Software Citation Today and Tomorrow

IDEAS-ECP Best Practices for HPC Software Developers Webinar Series
Date: April 18, 2018
Presented by: Daniel S. Katz

Q.1. You note that github might go away in the future, and we should use DOI on figshare or zenodo, but why should we expect those to last?

A.1. Figshare and Zenodo are archival repositories that are built and operated to be persistent, under the governance of the International DOI Foundation (IDF). The DOI and handle infrastructure has been designed for this purpose. On the other hand, GitHub is a platform for work with code, including versioning, social interaction, etc. It makes no claims to be persistent, and the history of previous coding platforms includes many that have been shut down, such as Google Code and Microsoft CodePlex, and some that still function but are not commonly used. While GitHub is today's standard, there's no reason to believe that it will remain so forever.

Q.2. Sometimes it is the case that when generating data for a paper, I use a version of the software that is a development version and/or does not correspond to a particular numbered release of the code. Is there any harm for creating a DOI for that specific commit? Can there be "too many" DOIs associated with a particular software repository?

A.2. This is a topic that had multiple answers. If your focus is on identifying the specific version of software that was used in a research project, then no. If you are a developer of the software and are worried about your citations being spread too widely, then currently yes. We hope that this latter worry will decrease over time as tools develop to bundle the citations of multiple versions into a set of citations to the overall project.

Q.3. What about forked repositories?

A.3. This falls into the discussion area in the talk about derived software. Citations should be made to the software that was actually used, not to software from which the used software was derived.

Q.4.a. Given the choices, Zenodo/Figshare etc., is DOE Code recommended? https://www.osti.gov/doecode

Q.4.b. I should add in the context of DOE funded software. Are there any metrics or indexing advantages for using one over the other. I ask because they are providing DOI and metadata, e.g., https://www.osti.gov/doecode/biblio/9801

Comment: Citing DOE CODE FAQ.> OSTI is a member of and registering agency for DataCite and has the authority to assign Digital Object Identifiers to software and code that are submitted by DOE and its contractors or grantees. The assigning and registration of a DOI for software is a free service provided by OSTI to enhance DOE's management of this important resource.

Comment: From the DOECode FAQ: "DOE CODE is the U.S. Department of Energy's (DOE) new software services platform and search tool for software resulting from DOE-funded research that provides functionality for collaboration, archiving, and discovery of scientific and business software funded by DOE. DOE CODE replaces the Energy Science and Technology Software Center (ESTSC)."

A.4. At least for the example given in Q.4.b, DOE Code seems to not be an archival repository, but more of a catalog of codes. Placing an entry in DOE Code is fine, and may be required for some who are funded by DOE, but it does not appear to take the place of archiving a specific version of the code in order to be able to cite that specific version.

Q.5. The principle of specificity leads us to perhaps citing specific versions but presently it's difficult to aggregate credit across versions. You mentioned some efforts to address this. Can you elaborate?

A.5. Recent metadata schema changes by DataCite (see the slides) allows metadata to be used to connect different versions of software. There is an enabling step; there is still a need for tools to aggregate citations, which the FORCE11 group is working to promote.

Q.6. In the area of citation, What is the BEST-practice for software authors TODAY?

A.6. The answer depends on what the author’s goals are. If the goal is just to get credit, the best thing right now is probably to write a software paper and to ask people to cite the paper. If the goal is also to support reproducibility, then archiving releases of the software and asking people to cite the correct release is probably a better choice. And we are working towards making the second answer the right answer for credit as well.

Q.7. Is there a standard format for listing the citation so that when it shows in google scholar and elsewhere it is understood that it is a software package and that it is maybe related to some other published product, either software or paper?

A.7. This is another area the FORCE11 group is working in, but today, the answer is generally no.

Q.8. Sometimes it is the case that when generating data for a paper, I use a version of the software that is a development version and/or does not correspond to a particular numbered release of the code. Is there any harm for creating a DOI for that specific commit? Can there be "too many" DOIs associated with a particular software repository?

A.8. As discussed previously, the potential harm is splitting citations across multiple software versions; however, we believe that tools to allow credit aggregation will be developed.