AMS 2013

What Does Peer Review of Data Sets Mean and What Roles do Data Archiving and Quality Control Have in the Process?

Matthew S. Mayernik[1], Mike Daniels[1], Christopher Eaker[2], Gary Strand[1], Steven F. Williams[1], Steven J. Worley[1]

1. National Center for Atmospheric Research (NCAR), 2. University of Tennessee-Knoxville, School of Information Sciences

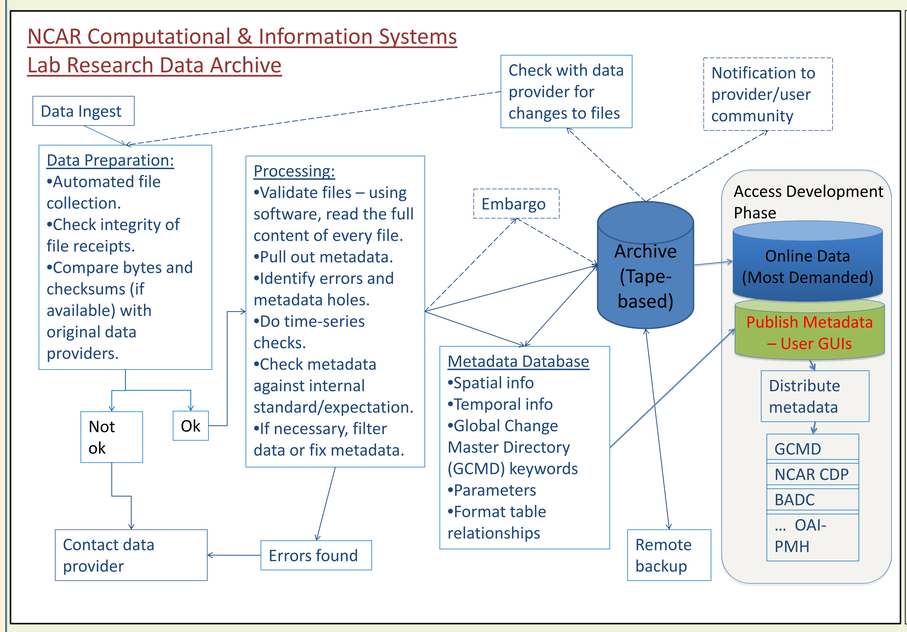
Data Peer Review

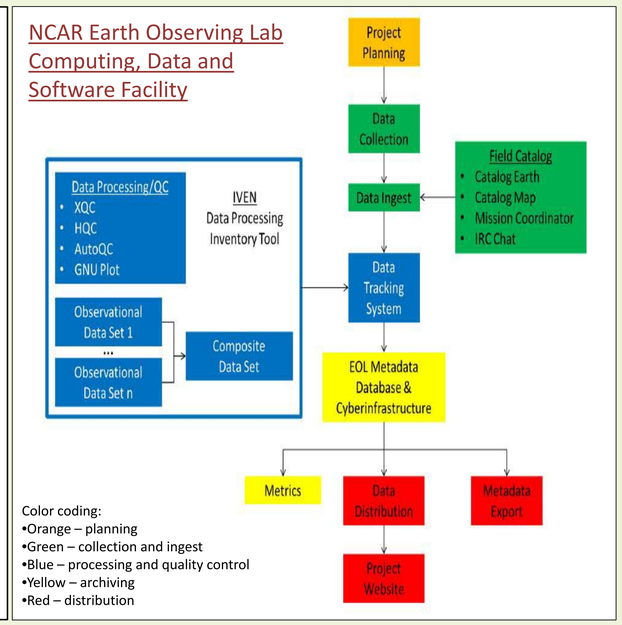
- Peer review is a central way to assess research quality
  - Cornerstone of professional reward structures, e.g. hiring, promotion, and tenure.
  - Integral to IPCC Assessment Report writing and review process
- How does peer review apply to data publication and citation initiatives?
- Peer review of growing volumes of digital data will increase the stress on the scholarly publication system.

Data Peer Review Challenges

- Data QC processes and software are very specific to data types, experimental designs, and systems
- What needs to be reviewed? All data versions? The metadata? Associated data papers? All of them?
- Different people have different expertise, e.g. scientists, data managers, software engineers. Would one reviewer be qualified to review all aspects of a data set?
- Human examination vs. Automated review
  - Human examination of data access interfaces, documentation, and metadata is essential to assess suitability for users.
  - Visual exams of data and metadata characteristics are often very important to identify systematic flaws.
  - If the data and metadata are published in standard form, readily available tools can be used to automate some data evaluation.
  - When possible, automation is desired to reduce the time and effort on the part of the human reviewer.
- Research timelines
  - Pre-publication review vs. post-publication review
    - Data users commonly find data errors that can only be found through intensive analysis
    - Repositories must have way to receive, evaluate, and respond to user-discovered errors
    - Reviewers from outside of a project need more time and often assistance from project members
  - There is a growing demand for real-time data. Quality control timelines have to balance researchers' desire to access and use the data, and the needs to run quality assurance processes.
- Peer review might be best conceptualized as review of the data collection, assessment, metadata, and archiving processes vs. review of the data themselves.

NCAR Data Repository Archiving Workflows





Repository workflows have data quality assessment processes integrated throughout the data ingest and archiving workflows

Repository Data Quality Control Processes

- Flag questionable or faulty data by creating new metadata. Always maintain original data.
- Provide mechanisms for feedback loops between users, the archive, and data providers.
  - Sometimes data quality problems are found by external users. External users are excellent data reviewers.
  - People who are knowledgeable about the project are more likely to find actual problems with the model and data, whereas users are likely to find smaller scale anomalies that may or may not be errors. Shared evaluation is sometimes required.
  - Maintaining records of what data sets were downloaded, and by whom, allows the data archive to inform data users of data updates and/or new versions.
- Compare data with other data, or model runs with other data/models
- Develop standard sets of diagnostics tools and methods over time
- Deploy reliable remote back-up and integrity check mechanisms
- Technical review vs. scientific peer review
  - For model data, use control runs to evaluate the model functionality
  - For observational field campaign data, keep "housekeeping parameters", like battery life, ambient or equipment temperature ranges, etc., to evaluate the equipment functionality
  - Is the data set well constructed, e.g. following conventions and standards?

This work conducted as part of the Peer REview for Publication & Accreditation of Research data in the Earth sciences project (PREPARDE)
http://proj.badc.rl.ac.uk/preparde