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

Matthew Addis
IT Innovation Centre
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

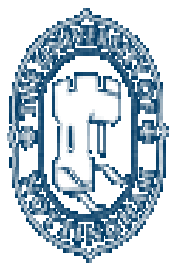
- Project timescales, funding and partners
- What problems is myGrid trying to solve?
- The myGrid approach
- Canned demo
- Web Services, Grid and the Semantic Grid
- Next Steps
- Further information

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



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

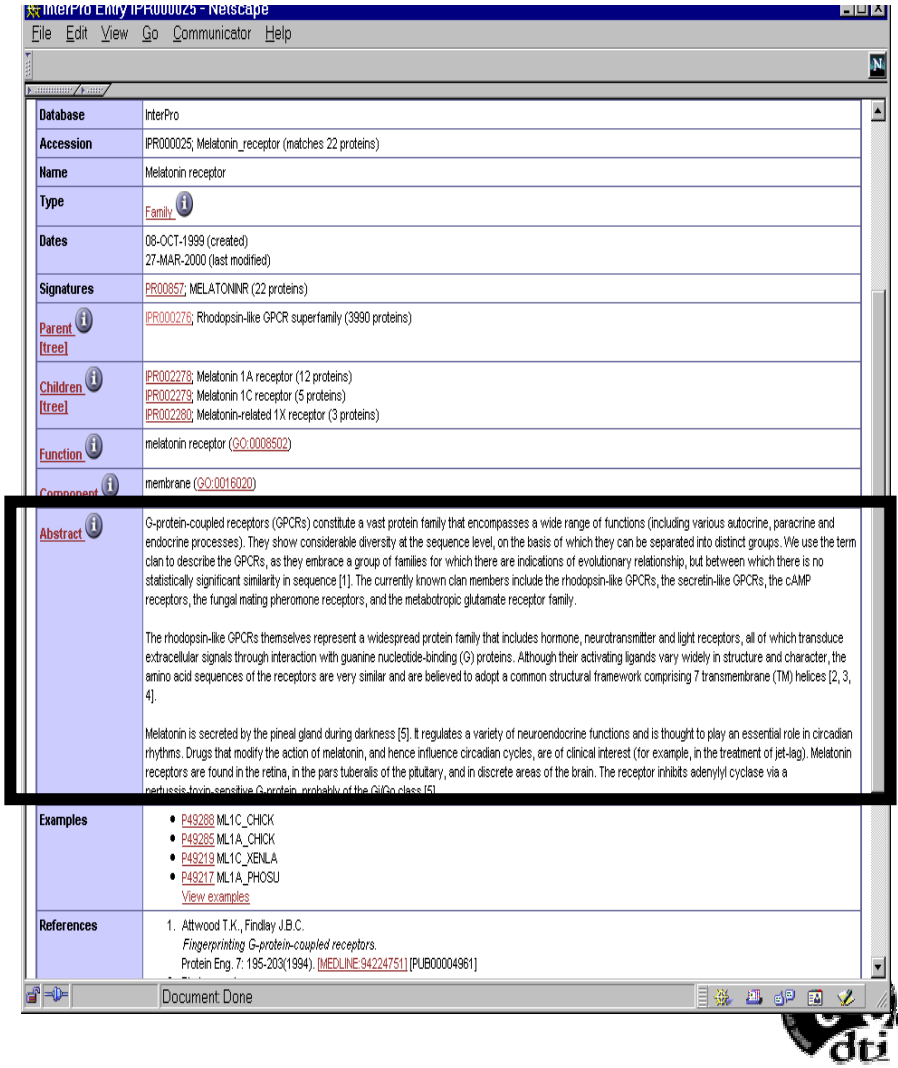
- EBI hosted tools 50+
  - Homology & Similarity
  - Prot. Function. Analysis
  - Structural Analysis
  - Sequence Analysis
  - Miscellaneous Tools
- EBI hosted databases 30+
  - Nucleotide Databases
  - Protein Databases
  - Proteome Analysis
  - Structure Databases
  - Microarray Database
  - Literature Databases

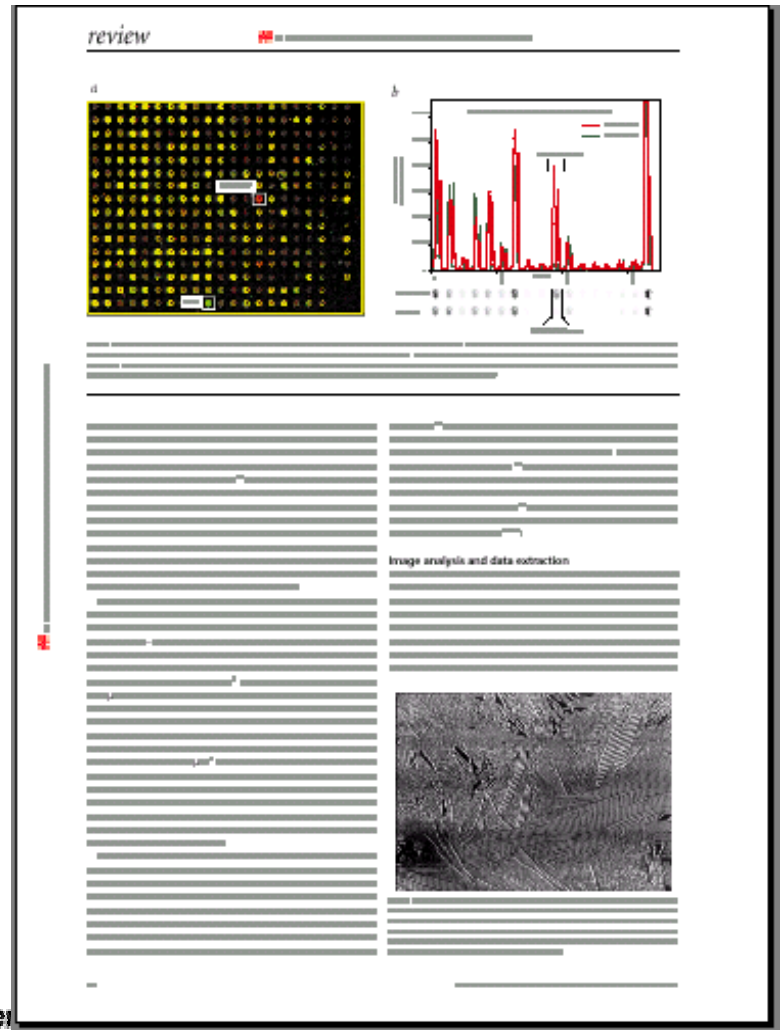
GRID OUTREACH

Problems: access mechanisms

Courtesy of Mark Wilkinson (BioMOBY)

