Creating surface temperature datasets to meet 21st Century challenges

Met Office Hadley Centre, Exeter, UK

7th-9th September 2010

White papers background

Each white paper has been prepared in a matter of a few weeks by a small set of experts who were pre-defined by the International Organising Committee to represent a broad range of expert backgrounds and perspectives. We are very grateful to these authors for giving their time so willingly to this task at such short notice. They are not intended to constitute publication quality pieces – a process that would naturally take somewhat longer to achieve.

The white papers have been written to raise the big ticket items that require further consideration for the successful implementation of a holistic project that encompasses all aspects from data recovery through analysis and delivery to end users. They provide a framework for undertaking the breakout and plenary discussions at the workshop. The IOC felt strongly that starting from a blank sheet of paper would not be conducive to agreement in a relatively short meeting.

It is important to stress that the white papers are very definitely not meant to be interpreted as providing a definitive plan. There are two stages of review that will inform the finally agreed meeting outcome:

1. The white papers have been made publicly available for a comment period through a moderated blog.
2. At the meeting the approx. 75 experts in attendance will discuss and finesse plans both in breakout groups and in plenary. Stringent efforts will be made to ensure that public comments are taken into account to the extent possible.

RETRIEVAL OF HISTORICAL DATA

Peter Thorne[1], Scott Woodruff[2], Rob Allan[3], Amy Luers[4], Albert Klein Tank[5], William Angel[6], Russ Vose[7], Stefan Bronnimann[8]

[1]CICS-NC, NOAA NCDC, Asheville NC, USA
[2] NOAA Earth System Research Laboratory, Boulder CO, USA
[3]Met Office Hadley Centre, Exeter, UK
[4]Google.org
[5]KNMI
[6] NOAA NCDC CDMP, Asheville NC, USA
[7]NOAA NCDC, Asheville NC, USA
[8] Univ. Bern, Switzerland

White paper topics given:

- what other known data sources exist that are not part of the current databases;
- other potential sources;
- data version reconciliation between data banks;
- digitised records that are not made available internationally;
- a practical model for the data rescue effort (e.g. whether one or more workshops are required, the best mechanism to solicit data release);
- what other efforts are currently under way (avoidance of duplication);
- and the potential of crowd sourcing digitisation

Databank introduction – proposed content and structure

We restrict our databank discussion (this and the other three databank related white papers #4-#6 to be discussed on the first day) to land station meteorological records. Although the focus of the workshop is surface air temperature records we recognise that it is important to the maximum extent practical to create a holistic database covering other meteorological parameters, which will then be of interest to additional researchers and stakeholders. The expense of handling original (e.g. paper) records can provide additional motivations for seeking the most complete possible digitisation (i.e. keying; accompanied ideally by imaging) of the land data and metadata. The images form an integral part of any modern data archaeology recovery and exploration activity (see also e.g. digitization guidance within WMO 2002).[1] A working assumption therefore is that records for all land station parameters would be recovered and archived wherever possible and not just temperatures. We may also consider thermohygrograph and barograph data. In general, with a change of the research focus from the mean state towards extremes and from the thermal regime to

[1] For historical ship logbook data, for example, the marine community struggles with similar prioritization questions (Wilkinson et al. 2010):
"Difficult cost-benefit decisions must often be made on the scope of information to be digitised. For example, many older ships' logbooks contain 'remarks' (e.g. on employment of the crew and detailed navigation information) not directly connected to the coincident meteorological or oceanographic observations, but nevertheless of potential interest to historians and other non-climatic research applications – but in many cases digitisation projects for climate research have omitted these for cost reasons (in this case, however, having the above images readily available can partly satisfy requirements from other disciplines)."

the water cycle and energy balance and with better numerical techniques becoming available (e.g., in the field of data assimilation), historical data need to be re-valued and often re-digitised. *What we are proposing, in effect, is a land ICOADS (http://icoads.noaa.gov/) databank to provide the data (and metadata) needed to meet the challenge of climate service requirements in the 21st Century.*

