BiCIKL Biodiversity Community Integrated Knowledge Library

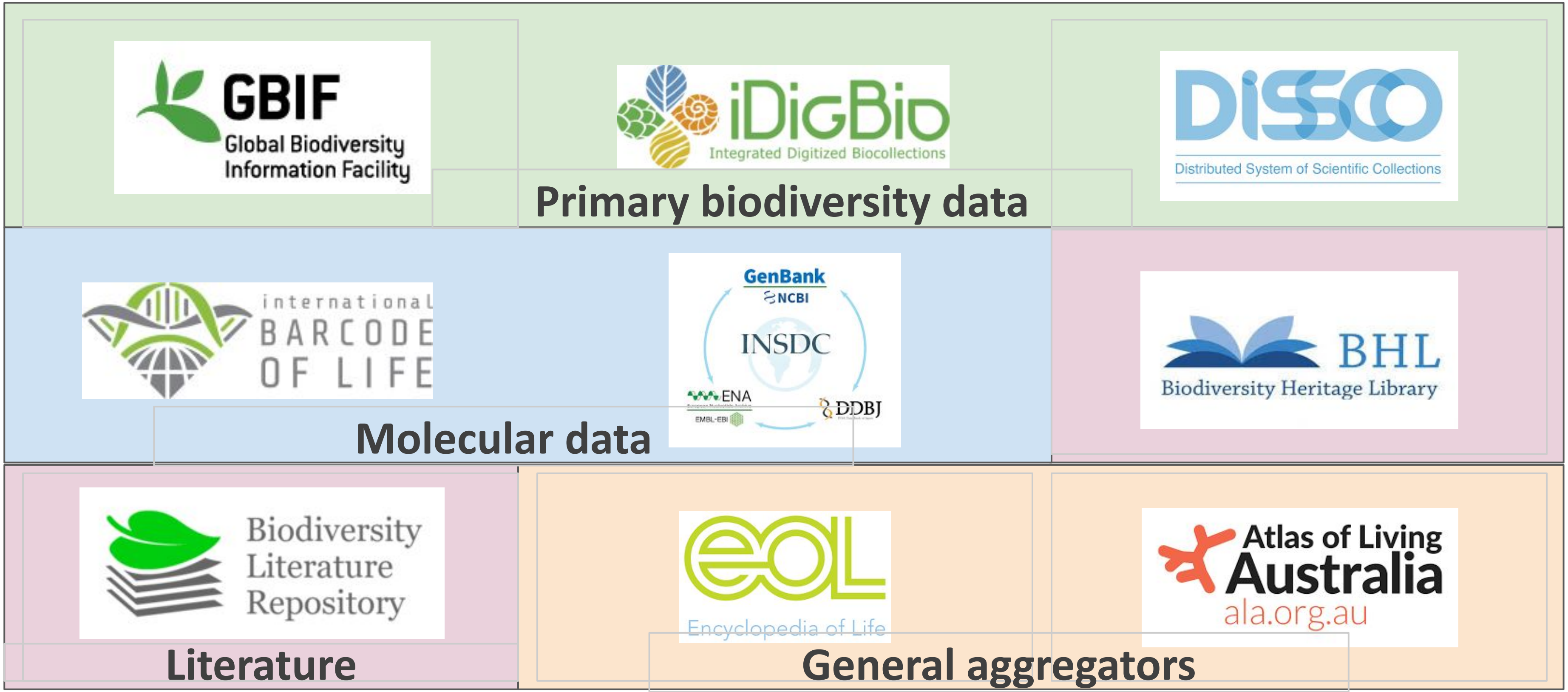
# Goals and Ambitions of the BiCIKL project

Lyubomir Penev § the BiCIKL Consortium

BiCIKL Kick-off Meeting
27 May 2021

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

# The three main realms of biodiversity data



Primary biodiversity data

GBIF Global Biodiversity Information Facility

iDigBio Integrated Digitized Biocollections

DiSSCo Distributed System of Scientific Collections

Molecular data

international BARCODE OF LIFE

GenBank NCBI INSDC ENA EMBL-EBI DDBJ

BHL Biodiversity Heritage Library

Literature

Biodiversity Literature Repository

General aggregators

eoL Encyclopedia of Life

Atlas of Living Australia ala.org.au

# The challenges

- Imbalances in **regional engagement** in biodiversity informatics.
- Uneven progress in **data mobilization and sharing.**
- Insufficient use of uniform **persistent identifiers** for data.
- Redundant and incompatible processes for **cleaning and interpreting data.**
- The absence of functional mechanisms for experts to **curate and improve data.**
- **Linking between the biodiversity data infrastructures is still in infancy.**

https://biodiversityinformatics.org

# BiCIKL brief profile

- **Biodiversity Community Integrated Knowledge Library**
- **Work programme:** Integrating Activities for Starting Communities (INFRAIA-02-20203)
- **Duration**: 3 years (1 May 2021-30 April 2024)

