Integration with DiSSCo – what are reasonable first steps?

Wouter Addink, wouter.addink@naturalis.nl

https://orcid.org/0000-0002-3090-1761

ICEDIG.EU

SYNTHESYS+ Synthesis of Systematic Resources a DiSSCo project

mobilise

DiSSCo prepare

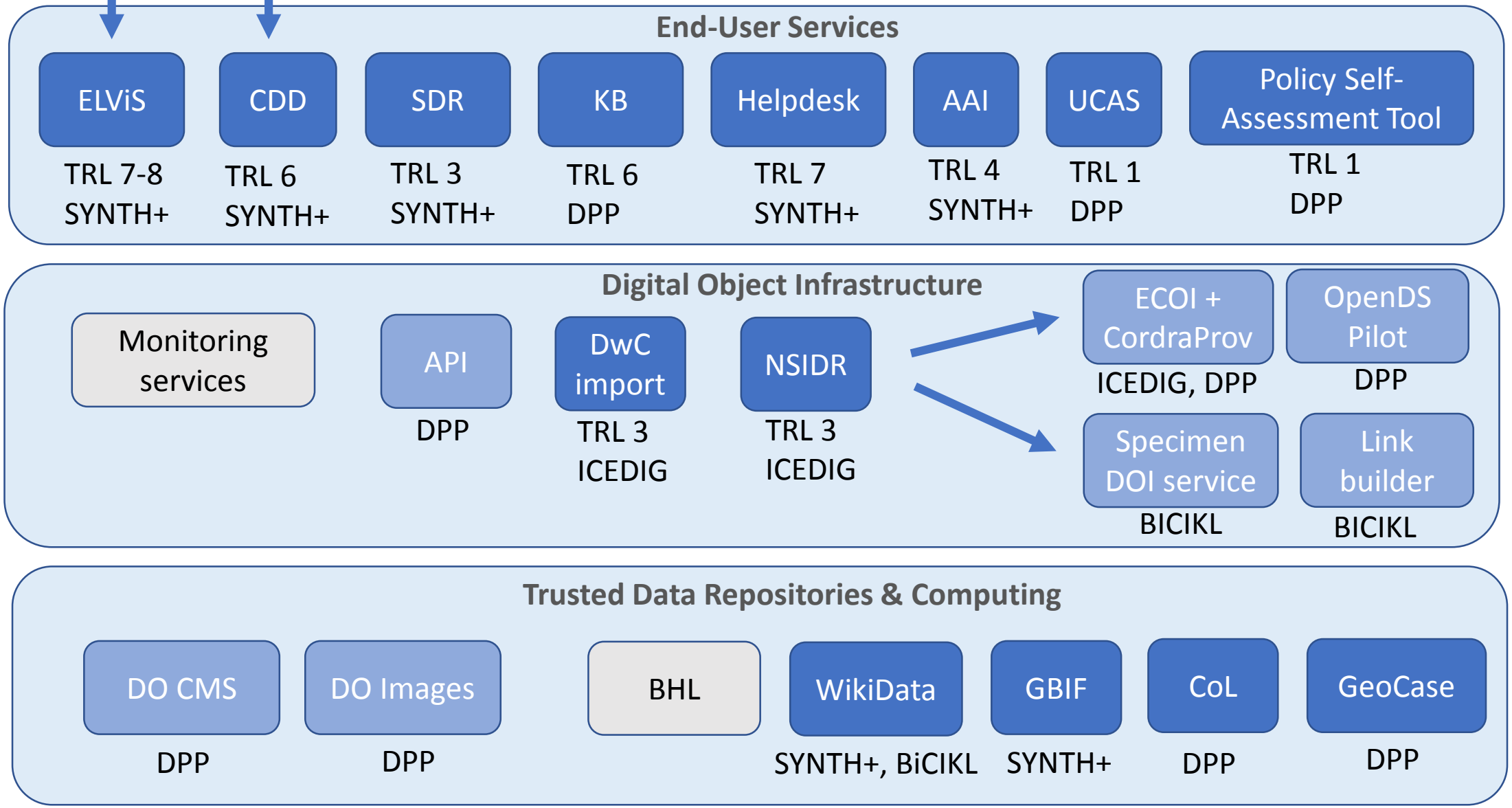
FAIR ENVRI

infrastructure overview

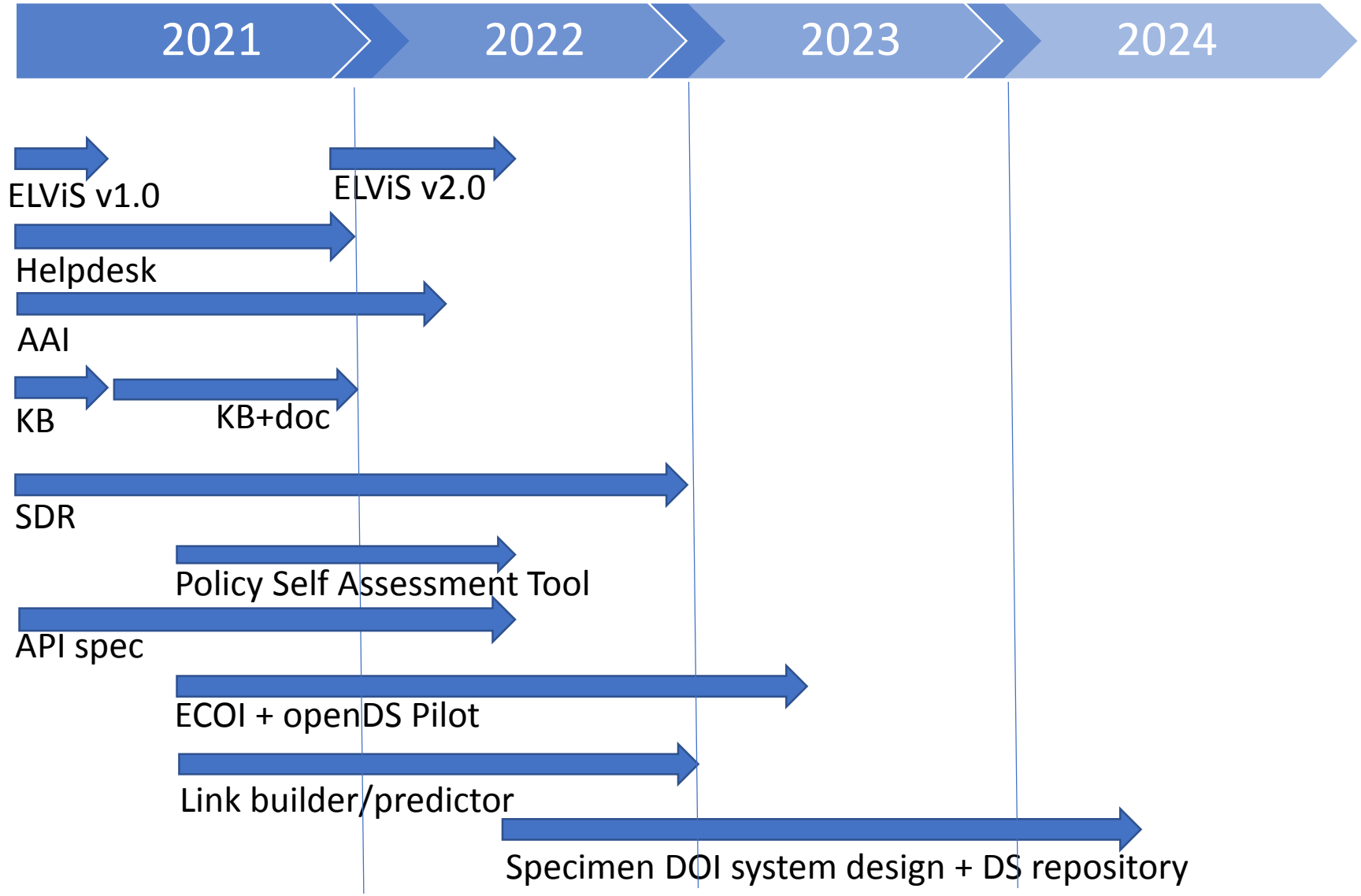
- Current status
- Planned

+ expertise?

