Data Archiving and Access Requirements Working Group (DAARWG)

Report to the
Science Advisory Board
National Oceanic and Atmospheric Administration

23 August, 2007

A reminder: WG Terms of Reference

- Provide scientific advice and broad direction to NOAA regarding the wide range of data, information, and products that NOAA should archive and how NOAA can best provide access to this information.
- The Data Archiving and Access Requirements (DAAR) Working Group will evaluate data archiving and access requirements from all of NOAA's observing systems and computational models, as well as non-NOAA information.

DAARWG membership

Roberta Balstad, Columbia University
David Blaskovich, International Business Machines (NOAA SAB Liaison)
Peter Cornillon, University of Rhode Island
Daphne G. Fautin, University of Kansas
Sara Graves, University of Alabama in Huntsville
Gary Jeffress, Texas A & M University
Phil Jones, University of East Anglia
Stephen Meacham, National Science Foundation
Anne Miglarese, Earth Data International, Inc.
Michael R. Mott, International Business Machines
Aaron J. Ridley, University of Michigan
Sami Saarinen, European Centre for Medium-Range Weather Forecasts
Roger Wakimoto, National Center for Atmospheric Research
Ferris Webster, University of Delaware
Brice A. Wielicki, NASA Langley Research Center

Two DAARWG archive issues

1. Multiple versions of datasets
   - NOAA archives data and data products for many reasons.
   - A policy on retention of multiple versions is needed.
2. CLASS & the NOAA archive
   - NOAA will benefit by clarifying the roles, responsibilities, and requirements of the participating elements of the archive in which CLASS will be a major element.

Issue 1: Archiving dataset versions

- Data are required by law to be archived, for activities that are central to NOAA's mission:
  - weather
  - climate
  - oceans
  - solid earth
  - space

What datasets are archived?

- Data and products are archived if they require extensive processing to reproduce.
  - Examples are:
    - output from computer models
    - satellite-derived products
    - radar
- Non-NOAA data that support NOAA's mission are archived.

What datasets are archived?

- Data are archived for regulatory purposes.
  - Examples are:
    - fisheries data
    - climate normals
- Data that are used for producing scientific assessments are archived.
  - Examples are:
    - *State of the Climate* reports
    - Data used by the Intergovernmental Panel on Climate Change (IPCC)

Multiple dataset versions

- Different versions of datasets proliferate
  - Datasets are modified as additional data become available.
  - Datasets are adjusted based on scientific discoveries.
  - Datasets are improved as the data are used and are more thoroughly scrutinized.
- How many dataset versions should be kept?

Dataset retention

- NOAA archives are expected to grow exponentially.
- Two major costs are associated with this growth:
  - storage
  - data stewardship
- Both costs can be reduced if archives can be reduced by minimizing the number of dataset versions.

What to keep?

- Which versions of a dataset should be kept, and which should be abandoned?
- Even if more than one version of a dataset is saved, sometimes NOAA finds that a needed version is no longer available.
- An example in responding to a recent Freedom of Information Act (FOIA) request follows.

FOIA request

- As shown on the next slide, NOAA received a 2007 request from an international researcher.
  - –He asked for the identification of the stations & the data as used in a classic 1990 paper on the climate effects of urbanization.
- The request couldn't be fulfilled as requested.

FOIA request

Re: Jones, et al., 1990. Assessment of the urbanization effects in time series of surface air temperatures over land. *Nature* 347, 169-172.

Dear sirs,
I request the following information in connection with this article co-authored by NCDC scientist Thomas Karl.

A)The identification of the stations used in the … Jones et al. 1990 networks

B)Identification of the stations used in the gridded network for comparison

C)The data as used by Jones et al for each of the above stations

This study continues to be relied upon and cited, including by the Intergovernmental Panel on Climate Change.

Thank you for your consideration.

FOIA request

- As it turned out, the data as used in 1990 were no longer available in the same form.
- After nearly 20 years, the original dataset had been augmented and modified.
  - Should NOAA have had a retention policy so that the dataset on which this paper was based would be still available 20 years later?
  - Who decides what will be important in 20 years?
  - Should there be a time limit on archiving multiple versions of datasets?

DAARWG conclusion

- NOAA should develop a retention policy for multiple versions of datasets.
- The policy should take account of
  - Benefits vs. costs
  - User needs
  - NOAA mission requirements
  - Legal and regulatory constraints
  - The NRC recommendations on archiving
- A useful first step would be a workshop involving users and NOAA data people.

Issue 2:
CLASS and the NOAA archive

- The Comprehensive Large-Array data Stewardship System (CLASS) is now being developed as the storage element in a NOAA archive.
- However, at the moment, the various elements in NOAA do not have a shared concept of what a NOAA archive would involve.

The DAARWG found…

…that because the scope of CLASS has evolved over time, there is widespread misunderstanding within NOAA of the CLASS role and purpose.

…a lack of understanding of the roles and responsibilities in creating and using a NOAA archive that is handicapping development.

…mistrust of CLASS by some NOAA elements whose active participation in the archive will be essential.

The DAARWG believes that…

…a NOAA-wide archive should incorporate data-originating and data-managing elements throughout NOAA.

…An archive will allow better use of data and information to meet NOAA mission objectives.

…the interacting roles of CLASS, data centers, centers of data, and legacy systems need to be clarified since they will be essential to the effectiveness of a NOAA archive.

…it's time for a fresh look at the data-system architecture that best meets NOAA's needs.

DAARWG conclusion

- NOAA should define its archive requirements
  - Based on those requirements, the roles and responsibilities of CLASS, data centers, centers of data, and involved legacy systems should be clarified.
  - A NOAA archive architecture group should be established to analyze and define archive elements and to track progress in achieving them.

DAARWG proposes that the SAB:

1. Recommend that NOAA develop a retention policy for multiple versions of datasets.
   - A useful first step would be a workshop involving users and NOAA data people.
2. Recommend that NOAA define its archive requirements
   - In order to clarify the roles and responsibilities of CLASS, data centers, centers of data, and involved legacy systems.
   - This process would be aided by creating a NOAA archive architecture group.

Looking ahead

- The National Research Council report on archiving and access should be released soon
  - the DAARWG awaits the report before looking at the following issues:
- Coping with data volume: data stewardship
  - Earth-science data records need scientific data stewardship
  - There are potential problems with archiving non-NOAA datasets

Looking ahead (cont.)

- Unified interdisciplinary NOAA standards and protocols
  - Global Earth Observation, Integrated Data Environment (GEO-IDE)
  - What architecture will best serve NOAA?
- Data Centers and Centers of Data
  - DAARWG will look at the role of Centers of Data in archiving & access