Databank Subgroup Teleconference #1. 29 October 2010

The first meeting of the Databank Subgroup was held via teleconference on 29 Oct 2010 from 1200-1345 UTC.

Members present:

John Christy, Univ of Alabama Huntsville
Rod Hutchinson: Australian Bureau of Meteorology
Albert Klein-Tank: KNMI
Bryan Lawrence: BADC
Jay Lawrimore, NOAA/NCDC
Jeremy Tandy: UK Met Office
Barbara Tencer, Univ of Buenos Aires, Argentina (Representing Matilde Rusticucci)
Peter Thorne, NOAA/NCDC/CICS
Steve Worley, National Center for Atmospheric Research

Major Decisions and Action Items:

Membership

Additional members need to be added to this Subgroup in key functional areas and regions of the world. To support data rescue activities Rob Allan (ACRE) will receive an invitation to join.

Need more representation from S. America to address data provision outside Argentina. **Steve Worley** to provide potential member from Brazil by **10 November**. **Barbara or Matilde** can provide potential member from Uraguay by **10 November**.

Need representation from Africa. Several names put forward. **Steve Worley and Jay L.** will continue to pursue and resolve by **15 November**. John Christy also offered to serve as representative given his experience there.

Databank Development

Databank will consist of data at monthly and daily timescales during at least Year 1 of project.
It will be housed at World Data Center-A, NOAA/NCDC.

Initial offer to use GHCN-D and GHCN-M datasets as foundation of databank was put on hold in favor of small test datasets that can be established from Stage 0 through Stage 3. **NCDC to pursue and establish a process for creating these test data sets before end of November. Goal to have test datasets available by end of December.**

Mirroring is not an immediate priority but one site has been offered by Steve Worley at NCAR.

Data will reside in ASCII format with NetCDF a high priority, especially as data of higher temporal resolution included. Relational Databases appropriate for metadata and to improve data access and visualization (to be addressed by Data Access and Visualization Team).

Data Provenance and Version Control

This is of highest priority. Processes and guidance need to be established early on.
This will be initiated by a Working Team: Initial contributions by Steve Worley, John Christy, and Jeremy Tandy. **Jay Lawrimore** to coordinate activity and address other resources at NCDC. To establish conference call for **latter half of November** to discuss direction of this working team.

Steve Worley to provide starting point by pulling from ICOADS data provenance and version control library of documents. **By 10 November.**

Data Rescue and Crowdsourcing

This will be an essential aspect and needs to receive considerable attention. Millions of images currently exist – major efforts over the past decade including NCDC's Climate Database Modernization Program have provided rich sources.

Data Rescue Working Team was established to address and to begin leveraging off burgeoning efforts currently focused primarily in marine community (oldweather.org). Team led by Peter Thorne and Rod Hutchinson with intent to bring in others including Rob Allan and possibly others including Dick Crouthamel and Manola Brunet. **Peter** to bring Rob Allan on board and progress with Working Team conference call **before end of November** to discuss data rescue approach.

Data Inventory

As a starting point, this Subgroup needs to develop an inventory of the data that currently exists.
This will begin with an inventory of GHCN-Monthly and GHCN-Daily data. **Jay and Matt Menne** will coordinate at NCDC, and Rod Hutchinson also has an inventory of data from Pacific that can be provided. **Inventory to be provided by 3 December.**

Providing an inventory of data in imaged format much more difficult and inventory currently limited. To the extent that an inventory of imaged data exists, it will be provided. **Peter Thorne to coordinate. Timeline TBD.**

Near real-time and historical data via WMO processes

Subgroup recognizes the need to address issues requiring WMO action (annual World Weather Record transition from decadal to annual, daily climate summaries via daily climate bulletins instead of synoptic, and inclusion of metadata in CLIMAT bulletins)

Jeremy Tandy to be involved via Expert Team on Metadata and Data Interoperability that he chairs.

