ARCHIVING NARSTO DATA SETS
February 6, 2002

Please check with the NARSTO QSSC Web site
(http://cdiac.esd.ornl.gov/programs/NARSTO)
to be sure you are using the most recent version of this document

Overview of the process of archiving NARSTO data

The Data Provider prepares a data set (collection of files) following the guidance below, and places the files on the NARSTO ftp site. The Provider then sends an e-mail message to Sig Christensen (christensen1@ornl.gov) with cc to Les Hook (hookla@ornl.gov) releasing the data set to the NARSTO Quality Systems Science Center (QSSC), and also identifying the data files' format and the format(s) of any ancillary or companion files. The NARSTO QSSC uses ftp or other procedures to get the files, and acknowledges receipt (assigning a Dataset ID). The QSSC then processes the data set (differently depending on the format), interacts as needed to resolve issues with the Provider, and transmits the data set to the NARSTO Permanent Data Archive.

Nomenclature

A **companion file** is a file named as such in one or more Data Exchange Standard files. It provides supporting information. Alternatively, a **companion file** is supporting documentation for data sets with data not in the Data Exchange Standard. This type of companion file will document the data file format, and provide statistics for variables contained in this format.

A **data set** is a collection of files to be sent together to the NARSTO Permanent Data Archive. These will be stored together under a single **data set name**.

A **NARSTO standard data file** is a file that follows the Data Exchange Standard, and which contains one main data table.

Data set creation (for NARSTO Archiving)

The Data Provider identifies a group of files to constitute a data set. A project, such as SOS Nashville 1999, will submit data in several to many data sets. A data set may contain as few as one data file, but ideally would contain multiple files, perhaps even hundreds of files. These files should be related in some reasonable way, such as: they contain data of a particular type, such as meteorology data or air chemistry data; or they were collected by the same Principal Investigator; or they were collected at the same location; or they were aircraft data, perhaps from the same aircraft.

The Data Provider (in coordination with the QSSC) may enter metadata for the data set into the Data and Information Sharing Tool (DIST).

The first part of the data set name is the study acronym. This is controlled; it has been determined by the Project Manager in consultation with the QSSC. If it was not provided to you, ask for it. The study acronym must also be used as the entry for the ***STUDY OR NETWORK ACRONYM** Key Phrase in the Data Exchange Standard, and the name of each file within the data set must begin with the study acronym (see examples below).

Data file creation (for NARSTO Archiving)

Data files are expected to be in the NARSTO Data Exchange Standard (DES). If another format is to be used, this needs to be discussed with the QSSC, preferably ahead of time. For other data formats, the provider will need to document the format, will probably need to interact with Permanent Data Archive staff, and may need to calculate statistics for the variables contained in the data set.

All of the data files in a data set must be in the same format (e.g., Data Exchange Standard). Please refer to the DES template on the QSSC web site. Each file name (whether data or companion) must be a two-part name (filename.ext). Filenames should use only alphanumeric characters, dash, plus, minus, or underscore characters. Note that data file names are visible/selectable on the Langley DAAC order system.

Data File Name Syntax

General Guidance

Following this syntax guidance will result in file names that will sort by study and contents and will contain enough specific information to enable data providers to keep files distinct and to enable data users to find files of interest.

Filenames must use only upper-case alphanumeric characters, dash, plus, minus, or underscore characters. They must not be longer than 57 characters total (including the suffix, e.g. ".csv").

[*STUDY OR NETWORK ACRONYM]_[*FILE CONTENTS DESCRIPTION-SHORT]
_[unique data file descriptor]_V1.csv

[15 chars max]_[15 chars max]_[22+ chars]_V1.csv (57 chars max)

Example: [EPA_SS_ATLANTA]_[PILS_IC]_[RESULTS_1999]_V1.csv

[ ] included for example only.
* STUDY..... and *FILE..... entries are the same as entered in the DES template used to create the formatted data file.

Examples of [unique data file descriptor] to fit in 22+ characters: event, place, P.I., averaging time, gridded spatial resolution, date range, start date, start date and days of data in file (e.g., yyyymmdd_nnn), quarter reported (e.g., Q1, Q1-Q2) or etc.

Note that date, time, latitude, and longitude ranges are shown for each data file on the Permanent Data Archive (Langley DAAC) data ordering page.

The last part of each filename before the extension will be:

Another underscore, a “V”, and followed by a version number-- “_V1" (initially “one”).

The extensions will follow prevailing conventions to identify the format of the file. NARSTO standard files (i.e., in the Data Exchange Standard) must have “csv” as the extension and no other format of files may be included in the same data set. The extensions for any companion files must be different from the extensions of the data files.

