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

A blue partition for marine and freshwater environments

Peter Woollard

Data Standards biocurator, ENA, EMBL-EBI

eosc | Blue-Cloud2026

A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters

Acknowledgements: **Stéphane Pesant,** Josie Burgin, Suran Jayathilaka and Guy Cochrane

EMBL-EBI

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

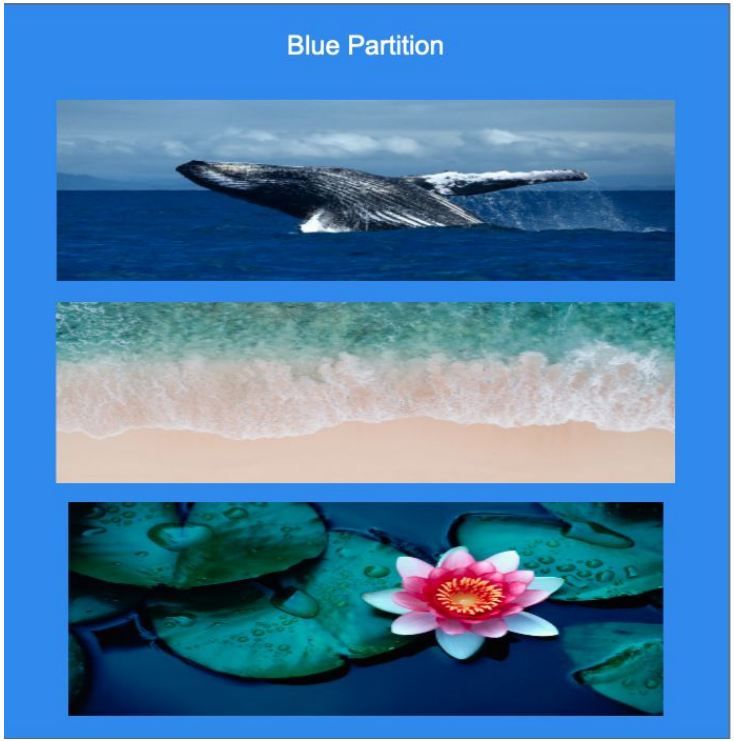
Is a sample marine?

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?



EMBL-EBI

The European Nucleotide Archive(ENA)

What is the ENA?

- A repository of the world's nucleotide data
- European node of INSDC
- Creators of tools for submission and retrieval
- Checklists are used to capture metadata, using minimum requirement and more detailed optional fields

Key Sample Provenance Attributes at ENA

Sample site fields	Control	Total samples with these	Comment
Taxonomy *"taxon_id"*	NCBI taxonomy	100% (23,724,884)	
Geography *"latitude"* and *"longitude"*	[0-9]+.[0-9]+	15% (3,588,320)	1,918 samples with start and end coordinates
Keywords *"Environment_biome"*	Free text + ENVO terms advised	6% (1,523,513)	Often allows determination if marine/freshwater biome
country	Controlled "country" list	62% (14,782,362)	(9,045 samples are classified as sea or ocean)

*ENA snapshot on 2023-03-10; 23,724,884 samples in total

Environment by Geography

Calculate if latitude and longitude are within geographic shapes

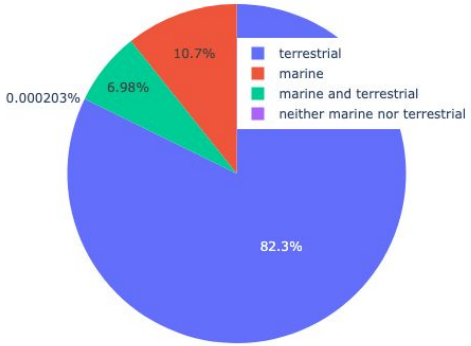
Examples of shapefile (multiple polygons with coordinates)

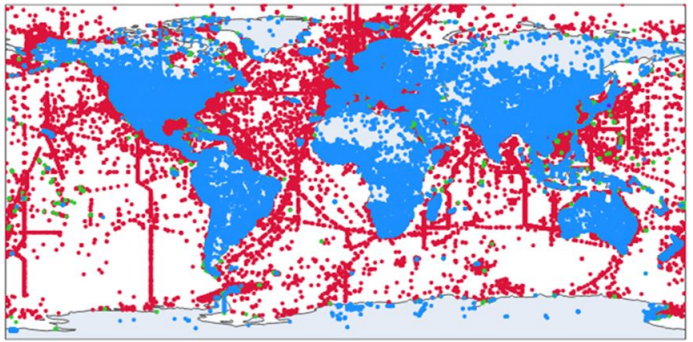
Intersect_EEZ_IHO_v4_2020.zip: Citation: Flanders Marine Institute (2020). The intersect of the Exclusive Economic Zones and IHO sea areas, version 4. Available online at https://www.marineregions.org/.https://doi.org/10.14284/402 The IHO polygon layer has a low resolution, but the EEZ polygon layer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

Longhurst_v4_2010.zip: Citation: Flanders Marine Institute (2009). Longhurst Provinces. Online at https://www.marineregions.org/. Unknown resolution, but likely coarse

GIS_hs_snapped.zip: Citation: Freshwater Ecoregions of the World(FEOW) hydrosheds (2008) https://www.feow.org/download

For samples with geographic provenances:





EMBL-EBI

Environment by Taxonomy

Environment flag by taxonomy experts

- marine
- brackish
- freshwater
- terrestrial

WoRMS World Register of Marine Species

Quick search... | Taxa | Literature

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, Subspecies *Calanus finmarchicus plumchrus* Karavaev, 189 Subspecies *Calanus finmarchicus glacialis* Jaschnov, 1 Subspecies *Calanus finmarchicus finmarchicus* (Gunneru
Environment	marine, ~~brackish~~, ~~fresh~~, ~~terrestrial~~
Original description	(of *Monoculus finmarchicus* Gunnerus, 1770) Gunnerus, *Kiobenhavnske Selskab.* 10 : 166-176., *available online at* htt
Descriptive notes	Distribution *C. helgolandicus*, a congener species of *C. fin* Distribution Arctic to West Indies
Taxonomic citation	Walter, T.C.; Boxshall, G. (2023). World of Copepods Databas https://marinespecies.org/aphia.php?p=taxdetails&id=104464
Taxonomic edit history	Date / action: 2004-12-21 15:54:05Z created; 2008-08-06 13:32:15Z changed; 2009-11-29 21:58:33Z changed; 2010-06-28 13:28:09Z changed; 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)

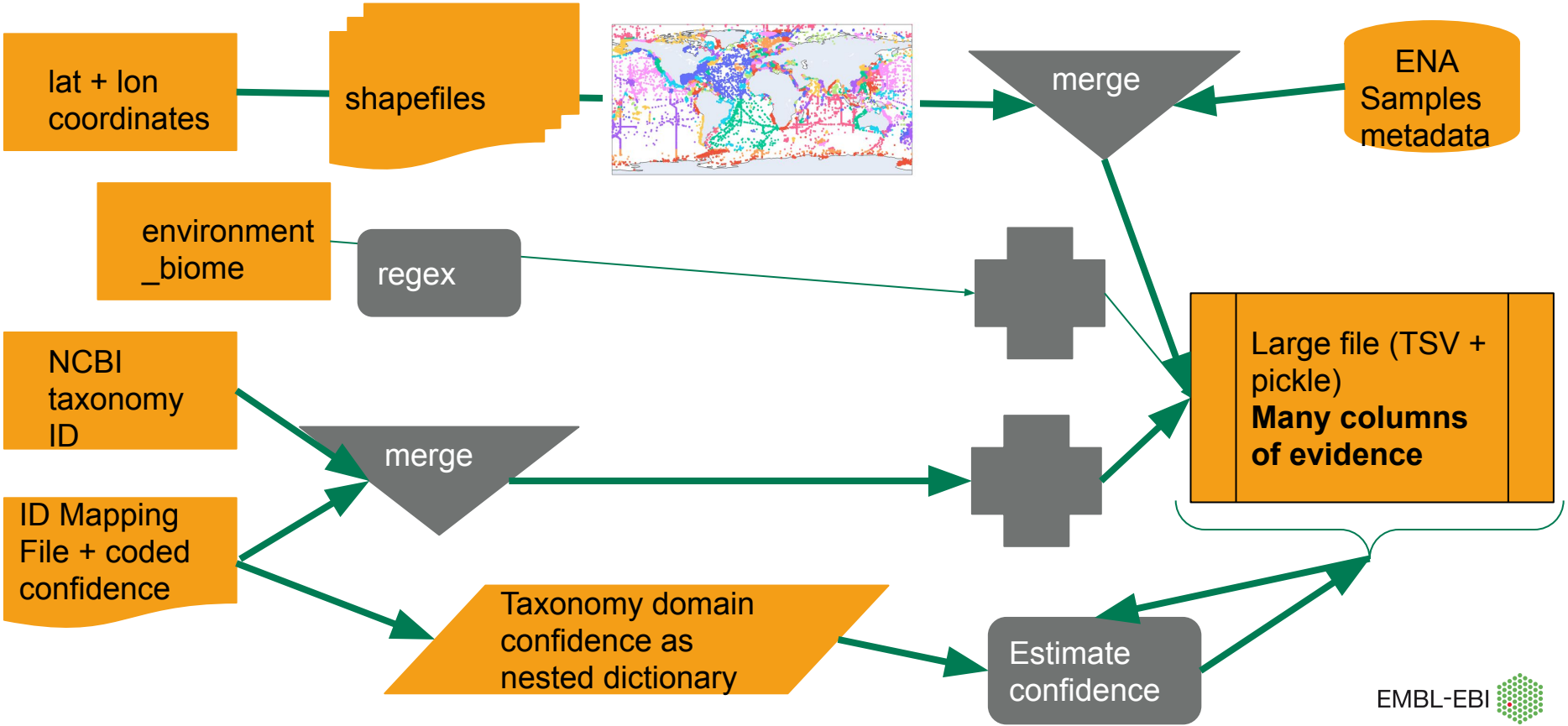

Calanus finmarchicu...

~189K taxa at NCBI
~78K (41%) taxa matched NCBI:WoRMS
~11K (14%) with data at ENA

Samples
~625K (3%) from matched taxa NCBI:WoRMS
~91K (15%) with geographic provenance

Slide adapted from:
Stéphane Pesant's

Workflow: Integrating Evidence of “Blue” Domain



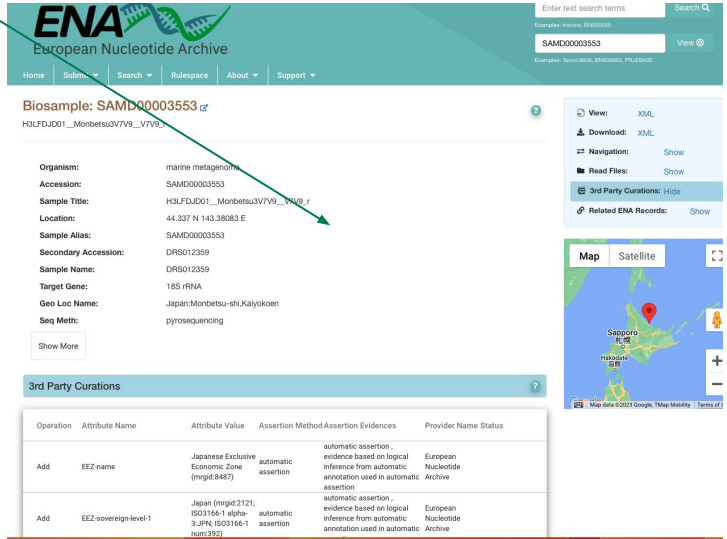
EMBL-EBI

Essential annotations for science & society

Clearinghouse

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
}
```

Thank you

The Blue-Cloud 2026 project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094227. The H2020 project Blue-Cloud received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue Cloud services, Grant Agreement No. 862409.

EMBL-EBI

BACKUPS

Abstract - 10mins lightning talk 7mins taking 3mins Q+A **(REMOVE THIS SLIDE BEFORE POSTING)**

Title: Providing Expanded Contextual Metadata for Biological Samples using Both Geographic and Taxonomic Factors
Authors: All for us.

The European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) is a long-standing database of record for nucleotide sequence data and associated metadata. The ENA has minimal required metadata standards for submitted records to balance the needs of the data generators/submitters and making the metadata as FAIR as possible for downstream users though recommended standards are not always utilised to their full potential and details can be left out.

There are nearly 200,000 marine samples alone within ENA and as part of the BlueCloud project (https://blue-cloud.org/) it was identified that there was a need to enhance the available specific metadata for marine and freshwater samples. By utilising user-provided geographic metadata, we can assert additional contextual metadata to enhance the existing sample records. Approximately 17% of all ENA samples have GPS coordinates. We have used the GPS coordinates to determine additional metadata, for example, the geographic political regions (e.g. countries and EEZs) and environment types (e.g. land and sea), via computational geometry. These were compared to existing submitter metadata provided with these samples. Additionally organism taxonomies were categorised with their likely marine or freshwater environment. The submitter, GPS and taxonomy insights were merged and compared. As expected much of the time there is clear cut metadata agreement, sometimes explainable differences and occasionally harder to explain or understand differences.

For the ENA and similar archives, submitter entered data is the record and so metadata cannot be changed substantively on the primary record without the approval of data owners. The extra contextual metadata is being added to the ELIXIR Contextual Data Clearinghouse see https://elixir-europe.org/internal-projects/commissioned-services/establishment-data-clearinghouse; the metadata will be programmatically available from https://www.ebi.ac.uk/ena/clearinghouse/api/. It will thus be straightforward to

Environment by Keywords

environment_biome

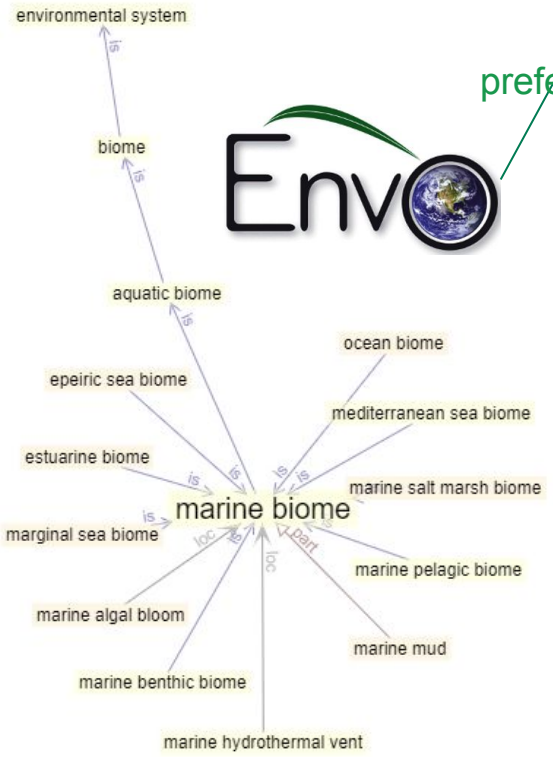
reality

preferred

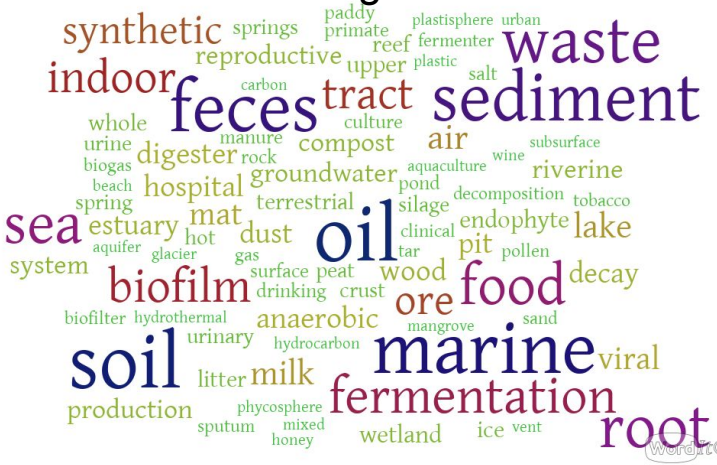
regex

marine/coastal/freshwater

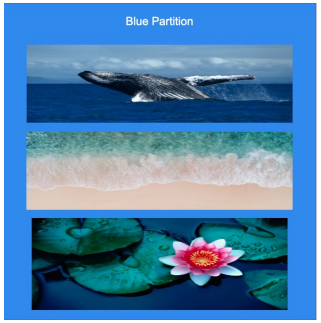
- marine benthic biome
- cropland biome
- freshwater biome
- coral reef
- ENVO:human-associated habitat
- forest
- Bos taurus
- ENVO:2100002
- area of cropland
- grassland
- mouse gut
- soil
- marine biome
- terrestrial biome [ENVO:00000446]
- lentic water body



Unclassified “metagenome” taxonomic terms



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

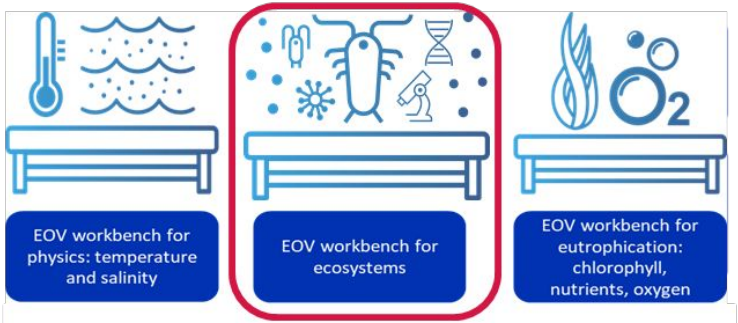
A3. EOVs

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

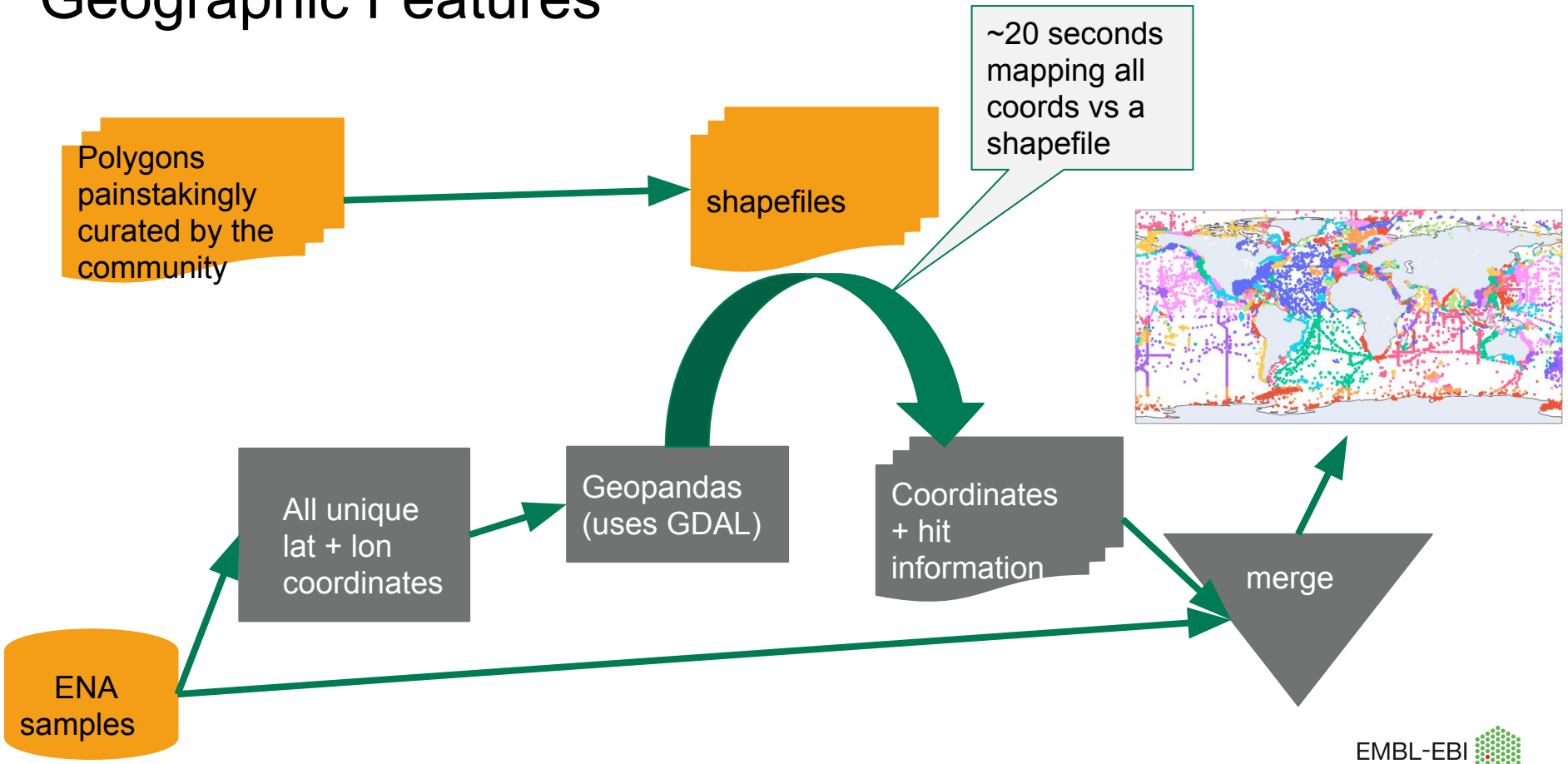


Essential Ocean Variables (EOV) Workbenches

"Developing and testing analytical Blue Cloud WorkBenches to generate high quality data products

EMBL-EBI

Mapping Latitude and Longitude Coordinates to Geographic Features



EMBL-EBI