# Building a Queryable Image Bank for Tracking Marine Health

**Authors:**
Jonas Mortelmans, Science officer, Flanders Marine Institute
Dias Bakeev, Analyst Developer, Flanders Marine Institute

Zooplankton and phytoplankton (Fig. 1) are two key groups of small aquatic organisms that play important roles in shallow coastal waters. Phytoplankton, consisting of microscopic algae, are primary producers that form the base of the food web in these ecosystems. Zooplankton, a diverse group of small animals, feed on phytoplankton and, in turn, serve as a food source for larger aquatic species. The dynamics of phytoplankton and zooplankton communities are closely linked to the health and functioning of shallow coastal waters and are sensitive to environmental pressures such as climate change, pollution, and habitat degradation.

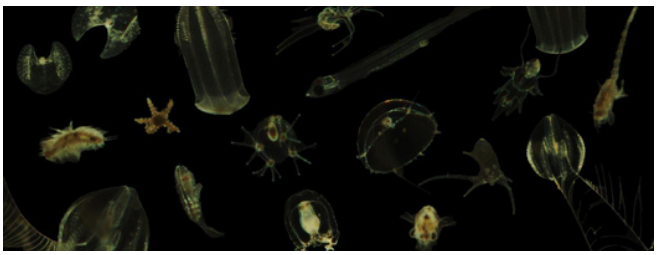


**Figure 1: typical zooplankton encountered in the Belgian part of the North Sea.**

**“The dynamics of phytoplankton and zooplankton communities are closely linked to the health and functioning of shallow coastal waters.”**

Tracking changes in phytoplankton and zooplankton populations and communities provides valuable insight into the impacts of ongoing environmental pressures and the effectiveness of conservation and management efforts. As such, the study of phytoplankton and zooplankton is crucial for understanding and addressing the challenges facing shallow coastal waters.

To keep track of these changes, the Marine Observation Centre (www.vliz.be/en/marine-observation-centre) is developing, deploying, and optimizing cost-effective, innovative, and integrated observation systems. These systems consist of a network of bio-sensors at sea and in the coastal area, in combination with monthly and seasonal multi-disciplinary measurement campaigns aboard the RV Simon Stevin in the Belgian part of the North Sea.

One sensor used for biodiversity measurements is the Video Plankton Recorder (VPR, Fig. 2), which captures underwater images of the encountered plankton.

**Figure 2: the Video Plankton Recorder (VPR) ready for deployment behind the Research Vessel (RV) Simon Stevin.**

The strength of this technique lies in the real-time data collection aboard the research vessel while sailing. Particles are photographed in the water in their natural environment. Simultaneously, precise field measurements (e.g., seawater temperature, salinity, turbidity, pressure), essential measurements of the depicted particles (e.g., length, height, circularity), and classification data (e.g., name of the particle that has been photographed) are stored alongside the photo image in associated fields. This yields an especially robust dataset important for both marine biologists and oceanographers.

**"6,087,636 images are stored in the BioSenseMongoDB, each image with over 45 associated fields. Studio 3T allows users to easily query those 6 million images and extract validated biodiversity data for downstream pipelines.”**

In order to easily manage such large image datasets, both the images and their associated metadata are stored in BioSenseMongoDB, a NoSQL database. MongoDB is well-suited for storing image data due to its flexible and scalable document-store data model. It can handle large amounts of unstructured data, including binary image data, without the need to define a schema beforehand. Additionally, MongoDB's efficient storage and retrieval capabilities, as well as its built-in support for handling GridFS (a specification for storing and retrieving files that exceed the BSON-document size limit of 16 MB), make it an ideal choice for large image library data storage. The built-in GridFS View provides a simple way to access the images directly. As of now, 6,087,636 images have been collected by the VPR and are stored in BioSenseMongoDB, with each image associated with 28 fields containing sample information, 7 fields of image-associated data, 6 fields of classification data, and 5 fields containing environmental information.