public sector

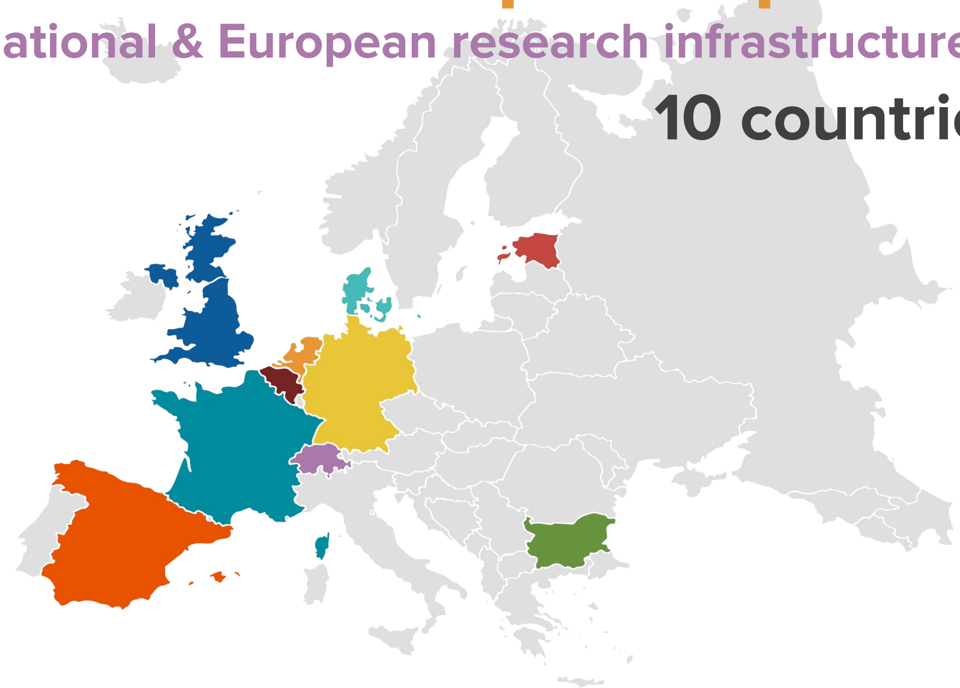
private sector

cross-disciplinary

14 participants

International & European research infrastructures

10 countries



BiCIKL

# The BiCIKL partners

PENSOFT®

Naturalis Biodiversity Center

PLAZI
TAKING CARE OF FREEDOM

Meise Botanic Garden

EMBL-EBI

elixir

CERN

CETAF
CONSORTIUM OF EUROPEAN TAXONOMIC FACILITIES

MUSÉUM NAL HIST NATURELLE

MUSÉUM
NATIONAL D'HISTOIRE NATURELLE

SIB
Swiss Institute of Bioinformatics

1632
UNIVERSITY OF TARTU

LifeWatch ERIC

GBIF
Global Biodiversity Information Facility

species 2000
Catalogue of Life

Biodiversity Information Standards
TDWG

BG BM
Botanischer Garten & Botanisches Museum Berlin

# BiCIKL Research Infrastructures

ARPHA XML

OpenBiodiv

DiSSCo

LifeWatch eInfra

Biodiversity and Ecosystem VREs

PlutoF

Zenodo

SIBiLS

TreatmentBank

BLR

Catalogue of Life

GBIF

ENA

MeiseBG

FUB-BGBM

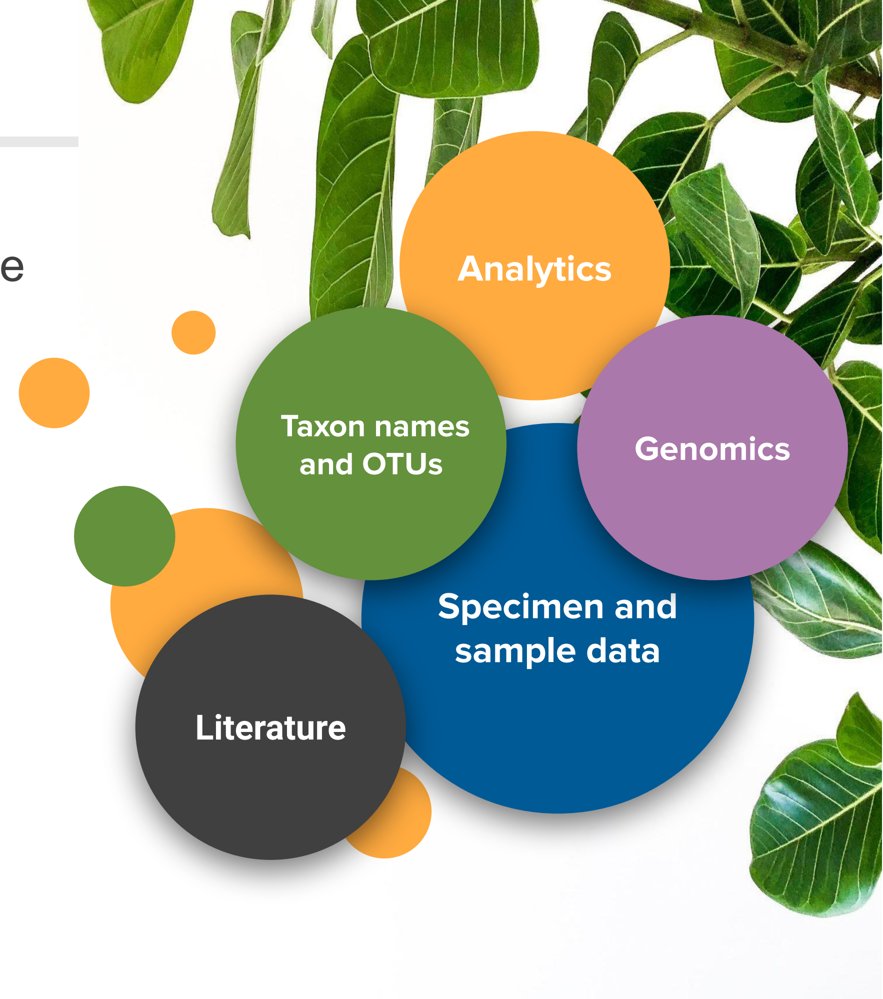
Europe PubMed Central

# The BiCIKL vision

BiCIKL aims to catalyse the culture change in the way biodiversity data are identified, linked, integrated and re-used across the research cycle. By doing so, BiCIKL helps to increase the transparency, trustworthiness and efficiency of the entire research ecosystem.

# Rationale

- Biodiversity **data deluge**:
  - > 500 million pages of published literature
  - > 2 billion specimens in collections
  - > 1.8 million species described
  - many billions of gene sequences
- How do we transform **raw data** and such from published narratives into **actionable knowledge**?
- How do we **link digital objects together?**
- Where and how do we **store, annotate, manage and use links** between data?



