BiCIKL

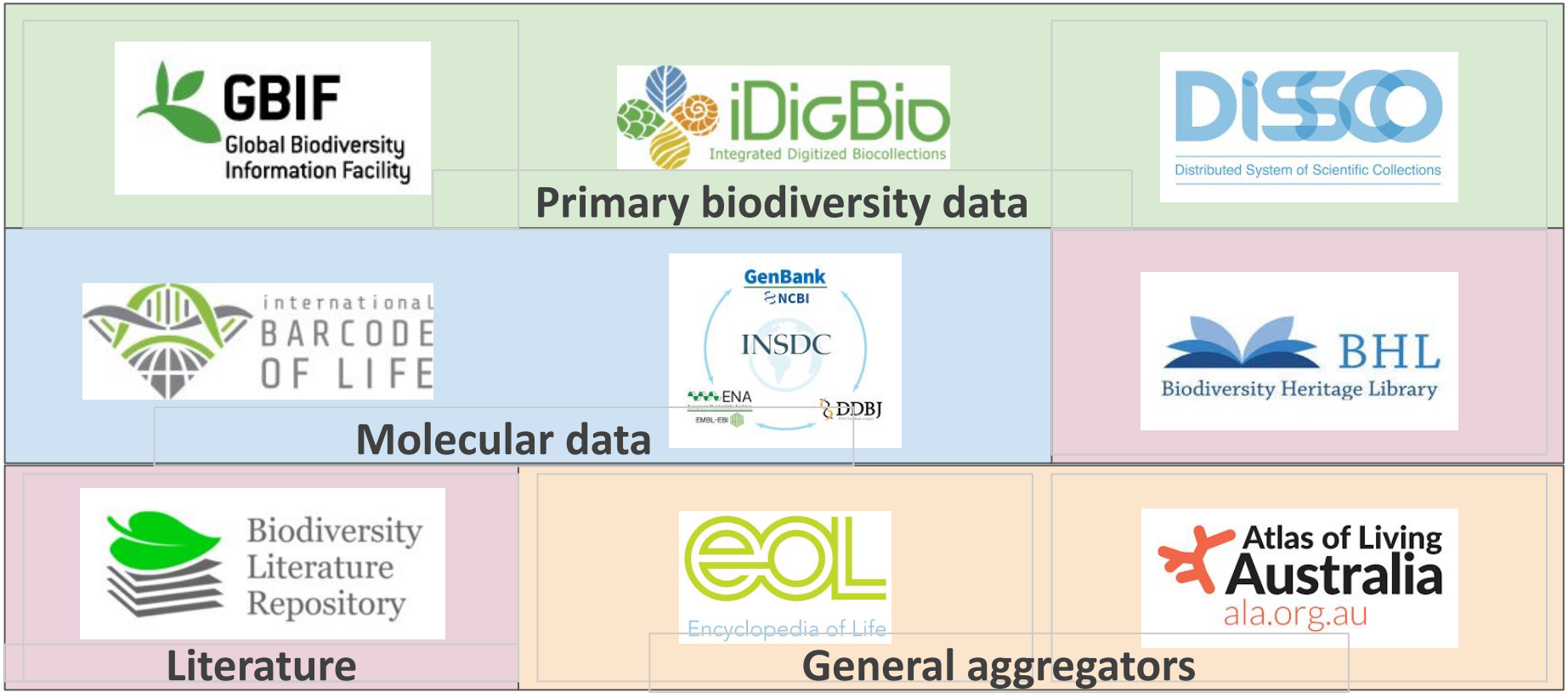
Biodiversity Community
Integrated Knowledge
Library

Goals and Ambitions of the BiCIKL project

Lyubomir Penev § the BiCIKL Consortium

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



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.**

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)

