The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web, or in the Cloud

Journal:	*Nucleic Acids Research*
Manuscript ID:	NAR-00282-Web-B-2013.R1
Manuscript Type:	7 Web Server Issue
Key Words:	scientific workflows, Web Services, data integration, distributed computing

The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web, or in the Cloud

Katherine Wolstencroft[1], Robert Haines[1], Donal Fellows[1], Alan Williams[1], David Withers[1], Stuart Owen[1], Stian Soiland-Reyes[1], Ian Dunlop[1], Aleksandra Nenadic[1], Paul Fisher[2], Jiten Bhagat, Khalid Belhajjame[1], Finn Bacall[1], Alex Hardisty[3], Abraham Nieva de la Hidalga[3], Maria P. Balcazar Vargas[4], Shoaib Sufi[1], Carole Goble[1]*

[1] School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
[2] Astrazeneca, Alderly Park, Macclesfield, SK10 4TF, UK
[3] Computer Science & Informatics, Cardiff University, Roath, Cardiff, CF24 3AA, UK
[4] Faculty of Science, University of Amsterdam, Amsterdam, 1098 XH, Netherlands

* Katherine Wolstencroft. Tel: +44 161 275 6276 Fax: +44 161 275 6236; Email: katherine.wolstencroft@manchester.ac.uk

ABSTRACT

The Taverna workflow tool suite (http://www.taverna.org.uk) is designed to combine distributed Web Services and/or local tools into complex analysis pipelines. These pipelines can be executed on local desktop machines or through larger infrastructure (such as supercomputers, Grids, or cloud environments), using the Taverna Server. In bioinformatics, Taverna workflows are typically used in the areas of high-throughput omics analyses (for example proteomics, or transcriptomics), or for evidence gathering methods involving text mining or data mining. Through Taverna, scientists have access to several thousand different tools and resources that are freely available from a large range of Life Science institutions. Once constructed, the workflows are reusable, executable bioinformatics protocols that can be shared, reused and repurposed. A repository of public workflows is available at http://www.myexperiment.org. This article provides an update to the Taverna tool suite, highlighting new features and developments since those reported in a previous Web Server issue of this journal.

INTRODUCTION

The quantity and heterogeneity of data in the life sciences has given rise to thousands of Web Services that provide methods for its analysis, retrieval and integration [1]. Large bioinformatics services providers, such as the EBI and the NCBI, routinely offer Web Service access to their resources in either REST or WSDL format [2]. These Web Services can be executed and combined into multi-step analysis pipelines, or workflows, using systems like Taverna [3].

Workflows are reusable informatics analysis protocols. The myExperiment workflows repository [4] provides a collection of workflows (written with Taverna, or with other systems like Galaxy [5] or Kepler [6]) for use by the bioinformatics community. This collection of workflows provides a rich resource for scientists developing new analysis methods. Workflows can be combined and modified to assemble new executable protocols, using published and established pipelines as components.

In practice, most Taverna workflows are composed from a mixture of distributed Web Services, local scripts and other service types (e.g. BioMart queries or R-Scripts) [7]. In some cases, Taverna workflows are *only* composed from local services, for example where data and service execution must remain behind a firewall (e.g. clinical or commercial data), or where the size of data means that

performance is significantly enhanced by reducing network traffic. For example, cloud installations of Taverna host the engine and services in the same cloud environment. Once an initial dataset has been uploaded, the workflow engine and the services in the workflow can receive references to that dataset, instead of the data itself.

The main advantage of using distributed services, however, is that the majority of computational processing in the workflow occurs remotely with the service providers. There is no requirement to install tools and data sources locally, which reduces local infrastructure and maintenance costs and enables rapid workflow development and testing. Consequently, genome-scale analyses can be performed regardless of local infrastructure, using distributed tools and resources.

A disadvantage of integrating third-party Web Services is the variable reliability of those services. If services are frequently unavailable, or if there are changes to service interfaces, workflows will not function correctly [8]. There is, however, large redundancy in web service functions, so the ability to identify reliable services and potential alternatives for non-functioning services is of great advantage. The BioCatalogue (http://www.biocatalogue.org) service registry provides this information, along with metadata descriptions of service inputs, outputs, dependencies and licenses.

