EMBL's European Bioinformatics Institute (EMBL-EBI)

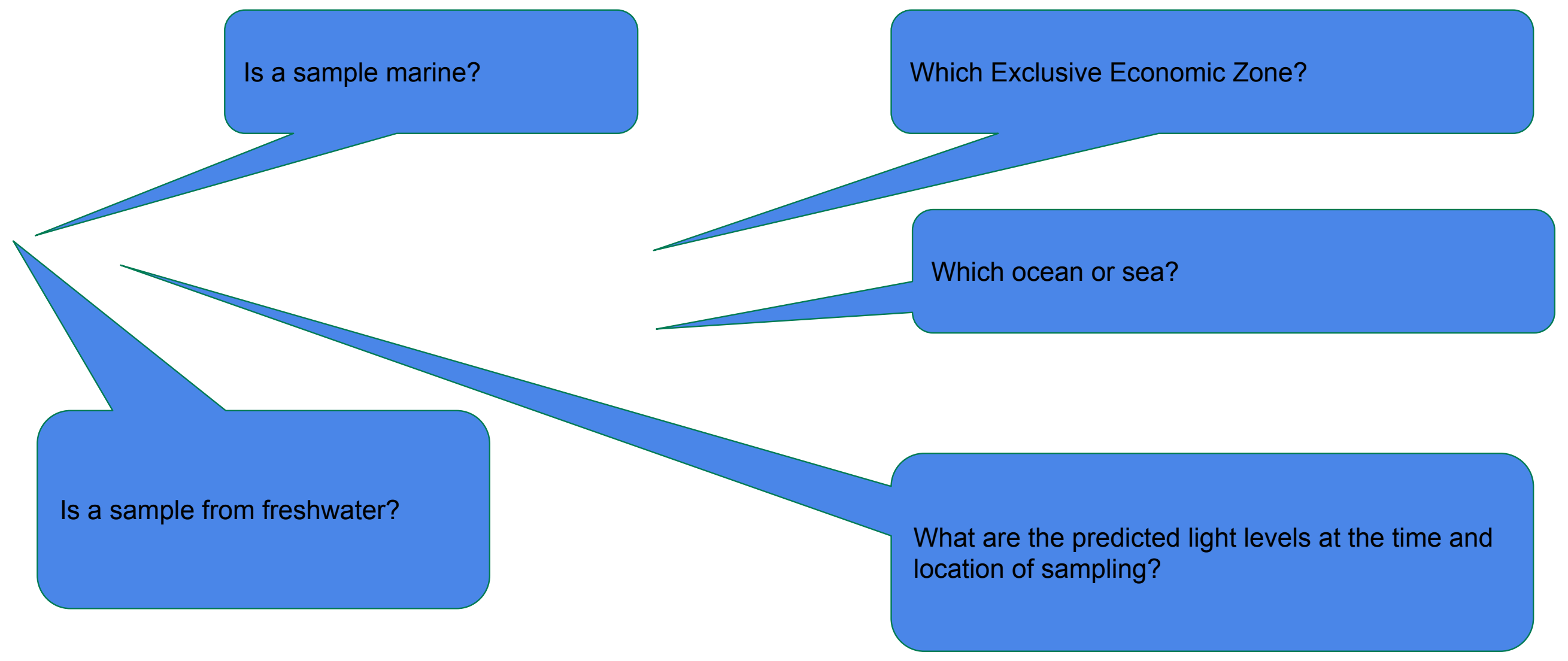
Augmenting the Environmental Context in Biological Samples using Geographic and Taxonomic Metadata

