Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Automating bibliometric analyses using Taverna scientific workflows: A tutorial on integrating Web Services

Arzu Tugce Guler[a], Cathelijn J.F. Waaijer[b], Yassene Mohammed[a], Magnus Palmblad[a,*]

[a] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands
[b] Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 11 May 2016
Accepted 13 May 2016
Available online 22 June 2016

Keywords:
Bibliometrics
Europe PMC
PubMed
Taverna
Scientific workflows
R
Biomolecular interactions

ABSTRACT

Quantitative analysis of the scientific literature is a frequent task in bibliometrics. Several large online resources collect and disseminate bibliographic information, paving the way for broad analyses and statistics. The Europe PubMed Central (PMC) and its Web Services is one of these resources, providing a rich platform to retrieve information and metadata on scientific publications. However, a complete bibliometric analysis that involves gathering information and deriving statistics on an author, topic, or country is laborious when consuming Web Services on the command-line or using low level automation. In contrast, scientific workflow managers can integrate different types of software tools to automate multi-step processes. The Taverna workflow engine is a popular open-source scientific workflow manager, giving easy access to available Web Services. In this tutorial, we demonstrate how to design scientific workflows for bibliometric analyses in Taverna by integrating Europe PubMed Central Web Services and statistical analysis tools. To our knowledge, this is also the first time scientific workflow managers have been used to perform bibliometric analyses using these Web Services.

1. Introduction

As science becomes more data intensive, access to data and the process of generating meaningful information from them become the main vehicle in the scientific process. In this process, the primary challenge is moving from generated or retrieved data to information. As in most fields, typical bibliometric analysis workflows require several discrete steps, each employing different software tools. Frameworks that allow users to efficiently but easily connect data access points to information generation play a key role here. However, it is not always straightforward to use a generic framework or design custom workflows every time a new analysis protocol is to be implemented. In the absence of a framework, users have to manually connect the inputs and outputs of individual steps through the entire analysis. This risks introducing errors and makes analyses difficult to reproduce, especially for other researchers.

Scientific workflow managers integrate several processing units to automate a data analysis procedure. They are field-independent, so analysis on data from any field, including bibliometrics, can be automated. Scientific workflows typically

* Corresponding author.
E-mail addresses: a.t.guler@lumc.nl (A.T. Guler), cathelijnwaaijer@gmail.com (C.J.F. Waaijer), y.mohammed@lumc.nl (Y. Mohammed), n.m.palmblad@lumc.nl (M. Palmblad).

http://dx.doi.org/10.1016/j.joi.2016.05.002

have inputs and outputs, where series of operations are performed on the inputs in order to produce the outputs. Thus various atomic processing units can be assembled to produce an analysis protocol that can run without manual intervention (de Bruin, Deelder, & Palmblad, 2012). On the other hand, reusability and reproducibility are also important for *in silico* experiments, facilitating collaboration and combining efforts. These are promoted by online scientific workflow repositories such as myExperiment (Goble et al., 2010). However, deciphering the hierarchical composition of a workflow, its control and connections could be difficult in a larger-scale workflow (Lu and Zhang, 2009). Taking a modular approach and defining the scope of each module in the workflow eases this process. Most of the freely available scientific workflow managers have a graphical user interface that helps to visualize the overall protocol, both when designing and when executing the workflow. Galaxy (Goecks, Nekrutenko, & Taylor, 2010), KNIME (Berthold et al., 2008) and Taverna (Oinn et al., 2004) are popular examples of such scientific workflow managers that also allow modular design. Automating an analysis consisting of several steps, such as in bibliometrics, using scientific workflow managers makes the process less laborious and decreases the risk of human errors. Scientific workflow managers follow a different paradigm than interactive software tools, such as the domain-specific (or perhaps domain-limited) BibExcel (Persson, 2016), Publish or Perish (Harzing, 2007) and Sci2 (Sci2 Team, 2009) though Sci2 certainly provides some aspects of the modularity and tool integration of the workflow managers.

