CCAMLR

Commission for the Conservation of Antarctic Marine Living Resources
Commission pour la conservation de la faune et la flore marines de l'Antarctique
Комиссия по сохранению морских живых ресурсов Антарктики
Comisión para la Conservación de los Recursos Vivos Marinos Antárticos

SCIENTIFIC COMMITTEE

SC-CAMLR-XXXV/BG/25 Rev. 1

12 October 2016

Original: English

# Developing the Secretariat's data management systems

CCAMLR Secretariat

EN

# Developing the Secretariat's Data Management Systems


## Abstract

The CCAMLR Secretariat's data management systems provide the secure infrastructure and repository for CCAMLR data which are submitted by CCAMLR Members to support the policy, scientific and administrative work of the Commission and Scientific Committee. These systems are being redeveloped as part of a long-term program of work which started in 2013. The work has focused on the development of structural and data elements of the enterprise data model, data warehouse, quality assurance, reference data, data extracts and metadata, and the appraisal of candidate systems for automated loading of, and reporting on, data submissions. These elements form the building blocks of the new system.

This paper provides the background, key achievements to date and work plan for the redevelopment of the CCAMLR data management systems. It updates Members on key tasks associated with transition to a new data warehouse, development of data extracts and metadata, and the establishment of a data management group. The paper also outlines the anticipated changes, and associated benefits for users of CCAMLR data. The work plan implements the advice of SC-CAMLR and its Working Groups on data traceability, system testing and evaluation, user training, data extracts and metadata, and establishing a data management group.

The user community can expect improvements in data quality assurance, database documentation and ease of use as the new system is progressively rolled out, including increased: integration across CCAMLR data; user-focussed documentation supporting data systems and CCAMLR data; availability of metadata libraries and data dictionaries; engagement, transparency and functionality relating to data submission; data quality assurance; dissemination of CCAMLR data and Secretariat support for data analytics and interrogation.


**Note: This revision includes a 2-year work plan as requested by WG-FSA-16 (see Table 2).**

## Table of Contents



# Introduction

1. The collective term 'CCAMLR data' refers to data which are submitted to the Secretariat by CCAMLR Members to support the policy, scientific and administrative work of the Commission and Scientific Committee and to give effect to Article XX of the Convention. CCAMLR data may also be submitted by Non-Contracting Parties participating in CCAMLR activities such as the Catch Documentation Scheme.

2. The CCAMLR Secretariat's data management systems provide the secure environment in which CCAMLR data are managed and made available to the CCAMLR community and other users in accordance with the rules for access and use of CCAMLR data (https://www.ccamlr.org/en/data/access-and-use-ccamlr-data). These data include structured data (data reported in fixed fields within a record or file, e.g. an observer logbook) and unstructured information (data or information that are not reported in a traditional row-column format, e.g. a document[1]).

3. The Secretariat's governance of CCAMLR data aspires to best practice and international standards to improve the quality and availability of data and information, ensure the confidentiality and integrity of data and information, promote the integration of data and information, support effective business processes and informed decision making through accurate data, and reduce Secretariat costs through efficient management of data and information.

4. As requested by WG-SAM-16, this paper provides background, key achievements to date and plans for on-going work associated with the redevelopment of CCAMLR's data management systems and CCAMLR databases. It updates Members on key tasks associated with transition to the data warehouse, data extracts and metadata, and the establishment of a data management group, and outlines the anticipated changes, and associated benefits, for users of CCAMLR data.

5. The user community can expect improvements in data quality assurance[2] (DQA), database documentation and ease of use as the new system is progressively rolled out, including increased:

[1] The Secretariat is also a repository for some physical data such as returned tags and otoliths.
[2] See Appendix B for an explanation of this term

