DiSSCo

Distributed System of Scientific Collections

Digital Specimens

Widening access to natural science collections

Alex Hardisty

Director of Informatics Projects, School of Computer Science and Informatics, **Cardiff University**

ICEDIG work package leader: Data Infrastructure, Design Alternatives and Economics

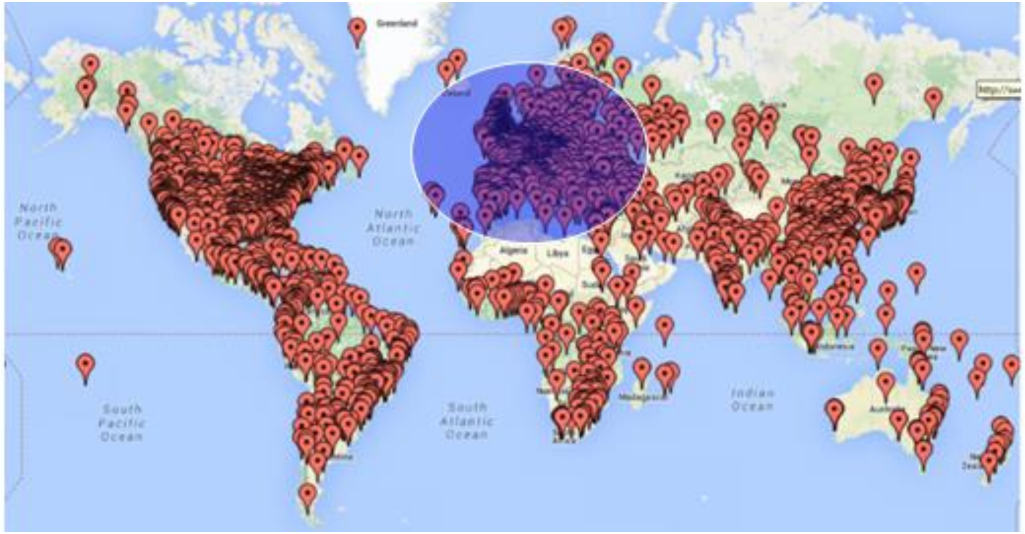
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ICEDIG.EU

ICEDIG project

Design Refinement Study for DiSSCo

European Natural Science Collections



- > **1.5 billion** specimens
- > **500** institutes/museums
- > **80%** of world's species
- > **5,000** scientists employed
- > **25,000** scientific visitors pa
- > **10 million** public visitors pa
- > **25 million** web visitors pa

