Databank Working Group

2011 Progress Report

October 2011

Current Members:

Jay Lawrimore (Chair)	- NOAA NCDC, USA
John Christy	- University of Alabama, Huntsville, USA
Waldenio Gambi de Almeida	- CPTEC/INPE, Brazil
Koji Ishihara	- Japan Meteorological Agency
Albert Klein-Tank	- KNMI, Netherlands
Matthew Menne	- NOAA/NCDC, USA
Matilde Rusticucci	- Univ of Buenos Aires, Argentina
Vyacheslav Razuvaev	- Russian Research Institute of Hydrometeorological Information
Madeleine Renom	- IFFC, Univ of the Republic, Montevideo, Uruguay
Jeremy Tandy	- UK Met Office, Exeter, UK
Peter Thorne (ex-officio)	- CICS-NCDC, USA
Steve Worley	- National Center for Atmospheric Research, USA

Ex-Members:

Rod Hutchinson	- Australian Bureau of Meteorology
Bryan Lawrence	- BADC, UK

New Members:

Meaghan Flannery	- Australia Bureau of Meteorology
David Lister	- Climatic Research Unit, East Anglia, UK
Albert Mhanda	- ACMAD, Niger
Jared Rennie	- NOAA NCDC, USA

October 2010 to October 2011 Objectives:

1) Invite members and establish working group communication including email lists, website and wiki.
2) Construct Terms of Reference for Databank WG.
3) Establish structure for a global temperature Databank and methods for data provenance.
4) Begin to populate Databank with Monthly and Daily timescale data.
5) Establish mechanism to support the collection of data; include data already digitized and non-digital through data rescue activities.

Objectives Met:

1) Invite members and establish working group communication including email lists, website and wiki.

Scientists from every WMO region were invited to join the Databank working group. Members participate on a completely voluntary basis and are responsible for providing leadership in identifying new sources of data and in providing expert guidance toward establishing and operating the Databank. Two individuals from Australia and the UK were replaced with other data experts from those countries, and there were additions later in the year from Africa and the U.S. The email list databank@surfacetemperatures.org was established to facilitated communications with all members between teleconferences which are held every 3 to 4 months.

A Databank Working Group page (www.surfacetemperatures.org/databank-working-group) is hosted on the ISTI website. This webpage provides background information on the purpose of the Databank and makes publicly available minutes of all teleconferences, documents that pertain to the development and operations of the Databank and its task teams.

The Databank WG launched a wiki to further facilitate communication between members and to serve as a reference for details on discussions and progress toward establishing the Databank. This wiki is open to all members as a means for tracking progress. Postings to the wiki can be made by any member using a unique login and password. http://editthis.info/intl_surface_temp_initiative/Main_Page .

2) Construct Terms of Reference for Databank working group

Databank working group Terms of Reference were written and agreed upon by all members. The TOR are hosted on the DWG website. The working group reported to the Steering Committee, giving a verbal progress report at each quarterly phone call and a written annual progress report.

3) *Establish structure for a global temperature Databank and methods for data provenance.*

The DWG agreed to focus efforts on Daily and Monthly timescale temperature data in keeping with an overall 6-Stage structure. Other elements and timescales will be collected if made available but will not be the focus of this effort for the foreseeable future.

STAGE 0: Digital image and hard copy
STAGE 1: Keyed in native format
STAGE 2: Converted into common format
STAGE 3: Consolodated master database
STAGE 4: Quality controlled derived products
STAGE 5: Homogenized products

The DWG is responsible for data collection and development activities including provenance and version control associated with establishing and maintaining Stage 0

through Stage 3 of the Databank. Development of quality controlled (Stage 4) and bias corrected (Stage 5) data will not be the responsibility of the DWG; criteria for assessment of quality control and bias correction methodologies as well as validation will be the responsibility of the Benchmarking and Assessment working group. However, the DWG will provide Stage 4 and 5 formatting and submission guidance and will be responsible for integrating these data back into the Databank as value added products.

The Databank is accessible from the Global Observing System Information Center (GOSIC) website (http://www.gosic.org/GLOBAL_SURFACE_DATABANK/GBD.html), and it is directly accessible at World Data Center-A (ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/) or its mirror site (ftp://ftp.meteo.ru/pub/data/globaldatabank/) which was established at World Data Center-B Oblinsk, Russia.

The DWB established as a high priority an effort to establish data provenance for all data to the greatest extent possible. While the working group recognized that for some data there is little or no information on its origin or history, a foundation was established for providing traceability of data through all stages of the databank. A Data provenance and version control task team was established and asked by the full WG to develop methods that would provide provenance and ensure version control of the Databank. The task team established Data Provenance Tracking Flags as the primary mechanism for documenting provenance and ensuring traceability in a manner consistent with the procedures established for ICOADS, although applied in a manner that meets the unique nature of land surface observations.

