SOCIB Balearic Islands Coastal Observing and Forecasting System

SOCIB Data Representations for In Situ Observations

SOCIB-Data Center Facility

Document type:	Specification
Date:	2022-08-02

Description:	SOCIB data representations for in situ observations
Authors:	P. Rotllán, X. Notario, I. Ruiz, M. Marasco
Supervision:	J. G. Fernández
Involved Personnel:	DCF, HFR
URL:	http://repository.socib.es/repository/entry/show?entryid=b8b357d6-7241-4353-ac53-8a4a85319042
Access:	Public

DOCUMENT DISTRIBUTION LIST

Date:	Distribution to:
2022-08-03	@internal

CHANGE RECORD

#	Date	Description	Author	Checked by
0.1	2020-10-07	Initial version	PR	XN, IR, MM, JGF
0.2	2020-10-13	Improved definitions	MM	JGF, PR
0.3	2020-10-19	Revision	PR, JGF	JGF
0.4	2022-02-23	Definition of platform, instrument and sensor concepts.	JGF, MM	ER
1.0.0	2022-06-01	Text polishing, review packing level concept, include related document	MM	

Table of Contents

1. Overview

SOCIB manages a set of different in situ observing platforms in the Western Mediterranean Sea, such as Coastal Station, Glider, HF-Radar, Oceanographic Buoy, Profiler drifter, Research Vessel, Sea Level, Surface drifter, Turtle and Weather Station. Data are collected, processed and therefore formatted as netCDF, according to common metocean community practices. SOCIB conventions follow the international guidelines established by Australia's Integrated Marine Observing System (IMOS), the Climate and Forecast NetCDF (CF) and OceanSites (OceanSITES).

2. Objective

The main aim of the document is to provide an overview of data types and classification undertaken at SOCIB according to the following criteria:

- Data model and data representation
- Data generation mode
- Data transformation
- Packing level
- Feature Type

SOCIB formatted data are classified depending on the processing chain/line or generation procedure (Real Time, Delayed Time, Delayed Mode), processing level (e.g. L0, L1, L2), feature types (e.g. time series, trajectory, profile) and dataset packing levels (e.g. deployment, aggregates). For more details on the netCDF format manual, please check the PUM_DCF_SOCIB-in situ-measurements-netcdf-format-manual (work in progress document).

3. In Situ Observations Data Representations/Categories

3.1. Data model and data representation

SOCIB platform network records multidisciplinary data, which are transmitted to the DCF in order to be quality-controlled, made Findable, Accessible, Interoperable, Reusable (following the FAIR principles). Data follow the SOCIB representation model in order to be included within the SOCIB Data Repository. The instrumentation data model is based on 4 main interconnected concepts:

- Platform: top level concept aiming to represent the multiplatform nature of SOCIB. Currently, a platform can be distinguished as follows:
  - group of integrated instruments physically mounted on a single device (RV) or having a common observing purpose (HF Radar, Stations, Mobims).
  - Identified single instrument (drifter, profiler, turtle, glider).

In order for a platform to be active, this needs to include a set of instruments at a specific time.

- Instrument: intermediate level concept which is related to the physical instruments installed inland or in the water (e.g. CTD, Currentmeter, Tide gauge, Barometer, Laboratory). An instrument is considered as active and able to collect measurements, if it is physically installed, and includes a set of sensors at the time of the installation.
- Sensor: lower level concept related to the physical sensors (e.g. ph, oxygen, conductivity), which directly collect the measurements needed for data generation. Sensors are considered active since those are installed on the instrument, and their state is operative at the time of the installation.
- Deployment: this key concept identifies the actual data production as soon as data enter into the SOCIB data pipeline. A deployment has a time range during which the instrument is not altered by human intervention. On the other hand, as any change is applied to the instrument and therefore to the data pipeline (e.g. for maintenance operations, uninstallation, data cleansing), the current deployment ends. The instrument has to be installed on a given platform as a requisite for a deployment to be considered valid. The deployment initial and end date indicates the maximum time range related to all the time series generated during that time. For more information on how to create a deployment, please refer to the PUM-DCF_instrumentation-database-processing-configuration.

