ICOADS, Data Provenance and Version Control

Steven Worley and Scott Woodruff, 9 November 2010

The International Comprehensive Ocean-Atmosphere Data Set (ICOADS) uses overarching version control and a two-tiered internal data source-tracking scheme to document data provenance.

The versioning of ICOADS has evolved over time and has now settled on a numerical scheme as shown in Table 1 (Table I, Woodruff et al. 2010). Currently, the most advanced ICOADS is Release 2.5. Each new Release is a significant new reprocessing and merging (joining multiple data sources following pre-conditioning and duplicate elimination) where the full period of record or a major piece of the historical time period undergoes data processing, and numerous assessments of the output data characteristics and Release-to-Release intercomparisons are conducted for validation purposes (e.g. http://icoads.noaa.gov/r2.5.html). It is not unusual for the first in-house evaluation of a new Release to expose undesirable outcomes. ICOADS data processing workflows are designed so that re-processing is relatively straightforward.

Table 1. ICOADS Release history (update of Table I in Worley et al. 2005). For each Release, any references, the total temporal coverage, any updates and extensions, and the composition (for Release 2.0 forward, only), in terms of the delayed-mode (DM) and real-time (RT; GTS-only) archives, are listed.

Release name	*References*	*Resultant period (issuance year)*	*Updates and extensions (issuance year)*		
Release 1	Slutz et al. 1985, Woodruff et al. 1987	1854-1979 (1985)	1980-91 (1987-92)[a]		
Release 1a	Woodruff et al., 1993	1980-92 (1993)	1992-93 (1995)[b] 1990-95 (1997) 1980-97 (1999)		
Release 1b		1950-79 (1996)	1970[c] (1999)		
Release 1c		1784-1949 (2001)			
Release name	*References*	*Resultant period (issuance year)*	*Main update scope (other amendments)[d]*	*DM archive*	*RT archive*
Release 2.0	Woodruff et al. 1998, 2003	1784-1997 (2002)	(none)[e]	1784-1997	
Release 2.1	Worley et al. 2005, Woodruff et al. 2005	1784-2002 (2003)	1998-2002 (1784-1997)	1784-1997	1998-2002[f]
Release 2.2		1784-2004 (2005)	1998-2004 (1784-1997)	1784-2004	
Release 2.3		1784-2005 (2006)	2005 (1999-2004)	1784-2004	2005
Release 2.4		1784-5/2007 (2007)[g]	1998-5/2007 (1784-1997)	1784-2004	2005-5/2007
Release 2.5	(this paper)	1662-2007 (2009)	1662-2007	1662-2007	(ongoing)[h]

[a] Following Release 1 (Slutz et al. 1985, Woodruff et al. 1987), "interim" products, constructed using simplified procedures and preliminary input data, were first issued in 1987 (covering 1980-86) and then extended, on an approximately annual basis, to finally cover 1980-91.
[b] Woodruff et al. 1993.
[c] Minor corrections for October-November 1970.

[d] Main update scope lists the temporal range of reprocessed or extended data, while "other amendments" refers to the temporal range impacted only by less significant changes (such as the addition of QC or metadata).
[e] New Release nomenclature adopted for the combination of Releases 1c, 1b, and 1a (no new data or products).
[f] March-December 1997 observational data were also processed.
[g] The official R2.4 period was extended with preliminary observational data (only) through 7/2007 (by 6/2008), through 12/2007 (by 12/2008), and through 12/2008 (by 2/2009).
[h] As discussed in the text, the preliminary observational data (only) are now updated monthly.