Jay L. also to coordinate with Stephan Bojinski at WMO regarding Expert Team on Data Representation and Codes and potential coordination already taken place. **By 5 November.**

Building Momentum

Following progress by this Subgroup's Working Teams (Data Provenance and Version Control; Data Rescue and Crowd Sourcing) this subgroup will address and plan a workshop to include this subgroups members and others who are key to this effort.

These meeting notes to be placed online on surface temperature .org site by **Peter Thorne by 3 November.**

Jay Lawrimore to begin developing Databank Scoping Document with executive summary for management. **Date TBD.**

Details of Discussion

The goal of the Databank Subgroup was put forward in advance of the meeting in short form.

To establish, maintain, and preserve a single universal databank of land surface meteorological observations (including variables other than temperature).The databank will include several "stages" beginning at the raw observation in written form (hardcopy) or voltage / digital count for electronic data and progressing through to a unified databank holding in a consistent format which accounts for merging* of records from the same station held by different agencies. The final aim is to include as many stages as possible for each data point with traceability and known provenance. The databank will consist of holdings from the individual observation level to monthly summaries. This will meet the needs of the climate services community for high quality, traceable, and fully accessible climate data.

**Merging will include bringing together various sources, while maintaining the source identification on each record, to create the longest possible time series at a nominal station position. Data fields from multiple records during overlapping periods will be reduced to a single record in later stages but the original source records will be retained in earlier stages.*

Team Membership: The initial set of team members was presented as follows:

WMO Region I:

TBD

WMO Region II:

Vyacheslav Razuvaev, Russian Research Institute of Hydrometeorological Information

Koji Ishihara: Japan Meteorological Agency

WMO Region III

Matilde Rusticucci: Univ. of Buenos Aires, Argentina

WMO Region IV

Matthew Menne: NOAA National Climatic Data Center
Steve Worley: National Center for Atmospheric Research
John Christy: University of Alabama Huntsville

WMO Region V

Rod Hutchinson: Australian Bureau of Meteorology

WMO Region VI

Albert Klein-Tank: KNMI
Bryan Lawrence: BADC
Jeremy Tandy: UK Met Office

Chair:

Jay Lawrimore: NOAA National Climatic Data Center

Ex-officio Member

Peter Thorne: NOAA National Climatic Data Center

There was broad agreement that membership should include those who could bring additional expertise in critical functional areas as well as those who could represent areas of the world not presently covered.

In particular the data rescue aspect. Several names were suggested and it was agreed that one additional person should be brought onto the Subgroup and others could be involved via a Data Rescue working sub-team activity. Rob Allan (ACRE) will receive an invitation to join the Databank Subgroup.

BT: Being associated with the University, they have to ask permission to release data. Also, they cannot represent or provide data from other countries in S. America. Recommend that other representatives be sought.

SW: Can provide a contact from Brazil who is heavily involved in data development, particularly on data rescue side.

There also was discussion regarding Databank Management – and the need for other representation in that area. There was general agreement that Matt Menne and Bryan Lawrence well represent expertise in that area.

Surface Databank Construction

There was reiteration and agreement that the data will exist in six stages:

Stage 0: Digital image and hard copy
Stage 1: Keyed in native format
Stage 2: Converted to common format
Stage 3: Consolidated Master Database
Stage 4: Quality controlled derived products
Stage 5: Homogenized and Benchmarked products

There is agreement that the databank would be housed at the World Data Center-A (WDC-A) and made accessible via the WDC-A (http://wdca-meteorology.org) website.

It was proposed by NCDC that the databank could begin with the Global Historical Climatology Network-Monthly (GHCN-M) and the GHCN-Daily datasets as the foundation for the databank. However, only Stage 2 and Stage 3 data exist for GHCN-M in particular. To a small extent Stage 1 data are available for GHCN-Daily. When GHCN-M was constructed in the early 1990s and Stage 1 data were collected from various sources they were stored on a server at Oak Ridge National Laboratory. In the days before good system backups, a server crash resulted in the loss of the Stage 1 data.