Problems: data isn't just numbers

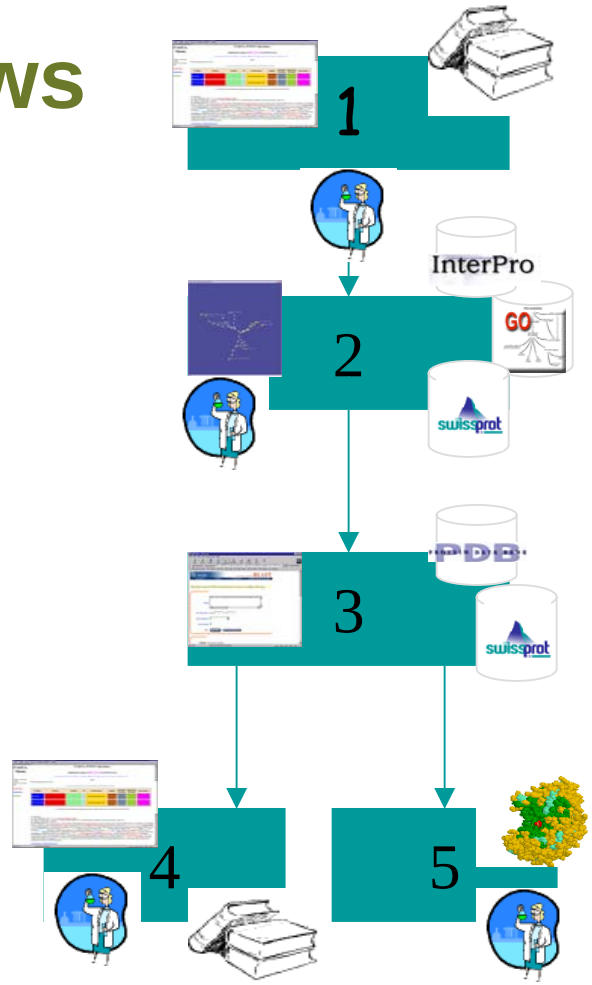




Problems: in-silico experiments are workflows

Circadian Rhythms

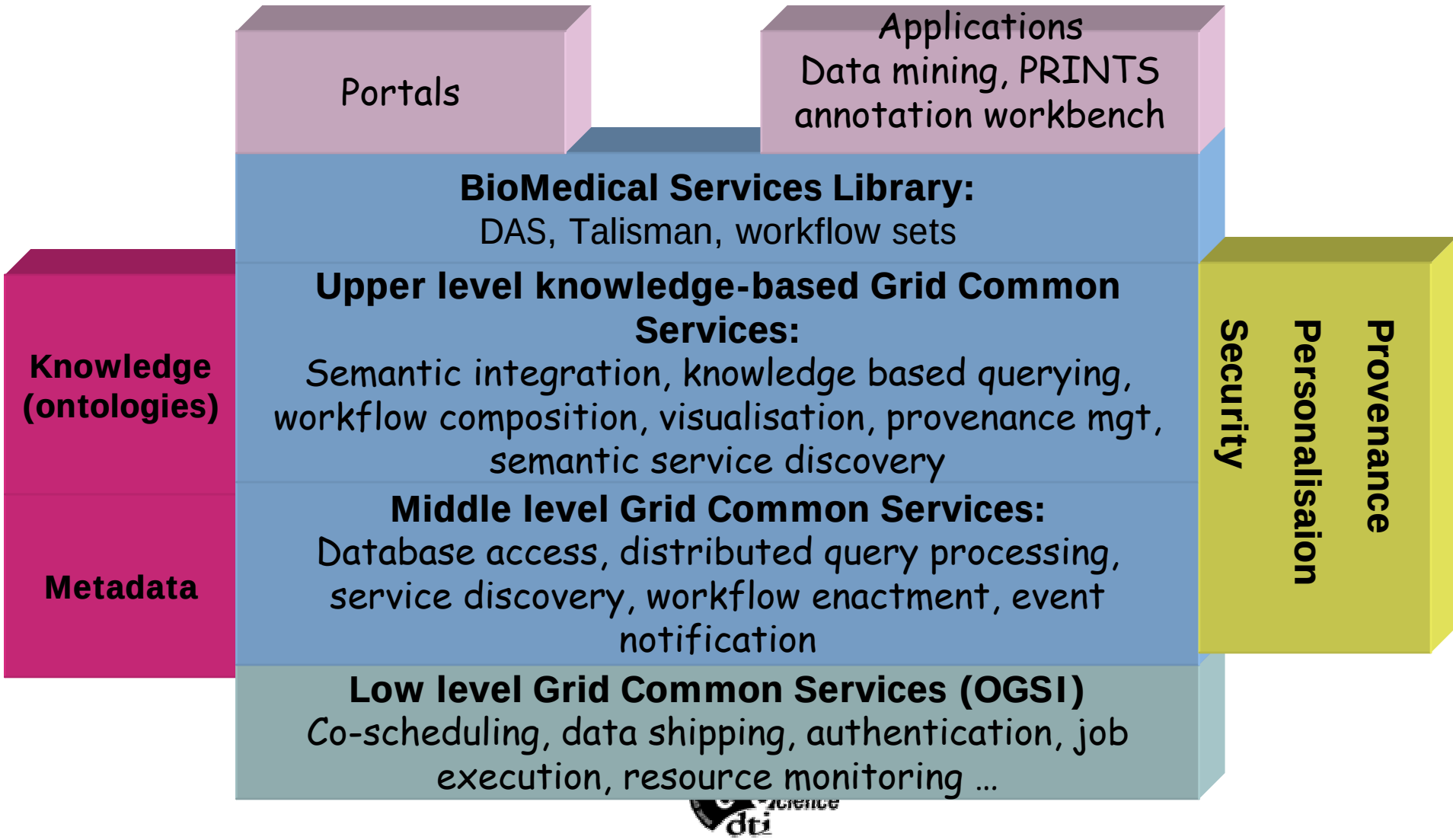
- Has anyone else studied the effect of neurotransmitters on the circadian rhythms in Drosophila?
- I've got a cluster of proteins from my experiment. How do their functions interrelate? And what are the proteins with a particular function?
- Is a structure known for my protein? What other proteins have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