EMBL-EBI

P. Woollard, S. Pesant, J. Burgin, S. Jayathilaka and G. Cochrane

ORCID:0000-0002-7654-6902, 0000-0002-4936-5209, 0000-0002-9818-1094, 0000-0002-0154-8807, 0000-0001-7954-7057

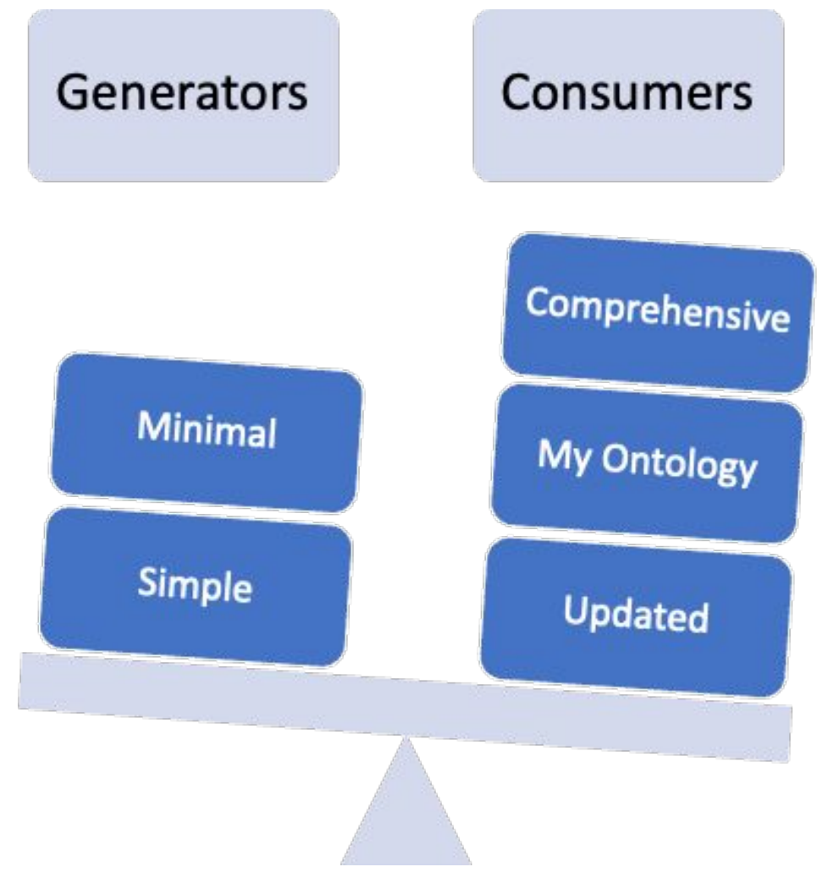
Challenge: determining the provenance of nucleotide data for marine and freshwater environments


Is a sample marine?
Which Exclusive Economic Zone?
Which ocean or sea?
Is a sample from freshwater?
What are the predicted light levels at the time and location of sampling?

European Nucleotides Archive(ENA)

- A repository of the world's nucleotide data
- European node of INSDC
- Creators of tools for submission and retrieval

Provide Checklists: with limited mandatory metadata



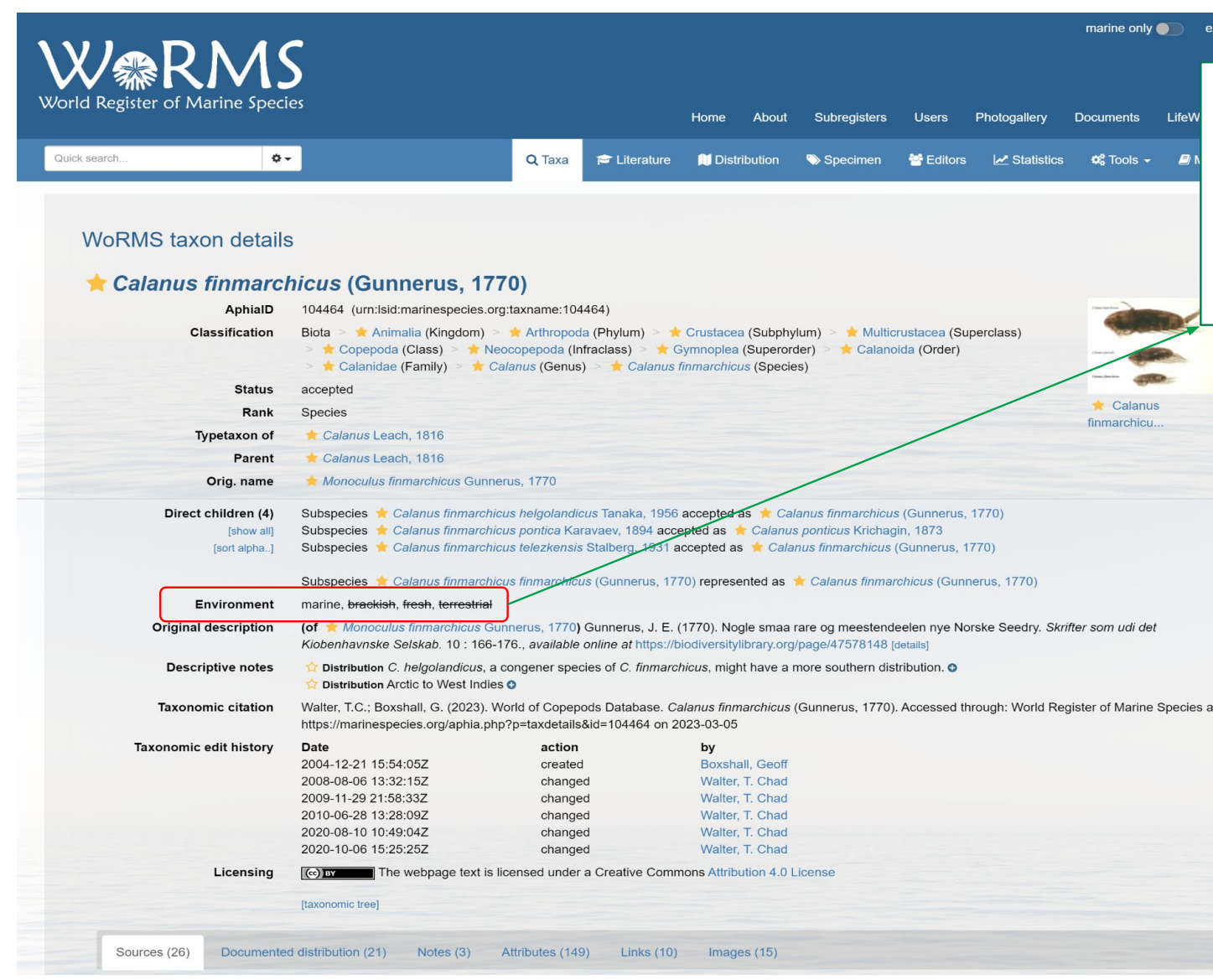
- balancing the needs of the different users
- science and technology all evolve.

