jmoore@ucar.edu 12/12/01

Dataset documentation and format Guidelines

Data format guidelines

An inherent flexibility of the JOSS data management system permits data in all different formats to be submitted, stored and retrieved from CODIAC. JOSS is prepared to work with the participants to bring their data to the archive and make sure it is presented, with proper documentation, for exchange with other project participants. In anticipation of receiving many data sets from the field sites in ASCII format we are providing guidelines below that will aid in the submission, integration and retrieval of these data. JOSS will work with any participants submitting other formats including NetCDF, AREA, HDF and GRIB to assure access and retrieval capabilities within CODIAC.

The following ASCII data format guidelines are intended to help standardize the information provided with any data archived for the project. These guidelines are based on JOSS experience in handling thousands of different data files of differing formats. Specific suggestions are provided for naming a data file as well as information and layout of the header records and data records contained in each file. This information is important when data are shared with other project participants to minimize confusion and aid in the analysis. An example of the layout of an ASCII file using the guidelines is provided below. Keep in mind that it is not mandatory that the data be received in this format. However, if the project participants are willing to implement the data format guidelines described below, there are some improved capabilities for integration, extraction, compositing and display via CODIAC that are available.

NAMING CONVENTION

A) All data files should be uniquely named. For example, it is very helpful if date can be included in any image file name so that the file can be easily time registered. Also include an extension indicating the type of file:

i.e. .gif = GIF image format file
.jpg = jpg image format file
.txt = Text or ASCII format file
.cdf = NetCDF format file
.tar = archival format

If compressed, the file name should have an additional extension indicating the type of compression (i.e. .gz, .z, etc.).

B) For Text (ASCII) files, the records should consist of both header records and data records. The header records at a minimum should consist of:

ASCII DATA FORMAT HEADER RECORDS SPECIFICATIONS

Standard header records should precede the data records within the file itself. The header records should contain the following information:

PI/DATA CONTACT =	Text [PI and data contact name(s) and affiliation(s)]
DATA COVERAGE =	Start/Stop time of continuous data or sampling interval (Use data/time format described below)
PLATFORM/SITE =	Text [e.g. C130, CITATION, BROWN, SK, KCO, DG, etc.]
INSTRUMENT =	Text [instrument name]
COORDINATES =	Fixed site coordinates (decimal degrees) or "mobile" (Text)
DATA VERSION =	Alphanumeric [unique ID (i.e. revision date, PRELIMNARY or FINAL]
REMARKS =	Text [PI remarks that aid in understanding data file structure and contents. Items such as file type, how missing and/or bad data are denoted or any other information helpful to users of this data]

Missing Value indicator - Text or integer [value used for data for missing information] (e.g. -99 or 999.99, etc)

Below Measurement Threshold - Text or Integer [Value used to signify reading below instrument detection threshold] (e.g. <0.00005)

Above Measurement Threshold - Text or Integer [Value used to signify reading at or above instrument saturation]

NOTE This type of header information cannot be contained within GIF and Postscript files. They will need to be submitted with attached files or separate documentation containing this information.

DATA RECORDS SPECIFICATIONS:

1) First data record consists of parameters identifying each column. Multiple parameter names should be shown as one word (e.g. thaw_depth or Leaf_area_index)
2) Second data record consists of respective parameter units. Multiple unit names should be shown as one word (e.g. crystals_per_liter)
3) Third data record begins actual data and consists of a date/time column followed by position coordinates (if mobile) and subsequent observations at that time and position.

! Date/time must be in UTC and recommended format is:
YYYYMMDDHHmmss.ss
where: YYYY= Year
MM = Month (00-12)
DD = Day (01-31)
HH = Hour (00-23)
mm = Minute (00-59)
ss = Second (00-59)
.ss = Decimal Second (unlimited resolution based on sampling frequency)

! For every mobile platform data set, position coordinates (i.e. latitude, longitude) should be expressed in decimal degrees for each data point. Altitude or elevation are given in appropriate metric units. This may be done by: (a) providing date/time of collection with position coordinates in each data record; or (b) providing date/time of collection for each data point in the submitted file, with an associated file containing date/time and location either from the platform navigation database or GPS file.

Latitude - Northern hemisphere expressed as positive or "N" and
Southern hemisphere expressed as negative or "S".
Longitude - 0-360^{o} moving east from Greenwich; west longitude goes from
180^{o} to 360^{o}; or
Eastern hemisphere expressed as positive or "E" and
Western hemisphere expressed as negative or "W".

NOTE – Position information in other grid conventions is acceptable but a conversion to latitude/longitude should be provided where practical.

