Data Identifiers and Citations Enable Reproducible Science

Curt Tilmes

Curt.Tilmes@nasa.gov

AGU FALL MEETING 2011
San Francisco, California, USA | 5-9 December

Why do we cite data?

- Link your research with the data you use -
  - Allow search/data mining tools to associate your research with the data and with other research using the same data
  - Track the impact of particular data sets
- Provide formal credit (and accountability) to data creators and providers
  - Acknowledge the contribution of the data, just as you would any other paper you cite.
- Allow users of your research more precise insight into the inputs into your process
- ***Reproducibility***

When scientific research is published, it should ***reference*** all data used in that research to a sufficient extent for ***others*** to ***obtain*** the data, ***reproduce*** the research and ***confirm*** the conclusions.

- Current state of practice for citation of Earth Science Datasets is poor to non-existent (but improving!)
  - Some have acknowledgements
    - "Thanks to NASA for data"
    - "Thanks to Fred who gave me some NASA data"
    - "Thanks to MODIS team for MODIS data"
  - Some reference specific data inline, with footnotes or in figure captions
    - Used data from Terra MODIS instrument
    - Used Collection 5 Land Surface Reflectance data from Terra MODIS
    - Used Collection 5 Land Surface Reflectance data from Terra MODIS downloaded on 2011-02-08
  - A few have started to actually include formal citations in references
    - Even those usually cite the dataset as a whole, not specific granules used in research.

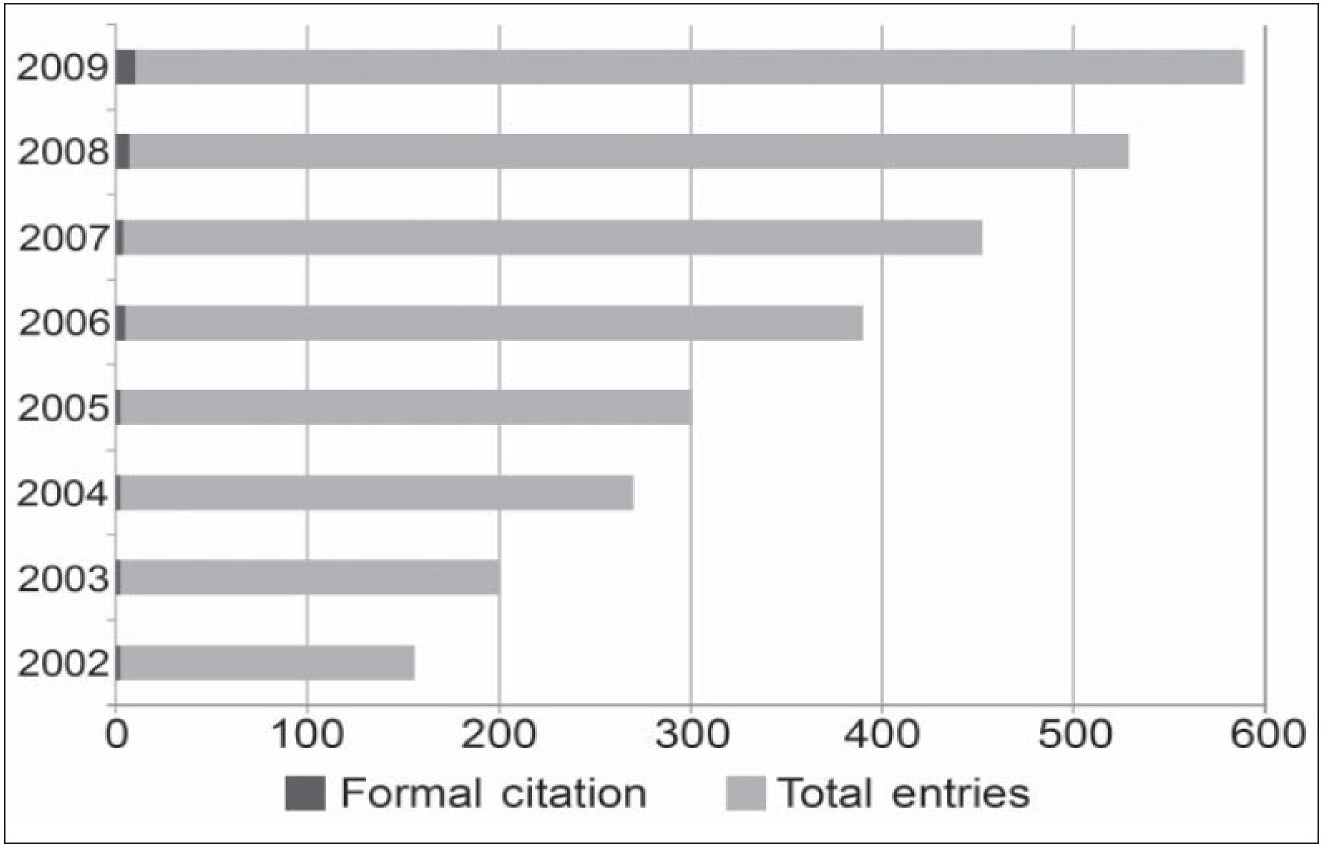

Fig 1. The National Snow and Ice Data Center distributes a variety of different snow cover products derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). The results of a quick analysis of how many scientific papers mention use of "MODIS snow cover data" (according to Google Scholar™) and how often the data sets themselves are formally cited show a huge disparity, illustrating the infrequency of proper data citation in practice. Moreover, the lack of data citation standards introduces the possibility that informal references to data do not point to the data set actually used.