Subsequent discussions led to a recommendation that this effort should begin with small sets of test data in Stage 0, 1, 2, and 3 format so that the team could start using and identifying potential issues that will come into play when the full databank is established.

So at this time the Subgroup's decision is to delay any effort to establish the databank with GHCN-M and GHCN-D, and focus on test datasets as described above. Starting points for this could be the U.S. Forts Data which was imaged and keyed by NCDC's Climate Database Modernization Program (CDMP). There is also data from Mexico that has been imaged and in process of being keyed - but not yet integrated into a dataset. PT also suggested the use of ocean data as an example. JL and PT to initially address.

Need for Mirror sites

Aspects of mirroring include political – will providing mirrored sites throughout various parts of the world help entice countries to provide their data if it will be accessible from a site closer to their country? There was general agreement that this will likely not be of help in motivating countries who have other reasons for not releasing their data. RH discussed issues in Pacific with many island nations being reluctant to release their data without charge.

Most important issues related to mirroring are need for offsite backup and concerns about performance. At this early stage it is difficult to say whether the system at WDC-A/NCDC would have difficulty handling the load of databank users. Regarding need for offsite backup – Data are routinely backed up at WDC-A/NCDC so data loss is no longer a concern. So the primary issue would be the ability to keep the databank running if systems were to go down at the primary site. Another benefit to a backup site would be the capacity to provide other modes of data access that are not necessarily available at the primary.

SW offered NCAR as a mirror site. Either Dark (backup only) or a site that mimics the primary and provides same or different modes of access.

Format of databank

There was general agreement that the data would exist firstly in ASCII format, but that there is clearly a need for NetCDF, particularly as the databank moves toward sub-daily data.

Regarding the use of relational databases – there was general agreement that they are best used for metadata but less so for the data itself – until reaching the point of integrating into data access and visualization tools. This will be the role of the future constructed Data Access and Visualization Subgroup. Decision is for now to focus solely on ASCII and NetCDF for the database.

There was also agreement that the focus of the Databank Subgroup would be on Monthly and Daily timescale data during at least Year 1 of this project. Sub-daily data to be addressed in the future.

Units: There is a need for defining how units will be handled. Recognizing that original source data is often in imperial units and others such as Beaufort Scale observations. This project needs to make decisions on issues such as how many digits to carry and rounding policy. Rounding may be patterned after the US National Data Stewardship Team recommendation that was recently summarized in a paper authored by Mike Palecki – round half up asymmetric.

Stage 1 Format: Some discussion of the wide use of Microsoft Excel by data providers. There is a need to convert from Excel to something like .csv because in years to come as technology evolves it may not be possible or easy to read a file in Excel format.

Data Provenance and Version Control

It was agreed that methods for maintaining Data Provenance (the process of tracing and recording the origins of data and its movement between databases or stages of data) requires careful planning. As part of this version control is an essential aspect of the databank and must take place across all six stages of data. Data at the station level will need to be versioned and updates to a station tracked by a version numbering system. Stations may exist in a variety of versions. While individual station updates may take place on a continual basis, it is envisioned that the collection and merging into Stage 3 data will take place at distinct points in time – potentially semi-annually or some regular basis.

It was recommended that a Work Team be established that could meet external to the broader Databank Subgroup with the intent of developing guidelines and processes to establish how data provenance will be ensured. The following people volunteered to participate on the team.
Steve Worley
John Christy
Jeremy Tandy
Jay Lawrimore will look to bring in someone from NCDC with version control expertise.

Steve offered to provide documentation from the ICOADS experience that may establish a foundation from which to approach version control and data provenance for this effort.

Data Rescue and Crowdsourcing

The ability to digitize the tens of thousands of imaged forms that exist will be key to this effort. Given the cost of paying (in the US 7 cents per key stroke) for digitization, it is imperative that other opportunities involving crowd sourcing be employed. PT is working with Zooniverse personnel to grow this capability. Is also interacting with IEDRO. (This has potential to create large numbers of versions for a single station – with 20 or more volunteers for example keying a single image – resulting in 20 or more versions for the Stage 1 data for this station.)