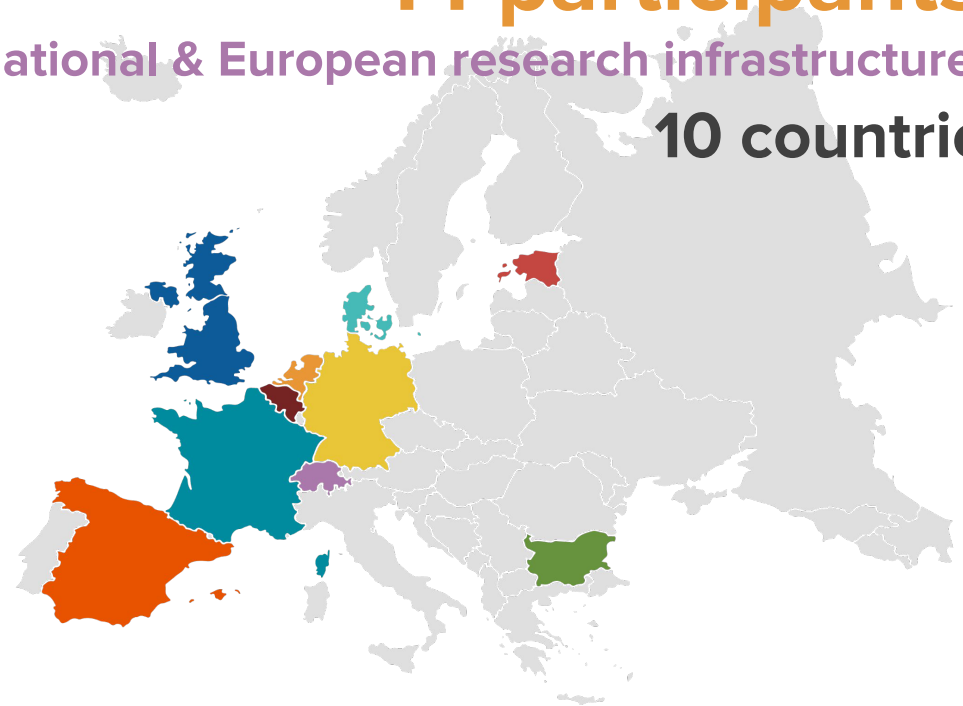
public sector private sector

cross-disciplinary

14 participants

International & European research infrastructures

10 countries



BiCIKL

The BiCIKL partners

BiCIKL Research Infrastructures

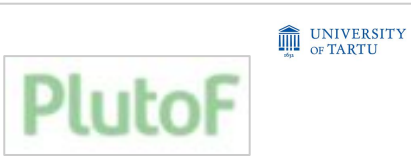
ARPHA XML

OpenBiodiv

DiSSCo

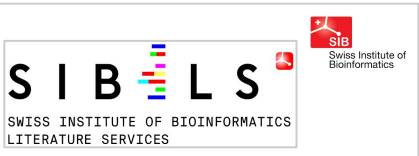
LifeWatch eInfra

Biodiversity and Ecosystem VREs



PlutoF

