Research data management: Where software meets data

A Research Data Management (RDM) "Green Shoots" Pilots Project Report by Christian T. Jacobs, Alexandros Avdis, Gerard J. Gorman and Matthew D. Piggott

Imperial College

This project was funded as part of Imperial College's RDM "Green Shoots" Programme. In 2014, the Vice-Provost, Research, approved an allocation of £100K for academically-driven projects to identify and generate exemplars of best practice in RDM, specifically frameworks and prototypes that would comply with key funder RDM policies and the College position. The call for projects outlined that frameworks could be based either on original ideas or integrating existing solutions into the research process, improving its efficacy or the breadth of its usage. There was an expectation that solutions would support open access for data; solutions that supported Open Innovation were strongly encouraged.

Six projects were funded, covering different disciplines, faculties and research areas. The projects ran for six months, finishing at the end of 2014. Project reports were made available in 2015.

For more information on the programme and projects please visit:

http://www3.imperial.ac.uk/researchsupport/rdm/policy/greenshoots

Imperial College London

Research data management: Where software meets data

Christian T. Jacobs, Alexandros Avdis, Gerard J. Gorman, Matthew D. Piggott; Imperial College London

Context

Computational science workflows are characterised by the interactions between both software and data. Collected data (e.g. tidal forcing conditions) is frequently given to scientific software (e.g. a numerical model) as input. This software then produces 'output data' (e.g. atmospheric flow fields or diagnostic quantities) which is analysed by the researcher to yield scientific results. In order to ensure reproducibility and recomputability of these results, the software, input data and output data should all be captured along with any provenance metadata. However, these components are often not published alongside the main scientific findings in journal articles.

This 'Green Shoots' RDM project investigated ways in which scientific software and data could be shared online at literally the 'push of a button'. It aimed to facilitate their publication in online, citable and persistent repositories by introducing a large amount of automation, which in turn motivates researchers and encourages the sharing and re-use of research outputs through a more open and easy-to-use scientific workflow.

Project Deliverables

The main deliverable of this project has been the development and release (under the GNU General Public License) of an open-source software library, PyRDM (github.com/pyrdm). This library is able to automatically publish software source code (stored under Git version control) and data to online repositories provided by Figshare (figshare.com). In return, Figshare yields a Digital Object Identifier (DOI) for each repository which can be used in journal articles to formally and properly cite research outputs. For example, the specific version of the software used to produce a given dataset/result is often not stated in journal articles; instead, it is usually the software's website or generic user manual that is cited. PyRDM is able to automatically determine the particular version of the software being used and publish it to a Figshare repository. The DOI that is minted can then be used to reference that repository to enable much better recomputability of the data. Metadata is also added to the repository automatically. This includes author information determined from the software's AUTHORS file to achieve proper affiliation, and the software's version identifier; if another researcher tries to publish the same version of the software, then the Figshare repository and DOI are simply re-used instead.

In order to demonstrate its functionality, PyRDM has been integrated into the Fluidity computational fluid dynamics code (www.fluidity-project.org). Researchers can perform a fluid flow simulation, and then run a Fluidity-specific publishing tool which uses the PyRDM library. This tool automatically determines the version of the software used to run the simulation, and uploads it to Figshare. The relevant input and output files can also be published by providing a minimal amount of information, such as their location on the computer, in the simulation's configuration file. Throughout the publication process, the DOIs that are minted are stored in the configuration file and the

simulation's metadata to (a) improve data provenance and (b) enable a new revision of the repository to be created if the data is updated at a later stage.

A more detailed description of PyRDM, its functionality, and an example of its application has been published in the Journal of Open Research Software (see [3]).

Recommendations

Throughout the implementation and use of PyRDM, several issues have been identified. For example, attempting to automatically affiliate software authors to an online repository can prove difficult due to lack of standardisation. For example, all authors must currently provide their Figshare author IDs in order for PyRDM to correctly attribute them to the software repository on Figshare; for a different service, another set of author information may need to be provided. The lack of standardisation was a problem during the development of this project. However, this situation is improving as a result of Figshare (and other organisations such as Symplectic (symplectic.co.uk)) recently adding support for ORCID (orcid.org) researcher IDs [1, 4], which are considered a more standardised way of identifying and attributing authors to research outputs. In the near future, it is hoped that support for authenticating via ORCID and using an ORCID ID when publishing via the Figshare API (instead of the web interface) will also be added.

Some research involves proprietary and/or private data which cannot be shared, but at the same time digital curation is important for the funders of the research. PyRDM is capable of publishing to private Figshare repositories. However, Figshare currently offers limited free private storage space (1 GB) for individual users which it is not generally large enough to store complete modern day simulations, for example. Moreover, the number of collaborators who can view/modify the private data is also limited. Figshare for Institutions, which allows members of an institution to store software and data privately in the cloud, may be a more suitable platform for larger-scale research data management. Its recent adoption by UK-based institutions such as Loughborough University [2] will hopefully further encourage institutions worldwide to integrate a more sustainable research data management framework that can cope with both public and private research outputs and a greater demand for collaboration.

References

[1] Figshare. figshare ORCID integration. http://figshare.com/blog, 2014. 2

[2] Figshare. Loughborough University, figshare, Arkivum and Symplectic announce pioneering research data management solution. http://figshare.com/blog, 2014.

[3] C.T. Jacobs, A. Avdis, G.J. Gorman, and M.D. Piggott. PyRDM: A Python-based library for automating the management and online publication of scientific software and data. Journal of Open Research Software, 2(e28), 2014.

[4] Symplectic. Elements integrates with ORCID. http://symplectic.co.uk/elementsupdates, 2014.