Zooniverse efforts which can be built upon are taking place now in the ocean community – with ship logs being keyed by crowdsourcing – see http://www.oldweather.org/

Sources of land surface imaged data include those at NCDC. NCDC maintains imaged data in archive using a system called EDADS (Environmental Document Access and Display System). It is principally designed for archive but also to display document images over the internet and contains approximately 40 million images of original weather records and documents organized in "libraries" within distinct categories. Additional information is available at http://www.ncdc.noaa.gov/oa/climate/cdmp/edads.html . At this time EDADS accounts are somewhat restricted because the interface is not designed to handle large numbers of people. Access is available to U.S. government employees and their contractors, educational institutions doing environmental research, and other researchers associated with NOAA projects. Members of the Databank team can be given access.

Imaged weather data are also available at the NOAA Central Library – See the Foreign Climate Data at http://docs.lib.noaa.gov/rescue/data_rescue_home.html . Imaged data from more than 60 countries is available in time periods that cover ranges from the 1830s through the 1970s with most data from the period prior to 1960.

Subgroup recommends establishment of a Working Subteam for data rescue. Led by Peter Thorne and including the following.
Rod Hutchinson
Potentially Stephan Bojinski (WMO)
Rob Allan (ACRE)
Dick Crouthamel (IEDRO)
Manola Brunet (Univ. Rovira i Virgili, Tarragona, Spain)

Offers of Data have already been made.
Provided in advance of this call: Vyacheslav Razuvaev (Slava) has prepared a list of countries from WDC-B from which it is likely possible to find data in documents stored at WDC-B. Slava has provided examples of scanned data from India and has scanned data for more than 50 countries. The quality of documents varies widely. This needs to be considered for potential use and way forward in addressing.

Others have offered data which exists in digital format and could be sent to WDC-A in Stage 1 format. No discussion of a timeline for doing this was discussed.

i. Offered at Exeter in September: 300+ stations from China (Qingxiang Li)
ii. Argentina (Rusticucci)
iii. KNMI (Klein-Tank)
iv. NCAR Archives (Worley)

SW recommended and Subgroup agreed that **we need to begin with development of an inventory of the data that currently exists.**
This will be relatively easy to provide for the GHCN-M and GHCN-D datasets. Jay L. will coordinate. RH also has an inventory of data from Pacific that can be provided.

Providing an inventory of data in imaged format much more difficult. Currently the small amount of inventory at NCDC in the CDMP program is limited to inventory by country and number of pages of imaged data. Little to nothing in the way of station information, elements, or years of data. To the extent that an inventory of imaged data exists, it will be provided. Peter Thorne to provide.

Discussion of interaction with WMO to address issues brought up in Exeter by Near-real time team
The principal issues that need addressing in the long term are:

Change World Weather Records from decadal to Annually (seek opinion of countries during 2001-2010 WWR effort was recommended during workshop)

Initiate daily CLIMAT bulletins

Metadata requirement on CLIMAT messages

JT suggested that the WMO CBS Expert Team on Data Representation and Codes (ET-DRC) led by Simon Elliot would be an appropriate place to engage. Stephan Bojinski may have already initiated contact. Jay L. to communicate with Stephan about this firstly.
Jeremy also offered to be involved as chair of Expert Team on Metadata and Data Interoperability (ET-MDI)

Establishing Momentum
PT suggested that meeting minutes be provided online via the surface temperature .org website http://www.surfacetemperatures.org/ . All agreed.

PT also proposed a workshop be considered to bring together Subgroup members and other key individuals who are essential to building momentum and moving this effort forward.

After some discussion it was generally agreed that this is important and should be considered in more detail after some headway is made by the Working Subteams. Once more is done to develop scoping document, to investigate approaches and learn a bit more about potential hurdles and possible methods of approach we will be in a better position to establish a meeting, its objectives, and the right participants.