A vast resource of biodiversity knowledge – past, present and future

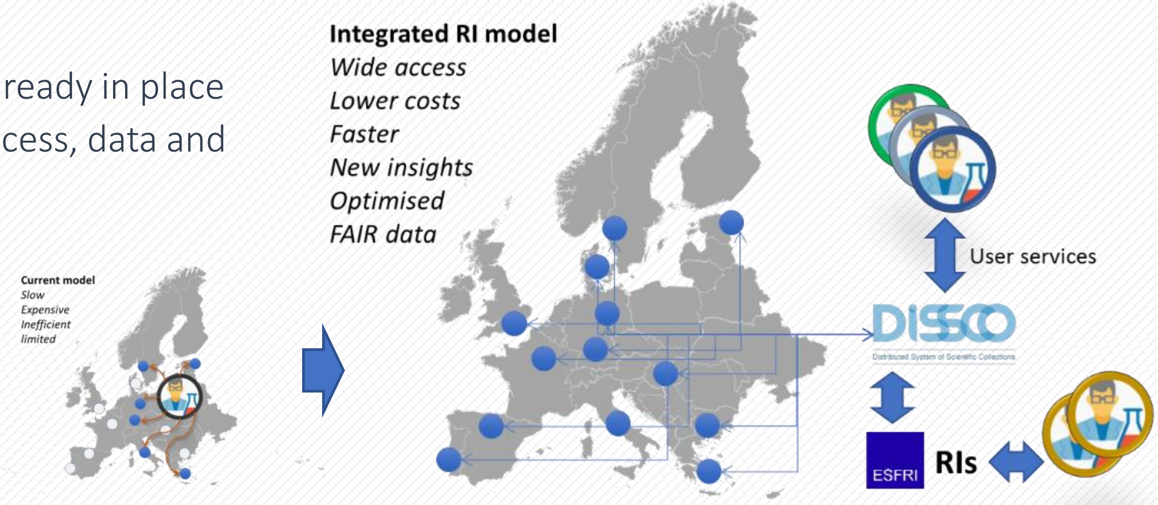
DiSSCo: A new European infrastructure

115 National Facilities

21 Countries

ESFRI Roadmap 2018

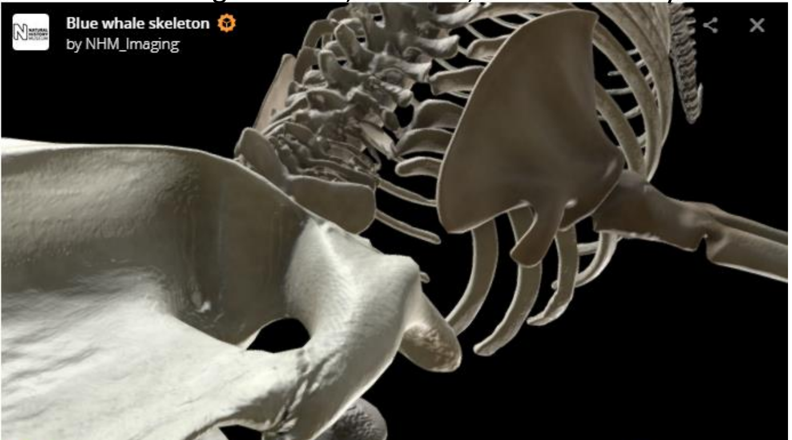
- **Largest ever** formal agreement between natural science collection facilities
- **Centralised governance** model already in place
- **Synchronisation** of facilities at access, data and policy level
- One European **virtual Collection**



Current model
Slow
Expensive
Inefficient
limited

Integrated RI model
Wide access
Lower costs
Faster
New insights
Optimised
FAIR data

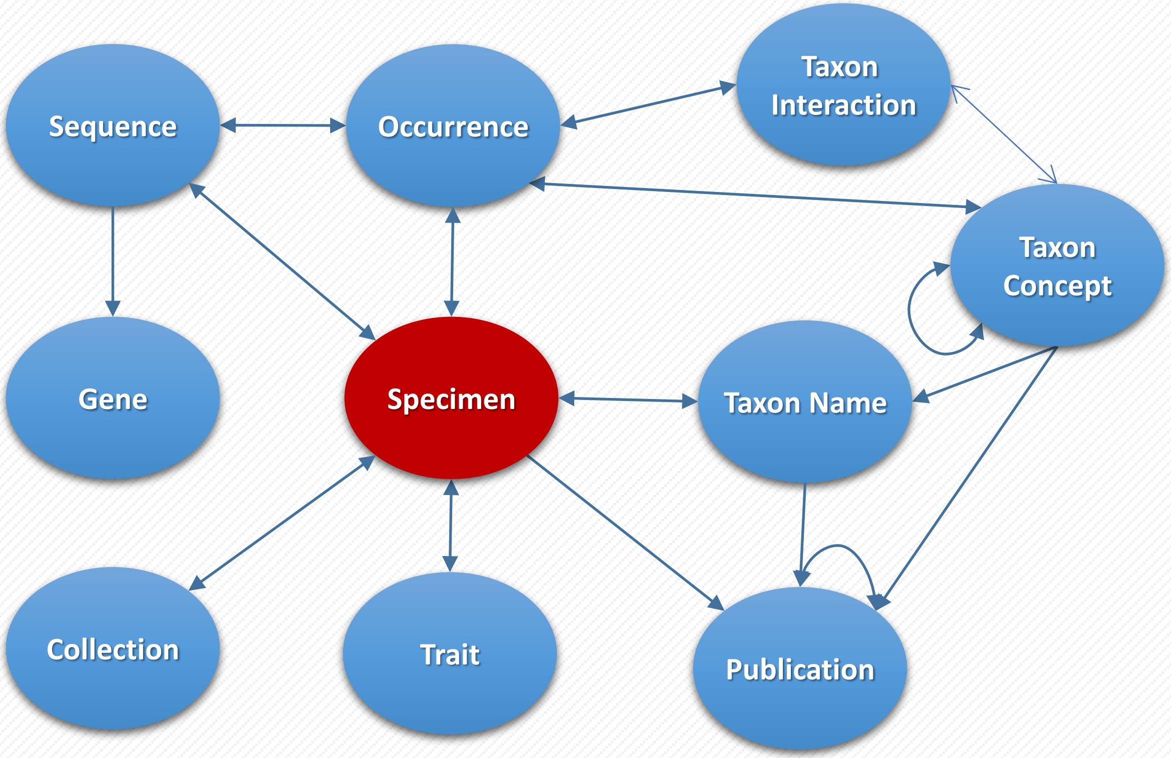
Pictures from Digitarium.fi, Picturae, Natural History Museum



DiSSCo

Distributed System of Scientific Collections

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



Specimens representations ('Digital Specimens') become the centrepiece of the DiSSCo knowledge base. They are used as anchoring points for diverse and dispersed data classes.

