Toolsets for Airborne Data (TAD): Improving Machine Readability for ICARTT Data Files

Amanda Benson Early[1], Aubrey Beach[2], Emily Northup[1], Dali Wang[4], John Kusterer[3], Brandi Quam[3], Gao Chen[3]

1. Science Systems and Applications, Inc., Hampton, VA 2. Booz Allen Hamilton, Inc., Norfolk, VA 3. Atmospheric Science Data Center, NASA Langley Research Center, Hampton VA

4. Department of Physics, Computer Science and Engineering, *Christopher Newport University, Newport News, VA*

Atmospheric Science Data Center

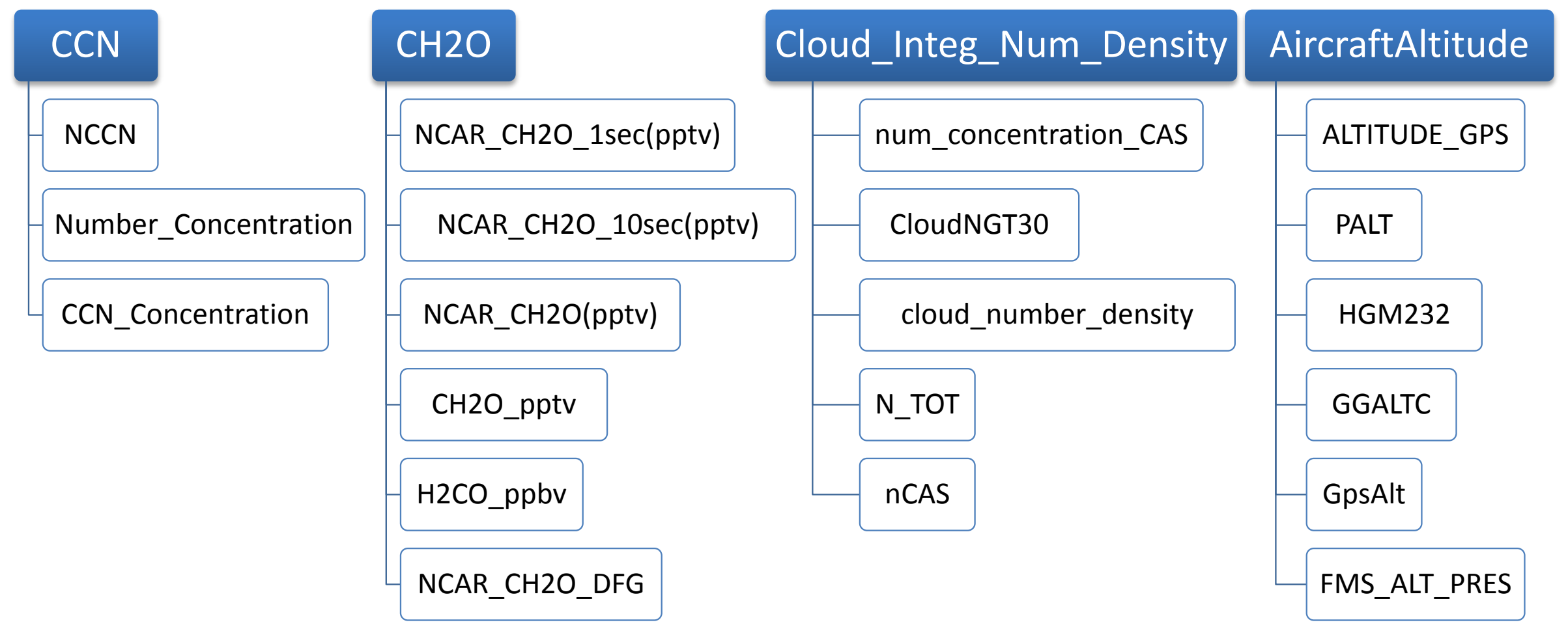
ASDC Introduction

The Atmospheric Science Data Center (ASDC) at NASA Langley Research Center is responsible for the ingest, archive, and distribution of NASA Earth Science data in the areas of radiation budget, clouds, aerosols, and tropospheric chemistry. The ASDC specializes in atmospheric data that is important to understanding the causes and processes of global climate change and the consequences of human activities on the climate. The ASDC currently supports more than 44 projects and has over 1,700 archived data sets, which increase daily. ASDC customers include scientists, researchers, federal, state, and local governments, academia, industry, and application users, the remote sensing community, and the general public.

