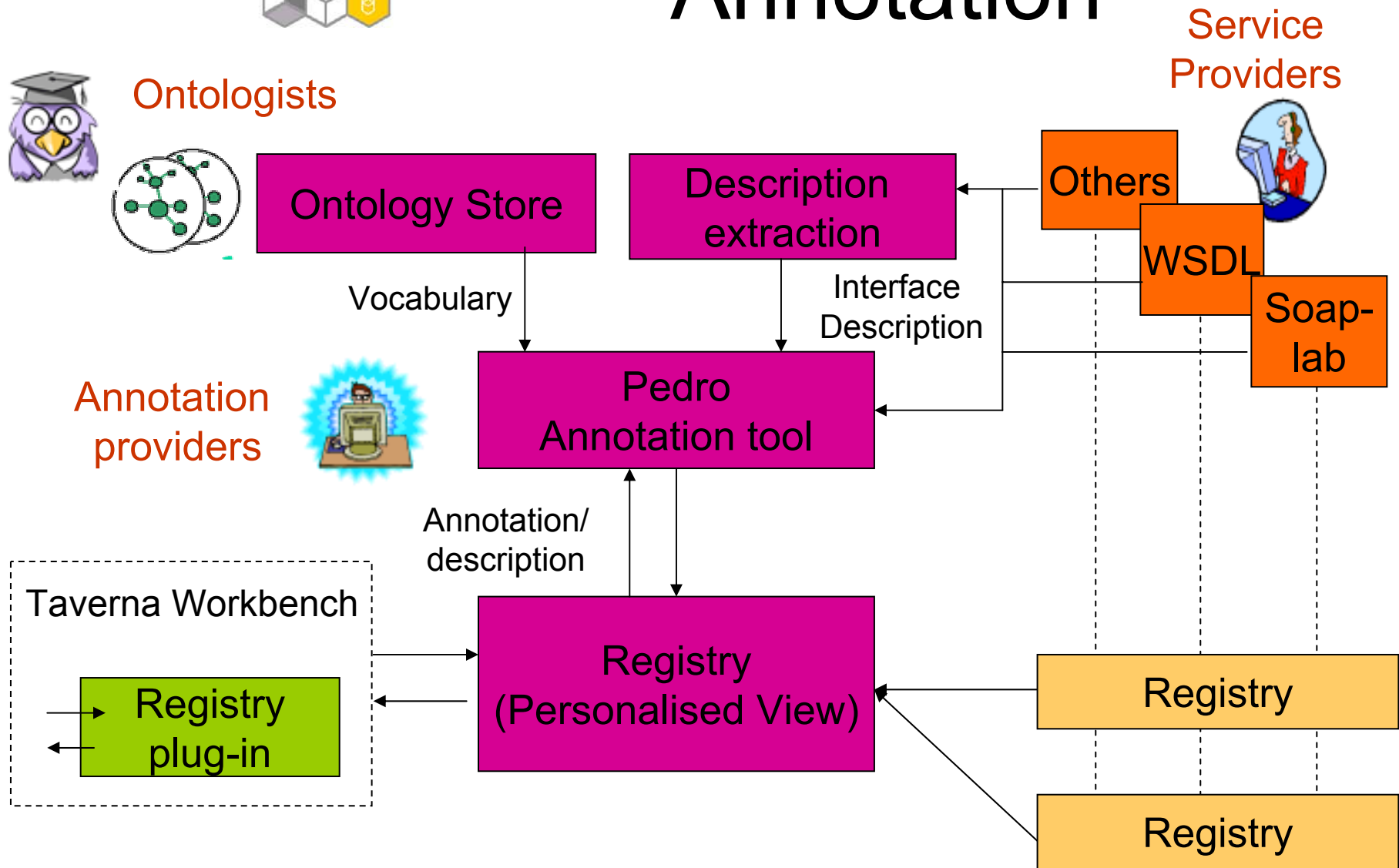
- Drag a workflow entry into the explorer pane and the workflow loads.
- Drag a service/workflow to the scavenger window for inclusion into the workflow



myGrid and Semantics

- Workflow and service discovery
 - Prior to and during enactment
 - Semantic registration
- Workflow assembly
 - Semantic service typing of inputs and outputs
- Provenance of workflows and other entities
- Experimental metadata glue
- Use of RDF, RDFS, DAML+OIL/OWL
 - Instance store, ontology server, reasoner
 - Materialised vs at point of delivery reasoning.
- myGrid Information Model

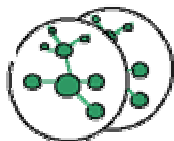
Annotation



Annotation

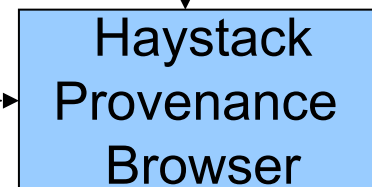
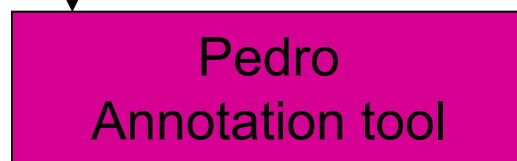


Ontologists



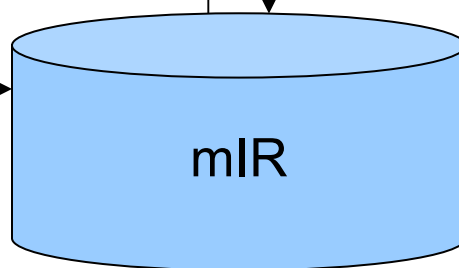
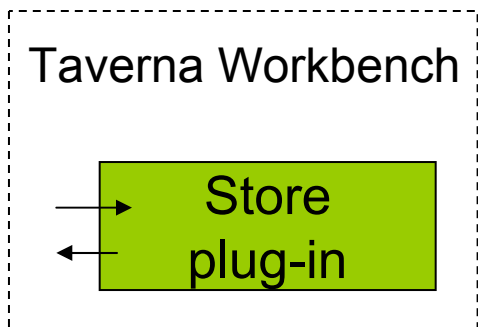
Vocabulary

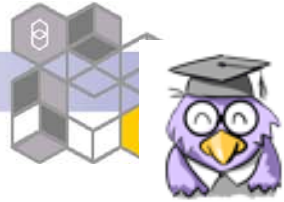
Annotation providers



Scientists

Annotation/
description

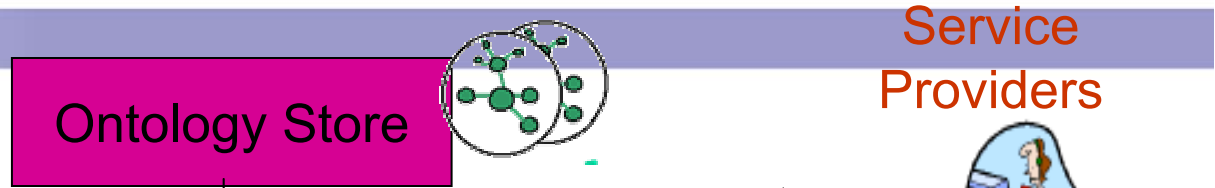
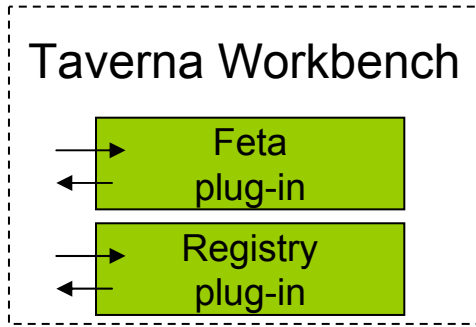




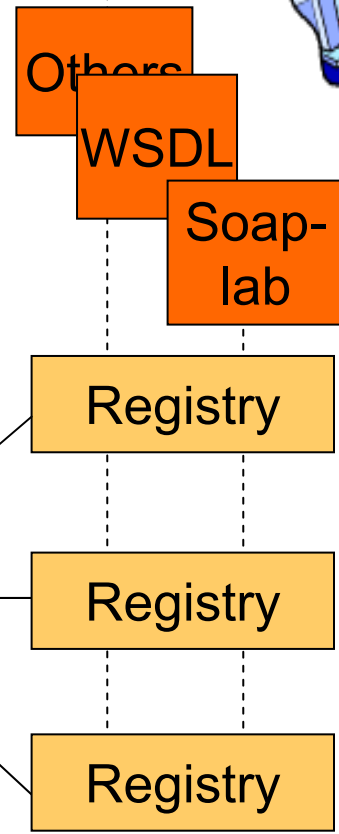
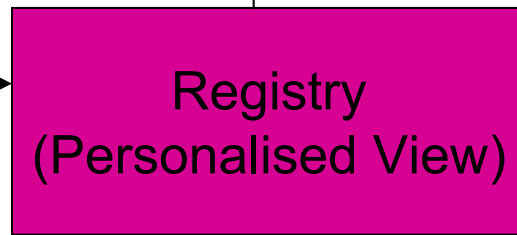
Ontologists

