Data Quality Strategy

SOCIB-Data Center Facility

Document type:	Activity Development Plan
Date:	2020-12-11

Description:	This document describes the general SOCIB data quality policy, intended to be a long-term, general framework, as well as the current specific objectives and plan. This strategy is reviewed yearly, and it aims to outline the decisions to make, actions to take, resources to apply and people to engage, to effectively improve the level of quality of the observational and model data produced at SOCIB either in the long and short term.
Authors:	J.G. Fernández, M. Marasco, P. Rotllán, M.A. Rújula
Supervision:	J. Tintoré
Involved Personnel:	SOCIB facilities
Keywords:	Data Quality, QC, SOP, Data Management Plan
URI:	https://repository.socib.es/repository/entry/show?entryid=e4e6d093-63ec-4fd4-9bf3-6da66ae92c4d
Access:	Public

www.socib.es

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

DOCUMENT VERIFICATION LIST

Date:	Checked by (name)	SOCIB division	Ref.

DOCUMENT DISTRIBUTION LIST

Date:	Distribution to:

CHANGE RECORD

#	Date	Description	Author	Checked by
0.9	2020-12-09	Initial version	JGF	
1.0	2020-12-11	Refinement	MM, PR, MAR	JGF

www.socib.es
Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

Index of contents

Acronyms in this document:

- *DM: Delayed mode*
- *DMP: Data Management Plan*
- *DOI: Digital Object Identifier*
- *DT: Delayed time*
- *QUID: Quality Information Document*
- *RT: Real Time (eq. near real time at SOCIB)*
- *SOP: Standard Operational Procedure*

www.socib.es
Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

1. Introduction

The mission of SOCIB is to operate a coastal ocean observing and forecasting system, a scientific and technological infrastructure that provides free, open, quality controlled and timely streams of oceanographic data, as well as data stewardship and long-term preservation. The ability to fulfill such a mission will, in part, be determined by the level of quality of the data collected, stored, and managed by SOCIB following the internal data management.

SOCIB's data quality strategy is not formulated in a single document but rather is embedded across a range of business information related to data quality.

This document describes the general SOCIB data quality policy (chapter 2), intended to be a long-term, general framework, as well as the current specific objectives and plan. This strategy is reviewed yearly, and it aims to outline the decisions to make, actions to take, resources to apply and people to engage, to effectively improve the level of quality of the observational and model data produced at SOCIB either in the long and short term.

Chapter 3 contains the set of tools and resources used to support the different areas of the data quality policy.

2. Data quality policy

The approach at SOCIB is based on the definition of the SOCIB data quality policy. SOCIB data quality policy establishes that data either generated or participated by SOCIB, must align with the following guidelines:

- Follow FAIR data principles.
- Implement Quality Control, following international standards, both in real time and delayed time.
- Meet scientific and operational requirements of both the SOCIB multiplatform system and forecasting systems (including research projects in which SOCIB is involved).
- Include data assets in the SOCIB Data Repository.
- Ensure the compliance of SOCIB metadata with international standards.

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

2.1. Stakeholders

In order to align with the SOCIB data quality policy, SOCIB should meet the data and metadata quality standards according to the following stakeholders:

- Internal, which will guarantee both the assessment and feedback on data assets in order to implement and maintain the related documentation..
- External (international and national), which will ensure the required data interoperability complying with well established standards..

2.2. Quality areas

The Data Quality Strategy represents guidance for SOCIB in terms of quality, based on the FAIR data guiding principles and aligned with SOCIB objectives and impacts. In this sense it directly and indirectly underlies several important data management activities.

The quality strategic initiatives are organized in three main areas and grouped in four cross-cutting activities (Figure 1).

- **Quality Assurance (QA):** which focuses on the process of generating data. This process allows us to provide confidence about the processes generating the data and metadata through checkers.

- **Quality Control (QC):** which focuses on the quality of the resulting data. This consists of the data flagging (revision/update QC tests) and reprocessing. The process involves an automatic or manual of technical activities to ensure the data quality.

- **Quality Assessment:** which focuses on evaluating the current status of the data.. The process trackstrack inconsistencies and other anomalies in the data. It involves the process of checking the data, examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that status-up-to date of data.

3. Data quality tools & resources

In order to support the above mentioned quality areas the following tools and resources are in place or will be improved with this strategy. The purpose of having software tools to display data and perform operations on specified subsets of the data and metadata minimises the user intervention.

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

Additional resources in the day-to-day handling operations life cycle help to assure and assess the quality of data.

3.1. (Meta) Data Checkers

The data and metadata checkers help to ensure the quality of the data and metadata requirements. These tools are intended to find possible errors and report them. Once the errors are notified, a correction attempt is made to ensure the best possible quality level of the data.

