Databank Working Group Teleconference #2. 5 April 2011

The second meeting of the Databank Working Group was held via teleconference on 5 April 2011 from 1200-1400 UTC.

Members present:
Waldenio Gambi de Almeida: CPTEC/INPE, Brazil
Meaghan Flannery, Australia Bureau of Meteorology
Albert Klein-Tank: KNMI, Netherlands
Jay Lawrimore, NOAA/NCDC
David Lister, Climatic Research Unit, East Anglia, UK
Matt Menne, NOAA/NCDC
Matilde Rusticucci, Univ of Buenos Aires, Argentina
Madeleine Renom: IFFC, University of the Republic, Montevideo, Uruguay
Peter Thorne, NOAA/NCDC/CICS
Steve Worley, National Center for Atmospheric Research

Action Items resulting from this conference call. Target date: end June/early July.

1. WG members have a goal of providing at least one new source of data in Stage 1 format (and Stage 2 if practical).
   Specific commitments
   a. MF: from Pacific and NZ.
   b. WA: Brazilian CPTEC data. And will talk to Brazilian Met Service.
   c. MRu: Will look into availability of CLARIS data from River Plate
   d. MRe: A number of stations from Uruguay, some recently digitized from beginning of 20th century.
   e. MM to provide data for US/N.A.
   f. AKT from European collection.
   g. SW from NCAR collection.
2. PT to develop data request letter for signature by CCl, GCOS Steering Committee, Databank, and Steering Committee leads.
3. JL to address pamphlet and poster to be produced to be produced by NCDC graphics. WG members to review.
4. JL to document and provide guidance for data submission
5. JL to develop terms of reference for the Databank. WG members to iterate and modify.
6. JL to refine Data Provenance Tracking flags and iterate with WG.
7. JL to consider additional options for databank mirror sites, such as WDC-B.
8. PT to follow up with UK rep to WMO Congress
9. PT to establish a blog and e-mail alias for the WG.

Other items for consideration during June conference call
Consideration of timing for next face-to-face meeting

Discussion Notes

Membership and Introductions

Four new members to the databank working group were introduced.
Waldenio Almeida, Meaghan Flannery, David Lister, and Madeleine Renom
Rod Hutchinson from Australia BOM is no longer an active member of the Databank working group but will remain engaged through the Data Rescue task team

Databank Development and Structure

Discussed deployment of a pilot databank which is now linked from the GOSIC website at http://www.gosic.org/GLOBAL_SURFACE_DATABANK/GBD.html . In addition to the pilot databank, links are also provided to NCDC's global daily and monthly (GHCN) datasets.

Reviewed the collection of data from US Forts, Mexico, Vietnam, and Spain that have been loaded to the Daily databank (Stage 0 and 1), and converted to the common Stage 2 format with Data Provenance Tracking.
ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/daily/ .

Stage 0 Data: There are imaged forms for 2 countries (Vietnam and Mexico) in the ./stage0 directory. 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. For Stage0 Vietnam data – all stations (31 of them) were provided in 7 separate subdirectories, and each station is segregated into individual subdirectories.

Stage 1 Data: The data for Mexico and Vietnam – all have been digitized and are provided in the ./Stage1 subdirectory. There are also data for 2 other countries (US Forts, and Spain) in the ./stage1 directory. Readme files are available for each Stage1 dataset in the respective country directories. The data for 22 countries in Spain (from 1800s) were provided by Manola Brunet and images are not yet available. Data from US Forts exist in a system currently accessible only internally to NCDC.
Stage 2 Data: The Stage 1 data have been converted to a common format and data provenance tracking flags assigned to each observation. This topic was discussed more in the Data Provenance task team section.

Discussion on encouraging contributions of both Stage 1 and Stage 2 data.
Recommendation from SW and general agreement that once we have solidified the format for Stage 2, the holders of stage 1 should be encouraged to undertake that reformatting conversion to Stage 2. In addition to greater support and global ownership for the databank effort, the providers are experts in their own format.

AKT indicated he could aid in transforming data for Europe.

However, there was concern expressed about the overhead required for such an effort. If this were a requirement it would limit the ability of many countries to participate.

All WG members agreed that it is important to encourage this while not making it a requirement. And WG members would act as ambassadors to aid in data collection. (More discussion of this below.)

Regarding submission of specific sources of data
Now that the databank has been initiated with 4 sources, we would like to move forward with collection of other source datasets. Several were offered.

