Scientific results traceability: software citation using GitHub & Zenodo

Charles Troupin, GeoHydrodynamics and Environment Research (GHER), Freshwater and OCeanic science Unit, University of Liège (Belgium), ctroupin@uliege.be
Cristian Muñoz, SOCIB (Spain), cmunoz@socib.es
Juan Gabriel Fernández, SOCIB (Spain), jfernandez@socib.es
Miquel Àngel Rújula, SOCIB (Spain), mrujula@socib.es

Background

Assigning unique identifiers to various types of objects is a common practice in the research field: data sets, scientists, papers, products... Nevertheless, the connection between the data and the final products or results is not always properly identified and documented, thus putting a damper on the reproducibility and traceability. This connection can be for example the execution of a numerical model or a software tool with a given set of parameters.

Having the code of a software tool in a version control system (VCS), with periodic releases with a unique and persistent identifier for each release contributes to the reproducibility.

The main reasons for assigning DOI to a software are:

- Reproducibility: any user shall be able to run the same experiment with identical parameters, the same dataset using the same model or software tool should obtain the same results, hence improving their credibility.
- Traceability: all the elements used in the analysis or experiment should be accessible, properly described and uniquely identified.

The present document focuses on real application of software citation in the frame of ocean data management and processing.

Citation in the context of ocean observation

Figure 1 describes the relationships between the Ocean Observation on one side and the corresponding scientific results on the other side. It is now frequent that the datasets coming from ocean observation are assigned a DOI. A good and illustrative example is the data citation related to the Argo profilers (see for example http://www.argo.ucsd.edu/Argo_DOIs_AST16.pdf).

The scientific results stemming from the combination of one or several datasets and different analysis methods are frequently published as articles in peer-reviewed journals, and these articles are citable using the corresponding DOI.

The intermediate part (blue items in the diagram) concerns the method or procedure from which results, products or outcomes are obtained, namely:

- the application of a numerical model using a given set of parameters (including boundary and initial conditions, forcing, …) and data or
- the utilisation of an analysis method, implemented as a software tool, on a given dataset.

A proper management of the considered piece of software and its different versions or releases is essential if one wants to guarantee the reproducibility of the results.

Procedure for the attribution of DOI to a software code

In the document "*Emerging standards for ocean data analysis and visualisation*" (submitted to ODIP), different options for software citation are analysed and compared. The following sections present an application using the Zenodo platform. That choice was based on these reasons:

- The code is free and open (https://github.com/zenodo/zenodo);
- It offers DOI versioning: either specific versions of a software or all the versions together can be cited.

- An account on ORCID or on GitHub is sufficient for login.
- There is no need to install anything, everything is stored, managed and backed-up on their servers.

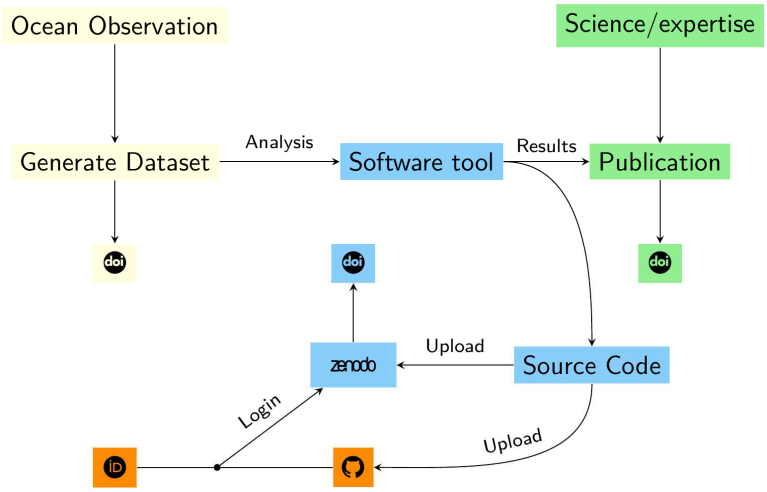

Figure 1: Workflow overview

Use cases developed within the ocean observing community

The DIVA interpolation tool

DIVA (Data-Interpolating Variational Analysis) is one of the reference tools developed within SeaDataNet. It is used to prepare the regional products, which consists of sets of gridded field obtained by spatial interpolation of all the available in situ observations collected by the data centers.

From 2008 on, DIVA code (mainly Fortran and bash) was managed via SVN and periodic releases were performed. In 2016, the code was migrated to GitHub, making it easier for the user to directly access all the versions of the code.

The GitHub repository (https://github.com/gher-ulg/DIVA) was connected to Zenodo, so that every release of the code would result in a new publication in Zenodo, accompanied with the corresponding DOI. With this approach, the DIVA users can know directly cite the version of the code they employed to prepare the climatologies (either in publications, technical reports or even in the netCDF files themselves).

The SOCIB glider toolbox

The toolbox consists of a set of MATLAB / GNU Octave programs designed to process the raw files provided by different types of gliders. The processing steps include the format conversion, the unit conversion, some corrections on the measurements (e.g. thermal lag) and lead to the creation of netCDF files and various figures (T-S diagrams, sections, …).

The code was managed using GitHub from the beginning (https://github.com/socib/glider_toolbox) and the coupling with Zenodo was performed in 2017.

Conclusions

A platform such as Zenodo constitutes a valuable instrument for scientists and data managers that want to ensure that their software tool is properly cited and identified, with an emphasis on the version of the code employed by the users.

Software citation strives to establish an explicit clearer relationship between:

- the data, more and more published and identified in public infrastructures.
- the result products, often described in peer-reviewed publications.

This methodology allows the description of the procedure that leads to results starting from the data.