This paper is an update of a previous Web Services NAR special issue [3]. Since this publication, the Taverna tool suite has undergone considerable changes and improvements, such as;

- The implementation of a new Taverna engine (currently version 2.4) which caters for the scalable processing of large data sets, and is capable of performing implicit iteration, looping and streaming of data.
- The ability to interact with new types of services in addition to WSDL Web Services, local scripts and BioMart data warehouses, in particular, RESTful Web Services, Grid Services, cloud services, R-scripts and distributed command-line scripts.
- The introduction of the myExperiment repository for sharing, reusing and repurposing workflows. Currently, myExperiment provides access to over 2600 workflows.
- The introduction of the BioCatalogue and Biodiversity Catalogue service registries for the discovery and use of Web Services. They currently contain over 2300 sets of Web Services, providing over 8000 service operations.
- The introduction of the Taverna Server, which allows workflows to be executed on remote computational infrastructure (such as clusters, Grids and clouds), or as components in other workflow systems, such as Galaxy [9].
- The Taverna Player, an interface for the Taverna Server to allow workflow execution from web browsers, or through third-party clients.
- The Interaction Service, which enables scientists to select parameters and data during workflow execution.
- The Taverna Provenance suite, which records service invocations, intermediate and final workflow results and exports provenance in the Open Provenance Model format (http://openprovenance.org/) and the W3C PROV model (http://www.w3.org/2011/prov).

- Improvements to the plugin architecture to enable easier code contributions and extensions, making it possible to extend and personalise the core functionality to suit individual scientists.

RUNNING TAVERNA WORKFLOWS WITH DISTRIBUTED SERVICES

Taverna Workflows can be designed and executed on local desktop machines through the Taverna workbench, or they can be executed through other clients or web interfaces, using the Taverna Server (or the Taverna command-line application). These alternative execution modes serve different types of workflow users. The first execution mode is through the Taverna workbench. The workbench is downloaded to a local machine and provides an environment for bioinformaticians to develop new workflows and test new analysis methods, by either developing workflows from scratch, or by composing them from existing workflows.

The second mode of execution is simple execution through the Taverna Server. The Server is an environment for serving finished workflows to a larger community of scientists. In this mode, a single installation of the Server provides access to a collection of workflows (normally through a web interface, called the Taverna Player). Regular users are not required to download or install any software and do not require any detailed knowledge of distributed computing or Web Services. One drawback of this mode is that users cannot alter workflows or add new workflows to the collection. However, when workflows are being provided as a service, they require hosting on larger production-grade infrastructure to support multi-user executions, session management, and potentially authenticated access.

The final execution mode is via a Taverna Lite installation. Taverna Lite provides an intermediate solution. It allows users not only to run workflows through the web, but also to upload new workflows from myExperiment and other sources. Consequently, Taverna Lite installations require user authentication, but no local software installation by regular users, since workflow execution also occurs on a server. The following sections describe how to execute workflows using both the workbench and the Taverna Server.

Using the Taverna Workbench

The Taverna workbench is freely available and can be downloaded from http://www.taverna.org.uk/. It runs on Windows, Linux and Mac OS X. Installation is a one-click download which must be unzipped on the local machine. For windows, there is also an installer wizard. There is no login required for the workbench, and the majority of third-party services in bioinformatics do not require a login. For those that do, Taverna allows credentials to be added at run-time, or to be stored in a purpose-built Credential Manager.

The Taverna workbench allows users to identify and combine services by dragging and dropping them onto the workflow design panel. The Taverna quick start guide (http://www.taverna.org.uk/documentation/taverna-2-x/quick-start-guide/) provides step-by-step

instructions on how to open and run existing workflows and how to design and run workflows from scratch. Figure 1 shows the Blast_Align_and_Tree workflow that features in the guide. It is part of the *Bioinformatics Workflow Examples* pack (available on myExperiment at http://www.myexperiment.org/packs/363.html). This workflow performs a *classic* phylogenetics analysis. From an input protein sequence, it performs a similarity search (BLAST [11]) against the UniProt database [12] and aligns similar sequences using ClustalW [13]. The alignment is then used to construct a phylogenetic tree, using the EBI ClustalW phylogeny service (the workflow metadata describes the experimental methods in more detail).

This workflow predominantly uses Web Services from the EBI, in both WSDL (Fig 1A) and REST (Fig 1B) format, as well as local scripts for formatting data and managing service compatibility (also known as *Shim* services - Fig 1C). Almost all workflows require shim services because the analysis services have not been specifically designed to work together. Therefore, they often have incompatible input and output formats.

