DATA 70202
Applying Data Science

University of Manchester Mentor:
Nuno Pinto

Industry Partner:
International Whaling Commission (IWC)

Student IDs:
10670994, 11557264, 11562596,
11604233, 14118011, 14126610

Word count: 7,010 words

TABLE OF CONTENTS

Executive Summary

The International Whaling Commission is a global body supporting the conservation of whales and the regulation of whaling activity, utilising global marine incident datasets to help inform decision making. However, inconsistencies in data formatting and collection methods have made it difficult to develop useful insights for stakeholders to develop evidence-based policies. Consequently, this project was undertaken to develop a comprehensive data strategy for these datasets, with its findings and recommendations highlighted in the report. Based on the review, the report proposes an integrated relational database model, bringing together currently separated data sources into a coherent schema. An ETL (Extract, Transform, Load) pipeline is also outlined, to support data transformation and migration. In addition to this, prototype data visualisations were created using Power BI, helping to demonstrate to stakeholders how a proposed centralised database can aid the effective exploration of geographic and temporal trends. These insights can then be used to inform policy and conservation efforts at the IWC. The report concludes with a set of recommendations for establishing robust protocols regarding governance and standardisation to ensure future data strategies avoid past mistakes. Collectively, these deliverables provide a roadmap to overcome current limitations in data utilisation, laying the foundations for a scalable future system to strengthen the IWC's capacity to protect whale populations worldwide.

1. Introduction

The International Whaling Commission (IWC), established in 1946, is the primary global body responsible for whale conservation and the regulation of whaling. A key part of the IWC's conservation work involves gathering and analysing data on human-related threats to cetaceans, including bycatch and entanglement, ship strikes, ocean noise, marine pollution and debris, and the impacts of whale watching practices (International Whale Commission, n.d.). These types of marine incidents pose substantial risks to cetacean populations. Systematic data collection helps to identify emerging threats, informs policy decisions, and facilitates international collaboration.

To support the IWC's objectives, this project focused on the assessment and documentation of relevant marine incident dataset. The scope of work was limited to consultancy, design, and analysis on sample data sources made available to the project team. Due to data sensitivity and restrictions on database access, direct implementation and system deployment were beyond the scope of the project. Consequently, the primary objective was to develop scalable and technical recommendations through documentation, rather than through direct implementation or system deployment. Regular weekly meetings with IWC representatives were held to ensure that the project remained aligned with stakeholder expectations and to incorporate ongoing feedback into the work process.

The project focused on three main datasets. The strandings dataset, derived from National Progress Reports (NPR) and regional databases captures key spatial, biological, and contextual information. The ship strikes dataset, compiled from regional and international sources, contains detailed information on vessel-cetacean collisions, including vessel attributes, incident descriptions, and outcomes. The bycatch dataset documents the accidental capture of cetaceans in fishing operations, which represents a major cause of mortality for many small cetacean species.

This project was structured around four key objectives provided by the IWC, which were followed to ensure alignment with the organisation's strategic priorities:

1. Identification, assessment, and documentation of data sources
   Each dataset was documented individually following a consistent structure, which included an overview of the dataset, data description, detailed feature definitions, identification of data quality issues, and corresponding recommendations.

2. Design of a data model, pipeline, and integration process

   A relational data model, Extract-Transform-Load (ETL) pipeline, and integration strategy were designed to standardise structure, improve data quality, and enable cross-dataset analysis across strandings, ship strikes, and bycatch records.
3. Exploration and development of data visualisation

   Exploratory data analysis using Python and ArcGIS generated static visualisations to reveal patterns. Two interactive Power BI dashboards were developed: one for whale stranding spatiotemporal distribution and species composition, and another for whale ship strike trends and risk areas.
4. Identification of data gaps and limitations

   Limitations in the datasets were documented and accompanied by actionable suggestions to improve completeness, standardisation, and long-term data governance.

This report is structured to reflect the four objectives and the consulting nature of the engagement. Section 2 provides contextual background and a thematic literature review on marine data management. Section 3 presents the results of the dataset review and assessment. Section 4 outlines the proposed data model and pipeline design. Section 5 focuses on the development and exploration of data visualisations. Section 6 highlights data gaps and offers strategic recommendations. Section 7 describes the project process and collaboration with IWC, while Section 8 reflects on the challenges and limitations of the non-implementation scope. The report concludes with a summary of proposed solutions and their anticipated value for future IWC data initiatives.

2. Thematic Literature Review

This section presents a thematic review of literature relevant to the technical and strategic challenges the IWC faces in managing marine conservation data. It explores key domains including environmental data integration, ETL pipeline design for biodiversity datasets, relational data modelling, visual communication in policymaking, and overarching data governance frameworks. The review aims to inform IWC strategies concerning the project at hand.

Environmental data integration and marine data management

The challenge of conserving marine life relies on the effective integration of datasets over significant time and space (Moudrý and Devillers, 2020). Benson et al. (2018) highlight how fragmentation of data can hinder ecological responsiveness. Instead, it is encouraged for organisations to cooperate on analysis and workflows, using data that fits within the FAIR principles (findable, accessible, interoperable, and reusable). The IWC's datasets available to project members are often inconsistent in format and metadata. Adhering to the principles suggested in this article can aid in the construction of an integrated management system that can leverage high-quality standardised data to enable more responsive conservation efforts for a corporation such as IWC. This can allow better development in policy and ecological insight

ETL pipeline design for biodiversity data

El Akkaoui, Vaisma and Zimányi (2019) propose Extract, Transfer, Load (ETL) models to consolidate data in preparation for decision making. Though not specifically linked to biodiversity models, their analysis concluded that these usually intricate and tedious processes should be implemented in data warehouses, but only where the quality of the pipeline design, measured by factors like accuracy, consistency and completeness, can be assured. The IWC's reliance on complex, heterogeneous marine datasets (e.g. strandings, ship strikes, bycatch) could benefit from the adoption of such models to enhance data reliability to support the integration into a single analytical system.

Relational Data Modelling

Morris (2005) highlights the importance of relational databases in biodiversity informatics to manage complex and hierarchical data structures. These structures can then facilitate the

representation of biological relationships between species, particularly marine life in IWC's case. However, the hierarchical nature seen in marine biological systems can present challenges when modelling, as Novotný and Wild (2024) describe that biodiversity databases can lack formalised structures for the specific evolutionary connections (taxa) between species. An alternative approach, the Closure Table model, is suggested instead to represent hierarchical data in relational databases, facilitating data normalisation, reducing redundancy while supporting scalability and data retrieval efficiency. An implementation of such relational data models can enhance the IWC's data analysis and integration of new sources and data types, ensuring consistency while also supporting complex queries.

Visual Communication in Conservation Policymaking

Visual communication plays a pivotal role in shaping public understanding and policy action in marine conservation. Scannell et al. (2023) emphasise that tools such as infographics, photographs, and videos can profoundly influence how environmental issues are emotionally perceived and cognitively processed. When effectively framed, such visuals can translate abstract or distant ecological threats into concrete, relatable narratives, thereby encouraging behavioural change and policy support.

McInerny et al. (2014) further argue that visualisation serves as a bridge between scientific data and stakeholder decision-making. Tools like interactive dashboards and thematic maps help condense complex datasets into digestible formats, enhancing their accessibility for non-specialist audiences, including policymakers and conservation practitioners.

Spatial visualisations, in particular, have emerged as essential instruments in environmental decision-making. Platforms using migration maps, heatmaps, or habitat overlays enable a more intuitive interpretation of issues such as ship strike hotspots or pollution corridors. As the Pew Trusts (2023) highlight, these spatial tools make environmental risks more tangible, thus better informing and guiding regulatory decisions.

For the IWC, such visualisation strategies hold considerable promise. Incorporating interactive dashboards and geospatial mapping tools into its data presentation frameworks can facilitate the communication of whale population trends, threat zones, and policy compliance metrics. This not only strengthens internal decision-making but also improves transparency and stakeholder engagement. The MarineTraffic platform, which enables live ship tracking,

provides a relevant benchmark for how such tools could be implemented in a marine regulatory context.

The literature collectively highlights the challenges the IWC faces in managing marine conservation data effectively across technical and organisational domains. Data integration and ETL pipelines have been seen as foundational for consolidating datasets, offering solutions to handling complex biological hierarchies In addition to this, visual communication strategies have been shown to enhance stakeholder engagement and policy impact. Together, these insights inform the project's recommendations by grounding them in established best practices and aligning them with the IWC's overarching goal of improving data-driven conservation outcomes.

3. Data Review and Assessment (Objective 1)

Objective 1 involved the identification, review, and evaluation of datasets related to cetacean strandings, ship strikes, and bycatch events managed by the International Whaling Commission (IWC). This task helped to understand the current data quality, structure, and readiness for integration.

3.1. Summary of Data Sources

The datasets reviewed for Objective 1 include a range of records related to whale strandings, vessel strikes, and bycatch events.

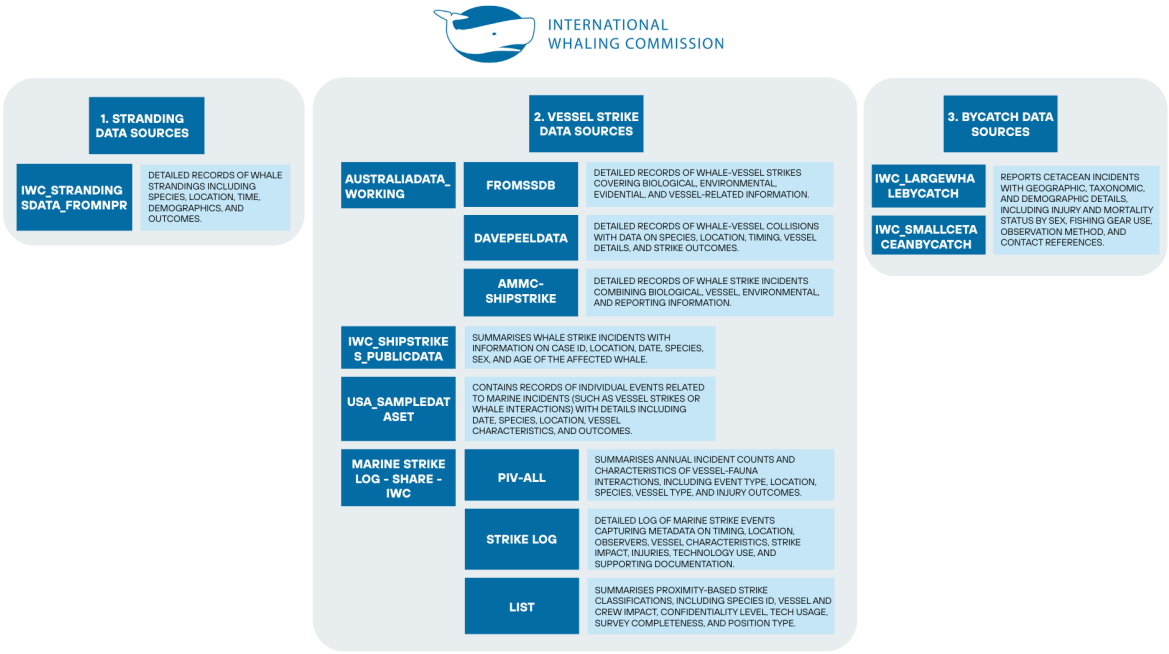

Figure 1: Overview of 11 Datasets Categorised under Three Data Sources

All datasets are provided in Microsoft Excel or CSV format. Several Excel files, namely AustraliaData_working, and Marine strike log - Share - IWC, contain multiple worksheets, each capturing distinct subsets of data or derived summaries.