Analytics

Taxon names and OTUs

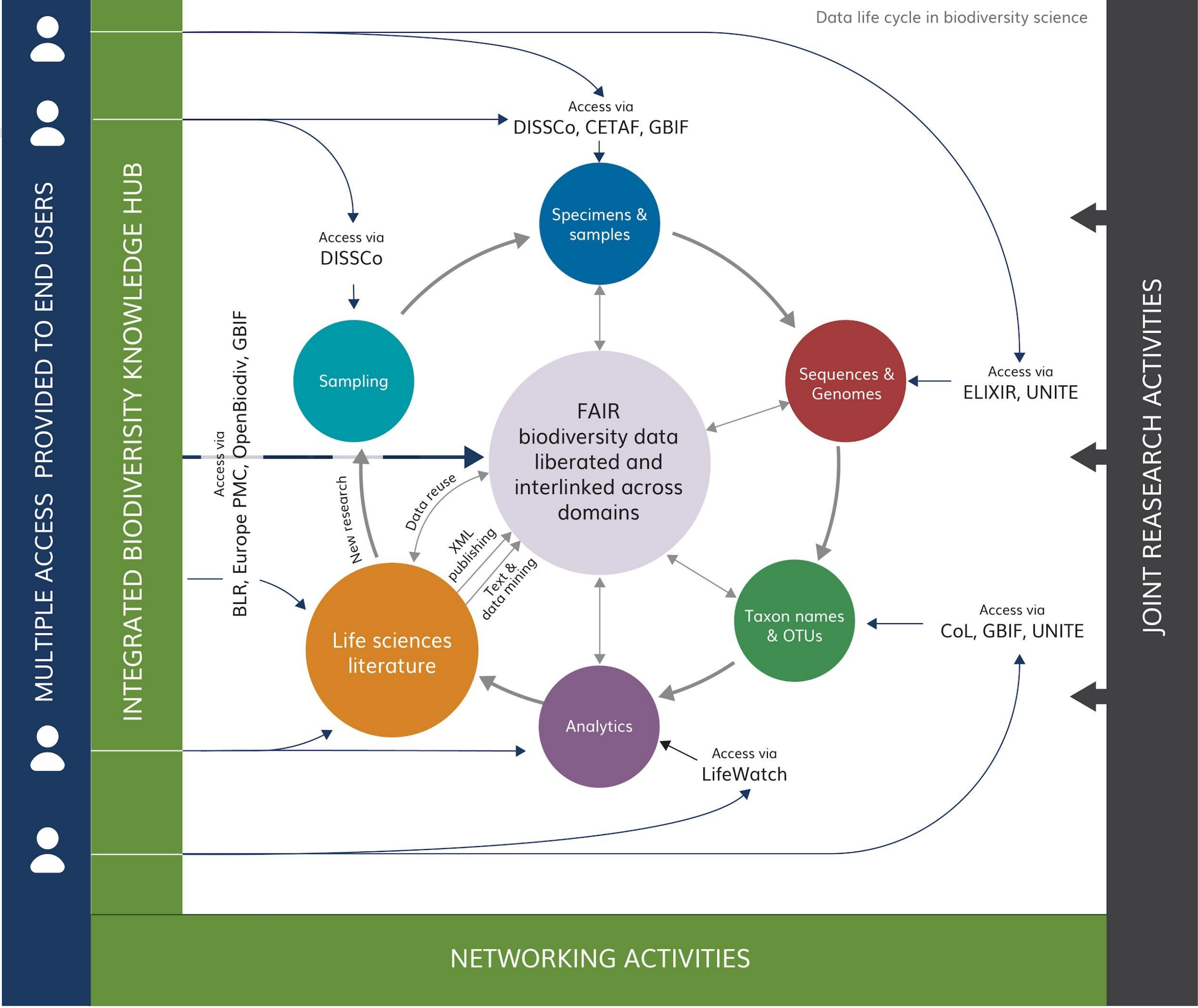
Genomics

Specimen and sample data

Literature

# Mission

1. **ACCESS** to **data, associated tools and services** at each stage & along the entire research cycle.
2. **LINKS** between: **specimens → genetic sequences → species → analytics → publications → biodiversity knowledge graph → re-use.**



Data life cycle in biodiversity science

MULTIPLE ACCESS PROVIDED TO END USERS

INTEGRATED BIODIVERISITY KNOWLEDGE HUB

JOINT REASEARCH ACTIVITIES

NETWORKING ACTIVITIES

Access via DISSCo, CETAF, GBIF

Access via DISSCo

Access via ELIXIR, UNITE

