biodiversa+
European Biodiversity Partnership

Workshop on biodiversity monitoring data interoperability and harmonisation

1st of September – 9am to 12pm CEST

REC

The plenary sessions of this meeting will be recorded and shared on the Biodiversa+ website and Youtube channel

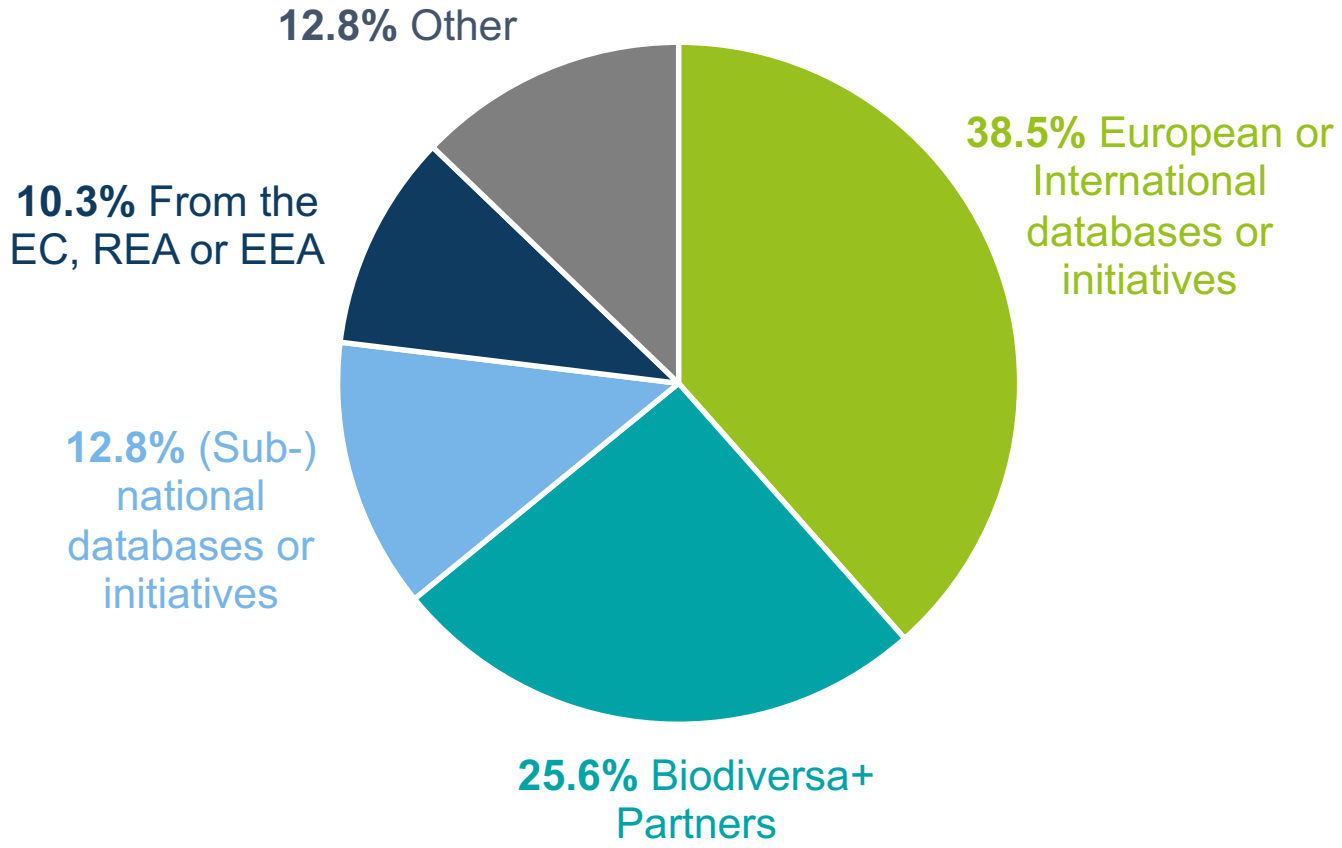
biodiversa+
European Biodiversity Partnership

Welcome words and introduction of the meeting

By Alberto Basset, MUR

Welcome to the workshop on biodiversity monitoring data interoperability and harmonisation

+ 40
Registered participants



Agenda of the workshop

- ✓ Introduction of the meeting and scene setting *by Alberto Basset, MUR and Hilde Eggermont, BelSPO, Biodiversa+ Chair and Coordinator*
- ✓ Data architecture and data flows in EuropaBON & EBVs, *by Dani Villero and Néstor Fernandez, EuropaBON*
- ✓ Presentation of data flows in GBIF and diversifying the GBIF model, *by Tim Hirsch, GBIF Secretariat*
- ✓ Presentation of Meta(data) structure and workflows in LifeWatch ERIC, *by Lucia Vaira, LifeWatch ERIC*
- ✓ DNA (meta)barcoding. BIOSCAN/iBOL Infrastructure developments, *by Rutger Vos, iBOL/BIOSCAN*
- ✓ Open discussion to identify barriers preventing data interoperability (i), factors encouraging data interoperability (ii) and possible support from Biodiversa+ (iii)
- ✓ Concluding words

Aims of Biodiversa+ in this context

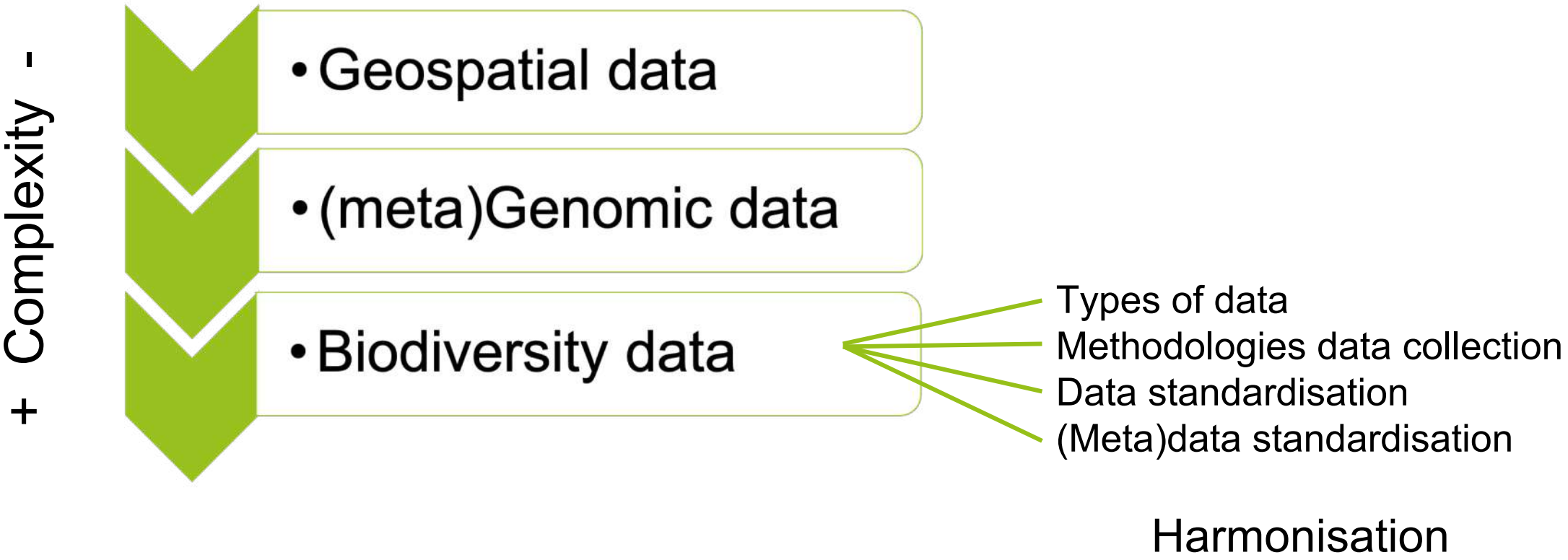
- ✓ Better understanding of data architecture & operating dataflow in European and international initiatives
- ✓ Exchange views with different actors about the level of harmonisation and interoperability the different databases
- ✓ Discuss how Biodiversa+ could help to harmonise databases and improve data interoperability, in order to support the overall objective of transnational biodiversity monitoring

To be developed:

✓ In connection with the biodiversity monitoring priorities at the national, European, international level;
✓ In connection with the running activities in major international initiatives and projects;
✓In connection with the running activities of the ESFRI landscape ERIs, mainly referring to the Biosphere domain.

Harmonise databases and data interoperability

Considering that



Next steps after this workshop

- ✓ Initial reflection of how to take up proposals from this WS in the Biodiversa+ workplan (possibly in collaboration with other key actors)

- ✓ Include the national perspective:
  - Showcases of (sub-)national data architecture plans and data workflows will be identified and mapped.

  - Possibly a follow-up workshop with presentations of (sub-)national showcases will be organised (in the coming months)

- ✓ A public report showcasing these national and global data architectures and workflows will be drawn. The objectives of this report will be to better understand the landscape (including its gaps), identify case studies enabling capacity building and knowledge transfer, identify good practices.

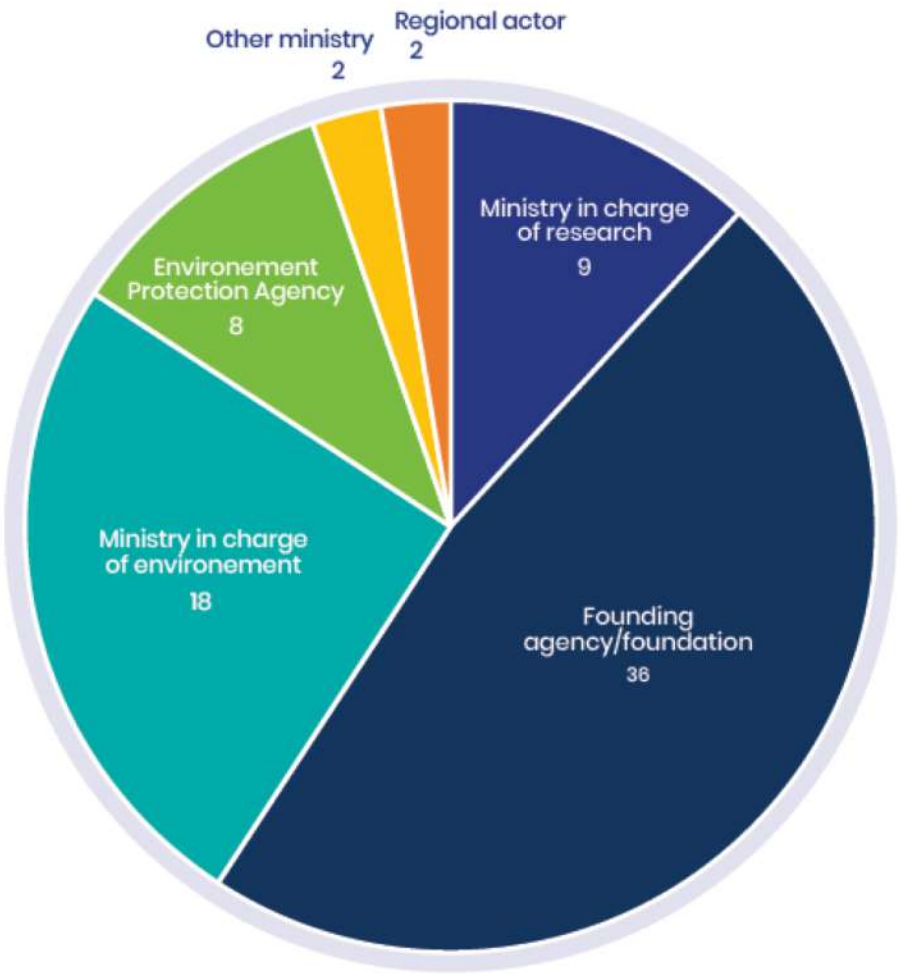
→ Building on relevant EuropaBON activities

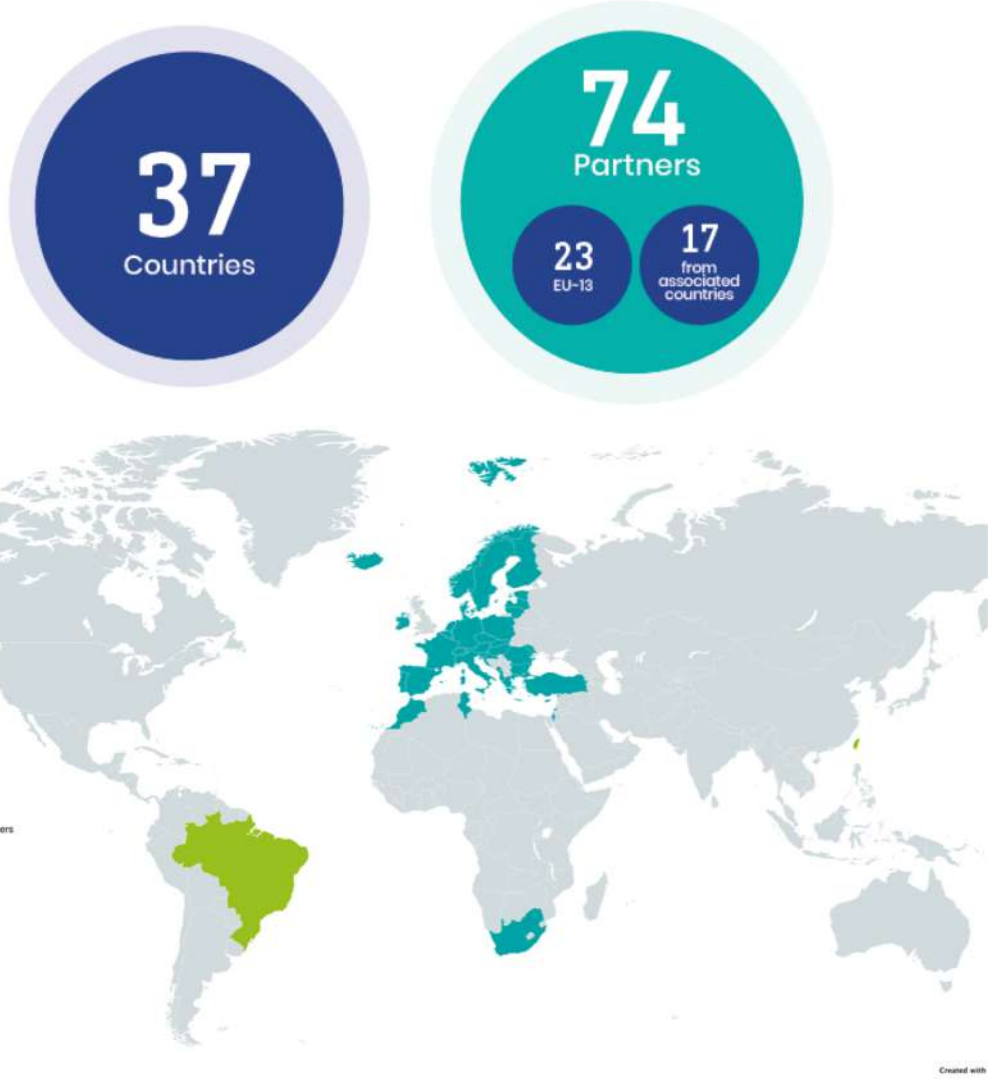
biodiversa+
European Biodiversity Partnership

Presentation of Biodiversa+

By Hilde Eggermont, BelSPO, Biodiversa+ Chair & Coordinator

Biodiversa+ membership



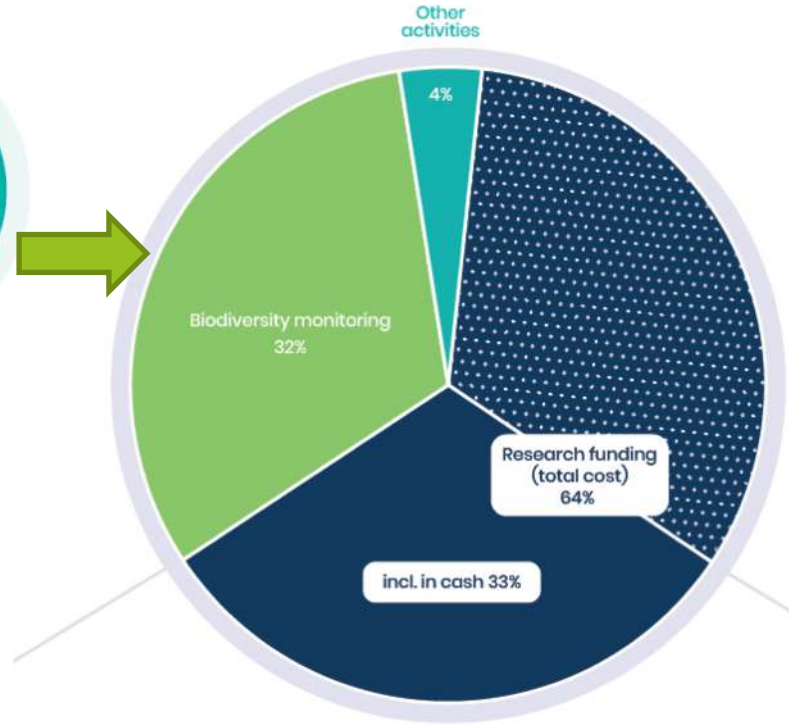


Biodiversa+ portfolio of activities and budget amplitude

Budget of >800 Mio € over 7 yrs, combining in-cash and in-kind resources from its Partners and including 165 Mio € by the European Commission

6 Calls planned

Typically **>42M€** (in cash) per call



Co-funded by the European Union

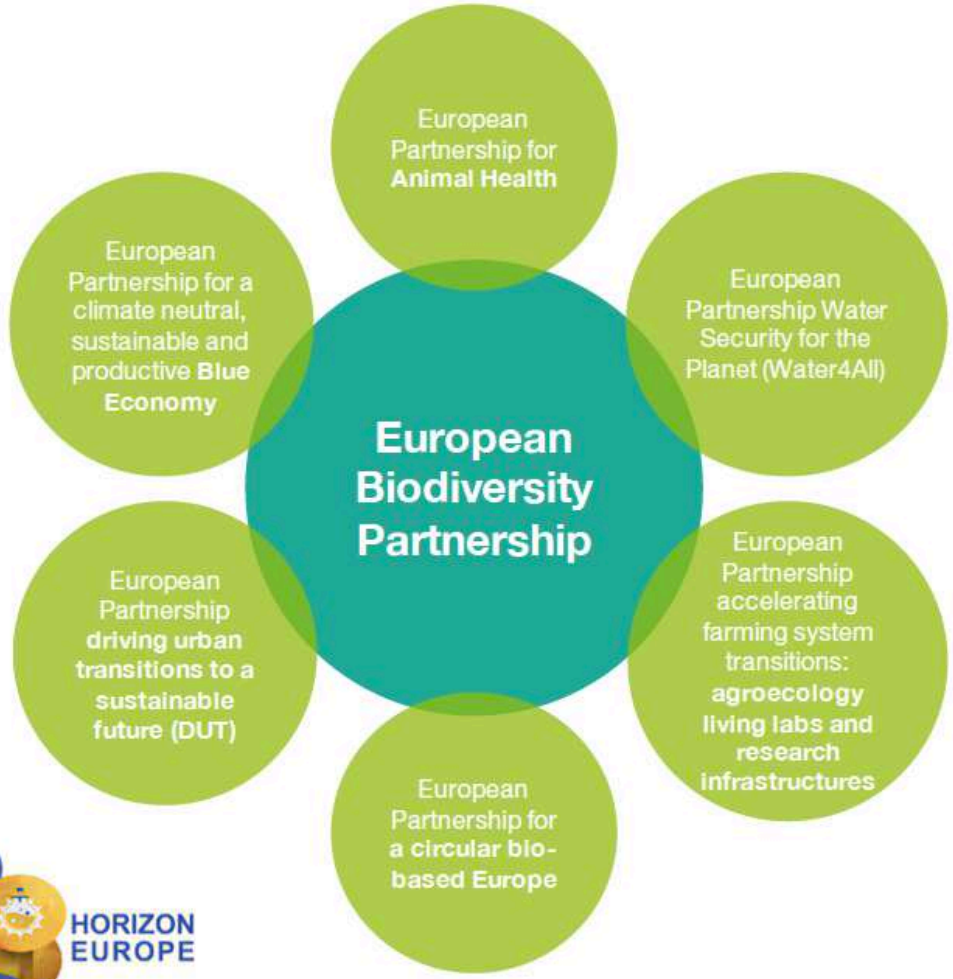
biodiversa+
European Biodiversity Partnership

Aligned with the policy context

EU Biodiversity Strategy 2030:

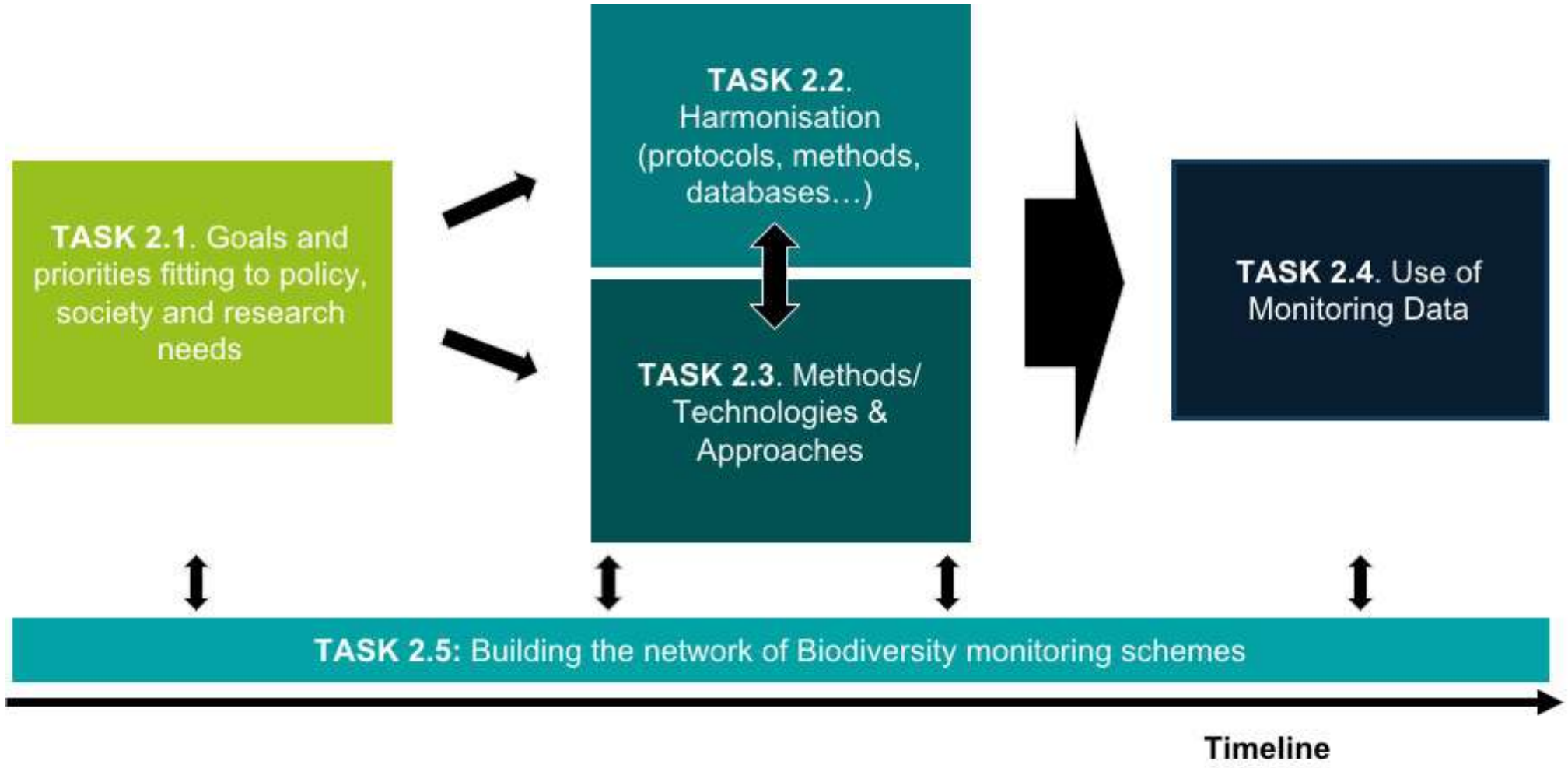
"… making the bridge between science, policy & practice..."

Building on existing and new initiatives



... and many more!

Biodiversity monitoring workstreams of activities



biodiversa+
European Biodiversity Partnership

Work on data architecture and data workflows in EuropaBON and EBVs

By Dani Villero and Nestor Fernandez, EuropaBON

EUROPABON
Europa Biodiversity Observation Network

Mapping and making current European monitoring initiatives accessible via a web-based database.