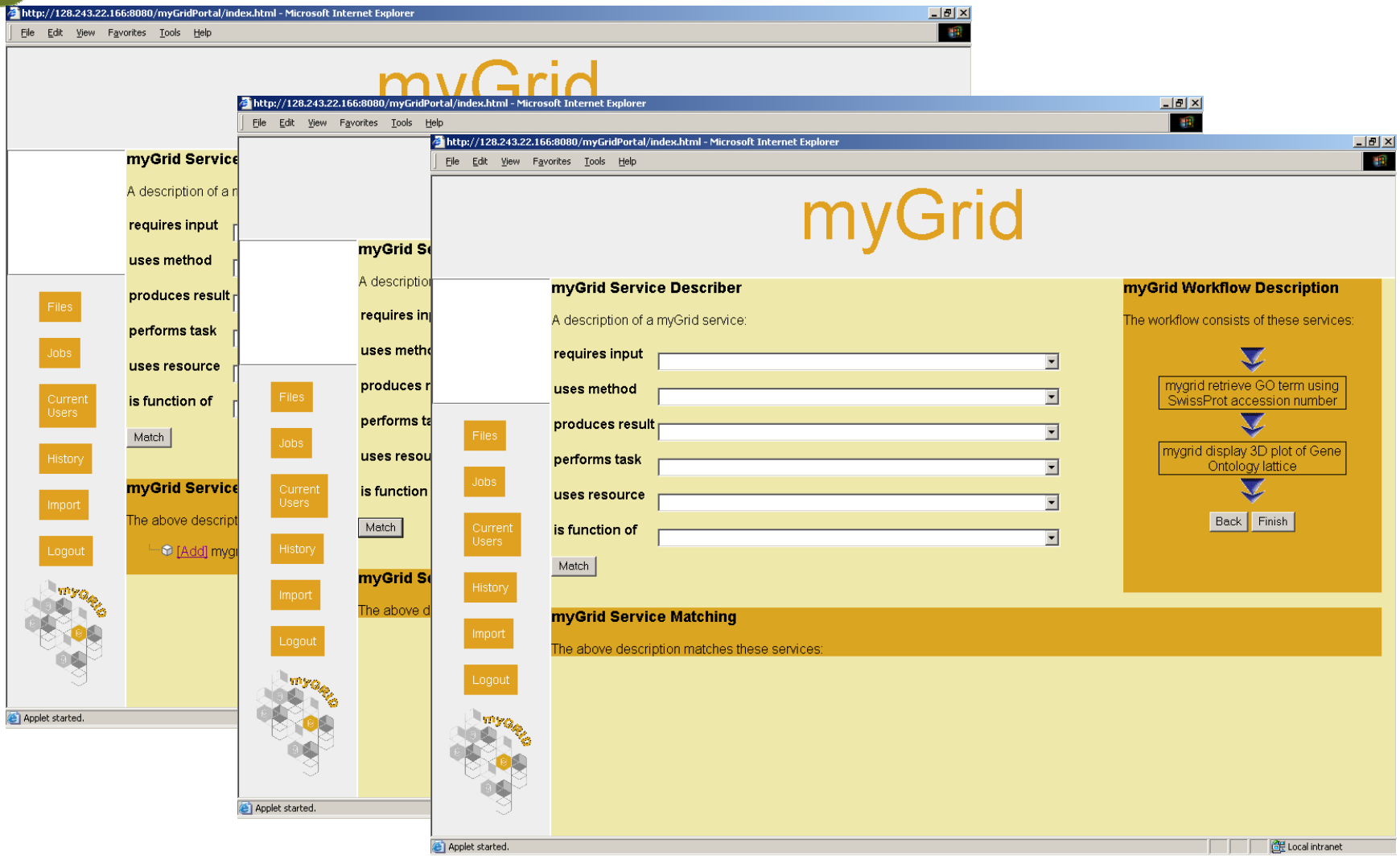
myGrid approach: marketecture diagram



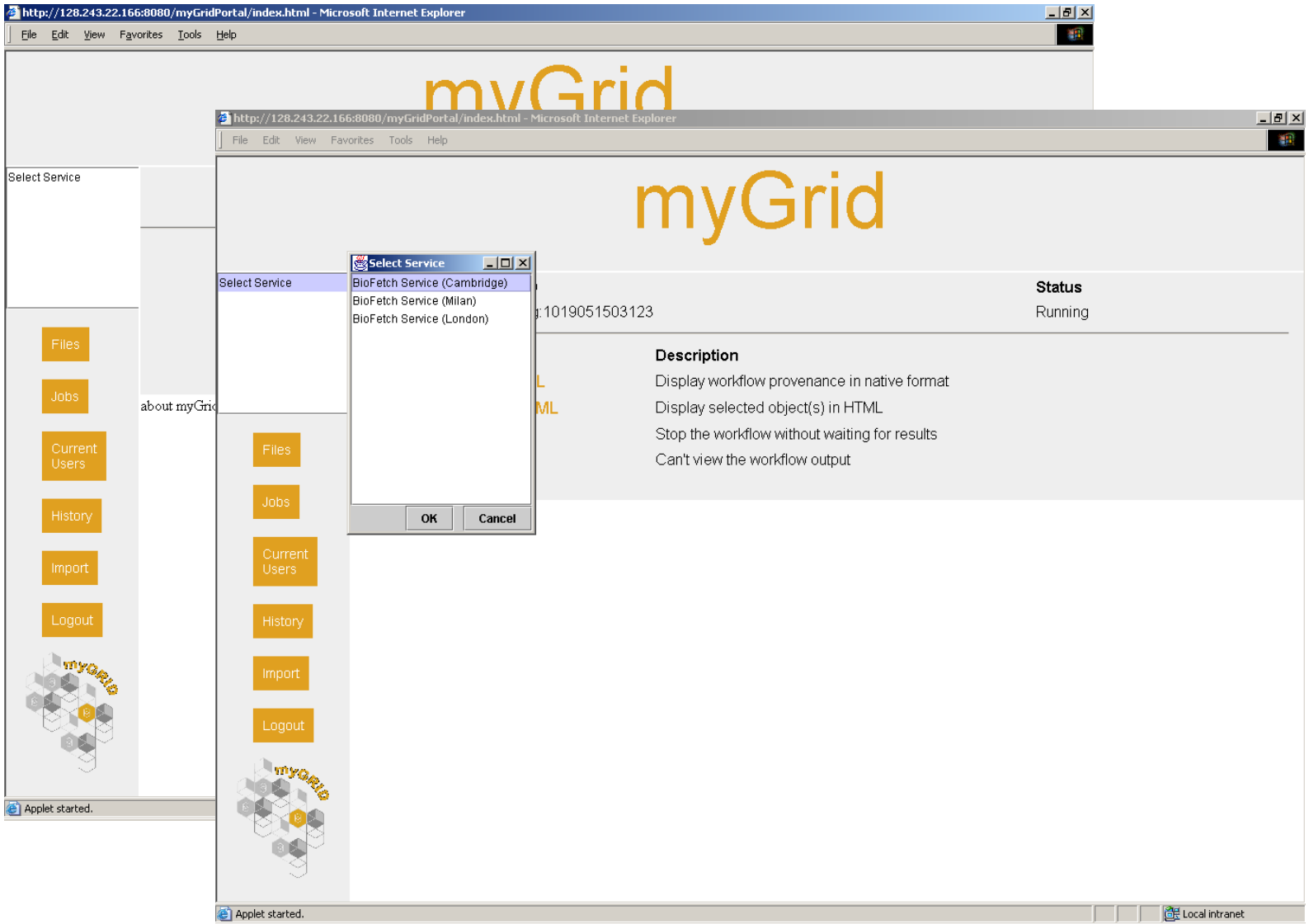
myGrid approach: modular set of services

