Data Citation in Astronomy

J.A. Hourclé
NASA-GSFC (Wyle)
joseph.a.hourcle@nasa.gov

DC[1] Data Citation Principles

Virtual Solar Observatory

Overview

Many astronomical observatories maintain ***bibliographies*** to both ***document the impact*** of their work and help ***justify their continued funding***. [1,2] These efforts can be ***labor intensive*** as significant human effort is required to discover and curate the links between the scientific papers and the data that was used as evidence.

These efforts do not scale well, and require focusing on only a subset of scientific journals.

To better deal with the issues of tracking cross-discipline data usage, many groups came up with ***guidelines and principles for data citation***. [3,4]

Recently, the research community came together to create a single set of principles for data citation that could be endorsed by all groups.[5] We believe that implementation of these principles can help to ***improve the scientific ecosystem*** by giving ***proper attribution*** to all contributors to data, ***improving transparency and reproducability***, and ***making data more easily reusable*** to both astronomers and other researchers.

Joint Declaration of Data Citation Principles

Preamble

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.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a number of groups. As we move into the next phase, we welcome your participation and endorsement of these principles.

Principles

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

These citation principles are not comprehensive recommendations for data stewardship. And, as practices vary across communities and technologies will evolve over time, we do not include recommendations for specific implementations, but encourage communities to develop practices and tools that embody these principles.

...

IF I INCLUDED THE WHOLE TEXT HERE, THE POSTER WOULD BE AN EYESTRAIN. TO READ & ENDORSE, VISIT:
http://force11.org/datacitation

Interpretation & Analysis for the Astronomy Community:

Preamble:

- Reproducible science relies on knowing the evidence (data) used
- Producing data is an important contribution to science
- Citing data is important for the scientific record & re-use of data

1. Importance

- Data should be part of the scientific & scholarly record.

2. Credit and Attribution

- There is no simple "author" for data and citing a "first results" or "instrument" paper doesn't give credit to people who come in later and give significant contributions.
- It is unrealistic to name hundreds of people in the citation string.

3. Evidence

- You should **link the data being used as evidence** near the claim being made; depending on the journal, this may be inline text, a footnote, or a caption to a figure or table.

4. Unique Identification

- We need **cross-discipline identifiers**.
- We are currently leaning towards DOIs (Digital Object Identifiers) at the "collection" (data set) level.
- DOIs would allow us to use existing bibliographic tools to track the use of data, **reduce the work needed to prepare telescope bibliographies**, and **find uses of our data by other communities**.

5. Access

- Citations do not need to link directly to the data. **DOIs should link to a webpage with info about the data**.
- These "landing pages" can be updated to **provide links to current documentation, software, related data** or whatever the community or PI feels is appropriate for the data.

6. Persistence

- Even if the data goes away (replaced by better data, removed due to security or budget, or lost by accident), the landing page remains, so there is **no gap in the scientific record**.
- This "tombstone page" should describe why the data was removed, and link to possible replacements or alternatives (eg, new calibrations)

7. Versioning and Granularity

- If there are formal releases, assign a DOI to each one, so researchers can **cite a specific version**. If not available, citations should **include an access date**.
- If you didn't analyze all of the data, **describe what portion you used** (eg, date ranges, spatial extent, filters, specific observing modes).

8. Interoperability and Flexibility

- Every journal / community cites things differently. The data citation community is working towards a **universal framework that each community can extend for their specific needs**. (eg, add metadata, interface w/ ADS to get publication lists, etc.)

Data "Landing Pages":

For this scheme to work for collections of large data, we need a level of indirection between the citation string in the article and the data itself. We assume the identifiers will resolve to a page with information about how to cite, access and use the data being referred to.[6]

This allows us to cite collections of any size, that may not be online (eg, glass plates, embargoed, or in dark archives) or no longer exist (older calibrations). Additional landing pages may be created to create different citable slices from the same overall collection.

