BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25474
doi: 10.3897/biss.2.25474

OPEN ACCESS

Conference Abstract

Digital Object Cloud for linking natural science collections information; The case of DiSSCo

Dimitrios Koureas[‡,§], Wouter Addink[‡], Alex Hardisty[|]

‡ Naturalis Biodiversity Center, Leiden, Netherlands
§ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
| School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

Corresponding author: Dimitrios Koureas (dimitris.koureas@naturalis.nl)

Received: 03 Apr 2018 | Published: 18 May 2018

Citation: Koureas D, Addink W, Hardisty A (2018) Digital Object Cloud for linking natural science collections information; The case of DiSSCo . Biodiversity Information Science and Standards 2: e25474. https://doi.org/10.3897/biss.2.25474

Abstract

DiSSCo (The Distributed System of Scientific Collections) is a **Research Infrastructure (RI)** aiming at providing unified physical (transnational), remote (loans) and virtual (digital) access to the approximately 1.5 billion biological and geological specimens in collections across Europe. DiSSCo represents the largest ever formal agreement between natural science museums (114 organisations across 21 European countries). With political and financial support across 14 European governments and a robust governance model DiSSCo will deliver, by 2025, a series of innovative end-user discovery, access, interpretation and analysis services for natural science collections data.

As part of DiSSCo's developing data model, we evaluate the application of **Digital Objects (DOs)**, which can act as the centrepiece of its architecture. DOs have bit-sequences representing some content, are identified by globally unique persistent identifiers (PIDs) and are associated with different types of metadata. The PIDs can be used to refer to different types of information such as locations, checksums, types and other metadata to enable immediate operations. In the world of natural science collections, currently fragmented data classes (*inter alia* genes, traits, occurrences) that have derived from the study of physical specimens, can be re-united as parts in a virtual container (i.e., as components of a Digital Object). These typed DOs, when combined with software agents

that scan the data offered by repositories, can act as complete digital surrogates of the physical specimens.

In this paper we:

1. investigate the architectural and technological applicability of DOs for large scale data RIs for bio- and geo-diversity,
2. identify benefits and challenges of a DO approach for the DiSSCo RI and
3. describe key specifications (incl. metadata profiles) for a specimen-based new DO type.

Presenting author

Dimitrios Koureas

(Presentation slides follow).

Digital Object Cloud for linking natural science collections information;
The case of DiSSCo

Dimitris Koureas, Wouter Addink, Alex Hardisty

Director of International Biodiversity Infrastructures, Naturalis Biodiversity Center
Coordinator, Distributed System of Scientific Collections (DiSSCo)
Chair, TDWG

ICEDIG project
Design Refinement Study for DiSSCo

Natural Science Collections

Leading Scientific facilities (research infrastructures)

Represent, through a well documented way, over 400 years of the planet's species diversity

Planetary library of genomic and chemical information & Reference material for all world's species

Unparalleled resource for current and future biodiversity discovery


JLG LIFTLUX 153-12

European Collections

European Collection facilities:

