SC/69A/GDR/01

Sub-committees/working group name: GDR

Data management strategy

IWC Statistics Department - Isidora Katara, Lydia O'Loughlin, Elsie Whittle

Papers submitted to the IWC are produced to advance discussions within that meeting; they may be preliminary or exploratory.

It is important that if you wish to cite this paper outside the context of an IWC meeting, you notify the author at least six weeks before it is cited to ensure that it has not been superseded or found to contain errors.

Data Management Strategy

IWC Statistics Department

Isidora Katara, Lydia Oloughlin, Elsie Whittle

The IWC Statistics and Modelling department is seeking to undertake data-driven digital transformation. To achieve this goal a data infrastructure and management strategy are critical. The strategy is based on the principles of Open Science and aims to streamline data management, optimise data dissemination, and advance the communication of the valuable data held in IWC.

Open Science

The data management strategy will follow the principles of open science, the movement towards accessible scientific research to the scientific community and the public. The 8 principles of open science[1] are FAIR Data, Research Integrity, Next Generation Metrics (beyond citation counts and journal impact), Communication, Citizen Science, Education and Skills, Rewards and Initiatives, and Collaboration. FAIR Data are:

> Findable – the wider academic community and the public can easily discover research outputs (including data and code),
> Accessible – the data have unique identifiers, metadata in clear language and access protocols,
> Interoperable – metadata and access to data follow known standards,
> Reusable – optimise the reuse of data to maximise their research potential.

We are revisiting our datasets to make them FAIR and are developing a data hub – a database and website of metadata – to expand accessibility, increase discoverability, and promote collaboration. Related dashboards and a communications plan encourage the participation of the public in scientific research and boost our educational role.

Main datasets

The Statistics and Modelling Department holds a wealth of historical and current data on large whales and small cetaceans that feed into stock assessments, management plans and communication of the status of whale populations with a wider audience. The information managed by the department includes:

- Aggregated historical catch data for large whales,
- Aggregated historical catch data for small cetaceans,
- Catch data for individual large whales,
- Abundance estimates discussed in ASI meetings,
- SOWER and POWER survey data,
- Mark-recapture data,
- Infractions,
- Whale welfare data.

[1] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

Other datasets held related to oil production, ageing data and programs (scripts, applications, and data) used for management and stock assessment purposes.

The data held by the Statistics and Modelling Department are complementary to data received through National Progress Reports and the ship strikes database. To optimize data integration and increase data value coordination within the Secretariat will be prioritized to ensure standardisation among current and future databases.

Data management components

The data management strategy is based on 3 components that mirror our objectives to

1. Coherently organise all data,
2. Facilitate data access and use,
3. Increase the visibility of the data.

The first component, the data hub, is the core of the data management strategy as it stores all the data in databases or standalone datasets and is simultaneously a working environment where data input processes are clearly and precisely defined. The second component is the dissemination of the data, where tools and systems are developed to allow access to the data or data outputs. The third component is the communication of the data consisting of an outreach strategy to raise awareness of the data.

Data Hub

A data hub is essentially a metadata database and related databases and datasets. The data hub consists of (1) a central data storage system, a repository, that allows storing relational databases and standalone datasets in PostgreSQL and CSV or Excel files respectively, and (2) a metadata database.

The databases currently under development are:

1. Abundances (the IWC abundances tables)
2. Survey data (SOWER and POWER data combined into one database)
3. Direct catch data (large whales)
4. Direct catch data (small cetaceans)
5. Indirect removals (bycatch and ship strike data derived from progress reports and other sources)

The datasets currently held relate to

1. Mark-recapture data
2. Infractions
3. Animal welfare (information on welfare during whaling and euthanasia operations)
4. Genetics
5. Age
6. Oil production data (to be incorporated into the catch data)

Each database and dataset have comprehensive documentation and metadata; the latter will be held in the metadata database to ensure that the data are FAIR: Findable, Accessible, Interoperable and Re-usable. The metadata are a general description of the data with a unique identifier, recording aspects of the data collection, updating, analysis, quality, and data use. Our metadata follows the

MEDIN[2] standard to ensure the use of standardised names and controlled vocabularies and will be held in a database that will be searchable to promote accessible data. Full documentation of every database and dataset, i.e., a description of the tables, their attributes and relationships, will also be available.