These datasets come from different regions and sources. They include details about whale species, event locations, types of incidents, and other variables. However, the quality and format of the data differ widely between sources, which is a common issue in ecological data collection (Michener and Jones, 2012).

3.2. Evaluation Method

Each dataset was assessed based on the following criteria adapted from biodiversity data management literature (Chapman, 2005; Hardisty, et al., 2013). These criteria were selected to evaluate the quality, usability, and integration readiness of the datasets for future analysis and database development.

The five criteria used are as follows:

1) Structure and format: This criteria considers whether the data are well-organised and easy to process, with correctly assigned data types and appropriate formats for key variables. Particular attention was given to fields that were misclassified (e.g., dates stored as strings) and to inconsistencies in value formats, such as various representation of "unknown" (e.g., "Unknown", "Not known", "NA", blank) were documented.
2) Completeness: The proportion of missing values in each column was measured to assess data coverage. Completeness is a critical indicator of a dataset's reliability, especially for statistical modelling and spatial analysis.
3) Consistency: The consistency of categorical values was reviewed, such as species names. Variations in spelling, format, and redundancy (e.g., overlapping fields) were noted as potential obstacles to integration and aggregation.
4) Documentation: No formal metadata audit was conducted. Instead, the meanings of each field were described based on file content and logical inference. Most datasets lacked standardised documentation such as data dictionaries, which limits usability for secondary users.
5) Integration potential: The review also considered whether key identifiers (species, location, date) were suitable for linking across datasets. Inconsistencies in naming or structural design were noted as barriers to smooth integration.

3.3. Common Problems and Causes

Several recurring issues were identified across the datasets, all of which affect their reliability and readiness for integration. One common problem was poor formatting in key fields; for example, dates were often stored as plain text and coordinates appeared in inconsistent or non-numeric formats, limiting their use in analysis. In addition, the same values were recorded in multiple ways, particularly for unknown entries, which were represented as "Unknown", "Not known", or simply left blank. This inconsistency

complicates cleaning and standardisation. High levels of missing data were also found in important fields, such as species identification, outcomes, and geographic location. Some columns contained only one repeated value or no meaningful variation, contributing little to analysis while increasing dataset complexity. Furthermore, many files included multiple worksheets with inconsistent structures, making them harder to align or merge.

These problems likely came from combining data from different sources without a shared structure. In some cases, manual data entry also introduced errors. Without standard rules for how to collect or enter data, mistakes and inconsistencies can easily occur and continue across files.

3.4. Summary of Findings and Recommendations

Overall, the datasets reviewed contain valuable content but require several improvements before they can be integrated or used reliably for research. Based on the assessment, the following actions are recommended. First, technical formatting should be corrected to ensure that key fields such as dates and geographic coordinates are stored in usable formats. This will improve compatibility with data analysis tools. Second, standardised terminology should be applied across categorical fields such as species, gear type, and outcome. Using consistent labels will reduce confusion and support data aggregation. Third, unknown values should be recorded using a consistent format, such as “Unknown” or NaN, rather than multiple variations that increase ambiguity. Structural issues should also be addressed by removing duplicate or low-value columns that do not contribute to analysis. In addition, each dataset should be accompanied by clear documentation, including field definitions and units, to support future interpretation. Lastly, preparing the datasets for integration will require common identifiers. By following these recommendations, the IWC datasets can be made more usable, reliable, and ready for integration. This will support future work in database development, data visualisation, and policy analysis.

4. Proposed System Design (Objective 2)

Objective 2 includes a relational data model, an ETL pipeline, and an integration strategy to standardize, clean, and unify datasets from multiple sources. This objective aims to create a database that supports cross-dataset analysis and IWC's data-driven decision-making.

4.1. Conceptual Data Model

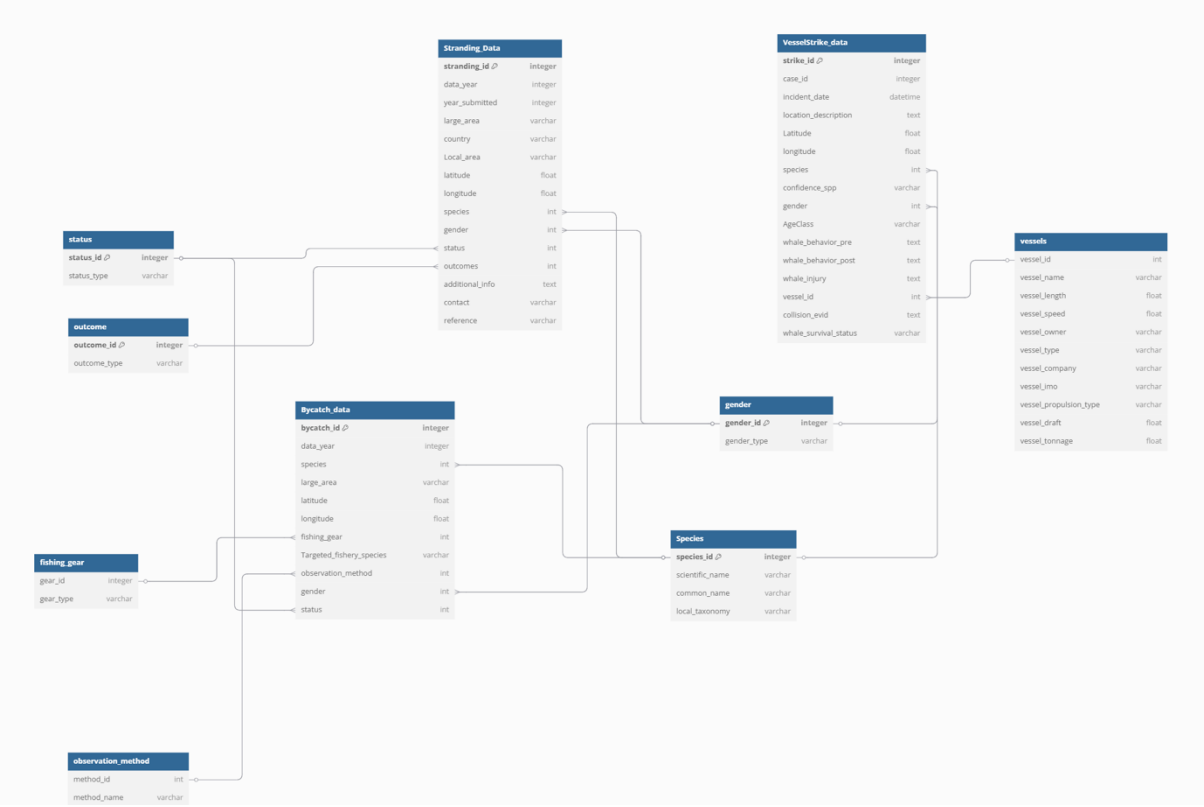

Figure 2: The data model structure

The model organizes the datasets into main tables (Stranding, Bycatch, Vessel Strike) and supporting reference tables (Species, Gender, Status, Outcome, Fishing Gear, Observation Method, Vessels).

This structure is designed based on the following rationale. First, it enhances flexibility by allowing new data types or categories to be added with minimal changes to the existing schema. Second, it promotes data integrity and consistency using foreign keys, which link main tables with certain fields in reference tables. Finally, it supports efficient querying

and cross-dataset analysis, enabling users to explore relationships across different marine incident types in a unified manner.

4.2. ETL Pipeline Proposal

Based on the theory from Kimball, R., and Caserta, J. (2004), the general process of constructing the ETL pipeline for IWC are as follows:

1) Extract: Read raw data from source files, handling different formats (CSV, Excel) and identifying specific sheets within workbooks.
2) Transform: Perform a series of cleaning and standardization steps tailored to each dataset. Common operations include:
    1. Data type conversions (dates, numbers, IDs).
    2. Standardization of categorical values (species, unknown indicators, locations).
    3. Handling missing data (identification, standardization, imputation/dropping strategy).
    4. Structural changes (splitting columns like Lat/Lon, reshaping wide data to long, decomposing large tables into normalized ones).
    5. Column management (renaming, dropping unnecessary columns).
    6. Ensuring data integrity (key validation, unit clarification).
3) Load: Insert the processed data into predefined tables within the target database, establishing relationships using primary and foreign keys where tables have been normalized.

The following figures illustrate the detailed ETL pipelines for each dataset:

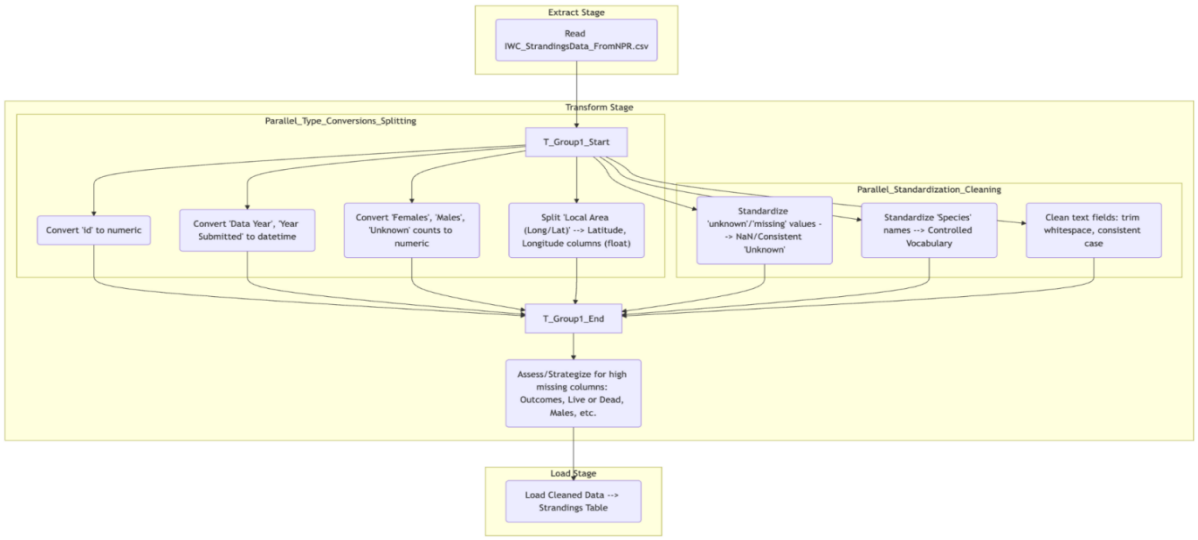

Figure 3: ETL Pipeline for IWC Strandings Dataset

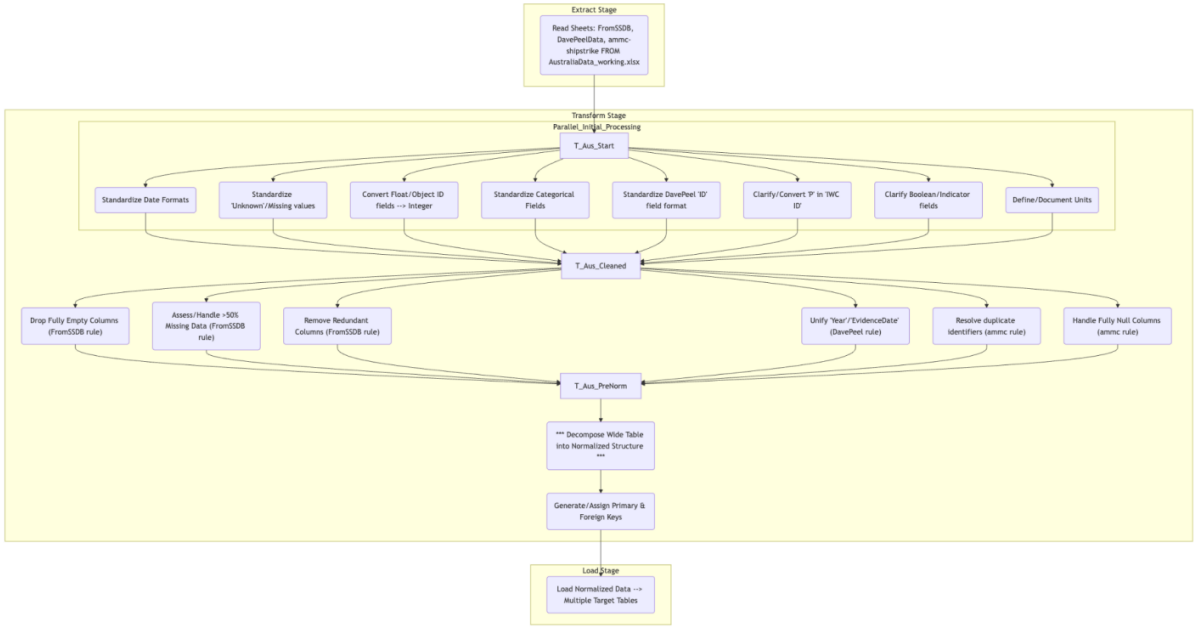

Figure 4: ETL Pipeline for Multi-Sheet Australian Incident Datasets

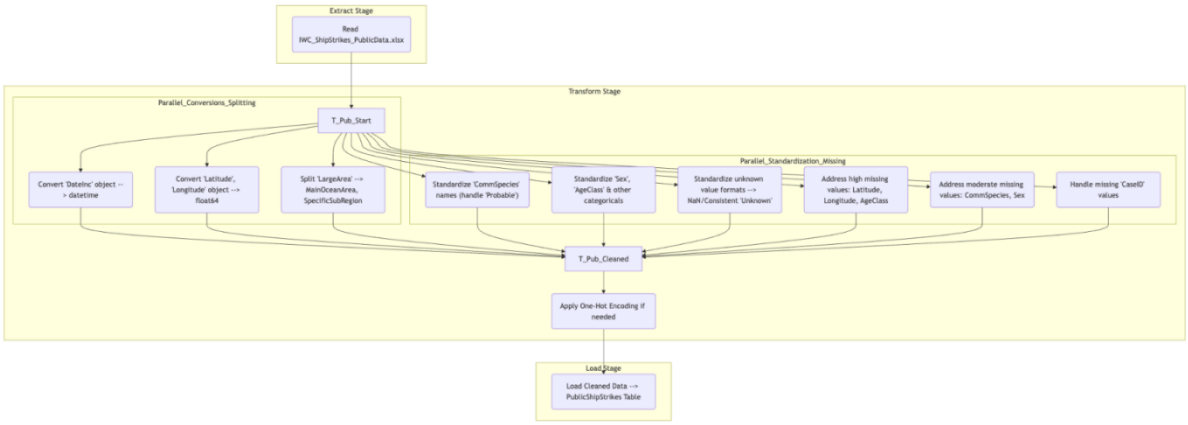

Figure 5: ETL Pipeline for IWC Public Ship Strikes Dataset

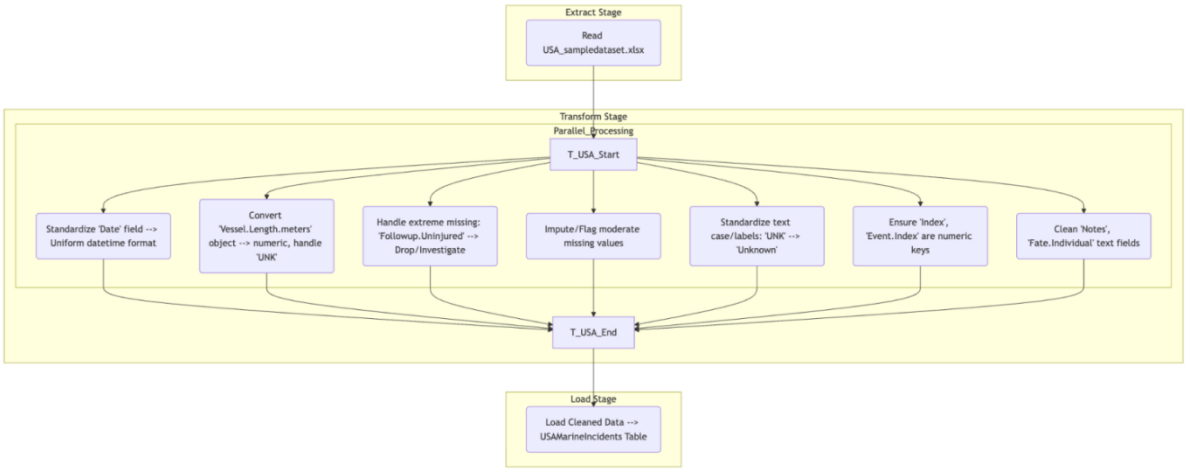

Figure 6: ETL Pipeline for USA Marine Incidents Dataset

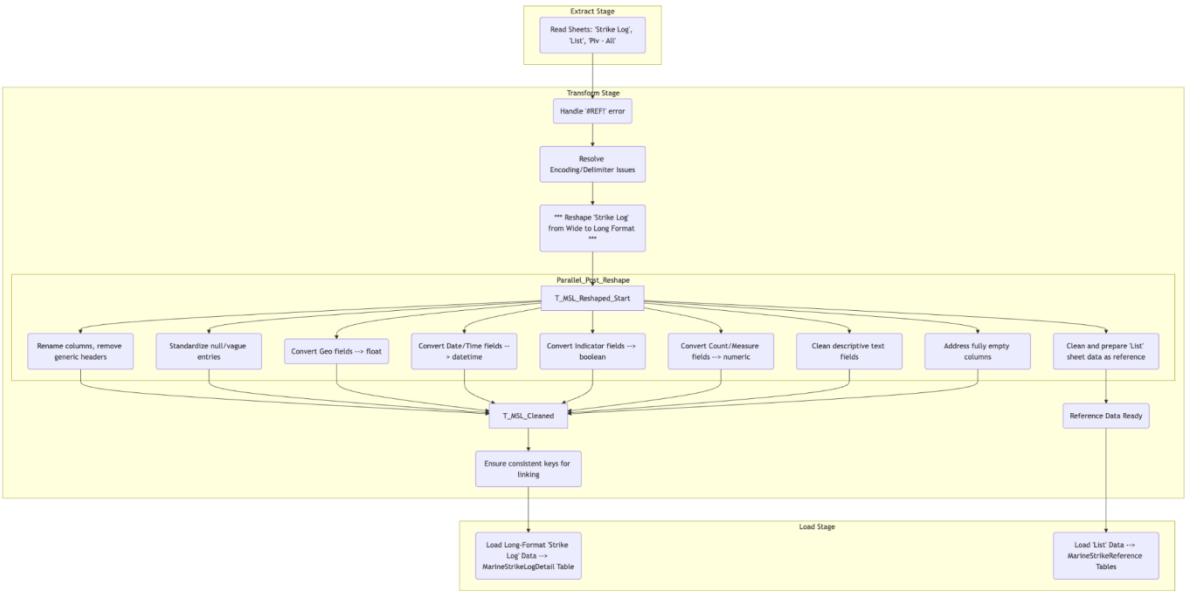

Figure 7: ETL Pipeline for Marine Strike Log Dataset

4.3. Data Cleaning Rules

During the ETL process, standardized data cleaning rules were applied. The rules include handling data types, missing values, and standardizing categorical variables.

For data types, date fields were converted to standardized datetime formats, while numeric fields, such as counts and IDs, were validated and transformed accordingly. Text fields were also cleaned. Extra spaces were removed to ensure the consistency of the whole database.

Missing values were handled by standardizing placeholders like blanks and “unknown” into unified formats such as "Unknown" or NaN. Fields with extreme missingness were either excluded or flagged.

Categorical values were standardized to enable integration. Species names were mapped to a controlled vocabulary, and other fields like gender, status, fishing gear, and observation methods were harmonized using lookup tables. Redundant entries were cleaned and merged to maintain clarity. Where necessary for advanced analysis or machine learning, categorical fields can further be converted into one-hot encoded formats.

Through these cleaning rules, all datasets were transformed into a structured and harmonized format suitable for integration into the unified database.

4.4. Integration Strategy

The integration pipeline draws inspiration from industry examples such as IBM's layered architecture model (IBM, 2010) and the operational flow outlined in DWG-based warehousing literature (Liu, 2001). Our flowchart models a simplified version of these, showing data flowing from raw ingestion through transformation and mapping to a unified schema.

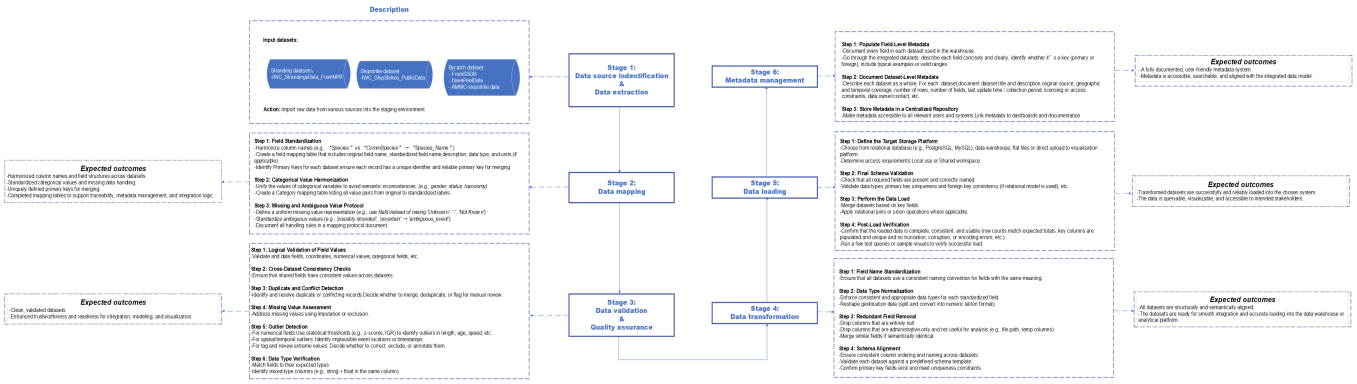

Figure 8: The conceptual structure of integration strategy

As illustrated in the diagram, the pipeline consists of six main stages:

Stage 1: Metadata Extraction and Initial Data Collection

Input datasets from various sources. Each dataset's schema and data characteristics are recorded to inform subsequent processing.