3.2. Standard Operational Procedures

The Standard Operation Procedure (SOP) documents describe the information on the creation of the deployment, processes and decisions identified for each role in SOCIB. The steps of this procedure are contained in the ASANA template, which will be handled by the facilities involved.

3.3. Data Management Plans

The SOCIB Data Management Plan (DMP) describes the data management life cycle for the data collected, processed and/or generated by SOCIB. The template aims to meet the requirements of different observing programs and platforms with the necessary flexibility and adaptability allowing the stakeholders to design a customized workflow, responding to their needs.

3.4. Processing applications

SOCIB's processing applications take raw data sent by the multi-platform system and processes them in order to produce data in NetCDF format files, with the corresponding processing levels.The data produced by the SOCIB's sensors network is processed by in-house developed applications, which are:

- Processing Application
- Glider Toolbox
- HF-Radar Toolbox
- Salinity Correction Toolbox

During the automatic process, some real-time QC tests are performed, which flag individual data bits with different quality levels. These tests are specified in this document: QUID_DCF_SOCIB-QC-procedures. They are based on several tests accepted by the relevant community and after checking a variable with quality flag is added, following the Argo Quality Flag scale. The values applied for Quality Control are mainly performed in the framework of operational oceanography programs. For example, the Recommendations issued by the EuroGOOS Data Management Exchange and Quality Working group (EuroGOOS Data-MEQ) are applied as well as

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

the quality control procedures from Argo community: Argo CTD and trajectory data, Bio-ARGO or EGO gliders NetCDF format, among others.

The main processing application has a Notification System that sends an email to the Data Center staff when a problem is detected. These issues are also stored in an internal database and remain open until they are fixed.

3.5. Instrumentation Web Application

The web application “Instrumentation” allows the management of the platforms, instruments, sensors and metadata linking the processing configurations and the data products in the repository. The related information is persisted in a relational database, called “management”. When raw data are processed to be transformed into a NetCDF file, the different processing toolboxes query the “management” database in order to fill in the metadata (global or variable attributes).

All changes in the metadata and configuration databases are tracked. Thus, when a problem occurs due to a human error changing the configuration, it is possible to check what was changed and who did it.

3.6. Data dissemination interfaces

Dissemination includes the activities related to the data being accessible within the SOCIB Data Repository. The vast majority of SOCIB data are archived, mapped into a Unidata Thredds Catalog, where those can be manually or semi-manually accessed. It can also be accessed in a more user-friendly way through the SOCIB Data Catalog.

In addition to those interactive accesses, and on top of the aforementioned Thredds Catalog, one RESTFull API enables machine-to-machine data access, the SOCIB Data API. This is considered SOCIB higher data dissemination level as it usually interfaces with major pan-European portals namely (1) the Copernicus Marine Service (In Situ TAC component), trough direct data requests or/and indirectly by means of the Hellenic Centre for Marine Research/HCMR (MONGOOs Mediterranean ROOS Data Center), and (2) EMODnet physics.

In addition to the above passive data dissemination fluxes, a parallel dissemination toward these ones is set up (typically via FTP syncing) to benefit from its expertise and networking. Examples at European level are Coriolis and EU HFR node. Same thing applies for metadata exchange (ie. SeaDataNet Cruise Summary Reports -CSR-).

These dissemination interfaces ease the profiling process, providing shortcuts to inspect the data and metadata.

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

4. Data quality activities

Four cross-cutting activities are defined within the quality areas (Figure 1).

