Targeting microbes in the digital ecosystem

Gaétane Sallard

In context of the lab-rotation III marmic - Max Planck Institute of Marine Microbiology

Goals

Targeting microbes in digital ecosystem of OBIS

- **How are datasets interlinked and what do the connections reveal about the data landscape?**
- Identify data gaps
- geographic focus areas and temporal coverage of datasets

1. Harvest microbial terms

2. Targeting data landscape

3. Building and visualising directed graph

4. Pinpoint dataset origins

1. Harvest microbial terms

Specific approach

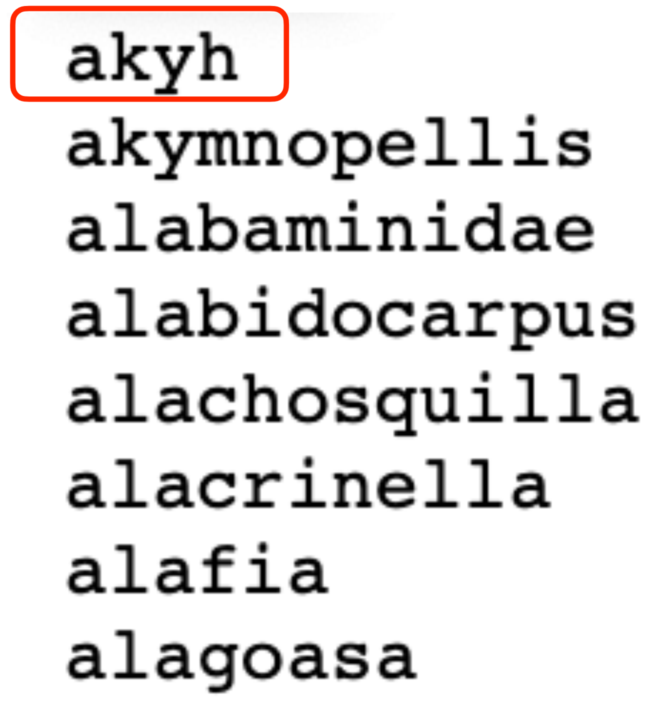
Retrieving all child clauses of the bacterial and archaean taxa from the SILVA database

- Retrieved all headers from the sequences of both bacterial and archaean domains
  - species epithet was removed and any duplicate generic names were eliminated
  - removed non-alphabetical characters from the list such as /*- and numbers
  - excluded names shorter than 4 characters and those containing more than 2 consecutive identical letters

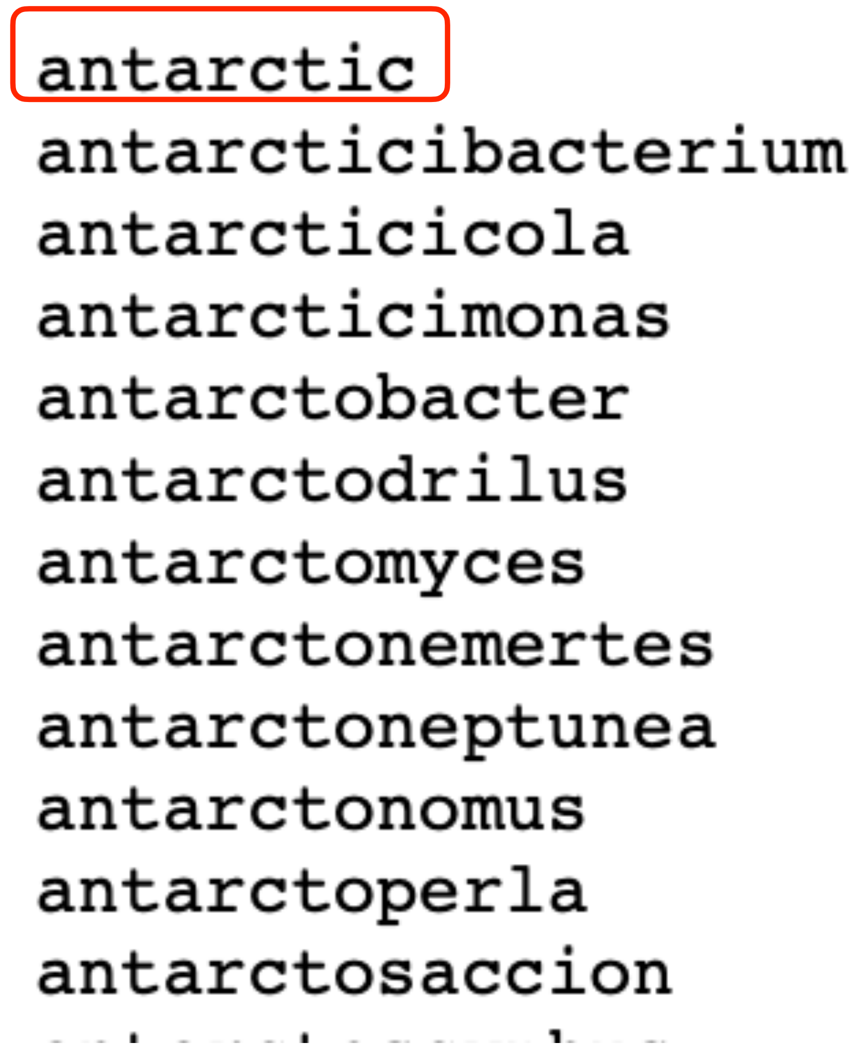
1. Harvest microbial terms

Specific approach

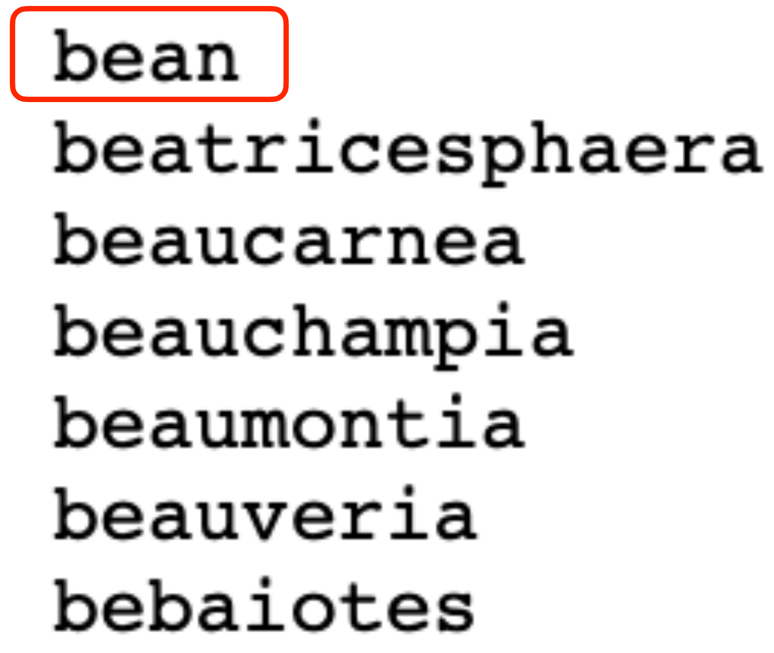
Retrieved taxonomic names were not always specific to microbes