SW indicated that there are many sources from NCAR that are not in NCDC's archives. PT indicated the importance of collecting all sources regardless of whether they may overlap other sources already in global archives. In many cases observations may differ between sources (e.g., 100+ ways of calculating monthly mean temperature).

MF indicated that there are many holdings in Australia but expressed concern about providing data that may already be widely available in datasets, such as GHCN.
JL made the point that having multiple versions of the same data is okay in Stage 1 and 2. It is in Stage 3 that the merging, duplicate removal and reconciliation is done.

AKT has ~2000 European stations with daily series available.

JL will need to provide an ftp site for which anyone can upload data and will develop a mechanism to pull the data into the databank.

WA: How will we ensure quality control of the data? JL: It is incumbent on the provider to provide the most accurate data and metadata, formats that match format descriptions, and ensure overall quality of the raw observations. If the provider has conducted quality control before submitting the data to the databank, that needs to be communicated with the submission. There are Data Provenance Tracking flags in Stage 2 that can be used to document such observations. Ideally the raw observations will be provided without replacement, but if obs have been replaced, the lineage of that will need to be traced.

Quality control as part of databank will take place, but not until Stage 4 – which, along with Stage 5 (bias corrections) is not under the purview of the Databank working group.

Issue of Need for Mirror Sites
PT asked if we should look to have a single site or mirror sites in each WMO region – for Better downloads and better acceptance globally.
Suggestion that WDC-B would be a good place to start.
JL to follow up with WDC-B, Vyacheslav Razuvaev.

Question about Metadata for Stage 3 data
JL: Databank is for collection of data and metadata. Collect, format, provenance, merging, and metadata.

There is the need to collect as much metadata as possible and include as part of data files in the same way that ICOADS has included as much metadata as possible. e.g., source id, sub-source id, platform. For this databank that will be accomplished through the use of Data Provenance Tracking (DPT) flags.

Data provenance tracking discussion

JL provided an overview of the 5 Data Provenance Tracking flags established by the Data Provenance and Version Control task team.

The five flags are: (1) *Stage 0 Source*, (2) *Stage 1 Source*, (3) *Data Type*, (4) *Mode of Digitization*, and (5) *Mode of Transmission/Collection*. It is possible to add additional flags (a 6^{th}, 7^{th}, 8^{th}, etc. type of DPT flag) whenever additional ones are needed. For example a 6^{th} DPT flag might be I*nstrument Type*. In addition, the information contained within each DPT flag can be expanded as necessary to provide as much information on an observation as possible.

MRe -- How would this cope with mixed format for stations that change, for ex., changing from keyed to auto collection?
Flags are by observation so mixed collection etc. can be acknowledged.

PT: Are there additional flags for monthly observations created from daily data?
This can be established through DPTs that are unique to the monthly dataset - need to cross-track daily / monthly.

MF: Is there flagging for non-standard practice?
Answer: There is the ability to provide that information through the incorporation of additional DPT flags.

SW: Stage 1 collection for a provider with QC flags – how would they record so it is carried forwards?
There are no current plans to provide the qc flags that data providers include with data submissions. PT suggested that a single binary indicator could be provided – identified invalid by provider could be included as a DPT.
JL reiterated that the 3^{rd} DPT flag already provides the opportunity to indicate that data was pre-qc'd by the data provider. An additional element can be added to that flag to indicate, Quality Controlled-Invalid.

AKT: Where does version control come in? When a replacement series is provided how do we track it?
PT: Version point is not just on the values, also the databank itself? Is the databank as a whole version pointed?
SW indicated that with ICOADS version controls at the Source and at the Dataset level. New source identifiers are used to indicate a new version of a source. For example, if NCAR were to provide version A1 of their data in 2005, the source identifier would indicate NCAR Vsn A1 as the source. If in 2009 they provided an update to their data and

it is now NCAR Vsn A2, that would be the source for any observation provided in that update.
For version controlling the full ICOADS, the dataset is given a new version number when new releases are made.
JL acknowledged that a similar method could be established for the databank – with the initial release of the databank in April 2012 being version 1.0.0 and future releases occurring on a periodic basis in a similar way to ICOADS. Regarding version controlling at the source level (for each observation), the DPT flags provide this capability.

AKT suggested that we need to pull through all of this information to stage 3.
JL: This will need to be addressed when Stage 3 data are produced – some months in the future following further development of Stage 2.

Data Acquisition Efforts

Issue regarding how the Databank WG can better publicize this effort.

PT: WMO is obvious. UK submission to WMO Congress provides a good opportunity. PT to follow up with UK rep to Congress.

