Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud, pages 1–4
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Store Scientific Workflows in SSHOC Repository

Cesare Concordia, Carlo Meghini, Filippo Benedetti
CNR - ISTI
Area della Ricerca CNR, via G. Moruzzi 1, 56124 Pisa, Italy
{cesare.concordia, carlo.meghini, filippo.benedetti}@isti.cnr.it

Abstract

Today scientific workflows are used by scientists as a way to define automated, scalable, and portable in-silico experiments. Having a formal description of an experiment can improve replicability and reproducibility of the experiment. However, simply publishing the workflow may be not enough to achieve reproducibility and re-usability, in particular workflow description should be enriched with provenance data generated during the workflow life cycle. This paper presents a software framework being designed and developed in the context of the Social Sciences and Humanities Open Cloud (SSHOC) project, whose overall objective is to realise the social sciences and humanities' part of European Open Science Cloud initiative. The framework will implement functionalities to use the SSHOC Repository service as a cloud repository for scientific workflows.

- **Keywords:** Research infrastructure components, scientific workflows, reproducibility

1. Introduction

Workflows were initially used in the business environment as a way to describe the flow of activities through an organization and were later adopted also for scientific applications. Today scientific workflows (Qin and Fahringer, 2012) are used by scientists as a way to define automated, scalable, and portable in-silico experiments. In recent years a number of studies have been made concerning the use of workflows in the Social Sciences and Humanities (SSH) scientific community (Turner and Lambert, 2015; Matthew and Shapiro, 2014). These studies, starting from the consideration that most SSH researchers create or reuse scripts (written in such programming languages as R, Python, Haskell etc) in their activities, introduce approaches on how to build scientific workflows for complex experiments, starting from these scripts. A researcher can consider scripts as building blocks and use a Workflow Management Systems (WMS) to: relate scripts using graphical notation, execute them, access and manage data, monitor processes and analyse results. In most cases it is not required a strong technical skill to build scientific workflows (Turner and Lambert, 2015), scripts can be seen as black boxes having input parameters and producing outputs, the user relies on the WMS functionalities to deal with many technical details. Scientific workflows are considered a way for researchers to formally describe complex scientific experiments, and it is becoming a widely adopted practice among researchers to publish scientific workflows, alongside with datasets, in order to enable reproducibility and replicability of experiments.

The Social Sciences and Humanities Open Cloud (SSHOC) project[1] aims at realising the transition from the current SSH landscape with separated e-infrastructure facilities into a cloud-based infrastructure offering a scalable and flexible model of access to research data and related services adapted to the needs of the SSH scientific community. In particular the project will generate services for optimal re-use of data by making data Findable, Accessible, Interoperable and Re-usable (FAIR). In this context it is important to provide a service enabling researchers to publish scientific workflows to enable reproducibility of experiments.

This paper describes a software framework that is being designed and implemented in the SSHOC project, to enable scientists and researchers in the SSH domains to use the SSHOC Repository as a repository for publishing the scientific workflows used in their experiments. The document first presents an overview of scientific workflows, then reports the major guidelines suggested in scientific literatures for storing and publishing workflows and in its last part presents the frameworks being developed.

2. Scientific Workflows Overview

A scientific workflow is a composition of interconnected and possibly heterogeneous scripts that are used in a scientific experiment. Scientific workflow languages provide statements to define the logic that relates calls of scripts; for certain processes, such as statistical analysis, a linear flow might be sufficient, but more complex flows may allow for parallel execution, event handling, compensation handling and error handling. According to (Barga and Digiampietri, 2008) a scientific workflow may be considered as a way to record the origins of a result, how it is obtained, experimental methods used, machine calibrations and parameters, etc. Examples of scientific workflows are: data chaining pipelines that gather and merges data from multiple sources, sequence of steps automating repetitive tasks (e.g. data access, data transformation), complex iterative chains of MapReduce jobs etc. Scientific workflows are created and managed using specific software frameworks called Scientific Workflow Management Systems (SWMS). An SWMS implements the execution of the scripts, manages the allocation of computational resources and the input and output of data ("data staging"), deploys software, cleans up

