OceanSITES

Data Providers' Guide

Version 1.3

February 27, 2020

Document History:

Date	Version	Author	Description
2012-08-09	0.1	Jing Zhou	Initial Draft
2013-05-30	0.2	Matthias Lankhorst	Edits during DMT meeting, Seoul
2013-12-27	0.4	Nan Galbraith	Renamed; added steps, changed reference to Data Format Specification
2019-08-21	1.2	Nan Galbraith	Prepare for approval
2020-01-15	1.3	Nan Galbraith	Add text re instruments, processing, and JCOMMOPS requirements

1 Overview

1.1 About OceanSITES

The OceanSITES program is the global network of open-ocean sustained time series sites, called ocean reference stations, being implemented by an international partnership of researchers. OceanSITES provides fixed-point time series of various physical, biogeochemical, and atmospheric variables at different locations around the globe, from the atmosphere and sea surface to the seafloor. The program's objective is to build and maintain a multidisciplinary global network for a broad range of research and operational applications including climate, carbon, and ecosystem variability and forecasting and ocean state validation.

All OceanSITES data are publicly available. More information about the project is available at: http://www.oceansites.org.

1.2 About this document

This document contains guidelines for providing metadata and data, describes the OceanSITES file naming scheme, and how to upload to GDACs. It should be used in conjunction with the Data Format Reference Manual. Intended users are OceanSITES data providers.

1.2.1 Technical Documentation available

Technical documentation of the OceanSITES system consists of three pieces

OceanSITES Data Format Reference Manual For data producers and users, formerly called the OceanSITES Users Manual, this document contains a description of the OceanSITES netCDF specification, code lists for required variables, and OceanSITES Data Management terms of reference.

OceanSITES Data Users' Guide For data users, formerly called How to Access OceanSITES Data, this document contains an outline of the Global Data Assembly Center (GDAC) data directory structure and ftp/opendap access, data use policy/license, list of sites, Data Assembly Centers (DACs), etc. It should be used in conjunction with the Data Format Reference Manual.

OceanSITES Data Providers' Guide For data producers: DACs and Principle Investigators (PIs); this manual.

1.3 Data Provider Obligations

An OceanSITES data provider is expected to read and understand this manual, the Data Format Reference Manual, and the netCDF specification it describes. OceanSITES participants are required to submit data in a timely fashion, with the understanding that these are the "best available" versions, and may be updated if improved versions become available. Data files should be in compliance with a published OceanSITES format specification.

To ensure that users of OceanSITES data are able to comply with the OceanSITES data use policy, as agreed by the OceanSITES Steering Team, data providers should use the netCDF global attributes "license" and "citation" to clearly state these requirements. **The citation should include any acknowledgments to the specific site operators as well as the following standard text:**

"These data were collected and made freely available by the international OceanSITES project and the national programs that contribute to it."

1.4 Disclaimer

OceanSITES data are published without any warranty, expressed or implied. It is the responsibility of the data provider to ensure that the data is appropriately described, and that any known flaws are documented. The user assumes all risk arising from his/her use of OceanSITES data.

OceanSITES data are intended to be research-quality and include estimates of data quality and accuracy, but it is possible that these estimates or the data themselves contain errors. It is the sole responsibility of the user to assess if the data are appropriate for his/her use, and to interpret the data, data quality, and data accuracy accordingly.

1.5 Feedback

Data providers are encouraged to communicate with the OceanSITES Data Management Team and to request clarification or other help as needed. Reporting of problems by both providers and users is welcomed, via the Data Management team and/or contact addresses listed in the data files and on the OceanSITES web page.

2. Data Management Overview

The OceanSITES data flow is carried out through three organizational units: Principal Investigators (PI), Data Assembly Centers (DAC), and Global Data Assembly Centers (GDAC). Oversight is provided by the OceanSITES Project Office at JCOMMOPS, the WMO-IOC Joint Centre for Oceanography and Marine Meteorology Observation Programmes Support.

Generally, OceanSITES publishes the "best" version available of each dataset, and does not maintain older versions once a given version has been superseded. Data is archived in snapshot form at the US National Centers for Environmental Information, NCEI (formerly NODC).
The OceanSITES data management team has developed an implementation of netCDF based on the community-supported Climate and Forecast (CF) standard, which supplies a standard vocabulary and some metadata conventions. The OceanSITES specification includes some requirements beyond the CF standard, for better data discovery and provenance information. In general, the PI provides the data and metadata information to a DAC; a PI may act as his/her own DAC. The DAC checks the metadata, and converts data and metadata into an OceanSITES file, and passes it on to a GDAC. The GDAC puts the files on its publicly accessible ftp server, and within 24 hours adds the file to its data catalog, and sends the file to the second GDAC to be added to its server.

3. Adding a site to OceanSITES

The following steps describe the basic approach for a PI or DAC to contribute data to the OceanSITES Project.

