Data Provenance and Version Control Task Team, Databank Subgroup. Conference Call #2
10 March 2011, 1500 UTC
In attendance:

Jay Lawrimore
Debra Braun
John Christy
Steve Worley
Jeremy Tandy

Primary purpose of call was to review pilot dataset and proposed data provenance tracking flags.

Spent some time initially going over the DAILY Stage 0 and Stage 1 data that are now available on the databank ftp site. ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/daily/stage0/, etc.

STAGE 0 Data: There are imaged forms for 2 countries (Vietnam and Mexico) in the ./stage0 directory. These were provided by NCDC's Climate Database Modernization Program (CDMP) – and are the only 2 sets of non-US daily surface digitized data that have so far come out of the CDMP program. There are thousands of non-US images on the Foreign Data Library website (http://docs.lib.noaa.gov/rescue/data_rescue_home.html ), but almost all of these have not been digitized into Stage1 digital data. (Data Rescue Subteam is tasked to address).

These images were exported out of NCDC's EDADS system (because external access is not provided to everyone). The Stage0 data for Mexico is voluminous, and only a sample of stage0 images (for a single station) are provided on the ftp site – for station Aguascalientes from January 1878 through Dec 1981. Jpg images by month are provided and there are multiple images for each month – Although the year and month are clear from each filename, it was found to be very difficult to determine what type of image was in each file – because there is a sequential number following the two-digit month (YYYY-MO_XXXX). This 4-digit number runs from 2719 in Jan 1878 through 7035 in Dec 1981. There are some months completely missing throughout the period of record, and the type of form and number of forms varies through time. There were as many as 12 different types of forms (images) for each month late in the record. Fewer forms earlier in the record. There was no easy way to determine what the form was from the 4-digit identifier. At the end of the record the first form in each month appeared to be hourly temperature for each day of the month and columns containing daily maximum and minimum temperature in Deg C to tenths. This was not the case earlier in the record, when it was more difficult to determine which form contained daily maximum and minimum temperature.

For Stage0 Vietnam data – all stations (31 of them) were provided in 7 separate subdirectories, and each station is segregated into individual subdirectories. Looking at the filenames within each subdirectory, it is not possible to tell which year, month and what type of form exists within each jpg file.

Stage 1 Data: The data for Mexico and Vietnam – all have been digitized by CDMP and are provided in the ./Stage1 subdirectory. There are also data for 2 other countries (US Forts, and Spain) in the ./stage1 directory. The data for 22 countries in Spain (from 1800s) were provided by local source (Manola Brunet) and images are not yet available. Data from US Forts exist on EDADS. Readme files are available for each Stage1 dataset in the respective directory.

For Mexico – all the Stage1 data were available on the ftp site in the Stage1 directory. However, the format had been found to not match what was provided in the Readme file. Follow up with the CDMP program ongoing. For Vietnam - the Stage1 data are easy to read using the Readme file and processing into Stage2 format already underway using the 5 Data Provenance Tracking Flags discussed in detail below. This DPTF process is proposed for all future Stage 2 data and is intended to encompass and allow for description and tracking of the data from its source.

Progress is being made toward creating a common formatted stage2 data for these subsets.

There was overall agreement from the subteam that the data provenance tracking flags (DPTF) are well designed. However – there was concern about the use of numbers in some instances and letters in others. It was agreed that we need to be consistent – all flags should be numeric. It is recognized that additional DPTF will likely be needed in the future.

However, Jeremy suggested that ideally - the Data Provenance Tracking metadata needs to provide better details and links to the Stage0 data.. specifically to the image file. There was not complete agreement on this with some others believing that it would be sufficient for the Data Provenance metadata to point to the directory where the Stage0 data resides – an in effect leave it to the user to identify the particular file containing the imaged data. Relating this to ICOADS – ICOADS data are provided starting at the Stage2 format. The metadata tracking flags simply provide information regarding the original source of each observation, but there is no direct link to the specific location of each observation. This remains an open issue. In the short term the DPTF metadata will indicate the source of the Stage0 data and its location.

Discussion regarding whether Stage2 data should be provided in ASCII and/or NetCDF formats. Steve provided an overview of the legacy of the ICOADS experience with ASCII. ICOADS remains in ASCII format largely because it was the only/best available format at the time ICOADS was initiated. Largely remains that way today as a result. The team generally believed that the use of NetCDF (CF convention) would be beneficial in standardizing access with the wider climate community (especially the climate forecast community). It is less likely that the international observational community will be as comfortable with NetCDF and would likely find it difficult to access at first. John pointed out that we would certainly need to provide a tool that would provide users with ability to convert from NetCDF to ASCII. Will need to get additional feedback from wider Databank Subgroup. In the mean time will proceed with creating both an ASCII and a NetCDF formatted Stage2 Pilot dataset.

Jeremy reported on ACRID project – mid-term deliverables have been made available. Success has largely been agreed upon scientific workflows for development of datasets such as HadCRUT. Information models to capture data still a work in progress. Open data provenance model no longer thought to be better than the ISO metadata standard – which is now the defacto metadata standard. (FGDC agreed to it as the international standard.)

Discussion of difference between use of "Data Provenance" versus "Lineage" – Jeremy, in reporting on status of ACRID project, suggested that we may want to distinguish between data provenance and lineage terminology. Subsequent details provided via e-mail.