[1] Parsons, et. al. "Data Citation and Peer Review"

- What do good formal citations look like today?
  - See the ESIP Federation Preservation and Stewardship Citation Guidelines
  - Zwally, H.J., R. Schutz, C. Bentley, J. Bufton, T. Herring, J. Minster, J. Spinhirne, and R. Thomas. 2003. *GLAS/ICESat L1A Global Altimetry Data V018, 15 October to 18 November 2003.* National Snow and Ice Data Center. Data set accessed 2011-07-21 at doi:10.3334/NSIDC/gla01.

- What do good formal citations look like today?
  - See the ESIP Federation Preservation and Stewardship Citation Guidelines
  - Zwally, H.J., R. Schutz, C. Bentley, J. Bufton, T. Herring, J. Minster, J. Spinhirne, and R. Thomas. 2003. *GLAS/ICESat L1A Global Altimetry Data V018, 15 October to 18 November 2003.* National Snow and Ice Data Center. Data set accessed 2011-07-21 at doi:10.3334/NSIDC/gla01.

Is that good enough?

Static vs. Dynamic Datasets

- Earth science remote sensing missions often have very long lifespans.
- Move to measurement based datasets makes these even longer, spanning multiple missions.
- ***Static*** dataset – A bunch of data go into the dataset and stay there.
- ***Dynamic*** dataset – New granules are frequently added to the 'end' of the dataset as time passes.
- For an operational mission, we also have operational issues that occasionally ***change*** older granules in the dataset.

- Basic configuration management works well for software.
- Any time the software is changed, we tag a snapshot with a revision number (v. 1.2.3) – We can go back and check out that version of the software, compare versions, etc.
- ***Data versioning*** is more complicated. The direct predecessors and the software that produced a given granule could have the same version, but due to changes 'up-stream' in the workflow, the data are different.
- Anytime a new granule is made, it has a ***distinct granule identifier***, even if it has 'equivalent' content..

- ***Reprocessing*** – Remake data granules in the best possible way

vs.

- ***Reproduction*** – Remake a product the “same” way it was made previously – goal is the create a new granule with equivalent content.

- ***Why reproduce instead of reprocess?***
  - Operational problems – disk crashes, data lost
  - Simply delete data that are less used to save disk space, “process-on-demand” when they are wanted (or not..)
  - End user trying to reproduce research.

- Scientists don't like things that change too frequently.
- We do “major” reprocessing in collections, batching up a bunch of changes at once.
- Could involve new calibration, new formats (hopefully minor changes..), new software versions throughout the chain.
- Each new collection should have a distinct identifier.
- The data content from old collections often get deleted, even if they are cited by published research.

- NASA ESDSWG and ESIP Federation study resulted in identifier recommendations [2] Duerr, et. al.
- We need a good way to distinguish particular granules from one another.
  - Globally unique, persistent identifier
  - ***UUID*** – Universally Unique IDentifiers provides a way for independent parties with no central coordination or registration to create such identifiers.
- We need a good way to reference datasets so they can be cited in scientific literature, and resolve back to an authoritative archive for that dataset.
  - ***DOI*** – Digital Object Identifiers provide a well-defined mechanism to attach an identifier to a digital object.
- We still need good identifiers to represent reproduced data, where matching granules have equivalent content.

Recommendations

- Data Users:
  - Read the Data Citation Guidelines and cite your data
  - Include the date/time a dataset was accessed
  - Describe the subset of granules you used

- Archives
  - Read the Data Citation Guidelines and recommend good citations.
  - Assign persistent identifiers for data granules and data sets
  - Provide a way to resolve a dataset access date/time to a particular set of granules
  - Provide a persistent description of provenance of data, even if you have to delete the data

- Journal editors and paper reviewers
  - Ensure the data are cited properly

- NASA EOSDIS has a plan to assign DOIs to all of their datasets.

- ESIP Federation to baseline citation recommendations, please read, use, encourage others to cite their data.

- DOIs are great, but particularly for a dynamic dataset, ***not sufficient*** to identify the precise granules used, especially if they no longer exist.

- We're working on more precise identifiers and recommendations to facilitate these use cases, please join us :

  ESIP Federation Preservation and Stewardship Cluster

References

[1] M. A. Parsons, R. Duerr, and J.-B. Minster. *Data Citation and Peer Review*. EOS Transactions, 91:297–298, August 2010.

[2] Ruth E. Duerr, Robert R. Downs, Curt Tilmes, Bruce Barkstrom, W. Christopher Lenhardt, Joseph Glassy, Luis E. Bermudez and Peter Slaughter. *On the utility of identification schemes for digital earth science data: an assessment and recommendations*, Earth Science Informatics, Vol. 4, Num. 3, 139-160, 2011, doi:10.1007/s12145-011-0083-6.

ESIP Federation Preservation and Stewardship Cluster

- http://wiki.esipfed.org/index.php/Preservation_and_Stewardship

Data Citation Guidelines

- http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations

Thank You!