DiSSCo

Distributed System of Scientific Collections

Building a world-class Research Infrastructure on RDA outputs

Dimitris Koureas
DiSSCo Coordinator

Director of International Biodiversity Infrastructures
Naturalis, Netherlands

Member, RDA TAB
@DimitrisKoureas

RESEARCH DATA ALLIANCE

Natural Science Collections support discovery and modelling of all life on earth

Europe: the global leader

55% of the world's assets with rich historical and global distribution

European Collections:

1.5 billion specimens

80% of world's species

5,000 Full Time Scientists

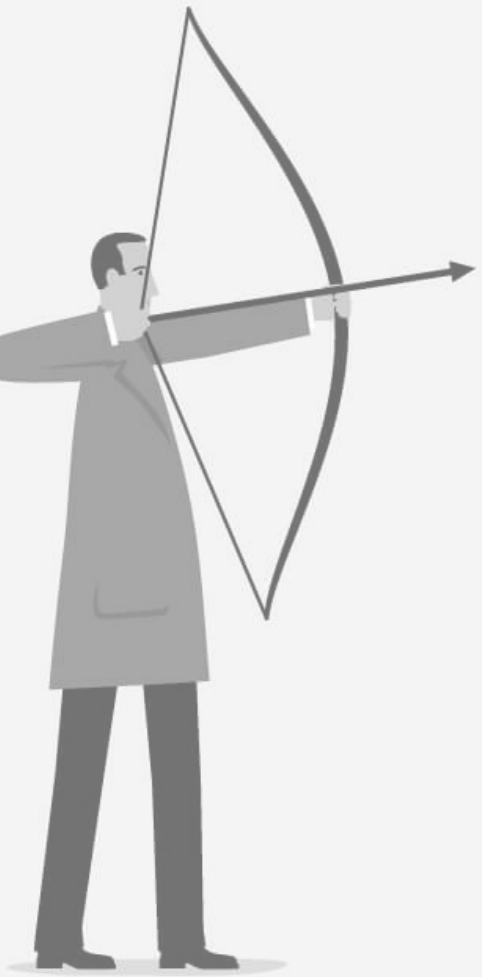25,000 scientific visitors pa

10 million public visitors pa
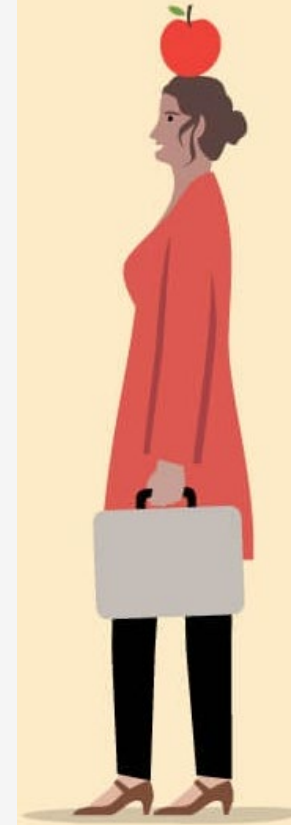
25 million web visitors pa

Trust lost when datasets
disconnect from:

context in which they were created,
or
communities who created them.

What builds

TRUST

in data?

Relevance

Provenance

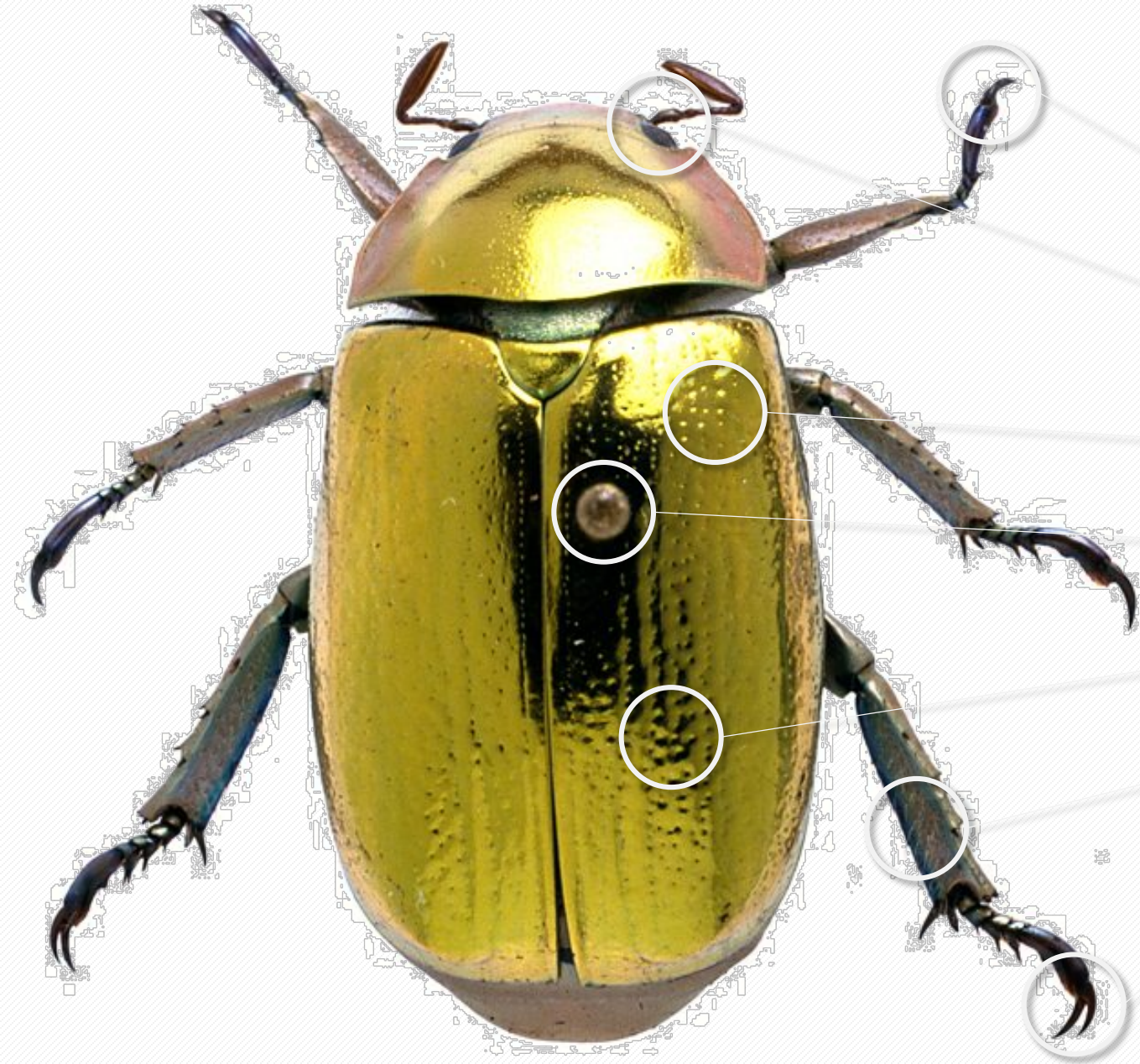*Attribution*

Completeness

Fitness-for-purpose

Agility

Branding (Datatyping)

What's in a Museum specimen?

