ASCII File Format Guidelines for Earth Science Data

Status of this RFC

This RFC provides information to the NASA Earth Science community. This RFC does not specify an Earth Science Data Systems (ESDS) standard. Distribution of this memo is unlimited.

Change Explanation

Version 1.1

- updated wording in some checklist items to better reflect the document text
- expanded UMT recommendations to include GMT

Copyright Notice

Abstract

This document lists recommended practices for formatting and describing ASCII encoded data files, such that the files will be self-describing and adhere to common conventions. Included recommendations address: General Structure, Header Information, Data Information, Location Information, Time Information, Missing Data, Limits of Detection, and Filenames. A reference checklist is provided in Appendix B.

1 Introduction

NASA Earth Science data systems produce and manage a wide range of data products that vary greatly in volume and complexity. Many of these data products were originally provided in ASCII file formats without following conventions established among members of the NASA Earth science community. Others follow community ASCII file format conventions that are widely used, but are not approved standards.

In some cases, data systems archive data in a more complex standard format (e.g., HDF-EOS, netCDF), but may offer these data in an ASCII file format for those users who prefer its simplicity.

This document provides recommendations for a minimum and necessary set of information to be contained in an ASCII file. These recommendations were originally developed by the ASCII Earth Science Data Systems Working Group (see the Authors section, below) and adapted and

revised by the ESDIS Standards Office.

By incorporating the recommended information, the ASCII files will be self-describing to the extent that a future data user will be able to decode and use the data in the file while needing to consult few, if any, external sources. These are intended as a basic set of ASCII conventions that are compatible with other, more detailed ASCII file format specifications like ICARTT – a current NASA standard format for airborne observations that has been widely used in airborne field studies.

Not all issues regarding machine readability of ASCII data are addressed here. All such discussion is instead left for a future working group.

2 Recommendations

2.1 Recommendation 1: General File Format Specification and Structure

Data **must** be stored using the American Standard Code for Information Interchange (US-ASCII) as defined in IETF RFC 20 [1] character-encoding scheme and organized as a matrix of rows and columns. The data section of the file comprises ASCII alphanumeric characters (including scientific expressions); and for widest interoperability, the file **should** use the standard US-ASCII character set, without extensions.

The file **should not** contain ASCII control characters as defined in section 4.1 of IETF RFC 20 [1] with the following exceptions: HT (Horizontal Tab), LF (Line Feed), and CR (Carriage Return).

Delimiters between columns of data **must** be one of the following: comma, horizontal tab, space, semi-colon, colon, or the ‘|’ character. Additional space characters **may** be used to aid readability. Visible characters (comma, semi-colon, colon, |) are preferred over white-space characters (space, horizontal tab). If fixed-width data columns are desired, multiple spaces between data columns **may** be used as padding and for the purpose of visual clarity (alignment) of the data. The delimiter chosen **must** be used consistently throughout the data section of the file. If the space character is used as a delimiter, missing data **must** be represented with a missing data value that is identified in the header (see section on missing data below).

Where the delimiter is a “white space” delimiter, particularly when it includes the space character, any data item containing embedded spaces **must** use an “escape mechanism” to ensure the embedded spaces are not interpreted as delimiters. This **should** include the use of matching quotation marks (“ or ‘) around the data item, use of the ‘\‘ character before each embedded space, or replacement of the space with a different character such as the underscore, ‘_’. Note that this same issue applies if including commas in a comma separated file or for any given

character chosen as a delimiter. Characters that are not meant to be delimiters but are the same as the delimiter character **must** be protected by an escape mechanism as described above. Where possible, the delimiter **should** be chosen in a way to avoid requiring the use of an escape mechanism.

This document uses the term "line" or "header line" to refer to information in the header section and "row" or "data row" to refer to information in the data section. Visually, a "line" and a "row" are the same, a series of characters followed by an end-of-line indicator.

Lines (and rows) **must** be separated by end-of-line (EOL) character(s), typically either a character pair such as CR LF or a single CR or LF. The end-of-line character(s) chosen **must** be used consistently throughout the entire file.

The final data row of the file **should** be terminated with the same end-of-line character(s) used throughout the file.

