An Audit of the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) Instrument Data Workflow from Ingest to Archival

10

NCAR UCAR EOL Earth Observing Laboratory

Development • Deployment • Data • Discovery

Janine Aquino[1], Steve Williams[1], Michael D. Daniels[1], Jean Hurst[1], Robert Rilling[1] and Kathryn Young[1] [1] Earth Observing Laboratory, National Center for Atmospheric Research, Boulder, CO, United States

Background

Researchers often require specialized suites of instruments in order to carry out the scientific field work needed to understand the complex, interdependent geophysical processes that govern our planet. The National Science Foundation provides the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) funding for the acquisition, operation, maintenance and upgrade of these instruments to support observational meteorology across the globe. **The data collected during these field campaigns are a legacy of NSF's investment**. Simple and timely community access to collected data and associated software tools, quality control, and preservation of the collected data for decades fosters scientific research well past the end of the field phase of a project.

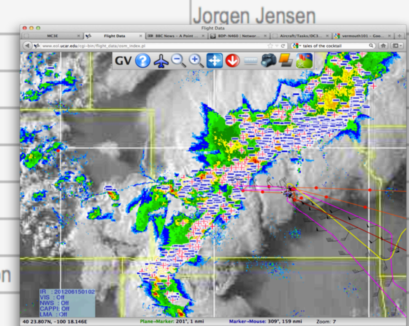
Support of EOL instrumentation is spread across three observational Facilities; the In-situ Sensing Facility (ISF), the Research Aviation Facility (RAF) and the Remote Sensing Facility (RSF). EOL's Computing, Data and Software (CDS) Facility has led the efforts to increase the coordination and stewardship of these data and processes across EOL.

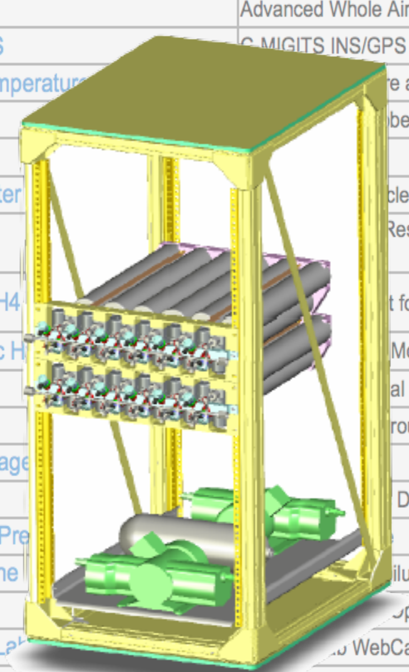
RSF Remote Sensing Facility · ISF In-Situ Sensing Facility · CDS Computing, Data and Software Facility