Genomic data

Biochemical data

Morphological data

Geographical data

Taxonomic Information

Species Interactions data

Ecological data

Jointly this represents all evidence for Biodiversity

GBIF
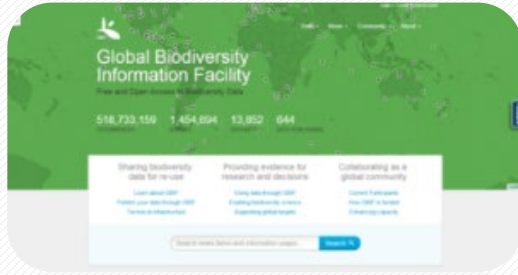
Collections-related
Data classes

Catalogue of LIfe

Re-unite and Serve

Genbank

GloBI

EoL - TraitBank

Plazi –
TreatmentBank

IPNI / Zoobank

Occurrence / Images

Taxonomy

Genomic information

Traits

Nomenclature

Literature - Treatments

Species Interactions
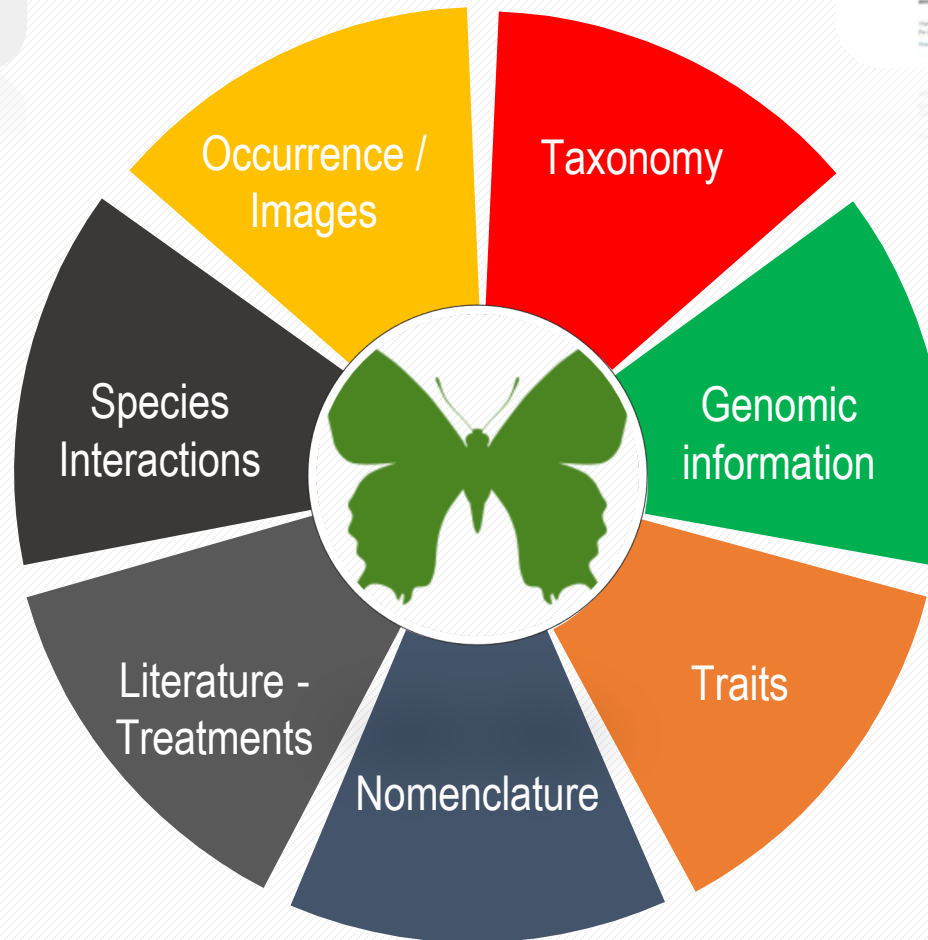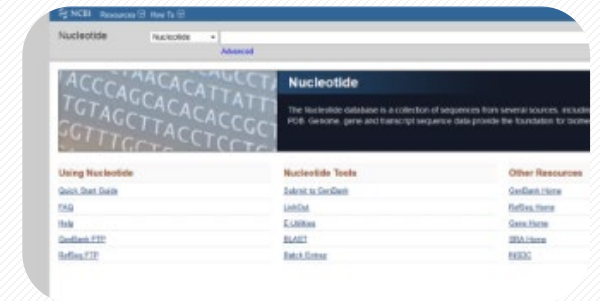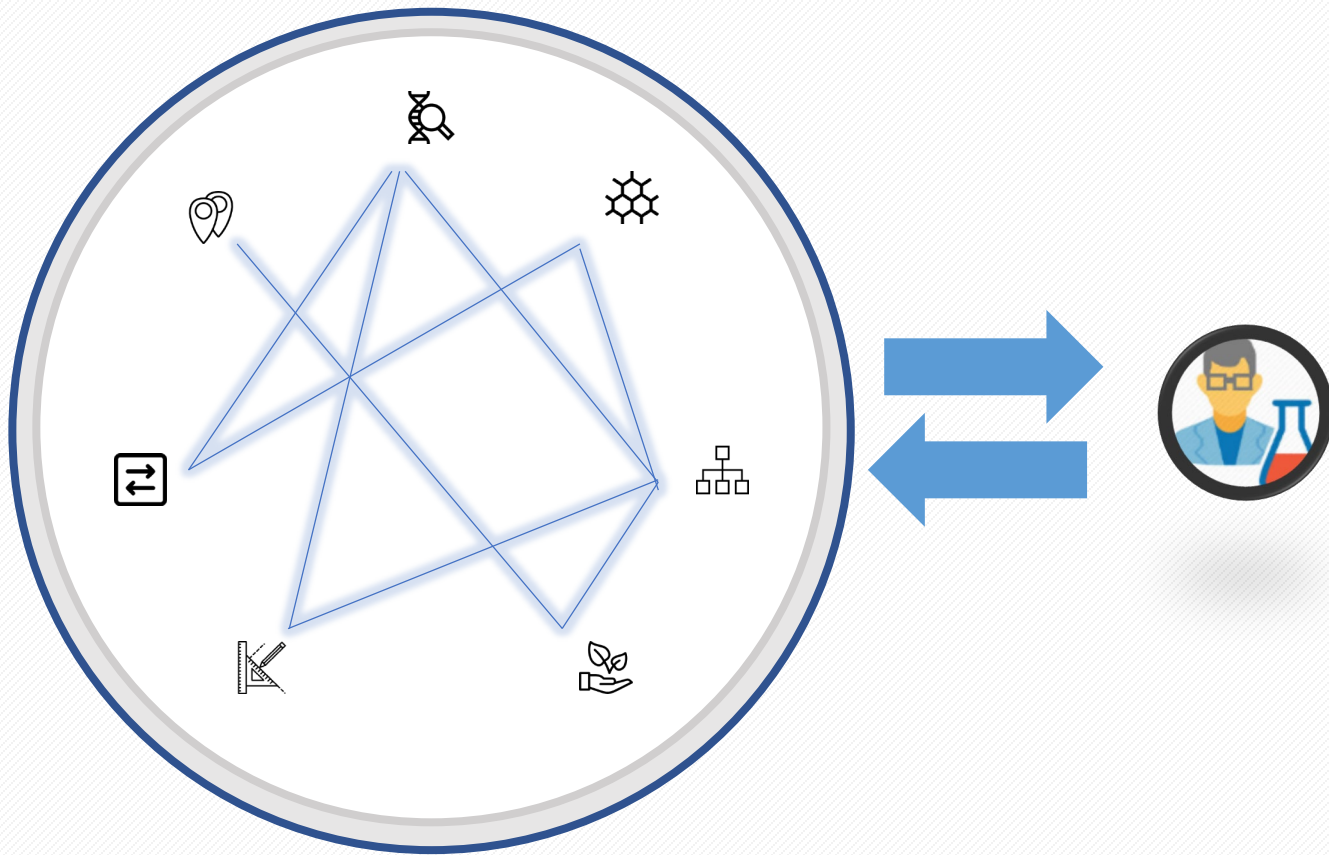
Unified evidence base, which enables semantic crosswalks between all data-types