Five (5) Data Provenance Tracking (DPT) flags are assigned to each observation within the Stage 2 data files. These flags provide information on the origins and types of Stage 0 and Stage 1 data. The 5 flags are: (1) Stage 0 Source, (2) Stage 1 Source, (3) Data Type, (4) Mode of Digitization, and (5) Mode of Transmission/Collection. Additional flags can be added as the need arises, and the information contained within each DPT flag can be expanded as necessary to completely define a new type of observation.

This task team also investigated the potential uses for Unique Identifiers (UIDs), which are now being implemented in the ICOADS marine dataset. While the unique nature of ocean observations creates a need for the use of UIDs the team determined that they would not benefit the land surface databank. Additional information is available at http://www.surfacetemperatures.org/databank/provenance-and-version-control-task-team .

4) *Begin to populate Databank with Monthly and Daily timescale data.*

Because of contributions from many members of the DWB and others throughout the international community, the Databank was populated with numerous sources of data at both the daily and monthly timescales during the past year. Daily data are available in Stage 1 and Stage 2 formats for the following and as shown in Figure 1.

- Australia
- Brazil
- Channel-islands
- ECA/KNMI
- Ecuador
- ISPD (International Surface Pressure Data)
    - IPY
    - Swiss
    - Sydney
    - Tunisia-Morocco
- Japan
- Mexico
- Pitcairnisland
- SACA/KNMI
- Spain
- Uruguay-INIA
- US Forts
- Vietnam

