Targeting microbes in the digital ecosystem

Report of Gaétane Magali Sallard
In the context of the Marmic Master Program, Lab Rotation III – Max Planck Institute of Marine Microbiology

During this lab rotation, our goals were to target microbes in the digital ecosystem of ODIS and to explore how datasets are interconnected and what these connections reveal about the data landscape. Additionally, we aimed to identify any data gaps and assess the geographic areas and time periods covered by the datasets. This report clarifies the workflow of the lab rotation and the steps taken to achieve these goals. This analysis was completed without prior knowledge of SPARQL, JSON-LD, or strong foundations in Python coding. The code was developed with the assistance of ChatGPT (v. 4.0), and all computational commands were executed on a local server. All codes can be found on the GitHub page, classified as in this document. All datasets can be found on Zenodo (see references).

1. Harvesting microbial terms

We identified specific microbial terms within the data landscape in the initial phase to accurately target microbial data.

Bacterial and archaean taxa from the SILVA database

As a first attempt to target microbially related data, we retrieved all the child clauses of bacterial and archaeal taxa from the SILVA database , that is all headers from the sequences of both domains of life. We saved this dataset in form of a textfile and curated the headers as follows:

- The species name's epithet was removed, and any duplicate generic names were eliminated.
- We removed non-alphabetic characters from the list, such as /*-_ and numbers
- Additionally, we excluded names shorter than four characters and those containing more than two consecutive identical letters.

Even after organizing the dataset[1], there were still some taxonomic names that were not specific to microbes. These names might have been used as acronyms for specific categories (e.g. "unio" for "uniola" or "unionicola"), as species identifiers like "urhd," or as part of the species name, such as "blood," "texas," “blueberry,” “antarctic” or even created purely for creative purposes by the namegiver, such as "unicorn".

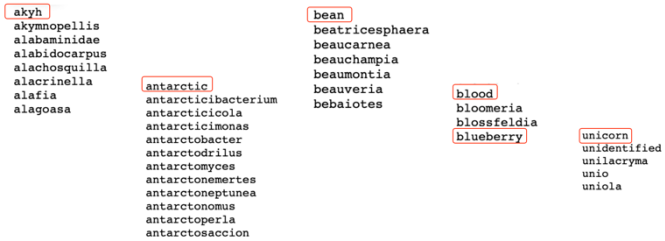


Figure 1: Extract from Sallard, G. M. (2024). SILVA database - all extracted bacterial and archaeal taxa (curated). The word unicorn appears to be a creative input of a taxonomic namegiver.

It is also worth mentioning that the complete dataset of all these names renders 31'703 entries.

Referring to the most common marine microbes in the SILVA database

Given the lack of specificity of certain words and the large number of term entries, it is impractical to manually review the entire dataset and individually remove non-specific words. Instead, we referred to the most common marine microbes as outlined in Figure 2 from Overmann, J., Lepleux, C. (2016)[2]. We then extracted specific microbial child clauses from the SILVA database and applied the same filtering rules as before. This reduced the entry list to 272 terms, which we could also visually check for non-specific taxonomic names (and found none)[3].

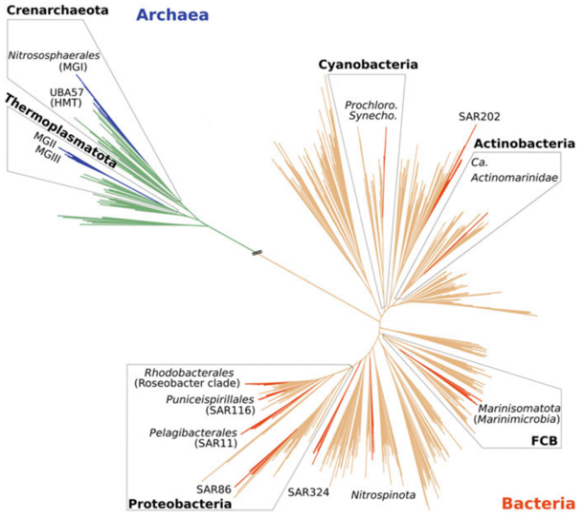


Figure 2: Most common microbial taxa in the marine environment. Overmann, J., Lepleux, C. (2016)[2]

```
Synechococcaceae
Synechococcales
Synechococcus
Synechocystis
Thermogymnomonas
Thermoplasma
Thermoplasmata
Thermoplasmataceae
Thermoplasmatales
Thermoplasmatota
Thermosynechococcaceae
Thermosynechococcales
Thermosynechococcus
Thysanoptera
Tolypothrix
Tracheophyta
Trichocoleus
Trichodesmium
Trichormus
Tychonema
Vampirivibrionia
Vampirovibrio
Vampirovibrionaceae
Vampirovibrionales
Wilmottia
Xenococcaceae
Xenococcus
```

