Data rescue task team meeting 4/11/11

Present: Peter Thorne (PT), Tom Ross (TR), Stuart Lynn (SL), Rick Crouthmael (RC), Rob Allan (RA), Manola Brunet (MB), Stefan Bronnimann (SB), Hermann Michael (HM),

Apologies received: Rod Hutchinson (in East Timor), Juerg Luterbacher, Masumi Zaiki (both technically unable to call in), Jay Lawrimore (on leave)

Agenda

1. Introductions (all)
2. Progress report
    a. Inventory of known efforts ongoing (see attachments)
    b. Zooniverse visit to NCDC
    c. Google.org discussions
3. Coordination of ongoing efforts
4. Pull through of data rescue to databank
5. More generic databank issues
6. AOB

Databank synopsis

The end aim of the work of this team and others is the creation of a single comprehensive repository of land meteorological data available without restriction. The data will have provenance tracking and be version controlled. It is envisaged to consist of four stages:

- Stage 0 hardcopy / image / digital count (for automated instruments)
- Stage 1 – keyed in native format
- Stage 2 – converted to common format with provenance flags
- Stage 3 duplicate merged

All stages will be available. Metadata to the extent available will be associated with the databank. Some initial data is available through www.gosic.org (http://www.gosic.org/GLOBAL_SURFACE_DATABANK/GBD.html) which you are encouraged to look at in advance of the call. The databank forms a part of the larger International Surface Temperatures Initiative (www.surfacetemperatures.org) but would also have a myriad of additional benefits far beyond this initiative.

Progress report

Inventory of known ongoing efforts

Group members have iterated an excel spreadsheet that gives a summary of known projects in the area of data rescue at the country / regional scale. ACRE/ISPD and

Stefan Bronnimann's group (the latter mainly for Upper Air records) each have a spreadsheet where this is split down to the station level.

Zooniverse visit to NCDC

Stuart Lynn of zooniverse visited NCDC for three days in March with the visit funded by the UK Foreign Office. Central to the visit was discussion of ways to use the zooniverse model to try to crowdsource digitization of the imaged records held first and foremost in the NOAA Foreign Data Library. The key challenges were seen to be the heterogeneity and density of many of the images and how to keep volunteers engaged. Possible gaming aspects or themes were discussed. Following up depends upon likely availability of funding. We are considering possible funding mechanisms but there is nothing concrete at present. It is planned to write up a brief summary with some potential pathways forwards this month if time permits. Other options that were suggested included the use of recaptcha or similar software to get digitization 'for free' but zooniverse do not have direct reach into recaptcha.

RA: Trying to work out how to get additional information which acts as a hook and keep people interested is going to be keyed. How do you get folks interested and keep them interested?
SL: Agree it's a very different challenge.
SB: Upper Air data that are imaged in NOAA FDL are being done for ERA-CLIM. We don't need to focus on any non-surface data. The UA data are technically most challenging.

General agreement that this is worth continuing to pursue. Recognition that onus is now on us (all of us) to fund these efforts.

Google.org discussions

Some initial discussions were undertaken with google.org around both digitization of imaged data and imaging of the 2000+ boxes containing various international holdings. A two side document was prepared and sent to google.org and a response is pending at this time. Discussions revolved around recaptcha type of approach to digitization and the use of some of the imaged records in an educational context with the digitization a side benefit. It was recognized that the volume of data and range of complexity, formats and languages means that there have to be multiple solutions. Key will be parsing the data in such a way that unnecessary redundancy is avoided.

ACTION: PT to follow up on this. *Done. Google are still interested in scoping but a definitive response is still some weeks away.*

Coordination of ongoing efforts

There are significant ongoing efforts at data rescue of surface data by numerous teams and organizations. There also exist numerous digital archives at present which the over-arching databank group is trying to reconcile. This leads to a real risk of redundancy of effort either multiple groups digitizing the same hardcopy record or hardcopy records already available digitally being archived (although note that this may not be entirely redundant if the provenance for the pre-existing data is dubious).

- Is it plausible to aim for a master listing to be held somewhere?
    - Technological solution?
    - Owner and associated overhead …
    - Does this group have the knowledge base or will it still be a finite sample of what the data status is?
- Is this group a sufficiently large sample of the community that it can act as a way of verifying data status for a given station and should we countenance being such a clearing house?
- Do we have a clear handle on what the publicly available databases already in digital form are? Should we look to collate those first so that we avoid rescuing data already readily available? There may exist many local / national / regional data archives that have yet to be bought in to the major global holdings.

TR: Redundancy is the issue. Don't want to duplicate efforts. Only a limited resource. Key is to know. Hourly, daily, monthly. Very smart effort is required.

SL: Don't waste people's time is the key – both for citizen scientists but also professionals.

RC: Web page offered by IEDRO to coordinate rescue efforts. Paper on shelves – started with image everything, no inventory to cross-compare. Now they inventory on paper and order, take photos of data types and then image only the data that isn't held in the databanks. Laborious process and the two step approach adds a delay.

RA: ACRE is trying to do this for surface pressure data. ISPD. Other variables in that archive. BADC is placing the images up for terrestrial soon. ACRE master list of several thousand stations will also go up including current status. We should link with this effort and others. Chile, India, Pacific focus with ACRE presently.

MB: Must ensure we avoid redundancy in scanning and digitization. We should be trying to put together such a list. WMO website to develop master listing? WMO DARE team to compile a comprehensive master listing? National level activities. We could be the team to lead it?