Digital Specimen: A dynamic "box" collecting links to all core information about a thing in one place

Images (2D, 3D)

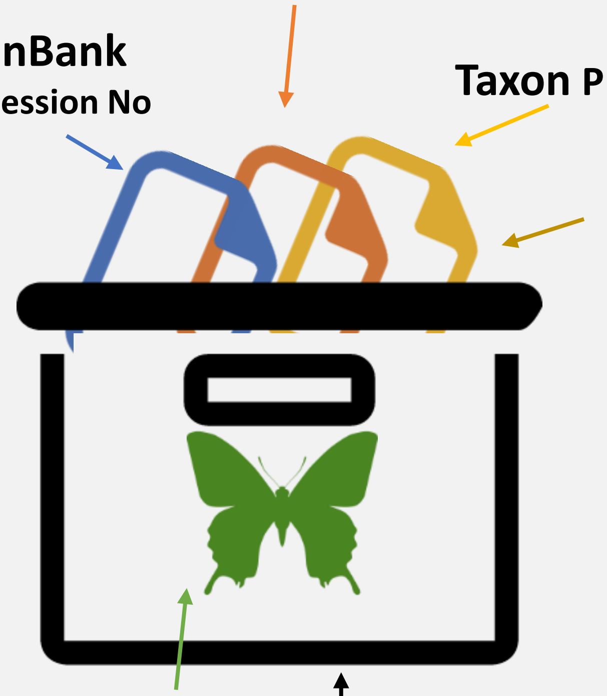
GET Image PIDs

GET Image metadata

GET Physical Object (PO) PID

GET PO PID metadata

DiSSCo
Distributed System of Scientific Collections



Occurrence ID

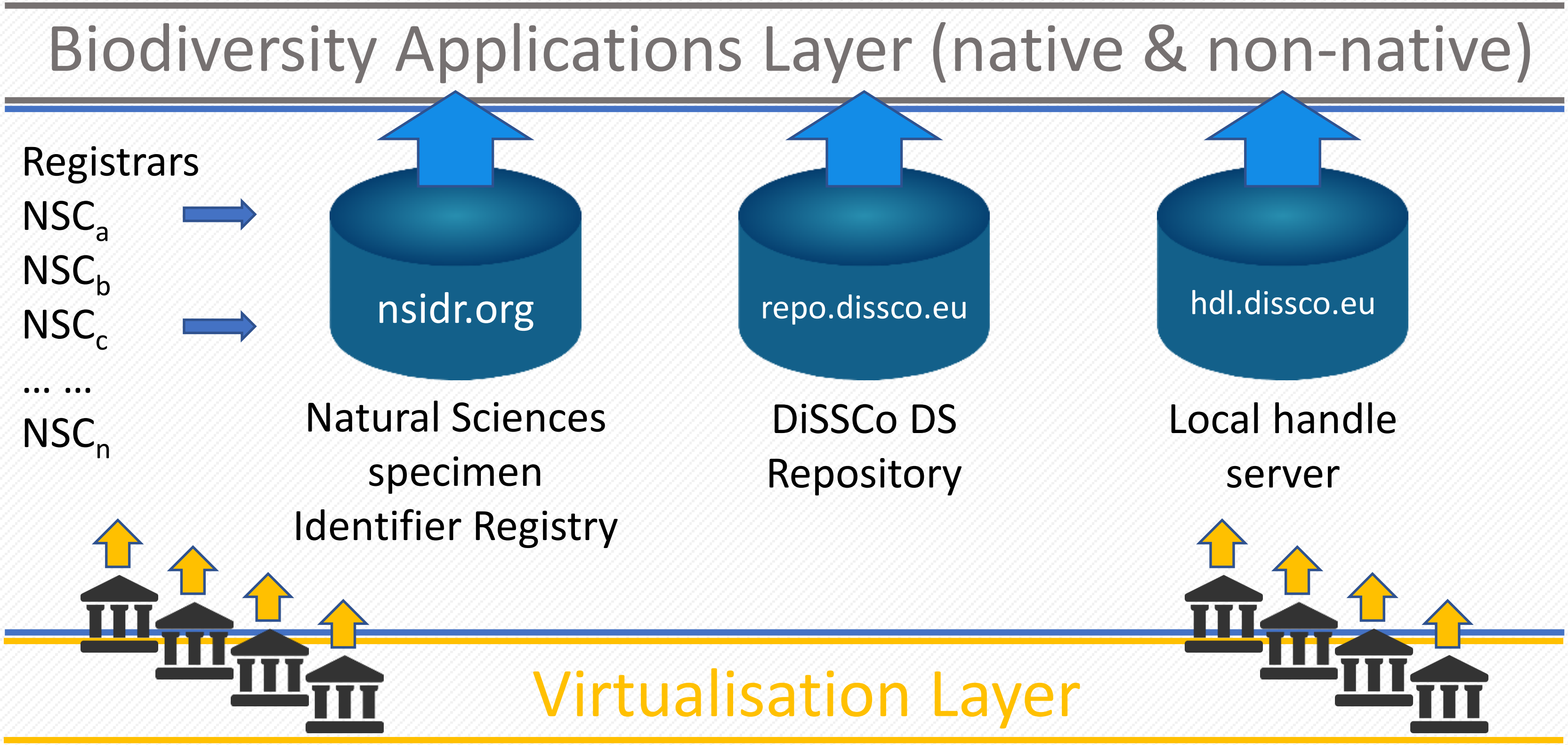
GenBank
Accession No

Taxon PID

PO PID
Image PIDs

Digital Specimen object

Digital Specimen (DS) objects layer inserted to unify natural science collections into a single data-driven European virtual Collection



Biodiversity Applications Layer (native & non-native)

Registrars
NSC_a
NSC_b
NSC_c
… …
NSC_n

nsidr.org

Natural Sciences specimen Identifier Registry

repo.dissco.eu

DiSSCo DS Repository

hdl.dissco.eu

Local handle server

Virtualisation Layer

Social and political concerns

- Tell a compelling added-value story, demonstrating value of handles in addition to HTTP URIs while playing down jargon and disabusing past associations
  - High value services from resolution of curated DS e.g.,
    - Retrieving images and/or annotation history of the specimen
    - Arranging to visit or for on-demand imaging of a specimen
  - Linking specimens to other related info, such as GenBank, etc.
- Accelerator mechanism will be applied
  - Digitisation of specimens is already well underway. Advances in mass digitisation can bring the cost down.
  - BUT
    - No “Natural Sciences specimen Identifier Registry” (nsidr.org) exists today
    - Specialist tools must be built to find and make links

Performance

- 1.5 billion specimens, 20 – 30 times as many links! Several hundred registrars and multiple link building tools
- Multiple object types
