Update on the forthcoming GBIF guide to publishing sequence-based biodiversity data

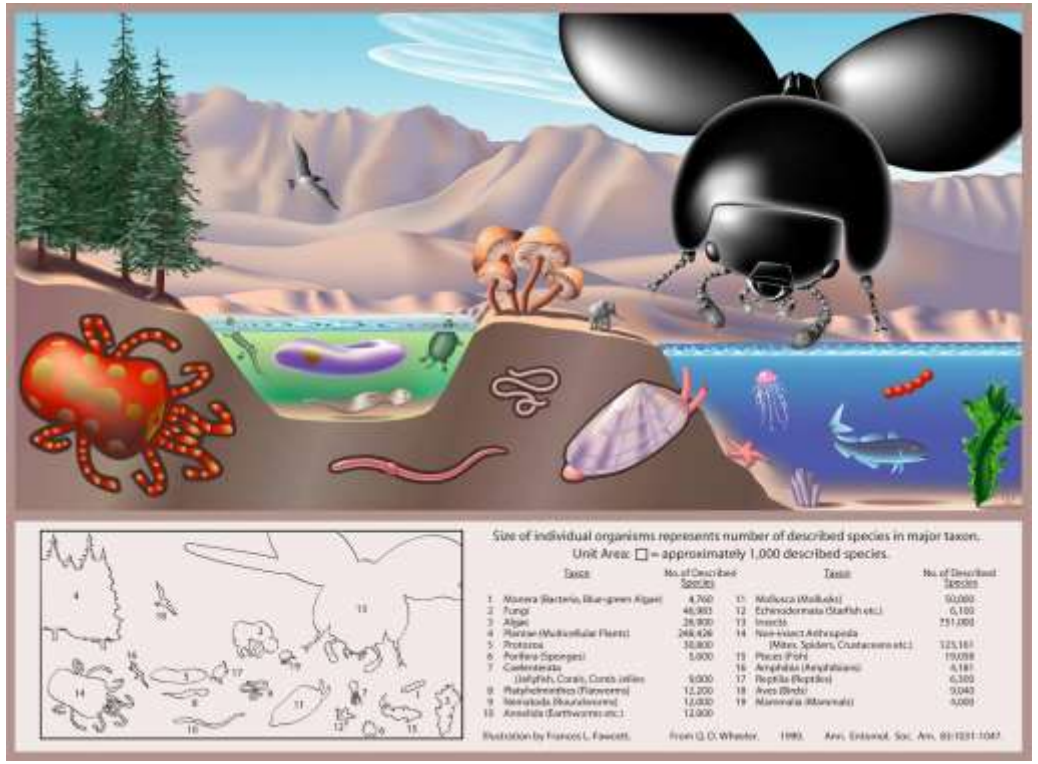
Dmitry Schigel | Scientific officer

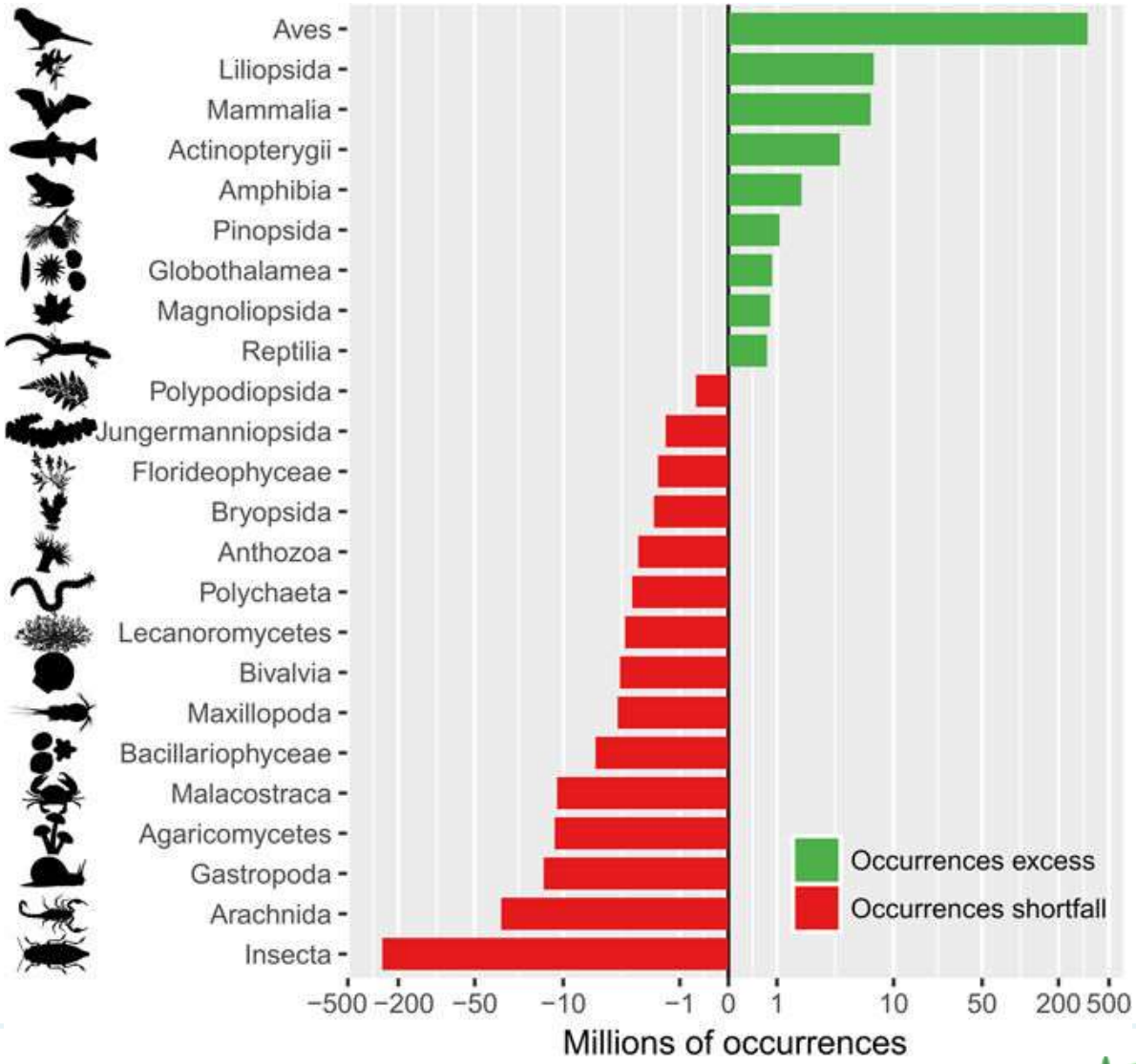
GBIF Global Biodiversity Information Facility