Key Sample Provenance Attributes at ENA

Sample site fields	Control	Total samples with these	Comment
Taxonomy *"taxon_id"*	NCBI taxonomy	100% (23,724,884)	• mandatory!
Geography *"latitude"* and *"longitude"*	[0-9]+.[0-9]+	15% (3,588,320)	• much automated cleaning by ENA technical team to provide this cleanly. • 1,918 samples with start and end coordinates
Keywords *"Environment_biome"*	Free text + sometimes ENVO terms	6% (1,523,513)	• Often allows determination if marine/freshwater biome • free text = harmonisation needed
country	Controlled "country" list	62% (14,782,362)	• (9,045 samples are classified as sea or ocean)

Implementation

- External resources inputs:
  - Species environment labels from the World Register of Marine Species(WoRMS)
  - Geography as polygons in shapefiles (WWF, OpenStreetMap.de, marineregetc.)
- A mainly python pandas based workflow was used to cope with 23 million samples worth of metadata



WoRMS World Register of Marine Species

marine only extant only

Home About Subregisters Users Photogallery Documents LifeW

Quick search...

Taxa Literature Distribution Specimen Editors Statistics Tools

WoRMS taxon details

★ *Calanus finmarchicus* (Gunnerus, 1770)

AphiaID 104464 (urn:lsid:marinespecies.org:taxname:104464)

Classification Biota > Animalia (Kingdom) > Arthropoda (Phylum) > Crustacea (Subphylum) > Multicrustacea (Superclass) > Copepoda (Class) > Neocopepoda (Infraclass) > Gymnoplea (Superorder) > Calanoida (Order) > Calanidae (Family) > *Calanus* (Genus) > *Calanus finmarchicus* (Species)

Status accepted

Rank Species

Typetaxon of *Calanus* Leach, 1816

Parent *Calanus* Leach, 1816

Orig. name *Monoculus finmarchicus* Gunnerus, 1770

Direct children (4) [show all] [sort alpha..] Subspecies *Calanus finmarchicus helgolandicus* Tanaka, 1956 accepted as *Calanus finmarchicus* (Gunnerus, 1770)
Subspecies *Calanus finmarchicus pontica* Karavaev, 1894 accepted as *Calanus ponticus* Krichagin, 1873
Subspecies *Calanus finmarchicus telezkensis* Stalberg, 1931 accepted as *Calanus finmarchicus* (Gunnerus, 1770)
Subspecies *Calanus finmarchicus finmarchicus* (Gunnerus, 1770) represented as *Calanus finmarchicus* (Gunnerus, 1770)

Environment marine, ~~brackish~~, ~~fresh~~, ~~terrestrial~~

Original description (of *Monoculus finmarchicus* Gunnerus, 1770) Gunnerus, J. E. (1770). Nogle smaa rare og meestendeelen nye Norske Seedry. *Skrifter som udi det Kiobenhavnske Selskab*. 10 : 166-176., available online at https://biodiversitylibrary.org/page/47578148 [details]

Descriptive notes Distribution *C. helgolandicus*, a congener species of *C. finmarchicus*, might have a more southern distribution.
Distribution Arctic to West Indies

Taxonomic citation Walter, T.C.; Boxshall, G. (2023). World of Copepods Database. *Calanus finmarchicus* (Gunnerus, 1770). Accessed through: World Register of Marine Species at: https://marinespecies.org/aphia.php?p=taxdetails&id=104464 on 2023-03-05

