ESDS-RFC-018
Category: Technical Notes
Updates/Obsoletes: None

Craig, Wagner, Cuddy, Vuu, Lewicki, Ilg, Larson

March 2009

Creating File Format Guidelines: The Aura Experience

Status of this Memo

This is Version 1.0 of the technical note on developing a file format for instruments with common data products.

Its distribution is unlimited.

Change Explanation

None

Copyright Notice

Abstract

This technical note describes the process that the Aura teams used to create file format guidelines for their data products.

Table of Contents

1.0 INTRODUCTION

When a common data file format is adopted and used by more than one team, customized software to read and manipulate individual data sets is minimized. Code reuse is maximized and less time is spent on the details of accessing the data. Documentation is also reduced. These principles apply both to data users, as well as data producers.

1.1 MOTIVATION BEHIND THIS TECHNICAL NOTE

This technical note describes the creation of a common file format developed and used by the individual teams working on the four instruments on NASA’s Aura satellite (HIRDLS, MLS, OMI and TES). Each of these teams was independent and there was no mandate that they all use a common file format. The decision to do so and the implementation of it was a grassroots effort that was accepted by all of the PIs and instrument teams’ leading scientists. This document describes the process used in developing the guidelines and the keys to its success. A brief summary of the guidelines’ features will be described, but a thorough description of the guidelines is contained in the document “A File Format for Satellite Atmospheric Chemistry Data” [1].

Early on in the Aura program, the Aura teams realized that common data and file formats would greatly facilitate the sharing of data. The teams agreed to use the HDF5/HDF-EOS5 data format. It was also necessary to concur upon specific details within the file itself. The teams agreed upon the names, data types and dimension order of fields. There was also agreement regarding the file-, group- and field-level attributes to include in each product file. A file-naming convention was also adopted.

Future NASA missions can use this technical note to learn from the process that the Aura instrument teams went through to develop their own set of guidelines. The process can be adopted and modified as needed.

2.0 BACKGROUND INTO THE GUIDELINES CREATION PROCESS

Throughout the entire process, the standard by which each guideline was weighed was whether it helped the end user to develop one universal reader to read the primary data within the Aura teams’ data files. Items which did not affect the reading of the data were not standardized. Compression is an example of an area where standardization did not take place. During reading of HDF files, uncompression is handled by the library itself, if it is needed, and the reading software does not need any modification to handle compressed versus uncompressed data. Another area which fell outside of the standardization process was instrument specific data. Instrument teams were free to add any additional fields that they required. Due to the direct access method of reading data within HDF files, additional information can be added to a file and it does not impact a reader, unless that data is required to be read.

2.1 HOW THE PROCESS BEGAN

In the winter of 1998-99 the Aura Project Scientist expressed strong desire for the instrument teams to form a Data Systems Working Group (DSWG) that would be a forum to aid in the exchange of information and sharing of common data issues. The TES Ground System Manager agreed to serve as the DSWG lead. It is possible that some of the guidelines work undertaken by the DSWG would have been done as a grass-roots effort. However, it might not have been as successful without the sponsorship of the DSWG and the Project Scientist.

At the time that the instrument teams were starting to think about their standard data products, an Aura instrument team scientist promoted the idea of standardizing data field names and units across the datasets produced by the four Aura instruments. He explored the concept with the other instrument teams via discussions at Aura DSWG meetings and via emails with DSWG members. After nurturing the idea of standardization among the instrument teams for a year, other demands on his time prevented him from continuing the effort and another Aura software engineer volunteered to take over leading the process.

Instrument teams were receptive to creating guidelines as they had heritage from the UARS platform (where a standard interface was used for all instruments) and the AMES format (where standardization was achieved with a number of different interfaces, each requiring a specialized reader). In contrast, there had not been any attempt at standardizing data file formats for either the Terra or Aqua platforms. It was apparent that if any standardization would be achieved, it would be directed by the instrument teams themselves. As the individuals involved in the discussions were instrument team scientists and software developers, each of them had a vested interest in making the standardization process work; their own reading of other teams' data after launch would be greatly simplified in addition to making it easier for their end users.