Zenodo



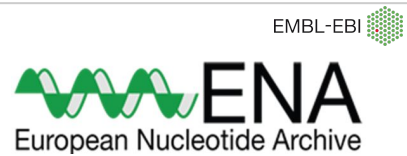
SIBiLS

TreatmentBank

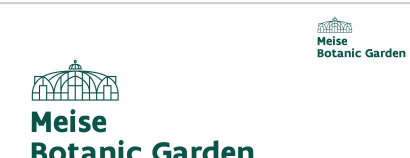
BLR

Catalogue of Life

GBIF



ENA



MeiseBG

FUB-BGBM

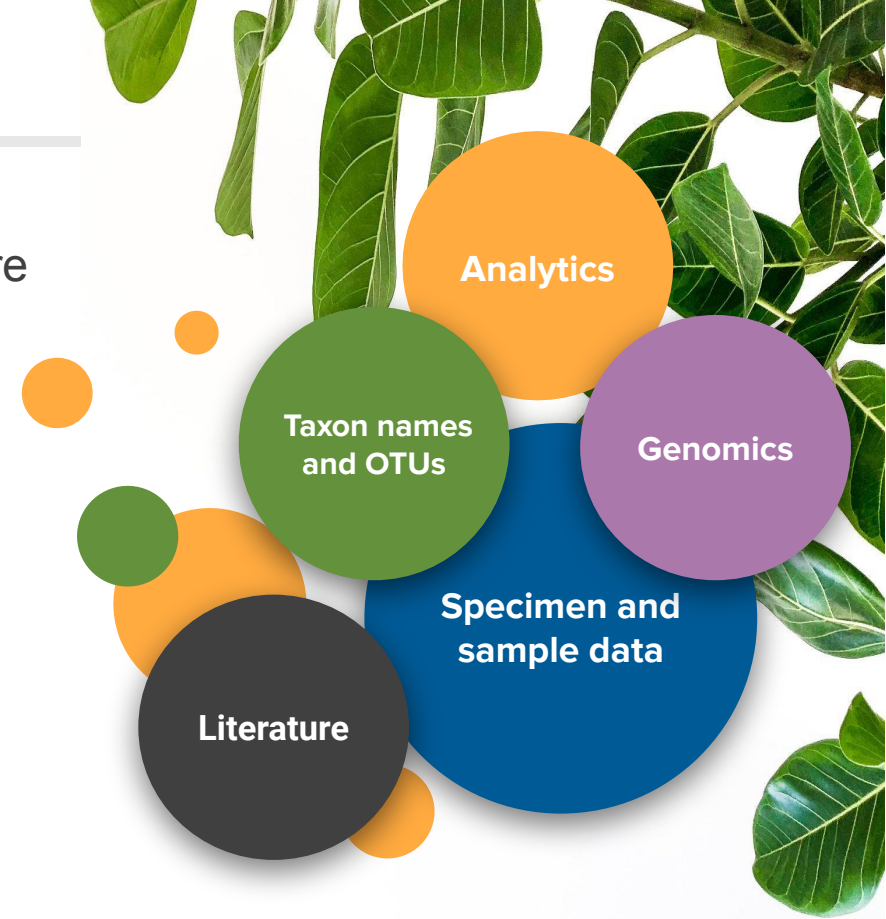
BiCIKL

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 lifecycle. By doing so, BiCIKL helps to increase the transparency, trustworthiness and efficiency of the entire research ecosystem.

