ICARTT Format Enhancements based on ESDSWG Recommendations

Status of this RFC
This RFC provides information to the NASA Earth Science community as well as the airborne instrument and data modeling communities. This RFC specifies proposed changes to a current Earth Science Data Information Systems (ESDIS) data format standard. Distribution of this memo is unlimited.

Change Explanation
This document is a new document that proposes additional guidelines and constraints on ESDS-RFC-019 *International Consortium for Atmospheric Research on Transport and Transformation (ICARTT) File Format Standards*.

Abstract
This document lists recommended guidelines, as derived from representatives of the airborne data community, for formatting and describing ICARTT data, such that the format will support more rigorous standards for metadata and substantially enhance machine-readability. Included recommendations address general file format specifications and structure for metadata, such as: variable standard name, version number, and time information structures.

1 Introduction

NASA Earth Science data systems manage data products that vary greatly in volume and complexity. Airborne field campaigns in particular generate a great wealth of observations, including a wide range of the trace gases and aerosol properties. The recent NASA Earth Venture (EV) Program has generated an unprecedented amount of aircraft observations in terms of the sheer number of measurements and data volume. ICARTT was designed to be a relatively simple ASCII standard data format, which allows the data files to be self-describing. The current standard has gained popularity among the airborne data user community and airborne instrument scientists because it is simple to use and is generally suitable for scientific research on climate change and air quality issues. However, over the years the use of the ICARTT format has revealed a number of issues that can complicate data usability. This ICARTT Refresh working group identified seven deficiencies of the current ICARTT V1.1 format by soliciting input from data providers, DAAC representatives, and end users. Through direct collaboration and use case analysis, we chose to concentrate on mitigating three of these seven deficiencies with the aim of substantially improve the ICARTT format. These ongoing discussions took into

consideration level of effort and possible impacts on data providers, data users, and DAACs. This document provides recommendations for necessary updates regarding current ICARTT file format standards. By incorporating these recommendations, new ICARTT files will improve upon the existing format thereby increasing airborne data accessibility and discovery. These recommendations are also intended to maintain backwards compatibility with current ICARTT parsers.

2 Recommendations

Recommendation 1: Extended Characters

Extended characters (umlauts, accents, etc.) need to be allowed in some parts of the file header and places in the file header where only ASCII characters are allowed need to be identified. The data section of the file will continue to be comprised of only ASCII characters.

Current Standard: The ICARTT format exclusively uses the ASCII character set

Recommendation: ICARTT files are text files, and therefore primarily use the ASCII character set. Within the header (metadata) section of a file, the ASCII character set may be extended using UTF-8 encoding (with some exceptions, as detailed in section 2.3.B[1]). The data section of a file, however, can only contain ASCII characters.

Given that ICARTT files require PI names and institution names, ICARTT files must allow PIs to spell their names correctly. This implies that ICARTT files must support extended characters (umlauts, accents, etc.) that are not part of the ASCII character set. The standard encoding for extended characters has become UTF-8. Its advantages over other encodings are that it is compatible with ASCII, it can represent an unlimited number of characters, and there is no ambiguity interpreting UTF-8 characters.

There are relatively few places in the ICARTT file header where UTF-8 characters would cause any problems. The primary exception is variable short names and standard names, where not even all ASCII characters should be permitted (see below). Also, the current file name requirements only allow ASCII characters; this is a requirement for some file systems and therefore needs to be maintained.

Recommendation 2: Variable Standard Name Structure

Add a required standard variable name entry before the long name in the variable name line.

Current Standard: Variable short name, variable unit, variable long name (optional)

Recommendation: Variable short name, variable unit, *variable standard name*, variable long name (optional)

The current ICARTT format requires that each dependent variable be listed on a single, comma-delimited line as page 10, section 2.3.B [1] of the ICARTT File Format Standards V1.1 document (ESDS-RFC-019v1.1).

Since each variable short name is chosen by the instrument Principal Investigator, this leads to a lack of standardization. Data users often need to adapt parsers for data files generated from different airborne platforms and campaigns. This was considered by the ICARTT Refresh working group to be the number one issue based on user feedback. In an effort to improve usability, standardization, and machine-readability, we propose adding a standard variable name entry prior to the optional variable long name. This document does not provide a list of standard variable names but does provide a structure to facilitate the use of standard variable names. The cultivation of a list of standard variable names may be the subject of a future working group.

For DAACS, the use of standard variable names will enable data search across instruments, platforms and missions, i.e., global searches for aircraft data. The data users will be able to use unified code to read the variable of interest from files generated by different airborne instrument PIs.

The standard name will be generated from a controlled list, the generation of which will be a collaborative effort of this working group and members of the user community. One particular effort involves a WMO working group for Atmospheric Composition Vocabulary Registry (TT-ACV). We will be in communication with TT-ACV and work to adapt a standard with broad community acceptance.