Figure 3: Extract from Sallard, G. Predominant Marine Microbial Taxa extracted from the SILVA Database

Building a microbial term index

In order to have a more general approach to target microbially related data in the ODIS graph, we decided to experiment with building a microbial term index. Our initial attempt involved using spaCy (https://spacy.io), a Python-based NLP language processing tool, and a predefined list of microbiology-related terms. The seed list of terms we used included 10 words: "microbe", "bacteria", "bacterium", "bacillus", "microflora", "microbial", "prokaryote", "protist", "archaea", "microorganism".

Using paper abstracts

We initially applied this method to various microbiology paper abstracts. Using spaCy, we processed the provided paper abstracts or chapters, breaking them into individual words. Then, each word was assigned different linguistic features, such as whether it was alphabetic, a stop word, or a part-of-speech tag. These terms were then filtered as follows to create a word list:

- The word had to be purely alphabetic
- The word must not be a common stop word
- The word had to be longer than three characters

SpaCy compared the generated word list to a set of word vectors that capture semantic similarity from the initial seed list. If a word matched one of these terms, it was considered relevant for the index entry. This approach typically yielded no more than seven terms, such as "archaea," "bacillus," "bacteria," "bacterium," "microbial," "microorganisms," and "protists."

Using the Wikipedia database

In the search of expanding our word pool, we turned to the Wikipedia database. Using the Wikipedia API library, we extracted marine microorganism related terms, complying with Wikipedias usage policies. The list of retrieved pages include topics such as "Marine microorganisms", "Marine microbiome", "Marine viruses", "Marine bacteria", "Bacterioplankton", "Bacterial motility", "Marine prokaryotes", "Marine archaea", "Marine protists", "Marine fungi", "Mycoplankton", "Marine microanimals", "Ichthyoplankton", "Marine primary production", "Algae", "Marine microplankton", "Marine microbenthos", "Sea ice microbial communities", "Hydrothermal vent microbial

communities", "deep biosphere", and "microbial dark matter".

We passed the combined text content from these Wikipedia pages into the spaCy model, which then produced a Doc object containing words and their associated linguistic information. After extracting these words from the Doc object, spaCy keeps track of their frequencies using a dictionary and filters out words that are not purely alphabetic, common stop words, and those shorter than four characters. These terms are each converted to lowercase and sorted based on their occurance in the combined text content[4].

Microbial-Related Term	Frequency
marine	697
bacteria	639
algae	343
species	310
organisms	309
ocean	281
cells	254
archaea	253
viruses	251
found	245
microbial	242

Figure 4: Extract from Sallard, G. Extracted microbial terms from Wikipedia - Marine Microbiology related pages (unfiltered)

We reviewed the full dataset[4] and narrowed it down to a subset of the 100 most frequent terms directly related to microbiology[5]. We excluded terms such as "marine", "organisms", "ocean", "life" as they are not strictly related to microbiology. Additionally, terms like "algae" and "plants" were removed since they are not microbial entities. We also avoided including verbs such as "found", "classify", and "form" because they may be related to sentences or methods in microbiology, but not strictly to microbial entities.

bacteria
archaea
viruses
microbial
microorganisms
fungi
bacterial
prokaryotes
protists
phytoplankton
eukaryotes
flagella
cyanobacteria
microbes

Figure 5: Extract Sallard, G. Extracted microbial terms from Wikipedia - Marine Microbiology related pages (refined).

2. Targeting the data landscape

Targeting the big picture: The ODIS data landscape

Using a SPARQL extension in Python, we constructed a query to search for datasets containing any of the specified marine microbiology terms from the created index in the dataset keywords. The query was sent to the ODIS endpoint, and the results were returned in JSON format. We processed the JSON response to extract relevant information for each dataset, including the subject (URL), name, description, and keywords. Out of the large dataset retrieved (565,676 nodes), we randomly selected fifty nodes to create a more manageable subset for further analysis. This subset was then converted into a Pandas data frame and saved as a CSV file to facilitate subsequent examination. Upon reviewing the random sample of 50 nodes, we found that many datasets were not directly related to marine microbiology. For example, a dataset focused on monitoring lobster egg development was included (https://catalogue.ogsl.ca/en/dataset/ca-cioos_3baf0a9e-de19-42b4-bcf7-1d51369333bd), which does not pertain directly to microbial data. Additionally, the keyword matching revealed inconsistencies. Some datasets contained keywords like "sea_water_electrical_conductivity" and "migrating birds," which were not part of the original terms extracted from Wikipedia.

