RO-Manager: A Tool for Creating and Manipulating Research Objects to Support Reproducibility and Reuse in Sciences

Jun Zhao[1], Graham Klyne[1], Piotr Hołubowicz[2], Raúl Palma[2], Stian Soiland-Reyes[3], Kristina Hettne[4], José Enrique Ruiz[5], Marco Roos[4], Kevin Page[6], José Manuel Gómez-Pérez[7], David De Roure[6], and Carole Goble[3]

[1] Department of Zoology, University of Oxford, Oxford, UK
jun.zhao, graham.klyne@zoo.ox.ac.uk
[2] Poznań Supercomputing and Networking Center, Poznań, Poland
piotrhol, palma@man.poznan.pl
[3] School of Computer Science, University of Manchester, Manchester, UK
soiland-reyes@cs.manchester.ac.uk, carole.goble@manchester.ac.uk
[4] Leiden University Medical Center, Leiden, NL
k.m.hettne, m.roos@lumc.nl
[5] Instituto de Astrofísica de Andalucía, Granada, Spain jer@iaa.es
[6] Oxford eResearch Center, University of Oxford, Oxford, UK
kevin.page, david.deroure@oerc.ox.ac.uk
[7] iSOCO, Madrid, Spain jmgomez@isoco.com

Abstract. In this position paper we present a lightweight command-line tool RO Manager, which provides a straightforward way for scientists to assemble an aggregation of their experiment materials and methods which can then be published and shared with colleagues or linked to scientific publications, to enhance the reproducibility and trustworthiness of experiment results. The tool is currently being tested by a small group of scientists from two different domains, who would like to preserve sufficient materials and information along with their scientific results in order to improve their reproducibility in the future.

1 Reproducibility and New Form of Digital Publication

There is a growing need for revolutionizing the existing practices of digital publishing, to take it beyond being a replicate of the paper form. The hypothesis is that digital papers can be greatly enhanced with additional features by making effective use of information technology, to accelerate the turn of knowledge [2]. The goals of these activities are multi-fold, ranging from enabling more efficient search and discovery of knowledge to supporting the reproducibility of science by promoting the sharing of data as well as tools and methods.

In this paper, we present an approach of aggregating and publishing a collection of auxiliary information together with experiment results, which can then

* The research reported in this paper is supported by the EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1).

be shared and linked in scientific publications in order to boost the reuse and reproducibility of these results. This aggregation of objects is represented using our Research Object (RO) model [1], which provides an aggregation structure for collecting essential resources related to experiment results along with publications. This includes not only the data used but also methods applied to produce and analyse that data, as well as auxiliary documents, scripts and software used in the research process. Built upon the RO model, we create a lightweight command-line tool called the Research Object Manager, or RO Manager. The goal of RO Manager is twofold: to ease the process of packaging necessary materials and methods together with experiment results in order to boost their reproducibility and hence reuse, and to ease the creation of new form of reproducible publications by making these aggregation objects sharable and citable.

Reproducibility of computational science has been widely explored in many existing scientific domains [3]. Recent efforts [8] have focused on building the tools and infrastructure to support the reproducibility of experiment results in publications. However, publishing reproducible papers requires preparation work prior to the final stage of experiment life cycle, and none of the existing work supports reproducible science from the early stage of the cycle. Neither is there an approach for automatically assessing and monitoring of the "health" of published materials and methods for supporting the reproducibility.

Existing studies have shown that the top barrier for the scientists to publish their results in a reproducible way is the time required for creating documentation [3]. Our RO Manager tool provides a lightweight solution for scientists to create a structured documentation about their reproducible experiment results in an environment most familiar to them, i.e. their local file systems, by simply executing a series of computer commands. Currently a manual validation process is commonly employed to validate the resources submitted by the scientists. However, this manual process is hard to scale and a continuous monitoring of the health of the aggregation (such as the accessibility of the aggregated resources) is entirely missing. The RO Manager tool takes one step further by providing a means to encode the requirements for the list of digital components to be submitted with experiment results in a machine-processable format so that we can evaluate that an RO contains all the necessary information required at the time of submission and monitor the health of these information.

These two gaps in supporting reproducible science and publication drove the design of our RO Manager tool, introduced in this paper.