SB: Really important to have a master list. Who does it is a secondary priority. We never know who does what otherwise. Many duplications in the past. Ongoing work. Once we have a master list it has to be updated. Annual updates?

HM: We need such a master list.

RC: Master list idea: we had approached WMO at a meeting some time ago. Limited success. Go through ACRE or other solution to do this.
RA: Not just NMSs calls for ACRE solution.
PT: Tech solution. Is there a way to allow identified users to update the list interactively?
PT: Can we augment that ACRE list? RA: Yes.

General agreement that if ACRE are willing to act in this capacity we should build on their effort and use their existing structure and protocols rather than pointlessly reinvent the wheel. Aim is to create an inventory for land data that is truly comprehensive and (quasi-)dynamic and propagate to the community. Having a single go-to resource is key. Need to be very careful to fully acknowledge that this is an ACRE led effort that the surface temperature initiative are augmenting rather than vice-versa.

ACTION: TT members to augment the ACRE list with any known activities at the station level before next call and send back to RA.

Pull through to the databank

- Is this a push process from the data rescuers or a pull process from nominal databank owners?
    - Databank Working Group lead Jay Lawrimore will prepare some guidance, largely aimed at pre-existing digital archives in the first instance so this is an opportunity to provide input.
- What is the mechanism?
    - ftp
    - cgi script
    - other?
- Should there be a realistic timeframe expectation on appearance in the databank?
    - Data rescuers have period of grace to undertake exclusive analysis?
    - Need to show pull through to keep people interested?
- What is it realistic to expect data rescuers to provide?
    - Minimum would be the digitized data but this would be seen as incomplete.
    - Aim would be the image of original and all intermediate steps and metadata

TR: Minimum requirements on metadata. Pre-existing digital archives. Data keyed and metadata is key thing to resolve a priori. Location of station has to be the absolute minimum otherwise it is useless information.

SL: The more metadata that comes is better from citizen science. Data format, what it is etc. which keeps the interest.

RC: As we get the paper data imaged we retain the image of the paper data on hard drives until we know what we want to do with it. Strip record precip data is being digitized by a program. Brings a question on barograph, thermograph at 5 minute intervals. Where do we archive? Is there interest on archival at this interval in the community?

RA: Data to ISPD is formats issue. Sending in multiple formats is a pain for the databank collators. Provide what you can and then transcribe? Images are being maintained locally by ACRE at BADC. Do we want the image repository?

PT: Can we use BADC as a perpetuity image repository rather than hold?

MB: Really desirable to have all the data. EURO4M, ERA-CLIM, MEDARE images of N. Africa being digitized from NOAA Foreign Data Library and other sources. Local law is an issue with making images available when putting all the info online in some cases. Need a period of grace to allow project rescuers to analyze for some data.

TR: Need to clarify legal aspects in advance.

SB: Images should be made available where possible. Level 1 data is also important. In the past we have had issues over website access. Append links to data owners. Convince of benefits of making data available.

HM: Make images available wherever possible.

It was agreed that until we saw guidance on how data were to be submitted to the databank we weren't in a position to provide specific feedback. Open question to databank working group: do you need images stored locally to the databank or will perpetuity links suffice?

More generic databank issues

Databank Working group will produce a brochure and / or poster. What do we want it to say regarding data rescue? Are we willing to propagate this at conferences / workshops?

RA: CD/DVD being prepared for WMO congress by IEDRO / ACRE.
RC: DVD on importance of data rescue and digitization given to PRs at congress. Discussion with Omar Baddour. Cover letter from Sec. Gen. and former director of NWS. Why is important from viewpoint of NMSs. Biggest fear is that paper data is still being thrown out.

PT: WMO has a lot on data rescue in congress. Publicity? Posters? Is there a single point of contact from the DVD materials?
RC: IEDRO is initial point of contact on DVD. Youtube, poster. Can put youtube video linked from anywhere. Looking at added Value added products software availability.

ACTION: RC to provide PT with link when done so the DVD can be linked from the surfacetemperatures.org domain.

MB: ETTCDI has software for analysis that could be made available.
PT: NCDC have offered creation of normal products (includes threshold exceedance etc.).
MB: Use side meeting at congress on data exchange. Poster for data rescue team? Master list pointer?
SB: Earth system science data journal. Could we put in a journal article?

Databank working group is working to increase visibility and engender input. Are there digital archives that you know of that we can suggest they pursue? Do you have personal connections that can make this happen? Should we collate a database of known digital archives or simply feed ideas up to Jay Lawrimore? Are you willing to promote the databank concept in your region?

ACTION: All to email PT and Jay Lawrimore with any digital holdings of which they are aware.

Is there anything that the databank working group or over-arching initiative steering committee can do to aid the work of this task team?

Any other business

RC: What IEDRO are doing is trying to find money much of the time. Need to get some expertise on board to try to solicit funding to do the work we do. Need to diversify funding avenues. How do we manage this?
RA: Work closely with social scientists and humanities to diversify resources.
RC: Clearing house for funding avenues?
TR: Not clear what the federal budget will be.

RC: ACMAD microfiche 1,000 African sites – having trouble getting these. They may be the only version of records left in the world for many of these sites.
PT: Albert Mhanda from ACMAD is on the steering committee.
RA: Key is to work with historical NMHSs and engage – helps in leveraging more modern data.

PT: Next call in about 8 weeks. Will schedule via doodle again.
ACTION: PT to schedule next call.