Linking Scholarly Literature to Research Data and Software - Lessons Learned in Astronomy

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

ABSTRACT
Publishing articles in scholarly journals is essential to furthering science. However, it is only one stage in the research cycle; one of the later stages, actually. After formulating a research goal, typically one or more proposals are written. When accepted, data is gathered, processed and compared to existing data and literature. Software plays an essential role in the data processing and comparison process. Often, before articles are published, papers and posters are presented at conferences. In short, there is a plethora of "products" associated with the entire research life cycle. An essential ingredient in this cycle is the ability to find literature and associated products that were used to create it. Reproducibility of and the ability to add to existing results should be the norm, not the exception, for robust scientific research. Discoverability of data products and software needs an infrastructure and a culture. Astronomy has a long tradition of sharing data, but there is still a lot of work to do. Having this infrastructure will not only support discoverability, it will also enable measurement of impact and proper attribution. Both elements will help in getting funding for future research, which completes the cycle.

Keywords
research data, research software, discovery, research impact, data and software citation, digital libraries.

INTRODUCTION
Being able to read the literature and access to the data and/or software discussed or used in the paper with just one click is obviously of great convenience for the end-user, in addition to providing evidence in support of the scientific arguments in the paper and the reproducibility of its results. Integration and indexing of this kind of associated information in a digital library, used by the research community, greatly enhances the discovery process. This is exactly what the NASA/Smithsonian Astrophysics Data System (ADS; Henneken et al., 2012) has been doing for the astronomy community for many years. Besides benefits to the research community, there are also benefits of this integration to the data archives, software repositories and projects providing their data to ADS. First of all, by having a bibliography integrated in the ADS database, it becomes trivial to retrieve metrics associated with it and thus evaluate the scientific impact of its datasets as well as using bibliographic-based analytics to gain insights on how the data or software is used in current research. Also, the act of linking literature with data products by itself has been shown to have impact. Studies by Henneken & Accomazzi (2012) and Dorch et al. (2015) have shown a "data sharing advantage" for papers which have links to data products, resulting in higher citation counts. Similarly, data re-use increases upon the publication of papers studying them (Winkelman et al. 2006), leading to an increase in archival research (White et al. 2009). In other words, well-linked data is more heavily used, and well-linked publications are more heavily cited, a win-win scenario, and the primary reason to have well-described, well-curated data products indexed in the ADS, or any other digital library for that matter. The linking of software to scholarly literature is still in its infancy (Allen et al., 2015; Muench et al., 2017).

CREATION STAGE
In order for data preservation to work, people need to be convinced to invest effort in depositing their data products and software in repositories like Dataverse, Zenodo, TheAstroData, the Astrophysics Source Code Library (ASCL) and Github. Many of these repositories assign persistent, unique identifiers (PDIDs) to support access. The PDID is often a DOI. Assigning a PDID helps discoverability and citability of the products (attribution). Astronomy has a long tradition of sharing data. For example, in 1972 the Centre de Donées Astronomiques de Strasbourg (CDS) was established to collect and distribute astronomical information. Publishers of major astronomy journals have shown willingness to innovate and participate in community initiatives (see e.g. the Policy Statement on Software of the American Astronomical Society, 2016). If a publisher requires data products used in publications to be available in a persistent manner, authors will find a way to meet them. A big challenge is attempting to unlocking or even reviving data that is not carefully indexed or stored. Attempts are made to unlocking this so-called "dark data" (Heidorn, 2008; Akers, 2013; Gallaher et al., 2015).

USE AND DISCOVERY STAGE
The research community uses the PDID to access the data products / software and carry out related research. This community also generates new publications using the PDID to reference the dataset or software. The ADS harvests and merges bibliographic data from multiple sources (arXiv, CrossRef, publishers, astronomy archives, ASCL, Zenodo). It also enriches metadata via text-mining of the full text sources (extract references, acknowledgments, keywords, plots and images). Furthermore, it generates and maintains citation and usage networks, and cross-correlates content. The ADS collects and maintains external links to publishers, archives (SIMBAD, VizieR, NED, MAST, ESO, etc.) and

incorporates bibliographies from institutes and archives. Important datasets are often described in "data" papers, but can also be available as electronic catalogs, e.g. from VizieR (the most complete library of published astronomical catalogues and data tables, with links to about 33,000 records in the ADS). Once in the ADS, they become easily discoverable, citable. Proposals contain early descriptions of current and ongoing science activities. They provide a direct link to existing or planned observations. In total the ADS has about 337,000 links to online data.