25 June 2020

GLOBAL BIODIVERSITY VS. DIGITALLY AVAILABLE DATA

Image: FL Fawcett in Wheller Ann. Entomol. Soc. Am. 1990





Troudet et al. Nature Scientific Reports 2017

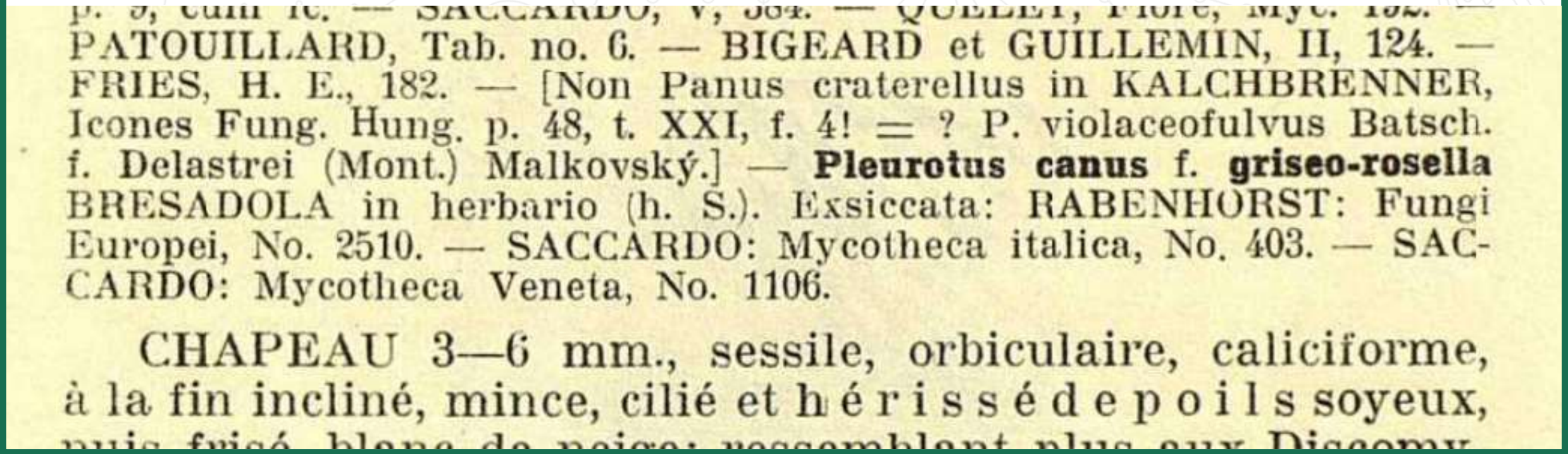
GBIF

CRYPTIC BIODIVERSITY IS UNDERREPRESENTED IN GBIF

- **Detection** difficulties: size, habitats, methods
  - Only emerging in natural history outside academia
- **Identification** difficulties: Collin's guide won't do!
  - Barcoding and molecular identification techniques
- **Attention** issues: few citizen science projects, lack of popularization, no flowers & feathers
  - Molecular citizen science, microbial social media
- **Isolation** of professional communities
  - Better interlinking happens as we speak
- **Technical** developments at GBIF.org are new
  - OTU and DNA work line to grow

BEYOND LATIN NAMES

UNITE and BOLD



PATOUILLARD, Tab. no. 6. — BIGEARD et GUILLEMIN, II, 124. — FRIES, H. E., 182. — [Non Panus craterellus in KALCHBRENNER, Icones Fung. Hung. p. 48, t. XXI, f. 4! = ? P. violaceofulvus Batsch. f. Delastrei (Mont.) Malkovský.] — **Pleurotus canus** f. **griseo-rosella** BRESADOLA in herbario (h. S.). Exsiccata: RABENHORST: Fungi Europei, No. 2510. — SACCARDO: Mycotheca italica, No. 403. — SACCARDO: Mycotheca Veneta, No. 1106.

CHAPEAU 3—6 mm., sessile, orbiculaire, caliciforme, à la fin incliné, mince, cilié et hérissé de poils soyeux,

UNITE'S MOLECULAR NON-LINNAEAN NAMES OF FUNGI

Classification

Kingdom Fungi

Phylum Basidiomycota

Class Agaricomycetes

Order Thelephorales

Family Thelephoraceae

Genus *Tomentella* Pers. ex Pat.

Species *Tomentella bryophila* (Pers.) M.J. Larsen, 1974

Synonyms
- *Botrytis bryophila* (Pers.) Sacc., 1885
- *Hypochnus obscuratus* P. Karst., 1896
- *Hypochnus pallidofulvus* (Peck) Burt, 1926
- *Hypochnus subferrugineus* Burt,

- SH002466.07FU
- SH002471.07FU
- SH002492.07FU
- SH002497.07FU
- SH002506.07FU
- SH002549.07FU
- SH002611.07FU
- SH002687.07FU
- SH009944.07FU
- SH036959.07FU
- SH037000.07FU
- SH037005.07FU
- SH037023.07FU

gbif.org

Get data Share Tools Inside GBIF Login

NEWS | 29 AUGUST 2018

Adding sequence-based identifiers to backbone taxonomy reveals 'dark taxa' fungi

Pilot project with northern European researchers enables inclusion of non-Linnaean 'species hypotheses' aimed at advancing scientific understanding of mycology and functional biodiversity

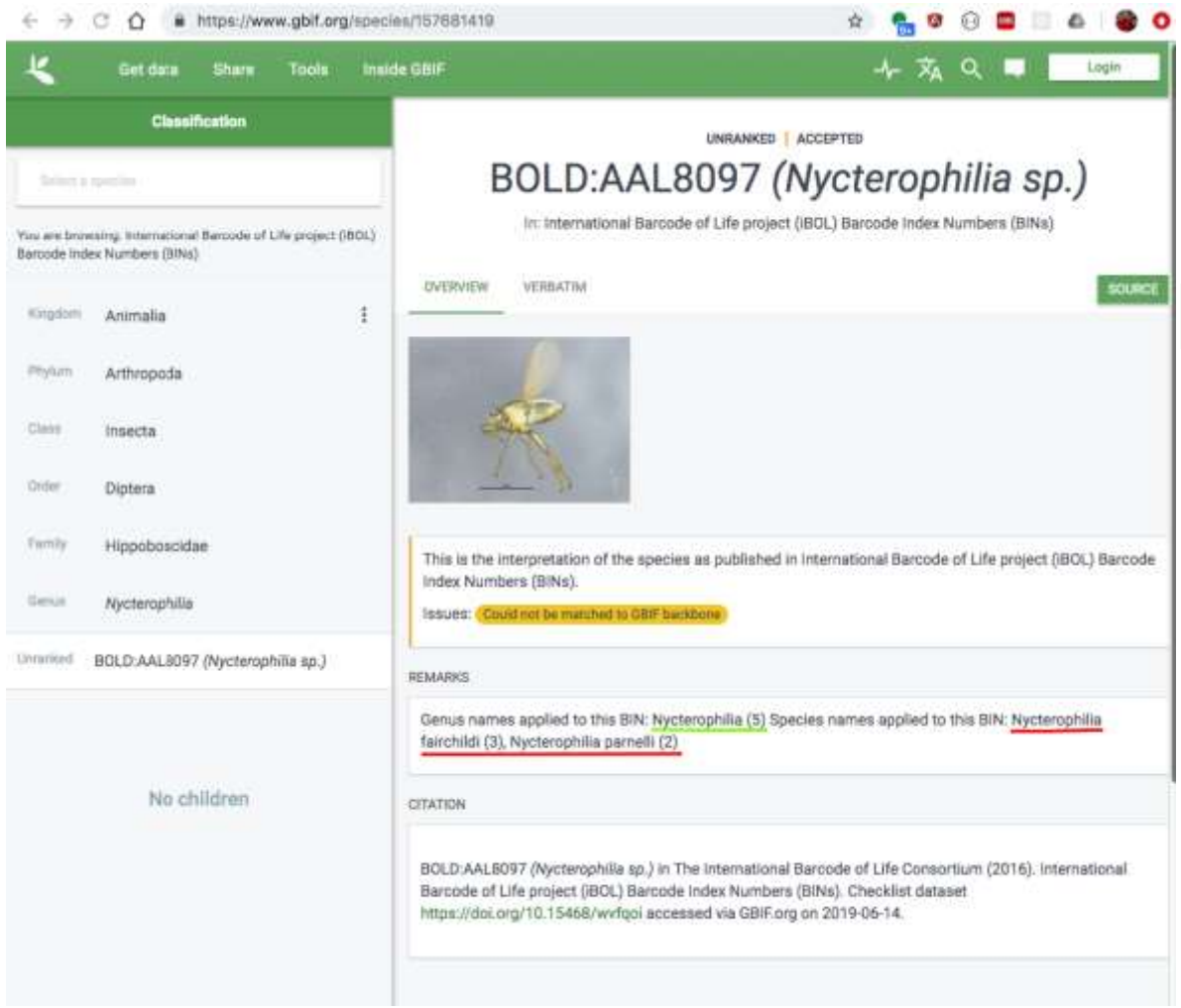
Hygrocybe conica, observed in Trondheim, Norway by Ole Reitan, via Norwegian Species Observation Service. Photo licensed under CC BY 4.0