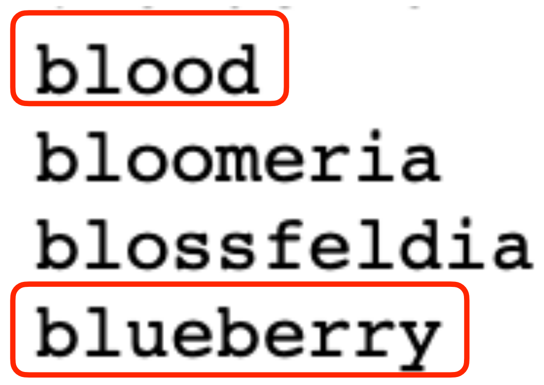
akyh
akymnopellis
alabaminidae
alabidocarpus
alachosquilla
alacrinella
alafia
alagoasa



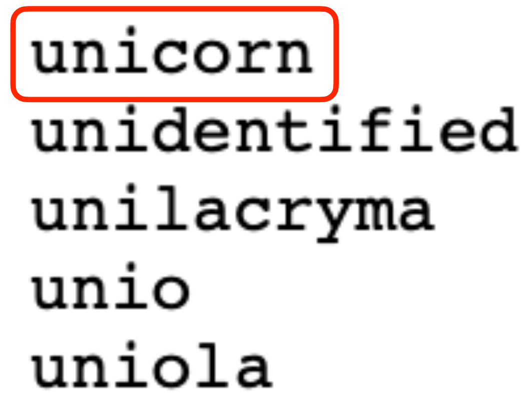
antarctic
antarcticibacterium
antarcticicola
antarcticimonas
antarctobacter
antarctodrilus
antarctomyces
antarctonemertes
antarctoneptunea
antarctonomus
antarctoperla
antarctosaccion



bean
beatricesphaera
beaucarnea
beauchampia
beaumontia
beauveria
bebaiotes



blood
bloomeria
blossfeldia
blueberry



unicorn
unidentified
unilacryma
unio
uniola

31’703 entries

1. Harvest microbial terms

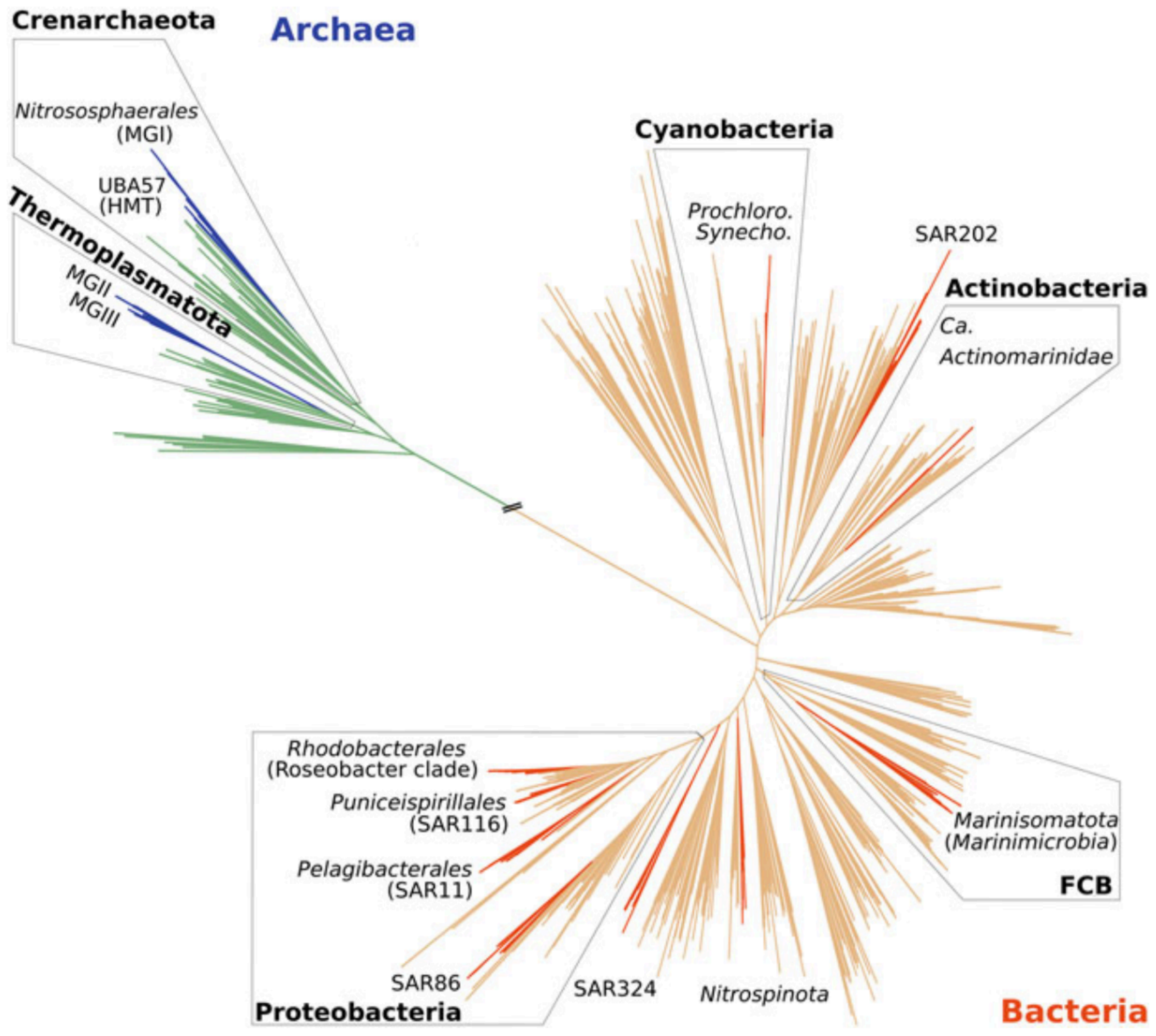
Specific approach

Referring to the most common marine microbes in the SILVA database

- most common marine microbes as described in study from

- Overmann, J. & Lepleux, C. **Marine Bacteria and Archaea: Diversity, Adaptations, and Culturability**. in *The Marine Microbiome* 21–55 (Springer International Publishing, Cham, 2016)

- *Retrieved all headers from the sequences of both bacterial and archaean domains*
  - *species epithet was removed and any duplicate generic names were eliminated*
  - *removed non-alphabetical characters from the list such as /*- and numbers*
  - *excluded names shorter than 4 characters and those containing more than 2 consecutive identical letters*



1. Harvest microbial terms

Specific approach

Narrowing down to specific names of most common marine microbes

```
Abobra
Acanthamoeba
Acaryochloridaceae
Acaryochloris
Acidiplasma
Acidiprofundales
Aciduliprofundaceae
Aciduliprofundum
Acrophormium
Aerosakkonema
Aetokthonos
Aliterella
Alkalinema
Alphaproteobacteria
Alteromonadaceae
Alteromonas
Amoebozoa
Amorphea
Anabaena
Anabaenopsis
Ancylothrix
Annamia
Aphanizomenon
Aphanotece
Archaeplastida
Arthronema
Arthrospira
Atelocyanobacterium
Bilateria
Burkholderiales
```

```
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
```

272 entries

1. Harvest microbial terms

Broad approach

Building a microbial term index using the spaCy for Python NLP tool

- **~~Using paper abstracts and chapters~~**
  - ~~input = seed list of 10 terms~~
  - ~~output = extension of relevant terms~~
  - ~~compares generated word list from abstract/chapter to seed list and checks for semantic similarities~~

