BioCatalogue: A Curated Web Service Registry for the Life Science Community

Carole A. Goble[1], Khalid Belhajjame[1], Franck Tanoh[1], Jiten Bhagat[1], Katy Wolstencroft[1],
Robert Stevens[1], Eric Nzuobontane[2], Hamish McWilliam[2], Thomas Laurent[2], Rodrigo Lopez[2]

[1]School of Computer Science, University of Manchester, Oxford Road, Manchester, UK
[2]EMBL European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD
Contact: Khalid.Belhajjame@manchester.ac.uk
http://www.biocatalogue.org/

Web services have gained a momentum as a means for packaging existing data and computational resources in a form that is amenable for use and composition by third party applications. They provide a well defined programming interface to integrate tools into applications over networks. Software applications written in various programming languages and running on various platforms can use web services to exchange data over the Internet. The life science community is certainly among the first adopters of web services. For example, *Nucleic Acids Research* describes 166 web servers [4].

However, one of the main issues that hinders the wide adoption and use of web services is the difficulty in locating the "*appropriate*" web service, i.e., the web service that performs the analysis the scientist is interested in. For example, Taverna[1], a workflow workbench that is popular within the life science community, provides access to over 3500 thousands web services that can be composed by scientists for constructing and enacting their in silico experiments. Manually browsing this list of web services can be time consuming. Furthermore, the descriptions of available web services are often poor providing little information to the scientist about their usefulness for the analysis s/he is after. This is a major obstacle to the users of these services, (biologists, bioinformaticians and tool providers) thus the community as a whole. It is also a frustration to the service providers, whose services are unknown, unused, poorly used or mis-used as a result. The story is one of confusion, frustration, fossilised scientific practice and wasted effort.

With the above issues in mind, the authors have recently initiated the BioCatalogue project. Unlike existing service registries, such as SEEKDA[2], which describe general purpose web services, BioCatalogue aims to build a web service registry dedicated to the life science community. BioCatalogue will provide a means by which a bioinformatician, for instance, can subscribe his/her favourite web service within the catalogue, e.g., by uploading the web service description file (WSDL). The Bioinformatician can then add more text to describe the web service and map the web service elements to terms in the myGrid service and domain ontology. Missing terms can be substituted with tags that can be subsequently mined for new ontology terms. Annotation dates and histories can also be part of the markup.

BioCatalogue Salient Functionalities

This section overviews the main functionalities that will be supported by BioCatalogue.

Integrated access to life science web services. BioCatalogue will act as a *one-stop-shop* the scientists can use to locate web services that implement the analysis relevant for their scientific experiments. BioCatalogue will provide an integrated access to life science web services: it will allow users to locate web services that are implemented by different providers and that are hosted by remote server. In addition to locating web services, users will be able to perform basic operations such as checking the accessibility of the web service, invoking it using sample inputs, and browsing the results returned, if any.

Rich Description of Web Services. Available web services are rarely described, and when they are, they are poorly described [1]. To enable scientists to effectively find a service, they need to know the *functional capabilities* supplied by the service, i.e., what it does, over which other resources it operates (e.g. which databases a tool searches over), the format of the inputs, the ranges for any parameters, the type of data they expect, the output

[1] http://taverna.sourceforge.net/

[2] http://www.seekda.com

formats and the type of outputs etc; *operational capabilities; reliability*, how up to date it is, if licences are required, its current version, its quality of service, failure rate, security policies, links to other services etc.; *provenance*, i.e., its history, originators, how is it sustained; and *reputation*, i.e., its standing in the community for its quality of a service, usefulness, popularity, anonymous and attributed recommendations, who else uses it and why etc.

In BioCatalogue, service descriptions will cater for the above information and will be encoded as metadata specified using annotations which associate web services to their respective descriptions. In its simplistic form, annotations will be textual descriptions or lists of keywords (tags). However, to enable their use by machines, as well as humans, a more controlled annotation mechanism will be supported. For example, service annotations will also be encoded in the form of associations that relate web service elements to concepts and properties defined in ontologies. A domain ontology that can be used for this purpose was built within the myGrid project[3] and which contains concepts that can be used for annotating and services from the domain of bioinformatics.

Curation of Service Descriptions. In bioinformatics, we are familiar with the idea of curated data as a pre-requisite for data integration. We neglect, often to our cost, the curation and cataloguing of the analysis, i.e., the web service that we use to compute our data. In this respect, and as pointed out earlier, to locate the appropriate web service, the scientists will need some information about the service. BioCatalogue will provide scientists with this information.

However, as shown by Belhajjame *et al.*, existing descriptions of web services are frequently erroneous or inaccurate [1]. This motivates the need for a curation process whereby existing service descriptions can be checked and augmented if necessary. In doing so, we envisage two mechanisms, automatic and manual. Automatic curation will be performed using tools that systematically verify the accuracy of service description before their publication. In this respect, we will be using the Quasar verification tool. This is a tool that was developed within the University of Manchester under the ages of the Quasar project [2] for detecting errors in semantic description of web services. Service descriptions will also be manually curated by experts in the life science domain as well as by the BioCatalogue's users who will be able to comment and refine existing descriptions as well as submit new ones.

Web Service Discovery. BioCatalogue will allow users to discover the web service using mainly the following mechanisms. *Keyword-based service retrieval* allows users to locate web service of interest by providing as input one or multiple key words that provide some information about the service, e.g., the name of service, its tag. *Advanced search capabilities* will also be supported by BioCatalogue. A user in this case, will provide more information about the functionality of the target web service, the kinds of data it takes as input and/or the format of the results it delivers, its reputation, reliability and so forth. Last but not least, we intend to couple the traditional "form-filling" search mechanisms with a modern facet-based "shopping" style web interface *à la* Amazon.

Interoperability. BioCatalogue will provide API that can be used by third party applications, such as myExperiment [3] and Taverna, to programmatically access the registry. For example, it will provides RSS feeds using which scientists can be kept up to date, e.g., if subscribed, a user will be notified when a new service that may be of relevance to his/her research is added to the catalogue.

This paper briefly presented the main requirements that the authors aims to cater for within the BioCatalogue project to improve the process by which web services in the life sciences are located, used, described, and validated. By creating a social environment to register, curate and search for web services and generating content that can be indexed by third party information providers, such as Google, web services should be easy to find and use.

References

1. K Belhajjame, SM Embury, NW Paton, R Stevens, and CA Goble. *Automatic Annotation of Web Services Based on Workflow Definitions*. in ISWC2006, LNCS 4273: 116-129 (2006)
2. K Belhajjame, SM Embury, NW Paton, R Stevens, and CA Goble. *The QuASAR project: Quality Assurance of Semantic Annotations for seRvices (A research project initiated by the University of Manchester)*, 2008. http://img.cs.man.ac.uk/quasar/index.php.
3. DD Roure, C Goble and R Stevens. *The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows*. FGCS, 2008. (In press)
4. JA Fox, SL Butland, S McMillan, G Campbell and BFE Ouellette. *The Bioinformatics Links Directory: a Compilation of Molecular Biology Web Servers*. Nucl Acids Res 2005 33: W3-W24

[3] http://www.mygrid.org.uk