+ facilities, resources, policies?



Timeline


2021
2022
2023
2024
ELViS v1.0
ELViS v2.0
Helpdesk
AAI
KB
KB+doc
SDR
Policy Self Assessment Tool
API spec
ECOI + openDS Pilot
Link builder/predictor
Specimen DOI system design + DS repository

Where to find the components

E-Service locations

- Mature e-Services (TRL 8-9): **Dissco.eu prefix** (elvis.dissco.eu, know.dissco.eu)
- Prototypes, demonstrators: **Dissco.tech/labs** (CDD,..)
- Digital specimen testing (not restricted to DiSSCo): **NSIDR.org**

DiSSCo GitHub (incl user stories):
https://github.com/DiSSCo

DiSSCo Digital transformation

1. Identify institutional technical+administrative contacts (institutional moderators)
2. Define technical requirements for minimal interoperability
3. Evaluate institutional technical readiness status, capacities and capabilities
4. Develop an overall integrated implementation plan
5. Implement the plan through dynamic interaction with the team of technical contacts (Enablement team) at all levels to align, coordinate and support each other in collective effort.

Planned Timeline for 2021

Proposed roles for ISTC in DiSSCo digital transformation

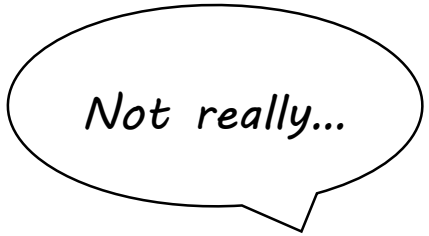
1. Make proposals and create community support for changes at international level with strategic guidance from the DiSSCo technical team
   - Example: change proposals for increased PID support and ‘what is it’ description in DarwinCore

2. Advise DiSSCo developers and create community consensus on implementation details
   - Example: What information does an institution need about a requester to give access to restricted data?

3. Organise workshops for the DiSSCo enablement team to build capacity and work out implementation details
   - Example: Organise a workshop to define and discuss small digital transformations (‘microchanges’) that DiSSCo institutions could achieve already this year with minimal technical knowledge and resources.

Darwin Core

- Widely used TDWG standard (1,6 Billion DwC records in GBIF)
- Ratified standard since 2009-10-09
- Maintenance group in place



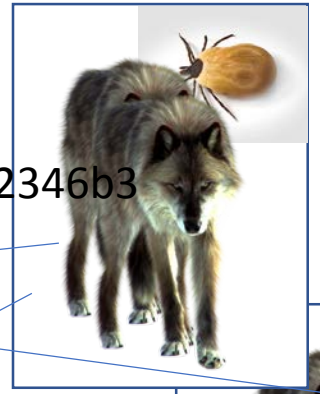
- The basic model – everything is an occurrence (MaterialSample, PreservedSpecimen, HumanObservation, etc) – is problematic
- Its basic classes (basisOfRecord) are problematic
- Lack of support for PIDs (resolvable identifiers)
- Not much controlled values (great for sharing, not for use)
- 88 open issues, mainly proposals for term changes or additions (https://github.com/tdwg/dwc/issues)

Comments about the basic Darwin Core model

DwC definition of an Occurrence:
An existence of an Organism (sensu http://rs.tdwg.org/dwc/terms/Organism) at a particular place at a particular time.

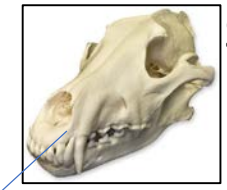
Example: A wolf pack on the shore of Kluane Lake in 1988.

