

# myGrid: An Open Platform for Data-intensive Post-Genomic Functional Analysis

Matthew Addis  
IT Innovation Centre  
[mja@it-innovation.soton.ac.uk](mailto:mja@it-innovation.soton.ac.uk)



## myGrid in a Nutshell

- “The key to bioinformatics is integration, integration, integration,”

*bioinformatics expert Jim Golden, Curagen spin-off 454 Corporation*

Bioinformatics: Bringing it all together, Nature 17 October 2002

- myGrid is all about integrating bioinformatics tools and data to support the scientific processes of *in-silico* experimentation

# Contents

- Project timescales, funding and partners
- What problems is myGrid trying to solve?
- The myGrid approach
- Canned demo
- Web Services, Grid and the Semantic Grid
- Next Steps
- Further information

## myGrid timescales and funding

- EPSRC eScience pilot project
  - Official start October 2001
  - Actual start January 2002
  - Planned end March 2005
- £3M EPSRC funding + contribution from industrial partners
- 16 RAs, 9 studentships
- Early days – still two thirds of the project to go...

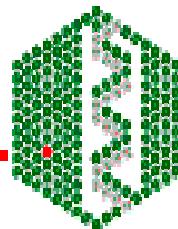
# myGrid partners



MERCK



EMBL  
European Bioinformatics Institute



# Problems: heterogeneity everywhere

- Seamless access to bioinformatics data sources and tools is not easy
  - Data formats
  - Data access mechanisms
  - Data annotations and interpretations
  - Analysis techniques and implementations
  - Service Providers
- Relatively few standards
  - GO
  - DAS
  - BioMOBY, I3C
- EBI hosted tools 50+
  - Homology & Similarity
  - Prot. Function. Analysis
  - Structural Analysis
  - Sequence Analysis
  - Miscellaneous Tools
- EBI hosted databases 30+
  - Nucleotide Databases
  - Protein Databases
  - Proteome Analysis
  - Structure Databases
  - Microarray Database
  - Literature Databases

# Problems: access mechanisms



Courtesy of Mark Wilkinson (BioMOBY)

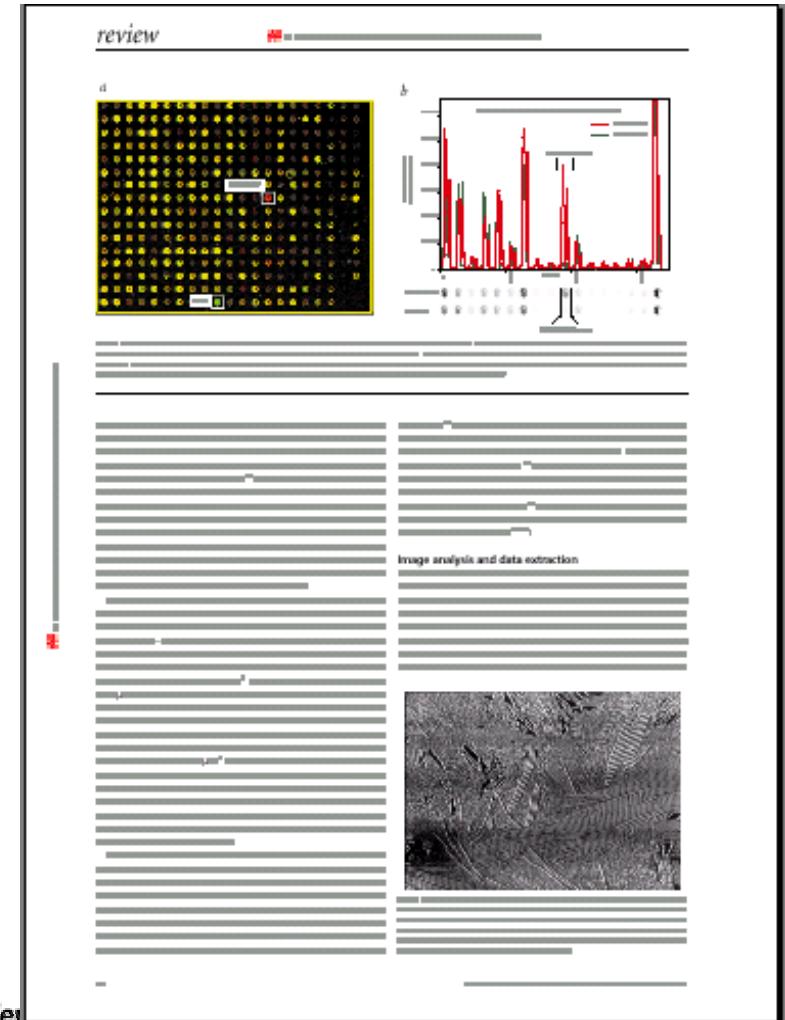
# Problems: data isn't just numbers

InterPro Entry IPR000025 - Netscape

File Edit View Go Communicator Help

Database	InterPro
Accession	IPR000025; Melatonin_receptor (matches 22 proteins)
Name	Melatonin receptor
Type	Family 
Dates	08-OCT-1999 (created) 27-MAR-2000 (last modified)
Signatures	<a href="#">PR00857</a> ; MELATONINR (22 proteins)
Parent  tree	<a href="#">PR00276</a> ; Rhodopsin-like GPCR superfamily (3990 proteins)
Children  tree	<a href="#">PR00278</a> ; Melatonin 1A receptor (12 proteins) <a href="#">PR00279</a> ; Melatonin 1C receptor (5 proteins) <a href="#">PR00280</a> ; Melatonin-related 1X receptor (3 proteins)
Function 	melatonin receptor ( <a href="#">GO:0008502</a> )
Component 	membrane ( <a href="#">GO:0016020</a> )
<b>Abstract</b>  <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 G<sub>i</sub>/G<sub>o</sub> class [5].</p>	
Examples	<ul style="list-style-type: none"> <li><a href="#">P49298</a> ML1C_CHICK</li> <li><a href="#">P49285</a> ML1A_CHICK</li> <li><a href="#">P49279</a> ML1C_XENLA</li> <li><a href="#">P49277</a> ML1A_PHOSU</li> </ul> <a href="#">View examples</a>
References	<ol style="list-style-type: none"> <li>Attwood TK., Findlay J.B.C. Fingerprinting G-protein-coupled receptors. Protein Eng. 7: 195-203(1994). [<a href="#">MEDLINE:94224751</a>] [<a href="#">PUBMED:004961</a>]</li> </ol>