- **Integration across CCAMLR data** - Integration provides the benefit of being able to systematically use data from various data collection systems to enhance the quality and value of each data source (e.g. vessel reported data, observer collected data, vessel monitoring and catch documentation systems), leading to more consistent, reliable and timely delivery of data.
- **User-focussed documentation supporting CCAMLR data systems and CCAMLR data** – CCAMLR data users require a deep understanding of the structure and meaning of the data being used. Gaining a better understanding of CCAMLR data starts with access to clear documentation of the datasets (entities), and associated data items (attributes) within each dataset, and should include the data source(s) and data collection regime(s), and the application of DQA. All uses of the same data item regardless of the source or purpose will use the same name.
- **Assurances regarding CCAMLR data quality** - Data users require full traceability of data from submission to data delivery. On submission of data, corrections and assumptions must be fully auditable. Modern data systems are expected to support traceability and reversibility. Automated and manual mark up of data (e.g. data corrections), auditable event logs, and the ability to undo/modify/redo any corrections all become metadata that are securely managed to provide a complete history of the data life cycle.
- **Availability of metadata** - Data are most valuable when they are accompanied by metadata[1] describing where the data came from, if and how the data were processed and transformed, how to interpret the data and how DQA was implemented. Metadata facilitates the CCAMLR datasets to be searchable, discoverable and potentially shareable (under the specified data security provisions).
- **Engagement, transparency and functionality of data submission** - The process of submitting data should be completely transparent and engage with data submitters. Data submission should be supported by standard operating procedures, data quality feedback, visualisations adding value to the submitted data, and training and documentation to ensure best practice and productivity gains are realised. Standard tools for the collection and transmission of data will be explored in particular the evolution of e-reporting platforms. Submitted data should be visible to the data provider with appropriate audit controls indicating who provided the data and when a submission was made.
- **Dissemination of CCAMLR data** - Data access must be convenient and systematic based on strict and agreed security protocols. As CCAMLR metadata increase in visibility, requests for access to secure data will undoubtedly increase. Data requests should be supported and facilitated with a rigorous and efficient approval process. Authoritative data (i.e. CCAMLR is the issuing authority), in particular CCAMLR management areas, will be accessible as a web service for assimilation by external users and automated (machine-to-machine) communication.
- **Support for data analytics and interrogation** - The growing demands of modern data analytics requires access to timely, integrated, quality assured and well understood data. Data must also be available in a consistent and timely machine readable form, and represent accurate and complete records of what is measured and reported. Commonly used dimensions of interest may be added to datasets in the form of derived data; for example CCAMLR season and geographical area based on dates and positions. Derived data will be clearly differentiated from source data. Given the complex analytical processing typically carried out with CCAMLR data, the provision of key routine analytical functions will be provided as documented, versioned R packages.

## Data Management Systems

6. The redevelopment of the Secretariat's data management systems involves a long-term program of work which spans structured data and unstructured information that support the policy, scientific and administrative work of the Commission and Scientific Committee.

7. This work is supported by the Secretariat's Data Management Strategy which aims to maintain high-quality data services to assist decision-making by the Commission, SC-CAMLR and working groups, and to support Secretariat services. The strategy promotes:

- Compliance with relevant international standards
- Secure data storage
- Efficient, error-free data processing and administration
- DQA
- Systems based on comprehensive data models and robust architecture
- Integration of data and business processes
- Timely and efficient access to data, derived data and outputs
- Feedback for data and process improvements.

8. The redevelopment was initiated following an independent review of CCAMLR's data management systems in 2011 (CCAMLR-XXX/05). The review identified various projects to facilitate the implementation of its recommendations. Work on these projects and related activities commenced in 2012 and resulted in wide-sweeping changes in the organisation of the Secretariat and the delivery of services, including *inter alia*:

- Re-development of the CCAMLR website using a content management system to provide a structured approach to web resources including CCAMLR information, meeting document submission, vessel registry and fishery notifications (https://www.ccamlr.org/). The website also supports an internal work flow for managing and translating web content, as well as preparing and translating circulars. In addition, a meetings server was developed to support meeting activities including the preparation and adoption of meeting reports (see https://meetings.ccamlr.org).
- Migration of the Secretariat's IT server infrastructure to virtualised instances of servers leading to improved efficiency in the management of network resources including web and data services. Part of that development included upgrading the SQL Server software and consolidating SQL instances.
- Development of an online GIS in collaboration with the British Antarctic Survey (UK). The GIS (https://gis.ccamlr.org/) provides a repository for managing and presenting spatial data for CCAMLR including statistical subareas and divisions, management areas, VMEs and conservation planning domains. The online GIS also allows users to view the latest extent of sea-ice, search for place names and zoom to specific areas, and authenticated users can download data layers and post community data.
- Review and implementation of a new Vessel Monitoring system (VMS).
- Review and current implementation of a new web-based system for the Catch Documentation Scheme.
- Development of a web-based Search and Rescue service for authorised users from Maritime Search and Rescue Coordination Centres (Conservation Measure 10-04) which queries VMS data and returns a list of fishing vessels in the vicinity of a maritime incident.

9. Much of the work completed to date is foundational and provides the platform for subsequent work on datasets that support the analytical and advisory functions of SC-CAMLR and its working groups. Work to date has focused on the development of structural and data elements of the Enterprise Data Model[1] (EDM), the Data Warehouse[1] (DWH), systems for transformation of submitted data to secure storage (termed Extract, Transform and Load[1]: ETL), DQA, reference data, data extracts and metadata, and establishing a data work flow for automated uploading of, and reporting on, data submissions. These elements are the building blocks of the redeveloped data management systems which are described in the following sections.

10. Current CCAMLR databases consist of a series of databases which have been developed at various stages in the past, and which implement various data formats, naming standards and DQA processes. The redevelopment will integrate the CCAMLR data in these databases, and will apply common formats and naming standards. In addition, the greater level of DQA which is being implemented during the redevelopment will also contribute to the quality assurance which is being applied to existing CCAMLR data.

11. A library of current CCAMLR data-related documents is listed at Appendix A (available on the WG-FSA meeting server, and on request from the Secretariat).

## Data Management Group

12. WG-SAM and WG-EMM agreed in 2016 that a Data Management Group (DMG) be established to provide a conduit between data users and the Secretariat in order to provide high-level input on the management and development of the CCAMLR databases including data extracts and data products (WG-SAM-16 paragraph 2.20 and WG-EMM-16 paragraph 6.21). A proposal for such a group, based on ICES's Data and Information Group, is outlined below for consideration by SC-CAMLR.

