

Review of ICARTT file format implementation and operational suitability

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. The ICARTT file format has since been widely accepted in the airborne field study community and used in recent major airborne studies sponsored by NASA, NSF, NOAA and international partners.

You are invited to review this Requests For Comment (RFC) in the context of your **implementation experience** with this data format specification and its **suitability for operational use**. You only need to answer questions applicable to you. Please send your completed review to:

spg-rfc-019@lists.nasa.gov.

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 have participated in aircraft field experiments over 20 years, participating in specifying data formats, implementing tools for reading and writing data, and using those tools for scientific analysis. I have no direct experience in implementing the ICARTT format, but I have had experience in implementing readers and writers for the Ames exchange file format which is very closely related to it. For the Ames format, my implementation was from scratch.

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 misses one important detail for implementors: what the character set or sets are acceptable for use in ICARTT files? "Plain text" is a rapidly fading concept these days. Does it mean simple ASCII? If so, then how are investigators to list themselves if they have diacritical marks, which are not included in ASCII? If not, then is the commonly-used ISO-8859-1 (Latin-1 Western European) to be used? Or the up-

and-coming UTF-8? Both are backwards-compatible with ASCII. Moreover both are used in modern operating systems, and typically the user has no idea that his or her documents are in anything other than ASCII. Will the file-checking software mentioned in the document choke if I give it a file that uses UTF-8 characters to represent some of my Chinese colleagues' names? If a Russian instrument team on a NASA aircraft submits its files using Cyrillic characters (which is “plain text” for them), is that acceptable for the archive? If not, then this document needs to say so explicitly.

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

In section 2.3.B, it is stated that delimiters (whether they be commas or the permitted-by-legacy spaces) should not be used anywhere else in the file. And specifically, “For delimiters to separate text, use underscores.” The intended meaning seems to be that header information that consists of lists must have the list elements be separated by delimiters, and that those delimiters must not be used within list items. But what is actually stated in the document is that delimiters cannot be used anywhere else in the file, not even in text comments. Thus, the examples in the document, where we encounter spaces and commas in for example the DATA_INFO field, do not conform to the format standard. This requirement must be reworded to state exactly what is meant.

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

It is unclear whether the date (and optional time) in the file name are truly the date (and time) at which the data begin. Section 2.2 seems to indicate this. But one can easily envision scenarios in which files from different instruments on the same flight end up with different dates in their file names. For example, if the aircraft takes off at 23:00 UTC on August 23, then instrument “A” might begin taking data during the ascent at 23:10 on that date, while instrument “B” waits until the aircraft reaches altitude and begins taking data at 00:30 UTC on August 23. Thus, instrument A's file for that flight would have a name like “A_DC8_20090823_RF.ict”, and instrument B's file for that flight would have a name like “B_DC8_20090824_RD.ict”. This could get confusing in an archive, especially after a sequence of flights occurring on back-to-back days. Such ambiguous circumstances can easily occur for campaigns conducted near the international dateline. Remember, the point of format standards and file naming conventions is to lessen confusion, not increase it.

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?*)

The ICARTT format is suitable for exchanging data between investigators participating in aircraft field experiments. See below for pros and cons.

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

For the kinds of data sets for which the ICARTT format is intended, it presents no significant implementation challenges. Certainly, data files that are going to be accessed frequently should be rewritten into a more efficient binary form. But for data exchange among a wide variety of investigators, a text format such as this is just the right kind of thing.

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?

I have no plans to use the ICARTT format, but if I were participating in a tropospheric aircraft mission that used the format, I would have no objection to using it.

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

I have not chosen to use the ICARTT format. The Ames format suffices for our needs and is the format used in the missions we have been participating in.

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

The ICARTT format, despite certain drawbacks, would meet my requirements for exchanging data with other investigators, especially those who are familiar with my field. It would not be suitable for high-performance applications or for anything more than occasional data access, because of the inefficiency of the text format.

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.

My users do not use the ICARTT format, but they have used the Ames format, which is closely related. Based on those experiences, I would say that users would have no more trouble with this format than with any other standardized exchange file format.

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

Being text-based, the ICARTT format is not efficient in its use of storage space, and any reader or writer is going to need to convert between character strings and binary numeric forms, which is also not efficient. But again, efficiency in day-to-day use is not a key issue here; the ICARTT format is for data exchange. So long as it is used for that, it presents no particular operational challenges or limitations.