~~never more than 7 broadened terms~~

- **Using the Wikipedia API Library**
  - input = combined text from various wikipedia pages related to marine microbiology
  - output = Doc object containing words and their associated linguistic information

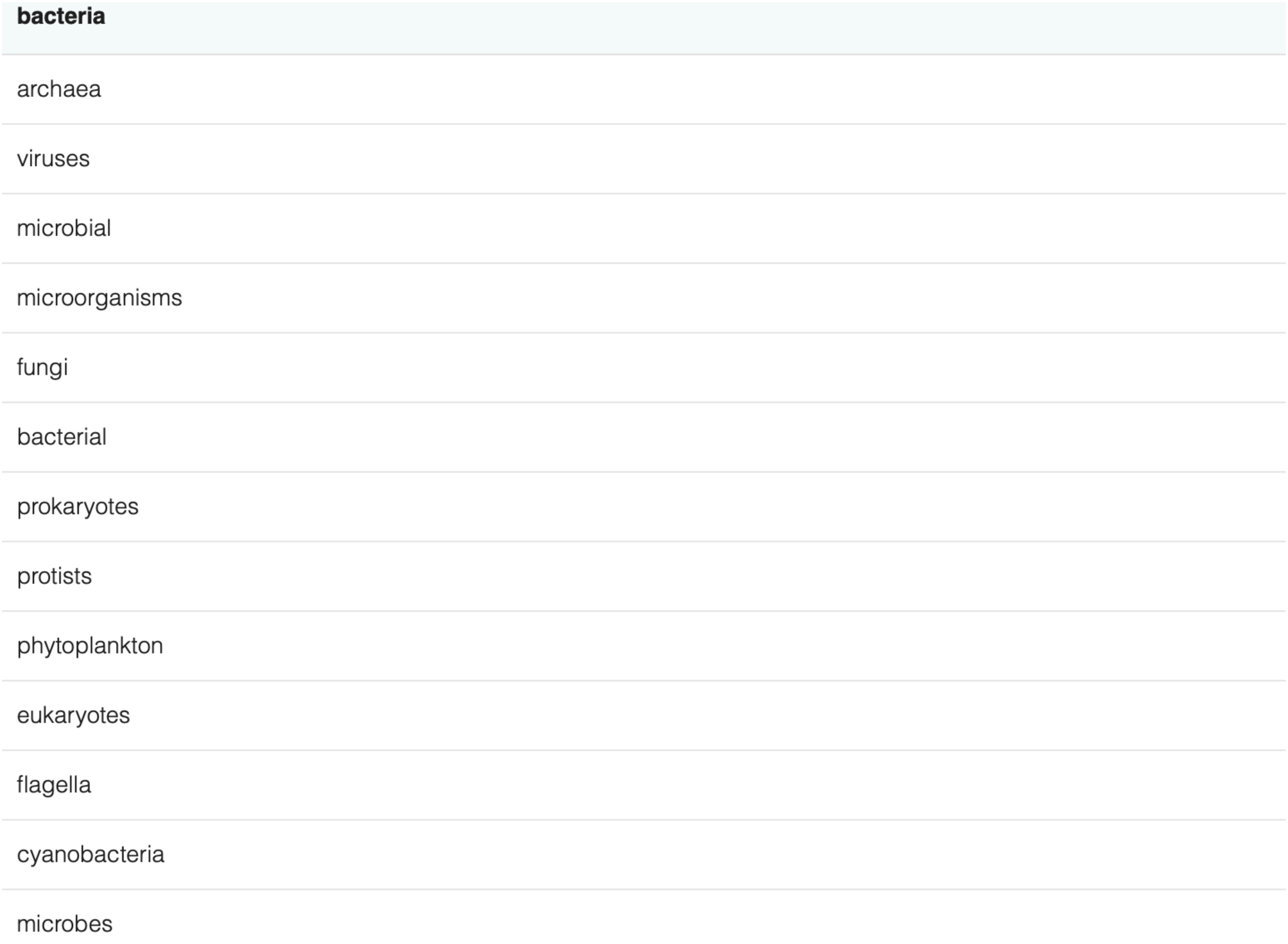
100 terms

1. Harvest microbial terms

Broad approach

Refined list of 100 terms in microbial index

- sorted terms based on frequency in the combined text document
- excluded non-specific terms such as “marine”, “organisms”
- excluded non-specific microbial entities such as “algae”, “plants”
- excluded verbs such as “found”, “classify”

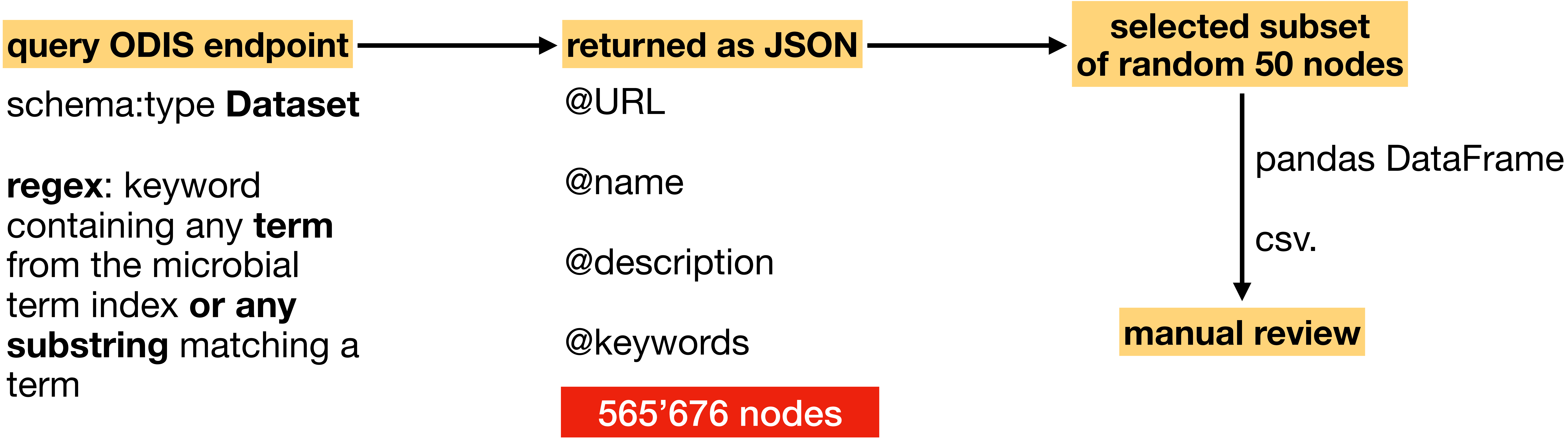


bacteria
archaea
viruses
microbial
microorganisms
fungi
bacterial
prokaryotes
protists
phytoplankton
eukaryotes
flagella
cyanobacteria
microbes