13. The DMG would be convened by the Secretariat and work virtually using the CCAMLR eGroup facility. The DMG would hold regular virtual meetings to review work plan progress, may meet physically on an opportunistic basis and where possible in association with a CCAMLR meeting. Topics of interest to the group and a wider audience would be posted on CCAMLR eGroups.

14. The proposed Terms of References for the DMG are to:

- Provide guidance across multiple disciplines regarding CCAMLR data and information, such as spatial, oceanographic, VMS, fishery operational and biological data, as well as expertise on metadata, vocabularies, user guidance and DQA.
- Provide advice and guidance on an annual CCAMLR data systems draft work plan
- Provide strategic input to CCAMLR data systems development and management including data policy, data strategy, DQA, technical issues, rules for use and access and user-oriented guidance and development of subject areas.

## Enterprise Data Model

15. The EDM is a model of the enterprise databases and provides the architectural foundation for the overall data and information management at the Secretariat, including:

- an overall data architecture plan used for designing the DWH
- a comprehensive data dictionary of logical data objects which will be available to data users, thereby enhancing the user community's overall knowledge and understanding of CCAMLR data
- an effective tool for communicating and sharing data and information
- an effective tool for detecting and resolving issues in existing systems and databases that may impact on data integration, DQA and effective use of data.

16. The development and documentation of the EDM is facilitated through the use of modelling software which supports the development of data model diagrams and metadata, and the implementation of the model in a relational database. The development to date has focused on five core subject areas for which the EDM is largely complete:

- Fishery notifications (exploratory fisheries and krill fisheries)
- Vessel operations (e.g. licensing, transhipments and VMS)
- Vessel catch and effort (e.g. haul-by-haul data)
- Observer Scheme (scientific observer data)
- Tagging Program (CCAMLR and national programs).

17. Other subject areas are under development and modelling has been done to varying extent for:

- Reference Entities (e.g. vessels, gear types, geographic areas)
- Data audit and DQA
- Conservation Measures
- Catch Documentation Scheme
- Vulnerable Marine Ecosystems
- CCAMLR Ecosystem Monitoring Program
- Marine Debris
- Meeting Management
- People and Organisations.

## Redevelopment of the CCAMLR Databases

18. In the current system, most CCAMLR data are systematically and manually validated and uploaded to the CCAMLR databases from MS Excel data sheets. CCAMLR data extracts are available to users on request, via email and subject to the rules for access and use of CCAMLR data. For external users, data are usually provided in an MS Access database with example queries to illustrate key relationships between data tables. Each data extract contains a subset of data which are available at the time of the request, and users often request updates at various times the working groups' annual cycle. This system will be phased out as the redevelopment work progresses.

19. The redevelopment will use web services, data warehouse concepts, new analytical tools, many of which are open source to change how the Secretariat processes and delivers data to the user

community. Data users will experience consistency and clarity in both data format and content, data extracts formatted as comma separated values (csv) and other open source data products.

20. The redevelopment of the CCAMLR databases is being done for each subject area as a staged development, and each subject area is being developed in consultation with subject area experts (e.g. the Secretariat's business users and external users) with advice from the SC-CAMLR and its working groups, and future advice from the proposed DMG. The redevelopment spans the entire life cycle of each subject area dataset, and a generic example of such a life cycle is provided in Table 1.

Table 1: Generic data life cycle and supporting data management systems

| Data life cycle | Data management system |
|---|---|
| Collection | CCAMLR data forms (future step: smart forms) |
| Submission | Automated on-submission data upload |
| DQA (pre-entry) > re-submission | On-submission DQA report (to submitter) |
| Load to staging database | Data registry |
| ETL DQA | DQA report > pass/fail and audit log |
| DWH and integration | ETL and audit report |
| Metadata | Metadata update and dissemination |
| Use | Data extracts, analysis and reporting |

21. The redeveloped CCAMLR databases will include (Figure 1):

- an integrated data registry
- staging databases (i.e. a transaction environment) for holding submitted data
- a Data Warehouse (DWH) serving the user community with high-level quality-assured data and related metadata (including information on quarantined data)
- improved data work flow allowing greater automation in data processing
- improved DQA including 'on-submission' data checks, strict data tolerances in the transaction database, and new processes implemented during the transfer of data to the data warehouse
- a data exchange gateway which will allow data to be submitted using CCAMLR data forms (including smart forms and online submissions), as well as other methods which meet the data exchange criteria. These other methods may include Members' institutional databases and vessel-based data reporting systems (e.g. e-logbooks)
- Automated data upload and on-submission DQA reports provided to data submitters (such as Technical Coordinators who currently receive a report for submitted observer logbooks). This step will include a DQA pass/fail evaluation, and any DQA issues will be automatically referred to the submitters for resolution and/or re-submission
- Data extracts and metadata which will be available to users at regular intervals, with information about new and changed data since the last update. Extracts will be updated at predefined intervals which correspond to the Commission and SC-CAMLR annual work cycles.
- a data portal will allow Members to access the data which they have submitted (including data submitted by a flagged vessel), as well as data extracts which would be provided subject to the authorised user acceptance of the rules for access and use of CCAMLR data. This usage will be logged
- Consultation with the proposed DMG with timely adoption of feedback received.