DiSSCo
Distributed System of Scientific Collections

ESFRI

117 National Facilities
21 Countries

- **Largest ever** formal agreement between natural science collection facilities

- **Centralised shared governance** model already in place

- **Synchronisation** of facilities at access, data and policy level

a new business model:
ONE EUROPEAN COLLECTION

- **One European Collection** of scientific assets

- **Common Collections development** strategy

- **Economies of scope and scale**

- **Monitoring impact** of collections (documenting ROI)

- **Specialisation strategies** (e.g. in alignment with national priorities, e.g. Smart Specialisation Strategies)

- **Joint Research Agendas**

Current model
Slow
Expensive
Inefficient
limited
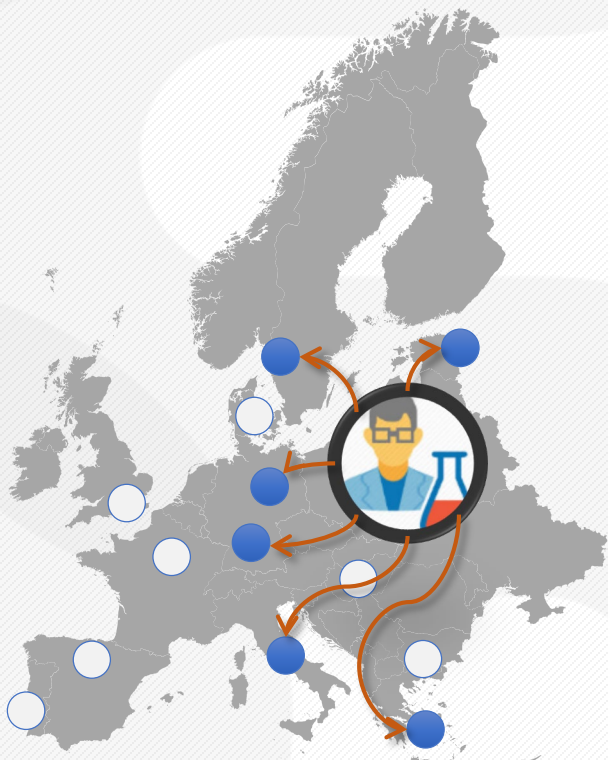
Integrated RI model
Wide access
Lower costs
Faster
New insights
Optimised
FAIR data

User services

DiSSCo
Distributed System of Scientific Collections

ESFRI RIs
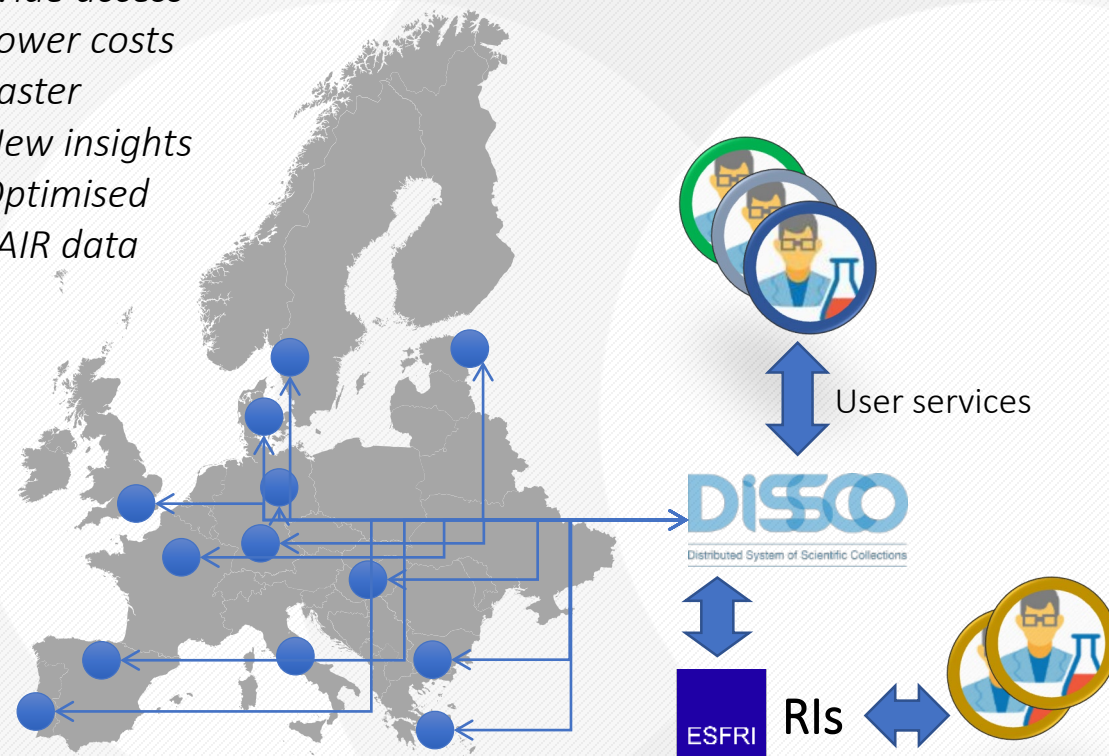
The first mass scale initiative to re-unite and serve genomic, chemical, geographical, morphological and taxonomic information and link it to collections objects

Case study – Invasive Alien Species

UN Sustainable Development Goals (Target 15.8)

EXAMPLE: Alligator Weed
(Alternanthera philoxeroides)
Negative impact on native species, ecosystem services and infrastructure