Access via CoL, GBIF, UNITE

Access via LifeWatch

Access via BLR, Europe PMC, OpenBiodiv, GBIF

Specimens & samples

Sampling

Sequences & Genomes

FAIR biodiversity data liberated and interlinked across domains

Taxon names & OTUs

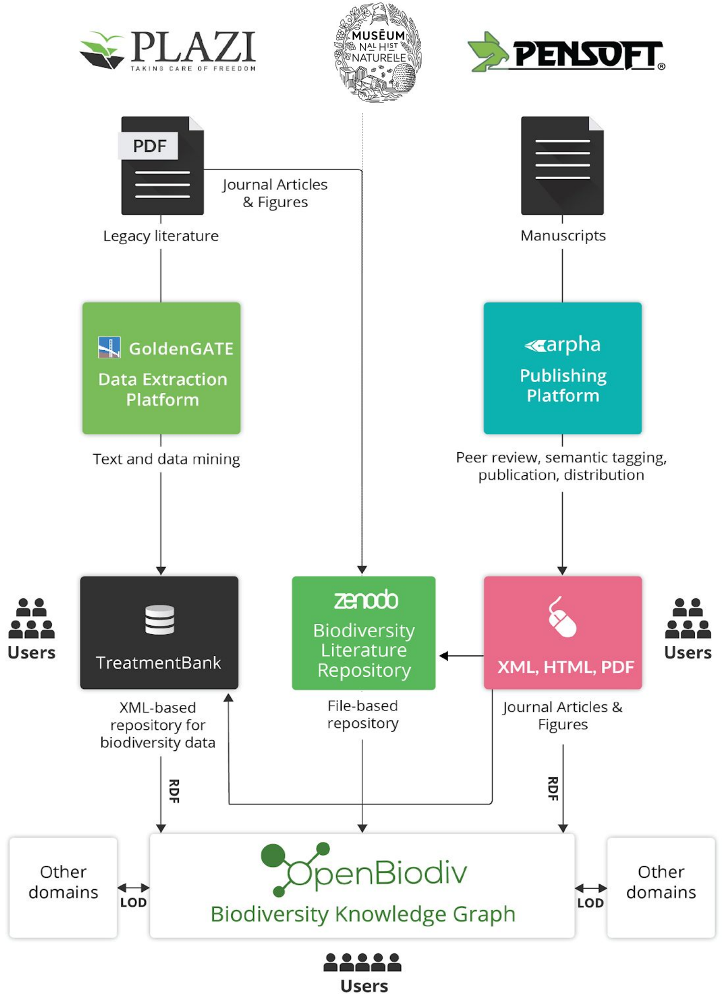
Life sciences literature

Analytics

# Special focus on literature

3. **Methods, tools** and **workflows** for **harvesting, liberating, linking, and re-using of sub-article-level data, extracted from literature**.

Data from both **legacy** (PDF-based) literature and **prospective** (XML-based) publishing come together.


