SC/69A/E/13

Sub-committees/working group name: E

Strandings Database Scope and Proposal

Iwc Secretariat

IWC Global Strandings Database Scope and Proposal

IWC Secretariat

Overview

The purpose of this paper is to provide a report of scoping work conducted with the goal of developing a global strandings database and to propose a data model that will support the development of such a database. A data model is an abstract model that organises elements of a database, data infrastructure and data flows and defines how each element relates to one another. To develop and implement a global strandings database, a data management framework is needed to ensure the quality and integrity of the data. This paper includes steps to be taken to integrate a data management framework within the proposed data model. Of course, the work to reach the goal of a global strandings database is sizeable and will need to be accomplished in phases based on priority and ease of the tasks to implement the data model. This paper provides a data model outline and outlines tasks by phase to reach the end goal of a global strandings database.

Purpose of a Global Database and IWC Background

The need to collate, standardise, store and disseminate strandings data at a global scale has been a topic of discussion by the Scientific Committee, the Conservation Committee and others outside the IWC. A global strandings database would store the largest spatial extent of strandings data currently available and would be usable by scientists, conservationists, governments and the public. By being easily accessible, a database will increase the value of the data and provide a global overview of strandings.

Through discussion, recommendations and endorsement, the value of collating and disseminating strandings data at a global scale has been made evident. At IWC68, the Commission endorsed several recommendations related to database development and management of strandings data. The 2020 recommendation (SC20130) by the Scientific Committee for the IWC Strandings Initiative and Secretariat to liaise intersessionally to discuss database development plans. Additionally, the workplan action (SC20131) to further consider the topic of database development and data standardisation as well as coordinate with on-going work on database development in the Ad Hoc Working Group on Databases was endorsed by the Commission at IWC68. Lastly, the 2021 recommendation (SC2185) by the Scientific Committee to evaluate how multiple existing and potential IWC databases (ship strikes, strandings, marine debris) could be successfully integrated to enable better comparison and cross-referencing of multiple data sources.

In the IWC Strandings Initiative Four-Year Work Plan 2021-2024, work area 3 provides clear activities for work towards increasing IWC data management, curation and visualisation on strandings data. These activities include identifying sources of existing data on cetacean strandings, evaluating methods to increase data reporting to IWC, evaluating data validation and curation procedures, and developing data analysis and visualisation methods.

Data Management Framework

To achieve the goal of a global strandings database, a data management framework is critical. To maintain the quality and integrity of the data, the framework applies data management strategies to the life cycle of the data. As seen below, this starts at the creation of the data and ends with sharing of the data.

Data Collecting

Acquiring data for a global strandings database involves collecting new data or obtaining existing data. Through fulfilling the Stranding Initiative Work Plan work area 5, capacity building of strandings response networks, data are collected by trained responders via established methods with controlled vocabularies. Additionally, the Strandings Initiative can ensure data collection procedures are standardised and useable by a wide-range of data collectors. These steps result in data that are standarised and easily entered into the database. Obtaining existing data focuses on constant collaboration with data holders to ensure data are well documented and correctly standardised to integrate into the database. Providing data holders with a data template is one method to assist with standardising data; similar to what currently is in place for the National Progress Report database.

Data Processing

Processing data involves various activities associated with the preparation of new or previously collected data. This includes activities to validate, transform and integrate data. The quality of data are assured through checks and inspections. These steps can be manual or automated within the data model and should be expanded upon established validation methods implemented by data providers. Transforming data involves converting data from one format to another to meet a requirement of the database in use. Data transformation steps should be well documented and automated as much as possible to remove errors. When possible, data integration should be automated to reduce errors. This could be via an application to enter data directly to the database or using forms to manually standardise the data for data entry.

Data Storing

Preservation of data in the data storage steps involves procedures used to ensure long-term viability and accessibility of data. This means using database software that is open source for long-term use, maintaining the latest software version for accessibility and storing data in the cloud to protect from data loss. Additionally, establishing backup procedures will protect the data from accidental data loss, corruption and unauthorised access.