Occ:12346b3



S12

S15



S16

Occ:12345axOcc:12346b3
Coll nr:112

S14

S13

5 specimens +
1 sample (catalogued in a separate catalog)

SA324

*So: there is a potential **many to many relationship** between Occurrences and Specimens!* When the specimen ID (globally unique version catalogeNumber) is filled in occurrence ID, you end up with 5 occurrence records instead of 2.

What you might expect:

occurrenceID	catalogNumber	recordNumber
Occ:12345ax	S12,S13,S14	112
Occ:12346b3	S12,S15	

occurrenceID: An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). But also: In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.

What you probably get:

occurrenceID	catalogNumber	recordNumber
InstA:CollB:S12	S12	112
InstA:CollB:S13	S13	112
InstA:CollB:S14	S14	112
InstA:CollB:S15	S15	112

One way to solve things is to create different objects and give each of these objects an identifier. This is very much in line with the FAIR Digital Objects idea and also seems to be the direction GBIF is thinking about It will take years to make this change though.

Another thing which makes things difficult is that we try to use one standard for both data created in the field, in an institution and in the lab.

Note:there is also an identifier for MaterialSample (as opposed to a particular digital record of the material sample).See http://rs.tdwg.org/dwc/terms/index.htm#materialSampleID

MaterialSample: A physical result of a sampling (or subsampling) event. In biological collections, the material sample is typically collected, and either preserved or destructively processed.

materialsampleID seems a better fit for specimen identifiers than occurrenceID

Current DwC classes (also recommended values for basisOfRecord)

- LivingSpecimen
- PreservedSpecimen
- FossilSpecimen
- HumanObservation
- MachineObservation

Issues:

1. As these define the origin of the record, a material citation is also a PreservedSpecimen, which is confusing. Proposal for new class Material Citation #329
2. Distinction between HumanObservation and MachineObservation is fuzzy at best: #314
3. The definition of MaterialSample is essentially the same as that for PreservedSpecimen #314
4. PreservedSpecimen, FossilSpecimen and LivingSpecimen would be better modelled as subclasses of MaterialSample #314, see also Dina model (but: is LivingSpecimen a MaterialSample or an Organism?
5. Occurrence terms like catalogNumber, otherCatalogNumbers, associatedSequences, and preparations would be better placed under MaterialSample because these have nothing to do with an Occurrence.
6. There are no fields for person identifiers like identifiedByID or recordedByID

Graph model basis of
2016 Darwin-SW (DSW) ontology
(version 1.0)

Steven J. Baskauf – Vanderbilt University
Campbell O. Webb – Arnold Arboretum of Harvard University

Basic class relationships laid out by Richard L. Pyle
http://lists.tdwg.org/pipermail/tdwg-content/2010-October/001703.html

Key:

namespace:property	property (italicized)
namespace:Class	instance of named class
⟶	property arc from subject to object
⟷	inverse property pair; arrow with property name shows direction

Colors:

ns:X ← ns:y	red=Darwin Core
ns:X ← ns:y	blue=Darwin-SW
ns:X ← ns:y	yellow=FOAF vocabulary