In terms of the duration of the historical record, it is highly likely that monthly records extend further back than do daily records than do synoptic (i.e. individual or "instantaneous") report records. This is because monthly records were easier to maintain than the instantaneous records. In terms of quasi-global coverage monthly records likely extend back to the mid to late 19th Century, daily to the mid-20th Century and synoptic data to the mid- to late-20th Century. For certain locations much longer records at each resolution will be possible. Whilst finer temporal resolution data can be averaged up to coarser resolution data, it is worth noting that despite formal guidance many countries and institutions have utilised their own methods to calculate daily or monthly statistics from the individual observations. There is therefore the risk of introducing non-climatic effects if uncoordinated or inconsistent attempts are made to backfill daily or monthly records. The *databank* should ultimately therefore seek to clearly track (where it is possible to determine) the source of computed values (e.g. most simply if they were computed before the data were provided to the international databank, or in the future the databank may include capabilities to consistently backfill data, which data could then be flagged to that effect). Another issue is that the ideas about the optimal way to derive e.g. daily averages from three times a day observations may change over time and may be different among scientists. If the sources are available we can allow for such differences.

A critical adjunct to the data themselves is metadata describing amongst others changes in instrumentation, siting and observing practices with time. Outside of a handful of countries the availability of this metadata to researchers is currently poor to non-existent. But metadata is a key step in building confidence in the presence of breaks in the station series and therefore an integral component of subsequent processing efforts to create homogeneous timeseries (see white paper #8 on homogenisation). In general, the available metadata have not been archived originally with this goal in mind. Metadata must be in a consistent and machine readable format to be useful for most purposes. It may include a summary of a weather station in the observation network, synoptic hours of observation, units of measured elements, observation precisions, observing instruments and environment conditions of the observation site, whether the station is manned or an automated weather system (AWS), among other elements. For more specifications and requirements, refer to WMO profile of WMO Core Metadata. Photographic and other evidence would also be useful but hard to make machine readable.

In writing this position paper we are making an implicit assumption regarding the over-arching structure of the databank which we propose herein should be akin to the commonly used satellite data product levels:

- Level 0 Digital imagery of original hardcopy or initial digital count for automated sensors
- Level 1 Version of the hardcopy as originally keyed or data converted to temperature in native format

Level 2 Converted into a common format[2]
Level 3 Integrated databank

Higher level products spawned from level 3 may be:
Level 4 Quality Controlled
Level 5 Homogenised

But we would want multiple independently derived versions of these levels and they may be distributed whereas the first four versions alluded to above would be an integral part of the raw databank. The proposed databank structure, and its relationship to other key components discussed in remaining white papers, is outlined in further detail in Figure 1.

