

Ontologies in biodiversity data management

Dag Endresen - GBIF Norway, University of Oslo



Illustration: GBIF data portal

UiO data managers network | 8 June 2023

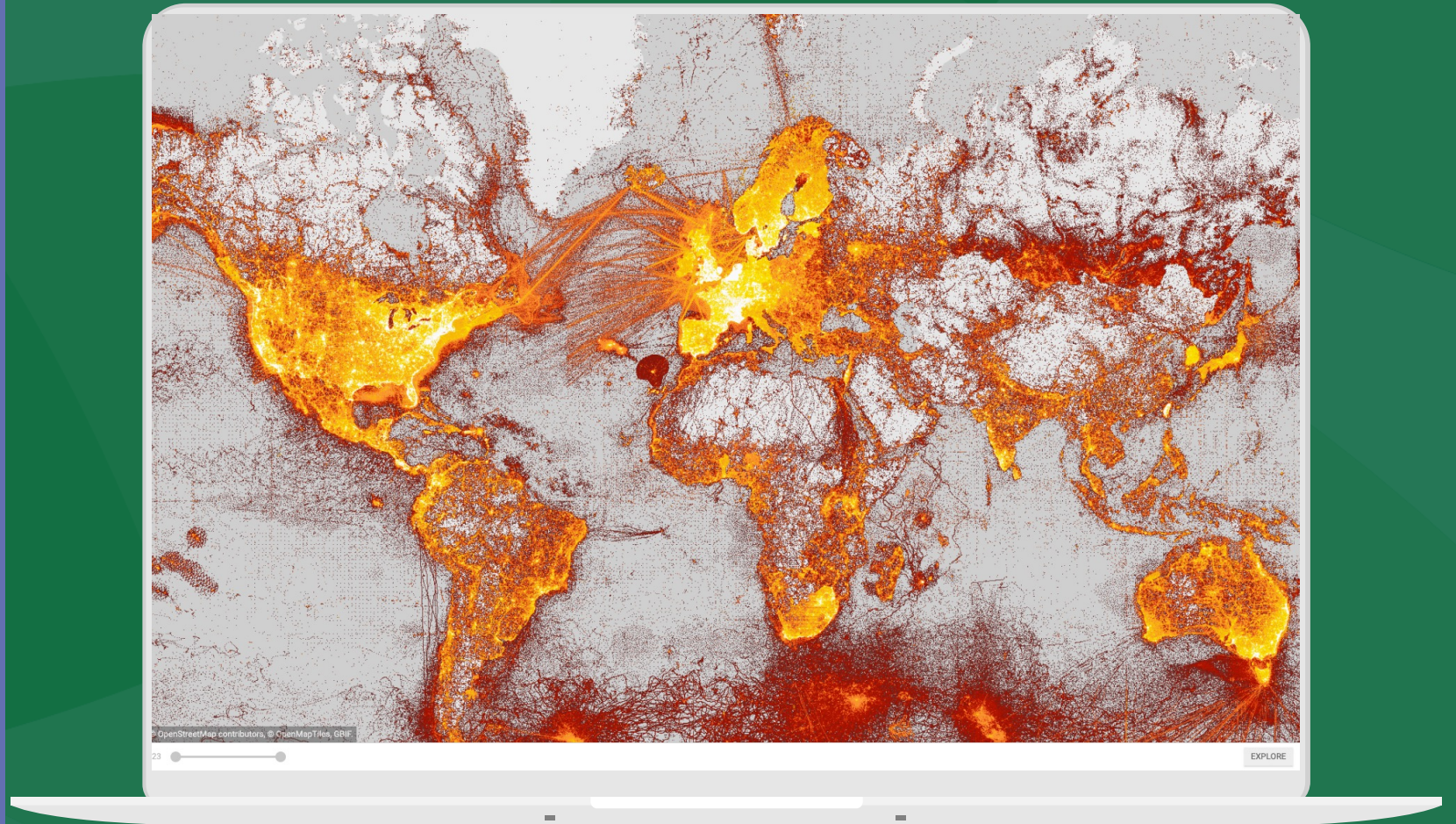
WHAT IS GBIF?

Intergovernmental network and research infrastructure

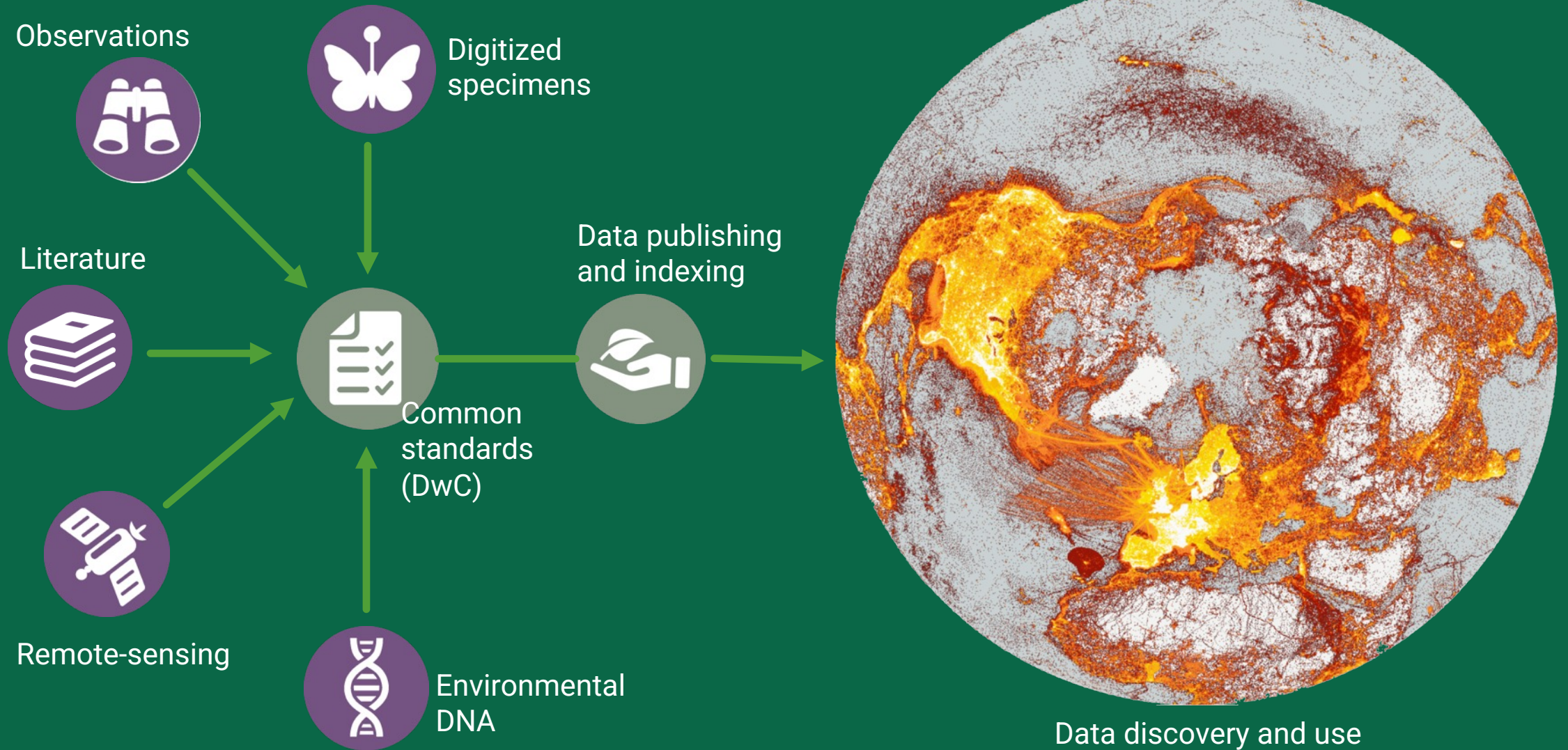
Provides anyone, anywhere, **free and open access to data** about all types of life on Earth

