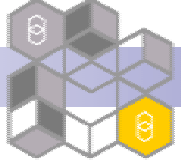


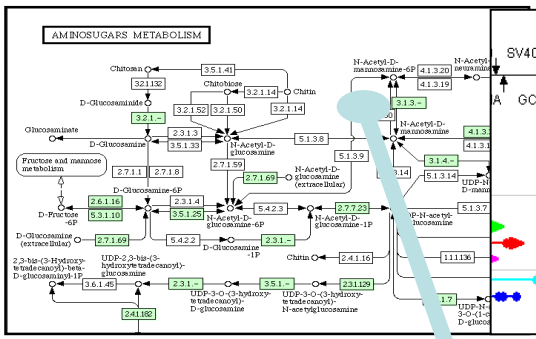
Treating “shimantic web” syndrome with ontologies

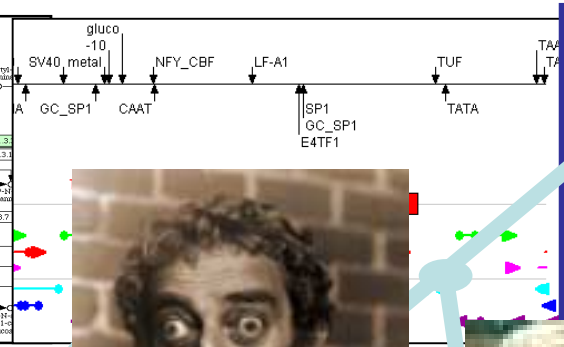
AKT workshop on Semantic Web Services

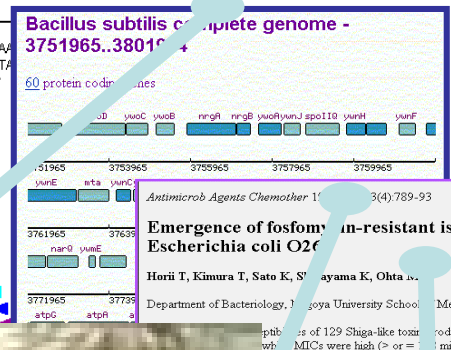
Duncan Hull, Robert Stevens, Phillip Lord, Chris Wroe and Carole Goble

