Editorial: **Citations to Published Data Sets**

Robert Cook

The last issue of Fluxletter (Vol. 1, No. 3) contained an editorial with suggestions and guidelines for co-authorship of papers representing the efforts of many flux tower investigators, such as in the collection of La Thuile Synthesis papers (Baldocchi and Vargas 2008). The FLUXNET community has evolved to the point where multi-site integration and synthesis is possible and necessary to advance our science. The community has a wealth of data and information on the "breathing of the biosphere" and integrated analysis of this information across many sites and biomes helps us understand processes controlling these fluxes. How can the contributions of each of the investigators providing data to these syntheses be appropriately recognized?

One of the suggestions from the editorial was to publish data sets at a data archive. There are many advantages to this suggestion.

A permanent data archive enables users to search for, access, and download published data sets. The finalized and published data sets can be cited, giving the data producers credit. Citations to these published data sets enable a student or a researcher to obtain the actual data files from the archive to reproduce the results from papers or to conduct further analyses. The scientific method requires that the information necessary to support published results be made available to other researchers. Publication of the data themselves makes available the information necessary to reproduce the findings. Published papers in AFM, JGR, or Nature provide a description of the methods, analysis, and results, but typically do not provide a way to access the data files themselves. There are a few notable exceptions. Some disciplines (e.g., biotechnology) require publication of data before a paper is published. ESA has a journal –Ecological Archives– that publishes data papers, supplements, and digital appendices for ESA journals.

Government agencies are under increasing pressure to show the benefits of the research they sponsor, both in terms of scientific findings—published papers—as well as data products.

"...A permanent data archive enables users to search for, access, and download published data sets. The finalized and published data sets can be cited, giving the data producers credit. ."

A recent US Government and Accounting Office Report summarizes the issues associated with loss of individual investigator's data and therefore the loss of some of the benefits of research, and outlines some solutions (http://www.gao.gov/products/GAO-07-1172).

Publication of scientific findings in journals has traditionally been the way the benefits were measured. In addition, publication of data sets in a long-term archive demonstrates the products of successful research. Subsequent use of the data sets in other papers as indicated by citations to these published data sets by other researchers in the coming years is also a measure of the value of sponsored research.

The ORNL DAAC, the long-term archive that houses the FLUXNET data, has a citation policy for finalized archived data: http://daac.ornl.gov/citation_policy.html

For example, this is our suggested citation for work conducted at an eddy covariance flux tower in Brazil:

Hutyra, L., S. Wofsy and S. Saleska. 2007. LBA-ECO CD-10 CO_2 and H_2O Eddy Fluxes at km 67 Tower Site, Tapajos National Forest. Data set. Available online at [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee. U.S.A. doi:10.3334/ORNLDAAC/860

The data citation gives credit to the data providers and gives users an opportunity to locate the data files. In addition, the ORNL DAAC, as publisher, gets credit for hosting data sets that are used to advance environmental research.

In the example above, Lucy Hutyra and her colleagues get credit for a data set citation--not weighted as high as a peer reviewed publication to be sure, but important in and of itself. This credit is one incentive to get data sets archived and shared. And it is measureable, so that research sponsors can see that that LBA data archived at the ORNL DAAC have been used in peer reviewed publications.

The ORNL DAAC has recently started assigning Digital Object Identifiers (DOIs) to the DAAC's collection of published data sets. The use of DOIs facilitates the ability of authors to cite data publications in refereed journals. Many journals now require DOIs when citing online material. DOI's enable users to locate a published data set regardless of where it is stored. This immutable identifier is part of the DOI system and the Hutyra data set can be found by searching for the DOI (10.3334/ORNLDAAC/860) or by searching at the DOI System's Web site: http://dx.doi.org/10.3334/ORNLDAAC/860.

For the archived data associated with a BOREAS flux tower, we suggest the following citation:

Wofsy, S. C., and A. Dunn. 2001. BOREAS Follow-On FLX-01 NSA-OBS Derived Data - NEE, GEE, and Respiration. Data set. Available on-line at

EDITORIAL cont. on page 5

Citations to Published Data Sets

cont. from page 4

[http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.
doi:10.3334/ORNLDAAC/603

For the Wofsy and Dunn (2001) data publication, the documentation about the data set requests that users seek updated and more recent versions of the data from the individual investigators. The citation and the DOI refer to this finalized and archived version of the data from 2001. This version is archived so that a record is kept of data associated with the BOREAS study and any publications associated with the data.

We encourage investigators to add data sets to their resume. This practice will enable investigators to show a product of their research and demonstrate to sponsors that they are complying with policies to share data products. And data product authors will be able to use standard citation tools to determine how many times their data set has been used by others in future publications.

As you know, preparing data for future use is viewed as a time-consuming and heretofore thankless task that can serve as a disincentive to data sharing. The benefits to an investigator are the publication of the finalized data themselves, credit and recognition for use of data by others, and demonstrating compliance with sponsors' data sharing policies. Publishing data at an archive also takes the burden off of the investigator for maintaining the information on-line and dealing with user requests and questions.

A scientist is recognized within the scientific community, proposal review panels, and Promotion and Tenure Committees based in part on publication in the peer reviewed literature. Until recently, preparation of finalized data to share with others has not been a part of this reward structure. Many sponsors of flux tower and other environmental research are using compliance with data policies as part of their proposal evaluations and the GAO report suggests that more sponsors will do so in the future.

Publishing data products and citing those products in the literature are relatively new practices, and we are seeing more citations to published data sets. Use of a standardized citation format along with immutable DOIs will promote this practice of publishing and sharing data sets. And this practice of publishing finalized data products will promote additional flux synthesis studies, advancing our understanding of the processes that control the "breathing of our biosphere," while giving credit and recognition to those who have made the observations.

contact: Robert B. Cook
Environmental Science Division.
Oak Ridge National Laboratory,
Oak Ridge, Tennessee, USA
cookrb@ornl.gov

This work was funded by NASA's Earth Science Data and Information Systems Program (NASA Project No. NNG04HU031).

Literature

U.S.GAO. 2007 . Climate Change Research: Agencies Have Data-Sharing Policies but Could Do More to Enhance the Availability of Data from Federally Funded Research. GAO-07-1172. Available on-line at http://www.gao.gov/products/GAO-07-1172.

Baldocchi DD, Vargas R. 2008. Whither co-authorship? FluxLetter 1(3) 3-4. Available on-line at http://www.fluxnet.ornl.gov/fluxnet/FluxLetter%20Vol1%20No3.pdf