Data Describing

Data describing involves ensuring data are accurately and thoroughly described using the appropriate metadata standards. It is currently suggested to use the Marine Environmental Data and Information

Network (MEDIN) metadata standards since the Statistics and Modelling Department is using these standards. Additionally, throughout the data lifecycle, documentation must be created and updated to reflect actions taken upon the data. Developing a data handling policy will fulfill the Stranding Initiative Work Plan item 3.4 to ensure all data steps are documented and following the established policy. When disseminating data, having metadata makes the strandings data findable and creating documentation provides data users with the ability to understand the data.

Data Disseminating

Developing a data-sharing policy provides a foundation for data access and use. This will include citations, identifying data sources, data use and limitations, and rules regulating data sharing. Creating levels of data dissemination allows flexibility in data use requirements from data sources, such as only sharing sensitive data for research purposes. To increase access and use of the global strandings database, data dashboards and data summaries will be developed.

Existing IWC Strandings Data and Data Use

Strandings data held by IWC vary in format based on year and by member nation submissions. The Scientific Committee has been collecting strandings data through the annual submission of National Progress Reports since the mid-1970s. Currently, the data are collected through the IWC Portal website (http://portal.iwc.int) and the strandings data from 2013 to present are in the National Progress Report database. Years 1998 to 2012 are in digital format on the IWC Archive website (archive.iwc.int) and all years prior to 1998 are available through request to the Secretariat. Strandings data are integrated into long-term, ongoing work to understand the status of whale populations within the IWC.

It should be noted that data in the National Progress Report database is skewed towards nations with consistent reporting since 2013 and therefore data gaps do exist. Within the database, 24 member nations have submitted 5,404 strandings records from 2013 to 2022. The top reported species are common dolphin, common bottlenose dolphin, striped dolphin, unidentified dolphin and humpback whale. The North Atlantic Ocean and North Pacific Ocean have the highest number of records.

When collecting data through the IWC Portal website or through a csv template, member nations fill out a form to report a stranding. The fields collected from the form and their descriptions are as follows:

- Data Year - Year stranding event occurred.
- Large Area – Large area to select from a dropdown list.
- Species – Species to select from a dropdown list.
- Country – Country stranding occurred in.
- Local Area – Description of area stranding occurred in.
- Local Taxonomy – The name this species is most commonly known by in the region.
- Local Area (Long/Lat) – Coordinates in decimal degrees describing a point or a polygon.
- Females – Number of female individuals stranded, please use 0 if none.
- Males – Number of male individuals stranded, please use 0 if none.
- Unknown – Number of individuals stranded of unknown sex, please use 0 if none.
- Information that may help to explain data – Any additional information or details on the event.
- Contacts – Contact details: name, email, phone numbers, address.
- References – Data source reference.

Overview of Data Sources and Data Users

Data Sources

Through meetings with strandings networks, regional intergovernmental organisations and data holders, and through online searches for strandings data, similarities and differences were noted. Some areas with similarities and differences between groups include data collection, storage, the final format of strandings data, data access and sharing policies. With a wide range of data sources and how data are handled, the data model must be developed for flexibility and scalability.

Data sources can be categorised by the level of data aggregation that has been accomplished to reach a final, sharable dataset. An example of a highly aggregated data source would be a regional entity that has gathered multiple strandings network data and has the data accessible in a standardised form. A minimally aggregated data source has data in multiple formats, older data are on paper data sheets or their location is unknown and newer data are digitised in a file. A moderately aggregated data source is somewhere in between the two categories. It should be noted that these categories are only created to assess which actions are best when working to integrate data into the data model and it does not reflect the quality or integrity of the dataset.

Highly aggregated data sources have data dashboards and online data access developed for their groups. Integrating data from these data sources could be manually accomplished by sending data files on a regular schedule to be added to the database or a high-tech data pipeline that utilises an API to connect the data source to the database could be implemented. There is flexibility and scalability when working with highly aggregated data sources and so it will be on a case-by-case basis and mostly limited by funding.

Moderately aggregated data sources have all data digitised and have high-level data available online to download a file or through contacting the data holder for a data file. To integrate data from these data sources manually, the data will need to be entered by the data holder or the IWC Data Manager. An automated method could be developed to run a script on the data file to transform and append data into the database, but that solution will need to be specialised for each data source.