- 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

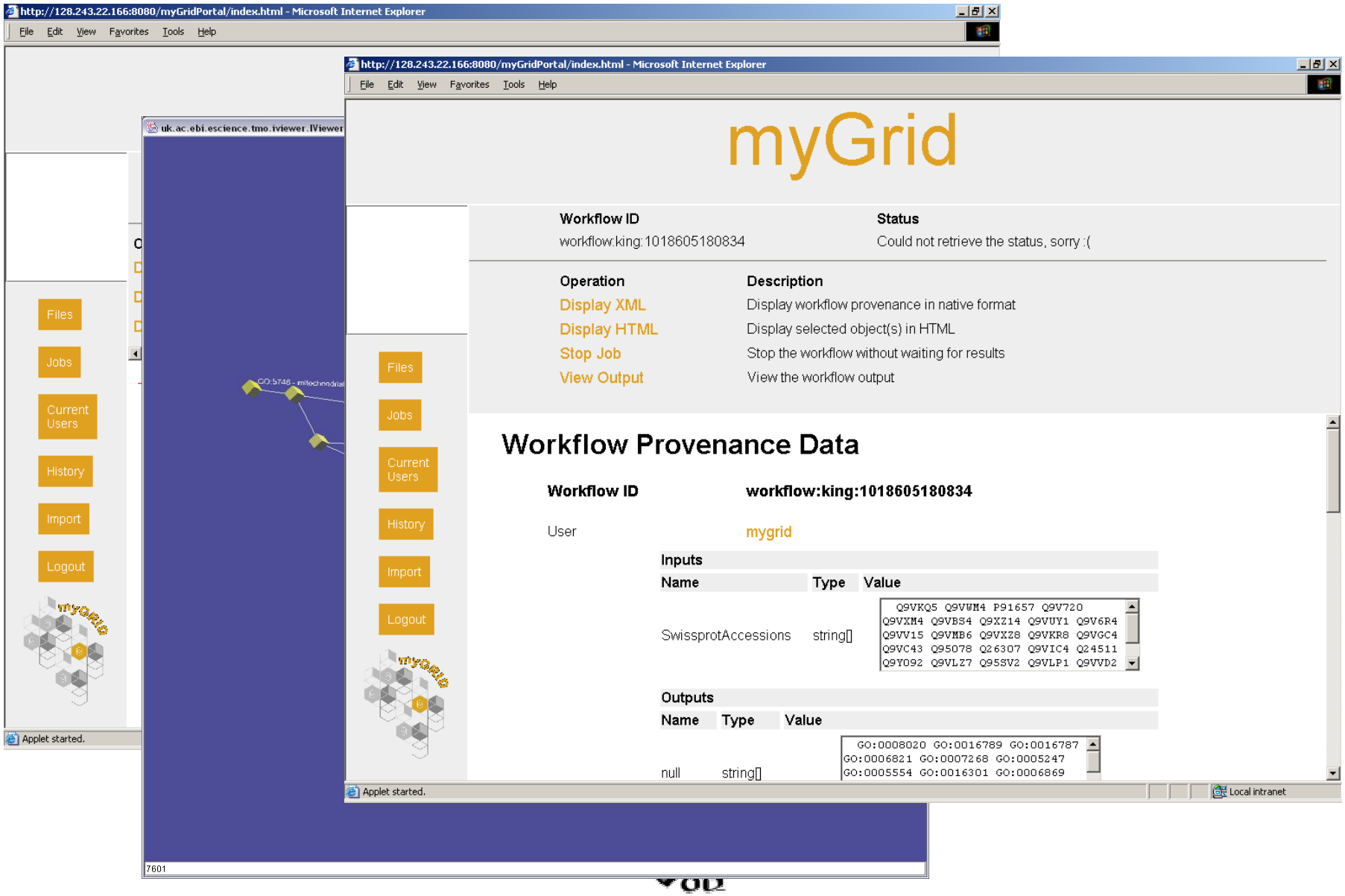
Example: describe the workflow



Example: execute the workflow



Example: view results

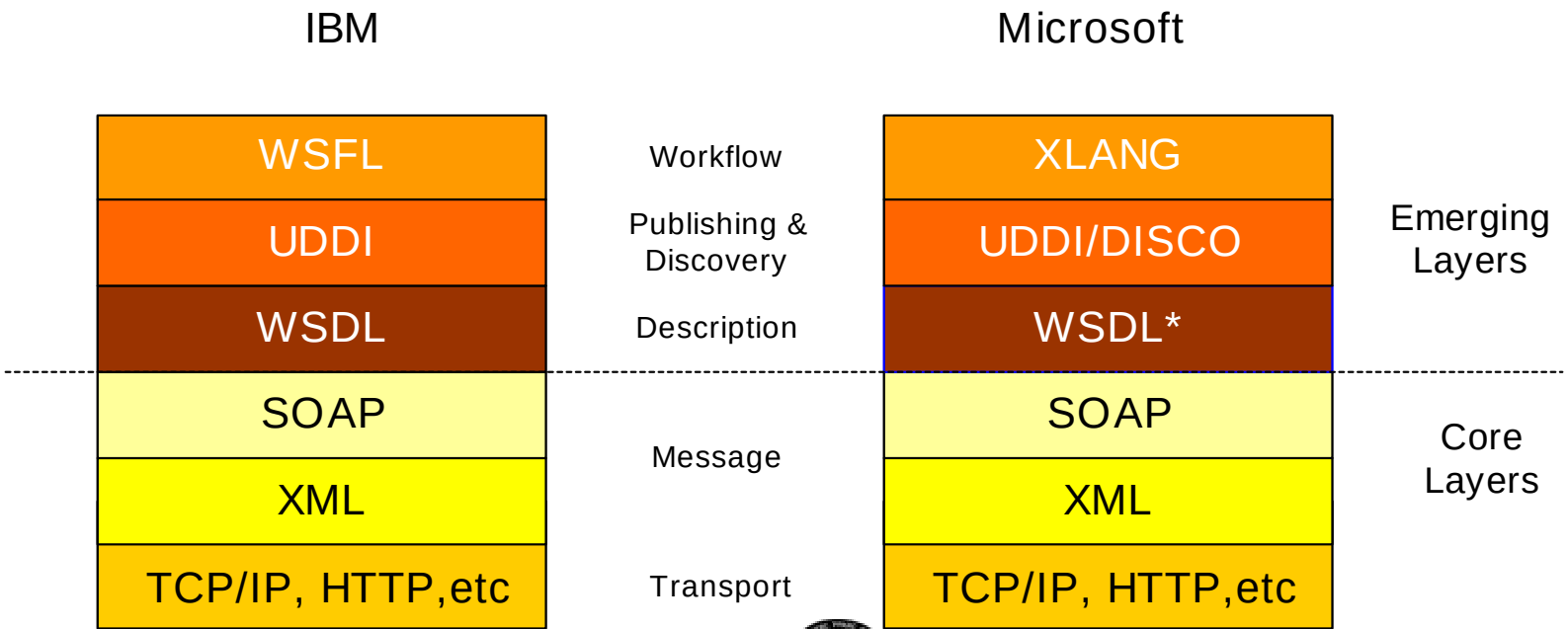


Extensions to example

- 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