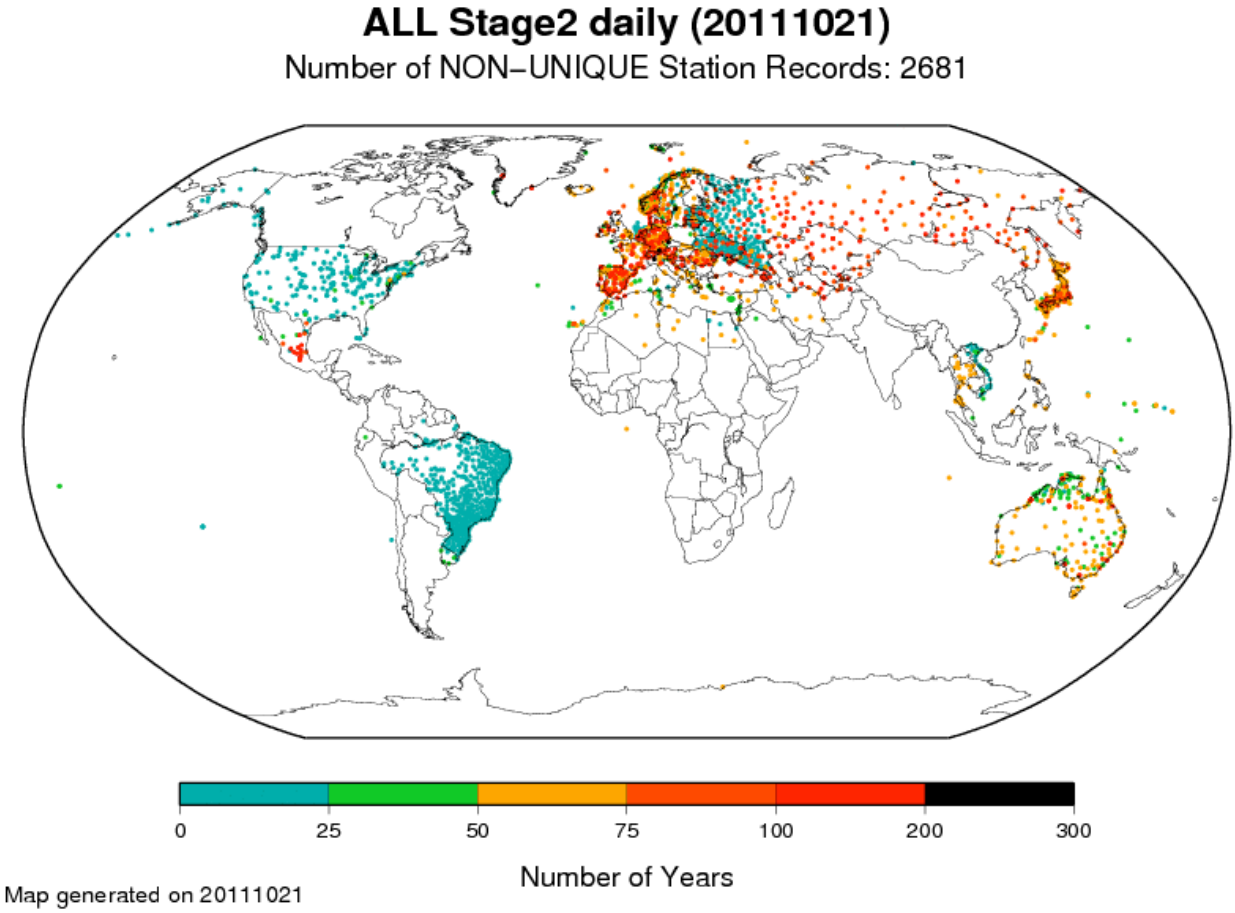

Figure 1. *Daily data submitted to the Databank by 21 October 2011, excluding data from the GHCN-Daily dataset or its 20 source datasets.*

Monthly summary data are available in Stage 1 and Stage 2 formats for the following and as shown in Figure 2.

- Antarctica-SCAR-reader
- Antarctica-South Pole
- Arctic
- Australia
- Canada
- Central Asia
- Colonial Era Archive
- East Africa
- GHCN-Monthly version 2
- GHCN-M version 2 source data
- HadCRU version 3
- Histalp
- Japan
- ECA/KNMI
- Russia
- UK Met Office historical
- World Weather Records

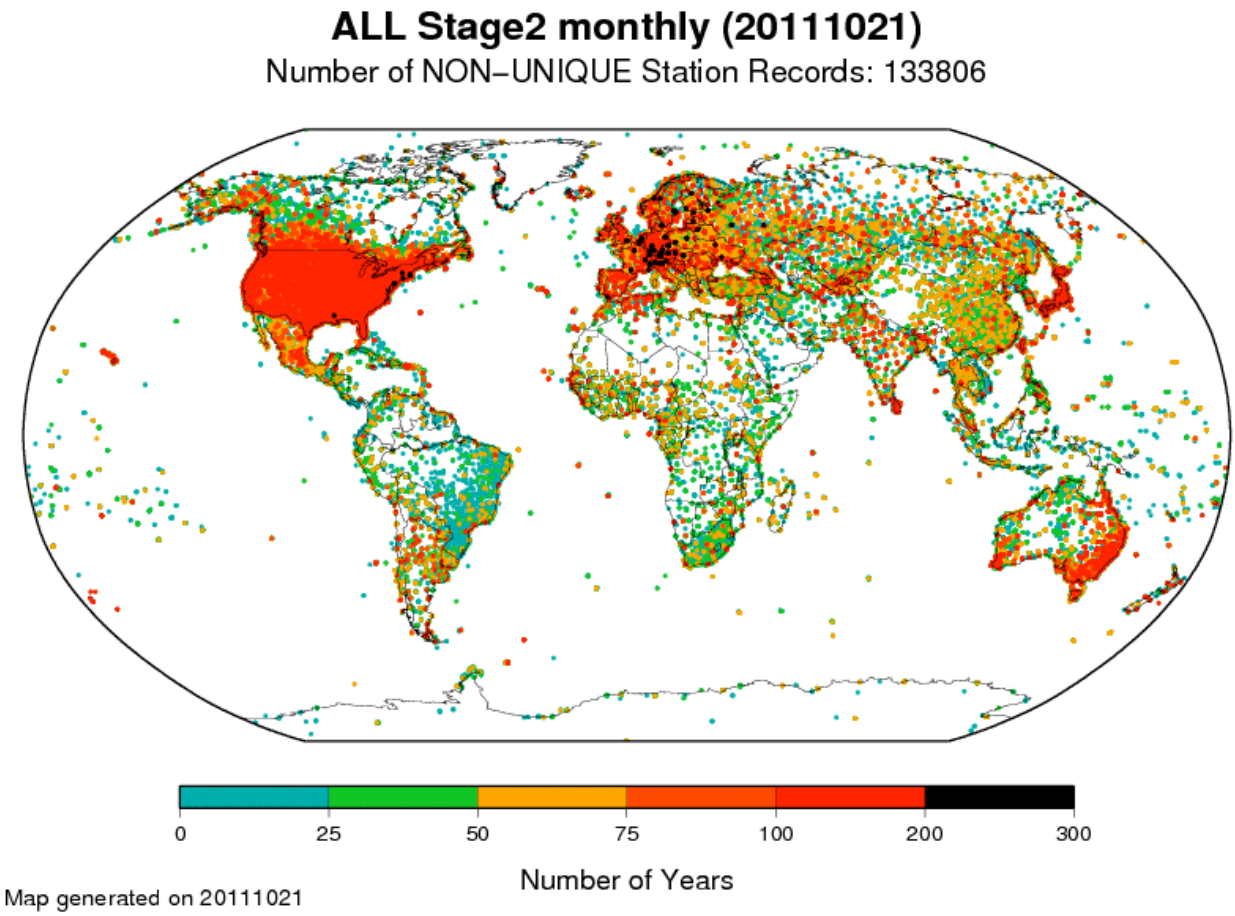

Figure 2. *Monthly data sources added to the Databank as of 21 October 2011. This includes daily data for which monthly summary data could be calculated.*

Each source dataset is held in its own subdirectory and a README file accompanies the data in the Stage 1 directory, providing basic information on the data provider and the data. An INVENTORY file is included in each source subdirectory within the Stage 2 directory structure. This file provides a list of each station for the particular source and other metadata including station id, name, location, elevation, and first and last year of data.