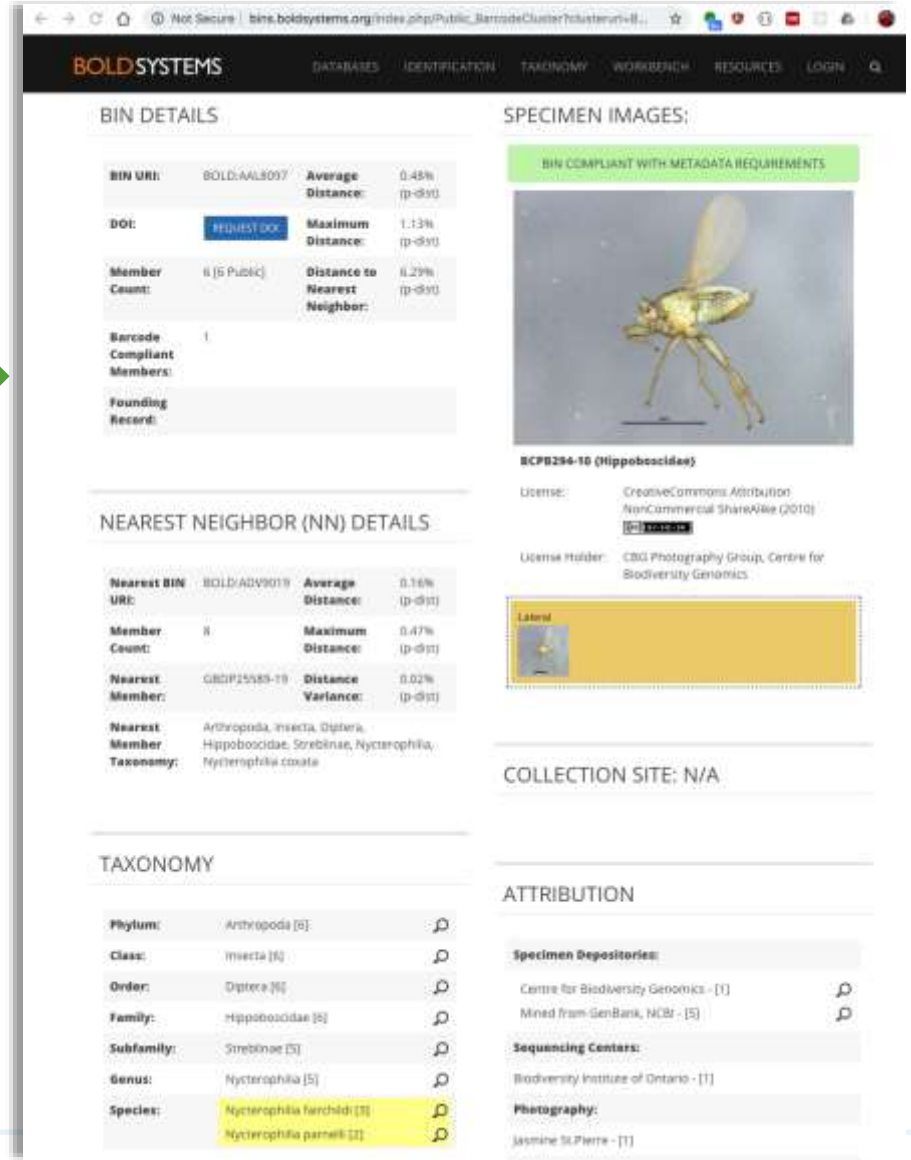
Until a few weeks ago, the GBIF backbone taxonomy fit snugly within a traditional model, classifying and ranking organisms' names using the system that Carl Linnaeus first outlined in *Systema Naturae* in 1735. By combining name-based information from dozens of different authoritative sources like the Catalogue of Life, IRMNG and the World Register of Marine Species (affectionately known as 'WoRMS'), the backbone provides a consistent means of organizing all species-related content on GBIF.org—like