IMPACT MEASUREMENT STAGE

Formalizing data/software citation supports giving credit where credit is due, enhances discoverability of data products and facilitates the compilation of metrics for data products (usage or citation based). These metrics are critical for evaluating a instrument, project, mission or grant, because they are a measure for productivity and impact. It takes community effort for data (and software) citation to work properly. Only with well-defined standards will data citation become as transparent as it is for publications. The Force11 community has formulated a set of "Data Citation Principles" (Smith et al, 2016). These principles are supposed to encourage a good practice of data citation, and, as a result, create a data and software citation culture. Measures for the impact of data products will only be meaningful when such a culture is in place. The metrics service of the ADS is an example of how impact measures are generated for the community. In addition to proper attribution of credit, being able to measure use of data and software can also be of vital importance to a project or mission: continued funding is often justified through use of data.

REWARD AND REUSE STAGE

Funding and research groups review publication and data set metrics. These metrics can influence the future funding of a researcher, project or instrument. In addition, having these metrics will help in providing scholarly credit and attribution to everybody who contributed to acquiring and creating the data (or software). Studies have found that publications linked to data products received higher citation rates, which, besides helping with the visibility of the data, is an additional incentive for authors to make their data public in an persistent way and to link to it in their publications. Only when there is a seamless network of repositories, publishers and discovery services, with transparent access to the creators of data and software, can a culture grow where data products are just a different type of publication (see e.g. Force11 declaration of data citation principles).

ACKNOWLEDGEMENTS

This work has been supported by the NASA Astrophysics Data System project, funded by NASA grant NNX16AC86A.

REFERENCES

AAS Editorial Board (2016). Policy statement on software. http://journals.aas.org/policy/software.html

Akers, K. G. (2013), Looking out for the little guy: Small data curation. Bulletin of the Association for Information Science and Technology 39, 58–59. doi:10.1002/bult.2013.1720390317.

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

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final]

Dorch, S. B. F., Drachen, T. M., & Ellegaard, O. (2015). The data sharing advantage in astrophysics. arXiv e-prints. arXiv:1511.02512

Gallaher, David, G. Garrett Campbell, Walter Meier, John Moses, and Dennis Wingo (2015). The Process of Bringing Dark Data to Light: The Rescue of the Early Nimbus Satellite Data. GeoResJ 6, 124–34. doi:10.1016/j.grj.2015.02.013.

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap, V., Mahabal, A., Siemiginowska, A., and Slavkovic, A. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology 10(4), e1003542. doi:10.1371/journal.pcbi.1003542

Heidorn, P. Bryan (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2), 280-299. doi:10.1353/lib.0.0036

Henneken, E. A., & Accomazzi, A. (2012). Linking to Data - Effect on Citation Rates in Astronomy. ASP Conference Series, 461, 763-766. arXiv:1111.3618

Henneken, E.A., Kurtz, M.J. & Accomazzi, A. (2012): The ADS in the Information Age - Impact on Discovery. Organizations, People and Strategies in Astronomy Vol. 1, Edited by Andre Heck, Venngeist, Duttlenheim, 253-263

Kohler, S. (2015). AAS Publishing News: Astronomical Software Citation Workshop. AAS Nova Highlights, 24 Jul 2015 (http://aasnova.org/2015/07/24/aas-publishing-news-astronomical-software-citation-workshop/)

Mooney, H. & Newton M. (2012). The Anatomy of a Data Citation: Discovery, Reuse and Credit. Journal of Librarianship and Scholarly Communication 1(1), eP1035. doi:10.7710/2162-3309.1035

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

Smith A.M., Katz D.S., Niemeyer K.E., FORCE11 Software Citation Working Group (2016): Software Citation Principles. PeerJ Computer Science 2:e86. doi:10.7717/peerj-cs.86

White, R.L., Accomazzi, A., Berriman, G.B., Fabbiano, G., Madore, B.F., Mazzarella, J.M., Rots, A., Smale, A.P., Storrie-Lombardi, L., and Winkelman, S. (2009). The High Impact of Astronomical Data Archives. In astro2010: Astronomy and Astrophysics Decadal Survey, no. 64. (http://www8.nationalacademies.org/astro2010/DetailFileDisplay.aspx?id=423)

Winkelman, S., Rots, A., Duffy, A., Blecksmith, S., & Jerius, D. (2006). Where Do the Data Go? An Analysis of Chandra Data Dissemination. ASP Conference Series, 351, 93-96.