PLAZI
TAKING CARE OF FREEDOM
MUSÉUM NAL HIST NATURELLE
PENSOFT
PDF
Journal Articles & Figures
Legacy literature
Manuscripts
GoldenGATE Data Extraction Platform
arpha Publishing Platform
Text and data mining
Peer review, semantic tagging, publication, distribution
Users
TreatmentBank
zenodo Biodiversity Literature Repository
XML, HTML, PDF
Users
XML-based repository for biodiversity data
File-based repository
Journal Articles & Figures
RDF
RDF
Other domains
LOD
OpenBiodiv
Biodiversity Knowledge Graph
LOD
Other domains
Users

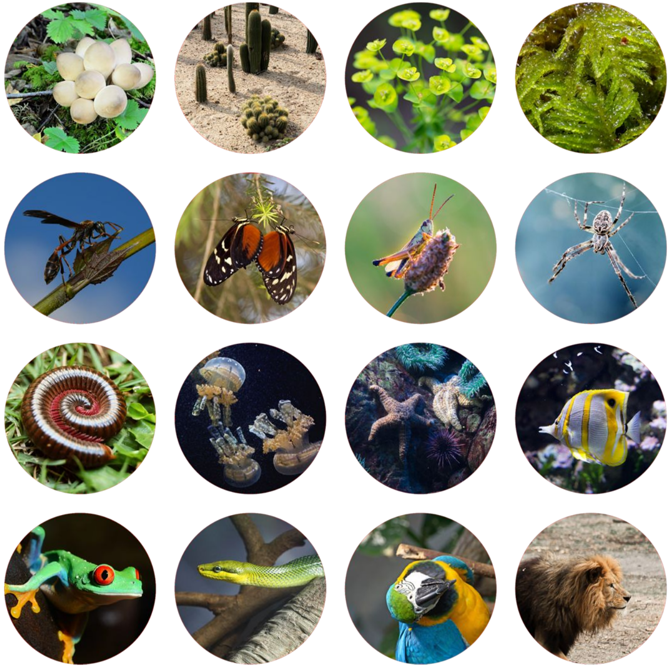
# Specific objectives

1. Develop and implement **open science** research practices

2. Harmonise **policies, standards and technologies** between the participating key infrastructures

3. Engage all actors and other stakeholders in the process of data upload/ingestion and **FAIR data delivery**

4. Improve **researchers' capacity** through enhanced digital skills in linking open data



# Specific objectives

5 Provide a **one-stop access point** to guidelines, standards, data and services via the newly developed Biodiversity Knowledge Hub (BKH)

6 Foster **joint research agendas** of European and international researchers

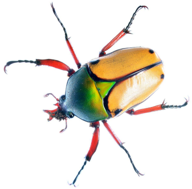
7 Support industrial innovation in building and implementation of next-generation, standards-aligned and **semantics-based publishing workflows**

# Specific objectives

8 **Liberate** and re-use the vast knowledge and **data imprisoned in literature**.

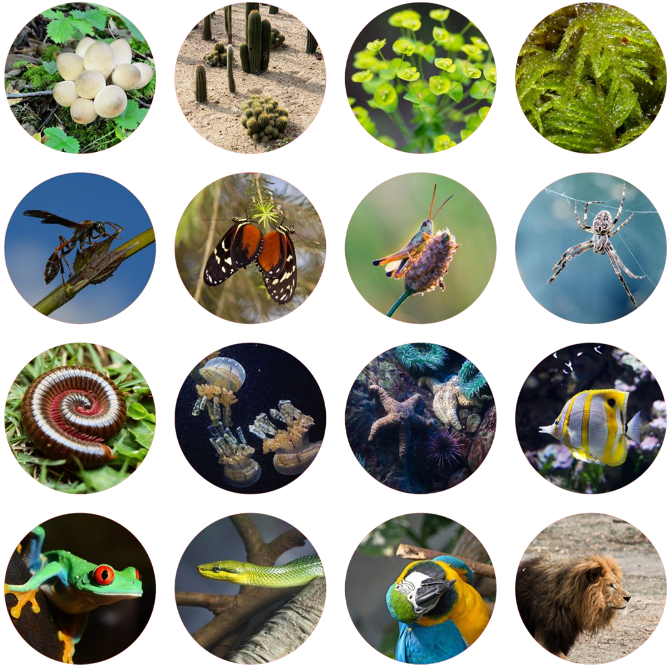
9 Support researchers' **access to the Linked Open Data** world through interoperable, AI-based, **FAIR Data Place (FDP)** interface, discovering & validating links between different resources.