Distributed research data infrastructure based on **national participant NODES**

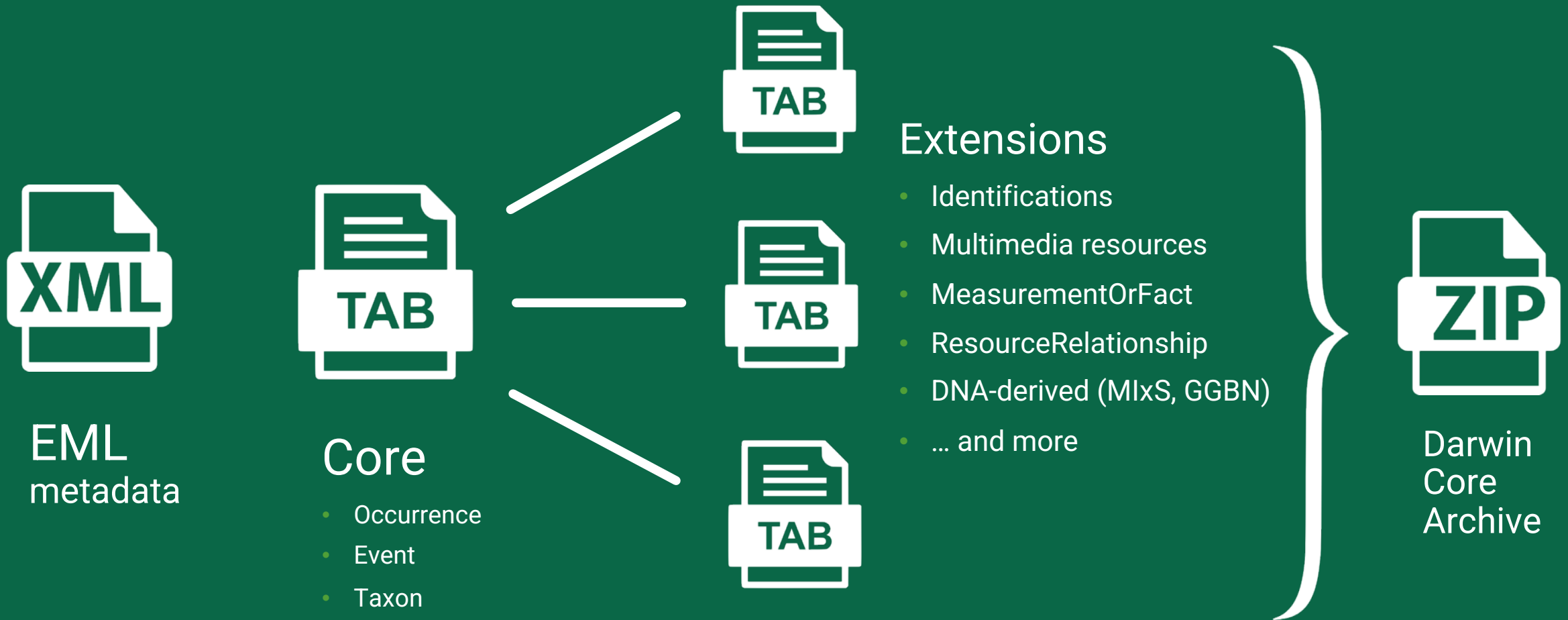
Secretariat in Copenhagen, Denmark



A WINDOW ON EVIDENCE ABOUT WHERE SPECIES HAVE LIVED, AND WHEN



DARWIN CORE ARCHIVE





Biodiversity Information Standards

TDWG

DARWIN CORE

Darwin Core is a data standard maintained by TDWG. It includes a **glossary of terms** (properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing **identifiers, labels, and definitions**. Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.

- Quick reference guide at <https://dwc.tdwg.org/terms/>
- Darwin Core Archive (text guide) at <https://dwc.tdwg.org/text/>
- Detailed list of DwC terms at <https://dwc.tdwg.org/list/>
- DwC is maintained in a GitHub repository : <https://github.com/tdwg/dwc>
- Darwin Core Q&A at: <https://github.com/tdwg/dwc-qa>



Biodiversity
Information
Standards

TDWG

DARWIN CORE CLASSES

- Record-level
- Occurrence *
- Organism
- MaterialSample
- Event *
- Location
- GeologicalContext
- Identification
- Taxon *
- MeasurementOrFact
- ResourceRelationship

Asterix (*) are supported as
“Cores” in GBIF
<https://rs.gbif.org/core/>

Most of the other classes
available as “Extensions” in GBIF,
meaning additional properties to
the “Core” records”
<https://rs.gbif.org/extension/>



Biodiversity
Information
Standards

TDWG

TDWG VOMAG (VOCABULARY MANAGEMENT STANDARDS)

- TDWG Standards Documentation Standard (SDS) (2017)
- TDWG Vocabulary Maintenance Standard (VMS) (2017)
- TDWG standards ratification process, review manager guidelines (2009, updated 2021)