Bioinformaticians

Service Providers



Vocabulary

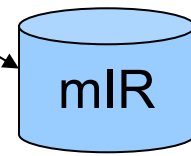


Workflow Execution



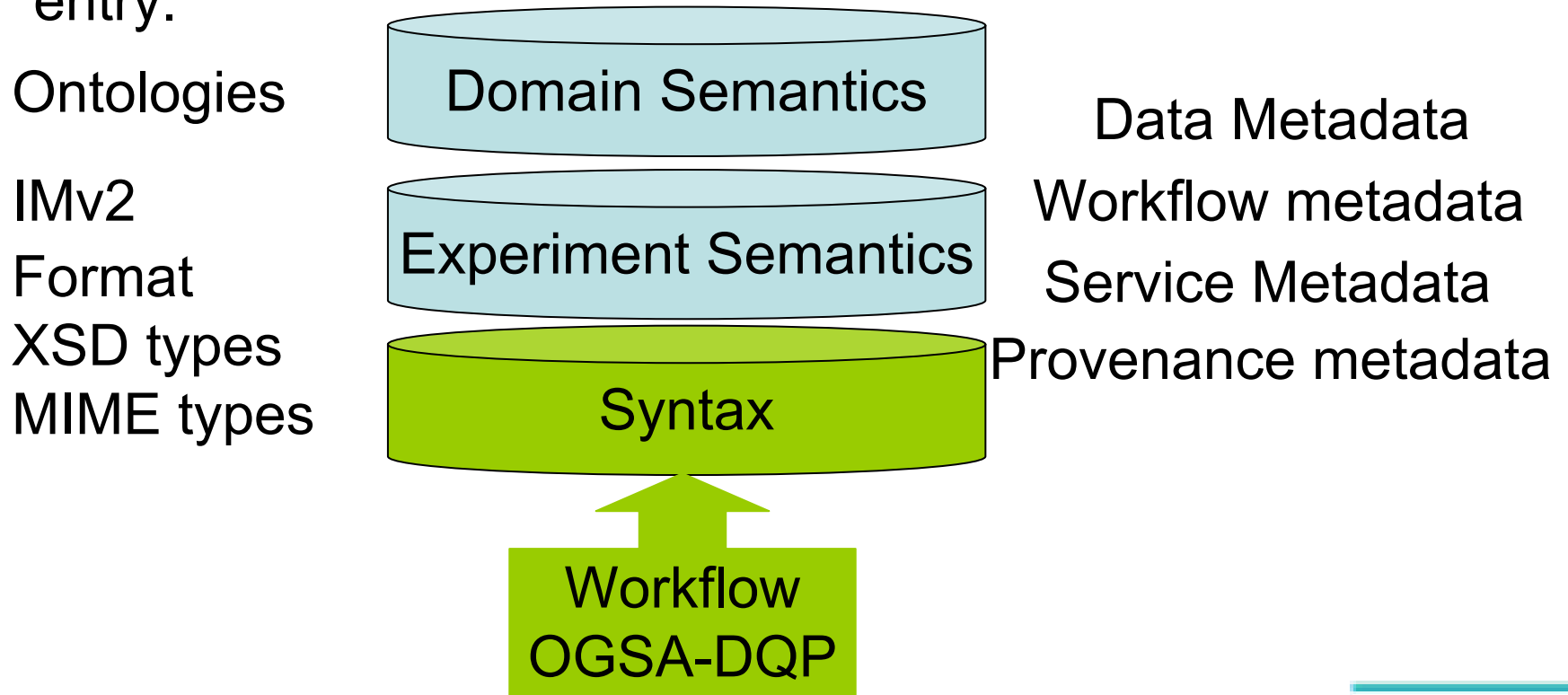
invoking

Store data & metadata

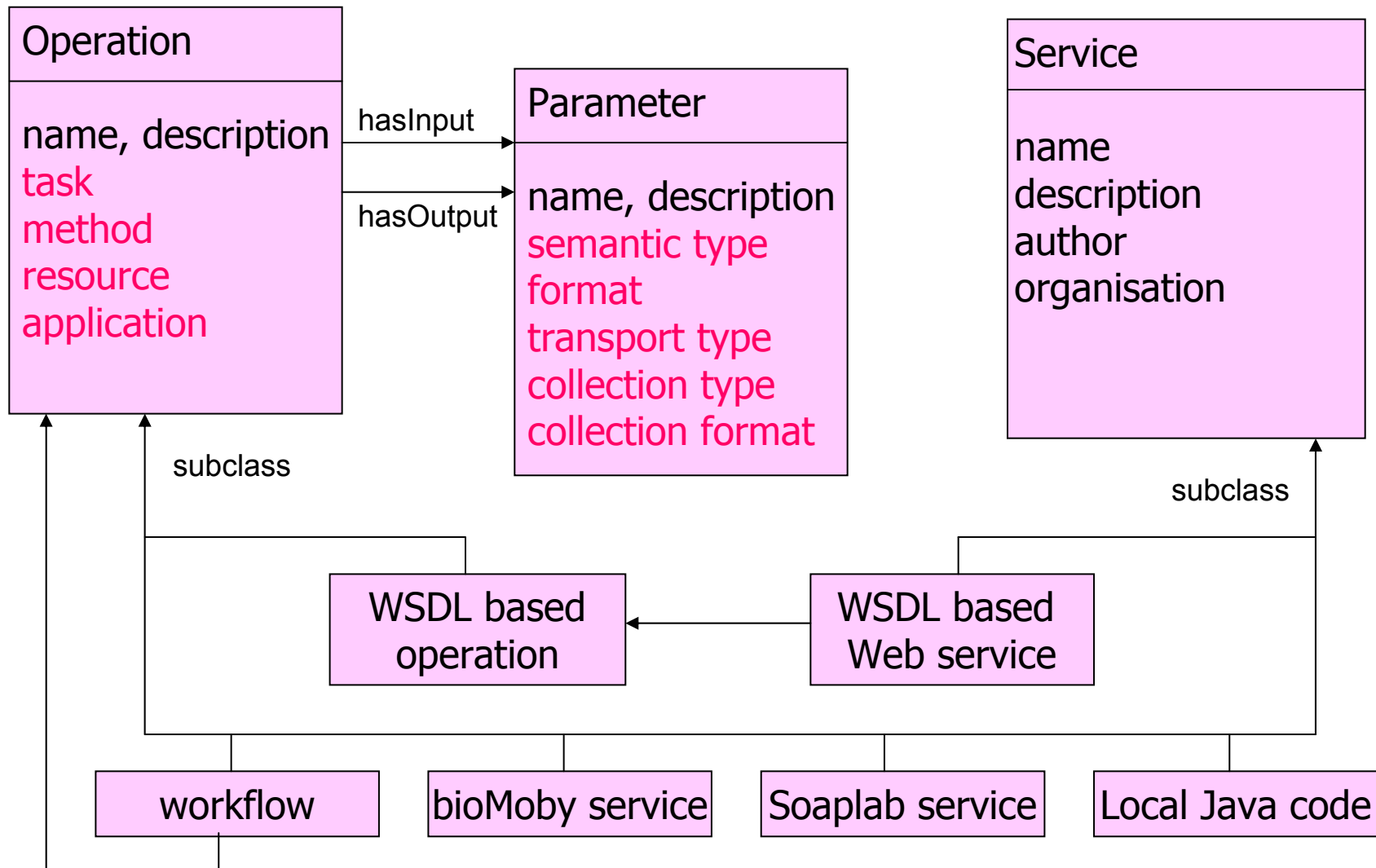


Layered Semantics

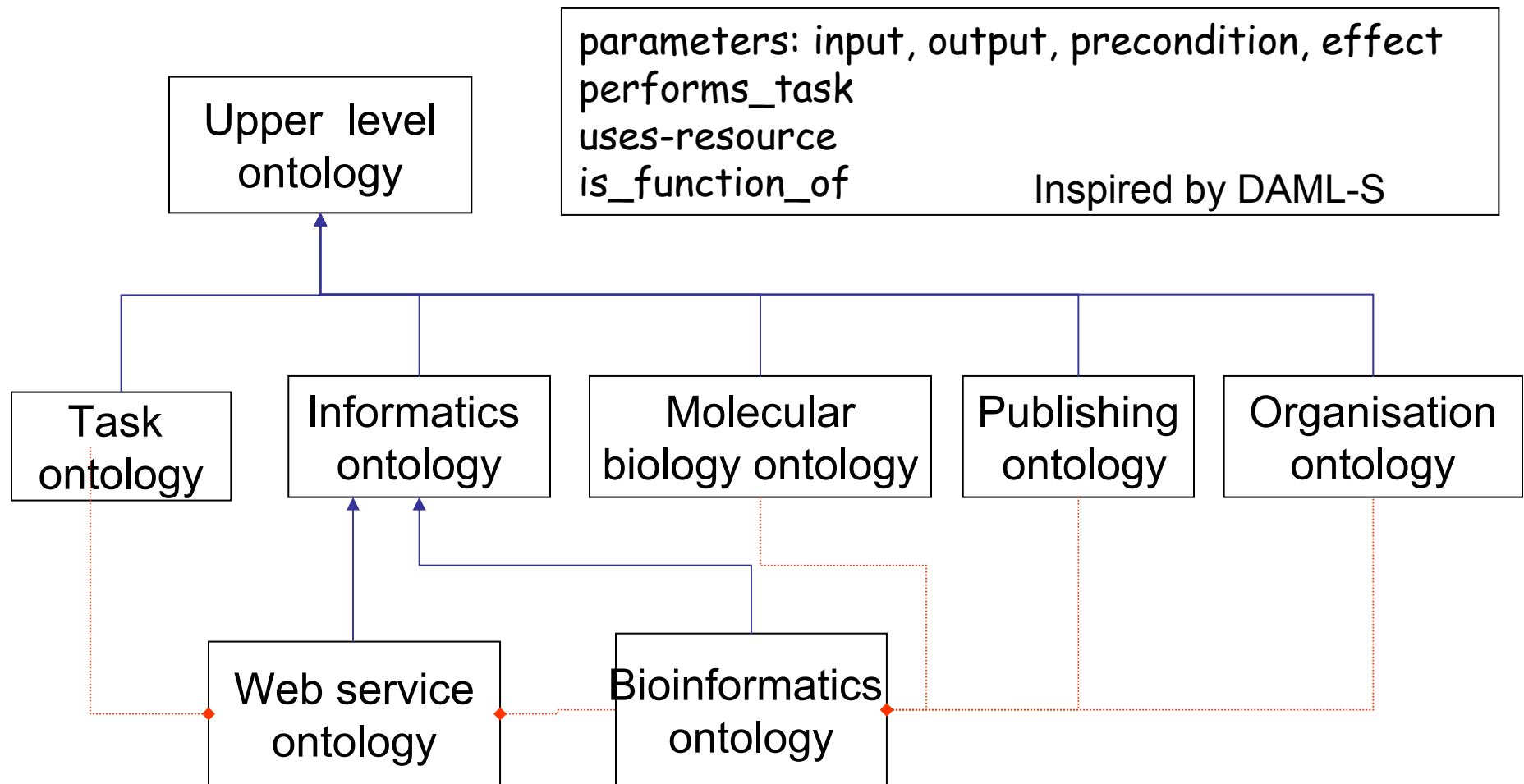
- Domain Semantics layered on top of domain neutral but scientific data model
- Reducing the activation energy, lowering barriers of entry.



Model of services



Service Ontology Suite



Current work: Joint development on an Open Biological Ontologies BioService Ontology. <http://obo.sourceforge.net/>

Workflow metadata

Three stages in lifecycle:

1. Workflow creation

- Service discovery

2. Workflow resolution

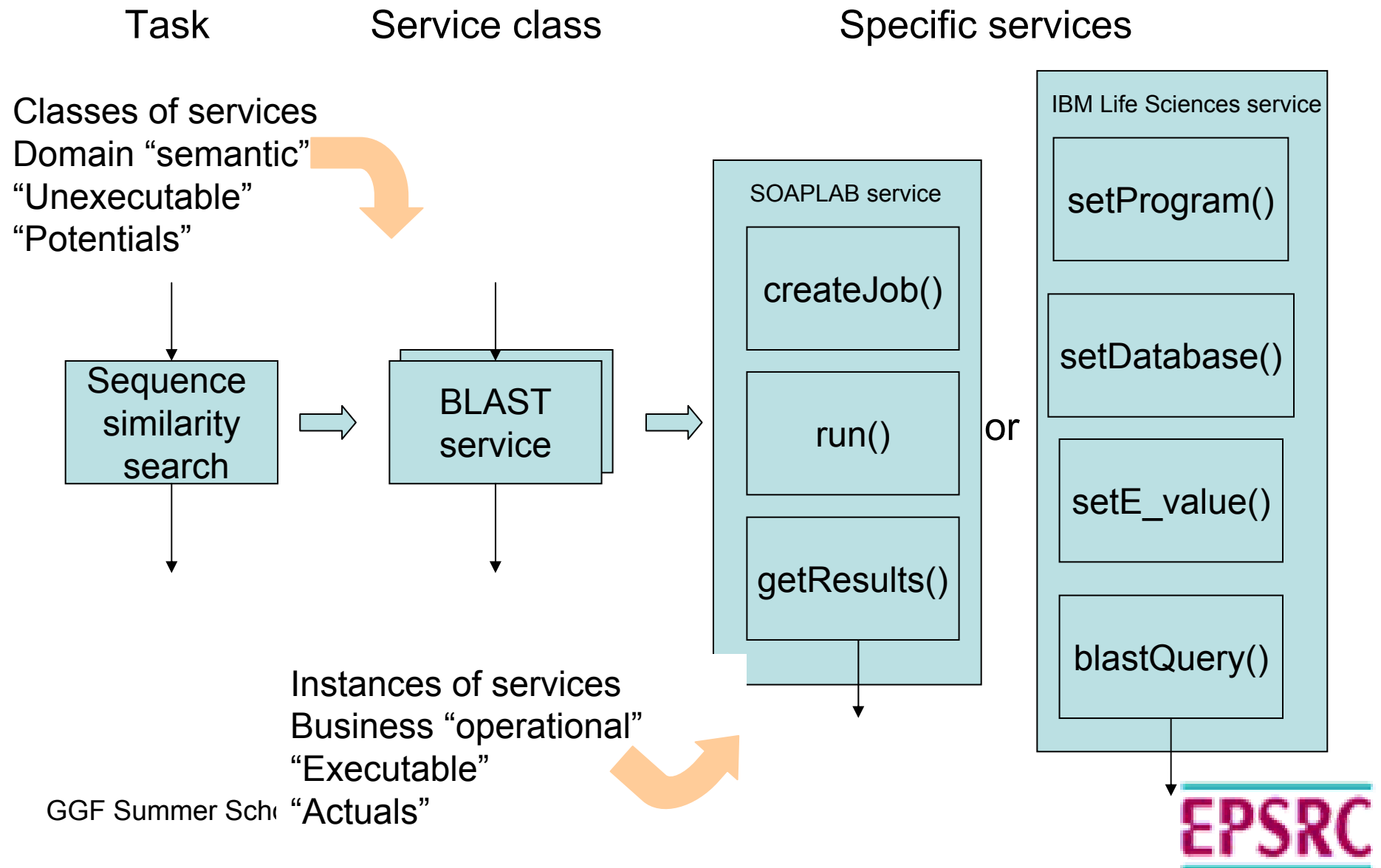
- Service selection

3. Workflow harmonization

- Reconciling parameters
- Format transformations
- Invocation and harmonization

Stage of invocation	DBJ BLAST service	Soaplab BLAST service
Creating a job	n/a	createEmptyJob()
Configuring the service	simpleSearch (program, database, query)	set_database(database, job)
Setting input data		set_query_sequence(qu ery, job)
Running the job		run(job)
Getting output data		getSomeResults(job)

