MARE/2020/08 - SI2.839444

DEVELOPMENT OF THE REGIONAL DATABASE FOR THE MEDITERRANEAN AND BLACK SEAS

This project has financed under the European Maritime and Fisheries Fund (EMFF)

Med&BS RDBFIS







Deliverable 3.1

Specifications for the RDB and final requirements including a Minimum Viable Product (MVP) definition

P. Carrara, M. Zilioli

Partners involved:

CNR, HCMR, COISPA, CIBM, NISEA, IFREMER

Table of Contents

Acronyms

AS-IS analysis	Analysis of the current state
CFP	Common Fisheries Policy
DC	Data Call
DCF	Data Collection Framework
DCRF	Data Collection Reference Framework
EU	European Union
EWG	Expert Working Group
FDI	Fisheries Dependent Information
GUI	Graphical User Interface
GFCM	General Fisheries Commission for the Mediterranean
ICES	International Council for the Exploration of the Sea.
JRC	Joint Research Centre
LP	Large Pelagic
MEDIAS	Mediterranean Acoustic Survey
MEDITS	MEDIterranean Trawl Survey
MS	Member States
MVP	Minimum Viable Product
RCG	Regional Coordination Group
PET	Protected, Endangered and Threatened species
RCG Med&BS	Regional Coordination Group of the Mediterranean and Black Sea
RDB	Regional database
RDBES	Regional Database and Estimation System
RDBFIS	Regional Database Fisheries Information System
SC	Steering Committee
SDEF	Standard Data-Exchange format
STECF	Scientific, Technical and Economic Committee for Fisheries
STREAM (project)	STrengtheningREgional cooperation in the Area of fisheries biological data
TO-BE analysis	Improved future state
UML	Unified Modeling Language
VMS	Vessel Monitoring System
WGRDBESGOV	Working Group on Governance of the Regional Database & Estimation System
WKRDB-EST	Workshop on Estimation with the RDBES data model

Introduction

Work Package 3 (WP3) is in charge of producing the Deliverable 3.1, whose objective is to illustrate the final requirements of the Med&BS Regional Database and Fishery Information System (MED&BS RDBFIS) by including the **definition of a Minimum Viable Product (MVP)** for the application. The content depends on the results of activities performed by WPs operating in upstream phases of the project (i.e., WP1: AS-IS analysis to describe the current setup and business needs; WP2: TO-BE analysis: what we need in future) as well as on the WP3 assessments pertaining the technological, human and policy dimensions of the system (i.e., use cases, RDBFIS governance and membership model, data policy).

In the timeframe of WP3 actions (from month 4 to month 8), the outcomes of WPs, which are hierarchical interconnected to WP3, constituted the basis of the present elaboration and are referenced in the following section "Contribution and Materials". For sake of completeness, the documents elaborated by WP3 are enclosed in **APPENDIX - Table of annexes**.

Contribution and Materials

This section highlights the information sources considered to produce D3.1, which are related to the documents and interactions (represented in Figure 1) with other WPs within the project and with actors relevant in the RDBFIS lifecycle (like data producers and stakeholders). Interactions with WP2 have been important; also discussions with the Regional Coordination Group (RCG) have been taken into account, as in the meetings with RCG, Member States (MSs) and DG MARE representatives expressed their needs and constraints with respect to RDBFIS.