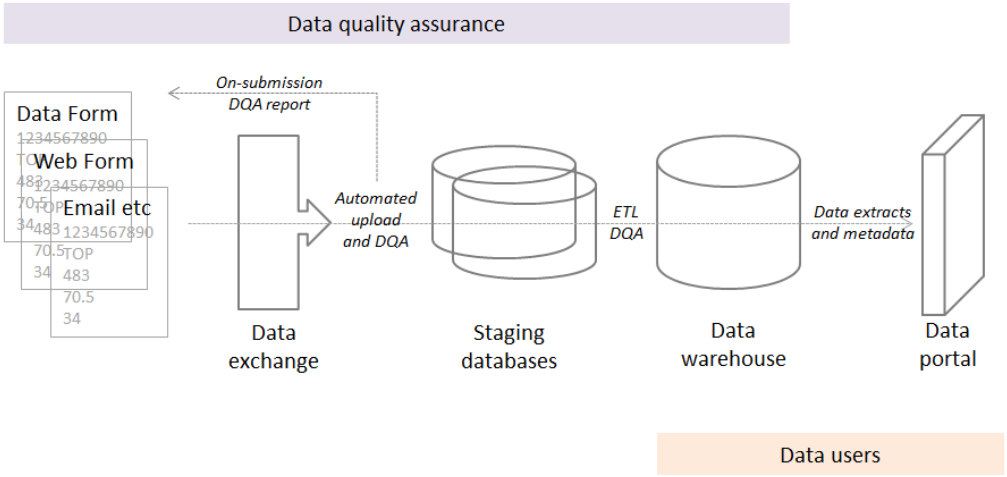



Figure 1: Concept for the redevelopment of the CCAMLR databases.
(DQA: data quality assurance; ETL: extract, transform and load)

22. The DWH is the central element in the redevelopment of the CCAMLR databases, and will:

- simplify the database architecture
- provide access to quality-assured and integrated CCAMLR data
- support an audit trail over the data life cycle
- be supported by comprehensive documentation and metadata.

23. The DWH architecture uses generic entities such as 'vessel catch' which holds catch data from all types of fishing (e.g. longline and trawl) in a single data table (previously, separate tables were used for holding data from longlining, trawling etc.). This approach simplifies the design of the DWH as well as the SQL scripts required to extract the data. In addition, the DWH architecture is robust and utilises parameterised attributes which allow new data requirements (e.g. new type of length measurement) to be added to the system without having to alter the structure of the EDM and DWH.

24. The DWH is populated (and updated) using the ETL process which extracts data from various source systems (staging databases), performs any required data transformations and loads quality-assured data into the DWH (Figure 1). The ETL is performed by a data management tool which runs at scheduled intervals to refresh the DWH with the latest available data. The process is automated, although some supervision is required to monitor the process and act on any DQA issues encountered during the ETL process.

25. CCAMLR data users will access and use the quality-controlled data in the DWH. This will include data extracts which will be updated at regular intervals commensurate with, for example, the annual work plans of SC-CAMLR and its working groups. It is proposed that the data extracts will be developed in consultation with expert users and the DMG, and each extract will provide the necessary data for a specific purpose such as the estimation of local biomass in data-poor fisheries (see WG-FSA-16/27 which solicits FSA feedback on the format and documentation accompanying the extracts available to users to support this work). Updated extracts will include metadata and a summary of what data have been added since the last update, and what data have been amended. Scheduled updates will allow users to base their analyses on the same version of CCAMLR data, and extracts will be disseminated to authorised users from an audited data portal (Figure 1).

## Redevelopment Roadmap

26. The redevelopment of the CCAMLR databases started in 2013 and was allocated a greater priority and resources in 2015 with part funding from the 'Korean special fund'. The work has focused on the development of structural and data elements of the EDM, DWH, ETL, DQA, reference data, data extracts and metadata, and establishing a data work flow which includes preliminary appraisal of candidate systems for automated uploading of, and reporting on, data submissions. These elements are the building blocks of the new system, and are being tested and implemented in selected subject areas.

27. Building on the considerable foundational work that has occurred in the past three years, the work plan for the end of 2016 through to 2018 is focused on completing the structural and data elements required to support the operation of the DWH and user-access to quality-controlled data extracts supported by metadata and documentation. The target DWH product of this 2-year work plan is the development of data extracts as guided by the DMG (or an alternative consultative arrangement).