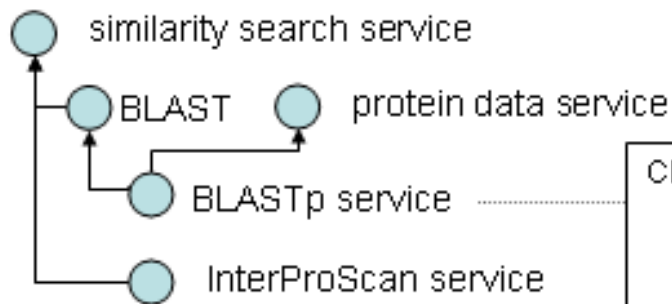
Tiered specifications



Stratified metadata

- Service Type and Class (OWL)

Extract of service classification



Example OWL description on which classification is based

```

Class(BLASTpService complete WebService
restriction(input someValuesFrom(Protein))
restriction(usesResource someValuesFrom(protein sequence database))
restriction(isFunctionOf someValuesFrom(BLAST)))
  
```

- Service Instance (RDF)

```

<profile:qualityRating>
  <profile:QualityRating rdf:ID="NCBI-BLASTn-Rating">
    <profile:ratingName>Recommendation</profile:ratingName>
    <profile:rating rdf:resource="http://www.mygrid.org.uk/quality_concepts.daml#recommended"/>
  </profile:QualityRating>
</profile:qualityRating>
  
```



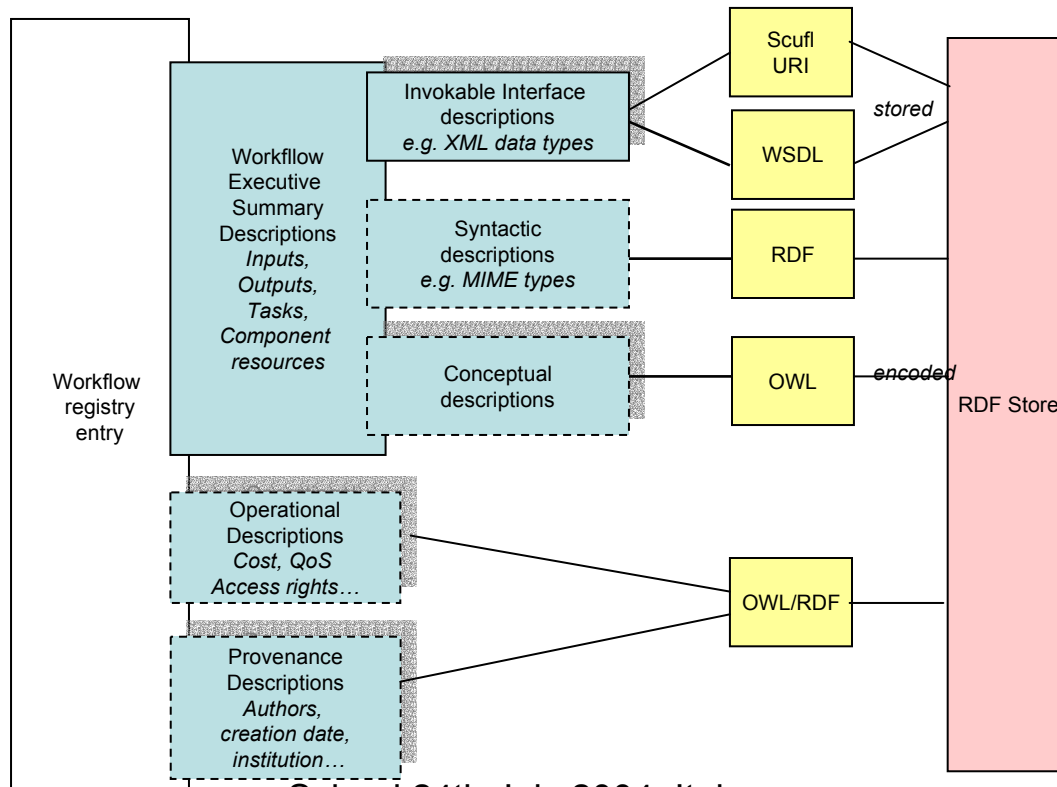
Seven types of service metadata

Conceptual
Configuration
Provenance
Operational
Invocation model
Interface
Data format

Service and Workflow registration

Workflow registration allows peer review and publication of e-Science methods.





- Description scheme
- RDFS & DAML+OIL / OWL ontologies of services & biology
- Based on DAML-S
- Reasoning over OWL descriptions
- Query over RDF
- Aim to have semantic discovery over public view on the web.



Reflections

- Multiple descriptions, multiple interfaces

- Users needs
- Machine needs

Service User	Human 	Machine 
Service provider 	UDDI style advertisements	Weak semantic descriptions Rewriting views
Human 	Elaborate Semantic descriptions Simplification views	Syntactic descriptions Interface descriptions Invocation descriptions Semantic mining

- The dimensions of Service Class substitution

- Biologists choose experimentally meaningful services and do not want “semantically similar” substitutions; only substituting one instance for another
- Experimentally neutral “glue” services that can be substituted are comparatively few

Reuse and Repurposing

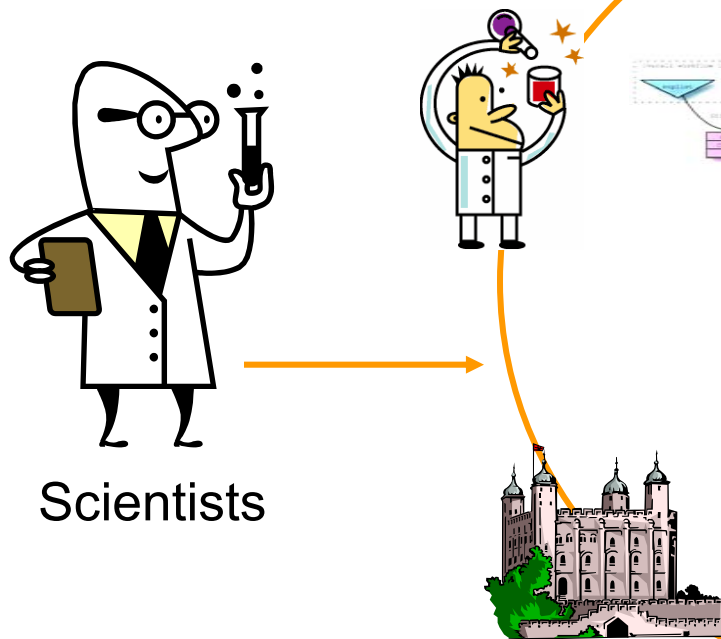
- Describing for reuse is challenging
 - Reuse depends on semantic descriptions and these are costly to produce
 - Describing for someone else's benefit
 - Reuse by multiple stakeholders
- Licensing workflows for reuse.
- Authorisation models
- But reuse does happen!
- Metadata pays off but it needs a network effect and there is a cost.

So far, Using Concepts

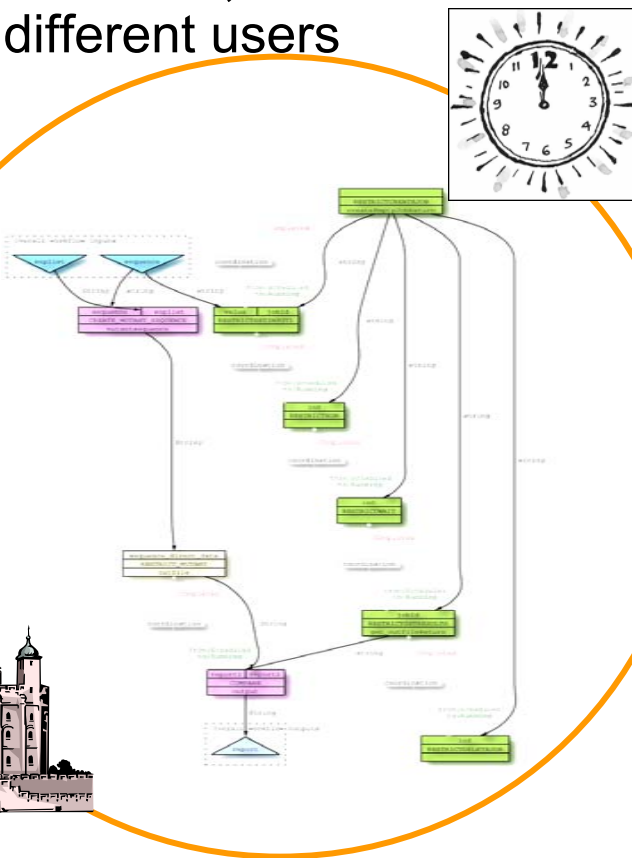
- Controlled vocabulary for advertisements for workflows and services
- Indexes into registries and mIR
 - Semantic discovery of services and workflows
 - Semantic discovery of repository entries
- Type management for composition
 - Semantic workflow construction: guidance and validation
- Navigation paths between data and knowledge holdings
 - Semantic “glue” between repository entries
 - Semantic annotation and linking of workflow

Provenance

Experiments being performed repeatedly, at different site, different time, by different users or groups;



Scientists



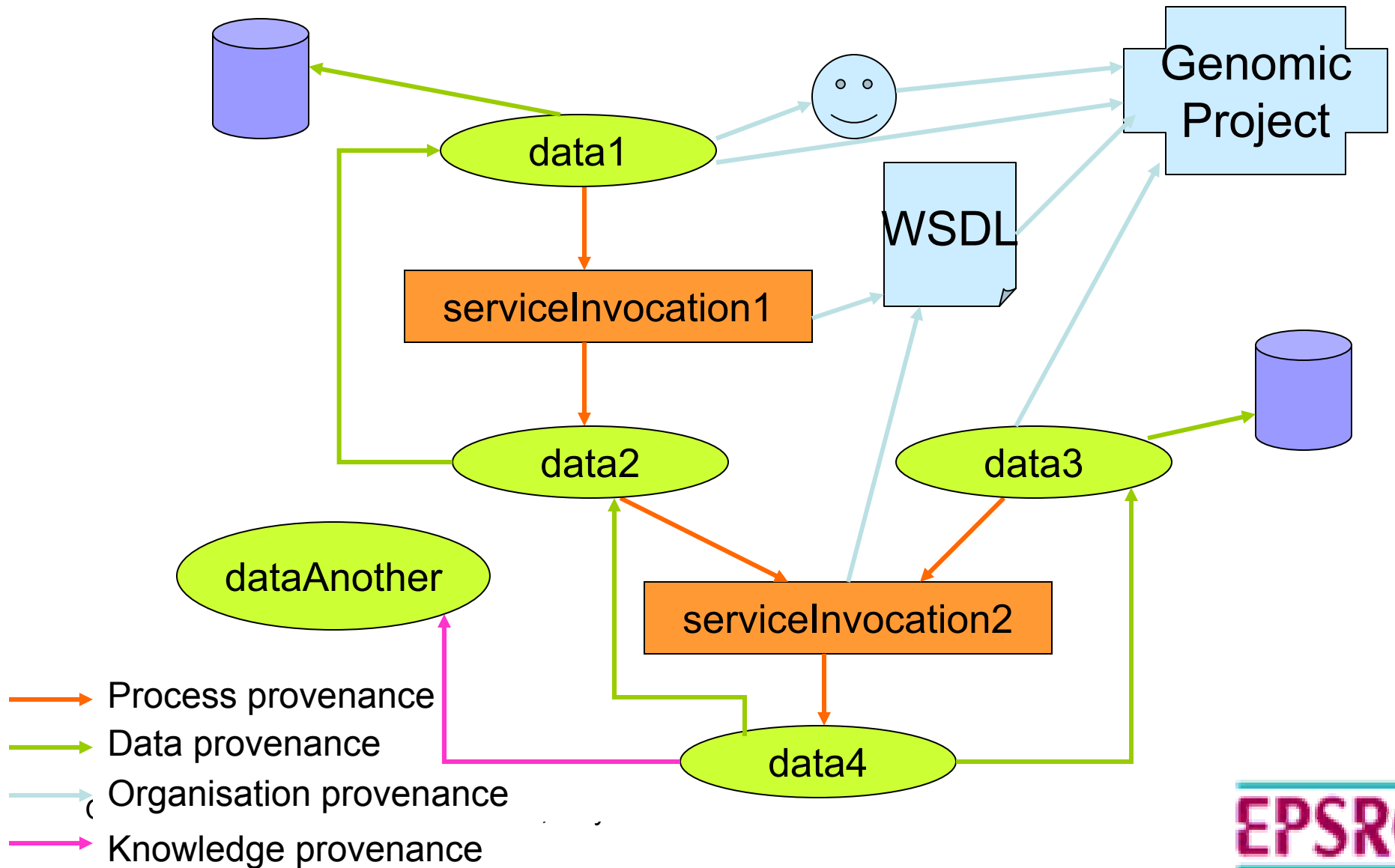
A large repository of records about experiments!!

