Data Provenance Conference Call – Summary Notes
20 Jan 2011, 1500 UTC
Jay Lawrimore
Debra Braun
John Christy
Steve Worley
Jeremy Tandy

As part of ensuring Data Provenance, version control and configuration management is required as outlined in white paper 6 from the Exeter meeting. http://www.surfacetemperatures.org/whitepapers .

Overarching themes.
Tracing the full history of the object (data) up to the present; Documentation; and Detail sufficient to allow reproducibility.

Our initial goal is to provide traceability from Stage 3 data sources of merged sources (such as GHCN or the global databank) back to the original Stage 0 data (paper or fundamental data stream such as digital counts or voltages). In many cases the original observations are not available so the best we can hope to do is traceability back to Stage 1 data (original keyed or first formatted version of data). In some cases even that will be difficult.

Decision to construct a data provenance model based on the ICOADS experience meshed with the GHCN structure and recognizing differences between ocean and land surface observations.

Jeremy provided information on a UK effort (the ACRID project) that began in December and is focused along the same lines as our goals. It was initiated in response to issues with data provenance that arose in late 2009. Jeremy promised to keep us abreast of their progress and will start by providing project documents and notes from their meetings – pending approval from the team (Approval given post call).

- Project website is http://www.cru.uea.ac.uk/cru/projects/acrid/
- Specifically ACRID - aims to develop an approach to exposing climate research data for re-use, through the adoption of linked-data principles for the data themselves. The data to be exposed within the project will be three major CRU datasets, but the methodology to be used will be deployable elsewhere too. Best practice data citation techniques will enable a seamless link with research publications. Mechanisms will be developed for capturing key provenance metadata, and for adapting previously developed climate science data models to integrate with data re-use standards (OAI-ORE) and emerging Cabinet Office guidelines for public data.
- Post call Jeremy provided the following
    - 1) The proposal [JISC-1409-CRU-RAL-FINAL.pdf]
    - 2) Description of the work packages [ACRID_workpackages_21Dec2010.pdf]
    - 3) Presentation used provide input info to kickoff meeting [ACRID_20101214-ppts.pdf]
    - 4) Official notes from the kickoff meeting in Dec [ACRID_Meeting14Dec2010_notes_v2.pdf]

- 5) His notes from kickoff meeting (a) that he wrote to keep his Met Office colleagues informed [ACRID project meeting at RAL.pdf], and (b) for his own records [ACRID project meeting - technical notes & other asides.pdf]
- 6) His notes from the 25 Jan 11 ACRID meeting [ACRID project meeting 25-Jan-2011.pdf]

Steven provided an overview of the ICOADS version control and data provenance procedures. See document entitled "icoads_prov_vers-v3.0.docx". It is structured around 3 primary types of metadata.

- DCK number: The primary description of the dataset.
- Source ID: to distinguish subsets within a DCK and dynamic cases where updated versions of a source are periodically received.
- Platform type: for ex., ships, moored buoys, drifting buoys.

2 Future additions planned as follows:

- Unique Report ID: Numbering of all records from 1, .. , ~295M in precise sequential archive.
- Release Number: String of numeric characters (e.g., 2.5 or 2.5.1). This is similar to GHCN versioning that's being implemented for GHCN-M v3.0.0. (GHCN uses a date stamp as well).

Each of the ICOADS data provenance elements are included within the dataset itself alongside each observation.

Jay provided an overview of the GHCN 3-flag format and how that might be used as a starting point for data provenance when combined with the ICOADS method. Mapping ICOADS to the current GHCN 3-flag format (Measurement, Quality, Source), would be as follows:

ICOADS	**GHCN**
Does not exist (?)	Measurement flag
Does not exist (?)	Quality flag
DCK Number	Source flag*
Source ID	Source flag*
Platform ID	*Does not exist*
Unique Report ID	*Does not exist*
Intra-Record Release #	V3.0.0.*Datestamp*

*The GHCN source flag contains a mix of the information in the ICOADS DCK number while not containing as much information. For example, the US COOP network of Summary off the Day data comes in at least 4 forms (Digitized from paper, Electronic transfer of web-based from Western Regional Climate Center composite once monthly as "official" observation, SHEF formatted and transmitted daily via SRRS as "preliminary" data, from High Plains regional Climate Center once daily as "preliminary" data). These are each listed as separate sources in GHCN-D. They would contain a single DCK number and multiple source IDs in ICOADS????

Discussion of ISO Standard 19115: Recommended that we use this standard. Deb cautioned that it could be used as a framework – but that's its complexity and byzantine nature makes much of it "optional". Probably 95% of it. Our pilot should reference back to the Standard but too early to impose ISO

vocabulary. From an IT perspective, the ISO standard is established as dataset hierarchies. Less metadata flags the further down the hierarchy.

ICOADS is not following an ISO standard – design dates back to the 1980s and NCDC's TD-1171 that was the standard at the time. No intention to repeat that now. Need to learn from ICOADS but don't force the land surface data into it unnaturally.

Initial Plan agreed upon by subteam:

- Investigate applicability of ISO 19115 standard for surface databank. Jeremy to provide pointers within ISO documentation.
- Deb to have side discussions with Steve to better understand ICOADS data provenance design.
- Subteam to follow progress of ACRID effort with evaluation of opportunity to leverage off of their data provenance and version control design. Jeremy to serve as liaison to those activities.
- NCDC to continue to work toward establishing the Stage 0 through Stage 3 pilot dataset.
- NCDC to begin developing a strawman for a recommended framework for data provenance based on the activities discussed today and to be discussed.

Next call: 10 March 2011 at 1500 UTC.