  - Digital specimen (DS), Specimen class, Collection type, Container type, Organisation type
- Achieving balance between fast presentation of informative registry records and the need to fetch and unpack comprehensive object content from a repository

Dynamic nature of DS demands extensibility

- DS have a set of mandatory and optional element types BUT DS are dynamic and can become more comprehensive over time.
- JSON facilitates a standard packaging format for exchanging DS and for extending DS with new information.
- All institutions’ software must understand the ‘standard’ information in a DS and the extension mechanism.
- Any institution can define and publish DS extensions to include new element types in DS in a way that allows any other institution to publish similar information in a mutually recognisable form.
- The extension mechanism specifies rules determining how to behave in respect of unrecognised or unsupported elements of DS.

Investing in handles

- Selecting an identifier scheme; buying into an invested, sustaining community
- Options for handles:
  1. Acquire top-level prefix from an MPA – XX in XX.NNNNN/
  2. Acquire second-level prefix – NNNNN in XX.NNNNN/
     - From Crossref, Datacite, ePIC, etc. Ideally, 4 digits.
- Rejected options
  1. Third level prefix e.g., from a Datacite member – too long!
  2. International Geo Sample Number (IGSN) – assumes physical PID and digital PID are the same. Doesn't work for natural science specimens.
- Main considerations:
  - Longevity/sustainability – 30 years at least
  - Flexibility of metadata in PID (registry) records – need PID Kernel Information Profile for Digital Specimens

DiSSCo

Distributed System of Scientific Collections

Questions on DiSSCo?

