# Bringing caBIG services together using Taverna

Aleksandra Nenadic, Stian Soiland-Reyes, Carole Goble
University of Manchester, Oxford Road, M13 9PL, UK

## 1 Introduction

The National Cancer Institute (NCI) from the US has built an integrated biomedical platform called caBIG in order to improve the US cancer research community's access to key bioinformatics data and tools. The caBIG programme was established in order to allow researchers to more efficiently discover, share, process and integrate disparate clinical and research data, with an ultimate goal of accelerating cancer research.

The backbone of the caBIG infrastructure is caGrid – a grid-computing service-oriented environment that integrates various data and analytical services and leverages the combined strengths and expertise of participating organizations in an open and federated environment. caBIG services are categorized as legacy, bronze, silver and gold based on their ability to interoperate with other caBIG services. Legacy does not guarantee any interoperability with external software components or data resources; silver services comply with standardized semantic advertising, discovery and use of the service; gold level is currently being defined by caBIG as an extension to the silver that will enable full syntactic and semantic interoperability of services. Most of the services currently deployed at caGrid are silver.

Being able to link caBIG services into a workflow is one of the key requirements for caBIG users, in order for them to automate the design and running of their virtual experiments. Taverna Workbench (henceforth Taverna) [1], a workflow design, management and execution system developed by the myGrid team (now based in Manchester, UK), has emerged as the primary candidate that will meet this need. The myGrid team has joined efforts with the caGrid workflow team from the US (funded by the NCI) in order to integrate caBIG services into Taverna and enable caBIG users to orchestrate data and analytical services in Taverna workflows. This also included integration with the caGrid's GAARDS security framework for interaction with secure caBIG services. As a side effect of the integration, caBIG services can be combined in mix-and-match style workflows with other existing third party services already accessible from Taverna. This paper describes the details of Taverna-caBIG integration – the extensions added to Taverna in order to be able to discover and invoke caBIG services from Taverna workflows.

## 2 caBIG services

caGrid is the underlying network architecture for caBIG aimed at enabling multi-institutional data sharing and analysis. It provides the basis for connectivity between various cancer community institutions in the US, allowing research groups access to rich collections of emerging cancer research data to support their individual investigations. Services deployed in caGrid include data and analytical services. Data services allow sharing of data defined in data models. Analytical services allow researches to analyse the data provided by data services.

caBIG services are implemented as WS-Resource Framework (WSRF) grid services using the Globus Toolkit. The WSRF is a set of Web services specifications that define what is termed as the WS-Resource approach to modelling and managing state in a Web services context. Web services must often provide their users with the ability to access and manipulate state, i.e. data values that persist across, and evolve as a result of, Web service interactions. WSRF defines conventions for managing state so that applications can discover, inspect, and interact with stateful resources in standard and interoperable ways, within the context of established Web services standards. Supporting WSRF Web services in Taverna was one of the key requirements for being able to invoke caBIG services – we give details of how this was achieved in the following section.

In addition, caBIG silver level compatible services are semantically integrated – all exposed API elements have runtime-accessible metadata that defines the meaning of the composing elements using controlled terminology and vocabularies. This metadata is stored in caDSR (cancer Data and Service Repository) and available to other services and applications. To achieve the semantic integration, design and implementation of caBIG services have to follow certain rules:

- Adherence to model-driven architecture: use UML to build platform-independent models that, through a series of transformations, are converted to platform-specific models and subsequently to executable code.
- Employing controlled vocabularies and metadata repositories to achieve runtime semantic interoperability, i.e. the ability to determine the context of data that is returned by a data or analytical service at runtime. This is critical to grids that need to perform federated queries across multiple services and need to determine the nature of the data on the grid. Controlled vocabularies used by caBIG services include SNOMED, MedDRA, GO Ontology and the NCI Thesaurus.