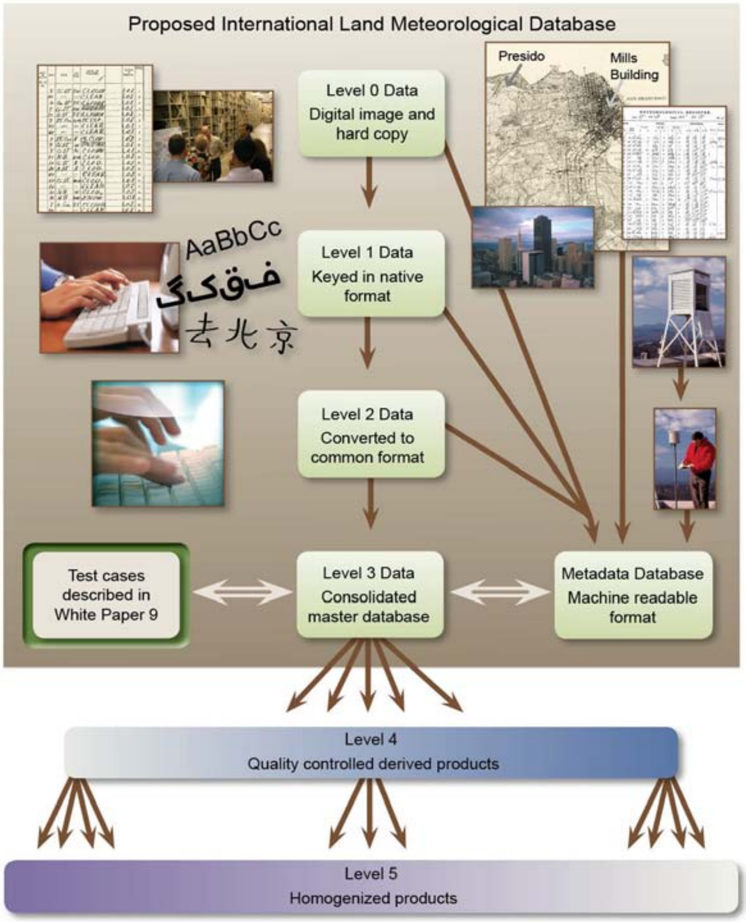

Figure 1. Proposed *databank* in relation to other downstream components of the international initiative. Components highlighted in light green constitute the proposed databank (figure courtesy of Deb Misch, NCDC Graphics team).

[2] We could consider the possible merits of development of a WMO-agreed format for level 2 data. In the marine sphere we are still working towards such WMO agreement, and for that purpose have developed a flexible format, including e.g. features to preserve “supplemental” (i.e. originating input) data. These features have proved extremely helpful to be able to recover from errors or omissions when data are inaccurately, or incompletely, translated from the input formats. See this long document for the details: http://icoads.noaa.gov/e-doc/imma/imma.pdf

For many stations there may exist multiple versions of level 0 and level 1 and perhaps even level 2 data and these may differ substantially in length and completeness as well as exhibiting substantially different behaviour or resolution characteristics (e.g. monthly, daily or synoptic). For example different versions of level 2 data might exist due to the lack of internationally standardised translations of ancient units[3]. The work in going to a level 3 product which most end-users would be encouraged to utilise in the first instance cannot therefore be under-estimated. For pre-existing records we may not have or be able to retrieve one or more of the precursor steps to level 3 data. In such a case hard decisions will be required as rejection of this data may badly compromise record completeness or spatial representivity or both. These aspects are discussed in more detail below and in accompanying first-day white papers, but as a key element of the databank design probably mainly it should be the end users who are empowered to make data selection decisions (e.g. through flags – but not actual rejection from the databank – indicating that data values are suspicious or good precursor data do not exist or have not yet been rescued). The ultimate choice made by the user will depend on the application in mind and the associated data requirements.

Databank in the bigger picture

The databank will constitute only one part of a bigger picture effort (levels 4 and 5 data, performance benchmarking, education, outreach and user tools and support) that is outlined in remaining white papers (and aspects of what follows are discussed throughout these). However, from a database engineering perspective this white paper's authors felt it is useful to consider up front in this first paper how the over-arching area could be managed.

One option that was discussed amongst the authors is a system of distributed management of datasets within a *DataSpace* including, but not limited to a central databank. This would involve having the multiple derived datasets managed independently in a virtual electronic "cloud" but accessed seamlessly within one working space – with collective visualization, collaboration and computation tools (Figure 2). From this space we would then optimally enable open access and contribution to the development of derivative datasets that would move from level 0 to level 5 (Figure 1). Each of the datasets within in the DataSpace could be have varying levels of public/private access for viewing and manipulation, and require tracking of version and QC.

[3] Also in the marine sphere carefully tested and well documented software libraries have been developed to share internationally for purposes of homogenizing e.g. ancient units translations as part of uniform format translations. We feel that a key underlying point is that the lower the level (e.g. 1-2) of the digital data, the more crucial it is that it be very carefully vetted and constructed, since it serves as the foundation for all subsequent work. This extends for example to keying, which should probably be done redundantly (or checked via Optical Character Recognition—OCR—where practical) to ensure greater accuracy.