Contact

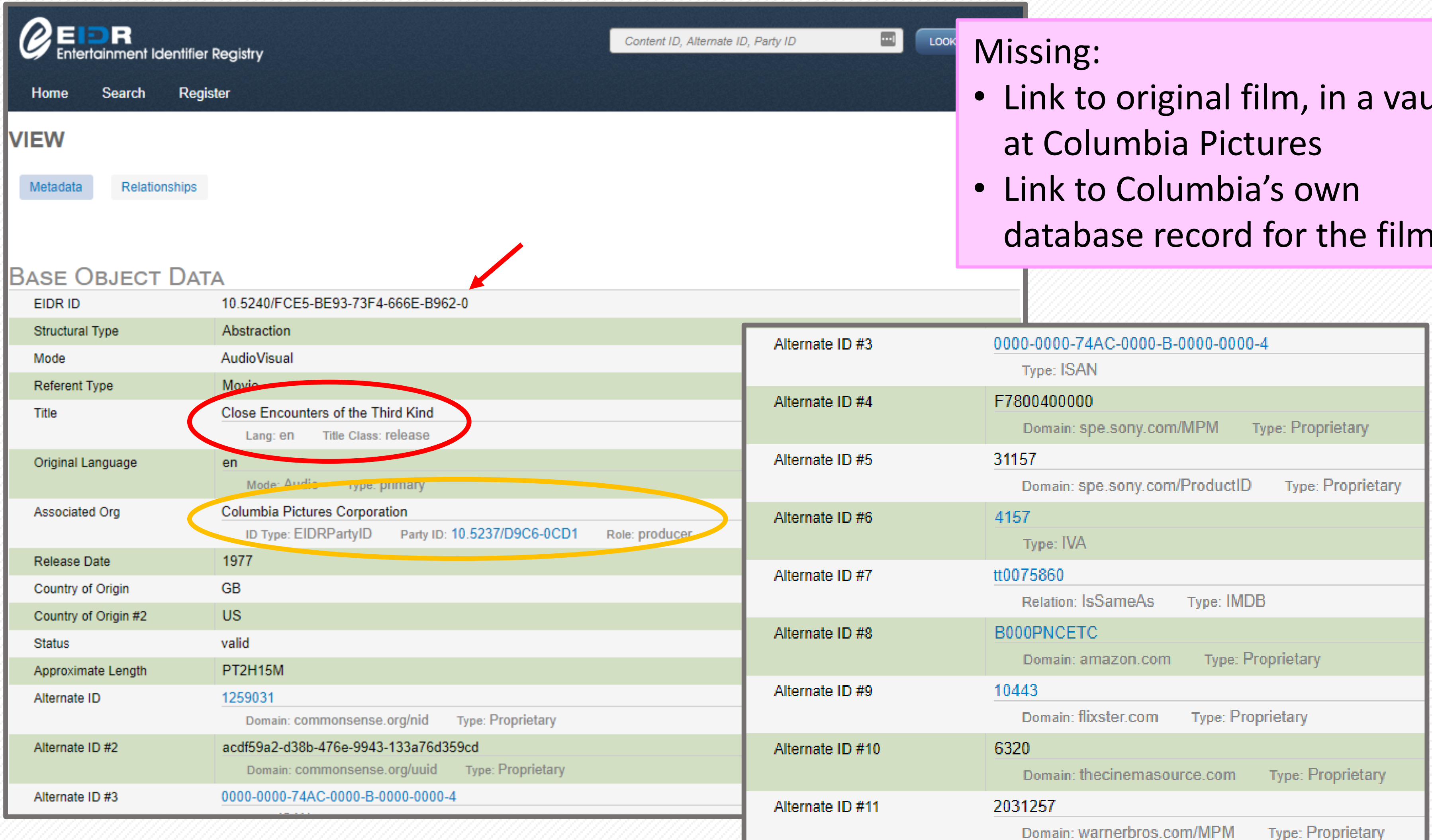
Dimitris Koureas @DimitrisKoureas

Wouter Addink @wouter99999

Alex Hardisty @AlexHardisty

CETAF stable identifiers (PURL) already in use

```
{
 "nsid": “21.nnnnn/20180904.000001",
  "type": "digital_specimen",
  "creators": [{+<expand for person details>}],
  "created": "2018-09-04T11:22:25.766698+00:00",
  "scientificName": "Toxodon platensis Owen, 1837",
  "specimen_records": [{"record_1":"http://data.nhm.ac.uk/object/34d2e921-01b5-40b7-8762-4269fac3c63d",
                        "record_2":http://data.nhm.ac.uk/dataset/darwins-fossil-mammals/resource/..."}],
  "institutionCode": "NHMUK",
  "collectionCode": "PAL",
  "catalogNumber": "PV M 100016",
  "recordedBy": "Charles R. Darwin",
  "eventTerms": [{"year":1833,"month":10,"day":10}],
  "locationTerms": [{"country":"Argentina","locality":"Cliff section on … …"}],
  "physical_specimen_pid": “PV M 100016",
  "annotations": [{"determinationNames":["Toxodon Owen, 1837","Toxodon platensis Owen, 1837"]}],
  "2d_images": [{
      "image_1": [{+<expand for links to hi-res and low-res two dimensional images / metadata>}],
      "image_2": [{+<expand for links to hi-res and low-res two dimensional images / metadata>}],
      "image_3": [{+<expand for links to hi-res and low-res two dimensional images / metadata>}]  }],
  "3d_images": [{
      "image_1": [{+<expand for links to hi-res 3-dimensional image / model>>}]  }]
}
```


EIDR
Entertainment Identifier Registry
Content ID, Alternate ID, Party ID
LOOK
Home
Search
Register
Metadata
Relationships

