Asclepias - Capturing Software Citations in Astronomy

Edwin A. Henneken
Smithsonian Astrophysical Observatory
60 Garden Street, Cambridge, Massachusetts, USA
ehenneken@cfa.harvard.edu

Alberto Accomazzi
Smithsonian Astrophysical Observatory
60 Garden Street, Cambridge, Massachusetts, USA
aaccomazzi@cfa.harvard.edu

Sergi Blanco-Cuaresma
Smithsonian Astrophysical Observatory
60 Garden Street, Cambridge, Massachusetts, USA
sblancocuaresma@cfa.harvard.edu

August Muench
American Astronomical Society
200 Florida Avenue, NW Suite 400, Washington, DC, USA
august.muench@aas.org

Lars Holm Nielsen
CERN, IT Department, Digital Repositories
Geneva, Switzerland
lars.holm.nielsen@cern.ch

ABSTRACT

There is general agreement that curating and preserving software products, and making them citable, are worthwhile efforts, as is linking software products with publications, in order to capture the full life cycle and to improve discoverability. This is definitely true in the physical sciences. However, there is currently no established standard or policy for citing software in scholarly publications. There are several reasons behind the current impasse: the lack of clear editorial standards and expectations on the part of the journal publishers; the difficulty in fitting software products in a model constructed for citing publications; the need to ensure a unique, persistent identifier to the software product used in the paper; and an antiquated view of the software as having a narrow scholarly impact on the paper itself. A proper solution will address all of these reasons. Even without an established way to cite software products, there is mounting evidence that a significant number of scientists want to acknowledge the contribution of software in their scholarly articles. This acknowledgment has taken one of two forms: via citations to a "software paper" or by mentioning a software product in the text of the article. Here we present an ongoing project called "Asclepias", a collaboration between the American Astronomical Society[1] (AAS), Zenodo[2] and the SAO/NASA Astrophysics Data System[3] (ADS). The goal of this project is to promote scientific software into an identifiable, citable, and preservable object. It is focusing upon the needs of two of the most important roles researchers play in the scholarly ecosystem: authors of scholarly manuscripts and developers of scientific software. Currently a technical framework is being built and work is done in promoting a set of social practices that will "fix" the problems associated with software citations.

Keywords

Software citation, research lifecycle, research impact, astronomy.

INTRODUCTION

The problem of software citation consists of a number of components (Allen et al., 2015; Rios, F., 2016). While writing a so-called "software paper" in a scholarly journal is one way of acknowledging the contribution of software to the research lifecycle, letting it serve as a proxy for citation purposes may not be enough to accumulate the proper amount of attribution. Besides the question whether software papers properly measure the scholarly impact, there is a more fundamental question whether these publications properly represent software products as entities in the research lifecycle. This is unlikely to be true for various reasons. For one, a scholarly publication can in no way capture the dynamic nature of a software product. Software development is, now more than ever, a collaborative process, often with a set of contributors that changes from release to release. Besides improperly representing the main contributors, software papers are also likely to be outdated, perhaps even already by the time of publication. In particular, where it comes to reproducibility of research data, it is crucial to be able to cite a specific version of software used. A user reading a research article that cites a software paper will be left with a poor set of options when trying to determine its usage in the paper:

[1] https://aas.org

[2] https://zenodo.org

[3] https://ui.adsabs.harvard.edu

they can go back to the original citation to the software paper (presumably an out-of-date description of the software package), or can try to locate the current version of the software package (presumably the latest version of the software, and not what was used in the study), or can try to contact the author of the citing articles asking for detailed information about what specific version was used. Mentioning software, in a separate section or footnotes, is a different way of acknowledging the use of software. Often, authors include URLs with these mentions. These mentions will not result in proper tracking of software, however. Where citations are being picked up because indexing services actively process article bibliographies, footnotes and "software sections" are not parsed by these indexing services. As a result, the intended goal of a mention, to properly attribute the use of software, is missed. It has been shown (Pepe et al, 2014) that the URLs included with mentions suffer from substantial "URL rot". Currently, the half-life of URLs in the journals published by the AAS is 10 years.