> **1.5 billion** specimens

> **80%** of world's species

> **5,000** scientists employed

> **16,000** scientific visitors pa

> **10 million** public visitors pa

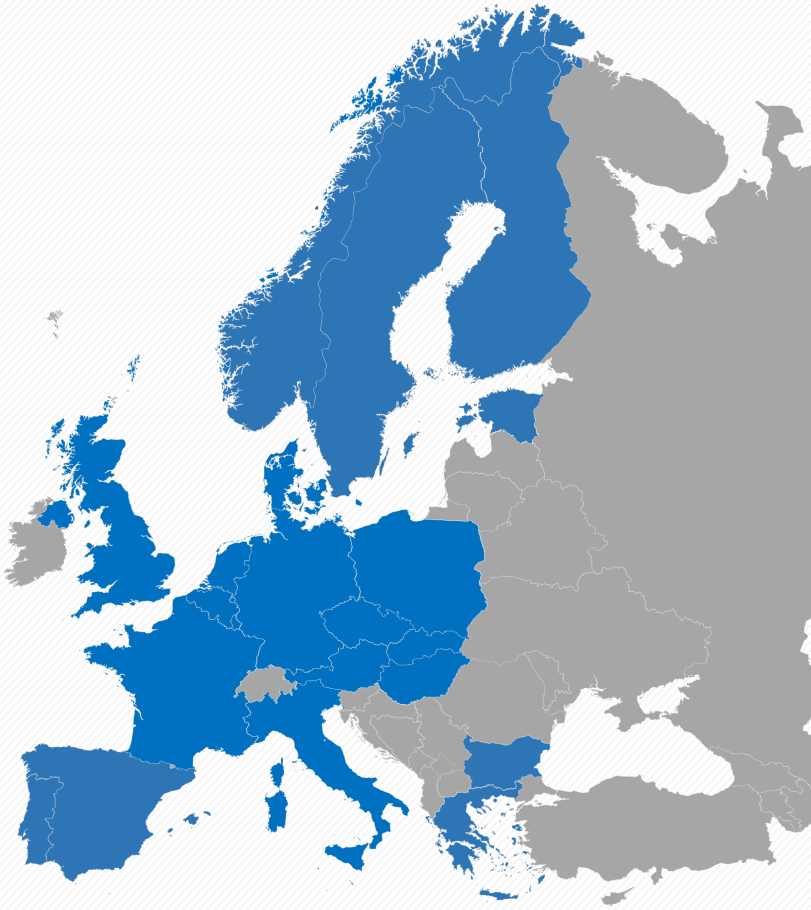
> **25 million** web visitors pa

DiSSCo: A new European infrastructure

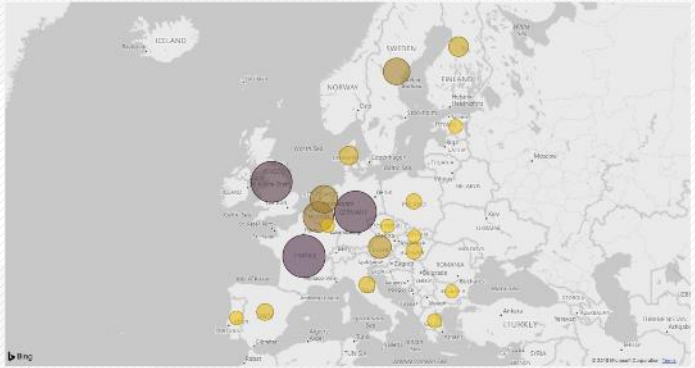
115 National Facilities

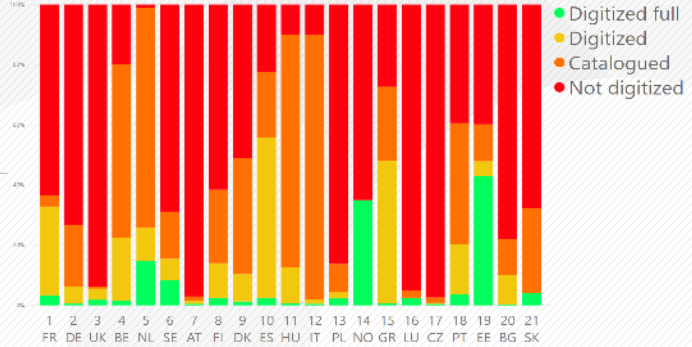
21 Countries

- **Largest ever** formal agreement between natural science collection facilities
- **Centralised governance** model already in place
- **Synchronisation** of facilities at access, data and policy level
- One European virtual Collection



Collection Size

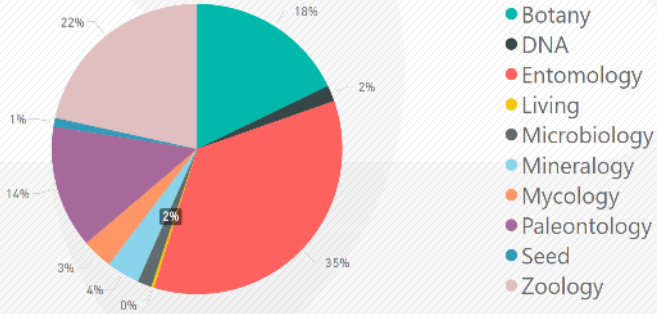




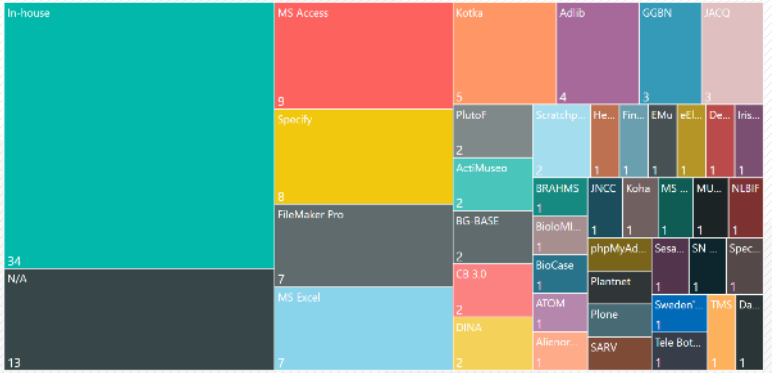
DiSSCo Collections

530 Million

The DiSSCo Collection



Collection Management Systems (CMS)

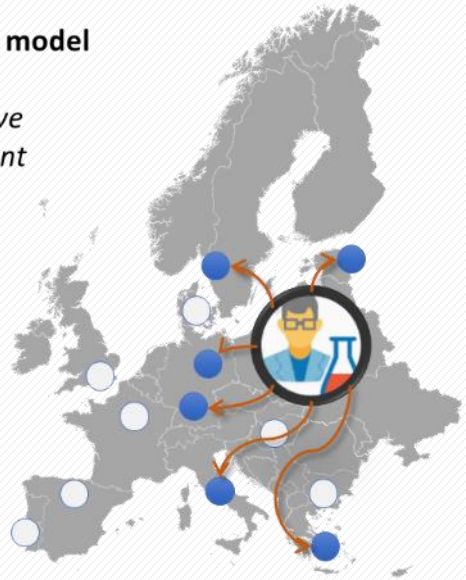


Lowering barriers for users

25,000 researchers *travel every year to physically access scientific collections and* ***800k objects*** *are packed and shipped (at an annual public cost of more than €70M)*