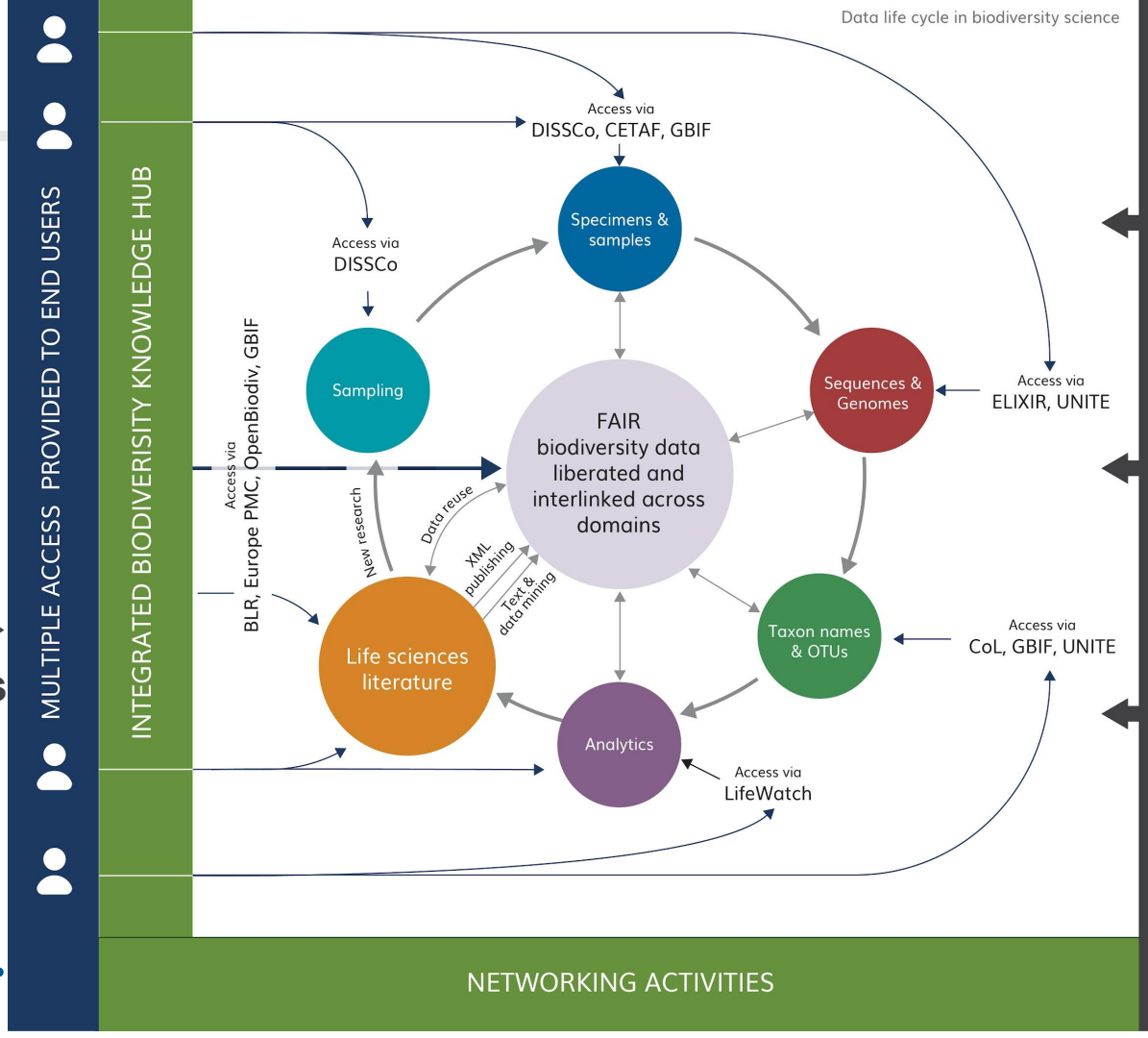
Rationale

- 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?



Mission

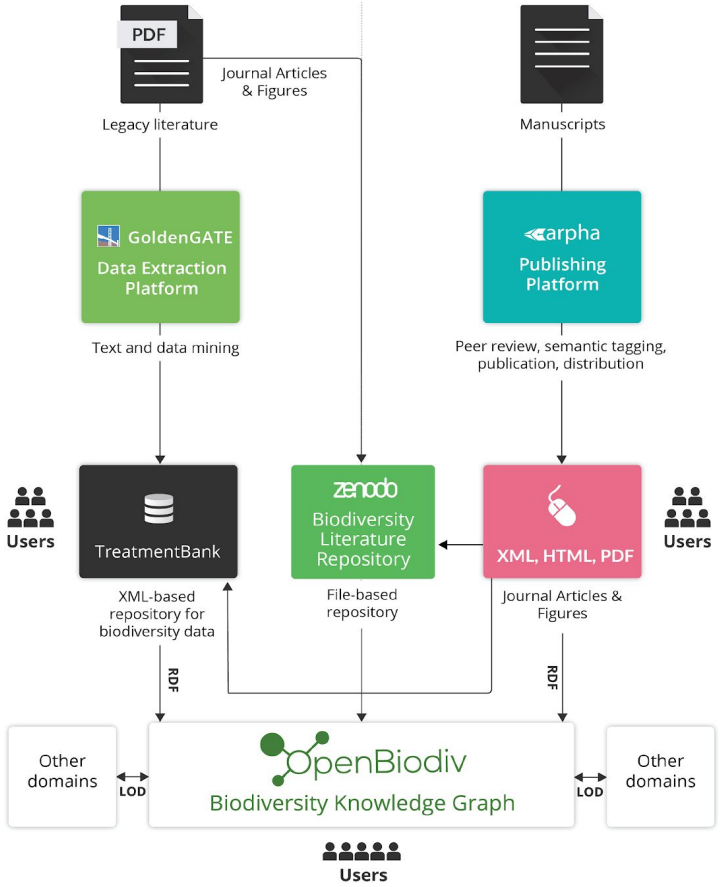
1. **ACCESS** to **data, associated tools and services** at each stage & along the entire research cycle.
2. **LINKS** between: **specimens → genetic sequences → species → analytics → publications → biodiversity knowledge graph → re-use.**