We have previously presented how scientific workflows can be used to solve simple bibliometrics problems, using Taverna Workbench (Guler, Waaijer, & Palmblad, 2016). Like any other scientific workflow manager, Taverna enables the user to integrate different types of components. What makes Taverna very useful for bibliometrics is that it already provides custom support for a number of tools and services that are easily adopted for performing such analyses, *e.g.*, R tools and XPath, Beanshell and WSDL services. The programming language R is primarily developed for statistical computing and visualization. Specific R plug-ins or packages expands its capabilities to machine learning, text mining and natural language processing (Feinerer, Hornik, & Meyer, 2008; Hornik, 2015). The XPath service is a user-friendly tool for creating XPath queries to parse XML documents by simply selecting nodes from an XML tree with a few mouse clicks. This is highly convenient, as most bibliometric databases can export information in XML format. For tabular formats, the Spreadsheet import service provides a similarly minimalistic tool for parsing tables. For general tasks, Beanshell services allow inclusion of scripts using a Java-like language. Last but not least, integrated support for Web Services allows Taverna workflows to directly communicate with remote databases using WSDL queries (Wolstencroft et al., 2013). As most Web Services use XML as the preferred message format, the Taverna XPath service is typically used to parse the results returned from Web Service calls.

An important functional aspect of Taverna is that iterations over individual processes or parts of the workflows are done implicitly by list handling. This feature provides great flexibility if a process or a sub-workflow has more than one input port. The user can specify whether the inputs are subjected to a “cross product” (all list elements in one input against all list elements in the other input) or a “dot product” (element-wise), or for processors with more than two inputs a combination of both; all while being able to define the order and precedence of the workflow operations on these input lists. A core set of built-in features and services provides basic list handling, such as flattening, merging a list to a string and removing duplicates.

Here we present a tutorial on how to use Taverna to build workflows that interact with the Europe PubMed Central Web Services. In principle, Taverna could interact with any Web Service that provide a SOAP or RESTful interface. The reason we are demonstrating the integration of Web Services in Taverna using PubMed rather than Scopus® or Web of Science™ (Falagas, Pitsouni, Malietzis, & Pappas, 2008) is that, among these three, PubMed is currently the only that provides a free Web Service interface. PubMed is also the most used bibliographic resource in the life sciences. In this tutorial, we show how to retrieve information using Web Services, how to parse this information, and how to use the various built-in Taverna services and processors to calculate and visualize the results. In principle, the same approach could be taken using other resources, provided that the user has access to them. We also made an example Taverna interface for connecting to the Thomson Reuters Web of Science™ Web Services and made this available on myExperiment (http://www.myexperiment.org/workflows/4705.html). We have built and tested the workflows in Taverna Workbench Bioinformatics 2.5.0, but in principle the workflows should run in any flavor of Taverna Workbench version 2.4.0 or later. For instructions on how to download and install Taverna, see www.taverna.org.uk. For Rshells to be executable in Taverna, R, RServe and required R packages must be installed and deployed (Williams, 2014).

2. Getting started: connecting to Europe PMC Web Services

Europe PubMed Central, or PMC (http://europepmc.org) is one of the leading databases for peer-reviewed life science literature, providing access to 30.4 million abstracts and 3.3 million full-text articles and metadata (December 14, 2015). The goal of Europe PMC is to “build open, full-text scientific literature resources and support innovation by engaging users, enabling contributors and integrating related research data” (The Europe PMC Consortium, 2015). This is achieved by providing access through a user-friendly Web interface, FTP, and SOAP and RESTful Web Service APIs. Here we will use the latter from within Taverna workflows. This is done as follows. First, the Europe PMC SOAP-based Web Services are imported into Taverna using “Import new services” in the Design pane using the WSDL http://www.ebi.ac.uk/europepmc/webservices/soap?wsdl. The available Web Services should now be listed as available in Taverna services menu. The 55-page Europe PMC SOAP Web Service Reference Guide (Europe PMC, 2015) describes all details of the API to these newly imported services. Although strongly recommended, it is not absolutely necessary to read the entire manual before starting to integrate Europe PMC Web Services from within Taverna. A Web Service component is simply added to a workflow by dragging it from the