https://www.gbif.org/news/2LrgV5t3ZuGeU2WIymSEuk

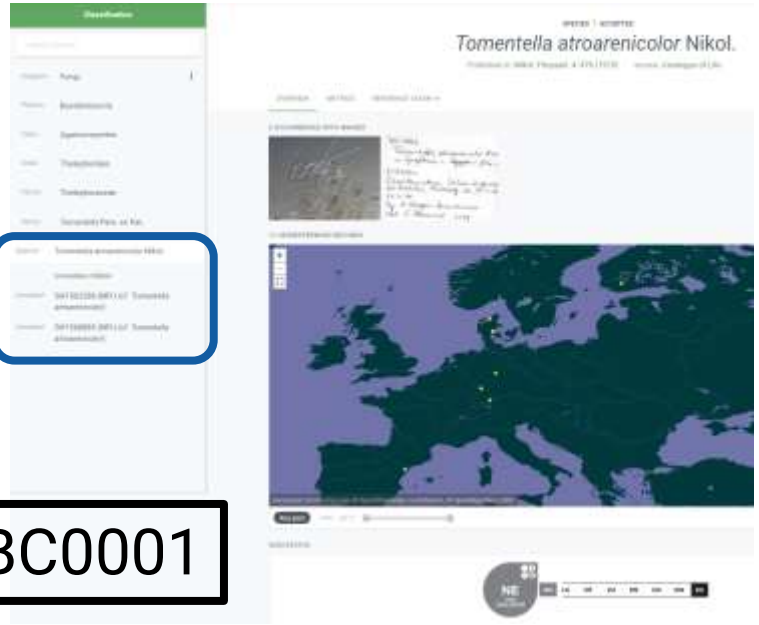
GBIF

BINs IN GBIF AS HANDLES FOR ANIMAL OTUs





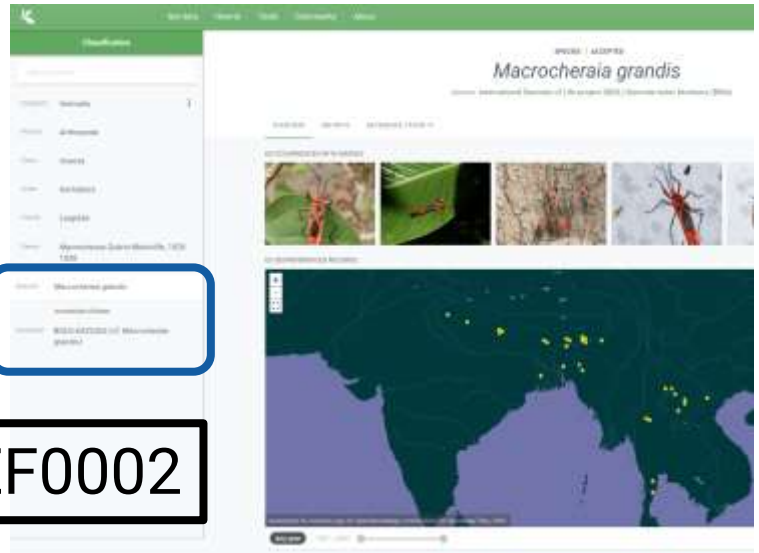
https://doi.org/10.15468/wvfqoi



SH ABC0001

OTU = **SH**,
Species hypothesis

GBIF
backbone taxonomy

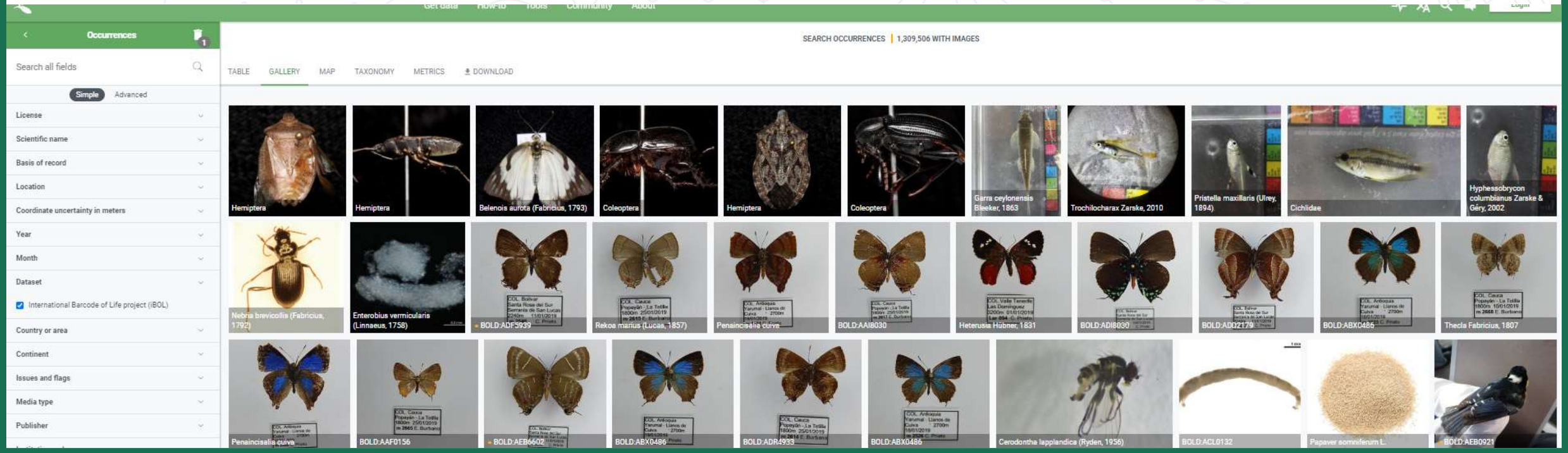


BIN DEF0002

OTU = **BIN**,
Barcode
identification number