The data management system is a workspace where data are updated (added or corrected) based on set processes of data input and quality assurance. Standalone applications will be developed to establish those processes and facilitate their application by users regardless of technical or data analytic knowledge. The applications will be written in R (R-shiny) and Python and will be shared through GitHub for others to download and run. Many of the applications will be based on existing DOS and Fortran scripts that will be 'translated' into R and Python and 'packaged' into the format of an application. Other applications will use AI and machine learning algorithms to expedite data digitisation where possible.

An additional component of the data hub is GitHub repositories relating to in-depth and comprehensive assessments. Using GitHub will ensure version control, increase code safety and promote collaboration. The programs, applications, scripts, input data and outputs of the assessments will be organised in these repositories.

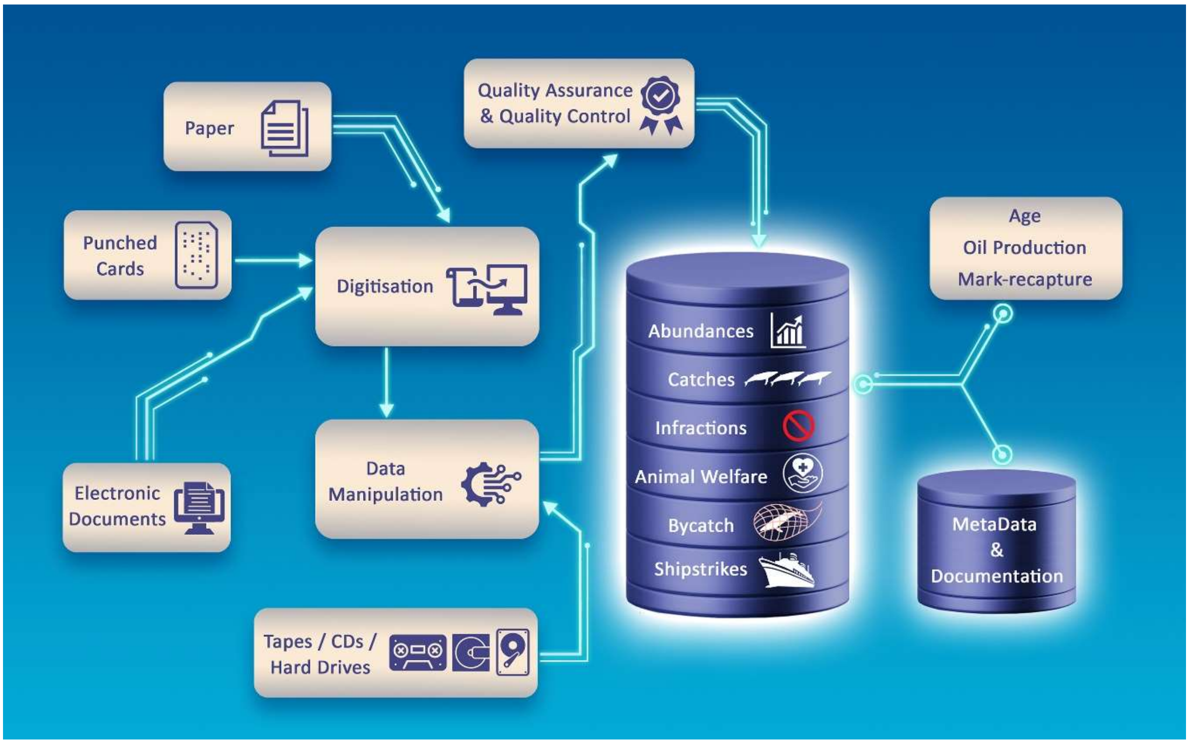

Figure 1 Schematic diagram of the data management system including processes of digitisation, quality assurance and control of the data, and the data hub (data repository and metadata).

Data Dissemination

The data dissemination plan is structured around a central set of web pages which will be part of the general IWC website. The landing webpage will provide an overview of available data and allow searching through the metadata. A further webpage will be dedicated to selected databases/datasets that will provide full documentation of the data, an application that will allow access to some views of the data including the functionality of producing maps and graphs based on

[2] https://medin.org.uk/

most common requests. Additionally, an option for direct read-only access to the database for scientists under specific conditions and through a formal request). The current option of using sharable folders of specific data views will remain available although the data will only be updated annually.