28. The work plan will be advanced by a series of projects which aim to progressively develop DQA, the DWH and routine production of data extracts for use in the work of the SC-CAMLR and Commission. The work plan[3] for each project is outlined in Table 2, and is summarised as follows:

1. Data Management Group (DMG) – subject to advice from SC-CAMLR, it is proposed to establish a DMG (or alternative consultative arrangement) to provide high-level input on the management and development of the CCAMLR databases including data extracts and data products (WG-SAM-16 paragraph 2.20 and WG-EMM-16 paragraph 6.21). This management project will facilitate the DMG work.
2. Platform for Managing DQA Rules – to establish a comprehensive centralised repository of DQA rules complete with definitions and documentation. Existing rules will be held in a repository on the Secretariat's Intranet[4], and new rules will be defined in consultation with the DMG and users.
3. Automated data load – to develop a system for the automated loading of submitted data and generation of feedback reports to data submitters. The work will focus on the automated loading of catch and effort reports (used for in-season monitoring), observer logbooks and tagging data, and vessel haul-by-haul data (C1 and C2 data). The project will develop a data loader for submitted data, implement DQA rules and develop and implement feedback reporting to data submitters.
4. Developing data extracts – to establish the mechanism for providing data extracts that will focus on the provision of DQA rules, metadata suitable for toothfish stock assessments. The work will review current DQA rules and implement additional rules, improve the DQA of CCAMLR data held in current databases. Prototype data extracts from all CCAMLR databases will be developed and tested by the Secretariat as well as external users. Feedback and advice from the DMG and data users will guide further work which will implement DWH data extracts suitable for toothfish stock assessments. Metadata and documentation, including training material, will also be developed (WG-FSA-15, paragraph 3.20).

[3] ~~A proposed 2016-2018 work plan outline is available from the Secretariat on request~~ The work plan has been added to rev.1 at the request of WG-FSA-16 (SC-CAMLR-XXXV/04, paragraph 7.7)

[4] To be available on-line to external users in various formats

5. DWH expansion – use the agreed DQA rules to review current content and determine additional datasets for DWH.
6. Algorithm linking vessels' fishing activities with observed activities – to implement a link between vessel-reported activities and observed activities in the DWH. The existing mechanism will be revised and developed as an ETL component.
7. Probabilistic algorithm linking tag recaptured fish with releases – to implement a new method for linking tagging data from recaptured and released fish in the DWH. This linking algorithm will replace the existing 'dynamic' linking algorithm.
8. Reference data – to enhance aspects of reference data for the DWH (for example, research block definitions and species taxonomy).
9. Data registry redevelopment – to redevelop the existing registry and establish an automated centralised registry for submitted data. The work flow associated with maintaining the registry will also be revised.
10. Data audit mechanism – to develop a mechanism and process for auditing data amendments (WG-SAM-15, paragraph 2.50). This will include implementing audit fields in all database tables, as well as developing a mechanism for logging amendments (e.g. changed values) in CCAMLR data.
11. Managing quarantined data – to develop a mechanism and process for managing quarantined data in the DWH. Quarantined data are CCAMLR data which have been deemed not suitable for analysis (SC-CAMLR-XXXII, paragraph 3.228; CCAMLR-XXXIII, paragraph 5.66).
12. Statistical Bulletin redevelopment – to develop a platform-independent application for the *Statistical Bulletin* to facilitate broad access to CCAMLR's fishery and trade statistics.
13. Data forms for observer, vessel and tagging data - the requirements for data forms will be reviewed, together with candidate platforms to support smart forms and/or e-logbooks.
14. Data portal – a scoping project to establish a web-based data portal for accessing DWH data extracts. The requirements for such a portal will be reviewed and analysed, in consultation with users, and a work plan will be developed.

29. This work plan addresses the advice of SC-CAMLR and its Working Groups on data traceability (WG-SAM-15, paragraph 2.50), DWH testing and system evaluation prior to rollout (WG-SAM-15, paragraph 2.51), user training materials and workshops (WG-FSA-15, paragraph 3.20), data extracts and metadata (WG-SAM-16, paragraph 2.17), and establishing a data management group ((WG-SAM-16 paragraph 2.20 and WG-EMM-16 paragraph 6.21). The work will result in users experiencing increased:

- Integration across CCAMLR data
- user-focussed documentation supporting CCAMLR data systems and CCAMLR data
- availability of metadata libraries and data dictionaries
- engagement, transparency and functionality relating to data submission
- assurances regarding CCAMLR data quality
- dissemination of CCAMLR data
- support for data analytics and interrogation.

Table 2: Work plan for the redevelopment of the CCAMLR data management systems from end of 2016 to 2018. (x scheduled time; > on-going from that time)

