Git-RDM: A research data management plugin for the Git version control system

Christian T. Jacobs

University of Southampton

Alexandros Avdis

Imperial College London

16 June 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00029
Software Repository: https://github.com/ctjacobs/git-rdm
Software Archive: http://dx.doi.org/10.6084/m9.figshare.3439742.v1

Summary

Many research funding agencies (Research Council UK 2015) and research societies (Royal Society 2012) are increasingly requiring that data from at least publicly funded research be made openly available, and with clear citations that describe provenance. These requirements have led to the proliferation of institutional repositories with universities maintaining a handful of data services, but also repository services capable of minting a persistent and citable Digital Object Identifier (DOI) (Technical Committee ISO/TC 46 (Information and documentation), Subcommittee SC 9 (Identification and description) 2012) for every published item. Figshare (figshare.com) and Zenodo (zenodo.org) are examples of the latter. Alongside data, software is also increasingly seen as a research output. This viewpoint necessitates not just open-source publication of code, but also provenance and attribution. While a DOI is an identifier of static items, many research teams use version control systems and services to organise their collective efforts and publish output, be that code or data. Popular examples include Git (Chacon and Straub 2014) and GitHub (github.com).

Git-RDM is a Research Data Management (RDM) plugin for the Git version control system. It interfaces Git with data hosting services to manage the curation of version controlled files using persistent, citable repositories. This facilitates the sharing of research outputs and encourages a more open workflow within the research community.

Much like the standard Git commands, Git-RDM allows users to add/remove files within a 'publication staging area'. When ready, users can readily publish these staged files to a data repository hosted either by Figshare or Zenodo via the command line; this curation step is handled by the PyRDM library (Jacobs et al. 2014). Details of the files and their associated publication(s) are then recorded in a local SQLite database, including the specific Git revision (in the form of a SHA-1 hash), publication date/time, and the DOI, such that a full history of data publication is maintained.

References

Chacon, S., and B. Straub. 2014. *Pro Git*. 2nd ed. Apress.

Jacobs, C. T., A. Avdis, G. J. Gorman, and M. D. Piggott. 2014. "PyRDM: A Python-based library for automating the management and online publication of scientific software and data." *Journal of Open Research Software* 2 (1): e28. doi:10.5334/jors.bj.

Research Council UK. 2015. "Guidance on Best Practice in the Management of Research Data."

Royal Society. 2012. "Science as an Open Enterprise: Open Data for Open Science."

Technical Committee ISO/TC 46 (Information and documentation), Subcommittee SC 9 (Identification and description). 2012. "ISO 26324:2012 Information and documentation – Digital object identifier system." International Organisation for Standardization.