Quality Assurance Framework

Institute full name:	Spanish Institute of Oceanography
Institute acronym:	IEO
Institute webpage:	http://www.ieo.es
Project full title:	Seguimiento y análisis de pesquerías
Project acronym:	SAP6
Project webpage:	http://www.proyectosap.es
Team full title:	Equipo de Muestreos SAP
Team acronym:	EM-SAP

Deliverable:

Data integrity, quality checkings

Deliverable Name	Data integrity, quality checkings
Deliverable No.	2
Date of preparation of this version:	22/05/2019
Authors:	Jose Rodríguez, Marco A. Ámez
Status (F: final; D: draft; RD: revised draft):	Final
Version:	1.0.0
Copyright	

Revision History

Version No.	Date	Details
1.0.0	*22/05/2019*	*1st final version*

List of abbreviations

Abbreviation	Definition
EM-SAP	Sampling team of SAP project
EU	European Union
QAF	Quality Assurance Framework
NVDP	Spanish official fisheries data (logbooks and sales notes)
OAB	On-board sampling network (ICES area)
RIM	On-shore sampling network (ICES area)
SIRENO	IEO database (Spanish acronym for "Integrated Monitoring of Ocean Natural Resources")

Table of Contents

1. Introduction **4**

1.1. About the Quality Assurance Framework 4

1.2. QAF version in English 4

2. Data Integrity general introduction **5**

3. Quality checks framework **7**

3.1. Introduction 7

3.2. Responsible 7

3.3. Conceptual frame quality checks review 7

3.4. Review methodology 10

3.4.1. Pre-dump revision 10

3.4.2. Dump 10

3.4.3. Post-dump revision 10

3.4.4. Post-dump corrections 10

3.5. Periodicity 12

4. Discrepancy checking (database - sampling sheets) **13**

4.1. Introduction 13

4.2. Responsible and time 13

4.3. Checks 13

5. Storage of sampling data **15**

5.1. Objective 15

5.2. Responsible and time 15

6. Main bibliography of interest **16**

1. Introduction

The present document has been developed by EM-SAP, in the frame of the SAP project and constitutes a deliverable as part of the overall Quality Assurance Framework.

1.1. About the Quality Assurance Framework

QAF describes the framework for the quality assurance of the team, and particularly the elements, actions and measures to be taken by the team members to ensure the quality of the project and its outcomes.

The main objectives of the QAF are:

- To ensure that all inputs of the EM-SAP are consistent; and
- To ensure that all outputs of the project are of the highest quality.

In order to ensure the achievement of the above objectives, the QAF aims to define procedures for ensuring best practices for every process undertaken by the team thus affecting to all process included in the team protocol.

The aim for the QAF is to be applicable to all EM-SAP activities in the long term.

The QAF will be a dynamic document throughout the lifecycle of EM-SAP, and will be updated as required.

1.2. QAF version in English

QAF addresses specific actions, measures and individual responsibilities and it's therefore produced and restricted to the members of the team. This includes the fact that EM-SAP works and produces most part of its work in Spanish.

Nevertheless, there is a need to transmit properly the processes and actions taken around the production, handling and management of the sampling information in a wider EU Data Collection context, where this team is incorporated.

Therefore, a reduced version of the QAF is produced in English thus allowing end-users, partners, reviewers and the European Commission to understand what and how is being done.

For further information or specific details please contact Jose Rodriguez (jose.rodriguez@ieo.es).

2. Data Integrity general introduction

Boritz definition (2004) is used hereby as a reference for the conceptualization of integrity. Data integrity is the maintenance of, and the assurance of the accuracy and consistency of, data. Integrity means an unimpaired or unmarred condition, entire correspondence of a representation with an original condition. Integrity is the "representational faithfulness of the information to the condition or subject matter being represented by the information". Being the core attributes of data integrity: completeness, accuracy and validity.

Following this approach, data integrity is an essential part of the data quality. Any poor data integrity practices undermine the quality of records and, therefore, the quality of final products.

It is thus necessary to take care about the data integrity process along the QAF to ensure data is recorded and keep recorded exactly as intended.

In the EM-SAP approach to establish a data integrity assurance system, both physical integrity and logical integrity are considered.

In the first one, elements associated with the management, physical and electronic storing are included, as well as the detection of erroneous data mostly related to human-induced errors -e.g. transcription from sheets to database- in this stage.

In the second one, different algorithms are applied to ensure the correctness or rationality of the data, thus including referential integrity, check constraints, etc.