| | Project | Component | Schedule | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 2016 | | | 2017 | | | | | | | | | 2018 | | | | | | | | | | | | | |
| | | | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| 1 | Data Management Group (DMG) | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Establish and facilitate the work of the proposed Data Management Group* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Establish the group | | x | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Convene meetings and develop and implement work program | | | | > | | | | | | | | | | | | | | | | | | | | | | |
| 2 | Platform for Managing DQA Rules | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Establish a comprehensive centralised repository of DQA rules complete with definitions and documentation* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Design and establish a centralised repository for DQA rules | x | x | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Define rules and populate repository | | x | x | x | x | x | > | | | | | | | | | | | | | | | | | | | |
| | | Workshop on DQA rules (Secretariat staff) | | | | x | | | | | | | | | | | | x | | | | | | | | | | |
| | | Consultation with DMG and external users | | | | > | | | | | | | | | | | | | | | | | | | | | | |
| | | Monitor implementation of rules | | | | > | | | | | | | | | | | | | | | | | | | | | | |
| 3 | Automated data load | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Develop a system for the automated loading of submitted data and generation of feedback reports to data submitters* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | -Catch and effort reports (in-season monitoring from 1-Dec-16) | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop data loader for submitted data | x | x | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Implement DQA rules for submitted data | | x | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop and implement feedback reporting | | | x | x | | | | | | | | | | | | | | | | | | | | | | |
| | -Observer and tagging data | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop data loader for submitted data | | | | | x | x | x | | | | | | | | | | | | | | | | | | | |
| | | Implement DQA rules for submitted data | | | | | | x | x | x | | | | | | | | | | | | | | | | | | |
| | | Develop and implement feedback reporting | | | | | | | | x | x | x | | | | | | | | | | | | | | | | |
| | -Vessel data (C1 and C2) | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop data loader for submitted data | | | | | | | | | x | x | x | | | | | | | | | | | | | | | |
| | | Implement DQA rules for submitted data | | | | | | | | | | x | x | x | | | | | | | | | | | | | | |
| | | Develop and implement feedback reporting | | | | | | | | | | | | x | | | | | | | | | | | | | | |
| 4 | Developing DWH data extracts | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Provision of data extracts from the DWH for toothfish stock assessments* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Review current DQA rules and implement additional rules | | | | x | > | | | | | | | | | | | | | | | | | | | | | |
| | | Populate DWH (including data amendments arising from DQA rules) | | x | > | | | | | | | | | | | | | | | | | | | | | | | |
| | | Test and validate DWH datasets | | | | | x | > | | | | | | | | | | | | | | | | | | | | |
| | | Develop and test data extract prototypes | | | | | | | x | x | | | | | | | | | | | | | | | | | | |
| | | Develop and document metadata | > | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop documentation and training material for transition and use | | | | | | | x | | | | | | | | | | | | | | | | | | | |
| | | Release data extract prototypes to DMG and users for evaluation | | | | | | | | x | | | | | | | | | | | | | | | | | | |
| | | Review feedback and advice from DMG and users | | | | | | | | | | | x | | | | | | | | | | | | | | | |
| | | Revise prototypes and implement DWH data extracts | | | | | | | | | | | | x | > | | | | | | | | | | | | | |
| | | Release DWH data extracts to authorised users | | | | | | | | | | | | | | | | | | | x | | | | x | | | |
| | | Develop status paper for working groups | | | | | | | | | x | | | x | | | | | | | | | x | | | x | | |

Table 2 (continued…)

| | Project | Component | 2016 | | | 2017 | | | | | | | | | 2018 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| 5 | DWH expansion | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Expand the content of quality-assured data in the DWH* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Review current content and determine additional datasets for DWH | | | | | | | x | > | | | | | | | | | | | | | | | | | | |
| | | Develop and implement data model for additional datasets | | | | | | | | | | x | > | | | | | | | | | | | | | | | |
| | | Define DQA rules for new datasets | | | | | | | | | | | x | > | | | | | | | | | | | | | | |
| | | Develop and implement ETL components for new datasets | | | | | | | | | | | | x | > | | | | | | | | | | | | | |
| | | Populate DWH (including data amendments arising from DQA rules) | | | | | | | | | | | | | x | > | | | | | | | | | | | | |
| 6 | Algorithm linking vessels' fishing activities with observed activities | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Implement a link between vessel reported activities and observed activites in the DWH* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Review, develop, test and implement algorithm | | | | x | x | | | | | | | | | | | | | | | | | | | | | |
| 7 | Probabilistic algorithm linking tag recaptured fish with releases | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Implement a new method for linking tagging data from recaptured and released fish in the DWH* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Further analyse user requirements | | | | | x | x | | | | | | | | | | | | | | | | | | | | |
| | | Develop data model | | | | | | x | | | | | | | | | | | | | | | | | | | | |
| | | Develop and test prototype algorithm | | | | | | x | x | | | | | | | | | | | | | | | | | | | |
| | | Review and implement algorithm | | | | | | | x | x | | | | | | | | | | | | | | | | | | |
| | | Develop paper for working group (WG-SAM) | | | | | | | | | x | | | | | | | | | | | | | | | | | |
| 8 | Reference data | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Enhance aspects of reference data for the DWH* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Implement research blocks in geographic areas | | | | x | | | | | | | | | | | | | | | | | | | | | | |
| | | Implement research blocks in DWH | | | | | x | | | | | | | | | | | | | | | | | | | | | |
| | | Analyse needs for taxonomic groups relevant to CCAMLR business | | | | | | | | | | | | | | | | | | x | | | | | | | | |
| | | Implement taxonomic groups in DWH | | | | | | | | | | | | | | | | | | x | | | | | | | | |
| 9 | Data registry redevelopment | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Implement an automated centralised registry for submitted data* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Review requirements for the redevelopment | | | | | | | | x | x | | | | | | | | | | | | | | | | | |
| | | Develop data model and implement | | | | | | | | | | | x | | | | | | | | | | | | | | | |
| | | Develop and implement work flow | | | | | | | | | | | x | x | | | | | | | | | | | | | | |
| 10 | Data audit mechanism | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Develop a mechanism and process for auditing data amendments* | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Implement automated audit fields in databases | | | | | x | | | | | | | | | | | | | | | | | | | | | |
| | | Analyse requirements for detailed audit mechanism | | | | | | | | | | | | | | | | x | | | | | | | | | | |
| | | Develop data model/mechanism for detailed audit | | | | | | | | | | | | | | | | | x | x | | | | | | | | |
| | | Implement detailed audit log | | | | | | | | | | | | | | | | | | | x | | | | | | | |
| | | Develop and implement work flow | | | | | | | | | | | | | | | | | | | x | | | | | | | |
| | | Populate audit tables (historic and current) | | | | | | | | | | | | | | | | | | | | x | > | | | | | |