NOTE - Having a common date/time stamp and common position coordinates in each data record will permit the ability to extract data and integrate multiple data records from different data sets. If two times are provided (e.g. UTC and local), they should be put at the beginning of each record.

- Preferred format for ASCII data files is space, comma or tab delimited columns, with a UTC date/time stamp at the beginning of each data record. If the data in the file are comma delimited, decimal places must be periods, not commas.

- All data files must contain variable names and units of measurements as column headings (if applicable).

- If, for some reason, the PI cannot provide the date/time in the format

shown above, it is important that the time be given in UTC. If local time is also supplied, a conversion to UTC must be provided. In addition to UTC and/or local time, other date/time formats (e.g. decimal days) can be used but must be fully documented.

- The internal format structure of the file should remain constant after the first data record to ensure continuity and permit plotting and graphing.

- Only COMPLETE replacement or updated data/metadata files can be accepted.

SAMPLE DATA SET (ASCII FORMAT):

The following is an example of an ASCII format data set in which the header precedes the reported data, and the data is organized in columns separated by spaces. Each column is identified by parameter and each parameter's units of measure are listed in the respective column. Also each row has a date/time of observation reported in Universal Time Coordinated (UTC) along with position coordinates. This data set organization is ideal for plotting and integration of various data sets. This data set format should be used whenever possible and could be easily produced automatically from a spread sheet computer program.

PI/DATA CONTACT= Doe, John (U of Hawaii)/ Doe, Jane (NCAR)
DATA COVERAGE = START: 19990821133500; STOP: 19990821135500 UTC
PLATFORM/SITE = C130
INSTRUMENT = C-130 External Sampler Data
LOCATION = mobile
DATA VERSION = 1.0 (10 March 1999), PRELIMINARY
REMARKS = National Center for Atmospheric Research, INDOEX
REMARKS = ppm values are mole fraction
REMARKS= nM/m3 at 25c and 101.3 kPa; DMS and NH4 in Parts per million (PPM)
REMARKS = Missing data = 99.9; Bad data = 88.8
REMARKS = Data point Date/Time provided in UTC

DATE/TIME	LAT	LONG	SAMPLE	NO2	CO	DMS	NH4
UTC	Deg	Deg	NUMBER	nM/m3	ppm	ppm	ppm
19990821133500.00	-43.087	263.116	E1.160.1	1000.65	200.67	345.98	2342.980
19990821133510.00	-43.090	263.120	E1.160.2	1003.45	200.60	349.76	2353.345

.
.
.
etc.

Data Documentation Guidelines

The documentation (i.e. the "Readme" file) that accompanies each project data set is as important as the data itself. This information permits collaborators and other analysts to become aware of the data and to understand any limitations or special characteristics of data that may impact its use elsewhere. The data set documentation should accompany all data set submissions and contain the information listed in the outline below. While it will not be appropriate for each and every dataset to have information in each documentation category, the following outline (and content) should be adhered to as closely as possible to make the documentation consistent across all data sets. It is also recommended that a documentation file submission accompany for each preliminary and final data set.

---TITLE: This should match the data set name

---AUTHOR(S):

-Name(s) of PI and all co-PIs
-Complete mailing address, telephone/facsimile Nos., web pages and E-mail address of PI
-Similar contact information for data questions (if different than above)

---FUNDING SOURCE AND GRANT NUMBER

---DATA SET OVERVIEW:

-Brief dataset description (to be used by the archive as an initial introduction to users.) - **This is CRITICAL!!**

-Time period covered by the data
-Physical location of the measurement or platform (latitude/longitude/elevation)
-Data source, if applicable (e.g. for operational data include agency)
-Any World Wide Web address references (i.e. additional documentation such as Project WWW site)

---INSTRUMENT DESCRIPTION:

-Brief text (i.e. 1-2 paragraphs) describing the instrument with references
-Figures (or links), if applicable
-Table of specifications (i.e. accuracy, precision, frequency, etc.)

---DATA COLLECTION and PROCESSING:

-Description of data collection

-Description of derived parameters and processing techniques used
-Description of quality control procedures
-Data intercomparisons, if applicable

---DATA FORMAT:

-Data file structure, format and file naming conventions (e.g. column delimited ASCII, NetCDF, GIF, JPEG, etc.)
-Data format and layout (i.e. description of header/data records, sample records)
-List of parameters with units, sampling intervals, frequency, range
-Description of flags, codes used in the data, and definitions (i.e. good, questionable, missing, estimated, etc.)
-Data version number and date

---DATA REMARKS:

-PI's assessment of the data (i.e. disclaimers, instrument problems, quality issues, etc.) Missing data periods
-Software compatibility (i.e. list of existing software to view/manipulate the data)

---REFERENCES:

-List of documents cited in this data set description