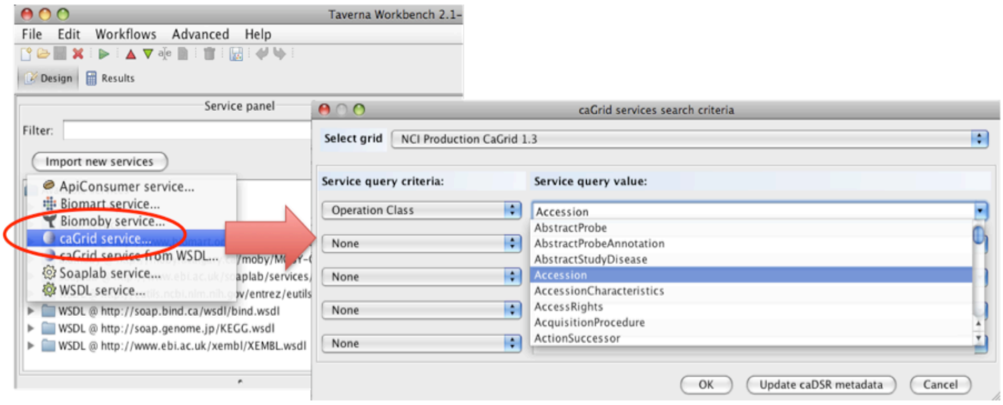
# 3 Support for caBIG services in Taverna

Even though caBIG services have been designed with interoperability in mind, currently the only way to orchestrate them together is though workflows [2,3,4]. In order to be able to combine caBIG services in Taverna workflows, the following extensions to Taverna had to be developed:

- Semantic discovery of caBIG services,
- Support for stateful WSRF-compliant caBIG services,
- Support for secure caBIG services.

## 3.1 Semantic discovery of caBIG services

The caBIG Service Discoverer plug-in (show in Figure 1) acts as a semantic (i.e. metadata-based) caBIG service searcher. It can find all caBIG services currently registered with the caGrid's Index Service. In addition, it can perform a more refined search making use of service metadata. As mentioned previously, all caBIG services are semantically annotated and this information is available through the caGrid's caCDR service. Semantic search enables users to combine full text search, simple text-based criteria such as specifying operation names or concept codes, and more complex criteria such as specification of point-of-contact information or research centre a service belongs to or UML class criteria for services inputs and outputs. For example, one can refine the search by defining criteria that will only search for services that return output of certain class type or that receive certain class type as input, e.g. a Protein or a Nucleotide class.



Select grid: NCI Production CaGrid 1.3

Service query criteria: Operation Class

Service query value: Accession

Figure 1. Semantic discovery of caBIG services

Once caBIG services are discovered, they are added to Taverna's palette of available services and are ready to be drag-and-dropped into workflows.

## 3.2 Support for stateful WSRF-compliant caBIG services

Taverna has already been capable of invoking non-WSRF Web services. However, since caBIG services follow the WSRF specification we had to adapt Taverna's Web service invoker to be able to cope with WSRF-compliant services. Essentially, all that Taverna knows about a Web service is the location of its wsdl document. The wsdl document is parsed in order to determine the available operations of the service, their inputs and outputs, encoding styles, bindings etc. and the service can then be readily invoked from inside a workflow. From the aspect of a wsdl document, there is nothing different between a WSRF-compliant and 'ordinary' Web service. So, in order to detect if a service is WSRF-compliant, Taverna searches for a special method called GetResourceProperty inside the wsdl document that will give this information away. Even though all caBIG services are WSRF-compliant, some of them do not implement the *GetResourceProperty* method, so additional methods that Taverna looks for include: *Destroy*, *SetTerminationTime* and *getServiceSecurityMetadata*.

If a Web service is detected to be WSRF-compliant, Taverna adds an extra input port called EndpointReference to such a service, which can be used to pass a reference to a resource (i.e. endpoint). Later on, when a workflow containing such a service is invoked, this special EndpointReference port can be used to pass a piece of XML that contains an endpoint reference as <wsa:EndpointReference> element. This element is extracted and inserted into the SOAP header according to the WS-Addressing specification. The service receiving it will understand and de-reference it accordingly to retrieve the resource in question.

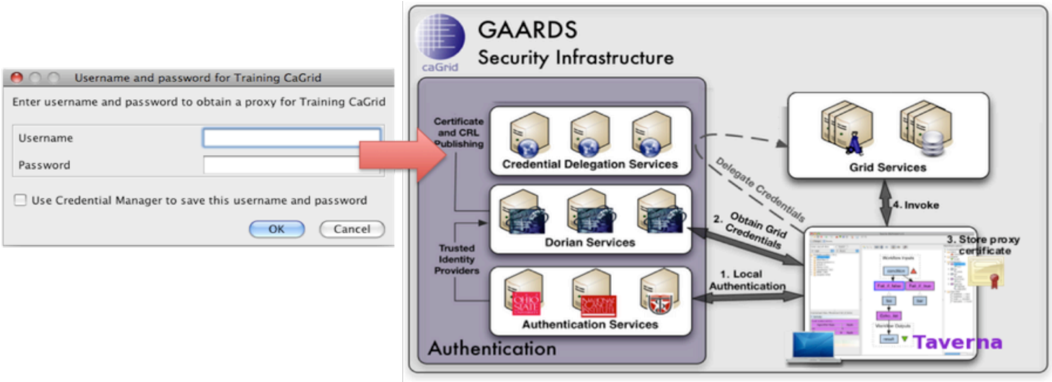
## 3.3 Support for secure caBIG services