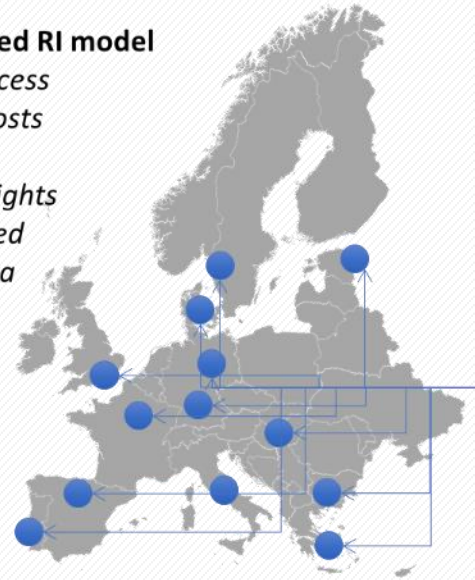
Current model

Slow
Expensive
Inefficient
limited



Integrated RI model

Wide access
Lower costs
Faster
New insights
Optimised
FAIR data



DiSSCo science services

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 data skills

A new business model

One European Collection

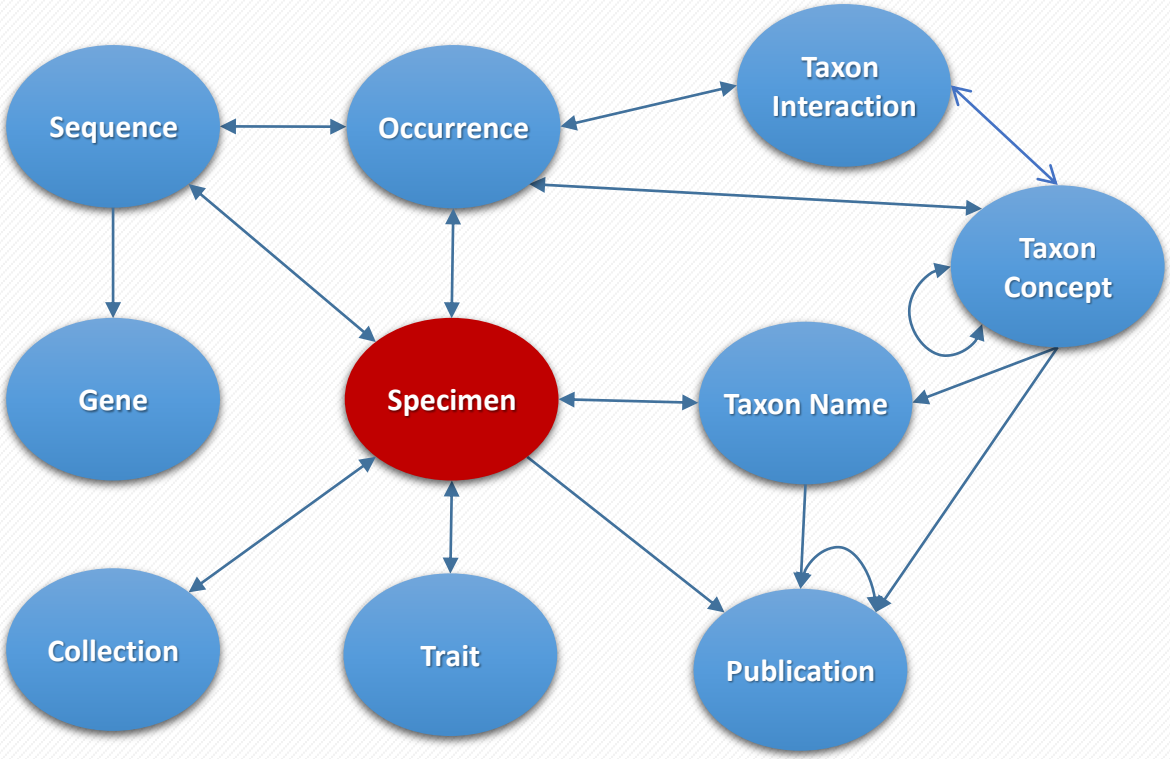
Synchronising 115 facilities:

- One European Collection of scientific assets
- European level strategy
- Economies of scope and scale
- **Monitoring impact of collections / Attribution**
- Specialisation strategies (e.g. in alignment with national priorities, e.g. Smart Specialisation Strategies)
- Joint Research Agendas

DiSSCo

Distributed System of Scientific Collections

All data classes **unambiguously linked** to the **physical objects** they derive from



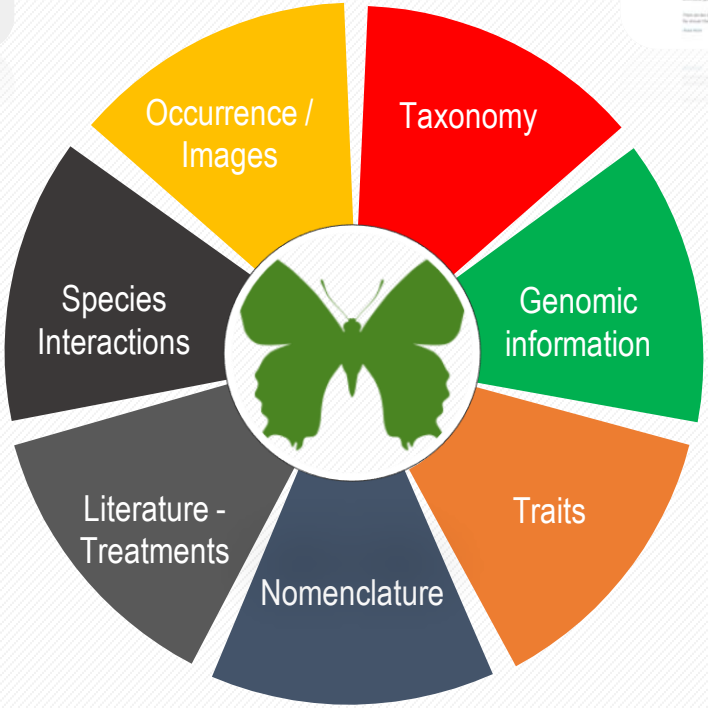
Specimens representations become the centrepiece of the DiSSCo knowledge base – They are used as anchoring points for disperse data classes