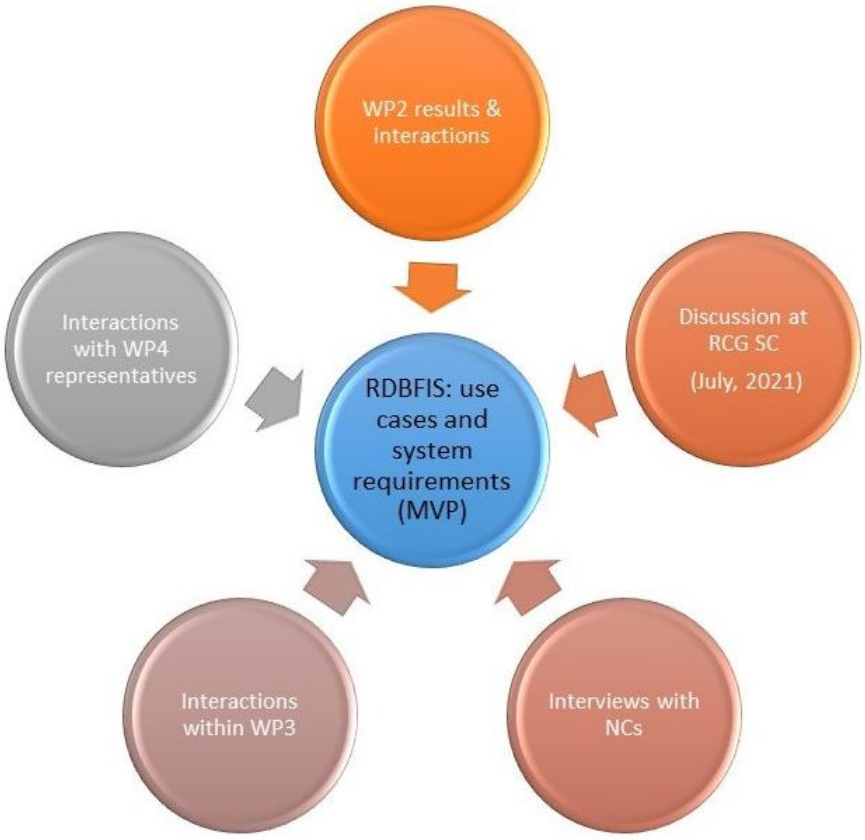

Figure 1. Activities contributing to the definition of RDBFIS use cases and requirements (including MVP)

In more detail, the specific documents considered so far are listed below:

- Milestone 2.1 - List of the RDB features needed to answer to the data collection submission and reporting obligations (month 4);
- Milestone 2.2 - Recommendations and requirements for the development and updates of the data validation and quality checking tools to be foresee for the RDB (month 6);
- RDBFIS main features and compatibility issues with RDBES (Communication of the SC of the Med&BS RDBFIS to RCG chairs and National Correspondents);
- Results of the RDBFIS bilateral meetings with MS (Presentation made the coordinator of the grant);
- Working document - Formats and tables to be implemented as MVP (provided by WP2/4);
- Working document – D6.1 Compilation and classification of quality checks at the national level (STREAM, MARE/2016/22)

RDBFIS specifications and final requirements

A first step in the activity of WP3 consisted in the definition of use cases, described in a document delivered as Milestone M3.1 “Use cases and list of actions defining the interactions between a user and the system to achieve a goal" (see **ANNEX V (RDBFIS_useCases_v0**) and **ANNEX VI (RDBFIS_useCases_v1**)). The use cases described in the current version are:

- Input user (e.g., MS representative, other authorized users)
    - Upload Sampling Data in RCG data format
    - Upload Sampling Data in ICES RDBES data format
    - Upload Landings Data in RCG data format
    - Upload Survey Data in MEDITS data format
    - Upload Survey Data in MEDIAS data format
    - Query-based export of detailed/aggregated data (Data Call format)
- Output user (e.g., end user representatives)
    - View/Export aggregated data in Data Call formats

Also, the M3.1 pinpoints other use cases (e.g. Upload aggregated data in data call formats (I.e., Med&BS, FDI, GFCM/DCRF), Export detailed data) that need to be formalised through the contributions of the experts or technicians of other WPs.

Use cases describe the typical interactions between the users of the system and the system itself; in the Unified Modeling Language (UML) formalism adopted in M3.1, interaction is represented as a sequence of steps (i.e., scenario). Scenarios present different outcomes (i.e., successes, failures, alternative pathways). In use case-speak, users are referred to actors and an actor is a role that a user plays with respect to the system. A single actor may perform many use cases; a use case may have several actors performing it and one person may act as more than one actor.
Use cases definition helps in identifying users, features, processes, constraints, inputs and outputs of the system. Moreover, the identification of ‘necessary’ use cases drives to the definition of the MVP (i.e., “describing the basic minimum features that would make the RDB operational”).
The following section describes the requirements derived from the contributions and materials presented above, with focus on the identified use cases.