- verification of data;
- “recipes” for experiment designs;
- explanation for the impact of changes;
- ownership;
- performance of services;
- data quality;

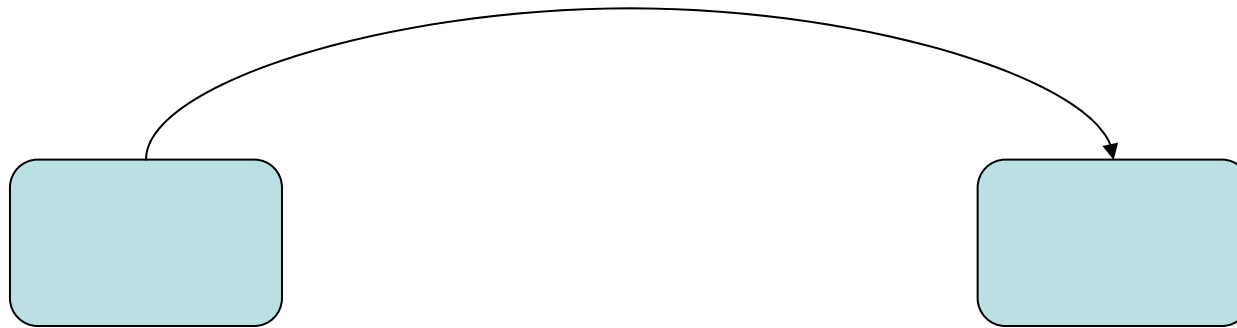




Provenance Web



Representing links

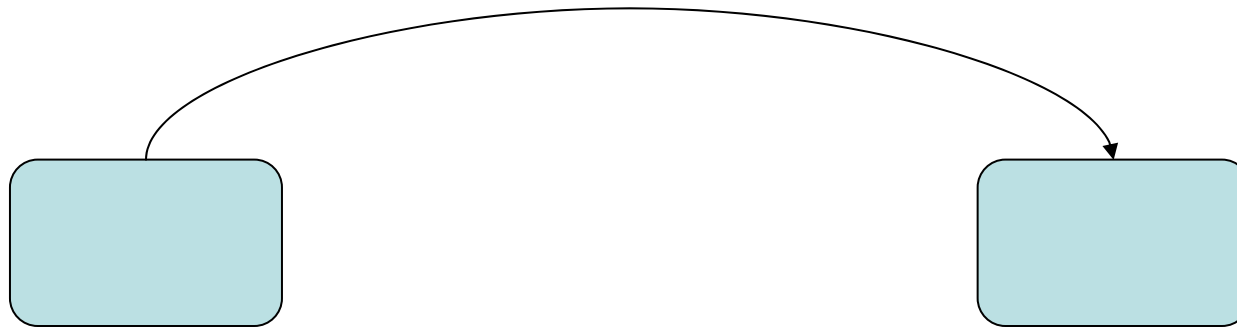


<urn:lsid:taverna.sf.net:datathing:45fg6> <urn:lsid:taverna.sf.net:datathing:23ty3>

- Identify each resource
 - Life science identifier: URI with associated data and metadata retrieval protocols.
 - Understanding that underlying data will not change

Representing links II

http://www.mygrid.org.uk/ontology#derived_from



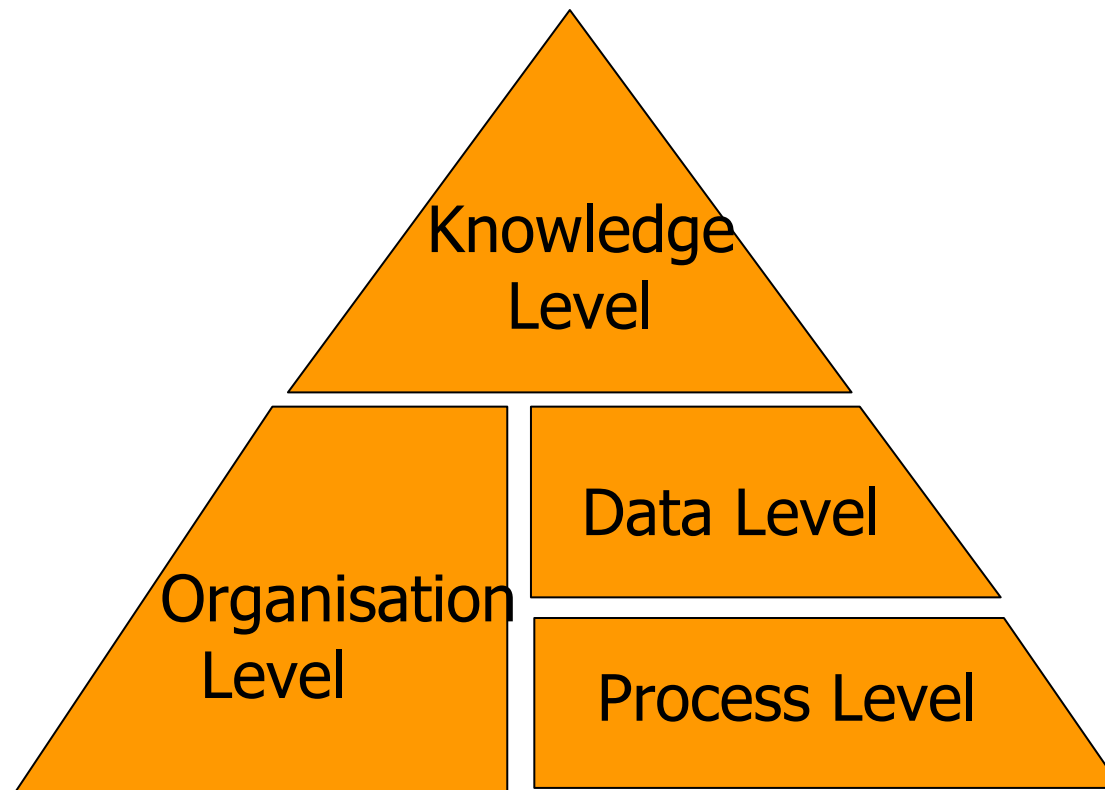
<urn:lsid:taverna.sf.net:datathing:45fg6>

<urn:lsid:taverna.sf.net:datathing:23ty3>

- Identify link type
 - Again use URI
 - Allows us to use RDF infrastructure
 - Repositories
 - Ontologies