Empty lines or rows (i.e., those that appear to be blank) **should not** be present in a file.

The data section **must** be preceded by the header section, as described in Recommendation #2.

2.2 Recommendation 2: Header Section

The data section **must** be preceded by at least one or more header lines. There is no restriction on how many header lines a data file can have. The file **must** clearly distinguish between header lines and data. This can be achieved in one of the following ways.

> (A) Each header line **must** be delineated by starting with the same specific character — e.g., a number sign (#) and no data line may begin with the same character.
>
> (B) The header section **must** begin and end with a specific pair of delimiters, such as lines containing the words "BEGIN HEADER" and "END HEADER".
>
> (C) The header section **must** end with a delimiter that clearly indicates it is the last line of the header, e.g., a line containing the words "END HEADER" or "BEGIN DATA".
>
> (D) The header section **must** begin with a number that is the total number of lines of header information (including that line). If there is additional information on the same line, the number of lines must be followed by a space or a comma or some other non-numeric character.

The header section is intended to contain documentation about the data and/or metadata needed for the file – such as the data content in each column, out of range and/or missing data values, errors, metadata, instrument documentation, etc. – thus making the file self-descriptive. As such,

at a minimum, the header **must** include at least one line that lists the unique names of the variables (i.e., the column names) separated by one of the delimiter options specified in Recommendation 1.

The variables and units of measurement for each column **must** be clearly and precisely defined in the header, preferably using units of measurement discussed in Recommendation 3. Where applicable, both long and short names and descriptions for each variable **should** be included. If missing data values (described in Recommendation #6) or any other flags values are used, one or more separate header lines **must** define these flags and values.

In order to accommodate a wide range of complexities, information to be included in the header will likely fall between two end member scenarios. Data products with a small number of variables (columns) that are more or less directly acquired by an instrument and require little documentation on data processing **should** include all metadata in the file header. Data products with a large number of columns and a complex processing flow (e.g., one that merges multiple L1B and L2 data products to derive new higher order data products, including uncertainties) **should** provide as much information as possible in the header. In cases where it is not practical to include specifics about data processing, etc., in a header, because this information requires many additional pages of documentation, this information can be maintained elsewhere and the header **must** reference this information. Since the data file may be used well past the tenure of a specific individual, this reference **should** be in the form of a DOI (Digital Object Identifier [4]) or URI [5] rather than simply listing a person's name or direct email address.

The following information **should** be provided in the header: A) Principal Investigator name and affiliation and contact information, in case there are questions about proper interpretation of the data; B) uncertainty information, if it is not reported in the data stream; C) the date of data collection; and D) the date of data processing. If the data were revised/updated, for purposes of record keeping and proper data use the header **should** include a record of the history of data revision.

If a DOI is available for the data in the file, it **should** be included in the header.

2.3 Recommendation 3: Data Representation

The data representation of a variable is defined by the units in which it was measured or derived. All data **should** be represented using units of measurement approved by the International System of Units (SI system), derived units (such as degree Celsius), or non-SI units accepted for use with SI (such as minute, hour, day, mixing ratio data).

Different research communities may have developed units of measurement and representations via long-standing convention. When data is represented in a community-specific manner, this

must be described in the header section.

In cases where the entire file consists of measurements taken from a single location, that location **must** be either included in each data row or be identified in the header (see Recommendation 4 for details pertaining to representation of geographic location information).

2.3.1 Point/Time Series Data

ASCII is often used for low-volume time-series data with parameters being measured sequentially (and/or simultaneously) in time. It is also suitable for two-dimensional (e.g., along-track) derived geophysical products that have been interpolated onto a common geographic location and/or time base. For these data, all rows **must** have an associated geographic location and/or time tag - as specified below. All variables in a row **must** correspond to the specified geographic location and/or time tag.

2.3.2 Profile/Gridded Data

This section applies to many categories —sometimes referred to as profile data, gridded data, or any generic two-dimensional dataset. Examples include height data all at a given time, for a sequence of times, e.g., Lidar data; or gridded data at the same time, for a sequence of times. These data all have a common independent variable - such as time or location. For each independent variable, there will be a series of row data with a second independent variable, e.g., altitudes. Each row will consist of all the data variables (dependent variables) that are valid at the values of these two independent variables.