2 What is RO Manager

RO Manager is a command line tool for creating, displaying and manipulating ROs. It is meant to provide a lightweight tooling for scientists to create ROs in an environment that is most familiar to them, i.e. their local file system, before publishing and sharing it in the open world. A command-line tool is the most lightweight choice for this purpose, which also provides the following additional advantages:

[8] `http://www.executablepapers.com/`

– Focus on the functionality of the tooling at the first stage of developments rather than graphical user interface (GUI) design.
– Provide users access control of their aggregation object before sharing it with the public or friends, which is crucial for scientists who want to protect their experiment resources before publishing them.
– Share the knowledge of its usage by simple shell script files, to demonstrate the usage of tool by executing a sequence of RO Manager comannds.

RO Manager is implemented as a Python program, using Python version 2.7 and available for download and installation at `https://github.com/wf4ever/ro-manager`. To date the RO Manager provides the following functionalities:

– Create and populate an RO: By executing the `ro create` command, the tool will automatically generate an RO structure in the local directory and a manifest file in RDF format to describe its content using RO ontologies [9].
– Annotate an RO or its component: Annotations can be provided directly, as values for specified attributes (title, type, etc.) or by attaching an existing RDF file to the metadata describing an RO, using the Annotation Ontology[10]. Some annotations can be automatically generated, describing who created the RO and when, while additional annotations, like document type, etc, have to be manually created, using the command `ro annotate`.
– Display the status of an RO and its annotations: All the annotations on the RO as well as on each of its components can be displayed by executing `ro annotations`.
– Evaluate the quality of the RO: We define the list of requirements for an RO to satisfy in a structured format, based on our Minim model [4]. Using this and the manifest file our evaluation component can assess whether an RO contains all the information required for supporting re-running an experiment or replicating a previous result, so that scientists can amend any missing resources before publishing their ROs.
– Publish an RO in a public RO repository: The resulting RO can be published in a web-based RO repository, becoming citable via a URI, which can be dereferenced either as an HTML page or a set of RDF descriptions, returned by our RESTful service API. We currently only support publication in our RO repository sandbox. We are working on supporting other existing public repositories for sharing reproducible experiment resources.

3 User Experiences of RO Manager

RO Manager has been presented to domain scientists as a workbench to create and manage ROs during the investigation phase of their research. The feedback from the scientists demonstrate the need for supporting the management of ROs prior to the final stage of research investigation. They also show a willingness to investigate time on RO creation in order to benefit from the evolution control and quality evaluation. Compared to a web-based user interface,

[9] `http://purl.org/wf4ever/ro#`
[10] `http://purl.org/ao/`

the scientists appreciate the flexibility of managing their data locally. Because the investigation and design phase involves a certain amount of modification to their initial experiment designs, they are very interested in adopting the RO evolution management functionality, in development at this moment, in order to help them track changes made in different versions of the ROs and analyze the impact on the reproducibility of the RO by these changes.

Although a command-line tool requires an initial learning curve, once the scientists got used to it they found it especially convenient for creating an RO from a bulk of resources. However, the current support for publishing and annotating ROs is less satisfactory. Additional editing or annotations might take place in a web-based space where RO is shared, which must be seamlessly synchronized with its local copies. Although some annotations can be automatically generated by the RO Manager, the majority of them must be manually created; as a command-line tool RO Manager is not the best tooling for this purpose.

4 The Vision of RO Manager

We position RO Manager as a local workbench for scientists to create and manipulate ROs, which can then be shared as either a resource on the Web or part of their newer, richer form of research publication. It is one small component among the big picture of supporting reproducible science. We would like to make use of existing annotation tools, to ease the creation of richer documentation. Particularly, we would like to support annotations at various granularities, and aggregate and retain existing annotations of an external resource (such as a script stored in web site or a web service) by its URIs. To promote the visibility of our resulting ROs as well reproducible science in general, we would like to work together with publishers and existing web sites dedicated for the sharing of reproducible experiment resources, to publish ROs and provide our enhanced support for assessing and monitoring their fitness for supporting reproducibility. Finally, we are working on migrating the functionalities of this tool to a Web-based interface, for users who are less influent with command-line tools, which will also provide some richer visualization of the content of the RO and its evolutions.

References

1. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems (2011)
2. Goble, C.A., Roure, D.D., Bechhofer, S.: Accelerating scientists' knowledge turns. In: Proceedings of The 3rd international IC3K joint conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. (2012), in press
3. Stodden, V.: The scientific method in practice: reproducibility in the computational sciences (2010)
4. Zhao, J., Gomez-Perez, J., Belhajjame, K., et al: Why workflows break-understanding and combating decay in taverna workflows. In: IEEE eScience. p. To appear (2012)