Biodiversity
Information
Standards

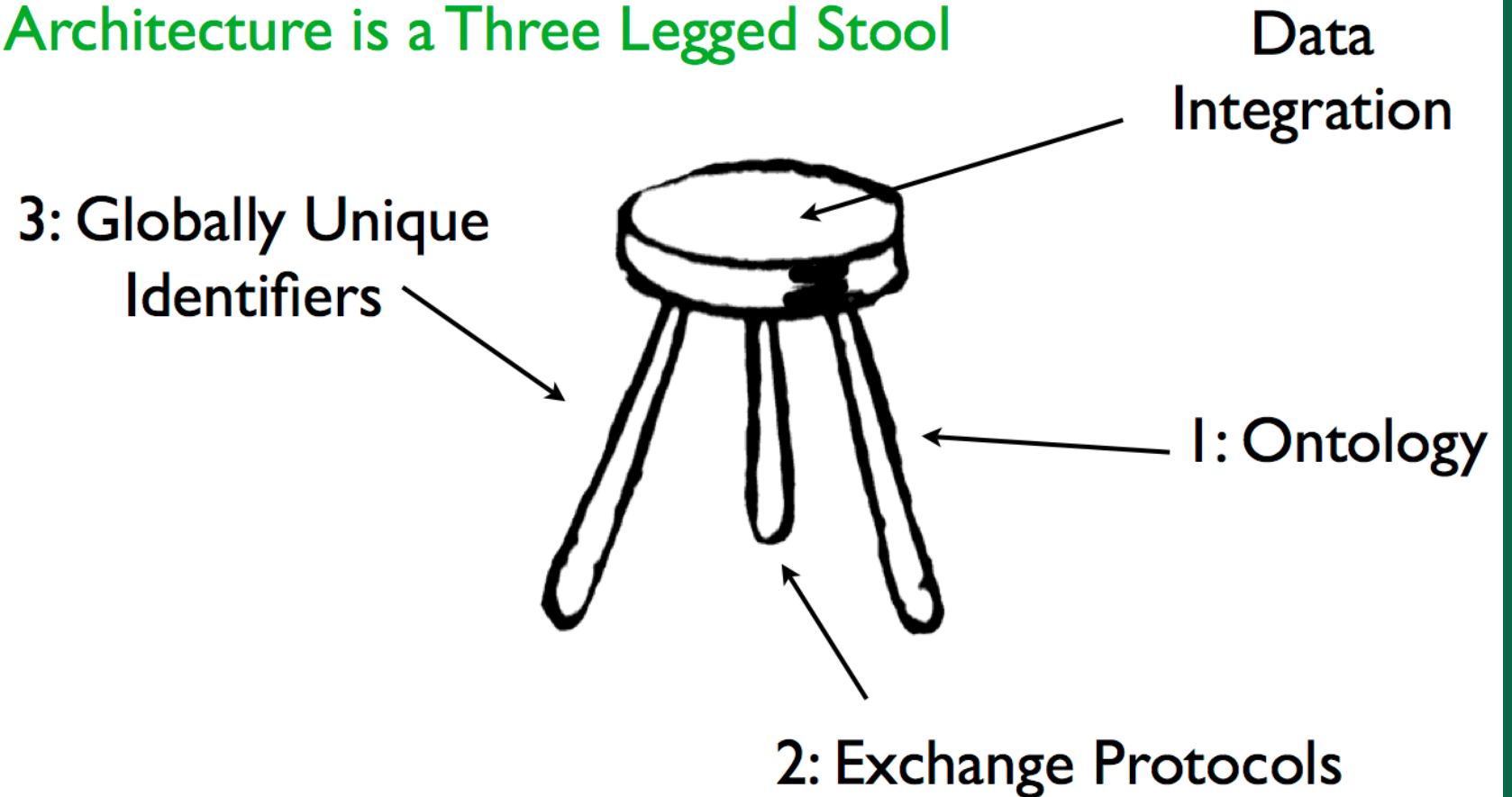
TDWG



TDWG TECHNICAL ARCHITECTURE | VOCABULARY MANAGEMENT

Vocabularies & ontologies are identified as one of the three core components in the **TDWG technical architecture**.

Architecture is a Three Legged Stool



VOCABULARIES/ONTOLOGIES

- Provide a shared understanding of what we mean when describing biodiversity entities.
- What kind of thing or property.
- A list of things we as a community can agree upon the meaning of.
- “*Concept repository*” with terms identified by URIs.

TERM LABEL VERSUS CONCEPT

Need to separate terms and term labels from the fundamental concepts behind the terminology.

BOX 1: The following syllogism will be familiar to students of Aristotelian logic



