Making Data Visible: Moving Towards Accessibility, Openness, and Transparency through Data Citations

Matthew Mayernik
Research Data Services Specialist
National Center for Atmospheric Research (NCAR)
University Corporation for Atmospheric Research (UCAR)

18th Biennial AMS/AGU Joint Heads and Chairs Meeting
UCAR, Oct. 19, 2012
Data

EarthCube

YELLOWSTONE
petascale computing for the atmospheric sciences

Science

nature

BIG DATA

SCIENCE IN THE PETABYTE ERA
Connecting data and publications

• Have you published data as an end in itself?
• Have you archived your data in a formal archive?
• Have you seen students having difficulty in locating data for research or course work?
Data Issues

• How to get data in the hands of researchers and students?
  – Finding data on the internet can be very difficult
  – Most data are discovered through the literature

• How to point to data sources?
  – Reproducibility and transparency of science and data
  – Digital data are largely linkable online

• How to reward data collection and compilation?
  – Data collection, management, and processing require significant expertise and effort
  – Professional credit structures rarely account for data work
From:
Motivation for Data Citations

• For Data Centers
  – Make data more findable and accessible for re-use
  – Measurements of data use through counting citations
    • Demonstrate value of data services
    • Help improve services

• For Data Creators/Collectors
  – Clearer understanding of how data are re-used
  – Get professional credit for collecting and managing data

• For Data Users
  – Connecting scholarship to underlying data
  – Increase transparency of data and science
UCAR Data Citation Initiatives

• Develop a coherent approach across the organization for:
  – Technical tools and methods
  – Policy/procedural protocols and standards

• Broad participation from UCAR/NCAR teams
Technical: DOIs for Digital Resources

- Digital Object Identifiers (DOIs)
  - DOIs provide a persistent locator for internet-based resources
  - Widely used for scholarly papers
  - Growing use for other kinds of digital resources
    - Data
    - Software
    - Educational materials
    - Gray literature (Reports, proceedings, etc.)
DOI for Data

http://dx.doi.org/10.5065/D6RN35ST resolves to http://www.earthsystemgrid.org/project/NARCCAP.html
Procedures for Citations & Identifiers

• Archives
  – Who should be assigning identifiers
  – How to assign identifiers to resources
  – How to provide recommended citations to users
    • *Recommendation*: Author. Release date. Title. Archive/Provider. Locator/Identifier. Data access date.
    • *Optional*: Version; Subset Used; Editor, Compiler, or other important role; Distributor, or other Institutional Role.

• Outreach to users
  – Make citation information easy to find/use
  – Proactively inform users about citations and identifiers
Citation Challenges

1. Diversity

2. Granularity

3. Version Control

4. Maintenance Over Time
UCAR/NCAR Progress

• Community and consensus building
• Assigned ~50 identifiers
• Producing technical reports
  – White paper on identifier and citation recommendations
  – Report from UCAR data citation workshop held in April
• Organizing citation implementations within individual groups
  – CISL RDA, VETS
  – EOL CDS
  – HAO MLSO
  – NESL CGD (CESM)
  – NCAR Library
Interest in Data Citations

• Federal agency emphasis - National Science Foundation (NSF), National Academies of Science

• 2009 American Geophysical Union (AGU) position statement endorses “the concept of publication of data, to be credited and cited like the products of any other scientific activity...”

• 2009 AMS Ad Hoc Committee on Data Stewardship Prospectus:
  
  “Develop a plan for citing data referenced in publications and preserving data links for the long term.”
NSF GEO Interest in Data Citations

“Dear Colleague Letter” issued on March 29, 2012

DATE: March 29, 2012

Subject: Data Citation in the Geosciences

Facilitating open and equal access to data and data sets is a fundamental operating principle of the Directorate for Geosciences (GEO), and the National Science Foundation (NSF) as a whole. Investigators are expected to share with other researchers, at no more then incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grants are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI D.4.