Economic costs of IAS for EU
€20 Billion / year
Kettunen et al. 2009

Urgent challenge

Institutional **collections**

Species **distribution & genomics**

Modelling / Prevention / Early detection

Facilities & information

Linked Data

Analysis / Interpretation Services

- *Climate data*
- *Ecological monitoring data*
- *Genomic information*

Other Research Infrastructures

15 LIFE ON LAND

DiSSCo — Distributed System of Scientific Collections
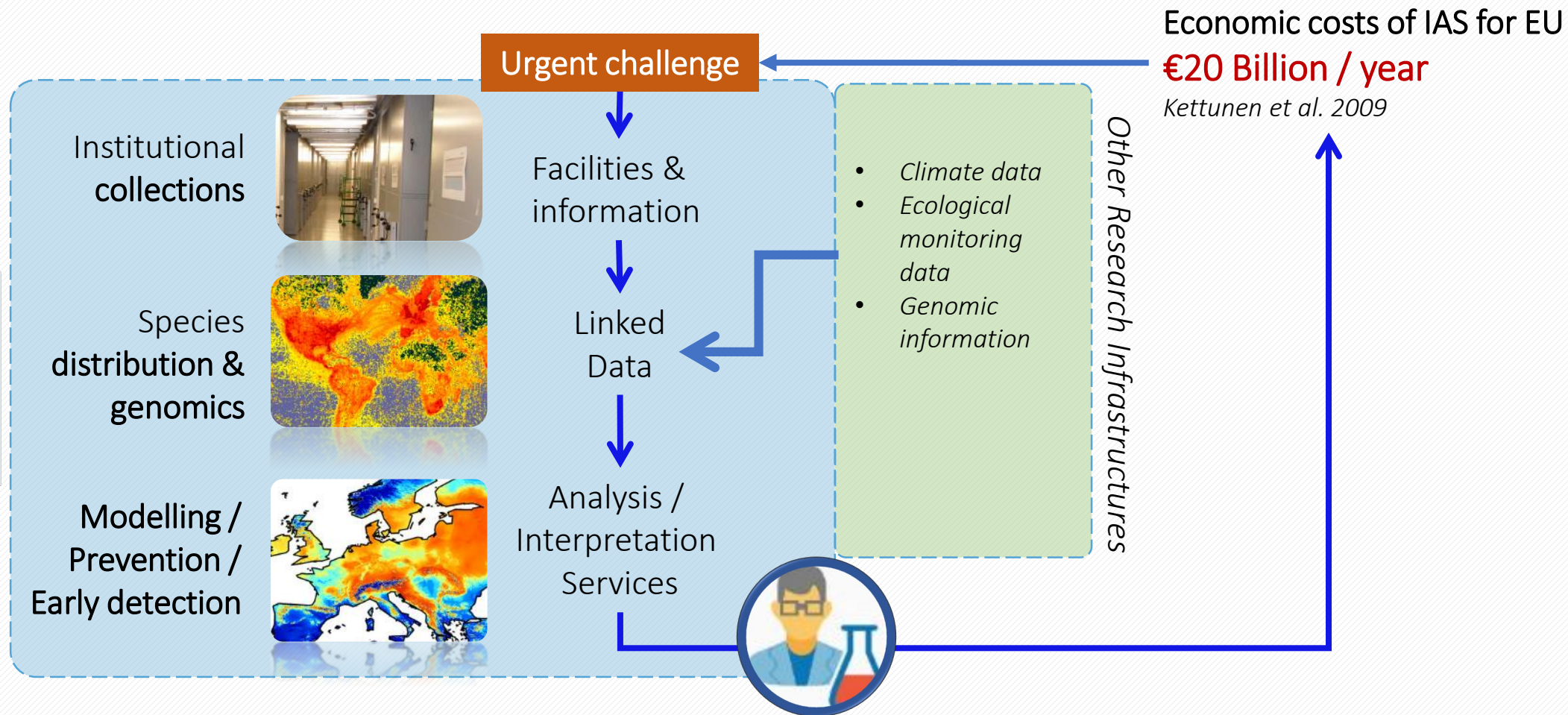
All data classes **unambiguously linked** to the **physical objects** they derive from
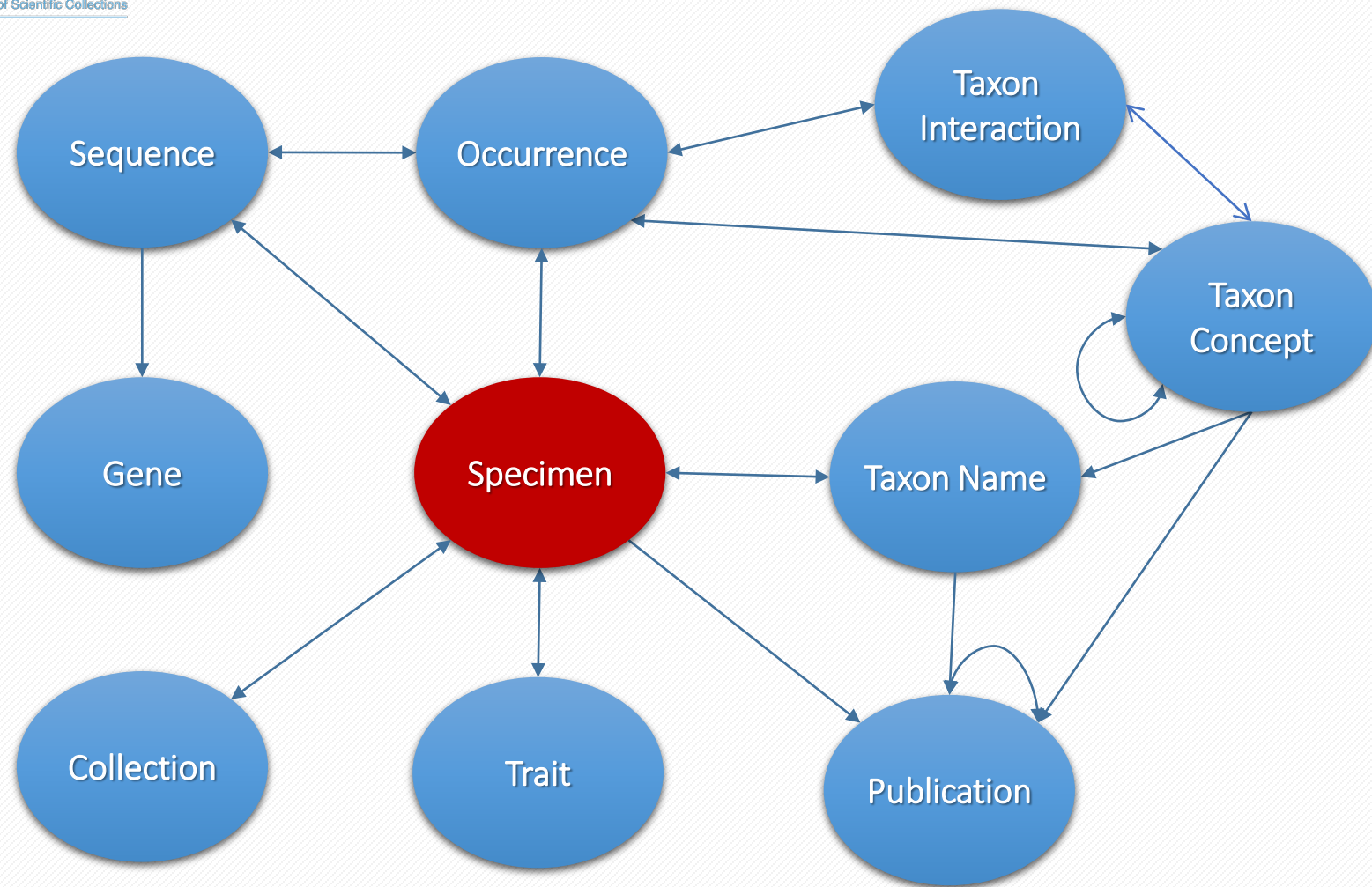
Specimens representations become the centrepiece of the DiSSCo knowledge base – They are used as anchoring points for disperse data classes