3.2. Data generation mode

The data pipeline of SOCIB netCDF Real Time (RT) and Delayed Time (DT) file production involves all the processing toolboxes (e.g. PUM_DCF_processing-application-user-manual and Glider Toolbox) and the Management database (i.e. PUM-DCF_instrumentation-database-processing-configuration). The first digest the raw data stored in the file system, and processes them, while the latter manages the deployments per instrument/platform and processing configurations. A post-processing procedure, run in delayed mode by the SOCIB scientific board, produces other netCDF files, which are carefully quality checked (i.e. Delayed Mode, DM). Data generation mode can be identified depending on the platform and instrument type (Table 1).
File generation is therefore classified as follows:

- RT: data produced since first raw data are transmitted by a set of automatic applications and procedures. RT data generation mode is applicable only if a communication channel enables the transmission of files as these are processed. Data are provided within 24-48 hours from acquisition on average after applying a basic level of automated data quality control.
- DT: data produced after recovery by a set of semi-automatic procedures. DT data generation mode is applicable only if the platform or sensors are supported by an internal memory to store collected measurements until their recovery.
- DM: data produced from DT (preferently) or RT datasets. This generation mode allows the production of more curated data. Additional data quality control, checks of suspicious observations and corrections are performed by a scientific assessment.

***Table 1**. The table shows the current relationship between platform/instrument type and data generation mode for the SOCIB data pipeline.*

Platform type	Instrument type	line type		
		RT	DT	DM
Coastal Station	Barometer	x		
	Conductivity and Temperature Recorder		x	
	Current profiler		x	
	Tide gauge	x		
	Waves recorder		x	
	Weather Station	x		
Oceanographic Buoy	Conductivity and Temperature Recorder	x		
	Currentmeter	x		
	Current profiler	x		
	Multiparameter probe	x		
	Oceanographic Buoy	x		
	Status	x		
	Temperature Recorder	x		
	Waves recorder	x		
	Weather Station	x		
Research Vessel	CTD		x	x
	Current Profiler		x	
	GPS	x		
	Thermosalinometer	x		
	Weather Station	x		
Sea Level	Barometer	x		
	Tide gauge	x		
Surface drifter	Surface drifter	x		
Profiler drifter	Profiler drifter	x		
Weather Station	Weather Station	x		
Glider	Glider	x	x	x
HF-Radar	HF-Radar	x		

3.3. Transformation (processing levels)

The process of re-formatting (raw to netCDF files) consists in implementing the originally transmitted/stored data (with additional derived and/or quality control variables) or/and reshaping them (trajectory into grids) independently from the data generation mode (RT, DT or DM). Files are stored as netCDF, according to the specifications of the SOCIB In-situ measurements NetCDF Format Manual (work in progress document). NetCDF files are implicitly linked through the different processing levels. For instance, a RT netCDF file should have a L0, and then a L1, and in some cases, a L1_corr and a L2. This chain is derived from the file name (see SPEC_DCF_SOCIB-netcdf-file-naming-convention). The SOCIB file processing levels are identified depending on the platform and instrument type (Table 2), and

can be classified as follows:

- L0: set of variables included within the raw data (e.g. cnv, ascii, binary).
- L1: set of variables which can include derived, and quality controlled variables (if Quality Tests available, reflecting data reliability according to SOCIB Quality Control flagging convention).
- L1_corr: set of variables which includes L1 and complementary variables whose corrections are performed by dedicated scientific teams). At time of writing, this processing level is tied to the DM data generation mode.
- L2: set of variables containing the processed L1 data in a gridded format (only for the glider data).

Table 2. *The table shows the current relationship between platform/instrument type and processing levels for the SOCIB data pipeline.*

platform type	instrument type	processing level types			
		L0	L1	L1_corr	L2
Coastal Station	Barometer	x	x		
	Conductivity and Temperature Recorder	x	x		
	Current profiler	x	x		
	Tide gauge	x	x		
	Waves recorder	x	x		
	Weather Station	x	x		
Oceanographic Buoy	Conductivity and Temperature Recorder	x	x		
	Currentmeter	x	x		
	Current profiler	x	x		
	Multiparameter probe	x	x		
	Oceanographic Buoy	x	x		
	Status	x	x		
	Temperature Recorder	x	x		
	Waves recorder	x	x		
	Weather Station	x	x		
Research Vessel	CTD	x	x	x	
	Current Profiler	x	x		
	GPS	x	x		
	Thermosalinometer	x	x		
	Weather Station	x	x		
	Laboratory	x	x		
Sea Level	Barometer	x	x		
	Tide gauge	x	x		
Surface drifter	Surface drifter	x	x		
Profiler drifter	Profiler drifter		x		
Weather Station	Weather Station	x	x		
Glider	Glider	x	x	x	x
HF-Radar	HF-Radar	x	x		