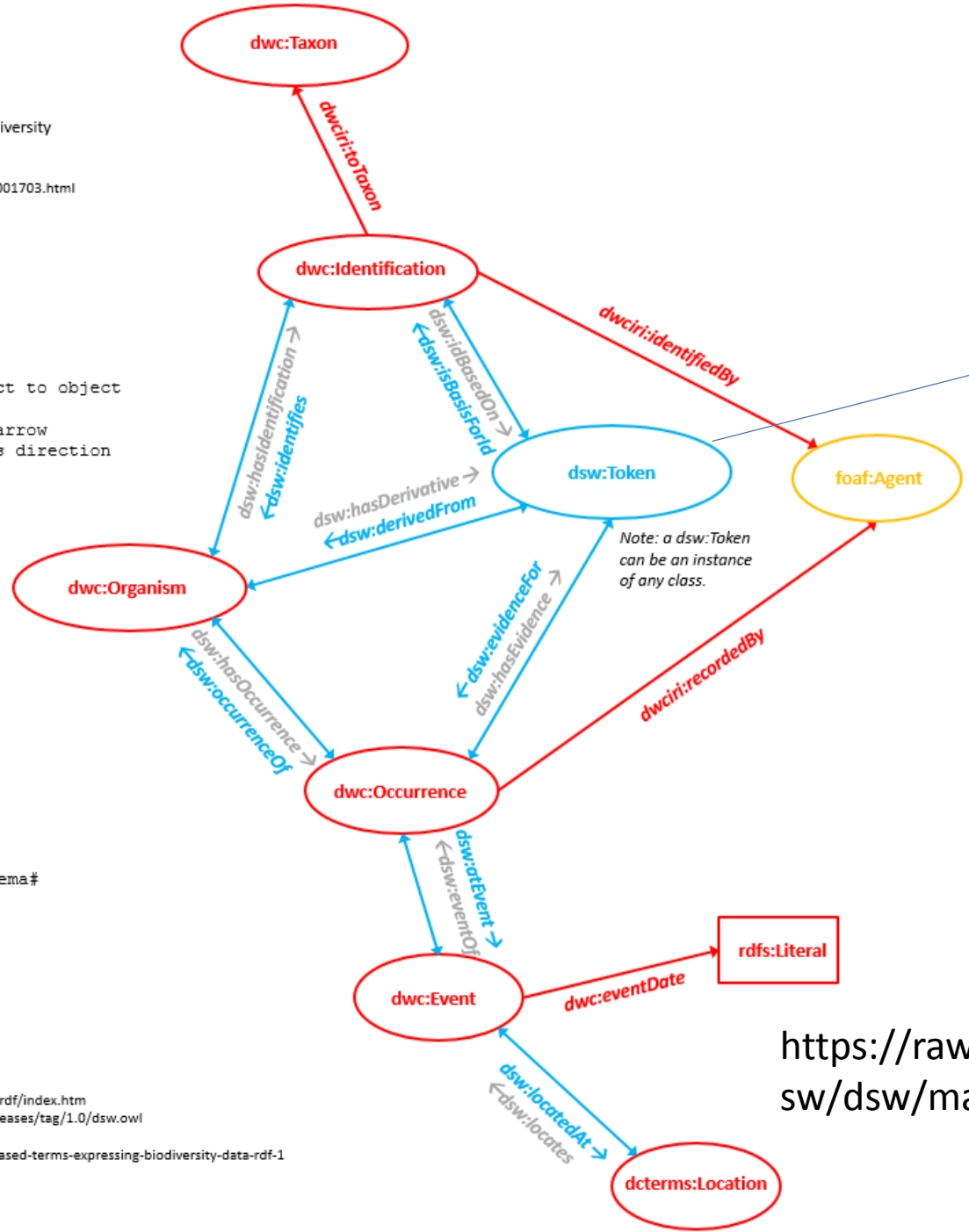
Gray represents the non-preferred predicate of an inverse pair.

Namespace abbreviations:

rdfs:	http://www.w3.org/2000/01/rdf-schema#
dwc:	http://rs.tdwg.org/dwc/terms/
dwciri:	http://rs.tdwg.org/dwc/iri/
dsw:	http://purl.org/dsw/
dcterms:	http://purl.org/dc/terms/
foaf:	http://xmlns.com/foaf/0.1/

Sources:

Darwin Core RDF Guide from http://rs.tdwg.org/dwc/terms/guides/rdf/index.htm
Darwin-SW version 1.0 from https://github.com/darwin-sw/dsw/releases/tag/1.0/dsw.owl
Darwin-SW is described in
http://semantic-web-journal.net/content/darwin-sw-darwin-core-based-terms-expressing-biodiversity-data-rdf-1



Tokens (provide some sort of evidence for an assertion)