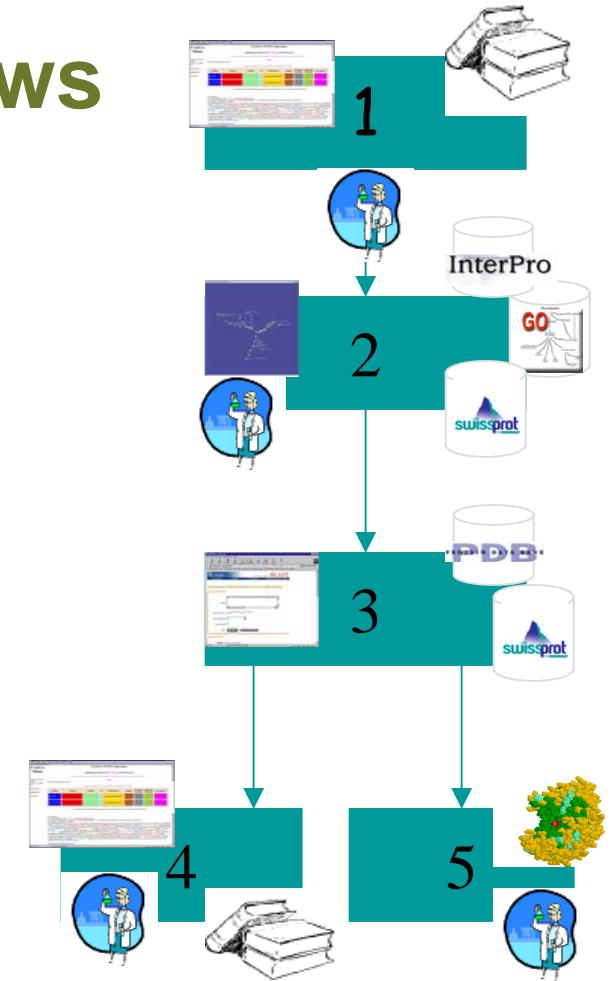
Document Done



# Problems: in-silico experiments are workflows

## Circadian Rhythms

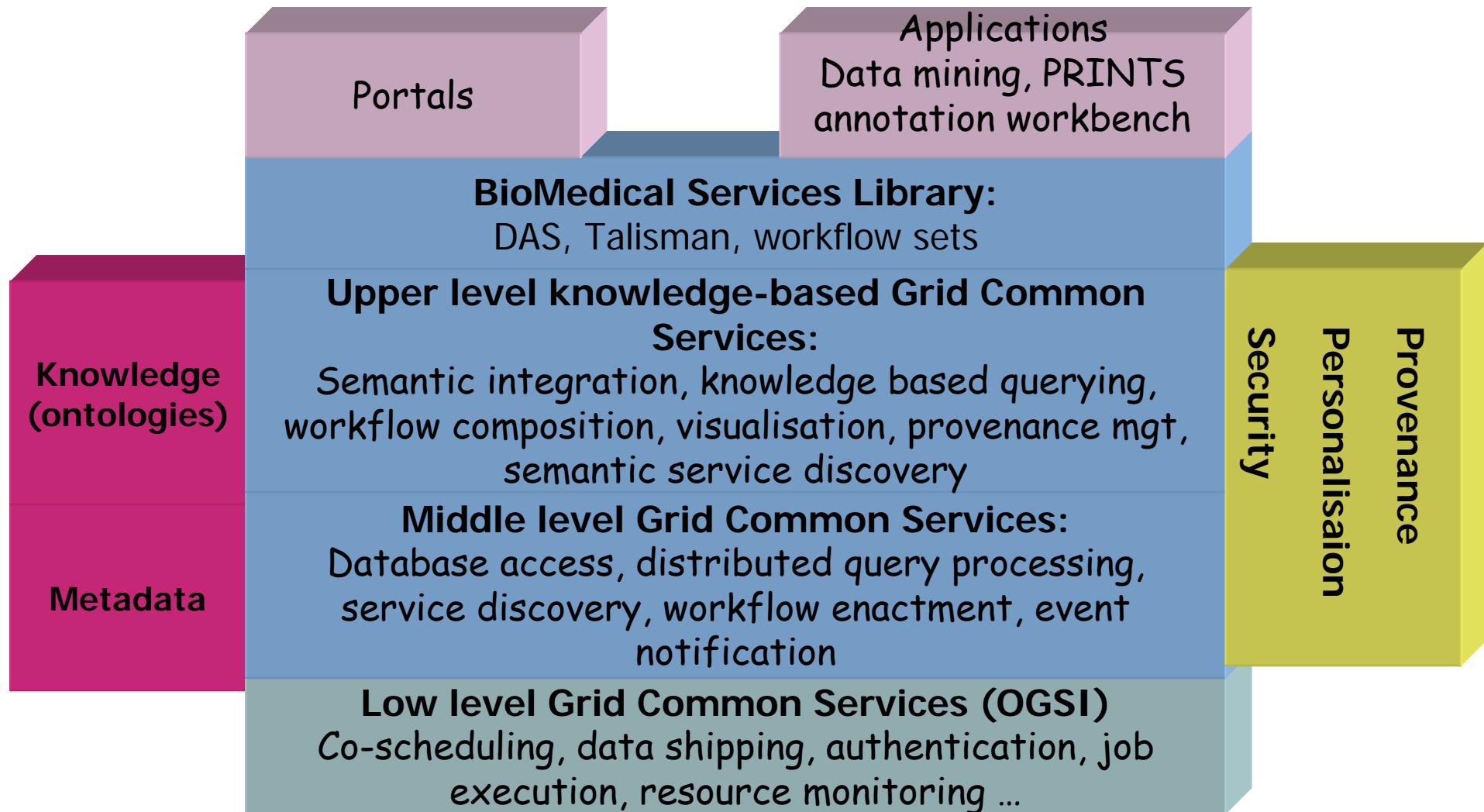
- Has anyone else studied the effect of neurotransmitters on the circadian rhythms in *Drosophila*?
- I've got a cluster of proteins from my experiment. How do their functions interrelate? And what are the proteins with a particular function?
- Is a structure known for my protein? What other proteins have a similar structure?
- What is known about the homologous protein?
- Can I build a homology 3D model?



# Problems: e-Science

- Personalisation
  - Who else has asked this question & can I use/adapt their approach?
  - I want to annotate and publish my process for use by others
  - I want to store and access my personal datasets
- Provenance
  - Which type, version and provider of BLAST did I use?
  - What was the workflow and the results at each stage?
  - I want to publish my findings as an protein annotation
  - Ownership, credit, trust, immutable and auditable data
- Change management and notification
  - When was P12345 last updated?
  - Has PDB changed since I last ran this workflow?
  - Has the data provenance changed?
  - Are there new or alternative services that I can use?