This discrepancy is due to the regex function in the SPARQL query, which enables partial and flexible matching of terms within the keywords. Any keyword containing a substring matching any of the specified terms will be included, even if it may not be strictly relevant.
For example, if the term "algae" is specified, the regex function will match keywords such as "sea_algae," "freshwater_algae," or even unrelated terms where "algae" appears as a substring. The "i" flag in the regex function makes the matching case-insensitive, further broadening the scope of matches. Additionally, the use of regular expressions with an OR operator (|) means that any keyword matching any part of the concatenated terms will be included. This broad matching criterion can lead to the inclusion of datasets with keywords that are related but do not exactly match the intended focus, such as "sea_water_electrical_conductivity."

Targeting the ocean biodiversity data landscape OBIS

We opted to use the OBIS (Ocean Biogeographic Information System) parquet file as a Pandas DataFrame for our data analysis instead of querying the entire ODIS endpoint (http://ossapi.oceaninfohub.org/public/assets/obis.parquet), as it is less broad and less compute intensive. OBIS is a comprehensive database with over 20 nodes worldwide, connecting 500 institutions from 56 countries. It contains more than 45 million observations of nearly 120,000 marine species, including bacteria. The integrated datasets allow to search and map data by species name, higher taxonomic level, geographic area, depth, time, and environmental parameters.

However, the OBIS parquet file only contains columns for unique identifiers (URI), type of entry (type), names (name), descriptions (desc), keywords (keywords), and providers (provider). We filtered the parquet file to only include entries of type 'dataset' and checked for the presence of the specific marine microbiology terms in either the keyword or description fields. This approach avoids the issues of broad and over-inclusive matching that we previously encountered when querying the ODIS endpoint.

3. Buiding and visualizing the directed graph

We used networkx for Python (https://networkx.org) to create a directed graph to visualize the connections between the filtered datasets based on matched terms. To keep the computational load manageable, we focused our analysis on the first 300 entries of the dataset. In the graph, each dataset was represented by a node with its unique URI, and we included additional properties such as the dataset name and matched terms. We established edges between the nodes using a function that linked datasets based on the co-occurrence of matched terms. The strength of these edges was determined by the frequency of term co-occurrence, meaning that datasets sharing more terms had stronger connections. Finally, we exported the graph to a GEXF file for further processing in Gephi.