Despite its importance within the quality assurance framework, there are not standard specific references for data integrity within the fisheries context, most part of the efforts including different elements of what achieving data integrity compliance means within their quality work without directly identifying those integrity elements. It is nevertheless straightforward to come forward other professional contexts where data integrity is deeply developed. ALCOA+ (US Food and Drug Administration, 2018) define data integrity standards/principles in a way that data is expected to be:

1. Attributable – Data should clearly show who acquired the data, when it was observed and recorded. Data must be attributable to the observer producing the data to track any changes occurring to the data.
2. Legible – Data stored must safeguard readability, should be easy to understand and recorded permanently.
3. Contemporaneous – Data should be recorded in the moment in which data are generated.
4. Original – Source data should be accessible and preserved in its original form. In case they have been modified, changes must be traced.
5. Accurate – Data should be free from errors, and conform with the protocol. Data must be routinely verified, through repeatable calculation or analysis.
6. Complete – full-length data
7. Consistent – all elements of the analysis are done in time in a expected sequence
8. Enduring – recorded in a permanent and mainatable form
9. Available – accessible for review and audit over the lifetime of the record.

EM-SAP is currently developing its integrity maintenance system as well as the QAF itself. To that purpose these principles are being helpful to change internal protocol procedures, adapt good practices and develop algorithms to detect integrity failures. Other sources, specially GFCM Data Collection

Framework (GFCM, 2017) are being used in order to strength EM-SAP QAF, enrich the indicators users and provide theoretical basis to the entire process.

When data integrity problems are found they are handled and solved according to the QAF. Also identification of the root cause allow the implementation of preventive measures (change in protocols, problems with observers, inclusion of new detection algorithms or analysis, etc).

Although the conceptual effort done to distinguish processes, most part of the QAF elements are linked at different stages. Hereby, integrity specific task are detailed, and other QAF elements detailed in different documents.

This document refers particularly to the the upload and data checking processes of the QAF.

3. Quality checks framework

Build accuracy checks into the design of the electronic system

3.1. Introduction

For the IEO, there is a subcontracted company who engage the samplers for the on-shore and on-board sampling. They are also in charge for the digitalization of the sampling information.

EM-SAP assumes the responsibility of the correct dumping into IEO's database (SIRENO) of the sampling information typed by the contracted company in a way that ensures the information saved complies with the criteria, format and protocol agreed in the team:

- Verification of the periodic dumping of samples in SIRENO.
- Standardization of the information placed under the criteria established by the sampling team.
- Errors correction.
- Review and maintenance of the integrity of the stored data.

In addition, during the process, EM-SAP proceed to identify recurring errors during typing. These errors are communicated to the outsourced company for the correct typing in the following months. The process allows the analysis of the problems in the reception of the data and its communication to the subcontractor company to avoid the repetition of the same problems.

The development of the process has led to the development of a framework to guarantee the quality of the information stored. The original initial controls (more focused on the format, accommodation in terms of master tables, etc.), are being expanded with elements such as the detection of outliers (e.g., landings weights), or taxonomic control (e.g., improbable identifications).

3.2. Responsible

The review process is carried out and coordinated by the data integrity manager.

At different stages it also involves the work of:

- The EM-SAP supervisors of each of the 4 RIM[1] areas.
- The EM-SAP supervisors of each of the 2 OAB[2] areas.
- Central computer service (SIRENO database, IEO, Madrid).
- The contracted company for sampling.

3.3. Conceptual frame quality checks review

There is a wide and continuously increasing number of checkings that can be made to the fisheries data to ensure data stored and, therefore, any subsequent use, have a proper quality.

[1] The on-shore sampling network (RIM) is divided into 4 geographical areas, each one of them with a supervisor in charge of the sampling.
[2] The on-board sampling network (OAB) is divided into 2 geographical areas, each one of them with a supervisor in charge of the sampling.

Checkings have been separated into 7 different groups, referring to the main purpose of the check. This work was originally based in the GFCM Data Collection Reference Framework (GFCM, 2017), and was adapted to our own needs and experience.

Every check developed controls a variable associated to one of the groups.

TYPE OF CHECKS

1. COHERENCE DATA SAVED

Logical relation between variables saved in database.

Examples of these checks are:

- samples with sampled weight but without lengths saved
- sampling weight equal to 0 but species with length data

2. CONFORMITY WITH MASTERS

Variables saved according to the masters.

Examples of these checks are:

- vessel code in official census (CFPO)
- port, gear, metier… according to masters

3. CONFORMITY WITH PROTOCOL

Some variables must be collected and saved in database according to the sampling protocol (mostly related to sampling of certain taxonomic group species).

Examples of these checks are:

- lengths samples of species aggregated in superior taxons following hierarchical agreed taxons. For example Lepidorhombus boscii saved as Lepidorhombus spp. in superior taxon and as L. boscii in lower one.
- mandatory sexed species doesn't been sexed

4. ACCORDING TO HISTORICAL DATA

Certain variables should be according the data stored in database from previous years.