Collections-related
Data classes

GBIF

Catalogue of LIfe

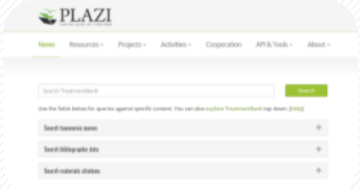
Re-unite and Serve



GloBI

Genbank

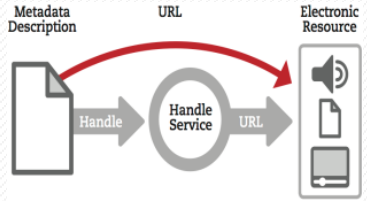
Plazi – TreatmentBank



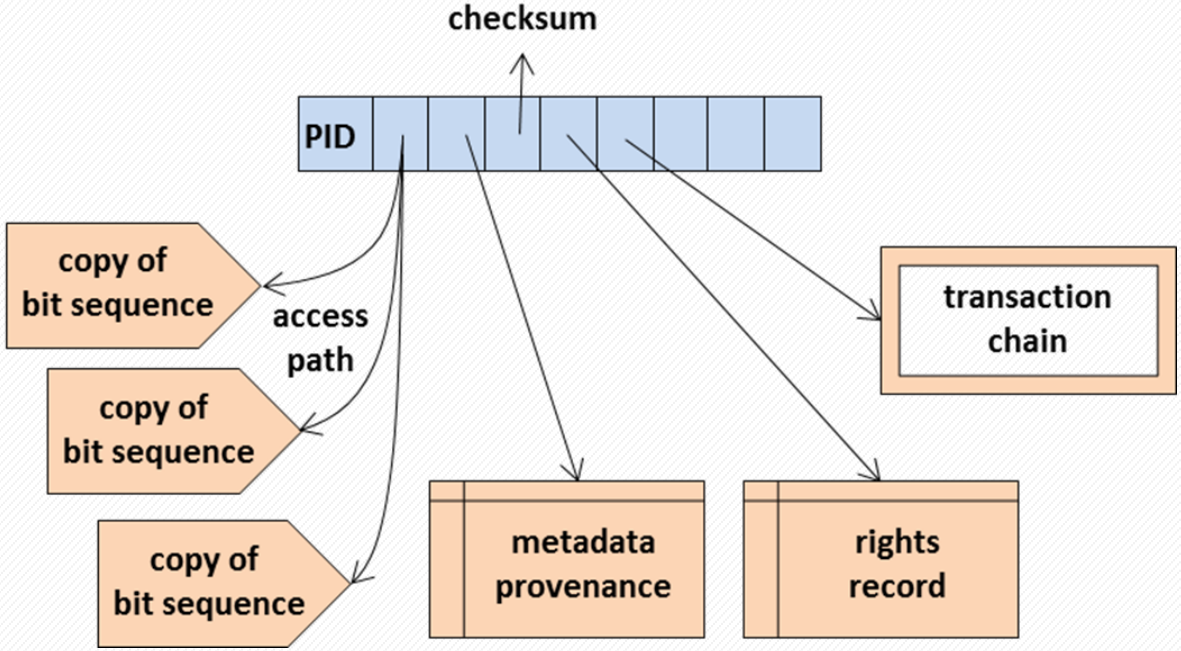
EoL - TraitBank

IPNI / Zoobank

Persistent Identifiers are crucial



- PIDs need to be persistent – we need to make them persistent (!)
- PIDs can help to identify, check authenticity, find copies, etc.
- PID record attributes can lead us to all entities of a DO, i.e. they can take a binding role
- PIDs can open the way to global virtualisation

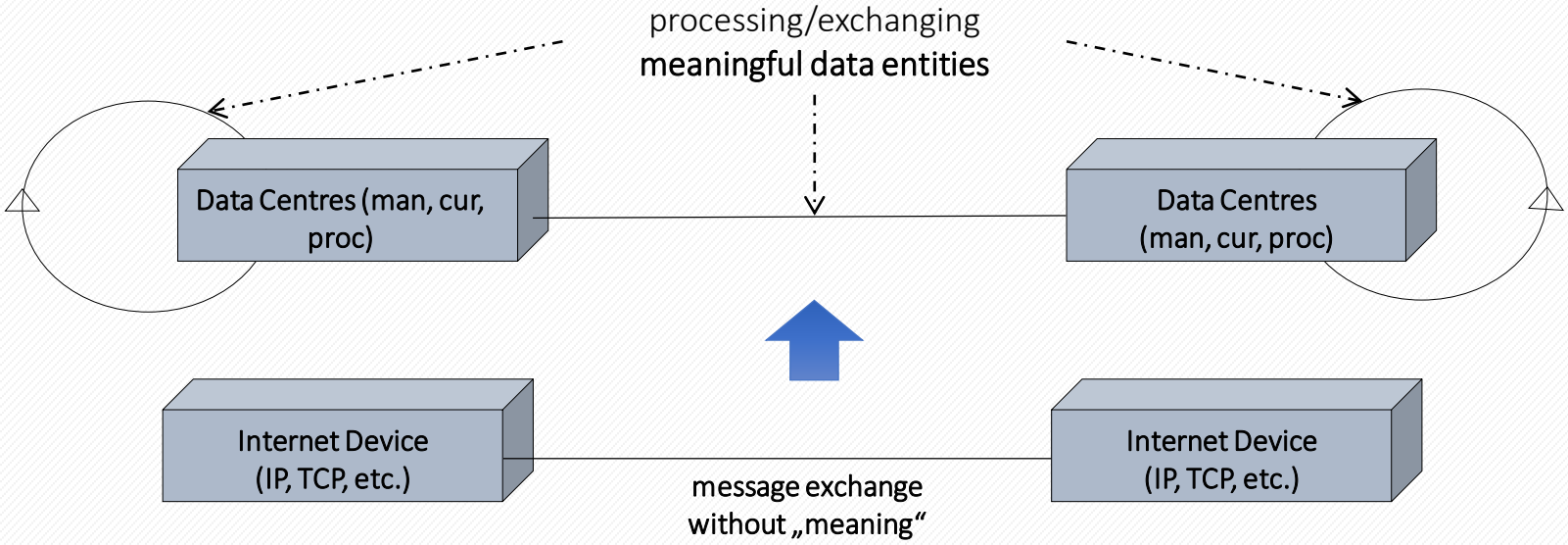


Developed in RDA Data Fabric IG

worked on by RDA Kernel Information WG

Digital Objects – looking back

- 1995 Kahn & Wilensky: DOs have structured bit sequence, persistent ID, key metadata (key metadata = one key-value pair to cover the PID)
- *something* was missing after Internet



Digital Objects (DO)