ICARTT Machine Readability Issues

Variable names are not standardized

- Different instrument Principal Investigators (PI) may name the same variable differently in a mission
- Different names exist for the same variable in different missions



Various names used for the same variable across different missions.

Date/time recording is inconsistent

- No requirements for naming time variables
- Not always a simple way of determining what the variable is actually measuring (start, stop, mid)

File structure varies greatly

```
52 1001
Collins, Don
Texas A&M University
Tandem Differential Mobility Analyzer
from the C130
NCAR MILAGRO Mission 2006
1 1
2006 03 04 2007 02 27
0
Start_UTC, sec
14
1 1 1 1 1 1 1 1 1 1 1 1 1 1
-9999 -9999 -9999 -9999 -9999 -
9999 -9999 -9999 -9999
End_UTC, sec
Mid_UTC, sec
Latitude, fractional degrees
Longitude, fractional degrees
```

```
41,1001
Anderson, Bruce E.
NASA Langley
Black Carbon Number Density
Measurements with a SP-2
DISCOVER-AQ
1,1
2014,07,20,2015,09,09
0
UTC_start,Secs after midnight,Time
of acquisition
3
1,1,1
-9999,-9999,-9999
UTC_end, Secs after midnight,Time
of acquisition
UTC_mid,Secs after midnight,Time
```