…

Targeting the big picture: the ODIS data landscape

- SPARQL extension for Python to query for @dataset keywords containing terms related to marine microbiology



query ODIS endpoint

schema:type **Dataset**

regex: keyword containing any **term** from the microbial term index **or any substring** matching a term

returned as JSON

@URL

@name

@description

@keywords

565’676 nodes

selected subset of random 50 nodes

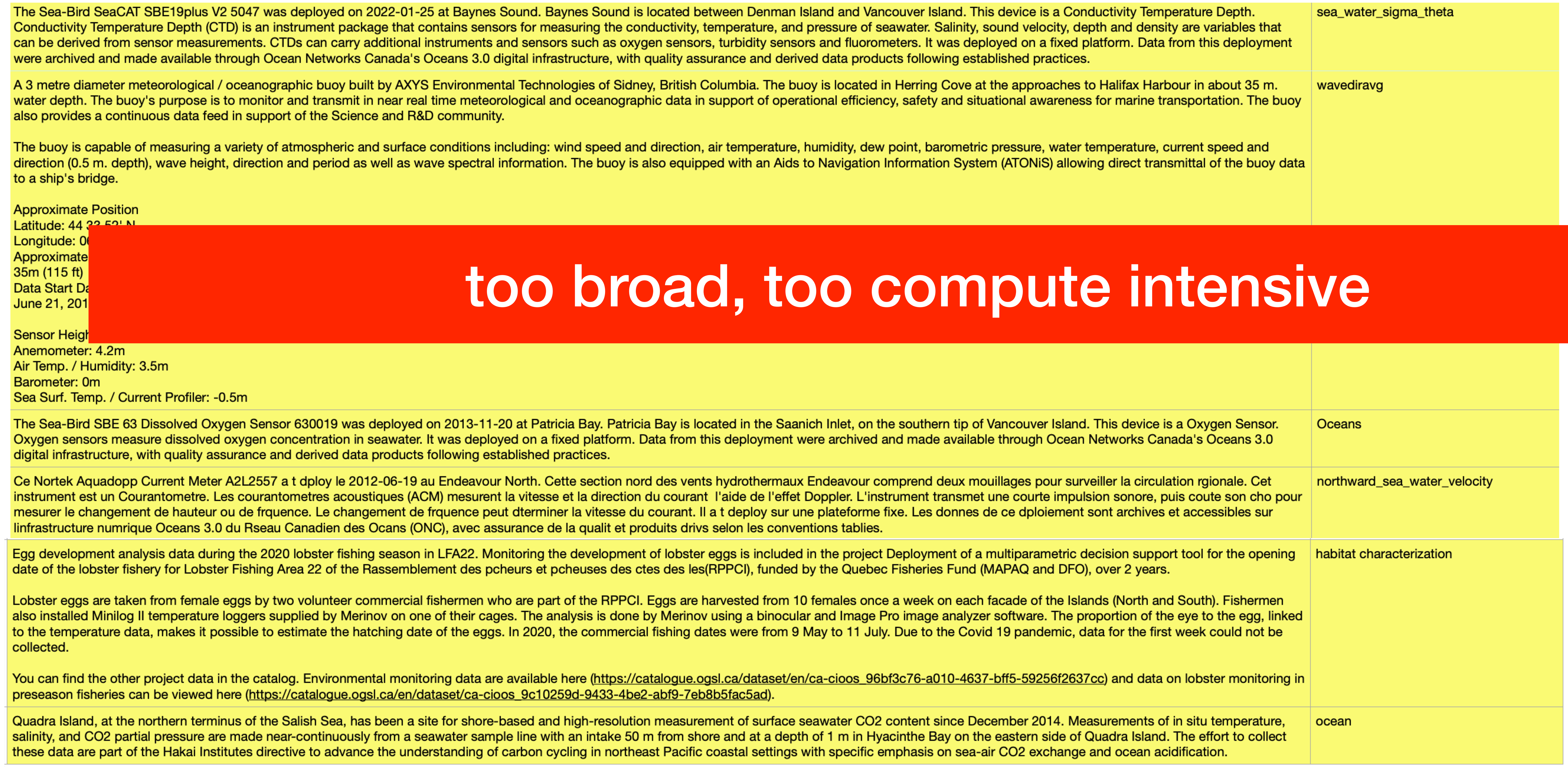
pandas DataFrame

csv.

manual review

2. Targeting data landscape

Manual review revealed the nesting of non-relevant datasets



The Sea-Bird SeaCAT SBE19plus V2 5047 was deployed on 2022-01-25 at Baynes Sound. Baynes Sound is located between Denman Island and Vancouver Island. This device is a Conductivity Temperature Depth. Conductivity Temperature Depth (CTD) is an instrument package that contains sensors for measuring the conductivity, temperature, and pressure of seawater. Salinity, sound velocity, depth and density are variables that can be derived from sensor measurements. CTDs can carry additional instruments and sensors such as oxygen sensors, turbidity sensors and fluorometers. It was deployed on a fixed platform. Data from this deployment were archived and made available through Ocean Networks Canada's Oceans 3.0 digital infrastructure, with quality assurance and derived data products following established practices.	sea_water_sigma_theta
A 3 metre diameter meteorological / oceanographic buoy built by AXYS Environmental Technologies of Sidney, British Columbia. The buoy is located in Herring Cove at the approaches to Halifax Harbour in about 35 m. water depth. The buoy's purpose is to monitor and transmit in near real time meteorological and oceanographic data in support of operational efficiency, safety and situational awareness for marine transportation. The buoy also provides a continuous data feed in support of the Science and R&D community. The buoy is capable of measuring a variety of atmospheric and surface conditions including: wind speed and direction, air temperature, humidity, dew point, barometric pressure, water temperature, current speed and direction (0.5 m. depth), wave height, direction and period as well as wave spectral information. The buoy is also equipped with an Aids to Navigation Information System (ATONiS) allowing direct transmittal of the buoy data to a ship's bridge. Approximate Position Latitude: 44 33.52' N Longitude: 0 Approximate 35m (115 ft) Data Start Da June 21, 201 Sensor Heigh Anemometer: 4.2m Air Temp. / Humidity: 3.5m Barometer: 0m Sea Surf. Temp. / Current Profiler: -0.5m	wavediravg
The Sea-Bird SBE 63 Dissolved Oxygen Sensor 630019 was deployed on 2013-11-20 at Patricia Bay. Patricia Bay is located in the Saanich Inlet, on the southern tip of Vancouver Island. This device is a Oxygen Sensor. Oxygen sensors measure dissolved oxygen concentration in seawater. It was deployed on a fixed platform. Data from this deployment were archived and made available through Ocean Networks Canada's Oceans 3.0 digital infrastructure, with quality assurance and derived data products following established practices.	Oceans
Ce Nortek Aquadopp Current Meter A2L2557 a t dploy le 2012-06-19 au Endeavour North. Cette section nord des vents hydrothermaux Endeavour comprend deux mouillages pour surveiller la circulation rgionale. Cet instrument est un Courantometre. Les courantometres acoustiques (ACM) mesurent la vitesse et la direction du courant l'aide de l'effet Doppler. L'instrument transmet une courte impulsion sonore, puis coute son cho pour mesurer le changement de hauteur ou de frquence. Le changement de frquence peut dterminer la vitesse du courant. Il a t deploy sur une plateforme fixe. Les donnes de ce dploiement sont archives et accessibles sur linfrastructure numrique Oceans 3.0 du Rseau Canadien des Ocans (ONC), avec assurance de la qualit et produits drivs selon les conventions tablies.	northward_sea_water_velocity
Egg development analysis data during the 2020 lobster fishing season in LFA22. Monitoring the development of lobster eggs is included in the project Deployment of a multiparametric decision support tool for the opening date of the lobster fishery for Lobster Fishing Area 22 of the Rassemblement des pcheurs et pcheuses des ctes des les(RPPCI), funded by the Quebec Fisheries Fund (MAPAQ and DFO), over 2 years. Lobster eggs are taken from female eggs by two volunteer commercial fishermen who are part of the RPPCI. Eggs are harvested from 10 females once a week on each facade of the Islands (North and South). Fishermen also installed Minilog II temperature loggers supplied by Merinov on one of their cages. The analysis is done by Merinov using a binocular and Image Pro image analyzer software. The proportion of the eye to the egg, linked to the temperature data, makes it possible to estimate the hatching date of the eggs. In 2020, the commercial fishing dates were from 9 May to 11 July. Due to the Covid 19 pandemic, data for the first week could not be collected. You can find the other project data in the catalog. Environmental monitoring data are available here (https://catalogue.ogsl.ca/dataset/en/ca-cioos_96bf3c76-a010-4637-bff5-59256f2637cc) and data on lobster monitoring in preseason fisheries can be viewed here (https://catalogue.ogsl.ca/en/dataset/ca-cioos_9c10259d-9433-4be2-abf9-7eb8b5fac5ad).	habitat characterization
Quadra Island, at the northern terminus of the Salish Sea, has been a site for shore-based and high-resolution measurement of surface seawater CO2 content since December 2014. Measurements of in situ temperature, salinity, and CO2 partial pressure are made near-continuously from a seawater sample line with an intake 50 m from shore and at a depth of 1 m in Hyacinthe Bay on the eastern side of Quadra Island. The effort to collect these data are part of the Hakai Institutes directive to advance the understanding of carbon cycling in northeast Pacific coastal settings with specific emphasis on sea-air CO2 exchange and ocean acidification.	ocean