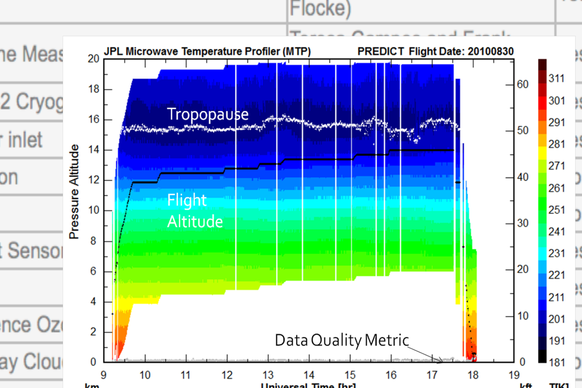
Data Audit

75 instruments have been audited so far

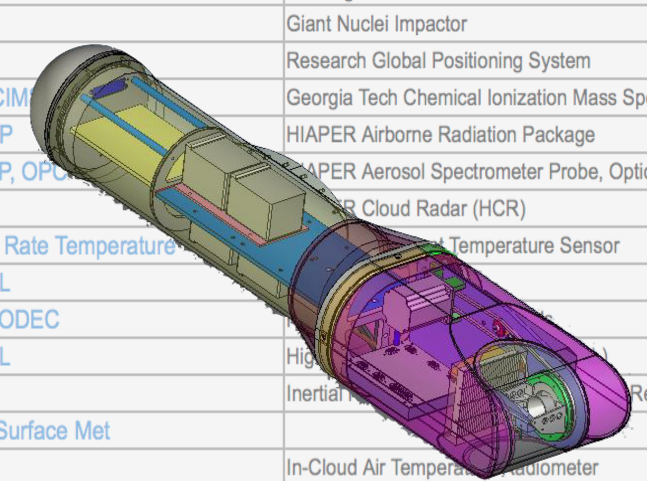
Title	Instrument	Lead Contact	Standard Data Path
2DC, 2DP	Two-Dimensional Optical Array Cloud Probe	Jorgen Jensen	Yes
3V-CPI	Three-View Cloud Particle Imager	Jorgen Jensen	No
449 Profiler	Wind Profiler		No
ADS	ADS		Yes
AO2	Atmospheric Oxygen Instrument		No
AVAPS Dropsonde	AVAPS Dropsonde		No
AWAS	Advanced Whole Air Sampler		No
C-MIGITS	C-MIGITS INS/GPS System		Yes
Cabin Temperatur	e at ADS Rack Location		Yes
CDP	be		Yes
Chat			No
CN Counter	cleus Counter (water or butanol)	Mike Reeves	Yes
CO	Resonance Fluorescence Carbon Monoxide	Teresa Campos (backup Frank Flocke)	Yes
CO2 & CH4	t for Airborne Meas		es
Cryogenic H	Model CR-2 Cryog		es
CVI	al Impactor inlet		es
DGPS	round station		o
Digital Image			o
DPx	Dew Point Sensor		es
Dynamic Pre			es
Fast Ozone	iluminescence Oz		es
Fast2D	Optical Array Clou		o
Foothills La	WebCam		
FSSP-100	ward Scattering Spectromet		
FSSP-300	Forward Scattering Spectrometer Probe, Model 300	Jorgen Jensen	Yes
GAUS	GPS Advanced Upper-Air Sounding System	Bill Brown	No
GISMOS	GNSS Instrument System for Multi-static and Occultation Sensing	June Wang, EOL	No
GNI	Giant Nuclei Impactor	Jorgen Jensen	No
GPS	Research Global Positioning System	Chris Webster	Yes
GT CIM	Georgia Tech Chemical Ionization Mass Spectrometer	Frank Flocke	No
HARP	HIAPER Airborne Radiation Package		
HASP, OPC	APER Aerosol Spectrometer Probe, Optical Partic		
HCR	R Cloud Radar (HCR)		
High Rate Temperature	Temperature Sensor		
HIMIL			
HOLODEC			
HSRL	Hig		
IRU	Inertial Reference		
ISS Surface Met			
ITR	In-Cloud Air Tempera Radiometer	Stuart Beaton	Yes
King	King (CSIRO) Liquid Water Sensor	Jorgen Jensen	Yes
KT19	Heimann KT19 Infrared Radiation Pyrometer	Julie Haggerty	Yes
LAMS	Laser Air Motion Sensor (LAMS)	Mike Spowart	Yes

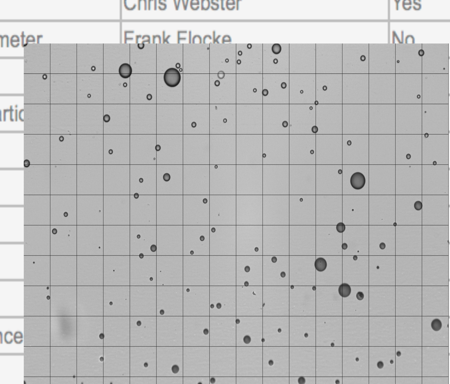






MTP Temperature Curtain Plot from PREDICT RF06 (8/30/10) into TS Fiona





Information Collected

Questions were devised to encompass the entire data lifecycle workflow of each instrument or other data source from acquisition to archive and distribution.

Field Collection · Initial transfer to local servers · Merging · Automated/Visual QA/QC · Plotting · Data /Software tracked in repository · Data<->Software Version matching · Metadata Ingest · Distribution · Final Documentation · DOIs · Post-processing · Metrics · Archival · Version tracking

Motivation

The goal of this audit is to ensure that all data and associated software are safeguarded throughout the data handling process, and that community standards of practice for data stewardship and software version control are followed.

Example instrument pages

Digital Imagery

Submitted by janine on Tue, 2013-11-05 16:47

Instrument: Digital Imagery

Standard Data Path: No

Collection & Ingest | Processing & QA/QC | Archiv

Archive format: MPEG (.mp4)

Final Data Arch… HPSS:/EOL/<yea…

Documentation… A README file f… to the project we…

Chat

Submitted by janine on Fri, 2013-…

Standard Data Path: No

Collection & Ingest | Process…

Collection: EOL runs a chatserver on rdcc.… by the EOL IRC Bot, which is d… provide chat functionality and l…

Ingest Format: TEXT (.txt)

In-Field Storage Location: The 'rbot' chat bot process on r… /cron_cp_logs_to_ftp.pl, is run i… only those log files that are acti…

In-Field Backup: rdcc.guest.eol.ucar.edu/pub/irc/logs

Metadata Ingest: In order to log a chatroom, someone with rdccadmin access has to add that room to the chatserver logging configuration. As part of that process, the user will be asked what project their chatroom is related to, and that information needs to be manually added to the chatroom mapping file /h/eol /dmg/HPSS_cronjobs/ingest/chat/trunk/chatrooms.pl, committed to svn, and copied to /h/eol/dmg/HPSS_cronjobs/chatrooms.pl. Also start and end dates for collection for a project need to be set in the same file.