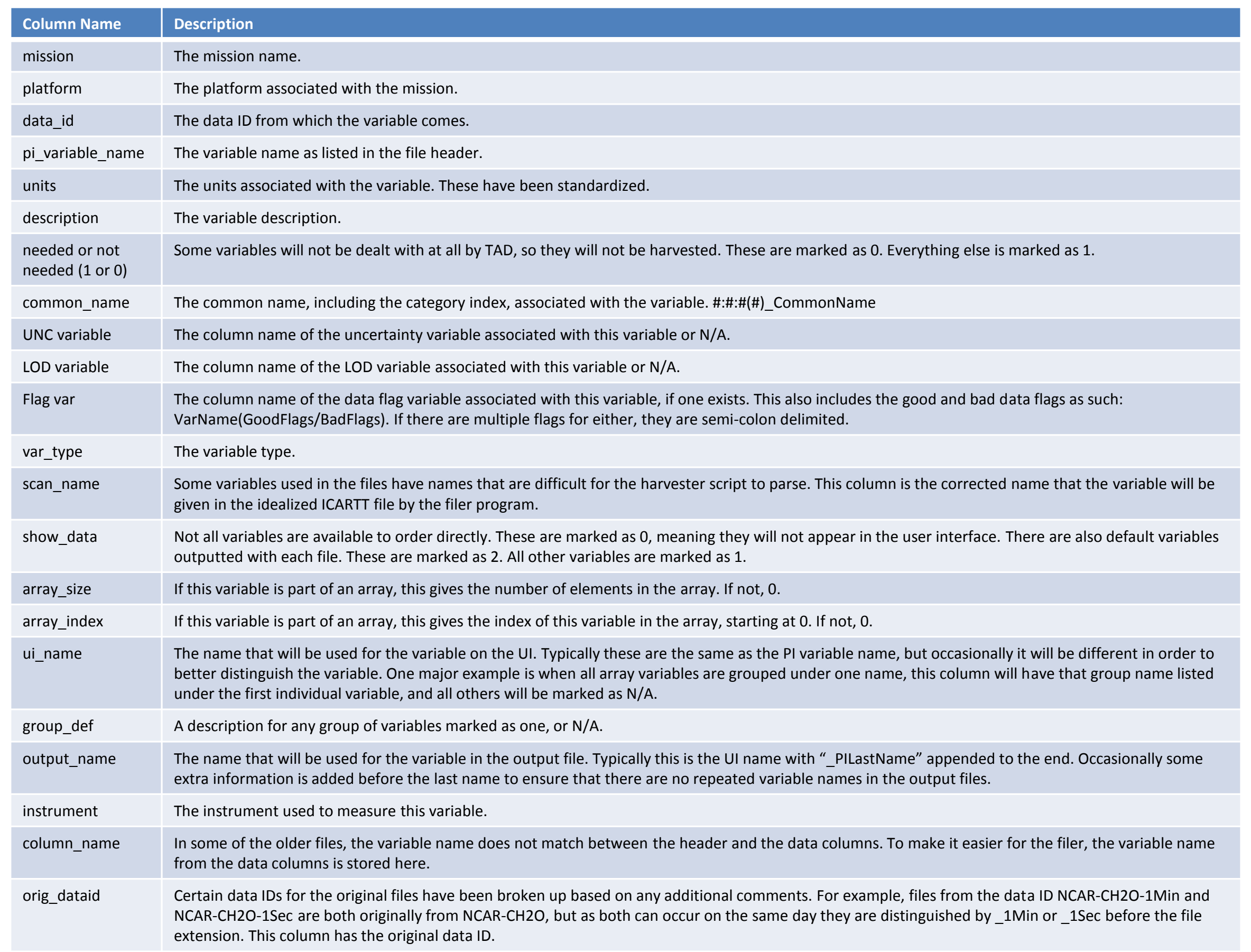
Airborne Tropospheric Chemistry Studies

NASA conducts airborne tropospheric chemistry studies, and has for over three decades. These field campaigns generate a great wealth of observations, including a wide range of the trace gases and aerosol properties. Even though the spatial and temporal coverage is limited, the aircraft data offer high resolution and comprehensive simultaneous coverage of many variables, e.g. ozone precursors, intermediate photochemical species, and photochemical products. The recent NASA Earth Venture Program has generated an unprecedented amount of aircraft observations in terms of the sheer number of measurements and data volume. The ASDC Toolsets for Airborne Data (TAD) design meets the user community needs for aircraft data for scientific research on climate change and air quality issues.