The workflow also contains several nested workflows (or subworkflows - Fig 1D), which can be downloaded individually from myExperiment, demonstrating that workflows can actually be components of other workflows. Some nested workflows control the retrieval of data from asynchronous services, using Taverna's looping mechanism (Fig 1E). The nested workflow is executed repeatedly until results are available. *Control link*s between the nested workflow output and downstream services pause the remainder of the workflow until all preceding results are available.

As the workflow runs, the results panel shows progress through the workflow and iterations over data. This view also displays any errors if there are problems with executions. The final workflow results show a list of similar protein sequences, the alignment of those sequences and a phylogenetic tree.

As of January 2013, myExperiment contains over 1800 Taverna workflows, many of which are in the field of Bioinformatics. The *Bioinformatics Workflow Examples* pack is a collection of workflows that demonstrate specific features of the workbench as well as a variety of bioinformatics analyses. Each workflow contains small sets of example input data, designed to produce results in a matter of minutes. Running these workflows on actual datasets may naturally take longer.

Running Workflows Through the Taverna Server

The Taverna Server and web interface running at http://tavlite1.biovel.eu contain the workflows from the Bioinformatics Example Workflows Pack, and additional workflows in the area of biodiversity. This Server was developed by the BioVeL project (the Biodiversity Virtual e-Laboratory http://www.biovel.eu/). It provides informatics workflows for analysing biodiversity, using third-party Web Services in a widespread "service network". There is no login required to use this server to execute a collection of public workflows, but only members of the BioVeL community have additional access to workflows under development and are able to submit new workflows to run on the server.

The BioVeL server instance is typical in that it provides a collection of workflows targeted to a particular research theme, and provides a mixture of public and restricted access workflows. BioVeL biodiversity analyses include workflows for phylogenetics, metagenomics, population modelling and ecological niche modelling. For example, the martrix population model workflow enables the analysis

of demographic data, such as age specific survival, generation time or net reproductive rate. The workflow was developed from already published R-Scripts, using Taverna's RServe service.

The functions of the workflows on the server can be explored through the *'Details'* links on the workflows page. The *'Run'* link allows workflow execution through the website. Each workflow is provided with default parameter values and example input data. The Blast_Align_and_Tree workflow has the same example data as in the myExperiment version and should therefore provide the same results thorough the web interface.

In general, the Taverna Server can be downloaded and configured to run with or without login restrictions. For a detailed description of installing the Server, see the beginners installation guide at: http://dev.mygrid.org.uk/wiki/display/taverna/A+Beginner%27s+Installation+Guide+to+Taverna+Server.

The Taverna Bioinformatics User Community

There are Taverna workflows spanning most fields of bioinformatics, including omics analyses (such as transcriptomics, proteomics and metabolomics [10,14-16]), text mining, biodiversity and data integration [17,18]. Taverna has also been adopted by a number of large-scale Life Science initiatives, such as the Virtual Physiological Human SHARE project, OpenTox [19] and caGrid [20]. Most published workflows are available from myExperiment, where the number of views and downloads demonstrates the propagation of methods through the community.

Taverna is an open source project. Community users have developed plugins to support specific user groups and tasks. For example, CDK-Taverna provides cheminformatics service support [21] and Tav4SB integrates tools for Systems Biology modelling [22]. Other plugins support the use of Taverna on specific Grid or cloud resources, such as the UNICORE plugin [23], which enables users to submit jobs from the Taverna Workbench to any UNICORE resource.

DISCUSSION AND FUTURE WORK

Analysing and processing the wealth of available data is a central concern in the Life Sciences. However, accessing distributed resources, that are sometimes incompatible and liable to change over time, is challenging. Continuing and future work with Taverna focuses on managing the heterogeneity and ever-changing nature of distributed services. This includes researching workflow preservation and methods for aggregating collections of data, protocols and provenance into digital bundles, termed *Research Objects*. It also involves developing methods for producing compatible collections of services by wrapping Web Services and *shim* services into plug-and-play *components*. By supporting service discovery, workflow design, reuse and execution, the Taverna tool suite enables the exploitation of distributed bioinformatics data and analysis methods.

ACKNOWLEDGEMENT

The authors would like to thank BioVeL, SCAPE, Helio, and all the members of the myGrid family (past and present), including the myGrid team, the PALs, code contributors and users.

