Guidelines for Archiving Data in the NARSTO Permanent Data Archive

May 2, 2006

Executive Summary

NARSTO archives air chemistry, meteorological, and related atmospheric data and metadata in a publicly accessible Permanent Data Archive (PDA). This document outlines how data are selected for archiving; identifies ways that Projects can foster archiving; lists items to consider when preparing data for archiving; and describes the archiving process.

The Project makes the primary decision about which data and metadata (information about the data) should be archived, and supports compiling data into one of a number of widely accepted data formats. The scope of this brief document is primarily limited to data that are reported in tabular data formats. Guidance is provided about what information to include and how to qualify and explain the data. Verification and documentation of the data and their format is done by the Project's data providers or by the NARSTO Quality Systems Science Center, depending on the format chosen. Preparation of the data for archiving emphasizes easy usability of the data by a variety of users, based only on the archived information.

The archiving process is described for a number of scenarios, and links are provided to additional sources of information and assistance.

Introduction

The NARSTO Quality Systems Science Center (QSSC) together with NARSTO sponsors provides a publicly accessible Permanent Data Archive (PDA) at the NASA Langley Atmospheric Sciences Data Center. Over the past 10 years the QSSC has developed extensive data management and quality assurance guidelines for use by all NARSTO members. NARSTO encourages Projects to utilize the NARSTO Data Exchange Standard (DES) format for ease of data evaluation, analyses, and sharing. In addition, we recognize that there are other widely accepted data exchange formats suitable for archiving. We provide the following general guidance for Projects planning to archive their data at the NARSTO PDA.

Twenty Year Test for Data and Documentation

NARSTO encourages scientists to document their data at a level sufficient to satisfy the well-known "20-year test". That is, someone 20 years from now, not familiar with the data or how they were obtained, should be able to find data of interest and then fully understand and use the data solely with the aid of the documentation archived with the data.

(Source: National Research Council, Committee on Geophysical Data, Solving the Global Change Puzzle, A U.S. Strategy for Managing Data and Information, National Academy Press, Washington, D.C., 1991.)

Characteristics of Project Data Management That Will Result in Successful Data Archiving

- Data Coordination roles and responsibilities clearly defined for Project
- Decision by Project about what data and products should be archived and the schedule for archiving. Project should consider archiving the following data products:
    - > Measurement data (point and gridded)
    - > Remote sensing products (data and images)
    - > Model products (source code, input data, and output data)
    - > Model comparison results
    - > Emission inventories and Transportation data
    - > Value added products, including
        - + Data that were used in publications
        - + Data manipulated to a common time base
- Funding sufficient for planning and implementing Project data management until archiving is complete

Characteristics of Good Archive Data File Formats

- Well defined and documented format
- Has established structures that are usable or adaptable for several data types
- Open-standard accessible format: ASCII file, space, tab, or comma-delimited (.csv) format
- Can be created and especially read without highly specialized software
- Consistent / standardized names, units, and metadata values based, when appropriate, on national or international standards
- While the scope of this guidance is limited to data that are reported in tabular data formats, these same principles generally apply to other data types and formats.

Characteristics of Good Archive Data Documentation

The PDA’s documentation provides the framework for subsequent data users to find, understand, and then appropriately use the data you have kindly agreed to share.

For efficiency of organization and retrieval, project data files are usually grouped by common themes into data sets for archiving (e.g., by measurement platform or site, data type, investigator, or the project as a whole).

- An informative data set title, general description, and metadata are added, which are the links that enable potential data users to find data of interest.
- Instructions for users to cite the data set and acknowledge the provider are always included in the archive documentation.

The data file metadata and supplemental data set documentation help users to understand and then to use the data.

- Individual data files may often contain sufficient metadata, comments, or notes to be self-documenting for both format and content.
- Data may be supported by Quality Assurance Project Plans, Final Project QA Reports, SOPs, publications, readme files, etc., that are identified as overall data set companion documents.

The metadata content of data file(s), plus data set companion documentation, plus the PDA's data set archive documentation, should meet the National Research Council's "20 year test". Consult with the QSSC for assistance in creating data sets and compiling and formatting archive documentation.