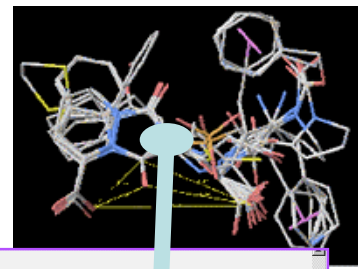
University of Manchester

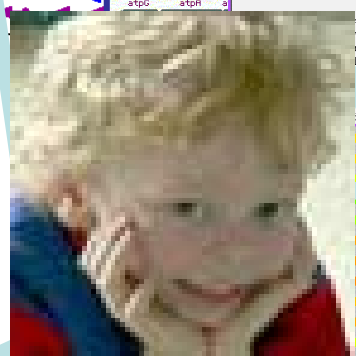
Data (Type) Intensive Bioinformatics

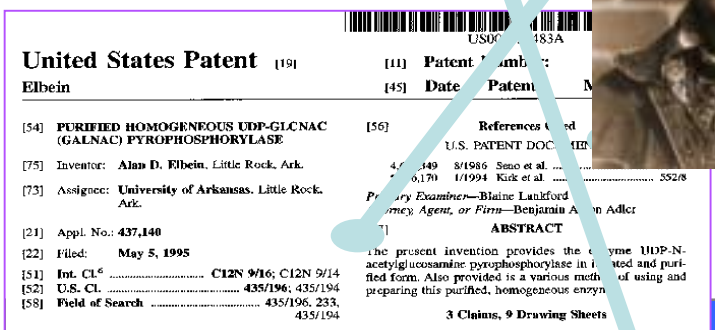


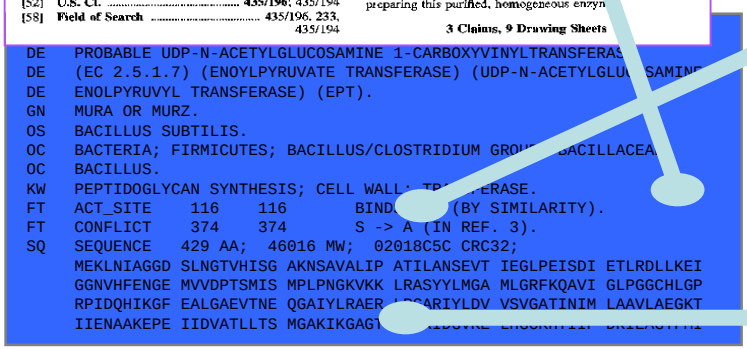


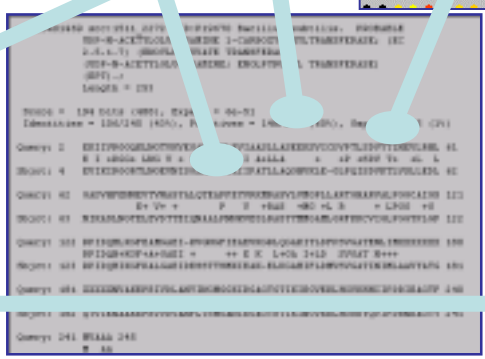


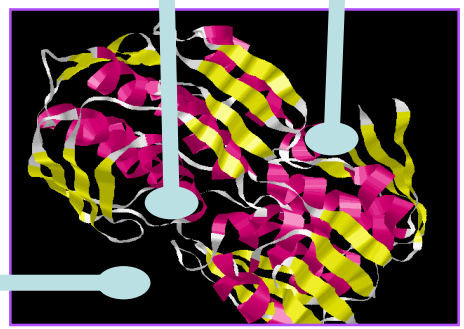


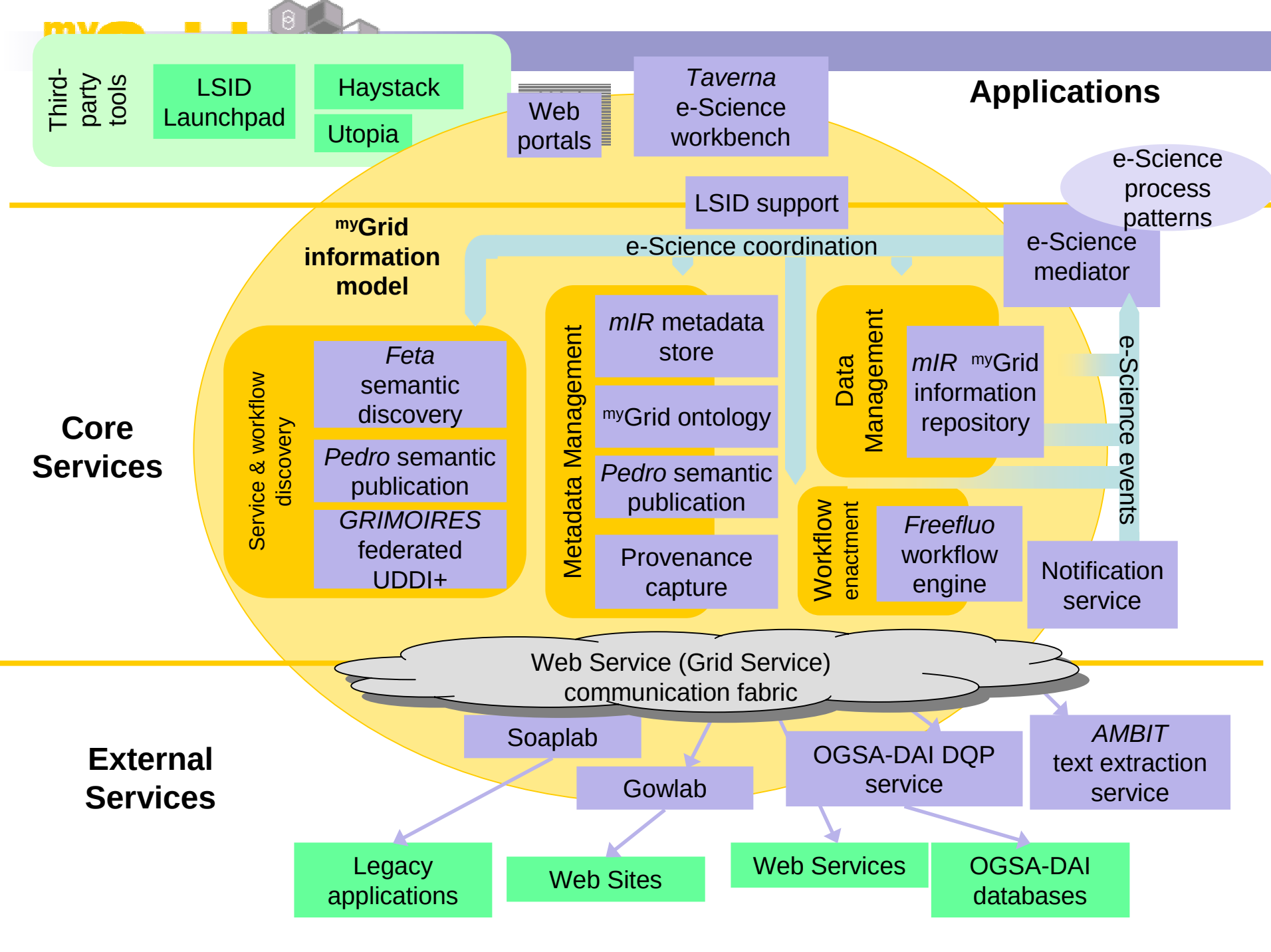






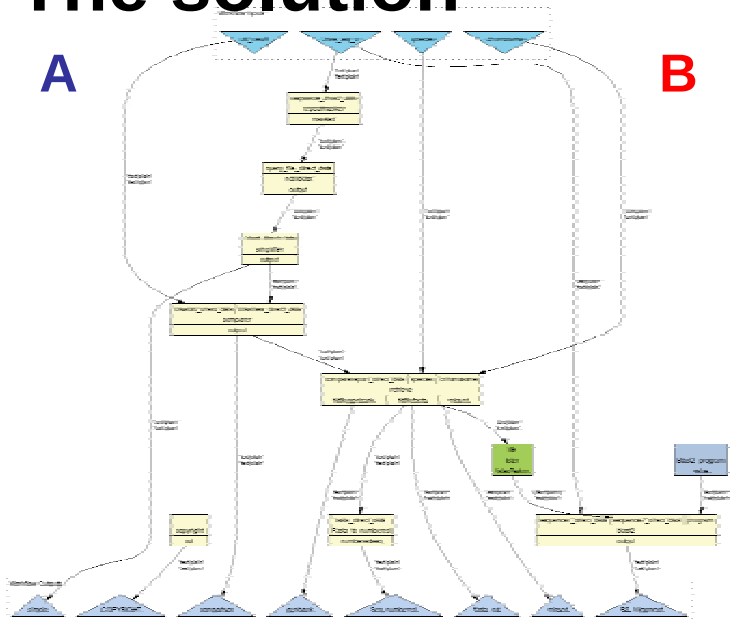


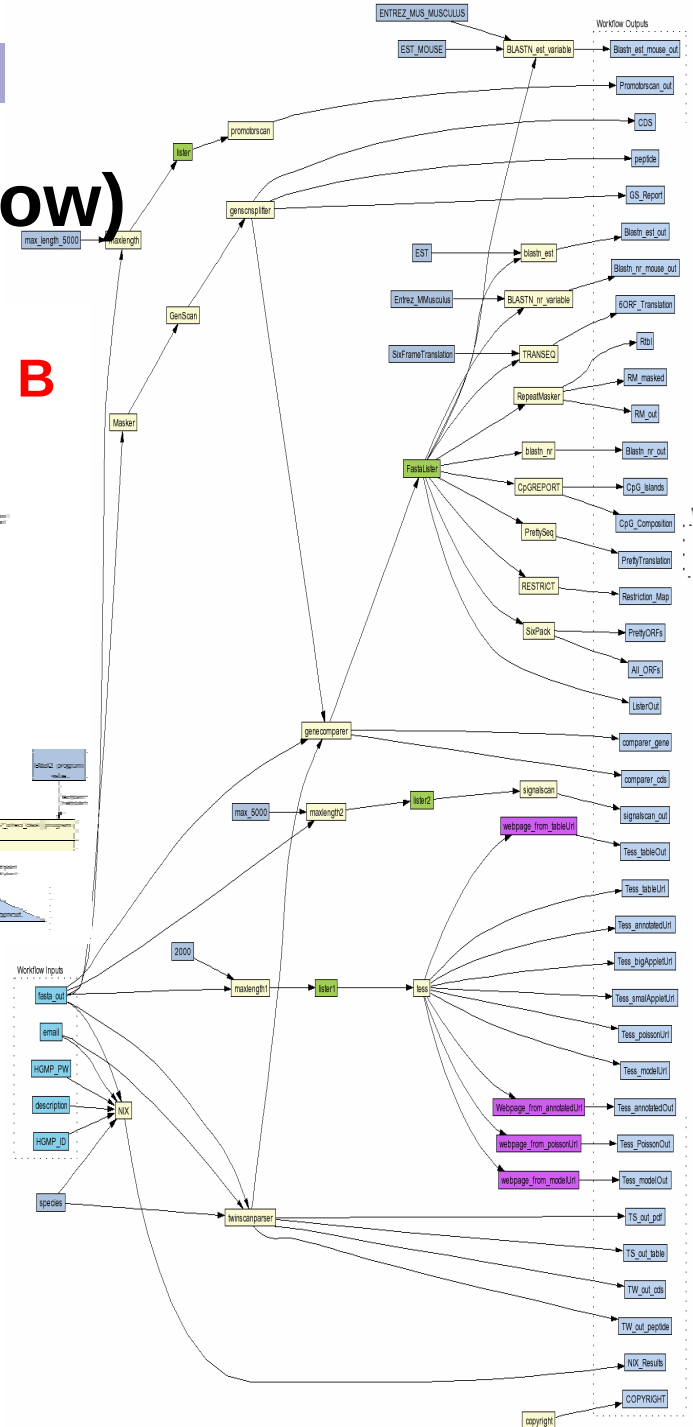


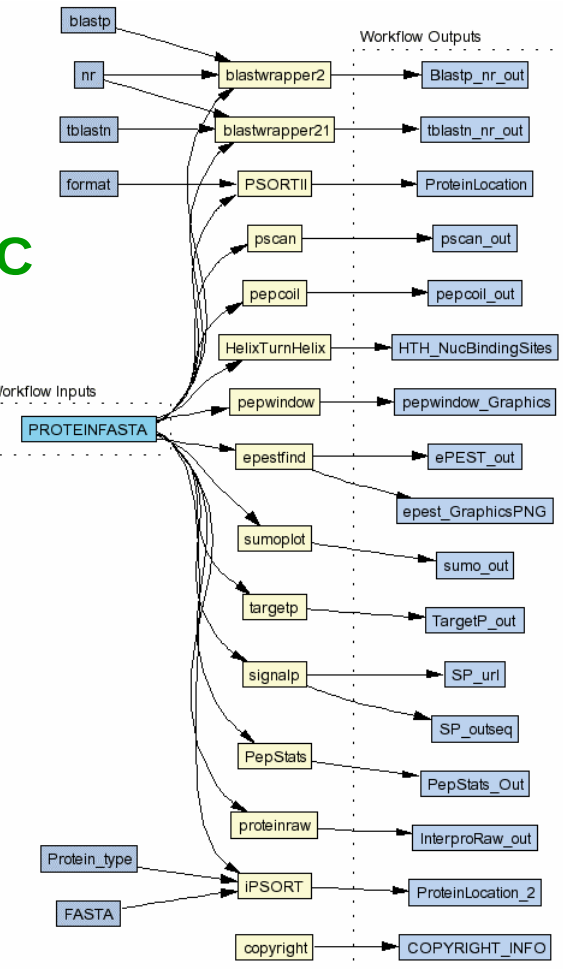

Third-party tools
LSID Launchpad
Haystack
Utopia
Web portals
Taverna e-Science workbench
Applications
e-Science process patterns
LSID support
e-Science coordination
e-Science mediator
myGrid information model
Core Services
Service & workflow discovery
Feta semantic discovery
Pedro semantic publication
GRIMOIRES federated UDDI+
Metadata Management
mIR metadata store
myGrid ontology
Pedro semantic publication
Provenance capture
Data Management
mIR myGrid information repository
e-Science events
Workflow enactment
Freefluo workflow engine
Notification service
Web Service (Grid Service) communication fabric
External Services
Soaplab
Gowlab
OGSA-DAI DQP service
AMBIT text extraction service
Legacy applications
Web Sites
Web Services
OGSA-DAI databases