Taxonomic edit history	Date	action	by
	2004-12-21 15:54:05Z	created	Boxshall, Geoff
	2008-08-06 13:32:15Z	changed	Walter, T. Chad
	2009-11-29 21:58:33Z	changed	Walter, T. Chad
	2010-06-28 13:28:09Z	changed	Walter, T. Chad
	2020-08-10 10:49:04Z	changed	Walter, T. Chad
	2020-10-06 15:25:25Z	changed	Walter, T. Chad

Licensing The webpage text is licensed under a Creative Commons Attribution 4.0 License

[taxonomic tree]

Sources (26) Documented distribution (21) Notes (3) Attributes (149) Links (10) Images (15)

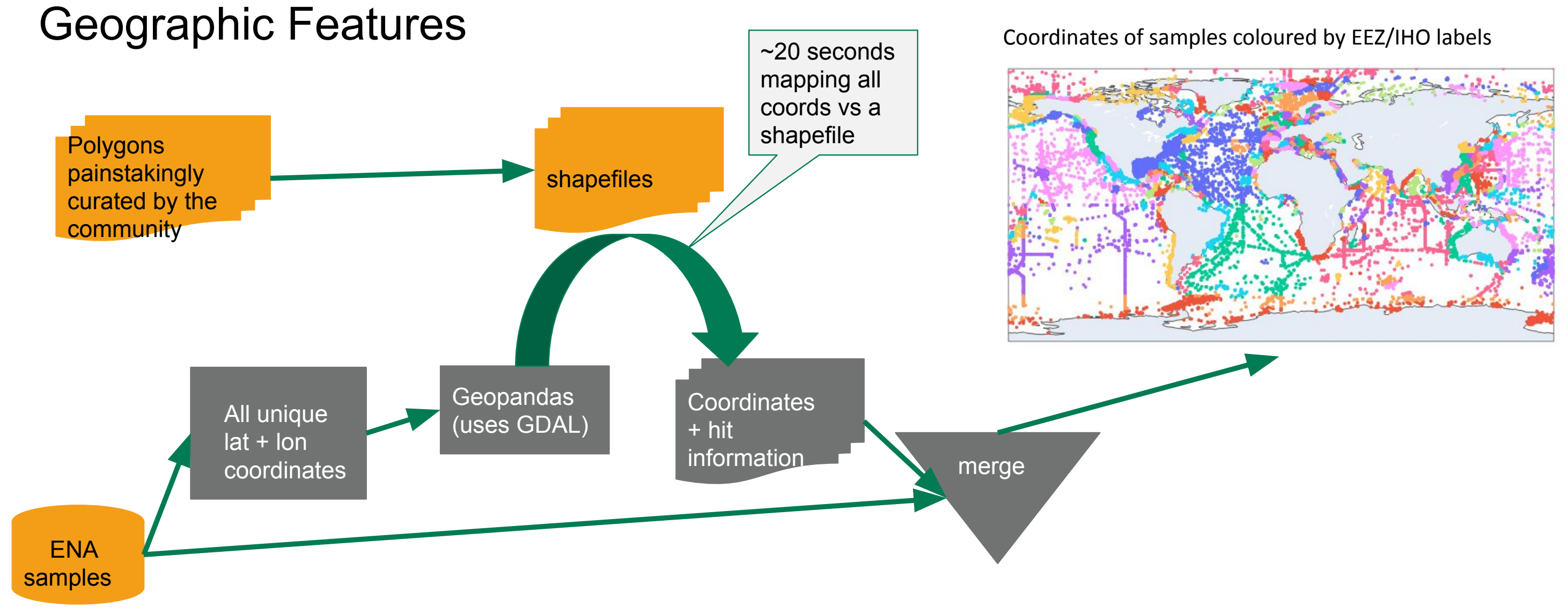
Calanus finmarchicu...

Environment flag by taxonomy experts
- marine
- brackish
- freshwater
- terrestrial

Example of a shapefile (multiple polygons with coordinates)

Intersect_EEZ_IHO_v4_2020.zip: Citation: Flanders Marine Institute (2020). The intersect of the Exclusive Economic Zones and IHO sea areas, version 4. Available online at https://www.marineregions.org/.https://doi.org/10.14284/402 The IHO polygon layer has a low resolution, but the EEZ polygon layer has a high resolution coastline (GSHHS).