Characteristics of Good Archive Data Reporting

- Values considered valid should be reported, although they may be qualified (flagged) for nonstandard sampling conditions, failing some statistical or QC criteria, etc.
- Values should be reported as measured, not "interpolated" or repeated to form a common time basis. The derived measures can be archived as value-added products.
- Below / above-detection limit values should be reported as measured and qualified (flagged), rather than substituting the detection limit, zero, or another value.
- User's prefer data files covering longer time periods, rather than single days, and prefer multiple related measurements in one file, rather than just one variable.

NARSTO Acceptance of Data for Archiving

Quality Level

Minimum of **Level 1,** which is a complete data set of specified quality that consists of research products subjected to quality assurance and quality control checks and data management procedures.

Metadata and Documentation

The QSSC would expect to see, documented either in the data file or data set companion documents, the following information:

- Project identification
- Data provider contact information
- Site / measurement platform attributes
- Sampling and measurement methods
- Compliance with quality assurance project plans verified
- Quality level

- Variable names / column headings
- Units / formats specified
- Dates and times
- Data qualification
    - > Preferred: data qualification flag assigned to each data value
    - > Alternative: statement of data quality is provided indicating overall quality of data
- Missing value and less/greater than detection limit value codes specified.
- Detection limits defined and specified. If unavailable or unknown this is documented.
- Uncertainty defined and specified. If unavailable or unknown this is documented.
- Explanation of any zero or negative values that may occur in the file.

Additional guidance can be provided by the QSSC.

Data Files

The QSSC would expect to see documentation of the verification steps that the data provider performed on the data files before submitting them for archiving.

Data Exchange Standard (DES) format files:

- The QSSC runs an automated "read and verification" QA program that checks the data file format and selected content vs. metadata reference tables. Summary statistics are calculated for each numeric variable. Time series plots are produced for each variable reported and altitude profile plots can be produced for vertical measurements.
- Any inconsistencies are reported to the Project data coordinator for possible updating of the data file.
- The data files are checked and updated in an iterative process.

-. Note that Projects can obtain the SAS code for this QA program.

Data files in other exchange formats:

- Project data coordinators would QA the data files and provide documentation of the verification task; either documented in the data file or described in the documentation.
- Verification could include automated checks of data file format and content.
- Verification could include calculating summary statistics for non-missing values of each numeric variable in the file. The statistics could include minimum, maximum, mean, standard deviation, n, number of missing values, and total number of records.
- Verification could include producing a time series plot for each variable reported and altitude profile plots for vertical measurements. Experience has shown that plots are valuable for identifying any remaining outliers and data gaps.

Summary Table of Archiving Process

The attached table shows general responsibilities of participants in the data archiving process for selected data file formats.

For additional information, contact Les Hook
[hookla@ornl.gov] and the NARSTO QSSC web site
[http://cdiac.ornl.gov/programs/NARSTO/]

Responsibilities of Projects and the NARSTO QSSC for Archiving Data in Various File Formats

	Data Management Activities Leading to Archiving				
Data File Format	**Data File, Metadata, and Documentation Preparation**	**Final Data File QA Evaluation**	**Review Project QA Verification**	**Compile Data Set Archive Documentation and Send All Files to Archive**	**Archive Data Set**
Data Exchange Standard (DES)[1]	**Project PI and Data Coordinator**	**QSSC (automated)**	NA	**QSSC**	**Atmospheric Sciences Data Center, NASA Langley Research Center**[3]
NASA Ames / ICARTT[2]	**Project PI and Data Coordinator**	***Project Data Coordinator***	**QSSC**	**QSSC**	
Other Data File Formats	**Project PI and Data Coordinator**	***Project Data Coordinator***	**QSSC**	**QSSC**	

[1] Data Exchange Standard Template, NARSTO QSSC Web site [http://cdiac.ornl.gov/programs/NARSTO/]

[2] ICARTT Data Management Implementation Plan, 1 December 2004
[http://www.al.noaa.gov/ICARTT/StudyCoordination/ICARTTDocs/DataManagement_plan.pdf]

[3] NARSTO Permanent Data Archive [http://eosweb.larc.nasa.gov/PRODOCS/narsto/table_narsto.htm]