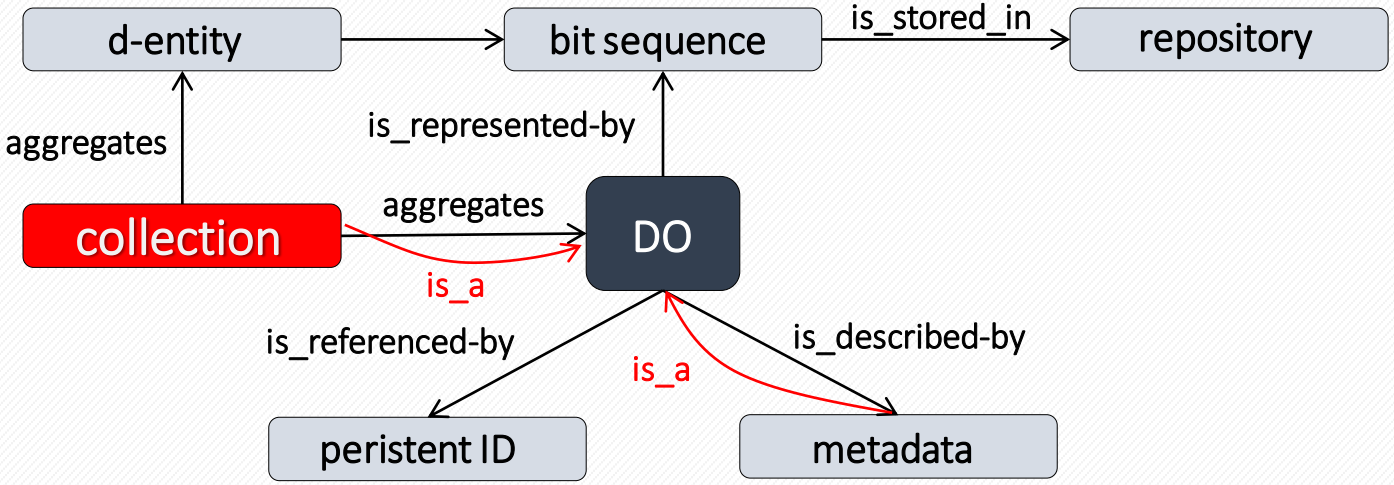
- Digital Objects are "**meaningful entities**" existing in the digital world of bits.
- **meaningful**: ability to talk about it, work with it, refer to it, cite it, etc.
- DOs can include data, collections, metadata, software, publications, workflows, configurations, categories, assertions, etc.

- **DOs have**
  - **Content** represented by (structured) bit sequences (stored somewhere)
  - **Name** (class)
  - **Properties**, which are described by metadata

- **DOs need to be actionable (capabilities list embedded at record level)**

Digital Objects are central for human and machine communication and we need to identify them

RDA Data Foundation & Terminology (2013/2014)



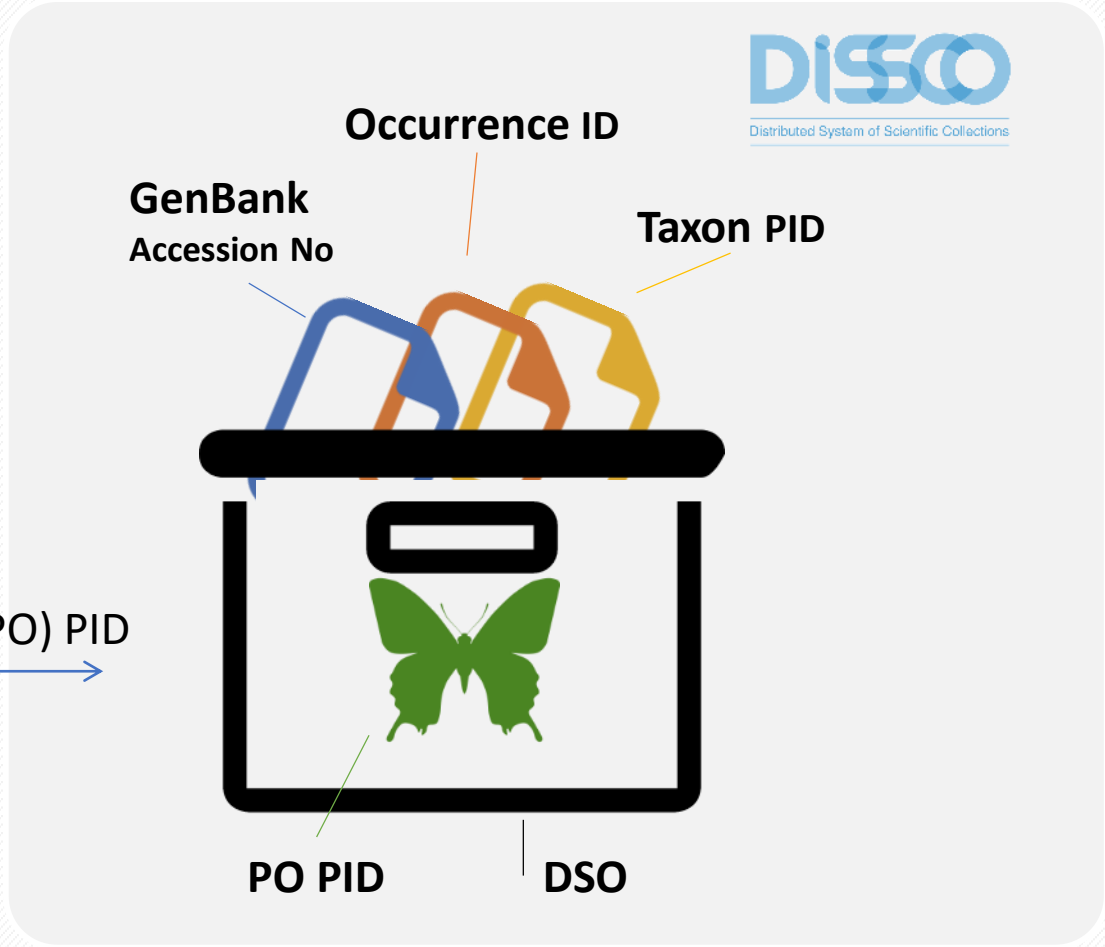
- if software/repository builders would follow this simple model for organising data much efficiency would be gained

- implemented by some communities to manage large collections from 2000 on (DOBES, ENES, etc.)