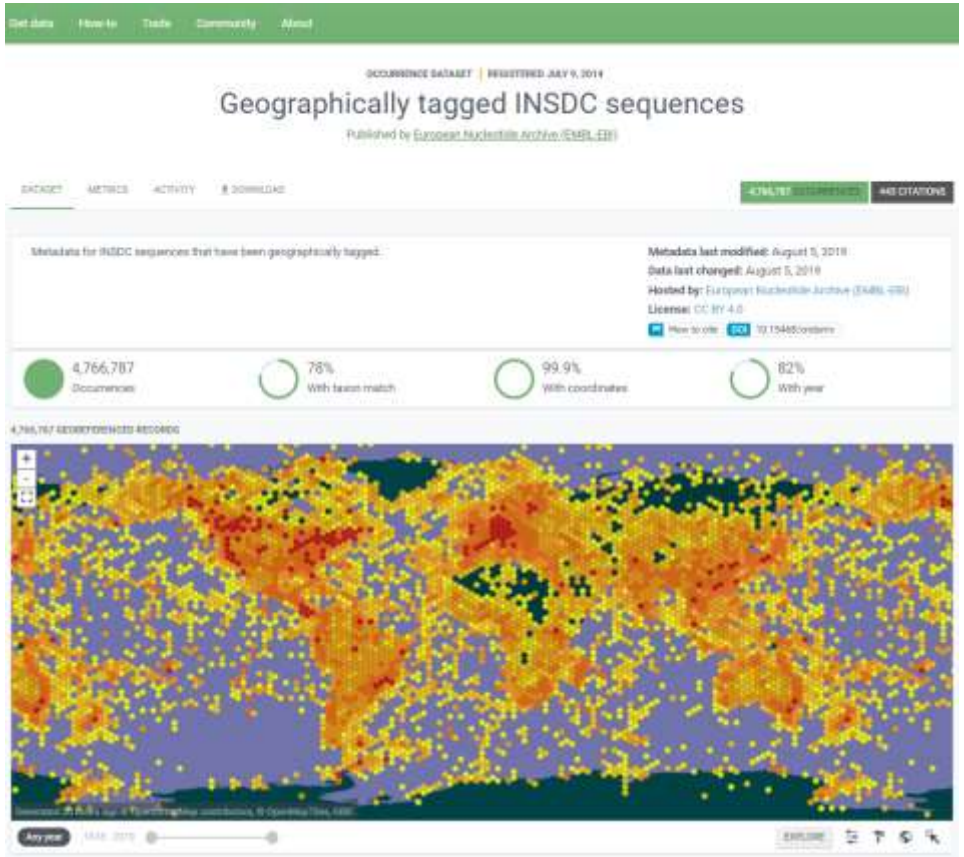
WILD SEQUENCES WITH COORDINATES

EBI: ENA and BOLD

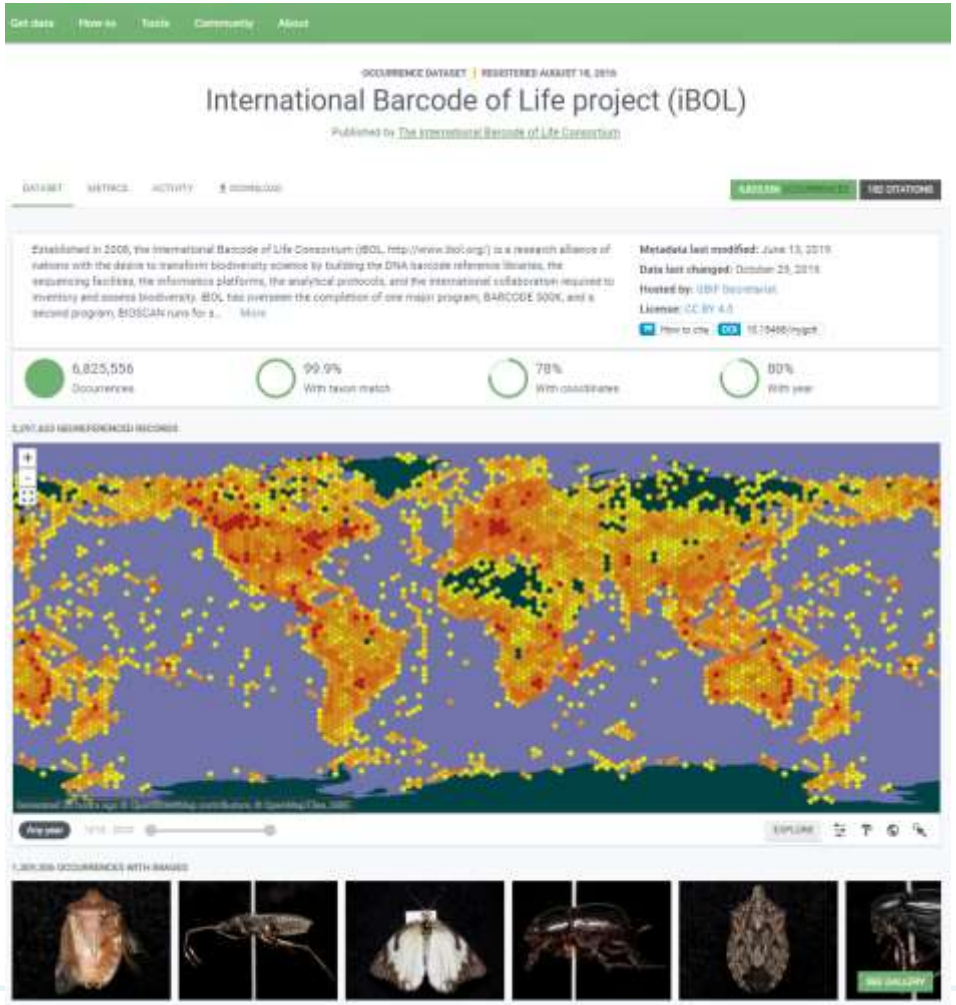


INDIVUDUAL SEQUENCES WITH COORDINATES

European Nucleotide Archive: 4.8M records



International Barcode of Life: 6.8M records

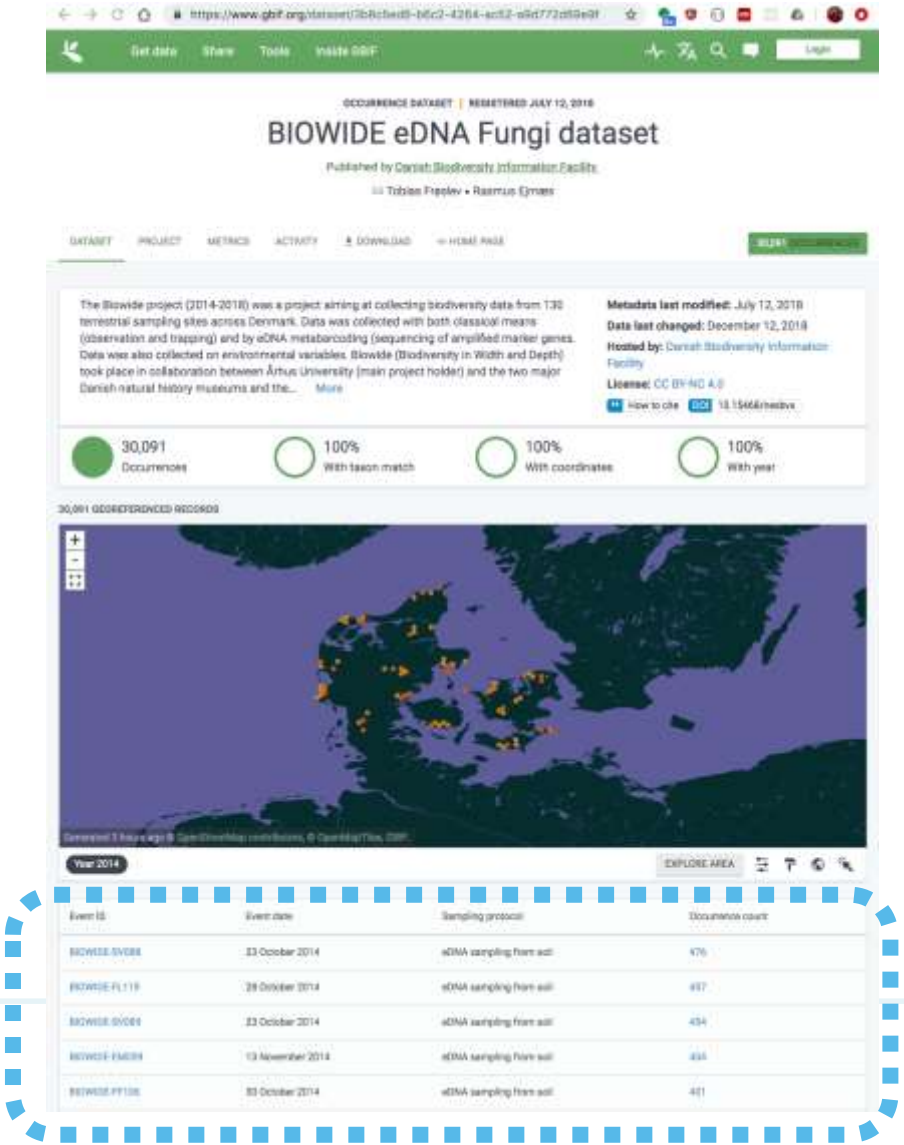


METABARCODING AND METAGENOMICS

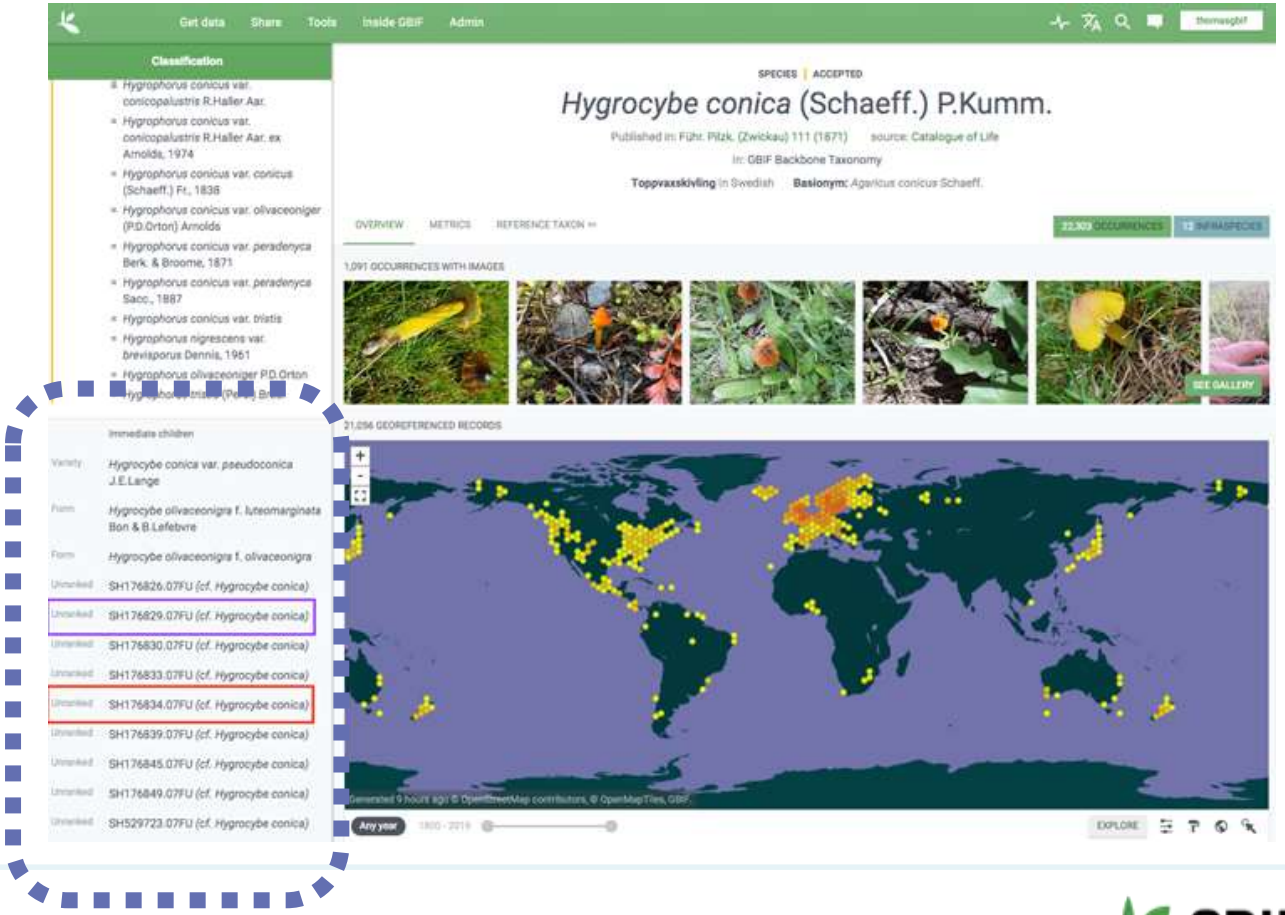
EBI: MGnify; BIOWIDE (Denmark)