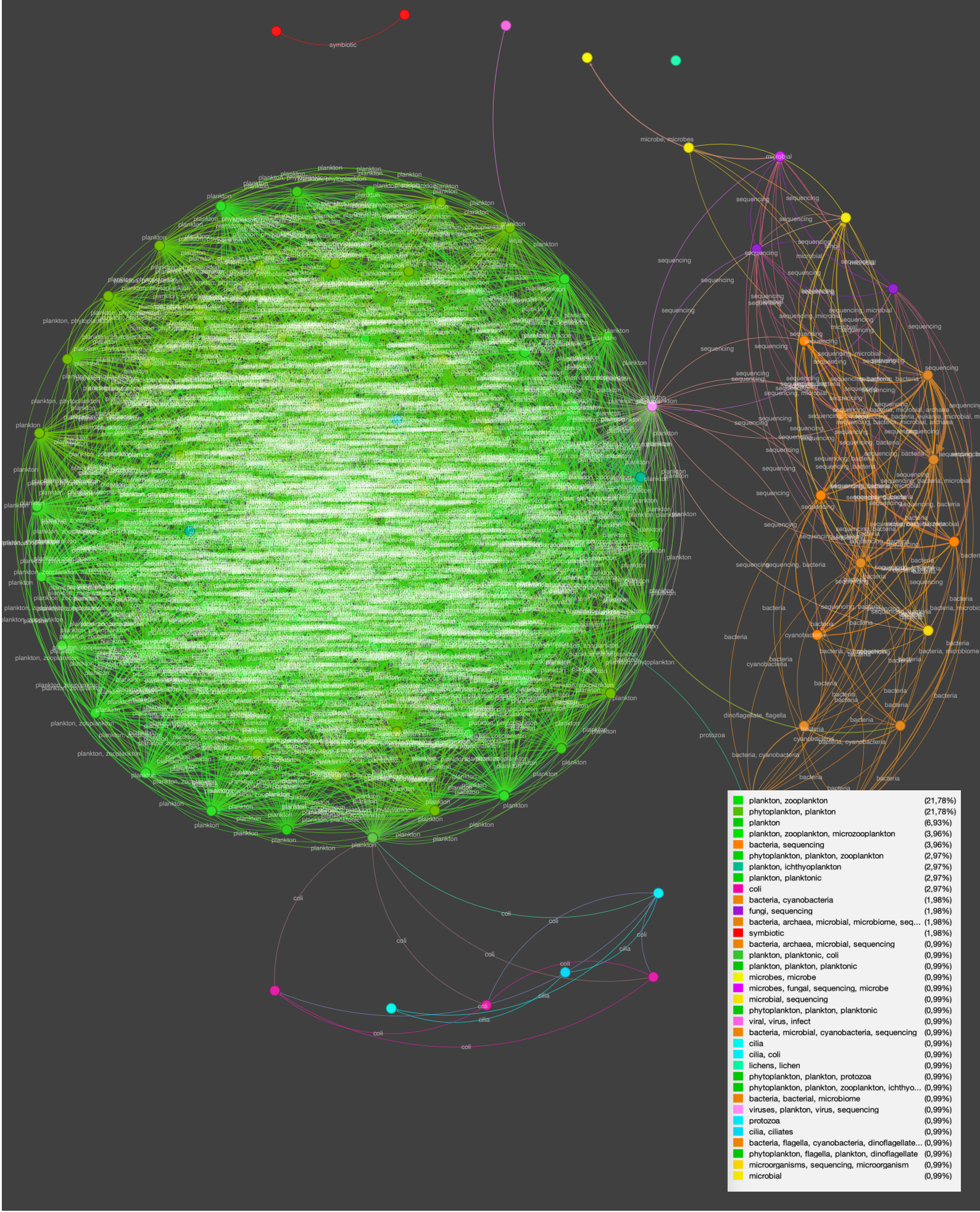
node: @id
node properties: @name @matched terms

edge: function that links datasets based on co-occurence of matched terms

strength of edges: determined by frequency of term co-occurence
e.g. 1 term co-occurs between datasets = weight 1.0; 2 terms = weight 2.0…

In Gephi (v.0.10.1, https://gephi.org), we adjusted the node colors to match similar terms. For example, green nodes represent plankton-related terms such as phytoplankton, zooplankton, planktonic, and ichthyoplankton. Virus or pathogenic-related terms were assigned a pink to purple color, microbial-related terms were represented in yellow, and ciliates and protozoans were denoted in blue, while bacteria were indicated in orange. The edges in the graph are partitioned based on the co-occurrence of matched terms between nodes and labeled accordingly. The weight of the arrows is determined by the frequency of these occurrences, with stronger connections represented by stronger arrows. Additionally, we used the Fruchterman-Reingold algorithm, a force-directed layout, to position the nodes of the graph such that all the edges are of more or less equal length and the graph structure is easy to perceive. The algorithm mimics a physical system in which nodes repel each other, similar to electrically charged particles. Meanwhile, edges function as springs, drawing connected nodes together.

Microbes with social relevance are the first to connect to OBIS



The graph reveals a large amount of data concerning plankton in the Ocean Biogeographic Information System (OBIS), as well as extensive bacterial sequencing data and Escherichia coli (E. coli). However, there is a noticeable lack of data for archaea, fungi, and symbiotic associations. Most of the data about plankton comes from studies on eutrophication and fisheries. This suggests the influence of human activities, such as fishing on data sharing a communicating with the industry and to OBIS. International regulations and non-governmental organizations (NGOs) may be accentuating the need to share ecological data to meet sustainable development goals (SDG 14: https://www.globalgoals.org/goals/14-life-below-water/). The study of pathogens, particularly E. coli, follows a similar pattern and is significant for human health and the microbiome. The abundance of E. coli data highlights its importance to social issues such as sewage leakages.
The insufficient data on archaea, fungi, and symbiotic associations may indicate a lack of communication and transparency in these areas of research. Archaea, often found in extreme environments, have unique metabolic pathways and biochemical properties that could offer insights into biotechnology and evolutionary biology. Similarly, studies on fungi and symbiotic associations are crucial for understanding ecosystem dynamics, yet there appears to be a lack of data flow in this area.
Overall, the first microbes to connect to OBIS seem to be those related to direct social relevance.