Special focus on literature

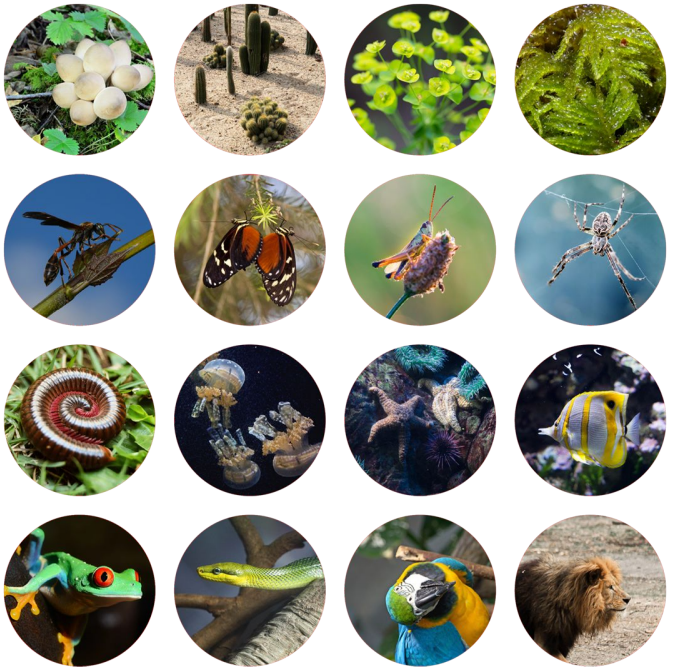
3. **Methods, tools** and **workflows** for **harvesting, liberating, linking, and re-using of sub-article-level data, extracted from literature**.

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



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

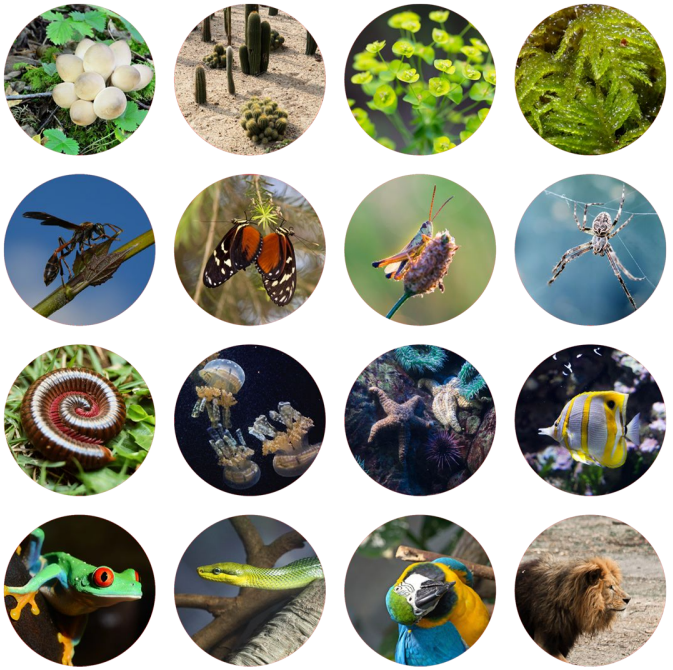
8 **Liberate** and re-use the vast knowledge and **data imprisoned in literature**.