too broad, too compute intensive

measurements for conductivity, T and Δd

buoy measurement for

dity…

oxygen sensors

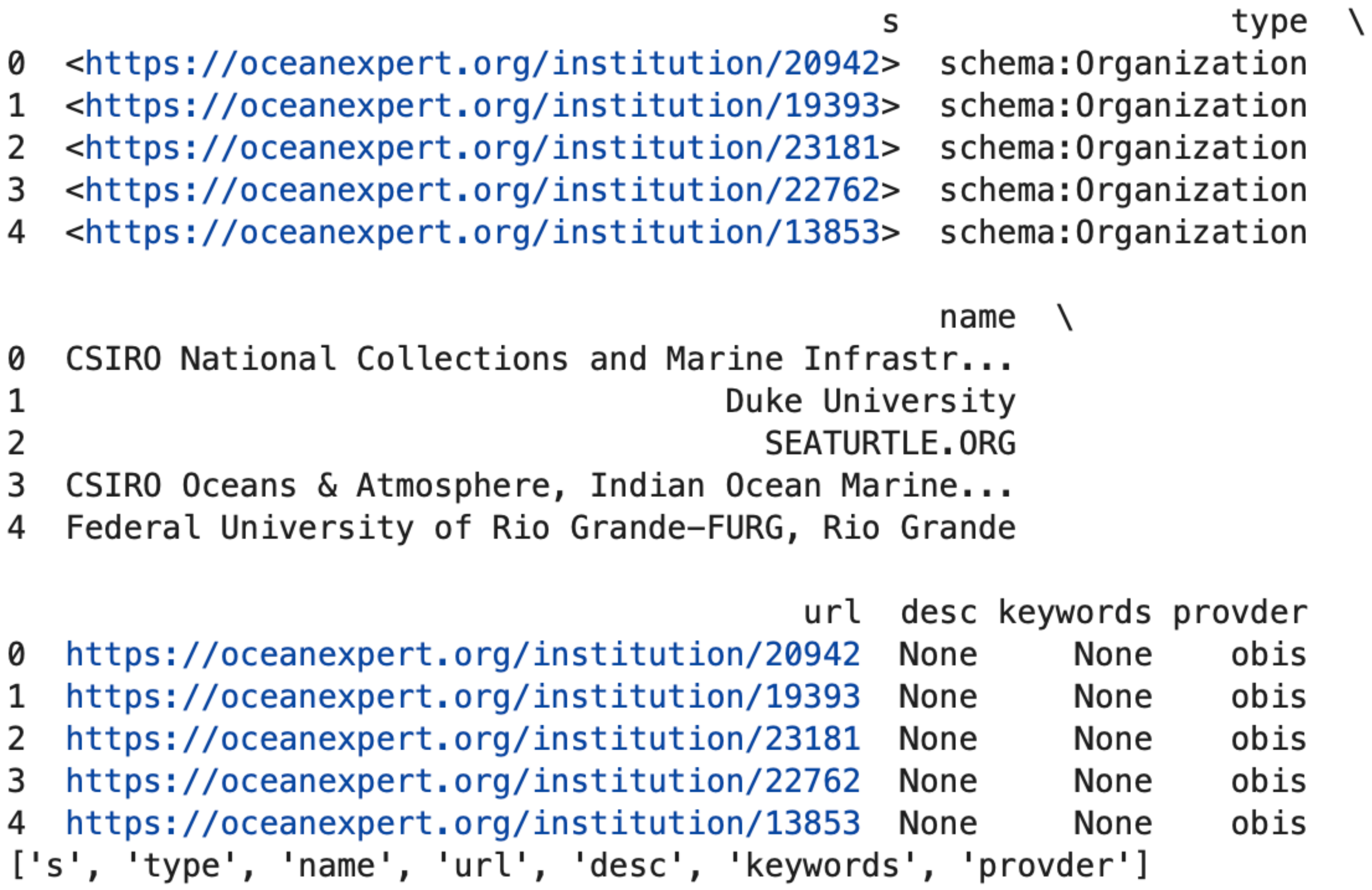
ocean circulations

lobster egg development

CO2 measurements

Targeting the ocean biodiversity data landscape OBIS

- Content of OBIS parquet file:
  - URI, URL
  - type of entry e.g. dataset
  - name
  - description
  - keywords
  - providers