Where applicable, if all values of a 2-dimensional grid, such as a latitude-longitude grid, are output without explicitly including the latitude or longitude, the header section **must** indicate which index runs faster.

The data gridding scheme (regular, rectangular, polar, etc.) **should** be described in the header.

2.4 Recommendation 4: Location Information

This recommendation is applicable to any location information contained in the header section or the data section.

Geo-located data products **must** have an associated geographic location — including latitude, longitude, and **should**, where applicable, have an associated elevation (or altitude).

Named geographic locations (e.g., “Station 14”, “Buoy Alpha”, “St. Louis”) **should** be accompanied by geographic coordinates.

Latitude and longitude **should** be reported in the format applicable to the coordinate reference system in use, e.g., decimal degrees with south latitudes and west longitudes represented as negative numbers (i.e., no N, E, W, S identifiers). The latitude/longitude (and elevation, if applicable) convention employed **must** be specified in the header, and the coordinate reference system and datum (if applicable) **should** also be included. To ensure consistency, elevation **should** be reported in meters. If applicable, the type of elevation measurement being used **must** be identified in the header section, since many types of elevation measurements are available and in common use (e.g., GPS altitude; radar altitude; pressure altitude, subsurface depth, derived/interpolated, etc.).

Precise geodetic information requires an underlying reference frame and reference ellipsoid. The reference frame, such as the International Terrestrial Reference Frame (ITRF) and its epoch (e.g., 2008) **should** be explicitly defined in the header section or in a separate document that accompanies the data file. The reference ellipsoid, such as WGS-84, GRS-80, or TOPEX **should** also be included. The header **should** document information about the type of location information that was used to create latitude, longitude, and elevation information. For example, some data products are created using real-time GPS feeds into a logging computer, while other products are created using post-processed DGPS or PPP solutions interpolated onto the time tags of the measurements and their variables.

For aircraft data, the header section **should** document the location of the GPS antenna on the aircraft and whether antenna positions have been used for the geolocation or if the locations in the data file reflect sensor/instrument positions.

2.5 Recommendation 5: Time Information

This recommendation is applicable to any time information contained in the header section or the data section.

All dates/times **should** be reported in UTC (or GMT) and the representation of dates and times **should** follow the ISO 8601 standard [2] (e.g., 'yyyymmddTHHMMSS.SSSZ'). If UTC (or GMT) is not used, the header **must** specify the time standard that is used (e.g., TAI or local time). If local time is used, time information **must** include the time zone name and offset. The time zone name **should** be taken from the IANA Time Zone Database [3]. Daylight savings time values **must** not be used.

Where possible, years **should** be represented with four digits.

If ISO 8601 is not used the time representation **must** be included in the header.

In data sets where time is an independent variable, all data points (variables) in a row **must** have

the same time stamp and consecutive rows **should** have monotonically increasing or unique time tags.

The time base information **should** be explicitly defined in the header and information regarding the time base **should** include the source of the time stamps (e.g., real time kinematic GPS, logging computer time).

Timestamps may be reported as UTC decimal seconds from the time at which measurements began (commonly as seconds past midnight). If so, the reported time **should** be monotonically increasing even when crossing over a date boundary. If this is the case, then the date **should not** roll over - so that when the UTC seconds are added to the start time information, the correct time tag is produced.

For some datasets, two time stamps are provided for a given row, indicating stop and start times for the measurement. When a single time stamp is used, it **should** be clearly defined in the header whether it is the start, stop, or the midpoint of the measurement period. Ideally the column will be given a descriptive name (e.g., START_TIME). If the measurements were made in irregular time intervals (or integration intervals), the start and stop time stamps **should** be reported.

The processing steps to derive certain higher-order geophysical products might not make it possible/practical to keep timestamps with the data product. In that case, an average time or the start and stop times reflecting the data collection window **should** be included in the header metadata.

If internal computer clocks have been used, the header **should** include whether or not these clocks have been synchronized to GPS time or some other time reference.