VIEW

Base Object Data

EIDR ID	10.5240/FCE5-BE93-73F4-666E-B962-0
Structural Type	Abstraction
Mode	AudioVisual
Referent Type	Movie
Title	Close Encounters of the Third Kind Lang: en Title Class: release
Original Language	en Mode: Audio Type: primary
Associated Org	Columbia Pictures Corporation ID Type: EIDRPartyID Party ID: 10.5237/D9C6-0CD1 Role: producer
Release Date	1977
Country of Origin	GB
Country of Origin #2	US
Status	valid
Approximate Length	PT2H15M
Alternate ID	1259031 Domain: commonsense.org/nid Type: Proprietary
Alternate ID #2	acdf59a2-d38b-476e-9943-133a76d359cd Domain: commonsense.org/uuid Type: Proprietary
Alternate ID #3	0000-0000-74AC-0000-B-0000-0000-4 Type: ISAN
Alternate ID #4	F7800400000 Domain: spe.sony.com/MPM Type: Proprietary
Alternate ID #5	31157 Domain: spe.sony.com/ProductID Type: Proprietary
Alternate ID #6	4157 Type: IVA
Alternate ID #7	tt0075860 Relation: IsSameAs Type: IMDB
Alternate ID #8	B000PNCETC Domain: amazon.com Type: Proprietary
Alternate ID #9	10443 Domain: flixster.com Type: Proprietary
Alternate ID #10	6320 Domain: thecinemasource.com Type: Proprietary
Alternate ID #11	2031257 Domain: warnerbros.com/MPM Type: Proprietary

Missing:

- Link to original film, in a vault at Columbia Pictures
- Link to Columbia's own database record for the film

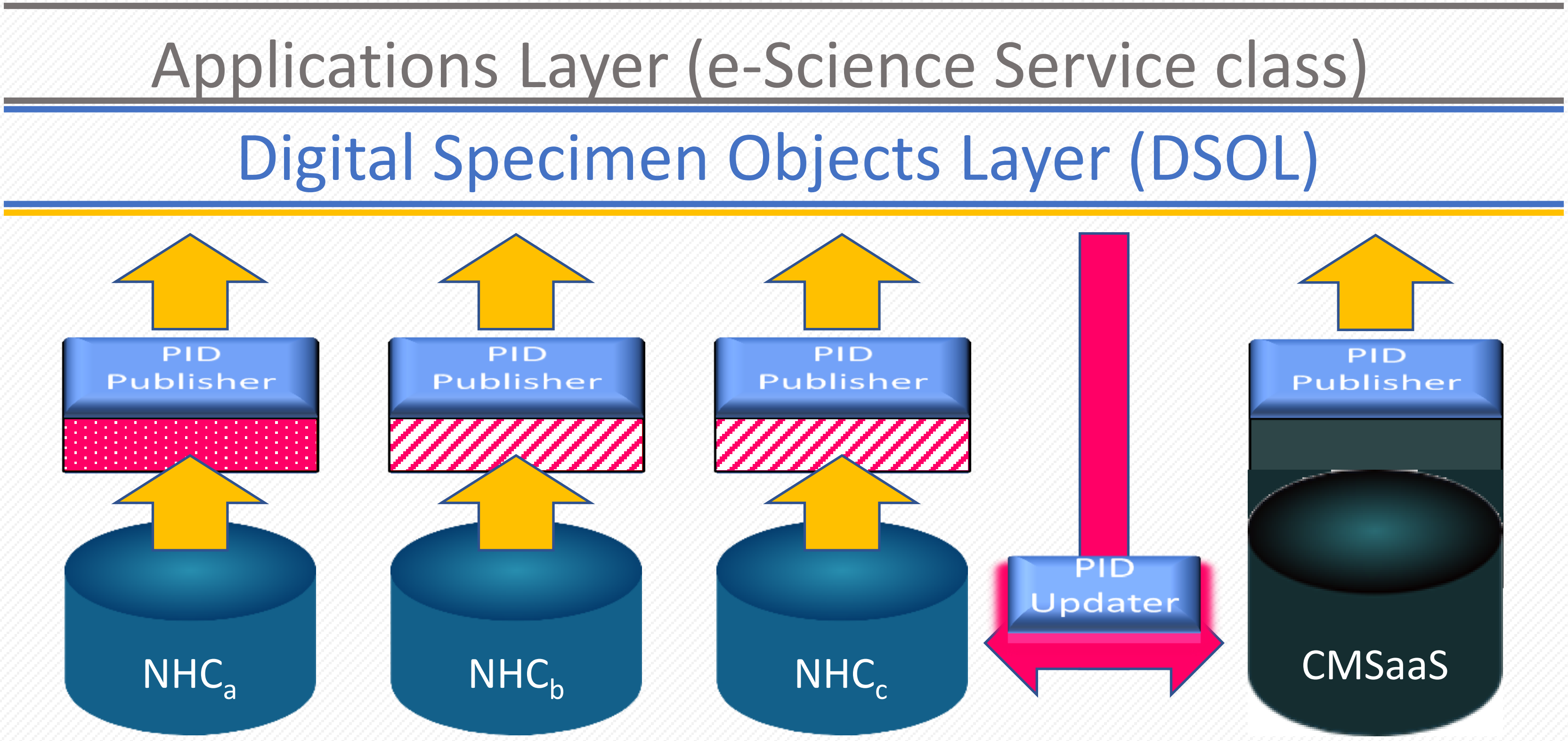
DiSSCo layers

Applications Layer (e-Science Service class)

