Scientific Workflows for Bibliometrics

Arzu Tugce Guler[1], Cathelijn J. F. Waaijer[2] and Magnus Palmblad[1]

[1]*a.t.guler@lumc.nl, n.m.palmblad@lumc.nl*
Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden (The Netherlands)

[2]*c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, Leiden (The Netherlands)

Abstract
Scientific workflows organize the assembly of specialized software into an overall data flow and are particularly well suited for multi-step analyses using different types of software tools. They are also favourable in terms of reusability, as previously designed workflows could be made publicly available through the myExperiment community and then used in other workflows. We here illustrate how scientific workflows and the Taverna workbench in particular can be used in bibliometrics. We discuss the specific capabilities of Taverna that makes this software a powerful tool in this field, such as automated data import via communication with Web services, smooth data extraction from XML by XPath and various data analyses and visualizations with the statistical language R. The support of the latter allows integration of a number of recently developed R packages for bibliometric analysis. A number of simple examples illustrate the possibilities of Taverna in the field of bibliometrics and scientometrics.

Conference Topic
Methods and techniques

Introduction
Information processing permeates the scientific enterprise, generating and organizing knowledge about nature and the universe. In the modern era, computational technology enables us to automate data handling, reducing the need for human labor in information processing. Often information is processed in several discrete steps, each building on previous ones and utilizing different tools. Manual orchestration is then frequently required to connect the processing steps and enable a continuous data flow. An alternative solution would be to define interfaces for the transition between processing layers. However, these interfaces then need to be designed specifically for each pair of steps, depending on the software tools they use; which compromises reusability. Whether the data flow is automated or done by the researcher manually, the latter still has to deal with many low-level aspects of the execution process (Gil, 2008).

Scientific workflow managers connect processing units through data and control connections and simplify the assembly of specialized software tools into an overall data flow. They smoothly render stepwise analysis protocols in a computational environment designed for the purpose. Moreover, the implemented protocols are reusable. Existing workflows can be shared and used by other workflows, or they can be modified to solve different problems. Several general purpose scientific workflow managers are freely available, and a few more optimized for specific scientific fields (De Bruin, Deelder, & Palmblad, 2012). Most of these managers provide visualization tools and have a graphical user interface, e.g. KNIME (Berthold et al., 2007), Galaxy (Goecks, Nekrutenko, & Taylor, 2010) and Taverna (Oinn et al., 2004). Not surprisingly, scientific workflows are now becoming increasingly popular in data intensive fields such as astronomy and biology.

In this paper, we describe the use of scientific workflows in bibliometrics using the *Taverna Workbench*. Taverna Workbench is an open source scientific workflow manager, created by

the myGrid (Stevens, Robinson, & Goble, 2003) project, and now being used in different fields of science. Taverna provides integration of many types of components such as communication with Web Services (WSDL, SOAP, etc.), data import and extraction (XPath for XML, spreadsheet import from tabular data), and data processing with Java-like Beanshell scripts or the statistical language R (Wolstencroft et al. 2013). Beanshell services allow the user to either program a small utility from scratch and towards a specific goal, or to integrate already existing software in the workflow. The R support is a particularly powerful feature of Taverna. Although R was initially developed as a language for statistical analysis, its widespread use has seen it adopted for many tasks not originally envisioned—a fate not unlike its commercial cousin, MATLAB. One such task is text mining. The R package *tm* (Feinerer, Hornik, & Meyer, 2008) provides basic text mining functionality and is used by a rapidly growing number of higher-level packages, such as *RTextTools* (Jurka, Collingwood, Boydstun, Grossman & van Atteveldt, 2014), *topicmodels* (Grün & Hornik, 2011) and *wordcloud* (Fellows, 2013). Similarly, there are many toolkits and frameworks for text mining in Java that could also be called from within a Taverna workflow.

An Example Workflow

We designed a simple workflow, *compare_two_authors* (see below), to generate a histogram for the number of publications over time and a co-word map for the titles of the two authors' publications. The workflow takes as inputs PubMed results in XML, the names of two authors, a list of excluded words and a minimum number of occurrences.