- MaterialSample
  - PreservedSpecimen
  - FossilSpecimen
  - LivingSpecimen
- Unvouchered Report
  - HumanObservation
  - MachineObservation
- Multimedia
  - StillImage
  - Sound
  - MovingImage
  - etc.
- MaterialCitation

https://raw.githubusercontent.com/darwin-sw/dsw/master/img/dsw-1-0-graph-model.png

Recommended actions for ISTC

Note: Next annual DwC update is planned 30 April: https://github.com/tdwg/dwc/milestone/14

1. Support the proposal for new class MaterialCitation (#329) so these can be clearly distinguished from PreservedSpecimen. Note: that is may also be recommended for addition in ABCD
2. Provide clear specimen identifier guidelines for CETAF &DiSSCo partners on which DwC fields to use for specimen identifiers, e.g. might promote materialSampleID instead of occurrenceID and should use a resolvable identifier such as a CETAF identifier instead of a 'DarwinCore Triplet', should be clear ion what to give an identifier (each individually curated object, which can be a single object or a lot)
3. Support the proposal for addition of identifiedByID and recordedByID terms: #101 and #102. These are already implemented by GBIF and enable the use of ORCID iD and WikiData identifiers.
4. The new Agent Actions DwC extension (https://tools.gbif.org/dwca-validator/extension.do?id=https://tdwg.github.io/attribution/people/dwc/AgentActions) would serve DiSSCo needs better than the addition of identifiedByID and recordedByID, since it also allows to distinguish between different agents doing the recording or identification (in the future most identification may be done by machines). ISTC should encourage and help the finalization of this extension so that it can be supported by GBIF soon.
5. Create a proposal for addition of a new term for digitalSpecimenID for the 2022 DwC update

‘What is it’ description of specimens

In MIDS, CD and DwC, terms are needed for describing what a record or object represents. These should be aligned between the specifications and have controlled values (and not too many – in iSamples they aim for max. 20 values and in DiSSCo dashboard development there was a similar aim).

- What does it represent (determines what information to expect to be associated with the object)
- What is it made of (important for geological specimen)
- How does it look like/is it mounted (what can an image recognition algorithm expect to be on the specimen image)
- How is it preserved (for biological specimens: which technique is used to prevent physical deterioration of non-living collections

The challenges:

1. have unambiguous distinct terms with no overlap between the controlled values
2. Have an agreed list of controlled values. Some terms may require a hierarchy of terms but to keep it simple key-value pair solution is preferred

Currently millions of ‘preserved’ values in GBIF

https://github.com/tdwg/mids/files/5842404/bq-results-20210120-122837-ydhq0a99j5dl.xlsx

Mix of material types, preparation types and other things

v_preparations	institutionCode	f0_
hb	MNHN	5916620
herbarium specimen of unspecified type	E	887309
Pinned	NHMO	322624
Herbarium specimen	MA	77258
Otholiths	NHMO	75511
skin	MIZPAN	35395
Planting	GBG	26041
herbarium specimen	MA	20979
Study skin	NHMO	19016
Preserved specimen	NHMO	16660
various	NHMD	13799
xy	MNHN	11496
Blown eggshell	NHMO	9777
Dried specimen	MA	8935
various - 1	NHMD	8506
lame	MNHN	6940
liquid-preserved material	E	6855
lame mince	MNHN	5717
Ethanol	NHMO	5581
Cranium	NHMO	4661

‘What does it represent’ description of specimens

- In DwC there are:
  - Record-level type: The nature or genre of the resource, must be value from DCMI type vocab, examples: StillImage, PhysicalObject, Text
  - basisOfRecord: The specific nature of the data record, examples: PreservedSpecimen, LivingSpecimen

- iSamples uses specimenType (kind of object) and materialType (what the object (specimen) is composed)

- NCD uses: collectionType :
  Archival | Art | Audio | Cell Cultures | Electronic | Facsimiles | Fossils | Genetic | Geological | Herbarium | Living | Manuscripts | Mineralogical | Observations | Preserved | Products | Specimens | Texts | Tissue | Visual

- GGBN uses MaterialSampleType (tissue, culture strain, specimen, DNA, RNA, Protein, environmental sample), PreparationType (leaf, muscle, leg, blood), PreservationType (dried, silica, alcohol, FTA card, tube, QIAsafe)

- GeoCase: faceted search on a limited list of specimen types, for example RockSpecimen, MineralSpecimen, SedimentSample

- **CD and MIDS: not yet decided** (MIDS: MaterialType, CD: basisOfCollection, ObjectClassification, ObjectType, PreservationMethod..)