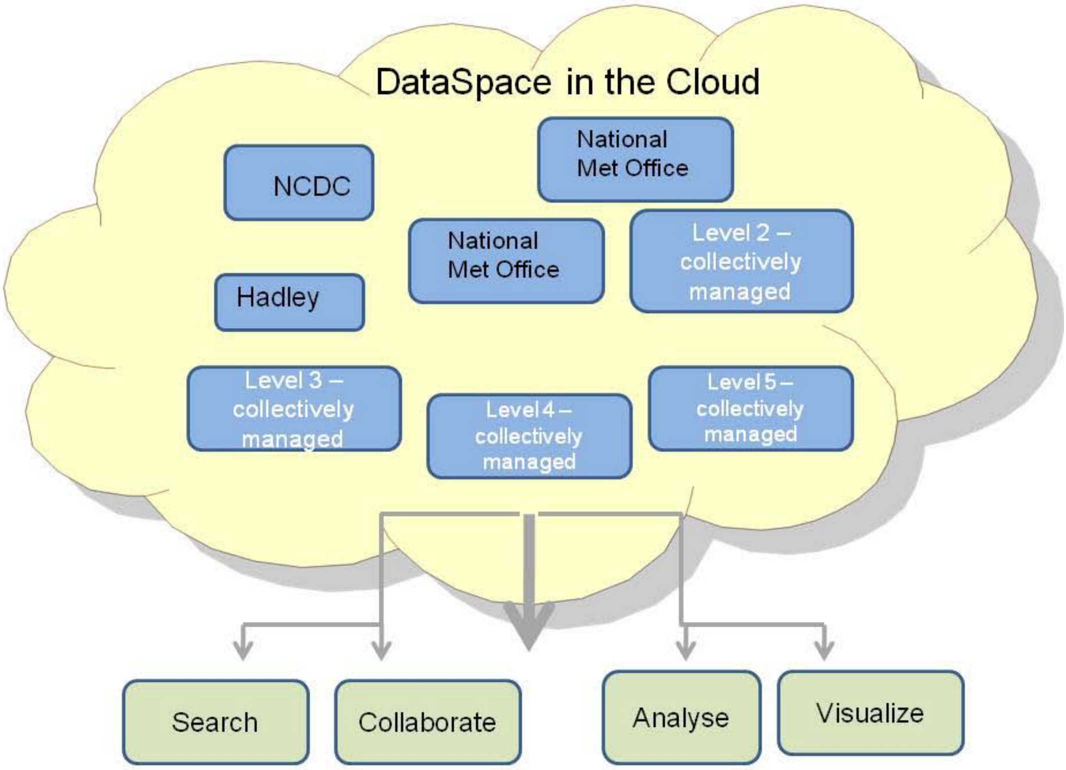

Figure 2. Proposed *DataSpace* in a cloud concept (courtesy of google.org).

A distributed cloud-based approach might help to:

1. Make this effort truly international and not owned by any single institution or entity – by providing each nation or institution the opportunity to manage and track the use of their contributed data
2. Enable broader access to data while providing tracking of how data is used – and potentially supporting micropayments for public uses as needed to meet needs to support data sharing and cost recovery.

We should also consider how these data could be best accessed by the climate services community as the global dataset is being improved. In other words, the climate adaptation and planning community should be able to access the most updated climate datasets available at any given time. This should be a living space.

Databank creation

There undoubtedly exist many paper records which are either available only in hard copy or in digital image form. Because paper records can be subject to deterioration, and to facilitate international distribution for research purposes, efforts should be made, building upon and partnering with pre-existing programs such as ACRE, to digitally image into robust archival-quality formats (e.g. tif, pdf) versions of hard copy only (or microfilm) records. Different options to organize and implement digitization efforts (level 0 to level 1) should be considered. In the ocean sphere for example, currently there are some institutionally funded efforts, some charitably funded efforts and initial efforts to “crowd source” (i.e. distributed outsourcing) of the digitization effort over the Internet. Crowd sourcing of digitization of the land data warrants further consideration. On the plus side it engenders public input, understanding and sense of ownership of the database. On the flip side there are overheads on some organization to quality check (or possibly independently digitize portions of) the data and prevent the possibility of deliberate manipulation by individuals or groups for whatever reason. What is obvious is that there is far more paper data than there is spoken resource to image and digitize it all. A caveat is required in that unless the process of taking the images, hosting them, provenance metadata (e.g. xml) generation and other aspects can be sufficiently streamlined and automated this creates a prohibitive overhead. This likely significantly restricts the number of possible hosting mechanisms for such efforts to a handful of technology organisations rather than science research groups in practice.