FUNDING

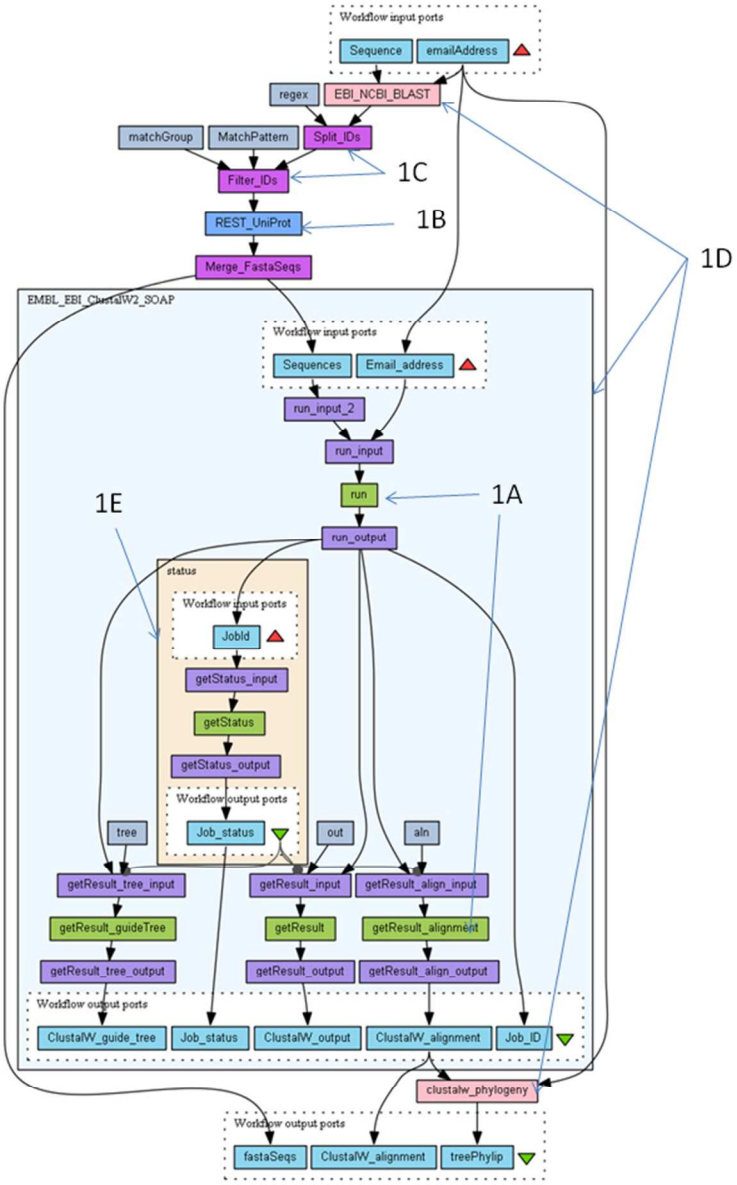
This work has been supported by the Engineering and Physical Sciences Research Council [EP/GO26238/1, EP/C536444/1] and the European Commission 7th Framework Programme [283359, 238969, ICT-2009.4.1]. Funding for open access charge: EP/G026238/1

REFERENCES

1. Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S. *et al.* BioCatalogue: a universal catalogue of Web Services for the life sciences. *Nucleic Acids Res*, **38**, W689-694. PMID:20484378
2. Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, **38**, W695-699. PMID: 20439314
3. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, **34**, W729-732. PMID:16845108
4. Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P. *et al.* myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res*, **38**, W677-682.PMID:20501605
5. Goecks, J., Nekrutenko, A. and Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11**, R86. PMID:20738864
6. Stropp, T., McPhillips, T., Ludascher, B. and Bieda, M. Workflows for microarray data processing in the Kepler environment. *BMC Bioinformatics*, **13**, 102. PMID:22594911
7. Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y. and Goble, C. (2012), *8th IEEE International Conference on eScience*. IEEE Computer Society Press, Chicago, USA.
8. Zhao, J., Gomez-Perez, J.M., Belhajjame, K., Klyne, G., Garcia-Cuesta, E., Garrido, A., Hettne, K., Roos, M., De Roure, D. and Goble, C. (2012), *8th IEEE International Conference on eScience*. IEEE Conputer Society Press, Chicago, USA.
9. Karasavvas, K., Wolstencroft, K., Mina, E., Cruickshank, D., Williams, A., De Roure, D., Goble, C. and Roos, M. Opening new gateways to workflows for life scientists. *Stud Health Technol Inform*, **175**, 131-141. PMID:22942004
10. Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. and Brass, A. (2007) A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Res*, **35**, 5625-5633. PMID:17709344
11. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402. PMID:9254694
12. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*, **41**, D43-D47. PMID:23161681
13. Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, **Chapter 2**, Unit 2 3. PMID:18792934
14. Ahmad, I., Suits, F., Hoekman, B., Swertz, M.A., Byelas, H., Dijkstra, M., Hooft, R., Katsubo, D., van Breukelen, B., Bischoff, R. *et al.* A high-throughput processing service for retention time alignment of complex proteomics and metabolomics LC-MS data. *Bioinformatics*, **27**, 1176-1178. PMID:21349866
15. de Bruin, J.S., Deelder, A.M. and Palmblad, M. Scientific workflow management in proteomics. *Mol Cell Proteomics*, **11**, M111 010595. PMID:22411703