[1] https://sshopencloud.eu

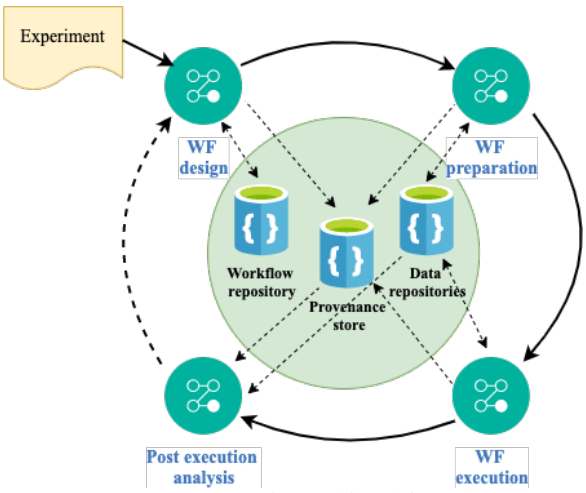

Figure 1 Scientific workflow life cycle

temporary data etc. Examples of SWMS are Kepler[2] and Taverna[3]. The life cycle of scientific workflows is composed of four main phases (Ludäscher et al, 2009): design, preparation, execution and post-execution analysis.

During the design and preparation phases, researchers may want to reuse pre-existing workflows (partly or as a whole) to create the new workflow. The SWMS provides functionalities to access local or remote[4] *workflow repositories*.

During execution phase, existing datasets are processed and new datasets can be generated. These datasets are accessed/stored by scripts, but the SWMS tracks these operations, and if necessary activates compensation handling procedures.

Every phase of a workflow life cycle generates provenance data, it is important to collect this data and store it. Provenance data of scientific workflows represents the entire history of the derivation of the final output of a workflow (Tan, 2007), it includes global configuration parameters, data propagation, data provenance of scripts, user annotations, performance and memory footprint etc. This data is used in the post-execution analysis phase: researchers evaluate data products and provenance information in order to validate the experiment. Provenance data is crucial to improve the reproducibility of workflows (Simmhan et al 2005), (Deelman et al. 2018).

3. Publishing Scientific Workflows

In principle, having a formal description of an experiment as a workflow (or as a script) can improve replicability and reproducibility of the experiments[5]:

- reproducibility: obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.
- replicability: obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

According to the above definitions, reproducing experiments involves using the original data and code, while replicating it involves new data collection and similar methods used by previous studies. Today it is an established behaviour in scientific communities to publish datasets used in experiments alongside with the scripts or workflows developed to process the datasets. However, according to many studies, this practice may not be sufficient to guarantee re-usability of datasets and reproducibility of experiments. This section focuses on issues on reproducibility of scientific workflows.

After February 2011 the journal Science adopted a policy that requires researchers to fulfil all reasonable requests for the data and code needed to generate results published in their papers. In a study, (Stodden et al. 2018) tested the reproducibility of results from a random sample of 204 scientific papers published in the journal after February 2011; they obtained data, scripts or workflows for 89 articles in their sample, and results could only be reproduced (with some efforts) for 56 articles, about 27% of total. In his study (Chen 2018) analysed all datasets published from 2015 to 2018 in the Harvard Dataverse[6] containing R scripts to reproduce results. His work concludes that 85.6% of stored R programs, when re-executed, generate several kinds of 'fatal' errors; only a subset of scripts runs correctly after debugging operations, while a significant number of scripts remains not usable. According to both studies a major reason for the reproducibility issues is the lack of provenance data, especially the lack information about the computational context of the scripts: library or external software packages dependencies, specific datasets versions, random or pseudo-random input values, etc.

The importance of capturing and storing provenance data to improve reproducibility of e-science experiments is outlined in several studies. In particular (Deelman et al.) clearly states that provenance data is necessary for reproducibility of scientific workflows.

4. The SSHOC Repository

One of the goals of SSHOC is to provide an European Open Science Cloud[7] (EOSC) repository service. An EOSC service can be defined as a resource that provide EOSC System Users with ready-to-use facilities. EOSC Services are supplied by a Service Provider in