PUBLISHING AND INDEXING DATA FROM eDNA SAMPLING EVENTS

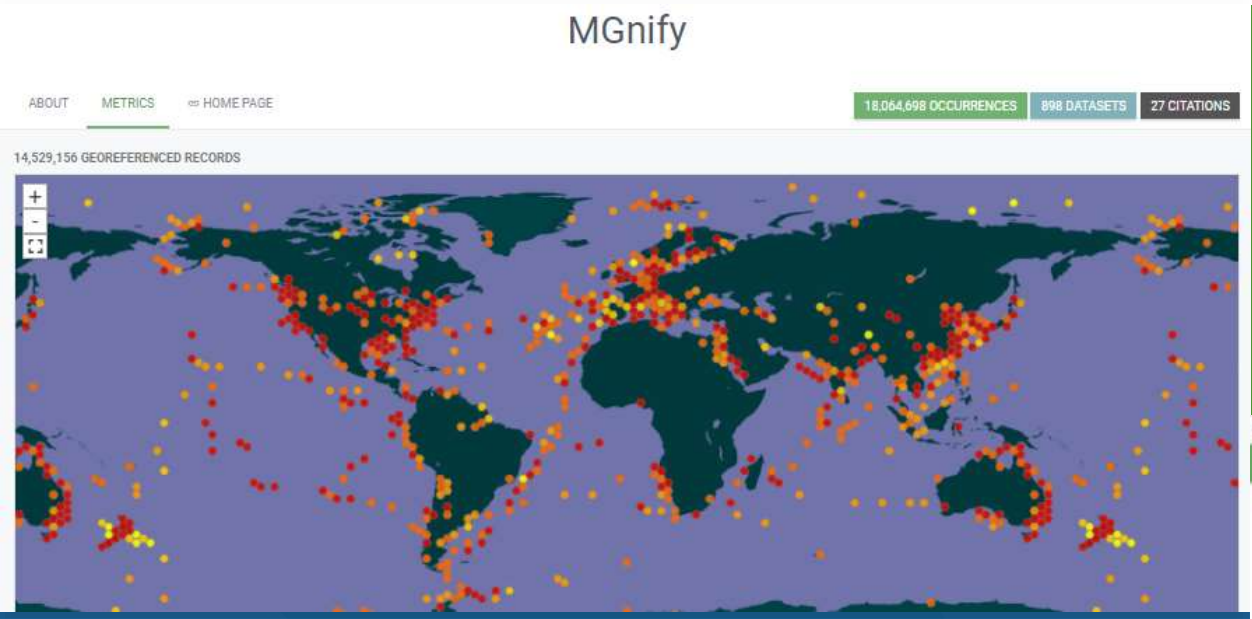
eDNA based dataset



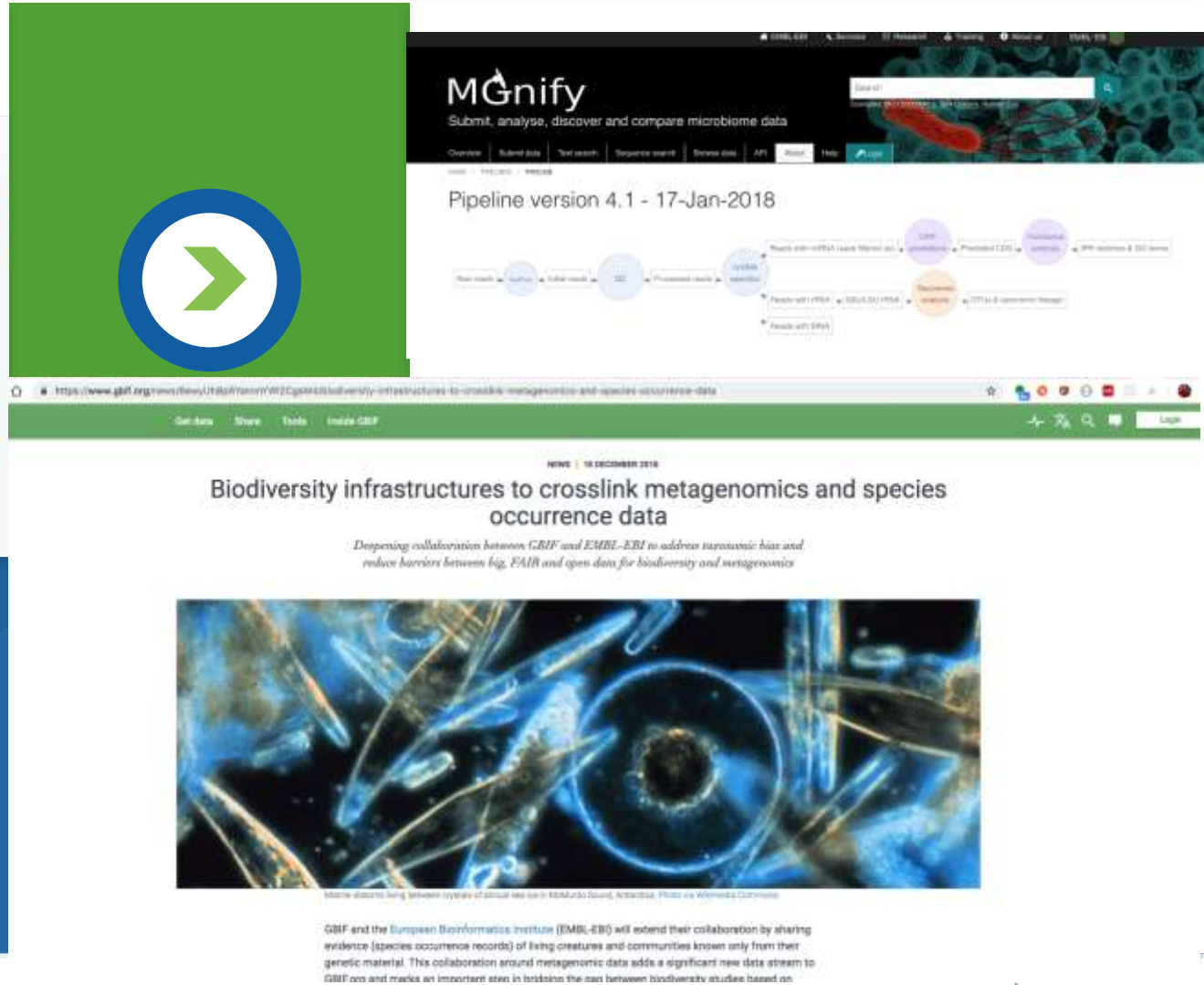
Hybrid Linnaean and OTU based taxonomy



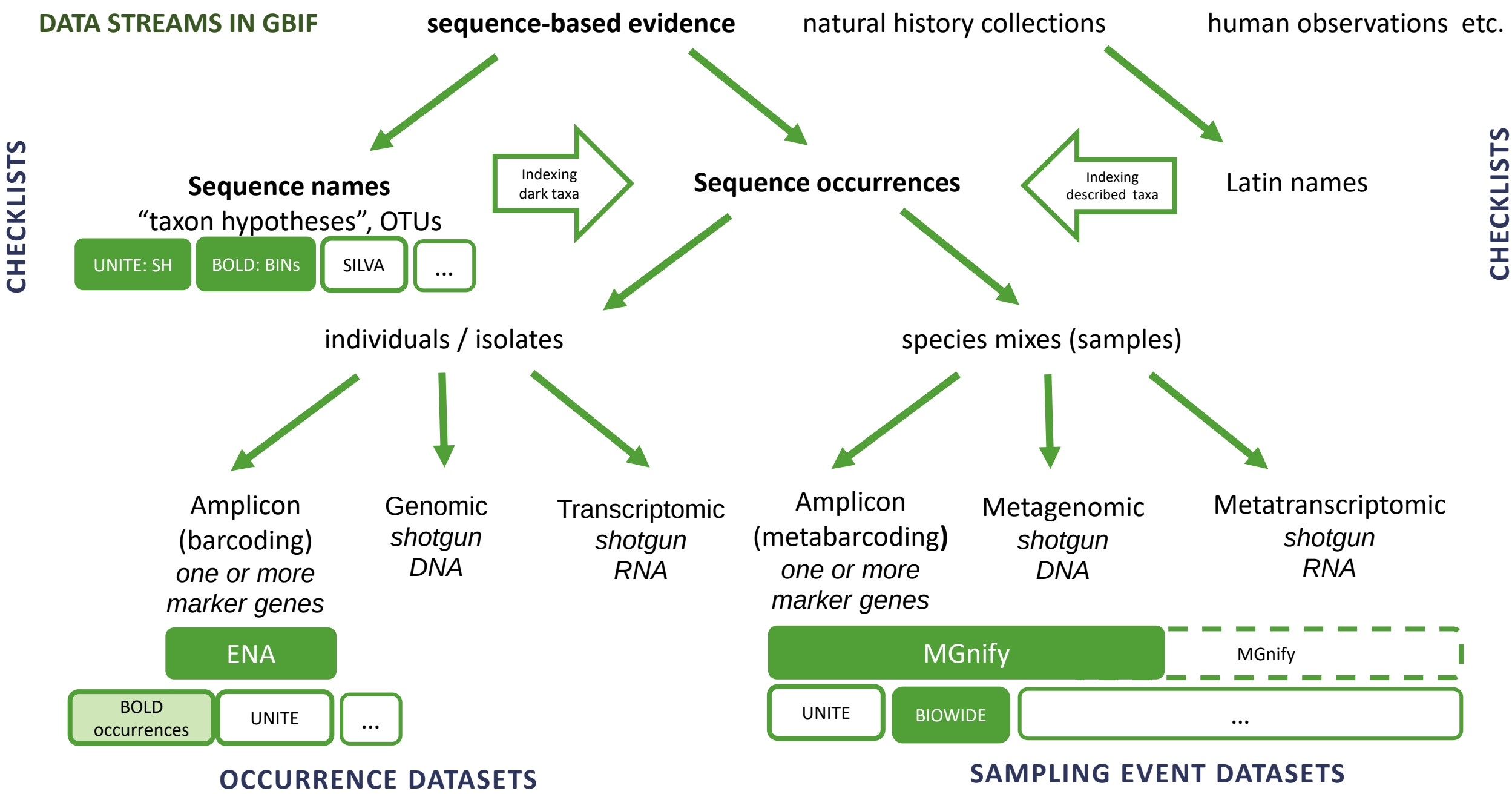
MGNIFY INCREASES COVERAGE OF UNDERREPRESENTED TAXA



898 datasets: 18.1M new records:
13.1M Bacteria
2.2M Chromista

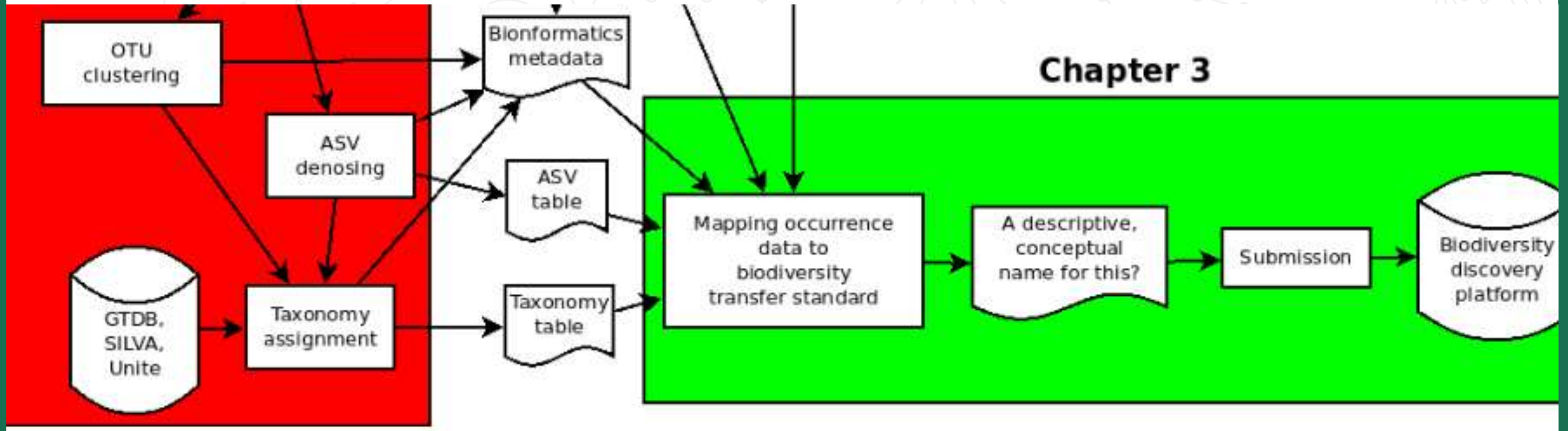


GBIF


DATA STREAMS IN GBIF
sequence-based evidence
natural history collections
human observations etc.
CHECKLISTS
Sequence names
"taxon hypotheses", OTUs
UNITE: SH
BOLD: BINs
SILVA
...
Indexing dark taxa
Sequence occurrences
Indexing described taxa
Latin names
CHECKLISTS
individuals / isolates
species mixes (samples)
Amplicon (barcoding) one or more marker genes