Workflows (dataflow)

The solution








```
ID   Q99NI3      PRELIMINARY;
AC   Q99NI3;
DT   01-JUN-2001 (TrEMBLrel. 17, Created)
DT   01-JUN-2001 (TrEMBLrel. 17, Last sequence update)
```

unique identifiers

We don't (always) have XML

This is

`xsd:string`!

```
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=C57BL/6J;
RX   PubMed=15100712; DOI=10.1038/sj.ejhg.5201174;
RA   Tipney H.J., Hinsley T.A., Brass A., Metcalfe K., Donnai D.,
RA   Tassabehji M.;
RT   "Isolation and characterisation of GTF2IRD2, a novel fusion gene and
RT   member of the TFII-I family of transcription factors, deleted in
RT   Williams-Beuren syndrome.";
RL   Eur. J. Hum. Genet. 12:551-560(2004).
RN   [2]
RP   SEQUENCE FROM N.A.
RA   Bayarsaihan D.;
RL   Submitted (MAY-2002) to the EMBL/GenBank/DDBJ databases.
DR   EMBL; AY014963; AAG41674.1; -.
DR   EMBL; AY116023; AAM48282.1; -.
DR   MGD; MGI:2149780; 1700012P16Rik.
DR   InterPro; IPR008938; ARM.
DR   InterPro; IPR004212; GTF2I.
DR   Pfam; PF02946; GTF2I; 2.
SQ   SEQUENCE   936 AA;  104577 MW;  28CB2994C0ABB1A9 CRC64;
     MAQVAVTTQP TDEPSDGRMV VTFLMSALES MCKELAKSKA EVACIAVYET DVYVVGTERG
     CAFVNARQDL QKDFAQHCQG EGLPEEKPLC LGNGEACPGE AQLLRRAVQD HFCLCYRKAL
     GTTAMVPVPY EQMLQDEAAV VVRGLPEGLA FQHPDNYSLA TLKWILENKA GISFAVKRPF
```

Gene name

Protein sequence

…no type system either

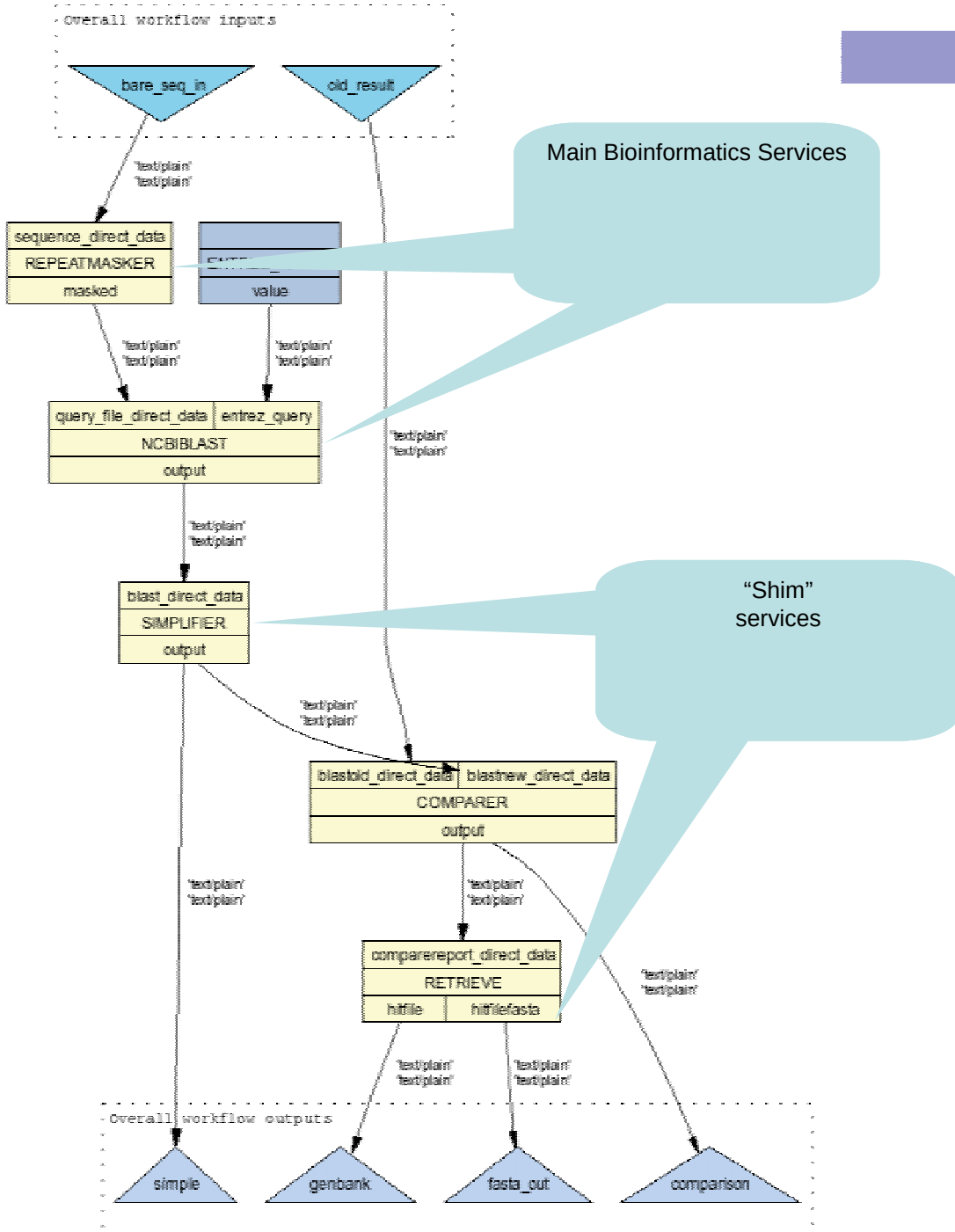
- To add insult to injury, we don't have a type system either
  - Probably never will do
  - Getting people to describe their data types using a global model (schema or ontology) is *hard*
- Especially when there is no obvious immediate benefit of doing so (see www.BioMOBY.org)
- myGrid accommodates 3rd party Web Services because bioinformatics is an *open* world