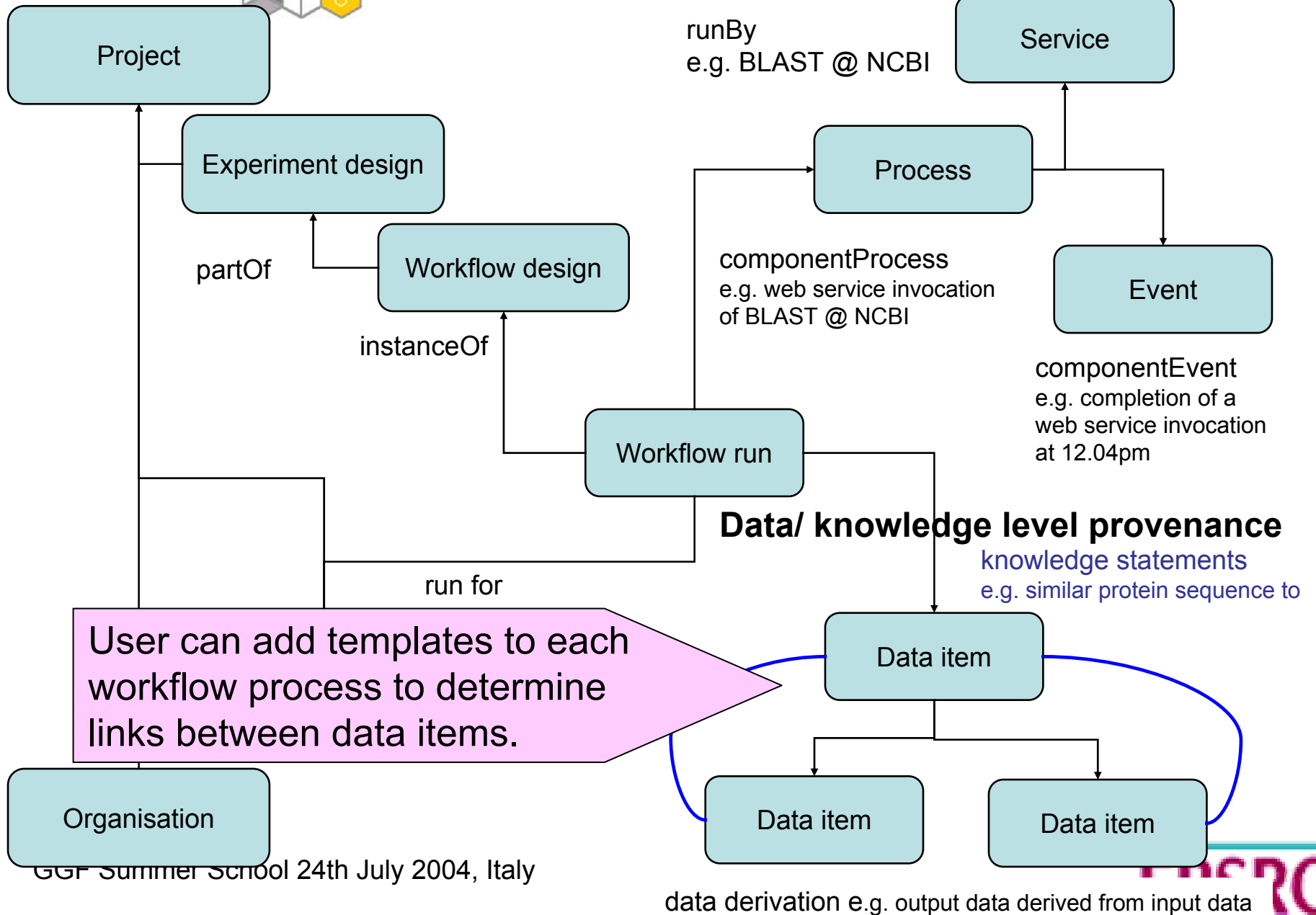
Provenance Pyramid



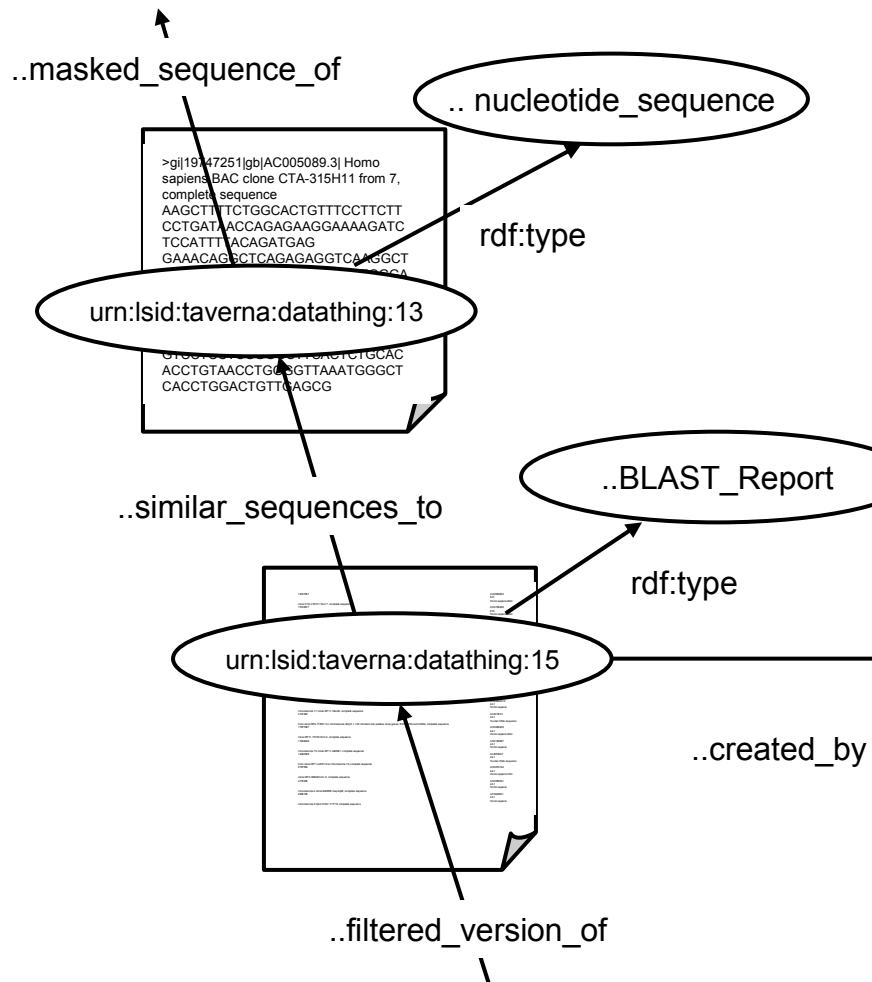


Organisation level provenance

Process level provenance



Provenance tracking



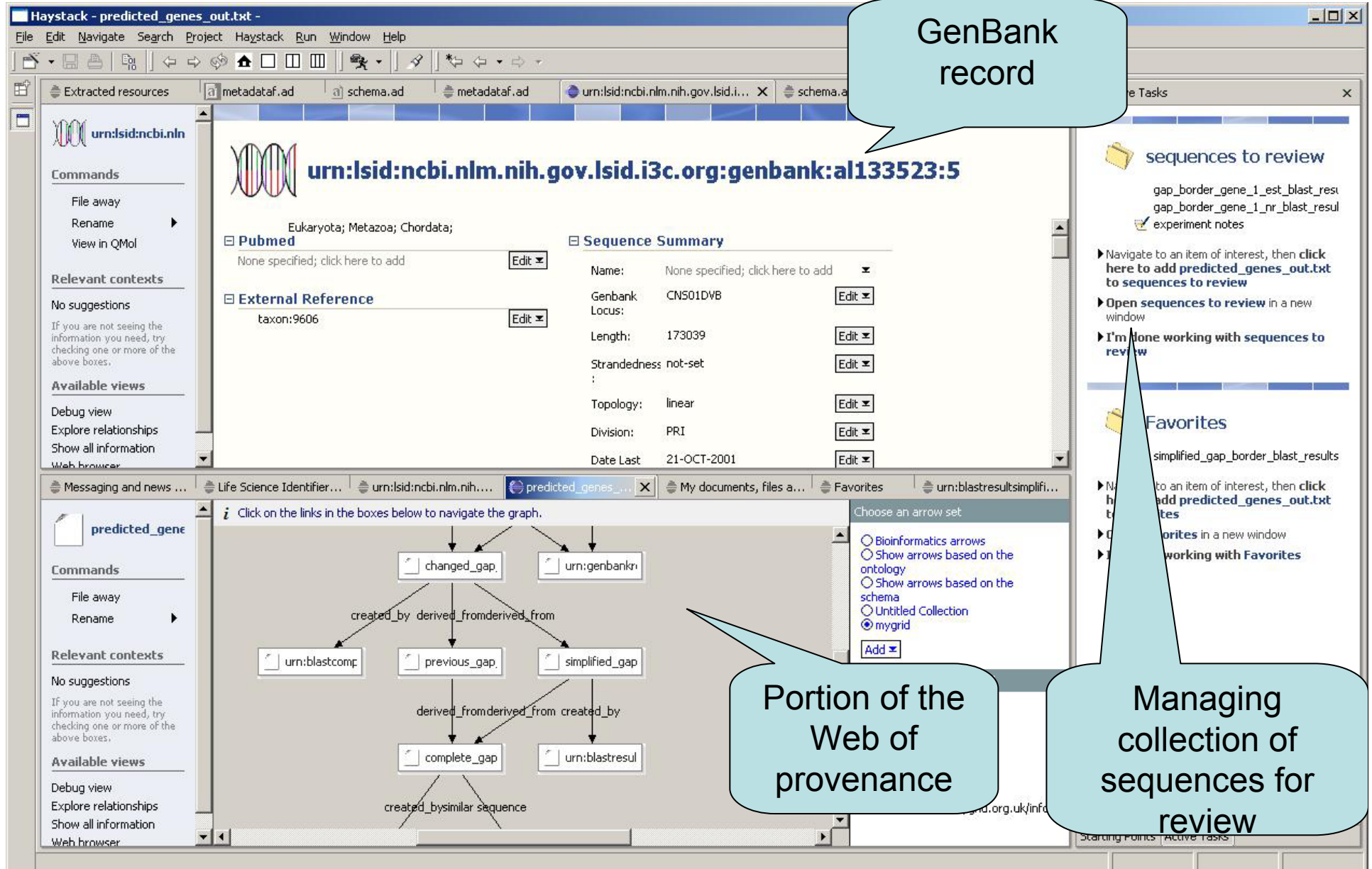
Relationship BLAST report has with other items in the repository

GGF Summer School 24th July 2004, Italy

- Automated generation of this web of links
- Workflow enactor generates
 - LSIDs
 - Data derivation links
 - Knowledge links
 - Process links
 - Organisation links

Other classes of information related to BLAST report

Haystack (IBM/MIT)




The screenshot displays the Haystack web application interface. The top window shows a GenBank record for `urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:al133523:5`. The record details include taxonomic classification (Eukaryota; Metazoa; Chordata), a Pubmed reference (taxon:9606), and a Sequence Summary with fields such as Name, Genbank (CNS01DVB), Locus, Length (173039), Strandedness (not-set), Topology (linear), Division (PRI), and Date Last (21-OCT-2001).

The bottom window displays a provenance graph with nodes representing different sequence states and relationships. The nodes include `changed_gap`, `urn:genbank:`, `urn:blastcomp`, `previous_gap`, `simplified_gap`, `complete_gap`, and `urn:blastresul`. Relationships are labeled with terms like `created_by`, `derived_from`, and `created_by similar sequence`. A legend on the right allows users to choose an arrow set for the graph, with options like `Bioinformatics arrows`, `Show arrows based on the ontology`, and `mygrid`.

Three callout boxes provide additional context:

- GenBank record**: Points to the top window displaying the sequence details.
- Portion of the Web of provenance**: Points to the provenance graph in the bottom window.
- Managing collection of sequences for review**: Points to the right-hand sidebar, which contains sections for `sequences to review` (listing `gap_border_gene_1_est_blast_res` and `gap_border_gene_1_nr_blast_res`) and `Favorites` (listing `simplified_gap_border_blast_results`).

Schema.ad Schema.ad urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12



urn:lsid:ncbi.nlm

Commands

- File away
- Rename
- View in QMol

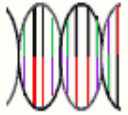
Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Available views

- Browse view
- Calendar



urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:ac009070:12

Pubmed

None specified; click here to add Edit ▾

urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:pred...

taxon:9606 Edit ▾

Sequence Summary

Name: None specified; click here to add ▾

urn:lsid:ncbi.r AC009070 Edit ▾

Extracted resources

Extracted resources

Commands

- Add existing items
- Add new item
- Clear collection/list
- Create checkbox aspect
- File away
- Rename

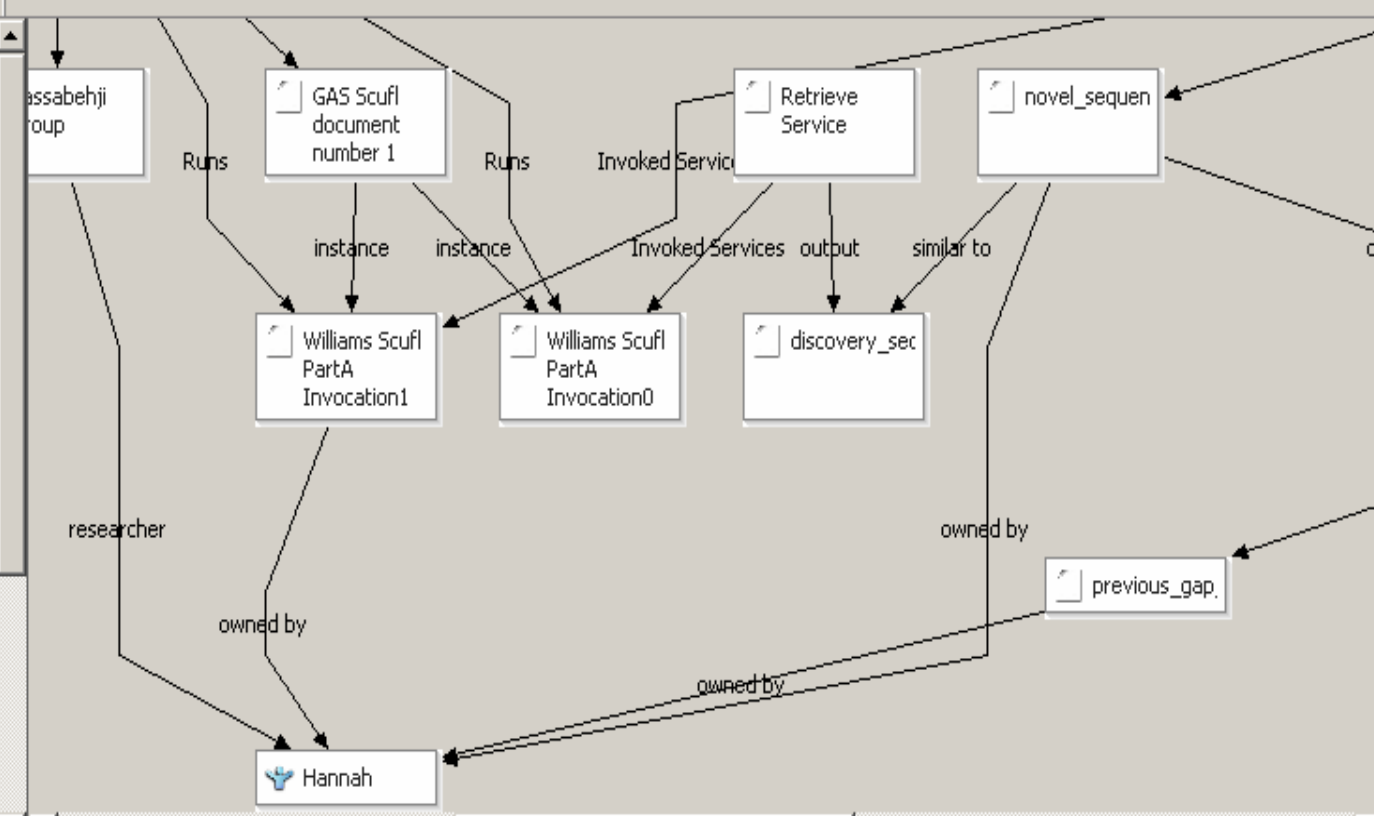
Relevant contexts

No suggestions

If you are not seeing the information you need, try checking one or more of the above boxes.

Available views

- Browse view
- Calendar



```

graph TD
    subgraph "assabehji group"
        A[assabehji group]
    end
    subgraph "GAS ScufI"
        B[GAS ScufI document number 1]
    end
    subgraph "Retrieve Service"
        C[Retrieve Service]
    end
    subgraph "novel_sequen"
        D[novel_sequen]
    end
    subgraph "Williams ScufI PartA"
        E[Williams ScufI PartA Invocation1]
        F[Williams ScufI PartA Invocation0]
    end
    subgraph "discovery_sec"
        G[discovery_sec]
    end
    subgraph "previous_gap"
        H[previous_gap]
    end
    subgraph "Hannah"
        I[Hannah]
    end

    A -- Runs --> B
    A -- Runs --> C
    B -- instance --> E
    C -- Invoked Service --> E
    C -- Invoked Services --> F
    C -- output --> G
    D -- similar to --> G
    E -- owned by --> I
    F -- owned by --> I
    G -- owned by --> I
    H -- owned by --> I
    
```

Choose an arr...

- Bioinformatic arrows
- Provenance Graph
- Show arrows based on the ontology
- Show arrows based on the schema

Add ▾

Available arrows for Provenance Graph

- ↳ Designs
- ↳ Invoked Services
- 11 more it...