METHOD

No single stakeholder can solve the software citation problem. It will need close collaboration between a publisher (AAS), a repository (Zenodo) and an indexing service (ADS) to be able to design and implement a solution that can be shown to "fix" the problems described in the introduction. The collaboration in the Asclepias project (Muench et al., 2017) is working on a solution consisting of a technical framework and the promotion of a set of social practices which will “fix” the problems associated with software citations. The following issues are envisioned to be addressed by this solution. First, software will now be treated as a true citable object by itself and its citation will become an encouraged practice in the publication of scientific papers. Secondly, this solution addresses preservation. Releases of software products and associated documentation will be deposited in a trusted repository (Zenodo), so that individual versions are archived as separate entities and proper authorship information is collected for each. Thirdly, identification will be made possible using a persistent identifier, specifically a DOI. Archived software releases will be assigned unique, persistent identifiers and associated metadata so that precise, persistent connections can be made between papers and software. Citations will include either a version-specific DOI or a "concept DOI", representing all versions (but always resolving into the record for the latest version). In the fourth place, software developers will be in full control of the process which determines the proper list of contributors for software packages on a release-by-release basis. This takes care of proper attribution. Lastly, cited software and its impact will be represented in discipline-specific indexing systems which are used by the targeted community for discovery and evaluation of scholarly content.

The Asclepias project focuses on astronomy. This discipline has a large community of developers, a culture of openly sharing code and data, and established and centralized infrastructure for communicating results. Instead of having to win an audience or build something entirely new, the focus will be on improving upon existing infrastructure to enable critical workflows that meet the needs of the two most important roles researchers play in the scholarly ecosystem: authors of scholarly manuscripts and developers of scientific software.

Journal articles are the primary way to distribute research results, yet authors often include no or only passing mention of what software packages were used in the research they describe. The Asclepias project will define submission workflows and promote new editorial policies that will encourage the robust identification of software used in a paper. In astronomy, this lack of detail commonly applies not only to other people’s code, but also to the methods, pipelines, and codes developed by astronomers for the purpose of reducing and analyzing the data in the paper itself, thus weakening transparency and hindering reproducibility. The Asclepias project will provide workflows to enable the publication and sharing of all code which can be legitimately be placed in the public domain.

IMPLEMENTATION

To “fix” the problem, a publisher will need to implement practices and tools that will support the integration of software citations into the scholarly publishing workflow. Specifically, this means the implementation of an authoring workflow for scholarly publishing that will be adopted for a significant fraction of astronomy and can be easily exported to other publishers and to other disciplines. From the view point of bibliographic indexing services, the implementation means the development of first-class support of a software framework to detect software citations and take all necessary steps to attribute and expose these citations. For repositories, implementation means interoperability with Github and similar codebases (like Bitbucket). The minting of DOIs is one step in this integration, and being able to send out notifications that (astronomy related) software has been released, including version updates, is another. All these steps are only possible if metadata standards for software packages are adopted. Within the Asclepias project, we will align our metadata development and software citation practices with the FORCE11 Software Citation Implementation Group[4] efforts. In addition to adopting metadata standards, the implementation phase also needs a means to automatically broker and efficiently curate software metadata. In order to support this, a thin, efficient layer of infrastructure will be developed, that processes events stemming from the release of software into repositories like Zenodo, and software

[4]https://www.force11.org/group/software-citation-implementation-working-group

citations discovered in journal articles by bibliographic indexing services like the ADS. It will associate software references and releases by version, and broadcast by unique identifier aggregated metadata about software citations. By automating this process this infrastructure will make efficient the otherwise manual process of tracking these software and corresponding citation changes. Last but not least, outreach to the community (authors and developers) is critical for the long-term success and sustainability of the project.

ACKNOWLEDGMENTS

The Asclepias project is funded through a grant from the Alfred P. Sloan Foundation to the American Astronomical Society, 2016. The ADS is operated by the Smithsonian Astrophysical Observatory under NASA Cooperative Agreement NNX16AC86A.

REFERENCES

Allen, A., Berriman, G.B., DuPrie, K., Mink, J., Nemiroff, R., Robitaille, T., Shamir, L., Shortridge, K., Taylor, M., Teuben, P., and Wallin, J. (2015). Improving Software Citation and Credit. eprint arXiv:1512.07919

Muench, A., Accomazzi, A., & Holm Nielsen, L. (2017). Asclepias: Enabling software citation & discovery workflows. Zenodo. doi:10.5281/zenodo.803474

Pepe, A., Goodman, A., Muench, A., Crosas, M., Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE 9,* 104798.

Rios, F. (2016). The Pathways of Research Software Preservation: An Educational and Planning Resource for Service Development. *D-Lib Magazine 22*, 10.1045/july2016-rios

.