Web Services don't quite fit together

- Many bioinformatics Web Services *nearly* fit together
- This is where shims come into play
  - Thin strips of metal used to align pipes or rails
  - Software components that transform between closely related data (either syntactically or semantically)

Shims

- ~10 / 30 Web Services are shims
- ready made WS
- homemade WS wrap legacy applications



Problems with shim solution

- Finding the right shim at the right time
  - Resource intensive: knowledge, time etc
  - Shims re-invent the wheel
    "Every biologist has written a BLAST parser"
  - For *n* services there are n^2 shims
  - In myGrid 600 services, therefore
    worst case scenario: 360,000 shims
- Like it or not, we're stuck with shims
- What do we do about it?

Research goals

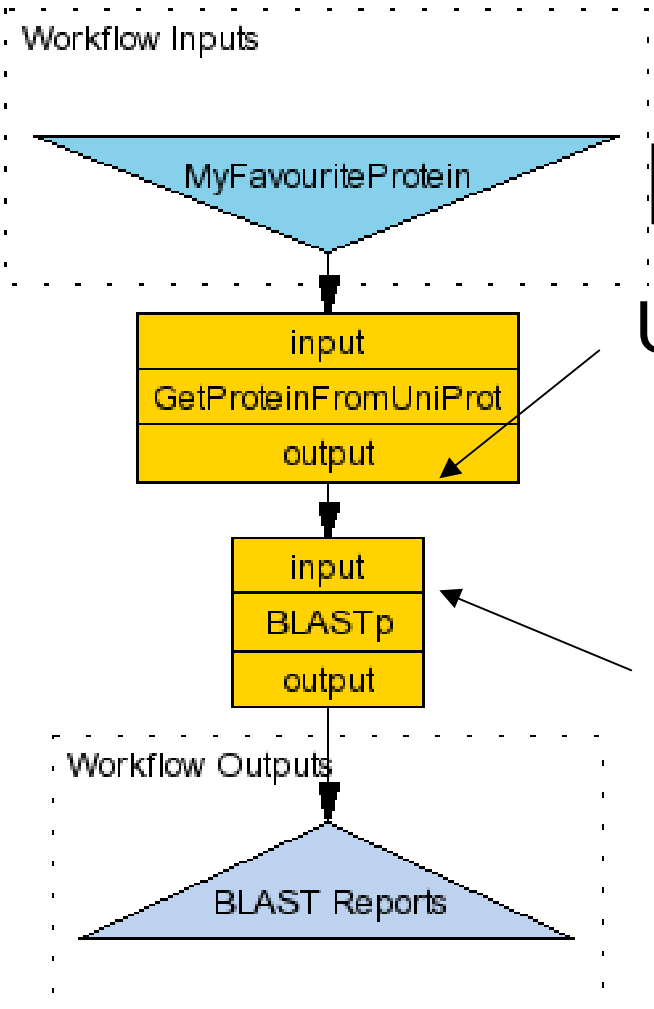
- Characterise the shims that are out there
- Classify and describe shims (and services they mediate between) using an ontology
- Create a library (not a factory) of shims
  - Reuse don't rewrite
- Model identifies mismatches
- Facilitate automation? It's not always safe…
- …or at least support and guide user selection of shims during workflow construction

Shim description and classification

Shim type	Input	Operation	Output	Description
Dereferencer	Identifier or Pointer	Dereference	The dereferenced resource	GenBank ID replaced with GenBank record
Syntax translator	Data represented in concrete representation, x	Translate	Data represented in an alternate concrete representation, y	SeqRet translates between representations of sequence data.
Semantic translator	DNA sequence	Translate	Protein sequence	Translate DNA into protein
Mapper	Unique identifier	Map	Unique Identifier	Maps between IDs. E.g. GenBank to EMBL
Parser	Record	Parse	Abstract syntax tree	Parse BLAST report.
Iterator	A set	Iterate	A member of a set	Iterate over members of a given set
Comparer	Two or more sets	Diff	Set of differences	Comparing BLAST reports notifies of new sequences
Accessor	Record	Access	Subset of record	Access a subset

Table 1. Examples of shims taken from the *my*Grid project. This partial classification is based on inputs, outputs and the task or operation performed.

