Geophysical Research Abstracts
Vol. 14, EGU2012-11999, 2012
EGU General Assembly 2012

Bridging data lifecycles: tracking data use via data citations

M. Mayernik, K. Kelly, M. Marlino, and M. Wright
National Center for Atmospheric Research (NCAR), Boulder, CO, United States (mayernik@ucar.edu)

The "lifecycle of data" encompasses data conceptualization, collection, evaluation, use, management, and preservation. Properly conceptualized, the idea of a "lifecycle of data" should also stress the desire to enable data to be re-purposed beyond their original use. In fact, much of the digital data informatics work in many disciplines focuses on this goal of making data available for secondary uses. In thinking about lifecycles of data as recursive phenomena, where data lifecycles are built upon each other in increasingly complex ways, understanding and tracing the use of data over time is of critical importance to curating data effectively over time.

At the National Center for Atmospheric Research (NCAR)/University Corporation for Atmospheric Research (UCAR) in Boulder, Colorado, USA, we are currently developing a formal organization-wide data citation policy and implementation. "Data citations" are increasingly seen as being critical to enabling scientific results to be traced back to their underlying data. Data citations also promote the transparency of scientific work by making data more discoverable, and enable scientists, data managers, and data centers to be credited for producing useful data.

NCAR/UCAR provides access to many thousands of data sets, including decade-to-century long time-series' of meteorological and atmospheric observations, climate model output data sets, and data sets from experimental field deployments. Our data citation initiative includes assigning Digital Object Identifiers (DOIs) to data held by NCAR/UCAR data centers, developing policies and procedures for using DOIs to trace the use of data over time, and enabling data users to cite data by providing them with citation information.

Assigning DOIs to data sets and promoting data citations are tasks that cross the data lifecycle. For example, data sets often go through many processes before they are made available online and the principal investigators of a project often produce the first publications from a data set before the data are made available online. Thus, it can be very difficult for data providers to decide when to assign a DOI to a data set. Secondly, DOI must be maintained over time, just as data sets must be maintained over time. Data providers have the responsibility to ensure that DOIs are always up-to-date. Additionally, data providers who assign DOIs to data must develop ways of counting data citations over time and providing robust links between data sets and publications to ensure two-way discovery of related resources.

Our talk will describe our ongoing development of policies, procedures, and technologies for assigning DOIs and promoting data citations at multiple stages of the data lifecycle. As methods for assessing the impact of data citations develop, these impact assessments can promote increased rewards for scientists who produce data and make those data available and usable as the building blocks for a secondary data lifecycle.