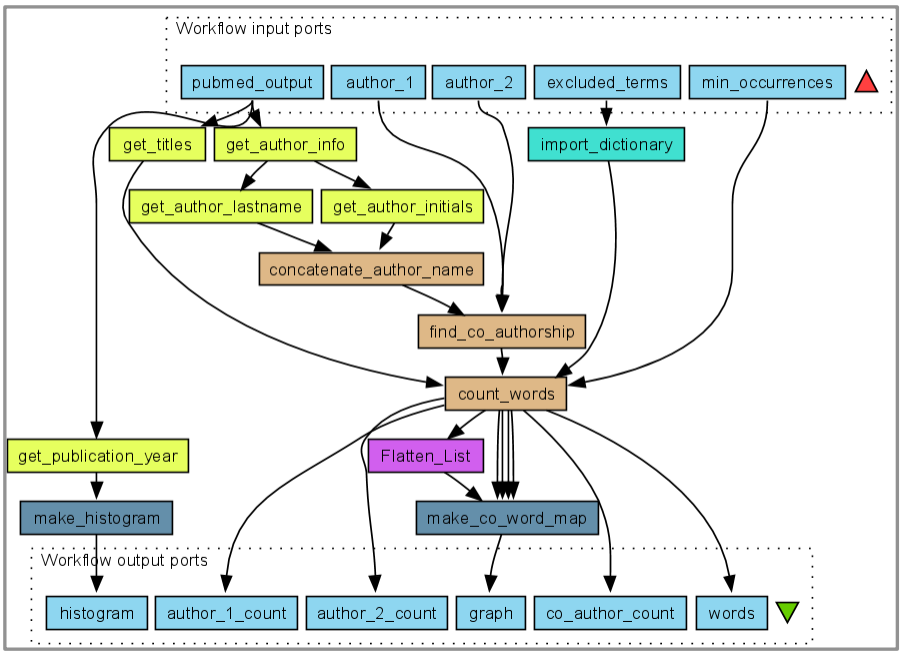

Figure 1. A workflow designed in Taverna for analyzing scientific output over time and comparing word usages of two authors.

The excluded terms are contained in a text file, so the *spreadsheet import* service in Taverna is used to extract each word in the file, line by line. The PubMed results are in XML format, and the extraction of publication years, titles and author names are done by *XPath* services. XPath is a query language for selecting elements and attributes in an XML document. The XPath service in Taverna eases this process by providing a configuration pane to render an XML file of interest as a tree and automatically generate an XPath expression as the user

selects a specific fragment from the XML (Fig. 2). The results of the query can either be passed as text or as XML to other workflow components.

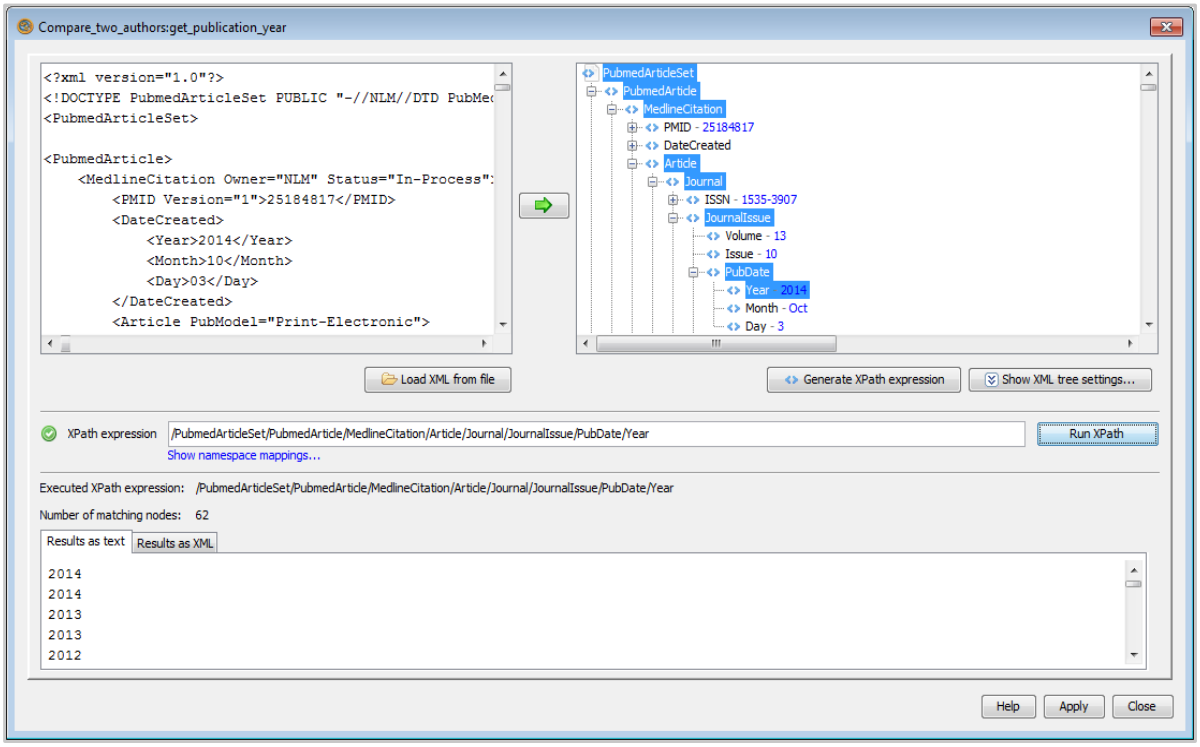

Figure 2. XPath configuration pane for extracting publication year from PubMed XML.

