

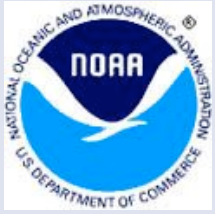


NOAA's Response to NRC Interim Report on Archiving Data

December 8, 2006

Zdenka Willis

National Oceanographic Data Center



Outline



-
- NRC Report - Preliminary Principles and Guidelines
 - NOAA's Response & Challenges
 - Questions/Comments



Preliminary Report Principles

The environmental and geospatial data collected by NOAA and its partners, including model output, are an invaluable resource that should be archived and made accessible in a form that allows researchers and educators to conduct analyses and generate products necessary to accurately describe the Earth System

- Response:
 - Data can't be precisely duplicated through re-collection; inherently valuable. Re-collection is expensive or impossible. Example: re-collection of data contained in WOD01 is ~ \$6.8B.
 - Legislative & Executive mandates dictate some archive and records retention requirements
 - Data handling standards are critical to ensure data integrity, integration, interoperability, and continuity, and to maximize accessibility.
- Challenges:
 - Data are collected from a wide variety of industries, people, places and platforms. Data collection methods & formats highly variable.
 - Difficult to enforce legislative mandates requiring data and records retention at all project/program scales; limited authority to enforce data delivery (including data products) for archive.
 - Current data ingest backlog with predicted growth in both data volume and user population.



Preliminary Report Principles

The decision to archive or continue to archive data or model output should be driven by its current or future value to society. The decision will need to take into account the cost to archive versus the cost to regenerate, as well as the costs of providing access to the data

- Response:
 - Current value derived from NOAA Program Observation Requirements and prioritization schema that ties observing parameters to societal needs for data and information.
 - NARA's Records Disposition Schedule provides for retaining original data for an additional 75 years after the decision to no longer actively archive data set.
 - No formal mechanism to analyze costs and benefits of regeneration versus long-term archive and access.
 - NOAA's National Climatic Data Center (**NCDC**) serves as NARA Agency Records Center. NNDCs make decisions regarding administration and disposition of NOAA data in accordance with NARA standards.
- Challenges:
 - Future value difficult to calculate and subject to debate.
 - Preserving potential future value and access requires consistent funding; fiscal constraints result in difficult trades.
 - There is no sole decision making authority, no consensus among repositories, and no formal decision process to determine what to archive or continue to archive. E.g., NCDC has extensive criteria; NGDC & NODC have general criteria and decisions are case by case.



Preliminary Report Principles

Funding for Earth system measurements should include sufficient resources to archive and provide ready and easy access to these data for extended periods of time. In particular, at the outset of undertaking an activity which will generate data or model output, end-to-end data management needs to be planned and budgeted.

- Response:
 - Publicly funded collection regimens should include cost plans for end-to-end data management needs. NOAA Administrative Order 212-15 addresses the process of data management planning but does not address the issue of funding.
 - Moving to end-to-end data management architecture (e.g., CLASS and GEO IDE).
 - NOAA systems should be included within the Planning, Programming, Budgeting and Execution System (PPBES).
- Challenges:
 - Expense of archive.
 - Keeping funding requests in-tact through the budget formulation cycle.



Preliminary Report Principles

All data that are well documented, are of known quality, and represent systematic collections or characterizations of the state of the environment should be archived in their most primitive useful form.

- Response:
 - Original parameter values from in-situ observing systems retained in archive. Suspect values discovered via quality control are flagged and corrections documented and included in archive.
 - Satellite derived data are archived in primitive form (level 1b) with raw data and calibration coefficients appended.
- Challenges:
 - Different QC methodologies; determination of quality.
 - Labor intensive to do QC – where does this occur?
 - Biological survey data is continually being refined making permanent archival difficult



Preliminary Report Principles

The decisions not to archive data permanently should only occur when the original and predicted purpose of the data has been satisfied, or when the cost of storing the data exceeds the cost of regeneration, and should be made in collaboration with the appropriate user communities.

- Response:
 - NOAA’s Records Disposition Schedule provides for retaining original data for an additional 75 years after the decision to no longer actively archive data set.
 - No current mechanism to analyze costs and benefits of regeneration versus long-term archive and access.
 - No single formal process to collaborate with appropriate user communities on decisions to not permanently archive data. Data submission agreements may result in periodic review of archival needs in context of mission focus and resource constraints.
 - NODC takes “safe” approach - preserves all data taken from credible sources.
- Challenges:
 - How to identify “appropriate user communities” and develop unbiased process for engagement?
 - No standard evaluation criteria to inform user community decisions to archive or delete.
 - Proclivity of scientists to want to keep all data.



Preliminary Report Principles



Metadata that completely document and describe archived data should be created and preserved to ensure the enhancement of knowledge for scientific and societal benefit.

- Response:
 - NOAA follows FGDC metadata standard.
 - NGDC data managers create, store and distribute metadata via the NOAA Metadata Management Repository (NMMR)
 - NODC data managers use an Archive Management System which includes an Accession Tracking Data Base that contains metadata. This system follows the Open Archive Information System (OAIS) Reference model and is FGDC compliant.
 - General philosophy: The more metadata the better as long as it is accurate and understandable.
 - Data centers do have data agreements but vary across NNDC.
- Challenges:
 - Data managers complete FGDC mandatory fields (ID & Metadata reference Information). Mandatory fields do not contain enough information to independently understand datasets.
 - No current agreement regarding applicability of all seven FGDC sections to NOAA data.



Preliminary Report Principles

NOAA's archival process should be designed to allow the integrated exploitation of data from multiple sources to answer environmental questions and support the total life-cycle aspects of individual data sets. This could potentially be accomplished through a distributed but federated archival system facilitated via a single user portal.

- Response:
 - NOAA is moving the Comprehensive Large Array-data Stewardship System (CLASS) to provide for the IT architecture to store and access NOAA data.
 - Data management still resident within the NNDC & other NOAA centers.
 - Database examples across NNDC: World Ocean Database; Pathfinder, Night Lights – individual products that support niche areas.
- Challenges:
 - The Ocean Action Plans (JSOST) and Ecosystem Task Force report call for high resolution coastal/near coast database which includes physical, statistical, demographic data in a single geo-registered format.
 - Putting databases online in GIS format without QC is not the same as labor intensive QC/QA and integrating data. NNDC will need to partner across NOAA with other centers of data to accomplish.



Preliminary Report Principles

Broad community representation is essential to establish the process whereby data proposed for archiving can be evaluated and prioritized in terms of scientific and societal benefits.

- Response:
 - NOAA uses an informal but structured process to engage community stakeholders in the evaluation and prioritization of data for potential scientific and societal benefits.
 - Community representation in NOAA's data handling process is achieved through data agreements, workshops, and user conferences. Currently, no clear link to systematic evaluation and prioritization of benefits.
 - NOAA established a Science Advisory Board (SAB) Data Archiving and Access Working Group. This SAB will help prioritize specific data sets and products and recommend access and archive strategies.
- Challenges:
 - Continued engagement with users.
 - Data policy with enforceable mechanisms to ensure access to federally funded data that is put in the archives.



Preliminary Report Principles



Scientific data stewardship should be applied to all archived information so it is preserved, continually accessible, and can be supplemented with additional data as discoveries build understanding and knowledge.

- Response:
 - Agree
- Challenges:
 - Costly



Questions/Comments