Raw Data Archive Location: /net/work/Projects/chatlog_recovery/logs

MTP

Submitted by janine on Mon, 2013-11-04 12:01

Instrument: Microwave Temperature Profiler

Standard Data Path: No

Collection & Ingest | Processing & QA/QC | Merging | Archive | Distribution

Local Storage Location: /h/eol/haggerty/data/<project>/MP

Post Processing: Calibration applied using target brightness temp and ambient measurement from aircraft sensors. Statistical retrieval method applied to convert measured brightness temperatures to physical temperature profiles. Retrieval method uses large set of historical radiosondes acquired for region and season; run through radiative transfer model to calculate corresponding brightness temperature profile. Retrieval coefficients generated from modeled brightness temperature profiles applied to measurements to obtain temperature profiles. Multiple iterations of retrieval method occur to obtain best data quality.

QA/QC: Primary quality metric is the MRI, which is a statistical measure of how well the measured brightness temperature profile matches a historical profile in the archive. The MRI indicates the suitability of the retrieval coefficients applies, and hence the quality of the retrieval.

Data conversion/reformatting: Processing -- multi-stage with multiple software components and significant human intervention (RAOBget, RAOBmanager, RCcalc, MTPbin, MTPsim, some interim calculations in Excel spreadsheets)

Data<->S/W version matching: Software is VisualBasic

Processing Instructions:

- Select an in situ temperature measurement for use in gain calculation.
- Acquire raobs near flight tracks plus large set of "similar" raobs.
- Generate final retrieval coefficients.
- Initial processing of each flight, generate additional retrieval coefficients as needed.
- Edit suspicious scans, e.g., during steep ascents, evidence of radio frequency interference, etc.
- Identify best match between measured brightness temperature profile and modeled brightness temperature profile
- Apply retrieval coefficients from best match to measured brightness temperature profile to obtain physical temperature profile
- Determine side lobe errors, apply corrections, reprocess.
- Determine errors arising from statistical retrieval process.
- Iterate as needed

Digital Object Identifiers (DOIs) in EOL

Federal agencies, professional societies, and research organizations in the geo-sciences are moving towards requiring researchers to **formally cite data and sources that led to a given research result**. Unique digital object identifiers (DOIs) for citation of EOL services are being developed to allow simple reference to EOL platforms and data in publications, create a simpler process for researchers to acknowledge use of specific EOL services, and, in conjunction with data workflow information and other internal tracking tools, allow EOL to uniquely track and identify the use of platforms, hardware, software, and data in the literature.

The ADS data workflow IS the Standard Data Path

Standard Data Path

Data are digitized where necessary and then recorded by the 'NIDAS' data acquisition system on the aircraft, then the resulting "raw" or "ADS-format" data files are processed by the 'nimbus' processing program to produce netCDF-format data files for archival.

For many projects, other measurements not acquired by NIDAS may be merged into these data files to produce more comprehensive final archive files. There is normally documentation included in the netCDF header describing each measurement.

This link provides information on the file formats, and this one gives information on some downloadable programs that may be useful when using the standard data products.

Factors

- Different standards of practice lead to redundancies, and not all staff are aware of all tools that can help them.
- Data handling processes were developed independently and are often based on historical precedents, so data workflows vary widely.
- Instrument PIs have varying backgrounds and technical expertise.

Best Practices

Recognizing that software tools and data formats evolve and become obsolete, we recommend the following best practices:

- **Use descriptive path components** when devising folder or path names to store data and software (we use year, project acronym & platform name).
- **Store data in widely-used standard formats**. The advantage of ASCII is that it can be read without software.
- **Store software under version control** such as Git or SVN.
- **Assign revisions to data and software** and be able to match data to the software used to generate it.
- **Collect metrics** on data distribution including user email so that you have a way to contact them when data changes or errors are found.

Benefits

- Data are easily accessible.
- Data and software are available for future use, and to support the reproducibility of published results.
- Data are citable.
- Common use of tools and standards of practice leads to increased efficiency and reduced redundancy, while preserving integrity, thus reducing costs.
- Data in widely-used standard formats are more easily integrated.

Resources and References

NCAR Earth Observing Laboratory Facilities & Instruments (https://www.eol.ucar.edu/instrumentation)
NCAR Earth Observing Laboratory Data & Software (https://www.eol.ucar.edu/data-software)
Eaker, Christopher, 2012: Data Audit and Analysis: Mapping the Data Workflow from Ingest to Archive. Final report to NCAR/EOL.
Downs, R.R., and R. S. Chen, 2011: Audit of a Scientific Data Center for Certification as a Trustworthy Digital Repository: A Case Study, Fall AGU Meeting, San Francisco, CA.
Lord, P., and A. MacDonald, 2003: Audit of a Scientific Data Center for Certification as a Trustworthy Digital Repository: A Case Study. prepared for The JISC Committee for the Support of Research (JCSR)
Jones, S., Ross, S., Ruusalepp, R., 2009: Data Audit Framework Methodology. draft for discussion, version 1.8, (Glasgow, HATII, May 2009)

www.eol.ucar.edu

NCAR National Center for Atmospheric Research · NSF