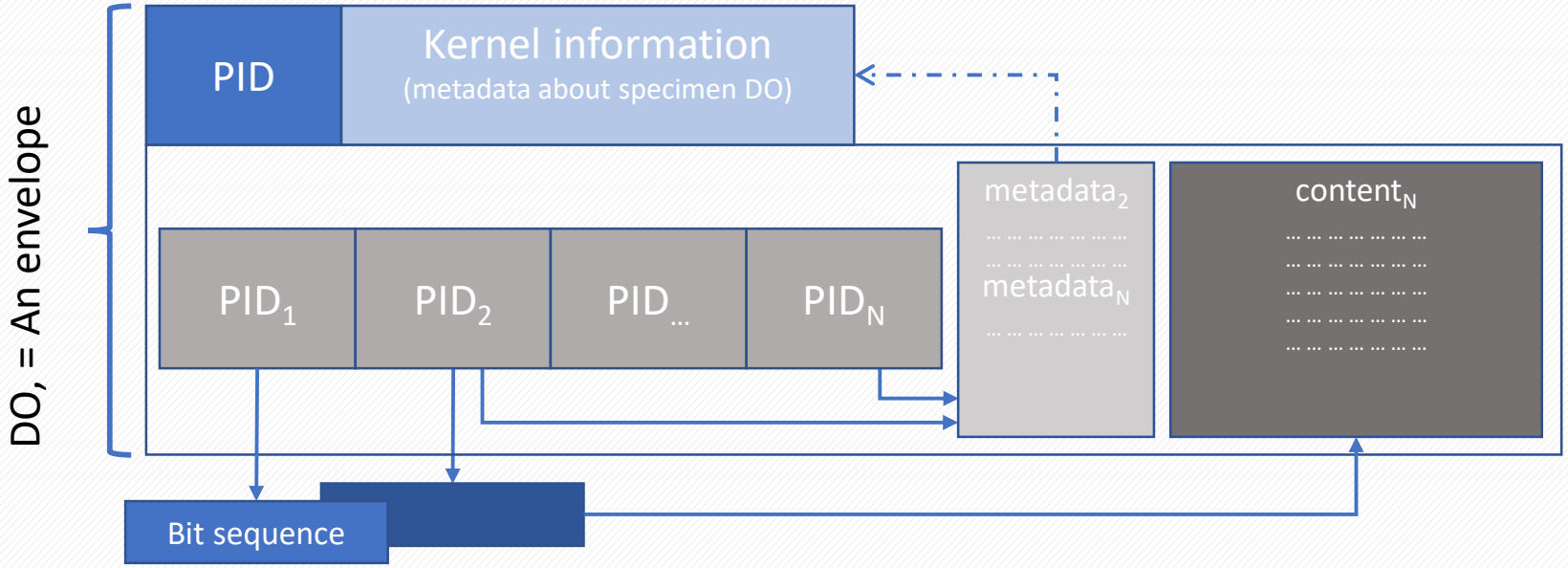
DSO: Digital Specimen Object
PO PID: Physical object PID

GET Physical Object (PO) PID

GET PO PID metadata



Structure of a Digital Specimen Object (DSO)



Why DOs approach is appropriate for re-uniting natural science collections-derived information

1

Specimens are atomic items

- Like journal articles, archaeological artefacts, DNA sequences, YouTube videos, taxon concepts, software programs, workflows, etc.
- **Deserve individual and unique identification to avoid ambiguity around use and interpretation**

2

Digital objects collect all core information about the thing in one place

- What it is, how it came into being, where it can be found, and pointers to other related things
- Editable but accuracy/authenticity can be controlled

3

A new kind of industrial object that pervades every aspect of our life today, a technical essence of a thing in cyberspace

- Virtual collection joined together through logical and temporal relations, networks, etc.

Further reading: 1) Yuk Hui, On the existence of digital objects; 2) Jannis Kallinikos et al., A theory of digital objects