Not all of caBIG services are freely available – some require user to log onto caGrid in order to be granted access. caGrid employs a security infrastructure called GAARDS (Grid Authentication and Authorization with

Reliably Distributed Services) developed on top of the Globus Toolkit and extending the Grid Security Infrastructure (GSI). The users' interactions with GAARDS is greatly simplified by enabling them to log onto caGrid using username and passwords provided by their affiliated institutions and then being issued with a 12 hour proxy certificate that brings the single sign on experience to the user for the lifetime of the proxy.

caBIG services can employ transport or message level security protections, or their combination. Transport level security implies that a service is invoked using https. Message level security refers to various security protections performed on SOAP messages according to the WS-Security and WS-SecureConversation specifications.

All caBIG services implement a special *getServiceSecurityMetadata* method that is used by Taverna to discover the security requirements of a service. Prior to invoking such a service, Taverna enables user to authenticate with their affiliated GAARDS Authentication Service and subsequently obtains a proxy on user's behalf from GAARDS Dorian Service (see Figure 2). Authentication Service provides a framework for authenticating users and issuing them with SAML assertions that confirm successful authentication. Dorian is a service for the provisioning and management of grid users accounts – it consumes SAML assertions and, based on them, issues users with proxy credentials. Dorian provides an integration point between external security domains and the grid, allowing users to use their existing credentials (external to the grid) to authenticate to caGrid.



Username and password for Training CaGrid

Enter username and password to obtain a proxy for Training CaGrid

Username

Password

☐ Use Credential Manager to save this username and password

Figure 2. Taverna-GAARDS interaction when invoking secure caBIG services

## 4 Conclusion

European eScience grid environments such as NorduGrid and EGEE [5,6] are already accessible from Taverna. Taverna-caGrid integration project has brought closer a grid from the US to European researchers and vice versa. It has shown how Taverna can be used as an interoperability platform for caBIG services and how they can be orchestrated into data processing pipelines by incorporation into Taverna workflows. It has also demonstrated how caBIG services can be combined with other existing third party services outside the caBIG world by using Taverna workflows, widening the range of data services and analytical pipelines available to caBIG users. The myGrid team has developed considerable expertise on caBIG services and their silver level compatibility guidelines that is ripe for further exploitation. The next steps are to capitalise on this strong foundation by enabling wider user adoption and hence ensuring a greater impact across caBIG community by enabling simpler workflow design and simpler workflow sharing by using a public repository of workflows such as myExperiment [7] inside the caBIG portal.

## References

[1] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, T. Oinn, Taverna: a tool for building and running workflows of services, Nucleic Acids Research, vol. 34, Web Server issue, pp. 729-732, 2006.

[2] W. Tan, I. Foster, R. Madduri. Scientific workflows that enable Web-scale collaboration: combining the power of Taverna and caGrid. IEEE Internet Computing. vol.12, no.6, pp. 30-37, 2008.

[3] W. Tan, K. Chard, D. Sulakhe, R. Madduri, I. Foster, S. Soiland-Reyes, C. Goble, Scientific workflows as services in caGrid: a Taverna and gRAVI approach. IEEE Conference on Web Services, 2009.

[4] W. Tan, P. Missier, R. Madduri, I. Foster, Building Scientific Workflow with Taverna and BPEL: a Comparative Study in caGrid, International Workshop on Engineering Service-Oriented, 2008.

[5] Zhou et al., Easy Setup for Parallel Medical Image Processing: Using Taverna and ARC, HealthGrid, 2009.

[6] K. Maheshwari, P. Missier, C. Goble, J. Montagnat, Medical Image Processing Workflow Support on the EGEE Grid with Taverna, IEEE Symposium on Computer Based Medical Systems, 2009.

[7] D. De Roure, C. Goble, R. Stevens, The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows, Future Generation Computer Systems, vol. 25, 2009.