Structure of a Digital Specimen Object (DSO)

PID

Kernel information
(metadata about specimen DO)

DO, = An envelope

| PID$_1$ | PID$_2$ | PID$_{...}$ | PID$_N$ |

metadata$_2$
… … … … … … …
… … … … … … …
metadata$_N$
… … … … … … …

content$_N$
… … … … … … …
… … … … … … …
… … … … … … …
… … … … … … …
… … … … … … …

Bit sequence

Data Fabric & Terminology IG

RDA
RESEARCH DATA ALLIANCE

Global Digital Object Cloud

Abstract global data architecture that is based on resolvable global persistent identifiers and globally accessible metadata
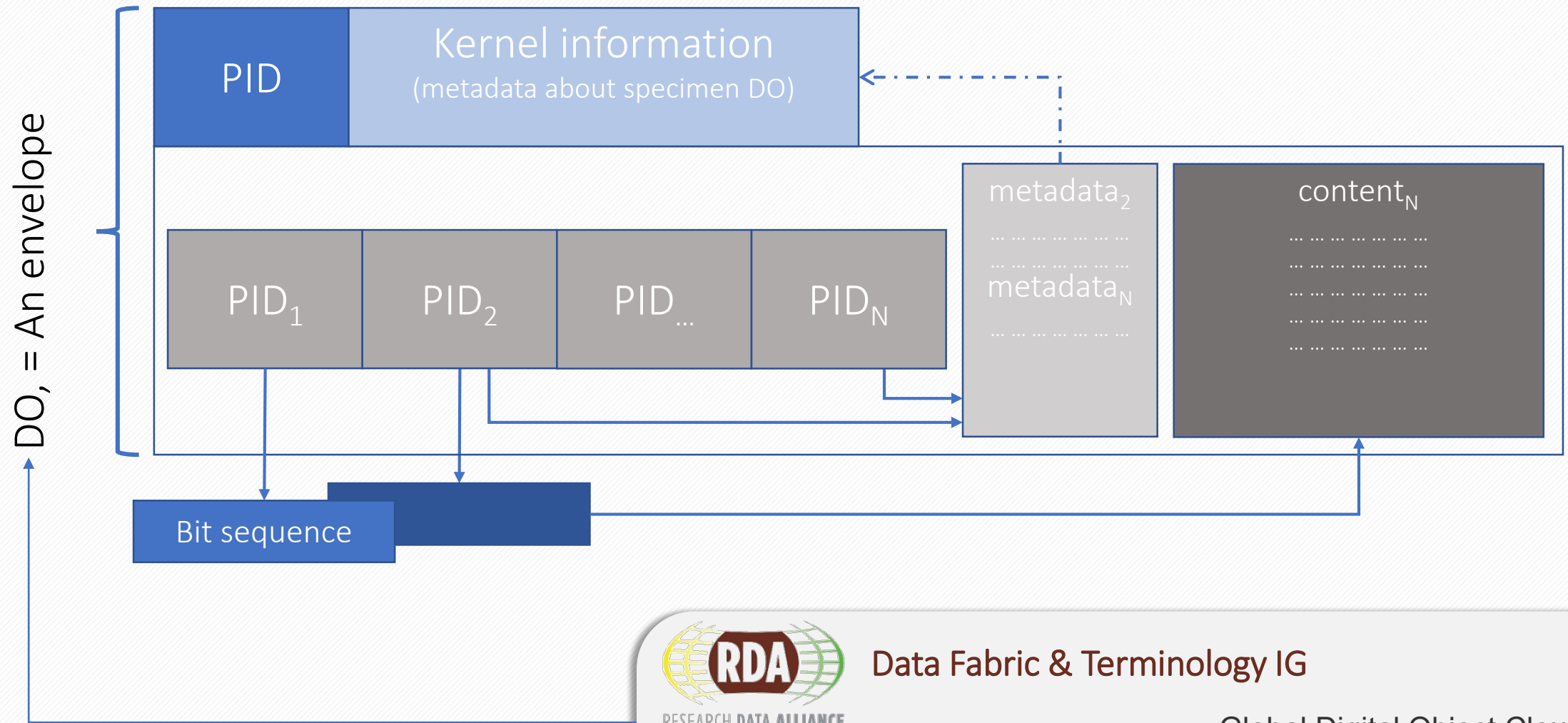
Digital Specimen: A dynamic "box" collecting links to all core information about a thing in one place

DiSSCo
Distributed System of Scientific Collections

Occurrence ID

GenBank
Accession No

Taxon PID

Images (2D, 3D)

