FAIR DIGITAL OBJECT AND RDA OUTPUT INCORPORATION IN DiSSCo's DESIGN

SHARIF ISLAM 0000-0001-8050-0299 @gnu111
Data Architect
Distributed Systems of Scientific Collections (DiSSCo) / Naturalis Biodiversity Centre (Leiden, The Netherlands)
RDA DATA FABRIC INTEREST GROUP MONTHLY MEETING
22 OCTOBER, THU, 2020

https://dissco.eu
120 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
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.
Current model
Slow
Expensive
Inefficient
limited

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

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

Digital Representation
FAIR Digital Object

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

An actionable knowledge unit
DATA AGGREGATION PATH

1. Individuals transcribing the original data & metadata (at the time of collection)

2. Copying the source data from field notes into a local database

3. Data clean up, data transformation, addition (e.g. geo-referencing, taxonomic identification)

4. Copy the data to regional/mid level aggregator

5. Copy the data to higher level aggregator
DATA AGGREGATION PATH

1. Individuals transcribing the original data & metadata (at the time of collection)

2. Copying the source data from field notes into a local database

3. Data clean up, data transformation, additional (geo-referencing, taxonomic identification)

4. Copy the data to regional/mid level aggregator

5. Copy the data to higher level aggregator
DATA COLLECTION

Collecting Activities And Processing

Lot sorting, taking pictures

Sub Sampling

Tissue Clipping DNA Extraction

Taxonomic Experts

Museum Specimen Collection Tissue, DNA

Sequencing Experts

Collection Manager

Molecular Database Barcode Database

Database Manager
Any specimen related data
- taxonomic, genomic, geographical...

Any application
- database, data repository, visualization, machine learning tools ...

Any platform
- cloud, desktop, mobile ...

→ How do we leverage globally distributed data with multifaceted applications in a scalable, long-term infrastructure?
Cornerstone: Digital Object (DO)

An important tenet of our approach is to move from a system-centric world to an information-centric world.

Hides “System” Implementation Details

Provides Uniform “Information” Interface

Reference: Giridhar Manepalli (CNRI) 2019
A Digital Object oriented approach for Natural Science Collections data:
Vision for an endless number of levels of abstraction
FAIR DIGITAL OBJECT
(historical context)

- Digital Object: “A digital object is a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a handle” – Kahn and Wilensky, 1995 (http://www.cnri.reston.va.us/k-w.html).

- FAIR (Findable, Accessible, Interoperable, Reusable) - Wilkinson et al. 2016.

- ”Digital Objects have the built-in capability to implement FAIR data” – George Strawn, 2019.

- FAIR Digital Object “..is a stable actionable unit that bundles sufficient information to allow the reliable interpretation and processing of the data contained in it” - De Smedt, K., Koureas, D. and Wittenburg, P., 2020.

Why FAIR Digital Objects for natural science collections data

- To use heterogenous data sources
- Need digital representations of physical specimens – unambiguous, persistent
- Different levels of abstraction
- Different actors and agents
- Collect and anchor core information about the specimen
- Information needs to be persistently and unambiguously linked to necessary context for interpretation and validation
- Standard and domain specific operations for specific applications
PIDs are pointers that resolve to location (URL) of the item e.g., DO itself, physical specimen, hi-res images, label information, tissue sample, DNA sequence, etc. Image credit: Alex Hardisty.
Physical Object

Digital Representation

FAIR Digital Object

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

An actionable knowledge unit
DIGITAL SPECIMENS

- Digital Specimen is a particular type of FAIR Digital Object.
- Digital Specimens are more than just digital representations. They act as digital containers and processable digital twins on the Internet for physical specimens in collections.
Holorchis castex

NADH dehydrogenase subunit 1 (ND1) gene, partial cds; mitochondrial.