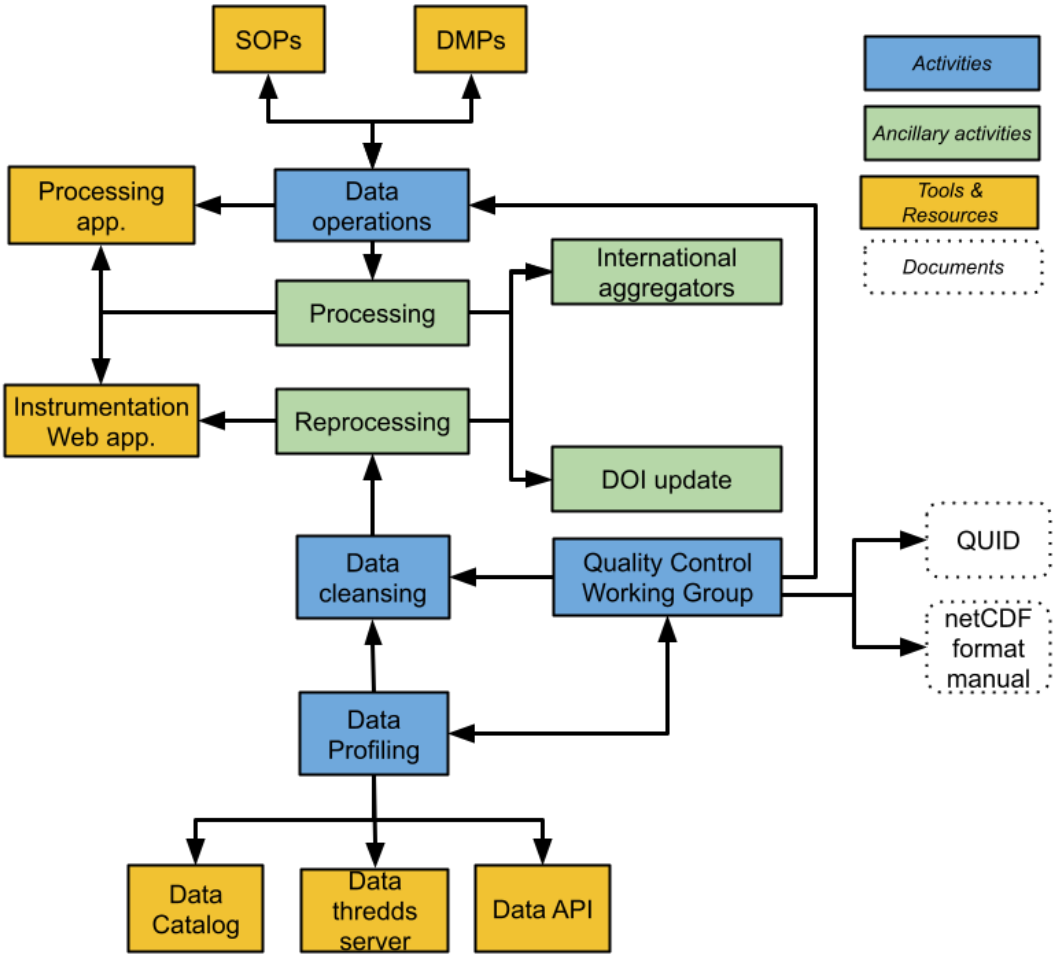

Figure 1- The scheme shows the relationships among SOCIB data quality main areas, data quality activities and related tools and resources.

4.1. Activity 1: Data profiling

It is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data. Quality of data is addressed by comparing the current status of data with the expected one through the profiling report.

- Objectives:
  - Improve the current method of work and implement tools for ensuring that non conformities are identified timely.
  - Perform data assessment.
  - Ensure that the link with the SOCIB facilities (data providers) are established.
- Plan
  - Review internal task management system open incidences.
  - Audit management database.
  - Provide metrics on data and metadata.
  - Develop a dedicated plan per facility.
  - Select priority topics.
  - Provide a list of tasks for data cleansing (Activity 2).

4.2. Activity 2: Data cleansing

The process is also known as data cleaning, and consists of correcting (or removing) corrupt or inaccurate records from a record set, table, or database. The procedure refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- Objectives:
  - Improve the current method of work and implement tools for improving data cleansing.
  - Perform data cleansing operations in response to both activity 1 and 4 outcomes and in response to incidences.
- Plan:
  - Improve Processing Application API to facilitate reprocessing of datasets.
  - Periodical cleansing operations are done from a backlog of data issues.
  - Versioning of data products with DOI.

4.3. Activity 3: Data operations

This activity is related to the data life cycle management and data operation procedures which link the facilities in SOCIB.

- Objectives:
  - Maintain stewardship of SOCIB data assets.
  - Foster collaboration with facilities to streamline data operations.
  - Improve SOPs per facility.
  - Improve DMP per facility.
- Plan:

Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79

- Finalise RT, DT and DM SOP procedures.
- Complete entry point for SOCIB DMPs.

4.4. Activity 4: Quality Control Working Group

The aim is to set up and lead a Quality Control Working Group (QCWG) in collaboration with the SOCIB facilities. In particular, the group aims to foster collaboration between facilities and the Data Center, share a common vision on Data Quality and increase the quality of the data. This activity implies defining tasks to be carried out in activities 1, 2 and 3.

- Objectives:
  - Resume previous WG.
  - Review existing tests.
  - Gather requirements from facilities.
- Plan:
  - Improve cooperation within QCWG
  - Elaborate global attribute manual
  - Implement QC tests
  - Update QC manual.

www.socib.es
Parc Bit. Ctra. Valldemossa, km. 7,4 · Edifici Naorte, Bloc A · Pl. 2a, P. 3 · 07121 Palma (Illes Balears, Espanya) · Tel: +34 971 43 99 98 · Fax: +34 971 43 99 79