GET Image PIDs

GET Image metadata

GET Physical Object (PO) PID

GET PO PID metadata

PO PID
Image PIDs

Digital Specimen object
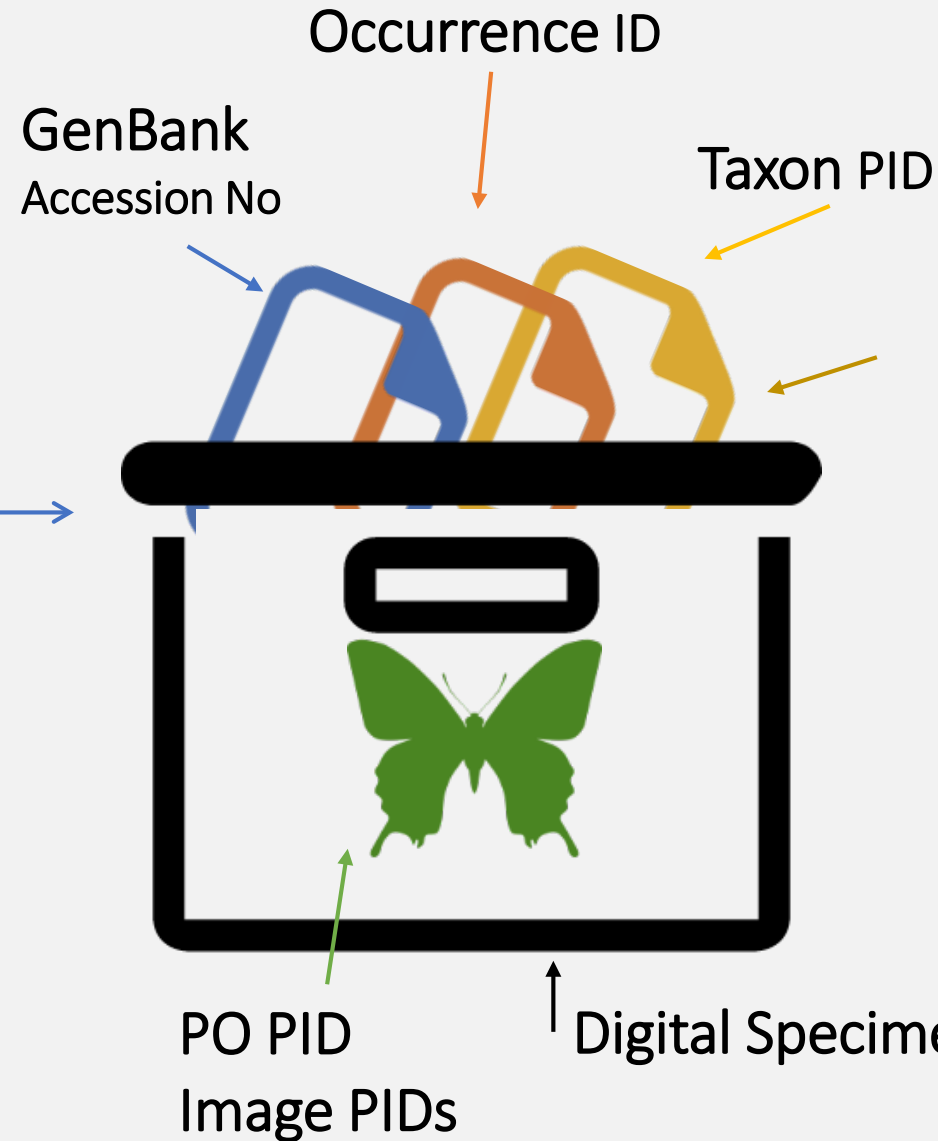
Physical Object → Digital Surrogate

- Genomic data
- Biochemical data
- Morphological data
- Geographical data
- Taxonomic Information
- Species Interactions data
- Ecological data

An actionable **knowledge unit**

How do we represent a Digital Specimen and what do we include in the PID metadata?

PID kernel metadata

? Type metadata

? Resource metadata

Digital Specimen

has — Taxon Name

has — Images
- has — Image Information
  - availableVia — iiif.io Image API
- isa — Lo-res
- isa — Hi-res

has — Publications
- isa — Taxonomic Treatment
  - isa — Journal Article

has — Occurrences
- isa — GBIF Record

has — Annotations

has — DNA Sequences

has — Label Data

has — Traits

← Specimen

← Resource type

← Resource

PID Kernel Information WG

PID Kernel Information Recommendation

Principles for structuring PID kernel information profiles

RDA RESEARCH DATA ALLIANCE

Transforming digital entities to meaningful entities and serving them

User

Web Services (specifications)

Data Layer

Digital Object

Data

Metadata (FAIR)

Context

Supplier

Research Data Collections WG

API Requirements Document
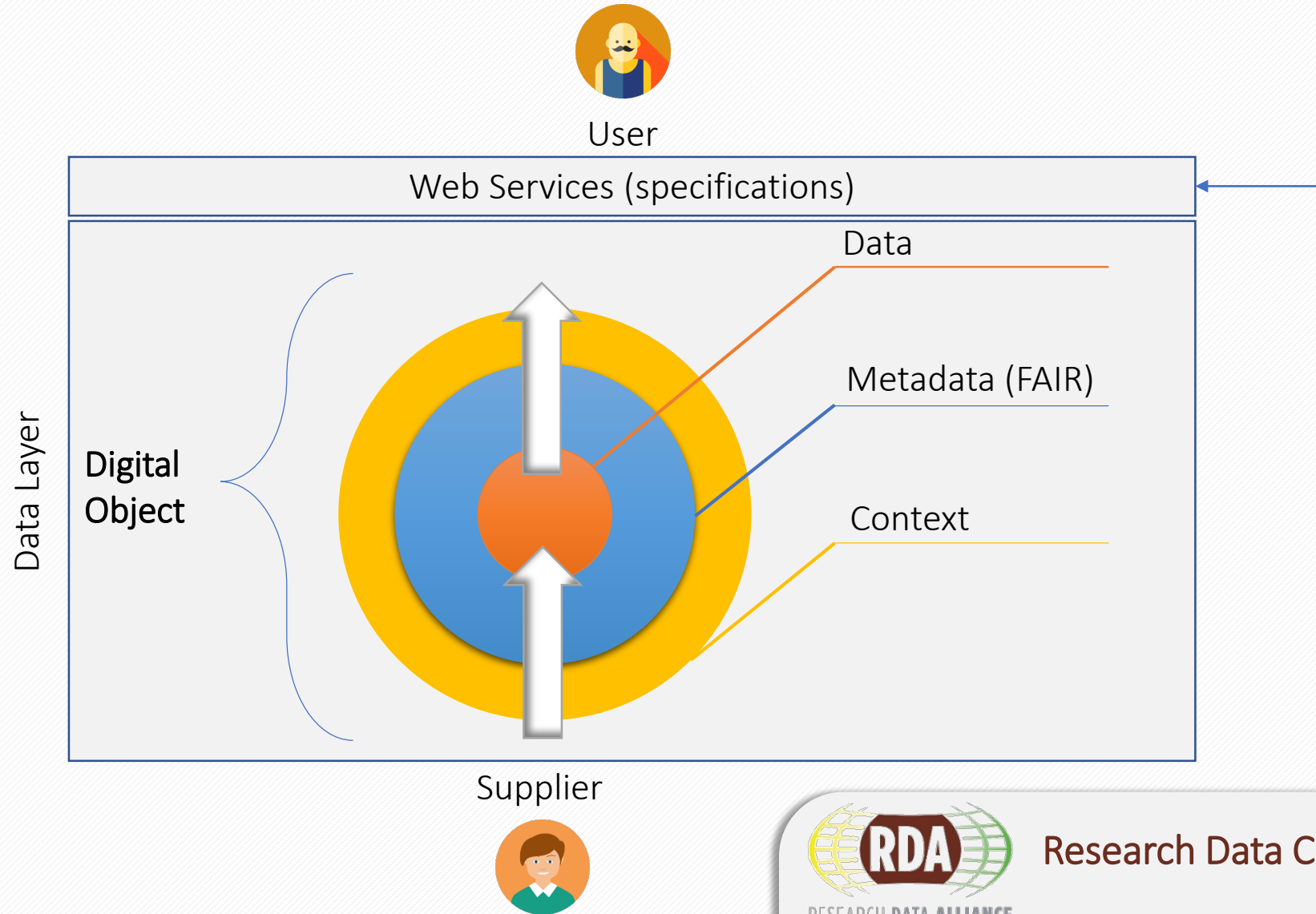
DiSSCo service portfolio by 2025

single entry point

1 e-Science services

A one-stop shop for services providing unified discovery, access, interpretation and analysis of complex linked data