Genomic shotgun DNA
Transcriptomic shotgun RNA
Amplicon (metabarcoding) one or more marker genes
Metagenomic shotgun DNA
Metatranscriptomic shotgun RNA
ENA
BOLD occurrences
UNITE
...
MGnify
MGnify
UNITE
BIOWIDE
...
OCCURRENCE DATASETS
SAMPLING EVENT DATASETS

THE GUIDE AND THE COURSE

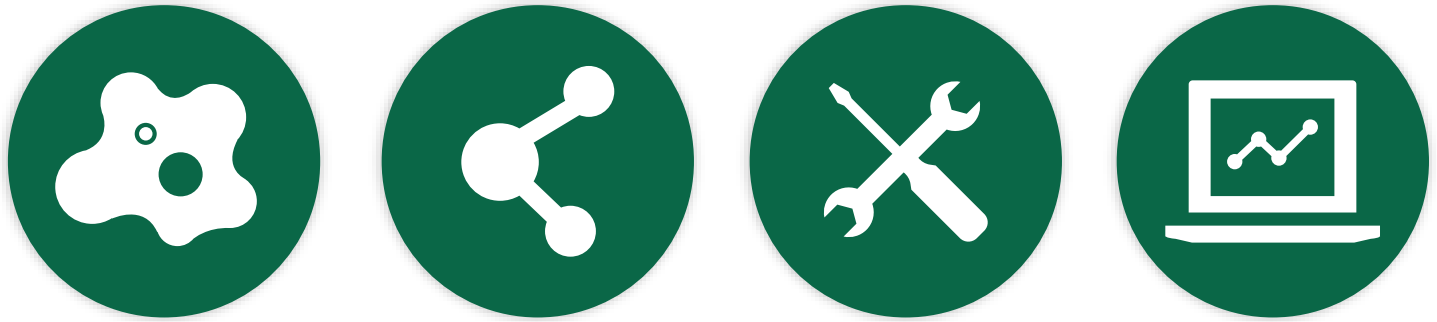
2020 development



NEW DATA PUBLISHING GUIDE COMING IN 2020: KEY FACTS

Publishing sequence-derived data through biodiversity discovery platforms

- 12 authors: Australia, Norway, Sweden, Denmark, UNITE, and GBIF
- Based on practical mapping and data publishing experiences
- Cross-platform
- About 50 pages
  - ❖ Introduction
  - ❖ Categorization
  - ❖ Mapping
  - ❖ Visuals
  - ❖ Future prospects
  - ❖ Resources: glossary, links, references