Digital Specimen Objects Layer (DSOL)

Virtualisation Layer

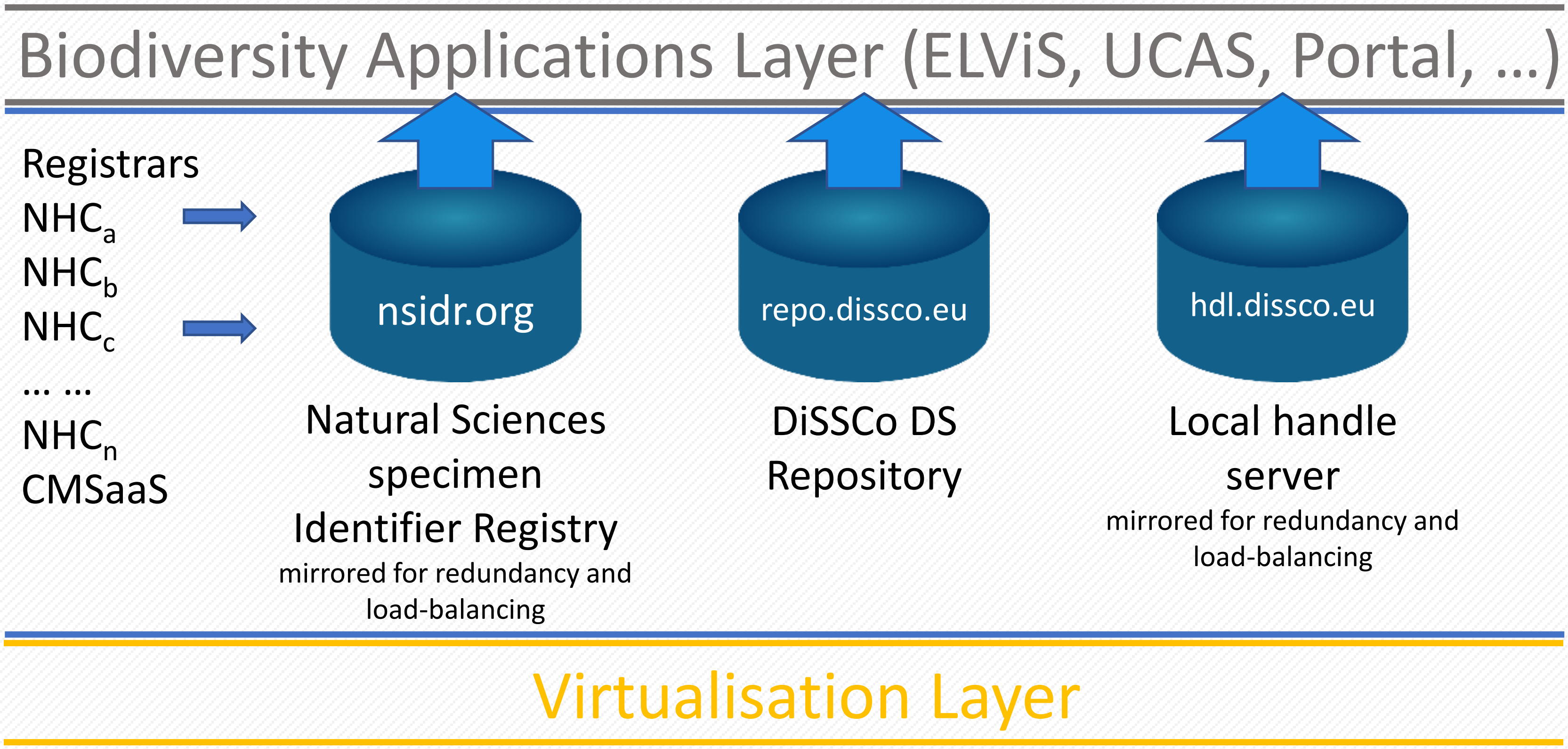
DiSSCo Virtualisation layer



Applications Layer (e-Science Service class)

Digital Specimen Objects Layer (DSOL)

Digital Specimen Objects layer (DSOL) inserted to unify natural science collections into a single data-driven European virtual Collection