Working with ICARTT

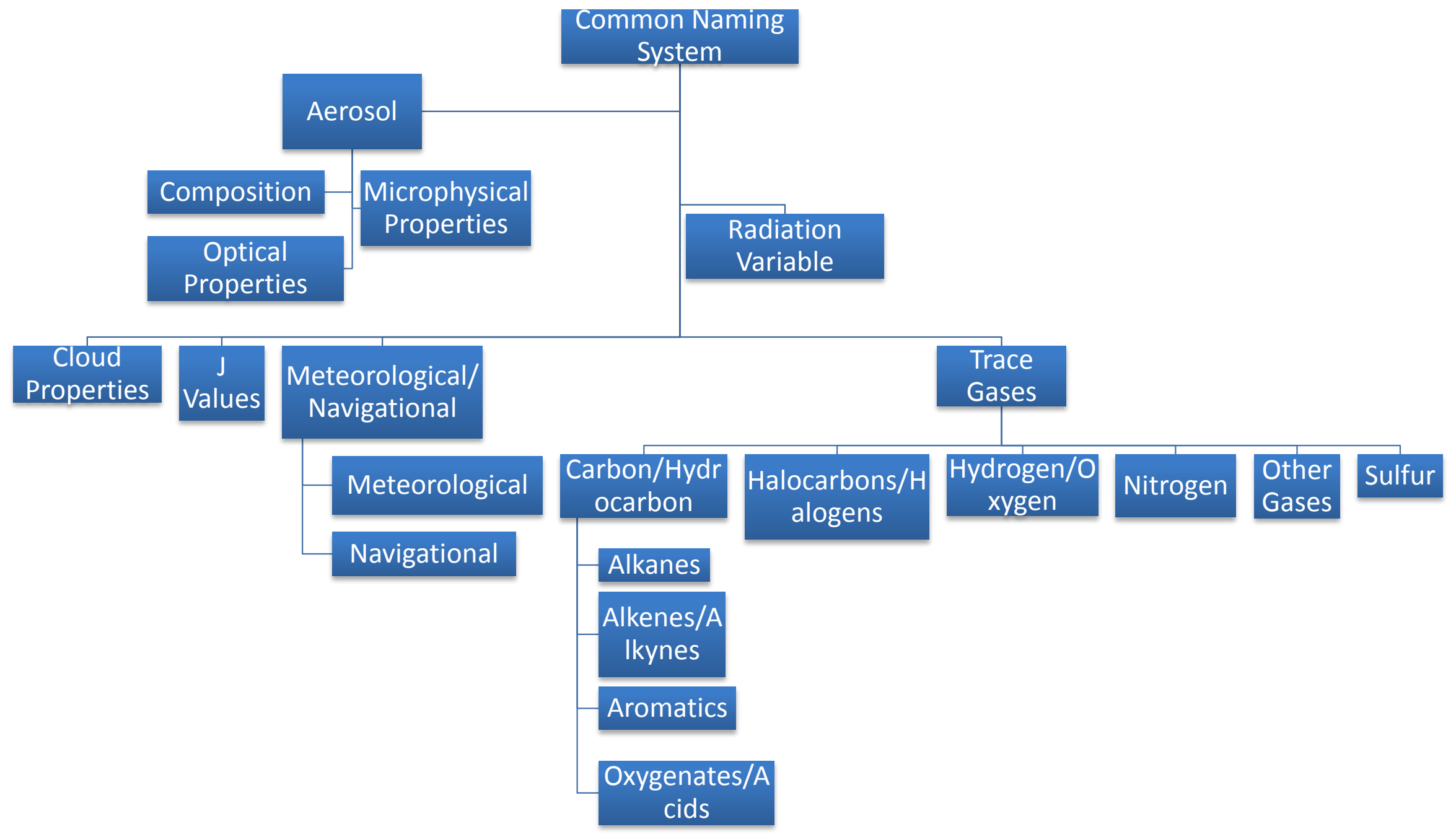
To compensate for the ICARTT file issues, files are first converted to an "idealized" format. The metadata is then stored in a comprehensive postgreSQL database. This is a three step process.

- **Step 1**. Map all metadata to an "extended map". Includes mapping each PI variable name to a common name.