Stage 2: Data Importing

During this stage, raw data files are imported. Different kinds of files like Excel and CSV are parsed and prepared for transformation.

Stage 3: Data Handling

Initial cleaning processes are applied. Moreover, fields with special structures (such as latitude/longitude) are normalized at this stage.

Stage 4: Data Harmonization

Schema-level reconciliation is performed, including mapping categorical variables (species, gender, status) to unified lookup tables.

Stage 5: Quality Assurance

Perform validation checks on the harmonized data, such as verifying key relationships and confirming valid field values.

Stage 6: Database Loading and Integration

The final processed datasets are loaded into the relational database according to the conceptual model. Lookup tables and foreign keys are used to establish links between datasets and support cross-dataset queries.

This step-by-step integration strategy helps the originally fragmented and inconsistent datasets to be transformed into a clean, standardized, and queryable database.

4.5. Metadata and Documentation Standards

According to the ISO/IEC 11179-1:2023 standard (International Organization for Standardization, 2023), metadata should clearly define the meaning, representation, and permissible values of data elements to ensure consistent understanding and management. In line with this principle, our database design metadata and documentation standards to support clarity, consistency, and long-term maintainability.

Specifically, categorical fields are standardized through lookup tables, which clearly define the values and ensure uniform classification across datasets. In addition, the database schema and field descriptions are carefully documented, including information about data types, field definitions, units, and sources. Time fields follow a standardized datetime format, and location fields are stored as latitude and longitude coordinates.

Furthermore, to facilitate future updates and ensure continued consistency, a data dictionary is recommended. This document should record field definitions, allowed values, sources, and update rules, particularly for dynamic categories such as species and observation methods. Together, these metadata practices help users accurately interpret the data and provide a clear framework for future data integration and maintenance.

4.6. Summary of Findings and Recommendations

The design in this chapter provides a unified and scalable solution for IWC's datasets. Through the relational data model, tailored ETL pipelines, data cleaning rules, and integration strategy, diverse datasets were standardized and harmonized. Metadata and documentation standards further ensure usability and future maintenance. It is recommended to maintain lookup tables, implement automated ETL processes, and establish regular data reviews to support ongoing integration and data-driven decision-making.

5. Visualisation concept design (Objective 3)

This task aims to develop interactive data visualisation products based on the integrated whale datasets to support the IWC's data-driven decision-making in whale conservation and management. Initially, Python will be used to conduct exploratory data analysis and generate static visualisations of the master table, revealing key data characteristics and trends. Building on this, two interactive dashboards will be designed and implemented: the first will focus on the spatiotemporal distribution and species composition of whale strandings, highlighting patterns across countries, years, and species; the second will present global trends and high-risk areas of whale ship strikes, offering visual support for risk assessment and management.

5.1. Methodology

This task begins with data preparation based on the cleaned and standardized integrated main tables.

Preliminary visualisation is conducted using Python and ArcGIS, generating histograms, time series plots, bar charts and geospatial heatmaps. These exploratory visualisations help reveal fundamental patterns in species distribution, stranding density, and temporal trends, guiding the design of the dashboards.

For dashboard development, Power BI is used to create interactive visualisation tools, consisting of two main dashboards. The first dashboard focuses on the species composition and spatiotemporal distribution of whale strandings. It integrates a species-region matrix and annual stacked trend charts, with filtering options by year and country to identify dominant species and distribution patterns across regions. The second dashboard addresses global trends and spatial risk patterns of whale ship strikes, incorporating annual incident trends, a global spatial distribution map, and a spatiotemporal heatmap. It supports multidimensional filtering by year, region, and species to assist in identifying high-risk areas and informing management strategies.

All dashboards are equipped with interactive features such as filters by time, country, and species. They also support tooltip details on hover and allow for exporting charts and data, enhancing usability and facilitating further analysis and reporting.

5.2. Exploratory visualisation of 3 main tables (Stranding, Bycatch, Vessel strike)

5.2.1. Stranding

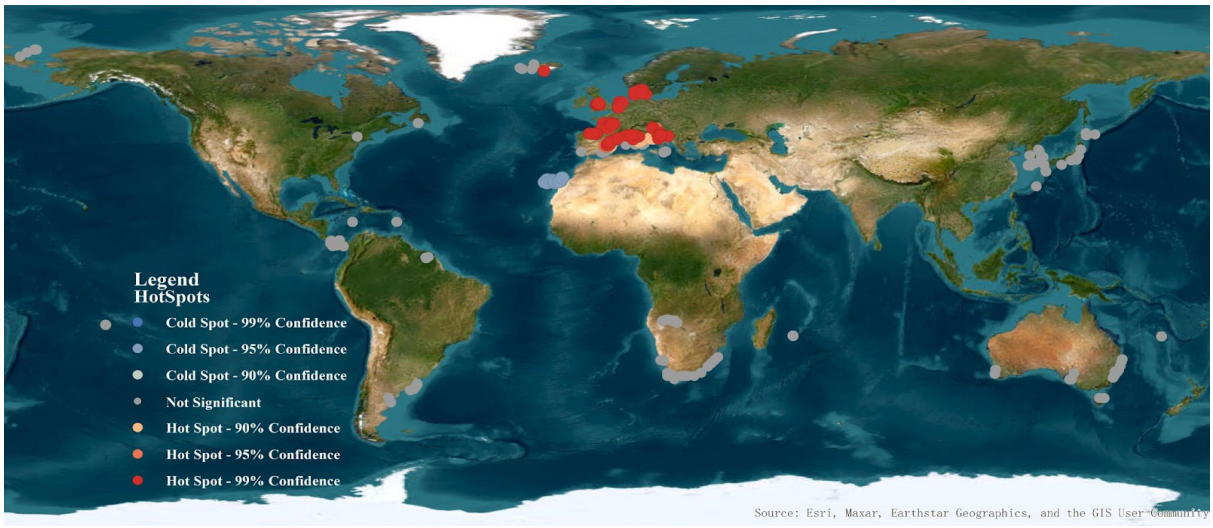

Figure 9: Hotspot Analysis of Whale Strandings

This figure presents a global spatial hotspot analysis of whale stranding events. Red areas (high concentration) are primarily along the western European coast and North Atlantic, blue areas (low frequency) include western South America and East Asia, and gray areas are statistically non-significant, with confidence levels at 90%, 95%, and 99%.

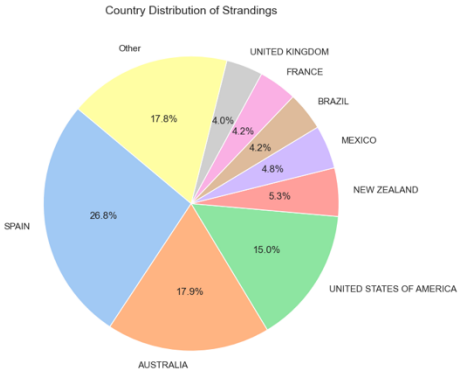

Figure 10: Country Distribution of Strandings(Groups <4% Merged as 'Other')

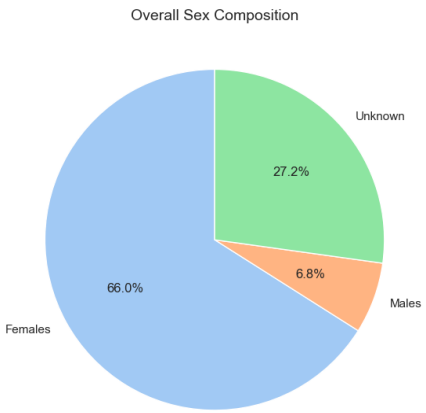

Figure 11: Overall Sex Composition of Stranded Individuals

Figure 10 shows whale stranding distribution by country. Spain leads (26.8%), followed by Australia (17.9%) and the United States (15.0%). Other contributors include New Zealand (5.3%), Mexico (4.8%), Brazil (4.2%), France (4.2%), and the UK (4.0%), suggesting regional risk or reporting variations.

Figure 11 presents the sex composition of stranded whales. Females dominate (66.0%), compared to males (6.8%), with 27.2% unidentified. This highlights potential behavioral differences or data collection gaps, emphasizing the need for standardized sex reporting.

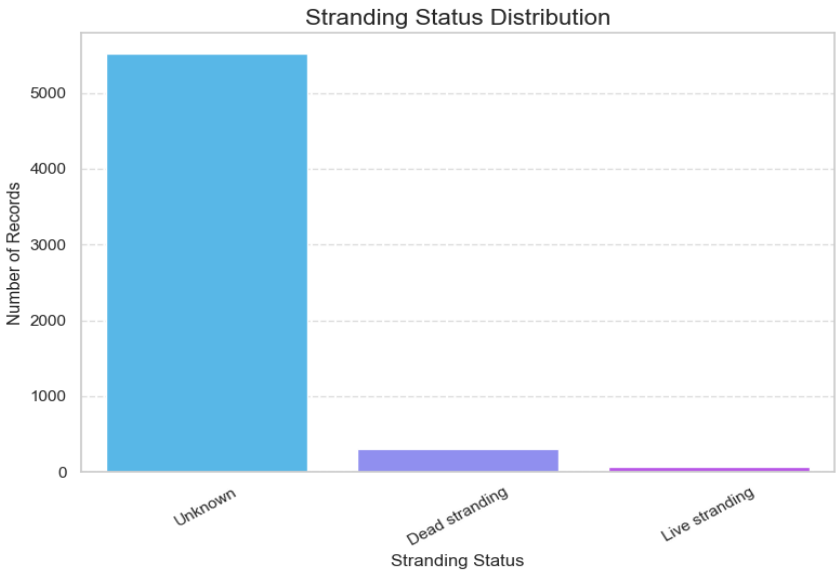

Figure 12: Stranding Status Distribution

This figure shows the distribution of whale stranding events by status. The “Unknown” category dominates with over 5,000 records, highlighting a major data gap. “Dead stranding” records are around 300, and “Live stranding” events are under 100.

5.2.2. Bycatch

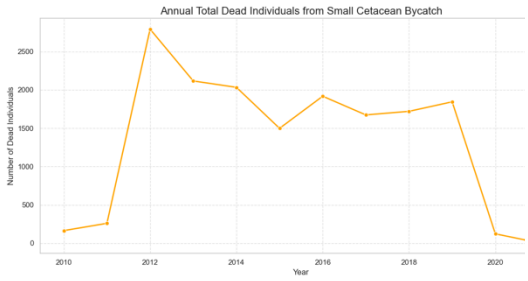

Figure 13: Annual Total Dead Individuals from Small Cetacean Bycatch

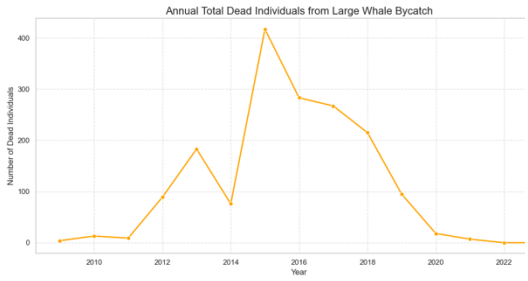

Figure 14: Annual Total Dead Individuals from Large Whale Bycatch

Figure 13 shows annual small cetacean bycatch deaths, peaking at over 2,700 in 2012. Deaths fluctuated between 1,500–2,200 thereafter, with a sharp decline post-2020, likely due to reduced fishing, COVID-19 disruptions, or incomplete data.

