

The ACRF Process (Abridged)

*For information on ARM and ACRF: <http://www.arm.gov/about/> and <http://www.arm.gov/acrf/>

ACRF collects data from over 150 instruments located around the globe. Instrument data plus external data and value added procedures make up over 1500 data streams. Data from the data streams are stored in an ACRF formatted data file. The ACRF data file is:

- netCDF
- 1 file per day per data stream
- contains all metadata need to describe the data in the file
- contains all relevant data fields
 - time
 - location
 - altitude
 - single and multi dimensional scientifically relevant (temp, rh, cloud base, etc.)
 - instrument relevant (case temp, dome temp, etc.)
 - all relevant data quality by field (min, max, delta, missing value, etc.)
- calibration information

The actual data is not currently loaded into any kind of database. The ARM archive has a database that keeps track of where the file is stored, the time range covered by the file, and some metadata. For the scientist to use this data, the scientist needs to know data stream he/she is looking for and the time range. The scientist then requests all the files within a date range, and receives all the daily netcdf files for the data stream within the date range. The scientist then need to read in all the data files into whatever program he/she uses to analyze the data.

ACRF Needs

ACRF needs to be able to load all data, associated data quality, and meta data into a query-able database to provide ARM data users with the capability to query, analyze, and manipulate ARM data. Resulting products should either be shipped to them as data files or stored in their own database working area. The database should support both users unfamiliar with the ARM data inventory, or unsure of exactly what they want by allowing the "what", "where", "when", and "how" of these actions to be described using vague, general terms.

The following are examples of the type of queries the scientists would run:

- Generic: Get me field(s) of interest from location(s) where algorithm(s) from time period(s)
 - field(s): This could be very specific like temperature from the aeri instrument to generic like meteorology where meteorology is defined within the database.
 - location(s): could be lat lon ranges, an ACRF site, or a description of an area (i.e. Southern Hemisphere).
 - algorithm(s): could be as simple as filtering where temperature is less than 80F, taking the difference of two values, or could be some complicated stored procedure
 - time period(s): could be a key word like summer, winter, fall, or some more complicated algorithm like burning season or monsoon season.

Examples:

- Get me meteorology from SGP C1 where we have optically thin clouds (clowd1turn) or clear sky (clearsky1long) and data is good from the summer of 2006.
 - SGP is the ARM Southern Great Plains site and C1 is the central facility.
 - meteorology is defined in the database as a set of data fields from a set of instruments.
 - optically thin clouds and clear sky are algorithms (stored procedures) run at the query time
 - clowd1turn and clearsky1long are algorithm names (or stored procedures) to use
 - data is good is based off of data quality flagging
 - summer is a stored procedure that is location based
- Get me cloud base height data from the TWP C2 site where it is not raining and data is good or questionable for all time.
 - cloud base height: over time and as the ARM program has matured data field names within and across data streams have changed. So cloud base height could be an alias for all field names over time or could be the current field name.
- Get me optical depths between 1km and 2km from NSA C1 where the temperature is above 50C and data is good during the spring, summer and fall of 1998

Prior to shipment or storage of the data, the user should be able to perform additional higher level data manipulation and analysis to create a specific data-set or results. This should include spatial and temporal averaging and interpolation to facilitate cross instrument analysis of data with different time intervals and height resolution, and data validation statistical algorithms that compare different versions of the same data.