SC/69A/GDR/02

Sub-committees/working group name: GDR

Data management - progress in database development

Isidora Katara, Lydia O'Loughlin, Marion Hughes, Sue Burkett, Elsie Whittle

Data Management: Progress in Database development

IWC Statistics Department

Isidora Katara, Lydia Oloughlin, Marion Hughes, Sue Burkett, Elsie Whittle

As part of the IWC statistics data management strategy, the development of relational databases has commenced aiming to form an integral part of the foreseen data hub. In line with open science principles, the databases are developed using PostgreSQL and related interactive dashboards will present basic outputs from each database on the IWC website.

Part of the database and dashboard development is undertaken by two groups of MSc students from the MSc Data Science Programme[1] of the University of Manchester, a collaboration initiated through the Office for National Statistics by the IWC projects coordinator.

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,
- Catch data for individual large whales,
- Abundance estimates discussed in ASI meetings,
- SOWER and POWER survey data

We have also started organising new datasets for

- Infractions and
- Whale welfare data

These datasets have facilitated reporting to the relevant IWC sub-commissions and identifying patterns to inform policy decisions. As more data become available, in-depth analyses will be possible.

Databases

The development of the databases involves building a relational data model, writing the SQL code that will create the database, the tables, and the links between them, writing full documentation of the tables, and putting together the metadata that will be available through the data hub (ref to management paper). Throughout the development of each database, a GitHub repository is used to store and track the database development.

Currently, the following databases are under development:

[1] https://www.manchester.ac.uk/study/masters/courses/list/11552/msc-data-science-computer-science-data-informatics/course-details/

- The Catch database combines data currently stored as aggregated historical catch data for large whales and data for individual large whales harvested. The data model, code, metadata and documentation can be found at https://github.com/intwhcom/IWC-catches-database-development.
- The Surveys database combines the SOWER and POWER data, currently held in a Paradox database and CSV files respectively. The data model, code and documentation can be found at https://github.com/intwhcom/IWC-surveys-database-development.
- The Abundances database holds the abundance estimates – and related information – that have been endorsed by the ASI, a subset of which is presented on the IWC website (https://iwc.int/about-whales/estimate). The data model (fig 1), code, metadata, and documentation can be found at https://github.com/intwhcom/IWC-abundances-database-development.