Location: 1..459

Serialised as JSON:
{
    "type": "DigitalSpecimen",
    "attributes": {
        "content": {
            "id": "20.5000.1025/486a7e883f14f88bba37",
            "creationDateTime": "",
            "creator": "",
            "midLevel": 2,
            "scientificName": "Holorchis castex Bray & Justine",
            "country": "New Caledonia",
            "locality": "Rocher a la voile",
            "decimalLat/Long": [22.3, 166.42],
            "recordedBy": "J L. Justine",
            "collectionDate": "2006-06-01",
            "catalogNumber": "2006.12.6.40-41",
            "otherCatalogNumbers": "NHMUK:ecatalogue:7072219",
            "institutionCode": "NHMUK",
            "collectionCode": "ZOO, Parasitic worms",
            "stableIdentifier": "https://data.nhm.ac.uk/object/e90b81bc-1642-47ca-b587-6aa8885cd6a0/1579046400000",
            "physicalSpecimenId": "013258549",
            "Annotations": {
                "Type status = paratype. Holotype = MNHN JNC 1848 – D1",
                "gbifId": "https://www.gbif.org/occurrence/1826086349",
                "literatureReference": "https://doi.org/10.5281/zenodo.175744",
                "treatmentbank": "http://tb.plazi.org/GgServer/html/03BA8782-5543-FFFB-D895-FD27FAE1DDAD",
                "enaBiosample": "None available",
                "enaSequence": "https://www.ebi.ac.uk/ena/data/view/FJ788436",
                "elements": []
            }
        }
    },
    "elements": [
        {
            "type": "DigitalSpecimen",
            "attributes": {
                "content": {
                    "id": "20.5000.1025/486a7e883f14f88bba37",
                    "creationDateTime": "",
                    "creator": "",
                    "midLevel": 2,
                    "scientificName": "Holorchis castex Bray & Justine",
                    "country": "New Caledonia",
                    "locality": "Rocher a la voile",
                    "decimalLat/Long": [22.3, 166.42],
                    "recordedBy": "J L. Justine",
                    "collectionDate": "2006-06-01",
                    "catalogNumber": "2006.12.6.40-41",
                    "otherCatalogNumbers": "NHMUK:ecatalogue:7072219",
                    "institutionCode": "NHMUK",
                    "collectionCode": "ZOO, Parasitic worms",
                    "stableIdentifier": "https://data.nhm.ac.uk/object/e90b81bc-1642-47ca-b587-6aa8885cd6a0/1579046400000",
                    "physicalSpecimenId": "013258549",
                    "Annotations": {
                        "Type status = paratype. Holotype = MNHN JNC 1848 – D1",
                        "gbifId": "https://www.gbif.org/occurrence/1826086349",
                        "literatureReference": "https://doi.org/10.5281/zenodo.175744",
                        "treatmentbank": "http://tb.plazi.org/GgServer/html/03BA8782-5543-FFFB-D895-FD27FAE1DDAD",
                        "enaBiosample": "None available",
                        "enaSequence": "https://www.ebi.ac.uk/ena/data/view/FJ788436",
                        "elements": []
                    }
                }
            }
        }
    ]
}
TECHNICAL AND COMMUNITY ALIGNMENT

**Recommendations**

**Guiding Principles**

*Resolvable via well-known resolver,
†Accessible via a type registry

DO = Digital Object
### DiSSCo Design Decisions and FAIR data lifecycle

<table>
<thead>
<tr>
<th>Adoption of Digital Object Architecture</th>
<th>Community Resource</th>
<th>Workflow</th>
<th>DiSSCo Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA Data Foundation and Terminology WG</td>
<td>All phases of the</td>
<td>Digital Specimen Architecture</td>
<td></td>
</tr>
<tr>
<td>Data Fabric and Terminology IG</td>
<td>data lifecycle</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Persistent Identifiers and Kernel Information</th>
<th>Community Resource</th>
<th>Workflow</th>
<th>DiSSCo Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA PID Kernel WG</td>
<td>- Meta-information about a digital object</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- DiSSCo (data) type registry</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- FAIR data principles
- Global discussions around FAIR Digital Objects
- Outputs from the Research Data Alliance (RDA) interest and working groups
- ENVRI FAIR, GOFAIR Initiative
## DiSSCo Design Decisions and FAIR data lifecycle

<table>
<thead>
<tr>
<th>Aggregation of digital objects</th>
<th>Community Resource</th>
<th>Workflow</th>
<th>DiSSCo Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA Research Data Collection WG</td>
<td>All phases of the data lifecycle</td>
<td>Digital Specimen Architecture</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata attribution and use of PROV entities</th>
<th>Community Resource</th>
<th>Workflow</th>
<th>DiSSCo Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA/TDWG Metadata attribution WG</td>
<td>Digitization, curation and maintenance of digital object</td>
<td>Digital Specimen and collection objects</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FAIR data maturity model</th>
<th>Community Resource</th>
<th>Workflow</th>
<th>DiSSCo Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA FAIR data maturity model WG, ENVRI FAIR, GOFAIR Initiative</td>
<td>All phases of the data lifecycle</td>
<td>Develop guidelines and specifications to assess FAIR implementation plan</td>
<td></td>
</tr>
</tbody>
</table>
DiSSCo Building Blocks

European Loans and Visits System
European Curation and Annotation System
Specimen Data Refinery
Collections Monitoring Dashboards
Knowledge base
Future e-services

DiSSCo core e-services

Collections Overview
Structured Institutional Collection holdings description

Describe institute collection holdings by dimensions + metrics (e.g. PreservationType, GeologicalTimeRang)
Aggregation of Digital Objects

Specimen Details
Digital Specimen
LESSONS LEARNT

- We need to focus on demos and targeted proof of concept works along with the design.
- Reaching out to the user community to identify edge cases (workshops, hackathons).
- Collaborate and learn from other domains/research infrastructures.
- Rinse, repeat – it is a marathon not a sprint (of course sprints are important too!)
ACKNOWLEDGEMENTS

- RDA Europe 4.0 (https://cordis.europa.eu/project/id/777388  Grant Agreement No. 777388).
- DiSSCo Prepare (https://www.dissco.eu/what-is-dissco-prepare/ ) is funded by the Horizon 2020 Framework Programme of the European Union (H2020-INFRADEV-2019-2020 – Grant Agreement No. 871043)
- Image attribution: By the author and the DiSSCo Technical Team (except where otherwise noted)
- Questions?