A limited amount of Stage 3 and Stage 4 data was added to the Databank on the Daily side. This was from the Global Historical Climatology Network-Daily dataset, which contains more than 25,000 stations with daily maximum and minimum temperature. This rich source of data is available for conversion to monthly summaries and future use on the Monthly side of the databank.

5) *Establish mechanism to support the collection of data; include data already digitized and non-digital through data rescue activities.*

To aid data collection activities the DWG developed a Data Submission Guidance document which provides instructions for data providers. It provides a description of the process for providing raw native digitized (Stage 1) data to the databank and encourages data providers to follow basic guidelines which will ensure data are accurately and efficiently added to the databank with the fewest complications. Information on essential and recommended metadata is also included.

A cover letter was prepared to accompany the data submission guidance document. This letter provides a brief overview of the International Surface Temperature Initiative and the role of the global databank within the wider context of the grand initiative. The cover letter and guidance document are available on the databank website: http://www.surfacetemperatures.org/databank .

Although those documents focus on data already available in digital format and most readily available, the DWG also established an effort to identify and collect sources of data through data rescue activities led by a Data Rescue task team. This task team includes members from multiple countries and data rescue activities.

The focus of this team is primarily on observations collected prior to the mid-twentieth Century and for which there likely exists as much data in non-digital form as exists in the current digital archives. The data rescue activities revolve largely around pulling through existing digitization efforts to the databank and attempting to ensure against redundancy of effort. This effort aims to leverage pre-existing programs where the credit for much of the data rescue clearly lies. An important area to capitalize on in 2012 will be the growing potential of Crowd Sourcing which has been used to great success by OldWeather.org in keying centuries old marine records. Tens of thousands of surface observation forms were imaged by the Climate Database Modernization Effort in the past 10 years. This provides a rich source of data that can be mined through global volunteer efforts aided by internet-based technologies.

Objectives Not Met:
None

Other Efforts and Achievements:

The opportunity to communicate the goals and objectives off the Databank working group has been taken at several international meetings. This is now aided by the preparation of a poster on the Databank which is available for inclusion in poster sessions and other venues of opportunity. Most recently this was included in a 4-poster display on the International Surface Temperature Initiative at the World Climate Research Programme Open Science Conference, which was held in Denver, Colorado, USA. This provided an opportunity to introduce many more people to the goals of the effort and led to the donation of several new sources of data.

A certificate of appreciation also is under development. This will convey to National Meteorological and Hydrological Services the gratitude of WMO's Commission for Climatology, the Global Climate Observing System, and the World Climate Research Programme for their support of the International Surface Temperature Initiative and broader efforts to meet 21st century needs for climate information.

2011 Annual Overview:

2011 was a year of successes for the global databank. From initial conception, a sound structure of design and implementation was begun. By the end of the year more than 20 sources of data had been collected and added to the databank as Stage 1 and Stage 2 data for both monthly and daily timescales. As the first year came to a close, discussions had begun for methods of data merging, and the coming year promises to be one of more successes as the DWG establishes a process for developing and launching the first version of the merged Stage 3 data. The working group has as its goal the launch of version 1 of the global databank in April 2012. Many things must happen before this occurs, but a continuation of the contributions from all members of the DWG is sure to make this goal a reality.

Objectives for October 2011 to October 2012:

1) Continue to add sources of Daily and Monthly timescale data to the Databank. Work with DWG members and others in identifying and collecting readily available sources of digital data.
2) Build upon Data Rescue activities and leverage crowd sourcing efforts to begin volunteer digitization of land surface records.
3) Complete position paper on version control and provenance available for public comment.
4) Develop an approved methodology for merging sources of data to create a monthly Stage 3 data product. Include a hierarchy of source data from which to build the merged dataset.
5) Launch version 1 of the Databank for monthly timescale data in April 2012, making all data, processes, and software freely available and accessible.
6) Complete and submit journal article describing version 1 of the Databank and its underlying principles.
7) Work with the Benchmarking and Assessment working group to expand opportunities for incorporating Stage 4 and 5 data into the databank.