System requirements

MED&BS RDBFIS ought to be a multicomponent, web-based information system consisting of:

- Front-end platform
  - Web-based Graphical User Interface (GUI)
  - Web pages for public contents
- Back-end platform
  - Data validation/processing layer
  - Data ingestion
  - Authentication and Authorization layer
  - PostgreSQL Database

User requirements

MSs were consulted in the context of WP2 activities (month 2 to 6) in order to collect, through semi-structured interviews, their main expectations in storing data for regional assessment by means of new tools. Discussions among the RDBFIS experts with the MS representatives have started in order to present the system under development and to investigate: (i) the existing systems used to support the DCF and datacall needs, (ii) the sampling scheme. Meetings have been held with Cyprus, Greece, Croatia, France, Italy, Slovenia, Bulgaria, Romania, Spain and Malta (see **ANNEX II (NCs-ICES-MCDA Meetings)**). These bilateral meetings are crucial to map the MSs requirements; nevertheless, they highlight so far reluctances and sometimes ambiguous demands of the input users as far as the routine adoption of the tool (e.g., evaluation of IT skills to perform analysis, deployment issues of software packages). In particular, some MSs initially seem to be unconvinced to store detailed data and others show some inertia to change their data management practices. On the other hand, other MSs demonstrate availability to use the new tool and express some preferences about the input formats to adopt.

Policy requirements

Policy requirements are detailed in documents listed in the following Annexes:

- Proposal for guidelines in the Med&BS RDB SC: collection phase (**ANNEX VII (Med&BS RDB SC guidelines v1**)),
- Members to be included in the SC (Membership model), (**ANNEX VIII (Med&BS RDB SC membership_v3**))
- Data policy for the Regional Database and Fishery Information System for the Mediterranean and Black Sea (**ANNEX IX (RDBFIS_dataPolicy)**).

Defining the Minimum Viable Product (MVP) and the different scenarios

MVP definition

The Minimum Viable Product (MVP)is the most pared down version of a new product (e.g., software, market/industry item) that can be still released and used[1]. The MVP identifies the basic features or "minimal requirements"[2] that are needed to satisfy early adopters, while the final, complete set of features is only designed and developed after considering feedbacks from the product's initial users. Testing is instrumental in the MVP-based development approach: It is the phase in which the actual behaviour of

[1] *https://www.techopedia.com/definition/27809/minimum-viable-product-mvp*
[2] *York, J. L. and Danes, J. E. "Customer Development, Innovation and Decision-Making biases in the Lean Startup." Journal of Small Business Strategy. Vol. 24(2), pp. 21-39, 2014.*

users with the product or service is observed and their experiences (e.g., successes/failures in achieving a goal) are recorded. The three key features of an MVP are:

1. It must have enough value that people are willing to use it
2. It demonstrates enough future benefit to retain early adopters
3. It provides a feedback loop to guide future development (something to be tested)

RDBFIS MVP definition

Taking into account the above definition, the RDBFIS MVP should guide the first implementation of the software that would make the system operational (i.e., it enables Med&BS RCG to perform fishery regional assessments, and MSs to fulfil to the data collection submission and reporting obligations). This condition is verified if the RDBFIS MVP is:

i) securely deployed and made accessible to the users through the HCMR servers
ii) testable at least by one representative per MSs and by one representative for each end users' category to achieve the tasks assigned to their respective users
iii) successfully adopted by all the MSs to populate the system with real data and used to automatically perform aggregations/analysis with data
iv) successfully adopted by Med&BS RCG to extract detailed/aggregated data through one unique access point.

The minimum features required in the RDBFIS MVP are classified below within three categories:

- the minimum set of interactions between the users and the system to be tested and performed[3]
- the minimum set of the database features
- the minimum set of functionalities/tools

Minimum set of database features

The minimum set of database features (i.e., data and sampling types/domains, aggregation levels, codes/reference lists) is essentially represented by the relational tables (i.e., main and parametric tables), which need to be implemented to build the RDBFIS database schema, allowing detailed/aggregated data ingestion and syntactical/integrity checks for the main domains covered by RDBFIS (i.e., commercial biological data, survey data). Each relational table corresponds to an ordered sequence of attributes and allowed values, specified in the data requirements bythe end-users of the system (i.e., Med&BS RCG, EU-DGMARE, FAO-GFCM).The organization of the RDBFIS database schema should allow the upload and management of both detailed and aggregated data.