2.2 PROCESS FOR DEVELOPING AND REFINING THE GUIDELINES

When the idea of standardizing the data field names and units was well received, it was proposed that the files be standardized as much as possible, to eliminate the need for custom readers. At that time a strawman draft was developed, documenting the guidelines for standardizing a Level 2 file. It was circulated to the DSWG members on each instrument team. In some cases, the document was forwarded by these members to others within their teams. These people then discussed the strawman with members of their instrument teams and responded back with comments. When instrument team members responded back, their comments were discussed and incorporated into the document and they were asked if they wanted to be listed as a coauthor on the paper. The coauthors included data managers, instrument team scientists and software engineers. More than one coauthor from each instrument team was allowed, and most teams had more than one. As items were not voted upon, but rather complete agreement from all of the coauthors was required, the number of coauthors per team was not constrained by potential voting inequities. When coauthors moved on to other projects, replacement members were selected by the instrument teams and the authorship list was updated accordingly.

An email list supporting the guidelines was also established and anyone who requested to be on it was included. This list was maintained by the guidelines leader. All discussions were initiated via the email list and anyone was free to express their views. The entire process was as open as possible, so that anyone who had an interest could respond as they saw fit.

At times, discussions became so complicated that consensus could not be reached via email. At those times, either a conference call was set up or the discussion was slated for a splinter meeting at the next Aura Science Team meeting. At either type of meeting, at least one coauthor from each instrument team was required to be in attendance. Anyone else who was interested in attending was also invited. The decision on whether to hold a conference call or to wait for the science team meeting was made based on how soon the next Science Team meeting was to be held and whether there was a need to reach a resolution quickly. Discussions were held and coauthors then returned to their instrument teams to gain final approval for any new guidelines.

When disagreements occurred, each team presented their reasoning for a particular viewpoint and discussion was held. If there was still disagreement, then possible "creative" solutions were brainstormed. Two such instances will be described. In the first case, the issue was how to store the vertical pressure coordinate values. Two teams felt strongly that the values should be stored as a global attribute and the other two teams felt equally strongly that they should be stored in a geolocation field. After discussion, both groups held firm to their views. Since these fields were small, the solution of including them in two places was put forth and the teams agreed to do that. A second case was how to label the data fields. For heritage reasons, MLS wanted to label their data fields with the generic name "L2gpValue" and the species name would be the name of the swath. The other three instrument teams wanted to label the data field with the species name. A compromise was achieved when MLS agreed to use the HDF5 alias feature and label their data fields with both L2gpValue and the species name.

There were times when agreement could not be found, and those areas were documented as well. One such instance is the names of the swaths. HIRDLS wanted to have all of their data in a single file with a generic swath name, while MLS wanted to have a separate file for each species and have the species identified by the swath name. Because of this fundamental difference and after much discussion, it was decided that there would not be agreement on this issue, and the swath name was documented as instrument dependent. In this particular case, as HDF-EOS allows one to query the swath names, an end user can write a code that still works with both types of files. The code is slightly more complicated than it would have been if agreement had been reached with the swath names.

Throughout the process, it was recognized that the development of guidelines was a completely voluntary effort.

If a team felt that a proposed standardization did not work for them at any time, they could decline. The goal remained to standardize as much as possible, however.

2.3 DOCUMENTING THE GUIDELINES

An overview of the key events in the development and documentation of the Aura Guidelines are highlighted in Table 1. It is important to note that the guidelines were not defined all at once, but were negotiated before the products they were supporting were completely developed. The guidelines evolved as teams started defining each data product.