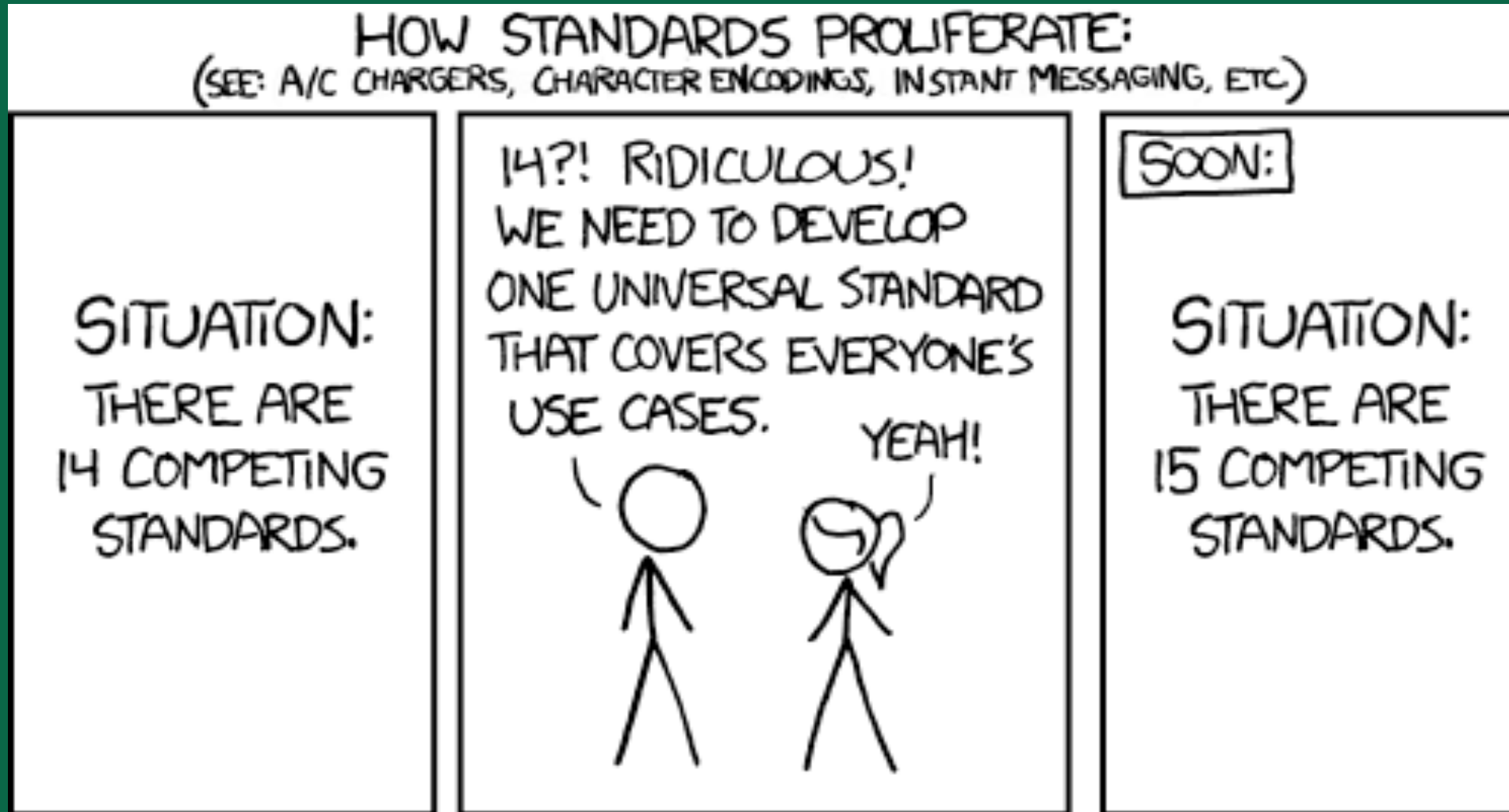
'man' is a 3-lettered word.

Socrates is a man.

Therefore, Socrates is a 3-lettered word.

The logical flaw is very obvious to a human reader, but a computer can easily be fooled if statements about a term are presented looking like statements about the concept represented by the term.

DATA STANDARDS



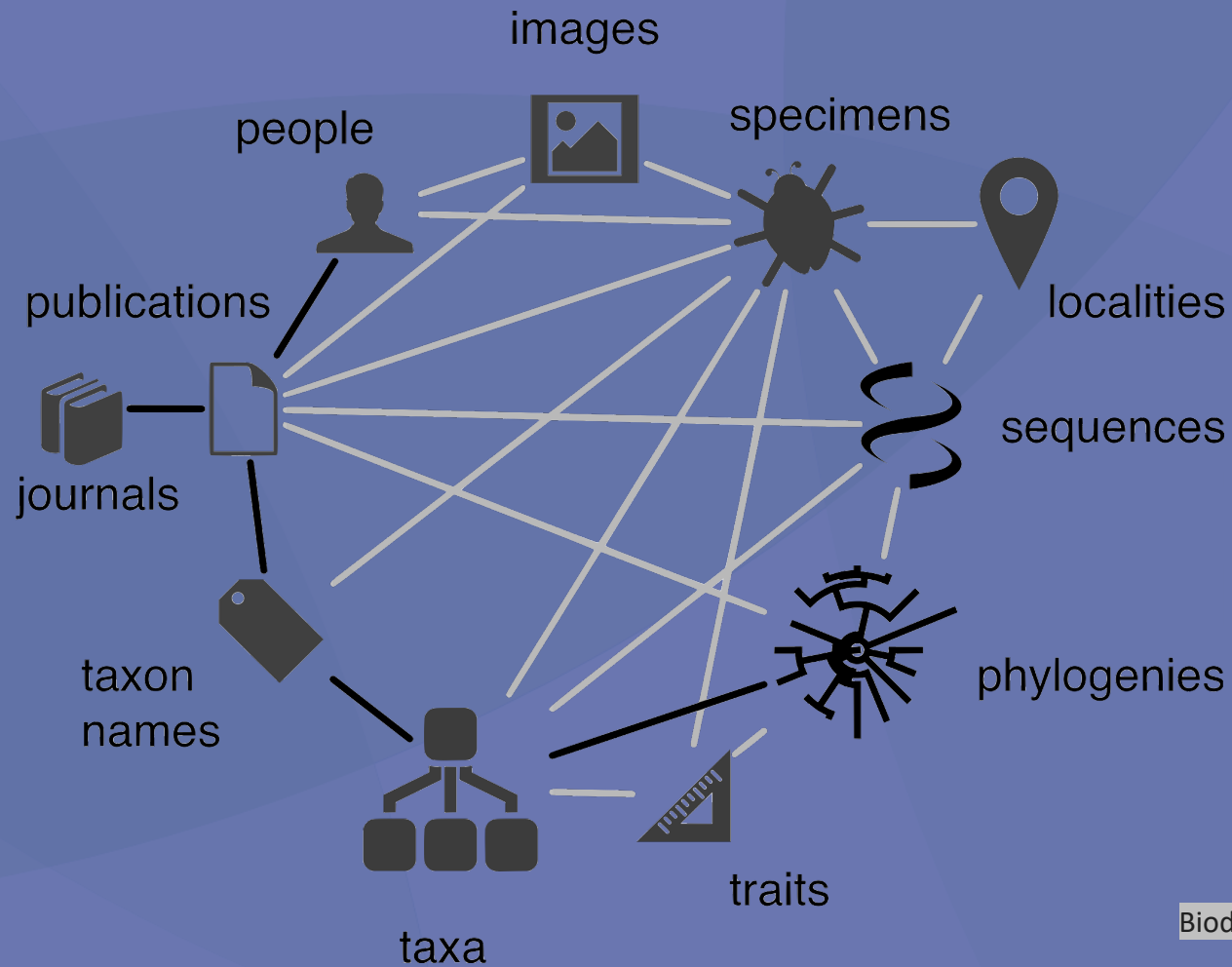
XKCD CC BY-NC 2.5
<http://xkcd.com/927/>

Important principle: *Re-use of terms from standardized terminologies wherever possible.*