The selection of the minimum set of tables to be implemented is listed below; the acronyms are in italics, grouped under the generic name of the data format (serialized with numbers) to which they are referred to. The parametric tables to assure integrity checks are required in the implementation plan and they enforce the syntactical and range constraints agreed by end users; they are not included in the list below but they can be obtained by WP4 following the data format documentation.

1. RCG data format

   Biological detailed data (for hierarchy levels of the relational tables, see M2.1):

   - Commercial Sampling data (*CS*)

[3] They correspond to use cases presented in section "RDBFIS specifications and final requirements"

Transversal aggregated data

- Commercial Fisheries Landings statistics (*CL*)
- Commercial fisheries Effort statistics (*CE*)

2. COST data format, implemented to facilitate the compatibility with the R tools developed in STREAM (not to be used as data input format by MS).

3. ICES RDBES data format

- *rdbes_bv*
- *rdbes_ce*
- *rdbes_cl*
- *rdbes_de*
- *rdbes_fm*
- *rdbes_fo*
- *rdbes_ft*
- *rdbes_le*
- *rdbes_lo*
- *rdbes_os*
- *rdbes_sa*
- *rdbes_sd*
- *rdbes_sl*
- *rdbes_ss*
- *rdbes_te*
- *rdbes_vd*
- *rdbes_vs*

Survey data formats (detailed data from acoustic and trawl surveys)

4. MEDIAS data format

- *medias_echosounder_param*
- *medias_processed_acoustic*
- *medias_surv_sset*
- *medias_surv_sset_bio*
- *medias_surv_sset_bio_spec*
- *medias_surv_sset_png*
- *medias_survey*
- *medias_survey_design*
- *medias_survey_identity*
- *medias_trawl_biodata*
- *medias_trawl_descr*
- *medias_trawl_haul*
- *medias_trawl_individual_biodata*

5. MEDITS data format-SOLEMON-Black Sea surveys

- *medits_ta*
- *medits_tb*
- *medits_tc*
- *medits_te*
- *medits_tl*

Datacall formats (aggregated data)

6. GFCM DCRF data format

- *dc_dcrf_task_ii1_landing*
- *dc_dcrf_task_ii2_catch*
- *dc_dcrf_task_iii_incidental_catch*
- *dc_dcrf_task_iv1_vessel_le15m*
- *dc_dcrf_task_iv2_vessel_over15m*
- *dc_dcrf_task_v1_fishing_effort*
- *dc_dcrf_task_v2_fishing_effort_gear*
- *dc_dcrf_task_v3_cpue*
- *dc_dcrf_task_vii2_length_data*
- *dc_dcrf_task_vii31_size_1st_matur*
- *dc_dcrf_task_vii32_maturity_data*

7. FDI (EU-DCF datacall format)

- *dc_fdi_a_catch*
- *dc_fdi_b_refusal_rate*
- *dc_fdi_g_effort*
- *dc_fdi_h_spatial_land*
- *dc_fdi_i_spatial_fe*
- *dc_fdi_j_capacity*

8. MED&BS (EU-DCF data format)

- *dc_medbs_alk*
- *dc_medbs_catch*
- *dc_medbs_discards_length*
- *dc_medbs_gp*
- *dc_medbs_landings_length*
- *dc_medbs_ma*
- *dc_medbs_ml*
- *dc_medbs_sra*
- *dc_medbs_srl*

Minimum set of functionalities/tools

The minimum set of functionalities/tools includes RDBFIS facilities which will allow users:

- to deliver/access data
- to validate and carry out data quality checks
- to perform analysis/automatic aggregation on data

They are described here following according to the tasks that they enable.

