Experiences with eScience workflow specification and enactment in bioinformatics

Darren Marvin
IT Innovation Centre
2 September 2003
All Hands Meeting
Contents

• What problem are we are trying to solve?
• The approach we’ve taken
• What we’ve built
• Integration of workflow technology into myGrid
• Who we’re working with
• What’s coming next
• Downloading and using our software
• Demonstrations
• Questions
What sort of biology problems is myGrid aiming to help solve?

- Graves’ Disease
  - Autoimmune disease of the thyroid in which the immune system of an individual attacks cells in the thyroid gland resulting in hyperthyroidism
  - Weight loss, trembling, muscle weakness, increased pulse rate, increased sweating and heat intolerance, goitre, exophthalmos
What sort of biology problems is myGrid aiming to help solve?

- Grave’s Disease is caused by the stimulation of the thyrotrophin receptor by thyroid-stimulating autoantibodies secreted by lymphocytes of the immune system.

- What is the molecular basis for this autoimmune response?
A biologist’s approach to the problem

- Combine lab biology and in-silico experiments
- Exploratory
- Ad-hoc
- Hypothesis driven
- Not prescriptive
- Bespoke processes
Example services: SoapLab

For each application
- CreateJob
- Run
- WaitFor
- GetResults
- Destroy
Example Services: Talisman

XML Scripts define a series of activities to perform
Workflow requirements

- Varying levels of abstraction
  - Let the biologist concentrate on the science not the technicalities of composing and invoking services
  - Stateful and script-driven services
- Workflow lifecycle
  - Authoring, enacting, validating, modifying
  - Publishing and sharing, which involves semantics, annotation, discovery and personalisation
- Provenance
  - What, where, when, how, who, why…
- Easy to use editing and enactment tools
  - open source is important to the Bio community
- Support for large datasets
The approach we’re taking

• Build something that people can use on a day-to-day basis within the bioinformatics and wider e-Science community
• Provide a basis for the research and demonstration of the benefits of new technologies (e.g. Semantic Web) in eScience
• Deliver tools and specifications in a form that can be easily taken further both during and beyond the end of the project
### Workflow ‘standards’ for Web Services

<table>
<thead>
<tr>
<th>Workflow Management Coalition</th>
<th>EDOC UML</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMG</td>
<td>EDOC UML</td>
</tr>
<tr>
<td>Web Services</td>
<td>Xlang WSFL BPEL4WS BPML WSCI</td>
</tr>
</tbody>
</table>

- Wrong level of abstraction, shifting sands, very few free tools, no explicit support for eScience (e.g. provenance, semantics)
What we’ve built: workflow engine

MyGrid Workflow Enactor

Workflow Enactment Core

BrokerFactory

MyGridBroker

Workflow Definition

conforming to a supported DiGraph based workflow schema

Workflow Language Extension Point

WSFL DiGraph Generator

Flow Entity

Flow Registry

Flow Command Handler - Deals with a Queue of commands for Flows

Create Flow from DiGraph

Workflow Commands (submit, delete, suspend...)

Task Message Handler - Resolves Ready Tasks within Core

TaskGraph

Task

Task Graph

Task Message Extension Point

MyGrid Task Monitor (Message Queue)

Accepts Task Message from

MyGrid Dispatcher

Remote Web Service

Assynchronous Messaging Support

Execution Architecture Extension Point

MyGrid Invocation API

RPC

Remote Web Service

MyGrid Task

Tasks for UDDI Lookup, WSDL Invocation, User Proxy...

Task Extension Point
What we’ve built: workflow workbench
What we’ve built: summary

• Taverna
  – build, edit and browse workflows
  – easy import of services and graphical view of workflows
  – integrated execution using enactor

• FreeFluo
  – parallel and sequential flows, data iteration, nested flows
  – web services, talisman, SoapLab
  – provenance and status reporting

• Deployment
  – available as easy to install desktop toolset
  – integrated within myGrid workbench
  – Enactor available as a Web Service and a Grid Service
Integration of workflow into myGrid

myView on the mIR

Workflow

Metadata about workflow

note about workflow
Who we are working with

- HGMP and EBI
- eHTPX
- Thorton Group at EBI
- ESSC at Reading
- Triana at Cardiff
What’s coming next (1)?

• Large datasets
  – Protocols: secure ftp, SOAP attachments
  – Intermediate staging of data
  – Streaming to and from files, Xpath

• Data model
  – How to deal with arbitrary complex types whilst remaining scalable to large datasets?

• Security
  – WS-Security
What’s coming next (2)?

- **Portal**
  - Workflow lifecycle
  - Requires semantics

- **Contextualised services**
  - Stateful interaction between client and server
  - Web Services standards emerging

- **Integration of local applications into a workflow**
  - Perhaps using Triana
Downloading and using our software

• Taverna
  – Graphical workflow authoring tool
    http://taverna.sourceforge.net
  – LGPL open source on SourceForge
  – User and developer documentation
    • Scufl language specification
    • Videos and examples

• FreeFluo
  – Workflow enactment engine
  – http://freefluo.sourceforge.net
  – LGPL open source on SourceForge
Demonstrations

- Building and enacting a simple workflow
- Use of stateful services
- Implicit data iteration
- Workflow composition
Questions?

Taverna: http://taverna.sourceforge.net
FreeFluo: http://freefluo.sourceforge.net
END