Table 2 (continued…)

| | Project | Component | 2016 | | | 2017 | | | | | | | | | | | | 2018 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
| 11 | Managing quarantined data | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Develop a mechanism and process for managing quarantined data* | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Exclude quarantined data from the DWH (interim step) | x | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Analyse requirements for using and managing quarantined data | | | | | | | | | | x | | | | | | | | | | | | | | | | | |
| | | Develop data model/mechansim for quarantined data | | | | | | | | | | | x | | | | | | | | | | | | | | | | |
| | | Develop paper for Commission | | | | | | | | | | | | x | | | | | | | | | | | | | | | |
| | | Revise and implement data model/mechanism in the DWH | | | | | | | | | | | | | | | x | | | | | | | | | | | | |
| 12 | Statistical Bulletin redevelopment | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Develop a web-based version of the Statistical Bulletin* | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Develop data model and implement | | | | | | | | | | | | | | | x | | | | | | | | | | | | |
| | | Develop and test data extract | | | | | | | | | | | | | | | | x | | | | | | | | | | | |
| | | Develop metadata | | | | | | | | | | | | | | | | x | | | | | | | | | | | |
| | | Develop web product | | | | | | | | | | | | | | | | | x | x | | | | | | | | | |
| | | Publish product annually | | | | | | | | | | | | | | | | | | | x | | | | | | | | |
| 13 | Data forms for observer, vessel and tagging data | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Evaluate the use of smart forms and e-logbooks for observer, vessel and tagging data* | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Review requirements for data forms | | | | | | | | | | | | | | | | | | | | x | x | | | | | | |
| | | Identify options for a reporting format/platform | | | | | | | | | | | | | | | | | | | | | | x | x | | | | |
| | | Develop paper for WG-FSA and SC-CAMLR | | | | | | | | | | | | | | | | | | | | | | | | x | | | |
| | | Develop work plan | | | | | | | | | | | | | | | | | | | | | | | | | | | x |
| 14 | Data portal | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | *Establish a web-based data portal for accessing DWH data extracts* | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | Analyse requirements for a data portal | | | | | | | | | | | | | | | | | | | | | | | | | | x | |
| | | Develop work plan | | | | | | | | | | | | | | | | | | | | | | | | | | | x |

# Appendix A: Current CCAMLR Data-Related Documents

CCAMLR Data management strategy

CCAMLR Data Modelling Guidelines

CCAMLR Data Quality Rules

CCAMLR Data Warehouse ETL System

CCAMLR Data Warehouse introduction

CCAMLR Database Redevelopment (WG-SAM-15-33 and WG-FSA-15-03)

CCAMLR Enterprise Data Model (including diagrams and dictionary)

CCAMLR Scheme of International Scientific Observation workflow Documentation

CCAMLR Secretariat Information Management Framework

CCAMLR Secretariat Strategic Plan 2015–2018

DAMA Dictionary of Data Management (2nd Edition)

DAMA Guide to the Data Management Body of Knowledge (1st Edition, 2010)

Report of the Independent Review of CCAMLR's Data Management Systems (CCAMLR-XXX/05)

# Appendix B: Data Terminology

**Enterprise Data Model (EDM)**

The CCAMLR EDM provides the architectural foundation for the overall data and information management at the Secretariat. The EDM incorporates comprehensive definitions of logical data elements such as entities, attributes, relationships, domains and business rules, and provides an organisation-wide perspective and integration of CCAMLR data. The core activities in developing the EDM include:

- Implementation of a data naming standard and modelling practices supported by an appropriate modelling tool
- identification and documentation of subject areas which facilitate modular development, and where each subject area broadly corresponds to a high-level CCAMLR business function
- development of logical data models for each subject area, with the process combining both top-down and bottom-up approaches which integrate best practice and existing data requirements
- development of a data dictionary.