Minimum set of user interfaces to allow users to interact with the system

The minimum set of user interfaces needed to start the MVP testing is reported below. Their development and release together with the other basic features of the system is a prerequisite of the MVP. They are:

- Web access to the system
- Access to the system facilities (upload/export) through authentication with credentials
- Upload data
- Export data

Minimum set of uploading procedures

The minimum set of uploading procedures to be developed will allow input users to support data submission required by different end users. Uploading procedures will be run through ad hoc GUIs, which need to be tailored on the input user profile (i.e., the subset of data tables the user is entrusted to submit in the RDBFIS). Procedures must enable the delivery of both historical data (i.e., data of previous data calls) and data that meet ongoing data calls, in different aggregation levels according to policy requirements.
Uploading facility interface has to provide a way for selecting one file at a time, to check the user input as well asto inform her/him on the outcome of the process.
Further specifications for uploading facility are the following:

- Exchange formats: The file format to upload data should be the Comma Separated Value (CSV). Naming convention, special characters allowed, mandatory headers (i.e., column names) should also be specified.
- Upload tools: Besides data upload, other tools have to be developed to allow the user to track the process:
    - Uploads history page: log box where a user can track the date and timestamp of each upload event
    - Submission status (maybe not compulsory)
- Editing/Deleting procedures: The uploading procedure must allow the user to edit/delete data in case of errors the user would amend.
- Validation tools and quality check procedures linked to upload: During uploading procedures the user must be allowed to use tools to validate the data (e.g. tools for the detection of formal errors on the data type as numerical, character, etc., allowed codes and ranges according to the end users data requirements) and to assess the quality (e.g. to evaluate inconsistencies with respect to all data stored in the same table; to evaluate coherence of spatial and temporal data) of data already present in detailed data. The validation tools and quality check procedures will be run after the upload of a file and before data are stored in the database. These procedures need to inform the user about the outcomes of the operations. The minimum validation and quality checks operations are:
    - syntactical checks and integrity constraints on detailed/aggregated data
    - reporting module (i.e., errors log file) of formal checks performed on the tables
    - a priori QCs for RCG sampling data format (CS) (developed in STREAM) to be updated/adapted
    - new a priori QCs for RCG sampling data format (CS) (summarizing the number of individual biological data, number of trips by port, trip position) (to be developed)
    - a priori QCs and reporting (i.e., errors log file) based on a new RoME package for MEDITS survey data
    - reporting module (i.e., errors log file) based on a priori checks for RCG sampling data format (CS) (developed in STREAM)
    - reporting module (i.e., errors log file) based on a priori checks for RCG sampling data format (CL) (developed in STREAM) to be updated/adapted

Minimum set of validation tools and quality checks tools (export phase)

Besides the minimum set of validation/quality check tools linked to upload, MVP should include tools to export both detailed and aggregated data. They are defined as *a posteriori quality checks* (i.e., scripts running on aggregated data which are extracted directly from the database by performing queries or derived by aggregation modules starting from detailed data). They are characterized by compulsory scores (i.e., mandatory/recommended/optional) to be agreed by the Steering Committee of the RDBFIS and they will be performed by output or input users before delivering data to the requesters. They are listed here following:

- a *posteriori* QCs applied on DGMARE-Med&BS data(developed in STREAM) to be updated/adapted
- a *posteriori* QCs for the Table A of FDI (to be developed)
- a *posteriori* QCs specifically designed for the G, H, I and J FDI tables aimed at verifying the consistency of transversal variables (effort and landing, also by rectangle) as well as the spatial and temporal coverage (to be developed)
- reporting module (i.e., errors log file) based on a *posteriori* checks for DGMARE-Med&BS datacall (developed in STREAM)
- reporting module (i.e., errors log file) based on new *a posterior* QC procedures on the FDI tables (G, H, I and J) aimed at verifying the consistency of transversal variables (spatial and temporal coverage)
- reporting module (i.e., errors log file) based on new a *posteriori* QC procedures specific for the GFCM DCRF tasks on biological information (to be developed)
- new R libraries of data validation and quality check functions (to be developed)

Minimum set of data processing tools and delivering to specific data calls tools

It is represented by:

- tools to automatically aggregate the detailed data previously stored in the database into the datacall formats (DGMARE Med&BS, FDI, GFCM)
- data processing tools developed to assist RCGs in performing other type of analysis (to be defined by WP4)

