

Review of ICARTT File Format Implementation and Operational Suitability

NASA Airborne Science Program Consolidated Response

NASA's Earth Science Data Systems Standards Process Group (SPG) is considering the International Consortium for Atmospheric Research on Transport and Transformation (ICARTT) file format for adoption as a community standard. The ICARTT file format was developed to fulfill the data management needs for the ICARTT campaign in 2004. This file format is text-based and composed of a header section (metadata) with critical data description information (e.g., data source, uncertainties, contact information, and brief overview of measurement technique), and a data section. Although it was primarily designed for airborne data, the ICARTT format proved to be practical for other mobile and ground-based studies and various data types.

In response to a request for comment (RFC) from the SPG, NASA's Airborne Science Program has collected and consolidated feedback and comments relevant to implementation experience and suitability for operational use.

Based on the small set of comments received to date from individuals without significant history of actually using the ICARTT format, the following summary statements can be inferred:

- There appears to exist a community of users for which this format was deemed useful in the 2004-2009 time period.
- Minor ambiguities and criticisms discussed here notwithstanding, the ICARTT format specification appears adequate for supporting users on an as-needed basis.
- The ICARTT format as it exists today is not adequate for supporting next-generation application infrastructure.
 - Extensibility and schema that enable automated verification are lacking in ICARTT.
 - ICARTT is not useful for realtime or online applications. In future automated systems without significant human-based data processing workflows, there is a need to consider seamless/automated transition between realtime and archival formats.
 - Scalable, extensible archives that support or facilitate random access and data discovery are needed in the future. This is an area of active research and development without clear solution paths, although HDF-EOS, HDF5 and NetCDF represent a major investment across scientific communities.

Implementation Experience questions:

1. (*Your background*) Describe in a sentence or two your overall implementation experience related to the proposed specification. (*e.g., specification implementer, tools developer, data provider, scientific analyst, science user, etc.*) Have you directly implemented the ICARTT format specification? Did you use a pre-existing software package, and if so, what did you use?

I am the lead software engineer at the NCAR Research Aviation Facility. I develop tools, formats, and standards for our use.
I have not used, nor have I evaluated this format prior to the current request.
In our work we deal with online distributed systems, primarily for situational awareness displays. Our solutions that use ascii data for distribution have migrated toward comma separated variables (CSV) with an external/seperate XML file containing metadata. Distribution mechanisms such as UDP and open source caching middleware such as DataTurbine enable use of commercial tools like web browsers, MSEXcel, Matlab, IDL, etc. The CSV approach also constitutes the dominant archive file approach used in REVEAL systems. We have not implemented any ICARTT-based archiving for the platform data systems.
I do not have any legacy experience with either the Ames or ICARTT formats. That said, we're giving the Science Community the option of retrieving the data from the PostGIS database being used on Global Hawk missions in a variety of formats, Ames and ICARTT included.
I have had to produce/post navigation data to an ICARTT archive at least once (most recently for the J-31's Milagro/Intex mission). The submission was for post flight data "archive" - in response to the aircraft scientist's request. Our facilities perspective has been to provide what science asks for (not vice versa). To that end, we have produced a variety of post flight formats for navigation data products (ICARTT, Ames as well as numerous "facility" formats). In my opinion, as a data provider, the emphasis should be on science requirements. Offering/supporting a variety of formats (IWG1, Ames, or ICARTT) is the optimal approach.

2. (*Completeness*) Does the specification provide all the detail you need to implement it in software? (*e.g., to read or write a data file; to implement the specification, a profile or extension; or develop a tool such as a format translator*) If not, describe what is missing in the specification.

The specification looks complete such that I could develop an ICARTT-based software tool.

3. (*Accuracy*) Do any parts of the specification contain inaccuracies, or internal inconsistencies? If so, please provide details.

Section 2, "File Format Specifications", suggests that 3D data "clearly cannot be represented as a single time series." My experience with RBNB suggests otherwise – for instance, any amount of data can be stored as a data "blob" associated with a timestamp (can be used, for instance, to store image data).
The top of p14 mentions "In these examples any continuation of lines from directly above has been indented for clarity", but this is not actually the case. To avoid line wraps, you may want to put the examples on landscape-oriented pages.
Section 2.3.B mentions that "Delimiters are commas only (spaces for legacy purposes only) and cannot be used anywhere else in the file. For delimiters to separate text, use underscores." There seems to be inconsistent use of this policy as shown in the provided examples: sometimes commas are used; sometime semicolons are used. Spaces are also used in text header fields w/o the use of underscores. In fact, in apparent contradiction to this policy, the last paragraph on page 12 states "When more than one value (or information) is to be written on the same line, separate the values using a semicolon."

4. (*Clarity*) Is any part of the specification ambiguous, or poorly explained? If so, please provide details.

Section 2.1.A doesn't mention how to handle data files that cross over multiple days; the specification doesn't mention this is possible until Section 2.3.B where it gives a side comment about "crossing over to a second day".

