Implementing W2Share: Supporting Reproducibility and Quality Assessment in eScience

Lucas A. M. C. Carvalho[1], **Joana E. Gonzales Malaverri**[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – University of Campinas (Unicamp)
Campinas – SP – Brazil

{lucas.carvalho,jmalav09,cmbm}@ic.unicamp.br

Abstract. *An open problem in scientific community is that of supporting reproducibility and quality assessment of scientific experiments. Solutions need to be able to help scientists to reproduce experimental procedures in a reliable manner and, at the same time, to provide mechanisms for documenting the experiments to enhance integrity and transparency. Moreover, solutions need to incorporate features that allow the assessment of procedures, data used and results of those experiments. In this context, we designed W2Share, a framework to meet these requirements. This paper introduces our first implementation of W2Share, which moreover guides scientists in step-by-step process to ensure reproducibility based on a script-to-workflow conversion strategy. W2Share also incorporates features that allow annotating experiments with quality information. We validate our prototype using a real-world scenario in Bioinformatics.*

1. Introduction

Reproducibility denotes the ability for a third party to reproduce results of an experiment using a possibly different setting, aiming at confirming or disputing the original experimenter's claims [Missier et al. 2016]. As we know, scientific transparency and integrity rely on the ability to reproduce experiments, e.g., for independent validation, adoption of procedures or building new solutions to move forward in a particular research domain.

Scripts and Scientific Workflow Management Systems (SWfMSs) are common approaches that have been used to allow the automation of processes and data analysis in experiments. Scripts are widely adopted in many disciplines to create pipelines in experiments, e.g., to clean and analyze a large amount of data. However, they are hard to understand, adapt and reuse. For this reason, several solutions have been proposed to help experiment reproducibility for script-based environments such as ReproZip [Chirigati et al. 2016], YesWorkflow [McPhillips et al. 2015] and noWorkflow [Murta et al. 2014]. Though those solutions help scientists to capture experimental details, they neither allow to fully document the experiment and nor add new additional information such as support quality assessment of the experiments. SWfMSs [Liu et al. 2015], on the other hand, help reproducibility by supporting scientists in the design and execution of their experiments, which are specified and run as interconnected (reusable) components. However, there is a gap between the script and the workflow communities. Moreover, workflows alone are not enough to ensure reproducibility.

Taking this overall scenario into account, we designed W2Share, a framework for retrieval and conversion of script-based experiments into executable workflows [Carvalho et al. 2016b]. The script-to-workflow conversion is based on a methodology

proposed by us [Carvalho et al. 2016a]. Reproducibility is enabled, via this methodology, by the adoption of Workflow Research Objects (WRO) [Belhajjame et al. 2015]. The WRO model allows the aggregation of resources, explicitly specifying the relationship between these resources and workflow using a suite of ontologies. A WRO encompasses information such as datasets and provenance traces related to the execution of workflows. In W2Share, a WRO also encapsulates the scripts that were transformed into workflows and quality annotations. Via W2Share, third-party users are thus able to understand the data analysis encoded by the original script and obtain the resources required to run or reuse the associated workflow and data.

This paper presents the first implementation of W2Share that implements script conversion and quality assessment. The prototype is available at http://w3id.org/w2share. It incorporates our work on Quality Flow, [Sousa 2015, Sousa et al. 2014] – a workflow-based computational framework for data quality assessment of scientific experiments. Our solution can be used in different situations such as: (i) publishing procedures and datasets related to an experiment; (ii) training members in a research group to gain skills in computational scientific procedures; and (iii) dynamically assessing the quality of experiments. This prototype was validated in a bioinformatics experiment. As discussed in the paper, through W2Share, we semi-automatically transformed a suite of R scripts into an executable workflow, which was annotated with quality information. Then, still under W2Share, several runs of this workflow were executed, each of which with potentially distinct quality annotations. The entire set (script, workflow, provenance traces, quality information) is encapsulated in WROs, that are stored in W2Share repository for WRO.