- Lineage:
  - source: ISO 19115:2006 geographic metadata
  - LI_Lineage entity description: "information about the events or source data used in constructing the data, or lack of knowledge about lineage"
- Provenance:
  - source Dublin Core Metadata Initiative - Terms (http://dublincore.org/documents/dcmi-terms/)
  - Term name: provenance [http://purl.org/dc/terms/provenance] "A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation"

You'll note that the dcmi-terms form an RDF vocabulary, so each term in the vocab can be explictly referenced.

Steve Worley responded with the following –

From the OAIS Archive Reference Model,
public.ccsds.org/publications/archive/650x0b1.PDF

Provenance Information: This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This gives future users some assurance as to the likely reliability of the Content Information. Provenance can be viewed as a special type of context information.

I do not find a definition for 'lineage' in the OAIS ARM, and maybe that is reasonable because their definition seems to combine the definitions provided by Jeremy.

We at NCAR, and I believe also at NCDC, discuss our archiving strategies in reference to the OAIS. So, that is a minor push toward following that definition.

I'm not an expert in this area, so I would defer to Jeremy and others at NCDC to help set this definition.

Deb to take the lead in getting more information on the issue -- If the OAIS definition has widespread use that would simplify things.

*Data Provenance – Metadata TrackingProposal (BASED ON RECOMMENDATION DURING CALL – ALPHA FLAGS #3 and #5 BELOW WILL BE CHANGED TO NUMERICS)*

Consistent with the procedures established for ICOADS (where all data reported at one time and location by an observing system is assigned 3 metadata fileds). For our purposes of dealing with land surface data, we have established 5 of what are currently referred to as Data Provenance Tracking Flags.

These DPT flags are assigned to each observation. At this time we are assigning these flags to data within the Stage 2 data files. It does not seem appropriate to attempt assigning these to the Stage 0 or Stage 1 data – as those stages are the initial paper, imaged, or engineering unit data (Stage 0) and native format keyed data (Stage 1).

However, information regarding the origins and types of Stage 0 and Stage 1 data are defined by 2 of the 5 DPT flags that exist within the Stage 2 dataset. The 5 flags are: (1) *Stage 0 Source*, (2) *Stage 1 Source*, (3) *Data Type*, (4) *Mode of Digitization*, and (5) *Mode of Transmission/Collection*. Additional information on these is provided near the end of this section.

Whatever format and flagging scheme that is established, we want to allow for additional flags (a 6th, 7th, 8th, etc. type of DPT flag) to be added whenever it's determined to be needed. In addition, the information contained within each DPT flag can be expanded as necessary to provide as much information on an observation as possible. For example a 6th DPT flag might be I*nstrument Type*.

Because of the large number of flags and the potential for new flags to be defined and added in the future – there are recommendations from others that we use NetCDF files for Stage 2 that follow the CF (Climate and Forecast) metadata convention. This is a format taking on growing acceptance in the climate community (particularly locally). NCDC is testing and evaluating the use of NetCDF for the Stage2 data.

The 5 tracking flags that are currently being used for the Stage 2 dataset are as follows and this provides examples of how they might be defined.

*Stage 2 Data Provenance Tracking Flags*

1. Stage 0 Source (100s: Offline paper -- 200s: Offline Images – 300s: Online Images)
    - 101: Paper, NCDC
    - 102: Paper, JMA
    - 103: Paper, Australian BOM
    - 201: Images, University Rovira I Virgili, Centre for Climate Change
    - 301: Images, Databank Stage 0 ftp site
    - 302: Images, EDADS website, NCDC

- 303: Images, NOAA Library website
- ….
- -99: Missing

2. Stage 1 Source (Describes source of native format data)
    - 100: NCDC International Collection
    - 101: High Plains Regional Climate Center
    - 102: NCDC DSI-3200
    - 103: NCDC DSI-3206
    - 104: University Rovira I Virgili, Centre for Climate Change
    - 105: NCDC CDMP digital archive
    - ….
    - -99: Missing
3. Type (This metadata field is needed because some data are pre-QC'd or Adjusted by Originator and the Raw observations are unavailable)
    - A: Raw
    - B: Quality Controlled by originator
    - C: Homog Adj. by originator
    - U: Unknown
4. Mode of Digitization (Describes method of digitization)
    - 101: Keyed, SourceCorp
    - 102: Keyed, CDMP
    - 103: Keyed, CDMP Forts Project
    - 104: Keyed, Local Originator
    - 000: Auto Collect
    - -99: Unknown
5. Mode of Transmission/Collection (Describes the process used to get the data to the databank)
    - A: Mail
    - B: E-mail
    - C: FTP
    - D: SRRS FTP
    - E: NOAAPort
    - F: NMHS Web Service
    - G: Telephone Modem
    - H: Direct Datalogger download/PDA
    - I: Other Satellite

Using these 5 flags, the data received and Processed into Stage 2 format would have the following flags:

- from Manola Brunet in Spain: 201/104/B/104/C
- Vietnam data and Mexican rescued by CDMP: 301/105/A/102/C
- Forts Data would have: 302/103/A/103/C

- COOP Data transmitted over WxCoder: 302/102/A/101/C
- COOP Data transmitted by HPRCC: -99/101/A/0/C