Minimum set of exporting procedures

The minimum set of exporting procedures to be developed will assist both input and output users to extract data from the database according to the access right granted to them (see data policy document). The exporting procedures to be assured are:

- query-based exporting procedure: this procedure will allow to export data previously stored in the database through pre-compiled queries. The procedure will allow users to apply filters directly on detailed and aggregated data by leaning on pre-compiled queries (i.e., list of operators or codes to apply to the data tables which will be accessed)
- R-based exporting procedure: this procedure will allow to automatically aggregate and formatting detailed data according to the data calls format that is selected. The procedure needs to be made available through an R module directly accessible from the exporting interface

Scenarios for the RDBFIS MVP

This section presents two scenarios for the RDBFIS MVP, i.e. different outcomes of the RDBFIS MVP testing (detailed in the following section "Iterative development roadmap for WP4"). Among the possible scenarios, we present those the project could most probably face, given some issues reported at the RCG meeting of September 7-9, 2021; in fact, this meeting highlighted that some MS representatives are not familiar with IT tools to perform quality checks/aggregation, and that actions to train MSs in this important task are necessary. WP4 planned training activities on routines which will be coordinated within STREAMLINE project and will be addressed to IT staff appointed for each MS.
At the beginning of meetings with MSs, some of them showed reluctance to share data; therefore another third possible scenario could be considered, whereby MSs do not upload data or do not upload all the data requested. However, in the following meetings with MSs and at the RCG meeting of September 7-9, 2021, after a thorough presentation of RDBFIS, no country has risen any objection in contributing to RDBFIS data upload. Therefore this third scenario has not been described so far.
In both scenarios we assume that the RDBFIS MVP is developed and operationalised. The degree (i.e., high or low) of appropriation (i.e., acquaintance and adoption) of the system's tools by their users, affects the chance that one of the two occurs.

Scenario A

The first scenario can take place when, after a successful testing phase, input and the output users acquire homogenous capacities to exploit the system's facilities. In particular Scenario A will take place if:

- input users will be enabled to routinely use both the validation and quality check facilities (i.e., performing error log interpretation, data correction and resubmission) in order to provide verified data to end users
- input users will routinely adopt facilities to automatically aggregate the detailed data in data calls formats
- output users will be enabled to run both the quality checks and the automatic aggregations of the detailed data by relieving MSs from the burden of performing 1) all the aggregations and 2) submit them through different platforms as it happens in this moment.

These conditions produce the maximum benefits for all users of the system, lead to the system population with robust data and information, and enable a flexible reuse of data uploaded.

Scenario B

The second scenario can take place when, after the testing phase, input and output users acquire heterogeneous capacities to exploit the system's facilities. In particular Scenario B will take place if:

- Input users will be partially enabled to routinely use both the validation and quality checks facilities. This condition is realized for example when data are verified for syntactical/integrity checks by all the input users but not all of them always succeed in checking data with R facilities
- output users will be partially enabled to run quality checks and the automatic aggregations.

These conditions produce a system where data stored are validated and controlled according to end users' specifications, by ensuring a first level of data quality for end users (i.e., syntactical checks). By granting the end users access to quality checks functions, the controls/aggregations that were not performed by input users, can be run by the end users if they are familiar/trained. This scenario does not prevent system functioning either data uploading itself. Nevertheless, it implies a longer follow up of the users' capacities and possibly further trainings and support.