10 Facilitate interdisciplinary research and **generation of new knowledge** through linking of FAIR data from different resources and domains



# The BiCIKL key products

1. A vibrant community equipped with tools for search of & access to **FAIR interlinked data**
2. Interlinked **corpora of knowledge**, used by biodiversity & related research domains
3. Automated tools & workflows for **data liberation & FAIR-isation** from literature
4. Semantic-based journal production workflows for **publication and re-use of FAIR biodiversity data**



# The BiCIKL Pillars

**Holistic targeted assembly of interlinked, machine-readable FAIR biodiversity data**

## Networking Activities

## Trans-national and Virtual Access

## Joint Research Activities

# The BiCIKL Work packages

## Networking Activities (NA) Pillar

**WP1**

NA-01 Coordination and interoperability of infrastructures through harmonisation of community policies, standards and guidelines

**WP2**

NA-02 Defining & co-designing the Biodiversity Knowledge Hub (BKH) and operational training

**WP3**

NA-03 Implementation, stakeholder engagement and outreach for the Biodiversity Knowledge Hub

## Trans-national and Virtual Access (TA+VA) Pillar

**WP4**

TA-01 Trans-national access to biodiversity infrastructure and services

**WP5**

VA-01 Virtual access to biodiversity infrastructure and services

## Joint Research Activities (JRA) Pillar

**WP6**

JRA-01 Liberation of data from literature, next-generation semantic publishing and delivery of FAIR data

**WP7**

JRA-02 Providing core access services and FAIR data on specimens and samples

**WP8**

JRA-03 A data foundation for connected molecular, natural history collections and taxonomic data

**WP9**

JRA-04 Delivering a trusted and evolving taxonomic framework for data integration

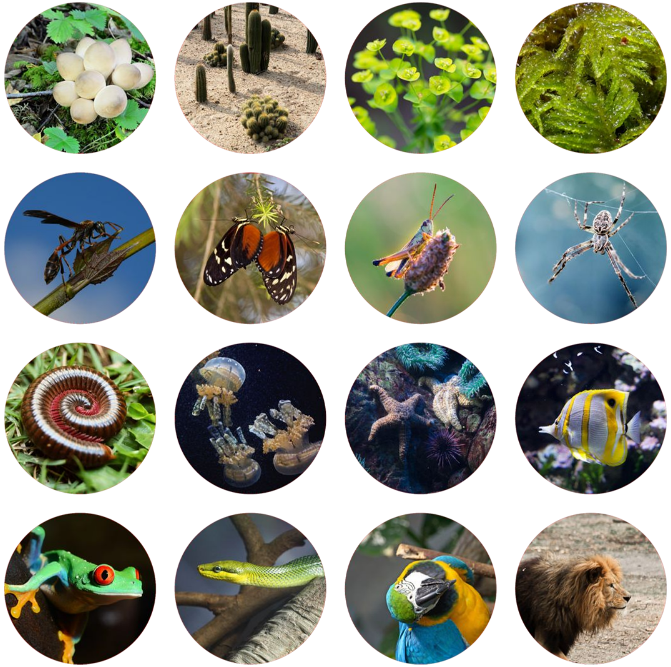
**WP10**

JRA-05 FAIR Data Place: linking, finding and access

**WP11 Project management**

# Networking (NA): 24.7 % of the budget

1. Standards & harmonisation of FAIR data linking between RIs
2. Training and capacity building
3. Communication, dissemination and outreach
4. Concept design of the Biodiversity Knowledge Hub (BKH)
5. Building and promotion of the BKH