**Data Warehouse (DWH)**

The DWH is a centralised database designed to service the business needs of CCAMLR. The DWH adheres to the EDM to ensure consistency of decision support activities across the enterprise. The DWH serves the user community with high-level quality-assured data and related metadata, integrates data from various sources and uses consistent naming standards and data formats.

The CCAMLR DWH will:

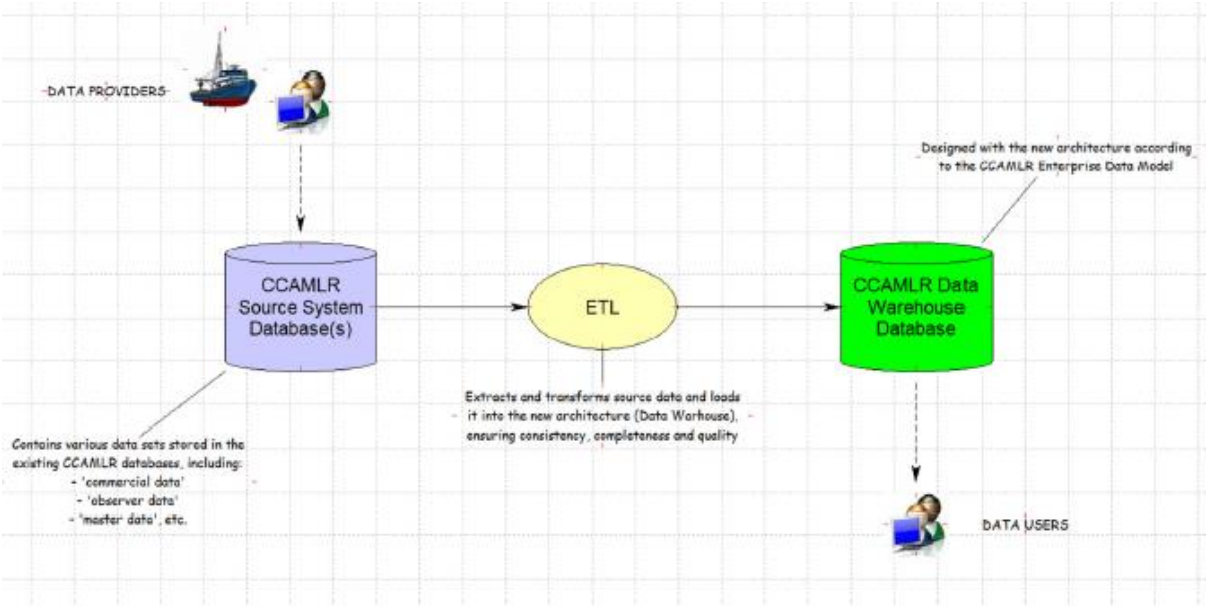
- Improve the integration of CCAMLR data
- Simplify the database architecture
- Improve DQA
- Provide data documentation and metadata

**Extract, Transform and Load (ETL)**

The DWH is populated (and updated) using a ETL process which extracts data from various source systems (staging databases), performs any required data transformations and loads quality-assured data into the DWH. Data transformations include any or all of the following:

- checking DQA
- checking the completeness of data and filling gaps if required and where possible
- harmonising potentially different ‘coding systems’ into a common framework
- mapping similar data types into a single consistent set of definitions.

The ETL is performed by a data management tool which runs at scheduled intervals to refresh the DWH with the latest available data. The process is automated, although some supervision is required to monitor the process and act on any DQA issues encountered during the ETL process.




**Data Quality Assurance (DQA)**

DQA is the process of profiling the data to discover inconsistencies and other anomalies in the data, and perform data activities such as removing outliers, missing data interpolation to improve the data quality. Three broad types of DQA are being developed for CCAMLR data:

- Atomic DQA which checks that the data are consistent with the formats and referential integrities defined in the EDM
- Logical DQA which checks that values in a dataset are internally consistent (e.g. chronology of fishing events, distance travelled, vessel dimensions).
- Analytical DQA which checks that reported values fall within expected ranges (e.g. length frequencies, catch distributions).

**Metadata**

Metadata (data about data) are the data context that explains the definition, control, usage and treatment of data content within a system, application or environment throughout the enterprise. Metadata are one of the key knowledge areas of the Secretariat's data management systems, and provide the context to CCAMLR data and information to make informed decisions, including:

- Where the data came from
- How the data are processed and transformed
- how to interpret the data
- how DQA was implemented

Four broad types of metadata are relevant to CCAMLR data and other enterprise data held by the Secretariat:

- Technical metadata define the objects and processes in databases and the DWH from a technical/structural perspective
- Process metadata describe the results of operations in the DWH, including the ETL process, description of DQA rules and changes and additions to data which are held in the DWH
- Business metadata describe the data and analyses used for specific business functions (e.g. toothfish stock assessment), including data sources, data meaning and what their relationships are to other data in the DWH, and how these data are analysed.
- Discovery metadata describe specific datasets which may be of interest and use to the scientific community, policy makers and the general public. These metadata are usually submitted to web-based global registers which promote the discovery and dissemination of information (e.g. GCMD http://gcmd.nasa.gov/).