2.6 Recommendation 6: Missing Data and Limits of Detection

There are cases where measurements cannot be made due to instrument or other related issues. If commas or other visible characters are used as the delimiter, the field in the data record that would normally include the missing measurement can be left absent. However, if using space or tab delimiters, the field in the data record **must not** be left empty or blank and instead, **must** include some designated value for the missing data.

The value(s) used to designate missing data **must** be described in the header. Missing data **should** be represented by numbers of enough magnitude to never be construed as actual data, such as -999 (or -99999, etc.) or be represented by a string such as NaN (Not a Number).

Data below (or above) a limit of detection (LOD) are not actually “missing” but do convey

useful information, when used to compute descriptive statistics. These conditions **should** be indicated by additional missing data flags that are substituted for the missing data values. If used, these flags and the values of the upper and lower LOD **must** be described in the header.

For example, the flag sometimes used for data values GREATER THAN some UPPER LOD (ULOD) is -7777 (or -77777, etc.), and the flag for data values LESS THAN some LOWER LOD (LLOD) is -8888 (or -88888, etc.).

If LLOD or ULOD values vary from point to point, they **should** be given in a separate column of data.

Any other flag values used in the data section **must** also be described in the header.

2.7 Recommendation 7: Filenames

Data files are often copied or moved from the organization that created them to other places. Names that adhere to organizational naming conventions might not be sufficiently descriptive or distinctive outside of the organization.

Different operating systems place different restrictions on the use of special characters within a file name. Use of special characters such as slash, backslash, and colon should be avoided.

Some communities have developed naming conventions, e.g., ICARTT [6].

Descriptive file names **should** be used to convey basic information about the data it contains. Unique file names **should** be employed.

3 References

3.1 Normative References

[1] http://tools.ietf.org/rfc/rfc20.txt
[2] http://www.iso.org/iso/home/standards/iso8601.htm
[3] https://www.iana.org/time-zones

3.2 Informative References

[4] https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS
[5] https://tools.ietf.org/rfc/rfc3986.txt
[6] http://www-air.larc.nasa.gov/missions/etc/IcarttDataFormat.htm#22

4 Authors

ASCII Earth Science Data Systems Working Group

Keith Evans, Working Group Technical Chair
Joint Center for Earth Systems Technology, NASA GSFC / UMBC
evans@umbc.edu

Aubrey Beach, NASA LaRC / Booz Allen
Gao Chen, NASA LaRC
Peter Leonard, NASA GSFC / ADNET Systems, Inc.
Emily Northup, NASA LaRC / SSAI
Anne Wilson, Laboratory for Atmospheric and Space Physics, U Colorado
et al.

Revised and adapted for ESO RFC format by ESO staff
eso-staff@lists.nasa.gov

Appendix A - Glossary

ASCII – American Standard Code for Information Interchange

CR – Carriage Return

DGPS – Differential GPS

DOI – Digital Object Identifier

ESDS – Earth Science Data Systems

ESDSWG - Earth Science Data System Working Groups

EOL – End of Line

EOS – Earth Observing System

ESO – ESDIS Standards Office

GMT – Greenwich Mean Time

GPS – Global Positioning System

GRS – Geodetic Reference System

HDF – Hierarchical Data Format

HT – Horizontal Tab

IANA – Internet Assigned Numbers Authority

ICARTT – International Consortium for Atmospheric Research on Transport and Transformation

IETF – Internet Engineering Task Force

ISO – International Organization for Standardization

ITRF – International Terrestrial Reference Frame

LF – Line Feed

LOD – Limit of Detection

LLOD – Lower Limit of Detection

netCDF – Network Common Data Form

PPP – Precise Point Positioning

RFC – Request for Comments

SI – International System of Units

TAI – International Atomic Time

ULOD – Upper Limit of Detection

URI – Uniform Resource Identifier

UTC – Coordinated Universal Time

WGS – World Geodetic System

Appendix B – Checklist For Earth Science Data Files in ASCII

This checklist is provided as a reference for creating ASCII data files that comply with these guidelines. The checklist is meant to be a guide only. The recommendations in the main body (indicated by "R#") are definitive.

