Getting Serious about a Community Bio-Service Catalogue

Carole Goble and Katy Wolstencroft
The Open Middleware Infrastructure Institute,
The School of Computer Science
The University of Manchester, UK
cgoble@manchester.ac.uk, kwolstencroft@manchester.ac.uk
http://www.omii.ac.uk

The rise of bioinformatics and the *in silico* experiment has revolutionised the Life Sciences. Biologists now share a global community with rich, publicly accessible data resources and analysis tools; currently 850 databases are publicly web-accessible [1]. To get the most value out of these resources, however, they need to be able to integrate, interrogate and mine heterogeneous data and associated knowledge from distributed sources.

myGrid (http://www.mygrid.org.uk) has developed the Taverna workbench which is a platform for accessing these distributed resources and providing the mechanical means of interoperating between them using workflows [2]. Workflows are an embodiment of the experimental protocol, to be repeated, reused, inspected and shared to improve and disseminate experimental best practice [3].

To enable scientists to design workflows, they have to discover services (and other prior workflows) and understand how to invoke them. Taverna enables access to 3000+ services that can become steps in a workflow: a bewildering, and increasing, number. To invoke a service the scientist must know the format of the input(s) the service is expecting. To combine services, they must also know the output formats. The heterogeneity of the bioinformatics domain and a lack of standard data format(s) means that describing services with simple typing is impractical. Describing the syntactic interface does not provide enough information for the user to successfully invoke the service. In many cases, each input or output is just a string. Services in the bioinformatics domain are provided with a varying amount of metadata describing their function and many have no descriptions at all. The service providers, decoupled from their consumers, have no obligation to supply such information and no great incentive either.

For myGrid, for other middleware developers like BioMOBY [4], for application developers in the community, for bioinformaticians, we need provisioned semantic annotations for the community's services, and a storage and discovery framework to support it. But we need more than that if we are to get serious:

Technical Infrastructure: myGrid service descriptions are produced by annotating services with terms from the myGrid ontology, stored in a central registry, GRIMOIRE. Services are found using the Feta discovery service [5]. We have piloted expert manual annotation tools augmented by automated tools using information extraction techniques.

Semantic Infrastructure the controlled vocabulary for our annotations is an suite of ontologies written in OWL, and deployed as RDFS, that describe the bioinformatics research domain and the dimensions with which a service can be characterised from the perspective of the *scientist*, not the software developer [6].

Tools: Feta uses a "query and respond" registry interface, deployed chiefly as a plug-in to the Taverna workbench. However, for many users an Amazon-style web based browser is preferable, so scientists can "go shopping" for web services; we have piloted BioBay that adapts shopping metaphors to science service discovery [7]. Going further, our myExperiment activity is designing and building a "MySpace-like" collaboration environment, in partnership with an international focus group of users, for light-touch hosting of the registry and a user-controlled workflow and data exchange platform [8].

Capture and Curation effort: For the semantic discovery framework to be effective, a critical-mass of service descriptions have to be present. We have employed a full-time curator for the provision and maintenance of these annotations and to design annotation tools for future service providers, and begun to

develop the Chameleon infrastructure to help cope with the impact of changes to the services and ontologies on the annotations [9].

Community buy-in: None of these services are owned or developed by us. They are developed by the community. We are developing a partnership with the European Bioinformatic Institute to improve the metadata for services provided by major suppliers *at source*, and propagate best practice *by example*. We know that there is a real enthusiasm for such a catalogue in the community.

Institutional support: Without institutional support for hosting the registry, the curation process, the engineering resources to develop the infrastructure and partnering with suppliers, the endeavour has no sustainability and no impact. The Open Middleware Infrastructure Institute gives us such a setting to make initiative realistic and work.

Its time to get serious about building and managing this catalogue. The outcome would be a living and useful resource, not just for bioinformaticians but also for those working in the Semantic Web for Life Sciences, or just the Semantic Web. We now have the parts, and the know-how, so lets Just Do It.

References

[1] Nucleic Acids Research, January 2006

[2] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: A tool for the composition and enactment of bioinformatics workflows," *Bioinformatics Journal*, vol. 20, pp. 3045-3054, 2004.

[3] C. Wroe, C. Goble, A. Goderis, P. Lord, S. Miles, J. Papay, P. Alper, and L. Moreau, "Recycling workflows and services through discovery and reuse," *Concurrency and Computation: Practice and Experience*, 2006

[4] M. D. Wilkinson, "BioMOBY - the MOBY-S Platform for Interoperable Data Service Provision," in *Computational Genomics Theory and Application*, R. P. Grant, Ed. Wymondham, U.K: Horizon Bioscience, 2004.

[5] Phillip Lord, Pinar Alper, Chris Wroe, and Carole Goble *Feta: A light-weight architecture for user oriented semantic service discovery* in Proc of 2nd European Semantic Web Conference, Crete, 29 May – 1 June 2005, Springer LNCS 3532

[6] Wroe C., Stevens R.D., Goble C.A., Roberts A., Greenwood. M. *A suite of DAML+OIL ontologies to describe bioinformatics web services and data*. International Journal of Cooperative Information Systems. special issue on Bioinformatics and Biological Data Management **12**(2):197-224, 2003.

[7] Qiuwei (sky) Yu, Investigating *the Application of Novel Browsing Techniques for Services for the Taverna Life Science Workbench*, MSc thesis, University of Manchester, UK, 2006

[8] http://www.myExperiment.org

[9] Wang Kaixuan *Handling ontology change in myGrid service discovery*, MSc thesis, University of Manchester, UK, 2006