There currently exist three major databases, one each at monthly, daily and synoptic reporting frequencies at the World Data Center (WDC) at the NOAA National Climatic Data Center (NCDC). These constitute probably the most complete databases in existence and arguably would sensibly form the initial baseline from which to start creating an augmented databank. There also exist numerous other national and international archives including records used as ingest to reanalyses products. An estimate from staff at NCDC was that at least as much data exists to be added to their databases as currently exists within them (Stott and Thorne 2010), an estimate also made by the ACRE project. Some of this will be data already in digital form which the rights holders currently do not allow to be used or which simply has not been incorporated yet. Much of it will be available in only paper form and need converting to a digital record to be usable in climate studies. A (finite) listing of these additional resources ("Known knowns") follows:

1. There is a wealth of paper archives from around the world dating back to the mid 1800s that have been internationally exchanged but have yet to be digitized. See http://docs.lib.noaa.gov/rescue/data_rescue_home.html for scanned images of these available data.
2. Some early land station records still reside in paper form at the National Archives and Records Administration (NARA), and possibly at other national archives in the US.
3. NCDC hosts over 2,000 (in many cases) large boxes containing historical data from many nations that have never been fully explored or exploited. CDMP is in the process of creating an inventory and comparing to NCDC digital holdings.
4. Besides NOAA, many other National Meteorological and Hydrological Services (NMHSs) and probably national archives hold additional data.
5. Some nations have large Mesonets (of AWS) which may not yet have long term records but will get there (e.g. Oklahoma Mesonet; GASIR, Mexico)
6. There are networks not run by National Met. Services or by non-govermental organisations e.g. there is a large Brazilian network run by a non-NMHS.
7. Reanalyses ingest fields.
8. An overview of regional activities in this field is given in WMO/TD No.1480

There are doubtless other sources unknown to the white paper authors that could and should be pursued. Some form of prioritization will be required. Most logically in the longer term this would be driven by balancing science or societal requirements against current data holding availability. There would be little point in prioritizing regions such as the contiguous United States which are already very data rich. In the initial phase it would be sensible to go after apparent low-hanging fruit: countries with large data holdings that we think we have a reasonable chance of agreeing to share the data and would lead to a step-change in data holding size, or early colonial data that may substantially improve coverage in the very earliest years to build momentum behind the effort before tackling more complex and challenging cases. Perhaps the best way forwards to prioritise would be to start with a questionnaire to experts (workshop participants, GCOS national offices, ACRE mailing list, International Surface Pressure Data Bank etc.) and to WMO Permanent Representatives as to their knowledge of what exists that is not yet digitised and how easy or otherwise it may be to get at.

The effort required in reconciling data sources (to go from level 2 to level 3) cannot be under-estimated. This may be required to extend a given station's record or to blend data held in different holdings using a different identifier nomenclature. Quite often data from apparently the same station will differ between separate archives. A good example of this is in efforts by Andrea Grant and colleagues to digitise very early weather balloon and kite data (Grant et al. 2009). Here data were often available from several sources and often differed substantially leading to substantial issues and an interesting metadata challenge. Early work by the same group for surface data (unpublished) suggests that similar issues pertain to early land records. In theory the best solution is to retrieve, from each country, it's most recent collection of historical data as well as the metadata to go with it (local ID, WMO id, latitude, longitude, station history etc.). Some work at NCDC has commenced on automating this reconciliation step, but it is not certain that it will work or be applicable to temperature records.

Another challenging issue is that in many cases land records were digitized e.g. decades ago using what may now be considered inferior techniques, including the possibility that important data and metadata elements were omitted or incompletely captured (sometimes owing to early technological limitations). In some cases the only record that may exist will be the digital record in its current form. Agreement as to mechanisms to consistently handle such cases and reconcile (sometimes competing) priorities is required.[4]

We know that many countries hold digital data that are not made available freely. Often this is because of political and / or financial imperatives. Although such data in theory should be covered by WMO Resolution 40 on data sharing they are often exempted (see white paper #5 on data policy). Furthermore, not all data were collected by NMHSs (including many historical data, which may not as clearly fall under Res. 40) and therefore logically fall under their purview. The marine

[4] The ocean community are facing similar problems as documented in Woodruff et al., 2010: "Work continues to actively catalogue, image, digitize, and ultimately convert digitized data into the IMMA format. However, these are all expensive tasks, and better methods are needed for prioritizing the value of specific collections, and the scope of digitization, for different climate applications, as well as for related research disciplines, including oceanography, fisheries, and ecology.