2. W2Share Instantiation

W2Share is an abstract generic framework to support executing and documenting experiments to enable their reuse and reproduction. As such, it can be instantiated in many different ways[1].

2.1. Overview

This section presents a specific instantiation, which moreover supports our methodology to guide scientists in the process of transforming scripts into workflows and their executions, with subsequent encapsulation into WROs. Figure 1 gives a high level overview of this instantiation, which is composed of three main modules: (i) **Script Converter** – responsible for guiding the scientist through the conversion of scripts into workflows; (ii) **WRO Manager** – responsible for creating, updating and exporting WRO bundles; and (iii) **Quality Flow** – responsible for annotating the workflow and provenance data with quality information, and creating quality according to users' needs.

These modules store and retrieve objects from the Knowledge Base (KB) using Semantic Web technology (in particular SPARQL queries). The KB encapsulates the WRO repository which includes scripts, workflows, provenance data, annotations, and input and output data; and the Quality Flow repository, which is responsible for storing quality dimensions, its metrics and creators. The KB is implemented using Virtuoso Open Source Edition[2].

[1]For brevity sake, we do not present the overall framework. For details see [Carvalho et al. 2016b].

[2]https://github.com/openlink/virtuoso-opensource

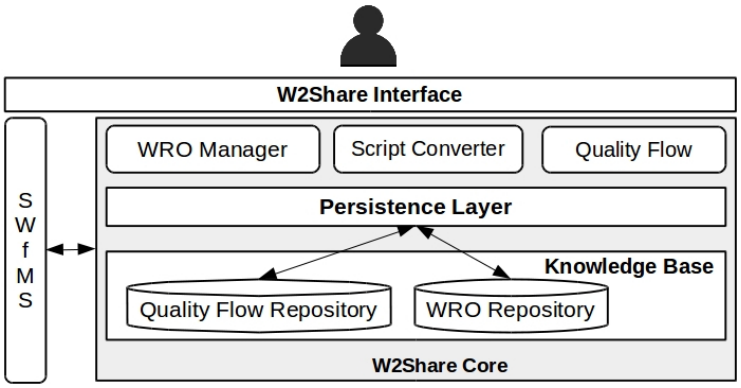

Figure 1. W2Share Software Architecture - workflow management is performed by the SWfMS.

The implementation uses a Model-View-Controller (MVC) architecture implemented using the PHP language and the Symfony Framework[3], a well established framework for web development. Among some of the tools and ontologies reused and/or extended in our software architecture are: YesWorkflow [McPhillips et al. 2015], Quality Flow [Sousa 2015, Sousa et al. 2014] and the WRO model ontologies [Belhajjame et al. 2015]. We also created an ontology for annotating data quality information (see section 2.3). These modules take advantage of the SWfMS to manage, design and execute workflows.

2.2. Script Converter

Script conversion works as follows. First, the scientist annotates the script with YesWorkflow tags (e.g., $@begin$ and $@end$ to delimiter a given task). This annotated script is fed into the YesWorkflow tool suite, integrated by us into W2Share. This suite produces a set of Datalog facts, and a visual rendering of the script as a workflow. Next, we developed code to process these facts to produce the executable workflow that corresponds to that visual rendering. From then on, this workflow can be executed, updated and tested using the underlying SWfMS. Our present code only allows creating workflows that can be executed in the Taverna system [Wolstencroft et al. 2013]. We chose Taverna due to its popularity and availability of tools to ease the process of creating an executable workflow, exporting provenance data and integrating with the WRO.

2.3. Quality Flow

In [Carvalho et al. 2016a], we state that there is a need to check the quality of the actual script-to-workflow conversion, and to assess the quality of the resulting workflow. In our solution, we adopt Quality Flow [Sousa et al. 2014] to allow scientists to annotate the elements that make up the workflow with quality information. By doing this, scientists are able to compute quality metrics related to data, processes, and overall execution, and request evaluation of specific quality metrics based on combining quality dimensions with provenance information generated at each run of the workflow.