We recognize that this working group lacks the ontological and semantic expertise to create a comprehensive list of standard variable names. The current plan of this working group is to initially adopt the WMO TT-ACV recommendations. Our confidence in WMO TT-ACV is based on the composition of the team, which includes atmospheric scientists and data ontologists. A future working group will be required to analyze those recommendations and solicit feedback from NASA airborne instrument and modeling community.

Recommendation 3: Formatting Variable Names

Limit all variable short and standard names to 31 characters and require that all start with a letter.

Current Standard: None

Recommendation: The formats of variable short name and the variable standard name are as follows: each name can be at most 31 characters in length composed of alphanumeric characters (A-Z, a-z, 0-9) and underscores (_). The first character must be a letter.

Variable short names are case-sensitive and must be unique within a given file (i.e. no two variables in a file can share the same short name). It is strongly recommended that short names should not be distinguished purely by case; that is, even if case is disregarded, no two short names should be the same. The short name must be unique within a given file (e.g., no two variables in a file can share the same short name), but standard names are not necessarily unique. The variable long name, however, can be as long as necessary and can contain any ASCII or UTF-8 character.

These names are case-sensitive, but it is recommended that names should not be distinguished purely by case, i.e., if case is disregarded, no two names should be the same. The short name must be unique within a given file (e.g., no two variables in a file can share the same short name), but standard names are not necessarily unique. The variable long name, however, can be as long as necessary and can contain any ASCII or UTF-8 character.

Recommendation 4: Version Number

Add a field to the end of the first header line for the version number.

Current Standard: Number of header lines, FFI

Recommendation: Number of header lines, FFI, *Version number*

It is envisioned that the ICARTT format will undergo further improvements to better fit the data products from future development of atmospheric measurements. Each new version of the ICARTT format could introduce new standards, which consequently would require users to update any programs used for reading these files. With the current file format, there is no specific way to indicate which file version is being used, making it difficult to determine which rules should be used for parsing.

To clearly identify the version of the ICARTT format, we suggest adding a version number that would be used to distinguish the specific standards aligned with that version. This would be added to the first header line, which currently contains the number of header lines and the File Format Index (FFI). This recommendation will be backward compatible with most current ICARTT file reader codes.

Placing this indicator at the start of the file would allow parsers to determine the required rules without having to dive deeply into the file. The version number will be represented by: “V##_YYYY”; whereas V## denotes the version number and YYYY indicates the year this version was approved.

Recommendation 5: Revision Identifier

Clarify the use of R#: data revision identifier.

Current Standard: The R parameter is not optional in the ICARTT data format. One must specify a data revision code that tracks updates to the data. This also requires documentation of those updates (e.g., new calibration, timing error, etc.) to be recorded in the file header (see section 2.3.B). For this we specify a revision identifier "_R#" in the filename where the underscore is a required element to separate the fields (this is needed for certain file checking software). The revision identifier "#" must match the revision identifier specified in the Normal Comments section of the file header (see section 2.3.B).

Recommendation: Update current document to distinguish between field (RA, RB, RC, etc) and preliminary and final data (R0, R1, R2, etc).

Over the course of a field campaign, "field" data files are generally created. Data exchanged during the field study are considered a special case since these data are typically "first look" and, due to time constraints, are not likely to have undergone the full scrutiny of the PI. Field data files are generally not made available to the public. These files should be deleted as soon as possible after the study and replaced with preliminary data files which will have some QA/QC performed.

For field data, revision identifier will start from letter "A", e.g., RA, RB, … etc. The field data revision identifier must be a single letter and should be capitalized in both the filename and header. For preliminary and final data, revision identifier will start from "0", e.g., R0, R1,…,R10,… etc. The preliminary and final data revision identifier must be at most two digits.

If filenames for field data need to be distinguished, campaigns may use revision identifers starting from letter "A", e.g., RA, RB, … etc. The field data revision identifier must be a single letter and should be capitalized. Otherwise, revision identifiers must be numbers starting from "0", e.g., R0, R1,…,R10,… etc. Numerical revision identifiers must be at most two digits.

Recommendation 6: Time Information

Require that the standard names for the date time variables be Time_Start, Time_Stop, and Time_Mid

Current Standard: None

Recommendation: short_name, unit, Time_(Start,Stop,Mid), long_name (optional)

The current ICARTT format requires that for any data reported at intervals longer than one second, both the interval start-time and interval stop-time must be reported, with optional mid-time reporting. Continuous data reported at intervals of one second or shorter is only required to report a single time column, although there is no requirement

as to which point in the interval is reported. While these requirements are rigorous, the non-standard variable names used for time reporting variables often create ambiguities, particularly when only one is used. This causes difficulties for merged (geo-located) data, as well as certain analyses. For merged data, choosing the wrong point in the reporting interval will change the merged values e.g. assuming that an interval is reported at the start time when it is actually reported at the mid time will incorrectly shift each merge interval by half a second. Given a typical aircraft speed of 100 to 200 m s-1, this will create a significant sampling location shift. Plume analysis requires precise sampling time, and often correlation analysis between chemical tracers, demanding the need for standard variable names for time reporting.

