Data Management Issues:
Summary of Recent Discussion on Data Format

Matthias Lankhorst

10. April 2008

(Comments arising during presentation added in red)

Existing OceanSITES User Manual v1.0

Past months & here:
Revised Manual with Complete Format Description (E-mail from Thierry Carval this March)—READ THIS TONIGHT!

Format is designed to accommodate:
- Subsurface mooring data (e. g. ADCP, T, S, Fluoro, Nuts)
- Surface buoy data (e. g. meteorology)
- Shipborne CTD, glider
- (Hopefully) data on sections and interpolated onto XYZT grids

NetCDF File Format:
File contains “attributes” and “variables”:

```
nc = netcdf('OS_MOVE-V404-1_200001_TS.nc', 'noclobber');

nc.title = ncchar(''OceanSITES in-situ data, MOVE mooring V404'');
nc.conventions = ncchar(''OceanSITES User Manual Version 1.0'');
nc.netcdf_version = ncchar(''Unknown'');
nc.creation_date = ncchar(''2008-01-30 05:00:00 UT'');
...total of circa 30-40 attributes

nc{'TEMP'} = ncdouble('TIME', 'LEVEL'); %% 69173 elements.
nc{'TEMP'}.FillValue_ = ...
nc{'TEMP'}.long_name = ...
nc{'TEMP'}.standard_name = ...
...total of circa 10 attributes

nc{'TEMP_QC'} = ncbyte('TIME', 'LEVEL'); %% 69173 elements.
nc{'TEMP_QC'}.FillValue_ = ncbyte(0);
nc{'TEMP_QC'}.long_name = ncchar(''Quality Flag for TEMP'');
nc{'TEMP_QC'}.conventions = ncchar(''OceanSITES reference table 2'');
```

Reminder on File Names:

OS_XXX_YYY_ZZZ.nc
or: OS_XXX_YYY_ZZZ_PARTx.nc

XXX: Site name actually called "platform code"
YYY: Configuration
ZZZ: Variables
PARTx: optional if files larger than 100 MB

Only three (or four) "_"

Bad: OS_V404_1_200001_CTD.nc

Good: OS_MOVE-V404-1_200001_TS.nc

actually not quite good because mooring name not listed in official catalogue

What should be done if depth (DEPH) parameters of different instruments in the same mooring mismatch?

Example: an ADCP at the surface delivers velocity at fixed, well-known depths, while a temperature sensor on the mooring line moves up and down as the mooring is tilted in the current.

Solution: Put data in separate files.

Done in recent manual update.

This was discussed controversely.

Define "parameters awaiting definition" as of manual v1.0 section 4.3! Check consistency with other naming conventions (Nan Galbraith).

More parameters to define:

- Sea water velocity in XY-coordinates (eastward and northward, resp.)

- Air temperature (or is it ATMP ? remove ambiguity with CAPH!)

Has been worked on (Thierry Carval, Nan Galbraith). Are we done yet? Cf. discussion led by Thierry later.

Manual section 4.3 references www.oceansites.org/data/units, but this page does not exist. Suggest to create this and have it include the standard parameter names and units.

Thierry: Status?

Global / General Attributes

Suggestion: Decide to NOT have separate metadata files. Include all info in global attributes of data file. Remove term "general attributes".

Consequence: attributes must be machine-readable and thus strictly standardized.

Advantage: Fewer files, saves time for data providers. Information is already there anyways.

Not agreed upon. There will be more metadata, cf. discussion led by Nan Galbraith later.

More attributes...

I find it confusing that some are CAPITALIZED and others not - do we want to change this?

Suggestion: all attributes non-caps, all variables and dimensions CAPS.

Conflicts between global/general attributes:

creation_date vs. DATE_CREATION
distrubution_statement vs. DATA_RESTRICTIONS
keywords_vocabulary exists, but not keywords
Shouldn't we have REFERENCE_DATE_TIME?
'institution' listed twice

Have been resolved in recent manual update. M. Lankhorst and T. Carval: double-check.

Explicitly state that all global/general attributes are of type ”char”, even those that contain numbers?

We want one human-readable time format in the attributes (previously two).

Done in recent manual update. M. Lankhorst and T. Carval: double-check.

Do we want *_QC for TIME, LATITUDE, LONGITUDE, DEPH? I think yes.

Has been included in recent manual update.

Update to manual: label and number "reference tables" more clearly (e.g. reference tables 1 and 3 each deserve a seperate caption).

Done. M. Lankhorst will double-check.

Some of the data files are rather big (>100 MB). Should we have a strict limit or at least a recommendation for file size (e.g. <50 MB)?

Suggestion: Come up with a strict limit. Suggest that files are split if they get too big (e. g. monthly).

Has been included in recent manual update.

Attributes for variables:

Manual should more clearly define which attributes are required for variables, e.g. *.valid_min, *.standard_name, *.QC_indicator, . . .

M. Lankhorst to propose update

How do I retrieve the "netcdf_version" for my software for the corresponding global attribute?

If variables “LATITUDE” and “LONGITUDE” are not there (they were optional acc. to manual), the corresponding “axis” attributes are also undefined (resulting in missing “X” and “Y” axes definitions).

Mandatory global attributes “latitude” and “longitude” do not make sense for gridded data or section data.

Suggestion: Make these variables mandatory and remove the “latitude” and “longitude” global attributes. However, keep global attributes “southernmost_latitude”, etc., because they allow for quick automatic screening and make sense for all (?) data types.

There are attributes other than, but similar to, “southernmost_latitude”, which conform to some standard. We should use these. Proposed by Mike McCann.

Likewise, require "DEPH" variable because it is the "Z" axis. (Somehow, pressure as a measured value does not make a good "axis".)

Require "DEPH_QC" to allow judgement of "DEPH".

Define a new value for *_QC to represent nominal depths?

Proposed 7 "nominal value", to be coordinated with Argo.

Remove global attribute "sensor_depth" because it makes no sense for CTD, gridded, and section data.

Suggest following global attributes:

author

Included in recent manual update.

Note that existing attributes on PI and contact addresses remain untouched.

Reg. “TIME” variable:

Should have an attribute with “reference_time” (e.g. 1950/1/1). This information is presently in the “long_name”, where it comes somewhat unexpectedly.

Should have a ”standard_name” in ref. table 3.

Add sentence in manual that clearly states “day zero is 1950/1/1”.

Reg. all variables:

Do we want an "accuracy" attribute? We already have "resolution" (I guess that means precision).

Lengthy discussion. General agreement that some uncertainty estimate is required. Proposed:

mandatory attribute "uncertainty" to all variables, which contains machine-readable options for absolute or relative uncertainty

optional variable PARA_ERROR for variables PARA with varying uncertainty