# myGrid approach: marketecture diagram



# myGrid approach: modular set of services

- Ontologies and knowledge-based services
  - Description and interpretation of data, tools and services
- Workflow authoring, publication and enactment
  - Service orchestration and process automation
- Directories and repositories
  - Services, workflows, data and results
- Portal
  - Easy access to all myGrid services
- Notification and event propagation
  - Change management
- Knowledge Extraction
- Security and trust



# Example: describe the workflow

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

myGrid

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

myGrid

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

myGrid

myGrid Service Descriptor

A description of a myGrid service:

requires input

uses method

produces result

performs task

uses resource

is function of

Files

Jobs

Current Users

History

Import

Logout

myGrid Service Matching

The above description matches these services:

myGrid Workflow Description

The workflow consists of these services:

```
graph TD; A[mygrid retrieve GO term using SwissProt accession number] --> B[mygrid display 3D plot of Gene Ontology lattice];
```

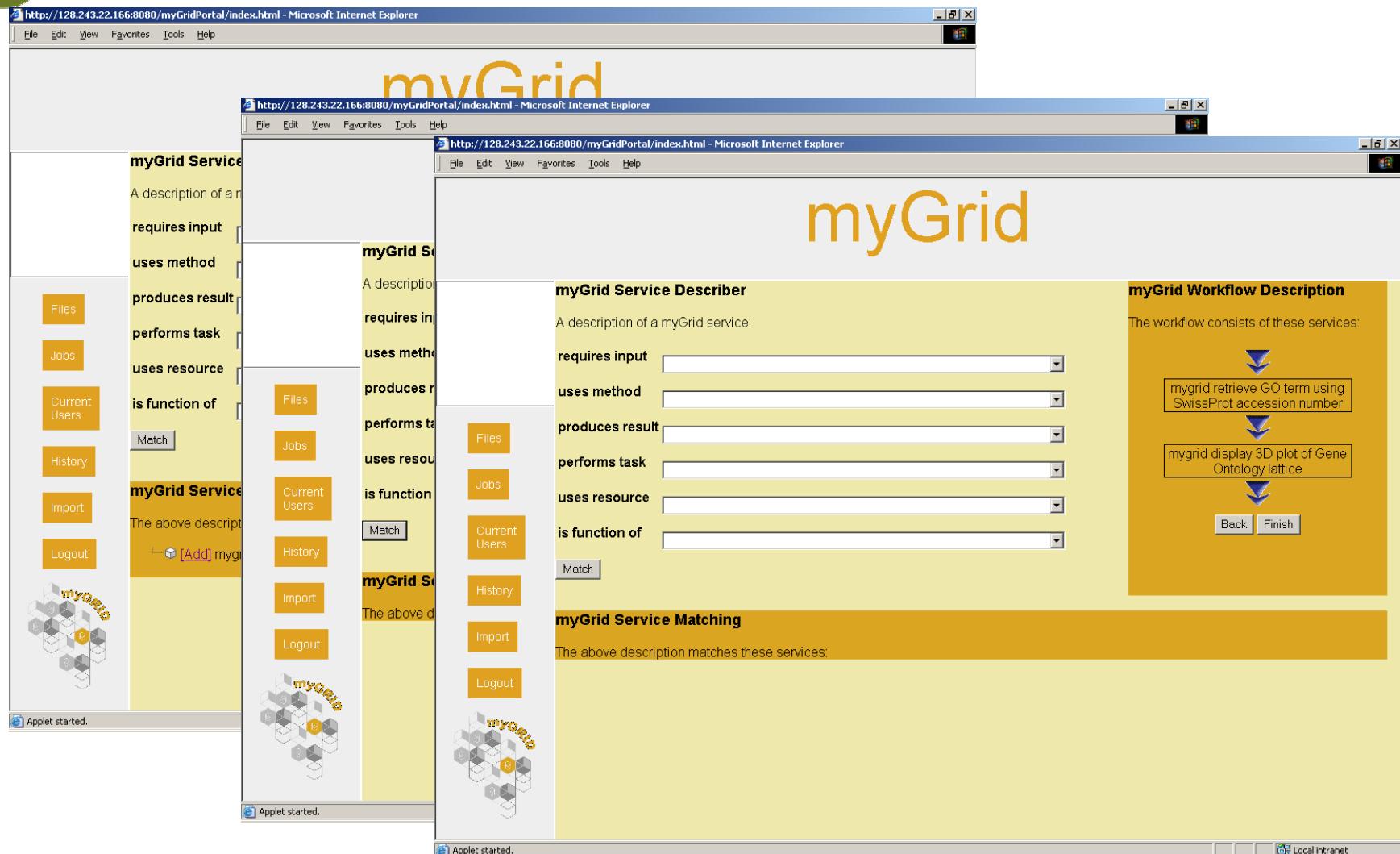
Back  Finish

Applet started.

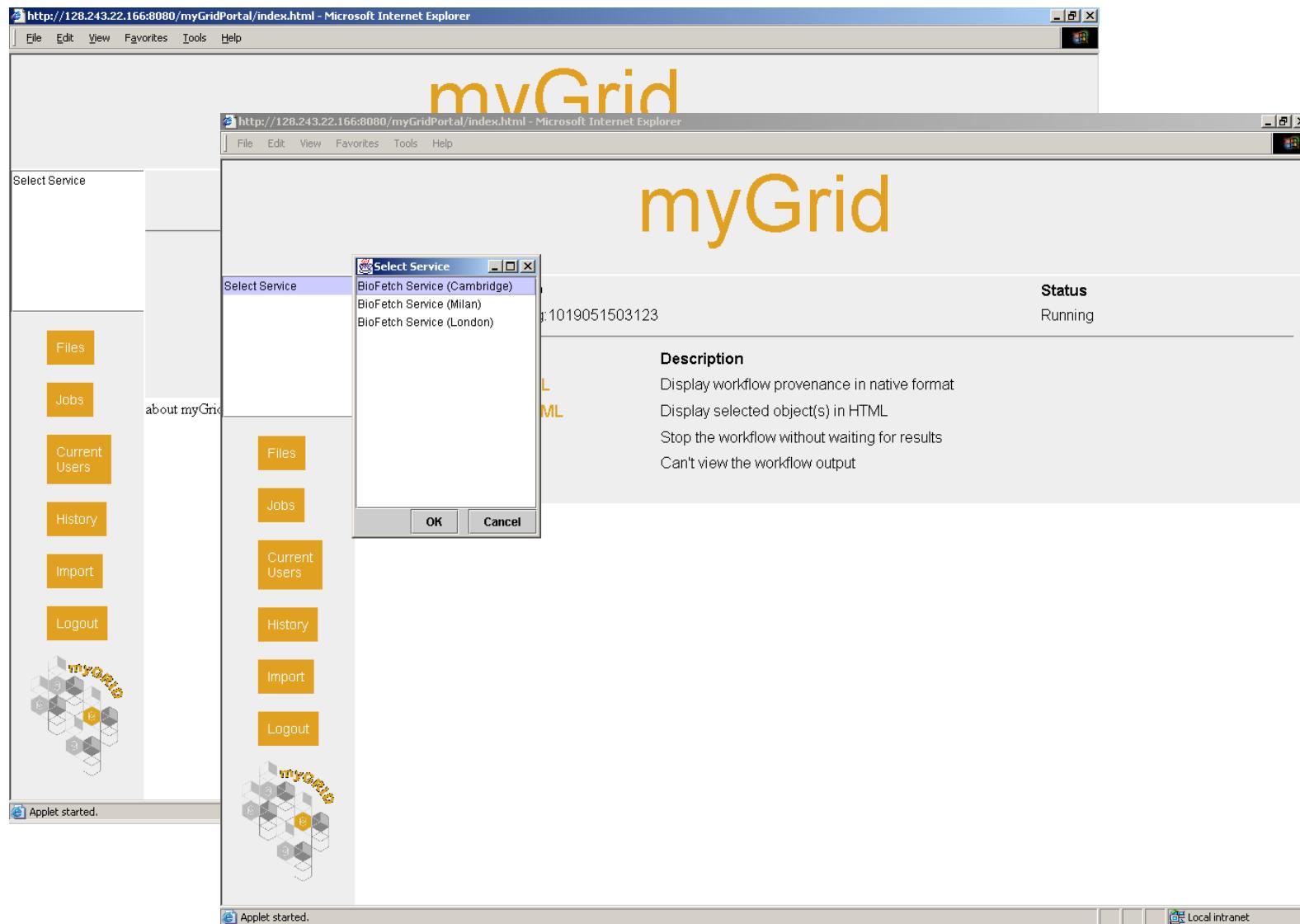
Applet started.