Examples of these checks are:

- outliers review in catches range by fleet stratum

5. FIELDS PROPERLY FILLED

Some variables must be mandatory in the database or in a certain range.

Examples of these checks are:

- variable number of rejections of samplings must be filled
- number of vessels must be greater than 0. Throw a warning when is greater than 2 (our fisheries rarely have trips with more than 2 ships; restricted to PTB activity)

6. EVIDENCE OF ERROR / EVIDENCE OF POSSIBLE ERROR

Through the data exploration, some checks evidence the possibility of errors committed in the sampling process.

Examples of these checks:

- theoretical weight of lengths sampled equal to 0
- sampling weight greater than landing weight

7. PLAUSIBILITY

The value of variables must me reasonable and probable. This indicator refers to the Principle of reality (Desnos' principle)[3].

Examples of these checks:

- detection of doubtful species
- outliers in length range of species

The number of checks is being continuously expanded, having at the beginning of 2019 60 checks.

[3] Accommodated from GFCM Data Collection Framework.

3.4. Review methodology

The process is carried out in several phases that involve different checks and treatments of the information received as described in the diagram below.

Basically the process consists in 4 steps:

3.4.1. Pre-dump revision

Once the data of the contracted company has been received, an R script is used to homogenize the information with respect to the existing one in SIRENO and to detect errors.

In addition, the files are exported to the appropriate format required for direct dumping into SIRENO. This process is carried out by the data integrity manager.

The R script can be obtained from the public GitHub repository:
https://github.com/Eucrow/IPDtoSIRENO

3.4.2. Dump

It consists of pouring information from the previous phase into SIRENO. It is carried out by SIRENO's computer service (mainly because some information can only be incorporated properly within the database, i.e. SOP weights).

3.4.3. Post-dump revision

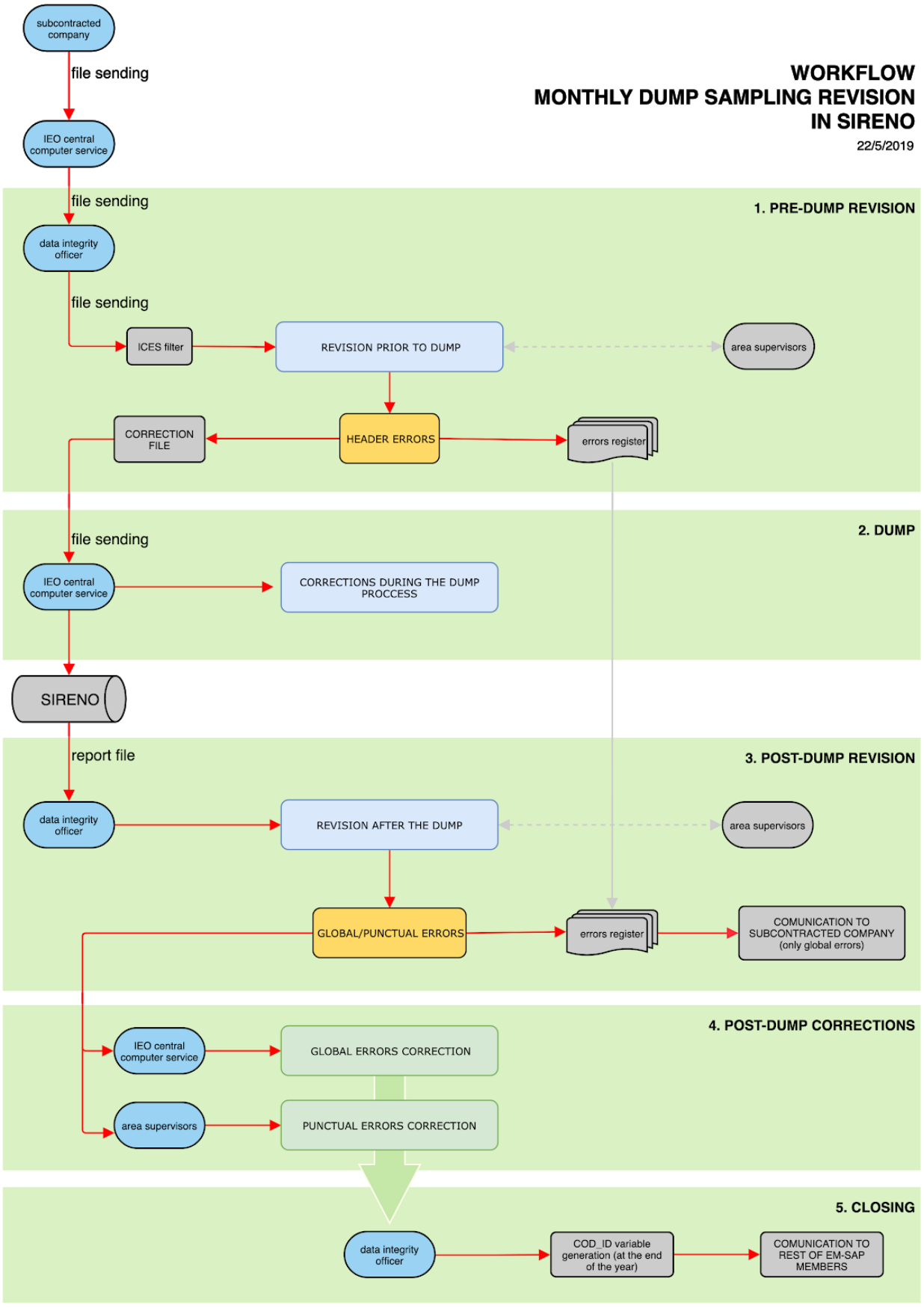
From the SIRENO output reports, an R script is used to detect errors and warnings.