[2] https://www.cct.lsu.edu/~sidhanti/classes/csc7700/papers/Ledashner05.pdf

[3] https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.1235

[4] E.g. myexperiment.org

[5] https://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm

[6] https://dataverse.harvard.edu

[7] https://www.eosc-portal.eu

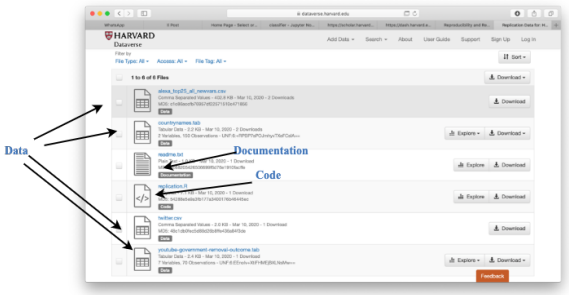

Figure 2 Example of datasets and scripts published in Harvard Dataverse

accordance with the Rules of Participation for EOSC Service Providers[8]. The SSHOC EOSC Repository service will provide SSH institutions without a repository service, such a facility for their designated communities. For organizations with limited technical resources, the service offers an opportunity to simply and effectively create an online repository. For organizations, which already provide archival solutions, this service can be used to set up a sharing and self-depositing environment for researchers in a user-centric manner.

The SSHOC EOSC Repository service is built upon the Dataverse software. The Dataverse is an open source web application designed to share, preserve, cite, explore, and analyse research data. Dataverse development is being coordinated by the Harvard's Institute for Quantitative Social Science (IQSS)[9]. Dataverse provides (among others) the following functionalities:

- A data citation with a persistent identifier (DOI)
- Standard metadata, plus custom metadata for journals
- Tiered access to data as needed: Fully Open, CC0, Register to access; Guestbook, Restricted
- Anonymous dataset review
- Versioning of datasets
- FAIR principles support
- Support for provenance (under development)[10]

Moreover, Dataverse allows integrations with other data services such as DataCite or ROpenScience. A Dataverse repository is a software installation, which hosts multiple virtual archives called *dataverses*. Each dataverse can contain several datasets, and each dataset contains descriptive metadata, code and data files. The Dataverse architecture implements the Service Oriented Architecture (SOA) principles, and provides APIs that can be used by developers to integrate micro-services on top. This last feature will

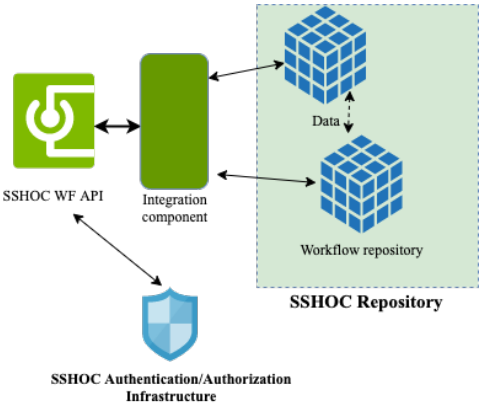

Figure 3 Overall architecture of SSHOC Workflow Repository API

be used as entry point for the software framework developed in this activity.

5. The SSHOC Workflow Repository API

The SSHOC Workflow Repository API is a software framework that can enable an SWFMS to use the SSHOC Repository as a workflow repository.

This software will implement a set of functionalities that can enable an SSH researcher to use the SSHOC Repository as

- a repository where to store and publish workflows alongside witjh the datasets used in her/his research
- a workflow repository that can be browsed and searched, for instance to re-use stored workflows in the design phase of new scientific workflows

Technically speaking the software will be composed of two main components (Fig. 3): an API publishing functionalities as Web Services and a middleware implementing the integration layer with the SSHOC Repository. A client application can use the SSHOC WF API to enable users to access workflows, download them, and execute or reuse to build new workflows.

A key challenge of the work is the definition of a data model for representing workflows. The general idea is to investigate a correct way to 'enrich' workflows description to improve both reproducibility of experiments and reusability of workflow as part of other workflows. As previously discussed data provenance has the potential to address a number of reproducibility issues. Provenance data for scientific workflows are collected by SWMS (observed provenance) and stored in local repositories or log files. The data provenance is currently mainly used to monitor the workflows behaviour and to enable an

[8] https://www.eosc-portal.eu/glossary
[9]https://www.iq.harvard.edu/product-development
[10]https://projects.iq.harvard.edu/provenance-at-harvard/tools

accurate post execution analysis. However, at the moment there is not yet a standard data provenance model for workflows (Delman et al. 2018), therefore in the first phase we have started to investigate the main approaches followed such as OPMW[11] or D-OPM (Cuevas-Vicenttín et al, 2012). They are based on W3C Open Provenance Model specification and describes workflows as graphs whose nodes are tasks and edges are relationships between tasks. These models provide very few specifications for provenance data and this could be an issue.