_____

Every individual data record (all data reported at one time and location by an observing system) is assigned three internal metadata fields before becoming part of ICOADS, a deck number (DCK), source identification (SID), and platform type (PT). These metadata elements have been a huge asset, without question, to the development team and the data users. They are permanent traceable references on each record that are used to guide merging strategies and assess impact of one source on other sources. (*In addition to these crucial fields, we plan to add two additional fields, Unique report ID (UID) and Release No. (RN) to the IMMA format fairly soon; see Annex.*)

An illustrative example is given next. Deck (DCK) is typically the first piece of assigned metadata; it is the primary description of the data. In the example there are two sources from Russia, arctic drift stations, a focused experimental period (FGGE), a collection from Japan, the NODC World Ocean Database, and a NCEP GTS BUFR data source (Table 2). Besides the assigned DCK number and description it is very useful to also document the starting and ending dates, and the number of records involved. For example DCK 780 = NODC World Ocean Database, the data span 1772-2005, and over 7 million records were included in ICOADS Release 2.5.

Table 2. R2.5 deck composition (update of Table II in Worley et al. 2005). For each deck number, the description, starting and ending years, and number of reports (in thousands) after final blending are listed. Table segment taken from Table II Woodruff et al. 2010

Deck	Description	Start	End	Rpts (K)
732	Russian Marine Met. Data Set (MORMET) (rec'd at NCAR)	1888	1995	7,873
733	Russian AARI North Pole (NP) Stations	1937	1991	98
734	Arctic Drift Stations	1893	1924	12
749	First GARP Global Experiment (FGGE) Level IIb	1978	1979	6
762	Japanese Kobe Collection Data (keyed after decks 118-119)	1889	1940	3,135
780	NODC/OCL World Ocean Database (WOD) (and formerly Atlas, WOA)	1772	2005	7,738
792	US Natl. Cntrs. for Environ. Pred. (NCEP) BUFR GTS: Ship Data	1998	2007	5,889

Source identification (SID) is used to distinguish subsets within a DCK (in general, with some exceptions discussed in ICOADS, 2010) and dynamic cases where updated versions of a source are periodically received (e.g. buoy data), or historical collections are reprocessed (Table 3). In the simple case where a DCK is static and homogeneous the SID description can be identical to DCK, as is the case for DCK 780

and SID 137, the NODC World Ocean Database. Differently, DCK 792 (NCEP BUFR GTS: Ship Data) has two SIDs representing distinct processing sources and time period of coverage within NCEP BUFR GTS data stream; see Table 3 SID 100 and 103. Again, documenting the date range and number of reports are key metadata.

Table 3. R2.5 source ID (SID) composition (update of Table III in Worley et al. 2005). For each SID number, the description, starting and ending years, and number of reports (in thousands) after final blending are listed. Table segment taken from Table III Woodruff et al. 2010)

SID	Description	Start	End	Rpts (K)
100	NCEP BUFR GTS: Operational Tanks: Converted from Original Message	1998	1999	2198
103	NCEP BUFR GTS: Dumped Data: Converted from BUFR	1999	2007	20,241
137	NODC/OCL 2005 World Ocean Database (WOD05) updated through 13 Dec. 2007	1772	2005	7,738

Another major metadata element is the platform type (PT) identification (Table 4.). Here we see the expected identifiers for ships, moored buoys, and drifting buoys. We also use PT to handle more subtle differences such as profiling ocean instrumentation. PT 10-12 distinguishes various instruments found in the World Ocean Atlas Database (Table 4).

Table 4. Sample of platform types (PT) used in Release 2.5 ICOADS

Code	Description
5	Ship
6	Moored buoy
7	Drifting buoy
10	Oceanographic station data (bottle and low-resolution CTD/XCTD data)
11	Mechanical/digital/micro bathythermograph (MBT)
12	Expendable bathythermograph (XBT)

As a final combined illustration, an ICOADS record with DCK=780, SID=137, and PT=12 is a record from the World Ocean Database, and is from a profile measured by an expendable bathythermograph. This metadata identification-triple is a robust set for tracking data record provenance. The scheme has been in place beginning with the first Release and the definition tables undergo modifications, generally expansion, with each new Release.

References

ICOADS, 2010: Archival of Data Other than in IMMT Format: The International Maritime Meteorological Archive (IMMA) Format. [http://icoads.noaa.gov/e-doc/imma/R2.5-imma.pdf].