Quality Flow is a software tool that allows domain experts to define their own quality dimensions and metrics for workflows and their components. Thus, a given work-

[3]http://symfony.org

flow (and even a given execution) may have multiple quality assessments, depending on each expert's point of view. Data quality metrics may be computed as simple equations on numeric values or more complex as a set of inference rules. On demand, these metrics are used to compute a dimension of quality, without need to modify the workflow structure. Quality dimensions and metrics can be defined at workflow creation or, little by little, as distinct scientists interact with the workflow. Such metrics define how to calculate quantitative quality dimensions like accuracy and efficiency or how to relate and summarize different qualitative quality dimensions like reliability, utility and so on. W2Share embeds these features from Quality Flow[4] responsible for managing, extracting and processing quality information. W2Share links the quality-annotated workflow with its runs and (provenance) traces. Distinct scientists can assess quality differently, defining distinct quality metrics for a quality dimension for a given piece of data or process. As a consequence, the result of a single experimental run can be assessed multiple ways. These distinct assessments are embedded into a WRO (see Section 2.4) to be published.

We created an ontology[5] to represent the quality information generated, thus allowing the adoption of semantic web technology integrated with the WRO ontologies and supporting the construction of inference rules to calculate the data quality metrics. Basically, the main entities of the ontology are: Quality Dimension, Quality Metrics and Quality Annotation. Quality Dimension describes quality properties, such as freshness or understandability. Quality Metrics defines functions to compute a specific quality dimension. Quality Annotation associates a quality dimension and quality metrics with a workflow and/or elements of the workflow on request.

2.4. WRO Manager

This module implements the last step in the methodology, aggregating all resources used or produced in an experiment and their quality annotations into a reproducible and reusable WRO bundle. The WRO management capabilities supported by W2Share include: (a) creating a WRO bundle; (b) exploring a WRO bundle created or uploaded into the system; (c) annotating resources; and (d) exporting a WRO bundle for publishing on other repositories or for sharing it directly with other scientists. We adopted the RO Manager tool[6] to create the WRO bundle files.

3. Case Study: DNA Methylation Microarray Analysis

Overview of the experiment Epigenome-Wide Association Studies (EWAS) examine the epigenetic status of many *loci* (a set of positions on a chromosome) for a set of individuals and assess whether any of these *loci* is associated with the phenotype of interest. To evaluate the feasibility of our instantiation, we implemented an EWAS experiment in W2Share. This experiment [Souza et al. 2014] was developed by the Biostatistics and Computational Biology Laboratory (BCBLab)[7] at Unicamp. It aimed at investigating how epigenetic marks change between two different tissues, prefrontal cortex and white blood cells, by assessing the DNA methylation profiles of control patients from a publicly available data set. The results were obtained comparing the profiles of these two tissues. The

[4]Though, we used only some Quality Flow's features, we maintained the same name in the text.
[5]http://w3id.org/w2share/ontologies/quality-flow.owl
[6]https://github.com/wf4ever/ro-manager
[7]http://bcblab.org/

steps involved in a typical differential DNA methylation analysis pipeline include: quality control, filtering, data exploration, normalization and statistical testing for identifying differentially methylated regions (DMR).

This experiment was implemented via a script (GSE37579_analysis) in the R language. The script uses as input data: (i) the Gene Expression Omnibus (GEO) accession number GSE37579[8]; (ii) symbol names of genes of interest; and (iii) names of the sample groups (tissue names) used to filter the public dataset. Data outputs at each run are files containing high-resolution graphic charts for manual inspection and files containing tables of the DMRs identified when compared the sample groups.

