Cyberinfrastructure solutions for scientific attribution: A dilemma of the digital age

Matt Pritchard, Cornell University, UNAVCO Board Member
On behalf of the 2009, 2010 and 2011 Boards of Directors, UNAVCO &
M. Meghan Miller, UNAVCO President

In our increasingly digital age, access to data is becoming faster and easier, with many positive consequences for scientific discovery, but also with challenges for protecting the data collectors and ensuring proper citation (e.g., Parsons et al., Eos v. 91, no. 34, 2010). These issues have been on the minds of the staff and those involved in the governance of UNAVCO (a non-profit consortium of institutions that facilitates collection and archiving of geodetic data) as they crafted a new data policy effective January 1, 2011. In this white paper, we describe the goals of the new data policy and outline some of the issues that arose when considering implementation that might be relevant for other research communities.

The new UNAVCO data policy states "All UNAVCO-facilitated GPS and other GNSS data and metadata must be archived at UNAVCO upon collection" (available at www.unavco.org/community/policies_forms/DataPolicy.html). The vision of the data policy is to create a world where all data is made publically available as soon as possible and is more restrictive than that required by the National Science Foundation's policy (www.nsf.gov/bfa/dias/policy/dmp.jsp). For many datasets that are collected by a large community of researchers, this vision is already achieved, as the data are immediately and readily available. What is new is that this data policy also applies to data collected by individual researchers -- also known as PI-led projects.

This open access to PI-led data requires that researchers follow a code of ethics where they do not publish a scientific article using someone else's data without appropriate co-authorship, permission, or proper citation if already published. The current policy allows exceptions to this model (i.e., a period of sponsor-approved exclusive use), but these must be justified in the proposal and evaluated in peer review. If justified and approved by peer review, the new policy allows for periods of exclusive use for continuous stations that were not allowed in the previous data policy. The period of exclusive use will be documented by the sponsor of the research project, typically in the award letter.

While crafting the data policy and this article, we have reviewed the ethical standards with regards to data availability and scientific publication from several organizations (links to these are available at: www.unavco.org/community/policies_forms/). While many of the ethical guidelines are not sufficiently explicit on the point of co-authorship, the ethical code mentioned in the previous paragraph is articulated in the policies of the journal Nature: "Manuscripts are sent out for review on the condition that any unpublished data cited within are properly credited and the appropriate permission has been sought" (www.nature.com/authors/policies/plagiarism.html). As the community voice for governance, we clarify that the phrase "properly credited" means that the collectors of the unpublished data are either co-authors of the submitted manuscript or have given their permission in writing to the authors for the use of the data in this

particular manuscript. For example, it would not be sufficient to take a picture of someone's otherwise unpublished data at a scientific conference and use it in a publication if while only citing the abstract of the presentation. Similarly, it is generally not acceptable to use unpublished data from a website in a peer-reviewed article when only the URL is cited.

There were several reasons that the new data policy was created. One was that the geodetic research community espoused the ideal of open data in the UNAVCO strategic plan of 2009-2013. While the UNAVCO data policy has always been for open data access, the new policy no longer allows for an automatic period of exclusive use. The consensus was that open data would foster new discoveries and provide the best value to the taxpayers on their investment to collect the data. Another was to simplify the previous data policy by creating a single classification of data instead of multiple categories. "Campaign" and "continuous" data are increasingly blurred as deployments of long yet finite term proliferate.

The new data policy will require active participation of data users, those who maintain the data archive, data providers, sponsors of funded research, and peer reviewers and editors of scientific publications. These requirements are essential to protect the "sweat equity" of those who spend the considerable time and effort to go to the field to make the measurements.

Responsibilities of the data collector: To submit data upon collection to the UNAVCO archive with proper metadata unless a period of exclusive use has been granted.

Responsibilities of the community of peer reviewers of scientific proposals and the sponsors of funded research: 1) Evaluate whether any requests for periods of exclusive use of the data are justified. There are many different issues to consider when evaluating requests for exclusive use, and the expertise of the peer reviewer will be called upon. Some possible justifications for a period of exclusive use might include the involvement of graduate students learning how to process the collected data and the requests of foreign collaborators. Certain experiments may require longer periods of exclusive use than others – for example, projects studying slow processes may require a longer period of exclusive use. 2) Assess whether the proposer has adhered to the UNAVCO data policy in previously supported work.

Responsibilities of the data user when writing a scientific publication: (1) The source of the data must be cited -- not just the data archive (e.g., UNAVCO) but (where appropriate) the PI who collected the data via a digital object identifier or doi. (2) Users of data that are identified in the archive as part of an ongoing project should consult the PI.

Responsibilities of the editors and peer reviewers of scientific articles: To ensure that all data used are properly cited – including the data archive and PI who archived the data, and determining if the PI's research project is still ongoing. Every reviewer and editor should pay special attention to the datasets being used and make sure that either the PI

who collected the data is involved in their publication or that the work is appropriately cited.

The Cyberinfrastructure challenge is to provide the tools for appropriate attribution and transparency. Responsibilities of those who maintain the data archive (UNAVCO) are to keep a long-term record of datasets and metadata including doi's and datasets that are still the subject of ongoing work by the PI's. The UNAVCO archive needs a new generation of tools to support authors using community and individual data sets these responsibilities.

The creation of the new data policy has provoked much discussion among members of the UNAVCO community, and we hope that this paper will contribute to this discussion. The guidelines and policies that we outline here are living documents that will change and improve with community advances and in implementation of Cyberinfrastructure solutions for appropriate attribution.