Minimally aggregated data sources have data in multiple formats and may not have historic data easily available. To account for the variability, multiple steps will need to be taken to integrate data into the database. For data that is digital, the method is similar to that of a moderately aggregated data source. For data that is in paper form, then time and support is required to digitise the data and integrate into the database.

Just as there is variation in data sources, the scoping process also identified a wide range of data sharing and access policies. One method is to share only high-level data of the stranding record and then attribute the record to a data source that interested researchers can contact for further detailed data. Another is that it is the countries that share the data rather than a strandings network not associated with a government entity. Similarly, this is how IWC collects strandings data through National Progress Reports submitted by member nations. These data are also high level with contact information for further detailed data. As work develops on designing and implementing a data model, ongoing conversations will be required to generate a data model that fits with the data sharing restrictions of each data source.

Data Users

To make the data within the IWC Global Strandings Database both useful and useable, the data model must be designed for the end users. Having data at a global scale has been valued by the Scientific Committee, the Conservation Committee and others outside the IWC. There will be a wide variety of users and purposes of the IWC Global Strandings Database. During scoping, the public, the scientific community, IGOs, NGOs and member nations were identified as currently using or interested in using strandings data; the purposes for data use ranged from scientific to educational to conservation to management. Some uses of the data within the database are expanded upon in Dr. Andrew Brownlow's Strategic Review for the IWC Strandings Initiative presented at SC68C. Data use ranged from guiding IWC Strandings Initiative work to identifying areas of increased incidence and areas of data gaps to informing risks to cetaceans at varying scales.