# Trans-national and Virtual (TA/VA) access: 20.5% of the budget!

ARPHA XML

OpenBiodiv

DiSSCo

LifeWatch eInfra

Biodiversity and Ecosystem VREs

PlutoF

Zenodo

SIBiLS

TreatmentBank

BLR

Catalogue of Life

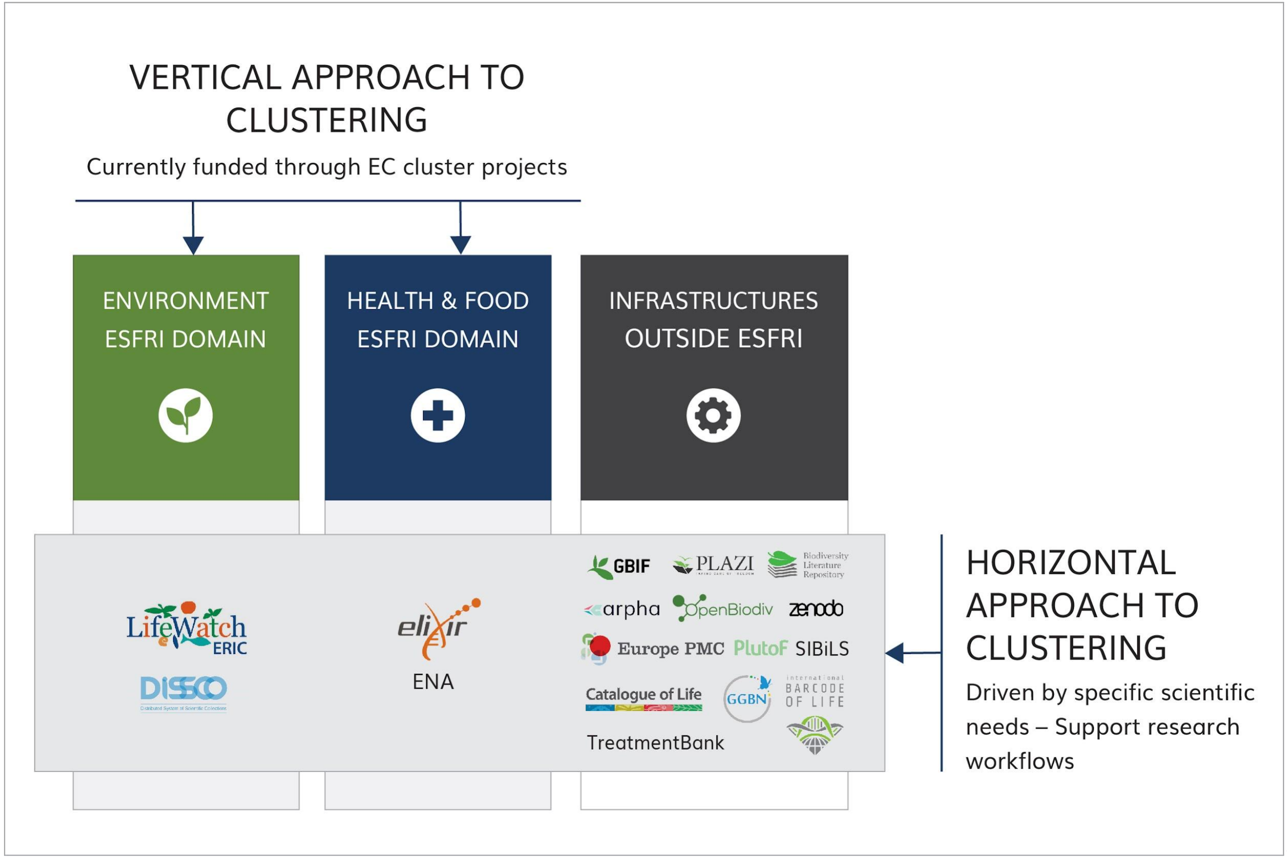
GBIF

ENA

MBG

FUB-BGBM

# Horizontal linkages between research infrastructures



VERTICAL APPROACH TO CLUSTERING

Currently funded through EC cluster projects

ENVIRONMENT ESFRI DOMAIN

HEALTH & FOOD ESFRI DOMAIN

INFRASTRUCTURES OUTSIDE ESFRI

ENA

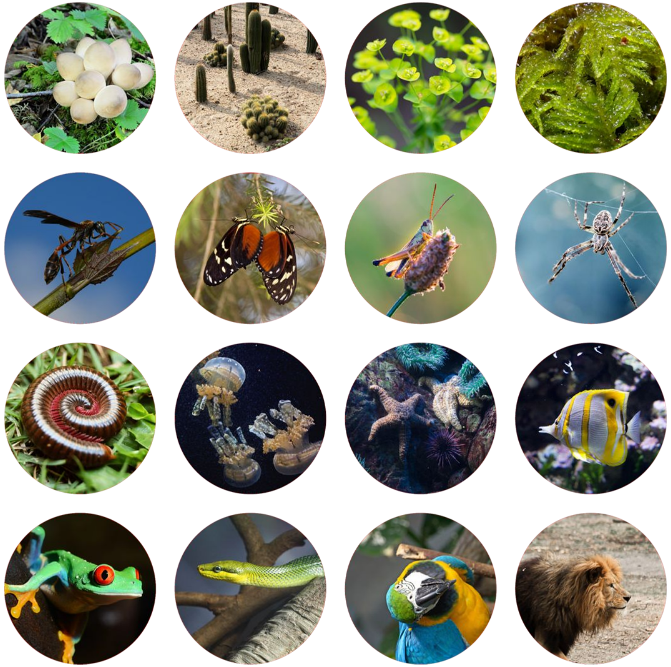
TreatmentBank

HORIZONTAL APPROACH TO CLUSTERING

Driven by specific scientific needs – Support research workflows

# Joint Research Activities (JRA): 43.3 % of the budget

1. Open APIs at each RI following community accepting standards
2. New tools and services at each participating RI **towards data linking with others**
3. Testing of access to linked data through TA/VA
4. Fair Data Place for search, access and storage of links between data



# The BiCIKL key question: What is ‘data linking’?

- **Linking between individual data records**
  - Through text string matches
  - Through persistent unique identifiers (PIDs)
  - Mostly uni- or bi-directional
  - Linking through literature (citations *sensu lato*)
- **Linked Open Data in the cloud**
  - Always through stable HTTP identifiers (URI)
  - Fully interoperable (RDF triples and other)
  - Machine-actionable
  - Multi-directional, anyone to anyone
- **High-level linking between two and/or many Research Infrastructures**