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

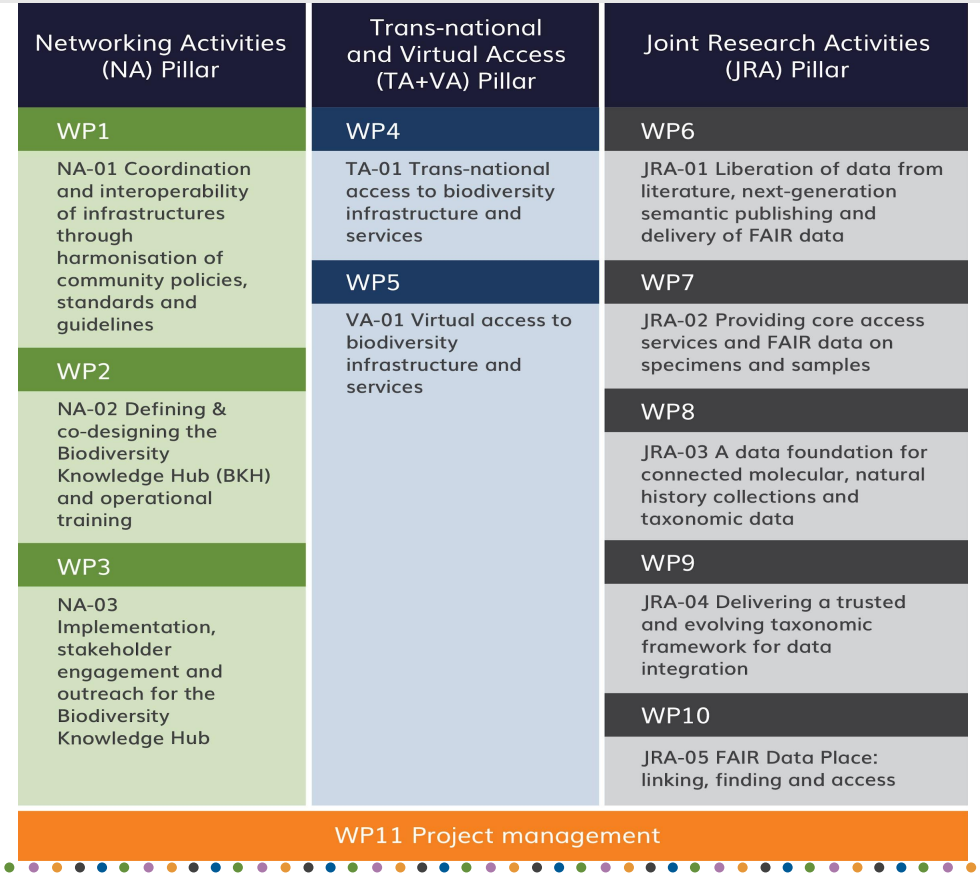
Holistic targeted assembly of interlinked, machine-readable FAIR biodiversity data

Networking Activities

Trans-national and Virtual Access

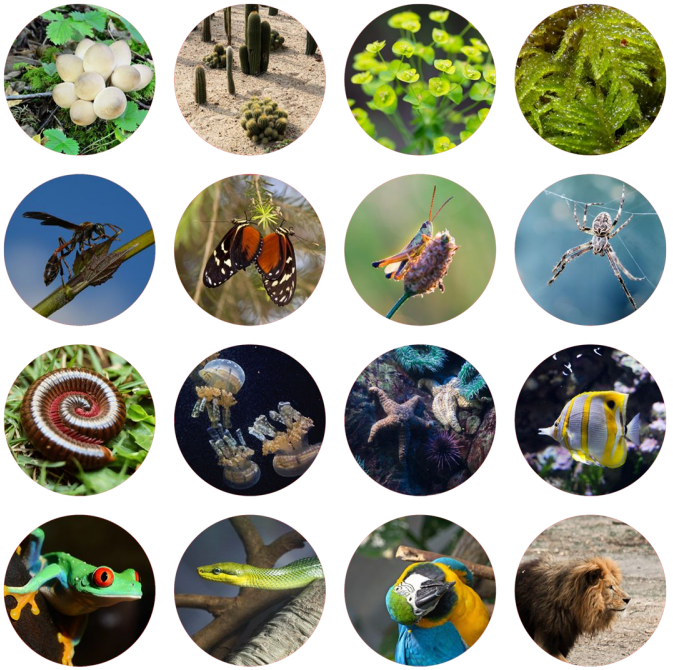
Joint Research Activities

The BiCIKL Work packages



Networking (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 access: 20.5% of the budget!