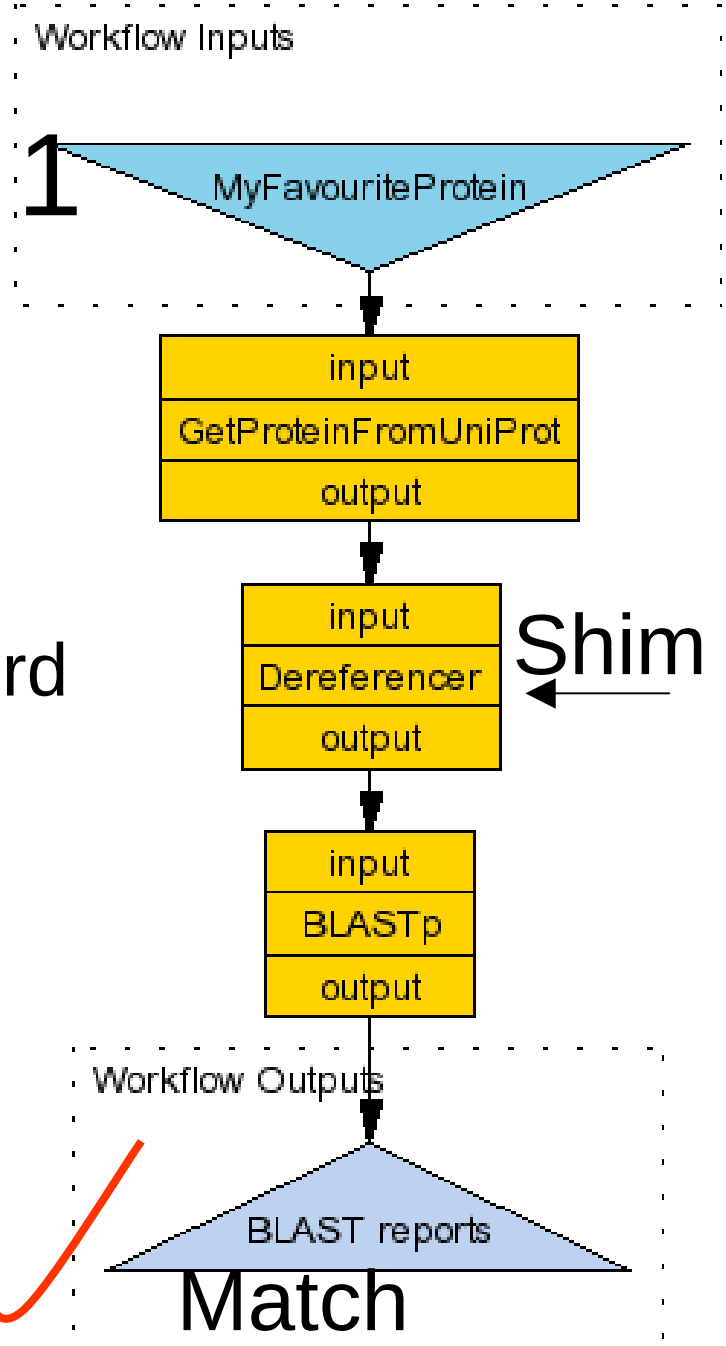
Example 1



UniProt_ID

identifies

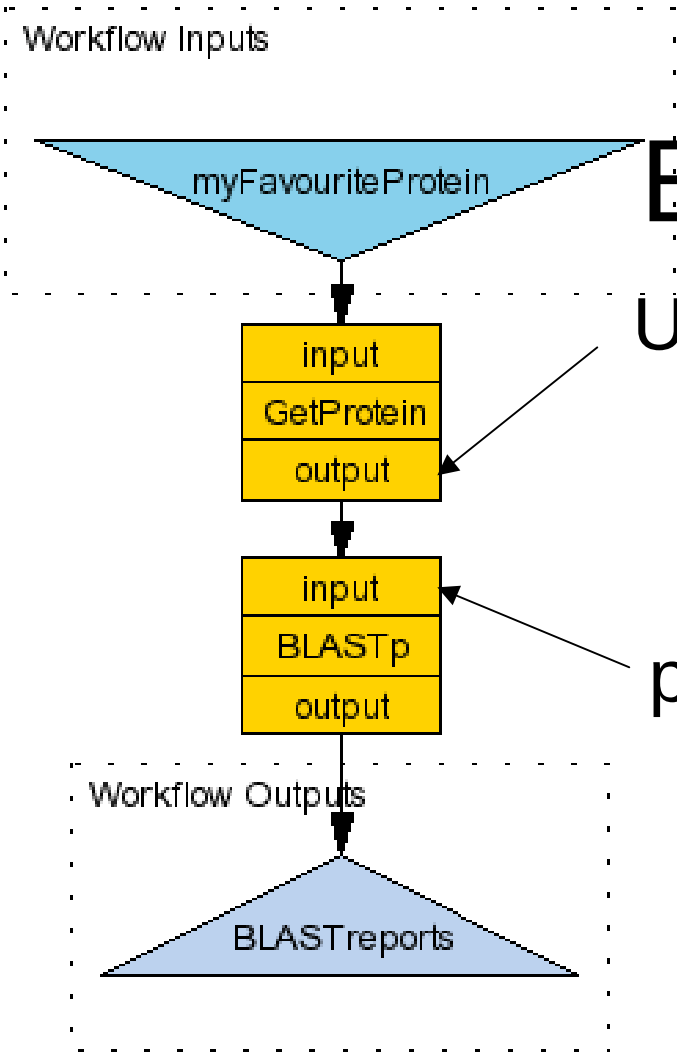
UniProt_Record



Mismatch

Match

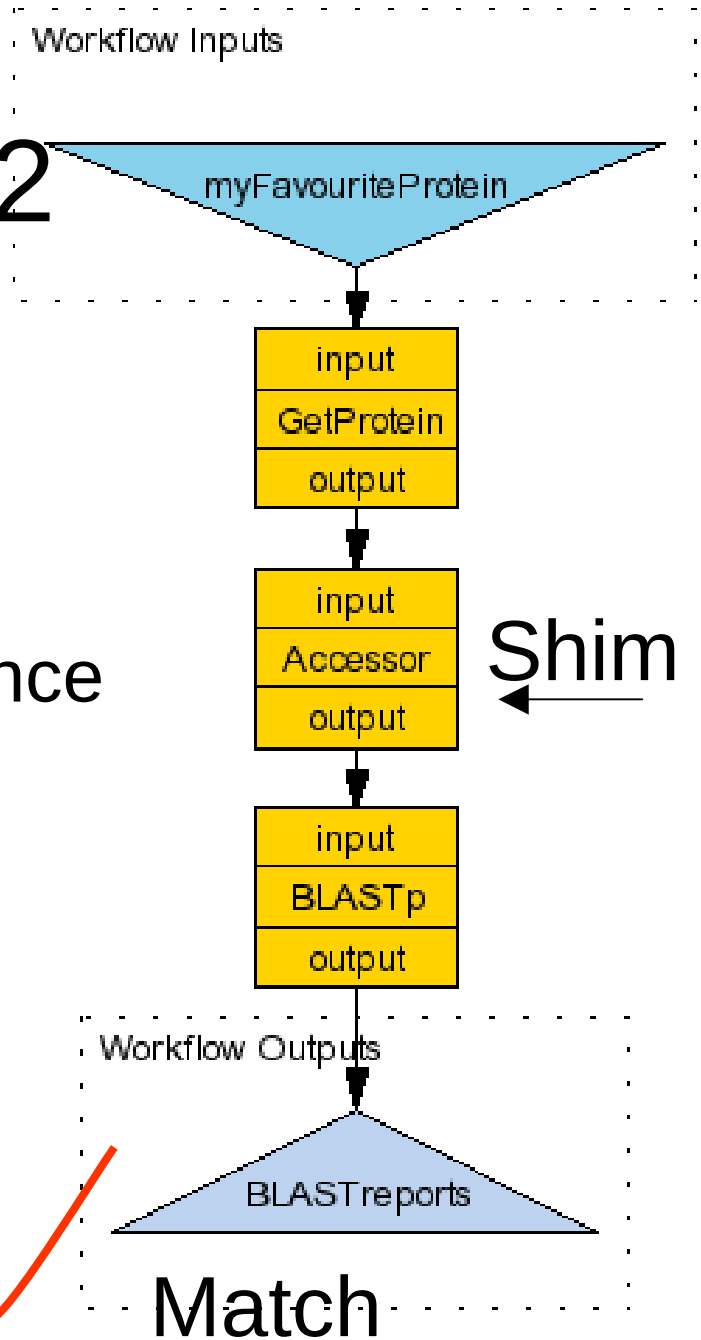
Example 2



UniProt record

hasPart

protein_sequence



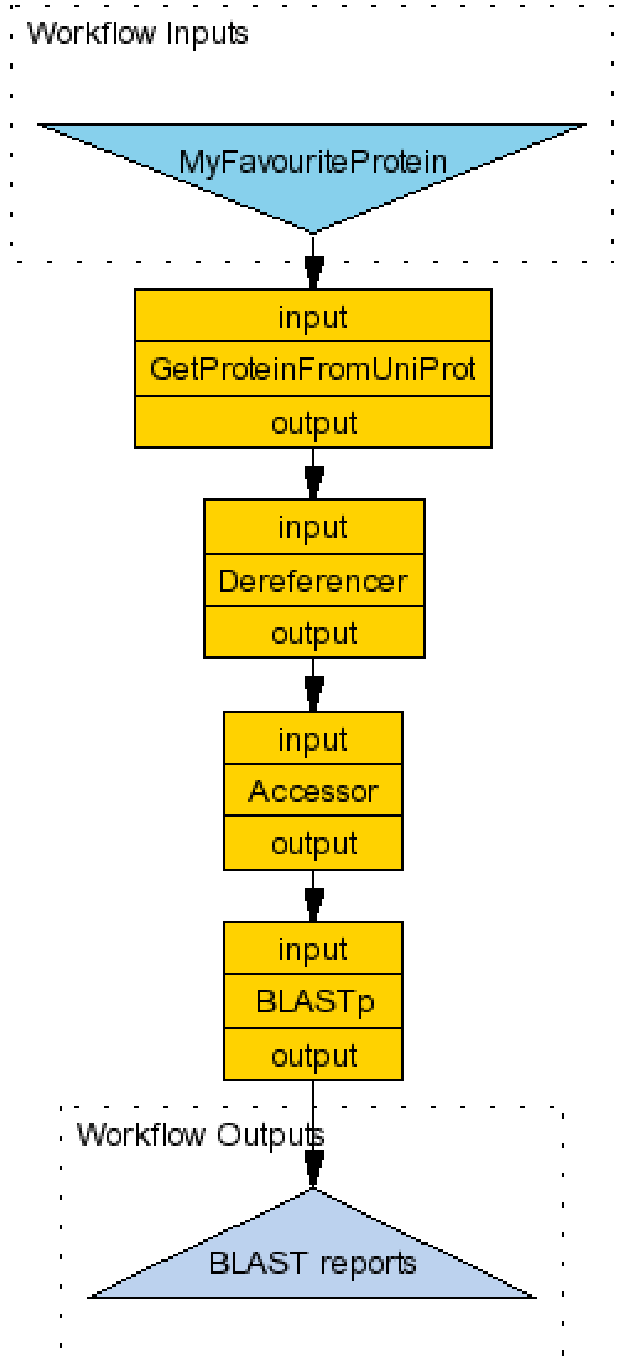
Shim

Mismatch

Match

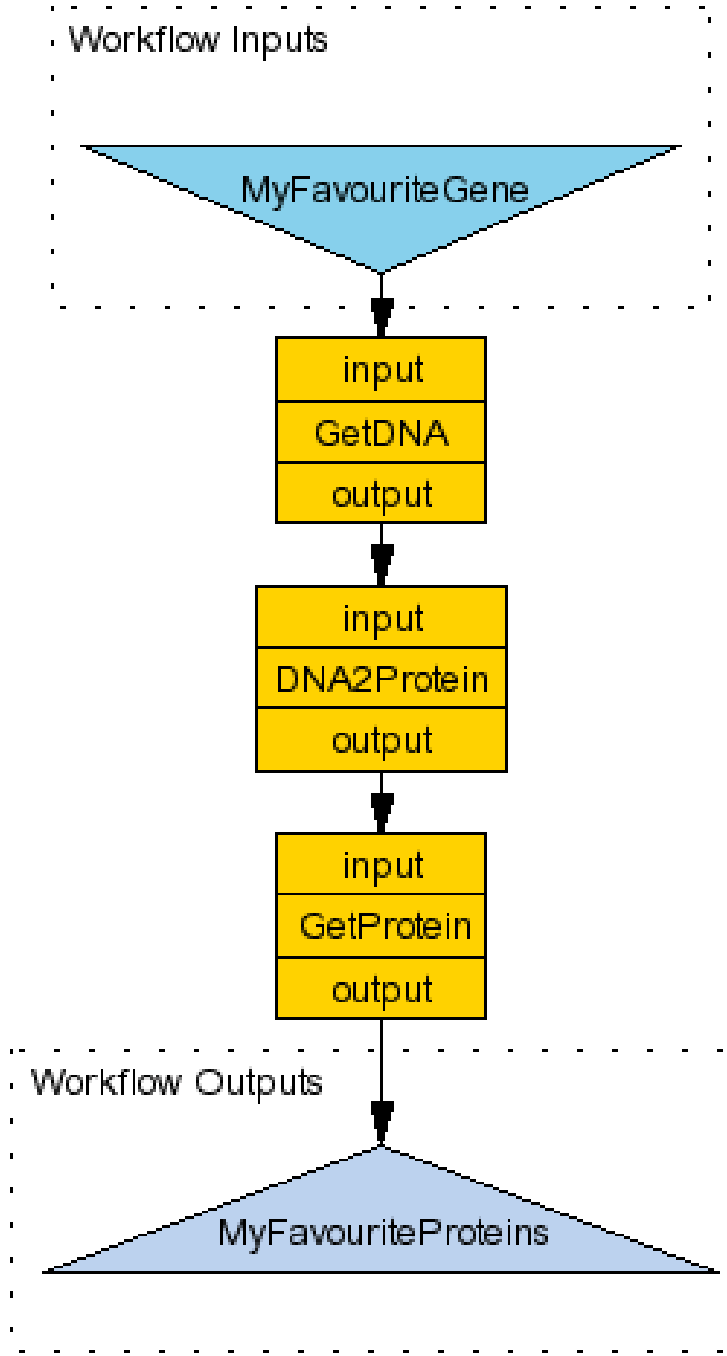
Example 3

- Some shims are multistep
  - Depending on how you describe them
- Investigating AI planning techniques for this



It’s not always safe!

- “Just because you can, doesn’t mean you should”
- Automation is context dependent
- change workflow, change the experiment (seriously annoy the biologist)



Conclusions

- Web Services in bioinformatics leave integration problems: Shims mediate
- Tractable number of shim *types*
- Semantically annotating *all* services reduces mediation to service discovery
- Some shims maybe be automated but it can be dangerous

Acknowledgements

PhD supervisors/advisors

Robert Stevens, Carole Goble, Phillip Lord, Chris Wroe, Alvaro Fernandes

myGrid users

Hannah Tipney, May Tassabehji, Medical Genetics team at St Marys Hospital, Manchester, UK. Andy Brass

Simon Pearce and Claire Jennings, Institute of Human Genetics School of Clinical Medical Sciences, University of Newcastle, UK

myGrid 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, Yike 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 Pocock Milena Radenkovic, Stefan Rennick-Egglestone, Peter Rice, Martin Senger, Nick Sharman, Robert Stevens, Victor Tan, Anil Wipat, Paul Watson and Chris Wroe.

Postgraduates

Martin Szomszor, 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)

The Wellcome Trust

myGrid

EPSRC Engineering and Physical Sciences Research Council