Patterns and Principles for Versioning of Research Data

Jens Klump[1]

Jens Klump[1], Mingfang Wu[2], Gerry Ryder[2], Julia Martin[2], Lesley Wyborn[2,3], Robert Downs[4], Ari Asmi[5]

[1] CSIRO Mineral Resources, Perth, Australia, jens.klump@csiro.au
[2] ARDC, Canberra, Australia, (mingfang.wu|gerry.ryder|julia.martin)@ardc.edu.au
[3] NCI, ANU, Canberra, Australia, lesley.wyborn@anu.edu.au
[4] Columbia University, New York, NY, USA, rdowns@ciesin.columbia.edu
[5] University of Helsinki, Helsinki, Finland, ari.asmi@helsinki.fi

INTRODUCTION

The demand for better reproducibility of research results is growing. It is becoming increasingly important for a researcher to be able to cite the exact extract of the data set that was used to underpin their research publication. However, while the means to identify datasets using persistent identifiers have been in place for more than a decade, systematic data versioning practices are currently not universally agreed upon, which continues to create challenges for the reproducibility and replicability of research.

Versioning procedures and best practices are well established for scientific software (e.g. (Fitzpatrick et al., 2009; Preston-Werner, May 29, 2011/2013)) and can be used to enable reproducibility of scientific results (e.g. (Bryan, 2018)). The related Wikipedia article gives an overview of software versioning practices ("Software versioning," 2019). The codebase of large software projects does bear some semblance to large dynamic datasets, which provides opportunities for leveraging software versioning practices to improve data versioning capabilities. Are, therefore, versioning practices for code also suitable for data sets or do we need a separate suite of practices for data versioning? How can we apply our knowledge of versioning code to improve data versioning practices?

THE NEED FOR DATA VERSIONING PRINCIPLES

The importance of data versioning practices has been recognised in different fields where the reproducibility of research is a concern, e.g. data citation, data provenance, and virtual research environments. The discussion of the recommendations published by the Research Data Alliance (RDA) Dynamic Data Citation Working Group (Rauber et al., 2016), as well as the work in other groups on data provenance and data citation, and in the W3C Dataset Exchange Working Group (Dataset Exchange Working Group, 2017), have highlighted that definitions of data versioning concepts and recommended practices were still missing. On the other hand, versioning procedures and standard practices are well established for scientific software and its concepts could be applied within other use cases to facilitate the goals of reproducibility of scientific results.

The RDA Data Versioning Working Group has worked with other groups within RDA and externally to develop a common understanding of data versioning and recommended practices. Agreement on versioning practices also is important for provenance tracking of a derived data set, and for improving capabilities to ensure that adequate credit and attribution are given to those people and institutions that have contributed to developing earlier versions of data products and services.

DATA VERSIONING USE CASES AND PRINCIPLES

Over the past two years, the RDA Data Versioning Working Group has collected numerous use cases of data versioning practices and extracted data versioning patterns. A key element that emerged from the analysis of the versioning use cases was the necessary distinction between revision, release, and manifestation of a dataset.

Versioning systems such as SVN (Fitzpatrick et al., 2009) or Git (Chacon & Straub, 2014) track the bitstream of a dataset. Whenever a dataset is changed, and thus its bitstream is changed, the resulting changes are considered to be a revision.

Tracking revisions is a technical process that may document the magnitude of the change but does not convey the significance of the change.

A dataset may undergo several revisions in its production process before it is considered to be "final" and is subsequently published as a data release. The significance of the new release will depend on the requirements of its designated user community. Procedures such as Semantic Versioning (Preston-Werner, May 29, 2011/2013) encode the significance of a change in the naming of the release. While revisions are a technical procedure, the release of a dataset stands at the end of an editorial process.

Sometimes datasets are published in different formats or encodings but are equivalent in their content. Following the model of Functional Requirements for Bibliographic Records (FRBR), both datasets can be seen as manifestations of the same intellectual work (Hourclé, 2009).

All three cases outlined above are currently subsumed under the term of "version", yet all three cases represent unique patterns and require different treatment with respect to identification, publication and citation. Based on the use cases and extracted patterns, we recommend the following data versioning principles:

- Management: Recognise identification and tracking of data revisions and data releases as an important component of data management
  - Establish a procedure and policy for consistent management of data revisions and releases.
- Identification: Be clear about which dataset is to be identified
  - Identify data revisions, consider issuing a new persistent identifier per revision and release
- Communication: Communicate the significance of the change to the designated user community of this dataset.
  - Concepts such as Semantic Versioning describe the significance of a version change
- Provenance: Track changes and record provenance information between revisions
  - Provenance information describes the changes that have been made to each newer revision. Display provenance information, attribution and credit on landing pages of each publically released dataset.
- Citation: Cite a specific data release
  - For each released dataset, have a clear recommendation, including a release number, on how to cite a dataset.

SUMMARY

The analysis of many use cases of data versioning by the RDA Data Versioning Working Group has allowed us to identify common usage patterns and derive principles for data versioning. Instead of subsuming all cases under the term "version", the Working Group proposes to differentiate between a dataset revision, a dataset release, and different manifestations of a dataset. The two key principles in data versioning are: (1) be clear about which dataset is to be identified, and (2) communicate the significance of the change to the designated user community of this dataset.

REFERENCES

Bryan, J. (2018). Excuse Me, Do You Have a Moment to Talk About Version Control? *The American Statistician*, *72*(1), 20–27. https://doi.org/10.1080/00031305.2017.1399928

Chacon, S., & Straub, B. (2014). *Pro Git* (2nd Edition). New York, NY: Apress. Retrieved from https://git-scm.com/book/en/v2

Dataset Exchange Working Group. (2017). Dataset Exchange Working Group [Working Group Pages]. Retrieved March 20, 2019, from https://www.w3.org/2017/dxwg/wiki/Main_Page

Fitzpatrick, B., Pilato, C. M., & Collins-Sussman, B. (2009). *Version Control with Subversion*. Sebastopol, CA: O'Reilly Media, Inc. Retrieved from http://svnbook.red-bean.com/

Hourclé, J. A. (2009). FRBR applied to scientific data. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1–4. https://doi.org/10.1002/meet.2008.14504503102

Preston-Werner, T. (2013). Semantic Versioning 2.0.0. Retrieved March 7, 2019, from https://semver.org/spec/v2.0.0.html (Original work published May 29, 2011)

Rauber, A., Asmi, A., van Uitvanck, D., & Pröll, S. (2016). *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)* (Technical Report). Denver, CO: Research Data Alliance. Retrieved from http://dx.doi.org/10.15497/RDA00016

Software versioning. (2019, March 6). Retrieved March 11, 2019, from https://en.wikipedia.org/w/index.php?title=Software_versioning&oldid=886437916