Biodiversity Semantics

BIODIVERSITY KNOWLEDGE REPRESENTATIONS



SPECIES OCCURRENCE | COLLECTION SPECIMEN | ENVIRONMENT SAMPLE

Species occurrence



Darwin Core Occurrence, Organism
BFO process

Collection specimen



Darwin Core PreservedSpecimen
BFO material entity
OBI specimen

Environment sample

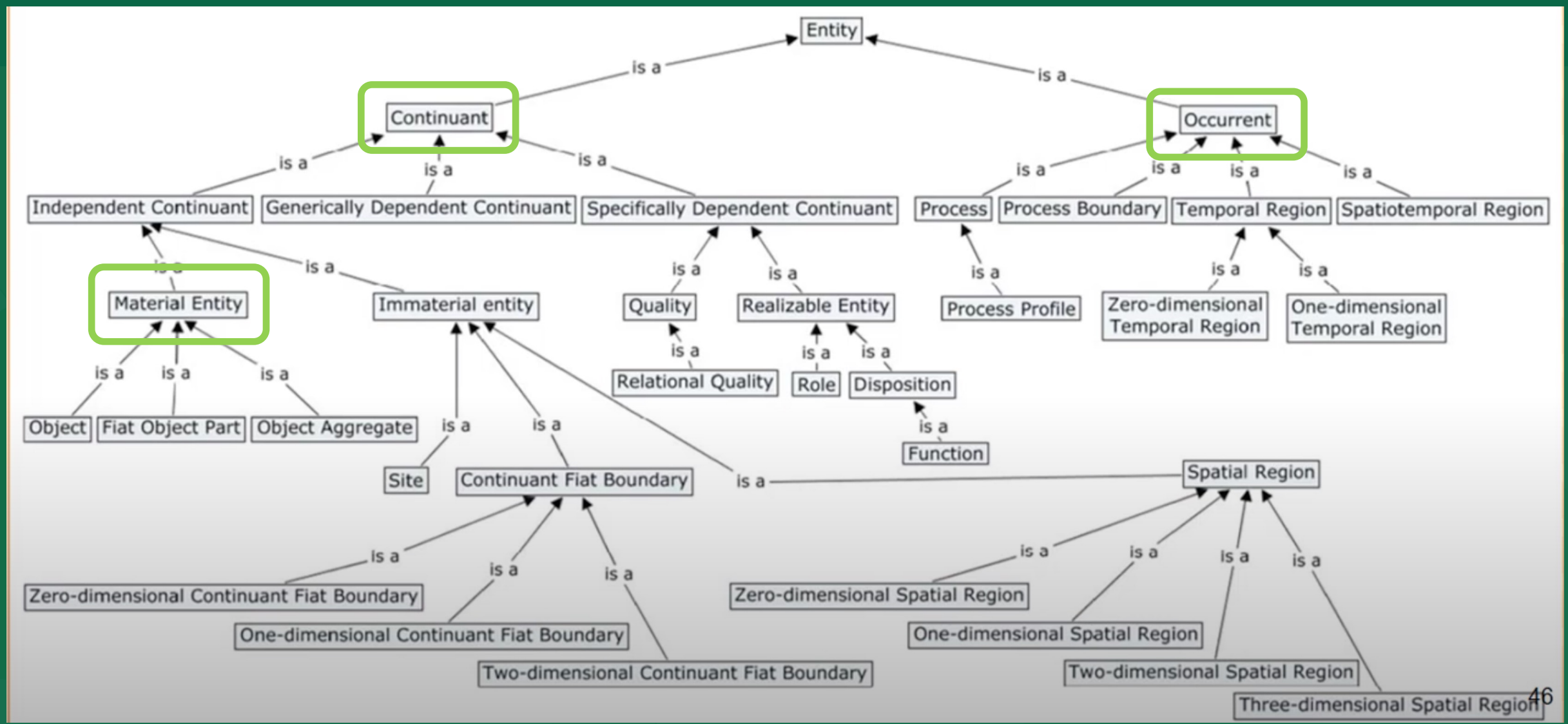


Darwin Core MaterialSample
BFO material entity

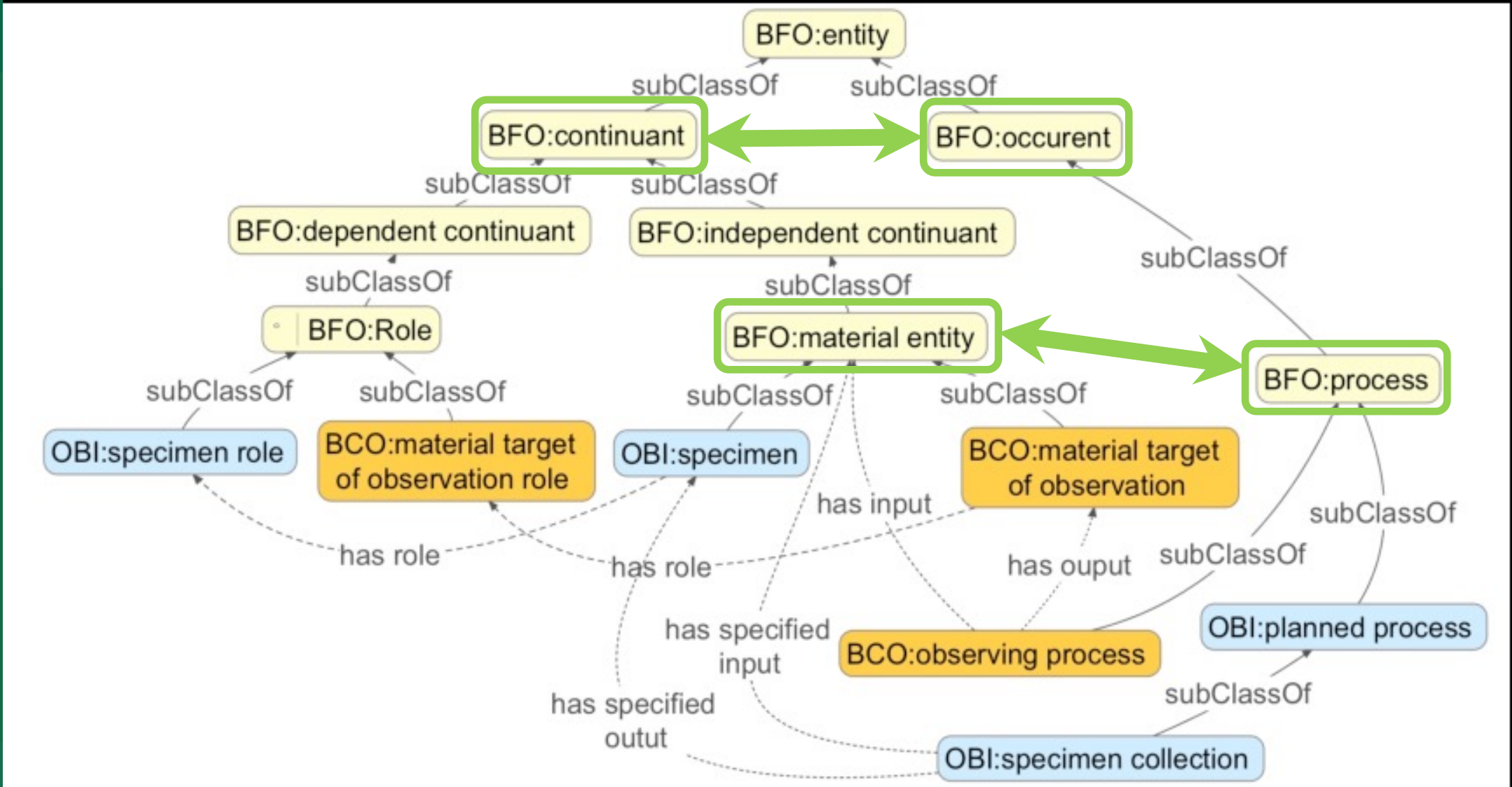
LINKING SEMANTIC MODELS USING TOP-LEVEL ONTOLOGY (2012)