Table 1
Development Events

DATE	KEY EVENTS
Late 1998-early 1999	DSWG established. Aura Project Scientist saw need for a DSWG and arranged for an Aura instrument team to start it up and lead it.
02/02/1999	Initial emails sent out to DSWG membership exploring the willingness to standardize field names and units
04/12/1999	**First DSWG meeting**. Informal discussion about standardizing field names and units. Logistics of sharing information between the teams discussed and implemented.
11/23/1999	Naming and Units strawman circulated between two instrument teams who had expressed interest
03/29/2000	Presentation at DSWG to gain support for standardization. Transfer of leadership.
05/01/2000	Level 2 standard product guidelines strawman sent out to all four teams
05/2000-04/2001	Numerous emails, a conference call and a DSWG meeting where guidelines were discussed and negotiated.
04/02/2001	V1.0 Released - ***Level 2 standard product guidelines***
08/27/2001	V1.1Released – Compromise on two locations for Pressure information and modifications to suggested attributes
04/2002	DSWG splinter meeting held and discussed mandatory attributes and file naming conventions
10/03/2002	V1.2 Released – ***Attributes now mandatory***. Draft of file naming conventions
10/16/2003	V1.3 Released – ***Elevated file naming conventions to accepted***. Added data type to field guidelines (missed in initial guidelines, documented after the fact). Added a couple more mandatory attributes
07/15/2004	**AURA Launch**
12/16/2004	V1.4 Released – Updates to only instrument specific dimensions and fields
09/09/2005	V1.5 Released – Instrument specific changes and editorial revisions.
10/2005-10/2006	Discussion on adding Level 3 Grid and Zonal Mean files to guidelines. Numerous emails, conference calls and splinter meeting at Aura Science Team Meeting
10/30/2006	V2 Released – ***Level 3 standard product guidelines added***

The initial strawman draft document was written by the guidelines leader in consultation with the instrument team scientist who conceived of the idea. It used key points from a white paper that the instrument scientist had written before resigning. Emails were exchanged during the handoff of leadership between the two people, and a short, six page document was generated that captured the understanding of both. At that time, the document was then circulated to all four teams for comments, additions and further refinement.

The guidelines document evolved over time. While some details were captured early on, as teams began implementing their data files other areas were added that were not initially considered. With each new guideline or modification, either a discussion was held on it or a draft document was composed and circulated for comment. After significant changes were made to the document, approval was required from every coauthor and then that version of the document was frozen and released to the public.

The official document was always stored at the guidelines leader's facility. When minor changes were indicated by a coauthor, the proposed changes were emailed to the leader and the leader incorporated them into the document. When the changes were more major, the document was handed off to the coauthor who wished to make the changes and then sent back to the leader when the editing was completed.

The guidelines started simply and grew as needed. The initial goal was to document what was required to standardize the Level 2 data files. The first version of the document contained requirements that allowed instrument teams to begin development. During development, as questions were raised or additional details were needed, they were addressed via email, conference calls or in a splinter meeting at an Aura Science Team meeting. With the success of the standardizing of the Aura Level 2 products, a request was made to standardize our gridded and zonal averaged Level 3 products as well as standardizing our file naming. Each of these processes again started with a strawman document being written by the instrument team who wanted to lead the effort and it was then refined.

Versions of the guidelines were baselined and officially released to the general community as changes to the requirements warranted. Before each public release, approval was required from every coauthor. At times, after the leader sent out a request for approval, follow-up reminders were needed before a response was received from everyone. One possible remedy considered was to take non-response as a sign of acquiescence. A conscious decision was made to require explicit agreement from every coauthor to settle an issue. Releases were postponed until everyone had time to adequately review and sign off on the document. As the guidelines creation process was designed to capture the agreement between the teams, it would have been detrimental to assume agreement when there might not have been.

The first official public release of the guidelines was put onto a web server and publicized to the community at large via the Aura DSWG email list and by announcements at the Aura Science Team meetings. The guidelines were also publicized at several HDF/HDF-EOS workshops. The web address for the currently released version of the guidelines has remained constant throughout the years, with new versions overwriting the previous ones. A second location on the web site was established for the current working copy of the guidelines. This site was not publicly advertised, but was where any coauthor could go to view the current revisions. This name also remained constant throughout the development of the guidelines. By handling the document versions in this way, the current released and working versions were always accessible to anyone on the team and were in a known location. Also because of the stability of the web location, the location was able to be announced for use by the general public.

Development of the first version of the guidelines maintained its momentum simply due to the fact everyone involved in the process saw the benefits and was eager to get it solidified before they started developing their standard data products. The need of instrument teams to develop their standard data products helped to keep the discussions going. Subsequent versions of the guidelines were created when any of the coauthors proposed changes or additional guidelines. As the teams developed their data products further, they discovered additional areas to standardize.