- **Errors**: there is an objective mistake and must be fixed.
- **Warnings**: there could be a mistake or an evidence of a possible error.

The R script can be obtained from the public GitHub repository:
https://github.com/Eucrow/revision_volcado_R

3.4.4. Post-dump corrections

Errors and warnings from the ‘3.4.3. Post-dump revision’ must be reviewed by the supervisor of the geographical RIM or OAB area, as it can’t be done directly by the data integrity manager nor automatically. It usually entangles revision of originally paper sampling sheets and/or communication with the sampler.



3.5. Periodicity

The process is done:

- **Monthly**. Once data file is received from the outsourced company.
- **Annually**. Before mid-February of the following year, phases 3. Dumped data revision and 4. Corrections in SIRENO will be repeated with all year data in order to apply the new checks that will be added to the overall process and make sure all the errors detected in the monthly revision has been corrected.

4. Discrepancy checking (database - sampling sheets)

4.1. Introduction

Data integrity also refers to the absence of unexplainable discrepancies between data in records dumped to the database/s and data in the original records registered by the samplers/observers on the market and on-board.

The discrepancy checking is the verification that the information in the database corresponds to the information collected in the sampling statements.

Part of this process is taken during the quality checks review where **errors** and **warnings** identified by the algorithm (3.4.3 Post-dump revision) must be reviewed by the supervisors, usually implying review of the original sampling sheets .

4.2. Responsible and time

Sampling sheets are received and checked by the responsible personnel during the month following their completion as soon as the originals are received.

This is done by the corresponding area supervisors.

- **RIM:**

  Port sampling, made by the subcontracted company and dumped to SIRENO by the EM-SAP. The check is carried out after the monthly dump sampling revision has been finished and the original sampling sheet has been received.

- **OaB:**

  In on-board sampling, where the computer record is made by the EM-SAP team, the check is made at a time other than typing.

4.3. Checks

To check following fields are verified:

Table SIRENO database	Variables
Head	Port, Date, Origin, ESTRATO_RIM, Ship, Type of sampling, Number of rejections.
Weights	Species, Category, Sex, Landed Weight and Sampled Weight.
Lengths	Species, Category, maximum and minimum size and sum of number of individuals.

After the discrepancy check, the checked sample box is marked in the sample header in the SIRENO database.

A procedure has been established in R for the check, in such a way that the revision of the variables is facilitated. It has been agreed the use of this procedure and the need to check only the variables considered to optimize working time.

5. Storage of sampling data

5.1. Objective

Following commented ALCOA+ guidelines all data recorded must be legible (readable) and permanent. Ensuring records are readable and permanent assists with its accessibility throughout the data lifecycle. This includes the storage of human-readable metadata that may be recorded to support an electronic record.

EM-SAP protocol:

- Upload and integrity management of the data in the database.
- Digitization of the original sampling sheets and archive.

5.2. Responsible and time

Supervisors in charge of each of the areas of the RIM and OAB.

Digitization is done in a month time after the sampling event.

6. Main bibliography of interest

Boritz, J. 2005. "IS Practitioners' Views on Core Concepts of Information Integrity". *International Journal of Accounting Information Systems*. Elsevier. Volume 6, Issue 4, December 2005, Pages 260-279.

GFCM, 2017. GFCM Data Collection Reference Framework (DCRF). Version: 2017.1

U.S. Food and Drug Administration. Pharmaceutical Quality/Manufacturing Standards (CGMP). December 2018.

Wikipedia contributors, 'Data integrity', *Wikipedia, The Free Encyclopedia,* 8 February 2019, 15:30 UTC, <https://en.wikipedia.org/w/index.php?title=Data_integrity&oldid=882362079> [last access 6 May 2019]