The data extracted by the spreadsheet import and XPath services is fed to a series of Beanshell components that find co-authorships and count co-occurrence of words in the extracted titles. Beanshell is a light-weight scripting language that interprets Java. In our workflow, the Beanshell services do simple operations on strings, such as concatenation of surnames and initials that are extracted separately using XPath (*concatenate_author_names*), matching strings to find co-authorships (*find_co_authorship*) and counting the number of words occurring in each title authored by one or both authors (*count_words*). The two authors' usage of the words, excluding *excluded_terms*, that appear at least *min_occurrences* times in total, are then used to draw a co-word map using the *igraph* (Csárdi & Nepusz, 2006) R package. It is generally up to the workflow designer what part of the workflow to code in Java (Beanshell), in R, or in third language called via the *Tool* command-line interface. More types are available for data connectors between R components (logical, numeric, integer, string, R-expression, text file and vectors of the first four types) than between Beanshell components, where everything is passed as strings. When dealing with purely numerical data, we recommend R over Beanshells within Taverna.

After all the necessary inputs are provided, the workflow is ready to be executed. In the Taverna Workbench *Results* perspective (Fig. 3), each completed process is grayed out to show the progress of the workflow run. The execution times, errors and results are also visible in this perspective.

We ran the workflow for two scientists active in our own field, mass spectrometry, Gary L. Glish and Scott A. McLuckey, whom we knew to have worked on similar topics and also co-authored a number of papers. However, the workflow will work on any two authors with publications indexed by PubMed. The co-word map in Figure 4 visualizes the co-occurrence of words in titles by the location and thickness of the connecting edge, while the relative frequency of usage by the two authors is indicated by the color (from white to gray).