2.4 VALIDATING THE GUIDELINES AND VERIFYING FILES

Prior to the release of V1.0, the guidelines document had been circulated within the instrument teams and with a representative from the Goddard Data and Information Services Center (DISC). Several people on the instrument teams as well as the Goddard DISC representative were familiar with developing files in HDF and HDF-EOS and lent their expertise to refining the initial draft guidelines document. V1.0 of the document described the Level 2 data files sufficiently for development of these data products to proceed.

As instrument teams developed their data files, they shared these files with the other Aura instrument teams. The teams then verified that the data file structure matched other teams' structures. Since the coauthors were not confident that the guidelines were defined adequately, this was an important part of the process. The sharing of data files validated the guidelines and verified the structure of the data files themselves. It was important to verify each type of data file as it became available, as each file was subject to implementation errors. All areas of non-uniformity turned out to be due to implementation errors rather than shortcomings of the guidelines. Errors were found and fixed in both the Level 2 and Level 3 files as they were developed.

Aura also had the use of a validation tool that was developed specifically to check Aura Level 2 data files for compliance. This tool was developed by a separate organization and relied on XML to define the structure of a specific data file. It was run on a sample data file to validate its structure and was useful for checking details within the file. While this tool was used at one point in the development process, its use did not catch on. There were several reasons which when combined led to it not being used further. The timing of the release of the tool was slightly after the teams had already performed most of the validating of their initial file formats. As the tool relied on a separate XML description file, there was an equal amount of time spent finding errors in the description file as well as finding errors in the data files themselves. In the long run, teams found it just as easy to validate the file formats by hand as opposed to using the tool for the incremental changes in future data file releases.

One glaring omission to the early versions of the document was the data type of fields. By the time this was discovered, teams had already developed their data files, but fortunately all teams had chosen to use the same data types. This guideline was actually documented after the initial data file development was completed.

3.0 GUIDELINES FEATURES

The first step in producing a set of guidelines is to identify the community of data producers and the fields which overlap between them. These overlapping fields are the only ones which need to be standardized. If a self-describing file format is used, individual data producers can include additional fields which are unique to their data products and the details need not be included in the guidelines, though they may be. Guidelines should be created to describe the individual data "boxes" and not necessarily address the contents of the boxes.

The use of a self-describing, binary format such as HDF, HDF-EOS or netCDF along with thoughtful selection of names and attributes, allows end users to understand the contents of a data file without further documentation. Another benefit of these data formats is that fields are available via direct access by name. This allows instrument teams to add additional fields to their data files, while software which is written to read the standardized fields will work without modification. End users can also efficiently read data from files using these formats, as they only need to read the fields in which they are interested.

3.1 ITEMS TO STANDARDIZE WITHIN A SELF-DESCRIBING DATA FORMAT

Every item not spelled out is subject to interpretation and alternate implementation. If a self-describing data format such as HDF, HDF-EOS or netCDF is being used, then standardization should include:

- Names of fields (including capitalization and spacing)
- Names and ordering of dimensions for each field
- Data types and sizes for each field (for instance integer, 32 bit)
- Attributes for each field and their types and definitions.

While the above list identifies the boxes with no constraint on their contents, additional benefits can be realized by standardizing the following contents as well:

- Units for each field
- Coordinates - The actual values of any fields which describe the location of data (such as latitudes if a gridded product, pressure levels, etc.)

4.0 LESSONS LEARNED

Highlights of the key points that Aura discovered throughout their development process follow:

- **Make sure every team is committed to the process at the outset, as there is a significant amount of time and compromise involved.** In Aura's case, the guidelines were nurtured at Aura Data System Working Group (DSWG) meetings where software developers and scientists from all of the instrument teams were in attendance. Representatives there indicated their willingness and desire to create a set of guidelines and were responsible for talking to their individual instrument teams to procure buy-in.
- **Acceptance of the effort needs to be at all levels of management.** While the members of the DSWG agreed to the creation of a set of guidelines, it was also crucial that the PIs and leading scientists endorsed the project.
- **Every team must have at least one dedicated author and representative**. It is crucial that every team be represented by at least one individual who is willing to spend the time and effort to work on the guidelines. This individual needs to communicate both with the guidelines group as well as the team's management to achieve agreement. Representatives should have an understanding of the basic file format and storage needs for their instrument. Also, if these individuals will be utilizing the data in the future, their willingness to find compromise will be enhanced.
- **The data fields which are in common between two or more instrument teams are the only ones which need to be standardized.** Since a generic reader was the goal, fields which were not shared with another instrument did not need to be discussed and standardized by the group.
- **Create a strawman draft**. One individual needs to take the lead to create a strawman draft. The main purpose of this draft document is to start the dialog and give a concrete framework which can be modified by the other teams. This draft should be treated as a first attempt and should be allowed to be extensively modified as needed. It should reflect the current agreement between the teams, however limited it may be at the time. For Aura, this process happened several times: once at the project's inception and at a couple other times when major features were added to the document. Each time, a different individual took the lead on this. The initial strawman document started out being only six pages long. Over the years it grew to over 50 pages.
- **Modify the document to incorporate every team's input**. The document needs to be "owned" by every team and not by one individual team. In Aura's case, the HIRDLS team took responsibility for maintaining the actual document, but it was routinely passed to other teams when an individual indicated the desire to make extensive modifications. When modifications were less extensive,

changes were emailed to the document maintainer and incorporated in that way.

- **Have a forum for gathering the team members interested in data issues together**. In Aura's case this was done with the DSWG. While Aura science team meetings typically gathered scientists together, the DSWG also included software engineers and data managers. Having a forum which included these team members was crucial in the development of the guidelines.
- **Communication is essential**.
  - At every Aura DSWG (which were held several times a year before launch), items which needed agreement were discussed. When extensive discussion was required, splinter meetings were held with at least one representative from each instrument team in attendance.
  - In between Aura DSWG meetings, email was the primary tool for discussion and reaching consensus. At times, it became apparent that a more involved discussion needed to take place and conference calls were used or the issue was tabled until the next DSWG meeting.
  - The email list contained both named and unlisted authors. Anyone who wished to be included on the email list was added to it. All discussions were sent using the general email list. When an agreement needed to be reached, everyone was entitled to respond, but authors whose names were on the document were required to respond.
  - Controversial items were discussed among the members and were then taken to their individual instrument teams for discussion and approval/disapproval. The results of these instrument team discussions were then reported back to the group.
  - Every major version release of the document was agreed upon by all of the named authors.
- **Allow flexibility**. Items which did not overlap with the other teams were not under constraint and teams were allowed to add features to their data files. The only items which the document standardized were the data fields which overlapped between instrument teams.
- **Communicate early.** Start creating guidelines as teams are beginning to identify their data products but before they are developing their files. It is much easier to reach consensus before individual team's decisions on how to store data have been made. Alternatively, if this process is attempted too soon, teams may not be as motivated to work on the guidelines nor as able to adequately define them. In Aura's case, from the time that the standardization was initially proposed until it was actively worked on was about a year. Active development started when the teams were beginning to think about their data file contents.
- **Be willing to compromise.** The benefit for developing a set of guidelines should outweigh the loss of any one team's "perfect data format". The inclusion of pressure in two locations within the Aura files is one such example of compromise.
- **Look for creative solutions to attain compromise.** The solution of using aliases to allow MLS to keep the name L2gpValue while also adhering to the guidelines is an example of this.
- **Appoint a dedicated group leader.** The leader must follow through on all issues, and leave no issues unresolved. This is most important when consensus is difficult to attain or when people are busy and are slow to respond.
- **Be willing to commit to the process for the long haul.** The process takes time and energy to see through to fruition. From Aura's experience, the discussion on guidelines started at the time that the instrument teams were just starting to develop their data products. The bulk of the work was done before launch (during a period of over 5 years) and additional modifications were made even after launch.
- **Document needs to be detailed.** Every item not spelled out completely is open to interpretation, leading to ambiguity and possible alternate implementation.
- **Use of a direct access, self-describing data storage library (like HDF and HDF-EOS) eases the**

standardization process. Because additional data within a direct access file does not affect the reading of the standardized data, instrument teams were free to add instrument specific information to their files. This simplified the standardization process as the entire file structure did not need to be agreed upon, but only the areas which overlapped between teams.

