Data rescue task team call 6/22/12 8EDT (12Z)

Present on call: Peter Thorne, Michael de Podesta, Jared Rennie, Rob Allan, Hermann Machel, Rick Crouthamel, Stefan Bronnimann

Apologies in advance: Manola Brunet (may be in another meeting, will try to join), Juerg Luterbacher (undertaking oral exams)

Welcome to Michael de Podesta - guest from steering committee

1. Review of action items from last call

Regarding crowdsourcing:
ACTION: PT/JL to send the surface proposal to RC.
Done.

Regarding certificate of appreciation:
ACTION: JL has e-mailed the certificate to the group.
Done. If anyone would find such a certificate useful please advise.

ACTION: RH to chase about what, if any data can be released from NIWA holdings.
Unknown status.
RA queried what issue was. PT clarified that it is over the right to rehost and repackage. The data are available.

ACTION: Stefan to ensure data is uploaded to databank.
Pending. Will be done. Swiss Digihom data is in the databank. Stefan will check to see if there is anything else to send.

2. Update on databank progress - Jay / Jared

The databank is almost at the point of first release. Its still not (quite) too late for ingesting of new submissions and we would encourage these still at this point. The databank working group have been engaged in development of the merge methodology. We will not discuss further than in the briefest of details on this call. However, for those interested the latest method and some breakdown of results can be found at: http://www.surfacetemperatures.org/databank/merge_update_20120606.pdf?attredirects=0 . Further discussions can be had offline with Jared / Peter / Jay.

The merge is an iterative process of comparing a master source with a candidate source. Sources have been ranked according to provenance, max/min availability and our knowledge as to whether QC / homogenization has been applied. The top source is thedaily holdings GHCND which we are designating as the daily stage 3 source. This ensures / enforces consistency between daily and monthly holdings. Geographical metadata is used to narrow potential matches. Data comparison metrics are used to provide additional discriminatory power. For any station in the candidate three eventualities can arise: 1. merging with an existing station, 2. pulling through as a new unique record and 3. withholding due to insufficient clarity as to what to do or because the record is redundant.

Lots of work is ongoing to quantify uncertainties and sensitivities before release. Release is likely to consist of a 'default' and several variants. This recognizes the vexed issue of the challenges and that there is no unique correct solution. The availability of variants will allow analysts to explicitly quantify downstream processing uncertainties in the context of merging uncertainties. Two papers are planned to result - one high level and one technical (we are considering Rob Allan's new geoscience data journal for that if appropriate). With the call in brief email were appended three figures. These gave an idea of coverage, length of record and US vs. Non-US record availability.

RA: With ACRE data we are going to be putting a batch to ISPD. Can make them available. Not clear how much temperature data there is and there are always other elements regardless.
JR: Have been taking any elements at this point, but only pulling through the temperature. Anything is fair game. Will take whatever you offer.
ACTION: RA to follow up with JR on data sharing this ACRE effort over e.g. dropbox.

3. Updates on data rescue activities

3.1 Attempts to garner funding from philanthropic sources to engage crowdsourcing capabilities - Michael

Michael has worked with various members of the databank working group to produce a two-side flyer with the aim of engaging philanthropic support for the creation of a crowdsourcing portal. He is working with a colleague with significant experience in this arena to try to push this forwards.
MdP: I would like to clarify how people see the balance between the imaging of the archived paper documents and the extraction of data (Level 2 & 3 I guess) from the existing imaged archives.
PT: They are two sides of the same coin. Citizen scientists need to see pull through to digital holdings and use, including novel analyses of their data otherwise engagement will fall off a proverbial cliff. So, its necessary to have resource to do pull through and some science analysis. Have viewed imaging as somewhat distinct in efforts to date - its supply-side to the crowdsourcers.
RC: Also a large amount of strip chart data rather than print / written values
RA: Can be automated?
RC: This is a somewhat complex problem.
MdP: oldweather.org has been done by small core mainly.
PT: We should look at the paper records first.
RC: Special niches can make a difference. Some people this makes a huge difference to their lives.
RA Link to BADC WWW site with images that ACRE is having scanned by the Met Office Archives:
http://badc.nerc.ac.uk/browse/badc/corral/images/metobs

RC: We have funding to bring microfiche (3 million pages) from Africa to image form. 1916 onwards. Issueues remain over then making the case to share these but rescue is, as always, the primary need. Rescue data before its too late. But digitization is something that needss work.
RA: Particularly good records from former French territories.