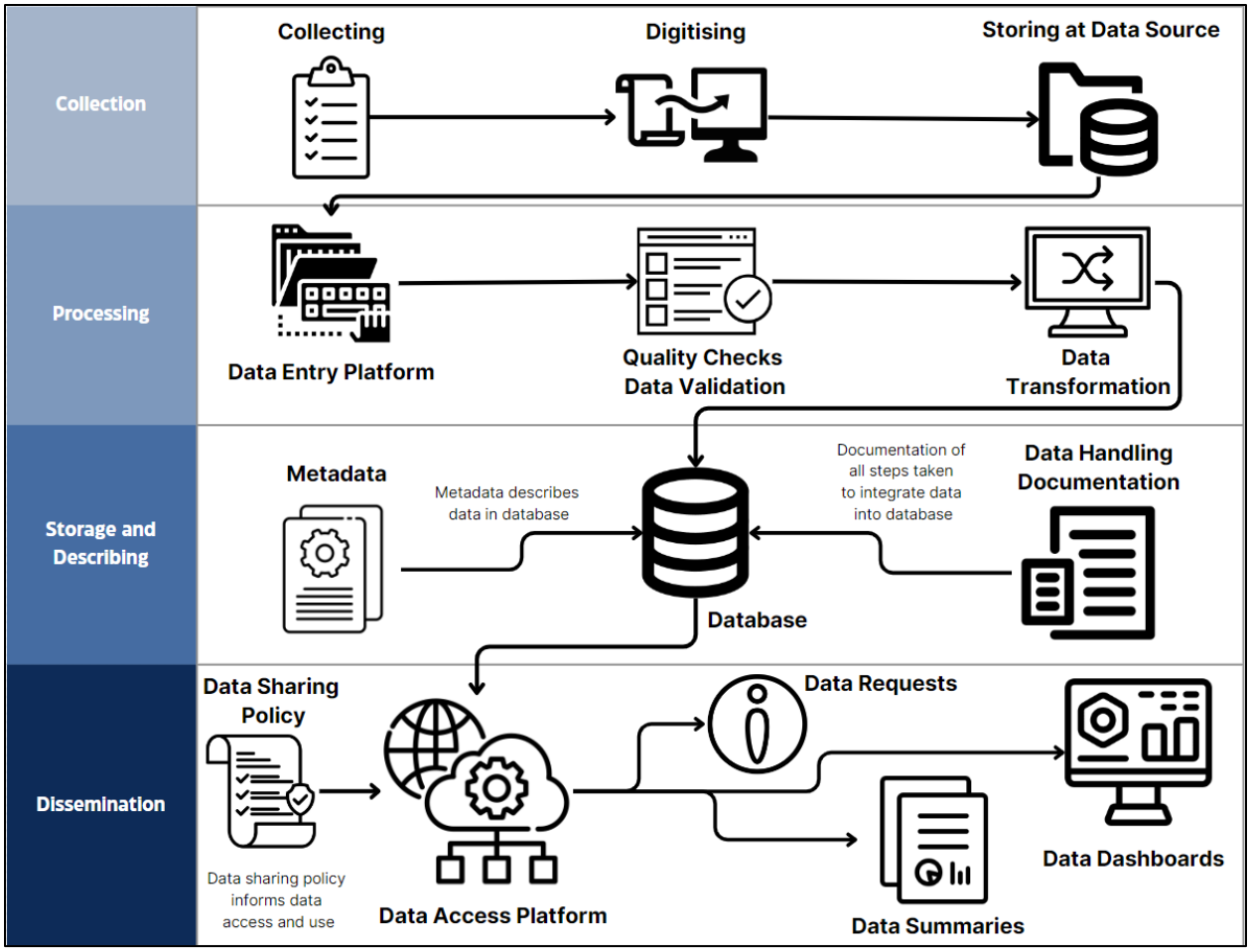
Proposed Data Model

A data model is an abstract model that organises elements of a database, data infrastructure and data flows. A diagram depicting the proposed data model is at the end of this section. The data flow is separated into four stages based on activities surrounding the data. The stages are Collection, Processing, Storage and Describing, and Dissemination.

In the Collection stage, data are collected by strandings responders, then digitised and stored in the data holder's database. A data platform is then used to integrate the data from the data source to the IWC Global Strandings Database. To make data entry easy to use, this platform could allow for uploading a data file or allow users to enter data through a form. In the Processing stage, several steps are taken to quality check and validate the integrated data as well as transform and standardise the data before being finalised in the database. These steps ensure the quality of the data to increase usability.

After processing, the data will be stored in a database using cloud-based systems that meet user needs by adding flexibility to scale up resources and storage without maintaining physical technology. To standardise data management at IWC and build off resources already implemented, the database management system will be either PostgreSQL or MySQL. Regular back-ups will be implemented to protect against data loss or corruption. To increase findability of data and make data easy to use, documentation and metadata will be developed to describe the data and the data handling steps taken.

The final stage is data dissemination where data are accessible and shared to meet user needs. A data access platform provides a foundation to access and share the data. This can be through direct requests for data, development of dashboards to dynamically explore the data and generation of data summaries to provide high level analysis of the data. To ensure various levels of data are shared with appropriate users, a data sharing policy will be utilised to clearly define data restrictions and limitations.



Proposed Methods

Phase 1

Timeline

- Months 0 to 18 from start of project.

Activities

- Digitise all IWC strandings data from National Progress Reports.
- Receive data from IGOs and other regional data holders to broaden data sources via email.
- Develop MySQL or PostgreSQL database with current dataset(s).
- Implement data management strategies to increase data quality.
- Develop website interface to interact with high level data.

Phase 2

Timeline

- Months 18 to 30 from start of project.

Activities

- Design data pipeline for higher tech users (ex: ACCOBAMS data via API)

- Develop data ingestion platform for countries and regions to submit data.
- Progress data access and sharing website based on user feedback and requests.

Phase 3

Timeline

- Months 30 to 36 from start of project.

Activities

- Implement data communication plan to increase awareness of database and data submissions.
- Continue developing features that increase data ingestion from harder to attain data sources.
- Continue progressing data access and sharing website based on user feedback and requests.

Next Steps

- Seek feedback from Strandings Expert Panel and the Scientific Committee and update proposal to incorporate feedback.
- Create a steering group to focus the work and ensure details are ready for endorsement.
- Continue scoping data sources and data users to incorporate feedback into proposal.
- Further develop timeline and phases to clarify scope and tasks.
- Explore funding sources to implement proposal.
- Seek endorsement from SC, CC and Commission to begin work.