We propose standardizing the format of the time variable standard names. This would allow both users and ICARTT parsers to correctly determine what a time variable is reporting without ambiguity.

For multi-dimensional data, i.e. File Format Index numbers other than 1001, this approach would be extended to the other independent variables. For example, where time might be reported as Time_Start, Time_Stop, and Time_Mid, other independent variables would be reported as IndepVar_Start, IndepVar_Stop, and IndepVar_Mid.

Recommendation 7: Normal Comments Section Formatting

Clarify the standards for the normal comment section including required keyword/value pairs and formatting for the actual keywords.

Current Standard: The current documentation requires that all keywords are included, however the scanning software only enforces UNCERTAINTY, U/L LODs, and REVISION. Keywords are case insensitive.

Recommendation: The normal comment section must include all keyword/value pairs listed in the ICARTT documentation [1]. Each keyword must appear at the start of the line, with no preceding spaces. Keywords must consist of capital letters and underscores and be followed by a colon and a space. The value corresponding to the keyword must appear on the same line, following the colon. Where no value exists, "N/A" should be used.

It is preferable to have the entire keyword/value pair appear on a single line, however, the value is allowed to wrap across multiple lines.

Recommendation 8: Multiple ULOD/LLOD Flags/Values

There is currently no clear documentation on how to provide multiple ULOD or LLOD values. The only relevant information is in the 2110 and 2310 addenda – but most of it is equally applicable to FFI 1001. Therefore, the information should be part of the main

ICARTT document, with slight changes to improve machine readability of LOD information and to maintain consistency with the rest of the ICARTT header.

Current Standard: None for all ICARTT files, although some information is provided for 2110 and 2310 files.

Recommendation: The value for the LLOD_FLAG keyword must consist of either a single flag value that applies to all dependent variables in the file or a comma-separated list of flag values with one flag value per dependent variable (where the order of the flag values corresponds to the order of the dependent variables). Each flag value must either be "N/A" or -888[8...], with enough 8s so that the flag value cannot be misconstrued as real data. The flag value must have at least three 8s and be at least one order of magnitude more negative than the most negative real data value in the corresponding dependent variable column (or file, if only one flag value is provided for the entire file).

The value for the ULOD_FLAG keyword must consist of either a single flag value that applies to all dependent variables in the file or a comma-separated list of flag values with one flag value per dependent variable (where the order of the flag values corresponds to the order of the dependent variables). Each flag value must either be "N/A" or -777[7...], with enough 7s so that the flag value cannot be misconstrued as real data. The flag value must have at least three 7s and be at least one order of magnitude more negative than the most negative real data value in the corresponding dependent variable column (or file, if only one flag value is provided for the entire file).

The value for the ULOD_VALUE/LLOD_VALUE keywords must consist of a single LOD value that applies to all dependent variables in the file or a comma-separated list of LOD values with one LOD value per dependent variable (where the order of the LOD values corresponds to the order of the dependent variables). Each LOD value must be one of the following: "N/A", a numeric value (i.e. the LOD is constant throughout the file), or the short name of a dependent variable (i.e. the LOD changes over time and is recorded in the indicated dependent variable column).

Note that the ULOD_VALUE/LLOD_VALUE for Time_Stop and Time_Mid variables must be "N/A", as they have no limit of detection.

For file format types 2110 and 2310: When identifying LOD flags or values using the comma-separated list method, the values for auxiliary variables must be provided first, followed by the values for primary variables. The order of the variables is maintained within each group.

3 References

[1] ICARTT File Format Standards V1.1:
http://www-air.larc.nasa.gov/missions/etc/ESDS-RFC-019-v1.1_0.pdf

4 Authors

ICARTT Earth Science Data Systems Working Group

Emily Northup, Working Group Technical Chair
Atmospheric Science Data Center, NASA LaRC
emily.a.northup@nasa.gov

Working Group Co-Chairs:
Gao Chen, NASA LaRC
Kenneth Aikin, NOAA ESRL
Christopher Webster, NCAR RAF

ESDIS POC:
David Batchelor, NASA ESDIS

Adapted for ESO RFC format by ESDIS staff

Appendix A – Glossary

Glossary of Acronyms
ASCII - American Standard Code for Information Interchange
DAAC – Distributed Active Archive Center
ESDIS – Earth Science Data and Information System
ESDSWG - Earth Science Data System Working Groups
ESO – ESDIS Standards Office
FFI – File Format Index
ICARTT – International Consortium for Atmospheric Research on Transport and Transformation
PI – Principal Investigator
WMO – World Meteorological Organization