# Why linking data

**A simple answer (among many others):**

Due to the enormous data deluge, especially in (meta)genomics, and the disruptive changes towards a digital world, it is not sufficient for even a renown taxonomist to say: “This is Species X”

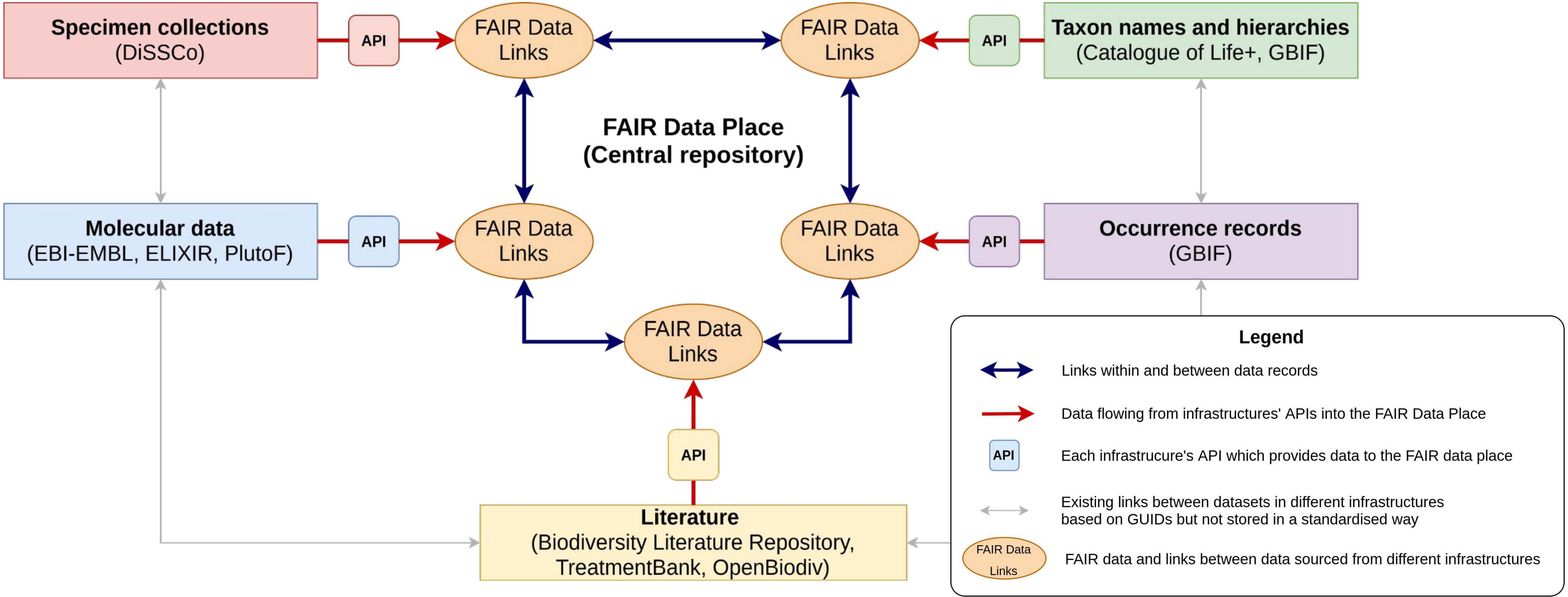
Rather, the reasonable statement would be: This is Species X, according to Treatment X, Specimen(s) XYZ and Sequence X, with a direct access to the data.

# Technological approaches to data linking

**Linking can be performed through / between:**

- Relational databases & Data warehousing
- Fair Data Objects (Open Digital Specimens)
- Linked Open Data (e.g. between RDF triples)
- Nanopublications
- Other?

# Where and how to link biodiversity data? Where and how to store and use these links?



**Specimen collections**
(DiSSCo)

**Molecular data**
(EBI-EMBL, ELIXIR, PlutoF)

**Taxon names and hierarchies**
(Catalogue of Life+, GBIF)

**Occurrence records**
(GBIF)

**Literature**
(Biodiversity Literature Repository, TreatmentBank, OpenBiodiv)

API

FAIR Data Links

**FAIR Data Place**
**(Central repository)**

**Legend**

- ↔ Links within and between data records
- → Data flowing from infrastructures' APIs into the FAIR Data Place
- API Each infrastrucure's API which provides data to the FAIR data place
- ↔ Existing links between datasets in different infrastructures based on GUIDs but not stored in a standardised way
- FAIR Data Links FAIR data and links between data sourced from different infrastructures

# Thank you for your attention and Good Luck, BiCIKL!

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