Alejandra Morán-Ordóñez

David Martí Pino, Dani Villero, Sergi Herrando, Lluís Brotons

CREAF

EXCEL·L SEVERO OCHOA

EBCC
European Bird Census Council
every bird counts

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101007492.

24/08/2022 ECCB Conference, Prague

Why is important to map monitoring efforts & workflows?

Biodiversity monitoring capability across Europe

- what taxa are currently (well) monitored?
- what is the geographic coverage of biodiversity data?
- what is the nature of monitoring programs?
- is the current monitoring capability temporally sustainable?
- cost-efficiency of current monitoring schemes




Credits: Gregoire Dubois

EUROPABON

Why is important to map monitoring efforts & workflows?

Biodiversity monitoring capability across Europe

- what taxa are currently (well) monitored? Gaps
- what is the geographic coverage of biodiversity data? Gaps
- what is the nature of monitoring programs? Bottlenecks
- is the current monitoring capability temporally sustainable? Bottlenecks
- cost-efficiency of current monitoring schemes Bottlenecks

Co-Design of EBVs, EESVs to track goals of (e.g.): EU restoration law, Biodiversity Strategy, etc.



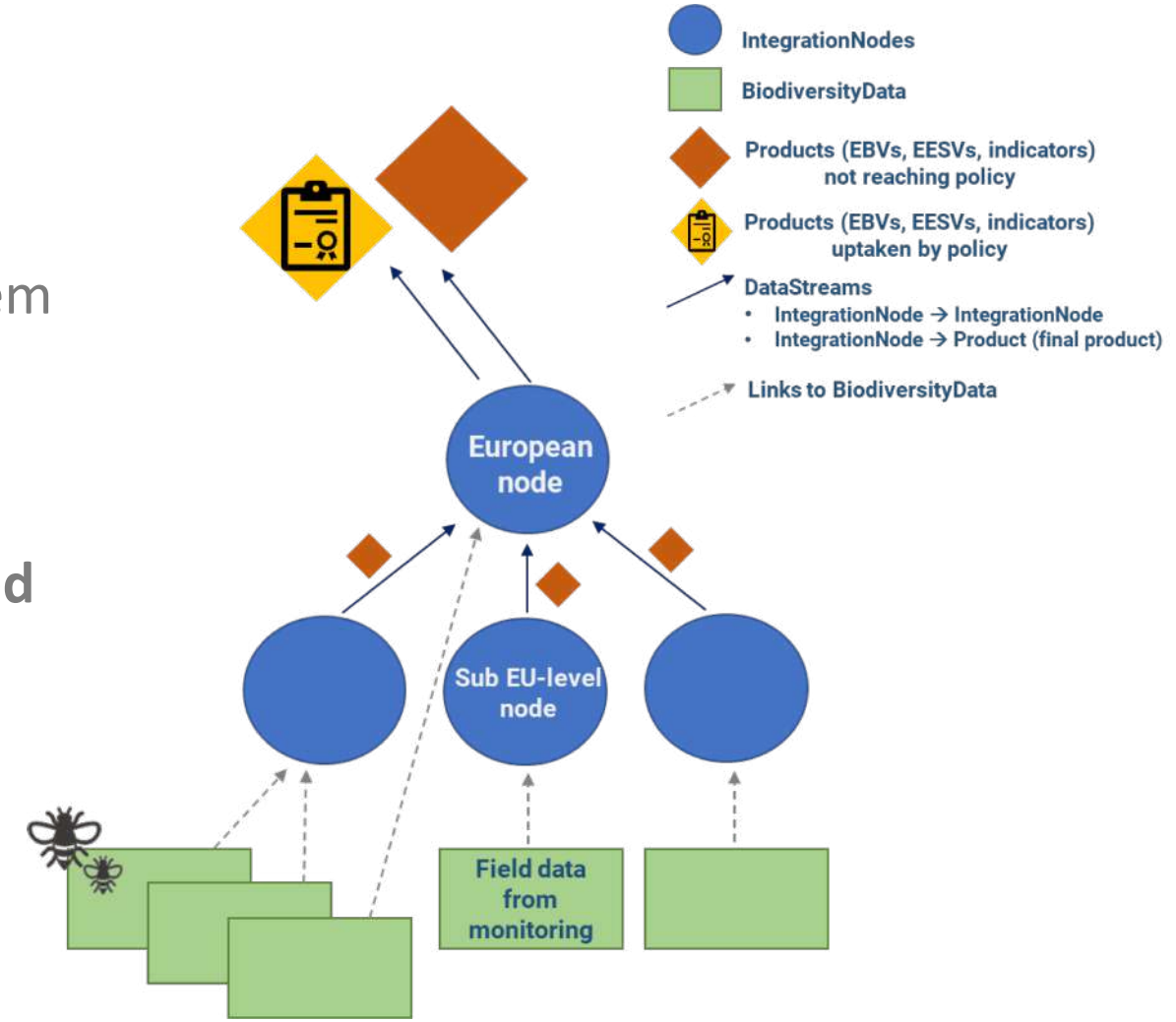

Credits: Gregoire Dubois

EuropaBON: Inventory of monitoring initiatives

Initiatives/projects/programs – NETWORKS - that **integrate biodiversity data** from monitoring schemes and process them to **produce** aggregated biodiversity data or Essential Biodiversity Variables (**EBVs**), Essential Ecosystem Services Variables (**EESVs**) and/or other European policy-relevant i**ndicators**.

Priority on **monitoring efforts that are coordinated at a transnational level** (combining bottom up and top down approach).

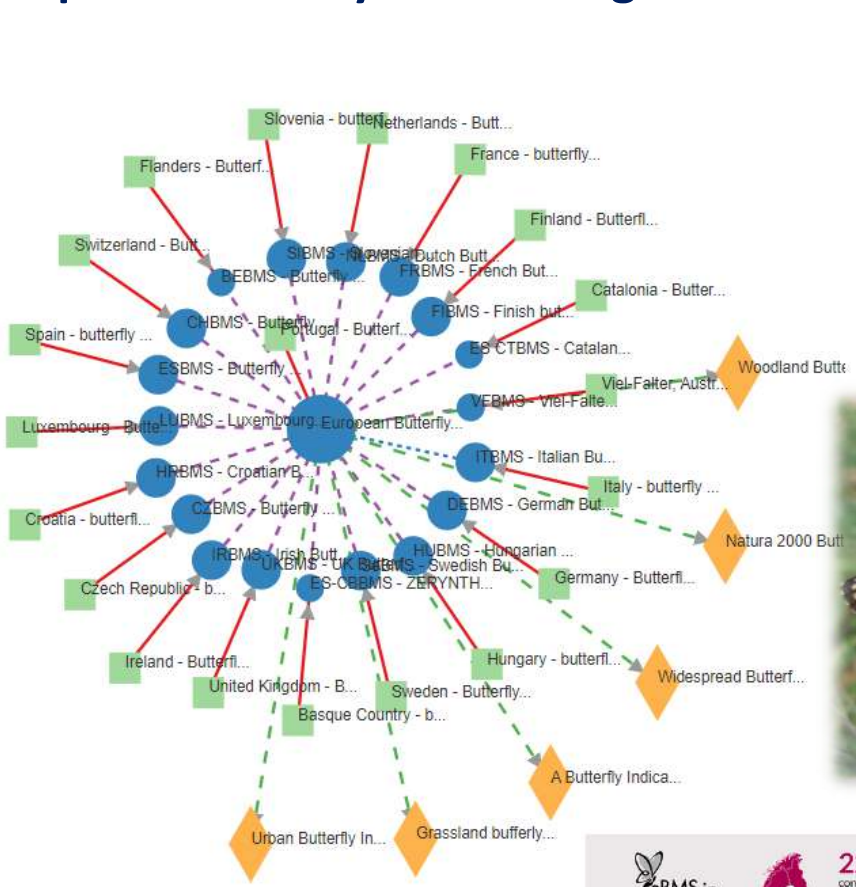
Focus on **data workflows** not only on data collection.



EUROPABON

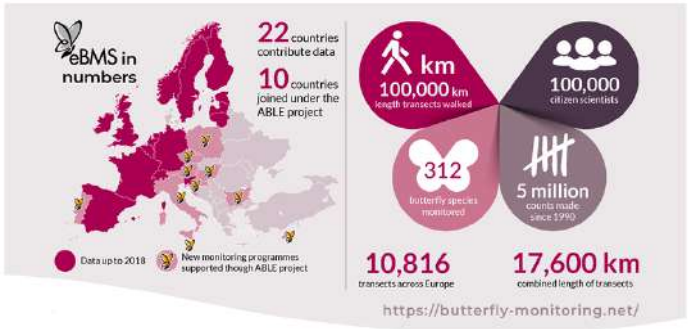
Integration initiatives: examples

European Butterfly Monitoring Scheme



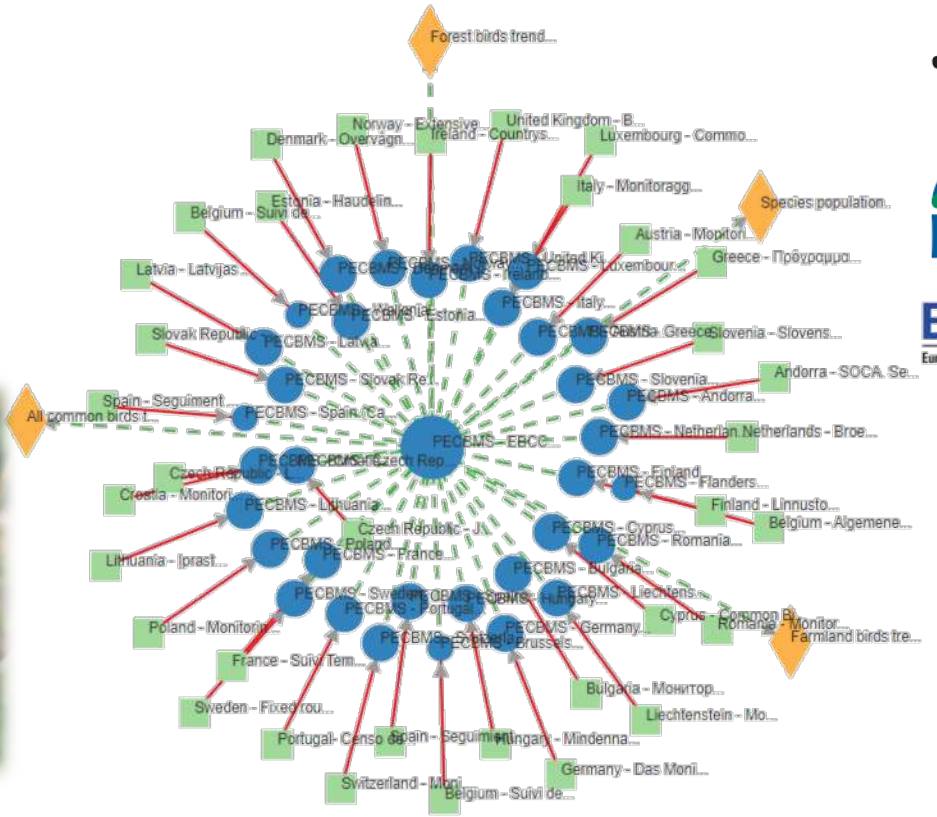
Butterfly CONSERVATION EUROPE

UK Centre for Ecology & Hydrology



EUROPABON

PanEuropean Common Bird Monitoring Scheme

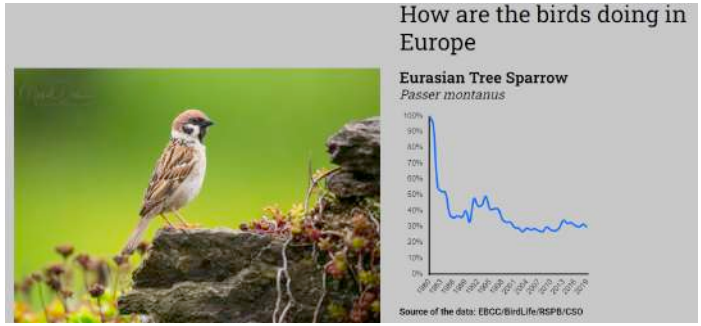


Pan European Common Bird Monitoring Scheme

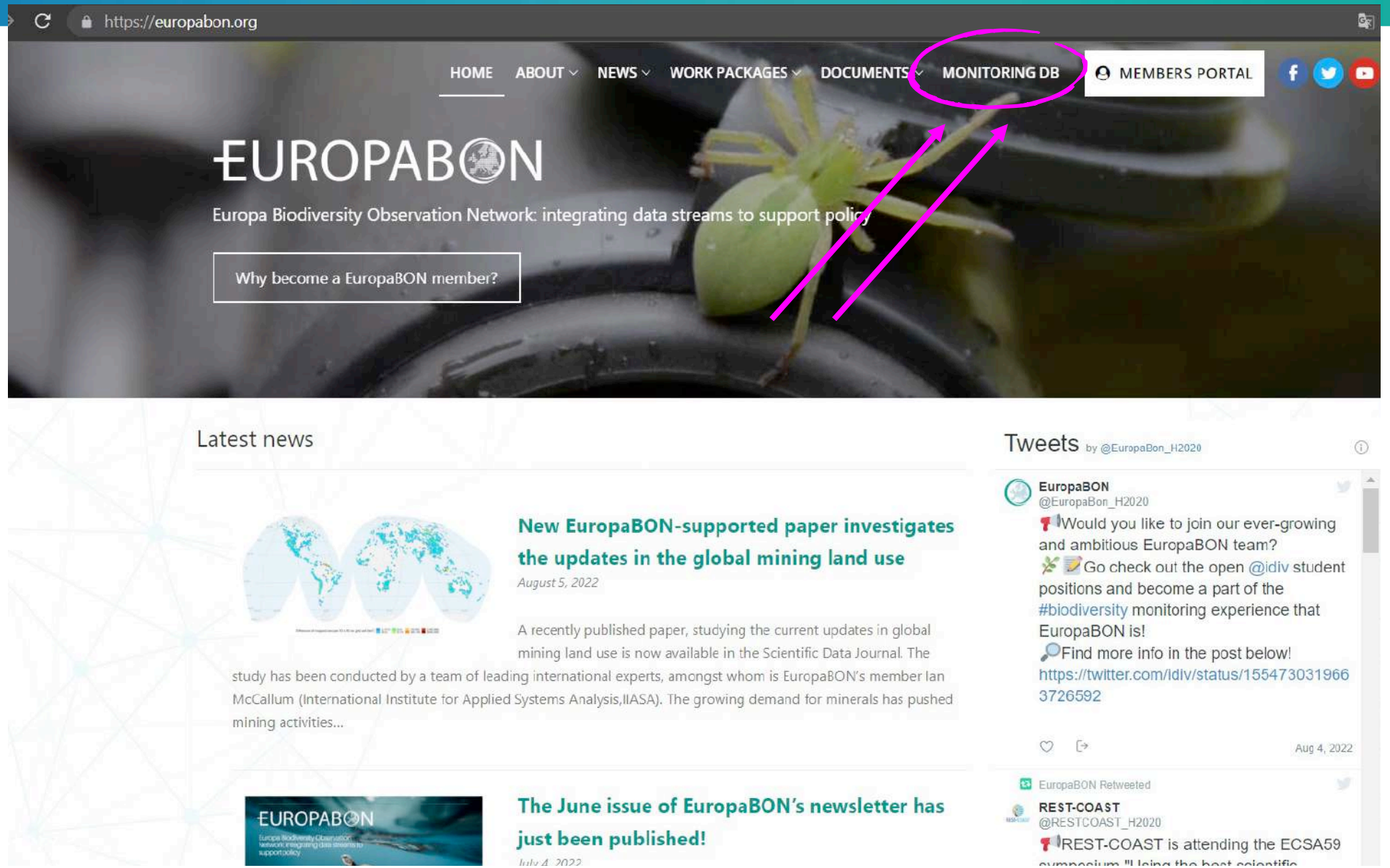
PECBMS

EBCC European Bird Census Council every bird counts

BirdLife INTERNATIONAL



Web-based database (open): access via europabon.org



EUROPABON

https://monitoring.europabon.org/

Web-based database (open)

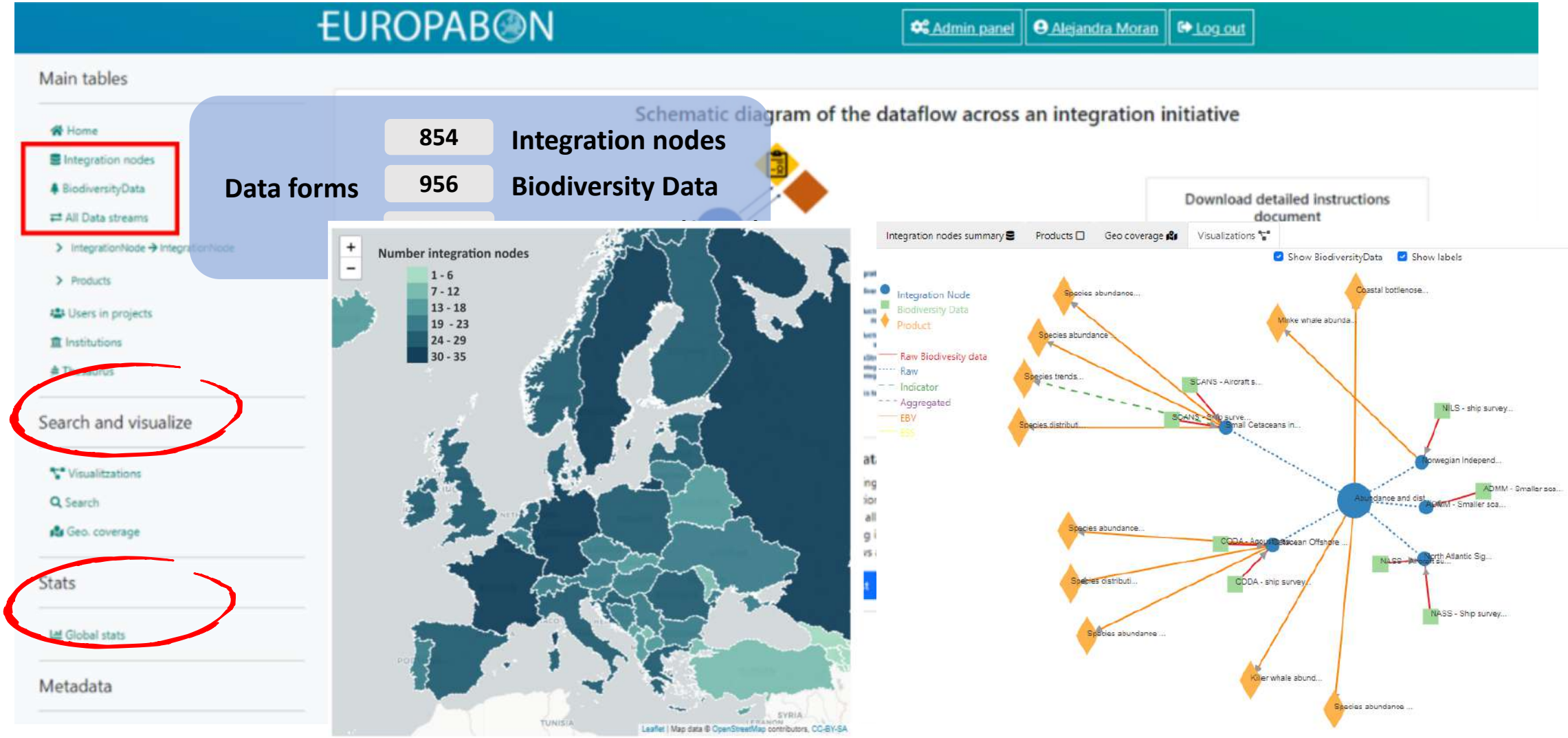
Website open: anyone can query content (permissions required for data input)

EUROPABON

https://monitoring.europabon.org/

Web-based database (open)

Website open: anyone can query content (permissions required for data input)



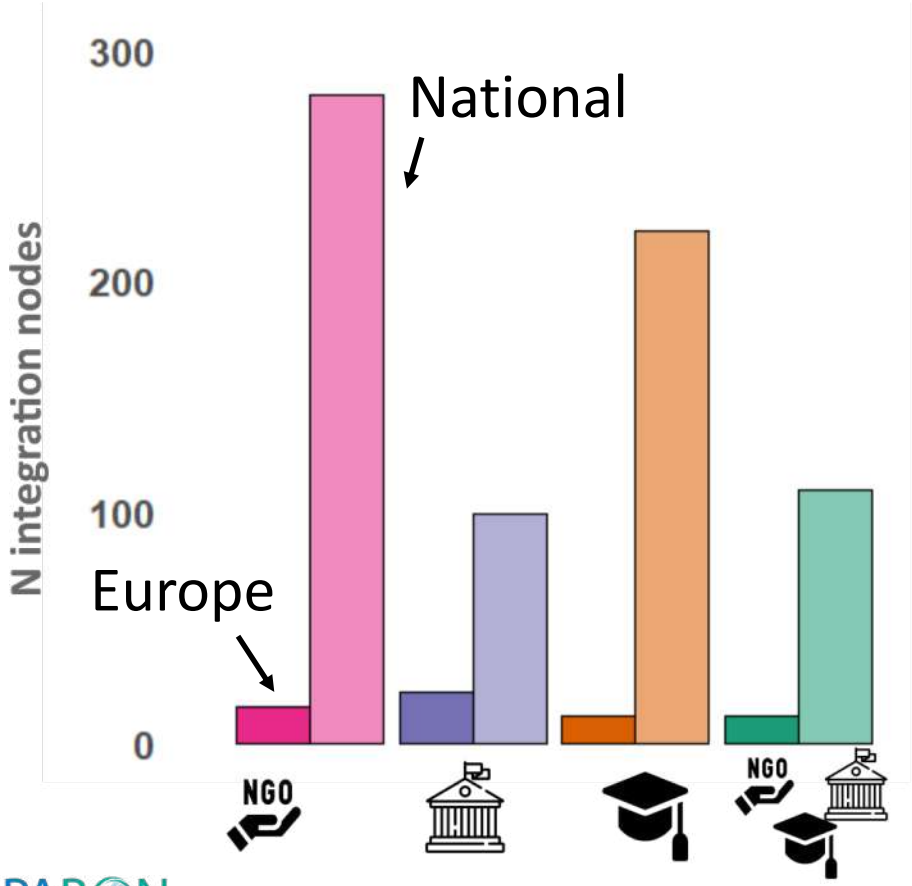
https://monitoring.europabon.org/

EUROPABON

Results: integration nodes

Who is integrating the data?

NGOs process and integrate most biodiversity data across Europe (38% of National integration nodes), followed by research institutions (35%) and governmental institutions (19,8%)



NGOs (e.g.): IUCN, Birdlife International, Wetlands International European Mammal Foundation, Ornithological Societies

Supra-/Governmental (e.g.): Joint Research Center (EC), Croatian Agency for Environment and Nature, Swedish Species Information Centre (SSIC), Centre Suisse de Cartographie de la Faune (CSCF)

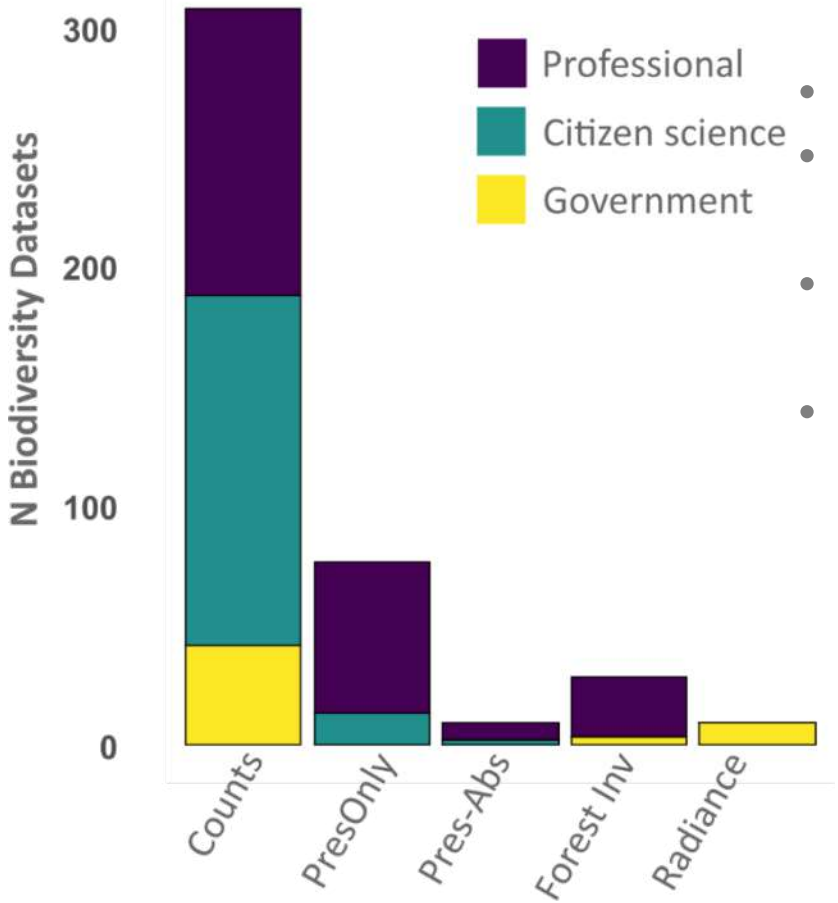
Research bodies (e.g.): Research Institute for Nature and Forest (INBO, BE), Institute of Zoology (Academy of Sciences of Moldova), Sapienza University of Rome (IT), University of Aarhus (DK), Museu de Ciències Naturals de Granollers (ES)

Number of integration nodes coordinated by NGOs, Governmental or supra-governmental institutions, Research bodies (including Universities) or by a combination of these (EuropaBON database numbers by end November 2021). In each group, the left column corresponds to European nodes and the right column to National nodes.

EUROPABON

Results: biodiversity data

Origins and data types of biodiversity monitoring data



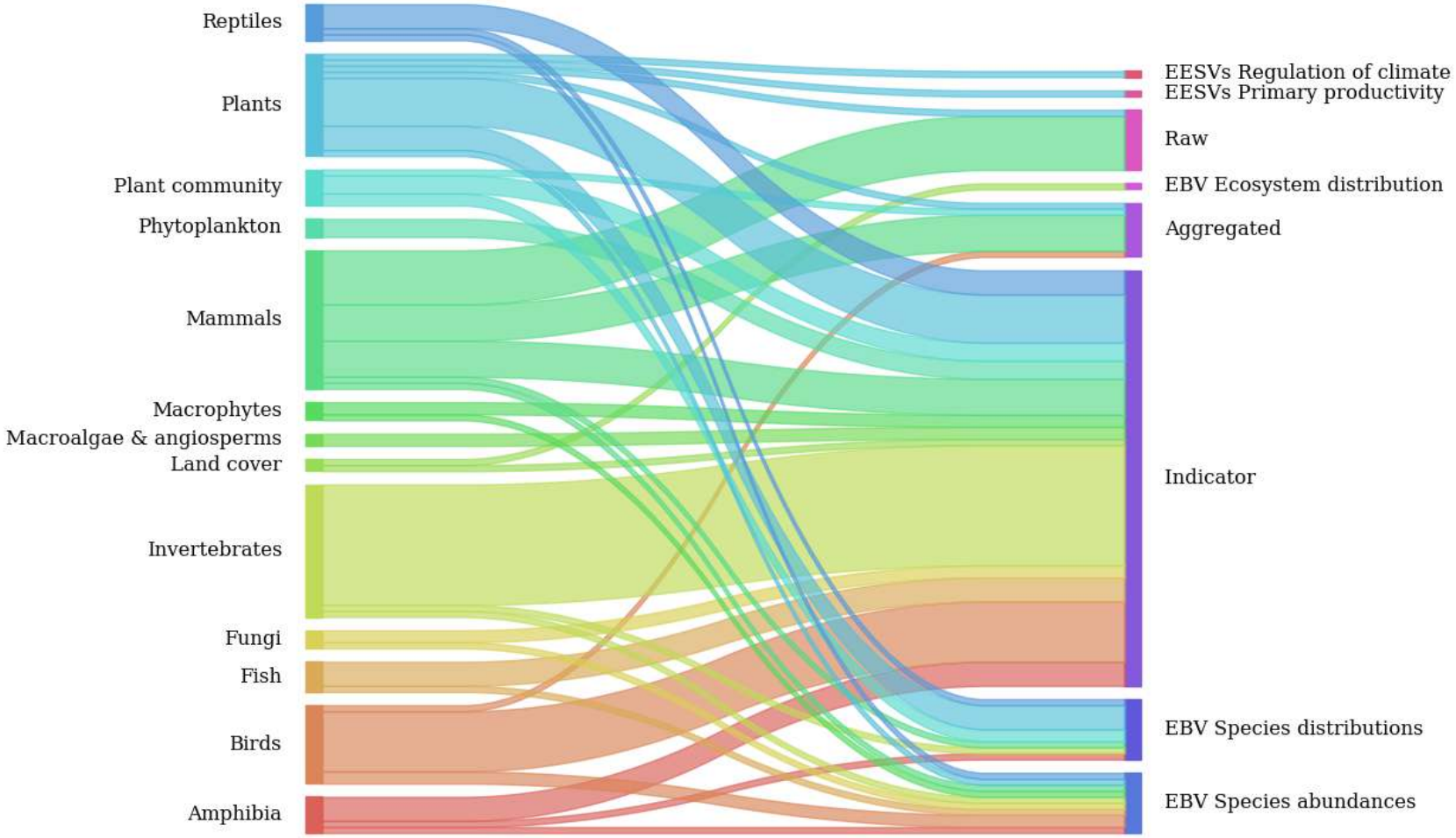
- **62 % corresponded to systematic monitoring programs**
- **Count data** for abundance estimation were the most common type of data collected **(80% of data)**
- The purpose of most biodiversity and environmental data sources gathered is **to advance ecological knowledge of species**
- **Only** a small fraction of all biodiversity and environmental data sources collected are **open access (10.6%)**

Number of Biodiversity Datasets categorized by their data origin (citizen science, governmental authorities, professional monitoring) and data type (counts, presence-only, Presence-Absence, Forest Inventory data and spectral radiance - remote sensing)

EUROPABON

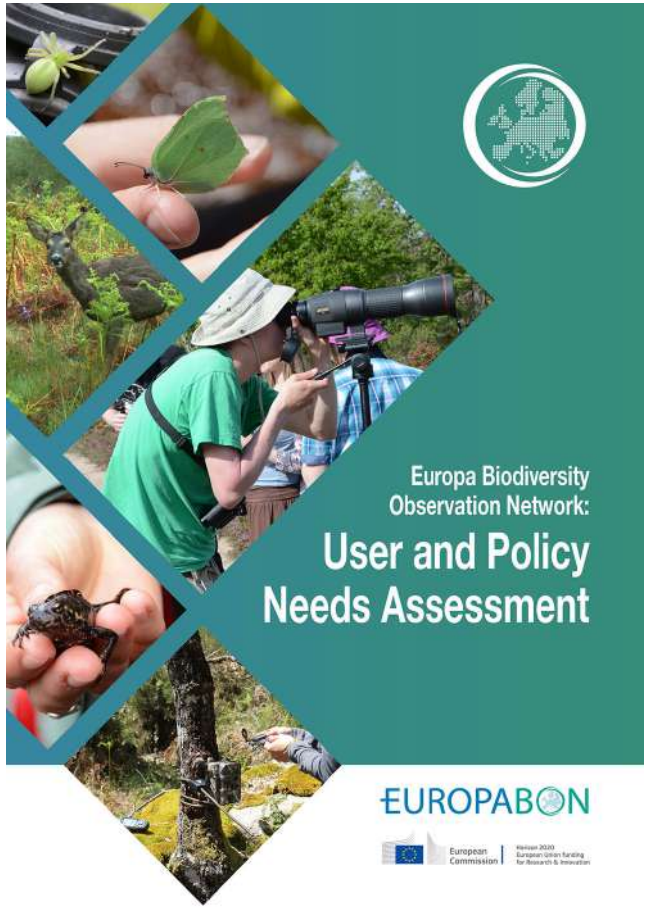
Data flows: European level

Only around 22 % of data flows have been documented to generate EBVs and EESVs: species distributions 55%, species abundance 17%, measures of taxonomic/phylogenetic diversity 7% and EESVs <1%

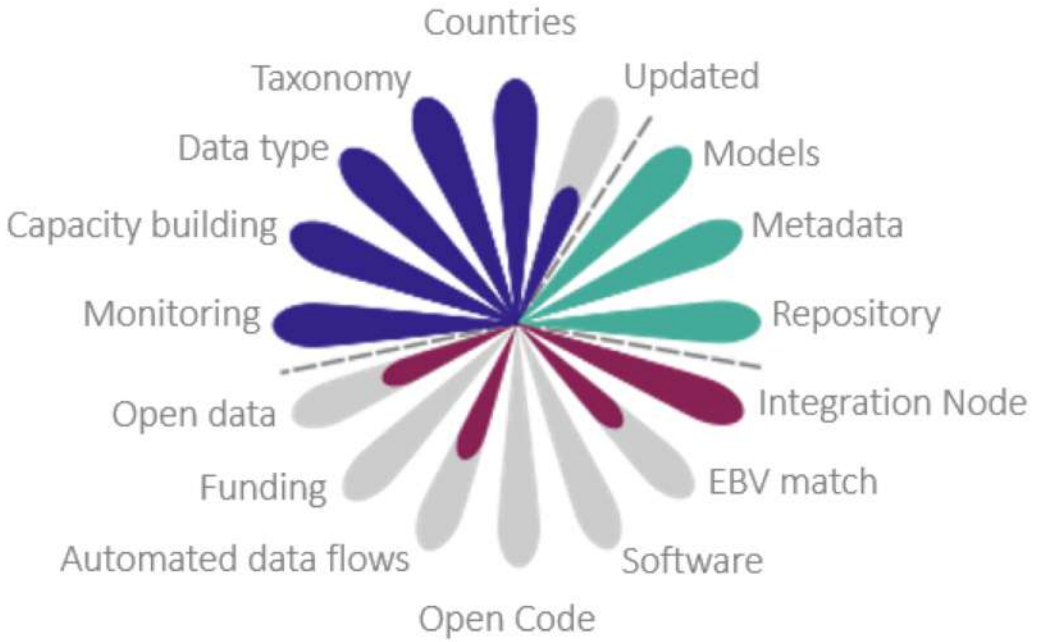


*Data flows of raw, aggregated or processed data (EBVs, ESSVs, or indicators) at the **European level** by taxonomic group*

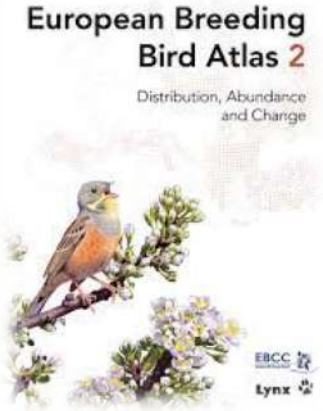
Next developments: identification of gaps & bottlenecks



EBV: species distributions all terrestrial birds



EBBA2
2013-2017



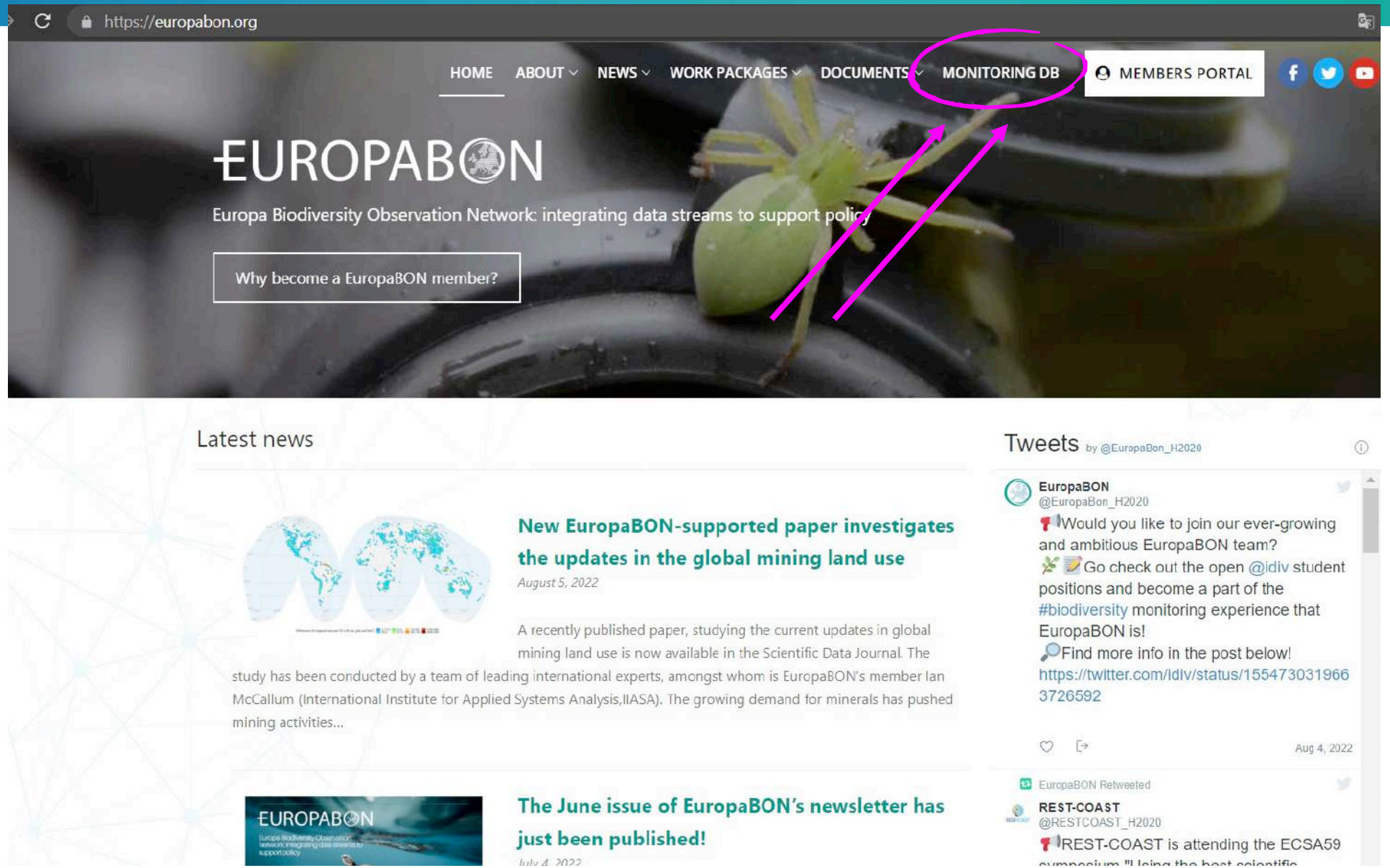
- Data collection & sampling
- Models, Interoperability, IT infrastructure
- Data Integration
- Bottlenecks

Thanks for your attention!

https://monitoring.europabon.org/
d.villero@creaf.uab.cat

EUROPABON
CREAF
EXCEL·LÈNCIA SEVERO OCHOA
EBCC European Bird Census Council every bird counts
European Commission
Horizon 2020 European Union funding for Research & Innovation

Web-based database (open): access via europabon.org



EUROPABON

https://monitoring.europabon.org/

German Centre for Integrative Biodiversity Research (iDiv)
Halle-Jena-Leipzig

Mainstreaming Essential Biodiversity Variables:

A Data Portal and a standard for spatiotemporal biodiversity data

Néstor Fernández, *Christian Langer, Luise Quoss, Miguel Fernández, José Valdez, Henrique M. Pereira*

portal.geobon.org

Group on Earth Observations
Biodiversity Observation Network

iDiv is a research centre of the
Deutsche Forschungsgemeinschaft

Mainstreaming Essential Biodiversity Variables

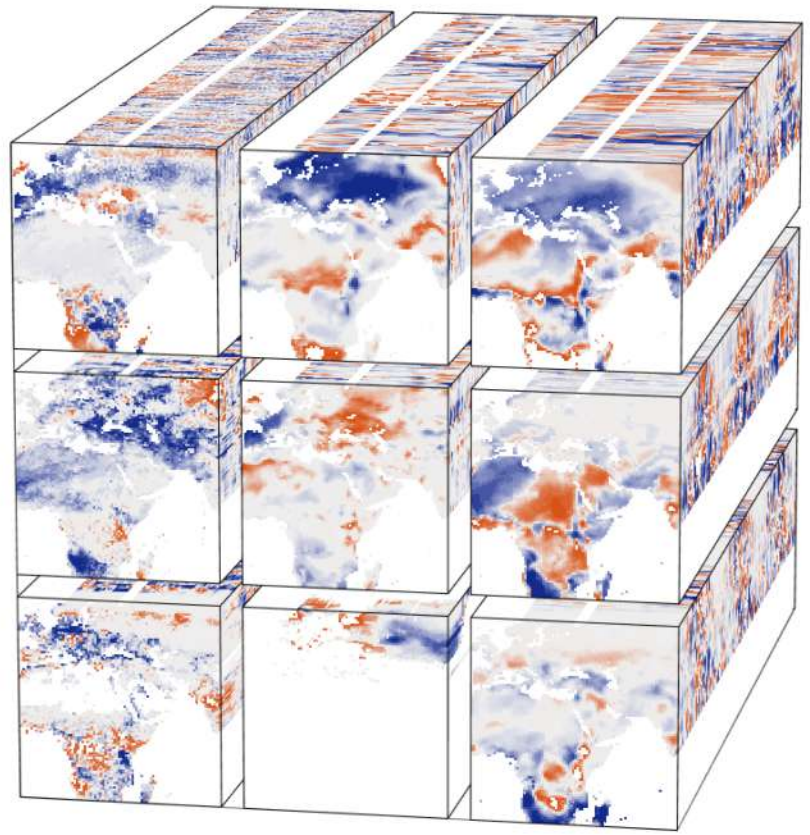
Biodiversity datasets are multidimensional but no standard exists for their organization

Many biological entities
Species, ecosystems, types of communities…

Different scenarios
(E.g. SSPs + RCPs)

Several biodiversity metrics
(E.g. alpha-diversity relative to pristine; to year; etc.)

Various metric estimates
(including uncertainty)

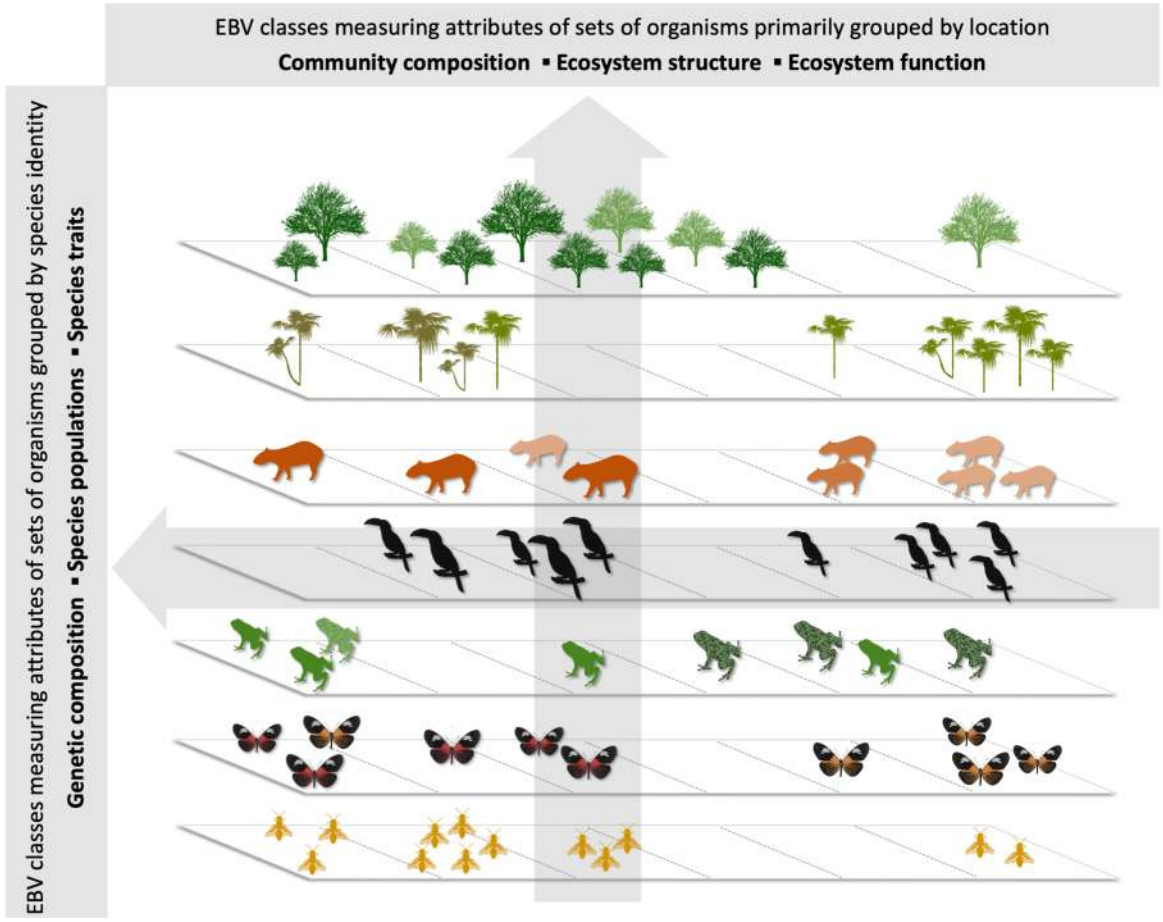


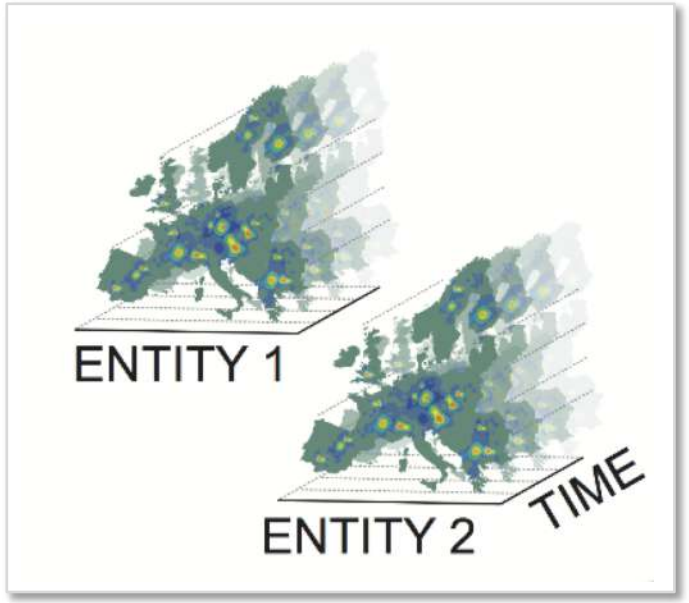
Picture produced by the Earth System Data Lab
www.earthsystemdatalab.net

iDiv

Mainstreaming Essential Biodiversity Variables

Comparable data outputs across EBV classes





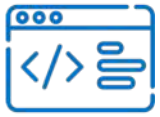
iDiv

Fernández, Ferrier, … & Pereira (in review)

Mainstreaming Essential Biodiversity Variables

Comparable data outputs across EBV classes

How should we organize such disparity of data?

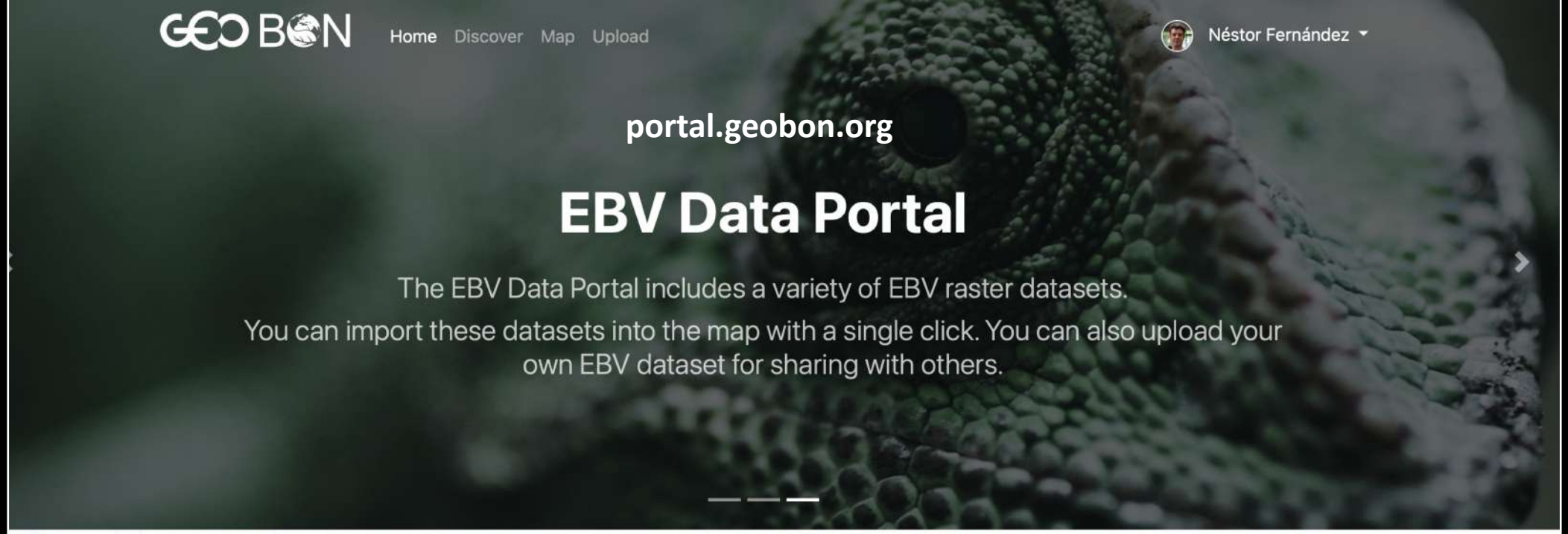


Consistent metadata

Interoperable

Easy access

iDiv



Portal statistics

Entity Type

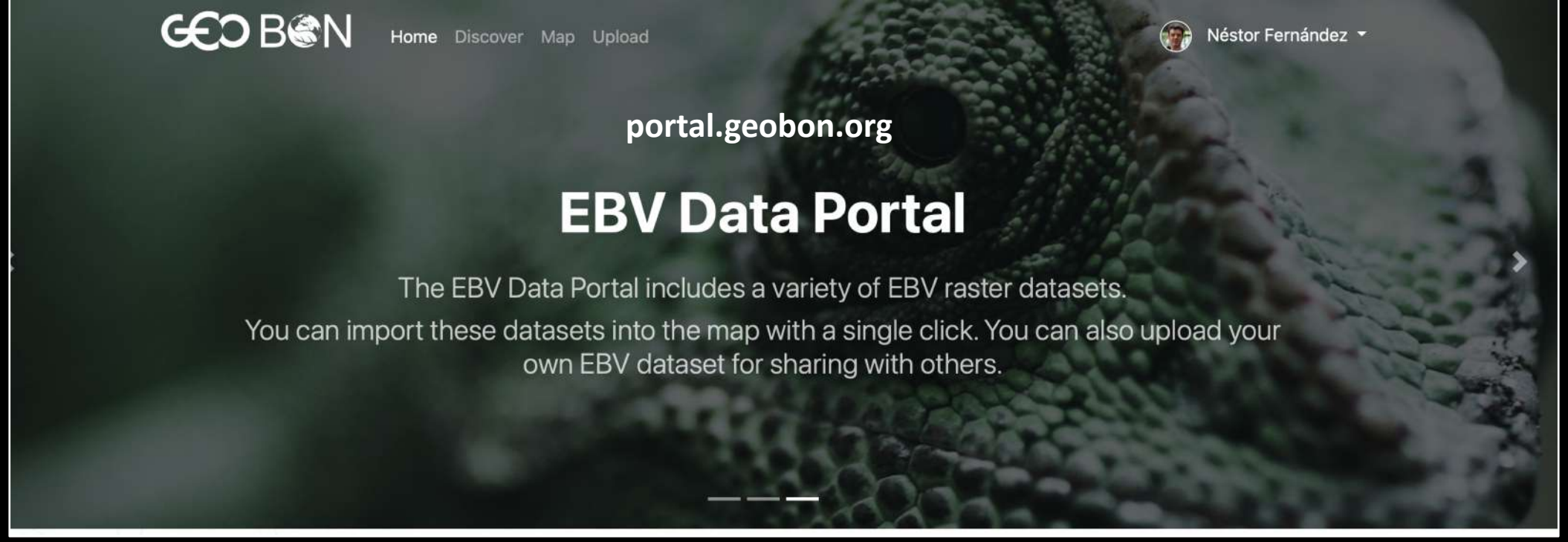
Spatial Domain

Environmental Domain

25 EBV datasets in total

Show All datasets

Search datasets: Search datasets...

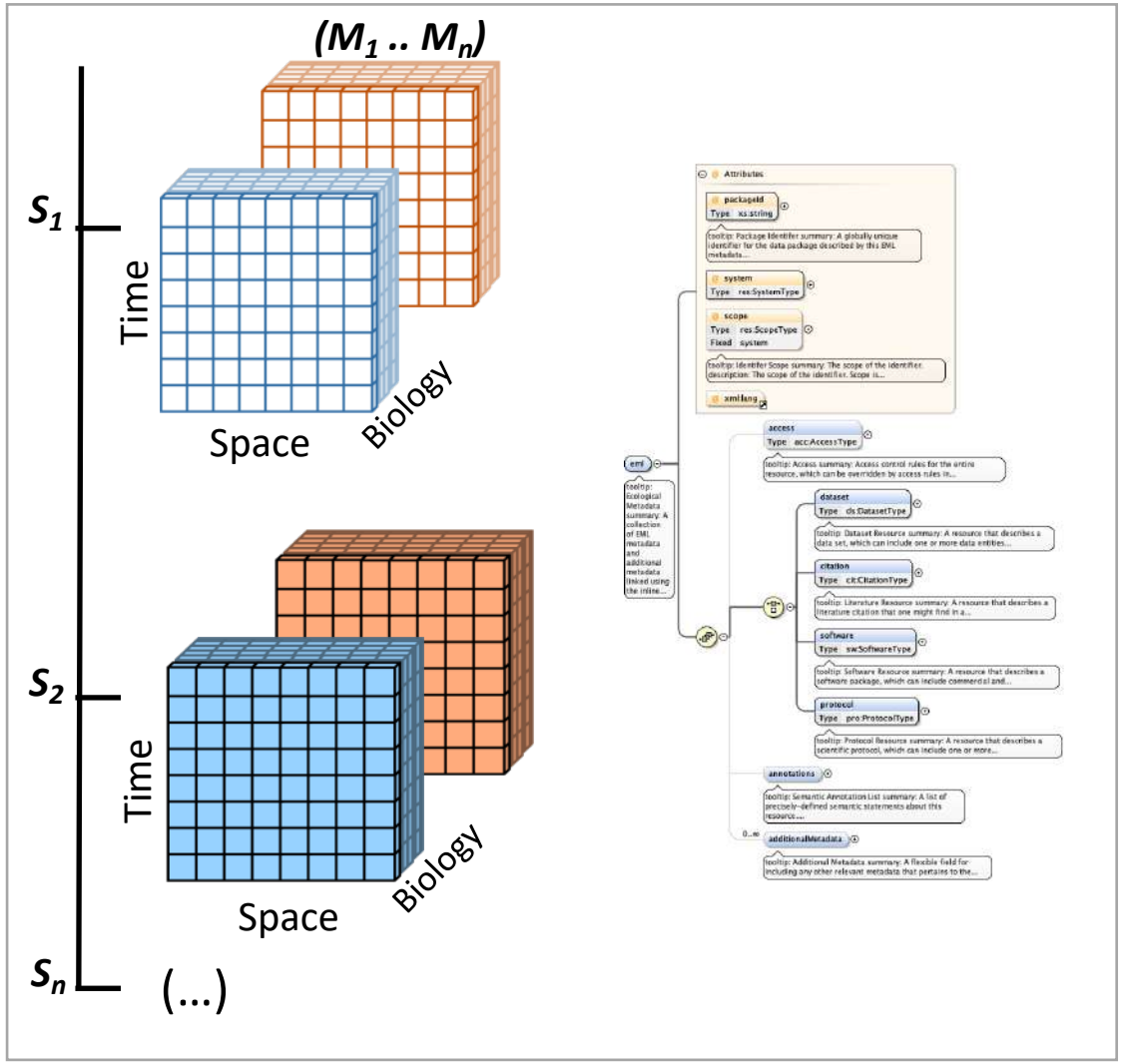


Datasets:

- Consistently structured across **thematic, spatial and temporal dimensions**
- **Consistently documented** in a way that maximizes usability
- **Traceable** (both resources and production pipelines)

Components of the EBV Cube Standard

- **NetCDF** specification with a **unified hierarchical structure** for organizing EBV data
- A **minimum information** specification using **ACDD** terms (with translation into **EML**)
  - *Self-described*
  - *Compliant with FAIR and GEOSS-DMP*



iDiv

Components of the EBV Cube Standard

NetCDF file structure

Groups

Scenario
(optional)

$S_1 \dots S_n$

Metric
(E.g. alpha-diversity)

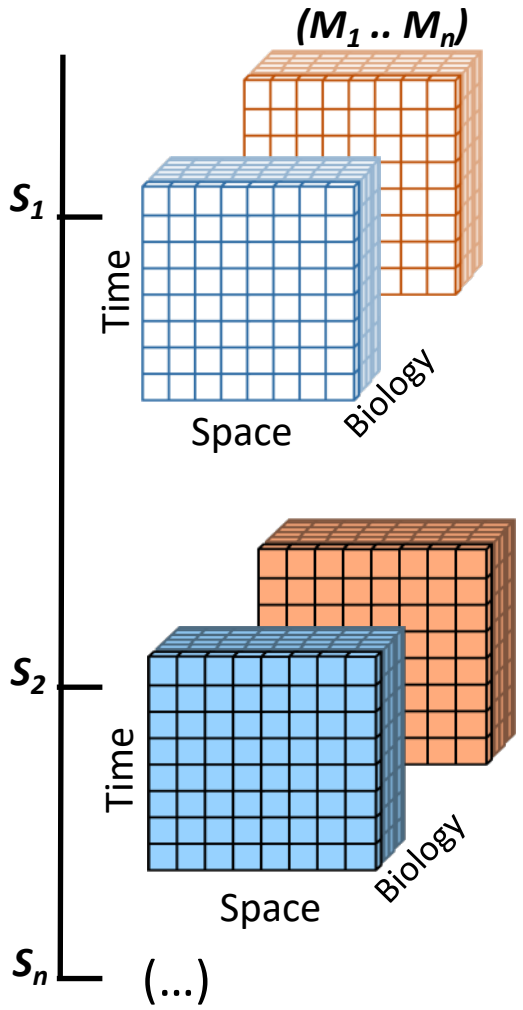
$M_1 \dots M_n$

Estimate
(E.g. relative to baseline)

$E_1 \dots E_n$

Dimensions

Entity x Space x Time



iDiv

Components of the EBV Cube Standard

Embedded metadata

ACDD 1.3

Attribute Convention for Data Discovery

1. Attribution
2. Publication
3. History and processing level
4. Geospatial and temporal attributes
5. Data, processing and models

EBV terms

6. EBV class, name, entities
7. Scenario and metric descriptions


```
"source": "sfsfsf",
"coverage_content_type": [
    "referenceInformation"
],
"processing_level": "N/A",
"project": "N/A",
"project_url": "N/A",
"creator": {
    "creator_name": "Abel Ramoelo",
    "creator_email": "N/A",
    "creator_institution": "Aarhus University",
    "creator_country": "Denmark"
},
"contributor_name": "N/A",
"license": "https://creativecommons.org/licenses/by-nc-sa/4.
"publisher": {
    "publisher_name": "Christian Langer",
    "publisher_email": "christian.langer@idiv.de",
    "publisher_institution": "German Centre for Integrative B
    "publisher_country": "Germany"
},
"ebv": {
    "ebv_class": "Species populations",
    "ebv_name": "Species distributions"
},
"ebv_entity": {
    "ebv_entity_type": "Communities",
    "ebv_entity_scope": "sfsfsfsf",
    "ebv_entity_classification_name": "N/A",
    "ebv_entity_classification_url": "N/A"
},
"ebv_metric": {
    "ebv_metric_1": {
        ":standard_name": "fsfsf",
        ":long_name": "sfsfsf",
        ":units": "sfsfsf"
    }
```


iDiv

Components of the EBV Cube Standard

Interoperable information standards

Attribute Convention for Data Discovery

- ❖ Standard for earth sciences data

Ecological Metadata Language

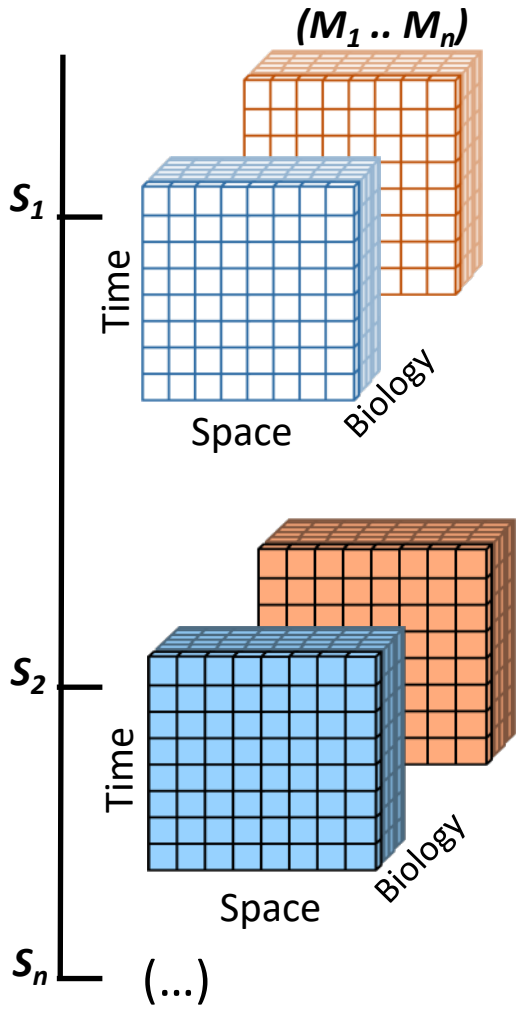
- ❖ LTER, NEON, DataOne…

.json

- ❖ Web / server applications

Implementation of multiple taxonomies

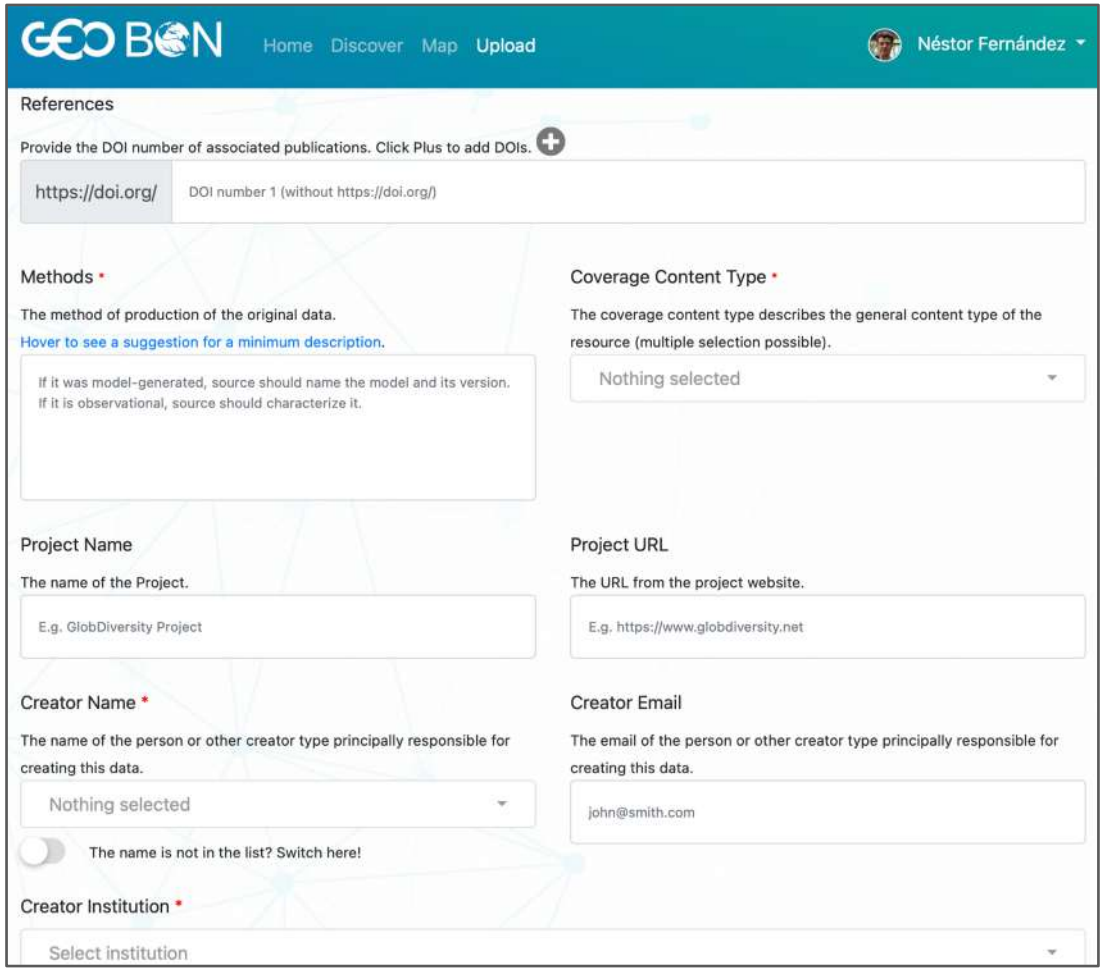
- ❖ GBIF backbone taxonomy



iDiv

EBV Cube tools

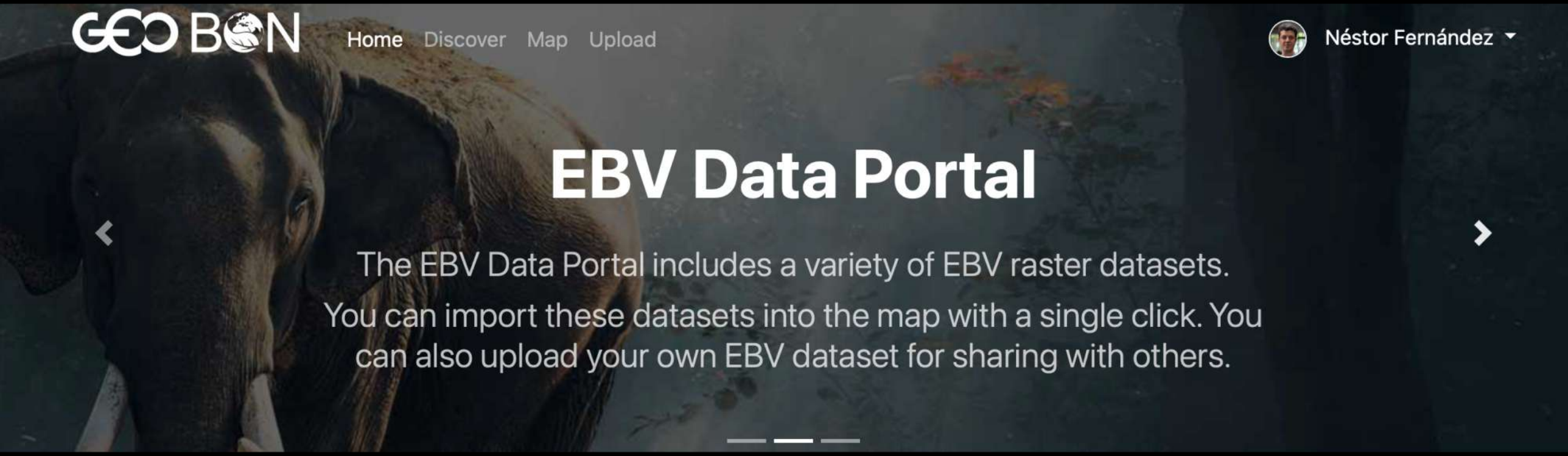
EBV Portal Upload tool



ebvcube

- Create NetCDF-EBV cubes
- Access the data
- Basic descriptions
- Visualization

iDiv



https://portal.geobon.org

Q&A

biodiversa+
European Biodiversity Partnership

Presentation of data architecture and data workflows in GBIF

By Tim Hirsch, Deputy Director, GBIF Secretariat

Data flows in GBIF and diversifying the GBIF data model

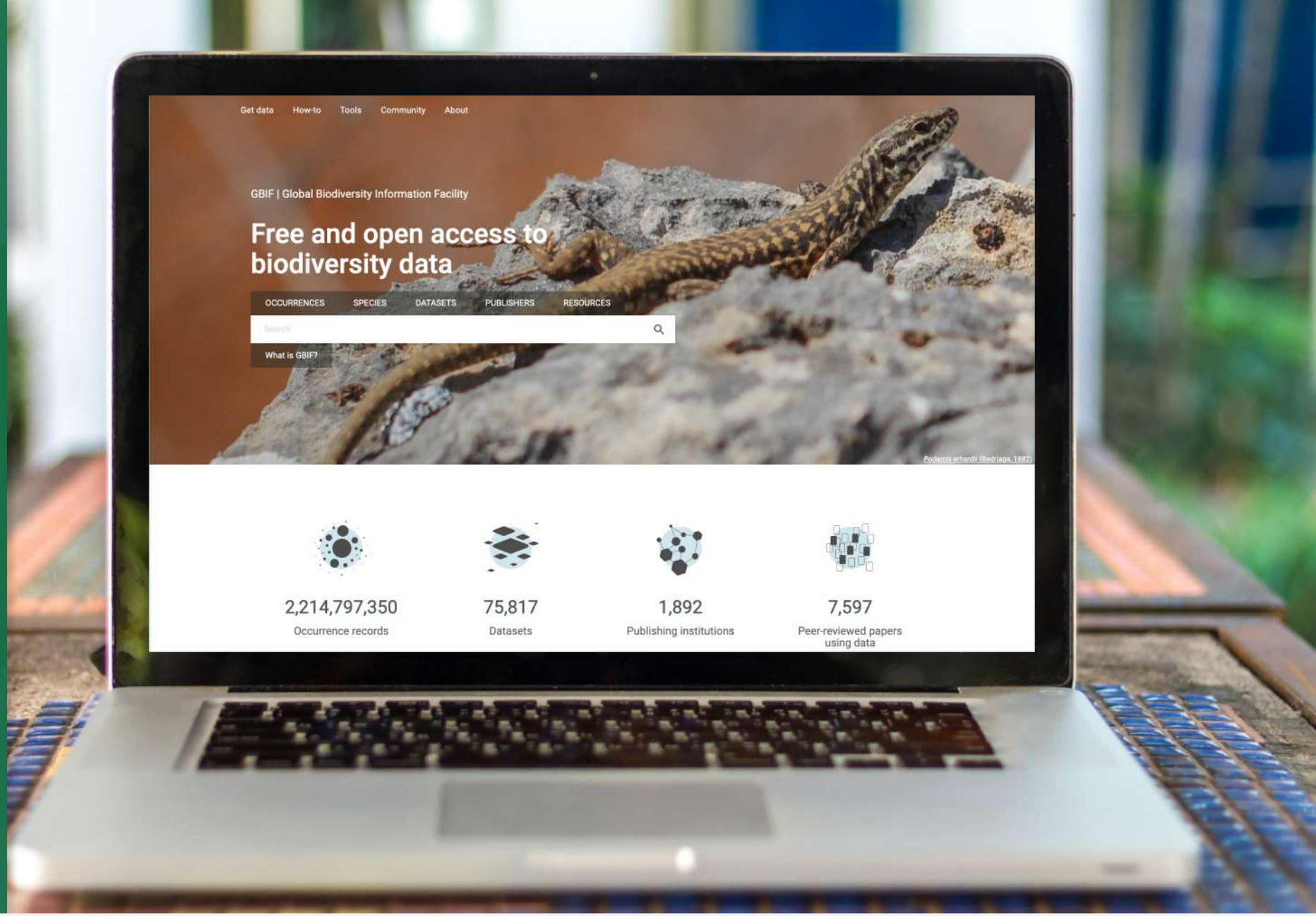
Tim Hirsch
Deputy Director, GBIF Secretariat

GBIF Global Biodiversity Information Facility

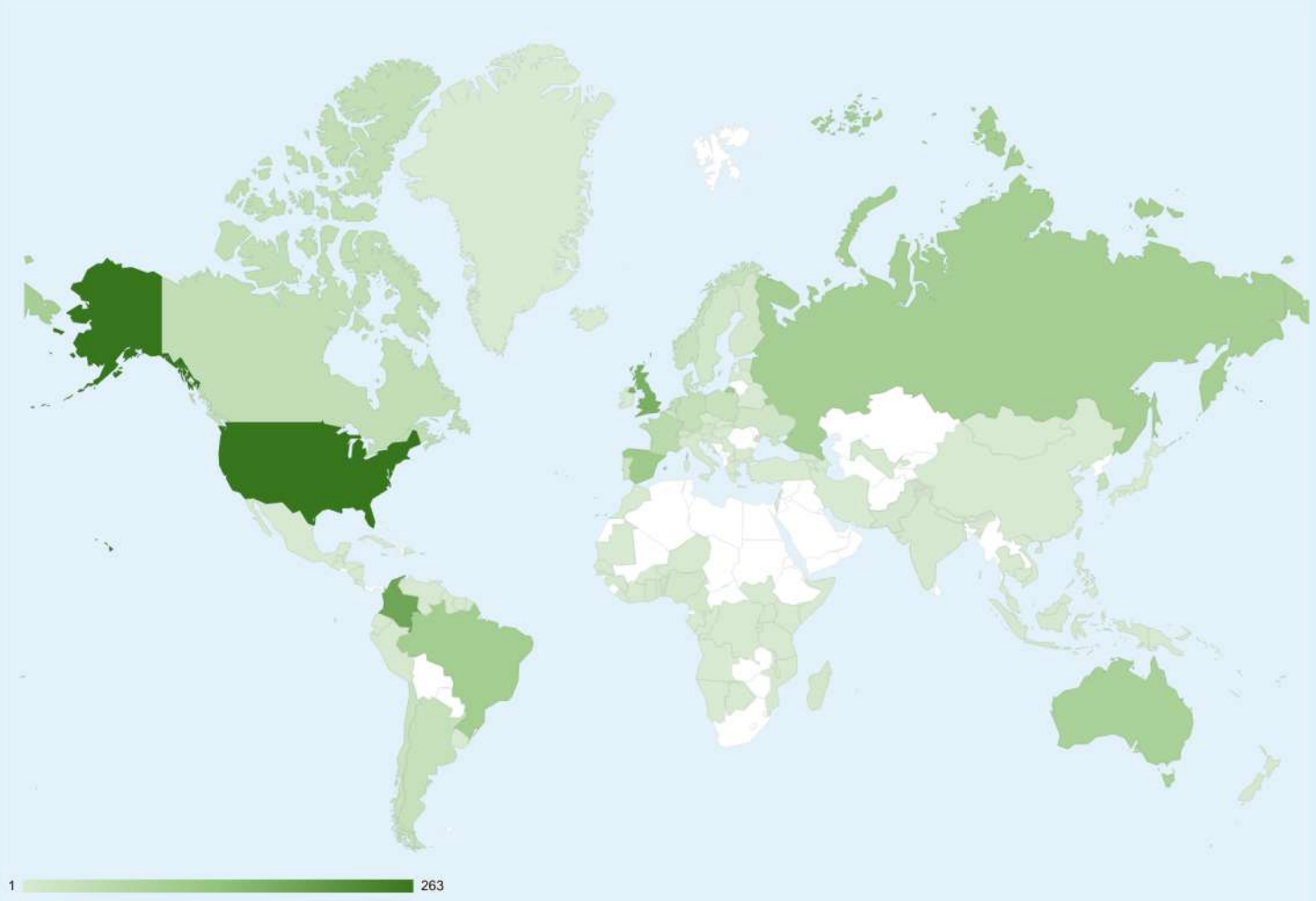
Open data

www.gbif.org

api.gbif.org



GBIF Network of Data publishing institutions 30 June 2022

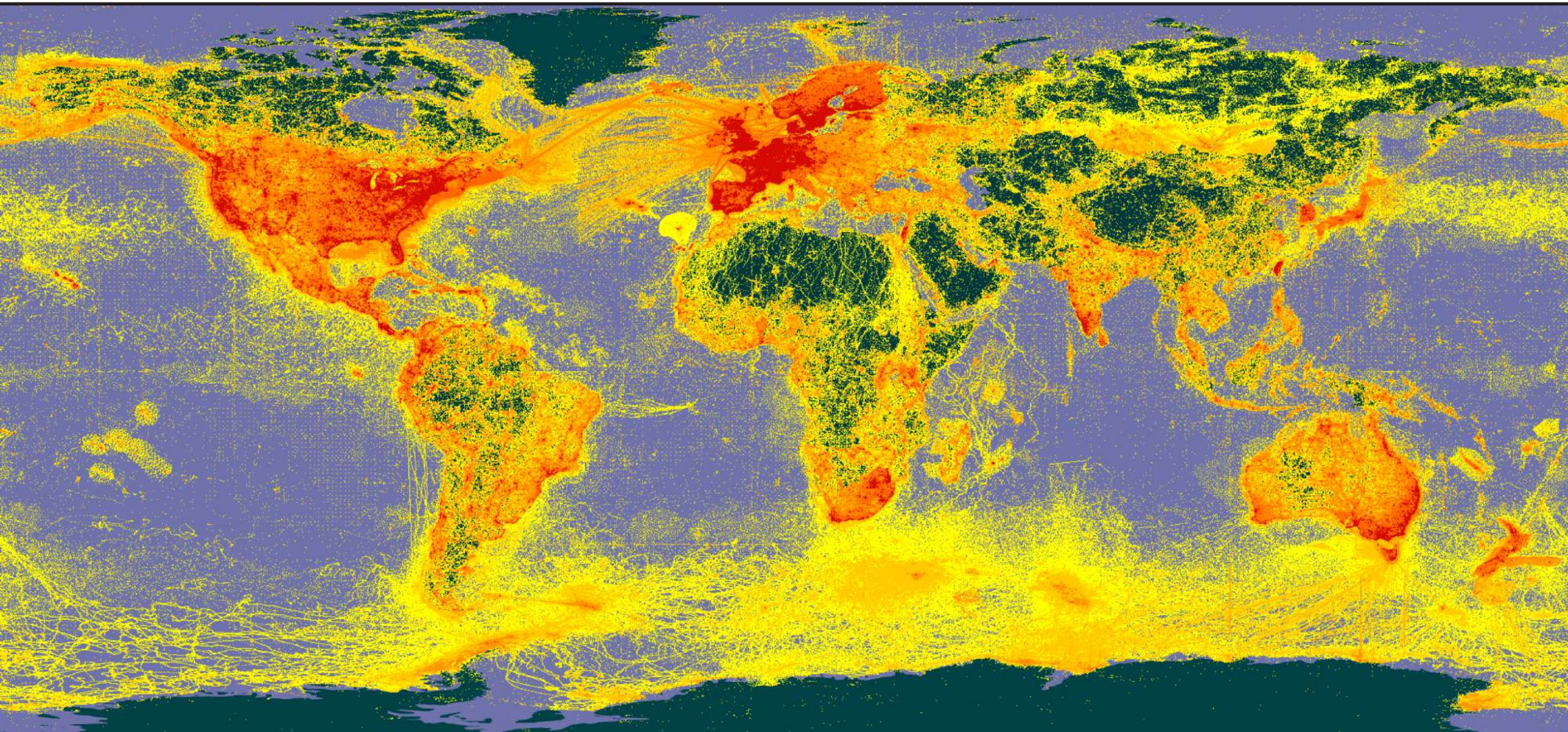


132 countries/territories with institutions sharing data through GBIF

Top 10 countries: number of data publishers		
1	United States	263
2	Colombia	183
3	United Kingdom	165
4	Spain	111
5	Brazil	100
6	Russian Federation	95
7	Australia	87
8	France	56
9	Canada	44
10	Netherlands	41

GBIF

Data From the GBIF Network 30 June 2022



Species occurrence records with MultiMedia evidence

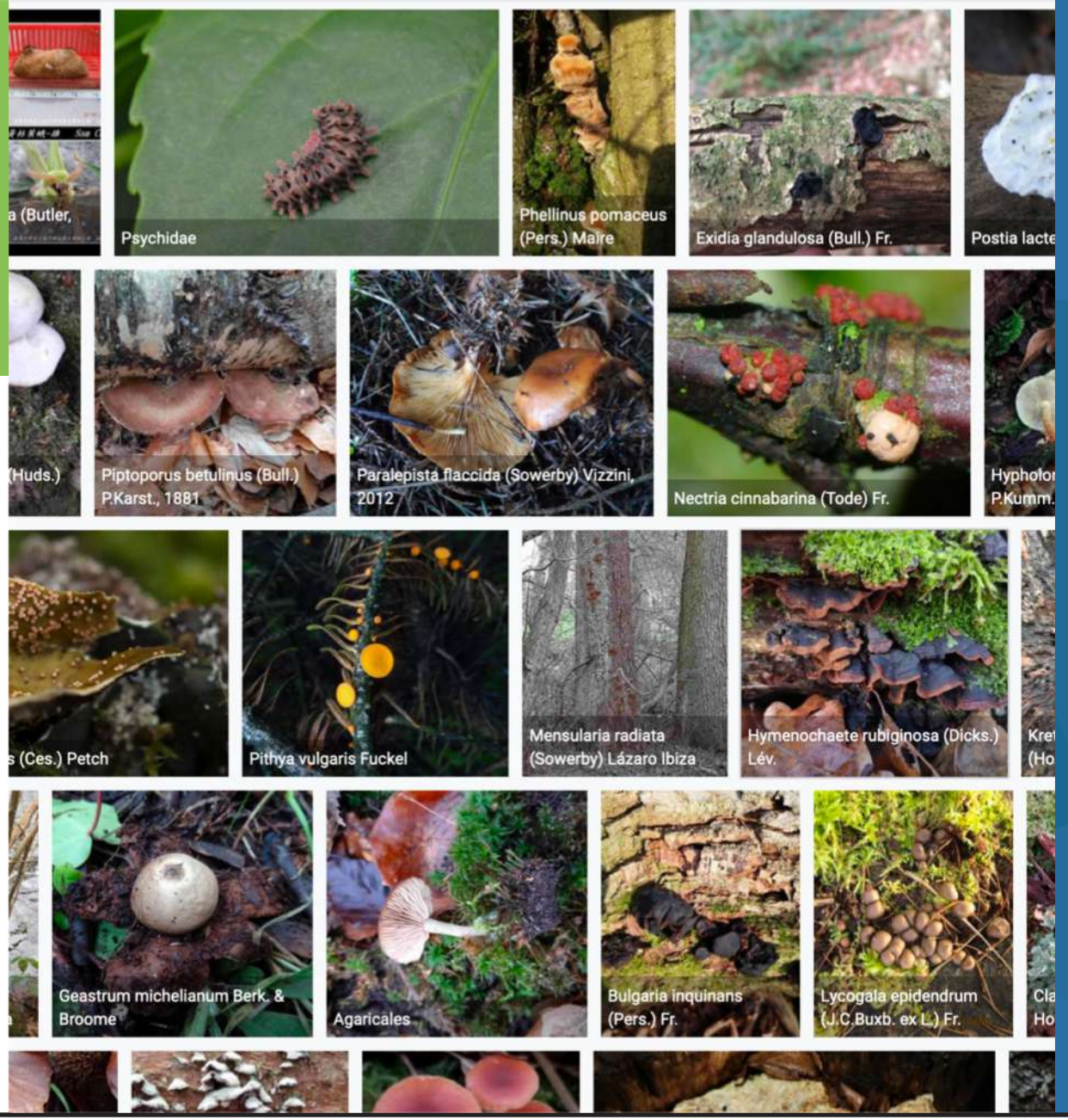
30 June 2022

107 million records with taxonomically identified images

- 61.4 million human observations
- 41.2 million specimens
- 2.3 million material samples
- 1.4 million fossil specimens

914,887 audio files

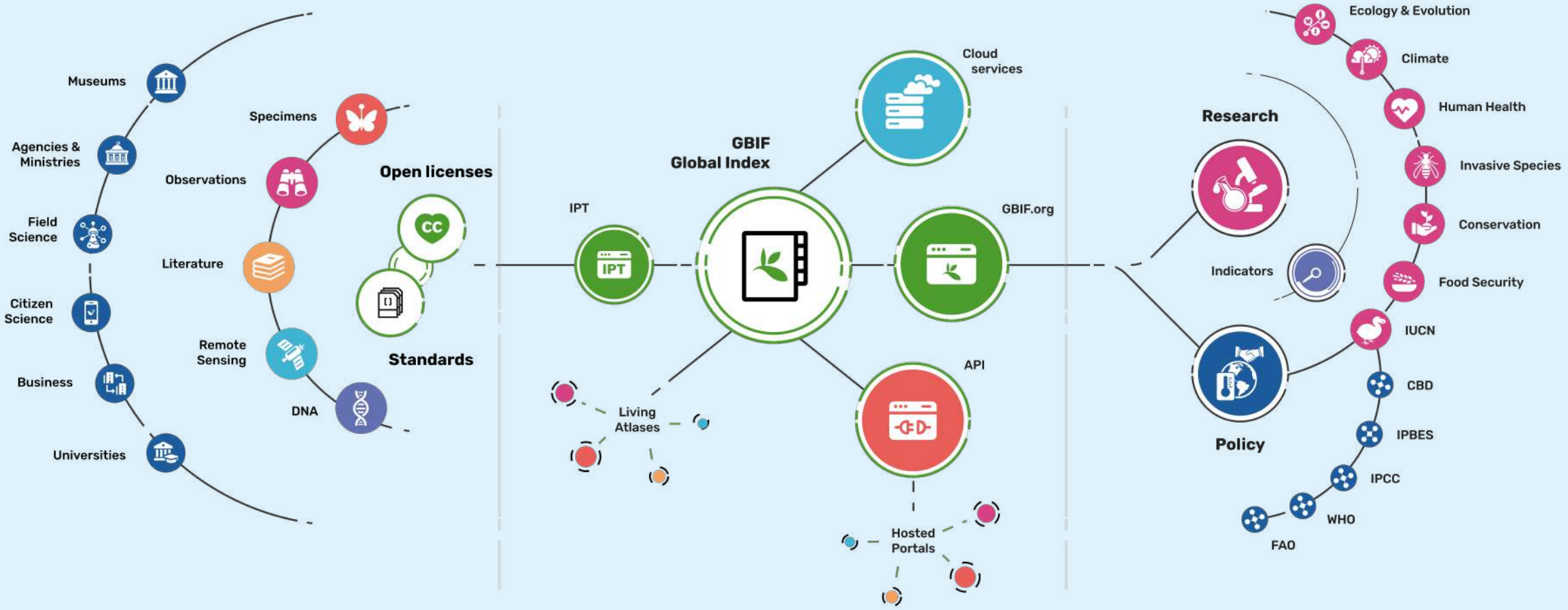
3,526 videos



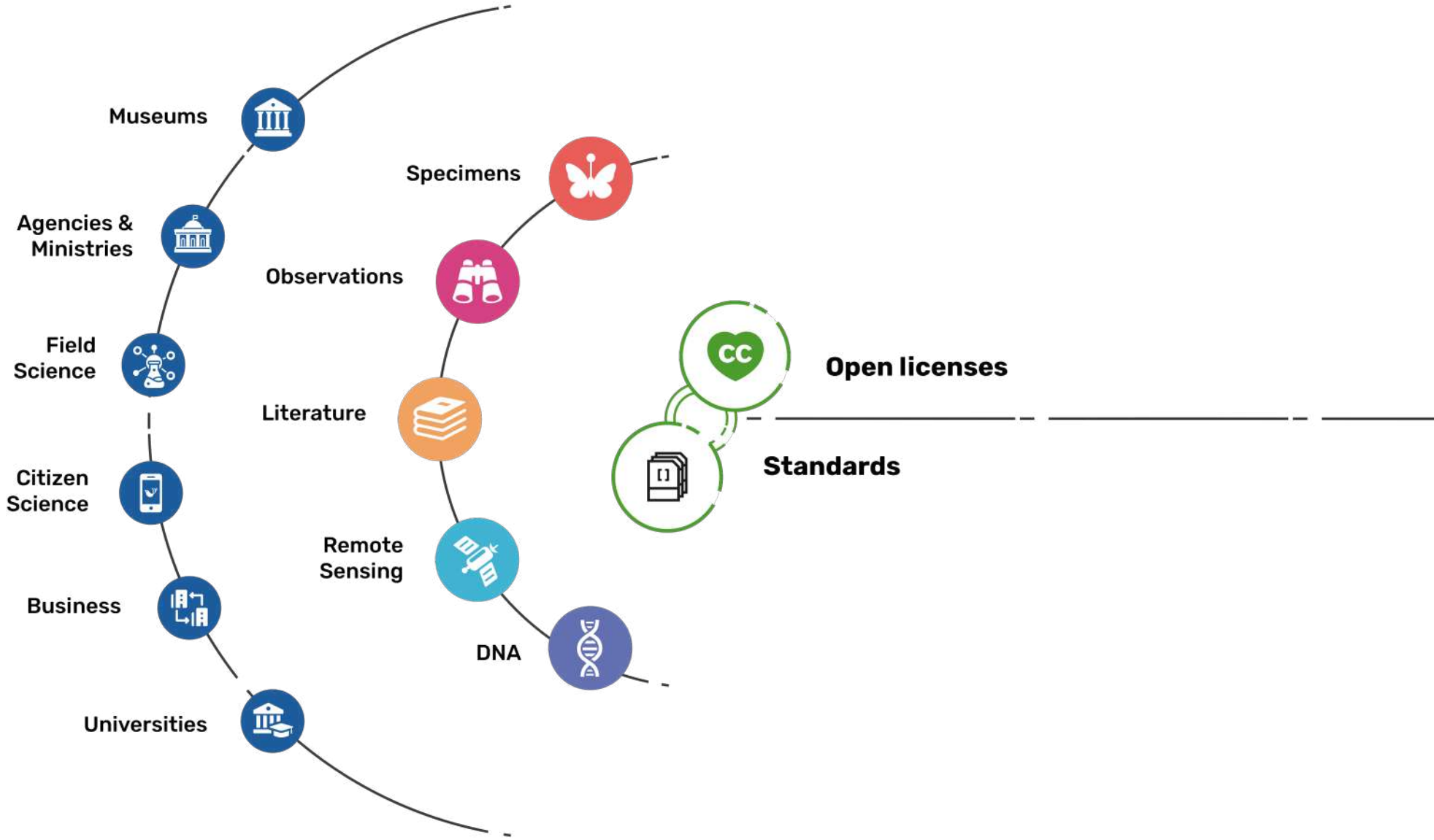
https://www.gbif.org/occurrence/gallery

GBIF

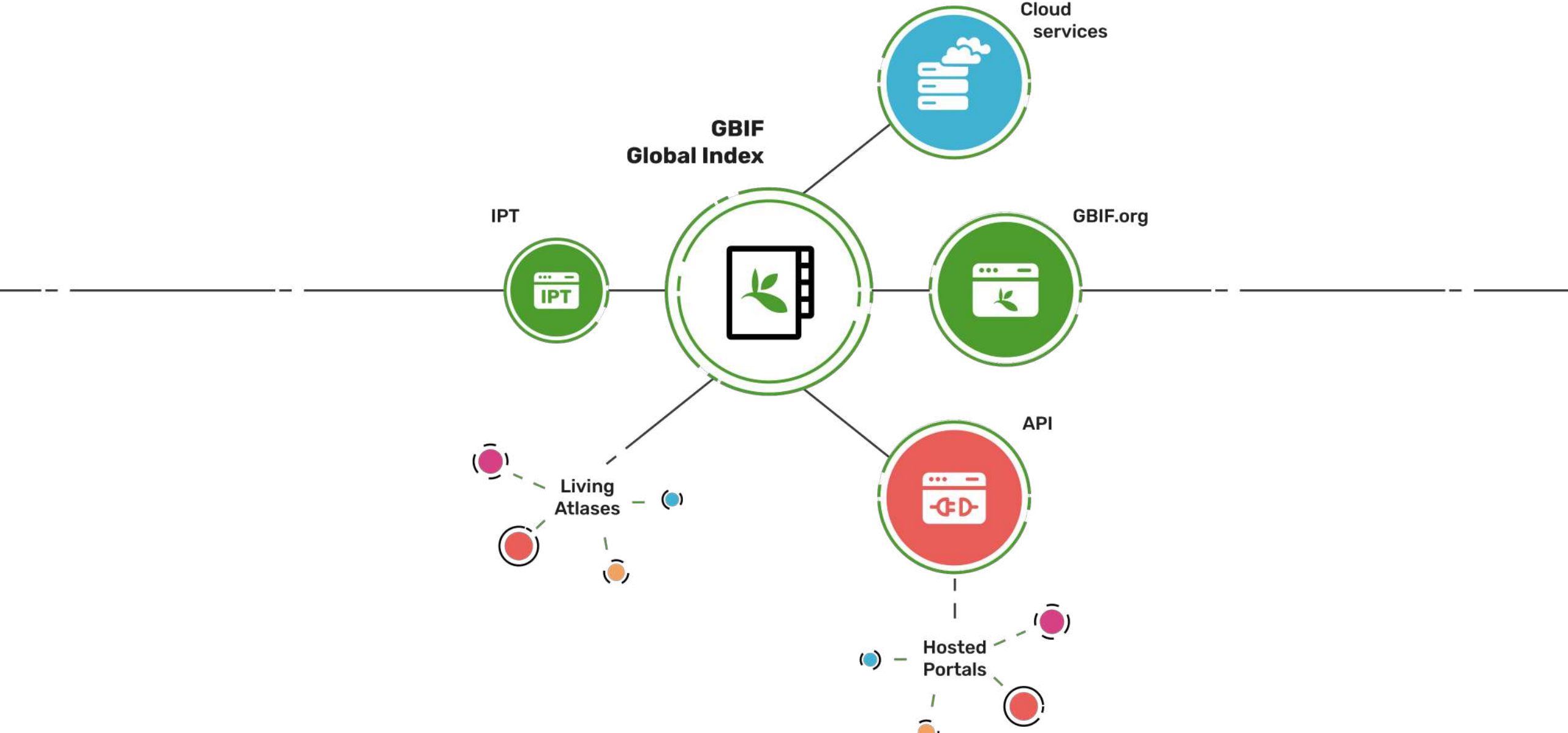
Providing biodiversity evidence for research and policy



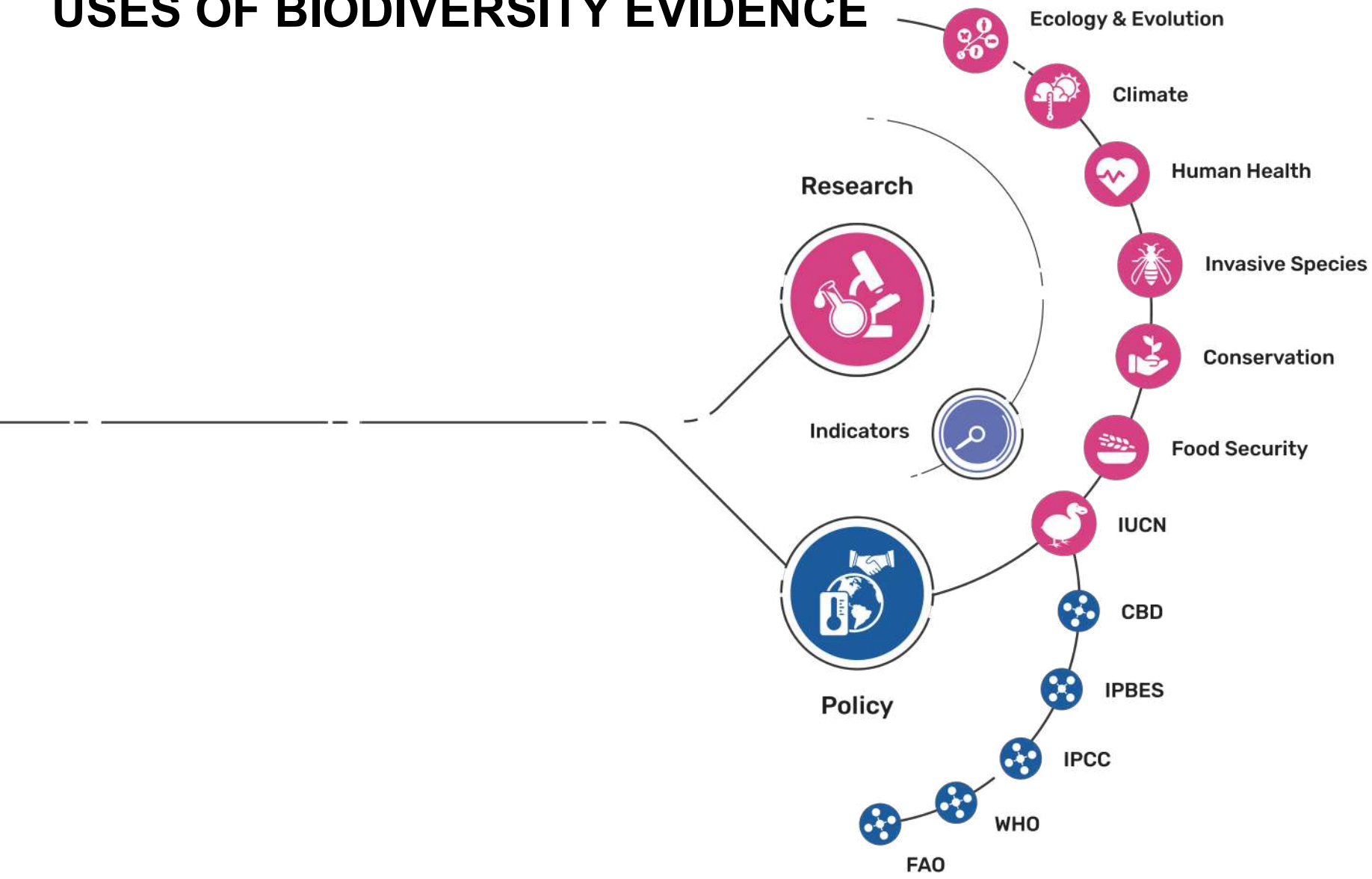
SOURCES OF BIODIVERSITY EVIDENCE



ACCESS TO BIODIVERSITY EVIDENCE

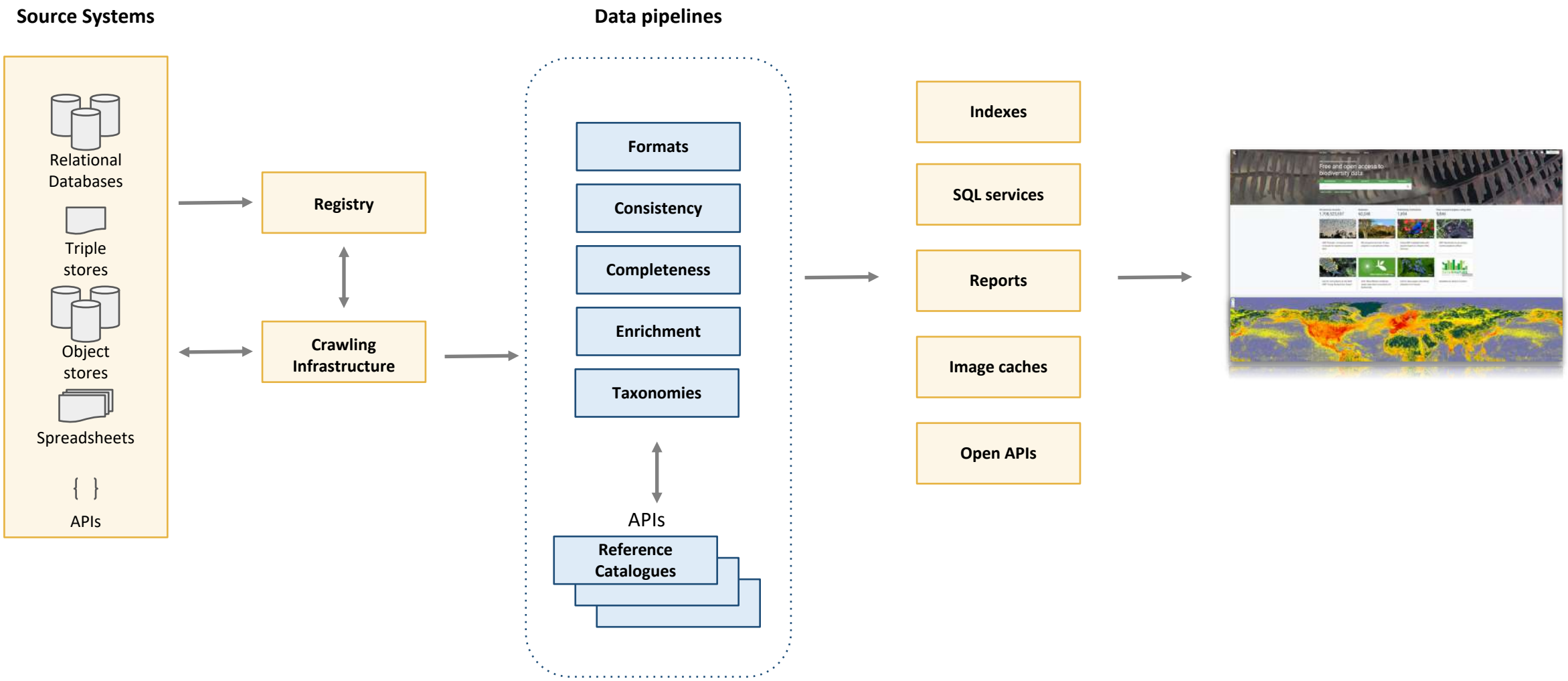


USES OF BIODIVERSITY EVIDENCE



https://www.gbif.org/science-review

Data warehouse



Built on open projects including Apache (Hadoop, Beam, Hive, HBase…) and Elasticsearch

Microsoft | Planetary Computer

Data Catalog Hub Applications Documentation Request access

Data Catalog

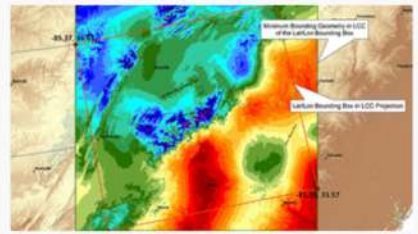
The Planetary Computer Data Catalog includes petabytes of environmental monitoring data, in consistent, analysis-ready formats. All of the datasets below can be accessed via Azure Blob Storage, and can be used by developers whether you're working within or outside of our Planetary Computer Hub.

Additional datasets

The following datasets are available on Azure, for use within or outside of the Planetary Computer Hub.

GBIF
Species occurrences shared through the Global Biodiversity Information Facility

Get GBIF occurrence data >



Daymet
Gridded temperature data across North America

Get Daymet temperature data >

Global Hydro Estimator
Global precipitation estimates

Get global precipitation data >

GOES-16
Weather imagery of the Americas

Get GOES-16 weather data >

Registry of Open Data on AWS

Global Biodiversity Information Facility (GBIF) Species Occurrences

biodiversity bioinformatics conservation earth observation life sciences

Description

The Global Biodiversity Information Facility (GBIF) is an international network and data infrastructure funded by the world's governments providing global data that document the occurrence of species. GBIF currently integrates datasets documenting over 1.6 billion species occurrences, growing daily. The GBIF occurrence dataset combines data from a wide array of sources including specimen-related data from natural history museums, observations from citizen science networks and environment recording schemes. While these data are constantly changing at GBIF.org, periodic snapshots are taken and made available on AWS.

Update Frequency

Snapshots of GBIF are taken on a monthly basis

License

This dataset is available under a CC-BY license and with the GBIF terms of use. Please refer to the GBIF citation guidelines when using this dataset.

Documentation

Documentation can be found here. You can learn more about GBIF here.

Managed By

The Global Biodiversity Information Facility (GBIF)

See all datasets managed by GBIF.

Contact

helpdesk@gbif.org

Resources on AWS

Description
GBIF species occurrence data in Parquet format (af-south-1 region).

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::gbif-open-data-af-south-1`

AWS Region
`af-south-1`

AWS CLI Access (No AWS account required)
`aws s3 ls s3://gbif-open-data-af-south-1/ --region af-south-1 --no-sign-request`

Explore
Browse bucket

Description
GBIF species occurrence data in Parquet format (ap-southeast-2 region).

Hosted portals building on subsets of data shared through GBIF

GBIF.us

Biodiversidad.co

SANBI-GBIF

Natural History Museum Rotterdam

Pacific Biodiversity Information Facility

Legume Data Portal

Virtual Herbarium Germany

Living Norway Ecological Data Network

https://www.gbif.org/hosted-portals

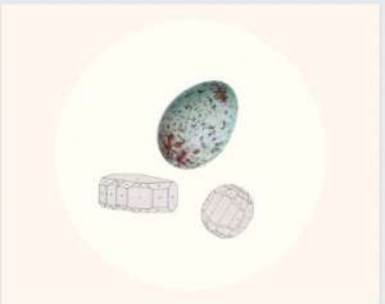
GBIF

Diversifying the GBIF Data Model

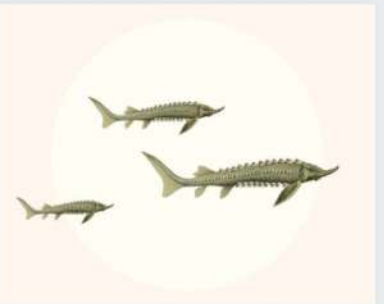
Explore the current ideas for a unified common model capable of supporting expanded data-publishing capabilities and potential directions for evolving the Darwin Core standard

▶ Community webinar recording

12 April 2022



▶ Intro presentation: Diversifying the GBIF data model



Slides: Diversifying the data model

Case studies

Introduction

eDNA barcoding

Camera trap

Tissue samples

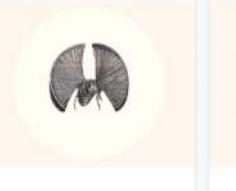
Automatic moth trap

Global malaise programme

iNaturalist

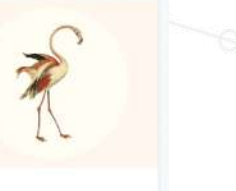
Specimens with media

Environmental and community measurements



Taxonomic treatments in literature

Malaise trapping for reference barcode collection



Taxonomic checklist

Ecological survey data exchange specification

Biotic interactions

Glossary

https://www.gbif.org/composition/HjlTr705BctcnaZkcjRJq/data-model

GBIF

THANK YOU

thirsch@gbif.org

Q&A

biodiversa+
European Biodiversity Partnership

Presentation of Meta(data) structure and workflows in LifeWatch ERIC

By Lucia Vaira, LifeWatch ERIC

Meta(data) structure and workflows in

Biodiversa+ WS biodiversity monitoring databases

Lucia Vaira | LifeWatch ERIC Service Centre, Lecce – Italy | lucia.vaira@lifewatch.eu

Outline

- LifeWatch ERIC
  - Organisation and governance
- The (meta)data interoperability need
  - The FAIR principles
  - Alignment with the EOSC Interoperability Framework
- Meta(data) structure and workflows in LifeWatch ERIC
  - The LifeWatch ERIC Metadata Catalogue
  - The LifeWatch ERIC EcoPortal
  - The interoperability in LifeWatch ERIC

LifeWatch ERIC

LifeWatch ERIC is a European Infrastructure Consortium that offers e-Science research facilities to scientists investigating Biodiversity organization and Ecosystem functions and services.

LifeWatch ERIC's mission: to be the leading worldwide provider of content and services for the biodiversity research community, creating new opportunities for large-scale scientific development, and enabling accelerated data capture and modelling thanks to the use of innovative technologies.

LifeWatch ERIC is a distributed research e-infrastructure consortium consisting of 8 EU Member States.

Its structure mirrors its nature, with central components (Common Facilities) located in 3 Member States (Spain, Italy and the Netherlands), and National Nodes in all 8 countries.

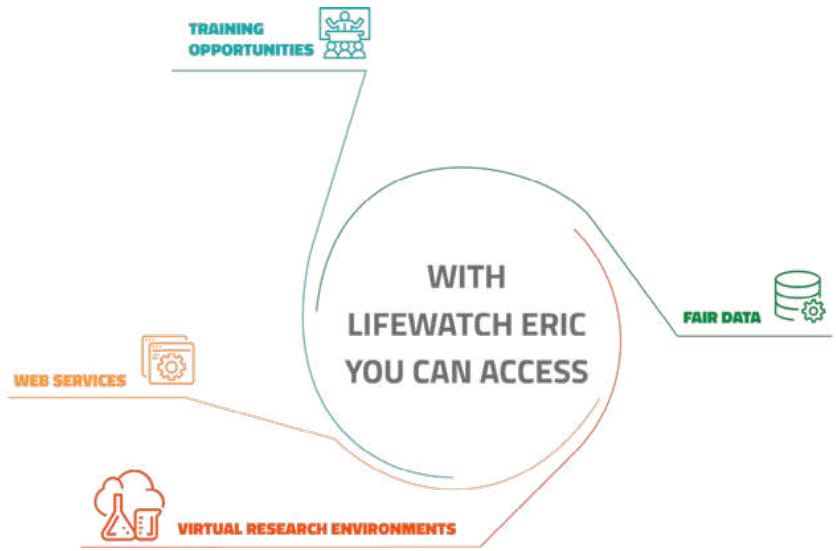
LifeWatch ERIC's current members are: Belgium, Bulgaria, Greece, Italy, the Netherlands, Portugal, Slovenia and Spain.

Slovakia participates as an Observer.

LifeWatch ERIC

With LifeWatch ERIC users can:

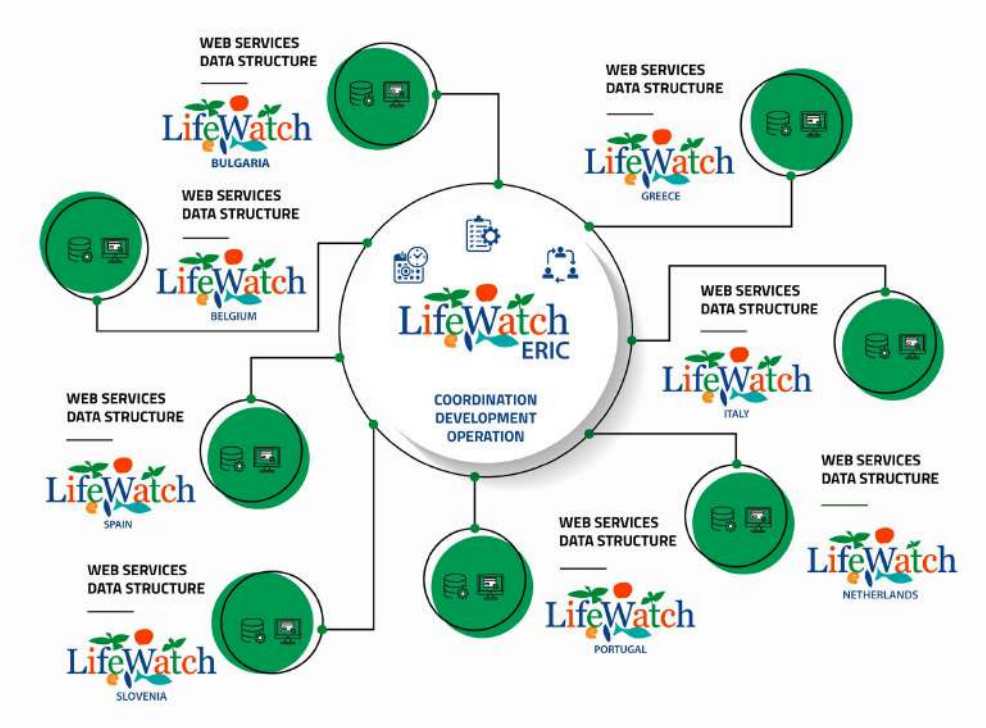
- find, access, work with and reuse FAIR data collected by science at a global level, of different typologies and scales. Advanced search functions, thesauri and ontologies are available on our catalogues to further combine data and generate new services;
- process and analyse data in our Virtual Research Environments (VREs);
- get trained in the use of our services, as well as on key scientific issues through Master's and PhD courses, summer schools, webinar programmes and educational initiatives;
- receive support through our Helpdesk and its experts specialised in different domains.



LifeWatch ERIC

Over the last 5 years the Common Facilities and Distributed Centres have gone from being isolated elements to fully integrated into LifeWatch ERIC.

Rules and policies are continuously being developed and updated to support this integration process.



The (meta)data interoperability need

Considering the FAIR principles, interoperability is considered in relation to the fact that "research data usually need to be integrated with other data; in addition, the data need to interoperate with applications or workflows for analysis, storage, and processing".

The following principles are proposed:

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

LifeWatch ERIC: achieving interoperability is essential to allow the federation of **all the LifeWatch ERIC assets** → provide added value for the final users.

Alignment with the EOSC Interoperability Framework

EOSC Interoperability Framework*: 4 main points of view and a set of recommendations has been proposed:

- **Technical**: the "ability of different information technology systems and software applications to communicate and exchange data", *EOSC IF*.
- **Semantic**: the "ability of computer systems to transmit data with unambiguous, shared meaning", *FAIRsFAIR deliverable D2.1*
- **Organisational**: the "way in which organisations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals", *EIF*
- **Legal**: "it deals, in particular, how data should be re-used", *EOSC IF*.

** European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., et al., EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, https://data.europa.eu/doi/10.2777/620649*

Alignment with the EOSC Interoperability Framework

Technical interoperability

1. The definition of a common security and privacy framework (including Authorisation and Authentication Infrastructure) is needed to ensure secure and trustworthy data exchange.

2. Easy-to-understand Service-Level Agreements for all resource providers.

3. Easy access to data sources available in several formats to facilitate overcoming their heterogeneity and allow integrating data across communities.

4. Coarse-grained and fine-grained dataset (and other research objects) search tools need to be made available.

5. Multiple service providers for different types of PIDs exist. As a result, sometimes the identifiers are not resolvable. A clear PID policy needs to be defined.

Alignment with the EOSC Interoperability Framework

Semantic interoperability

1. Clear and precise, publicly-available definitions for all concepts, metadata and data schemas should be available in the communities.

2. Semantic artefacts should be available preferably with open licenses and must have sufficient associated documentation.

3. There is a lack of common reference repositories of semantic artefacts.

4. A minimum metadata model should be adopted to ease discovery over existing federated research data and metadata.

5. There should be extensibility options to allow for disciplinary metadata allowing users to add annotations.

6. There should be clear protocols and building blocks for the federation/harvesting of semantic artefacts catalogues.

Alignment with the EOSC Interoperability Framework

Organisational interoperability

1. A clearly-defined governance structure that includes the governance framework that will deal with interoperability across organisations and disciplines.

2. A clear description of the “terms and conditions” and “acceptable use policies” that will rule the provisioned services.

3. The rules of participation recommendations should be completed with aspects related to interoperability.

4. Usage recommendations of standardised data formats and/or vocabularies, with their corresponding metadata.

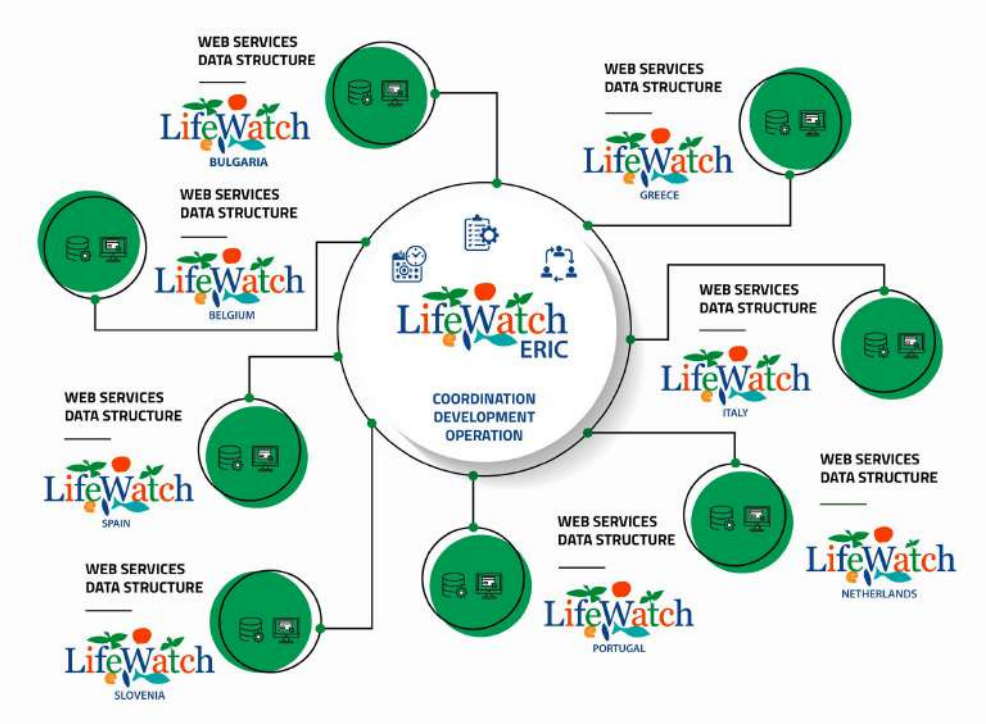
5. A clear management of permanent organisation names and functions needs to be provided (sustainability).

Alignment with the EOSC Interoperability Framework

Legal interoperability

1. Information on who holds the rights to the data, including any embedded data.
2. Metadata to be available without restrictions on (re-)use to facilitate the FAIRness of the data it describes.
3. Licensing requirements should be both human and machine readable and allow data providers and users to understand license compatibility.
4. Clear status of historic copyrightable datasets and metadata with no license or unclear licenses arrangements (orphan data).
5. Need to track the license evolution over time for datasets.
6. Need of harmonised policy and guidance to deal with patents or trade secrets disclosure.
7. Need to ensure adequate protection of personal data and general compliance with the GDPR and EU Member States' domestic law and guidelines.
8. Harmonised terms of use across repositories.

Meta(data) structure and workflows in LifeWatch ERIC



LifeWatch ERIC: central node for the coordination, development and operation.

Main goal: to increase collaboration within and among national nodes but also among external organisations to:

- enhance findability;
- reduce duplication;
- improve information consistency and quality.

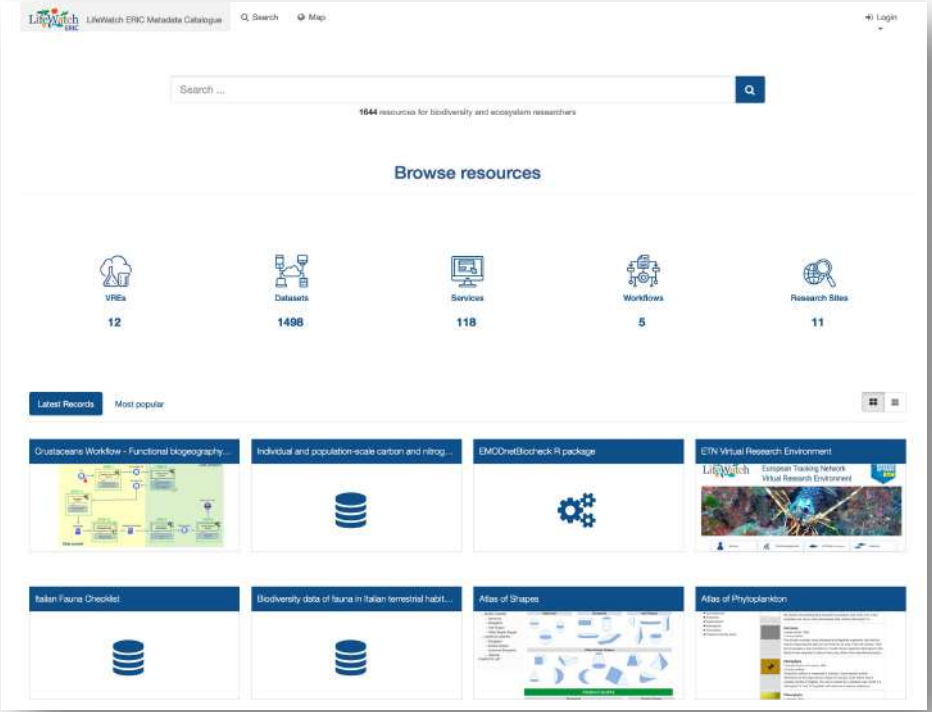
LifeWatch ERIC Metadata Catalogue

A standard-based information management system based on GeoNetwork 3.10.

It enables access to several resources from a variety of providers through descriptive metadata, enhancing and promoting the information exchange and sharing among organisations and research infrastructures.

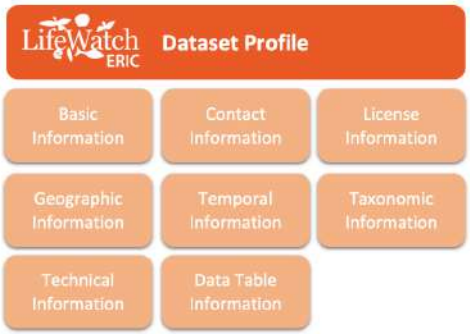
It allows to manage descriptive metadata related to **datasets** (EML 2.2.0 standard), **research sites**, **services**, **Virtual Research Environments**, and **workflows** (ISO 19139 standard). Metadata attributes can be optional/mandatory and can require single/multiple values.

It allows (upon validation and verification) the creation of Digital Object Identifiers (DOIs) for resources that do not have it, by exploiting the GeoNetwork – DataCite connection.



https://metadatacatalogue.lifewatch.eu

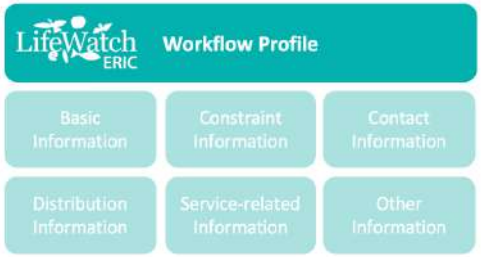
LifeWatch ERIC Metadata Catalogue



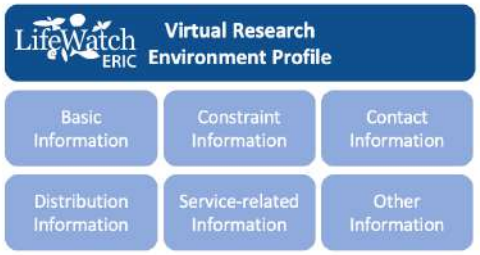
EML2.2.0 standard - **77** metadata fields

ISO 19139 standard - **34** metadata fields

ISO 19139 standard - **16** metadata fields



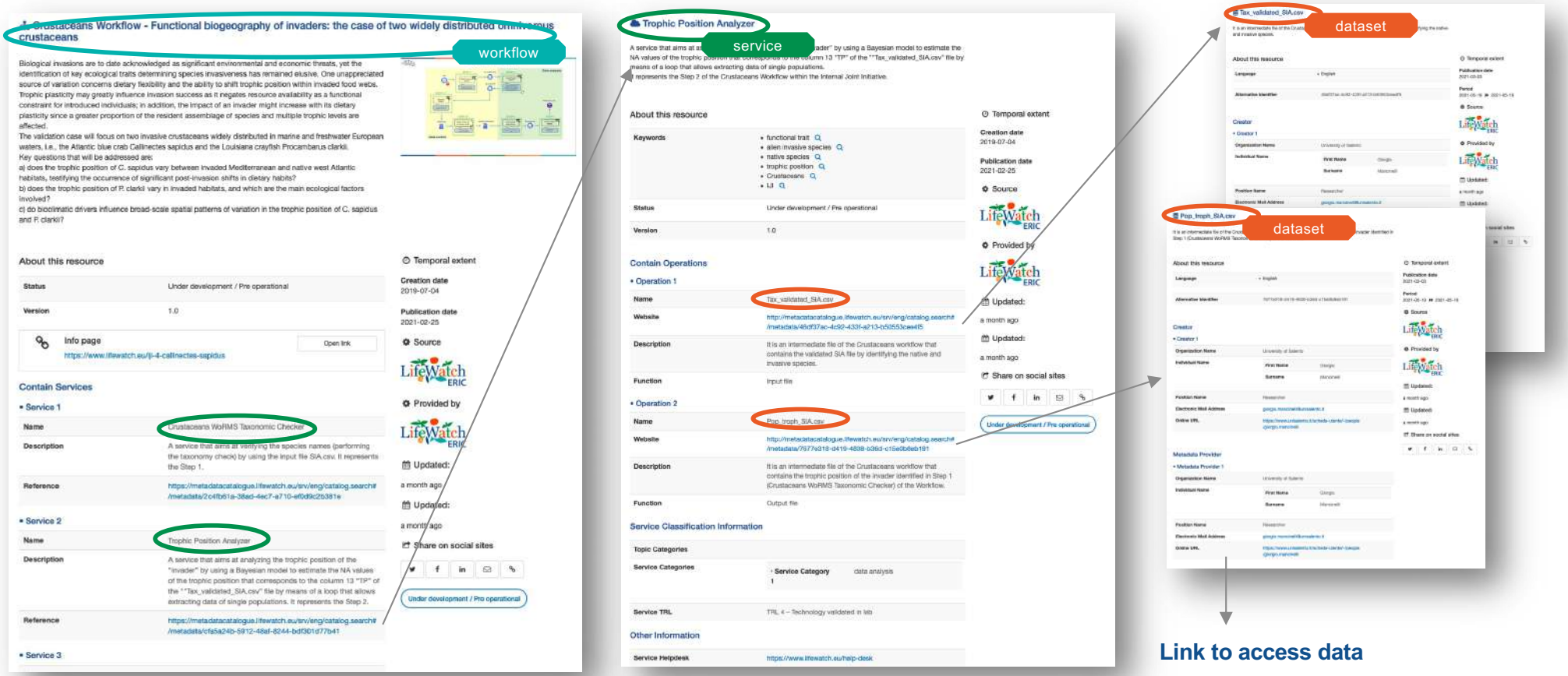
ISO 19139 standard - **24** metadata fields



ISO 19139 standard - **24** metadata fields

LifeWatch ERIC Metadata Catalogue

LifeWatch ERIC Metadata Catalogue



Link to access data

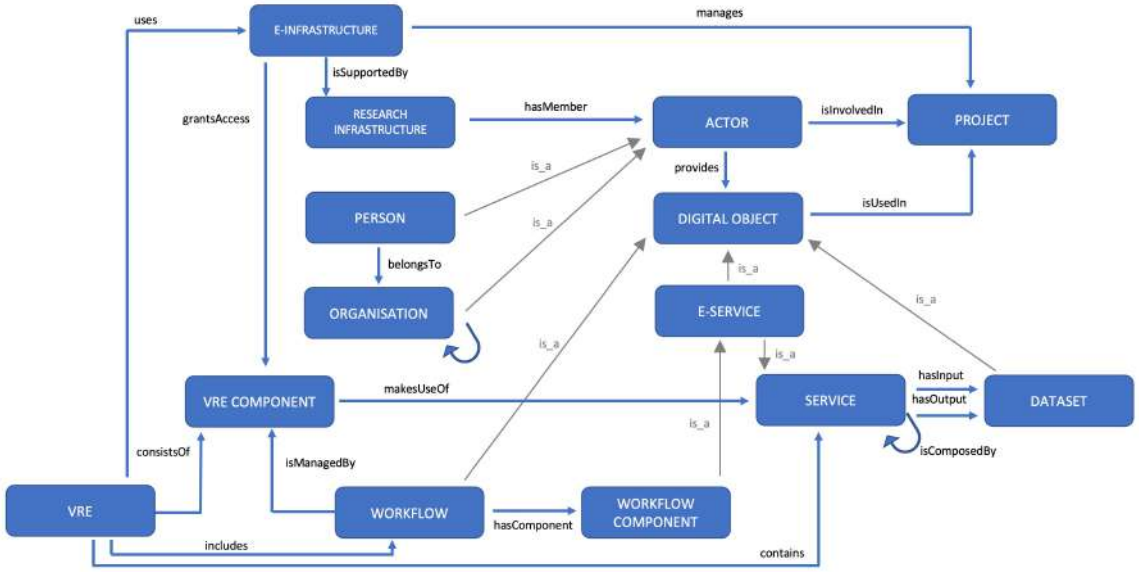
The LifeWatch ERIC Upper Ontology

The usage of the same terminology concepts is highly relevant for the interoperability in the LifeWatch ERIC context.

The resources hosted in the Metadata Catalogue exploit the work already done within the Working Group D of the LifeWatch ERIC Internal Joint Initiative where **LUPO, the LifeWatch ERIC UPper Ontology*** has been defined so that the key concepts (dataset, service, Virtual Research Environment, workflow, etc.) are well defined and agreed within the scientific community.

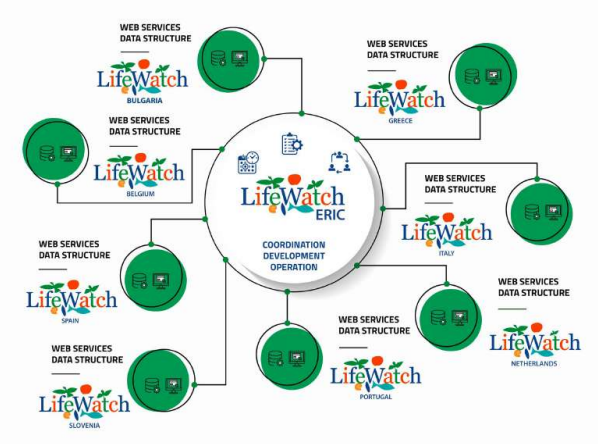
An upper ontology (top-level ontology, upper model, or foundation ontology) is an ontology which consists of very general terms (such as "object", "property", "relation") that are common across all the Research Infrastructure domains.

Main goal: to support broad semantic interoperability among a large number of domain-specific ontologies by providing a common starting point for the formulation of definitions.



* *LUPO Ontology: http://ecoportal.lifewatch.eu/ontologies/LUPO*

LifeWatch ERIC Metadata Catalogue
Harvesting workflows



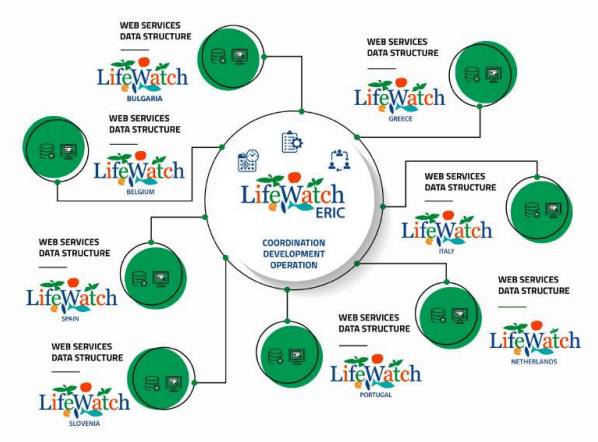
Harvesting procedures to import/export datasets via UI or API → humans + machine2machine

OAI/PMH*: our most used protocol due to the possibility to specify the exact endpoint, the prefixes, etc.

Used to harvest datasets metadata records from 4 LifeWatch ERIC national nodes

OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting

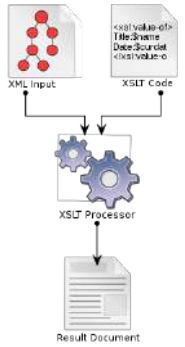
LifeWatch ERIC Metadata Catalogue
Mapping workflows



Mapping procedures to convert formats, map into LifeWatch ERIC metadata profile, align standards to the same version, etc.

XSLT*: our most used method due to the possibility to specify the desired output according to the specific input.

Used to map harvested datasets metadata records from 4 LifeWatch ERIC national nodes



XSLT: eXtensible Stylesheet Language Transformations

LifeWatch ERIC Metadata Catalogue
Next steps

New version of the Catalogue (by the end of 2023)

- Semantic search
- Semantic annotation
- Map features enhancement
- Metrics and statistics
- Versioning mechanism
- Editing tool improvement
- Validation mechanism improvement
- Editorial workflow improvement
- ...

LifeWatch ERIC EcoPortal

The use of semantic resources (vocabularies & ontologies), giving well defined meaning and understanding, ensure the exchange of information among machines and people.

Controlled vocabularies (CVs) play an important role in metadata standards → define the meaning of metadata elements and the allowed values.

Help to find relevant data, or provide information on how to interpret data and reuse it:

- the use of CVs helps to improve the data interoperability by facilitating the data interpretation and harmonization;
- the use of CVs improve the discovery, linking, understanding and reuse of research data within research communities;
- CVs can be used for the (meta)data annotation to comply with the FAIR principles.

Main issues:

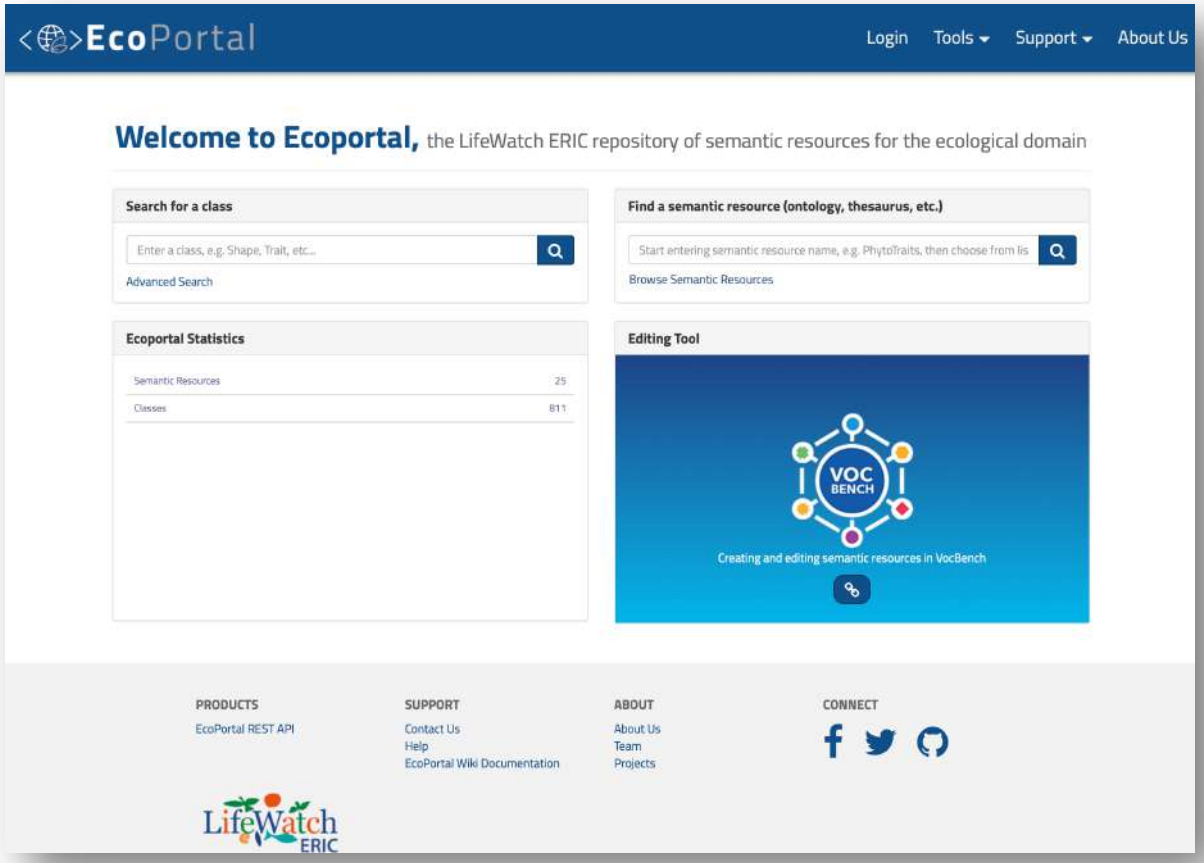
- Lack of semantic artefacts across communities (or with no open licenses and adequate documentation).
- Lack of common reference repositories of semantic artefacts like ontology catalogues .

LifeWatch ERIC EcoPortal

The LifeWatch ERIC repository of semantic resources (ontologies, vocabularies, thesauri, etc.) for the ecological domain.

A web-based application for supporting the community in the:

- identification and selection of semantic resources;
- creation, management, mapping and alignment of semantic resources in the ecological domain.



http://ecoportal.lifewatch.eu

The interoperability in LifeWatch ERIC status & challenges

Almost aligned with the EOSC Interoperability Framework.

Big challenge: overcome data heterogeneity and allow integrating data across communities.

- Research data available in multiple general-purpose formats (CSV, database dumps, JSON, XML, etc.) or community-based models (Darwin Core, NetCDF, etc.) → hard to align when reusing datasets across communities.
- Semantic interoperability issues also appear (lack of agreement in attributes or column headers, absence of headers or adequate documentation, etc.).
- There should be extensibility options to allow for disciplinary metadata that is typical for some research communities, allowing users/researchers to add annotations according to the established practices of their communities → LifeWatch ERIC Semantic Platform – *work in progress*
- But the possibility to integrate data across communities depends not only on the (meta)data structure and/or format but also on other aspects like the heterogeneity in the adopted tools and methods for collecting and processing data.

Thank you

Lucia Vaira | LifeWatch ERIC Service Centre, Lecce – Italy | lucia.vaira@lifewatch.eu

Q&A

biodiversa+
European Biodiversity Partnership

Presentation of data architecture and data workflows in BIOSCANEurope

By Rutger Vos, iBOL/BIOSCAN

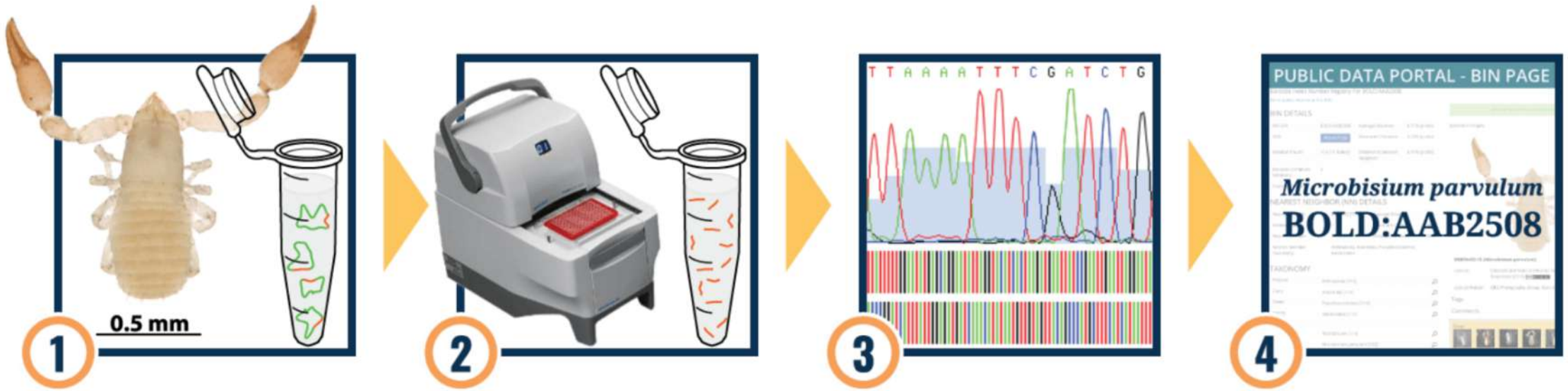
DNA (meta)barcoding for biodiversity monitoring

BIOSCAN/iBOL Infrastructural developments

Sujeevan Ratnasingham & Rutger Vos

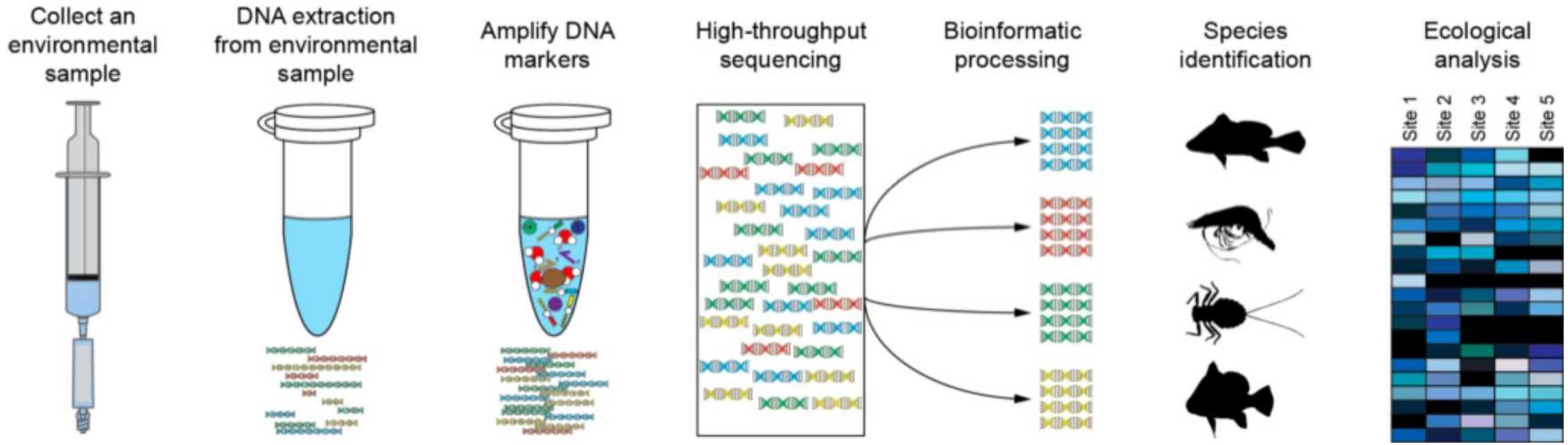
international
BARCODE
OF LIFE

DNA barcoding



- DNA extraction, amplification, sequencing and processing of (a) specific marker gene(s) from specimens from known species and those from novel species
- Broad data model including specimen records, images, sequences, and provenance
- Results in annotated reference data in which taxonomically-broad patterns of sequence diversity and specific identifications can be investigated
- In existence since Sanger sequencing, scaled up using high-throughput sequencing methods

Metabarcoding



- DNA extraction, amplification and high-throughput sequencing of mixtures, e.g. soil, water, air, but also gut contents, commercial products, etc.
- Requires annotated barcode reference databases to perform species identification.
- Results in lists of taxonomically identified entities (OTUs, ASVs), which can then be analyzed for alpha/beta/gamma indices of biodiversity.

Trends

Producers

- High-throughput sequencing (HTS) allows for broader taxonomic coverage at low cost compared to Sanger
- Using HTS for barcoding results in the production of multiple barcodes per specimen
- Most barcodes are produced through institutional work processes rather than *ad hoc*, small-scale studies by individual researchers.

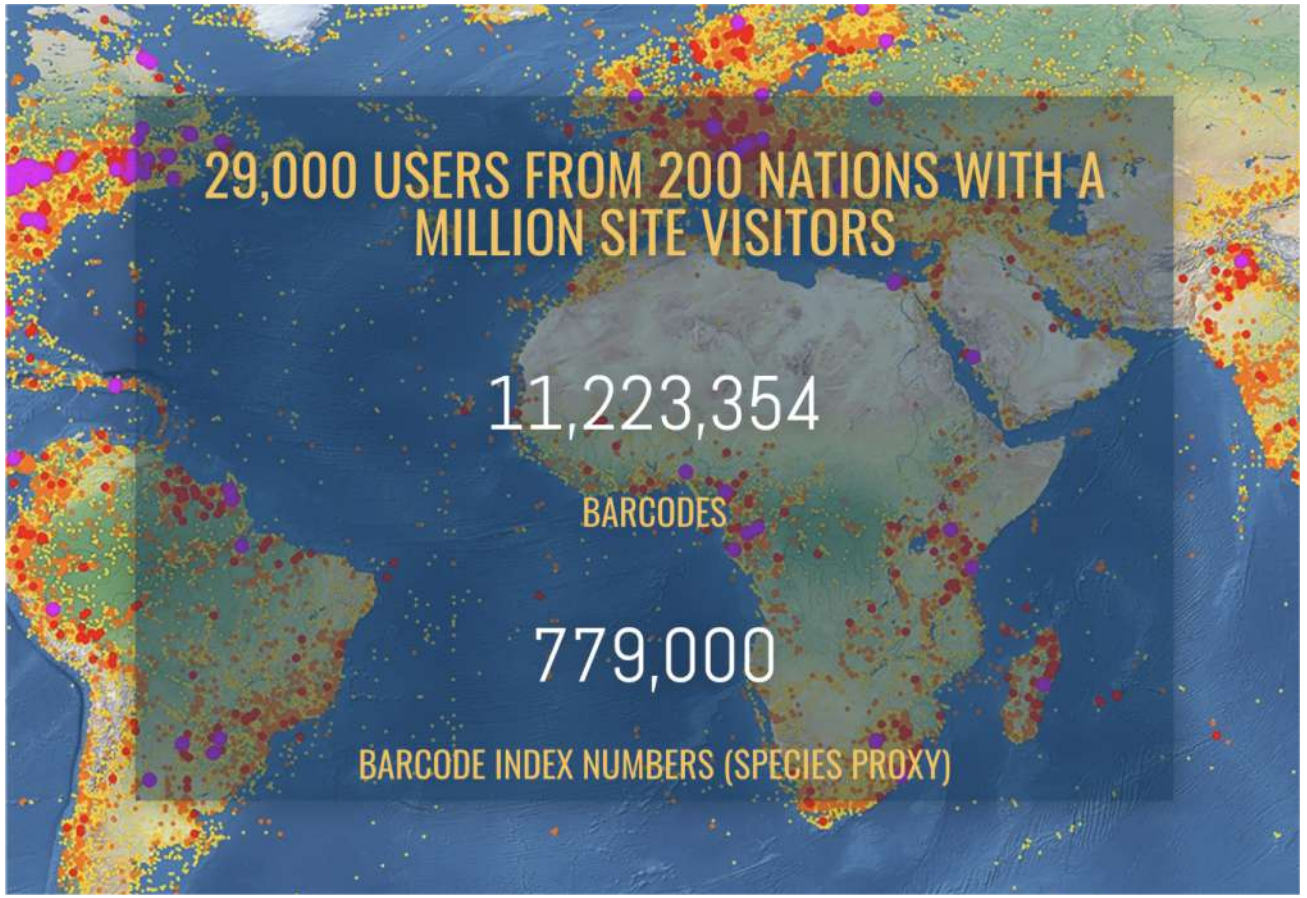
Consumers

- Large-scale, integrative projects (e.g. UNITE, ARISE) seek to harvest well-annotated sequences at scale
- Law enforcement
- Rising demand from corporations
- FAIR data considerations gain importance

iBOL

- Research alliance of nations
- Established in 2008
- Builds:
  - barcode reference libraries
  - data and process standards
  - steering and capacity building
  - international collaboration
  - **informatics platforms**

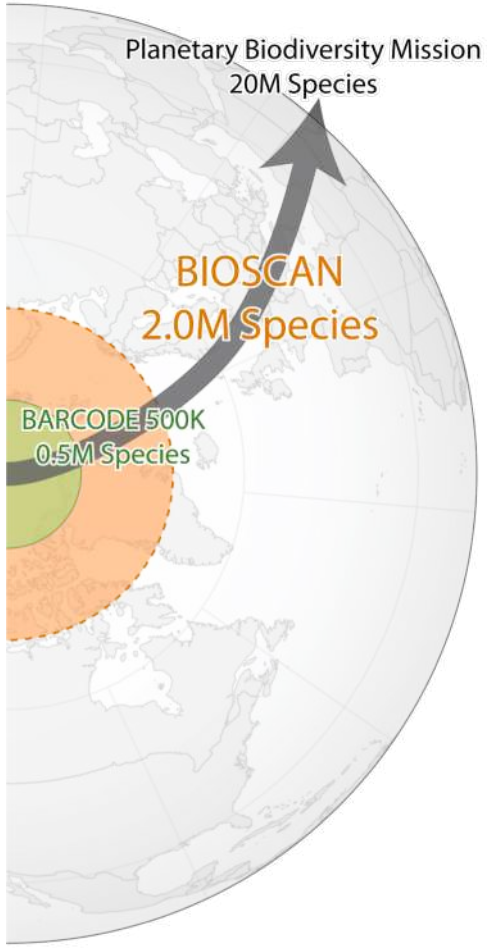
international BARCODE OF LIFE



BioScan

- Launched in 2019
- Focus on building baselines for biomonitoring
- Establishes the European node of iBOL
- Complementary with genome skimming (Biodiversity Genomics Europe)

BIOSCAN EUROPE



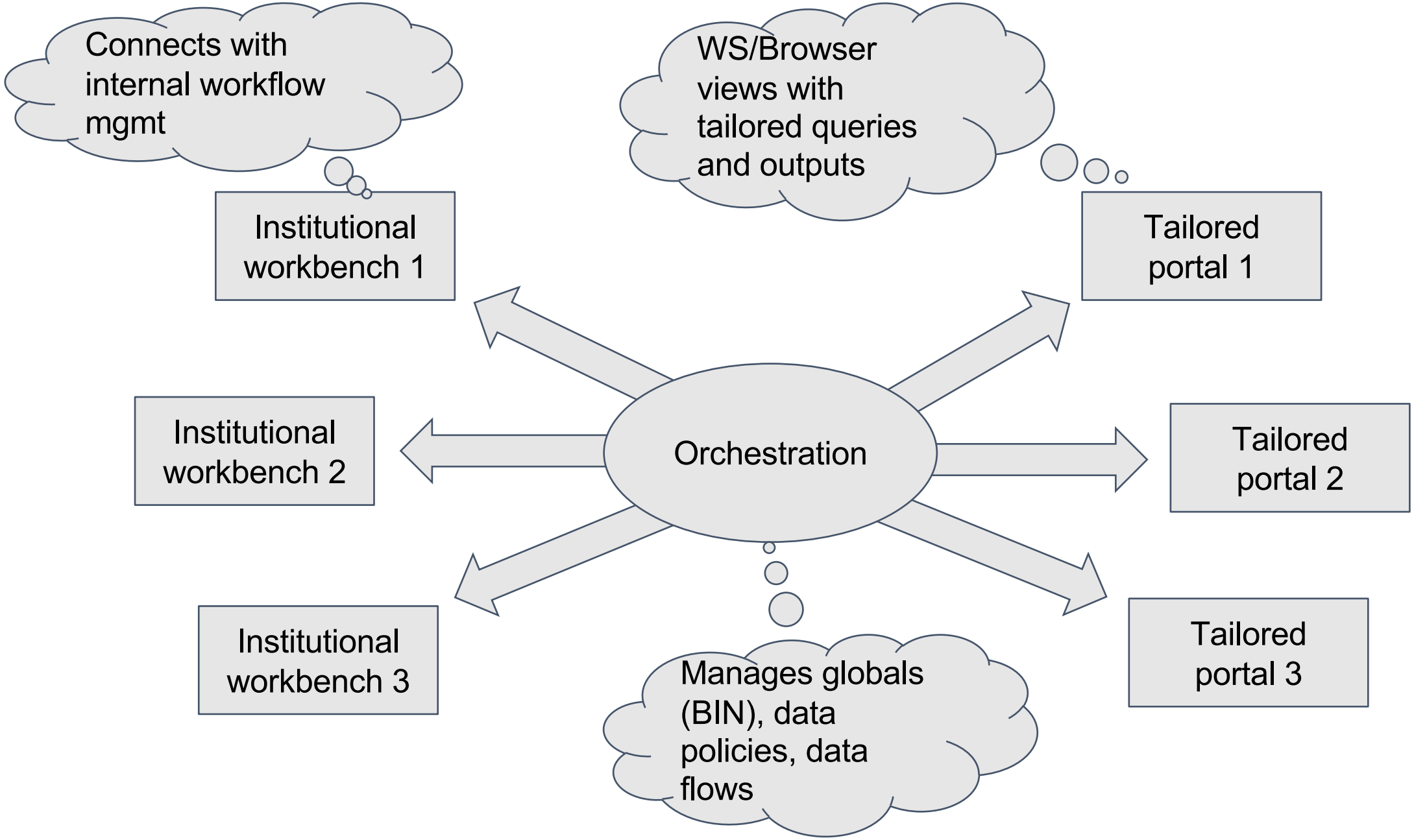
Towards a decentralized infrastructure for DNA barcodes

Back-end/producer considerations

- Higher data volumes make manual submissions to repositories cumbersome
- More of the preprocessing workflow is automated, e.g. in LIMS and workflow mgmt tools
- Need for tailored, automated submission infrastructure

Front-end/consumer considerations

- Larger-scale initiatives in data reuse, integration and curation
- Data collection moving from browser-based/manual to APIs
- Need for curated collections and data packages
- Need for tailored queries, with provenance and versioning


Connects with internal workflow mgmt
WS/Browser views with tailored queries and outputs
Institutional workbench 1
Tailored portal 1
Institutional workbench 2
Orchestration
Tailored portal 2
Institutional workbench 3
Tailored portal 3
Manages globals (BIN), data policies, data flows

Implementation

Planning

Q4-2022..Q3-2023

- Initial clone of BoLD deployed
- Initiation of OS code dev
- Development of governance
- Technical roadmap

Q4-2023-Q2-2026

- Roadmap implementation

Participants

- **CBG** - data producer
- **Naturalis** - data producer
- **EBI** - consumer
- **Unite** - consumer
- **Elixir** - community steward
- **RBGE** - community steward


Thank you for your attention

Q&A

biodiversa+
European Biodiversity Partnership

Open discussion session - How Biodiversa+ can help achieve effective data interoperability

biodiversa+
European Biodiversity Partnership

Objectives of this session

By Hilde Eggermont, BelSPO, Biodiversa+ Chair & Coordinator

Objectives

- Identify barriers preventing data interoperability
- Factors encouraging data interoperability
- Possible support from Biodiversa+

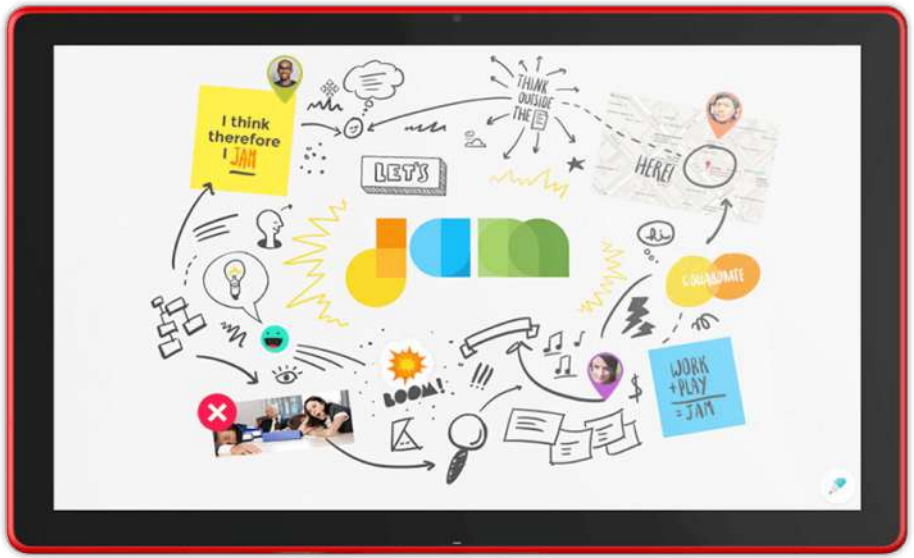
Support from Biodiversa+

- ✓ Enriching the EuropaBON work in relation to description of data flows (from raw observation, to indicators, to multiple uses);
- ✓ Mapping of (sub)national data architectures => identify best practices;
- ✓ Promoting best practices through capacity building activities;
- ✓ Promoting open data and standards for exchanging information, in line with the EU digitization strategy;
- ✓ Promoting use of European and global research infrastructures; ...

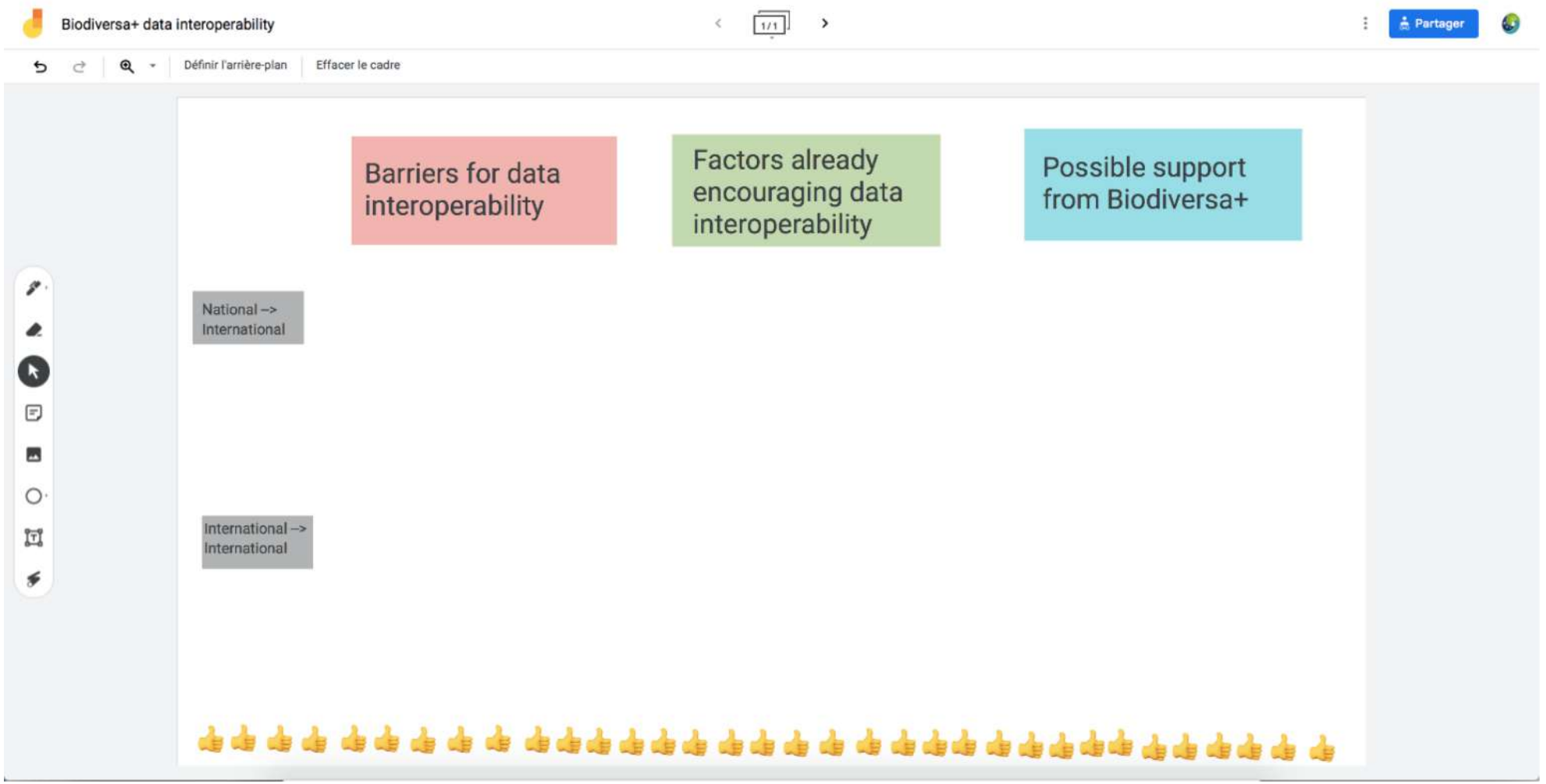
biodiversa+
European Biodiversity Partnership

How to use Jamboard?

By Cécile Mandon, FRB, Biodiversa+ Officer



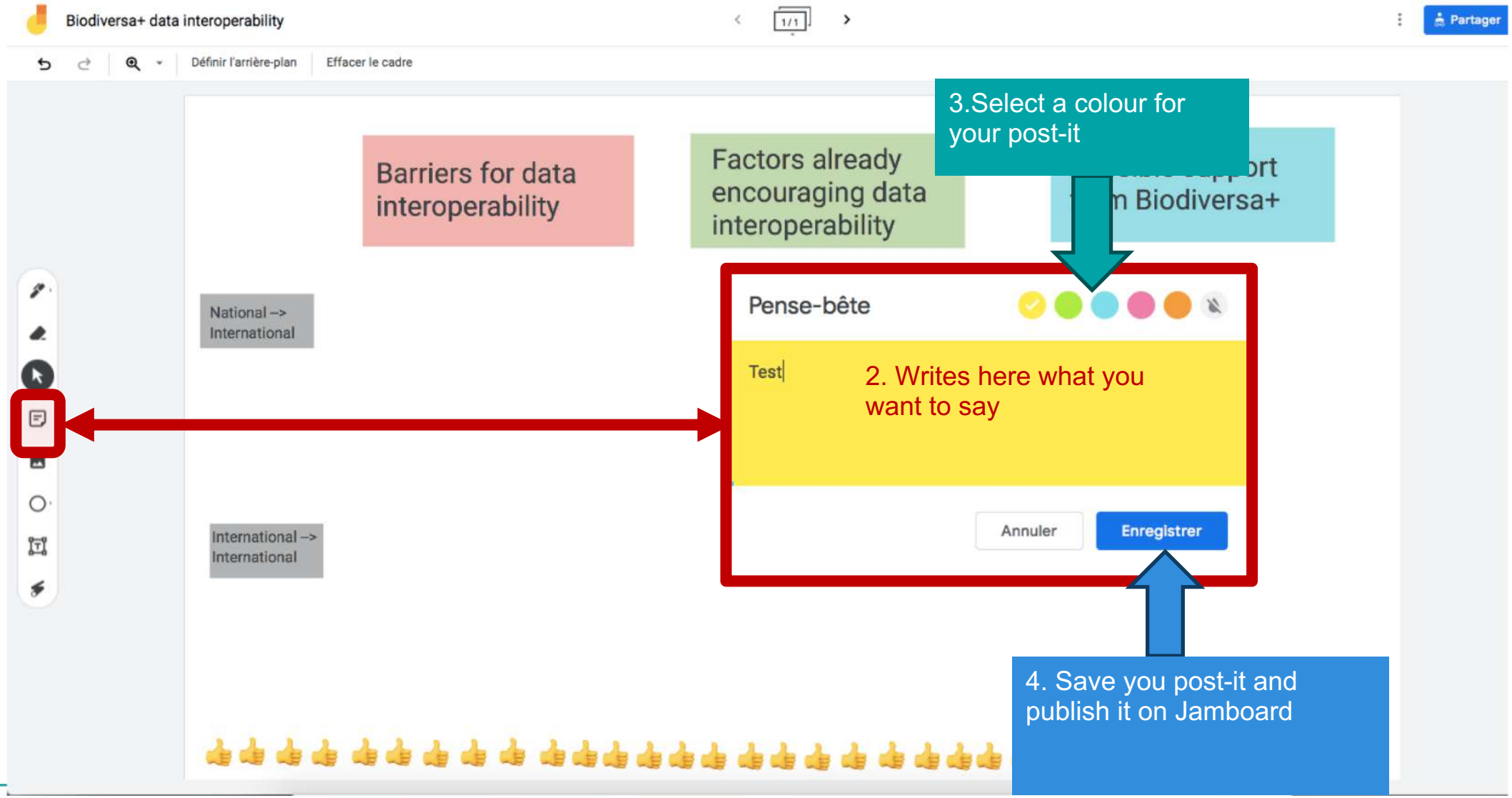
1. Click on the link we will share in the chat and join the board!



https://jamboard.google.com/d/1jNoW2BkdudgpgTILpGamW9h0d8bvpKi4HJzOZO3UJKw/edit?usp=sharing

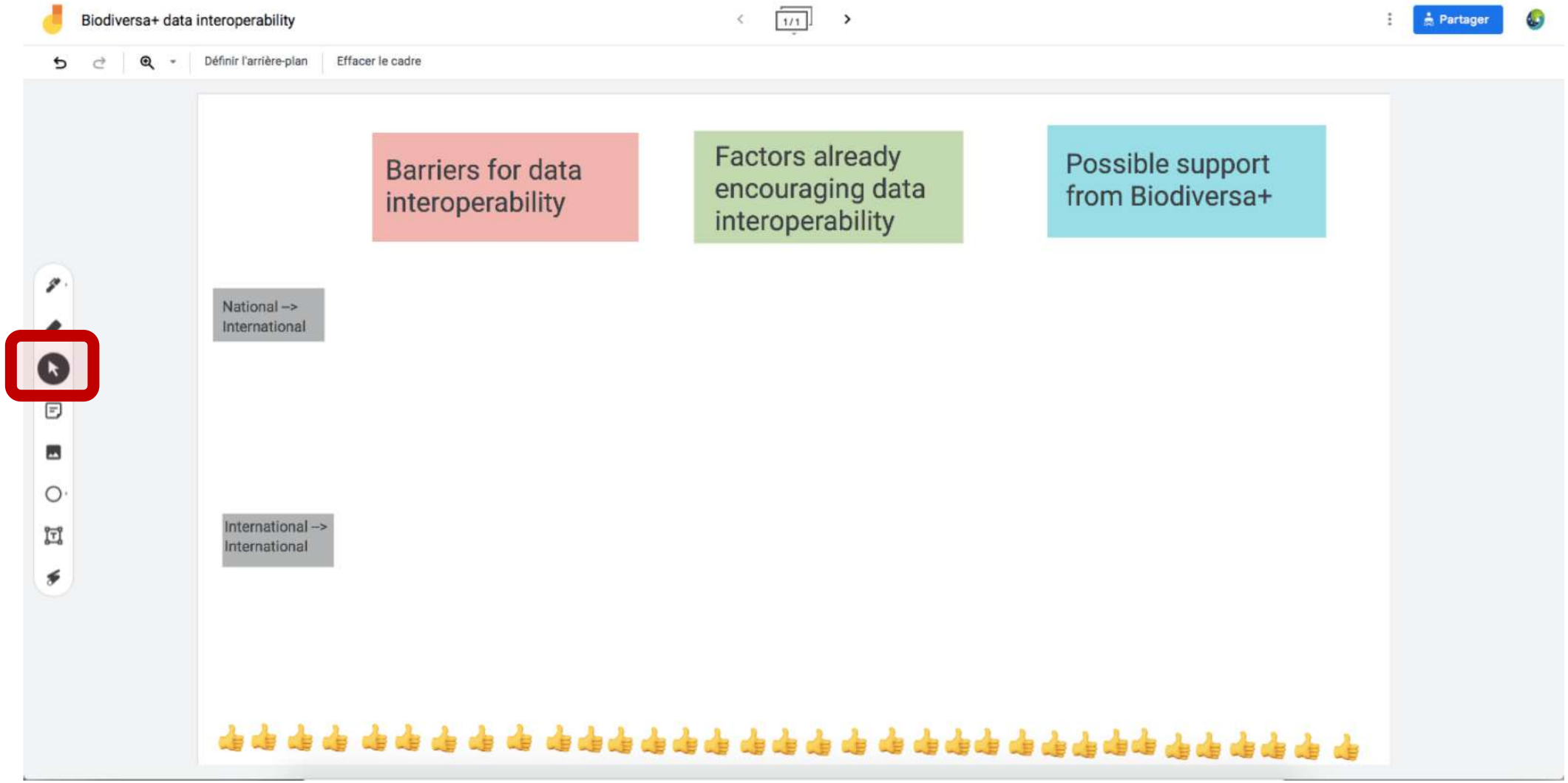
Add post-its

1. Click here to add a post-it



Move your post-it

1. Click here
2. Select your post-it by clicking on it
3. Move your post-it on the board



interoperabil … es to be tackle

Possible support from Biodiversa+

Use of new technologies to tackle not only interoperability but also their use and impact

Semantic artefact development requires people who know not only one institution's data and metadata architectures, but multiple ones. Exchange of experts for short period of times across organisations

Best practices in semantic artefact development/management/governance for establishing guidelines.

Proliferation of metadata standards and paradigms. Hard to make the "correct" choice.

…ecific funding an …pport for …ndards …velopment are …eded (TDWG, RI etc)

Can propose best practices for leveraging funds for the modest but non-negligible costs of standardising and sharing data, otherwise this effort is left out of budgets and plans

Adherence to the relevant standards is required in funding programs, including guidance and a list of standards (for information purposes, the list may not be comprehensive)

facilitate integration across sectors (agri, forest, nature, water etc.) to create co-benefits and improve cost effectiveness also in line with global and EU policy targets

harmonise data and methods across the heterogeneous research landscape and link with national, regional (EU) and global institutions

Help documenting the evolving landscape incl. key standards (decision tree?), and continue promoting interoperability and integration

Much existing data is collected in templates that lose the raw data, e.g. for EU reporting.

There is no clear guidance/mandate from funding agencies to adhere to specific standards that facilitate interoperability

Define clear tasks and goals for the different organisations and align these (exactly - who is doing what)

There will continue to be a proliferation of metadata standards. The issue is how to translate the information between standards

Biodiversa+ could coordinate the interaction between all these organisations/ initiatives to build something big that works

Machine observations have large data storage needs and IT infrastructure is not keeping up

Creation / application of standards (e.g. metadata); provision of a central register for those standards (one stop shop)

Although we have provenance tools, such as ORCID, ROR, DOIs etc they are still not being consistently used by the community

Proliferation of ontologies which are not easy to cross-walk. Requires initial human investigation and sense-making to build machine-readable repos.

Mandating data management plan and the use of certain data management strategies within Biodiversa+ projec

Define "playbooks" for data management at different stages of the cycle (collect, analyze, publish, etc.)

data interoperability and semantic meaning (mentioned by Life Watch): foster this to allow interdisciplinary use

actor together with the EU science service in make (BioAgora) and the KCBD to see how these infrastructures and biodiv monotoring (Monitoring centre on Biodiv on European level?) could be a pillar under the

Always needed: a one stop shop for collecting the relevant information needed in terms of monitoring initiatives and FAIR data (e.g. standards, protocols)

From the management perspective it is important to see what dataflows are/can be channeled into official reporting mechanisms

Integration of machine learning results, with human interpreted results has no standards

Machine learning processes are not always transparently documented (e.g. training sets)

Data collected by different agencies within government apply different standards - so biodiversity data interoperability can be seen as part of the process of mainstreaming.

A lack of harmonization on what to monitor

common guidelines to facilitate interoperability between monitoring data from the management and research infrastructures

Establish common best practices and guidelines

Awareness/ increase the understanding of existing tools / standards (uptake at the relevant levels)

Metadata review and validation protocols to ensure the quality and consistence of information workflows

Concepts such as interoperability, ontologies, semantics etc are still unfamiliar to many data collectors

facilitate implementation of domain specific standards and standards wich can be common across sectors

Too little attention to capacity building in the poorer areas of Europe

capitalise on increasing profile of / evolving techniques for 'explainable AI' with biodiversity- and provenance-specific funding calls

Identification of the minimum metadata sets of information that can be translated in all the major existing metadata schemas

provide knowledge on existing methods and standards for harmonisation and data interoperability to key actors and stakeholders (technical knowledge)

Biodiv+ can provide expertise in data management and interoperability

Conclusions of the workshop

By Alberto Basset, MUR and Hilde Eggermont, BelSPO, Biodiversa+ Chair and Coordinator

REC

The plenary sessions of this meeting will be recorded and shared on the Biodiversa+ website and Youtube channel

biodiversa+
European Biodiversity Partnership

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Thank you!

www.biodiversa.org

contact@biodiversa.org

BiodiversaPlus