3.1 Register with OceanSITES, Obtain Site and Platform Codes

A prospective PI may approach the OceanSITES Project Office to become a member. Contact information for the project is available on the project website, OceanSITES.org. Projects with sustained, high quality time series observations in open ocean locations are the main focus of the OceanSITES project.
Once approved, the PI works with the OceanSITES project office to determine meaningful site codes and platform codes for his/her platforms. An OceanSITES site is a defined geographic location where sustained oceanographic, meteorological or other observations are made. An OceanSITES platform is an independently deployable package of instruments and sensors forming part of the site. There can be several platforms (e.g. several moorings) at a single site. The OceanSITES catalogue maintained by the project office has a full list of site codes and platform codes. Those with data present at the GDACs are listed in the index file (described in section 3.3. below) which is harvested daily and available via the OceanSITES.org web pages.

3.2 Collect Data and Metadata

The OceanSITES Data Format Reference Manual documents a fairly rigorous content standard; we hope to have fully self-documenting data in the project, to ensure the maximum usefulness of the data in the long term.
Please include all metadata required by the specification, but also record any relevant additional information to the best of your ability. Items such as instrument manufacturer and model name, sensor specifications, processing steps (e.g. source of and value of magnetic declination corrections), and conventions (e.g. whether wind is in meteorological or oceanographic convention) should be included. Note any shortcomings in the collection methods or processing.

3.3 Generate OceanSITES-compliant Data Files

The DAC (or PI) assembles the data into files that comply with the OceanSITES Data Format Reference Manual, available on the OceanSITES web site. The OceanSITES Data Management Team can be contacted with formatting or content standard questions. A number of software tools can be provided upon request to assist with the data file assembly, such as example files and scripts to generate example files.
Certain parts of the Format Specification are used by the GDACs to generate an inventory 'index' file of our data holdings; other items are used by the JCOMMOPS metadata management system to populate their database and ensure that OceanSITES data files are properly represented in the JCOMMOPS portal. The following details of the specification are used in those processes, and so are of particular importance:

- **File Names**

A file naming convention assists us in organizing our data so it can be catalogued and discovered more easily. An OceanSITES netCDF file name for single deployment data is in the form:
OS_PlatformCode_DeploymentCode_DataMode_PARTX.nc

Please see the Data Format Reference Manual for a complete description of the file name fields.

• Global attributes used by GDACs

The GDACs maintain a data catalog that depends upon specific global attributes in the netCDF files. The following global attributes are required for that catalog:

site_code: site code from the OceanSITES catalogue
platform_code : platform code from the OceanSITES catalogue
date_update: file update or creation date
update_interval: update interval for the file (void, 'P1D')
data_mode: data mode (the same as appears in the file name)
time_coverage_start: start date of the data
time_coverage_end: final date of the data in the file
geospatial_lat_min: southernmost latitude
geospatial_lat_max: northernmost latitude
geospatial_lon_min: westernmost longitude between
geospatial_lon_max: easternmost longitude between
geospatial_vertical_min: minimum depth in meters for measurements
geospatial_vertical_max: maximum depth in meters for measurements

- **Global Attributes used by the JCOMMOPS Data Portal**

platform_deployment_date, platform_recovery_date: mooring dates
platform_deployment_ship_ICES_code, platform_recovery_ship_ICES_code:
platform_deployment_cruise_ ExpoCode, platform_recovery_cruise_ ExpoCode: Note:
ExpoCodes are composed of ICES ship code and cruise start date expressed as YYYYMMDD.

- **Variable Attributes used by the JCOMMOPS Data Portal**

sensor_SeaVoX_L22_code: use the SeaVox device ontology
sensor_data_start_date: required if different from time_coverage_start
sensor_data_end_date: required if different from time_coverage_end

Notes:

• All dates and times are in UTC. With the exception of the time field in the ExpoCode, the string representation of date and time must be compliant with ISO 8601, i.e. “YYYY-MM-DDThh:mm:ssZ”

• The update_interval complies with ISO 8601: ‘void’ for no scheduled updates, 'P1D' for daily, or ’PT12H’ for twice daily updates.

• Latitudes are specified as +/- 90 degrees; longitudes are +/- 180.

(+ indicates north or east, respectively)

3.4 Arrange for Data Upload, Upload Data

Before submitting data to a GDAC for inclusion in the OceanSITES repository, the data provider must request an account from one of the GDACS. The GDACs are at IFREMER, in France, and at NDBC in the USA; they run scheduled data synchronization, so data only needs to be uploaded to one site.
Typically, users upload data via FTP to a GDAC-provided directory. For near real-time data, files may be updated on a daily basis. Files are not placed directly into the public FTP directory, but into users’ areas; the GDACs populate the public FTP structure on a scheduled basis.

3.5 Check Status

Within 24 hours of data upload, new files will appear in the first GDAC’s public FTP, if processing is successful. The second GDAC server will synchronize the data files within the following 24 hours. Data providers should check that their files appear on the servers; they may also verify the contents, if desired, by downloading the released data files from the FTP servers.
The OceanSITES GDACs provide an inventory of data holdings in an ‘index file’ named oceansites_index.txt, in the top directory of the FTP site. Within 24 hours of data upload, there will be a line entry in the index file for each successfully processed data file. The index file is a comma-delimited text file; it is the responsibility of the data provider to check his/her data files in this index, to look for metadata errors.
The first field is the path from the top FTP directory to the corresponding data file. A typical OceanSITES data file would have a path like: DATA/SiteCode/FileName.nc Successive fields provide basic metadata about the file contents, taken from the metadata fields listed earlier. Please check the geo-spatial-temporal extents and the CF standard names of the data variables for errors, and discuss any problems with the GDAC or the Data Management Team.