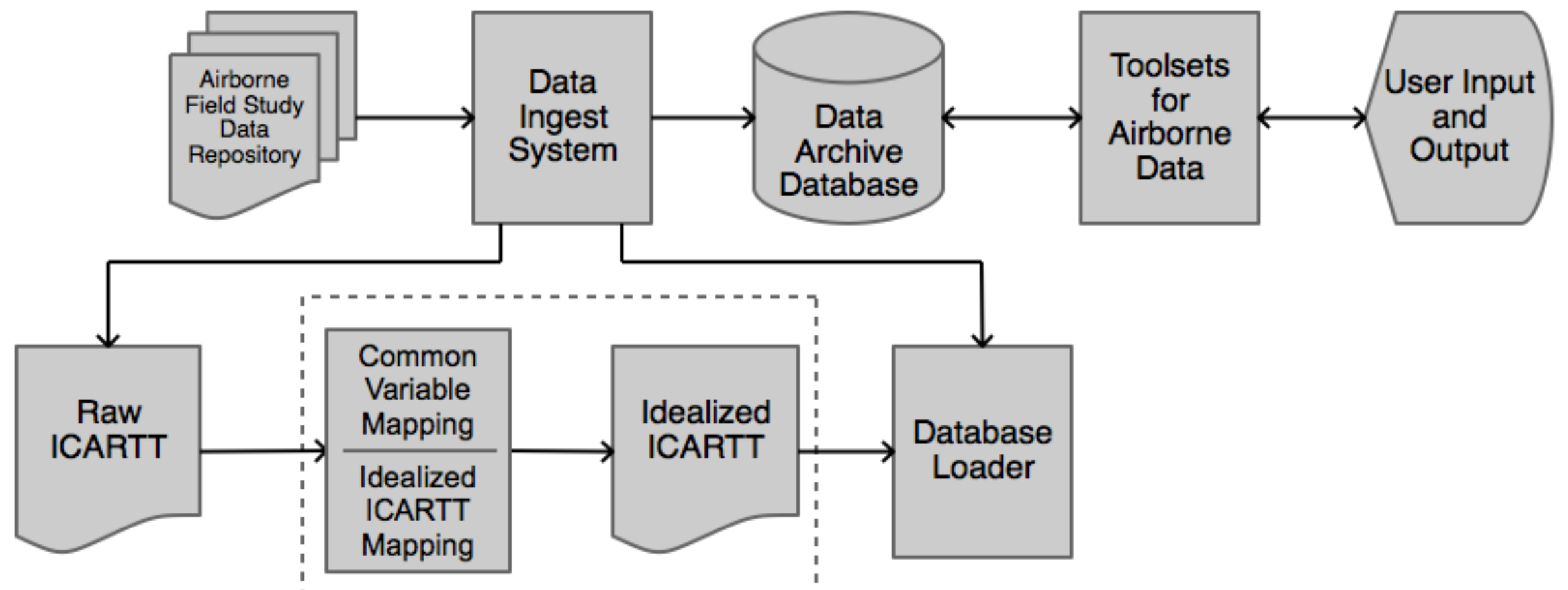
Column Name	Description
mission	The mission name.
platform	The platform associated with the mission.
data_id	The data ID from which the variable comes.
pi_variable_name	The variable name as listed in the file header.
units	The units associated with the variable. These have been standardized.
description	The variable description.
needed or not needed (1 or 0)	Some variables will not be dealt with at all by TAD, so they will not be harvested. These are marked as 0. Everything else is marked as 1.
common_name	The common name, including the category index, associated with the variable. #:#:#(#)_CommonName
UNC variable	The column name of the uncertainty variable associated with this variable or N/A.
LOD variable	The column name of the LOD variable associated with this variable or N/A.
Flag var	The column name of the data flag variable associated with this variable, if one exists. This also includes the good and bad data flags as such: VarName(GoodFlags/BadFlags). If there are multiple flags for either, they are semi-colon delimited.
var_type	The variable type.
scan_name	Some variables used in the files have names that are difficult for the harvester script to parse. This column is the corrected name that the variable will be given in the idealized ICARTT file by the filer program.
show_data	Not all variables are available to order directly. These are marked as 0, meaning they will not appear in the user interface. There are also default variables outputted with each file. These are marked as 2. All other variables are marked as 1.
array_size	If this variable is part of an array, this gives the number of elements in the array. If not, 0.
array_index	If this variable is part of an array, this gives the index of this variable in the array, starting at 0. If not, 0.
ui_name	The name that will be used for the variable on the UI. Typically these are the same as the PI variable name, but occasionally it will be different in order to better distinguish the variable. One major example is when all array variables are grouped under one name, this column will have that group name listed under the first individual variable, and all others will be marked as N/A.
group_def	A description for any group of variables marked as one, or N/A.
output_name	The name that will be used for the variable in the output file. Typically this is the UI name with "_PILastName" appended to the end. Occasionally some extra information is added before the last name to ensure that there are no repeated variable names in the output files.
instrument	The instrument used to measure this variable.
column_name	In some of the older files, the variable name does not match between the header and the data columns. To make it easier for the filer, the variable name from the data columns is stored here.
orig_dataid	Certain data IDs for the original files have been broken up based on any additional comments. For example, files from the data ID NCAR-CH2O-1Min and NCAR-CH2O-1Sec are both originally from NCAR-CH2O, but as both can occur on the same day they are distinguished by _1Min or _1Sec before the file extension. This column has the original data ID.