Applet started.

Local intranet



# Example: execute the workflow



The screenshot shows a Microsoft Internet Explorer window displaying the myGrid portal. The main page has a large orange "myGrid" logo. On the left, a sidebar menu lists "Select Service", "Files", "Jobs", "Current Users", "History", "Import", and "Logout". Below this is a "myGrid" logo icon. The main content area shows a workflow execution dialog box. The dialog has a title "Select Service" and a list of options: "BioFetch Service (Cambridge)", "BioFetch Service (Milan)", and "BioFetch Service (London)". The "BioFetch Service (Cambridge)" option is selected. To the right of the dialog, the workflow details are shown: "Workflow ID: 1019051503123", "Status: Running", and a "Description" section with the following text:  
Display workflow provenance in native format  
Display selected object(s) in HTML  
Stop the workflow without waiting for results  
Can't view the workflow output

# Example: view results

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

http://128.243.22.166:8080/myGridPortal/index.html - Microsoft Internet Explorer

File Edit View Favorites Tools Help

uk.ac.ebi.escience.tmo.iviewer.IViewer

myGrid

Workflow ID: workflow:king:1018605180834

Status: Could not retrieve the status, sorry :(

Operation:

- Display XML
- Display HTML
- Stop Job
- View Output

Description:

- Display workflow provenance in native format
- Display selected object(s) in HTML
- Stop the workflow without waiting for results
- View the workflow output

Workflow Provenance Data

Workflow ID: workflow:king:1018605180834

User: mygrid

Inputs:

Name	Type	Value
SwissprotAccessions	string[]	Q9VKQ5 Q9WM4 P91657 Q9V720 Q9VX4 Q9VBS4 Q9X214 Q9VUY1 Q9V6R4 Q9VV15 Q9VMB6 Q9VX28 Q9VKR8 Q9VGC4 Q9VC43 Q95078 Q26307 Q9VIC4 Q24511 Q9Y092 Q9VL27 Q95SV2 Q9VLP1 Q9VVD2

Outputs:

Name	Type	Value
null	string[]	GO:0008020 GO:0016789 GO:0016787 GO:0006821 GO:0007268 GO:0005247 GO:0005554 GO:0016301 GO:0006869

Applet started.

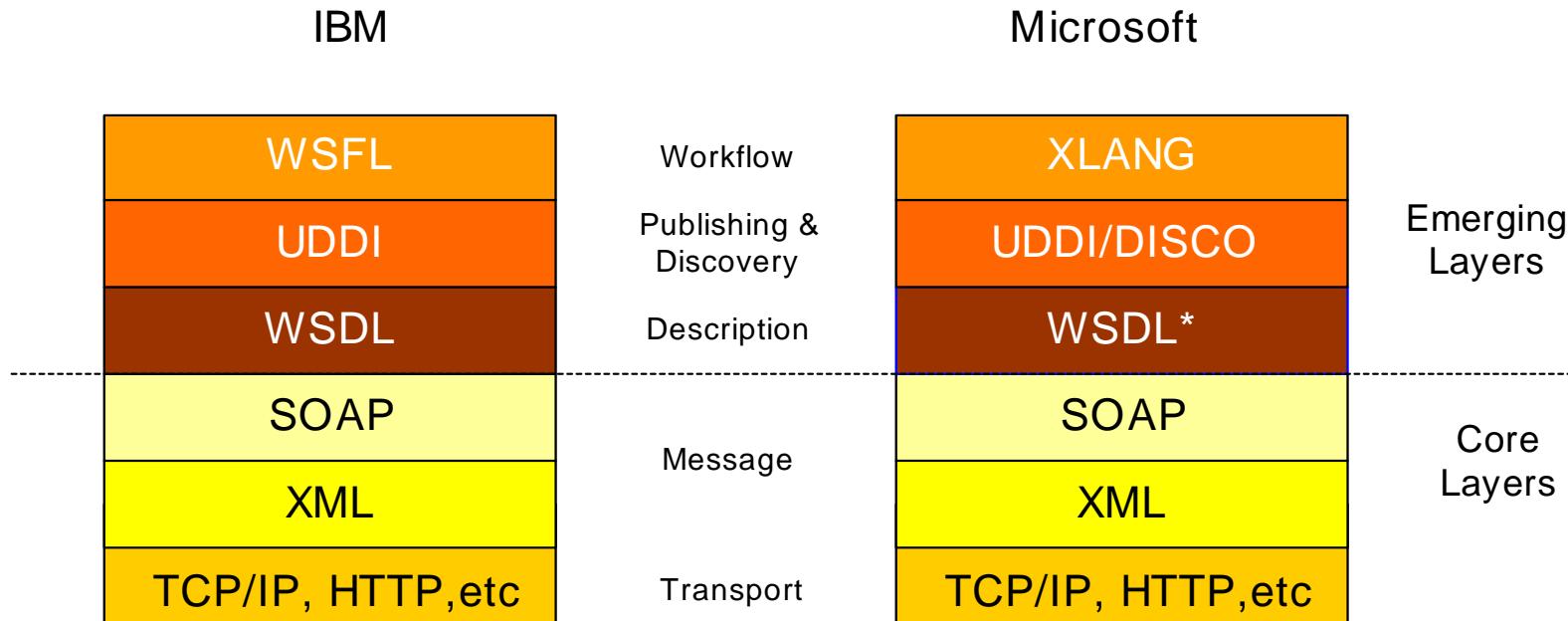
7601

## Extensions to example

- Knowledge extraction from literature
  - Extract medline references from SWISSPROT annotations
  - Retrieve journal abstracts
  - Analyse abstracts using NLP and ontologies
  - Automatically hypertext link literature
- Change management
  - User notified by EBI of changes to services
  - User (or their software agent) re-enacts workflow
  - New results compared with previous results
- More complexity: EMBOSS workflow
  - Combined two concurrent application flows
  - Executing 7 applications using 45 web service invocations

# Underpinning technologies: Web Services

- Platform, language and object model neutral
- Industry drive with a wide range of tools
- Rapid standardisation