GEO believes that the benefit to the over-arching scientific enterprise from access to data and data sets far outweigh the burden of time and resources to an individual investigator and his or her host institution. GEO encourages data citation as a means to achieve the desired operating state for the geosciences with open and equal access to data available to all interested parties at a reasonable cost.

Principles of data citation are at various stages of maturity and adoption among scientific and engineering communities. In a 2009 report, for example, the American Meteorological Society (AMS) was urged by its Ad Hoc Committee on Data Stewardship Prospectus to "develop a plan for citing data referenced in publications and preserving data links for the long term." The American Geophysical Union (AGU) has taken the position that "the scientific community should recognize the professional value of data activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer review of such public data."

While many policy and practical challenges remain to be resolved and implemented, the Directorate for Geosciences encourages members of the community to lead an evolutionary transformation to establish data citation within the geosciences as the rule rather than the exception.

The Australian National Data Service lists many references to the benefits of and practices for data citation (http://andn.org.au/dsa-data/resources.htm#Data_Citation_Benefits). Benefits include the acceptance of research data as a legitimately citable contribution to the scientific record; permitting results to be verified and re-purposed for future study; and enabling data citation metrics to be tracked, as is done with publications. Also, data citation is one mechanism for complying with the long-standing NSF policy of data sharing (see Award and Administration Guide, Chapter VI D.4, http://www.nsf.gov/pubs/policydocs/pappguide/nsf11091/eaag_6.jsp#V1D4). An example of a data facility that currently assigns Digital Object Identifiers (DOIs) to datasets that investigators submit to its repository is the NSF-funded data facility Integrated Earth Data Applications (IEDA, www.ieda-data.org). The DOI can then be used in publications to cite the data (e.g., see citation of the GSNRT Synthesis dataset in Ryan, W.P.D., et al., S-Cubed, 2009). The DOI resolves to the bibliographic metadata of the dataset.

Thomson Reuters

- Thomson Reuters Data Citation Index just released, Oct. 16, 2012*
  - Integrated into the Web of Science citation indices
  - Details of scope, functionality, and cost TBA

Data Citation Practices

- Data citations are uncommon
- Ex. “MODIS snow cover data” from NSIDC

Example from:
Data Citation Activity

- World Data Center for Climate/British Atmospheric Data Centre

- NASA

- Oak Ridge National Laboratory

- National Snow and Ice Data Center
Institutional Challenges

• The need to fill data curation gaps
  – DOIs and data citations have implications of persistence
  – Many scientists do not formally archive their data

• Persistence of data and citations. Ex. AGU policy*
  – “Data cited in AGU publications must be permanently archived in a data center or centers”
  – “Data sets that are available only from the author...may not be cited in AGU publications”

Archiving Data – Science Survey

Where do you archive most of the data generated in your lab or for your research?

- 50.2% Our Lab
- 38.5% University servers
- 7.6% Community repository
- 3.2% Other
- 0.5% It is not stored

Archiving Data – Science Survey

UCAR Data Citation Group Approach

• Not all data can/should be archived or cited
• Take a local and divisional approach to data archiving and citation decisions
• Formulate an institutional data policy
  – What should policy be with respect to archiving and citation?
  – Ex. should all data of long-term value be submitted to a repository? Data used to produce a publication?
Questions

• How does your institution support data management and archiving?

• How do data management and archiving fit within:
  – geoscience education?
  – professional reward structures (hiring, promotion, etc.)?

• How can data citation be promoted and enabled as a way to link publications to their data sources?
Thank You

- NCAR/UCAR data citation white paper draft available at: https://wiki.ucar.edu/display/DatacitePublic/

- Workshop at NCAR, April 5-6, 2012: “Bridging Data Lifecycles: Tracking Data Use via Data Citations”
  - Presentation slides: http://library.ucar.edu/data_workshop/
  - UCAR AtmosNews - “Don’t Forget the Data”: http://www2.ucar.edu/atmosnews/features/7138/dont-forget-data

Email:
mayernik@ucar.edu