Woodruff, S.D., S.J. Worley, S.J. Lubker, Z. Ji, J.E. Freeman, D.I. Berry, P. Brohan, E.C. Kent, R.W. Reynolds, S.R. Smith, and C. Wilkinson, 2010: ICOADS Release 2.5: Extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.* (in press) (doi:10.1002/joc.2103).

Worley, S.J., S.D. Woodruff, R.W. Reynolds, S.J. Lubker, and N. Lott, 2005: ICOADS Release 2.1 data and products. *Int. J. Climatol. (CLIMAR-II Special Issue)*, **25**, 823-842 (DOI: 10.1002/joc.1166).

Annex: Planned ICOADS Development of a Unique Report ID (UID); and Intra-record Release No. (RN) Tracking

A variety of IMMA format improvements are planned in the context of a 3-year research proposal recently submitted to the NOAA Climate Program Office, Climate Observations and Monitoring (COM). Among those, we will include a Unique Record ID (UID), in additional to improved ICOADS Release number (RN) tracking (i.e. within the IMMA records, as is not presently the case). The specific form of the UID is yet to be decided, and will be coordinated e.g. with reanalysis centers during the development process to solicit any additional ideas or technical considerations. The prospect has already been raised whether other developing or planned in situ international "comprehensive" archives (e.g. upper air, and land surface) should considering partly) compatible (as applicable) or at least technically similar schemes (e.g. so the UID interface with reanalyses to the various archives could be unified).

Initial assignment of the UID number to the archive
In the simplest form, the UID might be initiated for ICOADS starting with a numbering from 1, …, ~295M ($m_{R2.5}$) of all the records (in some precise sequential archive ordering to be defined, but probably temporally organized at the highest sort level, e.g. 1662, 1663, …) the R2.5 intermediate product (ICOADS, 2010) (however the possible advantages of having the UID be a smaller number instead local to an archive increment, such as year-month, should also be considered). The intermediate product, described in ICOADS (2010), contains all currently blended duplicates and other questionable reports, and from it a smaller finalized product (261M reports) without the dups etc. is constructed for most users (although the intermediate product is also available in the event advanced users wish to study the duplicate matching etc.).

Tentative plan for handling new records introduced during Releases/updates
Number these from $m_{R2.5}$ + 1 to $m_{R2.k}$ (k = new Release increment, for an additional set of records to be blended). When these data are blended into e.g. R2.5 data (i.e. records numbered 1, …, ~295M ($m_{R2.5}$), in the resulting blended data the UIDs will no longer be sequential (i.e. new UIDs will be interleaved into the old purely numeric sequence).

Possibility of merged (multi-source) reports in the future
Data sources such as the VOS Climate (VOSClim) project, in which unique data fields flow from up to three distinct data streams (GTS, delayed-mode, and NWP comparison) could benefit from merged records in the future. However, blending records could potentially be used more widely to improve the quality and completeness of the data more generally (e.g. GTS vs. delayed-mode). Probably blended records should receive a new UID (and DCK, SID).

UID format considerations (database and IMMA)

Base36 (alphanumeric) encoding (ICOADS 2010, Table 1) could potentially be used (both in the DBMS and IMMA, if desired, or the DBMS could decode the base36 values into integers). This could achieve at least a 50% reduction (?) in storage size. Dave Berry (UK NOCS) has implemented generalized software for this purpose, and different implementations appear possible either to maximize space savings, at the expense of CPU time, or vice versa.

Release No. (RN) format/update considerations
This could be a straightforward string of a few numeric characters (e.g. 2.5 or 2.5.1…, likely omitting the decimal points for storage efficiency in IMMA per the uniform representation conventions). This could be initially assigned at the same time as UID (both likely in expansions of the existing ICOADS attm C1). Then: (i) all records would have this field update at the time of a new Release, (ii) handling of preliminary and auxiliary data (near-real-time updates and any major new data blocks offered in advance of Release blending) to be decided.