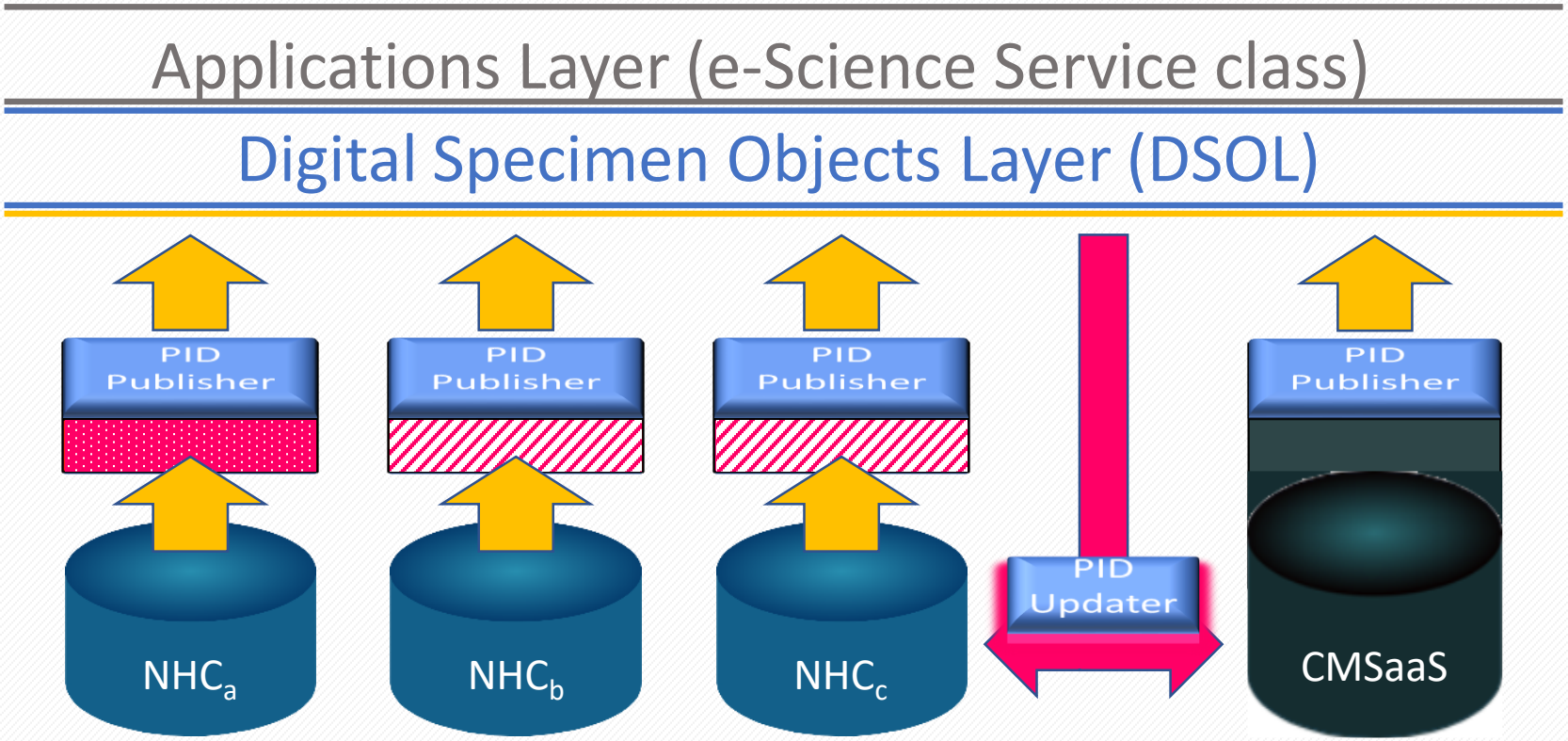
DiSSCo layers

Applications Layer (e-Science Service class)

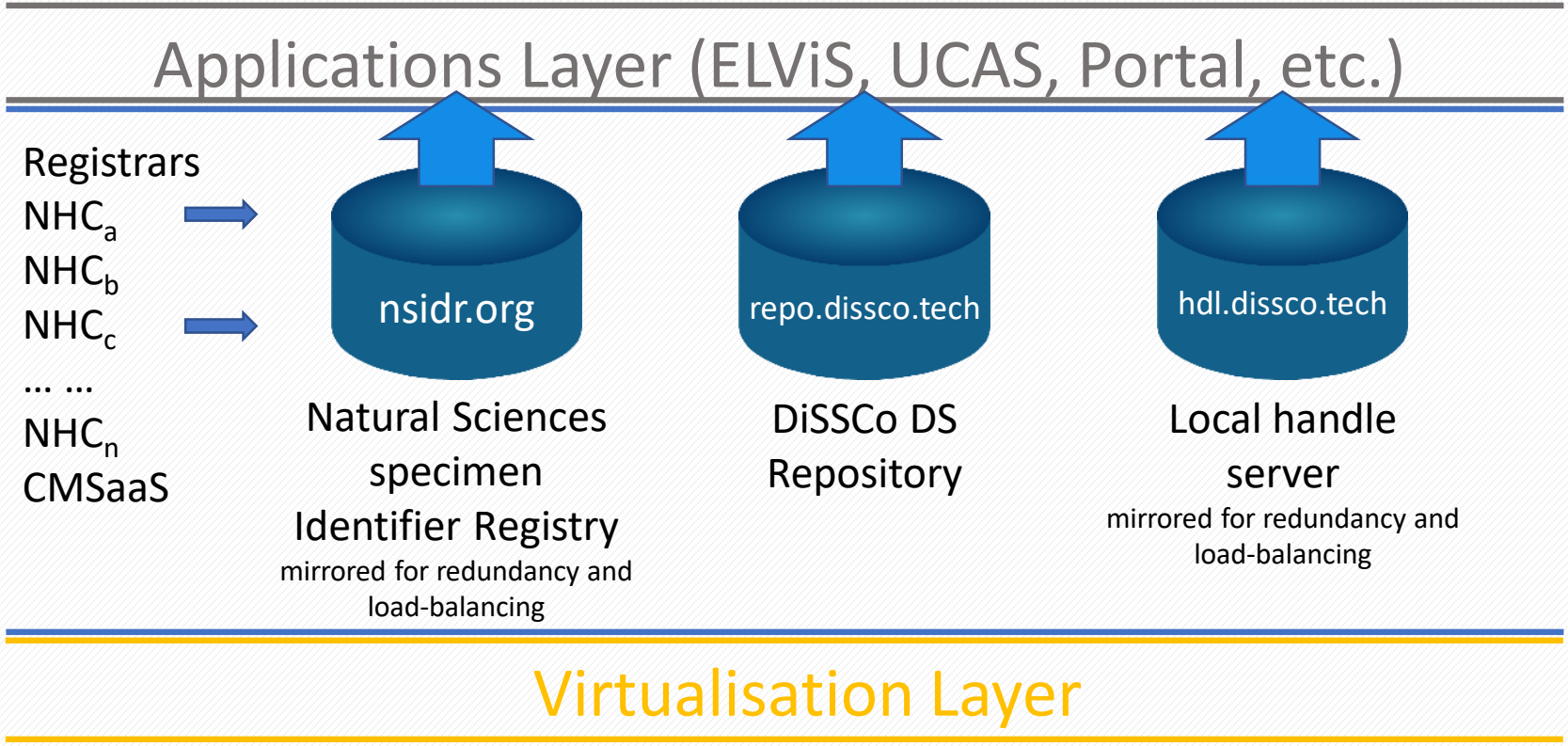
Digital Specimen Objects Layer (DSOL)