General Requirements

- ☐ Create files with separate header and data sections – R1
- ☐ Use a consistent delimiter between data values throughout the file – R1
- ☐ Use escape mechanism if the designated delimiter character appears in text or data – R1
- ☐ Separate lines of text and rows of data with end-of-line (EOL) character(s), used consistently throughout the file – R1

General Recommendations

- ☐ Use the standard US-ASCII character set, without extensions – R1
- ☐ Avoid ASCII control characters, except tab or EOL characters – R1
- ☐ Do not use empty lines or rows – R1
- ☐ Chose delimiter character to avoid need for escape mechanism – R1
- ☐ Terminate file with same end-of-line (EOL) character(s) used to separate data rows – R1
- ☐ Use unique, descriptive file names – R7

Header Section – required

- ☐ Clearly delineate header section as described in this document – R2
- ☐ List unique variable names (columns) – R2
- ☐ Define units of measure for each variable – R2
- ☐ Identify community-specific convention used for data representation and/or units of measure, if applicable – R3
- ☐ Specify conventions used for latitude, longitude, and elevation if applicable – R4
- ☐ Identify type of elevation measurement used, if applicable – R4
- ☐ Indicate which data grid index runs faster, if applicable – R3
- ☐ Reference any additional documentation needed to understand the data in the file, preferably by DOI – R2
- ☐ Specify time representation if not using ISO 8601 – R5
- ☐ Specify time zone and offset if using local time instead of UTC or GMT – R5
- ☐ Specify single location or time associated with all data in the file, if not specified with each data row – R3
- ☐ Define missing or out of bounds data fill values, any other flag values – R6

Header Section – recommended

- ☐ Provide as much metadata as practical, all metadata if possible – R2
- ☐ For each variable, provide long and short names and description – R2
- ☐ Describe gridding scheme used, if applicable – R3

- ☐ Define geographic reference frame and ellipsoid – R4
- ☐ Specify coordinate reference system, and datum if applicable – R4
- ☐ Document type of location information used – R4
- ☐ Document location of GPS antenna on aircraft, if applicable – R4
- ☐ Provide geographic coordinates for place name associated with data – R4
- ☐ Define time base information, including source of time stamps – R5
- ☐ Identify time zone from IANA Time Zone Database if using local time – R5
- ☐ Specify whether time stamps identify start, stop, midpoint or average of measurement period – R5
- ☐ For averaged or derived products, indicate data collection window – R5
- ☐ Indicate whether internal computer clocks are synchronized to GPS time or other – R5
- ☐ Provide principal investigator name and contact information – R2
- ☐ Provide uncertainty information – R2
- ☐ Indicate dates of data collection and processing – R2
- ☐ Provide a record of data revision – R2
- ☐ Provide data DOI if available – R2

Data Section – required

- ☐ Organize data as matrix of rows and columns – R1
- ☐ Provide geographic location and/or time tag (as applicable) for each data row or value – R3, R4, R5
- ☐ Use a designated flag value to indicate missing data when using space or tab delimiters – R1, R6
- ☐ Do not use daylight savings time if using local time instead of UTC or GMT – R5

Data Section – recommended

- ☐ Provide lat/lon in format applicable to coordinate reference system used – R4
- ☐ Provide elevation for each data row or value if applicable – R4
- ☐ Specify elevation in meters – R4
- ☐ Provide data in SI units, derived units (such as degree Celsius), or non-SI units accepted for use with SI (such as minute, hour, day, mixing ratio data) – R3
- ☐ Provide date/time in UTC or GMT, following ISO 8601 standard – R5
- ☐ If time is specified in seconds past some starting point (e.g., midnight) and measurements in the file span date boundaries, assure that time increases monotonically (>86400) and date does not change – R5
- ☐ Structure data so that consecutive rows have monotonically increasing or unique time tag where applicable – R5
- ☐ Represent years with four digits – R5
- ☐ Provide start and stop timestamps for measurements in irregular intervals – R5
- ☐ Indicate data above or below a limit of detection using a flag value – R6
- ☐ Represent flags (missing data, etc.) so as not to be construed as data – R6
- ☐ Provide a separate column for flag values that vary from point to point – R6