Figure 14 displays large whale bycatch deaths, peaking at over 400 in 2015. Numbers declined from 2016–2018, suggesting mitigation efforts. Post-2019, deaths remained low, possibly due to pandemic effects, delayed reporting, or data inaccuracies.

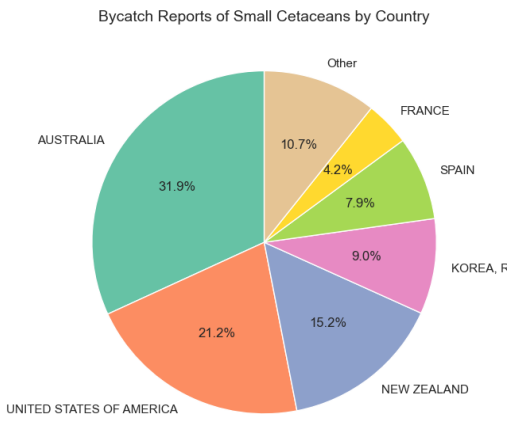

Figure 15: Bycatch Reports of Small Cetaceans by Country

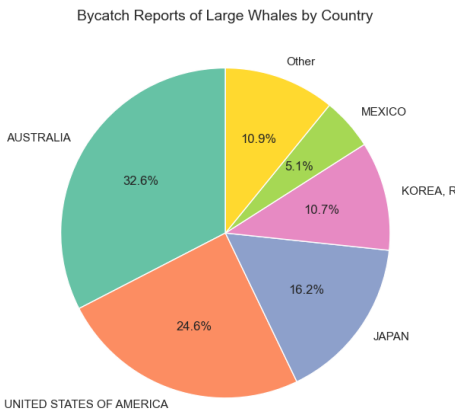

Figure 16: Bycatch Reports of Large Whales by Country

Figure 15 shows the country-wise distribution of small cetacean bycatch reports. Australia leads (31.9%), followed by the United States (21.2%) and New Zealand (15.2%). Other contributors include Korea (9.0%), Spain (7.9%), France (4.2%), and "Other" countries (10.7%), highlighting concentration in fisheries-intensive nations.

Figure 16 presents large whale bycatch reports by country. Australia ranks highest (32.6%), followed by the United States (24.6%), Japan (16.2%), Korea (10.7%), and Mexico (5.1%), indicating significant bycatch issues in key coastal fishing nations.

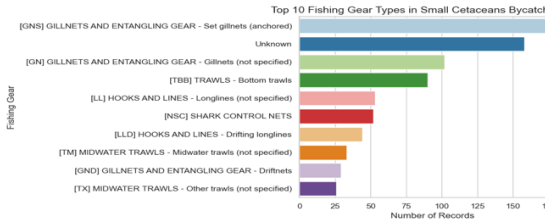

Figure 17: Top 10 Fishing Gear Types in Small Cetaceans Bycatch Records

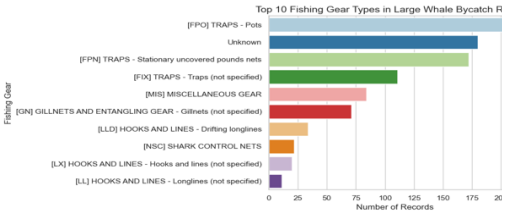

Figure 18: Top 10 Fishing Gear Types in Large Whale Bycatch Records

Figure 17 shows the top 10 fishing gear types in small cetacean bycatch. Set gillnets (GNS) lead with ~180 records, followed by "Unknown" gear (around 160 records). Other types include unspecified gillnets (GN), bottom trawls (TBB), and longlines (LL), emphasizing data gaps and the need for gear-specific mitigation.

Figure 18 presents the top gear types in large whale bycatch. Pots (FPO) lead with over 200 records, followed by "Unknown" entries. Other types include stationary pound nets (FPN), unspecified traps (FIX), and gillnets (GN), highlighting diverse fishing methods and the need for improved gear reporting.

5.2.3. Vessel strike

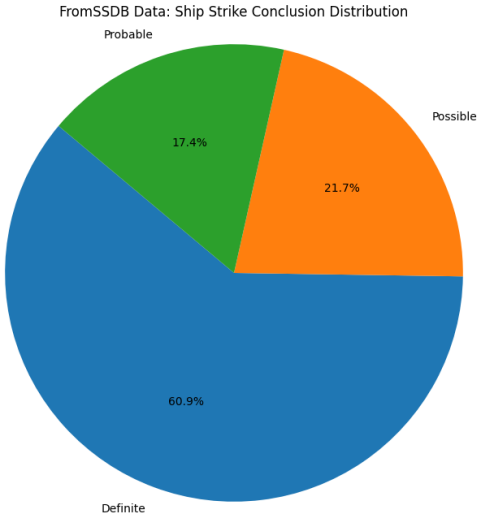

Figure 19: Pie Chart on Ship Strike Conclusion Distribution

This pie chart illustrates the distribution of ship strike cases by conclusion (Definite, Probable, Possible). Definite cases dominate (60.87%), followed by Possible (21.74%) and Probable (17.39%), indicating robust evidence for most incidents but highlighting areas for improved data collection.

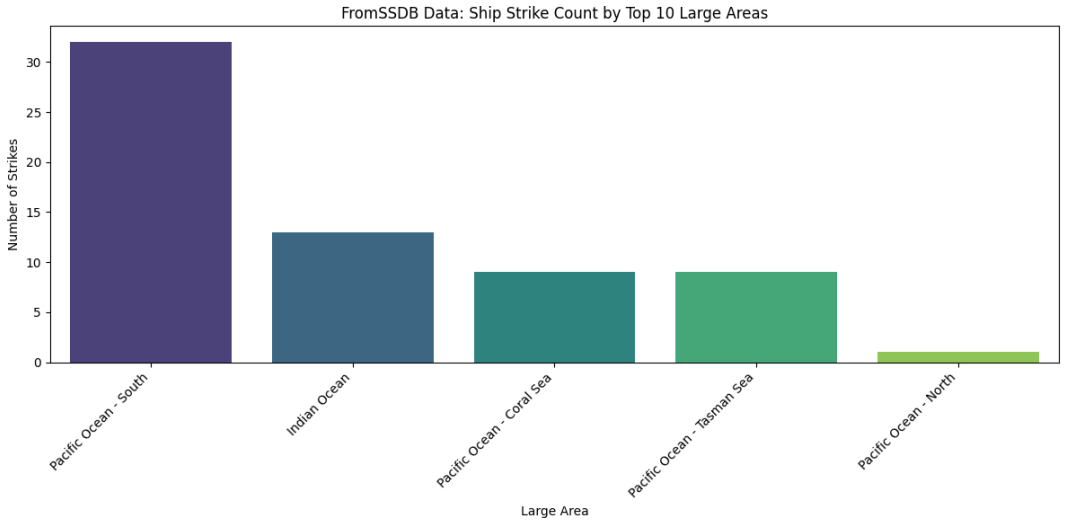

Figure 20: Bar Chart on Ship Strike Count by Top 10 Large Areas

This bar chart highlights the top 10 geographic areas with the highest ship strike counts. The Pacific Ocean - South leads with 23 cases, followed by the Indian Ocean and

Pacific Ocean - Coral Sea (7 cases each), identifying key hotspots for conservation focus.

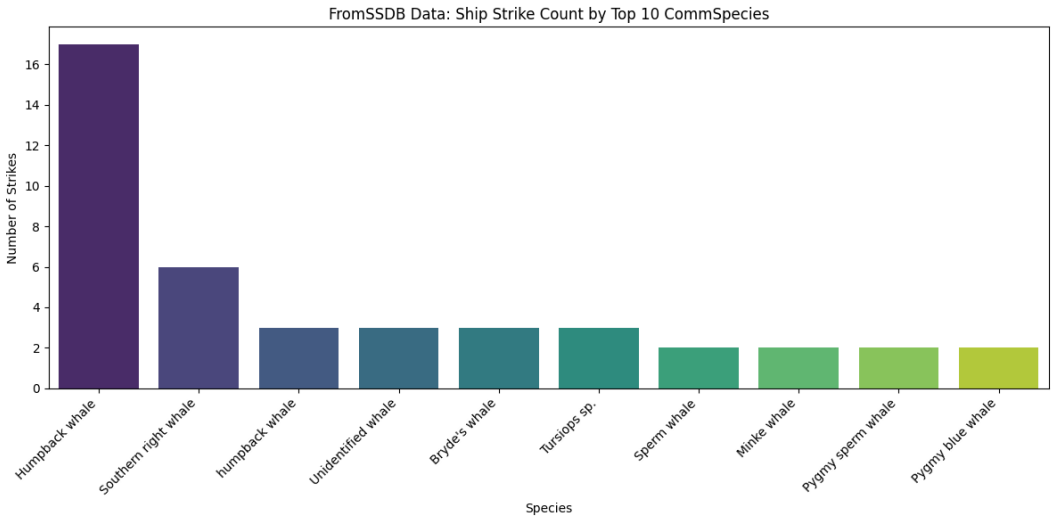

Figure 21: Bar Chart on Ship Strike Count by Top 10 Species

This bar chart presents the top 10 species involved in ship strikes. Humpback whales are the most affected (17 cases), followed by Southern right whales (6 cases), suggesting higher vulnerability due to migratory behavior or coastal distribution.

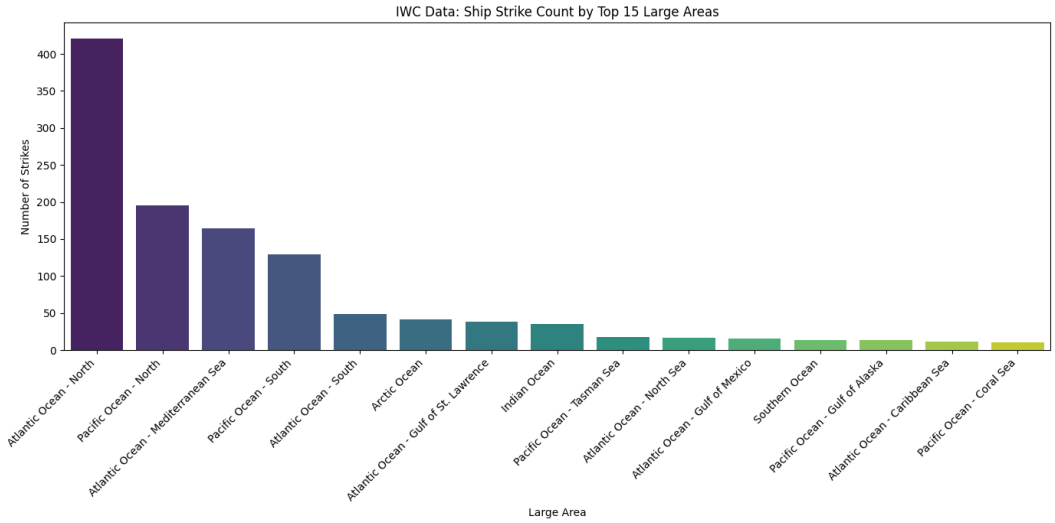

Figure 22: Bar Chart on Ship Strike Count by Top 15 Large Areas

This bar chart displays the top 15 large geographic areas with the highest number of reported ship strikes in the IWC Public Data. It identifies the major regions globally

where collisions between vessels and cetaceans are most frequently documented, highlighting key hotspots for conservation efforts.