- USA NSF funded RCN4GSC and BioSciCol project (Deck et al. 2012)
- DwC & MlXS – *Sample* versus *Specimen* versus *Occurrence*.
- RCN4GSC – Track DNA extracts and environment samples.
- BioSciCol – Track collection specimens (and processes acting on them).
- → Definitions linked to **BFO (basic formal ontology)**.
- Using a top-level ontology (here BFO) to unite previously disconnected ontologies
- BFO *continuants* (continuous existence) vs BFO *occurents* (processes unfolding through time)
- Inspired by EnvO (Environment ontology) designed under the BFO & OBO Foundry principles
- Aligned with GGBN (Global Genome Biodiversity Network)
- → Development of **BCO (Bio-Collections Ontology)** (Walls et al. 2014)
- → Adding new class MaterialSample (and materialSampleID term) to Darwin Core (2014)

BFO – BASIC FORMAL ONTOLOGY



BCO – BIOLOGICAL COLLECTIONS ONTOLOGY (WALLS ET AL. 2014)



SPECIES OCCURRENCE | COLLECTION SPECIMEN | ENVIRONMENT SAMPLE

Species occurrence



Darwin Core Occurrence, Organism
BFO process

Collection specimen



Darwin Core PreservedSpecimen
BFO material entity
OBI specimen



SPECIMEN VERSUS SPECIES OCCURRENCE

- *Specimen* is product of a **sampling process** that generates a **BFO material entity** (OBI specimen).
- A species **occurrence** is product of a **BFO process** that generates a **IAO information content entity**.
- Need **instance indentifiers** to track DNA extracts and collection specimens.
- These semantic specifications makes it possible (are required) to declare and/or infer **relations between** the material samples and species occurrences mediated through GBIF.



versus



TDWG MATERIAL (SAMPLE) TASK GROUP (2021-[2023])

- The purpose of this task group is to investigate and make recommendations on current shortcomings in the capacity to share, re-use, compare, and relate **physical objects** to one another and to other concepts, and further integrate with other sources of biodiversity data.
- Our main goal is to **enable physical object data sharing** as part of the TDWG corpus of standards.
- The outcome will be to **provide clear semantics for physical object classes and properties** with possible recommendations for changes to the class associated with some properties.



SPECIES OCCURRENCE | COLLECTION SPECIMEN | ENVIRONMENT SAMPLE



Environment sample

Darwin Core MaterialSample
BFO material entity

DIGITAL SEQUENCE INFORMATION / GENETIC SEQUENCE DATA

GBIF has developed data mobilisation guidelines for DNA-derived data available for reuse by GLIS. Based on **Darwin Core** and **MIxS** data standards. **DNA-detected biodiversity** information mobilization has high strategic priority in GBIF!

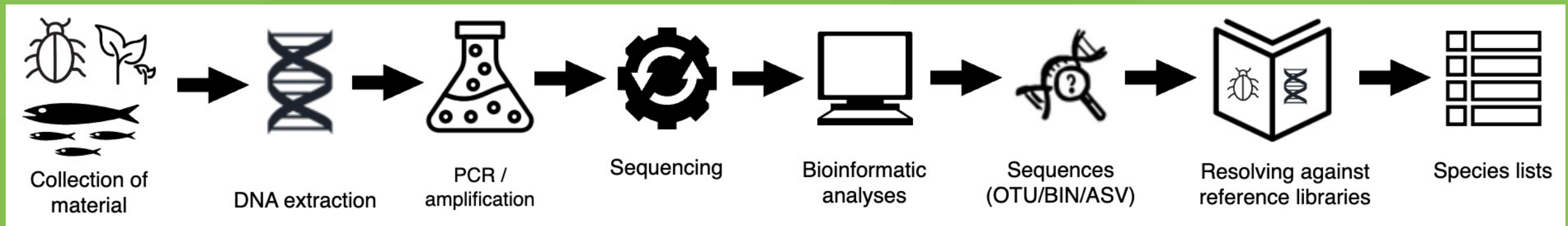
Publishing DNA-derived data through biodiversity data platforms

Anders F. Andersson · Andrew Bissett · Anders G. Finstad · Frode Fossøy · Marie Grosjean · Michael Hope · Thomas S. Jeppesen · Urmas Kõljalg · Daniel Lundin · R. Henrik Nilsson · Maria Prager · Cecilie Svenningsen · Dmitry Schigel – Version 3027b16, 2022-08-05 14:26:56 UTC