3.4. Dataset packing level

The collection of netCDF files resulting from each instrument/sensor unit are packed in dedicated datasets to allow the traceability between the instrumentation configuration of the deployment and the resulting data (see examples in Table 3). There are 2 levels for packaging these netCDFs files in the SOCIB Data Repository:

- Intermediate dataset:
  - Deployment: Collection of netCDF files (most of the cases, monthly) produced during the lifetime of a deployment regardless of their processing level or data mode. The corresponding datasets aligns with the Deployment concept explained in section 3.1. The overall time coverage of these datasets is constrained by maintenance or replacement operations of the source instrument unit.
    - Latest: Collection of netCDF files produced during a deployment, only including up to the last two months of the datasets.
  - Platform: collection of netCDF files (most of the cases, monthly) originated from more than one instrument of a platform. Differently from the previous datasets, the overall time coverage of these datasets is not constrained by maintenance or replacement operations of the instrument units.
  - Aggregated (currently per instrument type and platform): collection of netCDF files included in the deployment datasets aggregated over time on the same instrument type of the same platform, currently provided by the SOCIB Data API.
- Data Products: a collection of datasets packed together based on a common context (e.g. observing program, projects). It is mandatory that each dataset is included in only one data product, avoiding data overlapping between different contexts.

Table 3. *The table shows the current relationship between platform/instrument type and packing level dataset for the SOCIB data pipeline.*8. ANNEX 5. Current contributions to datasets per instrument and platform types

platform type	instrument type	Packing level datasets	
		deployment	platform
Coastal Station	Barometer	x	
	Conductivity and Temperature Recorder	x	
	Current profiler	x	
	Tide gauge	x	
	Waves recorder	x	
	Weather Station	x	
Oceanographic Buoy	Conductivity and Temperature Recorder	x	
	Currentmeter	x	
	Current profiler	x	
	Multiparameter probe	x	
	Oceanographic Buoy	x	
	Status	x	
	Temperature Recorder	x	

	Waves recorder	x	
	Weather Station	x	
Research Vessel	CTD	x	
	Current Profiler	x	
	GPS	x	
	Thermosalinometer	x	
	Weather Station	x	
	Laboratory	x	
Sea Level	Barometer	x	x
	Tide gauge	x	x
Surface drifter	Surface drifter	x	
Profiler drifter	Profiler drifter	x	
Weather Station	Weather Station	x	
Glider	Glider	x	
HF-Radar	HF-Radar	x	

3.5. Feature type

Data mapped to netCDF files match a defined sampling geometry representation. The geometrical representation is affected by the sampling capabilities of the instrument (discrete or continuous measurements) and platform orientation/mobility (vertical, horizontal, both, none), as shown in Table 4. SOCIB follows the CF convention, except for the grid, which represents data projected on a regular or irregular grid (e.g. High Frequency Radar platform).

Table 4. *The table shows the current relationship between platform/instrument type and feature type for the SOCIB data pipeline.*

platform type	instrument type	feature type					
		time series	time series profile	trajectory	trajectory profile	profile	grids
Coastal Station	Barometer	x					
	Conductivity and Temperature Recorder	x					
	Current profiler		x				
	Tide gauge	x					
	Waves recorder	x					
	Weather Station	x					
Oceanographic Buoy	Conductivity and Temperature Recorder	x					
	Currentmeter	x					
	Current profiler		x				
	Multiparameter probe	x					
	Oceanographic Buoy	x					

	Status	x					
	Temperature Recorder	x					
	Waves recorder	x					
	Weather Station	x					
Research Vessel	CTD				x		
	Current Profiler				x		
	GPS			x			
	Thermosalinometer			x			
	Weather Station			x			
	Laboratory					x	
Sea Level	Barometer	x					
	Tide gauge	x					
Surface drifter	Surface drifter						
Profiler drifter	Profiler drifter						
Weather Station	Weather Station	x					
Glider	Glider			x	x		
HF-Radar	HF-Radar						x