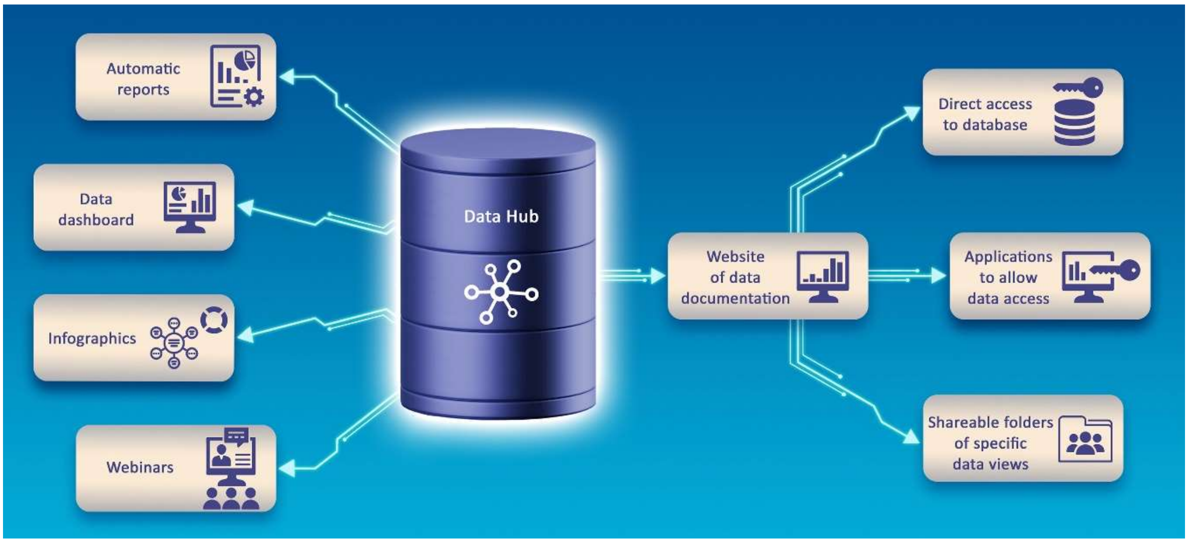
Automatic reports
Data dashboard
Infographics
Webinars
Data Hub
Website of data documentation
Direct access to database
Applications to allow data access
Shareable folders of specific data views

Figure 2 Schematic representation of the data dissemination and communication plans.

Data Communication

The data communication plan aims to update existing pages on the IWC website. Through direct links to the databases, the renewed webpages will work as data dashboards, i.e., customisable information management tools that visualise, analyse, and display key data points and indicators to monitor specific processes (e.g., trends of the abundance of a stock or the distribution of reported bycatch). Automatic reports will also be available as R markdown documents, customisable parameterised reports, supporting R, Python and SQL.

Infographics will be developed and shared at events where IWC is invited to communicate high-level information derived from the data. As the development of each database is completed, a launch will be organised to raise awareness of the data and encourage data access and use. The launch will be both internally through presentations to the Scientific Committee and externally, including giving webinars on platforms such as OCTO (Open Communications for The Ocean), the EBM Tools Network and USAID BiodiversityLinks.

Infrastructure

Data infrastructure includes various components, such as hardware, software, networking, services, and policies that capacitate data consumption, storage and sharing. Cloud databases are commonly used to store data, as an affordable data storage option with adjustable storage capacity. The applications can be hosted on a server as web apps with the advantage of easy access, publication and updates and the drawback of cost. Standalone applications can also be developed. Existing software such as Power BI has been suggested for data visualisations.

IWC is moving to cloud-based systems that meet user needs by adding flexibility to scale up resources and storage without maintaining physical technology; the data hub will be hosted on a cloud server (Microsoft Azure or Amazon Web Services). The databases are developed using PostgreSQL, a relational, open-source database used in data science, graphing, and AI industries. It supports spatial data analysis and is ideal for Python and R applications. To share data and information in an easily accessible and collaborative way, applications can be developed and shared through Shiny tools (e.g., shinyapps.io) and GitHub (e.g., gist.github.com) instead of static documents.

A comprehensive data-sharing policy will be developed to provide a framework for data access and use. This will include recognition requirements such as authorships, acknowledgements, or standard citations. Although transparency is pivotal, sensitive data need to be protected through the data policy. The policy will delineate rules regulating data sharing with different interested parties - the public, the scientific community, IGOs, NGOs and companies - and for different purposes – scientific, educational, for profit.