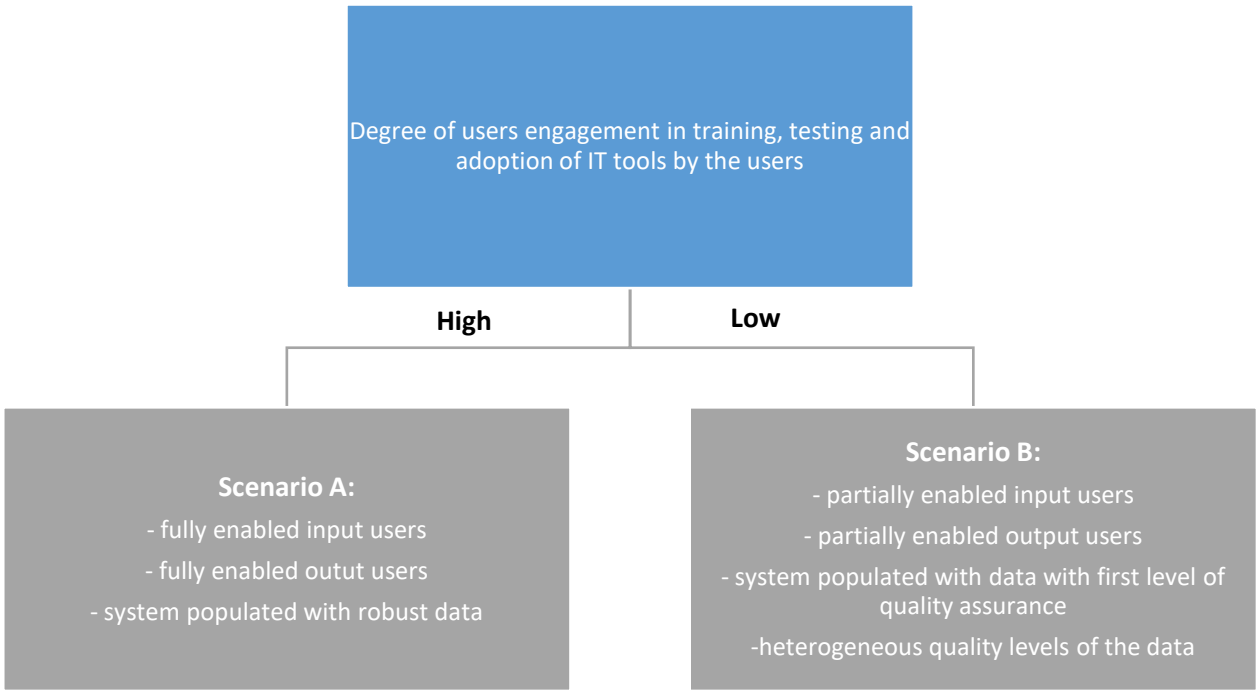

Figure 2. Dependencies between the degrees (high or low) of users' engagement in training, testing and IT tools adoption (blue box) and the scenarios (A and B, grey boxes)

Iterative development roadmap for WP4

This section illustrates a short list of checkpoints, which can mark and define the roadmap for WP4 development activities. The roadmap can be followed once the MVP will be finalised and used as version 0 of the system. In fact, the WP4 development team demonstrates to lean upon strong background works and tools derived from previous grants. Given the WP4 self-reliance cycle of checkpoints was outlined, will be iterated by WP4 once any new module is integrated in the previous version of the system

The checkpoints are the following:

1. WP4 requests real data for in-home testing (i.e., tests performed by WP4 components) of the current version of the system
2. WP4 perform tests
3. WP4 performs debugging/fixing/update of the system version
4. WP4 selects one representative for each user category (i.e., the testers)to perform controlled test on user permissions and workflows
5. WP4 organizes and supervises activities of the testers, who test the current version of the system
6. WP4 gathers feedback from testers
7. WP4 repeats point 3

8. WP4, supported by WP5, organizes workshop/s to present the system version to the users' community: facilities usage (e.g., diversified on user category or on facility) will be introduced by WP4 with simulation with real data; the whole community will agree on a formalized language through which users transmit feedbacks and new requirements
9. WP4 gathers feedback from workshops' participants
10. WP4 repeats point 3.

APPENDIX - Table of annexes

ANNEX description	ANNEX name
M3.1 - Use cases and list of actions defining the interactions between a user and the system to achieve a goal (month 7)	ANNEX V (RDBFIS_useCases_v0) ANNEX VI (RDBFIS_useCases_v1)
Proposal for guidelines in the Med&BS RDB SC: collection phase	ANNEX VII (Med&BS RDB SC guidelines v1)
Members to be included in the SC (Membership model)	ANNEX VIII (Med&BS RDB SC membership_v3)
Data policy for the Regional Database and Fishery Information System for the Mediterranean and Black Sea	ANNEX IX (RDBFIS_dataPolicy)