ARPHA XML

OpenBiodiv

DiSSCo

LifeWatch eInfra

Biodiversity and Ecosystem VREs

PlutoF

Zenodo

SIBiLS

TreatmentBank

BLR

Catalogue of Life

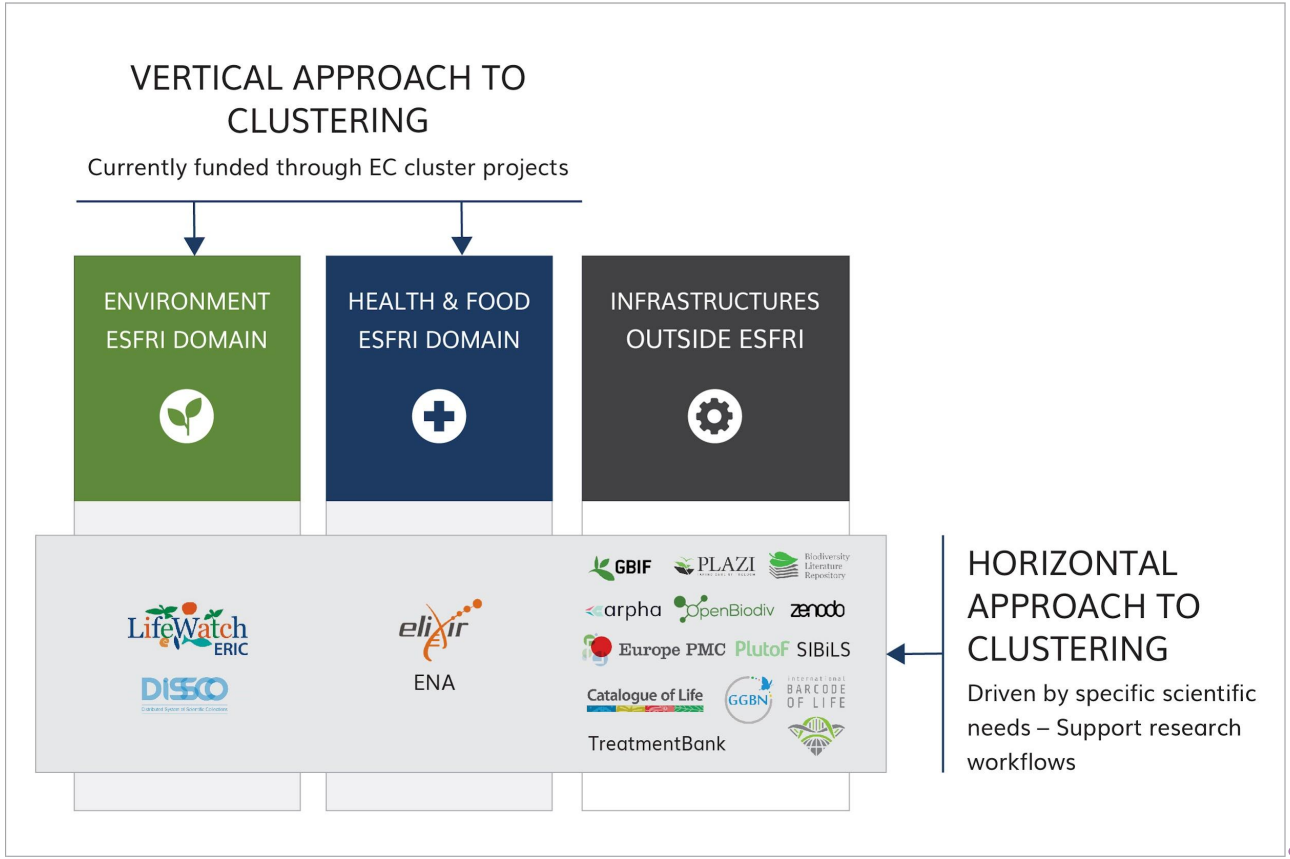
GBIF

ENA

MBG

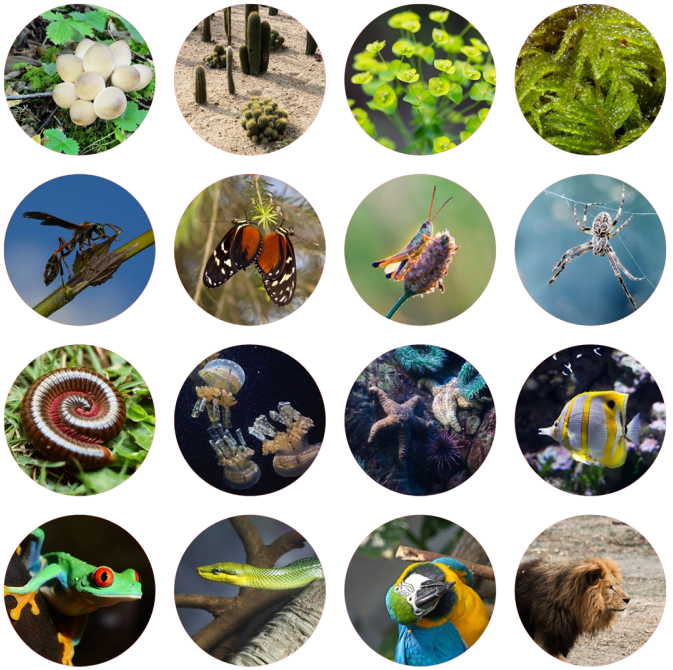
FUB-BGBM

Horizontal linkages between research infrastructures



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 