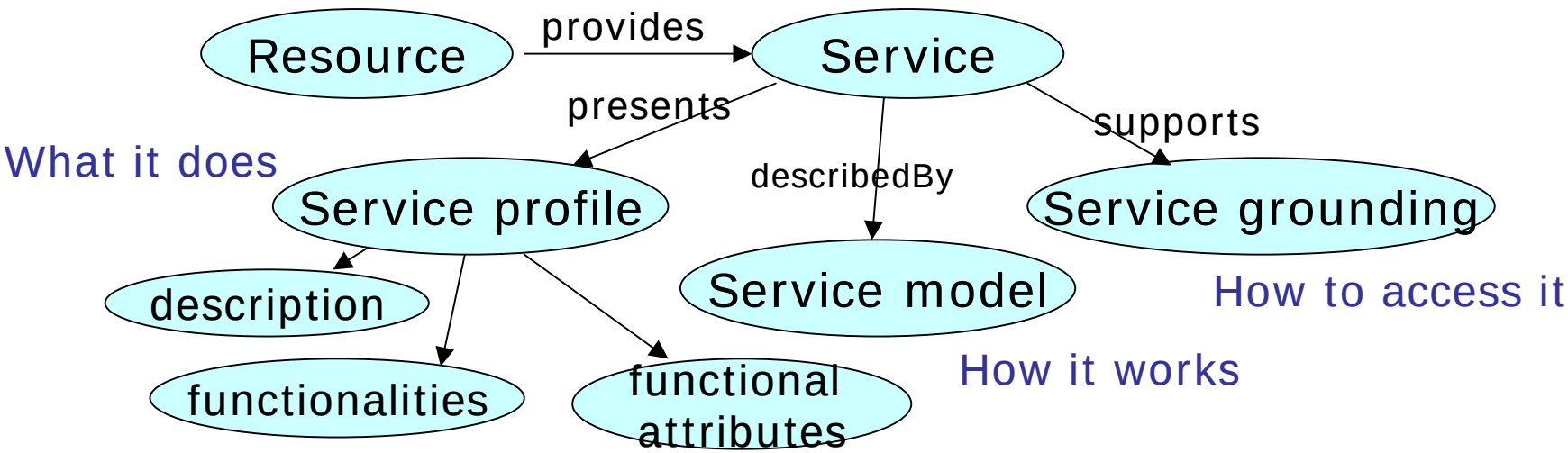


Semantics is the key to interoperability

- 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

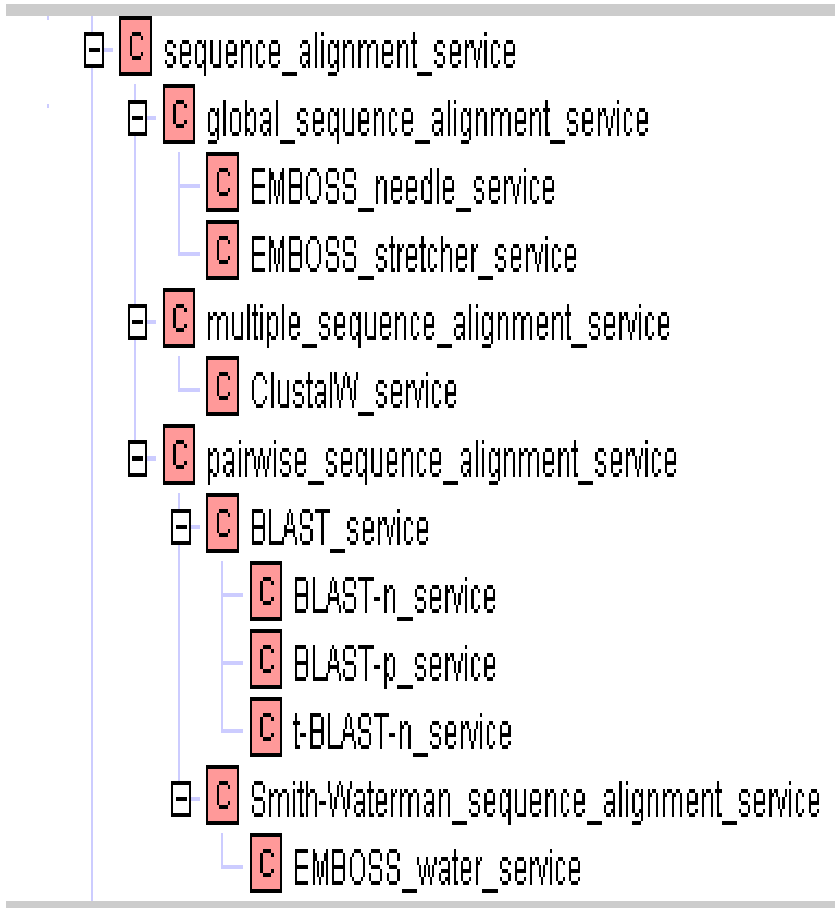
   BLASTn is a tool for computing sequence homology that uses the BLAST algorithm over nucleotides