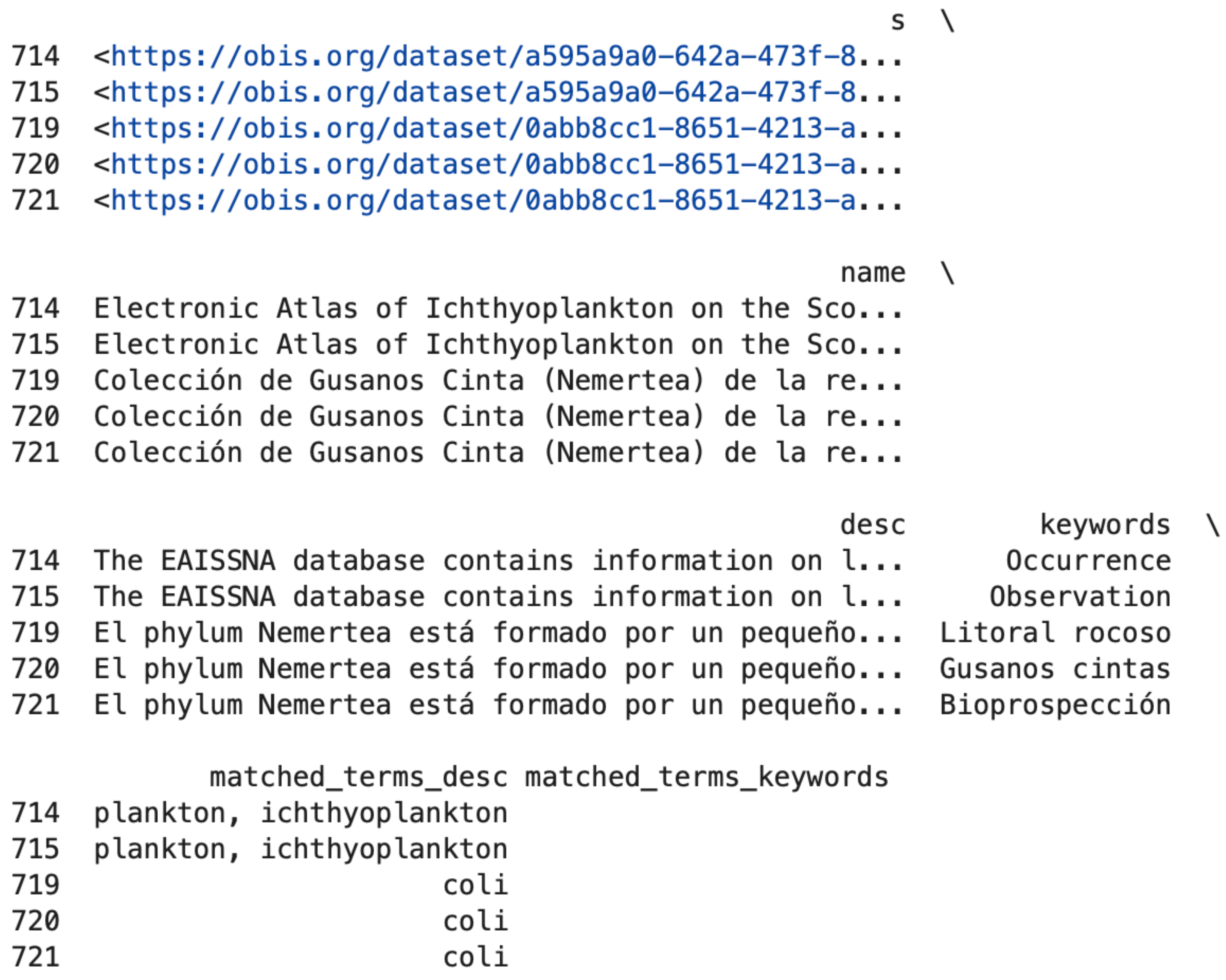


```
                                             s                 type  \
0  <https://oceanexpert.org/institution/20942>  schema:Organization
1  <https://oceanexpert.org/institution/19393>  schema:Organization
2  <https://oceanexpert.org/institution/23181>  schema:Organization
3  <https://oceanexpert.org/institution/22762>  schema:Organization
4  <https://oceanexpert.org/institution/13853>  schema:Organization

                                                name  \
0  CSIRO National Collections and Marine Infrastr...
1                                    Duke University
2                                      SEATURTLE.ORG
3  CSIRO Oceans & Atmosphere, Indian Ocean Marine...
4  Federal University of Rio Grande-FURG, Rio Grande

                                         url  desc keywords provder
0  https://oceanexpert.org/institution/20942  None     None    obis
1  https://oceanexpert.org/institution/19393  None     None    obis
2  https://oceanexpert.org/institution/23181  None     None    obis
3  https://oceanexpert.org/institution/22762  None     None    obis
4  https://oceanexpert.org/institution/13853  None     None    obis
['s', 'type', 'name', 'url', 'desc', 'keywords', 'provder']
```

2. Targeting data landscape

Filtering the OBIS parquet file for datasets relevant for marine microbiology

- filtered for type = **dataset**
- **exact matching** any of the specific terms of the microbial index within the keyword or description fields
- created a column for **matched terms** in **description** and matched terms in **keywords**
- saved as csv.



```
                                                     s  \
714  <https://obis.org/dataset/a595a9a0-642a-473f-8...   
715  <https://obis.org/dataset/a595a9a0-642a-473f-8...   
719  <https://obis.org/dataset/0abb8cc1-8651-4213-a...   
720  <https://obis.org/dataset/0abb8cc1-8651-4213-a...   
721  <https://obis.org/dataset/0abb8cc1-8651-4213-a...   

                                                  name  \
714  Electronic Atlas of Ichthyoplankton on the Sco...   
715  Electronic Atlas of Ichthyoplankton on the Sco...   
719  Colección de Gusanos Cinta (Nemertea) de la re...   
720  Colección de Gusanos Cinta (Nemertea) de la re...   
721  Colección de Gusanos Cinta (Nemertea) de la re...   

                                                  desc        keywords  \
714  The EAISSNA database contains information on l...      Occurrence   
715  The EAISSNA database contains information on l...     Observation   
719  El phylum Nemertea está formado por un pequeño...  Litoral rocoso   
720  El phylum Nemertea está formado por un pequeño...  Gusanos cintas   
721  El phylum Nemertea está formado por un pequeño...  Bioprospección   

            matched_terms_desc matched_terms_keywords  
714  plankton, ichthyoplankton                         
715  plankton, ichthyoplankton                         
719                       coli                         
720                       coli                         
721                       coli                         
```

Building blocks for the directed graph

- **networkx** for python→ provides tools to build and manipulate directed graphs and analyse their structure and dynamics
- focused on the **300 first entries** of the retrieved and filtered OBIS
- **summarised matched terms** from keywords and description into one single matched term column