Also important is the potential value of enhancing digital collections already included in ICOADS. There may be alternative data sources closer to original data than those previously used, or additional data elements such as early sea-ice observations."

meteorological community has benefited since the early 1960s from having an agreement (i.e. under the WMO Marine Climatological Summaries Scheme) for the exchange of digitized ship logbook data (as discussed in more detail e.g. Woodruff et al. 2010). Exploration of the feasibility of a similar agreement within WMO regarding release and exchange of historical and contemporary land data records would seem appropriate (see also white paper #5 on Data policy).

In terms of creating a truly international databank we should build upon the pre-existing databases (including the International Surface Pressure Databank) and expertise at NCDC and elsewhere rather than starting from scratch, and additionally build on the experiences of other data communities (e.g. marine as is already the case in this paper, radiosondes etc.). For political and practical reasons it is worth considering a greater internationalisation of the effort (including the possibility of mirrored data holdings at other archive centers internationally/WDCs). As discussed above, the creation of an international land surface databank that is not owned by a single institution would seem desirable, and novel approaches such as the DataSpace concept (Figure 2), potentially including hosting (or mirroring) through a non-governmental portal like google.org should be actively considered. Due attention should also be given to how this databank is officially recognised (e.g. formally through WMO). For example, in the marine sphere there has been some reluctance by countries to contribute historical data for ICOADS, without the assurance that those data and metadata (in some cases rescued at considerable expense to nations) would become part of a formal and permanent international archive. Therefore the Expert Team on Marine Climatology (ETMC) is developing a proposal for formal recognition of ICOADS through WMO and IOC (the Intergovernmental Oceanographic Commission), which would also allow for the possibility of other qualifying centers internationally mirroring the data (and products) and providing other complementary functions.

To engender data submission and input one or more (probably several) workshops would seem appropriate. The experience of the WMO Commission for Climatology (CCl/CLIVAR/JCOMM) ETCCDI[5] regional workshops and ACRE is important here. It is likely that we would be more successful in targeting efforts on data sparse regions and undertaking regional workshops. This has a higher overhead in terms of organization. Resources and a dedicated lead team would be needed in advance to make this a success. Otherwise a single large global workshop may engender input from many institutions. In reality a combination of these approaches may be required.

Current or recent specific activities that we should liaise with and ensure against duplication with are:

- WMO-CCl groups that have been active on retrieval of historical data / data rescue, in particular on a regional basis e.g. DAta REscue (DARE; http://www.wmo.int/pages/prog/wcp/wcdmp/dare/index_en.html), MEditerranean climate DAta REscue (MEDARE).
- The Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative for data rescue and facilitating reanalyses http://www.met-acre.org/

[5] The Expert Team on Climate Change Detection and Indices (http://www.clivar.org/organization/etccdi/etccdi.php) is joint among CCl, the Climate Variability and Predictability (CLIVAR) Program, and the Joint WMO/IOC Technical Commission for Oceanography and Marine Meteorology (JCOMM).

- The International Environmental Data Rescue Organization (IEDRO) http://www.iedro.org/
- NCDC's Climate Database Modernization Program (CDMP) http://www.ncdc.noaa.gov/oa/climate/cdmp/cdmp.html
- Data rescue efforts planned with the ERA-CLIM reanalysis
- Regional or national data rescue efforts, e.g. in Germany (KLIDADIGI), Portugal (SIGN), Italy (CLIMAGRI), Switzerland (DigiHom) etc.

We need to partner with rather than compete with these efforts, remaining open however to the possibility of facilitating new progress through wider and potentially even more integrated international efforts. What this effort should bring at least is some renewed momentum, which we hope will help open up access to a greater set of data, or at least achieve more seamless access globally (recognizing that some highly detailed historical data holdings at the national level are unlikely to become widely available in the foreseeable future).