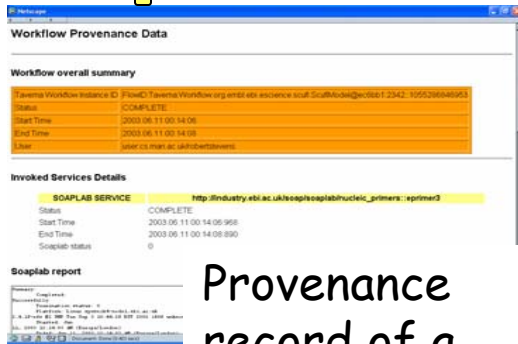
Reflections

- Visualisation of results usually domain specific
- Provenance browsing and querying needs to fit with that visualisation
- Generic graphical presentation limited to small, low complexity result sets
- Layered provenance for different purposes and different stakeholders
 - Detailed process for debugging and usage statistics for QoS
 - Data and Knowledge for the Scientist
- Migration with data objects

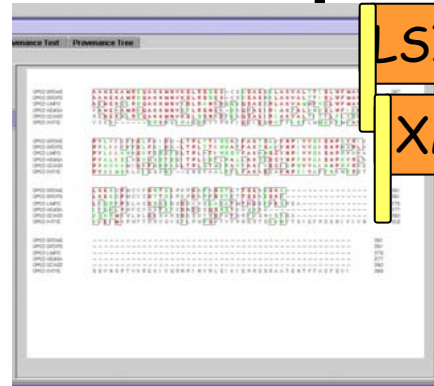


Map of Context

RDF



Provenance record of a workflow run



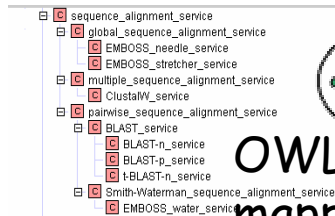
LSID

XML

PDF



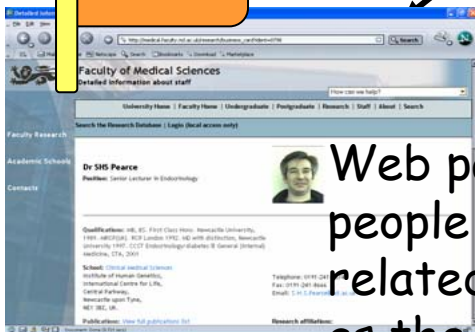
Literature relevant to provenance study or data in this workflow



OWL Ontologies mapping between objects



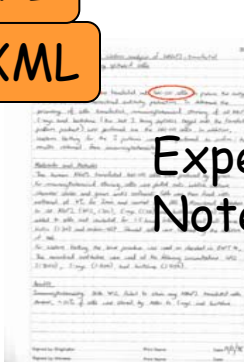
HTML



Web page of people who has related interests as the owner of the workflow

URI

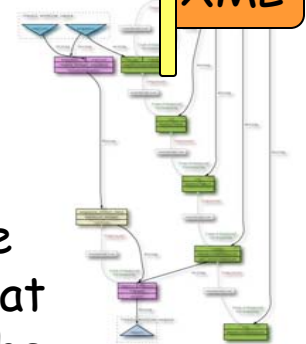
XML



Experiment Notes

LSID

XML



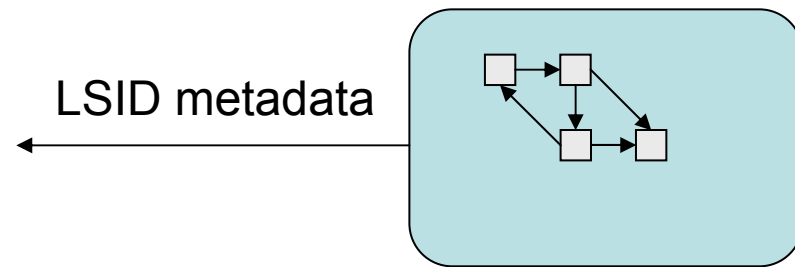
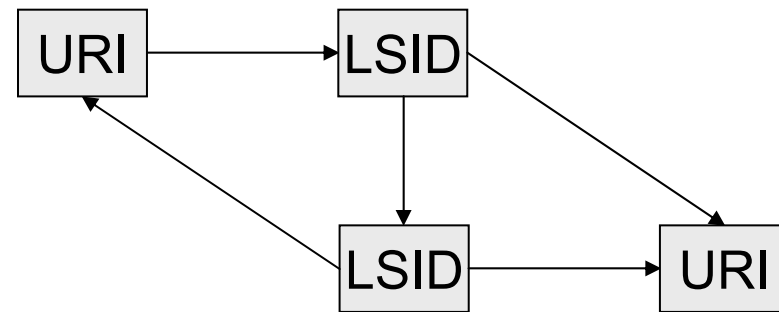
Interlinking graph of the workflow that generates the provenance logs

GGF Summer School 24th July 2004, Italy

Provenance metadata

- Outside objects
 - RDF store

- Within objects
 - LSID metadata.





Linked Provenance Resources

Port AffymetrixMapper
 Operation getAccessionNumber
 Status COMPLETE
 Start Time 2003.06.27 09:12:32.430
 End Time 2003.06.27 09:12:34.373

INPUT DATA SET		
Name	Type	Value
probeSetId	string	probeSetId

OUTPUT DATA SET		
Name	Type	Value
getAccessionNumberReturn	string	getAccessionNumberReturn
getAccessionNumberReturn	string	getAccessionNumberReturn

INPUT DATA

Name	Type	Value
probeSetId	string	probeSetId

OUTPUT DATA

Name	Type	Value
queryByIdReturn	string	queryByIdReturn

The subsumed concepts

Link to the log annotated with more general concept

Link to the log annotated with more specific concept

The subsuming concepts

INPUT DATA

Name	Type	Value
embID	string	embID

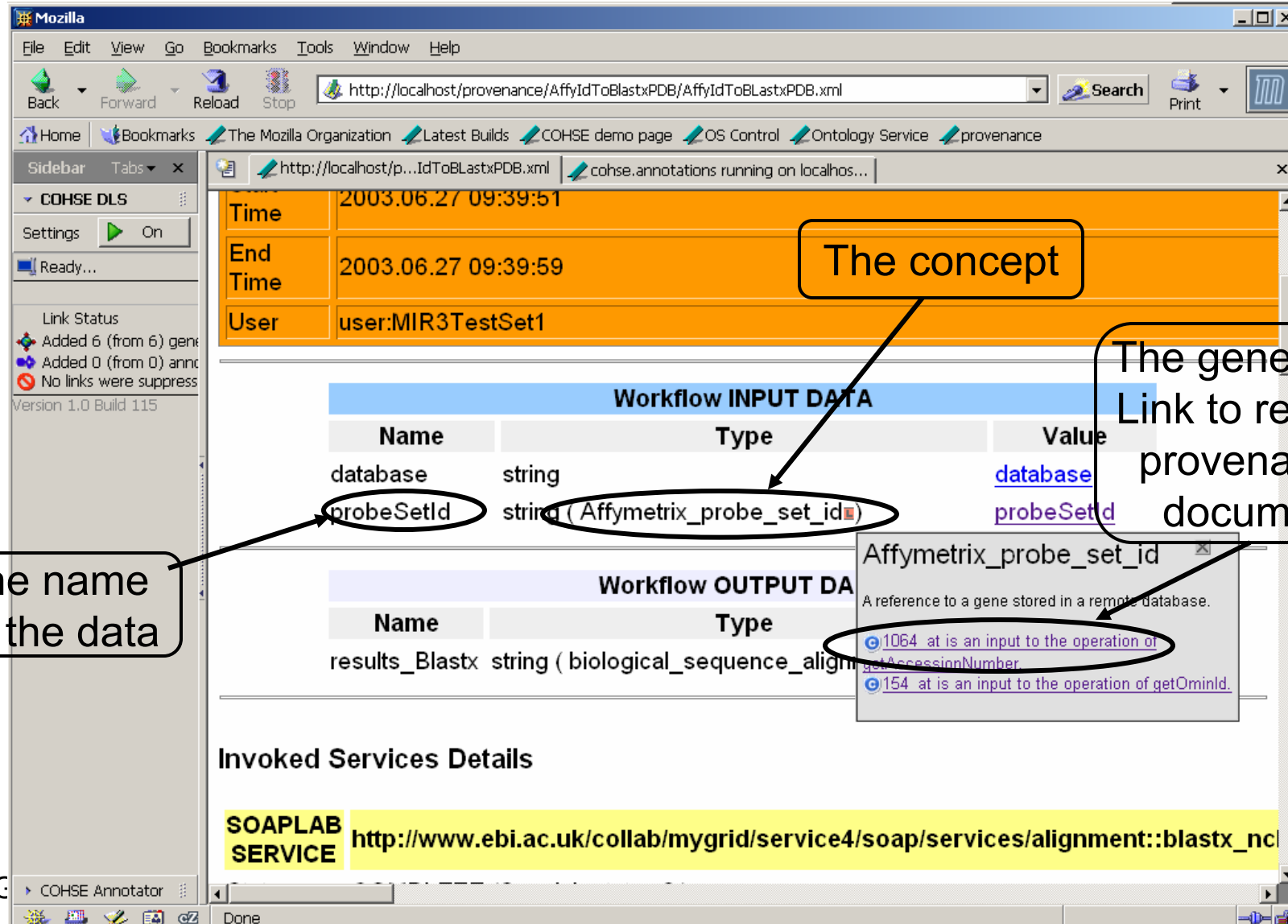
OUTPUT DATA

Name	Type	Value
embID	string	embID

Invoked Services

Status	URL	Service Name	Operation
COMPLETE	http://mygrid.ncl.ac.uk/axis/services/SrsEbiQuery?wsdl	SrsEbiQuery	queryByArrayIds
COMPLETE	http://mygrid.ncl.ac.uk/axis/services/SrsEbiQuery?wsdl	SrsEbiQuery	queryHgbaseByEn
COMPLETE	http://mygrid.ncl.ac.uk/axis/services/EmbSnpEdit?wsdl	SnpFeatureEmbEditor	mergeSnpData
COMPLETE	http://mygrid.ncl.ac.uk/axis/services/SrsEbiQuery?wsdl	SrsEbiQuery	queryById

Generating Links



The screenshot shows a Mozilla browser window displaying a web application. The address bar shows the URL: `http://localhost/provenance/AffyIdToBlastxPDB/AffyIdToBlastxPDB.xml`. The main content area is divided into several sections:

- Metadata:** A table with the following rows:

Time	2003.06.27 09:39:51
End Time	2003.06.27 09:39:59
User	user:MIR3TestSet1
- Workflow INPUT DATA:** A table with columns Name, Type, and Value.

Name	Type	Value
database	string	database
probeSetId	string (Affymetrix_probe_set_id)	probeSetId
- Workflow OUTPUT DATA:** A table with columns Name and Type.

Name	Type
results_Blastx	string (biological_sequence_alignm
- Invoked Services Details:** A section with a highlighted entry:

SOAPLAB SERVICE http://www.ebi.ac.uk/collab/mygrid/service4/soap/services/alignment::blastx_nc

Annotations and callouts are present:

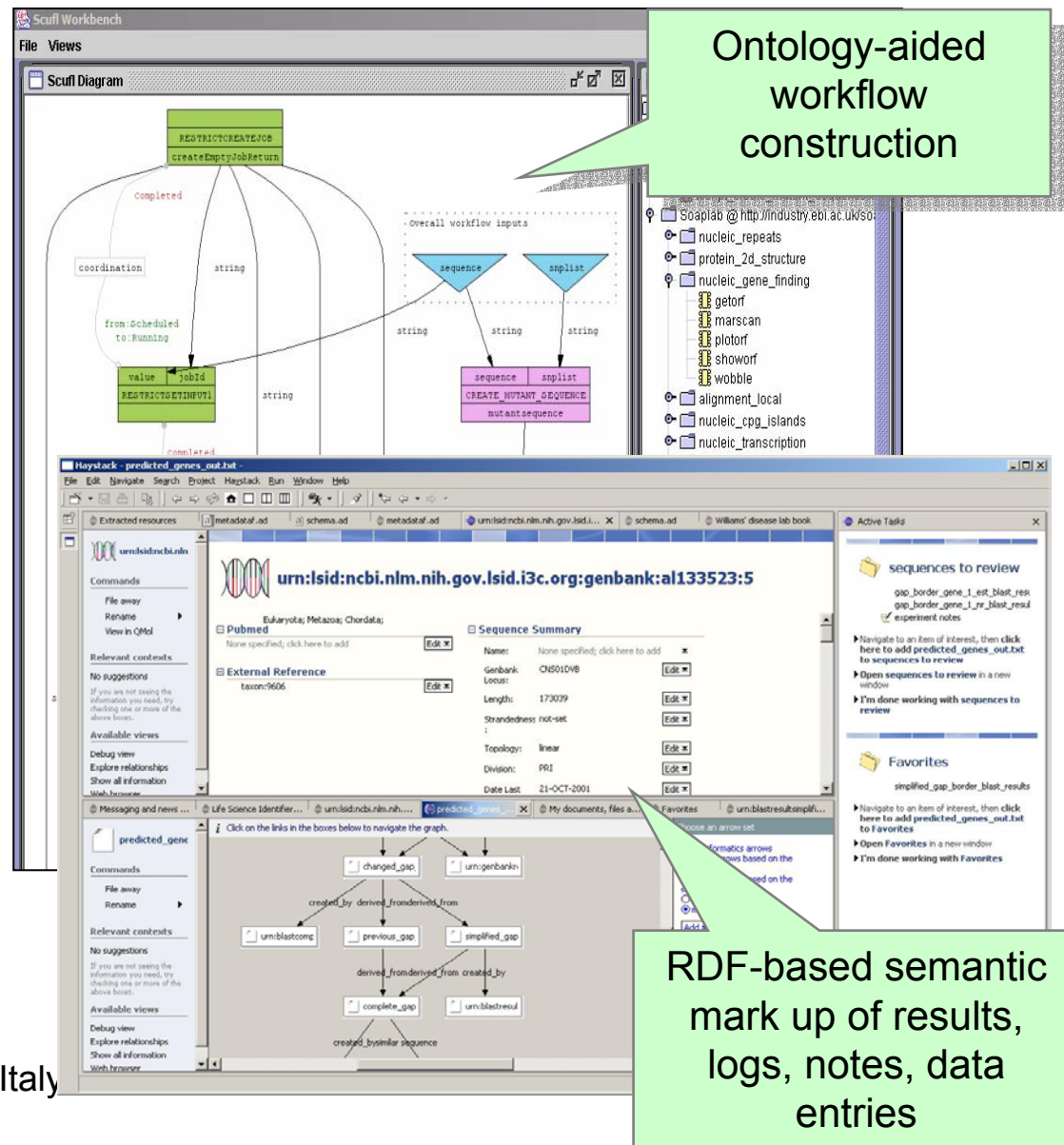
- A callout box labeled "The concept" points to the `probeSetId` entry in the Workflow INPUT DATA table.
- A callout box labeled "The generated Link to related provenance document" points to the `probeSetId` value in the Workflow INPUT DATA table.
- A callout box labeled "The name of the data" points to the `probeSetId` column header in the Workflow INPUT DATA table.
- A tooltip for the `probeSetId` value shows:

Affymetrix_probe_set_id
A reference to a gene stored in a remote database.
[1064](#) at is an input to the operation of [getAccessionNumber](#).
[154](#) at is an input to the operation of [getOminId](#).

Semantics

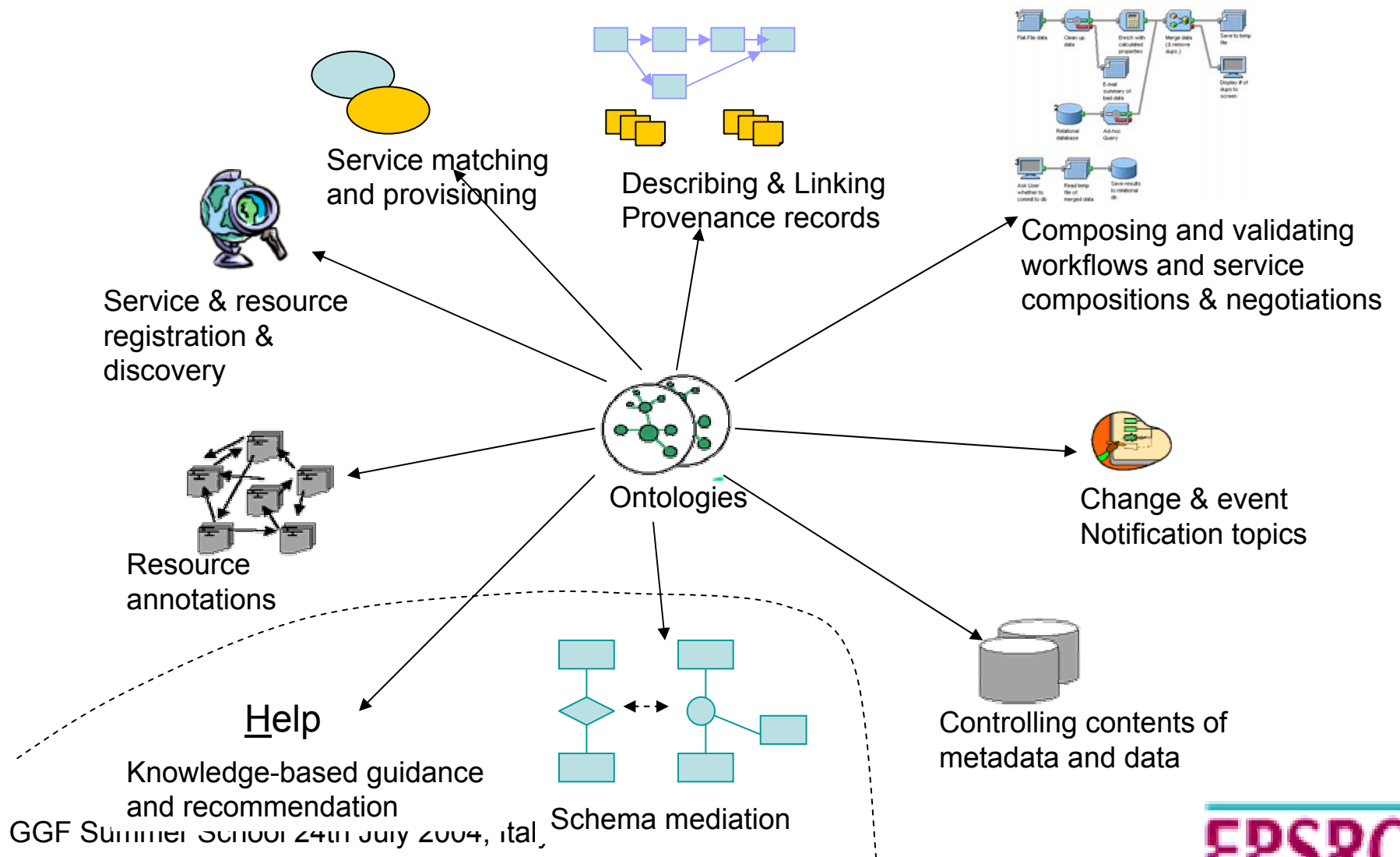
- RDF-based service and data registries
- RDF-based metadata for ALL experimental components
- RDF-based provenance graphs
- OWL based controlled vocabularies for database content
- OWL based integration of experiment

GGF Summer School 24th July 2004, Italy



The image displays two software interfaces. The top interface, Scuff Workbench, shows a workflow diagram with nodes like 'RESTRICTCREATEJOB' and 'CREATE_MUTANT_SEQUENCE' connected by arrows representing data flow. A green callout box points to this diagram with the text 'Ontology-aided workflow construction'. The bottom interface, Haystack, shows a detailed view of a sequence entry from GenBank (accession number al133523.5). It includes fields for Name, Genbank, Length, Topology, and Division. A green callout box points to a provenance graph below the sequence entry, which shows relationships like 'changed_gap', 'previous_gap', 'simplified_gap', and 'complete_gap' connected by 'created_by' and 'derived_from' relationships. This callout box contains the text 'RDF-based semantic mark up of results, logs, notes, data entries'.

Role of Ontologies



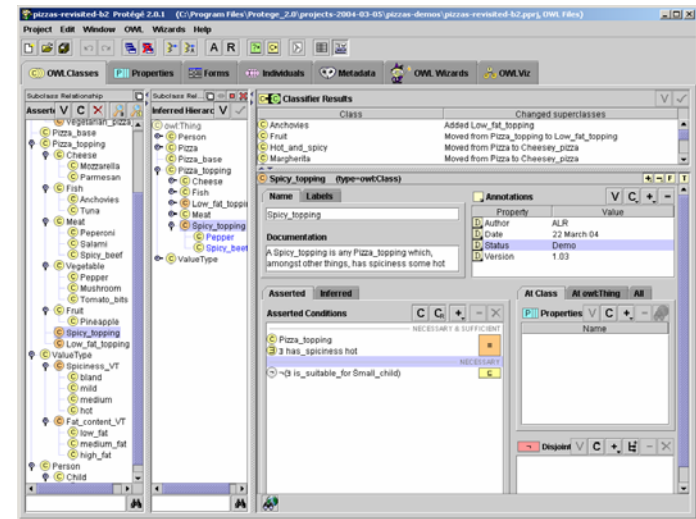
- Resource Description Framework
- W3C candidate recommendation (<http://www.w3.org/RDF>)
- Graphical formalism (+ XML syntax + semantics)
 - for representing metadata
 - for describing the semantics of information in a machine- accessible way
- RDFS extends RDF with “schema vocabulary”, e.g.:
 - Class, Property hasColleague
 - type, subclassOf, subPropertyOf
 - range, domain
- Statements are <subject, predicate, object> triples:



W3C Web Ontology language OWL



- The Ontology Language *de jour*
- Continuum of expressivity
 - Concepts, roles, individuals, axioms
 - From simple frames to description logics
 - Sound and complete formal semantics
- Supports reasoning to infer classification
 - Based on the SHIQ description logic





A pioneer of the...

The Semantic Grid is an extension of the current Grid in which information and services are given well-defined and explicitly represented meaning, better enabling computers and people to work in cooperation

Semantic Grid


Semantics in and on the Grid

The semantics of knowledge

- Semantic Grids
 - Grids and Grid middleware that makes use of semantics for its installation, deployment, running etc.
 - I.e. Semantics IN the Grid FOR the Grid.



Knowledge Grids

- A virtual knowledge base derived by using the Grid resources, in the same spirit as a data grid is a virtual data resource and a compute grid a virtual computer. Knowledge Grids include services for knowledge mining. 
- Semantics ON the Grid arising from



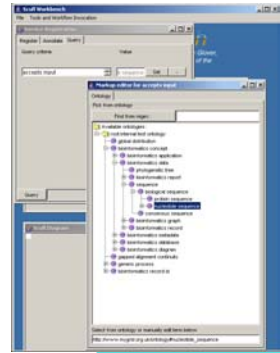
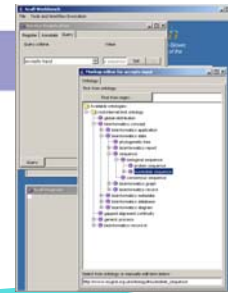
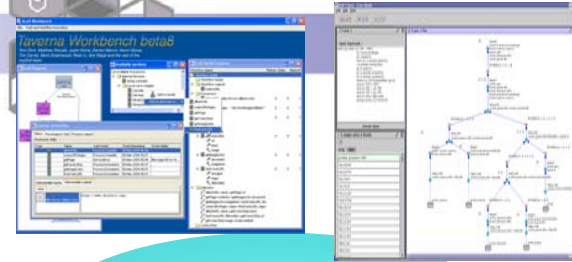
Roadmap

- Part 1
 - Application context
- Part 2
 - Architecture
 - Information and Workflows
 - Semantics and provenance
- Part 3
 - Wrap up



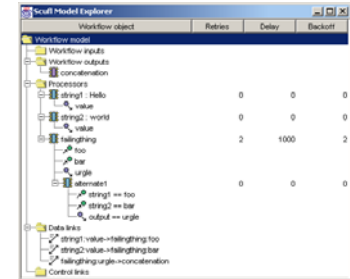
Key Characteristics

- Data Intensive, Up stream analysis
- Pipelines - experiments as workflows (chiefly)
- Adhoc exploratory investigative workflows for individuals from no particular a priori community
- **Openness – the services are not ours.**
- Low activation energy, incremental take-on
- Foundations for sharing knowledge and sharing experimental objects
- Multiple stakeholders
- Collection of components for assembly



Forming experiments

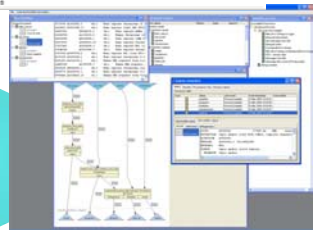
Personalisation



Discovering and reusing experiments and resources



Executing and monitoring experiments

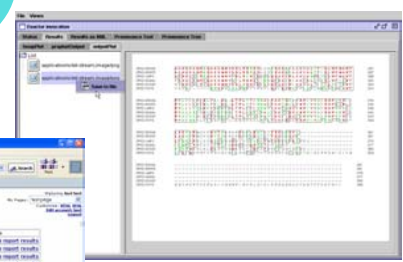
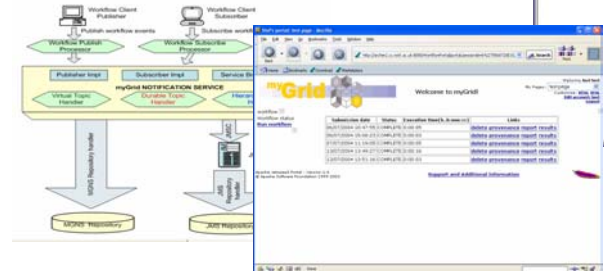
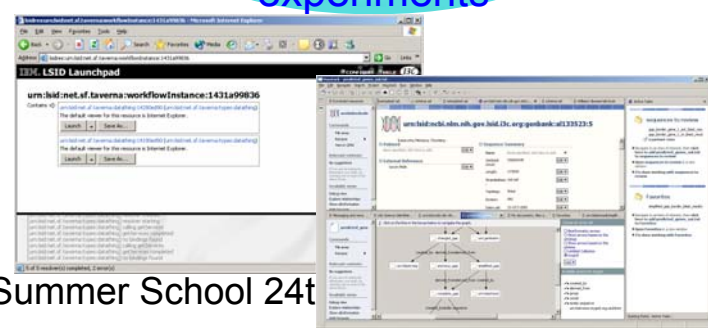


Sharing services & experiments

Managing lifecycle, provenance and results of experiments



Soaplab





Putting the user first

User-driven end to end scenarios essential

Whole solution that fits with them

Users vs Machines (vs Interesting computer science)

- Mismatch for information needs
 - ScufI instead of BPEL/WSFL
 - Layers of Provenance
 - Service/workflow descriptions for PEOPLE not just machines
 - Bury complexity, increasingly simplify
- Bioinformaticans **HARDLY EVER** want to have their services automatically selected
 - Except SHIMs, Replicas, User specified equivalences

Service providers and developers are users too!