‘What it is’ in Naturalis:preservedPart and Mount

‘What does it represent’ description of specimens

Recommended action for ISTC:

- Following further definition in CD and MIDS, create a recommendation for the terms to use and controlled vocabularies to use in CETAF/DiSSCo to describe “what it is”.

What information does an institution need about a requester to give access to restricted data

- Will be piloted in Synthesys+ JRA1 through AAI pilot
- Input needed from ISTC!

From ELViS development:

- Person name
- email address
- ORCID iD
- Current affiliation (home institution)
- Publications
- CV

Q1: ORCID does not include CV but includes biography, would that be sufficient?
Q2: Would students need their supervisors to access the data?
Q3: anything else needed?

Micro changes to implement this year by the institutions:
1. Institutional identifiers

Challenges:

1. To describe the collections in **Europe as one virtual collection**, a highly standardised description in dimensions is needed and **no overlap** in specimen between collection descriptions.
2. Finding existing collection descriptions from sources like GrSciColl (https://registry.gbif.org/collection/search) is a challenge because currently you cannot yet rely on an institutional identifier like ROR/GRID. Instead you have to find an **institution name** which can have many variants.
3. Finding published collection datasets in GBIF is a challenge, not only because you cannot rely on an institutional identifier but also because **institutions not always match to a data publisher in GBIF**.

Institution international name	**GBIF publisher name**
Bavarian Natural History Collections	Staatliche Naturwissenschaftliche Sammlungen Bayerns
Westerdijk Fungal Biodiversity Institute	CBS Fungal Biodiversity Centre
University of Jyväskylä	Jyväskylä University Museum - The Section of Natural Sciences
Estonian University of Life Sciences	PlutoF (with as one of the datasets: Estonian University of Life Sciences)
National Museum of Natural History	UMS PatriNat (OFB-CNRS-MNHN), Paris

Micro changes to implement this year by the institutions:
1. Institutional identifiers

ROR:

- Now supports the full metadata GRID metadata schema including organization parent-child relations
- But: still dependent on GRID: 1:1 copy. GRID now has a more strict interpretation and does not support identifiers anymore for institutions embodied in another organization.
- ROR will in the future be independent with its own policy, might support identifiers for all our institutions (but no guarantee)
- ROR has a working group working on a ROR extension for departments
  - https://ror.org/[ror-id]/[subunit-id] (stored centrally at ROR, through GitHub or WikiData)
  - More info: doi.org/10.5281/zenodo.4552755

Proposal:

- Let each institution identify if they have a ROR already, or an ROR for their parent
- With a list of missing RORs, see if we can get them for all institutions
- If not: use CETAF registry IDs, make sure there is also a WikiDate entry, and link to the parent ROR

Micro changes to implement this year by the institutions:
2. Harmonize institution and collection identifiers in GBIF

The issue:
In GBIF, if the institutionID, collectionID, institutionCode and collectionCode are the same for collections in both GRSCiColl and the datasets, then these can be linked

This enables:

- Deduplication of GriSciColl records
- Direct links between specimen datasets and collection descriptions in GBIF
- Combination of non-digitized collection information and digitized specimens in DiSSCo services