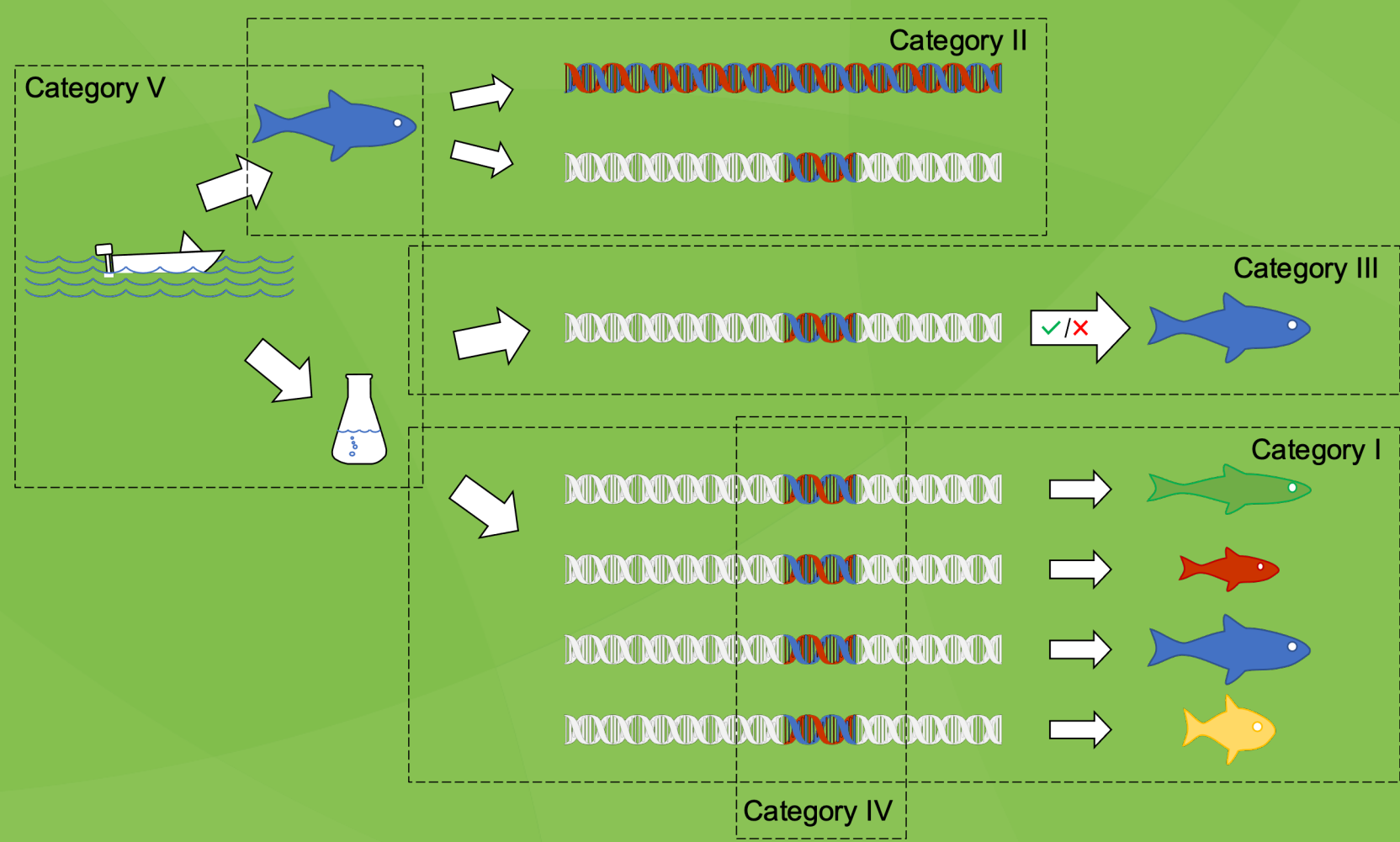
This document is also available in [PDF format](#) and in other languages: [français](#).



METABARCODING ENVIRONMENT DNA SAMPLING PROCESS



METABARCODING – DNA DETECTED SPECIES IDENTIFICATION





Agrobiodiversity data standards

CROP ONTOLOGY

Crop Ontology – is harmonizing semantics for agricultural field data (trait measurement experiments)



Illustration Dag Endresen

CROP DESCRIPTOR LISTS (MORE THAN 100 CROP STANDARDS, SINCE 1970)



Published Descriptor Lists

- Allium (E, S) 2000
- Almond (revised)* (E) 1985
- Apple* (E) 1982
- Apricot* (E) 1984
- Avocado (E, S) 1995
- Bambara groundnut (E, F) 2000
- Banana (E, S, F) 1996
- Barley (E) 1994
- Beta (E) 1991
- Black pepper (E, S) 1995
- Brassica and Raphanus (E) 1990
- Brassica campestris L. (E) 1987
- Buckwheat (E) 1994
- Capsicum* (E, S) 1995
- Cardamom (E) 1994
- Carrot (E, S, F) 1999
- Cashew* (E) 1986
- Chenopodium pallidicaule (S) 2005
- Cherry* (E) 1985
- Chickpea (E) 1993
- Citrus (E,F,S) 1999
- Coconut (E) 1992
- Coffee (E, S, F) 1996
- Cotton (Revised)* (E) 1985
- Cowpea* (E) 1983
- Cultivated potato* (E) 1977
- Echinochloa Millet* (E) 1983
- Eggplant (E, F) 1990
- Faba bean* (E) 1985
- Fig (E) 2003
- Finger millet* (E) 1985
- Forage grass* (E) 1985
- Forage legumes* (E) 1984
- Grapevine (E, S, F) 1997
- Groundnut (E, S, F) 1992
- Jackfruit (E) 2000
- Kodo millet* (E) 1983
- Lathyrus spp. (E) 2000
- Lentil* (E) 1985
- Lima bean* (E) 1982
- Litchi 2002
- Lupin* (E, S) 1981
- Maize (E, S, F, P) 1991
- Mango (Revised) (E) 2006
- Mangosteen (E) 2003
- Medicago (Annual)* (E, F) 1991
- Melon (E) 2003
- Mung bean* (E) 1980
- Oat* (E) 1985
- Oca* (S) 2001
- Oil palm (E) 1989
- Palmier dattier (F) 2005
- Panicum miliaceum and P. sumatrense (E) 1985
- Papaya (E) 1988
- Peach* (E) 1985
- Pear* (E) 1983
- Pearl millet (E, F) 1993
- Pepino (E) 2004
- Phaseolus acutifolius (E) 1985
- Phaseolus coccineus* (E) 1983
- Phaseolus lunatus (P) 2001
- Phaseolus vulgaris* (E, P) 1982
- Pigeonpea (E) 1993
- Pineapple (E) 1991
- Pistacia (excluding P. vera) (E) 1998
- Pistachio (E, F, A, R) 1997
- Plum* (E) 1985
- Potato varieties* (E) 1985
- Quinoa* (S) 1981
- Rambutan (E) 2003
- Rice* (E) 2006
- Rocket (E,I) 1999
- Rye and Triticale* (E) 1985
- Safflower* (E) 1983
- Sesame* (E) 2004
- Setaria italica and S. pumila (E) 1985
- Sorghum (E, F) 1993
- Soyabean* (E, C) 1984
- Strawberry (E) 1986
- Sunflower* (E) 1985
- Sweet potato (E, S, F) 1991
- Taro (E, F, S) 1999
- Tea (E, S, F) 1997
- Tomato (E, S, F) 1996
- Tropical fruit* (E) 1980
- Ulluco (S) 2003
- Vigna aconitifolia and V. trilobata (E) 1985
- Vigna mungo and V. radiata (Rev.)* (E) 1985
- Walnut (E) 1994
- Wheat (Revised)* (E) 1985
- Wheat and Aegilops* (E) 1978
- White Clover (E) 1992
- Winged Bean* (E) 1979
- Xanthosoma* (E) 1989
- Yam (E, S, F) 1997