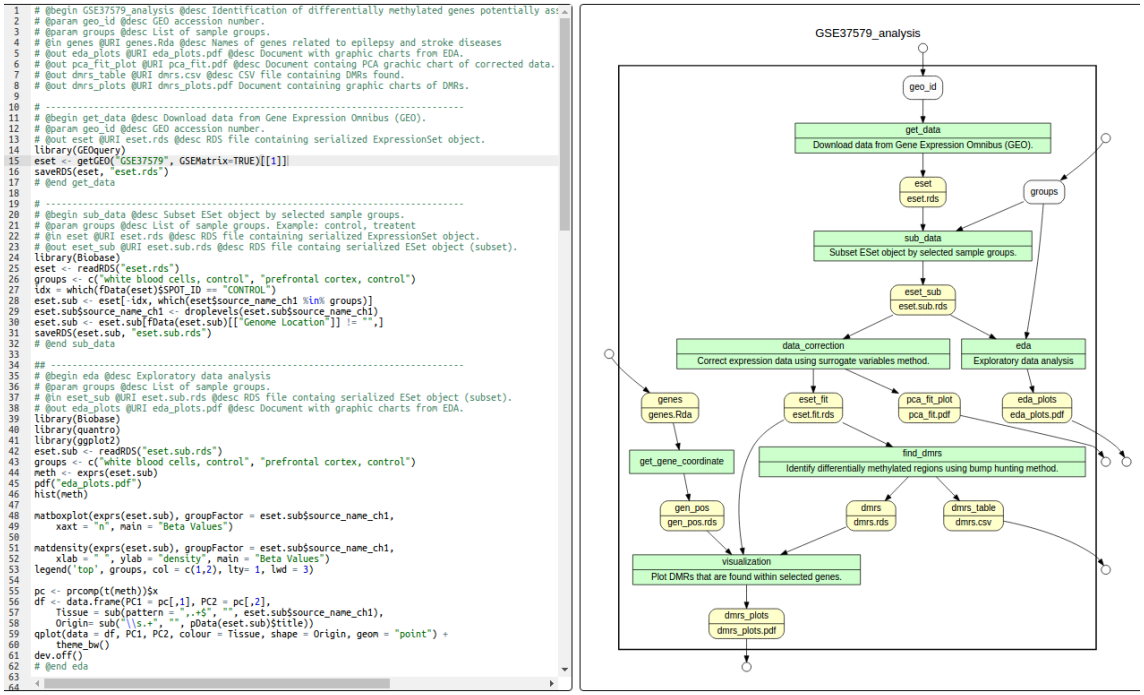

Figure 2. W2Share Editor for converting script into workflow.

Generating the executable workflow We first annotated script GSE37579_analysis using W2Share (see section 2.2). Figure 2 illustrates the annotated script (on the left side) and the abstract workflow generated (on the right side). Lines 11 to 17 from the script show tag $@begin$ identifying *get_data* activity, $@desc$ describing the activity as "Download data from Gene Expression Omnibus (GEO)", $@param$ and $@desc$ describing the parameter *geo_id* as "GEO accession number." and $@out$, $@uri$ and $@desc$ specifying the file *eset.rds* to the output port *eset* and describing this port as "the eset RDS file containing serialized ExpressionSet object". This first activity of the script is represented in the abstract workflow as the first green box, where the first row shows *get_data* and the second one, the description of the activity obtained from the $@desc$ tag.

W2Share next automatically creates the corresponding executable workflow[9]. In the current version of W2Share, we run the workflow using the Taverna workflow system

[8]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37579

[9]Our implementation generates Taverna workflows

– and thus, we capture all provenance information provided by Taverna.

Annotating with Quality information At any point, different scientists can annotate the workflow and components with quality information. Figure 3 shows a form to annotate the process *find_dmrs* with quality information.

As an example of how this form works, consider that at some point, a scientist $Lucas$ defined the *accuracy* quality dimension. Scientist $Joana$ retrieves from W2Share the information about GSE37579_analysis. $Joana$, then, annotates the process *find_dmrs* with the *accuracy* dimension value 0.85. The same scientist defines that the metrics to compute *accuracy* are $(correctly\_identified\_methylate\_regions\ /\ total\_selecting\_samples)$. At present, quality dimensions and metrics are stored as textual descriptions. This allows any kind of dimensions and metrics to be defined using W2Share. However, our approach still requires of manual specification of quality information. To overcome such a limitation, as future work, we intend to design a semi-automatic approach to define and assess the quality metrics associated with a dimension.