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…

.GEFX →

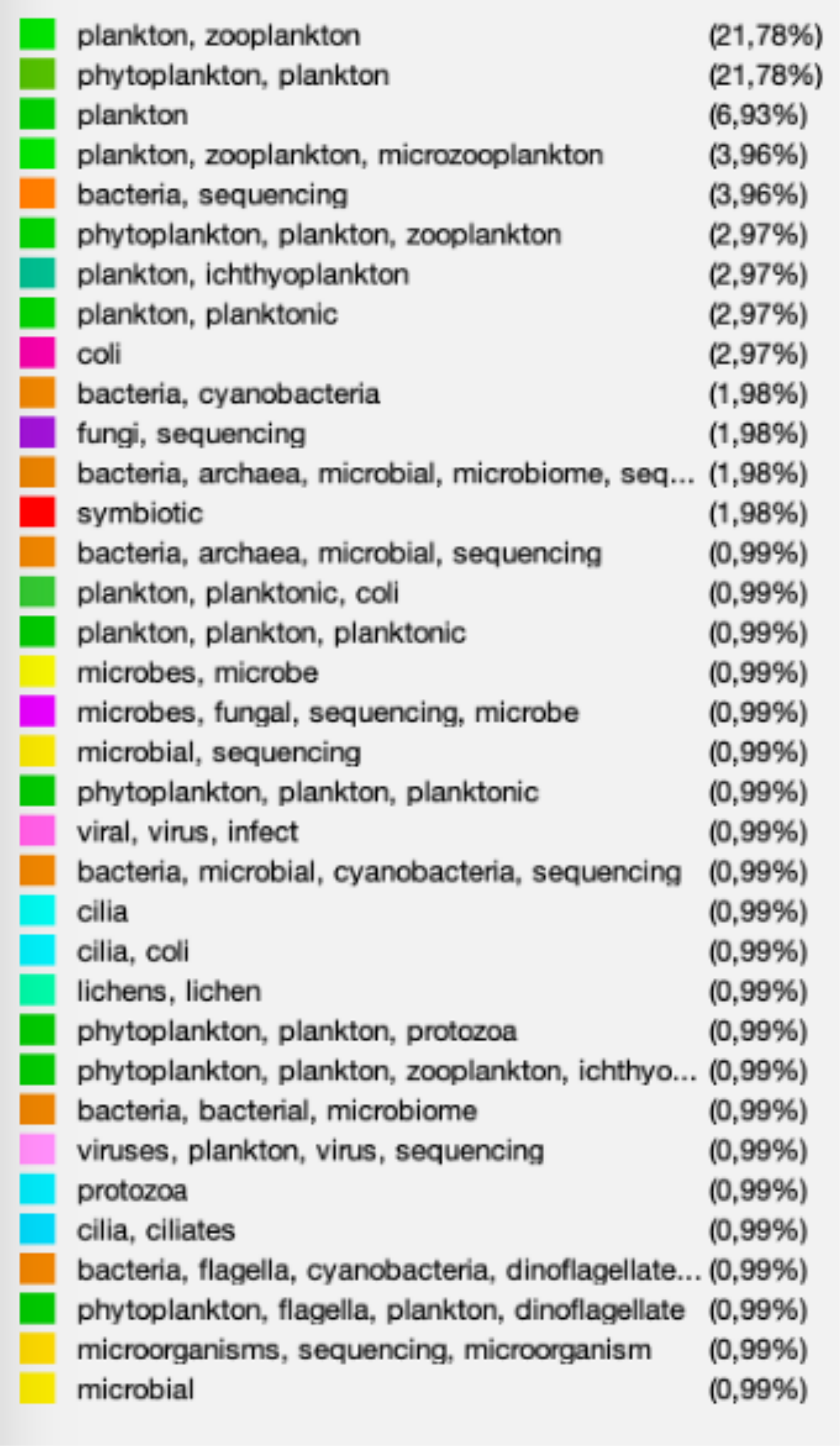
Gephi

v.0.10.1

3. Building and visualising directed graph

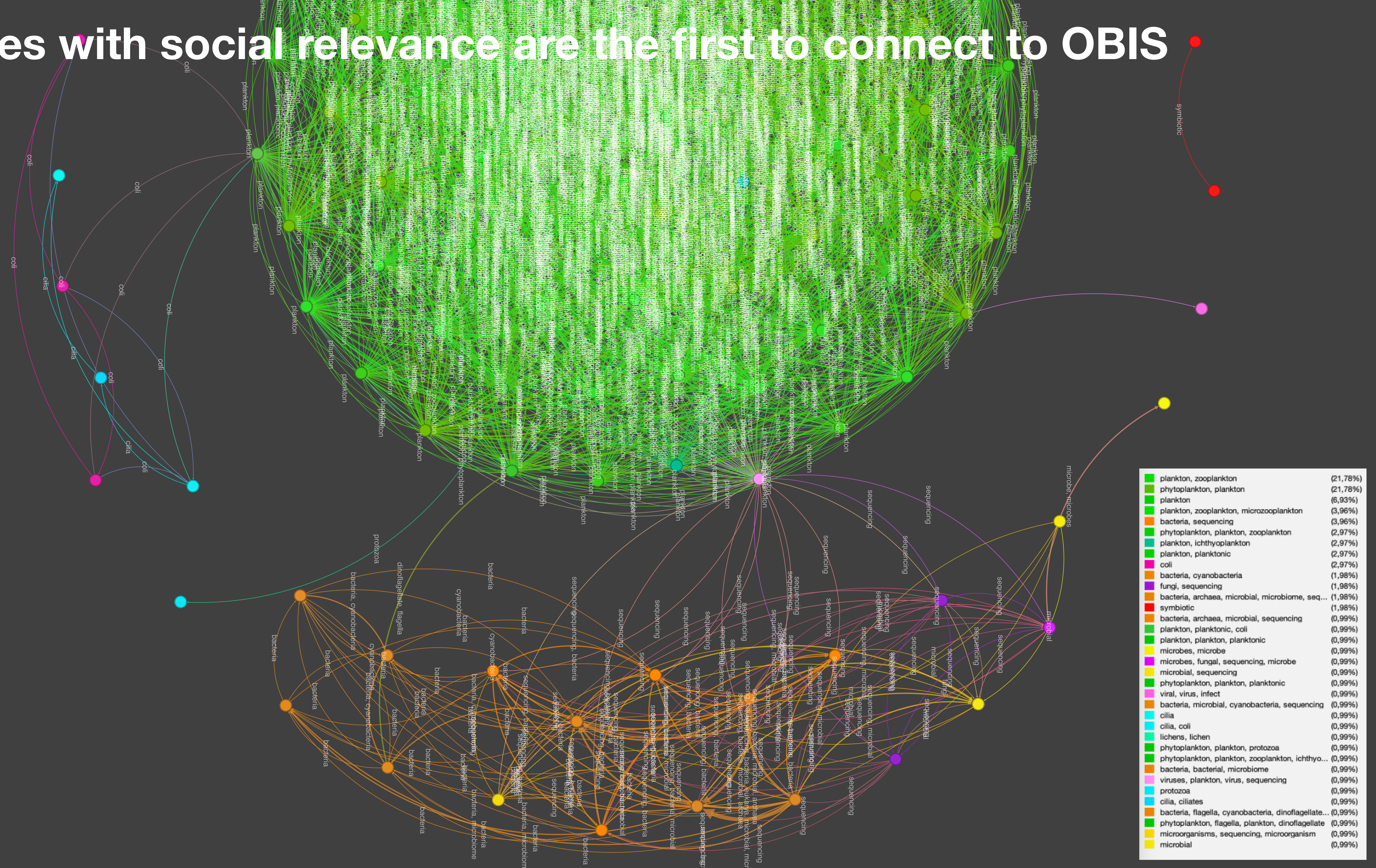
Customising Gephi settings for optimal graph visualization

- Fruchterman-Rheingold algorithm
- adjust node colours to match similar terms
- edges partitioned based on co-occurence of matched terms
- adjust weight of arrows based on frequency of co-occurrence



🟩	plankton, zooplankton	(21,78%)
🟩	phytoplankton, plankton	(21,78%)
🟩	plankton	(6,93%)
🟩	plankton, zooplankton, microzooplankton	(3,96%)
🟧	bacteria, sequencing	(3,96%)
🟩	phytoplankton, plankton, zooplankton	(2,97%)
🟩	plankton, ichthyoplankton	(2,97%)
🟩	plankton, planktonic	(2,97%)
🟪	coli	(2,97%)
🟧	bacteria, cyanobacteria	(1,98%)
🟪	fungi, sequencing	(1,98%)
🟧	bacteria, archaea, microbial, microbiome, seq...	(1,98%)
🟥	symbiotic	(1,98%)
🟧	bacteria, archaea, microbial, sequencing	(0,99%)
🟩	plankton, planktonic, coli	(0,99%)
🟩	plankton, plankton, planktonic	(0,99%)
🟨	microbes, microbe	(0,99%)
🟪	microbes, fungal, sequencing, microbe	(0,99%)
🟨	microbial, sequencing	(0,99%)
🟩	phytoplankton, plankton, planktonic	(0,99%)
🟪	viral, virus, infect	(0,99%)
🟧	bacteria, microbial, cyanobacteria, sequencing	(0,99%)
🟦	cilia	(0,99%)
🟦	cilia, coli	(0,99%)
🟩	lichens, lichen	(0,99%)
🟩	phytoplankton, plankton, protozoa	(0,99%)
🟩	phytoplankton, plankton, zooplankton, ichthyo...	(0,99%)
🟧	bacteria, bacterial, microbiome	(0,99%)
🟪	viruses, plankton, virus, sequencing	(0,99%)
🟦	protozoa	(0,99%)
🟦	cilia, ciliates	(0,99%)
🟧	bacteria, flagella, cyanobacteria, dinoflagellate...	(0,99%)
🟩	phytoplankton, flagella, plankton, dinoflagellate	(0,99%)
🟨	microorganisms, sequencing, microorganism	(0,99%)
🟨	microbial	(0,99%)

