The Case for Data Management

Ruth Duerr
National Snow and Ice Data Center

Overview

- Preserving the scientific record
- Facilitating science through data sharing
- Return on investment
- Enhancing your reputation
- Responding to agency requirements

The Scientific Record

- The scientific record is an aggregation of:
  - Scientific journals.
  - Conference presentations and proceedings.
  - Technical reports and pre-prints.
  - The underlying data, software, and other evidence to support published findings.
- This aggregation is highly distributed across:
  - Libraries, archives, and museums.
  - Data centers.
  - For-profit publishers.
  - Investigator web sites.
  - Investigator file cabinets and hard drives.

Purpose of the Scientific Record

- *Communicating findings, hypotheses, and insights* from one person to another, across space and time.
- *Organizing scientific communities*
  - Establishing common nomenclature and terminology.
  - Connecting related work.
  - Developing disciplines.
- *Documenting, managing,* and *resolving* controversies and disagreements.
- *Establishing precedence* for ideas and results.
- *Offering evidence for the quality and significance* of scientific work through citation and bibliometrics.

Challenges to the Scientific Record

- Two major challenges to the scientific record
  - Increasing complexity of experiments and data cause the linkages between evidence and writings to become more complex and elusive.
  - Increasing rate of the growth of the literature and data.
    - Disciplines and sub-specialties branch and grow continuously.
    - Tools and practices were developed to help manage literature: specialized journals, citations, indices, review journals and bibliographies, managed vocabularies, and taxonomies in various areas of science.
    - These tools and practices are either non-existent or just beginning for data.

Preservation of the Scientific Record

- Tenets of the scientific record
  - That scientific products are trustworthy.
  - That scientific products enable results to be reproducible and/or transparent.
- Preserving the scientific record: Data considerations
  - Are data stored in a trustworthy institutional setting?
  - Are data documented in a way that ensures understandability, reproducibility, and transparency over time?

Facilitating science through data sharing

- Solving modern complex scientific problems typically requires access to a wide variety of multi-disciplinary data.
- Even small data sets produced by a researcher or a small research team can be useful to a broader community.

Return on investment

- Who should own the data from publicly funded research?
  - Studies have shown that making publicly funded Earth science data openly and freely available drives economic growth.
  - Similarly, studies have shown that there are positive public health and well-being impacts if publicly funded data is available.
- What’s the return on your investment if your data is lost due to poor data management practices?

Scientific Reputation

- Reputation is central to the scientific community.
- Researchers build a reputation by producing valuable results, contributing constructively to scientific debates, and being good colleagues.
- Peer-recognition influences one's employment opportunities, promotion at work, and ability to win further research funding.

Reputation and Data - Why

- Data re-use is growing in importance in almost all scientific fields.
  - Data re-use depends on the availability of trust-worthy data sets.
  - Trust in data is highly connected to the reputation of the data collectors and data archives.
- Having a reputation for collecting and sharing high quality and well documented data makes it more likely that:
  - Other researchers will use your data.
  - Other researchers will cite your data.
  - Other researchers will share their data with you.

Reputation and Data - How

- How to get a reputation for data management?
  - Make data openly accessible by submitting to open data archives.
  - Provide comprehensive metadata.
  - Answer questions from data users in a timely manner.
- How to ensure that reputations for data management can grow?
  - Provide proper attribution when you use data collected by someone else.
  - Cite data sets in your reference lists.
  - Teach proper data management and data attribution to new scientists.

Overview

- Most agencies have or are developing data management policies or guidelines.
- As an investigator:
  - You may need to discuss data management in your proposals in order to obtain funding.
  - You may need to follow agency mandates in regards to managing or archiving the data you generate in order to retain funded status.

Agencies with Proposal Requirements

- NSF requires a two page data management plan submitted with all proposals.
- Many NASA Earth Science solicitations require a discussion of data management in submitted proposals.

Agencies with Data Management Mandates

- Many NSF directorates and even specific solicitations have specific requirements for what data should be archived and where the data should go
- All NASA Earth Science science missions, projects, and grants and cooperative agreements require a data management plan
- NOAA offices, contractors and partners receiving NOAA funding must manage environmental data in compliance with Federal requirements and directives

Acknowledgements

- Matt Mayernick
  - National Center for Atmospheric Research, NCAR

References and Resources

- Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen Moorman D, Uhlir P, and P Wouters. 2004. “Promoting Access to Public Research Data for Scientific, Economic, and Social Development,” *Data Science Journal* Volume 3,.
- Hanson, B., A. Sugden, and B. Alberts. 2011. “Making data maximally available.” *Science* 331(6018): 649. http://dx.doi.org/10.1126/science.1203354
- Lynch, C. 2009. “Jim Gray’s Fourth Paradigm and the Construction of the Scientific Record.” In *The Fourth Paradigm: Data-Intensive Scientific Discovery,* edited by T. Hey, S. Tansley, & K. Tolle,137-146*.* Redmond, WA: Microsoft*.* http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_lynch.pdf
- Piwowar HA, Day RS, Fridsma DB, 2007 “Sharing Detailed Research Data Is Associated with Increased Citation Rate”. PLoS ONE 2(3): e308. doi: 10.1371/journal.pone.0000308

References and Resources

- Uhlir, P.F. and P. Schröder. 2007. “Open data for global science.” *Data Science Journal* Volume 6. http://www.jstage.jst.go.jp/article/dsj/6/0/OD36/_pdf
- Costello, M. J. 2009. “Motivating Online Publication of Data.” *BioScience* 59(5): 418-427. http://www.jstor.org/stable/10.1525/bio.2009.59.5.9
- Milinski, M., D. Semmann, and H.-J. Krambeck. 2002. “Reputation Helps Solve the ‘Tragedy of the Commons’.” *Nature* 415(6870): 424-426. http://dx.doi.org/10.1038/415424a
- Barton, C., R. Smith, and R. Weaver. 2010. “Data practices, policy, and rewards in the information era demand a new paradigm.” *Data Science Journal* 9(12). http://www.jstage.jst.go.jp/article/dsj/9/0/IGY95/_pdf

References and Resources

- UC3 overview of Federal Funding Agency Data Management and Sharing Policies: http://www.cdlib.org/services/uc3/datamanagement/funding.html
- Agency Guidance:
  - NSF Guidance: http://www.nsf.gov/bfa/dias/policy/dmp.jsp
  - NASA Data Policy: http://science1.nasa.gov/earth-science/earth-science-data/data-information-policy/
  - NOAA Administrative Order 212-15: http://www.corporateservices.noaa.gov/~ames/NAOs/Chap_212/naos_212_15.html

References and Resources

- NSF Data Policy

  NSF. 2011. “Award and Administration Guide.” http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4.

- NSF Data Management Plan Requirement

  NSF. 2011. “Grant Proposal Guide, Chapter II.C.2.j.” http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.

- NSF Directorates with specific guidance

  NSF. 2011. “Dissemination and Sharing of Research Results.” http://www.nsf.gov/bfa/dias/policy/dmp.jsp.