Above: Extended map format



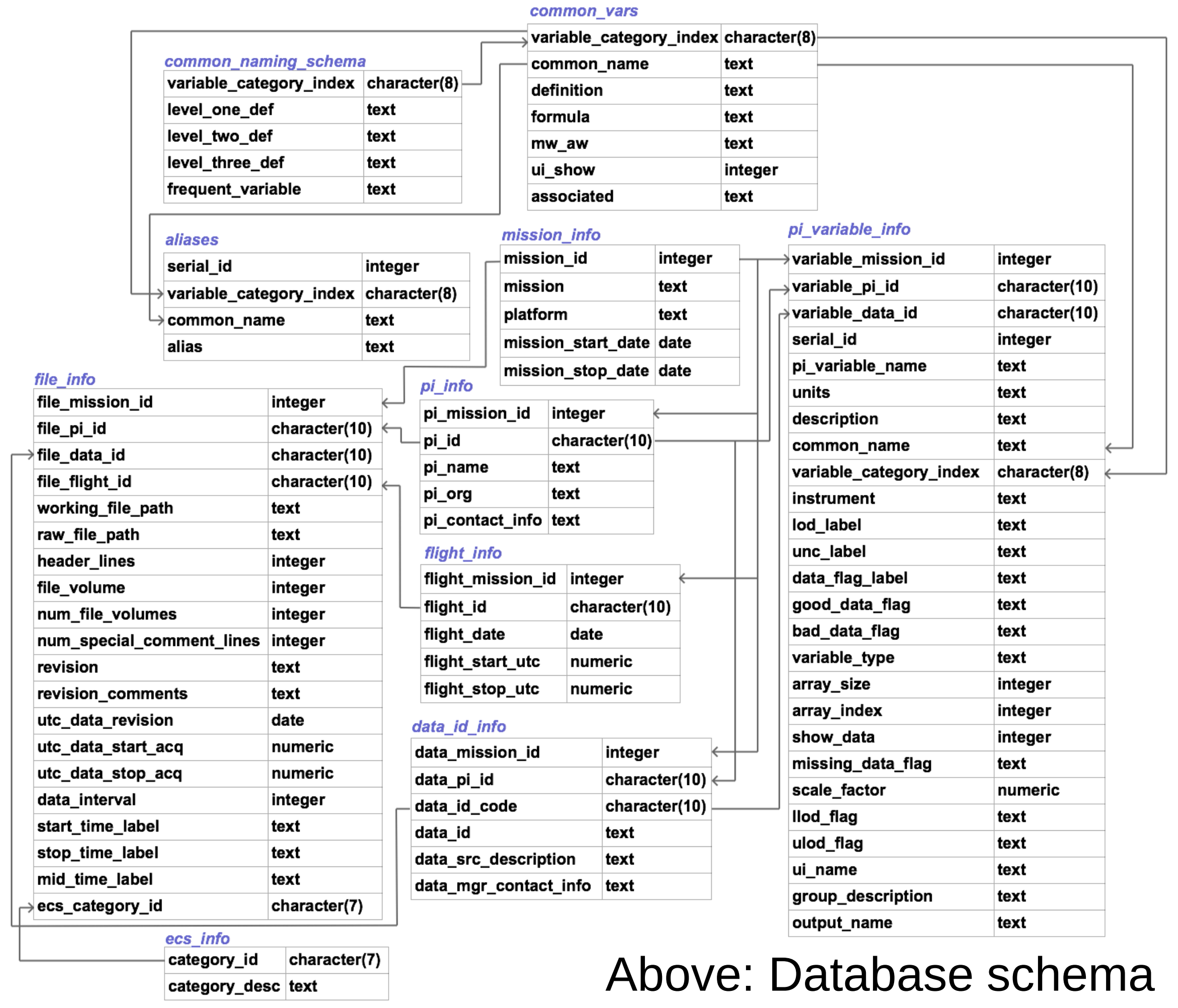
The Variable Common Naming System links the non-standardized PI variables. It consists of six groups based on the physical and chemical properties of the measurements. This system is scalable to properly handle future measurements and was created in consultation with GCMD and the airborne community at large.