Virtualisation Layer

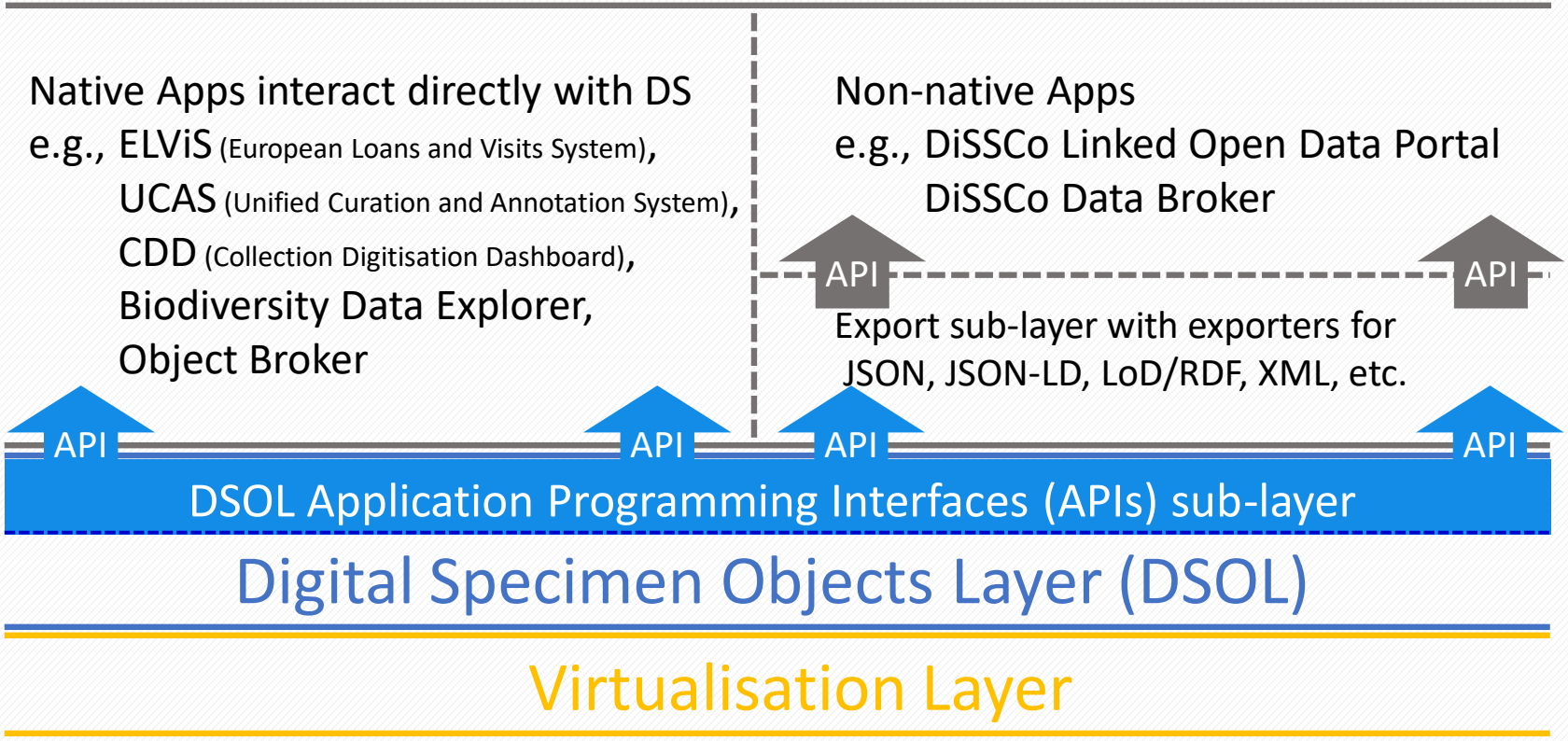
DiSSCo Virtualisation layer



DiSSCo Digital Specimen Objects layer (DSOL)



DiSSCo Applications layer (ELViS, UCAS, Portal, etc.)



Essential components already established & used

- Identifiers and resolution system: Handle System
  - reliable, mature system with organizational backing
- Data Types: registries and concepts as discussed by RDA DTR
  - ready to use
  - small-scale demonstrators exist

Further components: evaluate and adapt

- Digital Object Repositories
  - evolve from current repositories
- Digital Object Interface Protocol (DOIP)
  - specification exists, needs practical evaluation
- Digital Object Registries
  - overarching registries for searching
  - concept needs to be sharpened, relation with repositories
- Mapping/Brokering software and services
  - concepts, capabilities, implementations

C2CAMP

Corporation for National Research Initiatives

MPCDF
Max Planck Computing & Data Facility

CSC – IT Center for Science

DEEP CARBON OBSERVATORY

ICOS
Integrated Carbon Observation System

INDIANA UNIVERSITY BLOOMINGTON

CLARIN
Common Language Resources and Technology Infrastructure

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

NCAR

DISSCO
Distributed System of Scientific Collections

JÜLICH
FORSCHUNGSZENTRUM

Computer Network Information Center
Chinese Academy of Sciences

doi

NATURAL HISTORY MUSEUM

DIRISA

Naturalis
Biodiversity
Center

DKRZ
Deutsches Klimarechenzentrum

is-enes

Questions on DiSSCo Technical Architecture?

Contact

Wouter Addink @wouter99999

&

Alex Hardisty @AlexHardisty