- **Exchange data sets early on.** Problems with individual interpretations may surface when data files are exchanged. Implementation mistakes can also be found when files are shared with other teams. The exercise can improve the confidence and commitment of all parties involved by demonstrating that the standard is on the right track, and allowing improvements to be made.
- **The guidelines were a voluntary effort.** At no time did the guidelines ever become a mandatory effort. This allowed each instrument team the ability to pull out at any time and, because of this, there was not a feeling of coercion.

5.0 CONCLUSIONS

Aura took the initiative and developed their own set of guidelines to constrain their data files. A balance was achieved between presenting the data which was in common in a standardized way and also allowing flexibility for additional instrument specific information within each data product. Because of this effort, generic readers could be written to read the standardized data from any Aura instrument. Future instruments can take note of the procedures that Aura used and develop their own guidelines for their instruments or use the Aura Guidelines as they stand.

6.0 INFORMATIVE REFERENCES

[1] Craig, C., P. Veefkind, P. Leonard, P. Wagner, C. Vuu, D. Shepard (May 2008), A File Format for Satellite Atmospheric Chemistry Data Based On Aura File Format Guidelines, http://www.esdswg.org/spg/rfc/esds-rfc-009/ESDS-RFC-009.pdf. The same document is also available on the science team web site http://www.eos.ucar.edu/hirdls/HDFEOS_Aura_File_Format_Guidelines.pdf.

[2] Details on HDF5 can be found by visiting the web site: http://hdfgroup.com/

[3] Details on HDF-EOS5 can be found by visiting the web site: http://www.hdfeos.org/

7.0 AUTHORS' ADDRESSES

Cheryl Craig
cacraig@ucar.edu
On HIRDLS instrument team
National Center for Atmospheric Research
1850 Table Mesa Dr.
Boulder, CO 80303 USA

Paul Wagner
Paul.A.Wagner@jpl.nasa.gov
On MLS instrument team
Mail Stop 183-701, Jet Propulsion Laboratory
4800 Oak Grove Dr.
Pasadena, CA 91109 USA

David Cuddy
David.Cuddy@jpl.nasa.gov
On MLS instrument team
Mail Stop 183-701, Jet Propulsion Laboratory
4800 Oak Grove Dr.
Pasadena, CA 91109 USA

Christina Vuu
cvuu@mls.jpl.nasa.gov
On MLS instrument team
Raytheon Information Solutions
299 N. Euclid Ave, Suite 500
Pasadena, CA 91733 USA

Scott Lewicki
Scott.Lewicki@jpl.nasa.gov
Formerly on TES instrument team
Mail Stop 301-225, Jet Propulsion Laboratory
4800 Oak Grove Dr.
Pasadena, CA 91109 USA

Doug Ilg
dilg@sfa.com
Formerly on OMI instrument team
Global Strategies Group (North America) Inc.
2200 Defense Highway, Suite 405
Crofton, MD 21114 USA

Steve Larson
Steven.A.Larson@jpl.nasa.gov
Formerly on TES instrument team
Mail Stop 301-490, Jet Propulsion Laboratory
4800 Oak Grove Dr.
Pasadena, CA 91109 USA

8.0 APPENDIX - GLOSSARY OF ACRONYMS

Acronym	Description
DISC	Data and Information Services Center
DSWG	Data System Working Group
HDF	Hierarchical Data Format
HDF5	Hierarchical Data Format Version 5.X
HDF-EOS	Hierarchical Data Format for the EOS mission
HDF-EOS5	Hierarchical Data Format for the EOS mission Version 5.X
HIRDLS	High Resolution Dynamics Limb Sounder
MLS	Microwave Limb Sounder
NASA	National Aeronautics and Space Administration
netCDF	Network Common Data Form
OMI	Ozone Monitoring Instrument
TES	Tropospheric Emission Spectrometer
XML	Extensible Markup Language