4. Pinpoint dataset origins

We wanted to further explore the extensive phytoplankton data and pinpoint its origins. However, as the OBIS parquet file (http://ossapi.oceaninfohub.org/public/assets/obis.parquet) lacks geographic coordinates, we had to find alternative methods to identify its sources. To do this, we filtered our nodes to include only those with terms related to phytoplankton. Then, we used the ODIS data mapper (https://mapper.obis.org) to pinpoint the origins of the datasets.

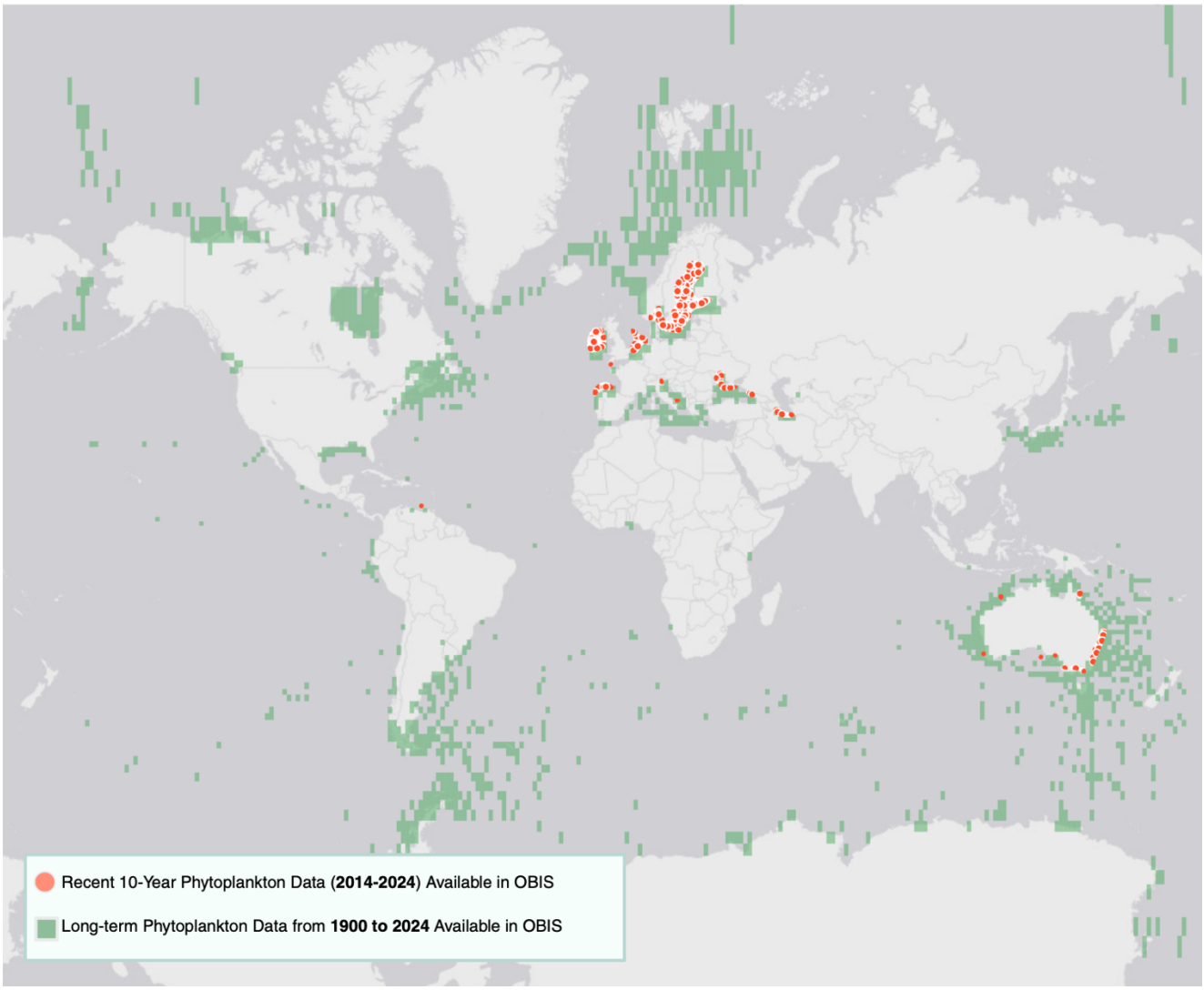


Figure 6: Long-term and short term phytoplankton data communicated to OBIS

The map shows green areas with long-term phytoplankton data from 1900 to 2024 available in OBIS across the globe. The red points indicate recent data collected over the past ten years. It's intriguing to note that China, Indonesia, and India, the top three fishing countries and largest fish producers globally, are not prominently featured on the map. This absence could be due to various factors like differences in national research priorities and data sharing policies.