Mapping Latitude and Longitude Coordinates to Geographic Features


~20 seconds mapping all coords vs a shapefile
Polygons painstakingly curated by the community
shapefiles
All unique lat + lon coordinates
Geopandas (uses GDAL)
Coordinates + hit information
merge
ENA samples

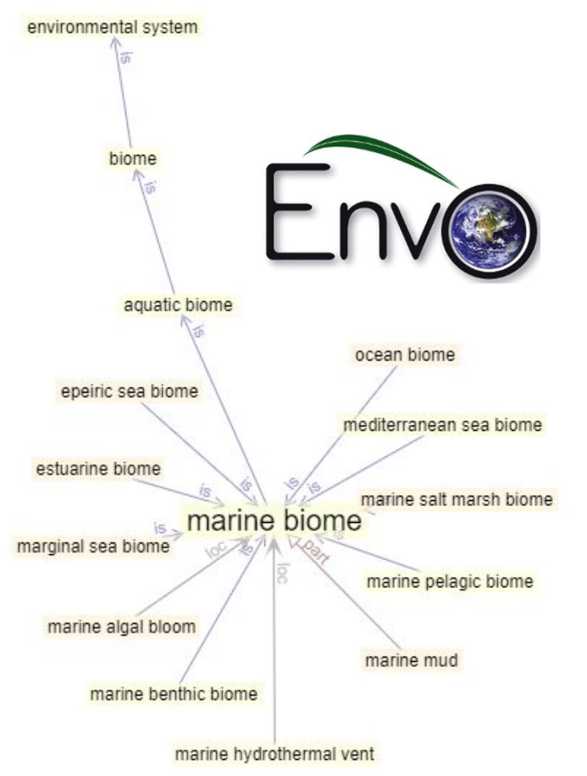
Coordinates of samples coloured by EEZ/IHO labels

Environment by keywords

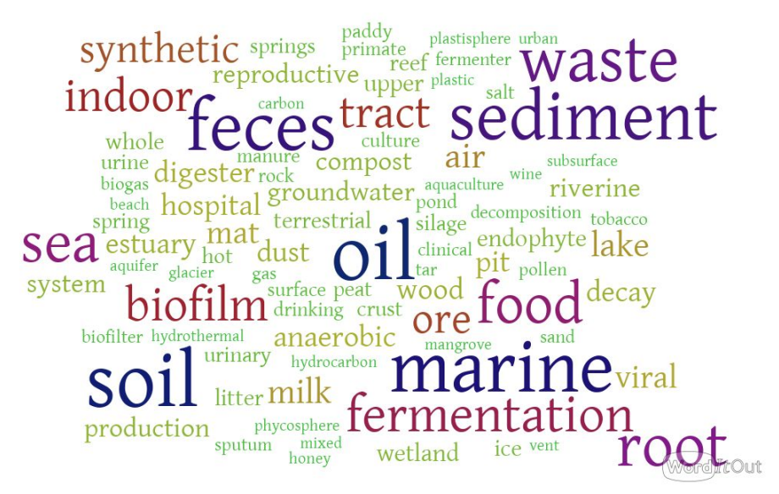
environment_biome examples

(ENVO or free text)

marine benthic biome
cropland biome
freshwater biome
Human gut
village biome
Gut
ENVO:dense settlement biome
ENVO:00009003
coral reef
ENVO:human-associated habitat
forest
ENVO:2100002
Human
area of cropland
grassland
mouse gut
soil
marine biome
terrestrial biome [ENVO:00000446]
lentic water body

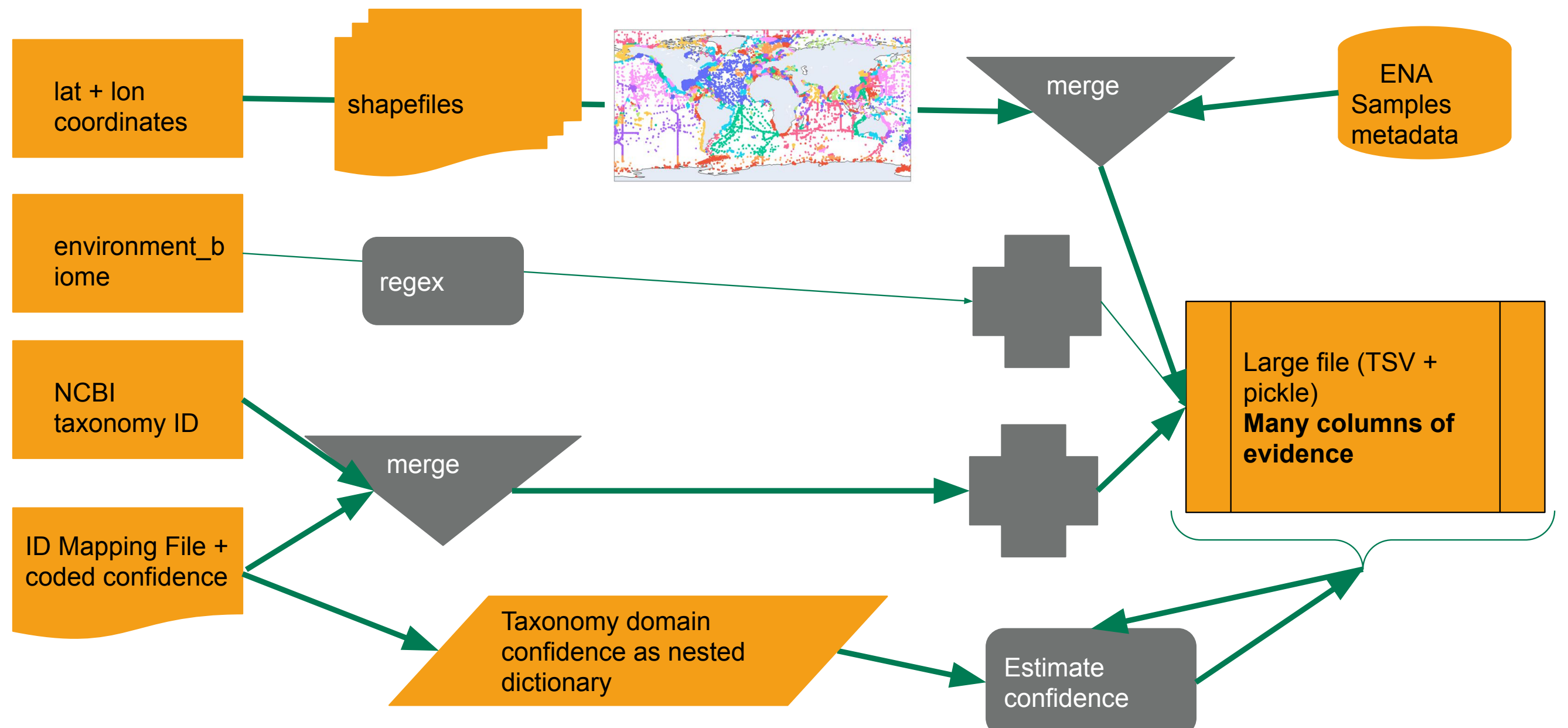


Unclassified "metagenome" taxonomic terms


synthetic springs paddy primate plastisphere urban reef fermenter waste
reproductive upper plastic salt
indoor carbon feces tract sediment
whole culture manure compost air subsurface
urine digester rock wine
biogas groundwater aquaculture riverine
beach hospital pond decomposition tobacco
spring terrestrial silage endophyte lake
sea estuary mat hot dust oil clinical pit
aquifer glacier gas tar pollen
system surface peat wood decay
biofilm drinking crust ore food
biofilter hydrothermal anaerobic mangrove sand
urinary hydrocarbon
soil litter milk marine viral
production phycosphere fermentation
sputum mixed honey wetland ice vent root

Merging Evidence

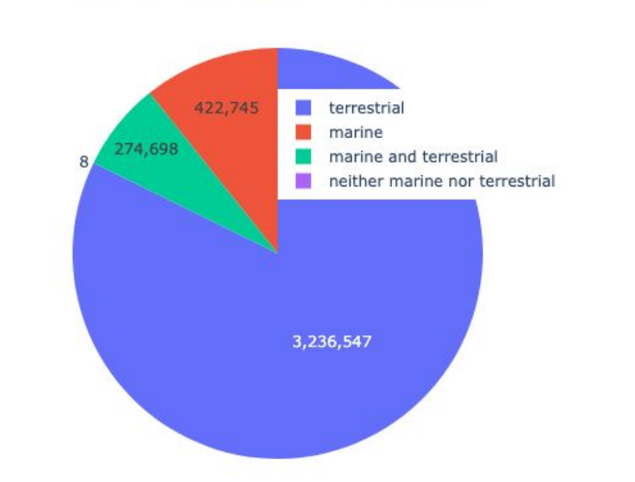
Workflow: Integrating Evidence of "Blue" Domain



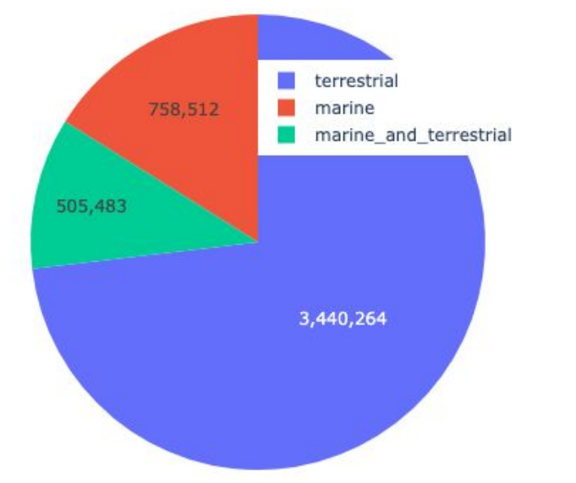
Combined coordinates, taxonomy and env_biome

- Just counts

location_designation sample counts



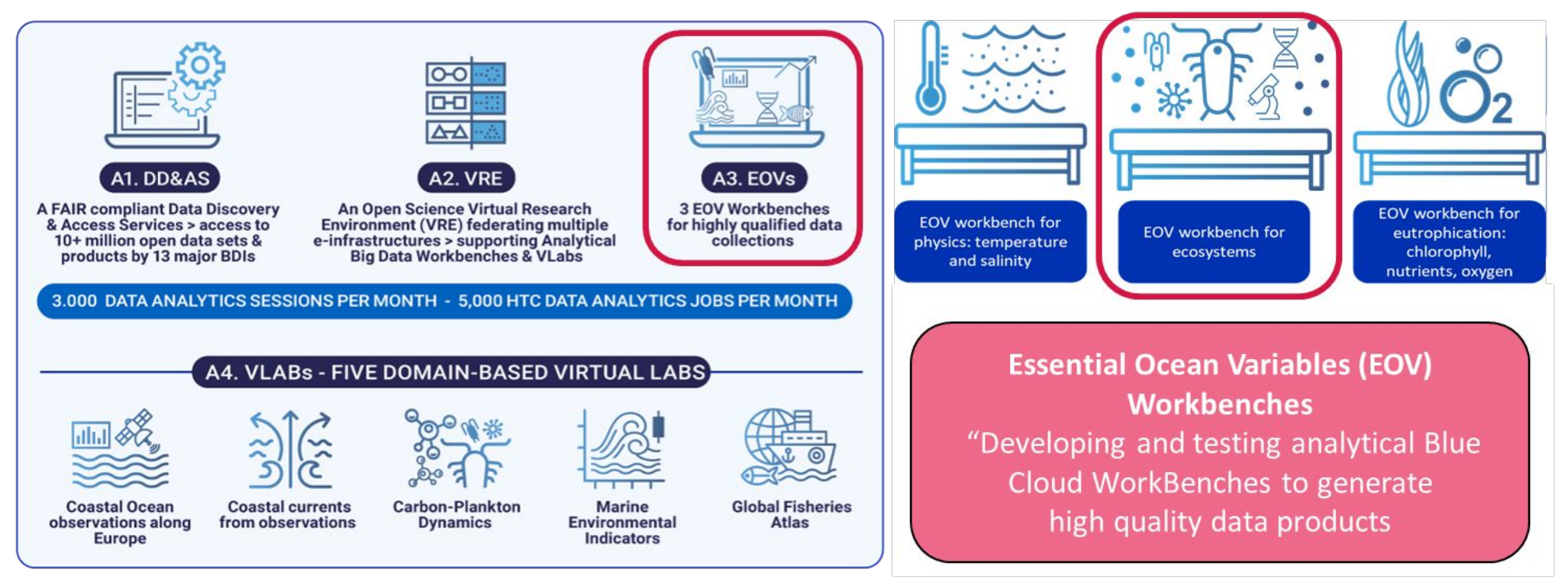
combined_location_designation sample counts



With the taxonomy info. in particular the marine domain massively increased in terms of sample coverage

Serving relevant marine & freshwater data

eosc Blue-Cloud


A1. DD&AS
A FAIR compliant Data Discovery & Access Services > access to 10+ million open data sets & products by 13 major BDIs
A2. VRE
An Open Science Virtual Research Environment (VRE) federating multiple e-infrastructures > supporting Analytical Big Data Workbenches & VLabs
A3. EOVs
3 EOV Workbenches for highly qualified data collections
3.000 DATA ANALYTICS SESSIONS PER MONTH - 5,000 HTC DATA ANALYTICS JOBS PER MONTH
A4. VLABs - FIVE DOMAIN-BASED VIRTUAL LABS
Coastal Ocean observations along Europe
Coastal currents from observations
Carbon-Plankton Dynamics
Marine Environmental Indicators
Global Fisheries Atlas
EOV workbench for physics: temperature and salinity
EOV workbench for ecosystems
EOV workbench for eutrophication: chlorophyll, nutrients, oxygen
Essential Ocean Variables (EOV) Workbenches
"Developing and testing analytical Blue Cloud WorkBenches to generate high quality data products