For a single datapoint that represents data that is integrated over time, how do you define location? Do you give the location at the start/mid-point/end? Or maybe give location at 1Hz intervals over the course of time used to derive the integrated quantity?

Section 2.2 ("File names") mentions "All fields not in square brackets are required and are described as follows". However the list that follows describes both the required and optional fields. A better wording might be "Fields in square brackets are optional. Fields are described as follows:"

The last paragraph on page 8 is supposed to give a simple example of the use of Revision from the ICARTT study. However, this paragraph is more confusing than anything; doesn't provide a clear, simple example. I find the mixed use of letters and numbers to indicate different meanings of Revision (for instance, "RA" and "R0") confusing/not clear.

I do not understand the meaning of the following sentence on page 10: "For example, if measurements are reported as 1 second (or subsecond) intervals, then stop time and mid-point time need not be included as data columns provided all time intervals in the measurement period are accounted for by inclusion of the missing data flag(s)."

Issues in Section 2.3.B:

I assume each "bullet" in the list represents its own line in the data file? If so, this should be clearly stated.

The spec mentions "All intervals longer than 1 second must be reported as Start and Stop times". Does this mean that for data where intervals are >1 second, the file needs to include start/stop columns? Seems very inefficient and unnecessary for many cases; if a single point time alone can describe the data, why dictate that stop time needs to be included?

Not sure I understand the use of the "FLAG" field versus the "VALUE" field for ULOD and LLOD. For ULOD, does this mean that when we see the ULOD_FLAG in the data ("-7777", for instance) then it means the data at this time was above the value indicated by the ULOD_VALUE field?

As discussed in Section 2.5, it isn't clear to me why ICARTT doesn't support satellite data. Why isn't "gappy" data supported by specifying Data Interval 0 and then the start/mid/stop time for each blob of collected data?

5. (*Usefulness*) How well does this specification meet your information sharing needs? (*e.g., does it work well with the data types and data manipulations in your application? Does it properly represent your datasets? What are the pros and cons of this data format?*)

ICARTT meets most if not all of our time-series data types; scalars, and histograms. Being an ASCII format it does not meet our criteria for large datasets (300-500 variables). We find ASCII formats such as Ames and ICARTT useful for accepting data from investigators to integrate into our larger netCDF datasets.

Text formats are nice for human readability, but not necessarily for making a compact file for long-term storage or transmission over constrained BW networks.

The "Introduction" lays out the need for a standard format to share data amongst diverse groups and also the need for long-term storage. For airborne science data, why not investigate the formats (HDF, for example) that are currently used for long-term storage by NASA EOS?

Good to see that this standard supports multiplexed data (as long as all quantities share a common time-base).

Sect 2.1.c(ii): The use of "-999999..." to represent missing data is wasteful for an ascii format, subject to misinterpretation, and arguably antiquated. Why wouldn't consecutive commas communicate the same information? This section is discussing the existing mechanisms for communicating "NULL" (unknown or missing data) and two classes of Not a Number (NaN) (upper and lower limits of detection)

6. (*Implementation*) What implementation challenges does the proposed standard present? Please provide details, if any.

No API for marshalling/unmarshalling data is provided.

Format is not supported by many tools – in fact, the only tool mentioned is a NASA-sponsored online tool to scan ICARTT files.

While this format could be used for more general data storage/sharing, the language of the standard (as well as the intended audience) is specifically for the airborne science community. For example, the use of "Launch" in the file name and the assumption that most parameters will be about 1Hz.

I would rather use absolute timestamps rather than time from the start of the day.

No xml schema or DTD exists.

Operational Suitability questions:

7. Do you currently use or plan to use the ICARTT format in a production setting? Do you plan to distribute data in this format to science collaborators and other researchers?

No, we distribute in binary/netCDF as we find ASCII to be poorly suited for large datasets. We currently provide translators that go in and out of NASA Ames DEF. These are mostly for internal use. I am considering switching these translators to use ICARTT instead, as the Ames DEF lacks formal specification for variable 'units'.

No current plans. Although we are currently examining alternative data formats for archiving RBNB [*online/live data distribution middleware, www.dataturbine.org*] data, with my current understanding of the ICARTT format, I would not recommend using this specification for RBNB.

8. Why do you choose to use the ICARTT format over other data formats for your applications?

No comments received

9. Does the ICARTT file format meet your requirements for storing and accessing data?

No.

1) ASCII does not allow random access to data.

2) The ICARTT header is not extensible to allow for future growth. e.g. The global attributes for ICARTT were shoe-horned into the comments field of the Ames format. How would a new attribute for a variable included?

My OPINION is that bundling the data and metadata together like this, as opposed to, say, the IWG1 data and separate metadata files, seems less than ideal.

10. Have you or your users encountered any difficulty when using some of the data in the ICARTT format? If you have, please provide a brief description of your experience.