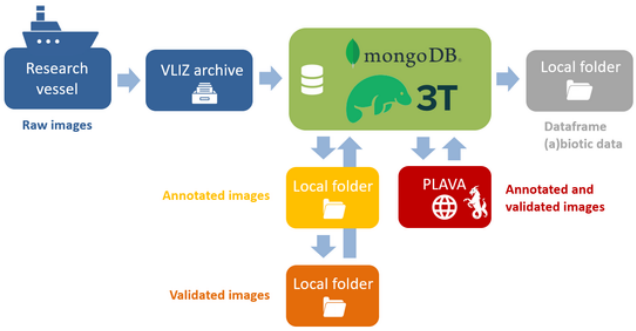



**Fig 3. Simplified data flow as used for the Video Plankton Recorder (VPR)**

A VPR Python tool and Symfony user interface (MongoDB Uploader Tool, MUT) were developed to allow scientists to easily import and access data from the VLIZ archive (see Fig. 3) into BioSenseMongoDB.

Image recognition algorithms are then run on data in BioSenseMongoDB, which will predict the taxonomic name for each image.

Moreover, an in-house tool was developed for validation of the collected image data (the Plankton Validation Tool; PlaVa). This tool, PlaVa, is a Graphical User Interface (GUI) for image collections from the BioSenseMongoDB server. It allows for easy and structured labeling/validation of images in the database, without the need for coding or downloading the images locally. Any change in fields will directly be stored again in BioSenseMongoDB.

Studio 3T's Aggregation Editor then allows users to easily query BioSenseMongoDB and extract validated biodiversity data for downstream pipelines. Studio 3T is useful for reviewing and managing the image data in BioSenseMongoDB for quality control and IT-developments.

The Reschema tool also proves invaluable in resolving challenges related to MongoDB schema updates, particularly in the context of metadata field names.

When modifications are required, the tool provides a systematic solution, preventing inconsistencies that might arise during the update process.

By helping enforce a structured approach to schema changes, Reschema minimizes the risk of data discrepancies and ensures the seamless evolution of the database. This capability not only simplifies the task of updating metadata field names but also plays a crucial role in maintaining data integrity, ultimately contributing to a more robust and reliable database.

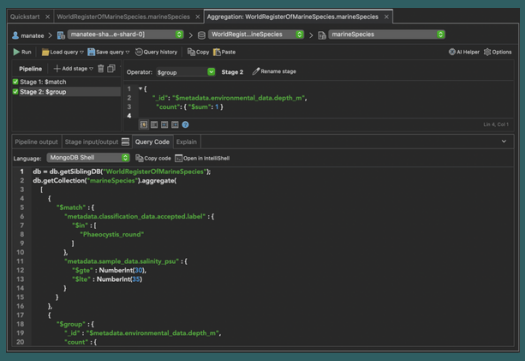
The Reschema tool thus emerges as a key resource for addressing schema-related issues and streamlining the data.

More information:
www.lifewatch.be
https://www.marinespecies.org/
www.vliz.be
www.studio3t.com

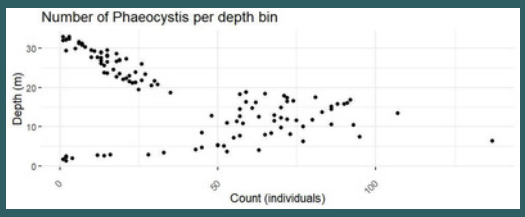
**Example**
Scientists typically query BioSenseMongoDB because they want to know:

**"I want to know the number of specimens, of a specific species [e.g., Phaeocystis], recorded in specific conditions [e.g., a specific month, specific environmental conditions, ...]; and at what depth are these recorded?"**

We can then retrieve this information with an aggregation in Studio 3T.



This two-stage aggregation yields a data frame with occurrences per depth bin, which are then also easily sent to downstream pipelines for visualisation (in this case we used Rstudio).




**Conclusion**

In conclusion, the integration of Studio 3T into the management and analysis of the BioSenseMongoDB database has significantly improved the efficiency and reliability of handling large-scale image datasets related to plankton populations.
The multiple querying capabilities of Studio 3T enable researchers to efficiently extract validated biodiversity data for downstream pipelines, enhancing the overall data analysis process. The Reschema tool proves to be instrumental in maintaining data integrity during MongoDB schema updates, specifically addressing challenges related to metadata field names. While precise metrics on time savings may vary, the streamlined processes facilitated by these tools undoubtedly contribute to increased productivity.
This case study exemplifies how the combination of innovative observation systems, sophisticated database management, and powerful querying tools plays a pivotal role in advancing marine biology research, providing valuable insights into the dynamics of plankton communities and their responses to environmental pressures.