These are examples of possible filenames constituting part of a data set:

SOS99NASH_SURF_MET_DATA_CORNELIA_FORT_V1.csv
SOS99NASH_SURF_MET_DATA_OPRYLAND_V1.csv
SOS99NASH_SURF_CHEM_DATA_CORNELIA_FORT_V2.csv

SOS99NASH_SURF_CHEM_DATA_SOP_17_V1.pdf
SOS99NASH_SURF_MET_DATA_SOP_19_V1.pdf
SOS99NASH_SURF_MET_STUDY_DESCRIPTION_V1.txt

(The .pdf files and the .txt file are ancillary files containing supplementary metadata; each would be named in one or more of the Data Exchange Standard files, using ***COMPANION FILE NAME/FORMAT/VERSION** Key Phrases.)

Project Specific Guidance

Through the coordination efforts of your Project Data Manager, and in consultation with the QSSC, a more project-specific file name syntax may be specified. The file names must sort by study and contain enough specific information to enable data providers to keep files distinct and to enable data users to find files of interest.

Several of the general syntax rules apply.

Filenames must use only upper-case alphanumeric characters, dash, plus, minus, or underscore characters. They must not be longer than 57 characters total (including the suffix, e.g. ".csv").

[study or network acronym/identifier]_[unique data file descriptor]_V1.csv

[15 chars max]_[unique data file descriptor]_V1.csv (57 chars max)
The study or network acronym/identifier may be the *STUDY ACRONYM as entered in the DES template or a shorter identifier that will be consistently applied across the study by all data providers.

Example: [PAC2001]_[SLPK_JRB_MASS_TEOM_200108D60]_V1.csv

[ ] included for example only.

The last part of each filename before the extension will be:

Another underscore, a "V", and followed by a version number-- "_V1", (initially "one").

The extensions will follow prevailing conventions to identify the format of the file. NARSTO standard files (i.e., in the Data Exchange Standard) must have "csv" as the extension and no other format of files may be included in the same data set.

The extensions for any companion files must be different from the extensions of the data files.

In addition, to enable data users to interpret the project-specific file names, the data provider must prepare a "readme" file explaining the file naming convention.

Data set delivery to the NARSTO QSSC

Data sets must be delivered via ftp (File Transfer Protocol). (Attaching data files to e-mail messages can be problematic.) The person transmitting the data set, logs in to their project's area on the NARSTO ftp site, changes to the data_staging subdirectory, and creates there a subdirectory reflecting the data set name (e.g., surf_met+chem). He or she then ftps the data files and the ancillary files to this subdirectory. Companion files (i.e., required format documentation and summary statistics when the data sets are not in the Data Exchange Standard), are to be put into a subdirectory of the data set's directory, named "companion_files". Once staged, send an e-mail message to Sig Christensen (christensen1@ornl.gov), with cc to Les Hook (hookla@ornl.gov), releasing the data set to the NARSTO Quality Systems Science Center (QSSC), and also identifying the data files' format and the format(s) of any ancillary or companion files. In the e-mail, also name the subdirectory(ies) you created for the files, and include a list of the files by subdirectory. The QSSC will send an acknowledgment e-mail, including a Dataset ID. Keep note of this data set identifier in case file revisions are later needed.

File Resubmissions

Any resubmissions of a data file(s) during initial processing (during QA checks prior to archiving), will need to reference (in the transmittal e-mail) the Dataset ID that the QSSC provided in its acknowledgment. The revised file name and version number are the same as originally used.

Any resubmissions of a data file(s) after archiving will again need to reference (in the transmittal e-mail) the Dataset ID that the QSSC provided in its acknowledgment. The revised files will need to include the same file name but the version number will be incremented.

Permanent Data Archive data set names

Data sets in the PDA have names the first parts of which are controlled. The "data set name" from DIST is used, with NARSTO placed in front.

Data set names therefore always begin with "NARSTO". The second part of the data set name is a standardized reference to the project (*STUDY OR NETWORK ACRONYM), and sometimes to a subproject (e.g., CE-1996; SOS99NASH; EPA_SS_HOUSTON). Following this, the data provider adds a description of the collection of files in the data set. An example:

NARSTO SOS99NASH Surface meteorology and chemistry data from Cornelia Fort Airpark

Bear in mind that some PDA screens currently truncate the displayed data set name to 40 characters. While it is always possible to view the full name, it may help users find data if the first part of the data provider's description both contains the key descriptive material and is constructed to make the description distinctive.

Permanent Data Archive file names

Same as **Data File Name Syntax**.