Japan makes significant contributions to data collection and communication, particularly around the island of Kyushu and the Izu Islands, as part of its extensive maritime research programs like JAMSTEC (Japan Agency for Marine-Earth Science and Technology). Additionally, Peru is renowned for producing large amounts of fish meal and fish oil, and its northern coast also gathers important datasets related to phytoplankton.

A surprising contributor to phytoplankton data is Australia. Despite not being a major fishing nation, it covers a significant area of phytoplankton-related waters. This could be due to its strong emphasis on biodiversity research, driven by the country's academic institutions and environmental agencies such as the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Australian Institute of Marine Science (AIMS).

Europe has comprehensive data coverage in the Mediterranean and Baltic Seas, largely due to collaborative efforts within the European Union. Programs such as Horizon Europe support marine research and sustainable marine management initiatives. The EU prioritizes marine biodiversity, pollution control, and sustainable fisheries, resulting in a valuable dataset that supports scientific research and policy-making.

What is concerning is that when we analyze the data measured and available over the last ten years, we can only find significant datasets from the European Union, Australia, and a single point on Curaçao.

The European Union has put in place strict marine environmental policies through the Marine Strategy Framework Directive (MSFD). This directive requires member states to ensure their marine waters achieve "Good Environmental Status". It involves thorough monitoring and data collection on marine biodiversity, including phytoplankton, to assess and manage the health of marine ecosystems (OECD).
Similarly, The Integrated Marine Observing System (IMOS) in Australia is a national initiative supported by the Australian government. It provides comprehensive and sustained monitoring of the marine environment. IMOS collects, manages, and provides access to marine data, including phytoplankton observations, to support scientific research and policy development (CSIRO). On the other hand, the coastal areas of Japan were greatly affected by the 2011 earthquake and tsunami. This may have led to a shift of resources from environmental monitoring to reconstruction and disaster management. Several South American countries, including Peru and Argentina, may not have the required funding and infrastructure to sustain continuous long-term phytoplankton monitoring programs. These nations may be prioritizing terrestrial environmental issues, such as deforestation and agricultural impacts, over marine research. In the United States, there are extensive marine research programs. However, these agencies such as NOAA have their own data flow and are not obligated to upload to OBIS.

5. Further improvements

We successfully met the primary objective of focusing on microorganisms in the digital environment of OBIS. We uncovered significant insights about the data environment and the information being transmitted to OBIS. However, when creating graphs and focusing on specific terms from our microbial index, we need to take into account the data quality from our partners. If a partner has a detailed description for their dataset, it is more probable that we will find a match with one of the terms in the microbial term index. This highlights the need for standardization in the transfer and communication of data.
We should also consider reducing the significance of closely associated terms like zooplankton and phytoplankton, as their connection leads them to frequently appear together and clutter up the graph. In addition, we should utilize graph analysis tools to delve into the graph structure. This involves using module detection tools to identify strongly interconnected clusters and nodes with the highest and lowest degrees, as well as the most isolated ones. Furthermore, the integration of spatial data in the OBIS parquet file should be prioritized for improvement.

It seems that there are significant gaps in our knowledge regarding the location of data. OBIS does not have a global agreement that mandates data to be uploaded to its servers. However, we could reach out to agencies where data gaps have been identified and encourage them to link to their data servers, thereby making their data known and accessible.

6. References

1. Sallard, G. SILVA database - all extracted bacterial and archaeal taxa (unfiltered). at https://doi.org/10.5281/zenodo.12653481 (2024).
2. Overmann, J. & Lepleux, C. Marine Bacteria and Archaea: Diversity, Adaptations, and Culturability. in *The Marine Microbiome* 21–55 (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-33000-6_2.
3. Sallard, G. Predominant Marine Microbial Taxa extracted from the SILVA Database. at https://doi.org/10.5281/zenodo.12653939 (2024).
4. Sallard, G. Extracted microbial terms from Wikipedia - Marine Microbiology related pages (unfiltered). (2024) doi:10.5281/zenodo.12571309.
5. Sallard, G. Extracted microbial terms from Wikipedia - Marine Microbiology related pages (refined). at https://doi.org/10.5281/zenodo.12651803 (2024).