Question from MF regarding whether 3rd party data will be included? Yes, certainly like to use so long as it's free of usage restrictions. Broaden horizons as far as possible.

How to do this at the Regional Level?

AKT: Can use European contact points to raise attention for the databank.

A pamphlet and poster were agreed to be good avenues to increasing awareness of the Databank effort during workshops and conferences.

JL to address pamphlet/poster with NCDC graphics team.

Discussion of Data Submission
JL needs to document the data submission process. Will limit to 1 or 2 pages.

Discussed need for Databank members from each region to take on action to identify sources and work towards getting them into the databank.

MF volunteered to be responsible for Region V (Pacific and NZ).
WA: There will be no problem providing Stage 1 data - Brazilian CPTEC data. Also will talk with Brazilian Met Service.
MRu: CLARIS data from River Plate may be available (e.g., Argentina, Uruguay, Bolivia) May be able to expand and try to improve S. America. A formal letter to provide would be very beneficial.

PT suggested that a letter from the head of CCl (Tom Peterson) and possibly GCOS Steering Committee (Adrian Simmons). PT to take this action.

MRe: Once provided with the letter will present to the weather services for the data to be released. 5 stations from Uruguay available from a library. 17 from NWS 17 and 11 digitized from beginning of Century.

Data Rescue Discussion

PT briefed his activities as lead of the Data Rescue task team.
Task team now consists of 15 members. Primary activity has been circulation by members of various inventories of current and planned data rescue efforts broken down to country / station level but these have yet to be tied together in a useful way.
The major effort of this team is coordinating efforts among the many data rescue activities underway worldwide.

NCDC hosted a visit by zooniverse which focused upon crowdsourcing the digitization of imaged data. Several potential avenues were discussed. There are two principal technical challenges:
Creating a versatile system that enables the reading of multi-format data in a cost-effective way.
A method to ensure user engagement to catch people's attention for long enough

These issues are resolvable so current focus is on pursuing funding for this effort.

Google also expressed interest in following up with a view to looking at various options including education and user verification to get involved. A two side summary was prepared and sent and a response is pending.

Discussion of Steering Committee's Proposed Implementation Plan

JL provided an overview of the Databank portion of the plan and reviewed the timetables in section 3.5.
There was general agreement with the timetables. The biggest is the April 2012 target for releasing version 1 of the Databank.
PT: Need to have people working and bringing into the databank so version 1 is demonstrably new and an improvement. 2012 is a challenge but in terms of momentum plausibly the last date that works. Need to ensure it is comprehensive. WG members are the experts. Concentrate on 3.5 the workplan.

Requests from Steering Committee provided by PT

Databank Terms of Reference: JL to develop terms of reference and circulate to WG members before next conference call.

PT will set up group e-mail alias.

Marketing. WMO UK submission, each act as ambassadors, and letter to come from us countersigned by Jay, Peter, Tom, Adrian.

Activities for coming 2 to 3 months prior to next Databank conference call

1. WG members have a goal of providing at least one new source of data in Stage 1 format (and Stage 2 if practical).

   Specific commitments

   a. MF: from Pacific and NZ.
   b. WA: Brazilian CPTEC data. And will talk to Brazilian Met Service.
   c. MRu: Will look into availability of CLARIS data from River Plate
   d. MRe: A number of stations from Uruguay, some recently digitized from beginning of 20th century.
   e. MM to provide data for US/N.A.
   f. AKT from European collection.
   g. SW from NCAR collection.
2. PT to develop data request letter for signature by CCl, GCOS Steering Committee, Databank, and Steering Committee leads.
3. Pamphlet and/or poster to be produced. JL and PT will take lead.
4. JL to document and provide guidance for data submission
5. JL to develop terms of reference for the Databank. WG members to iterate and modify.
6. JL to refine Data Provenance Tracking flags and iterate with WG.
7. JL to consider additional options for databank mirror sites, such as WDC-B.
8. PT to follow up with UK rep to WMO Congress
9. PT to establish a blog and e-mail alias for the WG.

Discussion of 2nd face-to-face meeting as follow-up to Exeter

PT suggested it be limited to databank and that it come four or so months before the initial databank released.

There was some concern about how rapidly that is approaching and that there are still many things to work out. That the WG would have a better feel for the timing by the next conference call. It was agreed to revisit the timing at the next call.

Next Conference Call

Tentatively scheduled for late June/early July. JL to send out doodle query in May. Will be in touch via e-mail and the blog between now and then.