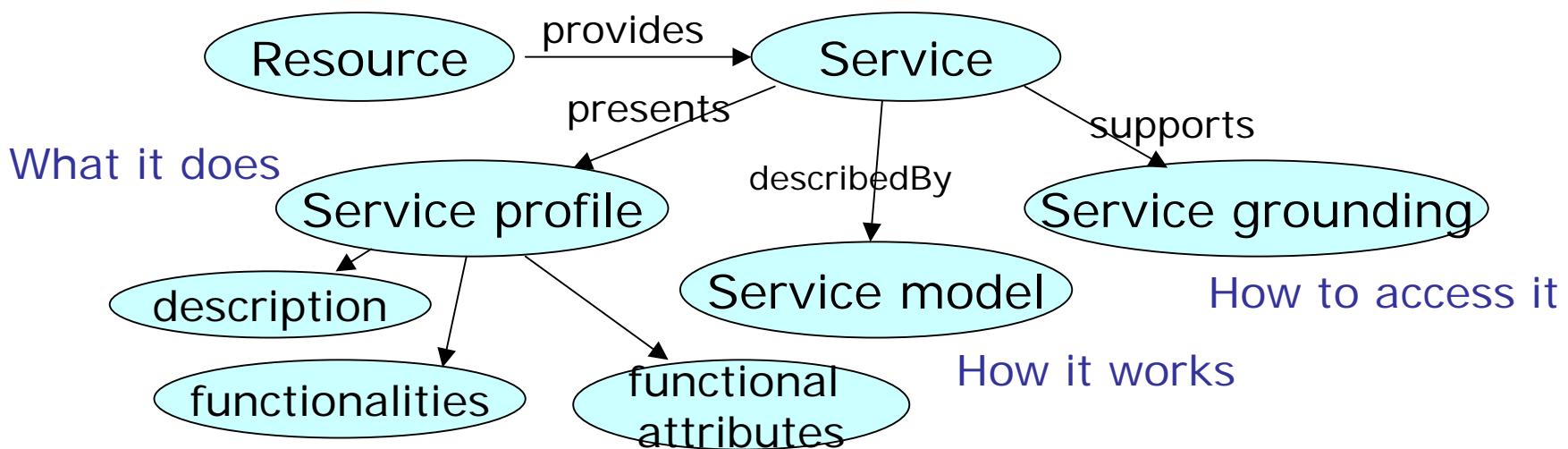


# Semantics is the key to interoperability

- Web Services and XML only provide syntax and a framework for communication
  - Still need human readable specifications
  - Still need software developers to write scripts
- Semantics are needed to allow software to ‘understand’ the data and the function of services
  - Improves service discovery by guided searches and inference
  - Substitution of alternative services that have the same function
  - Automatic generation of clients that use and orchestrate services
  - Automatic selection and application of data transformations

# Ontologies

- Ontologies: the shared and common understanding of a domain often in the form of a taxonomy
- Applies to services and people as well as biology
- Can be reasoned over by software



# myGrid layered service ontology

## 1. Class of service

protein sequence alignment,  
protein sequence database.

## 2. Specific services

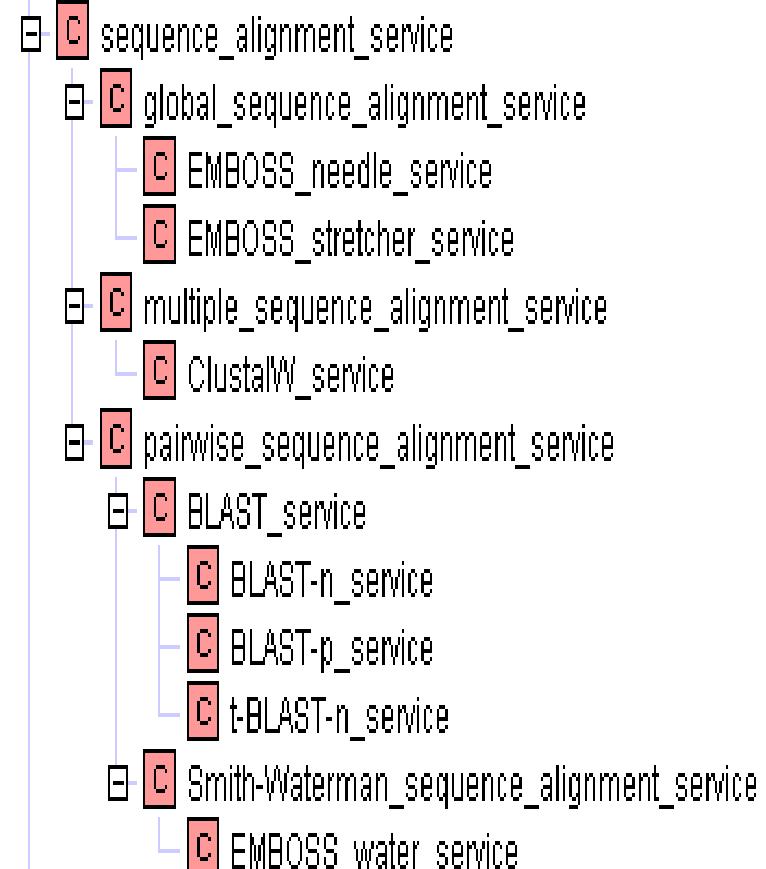
BLASTn is a tool for computing  
sequence homology that uses the  
BLAST algorithm over nucleotides