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?*)

The ICARTT format is flexible and suitable for representing the 1-D, 2-D and 3-D data sets that are likely to occur in the setting of an aircraft field experiment. Both scalar quantities and their uncertainties are represented. Vector or tensor quantities must be represented in terms of their individual components, which is a little awkward, but not unreasonably so.

For uses outside of the aircraft mission environment, this format seems less useful. For example, the specification document concentrates on the time variability of the measurements, so that investigators are generally required to specify a start time, a stop time, and a mid-point time for each measurement. This makes sense for a probe moving through the atmosphere at hundreds of meters per second. But the same arguments this document makes for such detailed descriptions of time sampling, can also be made for spatial sampling. There are contexts within which an investigator should specify the starting spatial position, ending position, and mid-point of a static measurement, and where temporal variations are negligible.

The structure of the file should be easy enough for investigators from various fields to grasp. However, the lack of a standard labeling convention for the identification of measured quantities, units, etc., is a drawback for interdisciplinary studies. Will an oceanographer looking at this file have any idea what “Klet” means? Will a tropospheric investigator who examines the archives 30 years from now recognize every label used? Of course, this is more of a failing of the current state of metadata in earth sciences, not a fault of this ICARTT format.

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.*)

None planned.

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

None planned. There will probably be at most no more than a handful of users in my immediate organization. The Ames format is more widely used.

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

The ICARTT format is clearly and strongly dependent on the old NASA Ames exchange file format. From the look of things, the ICARTT mission planners apparently took the Ames format, relaxed some of its requirements, and added some new conventions (*e.g.*, the keyword-value pairs in the “Normal Comments” section). This in itself neither

enhances nor detracts from the format's appropriateness for aircraft field experiment data. However, some of the differences do not seem very well thought out. For example, the ICARTT format relaxes the recommendation of the original Ames format specification that data values be scaled to integers, and consequently ICARTT recommends that the scaling factors listed in the header should all be 1.0. But the scale factor line is retained in the file header, instead of moving it into the normal comments section as an optional keyword alongside such keywords as LLOD_VALUE and ULOD_VALUE.) A desire for backwards compatibility with the Ames format would explain this, but ICARTT sacrifices backwards compatibility by recommending the use of commas instead of spaces as delimiters. The format has a number of oddities like this, where an element of the Ames format has been rendered obsolete or unnecessary, but has nonetheless been retained as vestigial elements in ICARTT.

As a participant in field experiments, I can easily live with such quirks in this format, but in my opinion the specifications document itself really does need a lot more work. It reads like an informal writeup intended for a group of colleagues working together, with similar backgrounds and attitudes. The use of phrases like “your mission” and “your project” betray the document's origins. There is a tone of “we're all colleagues here, and we know what we mean when we use such-and-such a term.” That may be fine for colleagues in a field mission, but it is not sufficient for a formal standards document, which should not presuppose a common background. There is for example no definition of what is meant by “auxiliary variable” and how it differs from “primary variables”--- terms borrowed from the Ames format without definition or explanation here. The authors make a passing comment about the dense, almost mathematical language used in the Ames format specification (“What this means in English is...”, on page 10), but in a standards document it is just this kind of precision of definitions and usage that is required. Section 2.5 attempts to expand the format into other data types that do not really fit the aircraft instrument paradigm, and the results sounds too much like “We don't know precisely how to handle these data sets, but if you investigators come up with something reasonable we'll say it conforms to the standard.” Again, that may work well during a field experiment but not in a format specification standard. Informal understandings simply do not withstand the nasty little edge cases that crop up.

In summary, I would approve of including the ICARTT data format as a NASA Earth Science Data Systems Standard, because it is useful for data exchange among colleagues, and because many valuable data sets have been archived using it. It is not the best thought-out format that there is, but it does no actual harm. It might be adopted as one of the heritage formats in SPG. However, I would urge that the format specification document be completely rewritten to be much more rigorous and unambiguous. The informal, investigator-friendly descriptions and examples could be included in an appendix.

(And by the way, the proper and up-to-date reference for the Ames format is http://espoarchive.nasa.gov/archive/docs/formatspec_2_0.html, not <http://cloud1.arc.nasa.gov/solve/archiv/archive.tutorial.html> as given in the document.)