16. Smedley, D., Swertz, M.A., Wolstencroft, K., Proctor, G., Zouberakis, M., Bard, J., Hancock, J.M. and Schofield, P. (2008) Solutions for data integration in functional genomics: a critical assessment and case study. *Brief Bioinform*, **9**, 532-544. PMID:19112082
17. Kano, Y., Dobson, P., Nakanishi, M., Tsujii, J. and Ananiadou, S. Text mining meets workflow: linking U-Compare with Taverna. *Bioinformatics*, **26**, 2486-2487. PMID:20709690
18. Maglogiannis, I., Goudas, T., Doukas, C. and Chatziioannou, A. A Collaborative Biomedical Image Mining Framework: Application on the Image Analysis of Microscopic Kidney Biopsies. *IEEE Trans Inf Technol Biomed*. PMID:23076078
19. Hardy, B., Douglas, N., Helma, C., Rautenberg, M., Jeliazkova, N., Jeliazkov, V., Nikolova, I., Benigni, R., Tcheremenskaia, O., Kramer, S. *et al.* Collaborative development of predictive toxicology applications. *J Cheminform*, **2**, 7. PMID:20807436
20. Tan, W., Madduri, R., Nenadic, A., Soiland-Reyes, S., Sulakhe, D., Foster, I. and Goble, C.A. CaGrid Workflow Toolkit: a Taverna based workflow tool for cancer grid. *BMC Bioinformatics*, **11**, 542. PMID:21044328
21. Truszkowski, A., Jayaseelan, K.V., Neumann, S., Willighagen, E.L., Zielesny, A. and Steinbeck, C. New developments on the cheminformatics open workflow environment CDK-Taverna. *J Cheminform*, **3**, 54. PMID:22166170
22. Rybinski, M., Lula, M., Banasik, P., Lasota, S. and Gambin, A. Tav4SB: integrating tools for analysis of kinetic models of biological systems. *BMC Syst Biol*, **6**, 25. PMID:22480273
23. Sonja Holl, O.Z., Martin Hofmann-Apitius (2011), *IEEE Services*, Washington DC, USA.

Figure 1: The Blast_Align_and_Tree workflow is an example Taverna workflow (available from http://www.myexperiment.org/workflows/3369.html), which performs a phylogenetic analysis. From an input protein Sequence, it performs a similarity search against the Uniprot Database [12], using BLAST [11] and aligns similar sequences using ClustalW [13].The alignment is then used to construct a phylogenetic tree, using the EBI ClustalW phylogeny service. The workflow shows 1A = WSDL Web Services, 1B = REST Web Service, 1C = Shim service, 1D = nested workflows (both expanded and collapsed), 1E = asynchronous service looping.



The Blast_Align_and_Tree workflow is an example Taverna workflow (available from http://www.myexperiment.org/workflows/3369.html), which performs a phylogenetic analysis. From an input protein Sequence, it performs a similarity search against the Uniprot Database [12], using BLAST [11] and aligns similar sequences using ClustalW [13].The alignment is then used to construct a phylogenetic tree, using the EBI ClustalW phylogeny service. The workflow shows 1A = WSDL Web Services, 1B = REST Web Service, 1C = Shim service, 1D = nested workflows (both expanded and collapsed), 1E = asynchronous service looping.