Scientific Data and Preservation – Policy Issues for the Long-term Record

by Vera Sarkol (CWI)

In order to keep open data accessible into the future, academics and librarians need to consider long-term preservation.

From open access of publications the trend is now expanding to open science, and, with that, open data. The progress of our communal knowledge is dependent on previously discovered truths, and therefore the data has to be openly available to the extent that others can find, understand and use it [1]. The concepts of 'openness' and 'preservation' are inextricably linked if we want to secure a continuous record of the path of discovery. The job of maintaining these records falls to the national or institutional libraries and repositories.

Many funders, such as the Netherlands Organisation for Scientific Research (NWO), are developing policies for data and software management which address openness and preservation. This puts some pressure on the issue, and it is the right place to raise the question of cost for documenting and depositing the artifacts, in terms of workload and resources. The most important challenges for long-term policy are selection, findability, and reusability.

Selecting what to preserve

Ideally we would preserve and make available every scientific artifact, but in reality this is neither feasible nor desirable [L1]. Constraints of size or legality will of course hinder preservation. Other constraints are the time it costs to properly document and describe datasets and software, and the environmental cost of storage. Therefore data that can easily be replicated or code that only serves to illustrate an algorithm does not necessarily need to be preserved. For now it is a good principle to preserve artifacts that underlie publications, but if in the future the boundaries of publications as the unit of scientific knowledge blur (e.g., if preprints and post-evaluation get integrated into the process), academics and librarians together will have to develop other criteria for selection.

Replication packages

Storing only data or software is no guarantee that a finding can be replicated if crucial information is missing. To avoid this problem NWO will soon make replication packages mandatory. This means that along with the dataset or program, the metadata, identifier and provenance information should be stored, as well as the software and hardware, or at the very least a description. However, even with that information, complex dependencies or outdated software packages may still prevent replication.

One project that provides a solution to this problem is being developed at CWI: Snakemake [2]. This is a text-based workflow management system that was originally developed for bioinformatics but could be suitable for other fields of research as well. Using a domain specific language, Snakemake aims to formalise an analysis workflow, including a specification of used software packages. Upon execution of a workflow, software packages are deployed automatically so that an analysis is reproducible without extra work.

Figure 1: Preserved knowledge of fish, from Adriaen Coenensz' 'Visboeck', 1579, the National Library of the Netherlands. Location: KB Den Haag, KW 78 E 54 fol. 346r

More drastic problems will occur when hardware becomes outdated. One possible way forward could be virtualisation [L2], where the old environment is emulated to access preserved scientific artifacts. However, at some point hardware may undergo such a large change that this too is no longer a viable option. It is necessary for the community to start thinking about what to do when that occurs.

Licences

A large variety of data and software licences are currently in use, sometimes prescribed by journals or repositories. When interoperability becomes more prominent these licences may not interact well with each other and this may lead to datasets being unable to recombine. Another problem is that not everyone has licences to proprietary (legacy) software and operating systems used. One solution is to fully commit to open-source. Another possibility is to licence everything to the public domain and that licenced legacy software is kept running centrally, for instance by national heritage institutions [L2] (the National Library in the Netherlands, for instance).

Findability of preserved software

For long term findability, citing an URL for a program in an article is not sufficient, since the content of an URL can easily change. More durable would be to give all scientific artifacts, including software, a persistent identifier which is a unique code given to an object by an organisation, irrespective of its location. The DOI has become the academic standard and thus may be expected to be maintained the longest. The version of a program that underlies a publication should be deposited in a repository and receive a DOI. For instance, Zenodo provides this service and is integrated with GitHub. Getting a DOI will make it easier to find the right version of the software with the publication, but it will also make it easier to find and cite the work for the broader (academic) community and funding agencies [3].

Conclusion

From a library's perspective the goal is to make the record of scientific knowledge as permanent as it was when there was only paper. While progress is being made, international consensus on a number of issues needs to be achieved. Consultation at a European level is necessary to establish guidelines for the long-term preservation of open data and software.

Links:

[L1] https://www.esciencecenter.nl/pdf/Software_Sustainability_DANS_NLeSC_2016.pdf

[L2] https://www.unesco.nl/sites/default/files/dossier/report_girona_session_persist.pdf

References:

[1] M. D. Wilkinson et al.: "The FAIR Guiding Principles for scientific data management and stewardship", Scientific Data 3:160018, 2016. http://dx.doi.org/10.1038/sdata.2016.18

[2] J. Köster, S. Rahmann: "Snakemake – A scalable bioinformatics workflow engine", Bioinformatics 28(19): 2520-2522, 2012. http://dx.doi.org/10.1093/bioinformatics/bts480

[3] A.M. Smith et al.: "Software Citation Principles", PeerJ Preprints, 2016. http://dx.doi.org/10.7287/peerj.preprints.2169v2

Please contact:

Vera Sarkol
CWI Information & Documentation
+31(0)205924051
vera.sarkol@cwi.nl