2 Physical and remote access services

A universal harmonised physical access service and digitisation on demand service

3 Support & Training services

Integrated user support desk and implementation of multi-modal training programmes to enhance skills & competencies

Simple DiSSCo implementation timeline

| Proposal Phase 2015 - 2018 | Preparatory Phase 2018 - 2021 | Construction Phase 2020 - 2023 | Transition Phase 2023-2024 | Operation 2025 – 2035 |

Governments political and financial commitments

Optimum Implementation Readiness Level (IRL)

Infrastructure in place

Legal entity in place
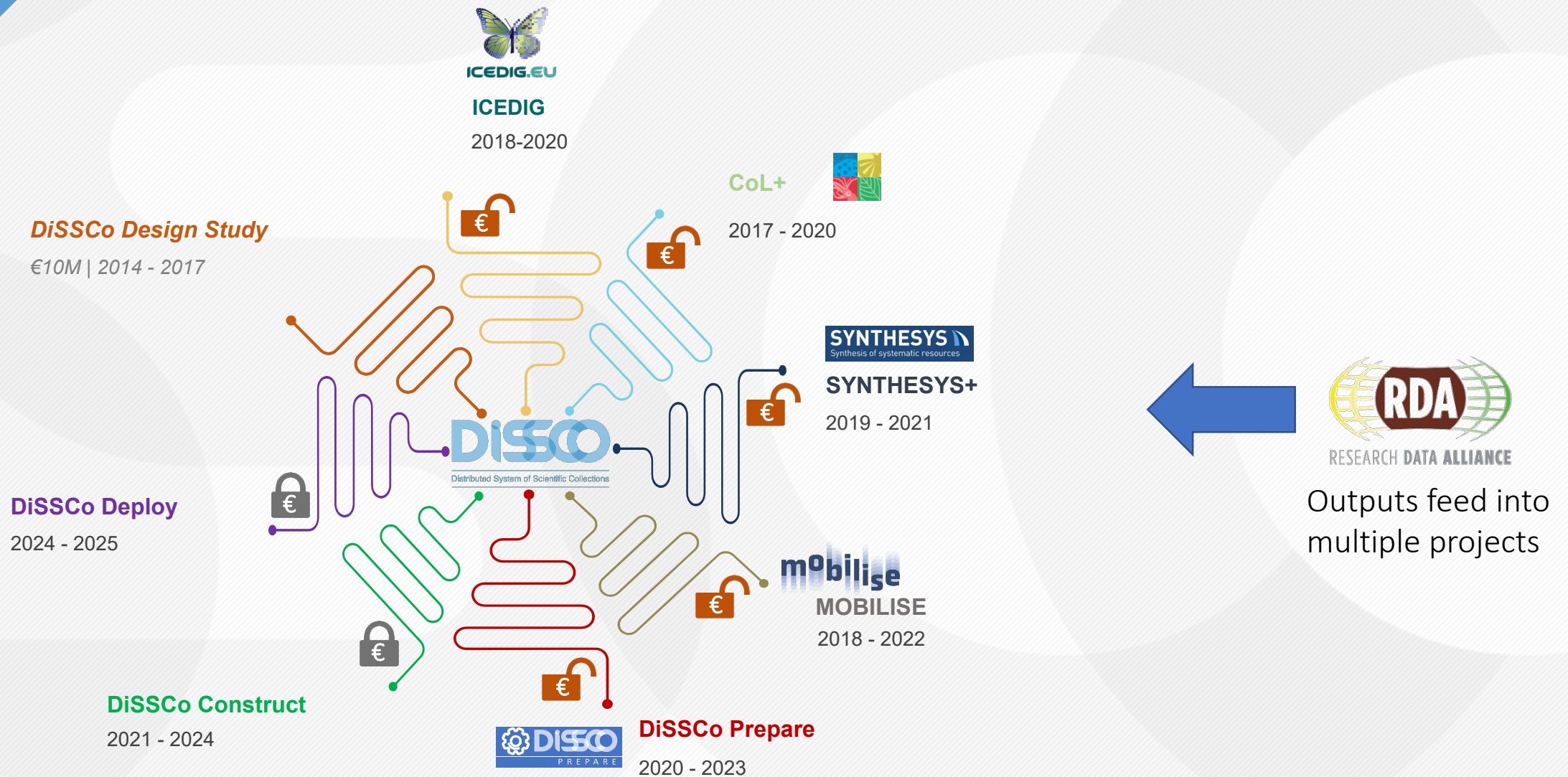
Full Service portfolio