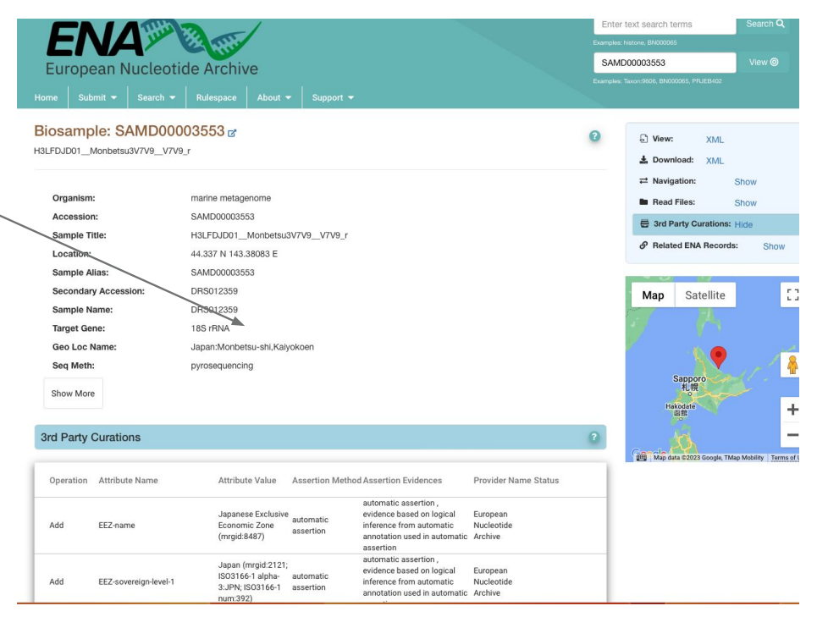
Informing

Essential annotations for science & society

curl "https://www.ebi.ac.uk/ena/clearinghouse/api/curations/16f53cc2-532c-486e-b8b1-b07ddb05f93b"
2>/dev/null | jq

```
{
  "curations": [
    {
      "id": "16f53cc2-532c-486e-b8b1-b07ddb05f93b",
      "recordType": "sample",
      "recordId": "SAMD00003553",
      "attributePost": "EEZ-name",
      "valuePost": "Japanese Exclusive Economic Zone (mrgid:8487)",
      "attributeDelete": false,
      "assertionMethod": "automatic assertion",
      "assertionEvidences": [
       {  "identifier": "ECO:0000203",        "shortForm": "ECO_0000203",
         "label": "automatic assertion },
       { "identifier": "ECO:0000366",        "shortForm": "ECO_0000366",
         "label": "evidence based on logical inference from automatic annotation used in automatic assertion" }
      ],
      "assertionSource": "",
      "assertionAdditionalInfo": "confidence:high; evidence:sample coordinates within EEZ shapefile",
      "providerName": "European Nucleotide Archive",
      "providerUrl": "https://www.ebi.ac.uk/ena/browser/home",
      "submittedTimestamp": "2023-04-05T07:09:55.109+0000",
      "updatedTimestamp": "2023-04-05T07:09:55.109+0000",
      "suppressed": false
    }
  ],
  "totalAttributes": 1
}
```

elixir Clearinghouse



ENA European Nucleotide Archive

Home Submit Search Rulespace About Support

Biosample: SAMD00003553

H3LFDJD01__Monbetsu3V7V9__V7V9_r

Organism: marine metagenome
Accession: SAMD00003553
Sample Title: H3LFDJD01__Monbetsu3V7V9__V7V9_r
Location: 44.337 N 143.38083 E
Sample Alias: SAMD00003553
Secondary Accession: DRS012359
Sample Name: DRS012359
Target Gene: 18S rRNA
Geo Loc Name: Japan:Monbetsu-shi,Kaiyokoen
Seq Meth: pyrosequencing

Show More

View: XML
Download: XML
Navigation: Show
Read Files: Show
3rd Party Curations: Hide
Related ENA Records: Show

Map Satellite

3rd Party Curations

Operation	Attribute Name	Attribute Value	Assertion Method	Assertion Evidences	Provider Name	Status
Add	EEZ-name	Japanese Exclusive Economic Zone (mrgid:8487)	automatic assertion	automatic assertion , evidence based on logical inference from automatic annotation used in automatic assertion	European Nucleotide Archive	
Add	EEZ-sovereign-level-1	Japan (mrgid:2121; ISO3166-1 alpha-3:JPN; ISO3166-1 num:392)	automatic assertion	automatic assertion , evidence based on logical inference from automatic annotation used in automatic assertion	European Nucleotide Archive	

2022 UN BIODIVERSITY CONFERENCE
COP15 - CP/MOP10-NP/MOP4
Ecological Civilization-Building a Shared Future for All Life on Earth
KUNMING – MONTRÉAL

Life in harmony, into the future
COP10/MOP5 AICHI-NAGOYA, JAPAN 2010

www.ebi.ac.uk

eosc Blue-Cloud2026
A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters

elixir Clearinghouse

WoRMS World Register of Marine Species

EMBL's European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus, Hinxton,
Cambridgeshire. CB10 1SD. UK.
T +44(0)1223 494 444