REALIZING THE LIFE SCIENCE GRID WITH TAVERNA
GRID?

In general a grid system is, or should be:

“A collection of resources able to act collaboratively in pursuit of an overall objective”

A life science grid is therefore:

“A collection of resources able to act collaboratively to solve a problem in the life science domain”
Life sciences

- Massive diversity of
  - Information classes
  - Services
  - Data
  - Problems
- Relatively small data sizes
- Relatively small computational load
- Challenge is complexity and heterogeneity
- Much scientific work is exploratory
  - Environment must be flexible and easy to reconfigure
  - Environment must provide facilities for provenance capture
Current inventory?

- **Existing diverse services**
  - Web based, SOAP services, custom protocols such as BioMoby etc.

- **Existing data resources**
  - Relational, unstructured flat file, XML
  - May or may not be exposed through some kind of service interface i.e. SRS, BioMart

- **Existing user communities**
  - Large well funded service and research projects with substantial IT support
  - Small groups with no IT support, little funding but interesting problems
Our life science grid:

“A collection of existing legacy and novel tools and databases exposed through a variety of technologies able to act collaboratively to solve a problem posed by an ‘IT naïve’ user in the life science domain across the public internet and with little or no technological support and as inexpensively as possible.”

Users typically have no control over services (provided by 3rd parties) so create a client side integration platform.

Should be accessible to an unsupported PhD student with standard networking, a three year old PC and no dedicated IT support.
Taverna workflow workbench - http://taverna.sf.net
Taverna also provides:

- **Service discovery**
  - Free text search over ‘known’ services.
  - Semantic search over service repository, relies on manual service annotation and submission of those annotations to the repository.

- **Provenance tracking**
  - Lineage tracking of result data.
  - Automatic semantic annotation of data from service annotations.
  - Possible as the workflow engine creates a ‘managed environment’ with an overview of all data movement.

- **Result visualization**
  - Common renderers included in base distribution include 3d structure, images, graph rendering

- **Extensibility**
  - New service classes
  - New renderer types
  - New UI elements
Project management

- Funded through the Open Middleware Infrastructure Institute (OMII-UK) as part of the myGrid project run by Carole Goble
- Four years old, funding secured through 2008 and beyond.
- Development team at Manchester, UK
- Wide group of ‘friends and allies’ across the world particularly within UK eScience
- Implemented in Java, released under LGPL licence.
Work with us! Please?

- Taverna2 under development, delivery by the end of 2007
- Rewrite of Taverna to support, amongst other things:
  - Integration with grid technologies through a set of new extensibility points
  - Transient VO management (short lived virtual organizations, 20 second upwards lifetime!)
  - More sophisticated computational model
  - Massive scalability, pipelining of nested token streams, single threaded execution model, transparent reference passing architecture
  - Monitoring and steering of running processes with arbitrary granularity through an extension point
- Implement extensions to interface to your GRID
- Get a free and well supported rich client portal for non expert users
- Access otherwise out of reach user communities
myGrid Acknowledgements

Carole Goble, Norman Paton, Robert Stevens, Anil Wipat, David De Roure, Steve Pettifer

OMII-UK
- Tom Oinn, Katy Wolstencroft, June Finch, Stuart Owen, David Withers, Stian Soiland, Franck Tanoh, Matthew Gamble

Research
- Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, Antoon Goderis, Alastair Hampshire, Qiuwei Yu, Wang Kaixuan,

Current contributors
- Martin Szomszor, Duncan Hull, Jun Zhao, Pinar Alper, Antoon Goderis, Alastair Hampshire, Qiuwei Yu, Wang Kaixuan,

User Advocates and their bosses
- Matthew Pocock, James Marsh, Khalid Belhajjame, PsyGrid project, Bergen people, EMBRACE people

Past Contributors

Industrial
- Dennis Quan, Sean Martin, Michael Niemi (IBM), Chimagtica,

Funders
- EPSRC, Wellcome Trust

OMII-UK, eScience All hands meeting 2006
Looking forward – transient grids?

- Science varies widely in scale both in space (CPU cycles required, storage, numbers of services etc) and time (duration of collaborations, stability of VO membership)
- Current grid infrastructure is focused on projects with large spatial and temporal scale
- Does this existing work map well to scientific problems with different characteristics, especially different temporal characteristics?
Enactment group structure

User agent(s)
- Act as proxies to a subset of user capabilities
- Provide security services i.e. message signing
- Embeds security policy

Workflow Engine Instance
- Orchestrates dataflow logic as defined by workflow
- Provides monitoring and steering logic

Data Manager Peer
- Provides access to local and remote data by reference
- Unification layer over other data services