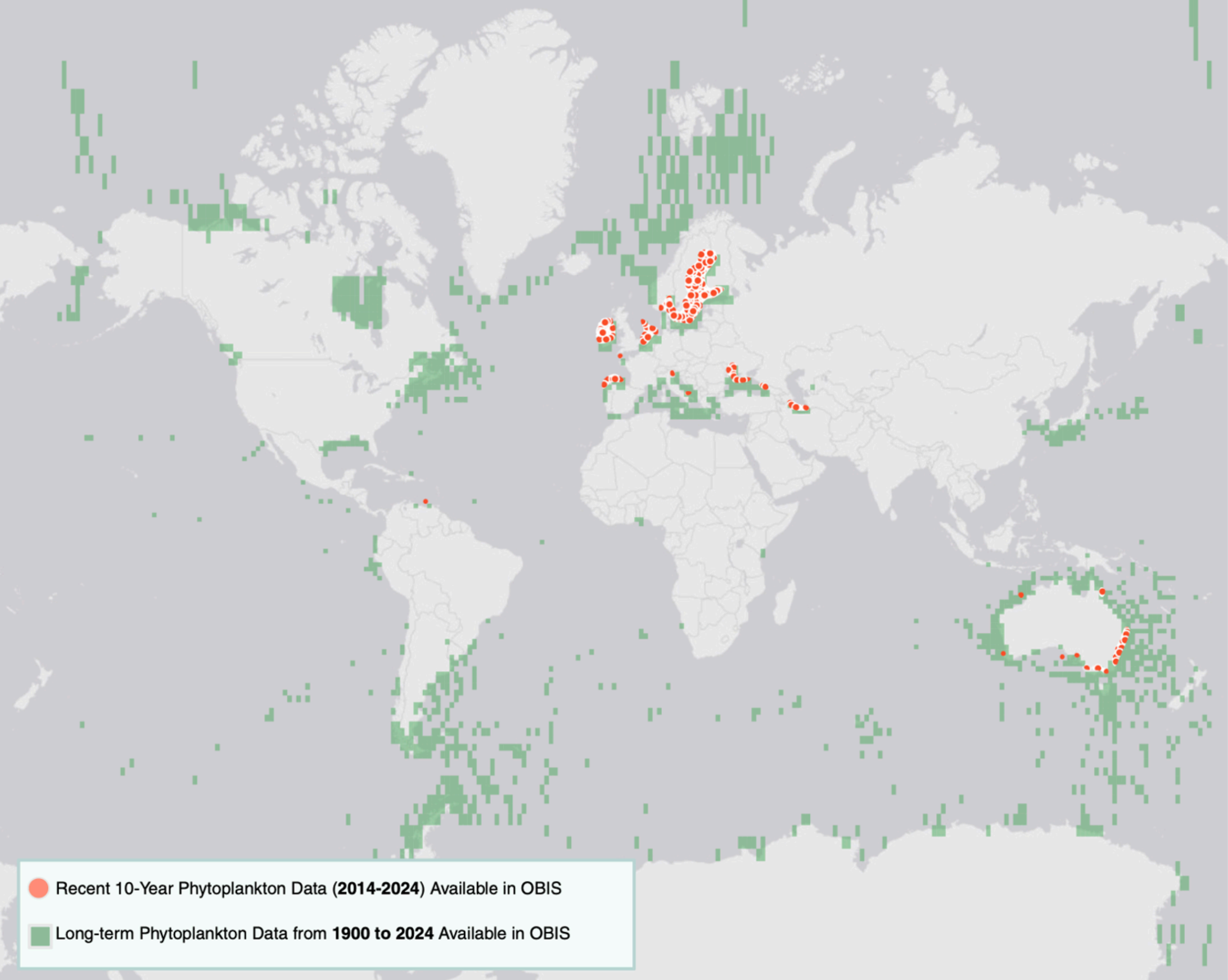
Microbes with social relevance are the first to connect to OBIS



🟩	plankton, zooplankton	(21,78%)
🟩	phytoplankton, plankton	(21,78%)
🟩	plankton	(6,93%)
🟩	plankton, zooplankton, microzooplankton	(3,96%)
🟧	bacteria, sequencing	(3,96%)
🟩	phytoplankton, plankton, zooplankton	(2,97%)
🟩	plankton, ichthyoplankton	(2,97%)
🟩	plankton, planktonic	(2,97%)
🟪	coli	(2,97%)
🟧	bacteria, cyanobacteria	(1,98%)
🟪	fungi, sequencing	(1,98%)
🟧	bacteria, archaea, microbial, microbiome, seq...	(1,98%)
🟥	symbiotic	(1,98%)
🟧	bacteria, archaea, microbial, sequencing	(0,99%)
🟩	plankton, planktonic, coli	(0,99%)
🟩	plankton, plankton, planktonic	(0,99%)
🟨	microbes, microbe	(0,99%)
🟪	microbes, fungal, sequencing, microbe	(0,99%)
🟨	microbial, sequencing	(0,99%)
🟩	phytoplankton, plankton, planktonic	(0,99%)
🟪	viral, virus, infect	(0,99%)
🟧	bacteria, microbial, cyanobacteria, sequencing	(0,99%)
🟦	cilia	(0,99%)
🟦	cilia, coli	(0,99%)
🟩	lichens, lichen	(0,99%)
🟩	phytoplankton, plankton, protozoa	(0,99%)
🟩	phytoplankton, plankton, zooplankton, ichthyo...	(0,99%)
🟧	bacteria, bacterial, microbiome	(0,99%)
🟪	viruses, plankton, virus, sequencing	(0,99%)
🟦	protozoa	(0,99%)
🟦	cilia, ciliates	(0,99%)
🟧	bacteria, flagella, cyanobacteria, dinoflagellate...	(0,99%)
🟩	phytoplankton, flagella, plankton, dinoflagellate	(0,99%)
🟨	microorganisms, sequencing, microorganism	(0,99%)
🟨	microbial	(0,99%)

Identifying sources of extensive phytoplankton datasets in OBIS

- missing geolocalization data in the OBIS parquet file
- instead used OBIS data mapper (https://mapper.obis.org/)
- filtered nodes for ones containing matched term for phytoplankton
- localised these nodes in the OBIS tool map
- compared data from 1900-2024 with data from the past 10 years



● Recent 10-Year Phytoplankton Data (**2014-2024**) Available in OBIS
🟩 Long-term Phytoplankton Data from **1900 to 2024** Available in OBIS

Areas of improvement

- Consider quality of data from partners for e.g. long description has a higher chance of matching one of the terms in the microbial term index
- down-weight closely related terms to avoid clutter
- Do some graph analysis:
    - module detection tools to look into graph structure
    - identify strongly interconnected clusters and Identify nodes with the highest and lowest degrees and the most isolated ones
- integration of spatial data in the OBIS parquet file