4. Data File Changes and removals

If a file needs to be updated, the DAC can upload the new version via FTP, and the GDAC processing will replace the existing file on the public server with the new version within 24 hours. You may check the index file to be sure the update was successful.
DACs do not have the ability to directly remove data files from the public-facing FTP directories. To remove a data file from the GDAC FTP servers, simply upload a zero-length data file with the intended file name to the respective FTP account. The GDAC processing programs will remove the corresponding data file from the GDAC FTP server within 24 hours, and from the other GDAC FTP server within the next 24 hours after synchronization.
Note that real time data should be replaced with delayed-mode data as soon as the latter is available. To achieve this, simply upload an empty file with the same name as your real time data file, when you're uploading your delayed-mode data.

5. Metadata and Quality Issues

5.1 Documenting instruments, processing and calibrations

The OceanSITES project does not require the use of any specific instruments, processing systems or algorithms; these are instead left to the best judgement of the PIs and DACs. The project does, however, require that data files be self-documenting, so that a data user can understand the contents, from the collection methods, to the processes, calibrations and constants that have been applied.

5.1.1 Documenting instrumentation

Instrument characteristics make datasets useful in different ways; knowing the precision, accuracy and expected drift characteristics of the instrumentation is critical to the correct use of any data. The Data Format Reference Manual contains guidelines for different methods of documenting instrumentation.

OceanSITES has adopted the SeaVox Device Catalogue vocabulary for instruments and sensors, L22, which is available online; the URL is found in Appendix 1. Previous versions of the format allowed data providers to describe instrumentation without a standard vocabulary, but as of 2020, the L22 identifiers must be used, for the sake of interoperability.

The required terms to describe instruments are model, manufacturer, and SeaVoX_L22_code. For some instruments, such as ADCPs, is also critical to include user-programmable settings, such as instrument deployment depth, instrument orientation, time between ensembles, pings per ensemble, bin length, number of bins, percentage good threshold, automated data rejection (i.e. fish detection algorithms). For other types of instruments, it may be worthwhile to include calibration information, model version numbers, or information on

firmware; it is up to the PI and the DAC to determine what a user would need to know about a specific instrument used.

5.1.2 Documenting processing

The global attribute 'history' is defined in the NetCDF Users Guide. It should explicitly describe processing done to the overall data, including dates. This attribute can contain information about processing done to data from individual sensors, or to the time base or other coordinate variables. Care should be taken when combining data from multiple individual sensor files to keep all the information in the history attribute from all input files.

Any standard suites of algorithms, i.e. TEOS-10 for salinity, or the COARE Bulk formula for air-sea fluxes, should be cited in the history attribute.

For variables that are processed individually (rather than globally), a similar history should be provided. Also include a description of any constants used, such as magnetic variation correction for winds and currents, algorithm for salinity calculations (e.g. PSAL:reference_scale = "PSS-78" ; and PSAL: algorithm = "TEOS-10" ;). Note that netCDF code does not handle global and variable attributes with the same names as one might expect; to avoid problems, the data provider may choose to use a slightly different name for any history attributes at the variable level, e.g. psal_history.

6. Appendices

6.1 Appendix 1: Further Information, links, tools

- OceanSITES website: http://www.oceansites.org
- ICES Ship Codes, used in platform_deployment_ship_ICES_code, platform_recovery_ship_ICES_code: https://ocean.ices.dk/codes/ShipCodes.aspx, or at https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/C17/
- The SeaVoX (SeaDataNet and MarineXML Vocabulary Content Governance Group) vocabularies, served at BODC, contain terms for some of our attributes:
  - Sensors and instruments: use the **L22** Device Catalogue https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/L22/
  - sea_area: use the **C19** Sea Areas vocabulary https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/C19/
  - source: use the SeaVoX Platform Categories vocabulary https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/L06/
- EPSG, used for the coordinate reference frames: http://www.epsg.org/
- ISO8601, used for date and time: http://en.wikipedia.org/wiki/ISO_8601
- WMO IDs, used for wmo_platform_code attribute: http://wmo.int/pages/prog/amp/mmop/wmo-number-rules.html
- NetCDF
  - Users Guide: https://unidata.ucar.edu/software/netcdf/docs/user_guide.html
  - Best Practices: https://unidata.ucar.edu/software/netcdf/docs/BestPractices.html
- CF, the netCDF Climate and Forecast Metadata Convention: http://cfconventions.org/
- Unidata udunits: https://unidata.ucar.edu/software/udunits/
- US NOAA-NCEI (formerly NODC) netCDF Templates: http://www.nodc.noaa.gov/data/formats/netcdf/
- ACDD, the Attribute Convention for Dataset Discovery: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery