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

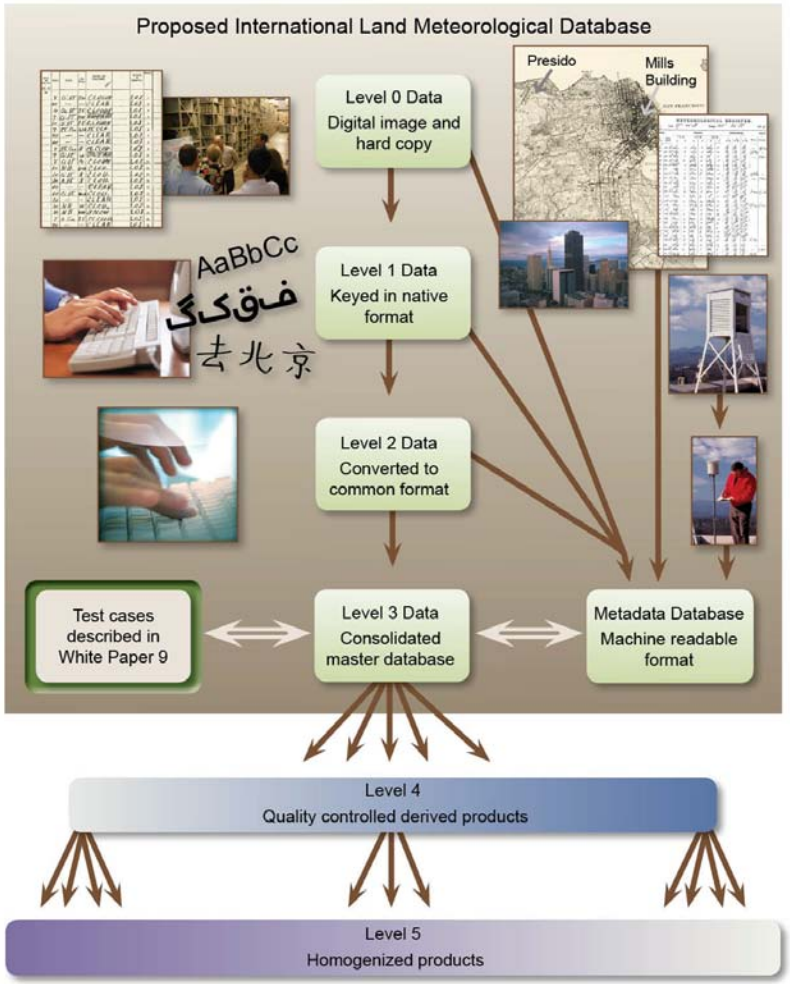
[5]KNMI

[6] NOAA NCDC CDMP, Asheville NC, USA

[7]NOAA NCDC, Asheville NC, USA

[8] Univ. Bern, Switzerland

What might the databank look like?



Current 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.