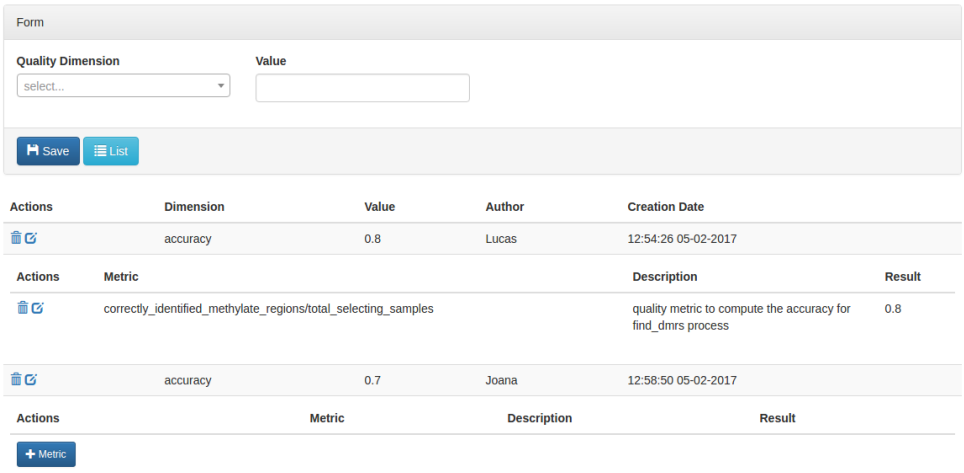

Figure 3. Quality Annotation form on W2Share.

Generating the WRO Finally, the scientist generates the WRO, which is stored in the corresponding W2Share repository. This output is available at http://w3id.org/w2share/usecase-bresci2017.

4. Related Work

There is a variety of work addressing reproducibility of script-based experiments. For instance, ReproZip [Chirigati et al. 2016] helps users to capture all the necessary components, environment variables and library dependencies used to execute data analysis, bundling them in a single, distributable package. However, unlike W2Share, ReproZip does not have a web interface to allow the exploration and annotation of the resources included in the package. Our solution provides a friendly web user interface to explore WRO bundles. noWorkflow [Murta et al. 2014] captures the execution provenance of Python scripts to support reproducibility. Our solution is not limited to a specific script language and we use the SWfMS to capture the provenance data related to the experiment execution. Furthermore, ReproZip and noWorkflow lack features to allow scientists to understand main script components to learn about the experiment. YesWorkflow

[McPhillips et al. 2015] is a tool that allow scientists to annotate their scripts to generate a workflow-like graphic view of the data analysis carried out by the script. In our work, we extend YesWorkflow features to allow transforming scripts components into executable workflow elements, storing the scripts and the abstract workflows into a WRO for reproducibility.

We use WRO as a component to package and publish the information of the experiment on W2Share. myExperiment [De Roure et al. 2007] and ROHub [Palma et al. 2014] are public web repositories that store Research Objects. ROHub is a repository for general-purpose Research Objects, whereas myExperiment stores WROs. Both repositories lack features such as annotation of quality information, available in W2Share, and none of them have a focus on exploring the provenance data of executions. Our solution is not limited to providing a repository of WROs, but also a system able to manage and enrich the WROs with quality information, activity descriptions, among others. The quality assessment facilities in W2Share are based on Quality Flow [Sousa et al. 2014]. We adapted Quality Flow to semantically represent its data and to work in an integrated manner with the modules of W2Share.