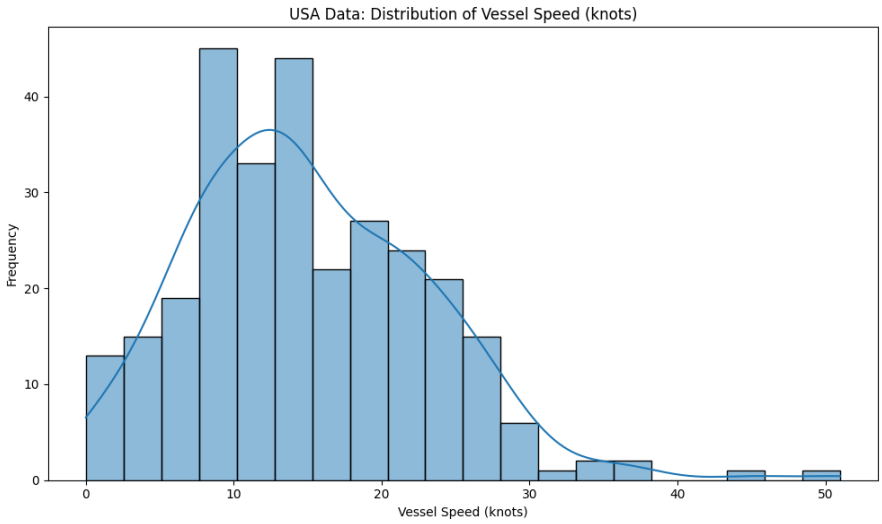

Figure 23: Histogram on Distribution of Vessel Speed (knots)

This histogram shows the frequency distribution of reported vessel speeds in knots at the time of ship strikes in the USA dataset. It indicates the range and common speeds at which these collisions occur, providing insights into speed-related risk factors.

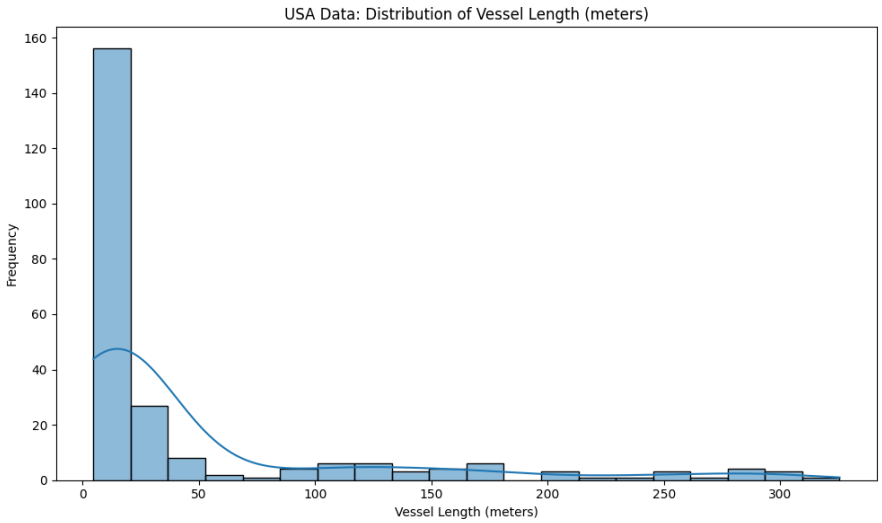

Figure 24: Histogram on Distribution of Vessel Length (meters)

This histogram illustrates the frequency distribution of reported vessel lengths in meters for vessels involved in ship strikes within the USA dataset. It provides insight into the size characteristics of the vessels implicated in these incidents.

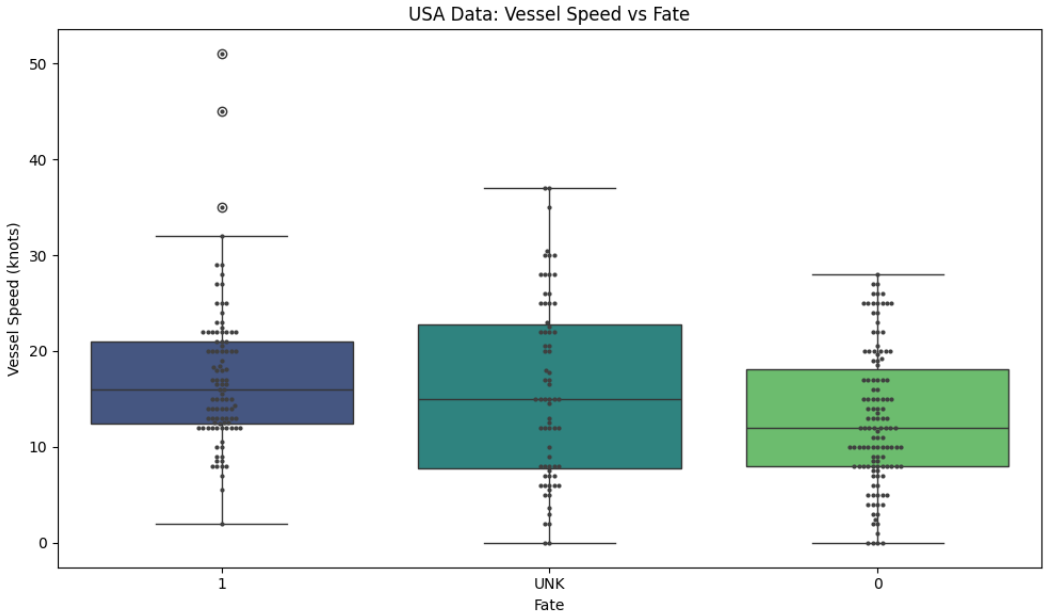

Figure 25: Box Plot/Swarm Plot on Vessel Speed vs. Fate

This plot combines box plots and a swarm plot to visualize the relationship between vessel speed and the outcome ('Fate') of ship strikes in the USA dataset. It allows for comparison of speed distributions across different fate categories (e.g., fatal vs. non-fatal), aiding in risk assessment.

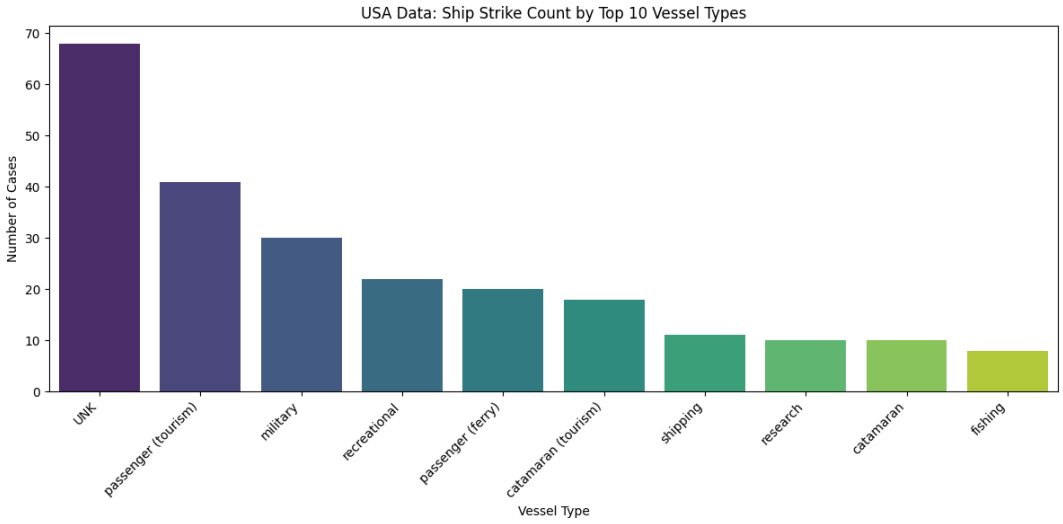

Figure 26: Bar Chart on Ship Strike Count by Top 10 Vessel Types

This bar chart displays the top 10 most frequent vessel types reported as being involved in ship strikes within the USA dataset. It highlights the categories of vessels most commonly associated with these incidents, informing targeted mitigation strategies.

5.3. Dashboard design and implementation

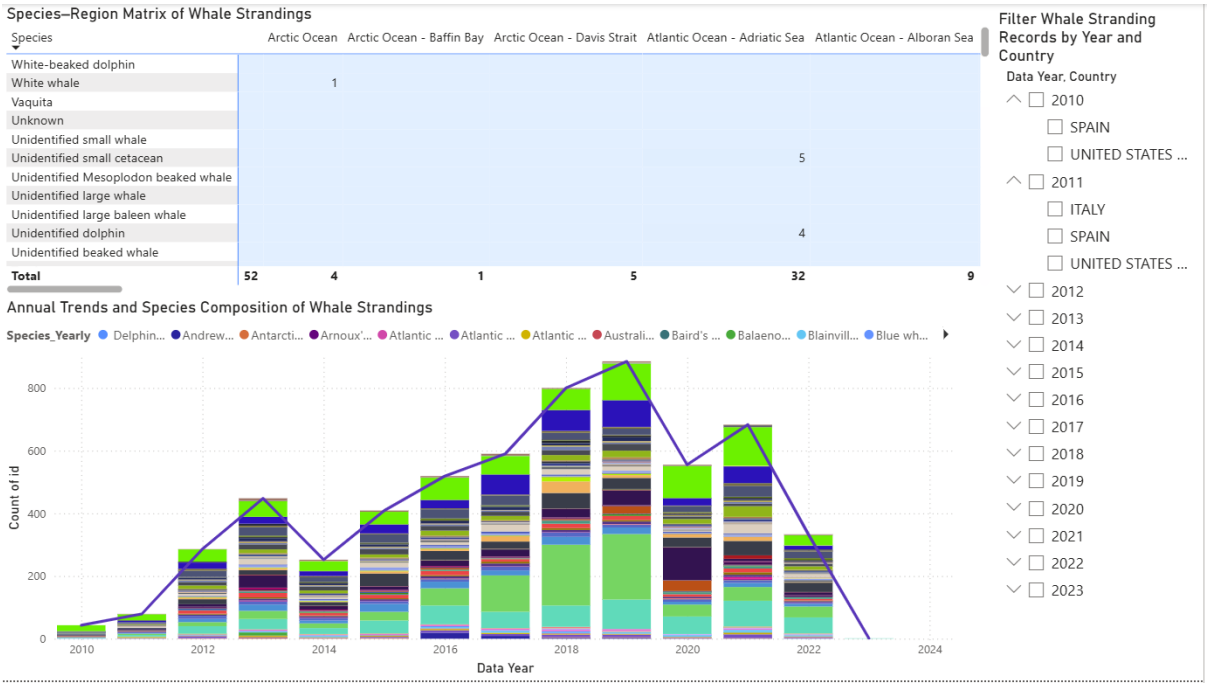

Figure 27: Dashboard on spatiotemporal distribution and species composition of whale strandings

This dashboard presents a comprehensive view of global whale stranding events by integrating species identity, ocean region, and temporal trends. Based on stranding data, it allows users to explore patterns in the frequency and distribution of strandings across both space and time, helping to identify high-risk areas and vulnerable species for conservation attention.