The SWMS Apache Taverna will be used to create a reference implementation for a client of the SSHOC WF API.

The Apache Taverna is an open source and domain-independent Scientific Workflow Management System, the data model used by Taverna is compatible with the W3C Open Provenance Model.

In particular there will be developed a plugin to enable Taverna Workbench users to use the SSHOC Repository. The Taverna Workbench is a tool that enables users to create, configure, execute and manage Taverna workflows, using a GUI. It is designed as a plugin platform, this means that its functionalities can be extended by installing new plugins.

The Taverna-SSHOC Repository plugin will be initially internally used to test developed software, and in a later stage it will be released via a public Maven repository to enable SSH scientists using Taverna to use its functionalities. The SSHOC WF Repository API and the plugin will be developed using Java based technologies.

6. Conclusion

This paper has presented a software framework for enabling SSH researchers to use the SSHOC Repository to store and publish scientific workflows. The software framework will technical implement the integration layer between the SSHOC Repository and a generic SWMSs, thus enabling users to store and access workflows, improving reproducibility of experiments and re-use of code. This activity is in progress: the design of the software is completed and design documents is going to be released in the following months. A first (alpha) release of the SSHOC WF API has been developed and deployed on development servers and is currently being tested. Technical documentation of the Web Services are available on line[12] while the source code will be published on SSHOC development repository.

7. Bibliographical References

Bertram Ludäscher, Mathias Weske, Timothy McPhillips, and Shawn Bowers. Scientific workflows: Business as usual?,7th Intl. Conf. on Business Process Management (BPM), LNCS 5701, Ulm, Germany, 2009 DOI: 10.1007/978-3-642-03848-8_4

Chen, Christopher Coding Be eR: Assessing and Improving the Reproducibility of R-Based Research With containR (2018). http://nrs.harvard.edu/urn-3:HUL.InstRepos:38811561

Deelman, E., Peterka, T., Altintas, I., Carothers, C. D., van Dam, K. K., Moreland, K., … Vetter, J. (2018). The future of scientific workflows. The International Journal of High Performance Computing Applications, 32(1), 159–175. https://doi.org/10.1177/1094342017704893

Gentzkow, Matthew and Jesse M. Shapiro. “Code and Data for the Social Sciences: A Practitioner's Guide.” (2014).

J. Qin and T. Fahringer, editors. Scientific Workflows – Programming, Optimization, and Synthesis with ASKALON and AWDL. Springer, Berlin, Germany, Aug. 2012.

Record, 34(3):31, 2005.

Roger S. Barga Luciano A. Digiampietri Automatic capture and efficient storage of e-Science experiment provenance. Concurrency Computat.: Pract. Exper. 2008; 20:419–429

Simmhan Y L, Plale B, and Gannon D. A survey of data provenance in e-science. ACM SIGMOD

T. Pasquier, M. K. Lau, X. Han, E. Fong, B. S. Lerner, E. Boose, M. Crosas, A. Ellison, and M. Seltzer, “Sharing and Preserving Computational Analyses for Posterity with encapsulator,” ArXiv e-prints, Mar. 2018.

Tan, W. C. Provenance in Databases: Past, Current, and Future. IEEE Data Engineering Bulletin, 30(4):3–12, Dec. 2007.

Turner, K.J., Lambert, P.S. Workflows for quantitative data analysis in the social sciences. Int J Softw Tools Technol Transfer 17, 321–338 (2015). https://doi.org/10.1007/s10009-014-0315-4

V. Cuevas-Vicenttín, S. Dey, M. L. Y. Wang, T. Song and B. Ludäscher, "Modeling and Querying Scientific Workflow Provenance in the D-OPM," 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, Salt Lake City, UT, 2012, pp. 119-128.

V. Stodden, J. Seiler,and Z. Ma, “An empirical analysis of journal policy effectiveness for computational reproducibility,” Proceedings of the National Academy of Sciences, vol. 115, no. 11, pp. 2584–2589, 2018.

[11] https://www.opmw.org

[12] http://146.48.85.197/Dataverse_tool-0.0.1-SNAPSHOT/swagger-ui.html#/