These pages may give credit to a longer list of contributors, link to related data, give additional metadata to allow indexing by search engines, or provide tools to interact with the data.

Citation Examples:

Formatting of the citation strings will vary by journal. A basic format might be:

Creator, Publication Year. "Title", Archive, Subset, Version. Identifier.

Note that all DOIs given are bogus; these are simply to provoke discussion:

NASA/SDO and the AIA Science Team, 2010. "SDO/AIA 171 Ångstrom, Level 1 Intensity Images", VSO, 2012-01-05 to 2012-01-07, accessed 2013-08-17. `http://dx.doi.org/10.example/sdo.aia.171.lev1`

NASA and ESA, 1996. "SOHO/MDI, Level 1.8 96min Doppergrams", SHA, 200arcsec patch around ARs 10652 and 10653, accessed 2005-11-01. `http://dx.doi.org/10.example/soho.mdi.v_96m_lev18`

STScI and AURA, 2012. "HST Wide Field Camera 3 DR6.1", Hubble Legacy Archive, http://dx.doi.org/10.example/hst.wfc3.dr6.1

Adelman-McCarthy, J., Agueros, M.A., Allam, S.S., et al. 2007, "Sloan Digital Sky Survey DR5". `http://dx.doi.org/10.example/sdss.dr5`

Dawson, K.S., Schlegel, D.J., Ahn, C.P, et al. 2012, "SDSS-III DR9 Baryon Oscillation Spectroscopic Survey". `http://dx.doi.org/10.example/sdss.dr9.boss`

As the groupings, titles and creators are set by those distributing the data, each archive or PI team can define whatever groupings are must useful to track the usage of their data. Authors then cite those groupings in a consistent manner.

How to implement:

Thanks to efforts from others representing groups with similar 'big data' collections, the policies should be implementable by our community. To proceed, we must:

1. **Determine our groupings of data[7]**
   - document the groupings
   - assign identifiers.
2. **Create 'landing pages' for the data**
   - Link to the data & documentation
   - Curate them over the long term
3. **Develop standards to describe subsetting in the citation string**
   - by spatial & temporal extent
   - by other common subsettings used in solar, planetary, astronomy, etc.
   - all others go into 'extended methods'
4. **Create tools for easy data citation**
   - don't place the burden on the scientists
   - 'receipt' when downloading data
   - software to identify FITS files on disk
5. **Outreach on importance of citation**
   - as documentation of the science
   - for continued project & archive funding
   - goal is more & better science

(Unresolved) ***Where should the landing pages be hosted?***

Should this be a society effort (AAS, EAS), a broader-scoped third party (ADS, arXiv), or does each discipline (astro, solar, planetary, etc.) run their own? Can we use existing system (zenodo, figshare) for this?

References

[1] Accomazzi, Hennekin, Erdmann & Rots, 2012. "Telescope bibliographies: an essential component of archival data management and operations". `http://adsabs.harvard.edu/abs/2012SPIE.8448E..0KA`

[2] Bishop, Grothkopf & Lagerstrom, 2012. "Best Practices for Creating an Observatory or Telescope Bibliography". `http://iau-commission5.wikispaces.com/file/view/Best+Practices+Final.pdf`

[3] National Research Council, 2012. "For Attribution--Developing Data Attribution and Citation Practices and Standards". `http://www.nap.edu/catalog.php?record_id=13564`

[4] CODATA, 2013. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data". `http://dx.doi.org/10.2481/dsj.OSOM13-043`

[5] 2014. "Joint Declaration of Data Citation Principles". `http://www.force11.org/datacitation`

[6] Hourclé et.al, 2012. "Linking Articles to Data". `http://virtualsolar.org/citation`

[7] Wynholds, 2011. "Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects". `http://dx.doi.org/10.2218/ijdc.v6i1.183`

The author is unable to attend this conference. Thank you to Sally Bosken for bringing & putting up this poster.

For links & more information, visit
`http://virtualsolar.org/citation`