Biodiversity Applications Layer (ELViS, UCAS, Portal, …)

Registrars

NHC$_a$

NHC$_b$

NHC$_c$

… …

NHC$_n$

CMSaaS

Natural Sciences specimen Identifier Registry

mirrored for redundancy and load-balancing

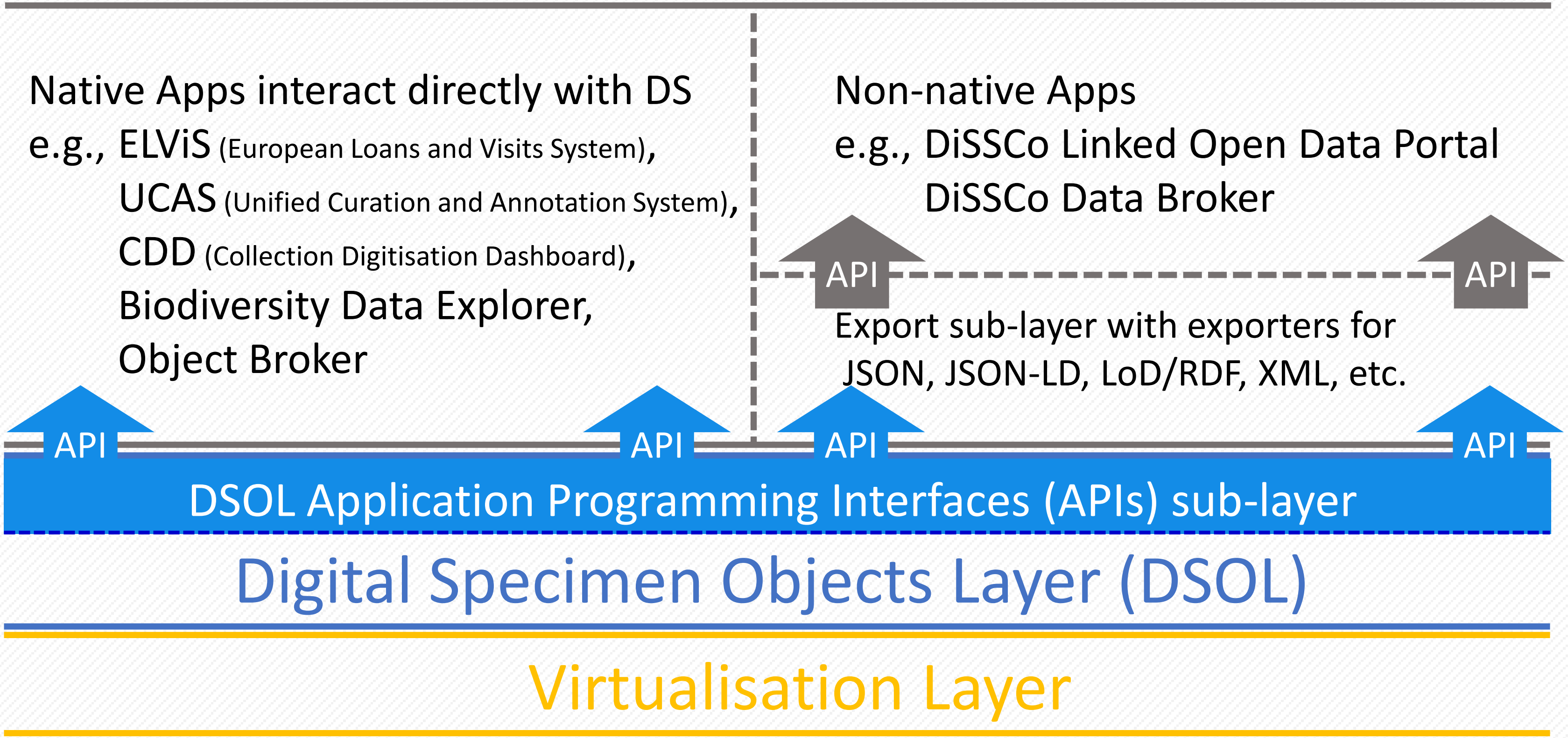
DiSSCo DS Repository

Local handle server

mirrored for redundancy and load-balancing

Virtualisation Layer

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



Native Apps interact directly with DS
e.g., ELViS (European Loans and Visits System),
UCAS (Unified Curation and Annotation System),
CDD (Collection Digitisation Dashboard),
Biodiversity Data Explorer,
Object Broker

Non-native Apps
e.g., DiSSCo Linked Open Data Portal
DiSSCo Data Broker

Export sub-layer with exporters for
JSON, JSON-LD, LoD/RDF, XML, etc.

DSOL Application Programming Interfaces (APIs) sub-layer

Digital Specimen Objects Layer (DSOL)

Virtualisation Layer

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