Preparatory Phase 2018 - 2021
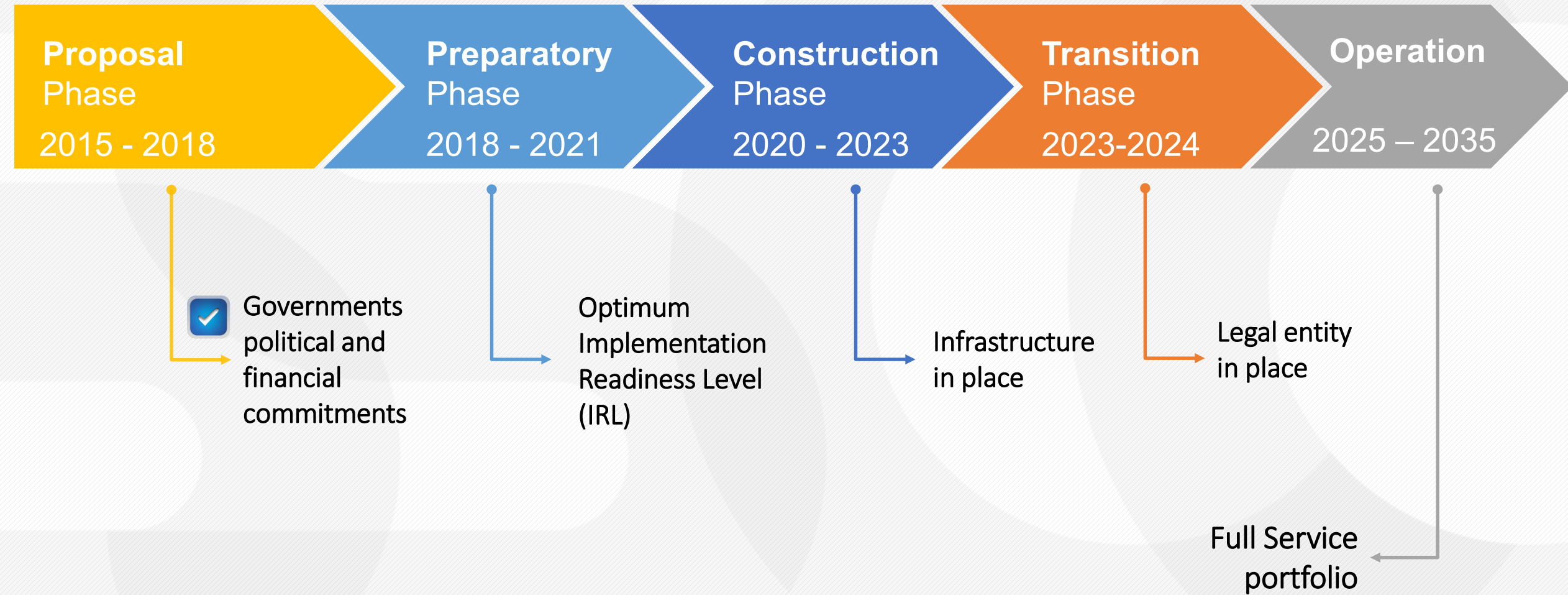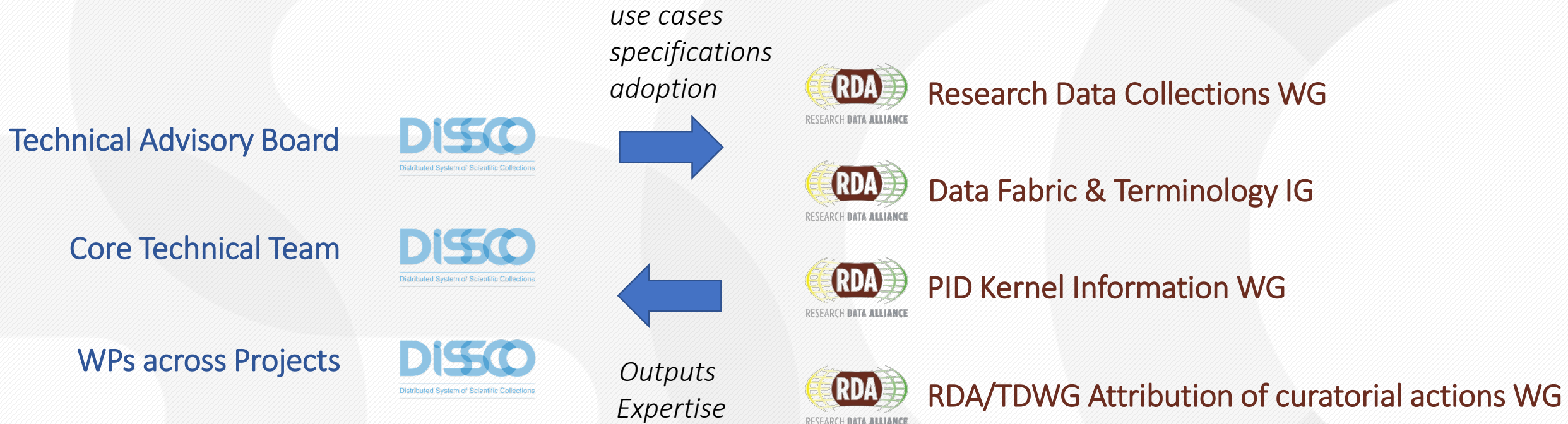
Implementation Readiness Levels (IRL)
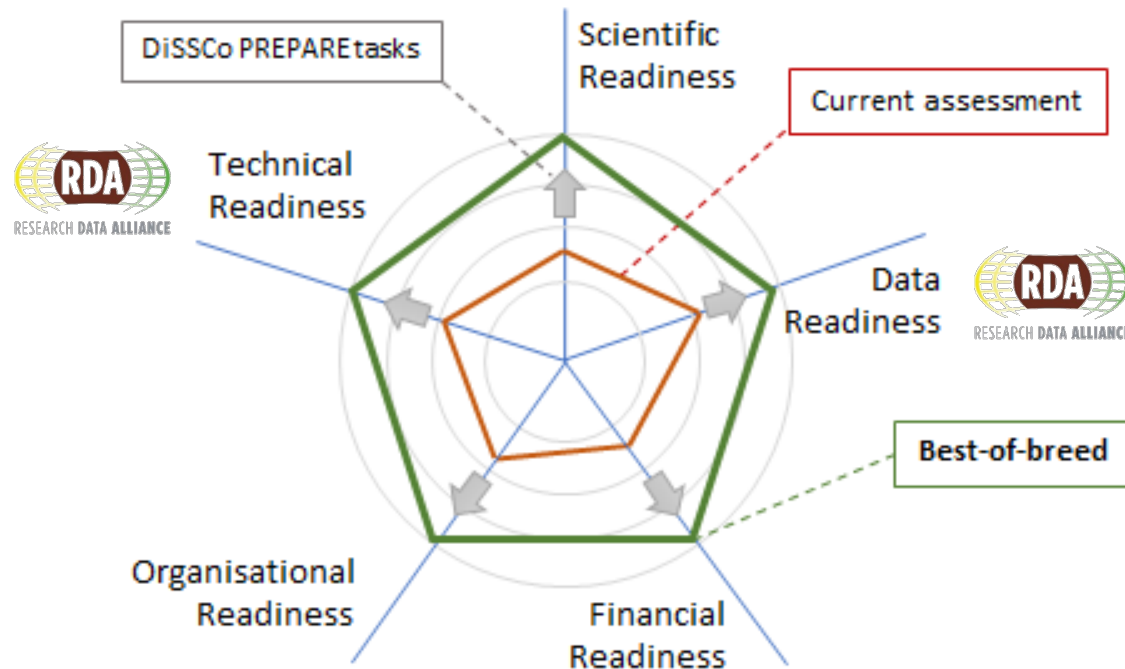
A new Data Research Infrastructure provides opportunities for a new approach to our community challenges

Preparatory Phase

Preparing through a complex programme

ICEDIG
2018-2020

CoL+
2017 - 2020

DiSSCo Design Study
€10M | 2014 - 2017

SYNTHESYS+
2019 - 2021

DiSSCo Deploy
2024 - 2025

MOBILISE
2018 - 2022

DiSSCo Construct
2021 - 2024

DiSSCo Prepare
2020 - 2023

Outputs feed into multiple projects

Interfaces between DiSSCo and RDA

Technical Advisory Board

Core Technical Team

WPs across Projects

use cases
specifications
adoption

Outputs
Expertise

Research Data Collections WG

Data Fabric & Terminology IG

PID Kernel Information WG

RDA/TDWG Attribution of curatorial actions WG

But the most valuable interaction comes through sharing of People

One world – One collection

Find out more at www.dissco.eu