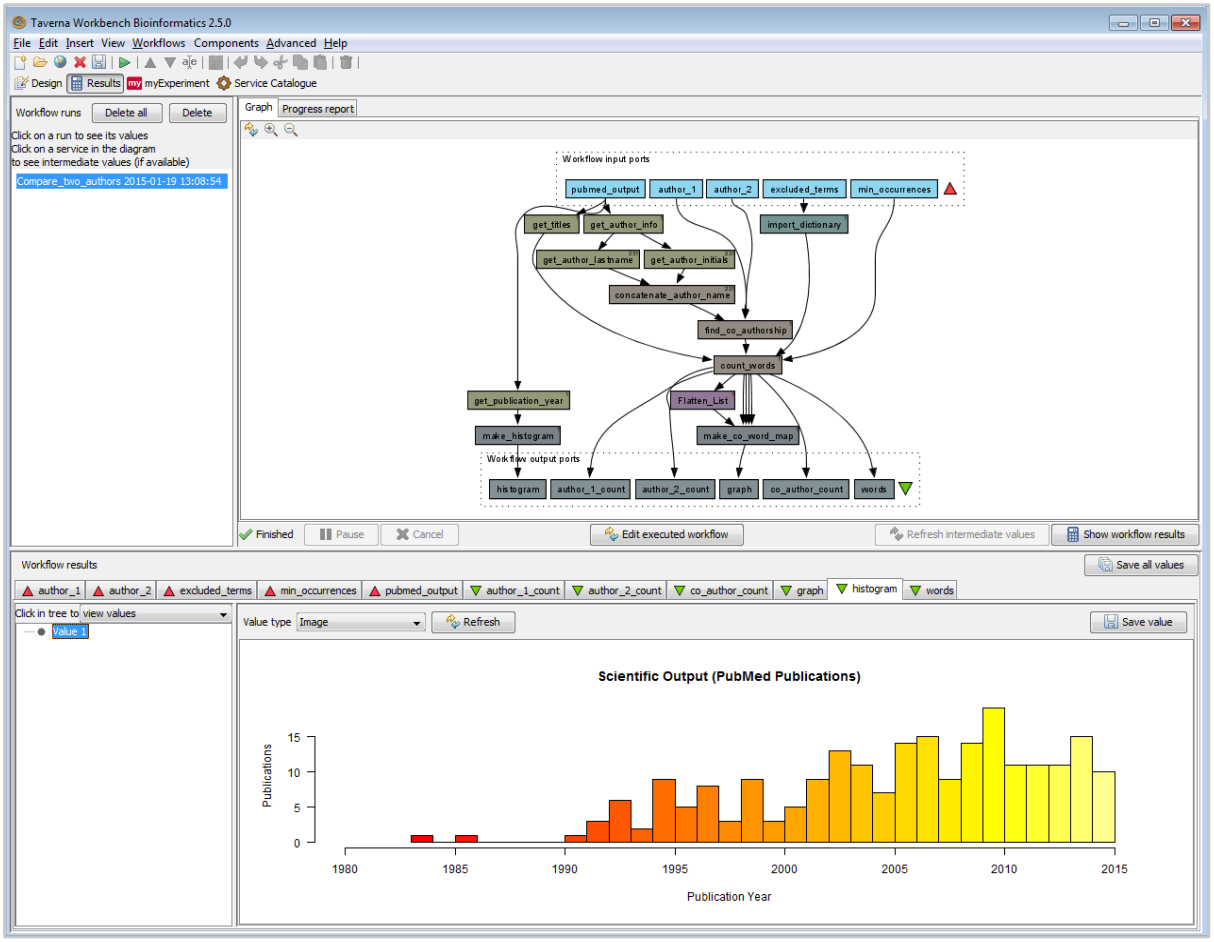

Figure 3. Workflow progress and output in the Taverna workbench Results perspective.

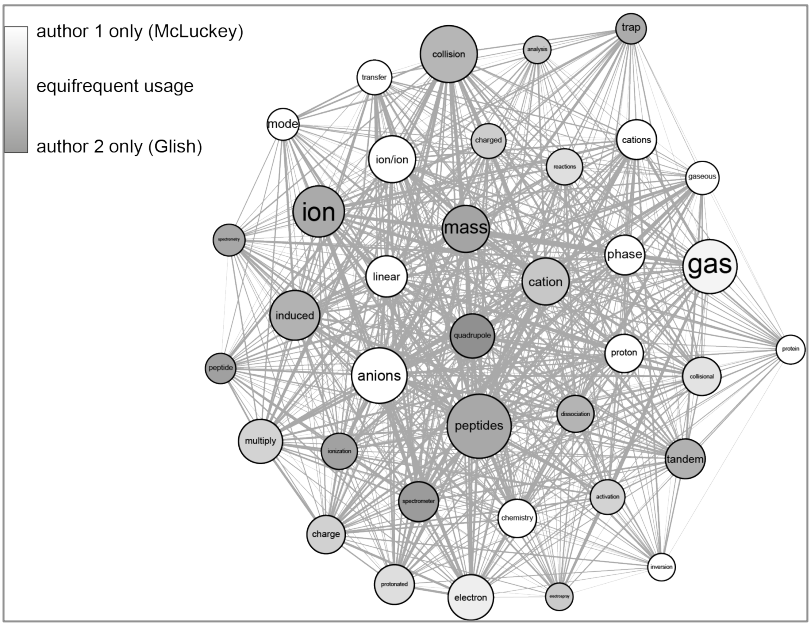

Figure 4. Co-word map output from the *compare_two_authors* workflow.

Connecting to Web Services and External Databases

Automatically generating networks directly from online data is also possible in Taverna workbench. Taverna can invoke WSDL (Web Services Description Language) style Web services given the URL of the service's WSDL document. The WSDL is an XML-based interface description language often used together with a SOAP (Simple Object Access protocol) to access the functions and parameters of a service. Many bibliographic resources are available through Web services, such as Web of Science (WoS). Some services, including the WoS, require authentication. An entire bibliometric study can be contained inside a single Taverna workflow that takes the user queries, or questions of the study, generate the Web service requests, execute these, retrieve the data and proceed with further (local) bibliometric and statistical analysis, and visualization.

A Taverna workflow that invokes WSDL services from WoS to automatically execute a query may look like in the figure below. This Taverna workflow takes as input common search parameters and a generic WoS query string, and passes these to the Web service via the WoS WSDL interface. Values that have only one possible value, such as the language (English, "en") are here hard-coded in the workflow as *Text constants*.