CROP DESCRIPTORS -- EXAMPLES

- **Plant height** (wheat)
 - measured from ground level to the top of spikes, excluding awns. Average of 5 randomly selected plants
- **Plant Height** (potato)
 - The plant height of the main stem
- **Plant height** (oat)
 - Measure from ground level to the uppermost leaf ligule.
- **Plant height** (*Brassica*)
 - maximum plant height from ground to tip of terminal raceme
- The **same descriptor name** (Plant height).
- But **different definitions**.
- → need clear definitions
- → linked through clear semantics



CROP DESCRIPTORS -- EXAMPLES

- **Plant height** (wheat) → **CO_321:0000020**
 - measured from ground level to the top of spikes, excluding awns. Average of 5 randomly selected plants
- **Plant Height** (potato) → **CO_330:0000756**
 - The plant height of the main stem
- **Plant height** (oat) → **CO_350:0000233**
 - Measure from ground level to the uppermost leaf ligule.
- **Plant height** (*Brassica*) → **CO_348:2100125**
 - maximum plant height from ground to tip of terminal raceme



- clear definitions
 - linked through clear semantics
-
- https://cropontology.org/term/CO_321:0000020
 - https://cropontology.org/term/CO_330:0000756
 - https://cropontology.org/term/CO_350:0000233
 - https://cropontology.org/term/CO_348:2100125

IN SITU CWR DATA STREAMS ARE ALREADY FLOWING THROUGH GBIF

The Plant Treaty, PGRFA-, and GBIF experts (including national nodes) can work together on the **implementation of the descriptors for CWR** conservation *in situ*.

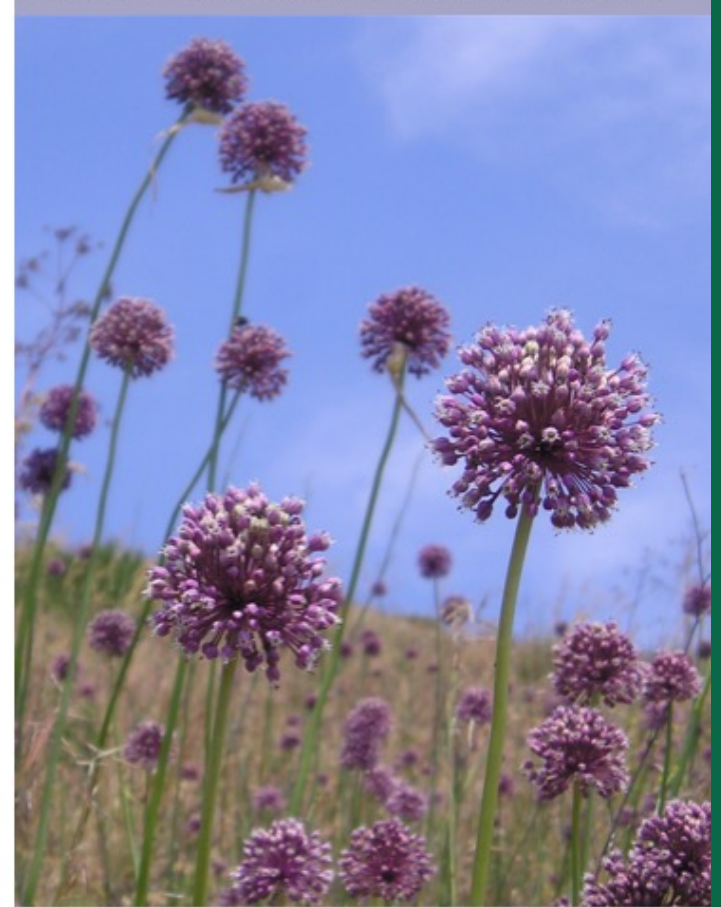
Including **harmonization with Darwin Core** and the development of extensions for data publishing services.

Many crop wild relatives conserved *in situ* are already an **integrated part of the GBIF** portal. (CWRs are an estimated 21% of the flora).



Core descriptors for *in situ* conservation of crop wild relatives v. 1

Imke Thormann, Adriana Alercia, Mohammad Ehsan Dulloo



DESCRIPTORS FOR CROP WILD REALTIVES CONSERVED IN SITU

Alercia, A., López, F., Marsella, M., and Cerutti, A.L. 2021. Descriptors for Crop Wild Relatives conserved in situ (CWRI v.1). International Treaty on Plant Genetic Resources for Food and Agriculture. Rome. FAO.

<https://doi.org/10.4060/cb3256en>

→ Under development: CWR ontology



DESCRIPTORS FOR
**Crop Wild
Relatives**
conserved
in situ

BIODIVERSITY DIGITAL TWIN (BIODT.EU)

GBIF and GBIF Norway participate in the Horizon Europe development of digital twins for biodiversity (2022-2025).

*Including a **digital twin for the crop wild relative genepool** of grasspea (*Lathyrus*).*

*Including a **digital twin for DNA-detected biodiversity** in poorly known habitats.*

*Including development of **interoperability between community data standards** – through ontologies.*



THANK YOU

Dag Endresen

GBIF Node Manager for Norway

UiO Natural History Museum



www.gbif.org