Linking Scholarly Literature to Research Data and Software

Lesson Learned from Astronomy

Edwin Henneken

SAO/NASA Astrophysics Data System

Smithsonian Astrophysical Observatory - Cambridge, MA

CREATE

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* (data), the *Astrophysics Source Code Library* (ASCL) and *Github* (software). 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 Données 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. If a publisher requires data products used in publications to be available in a persistent manner, authors will find a way to meet them. The more transparent the process, the higher the adoption rate will be.

USE / DISCOVER

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 Astrophysics Data System (ADS) harvests and merges bibliographic data from multiple sources (arXiv, CrossRef, publishers, astronomy archives, ASCL). 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 (arXiv & published paper, translations, re-publications, VizieR catalogs, observing proposals). 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 (close to 10,000 records). Once in the ADS, they become easily discoverable, citable. This is how the astronomy community has dealt with "data citation" all along. 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 ~296,000 links to online data.

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.

Dataverse: <https://dataverse.harvard.edu/>, Zenodo: <https://zenodo.org/>, TheAstroData: <https://dataverse.harvard.edu/dataset/cfa>, ASCL: <https://ascl.net/>, Github: <https://guides.github.com/activities/citable-code/>

Mooney, H. & Newton M. (2012), "The Anatomy of a Data Citation: Discovery, Reuse and Credit", *Journal of Librarianship and Scholarly Communication* 1(1): eP1035

Evans, P. L., A. Strollo, A. Clark, T. Ahern, R. Newman, J. F. Clinton, H. Pedersen, C. Pequegnat (2015), "Why seismic networks need digital object identifiers", *Eos*, 96, doi:10.1029/2015EO036971

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>]

Henneken, E & Accomazzi, A. (2012), "Linking to Data - Effect on Citation Rates in Astronomy", *ASP Conference Series*, 461, p.763

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(8): e104798

MEASURE

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". These principles are supposed to encourage a good practice of data citation, and, as a result, create a data citation culture. Measures for the impact of data products will only be meaningful when such a culture is in place, just like it is for scholarly publications. The metrics service of the ADS, together with e.g. telescope bibliographies maintained by observations, is an example of how impact measures are generated for the community.

REWARD / REUSE

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 a persistent way and to link to it in their publications. The preamble of the Force11 declaration of data citation principles says: "Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse". 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.

@adsabs <https://ui.adsabs.harvard.edu> ads@cfa.harvard.edu

The ADS is operated by the Smithsonian Astrophysical Observatory under NASA Cooperative Agreement NNX16AC86A