Recommendations

1. A formal governance is required for the databank construction and management effort that will also extend to cover other white paper areas on the databank. This requires a mix of management and people with direct experience wrestling with the thorny issues of data recovery and reconciliation along with expertise in database management and configuration management.
2. We should look to create a version 1 of the databank from current holdings at NCDC augmented by other easily accessible digital data to enable some research in other aspects of the surface temperature challenge to start early. We should then seek other easier targets for augmentation to build momentum before tackling more tricky cases.
3. Significant efforts are required to source and digitise additional data. This may be most easily achieved through a workshop or series of workshops. More important is to bring the ongoing and planned regional activities under the same international umbrella, in order to guarantee that the planned databank can benefit from these activities. The issue is two-fold: first the releasing of withheld data, and secondly the digitising of data in hard copy that is otherwise freely available.
4. The databank should be a truly international and ongoing effort not owned by any single institution or entity. It should be mirrored in at least two geographically distinct locations for robustness.
5. The databank should consist of four fundamental levels of data: level 0 (digital image of hard copy); level 1 (keyed data in original format); level 2 (keyed data in common format) and level 3 (integrated databank/DataSpace) with traceability between steps. For some data not all levels will be applicable (digital instruments) or possible (digital records for which the hard copy has been lost/destroyed), in which case the databank needs to provide suitable ancillary provenance information to users.
6. Reconciling data from multiple sources is non-trivial requiring substantial expertise. Substantial resource needs to be made available to support this if the databank is to be effective.
7. There is more data to be digitised than there is dedicated resource to digitise. Crowd-sourcing of digitisation should be pursued as a means to maximise data recovery efficiency. This would very likely be most efficiently achieved through a technological rather than academic or institutional host. It should be double keyed and an acceptable sample check procedure undertaken.

8. A parallel effort as an integral part of establishing the databank is required to create an adjunct metadata databank that as comprehensively as feasible describes known changes in instrumentation, observing practices and siting at each site over time. This may include photographic evidence, digital images and archive materials but the essential elements should be in machine-readable form.

9. Development may be needed of formalized by new WMO arrangements, similar to those used in the marine community, to facilitate more efficient exchanges of historical and contemporary land station data and metadata (including possibilities for further standardization).

10. In all aspects these efforts must build upon existing programs and activities to maximise efficiency and capture of current knowledge base. This effort should be an enabling and coordination mechanism and not a replacement for valuable work already underway.

References

Grant, A, S. Brönnimann, T. Ewen and A. Nagurny, 2009: A new look at radiosonde data prior to 1958. *J. Climate*, **22**, 3232-3247.

Stott, P. A., and Thorne, P. W., 2010: How best to log local temperatures? *Nature*, **465**, 158-159

Wilkinson, C., S.D. Woodruff, P. Brohan, S. Claesson, E. Freeman, F. Koek, S.J. Lubker, C. Marzin, and D. Wheeler, 2010: RECovery of Logbooks And International Marine Data: The RECLAIM Project. *Int. J. Climatol.* (in press), DOI: 10.1002/joc.2102 [pdf available from: ftp://ftp.cdc.noaa.gov/pub/Public/swoodruff/].

WMO, 2002: 1. Report of the CLICOM-DARE (San José, Costa Rica, 17-28 July 2000), 2. Report of the International Data Rescue Meeting (Geneva, 11-13 September 2001). WCDMP-No. 49 (WMO-TD No. 1128) [http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/documents/WCDMP49_DAREReports.pdf].

Woodruff, S.D., N. Scott, D.I. Berry, M.A. Bourassa, E. Charpentier, S. Gulev, H. Haar, E.C. Kent, R.W. Reynolds, G. Rosenhagen, M. Rutherford, V. Swail, S.J. Worley, H.-M. Zhang, R. Zöllner, 2010: Surface In situ Datasets for Marine Climatological Applications. In *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society* (Vol. 2), Venice, Italy, 21-25 September 2009, Hall, J., Harrison D.E. & Stammer, D., Eds., ESA Publication WPP-306 (in press) [http://www.oceanobs09.net/blog/?p=452]