The species–region matrix highlights how certain whale species are disproportionately represented in specific ocean zones, revealing spatial concentrations and ecological exposure. The stacked column and line chart below visualizes the yearly evolution of stranding events from 2010 to 2023, showing both the overall trend and the changing species composition over time. On the right, the year–country slicer enables targeted filtering, allowing users to isolate and examine stranding events for specific periods and nations.

Designed for researchers, marine conservation organizations, and policy makers, this dashboard serves as an analytical tool for understanding whale stranding phenomena. It

supports better-informed decisions in rescue planning, coastal resource allocation, and strategic prioritization for species protection at national and global levels.

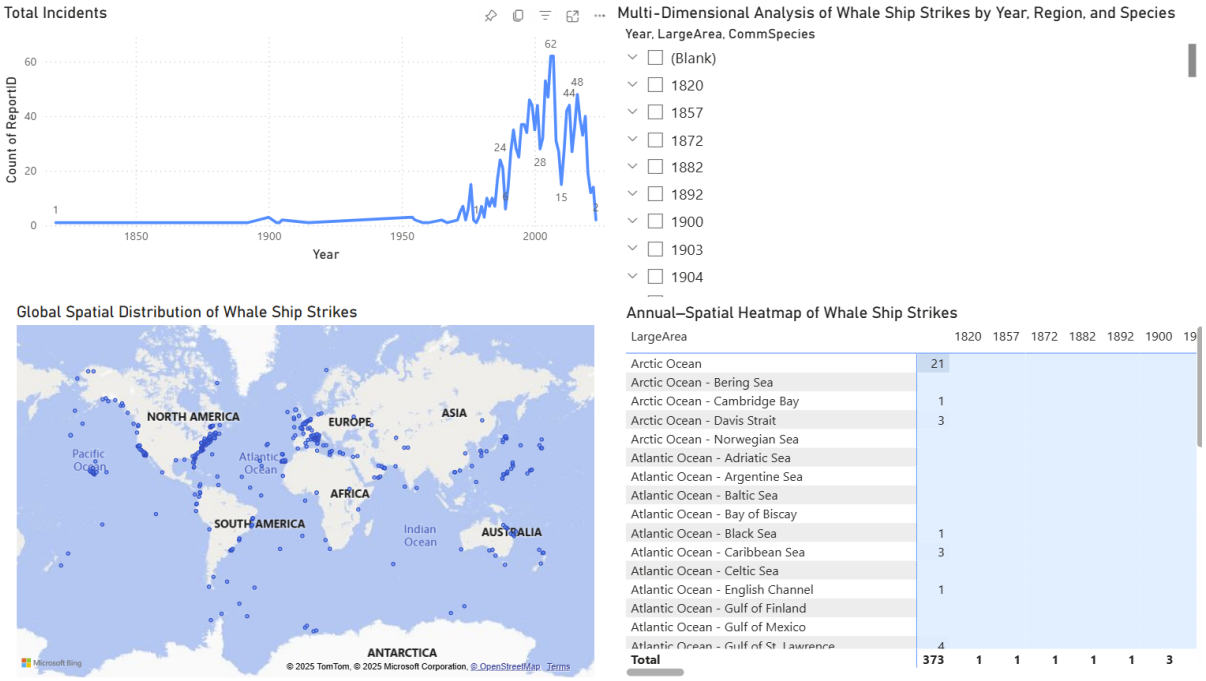

Figure 28: Dashboard on global trends and spatiotemporal risk patterns of whale ship strikes

This dashboard presents a multi-dimensional analysis of whale ship strikes based on publicly available data from the IWC. It integrates temporal, spatial, and species-level dimensions to support the investigation of long-term trends, regional concentrations, and species-specific risk exposure.

The top-left line chart displays the total number of reported ship strike incidents over time, highlighting the sharp increase in reports since the mid-20th century. The bottom-left map visualizes the global spatial distribution of ship strike events, indicating geographic hotspots across major oceans, particularly in high-traffic coastal regions.

On the right, the heatmap reveals the cross-distribution of incidents by year and ocean region, helping to identify spatiotemporal hotspots. The slicer above allows users to filter the data by year, region, and species to focus the analysis on specific dimensions of interest.

Together, these components provide a comprehensive platform for researchers, conservationists, and policymakers to evaluate risk patterns, assess conservation outcomes, and develop region-specific maritime mitigation strategies.

5.4. Outcomes and insights

Objective 3 successfully delivered interactive visualisation tools that transform integrated cetacean data into accessible and insightful formats. Through Python and ArcGIS-based exploratory analysis and the development of two Power BI dashboards, the project uncovered meaningful patterns in whale strandings and ship strikes across time, geography, and species. These visualisations provide the IWC with an evidence-based foundation for identifying priority regions and species, facilitating targeted conservation and management strategies.

The dashboards enhance stakeholder engagement by enabling intuitive data exploration and promoting transparency in marine mammal monitoring. Looking forward, future enhancements may include predictive modelling, effort-standardised comparisons, integration of external datasets (e.g., environmental or shipping data), and improved user interfaces tailored to different stakeholder needs. Collectively, this objective demonstrates the critical role of visual analytics in supporting the IWC's data-driven decision-making and long-term conservation goals.

6. Data Gaps and Strategic Recommendations (Objective 4)

This section identifies key gaps in the coverage and scope of IWC datasets, focusing on five dimensions—geographic, temporal, species, metadata, and structural integration. These dimensions were informed by the *Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders* published by the Global Biodiversity Information Facility (GBIF) (Ariño, Chavan & Otegui, 2016). The goal is to evaluate how well existing data reflect global cetacean incidents and to inform recommendations for improving completeness and standardisation.

6.1. Geographic Gaps

Spatial analysis reveals a strong imbalance in data submissions, with concentration in a few countries and ocean basins. At the national level, Australia, the USA, Spain, and New Zealand contribute disproportionately to the records, while countries across West Africa and Southeast Asia are notably underrepresented. At the ocean basin level, the Atlantic and Pacific Oceans are relatively well covered, whereas the Indian and Arctic Oceans have sparse reporting. These disparities are consistent across stranding, vessel strike, and bycatch datasets. To improve geographic representativeness, we recommend targeted data sharing and engagement initiatives with underreported regions.

6.2. Temporal Gaps

Temporal coverage is inconsistent across datasets. Many records cluster post-2010, while earlier periods (especially pre-2000) are poorly represented. Some datasets, such as DavePeelData or the Marine Strike Log, are primarily retrospective, offering historical but uneven coverage. Others, like AMMC, reflect narrow reporting windows. Reporting delays, identified through comparison of event and submission years, show irregularities including negative delays. Recommendations include retrospective data retrieval, better metadata flagging, and clearer differentiation between historical and real-time reporting.

6.3. Species Gaps

Species reporting is dominated by high-profile and coastal species such as humpbacks, bottlenose dolphins, and common dolphins. In contrast, rarer or offshore species, particularly beaked whales and small odontocetes, are often underreported or absent. Many entries are vague (e.g., “unidentified dolphin”), especially in older records, indicating limited taxonomic resolution. These patterns are consistent with known biases in stranding data, where elusive or pelagic species such as beaked whales and melon-headed whales are less likely to strand or be detected due to their offshore habitat and small population sizes (Williams et al., 2011). Grouping species with fewer than three records under “Other species” helps visualise this gap.

6.4. Metadata Gaps

Across all datasets, key metadata such as age, sex, coordinates, and incident outcomes are frequently missing or marked "Unknown." Inconsistent terminology and unstructured formats further complicate cleaning and analysis. Some datasets lack versioning, coded field definitions, or use obsolete columns. These challenges reflect issues in biodiversity data management, where inconsistent terminology, vague definitions, and heterogeneous schemas are known to limit data usability and integration (Chapman, 2005). Standardised metadata practices and field validation protocols are recommended to enhance data quality and traceability.

6.5. Structural and Integration Gaps

There are differences in schema, for example, Species, CommonName, CommSpecies. Field availability create challenges for integrating datasets. No common identifier exists across datasets (e.g., strandings vs. ship strikes), and similar fields are often stored in different formats. Without a unified structure or controlled vocabulary, cross-dataset linkage and large-scale synthesis are difficult. A harmonised schema and minimal required fields should be developed for future data consolidation.

6.6. Summary of Findings and Recommendations

The analysis of IWC datasets revealed persistent gaps across geographic, temporal, species, metadata, and structural dimensions. Data submissions are heavily skewed toward certain countries and ocean basins, with large geographic regions underrepresented. Temporally, coverage is uneven and often delayed, particularly before 2000. Species records show bias toward a few well-known cetaceans, with vague or unidentified entries limiting taxonomic insights. Metadata fields such as age, sex, location, and outcome are frequently missing or inconsistently formatted, reducing data usability. Structurally, inconsistent schemas and field names hinder dataset integration. To address these gaps, we recommend: (1) expanding partnerships to improve geographic and temporal coverage; (2) standardising species classification and encouraging expert verification; (3) implementing controlled vocabularies and data validation rules for metadata; and (4) adopting a harmonised schema with minimal required fields to support cross-dataset integration and scalable analytics.

7. Project Process and Management

The IWC data documentation project was conducted using a structured, phased process designed to meet four defined objectives. Given that the project was not implementation-based but rather documentation-focused, the team placed strong emphasis on structured and time-bound approach supported by clear task assignments, a colour-coded Gantt timeline (Figure 29) to structure and track this timeline.