3.2 NCDC archive data discovery and imaging - Jay / Peter

NCDC recently started trying to inventory and image some of the international holdings in their basement. Just over half of the boxes have been inventoried at this time. A number of records have been imaged. These images will shortly be posted online on the databank ftp area. We will then promote their availability and hope some of the data can get digitized by third parties. For now further progress is slow but we are trying to engage all potential avenues to regain momentum.
Do you have a headline cost for imaging?
This is vexed. We are working on getting volunteer effort but there are legal issues to so doing. So it may have to be paid work to acheive this. It would be competitive tender.

3.3 New data rescue ingested into databank - Jared

Juerg Luterbacher has provided substantive data holdings which are in the process of being added to the databank and will be included in the first version release. 17 Stations from the 19th Century / Early 20th century.

3.4 Updates from TT members on relevant activities (round the call in, those unable to attend can edit the pad in advance or email submissions)

An update on the joint EURO4M and MEDARE effort for enhancing data availability and accessibility over southern and Middle East Mediterranean countries and some constraints on data accessibility.

Under the EU EURO4M project (http://www.euro4m.eu/Data_archaeology_in_the_Mediterranean_region.html), the C3 (Centre for Climate Change at URV) is recovering and digitising ancient fractions of daily data for 65 (67) temperature (precipitation) locations and 36 stations with SLP data at the hourly scale from several on-line repositories and physical archives. The fractions of these series covering an elapsed period between the 2nd half of the 19th century till around the 1970s. First stage of this effort is being the recovery and digitisation of the network, which is expected to be finished on late Oct/2012. In parallel, we are passing on to the series digitised a complete QC, which will be over also by the end of Oct. All these data recovered/QC'ed under EURO4M will be made freely available to ISTI and to other global and regional databanks once the whole exercise is over (before March 2014, although most of the fractions recovered will be sent before 2013 ends). The Second stage of the effort is intended to merge the ancient fractions with more recent observations already available in other accessible databanks (e.g. ECA&D, AEMET data over north Morocco) and from NMHS in the region via a data exchange exercise in order to homogenise some of the longer and more continuous records. This task is expected to be finished in late March 2013, but for some Libyan and Jordanian records will last longer, likely.

Linked to MEDARE, we have been proposing to all the NMHS in the targeted Med sub-region a data exchange exercise with some success, since the Algeria, Cyprus, Libya and perhaps Jordan NMHSs have agreed to exchange with us their present digitised data by our digitised old records. However, accessibility to the series developed when they include recent data from NMHS will be likely restricted to their use in EURO4M and to the MEDARE databank. This has been the restrictions put by the NMHS that have agreed to exchange the data, since data sharing and accessing is still a big issue in this region. Finally, we have past on to CDMP staff a comprehensive list with the problem encountered with poorly scanned holdings. Regards, Manola

Update from ERACLIM: Digitizing will be finalized end of June. Metadata on ERACLIM data rescue metadata base (mostly upper-air): http://www.oeschger-data.unibe.ch/metads/
Username and password available from hosts.
Not much surface temperature, mostly upper-air.
ACTION: Jared and Stefan to engage on ERA-CLIM data rescue mid-July

RA will share some data with Jared. October deadline but some may be available in time for 1st release of databank (see action above)

Hermann can deliver c. 100 stations. Will follow up with Jared Rennie.
ACTION: Jared Rennie to follow up with Hermann Machel on this German data

Rick - Sharon Leduc went down to Bolivia - the Bolivian Met service will be working with IEDRO on data rescue and digitization.

PT - Central American digitization - follow up with RA.
ACTION: PT to advise Rob Allan of the Central American effort.

SB: Thanks for emails of support on Meteoswiss.

4. Updates on WMO GFCS discussions

Caveat emptor: this is all still to be discussed and agreed. Current draft texts have encouraging noises viz. coordination of and promotion of data rescue activities and also the sharing of historical data archives. What comes to pass should be clearer by the next call. This is simply noted as a potential brighter dawn tomorrow point for now.

RA: We (me and RC) went to a meeting in Geneva.

5. AOB

RC: Data sharing is seen as a valuable commodity. Data on paper is not worth the paper its written on, give us the data and make it available. Then give something in return to make value added products.
PT: Climdex software may be available. PT will chase up next week with Lisa Alexander.
ACTION: PT to ascertain what Climdex software would be available to IEDRO and advise RC.

Suggest next call in September when summer holidays are done and dusted.