3. Instance description of specific services

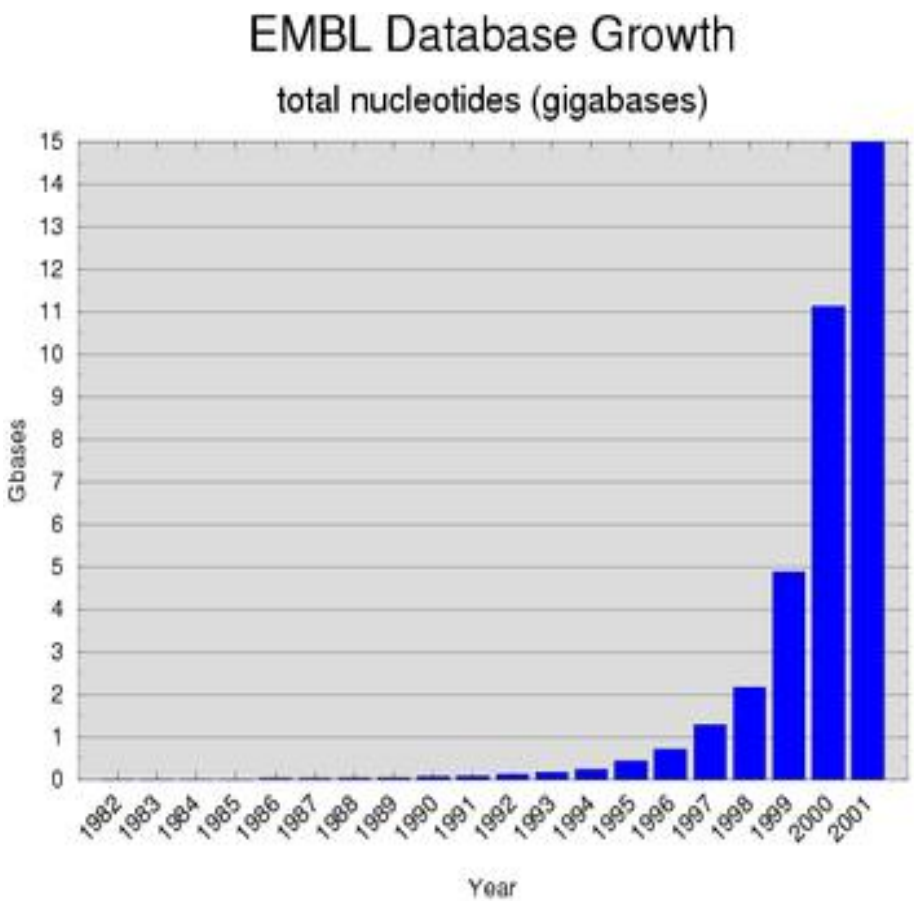
   BLASTn service is provided by EBI

4. Description of invoked instance of a service

   BLAST as executed for a particular workflow

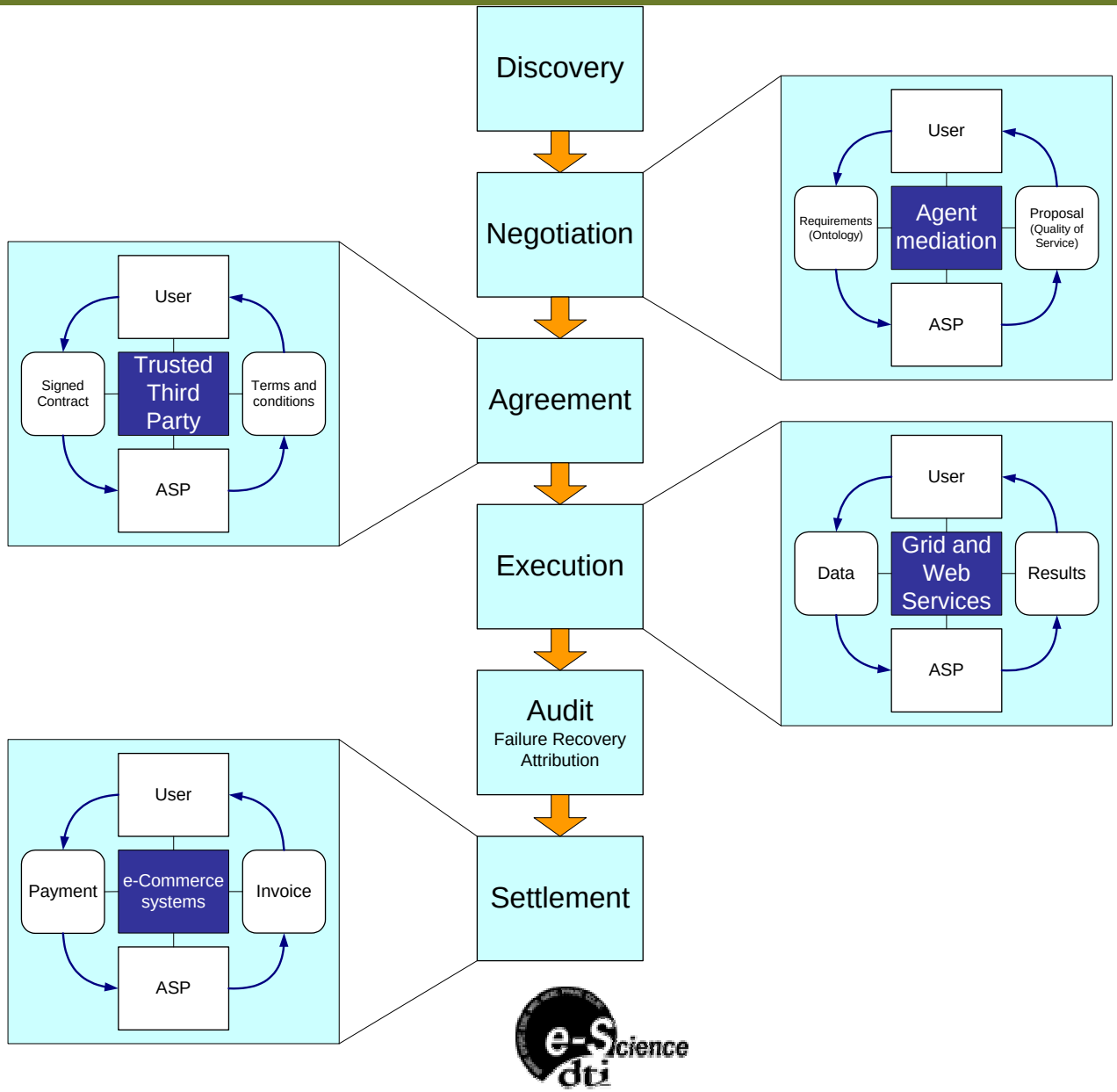


Underpinning technologies: Grid



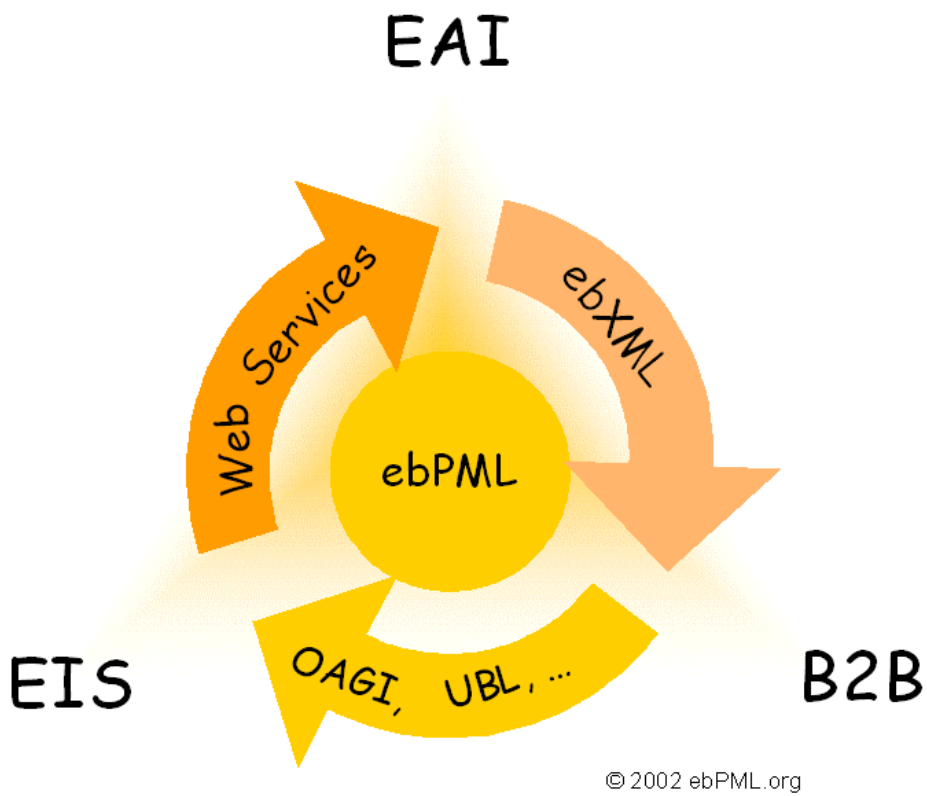
- Can't continue to centralise all resources in one place
  - Public service providers
  - Commercial users
- Large data volumes
  - Transfers and staging
  - Distributed query
- Computationally intensive
  - Resource management