Record

Term	Interpreted
Institution code	RMNH Naturalis Biodiversity Center, Zoology Collections
Basis of record	Preserved specimen
Collection code	PISC Pisces collection

https://www.gbif.org/occurrence/171309909

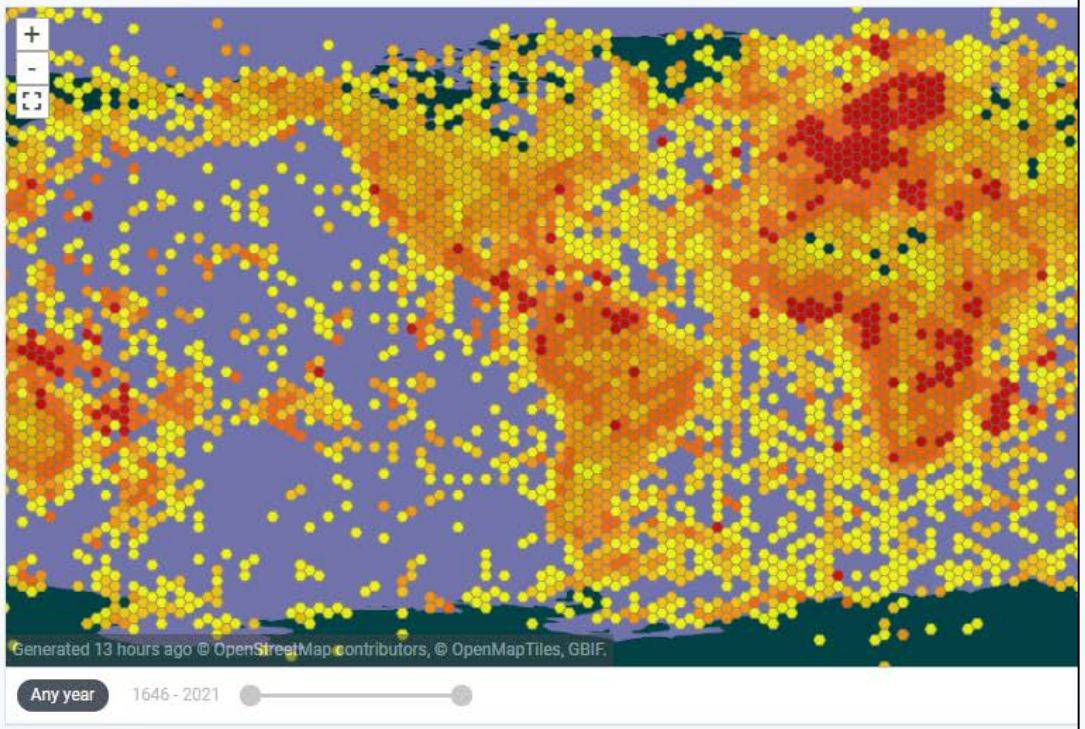
Micro changes to implement this year by the institutions:
3. Identify datasets in GBIF and GeoCase by checking a prepared list

NETWORK | REGISTERED APRIL 9, 2021

Distributed System of Scientific Collections (DiSSCo)

SUMMARY PUBLISHERS DATASETS METRICS HOME PAGE

13,057,776 GEOREFERENCED RECORDS



584 datasets from 55 publishers (next update: 62 DiSSCo publishers and 761 datasets)

OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Institution match fuzzy	17,623,566
Institution collection mismatch	12,426,358
Institution match none	8,226,791
Occurrence status inferred from individual count	7,124,415
Collection match none	6,993,839
Ambiguous collection	4,971,172
Geodetic datum assumed WGS84	3,898,773
Taxon match higherrank	2,464,444
Ambiguous institution	1,719,793
Taxon match none	1,513,558

Other Micro changes to implement this year by the institutions:

- Identify collection catalog datasets not yet shared with GBIF and GeoCase
  - Use European IPT installation to add to GBIF
- link data providers to institutional identifiers and add these to IPT metadata
- Provide CETAF register data including collection descriptions or at least names of the collections and collection codes
  - Potential issue: the register is much more detailed than the information needed for DiSSCo – provide clear guidance, also in the tool

Proposal for ISTC

- Prepare these micro changes further and organize a workshop to present and discuss these with the institutional moderators before summer.
- Anything missing that should be added?