## 3. Instance description of specific services

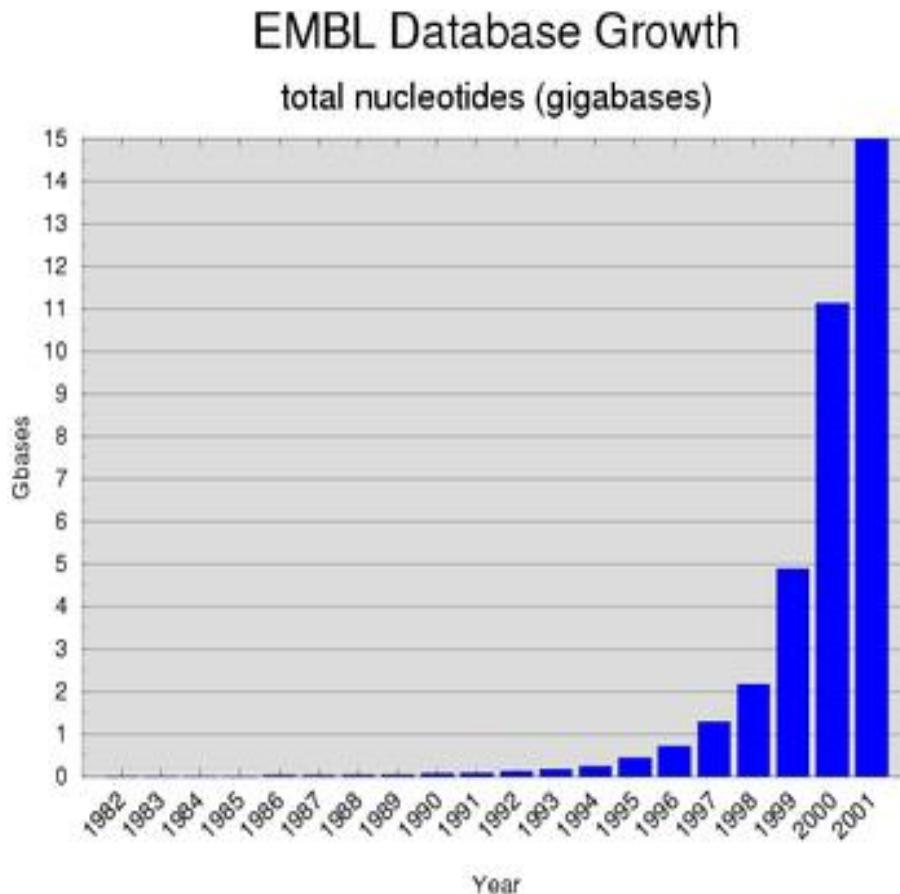
BLASTn service is provided by EBI

## 4. Description of invoked instance of a service

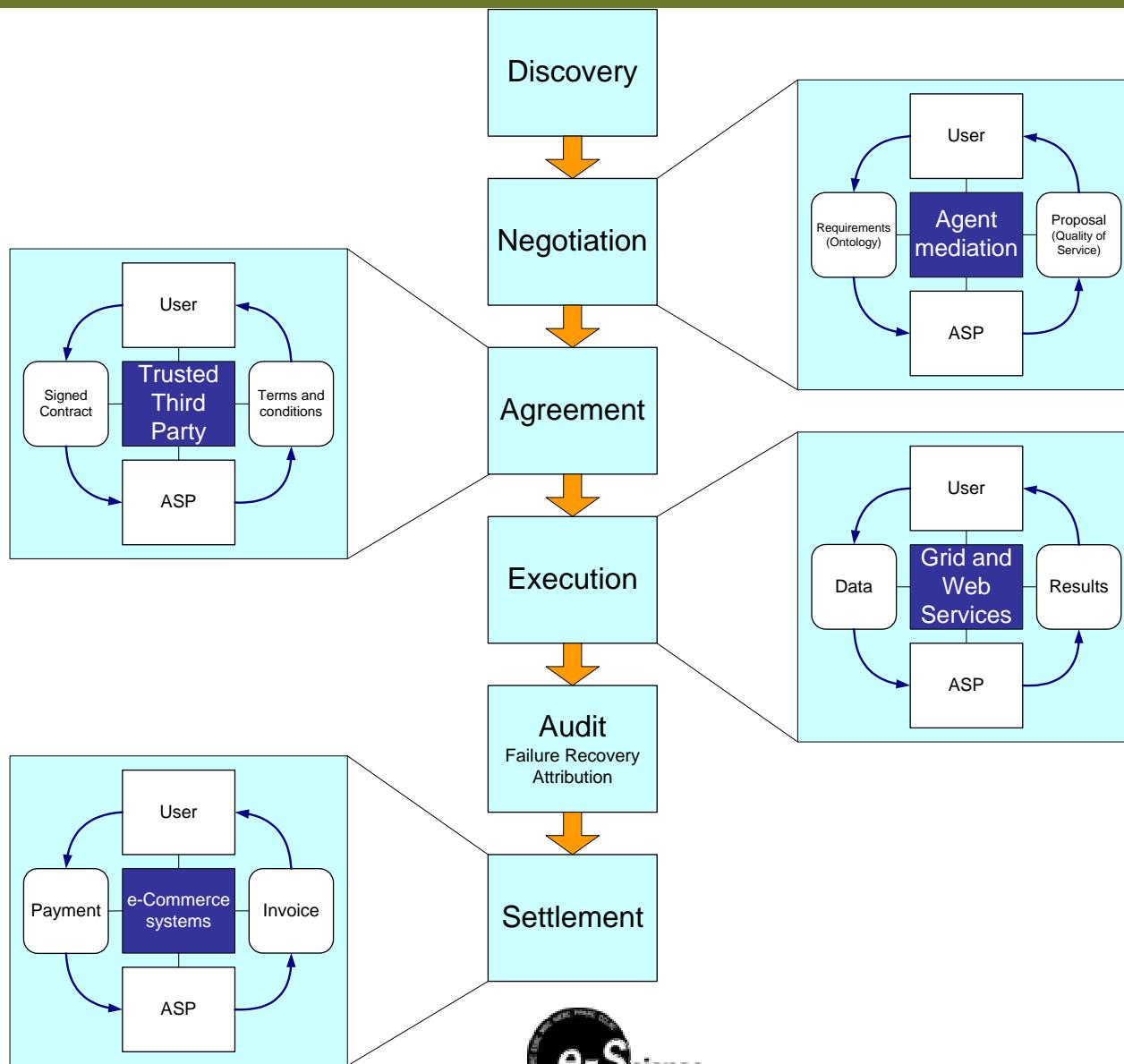
BLAST as executed for a particular  
workflow



# Underpinning technologies: Grid

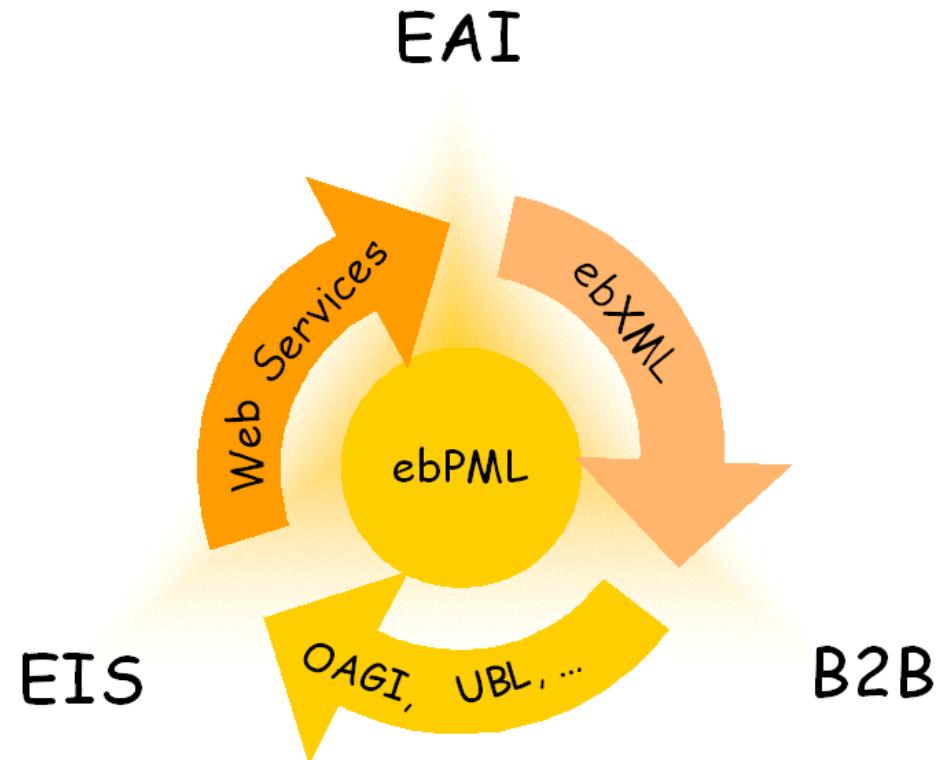


- Can't continue to centralise all resources in one place
  - Public service providers
  - Commercial users
- Large data volumes
  - Transfers and staging
  - Distributed query
- Computationally intensive
  - Resource management