Processes for outsourcing

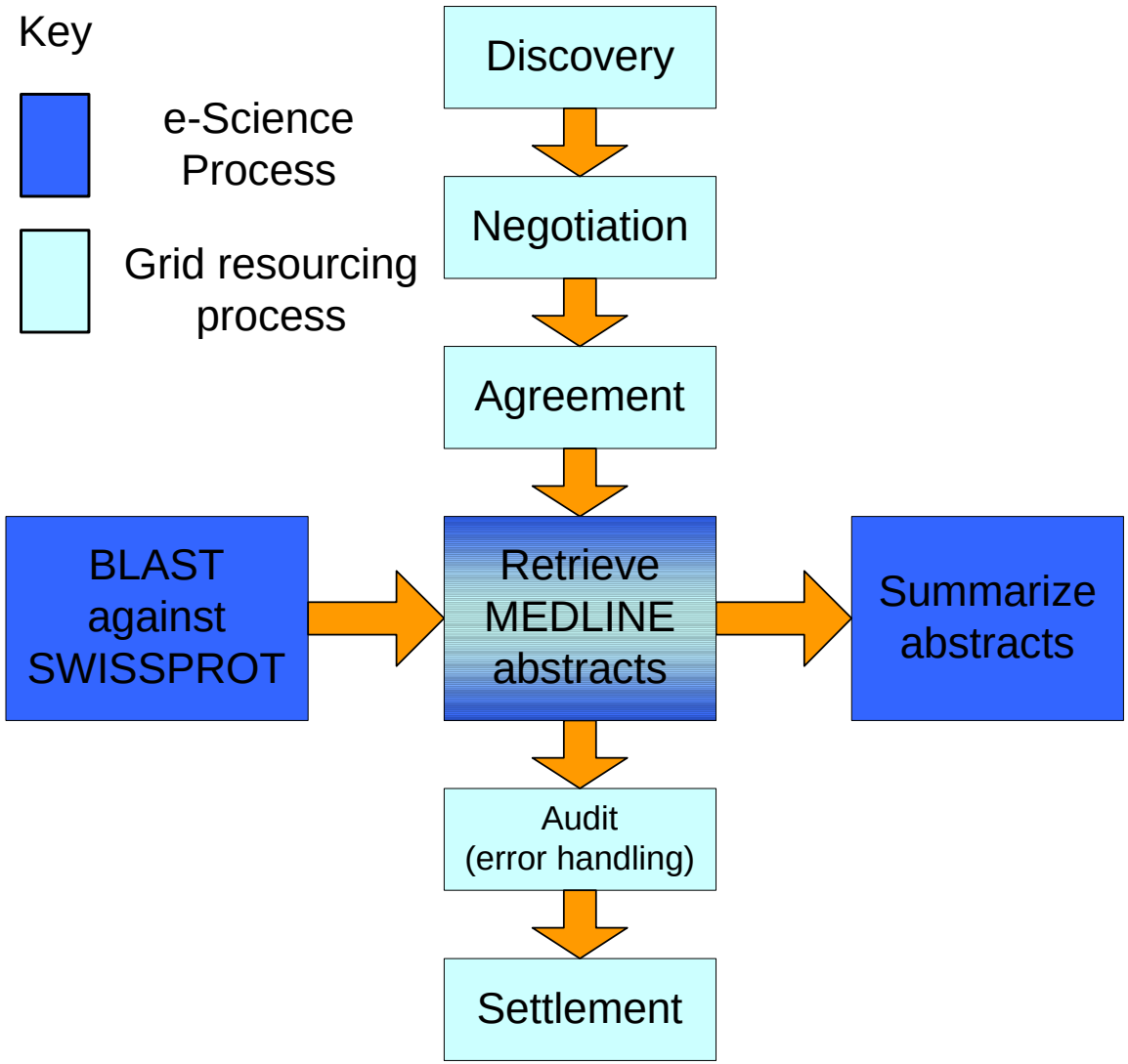


Underpinning technologies: workflow

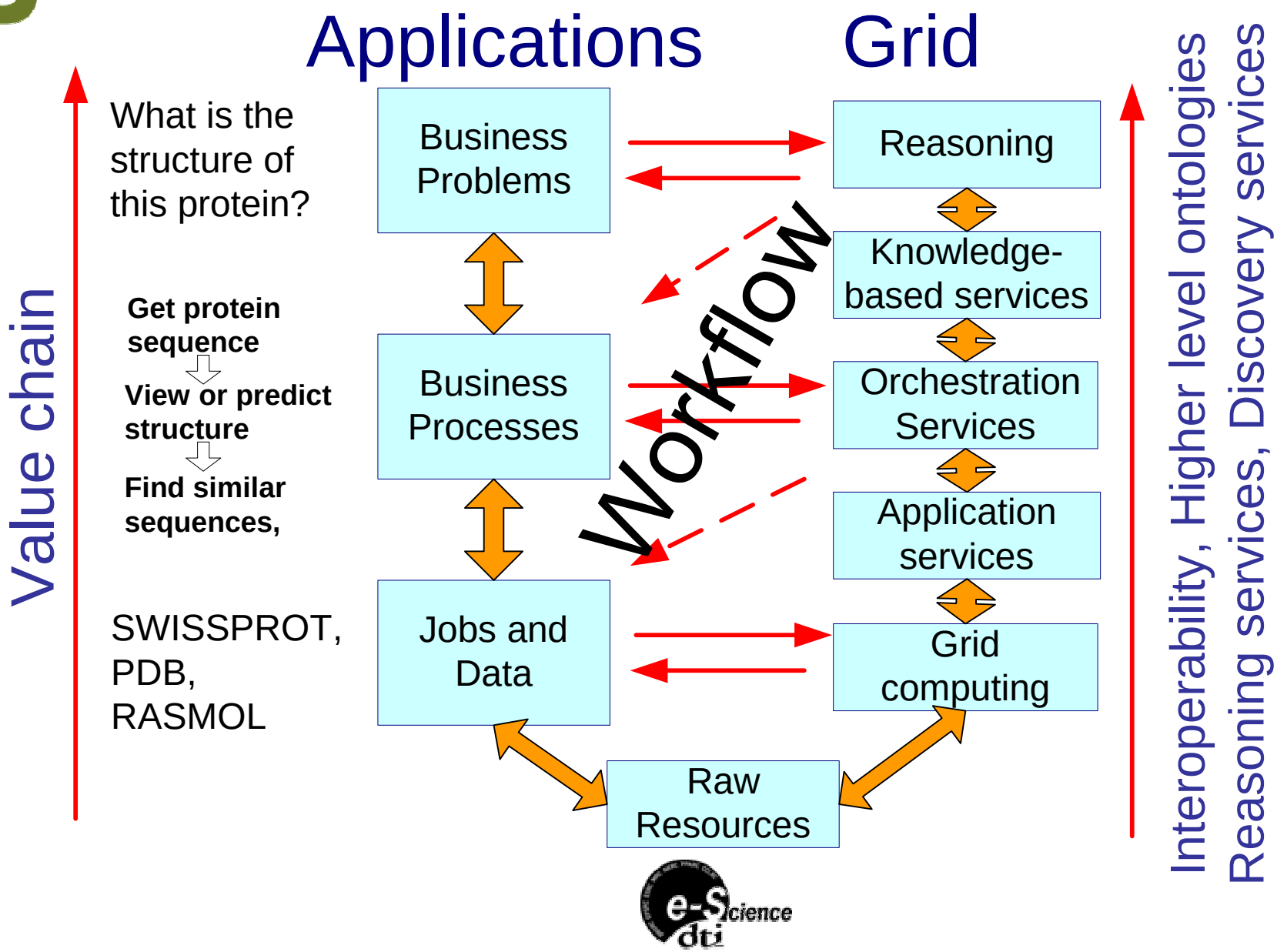
Workflow Management Coallition	WfMC
OMG	EDOC UML
Web Services	Xlang WSFL BPEL4WS BPML