5. Conclusion and Future work

This paper presented our implementation efforts to instantiate the W2Share framework, which supports scientists in transforming their scripts into executable workflows, and assessing the quality of workflow runs. Our instantiation is enhanced with Quality Flow's features to annotate the workflows, the processes and the output data with quality information. This allows users to assess the quality of an experiment and create WROs extended with quality information to be consumed by other scientists. To the best of our knowledge, this is the first attempt to fully integrate use-tailored dynamic quality assessment to a reproducibility environment. Our implementation strived to reuse software tools, standards and ontologies developed by the scientific community. As such, the incorporation of quality assessment features is one of our major contributions of this work. Nevertheless, much need to be done to meet full support to reproducibility and quality management. One possible direction is to ensure support to data citation standards. Full reproducibility may moreover encompass preserving the original execution environment (e.g., variables and software configuration).

Automatic comparison of the quality of the experiment results based on the original script and the workflow is also left as future work. Finally, we desire to explore Common Workflow Language (CWL)[10] to create executable workflows to use a standard that works across multiple SWfMS.

Acknowledgements

Work partially financed by FAPESP (2014/23861-4), CAPES (1658841), FAPESP CCES (2013/08293-7) and CNPq/INCT in Web Science (557128/2009-9). We thank professor Benilton Carvalho and Welliton Souza from the BCBLab at Unicamp for their valuable support to construct and validate the case study.

[10]http://www.commonwl.org/

References

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., et al. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:16–42.

Carvalho, L. A. M. C., Belhajjame, K., and Medeiros, C. B. (2016a). Converting scripts into reproducible workflow research objects. In *Proc. of the IEEE 12th Int. Conf. on eScience, October 23-26*, pages 71–80, Baltimore, MD, USA. IEEE.

Carvalho, L. A. M. C., Silveira, R. L., Pereira, C. S., Skaf, M. S., and Medeiros, C. B. (2016b). Provenance-based retrieval: Fostering reuse and reproducibility across scientific disciplines. In *Proc. of the 6th IPAW, June 7-8, 2016*, pages 183–186. Springer.

Chirigati, F., Rampin, R., Shasha, D. E., and Freire, J. (2016). Reprozip: Computational reproducibility with ease. In *SIGMOD Conference*, pages 2085–2088. ACM.

De Roure, D., Goble, C., and Stevens, R. (2007). Designing the myexperiment virtual research environment for the social sharing of workflows. In *IEEE Int. Conf. on e-Science and Grid Computing*, pages 603–610. IEEE.

Liu, J., Pacitti, E., Valduriez, P., and Mattoso, M. (2015). A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, 13(4):457–493.

McPhillips, T., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., et al. (2015). Yesworkflow: A user-oriented, language-independent tool for recovering workflow information from scripts. *Int. Journal of Digital Curation*, 10(1):298–313.

Missier, P., Woodman, S., Hiden, H., and Watson, P. (2016). Provenance and data differencing for workflow reproducibility analysis. *Concurrency and Computation: Practice and Experience*, 28(4):995–1015.

Murta, L., Braganholo, V., Chirigati, F., Koop, D., and Freire, J. (2014). noworkflow: Capturing and analyzing provenance of scripts. pages 71–83.

Palma, R., Hołubowicz, P., Corcho, O., Gómez-Pérez, J. M., and Mazurek, C. (2014). Rohub—a digital library of research objects supporting scientists towards reproducible science. In *Semantic Web Evaluation Challenge*, pages 77–82. Springer.

Sousa, R. B. (2015). Quality flow: a collaborative quality-aware platform for experiments in escience. Master's thesis, Institute of Computing - University of Campinas.

Sousa, R. B., Cugler, D. C., Malaverri, J. E. G., and Medeiros, C. B. (2014). A provenance-based approach to manage long term preservation of scientific data. In *2014 IEEE 30th Int. Conf. on Data Eng. Workshops (ICDEW)*, pages 162–133. IEEE.

Souza, W., Carvalho, B., Dogini, D., and Lopes-Cendes, I. (2014). Identification of differentially methylated genes potentially associated with neurological diseases. In *ASHG 64th Annual Meeting, October 18-22, 2014*. American Society of Human Genetics.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., 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.