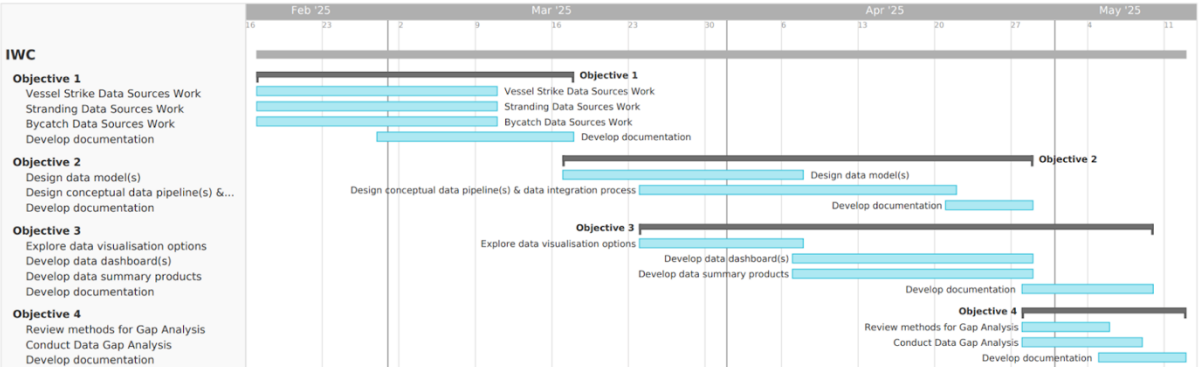

Figure 29: Timeline of Tasks and Deliverables for the IWC Project, Organised by Objectives 1- 4 (February to May 2025)

7.1. Objective-Based Planning and Structure

Each objective was broken down into specific tasks with defined start and end dates. The schedule ensured that outputs from earlier objectives informed later ones. For instance, data source reviews under Objective 1 directly shaped the design of data models in Objective 2.

As shown in Figure 29, the work began in mid-February with a detailed scoping and planning phase.

Objective 1: Data Source Assessment (17 Feb - 17 Mar) Activities included reviewing existing datasets related to strandings, ship strikes, marine debris, and bycatch. Team members documented source formats, completeness, and schema alignment using standardised profiling templates.

Objective 2: Data Model and Pipeline Design (17 Mar - 28 Apr) This phase involved drafting normalised relational database schemas using entity-relationship diagrams, defining controlled vocabularies, and proposing ETL (Extract,

Transform, Load) workflows. Pipeline diagrams and data validation rules were completed during the second half of this phase.

Objective 3: Visualisation Planning (24 Mar - 9 May)
Mock dashboards, exploratory chart templates, and layout designs were developed using sample datasets. This included visuals for strandings distribution, bycatch severity, and data coverage by region.

Objective 4: Gap Analysis (28 Apr - 12 May)
Findings from Objectives 1 and 2 were consolidated to produce a structured gap report highlighting missing fields, inconsistencies in taxonomies, and regional imbalances in data availability.

7.2. Workflow and Management Approach

The project applied a simplified agile workflow, with development activities structured in overlapping phases. Each task followed a "review, build, document" cycle, supported by regular meetings and shared documentation. Collaborative tools and version control systems enabled real-time coordination and parallel workstreams.

Workload distribution was managed equitably among team members to ensure balanced contributions across all objectives. After the completion of individual components, a cross-review process was conducted to provide internal feedback and quality assurance before submitting outputs to the IWC.

Gantt chart planning helped maintain transparency and time control (Figure 29). Task dependencies were mapped to avoid delays and ensure that inputs from earlier stages were ready for subsequent workstreams.

7.3. Meetings with the Industry Partner (IWC)

Initial scoping and alignment meetings were held weekly with a IWC representative to clarify expectations, constraints, and documentation requirements. While regular contact was limited due to time zone differences, connectivity, and team members availability, key feedback was provided on schema design decisions and data prioritisation. Internal team meetings were used to monitor progress, troubleshoot issues, and coordinate task handovers across objectives.

7.4. Project Tracking and Workflow Tools

Project tracking with the IWC was handled through a shared Gantt chart and a shared drive containing all project outputs. Internally, the team used collaborative tools such as Google Docs and Microsoft Word (via shared OneDrive links) to coordinate writing and facilitate parallel editing. The workflow was organised into weekly sprint cycles, each with its own set of deliverables and internal deadlines.

8. Reflections and Limitations

8.1. Reflections

In this project, we have deepened our understanding of the processing and visualization of environmental data. At the beginning, we realized that data cleaning is not only about removing erroneous data but also about ensuring structural compatibility between different datasets. By learning and applying the ETL (Extract, Transform, Load) process, we constructed a clear pipeline for data processing so that raw data could be transformed into an analyzable format. In terms of visualization, we explored how to transform spatio-temporal data into meaningful patterns. In addition to using static charts, we experimented with designing interactive dashboards to present trends hidden in the data in a more intuitive and engaging way. These explorations greatly enhanced the interpretability of our results and were particularly effective in revealing the relationship between vessel activity and cetacean distribution.

More importantly, with the background of whale conservation, we gained a more comprehensive understanding of the realities of human-wildlife conflict in the marine environment. We came to see that such ecological challenges are not only biological in nature but also closely linked to regulations, technology, and human behavior. This experience made us recognize that working with environmental data is not only a technical task, but may also become a responsibility—one that requires us to communicate our findings clearly and responsibly to stakeholders, in the hope of supporting conservation efforts in a meaningful way.

8.2. Limitations

In this project, there are some limitations, mainly in terms of methodology, data sources, and resources.

First, this project did not use more complicated machine learning models or biological modeling methods for data analysis and prediction. This limited our ability to mine the data for deeper patterns. Traditional data analysis methods can provide some insight, but more advanced techniques, such as deep learning or population dynamics models, could have more accurately predicted changes in whale behavior and activity and revealed more complex ecological relationships. Additionally, the project relies heavily on publicly

available data from various countries. Due to different levels of motivation in whale conservation and vessel activity monitoring across countries, some provide exhaustive data, while others have relatively scarce or incomplete data. This unevenness in data sources led to limitations in our analysis, especially when making global comparisons, where missing or incomplete data in certain regions impacted the generalizability and accuracy of the results.

Moreover, the time and resources available for the project also affected the results. Due to time constraints, the team did not have sufficient opportunity to conduct more in-depth background research and expert interviews, leading to a less comprehensive understanding of whale conservation and marine ecosystems. If more time had been available for literature review, expert interviews, and obtaining additional relevant data, the depth and accuracy of our analysis could have been improved. We also lacked key datasets, such as marine group data. This data gap prevented us from presenting a complete picture of the ecosystem's complexity, which may further limit the accuracy and reliability of our conclusions.

9. Conclusion

This report has delivered a research-informed assessment of the IWC's marine incident datasets, providing insights and practical recommendations for improving data integration, standardisation and reusability. The project addressed four key objectives that align with the IWC's long-term goals concerning data management and coordinated future strategies.

An extensive review of the dataset provided helped build an understanding of the structure, completeness, and integration potential of key data. Common issues found included (but were not limited to) poor formatting, lack of metadata and incompatible schemas. This review helped project members understand the scope of the task at hand before moving on to the design of a relational database model. The proposed unified model with an ETL pipeline offered modular and scalable solutions to these problems, with its unique infrastructure increasing data quality and providing greater analytical potential across previously separated datasets.

The exploration of data visualisation tools helped further enhance decision-making capabilities and possible engagement. Combining Power BI with real data demonstrated how an interactive dashboard can help identify hotspots, temporal trends, and increase data accessibility to internal stakeholders and the public. These outputs also reflect best practices highlighted in the literature review, which emphasise the importance of clear visual framing and data accessibility in environmental policymaking.

As objectives were completed, pathways for future improvement were unveiled, with strategic recommendations made to address gaps in protocol adherence and general data practices. Although constrained by the lack of implementation scope and availability of datasets, the project demonstrates foundational steps the IWC can take to transform their currently flawed system into something coherent, queryable and insightful. With modest investment into infrastructure, documentation and training, there is potential to significantly enhance support of the IWC's mission to protect cetaceans through informed, data-driven action.

References

Ariño, A. H., Chavan, V., & Otegui, J. (2016). Best practice guide for data gap analysis for biodiversity stakeholders. GBIF Secretariat, Copenhagen, Denmark.

Benson, A., et al. (2018). 'Integrated Observations and Informatics Improve Understanding of Changing Marine Ecosystems', *Frontiers in Marine Science*, 5(428). DOI: 10.3389/fmars.2018.00428.

Chapman, A.D. (2005). 'Principles of Data Quality', *Global Biodiversity Information Facility*. DOI: 10.15468/doc.jrgg-a190.

El Akkaoui, Z., Vaisman, A.A. and Zimányi, E. (2019). 'A Quality-based ETL Design Evaluation Framework', *The 21st International Conference on Enterprise Information Systems (ICEIS),* pp. 249-257. DOI: 10.5220/0007786502490257.

Hardisty, et al. (2013). 'A decadal view of biodiversity informatics: challenges and priorities'. *BMC Ecology*, 13(16). DOI: 10.1186/1472-6785-13-16.

IBM. (2010). *The IBM data warehousing architecture*. IBM Redbooks. Available at: https://www.redbooks.ibm.com/redbooks/pdfs/sg247138.pdf. (Accessed: 1 May 2025)

International Organization for Standardization (ISO). (2023). *ISO/IEC 11179-1:2023 — Information technology —Metadata registries (MDR) — Part 1: Framework*. ISO. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:11179:-1:ed-4:v1:en. (Accessed: 2 May 2025)

International Whaling Commission (n.d.). *International Whaling Commission*. Available at: https://iwc.int/en/. (Accessed: 1 May 2025)

Ison, S., et al. (2024). 'The role of visual framing in marine conservation communication'. *Ocean & Coastal Management*, 248. DOI: 10.1016/j.ocecoaman.2023.106938.

Kimball, R. and Caserta, J. (2004). *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley.

Liu, L. (2001). 'A case of data warehousing project management'. *Decision Support Systems*, 31(1), pp. 3-20. DOI: 10.1016/S0378-7206(01)00137-9.

McInerny, G.J., et al. (2014). ‘Information visualisation for science and policy: engaging users and avoiding bias’. *Trends in Ecology & Evolution*, 29(3), pp.148-157. DOI: 10.1016/j.tree.2014.01.003.

Michener, W.K. and Jones, M.B. (2012). ‘Ecoinformatics: Supporting ecology as a data-intensive science’, *Ecological and evolutionary informatics*, 27(2), pp. 85–93. DOI: 10.1016/j.tree.2011.11.016.

Morris, P. J. (2005). ‘Relational Database Design and Implementation for Biodiversity Informatics’, *The Academy of Natural Sciences*, 7, pp. 1-16. Available at: https://www.athro.com/general/Phyloinformatics_7_85x11.pdf. (Accessed: 2 May 2025)

Moudrý, V. and Devillers, R. (2020). ‘Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data’. *Ecological Informatics*, 56. DOI: 10.1016/j.ecoinf.2020.101051.

Novotný, P. and Wild, J. (2024). ‘The relational modeling of hierarchical data in biodiversity databases’. *Database*, 2024. DOI: 10.1093/database/baae107.

Williams, R., et al. (2011). ‘Underestimating the damage: interpreting cetacean carcass recoveries in the context of the Deepwater Horizon/BP incident’, *Conservation Letters*, 4(3), pp. 228-233. DOI: 10.1111/j.1755-263X.2011.00168.x.

Zimmer, A. and Meyer, K. (2023). From pixels to policy: How maps inform climate and conservation decisions. *The Pew Charitable Trusts*. Available at: https://www.pewtrusts.org/en/research-and-analysis/articles/2023/05/25/from-pixels-to-policy-how-maps-inform-climate-and-conservation-decisions. (Accessed: 2 May 2025)

Appendix

GitHub Repository: https://github.com/Rachel-XMR/Data-Model-for-IWC-Data-Holdings