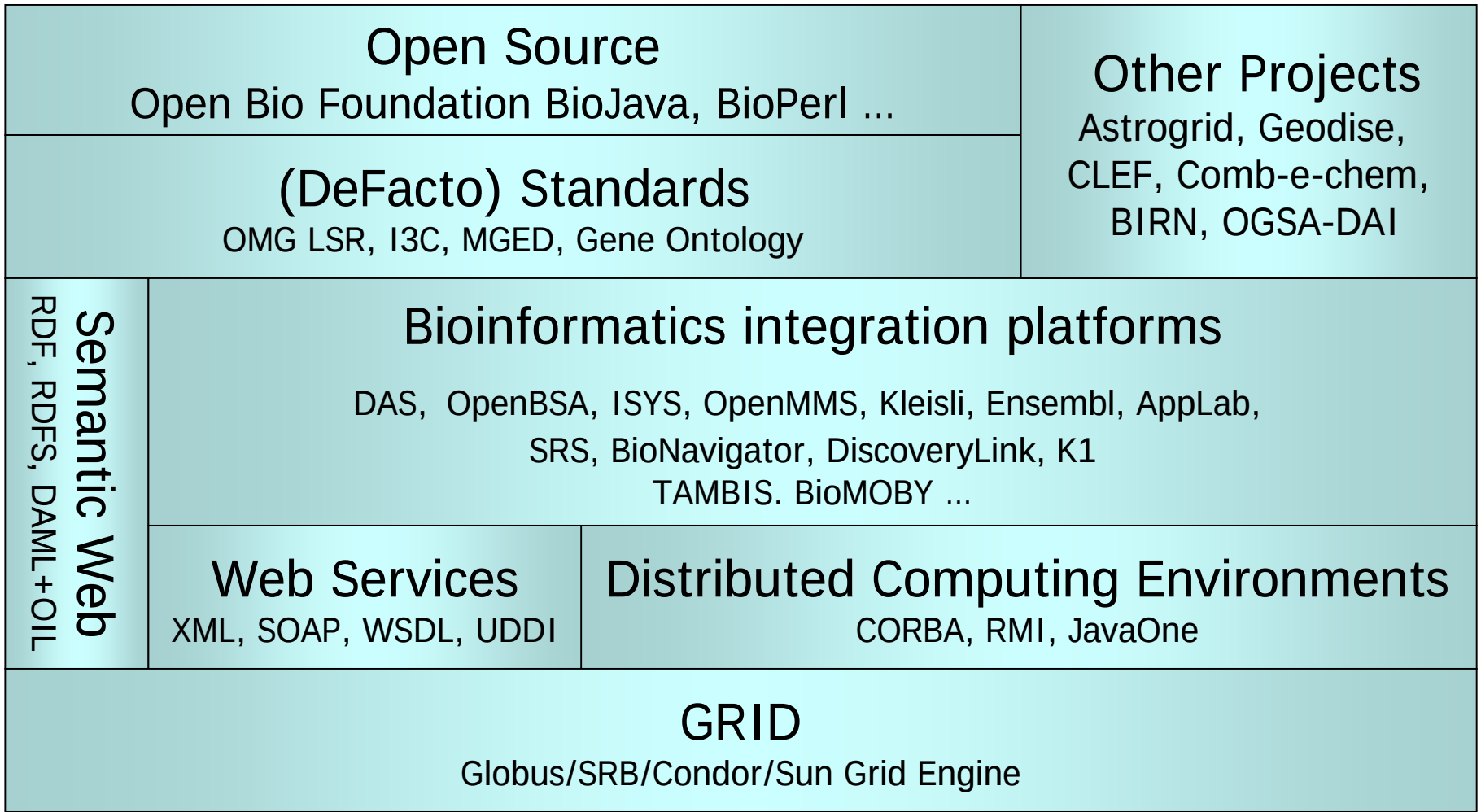
Combining workflow and the grid



Bringing it all together



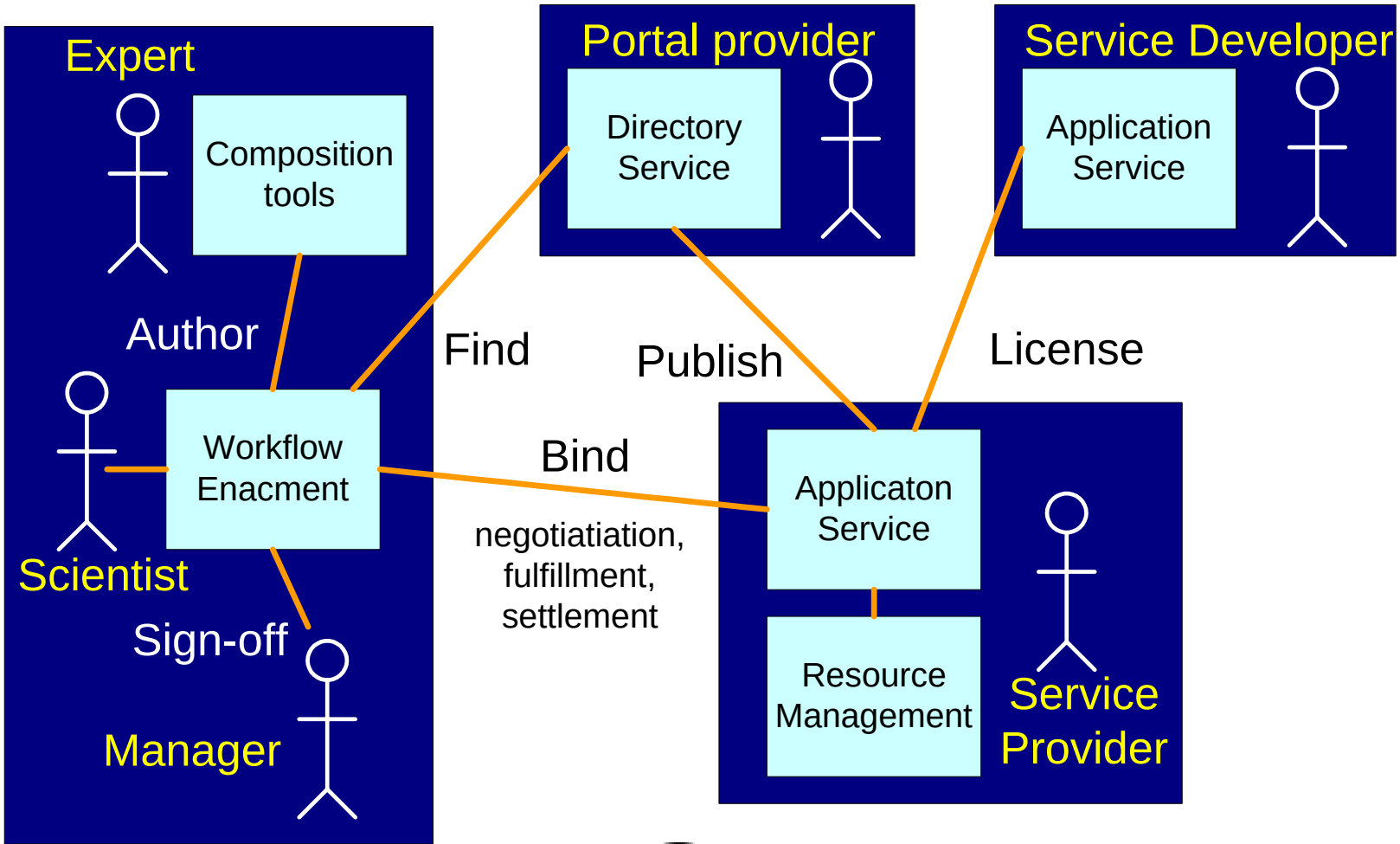
Next steps (1)


Open Source
Open Bio Foundation BioJava, BioPerl ...
Other Projects
Astrogrid, Geodise, CLEF, Comb-e-chem, BIRN, OGSA-DAI
(DeFacto) Standards
OMG LSR, I3C, MGED, Gene Ontology
Semantic Web
RDF, RDFS, DAML+OIL
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)


Expert
Composition tools
Author
Workflow Enacment
Scientist
Sign-off
Manager
Portal provider
Directory Service
Service Developer
Application Service
Find
Publish
License
Bind
negotiatiation, fulfillment, settlement
Applicaton Service
Resource Management
Service Provider

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
- IT Innovation
  - Grid Projects http://www.it-innovation.soton.ac.uk/grid
  - Matthew Addis mja@it-innovation.soton.ac.uk