No comments received

11. What operational challenges or limitations does ICARTT present? Please provide details.

No comments received

12. What benefits does the ICARTT file format present? Are there any drawbacks to using this file format? (*e.g., Does it offer the flexibility you want to package the data types in your applications? Does it facilitate interdisciplinary studies?*)

Drawbacks are [previously noted].

Benefits include:

- It is a relatively simple format
- The format has been tested/used in the "real-world" by teams composed of international and cross-department participants.
- The format appears to have a certain measure of acceptance by the community.

13. How much data do/will you provide or archive in the ICARTT format? (*Number of distinct data sets, total data volume, number of files.*)

We will not archive data in this format. We will fulfill specific requests for it.

For the large imagery data sets that we mostly work with, I believe that EOS-HDF will continue to be the NASA standard.

14. How many users do you have or expect to have for data in the ICARTT format, and what is your expected user community?

No comments received

15. (*User comments*) Any additional comments, observations or criticisms of the ICARTT format and the RFC can be provided here.

This format does not lend itself well to raster data from remote sensing instruments. NASA has invested a lot of effort into EOS-HDF, and that file format is supported by most commercial image processing packages.

<p>The location information lacks sufficient precision for some applications. (Unless I miscalculated, Decimal Degrees to 5 decimal places is about 3.6 feet.) Airborne systems can generate sub-meter data. Overlay of geographic data with only about a meter precision will generate unacceptable errors at boundaries of pixel groups, worse for vectorized data</p>
<p>I haven't looked in detail at all the specification, but it does not seem to be useful for mapping applications.</p>
<p>I would call this format 'good enough', but not a next generation format; specifically referring to the header portion. A next generation format would move to a derivative of XML or some such. ICARTT would provide for an incremental improvement to the original NASA Ames format (more complete time representation, units formalization, additional required specific global attribute information, and using a comma to separate the data instead of white-space), and provide the least disruption to NASA legacy with Ames DEF.</p>
<p>Identifying time as the independent variable is good for time-sequenced test data. Allowing fractional-second times is good.</p> <p>The start, stop, and mid-time is overkill for many scenarios but it is optional to include all fields and thus can efficiently adapt to various time stamping needs.</p> <p>The stipulation that "All intervals longer than 1 second must be reported as Start and Stop times" seem arbitrary and unnecessarily constraining.</p> <p>Combining the header (metadata) with sensor time-series data in the same file may be handy for post-flight data reduction but is not ideal for real-time data transmittal. I.e. one would want to incrementally update sensor data and not retransmit header info. A single standard that supported both in-flight data transmittal and post-flight data archival would be desirable.</p> <p>The row-by-row time,sensor data format is not too far removed from the IWG1 format. Perhaps there could be some merger of the formats; e.g. ICARTT headers with the ability to define a separate file of IWG1 sensor data in an open-ended number of time-stamped rows.</p> <p>The requirement that missing data be denoted by -99999 fields is cumbersome. It can work reasonably well with post-flight data when you have the advantage of hind-sight. But during real-time there are times when you cannot tell the difference between gaps and data transmission delays (i.e. the data might show up but late).</p> <p>Although text-based, the format is reasonably efficient. The argument to make it more "modern" (e.g. XML) would be consistent with it being only a post-flight format. However, if it is to become an in-flight data format as well, the terse/efficient comma delimited rows of time-stamped sensor values would work well. This is another argument to extend this format to support real-time data transmittal.</p> <p>The multi-dimensional formats presume some knowledge of Ames file formats, and are less relevant to my own experience in handling test data. The given example of this shows lots of -999999 fields which is arguably inefficient. How about empty fields instead (e.g. 1,2,,4,,,7,8)?</p> <p>Other formats, such as HDF5 and NetCDF, seem to overlap much of the same need, particularly from the perspective of post-flight data archive formats. What is the motivation and advantages of the proposed ICARTT format relative to these other widely used formats?</p> <p>Based upon some recent thinking, I see two priorities in selecting a test data format: A real-time format that can also used as the post-test archive format The format should be easy to use/understand, efficient, and have broad acceptance</p> <p>On the latter broad-acceptance note, this could be either via finding an existing format or by dictating a new format. The latter is hard to do, which makes me question the value of yet-another format.</p>
<p>The file naming conventions as described implement a practice of encoding metadata into a string (up to 127 characters) used for the filename. This is advantageous for managing large numbers of ICARTT files but validation mechanisms are not discussed. File naming must be strictly enforced to be effective. Outside of the .ict suffix, I wonder if the file naming convention should be outside the scope of the file format specification. Specifically, naming conventions address the management of data files for a field campaign would generally include data that comes in multiple formats, not all airborne data sources, etc. The electronic records management strategy for any measurement campaign is a challenge that is at a higher level than the file format specification.</p>

Contributors:

This response to the RFC intended to represent assist the SPG in their assessment if the ICARTT format via collecting inputs from people associated with NASA's Airborne Science Program and the Interagency Working Group for Airborne Data and Telecommunications systems (IWGADTS).