Defining the Data Citation Problem in the DataNet Context

John Kunze, California Digital Library, University of California Office of the President, Oakland, CA
Robert Cook, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN
Patricia Cruse California Digital Library, University of California Office of the President, Oakland, CA
Carol Tenopir, School of Information Sciences, University of Tennessee, Knoxville, TN
Todd Vision, Department of Biology, University of North Carolina, Chapel Hill, NC
William K. Michener, University Libraries, University of New Mexico, Albuquerque, NM

While data-driven Earth science research is usually recognized by its presence in the published literature, there is no mainstream publishing system for the original data behind that research. Establishing the practice of publishing the data themselves, along with descriptive metadata, would bring a number of benefits for the advancement of science. The published data could be discovered, accessed, understood, and used to verify the findings of the original publication or could be used by themselves or combined with other data to address new hypotheses.

One obstacle in the publication of data products is the lack of effective data citation practices. Citation of published and archived data can provide a bridge from the traditional literature to directly supporting evidence. The citations to data products that are already appearing, however, have significant gaps.

Citations to published data should acknowledge the data collectors, thereby providing a means to recognize the professional value of data production. The citation should include a machine readable unique identifier so that readers can find, obtain, and understand the data. Ideally, published data should have a permanent identifier that can be resolved far into the future. Without long-term stewardship, this data and the scientific record that builds upon them are at risk.

The components of a data set citation, while similar in many ways to those of a traditional citation, need to be defined. Because data products can be revised for any number of reasons, a citation must refer to the precise version that was intended. Furthermore, authors should be able to cite at an appropriate level of specificity; for example, with a data product consisting of multiple files, it may be desirable to cite either the entire data product or one particular file within it.

For DataONE (Observation Network for Earth), we are evaluating the current landscape of citations for data products and working with the international community to develop and promote best practices for data citation.