s*ensu 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”

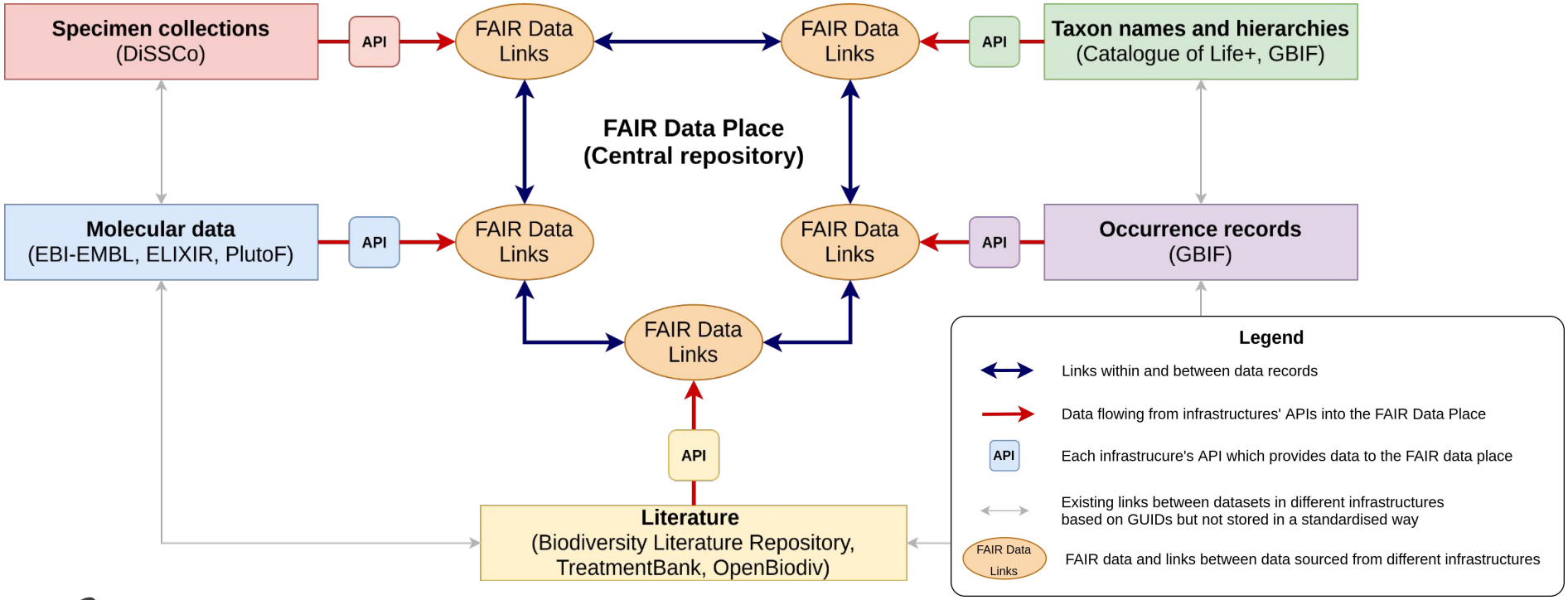
Rather, the reasonable statement would be: This is Species X, according to Treatment X, Specimen(s) XYZ and Sequence X, with a direct access to the data.

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?



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