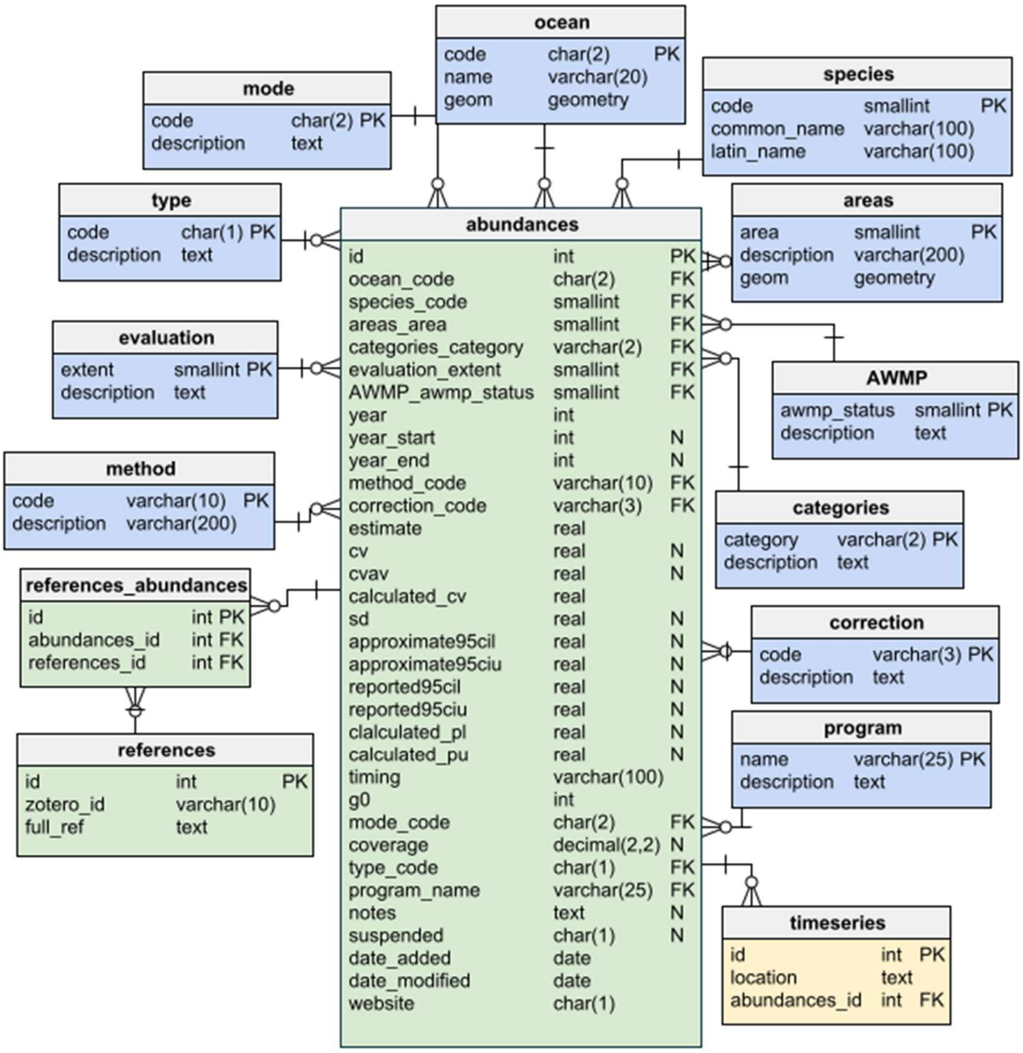

Figure 1 Draft schema of the abundances database. The data model can also be found at https://github.com/intwhcom/IWC-abundances-database-development.

Currently, the IWC GitHub repositories are not publicly available. The repositories above will be made public for the duration of the SC meeting to facilitate communication. To get access to the repositories after the SC meeting please send your GitHub name to Statistics@iwc.int along with which repositories you want access to.

Fig. 2 shows the progress made intersessionally towards the development of the databases. During the **planning** phase the scope of the database and the requirements for the database development, in terms of resources and infrastructure, are determined. The **analysis** refers to examining the development process including data processing, understanding data items and the relationships between them. Based on the analysis the data model is **designed**. The **implementation** phase includes the construction of the actual database on a server and any visualisations of the data. Finally, **maintenance** is an ongoing, regular task, including backups, corrections, and managing access.

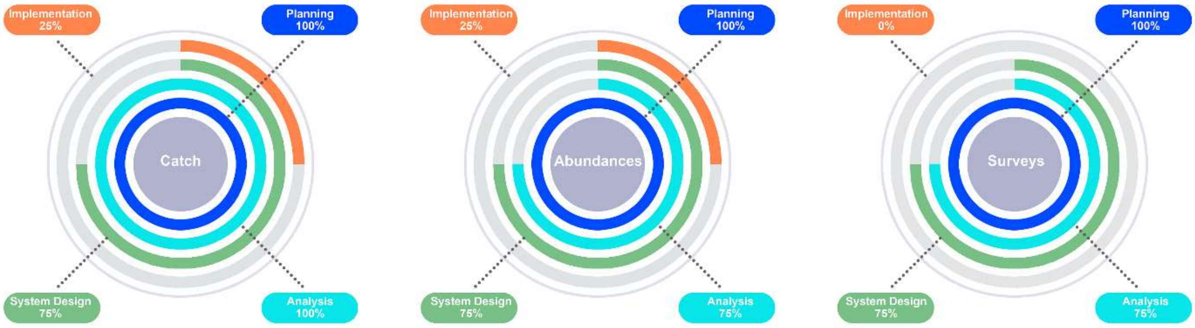

Figure 2 shows the progress made towards the development of each database, the Catch database, the Abundances database, and the Surveys database. Planning, Analysis and System Design are close to completion.