Security

- Single sign-on to myGrid services
- Credentials mapping to external services (though most are open and free)
- Policy-driven authorization
- Solutions?
 - PERMIS, Shibboleth, WS-Security, XACML, SAML
 - FAME/PERMIS, SAM

Reuse

- Describing for reuse is challenging
 - Reuse depends on semantic descriptions and these are costly to produce
 - Describing for someone else's benefit
 - Reuse by multiple stakeholders
- Licensing workflows for reuse.
- Authorisation models
- But reuse does happen!
- Other genomic disorders (e.g. sick cows)
- Metadata pays off but it needs a network effect and there is a cost.



Personalisation

- Dynamic creation of personal data sets.
- Personal views over repositories.
- Personalisation of workflows.
- Personal notification
- Annotation of datasets and workflows.
- Personalisation of service descriptions – what I think the service does.

Standards

- By tapping into (defacto) standards (LSID, RDF, WS-I) and communities we can leverage others results and tools
 - Haystack, Pedro, Jena, CHEF/Sakai.
- The Grid standards are confusing and volatile
 - The choice of vanilla Web Services was good.
 - We didn't jump to OGSI. We won't jump to WSRF until its necessary.
- And workflow standards have been untimely.



Where is the WSRF?

There isn't any – vanilla Web Services



Computational processes

- Most service are quick pipes
- Long running services
 - Gene expression clustering service in Hong Kong
 - parking the data at a URL & notification through polling or email (GridFTP, event notification, data staging!)
 - Integrative Biology e-Science pilot follow-on to include simulation services
 - High throughput BLAST with NCBI update profile
- Stateful interactions

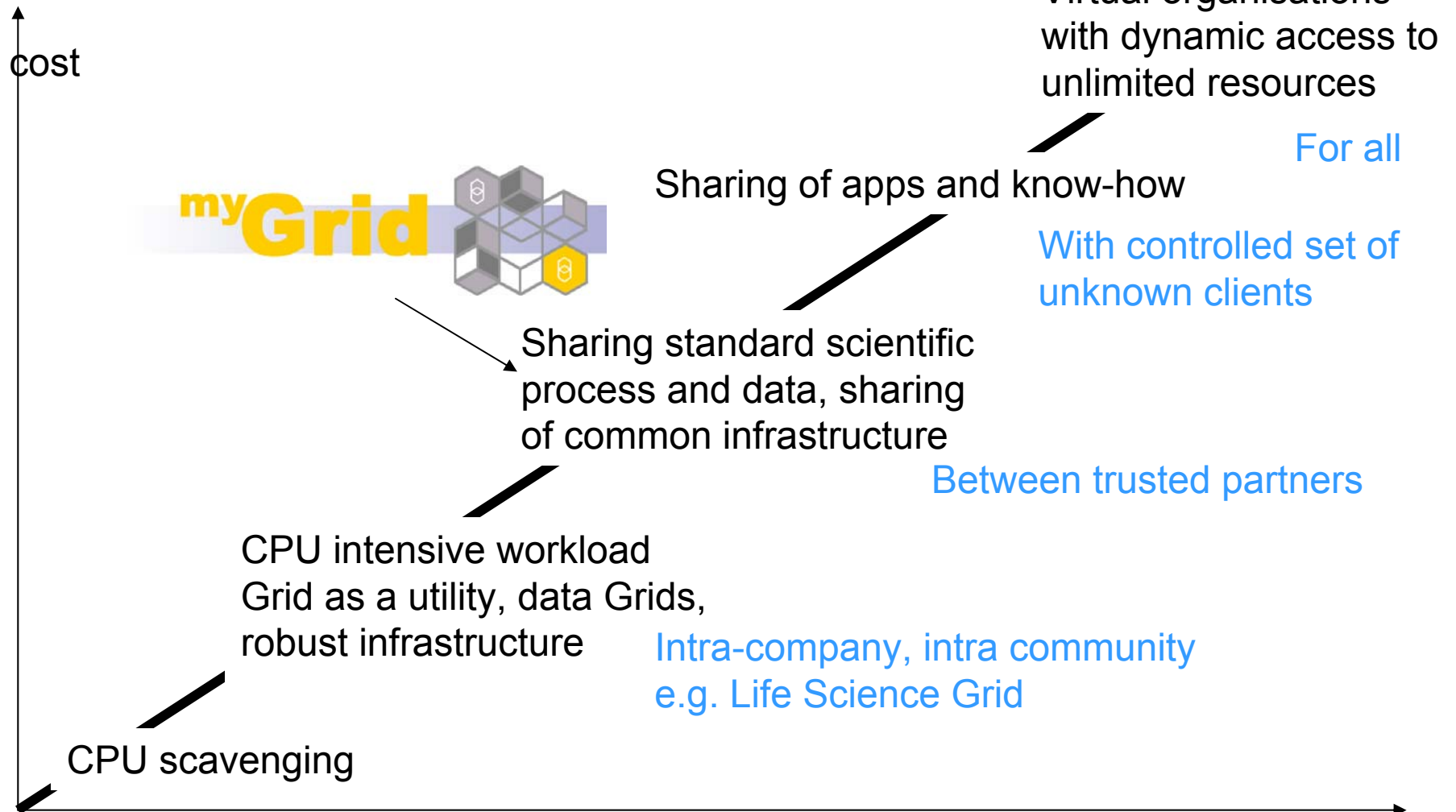


Observations

- Show stoppers for practical adoption are not technical showstoppers
 - Can I incorporate my favourite service?
 - Can I manage the results?
- Service providers are a bottleneck
- For every user dedicate a technologist.
- Caution against technology push.
- Rapid prototyping, deployment, feedback crucial.



Grid Computing trajectory





Acknowledgements

An EPSRC funded UK eScience Program Pilot Project



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



Core

- Matthew Addis, Nedim Alpdemir, Tim Carver, Rich Cawley, Neil Davis, Alvaro Fernandes, Justin Ferris, Robert Gaizaukaus, Kevin Glover, Carole Goble, Chris Greenhalgh, Mark Greenwood, Yikun Guo, Ananth Krishna, Peter Li, Phillip Lord, Darren Marvin, Simon Miles, Luc Moreau, Arijit Mukherjee, Tom Oinn, Juri Papay, Savas Parastatidis, Norman Paton, Terry Payne, Matthew Pokock Milena Radenkovic, Stefan Rennick-Egglestone, Peter Rice, Martin Senger, Nick Sharman, Robert Stevens, Victor Tan, Anil Wipat, Paul Watson and Chris Wroe.

Users

- Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK
- Hannah Tipney, May Tassabehji, Andy Brass, St Mary's Hospital, Manchester, UK
- Steve Kemp, Liverpool, UK

Postgraduates

- Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, John Dickman, Keith Flanagan, Antoon Goderis, Tracy Craddock, Alastair Hampshire

Industrial

- Dennis Quan, Sean Martin, Michael Niemi, Syd Chapman (IBM)
- Robin McEntire (GSK)

Collaborators

- Keith Decker
GGF Summer School 24th July 2004, Italy



<http://www.mygrid.org.uk>

Tutorial

<http://twiki.mygrid.org.uk/twiki/bin/view/Mygrid/NeSCmyGridTutorial>

Publications

- P Lord, C Wroe, R Stevens, CA Goble, S Miles, L Moreau, K Decker, T Payne, J Papay, *Semantic and Personalised Service Discovery* in Proceedings IEEE/WIC International Conference on Web Intelligence / Intelligent Agent Technology Workshop on "Knowledge Grid and Grid Intelligence" October 13, 2003, Halifax, Canada.
- J Zhao, CA Goble, M Greenwood, C Wroe, R Stevens *Annotating, linking and browsing provenance logs for e-Science* in 1st Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data, Florida, USA, October 2003
- C Wroe, R.D. Stevens, CA Goble, A Roberts, M Greenwood *A suite of DAML+OIL ontologies to describe bioinformatics web services and data.* International Journal of Cooperative Information Systems. Special issue on Bioinformatics and Biological Data Management 12(2):197-224, 2003.
- C Wroe, CA Goble, M Greenwood, P Lord, S Miles, L Moreau, J Papay, T Payne *Experiment automation using semantic data on a bioinformatics Grid*, IEEE Intelligent Systems, Jan/Feb 2004
- J Zhao, C Wroe, CA Goble, R Stevens, D Quan, M Greenwood, *Using Semantic Web Technologies for Representing e-Science Provenance* in Proc 3rd International Semantic Web Conference ISWC2004, Hiroshima, Japan, 9-11 Nov 2004.
- C Wroe, P Lord, S Miles, J Papay, L Moreau, C Goble *Recycling Services and Workflows through Discovery and Reuse* to appear in Proceedings UK e-Science All Hands Meeting Nottingham, UK, 1-3 September, 2004.

- T Oinn, M Addis, J Ferris, D Marvin, M Senger, M Greenwood, T Carver, K Glover, Matthew R. Pocock, A Wipat, P Li. *Taverna: A tool for the composition and enactment of bioinformatics workflows* accepted for Bioinformatics Journal, 16 June 2004
- T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Greenwood, C. Goble, A. Wipat, P. Li, T. Carver *Delivering Web Service Coordination Capability to Users* In Thirteenth International World Wide Web Conference (WWW2004) pp. 438-439, New York, May 2004.
- M Addis, J Ferris, M Greenwood, D Marvin, P Li, T Oinn and A Wipat *Experiences with eScience workflow specification and enactment in bioinformatics*, Proceedings of UK e-Science All Hands Meeting 2003, pages 459-467
- M.N. Alpdemir, A. Mukherjee, N.W. Paton, P. Watson, A.A.A. Fernandes, A. Gounaris and J. Smith *Service-based Distributed Querying on the Grid* in the Proceedings of the First International Conference on Service Oriented Computing, 15-18, December 2003 Trento, Italy. Springer.
- J. Smith, A. Gounaris, P. Watson, N.W. Paton, A.A.A. Fernandes and Rizos Sakellariou *Distributed Query Processing on the Grid* in International Journal of High Performance Computing Applications, Volume 17, Issue 04, November 2003

Publications

- R. Stevens, H.J. Tipney, C. Wroe, T. Oinn, M. Senger, P. Lord, C.A. Goble, A. Brass and M. Tassabehji *Exploring Williams-Beuren Syndrome Using myGrid* to appear in Proceedings of 12th International Conference on Intelligent Systems in Molecular Biology, 31st Jul-4th Aug 2004, Glasgow, UK.
- C.A. Goble, S. Pettifer, R. Stevens and C. Greenhalgh *Knowledge Integration: In silico Experiments in Bioinformatics* in *The Grid: Blueprint for a New Computing Infrastructure Second Edition* eds. Ian Foster and Carl Kesselman, 2003, Morgan Kaufman, November 2003.

R. Stevens, A. Robinson, and C.A. Goble *myGrid: Personalised Bioinformatics on the Information Grid* in proceedings of 11th International Conference on Intelligent Systems in Molecular Biology, 29th June–3rd July 2003, Brisbane, Australia, published *Bioinformatics* Vol. 19 Suppl. 1 2003, pp302-304.