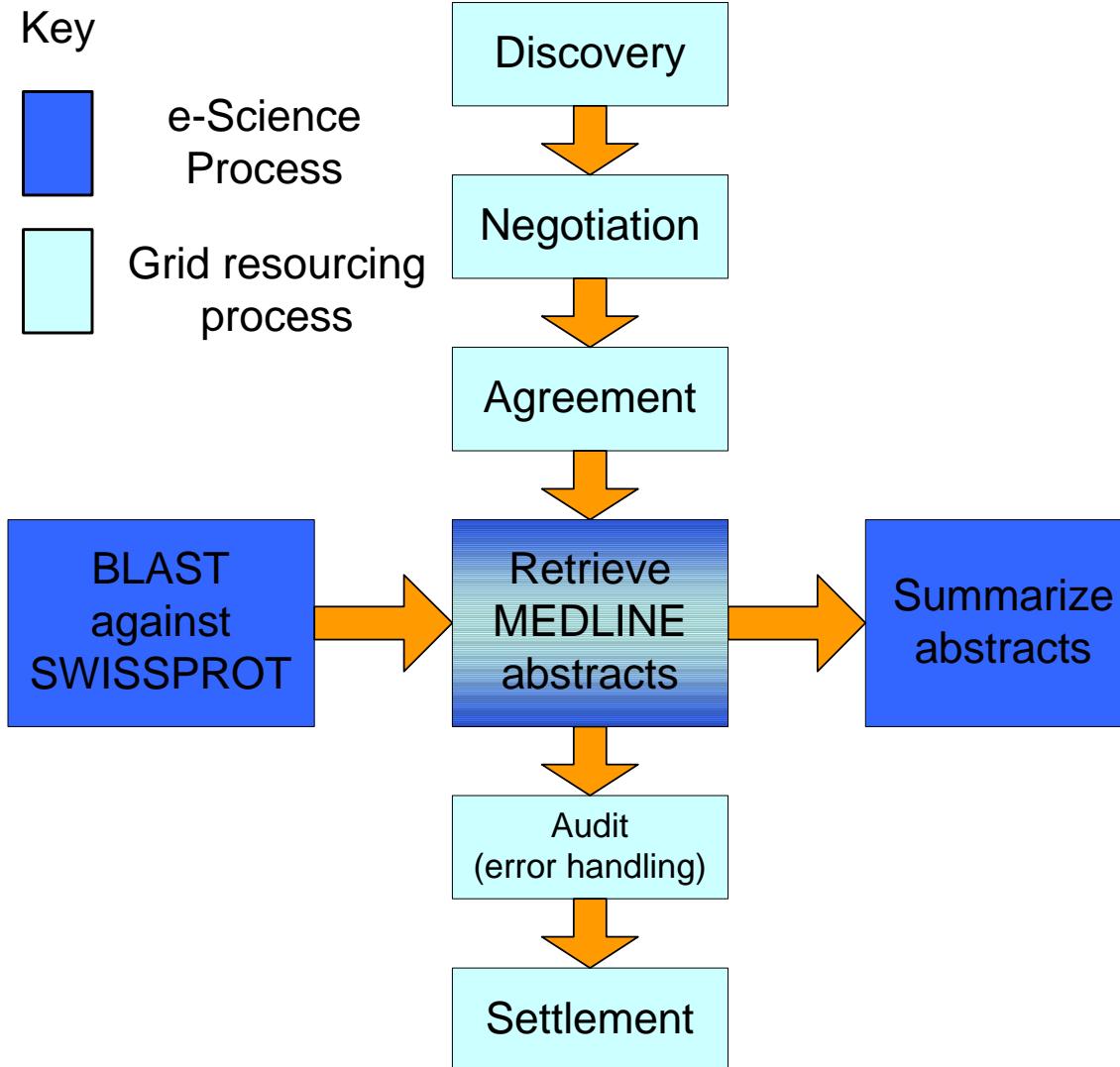


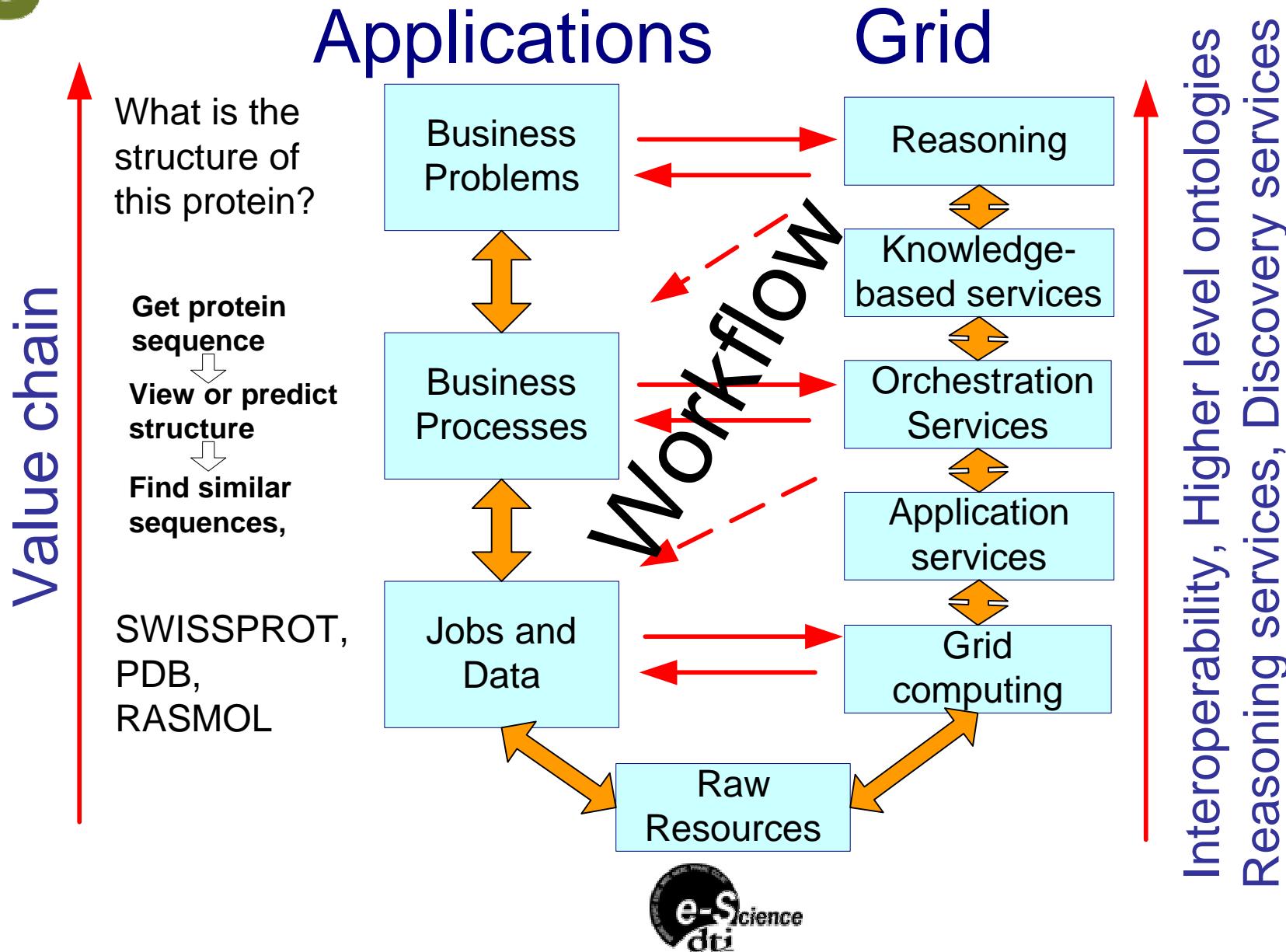
# Underpinning technologies: workflow

Workflow Management Coalition	WfMC
OMG	EDOC UML
Web Services	Xlang WSFL BPEL4WS BPML



© 2002 ebPML.org





# Next steps (1)

## Open Source

Open Bio Foundation BioJava, BioPerl ...

## (DeFacto) Standards

OMG LSR, I3C, MGED, Gene Ontology

## Other Projects

Astrogrid, Geodise, CLEF, Comb-e-chem, BIRN, OGSA-DAI

## Bioinformatics integration platforms

DAS, OpenBSA, ISYS, OpenMMS, Kleisli, Ensembl, AppLab, SRS, BioNavigator, DiscoveryLink, K1 TAMBIS, BioMOBY ...

## Web Services

XML, SOAP, WSDL, UDDI

## Distributed Computing Environments

CORBA, RMI, JavaOne

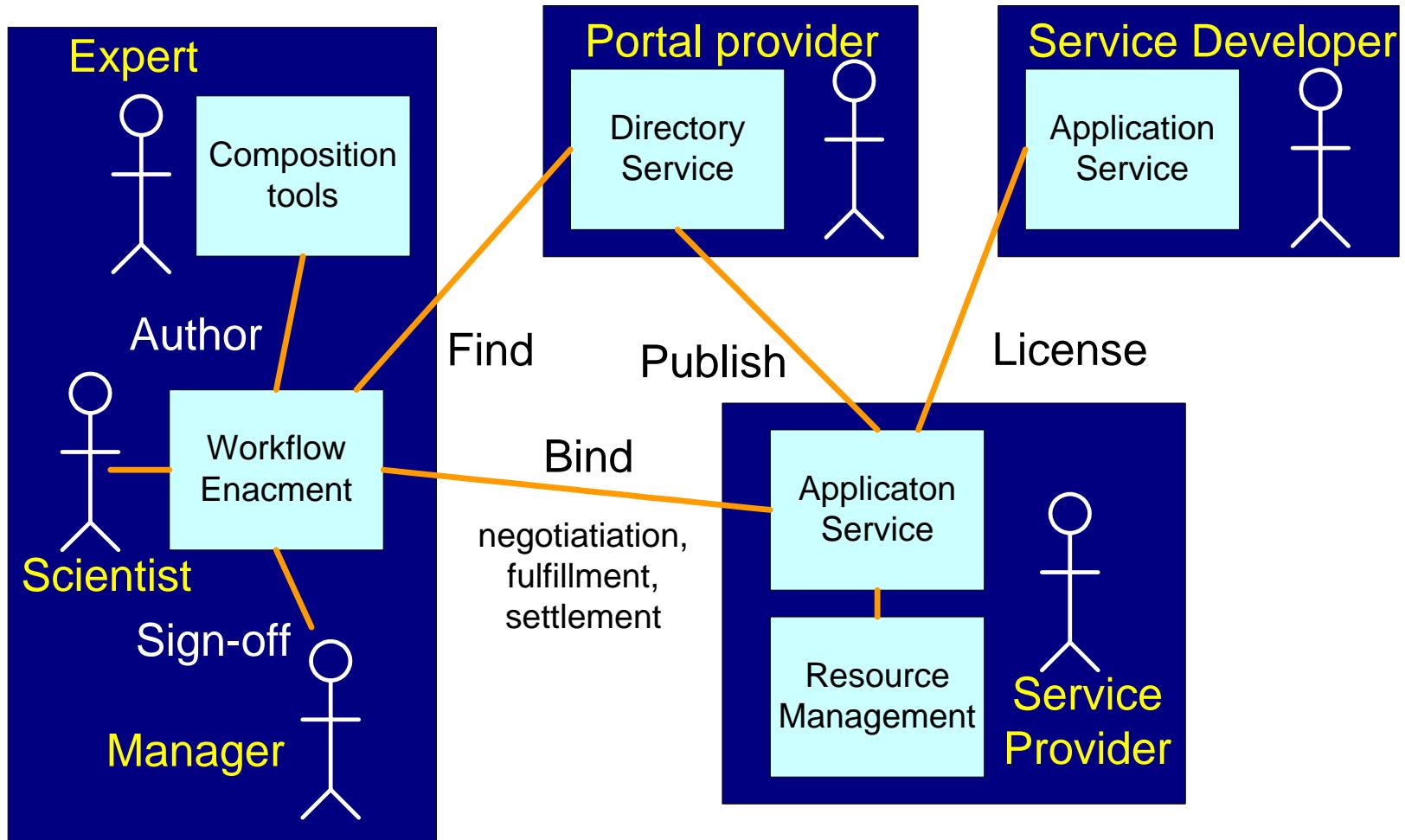
## GRID

Globus/SRB/Condor/Sun Grid Engine

## Next steps (2)

- Develop and execute real world use cases
  - More complexity
  - Interaction with the user and their local tools
- Continue to research and develop infrastructure
  - Provenance and security
  - Personalisation
  - Distributed Query
- Release myGrid to the community as open-source
  - Spring 2003

# Next steps (3)



# Conclusion

- myGrid aims to develop middleware to integrate bioinformatics tools and data in a way that supports the scientist
- The setting is bioinformatics but the results are intended to be generally applicable to e-Science.
- A mix of standard, vanguard and bleeding edge technologies, advanced development and (some) research.
- Semantics and workflow are key to making the most of Web Services and Grid technologies

## More information

- myGrid
  - Project Web Site <http://www.mygrid.org.uk>
  - Email myGrid [mygrid@cs.man.ac.uk](mailto:mygrid@cs.man.ac.uk)
- IT Innovation
  - Grid Projects <http://www.it-innovation.soton.ac.uk/grid>
  - Matthew Addis [mja@it-innovation.soton.ac.uk](mailto:mja@it-innovation.soton.ac.uk)