- **Step 2**. Convert to the idealized format via the automated filer program



Above: Workflow for the automated filer program

- **Step 3**. Extract metadata to the database



common_naming_schema

variable_category_index	character(8)
level_one_def	text
level_two_def	text
level_three_def	text
frequent_variable	text

common_vars

variable_category_index	character(8)
common_name	text
definition	text
formula	text
mw_aw	text
ui_show	integer
associated	text

aliases

serial_id	integer
variable_category_index	character(8)
common_name	text
alias	text

mission_info

mission_id	integer
mission	text
platform	text
mission_start_date	date
mission_stop_date	date

pi_variable_info

variable_mission_id	integer
variable_pi_id	character(10)
variable_data_id	character(10)
serial_id	integer
pi_variable_name	text
units	text
description	text
common_name	text
variable_category_index	character(8)
instrument	text
lod_label	text
unc_label	text
data_flag_label	text
good_data_flag	text
bad_data_flag	text
variable_type	text
array_size	integer
array_index	integer
show_data	integer
missing_data_flag	text
scale_factor	numeric
llod_flag	text
ulod_flag	text
ui_name	text
group_description	text
output_name	text

file_info

file_mission_id	integer
file_pi_id	character(10)
file_data_id	character(10)
file_flight_id	character(10)
working_file_path	text
raw_file_path	text
header_lines	integer
file_volume	integer
num_file_volumes	integer
num_special_comment_lines	integer
revision	text
revision_comments	text
utc_data_revision	date
utc_data_start_acq	numeric
utc_data_stop_acq	numeric
data_interval	integer
start_time_label	text
stop_time_label	text
mid_time_label	text
ecs_category_id	character(7)

pi_info

pi_mission_id	integer
pi_id	character(10)
pi_name	text
pi_org	text
pi_contact_info	text

flight_info

flight_mission_id	integer
flight_id	character(10)
flight_date	date
flight_start_utc	numeric
flight_stop_utc	numeric

data_id_info

data_mission_id	integer
data_pi_id	character(10)
data_id_code	character(10)
data_id	text
data_src_description	text
data_mgr_contact_info	text

ecs_info

category_id	character(7)
category_desc	text

Above: Database schema

Toolsets for Airborne Data

- TAD is focused on *in situ* observational data, which represent the majority of the airborne measurements in the Atmospheric Composition Focus Area
- TAD draws on aircraft data holdings at the ASDC to create a data discovery tool that generates on-the-fly weighted averages of derived value-added products for researchers.
- Automated parsing tools convert ICARTT files to an idealized format for TAD ingest

Idealized ICARTT Format

The idealized ICARTT format, a restructured ICARTT Data file, improves machine readability to sustain the TAD system. The idealized format lessened the complexities with Airborne data. The advantages to the idealized format are given below.

- Consistent delimiter for scale factors, missing data flags, variables, and data.
- Data interval of 0 or 1 only.
- Time variables always reported for start, stop, and mid. Null values used as placeholders when necessary.
- Short name, unit, and long name listed for all non-time dependent variables.
- Stricter rules for variable names, i.e. no commas or spaces
- Standardized delimiters for LOD flags.
- All normal comments listed. N/A used where necessary.
- Revision information listed chronologically.
- Guaranteed variable consistency between header and columns

More Information

TAD
Toolsets for Airborne Data

https://tad.larc.nasa.gov

Special Thanks

NOAA ESRL
GCMD
WWW-AIR