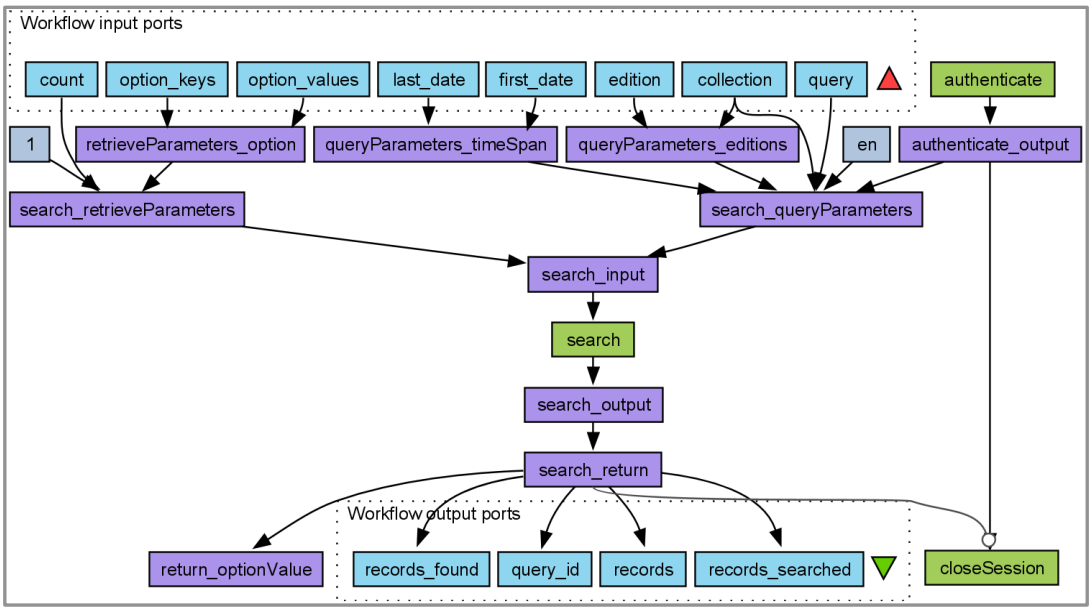

Figure 5. A simple workflow for retrieving bibliometric data using Web services.

Future Work

The use of scientific workflows in bibliometrics is still in its infancy. Modules that accomplish basic bibliometric tasks could be designed and combined in various ways for different studies, thus benefiting from modularity and reusability of scientific workflows. As mentioned above, the direct support of R inside Taverna workflows is particularly useful for bibliometrics. A number of R packages for bibliometric analysis have recently been released, ranging from simple data parsers such as the *bibtex* package (Francois, 2014) for reading BibTeX files to libraries or collections of functions for scientometrics, such as the *CITAN* package (Gagolewski, 2011). The latter package contains tools to pre-process data from several sources, including Elsevier's Scopus, and a range of methods for advanced statistical analysis. The *igraph* package itself comes with some functions specifically for bibliometric analysis, e.g. *cocitation* and *bibcoupling*. Clustering or rearranging the graph spatially so that strongly connected words appear closer together is possible with *igraph*, but may also be assisted by other packages. More crucially, the example workflow here does not yet

implement any advanced text mining functionality for homonym disambiguation or natural language processing. The *openNLP* R package provides an interface to openNLP (Hornik, 2014) and may be used to extract noun phrases and clean up the co-word maps.

Several of our Taverna workflows for bibliometrics and scientometrics, including the two workflows in Figure 1 and Figure 5, can be found in the myExperiment (Goble et al., 2010) group for Bibliometrics and Scientometrics (http://www.myexperiment.org/groups/1278.html). As always, we are grateful for any feedback on these workflows.

Acknowledgements

The authors would like to thank Dr. Yassene Mohammed for technical assistance and Thomson Reuters for granting access to the Web of Science Web services lite.

References

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (pp. 319-326). Heidelberg: Springer.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*, 1695.

De Bruin, J. S., Deelder, A. M., & Palmblad, M. (2012). Scientific Workflow Management in Proteomics. *Molecular & Cellular Proteomics*, *11*, M111.010595–M111.010595.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54.

Francois, R. (2014). bibtex: bibtex parser. R package version 0.4.0. Retrieved from http://CRAN.R-project.org/package=bibtex.

Fellows, I. (2013). wordcloud: Word Clouds. R package version 2.4. Retrieved from http://CRAN.R-project.org/package=wordcloud.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692.

Gil, Y. (2008). From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. *To Appear in Scientific Programming*, *16*(4), 1–25.

Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., & De Roure, D. (2010). myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, *38*(May), 677–682.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*, R86.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*, 1–30.

Hornik, K. (2014). openNLP: Apache OpenNLP Tools Interface. R package version 0.2-3. Retrieved from http://CRAN.R-project.org/package=openNLP.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E,. & van Atteveldt, W. (2014). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.4.2. Retrieved from http://CRAN.R-project.org/package=RTextTools.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., & Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054.

Stevens, R. D., Robinson, A. J., & Goble, C. a. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics (Oxford, England)*, *19 Suppl 1*(1), i302–i304.

Wolstencroft, K. et al. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research* 41(W1), W557-W561.