PUBLISHING SEQUENCE-DERIVED DATA: THE “LEARN” SECTION

2. **Introduction**

    2.1 Rationale

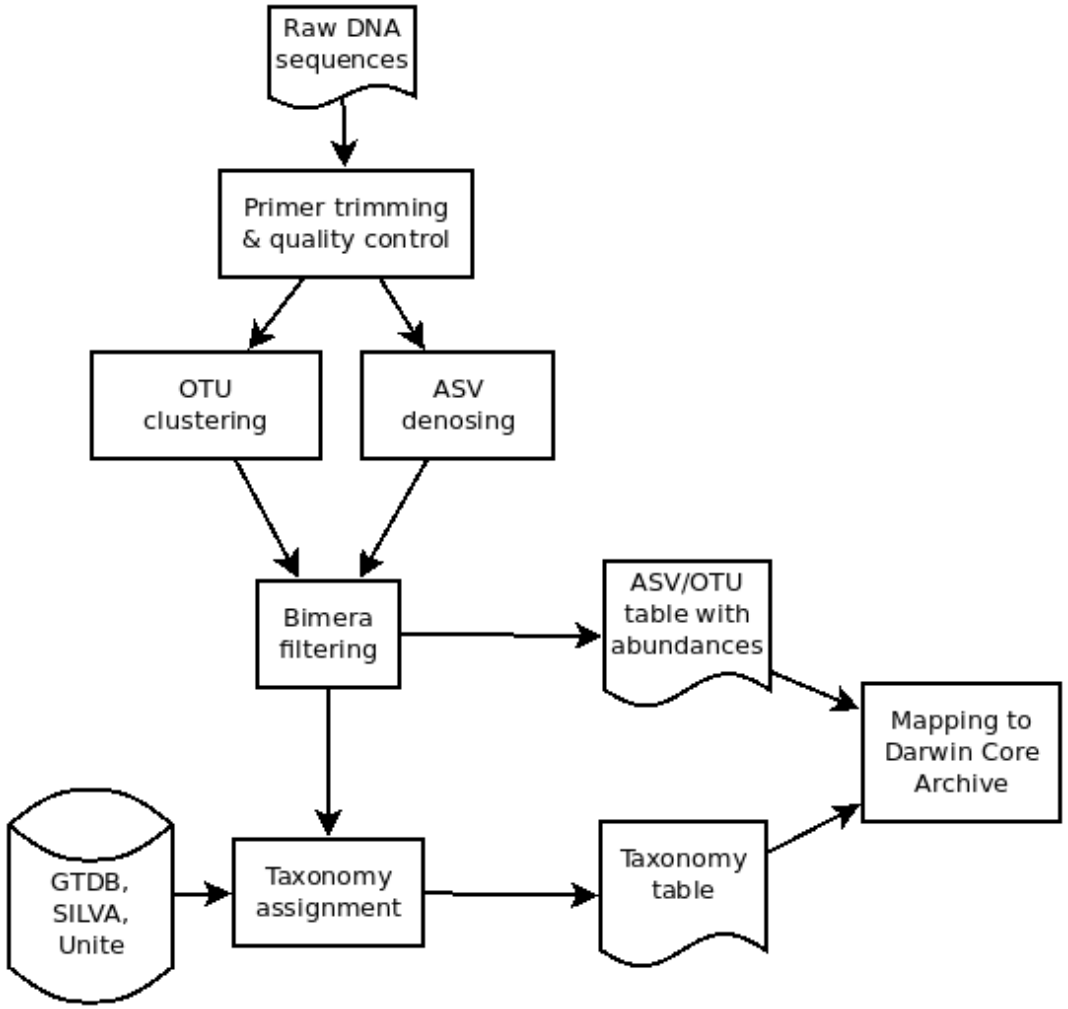
    2.2 Target audiences

    2.3 Introduction to sequence derived occurrence data

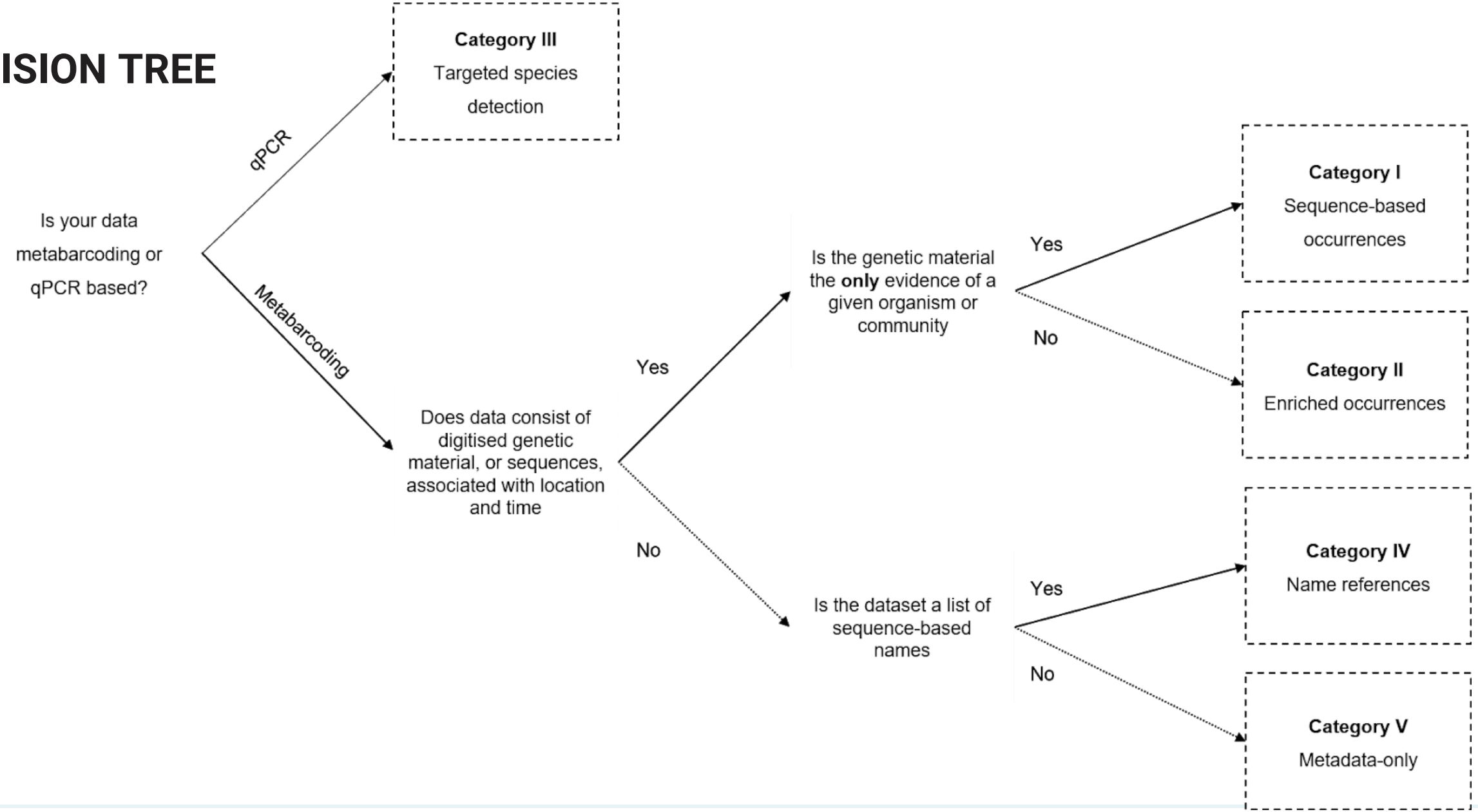
    2.4 Introduction to biodiversity publishing

    2.5 Processing workflows – from sample to ingestible data

    2.6 Taxonomy of sequences



DECISION TREE



PUBLISHING SEQUENCE-DERIVED DATA: THE “DO” SECTION

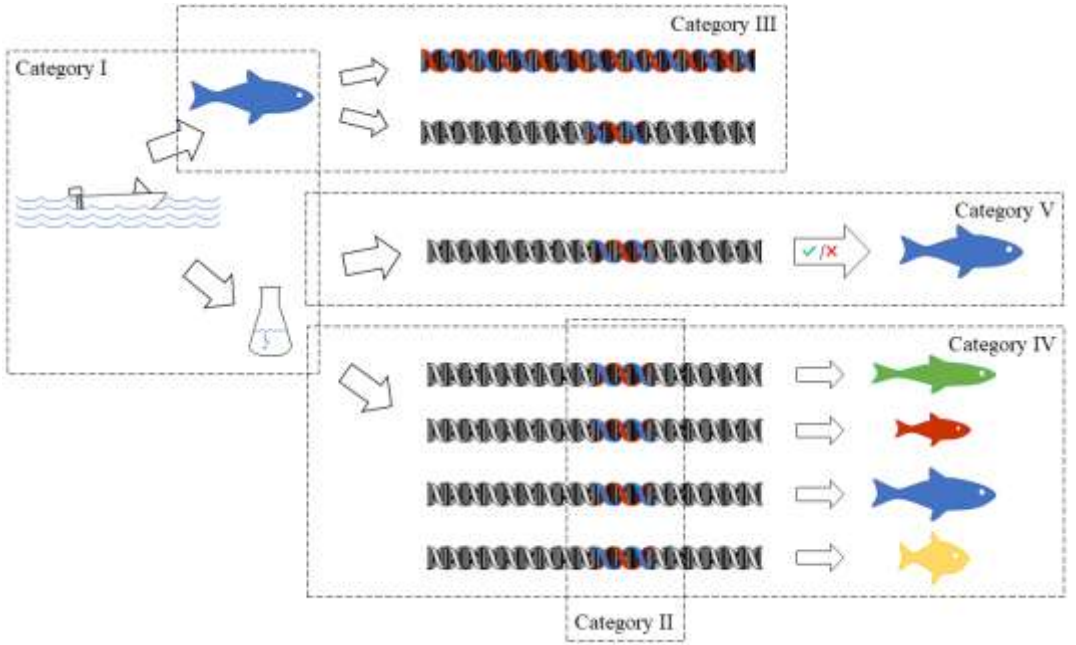
3. Data packaging and mapping

3.1 Categorisation of data - from samples to processed data

- **I** Sequence-based occurrences
- **II** Enriched occurrences
- **III** Targeted species detection (qPCR)
- **IV** Name references
- **V** Metadata-only

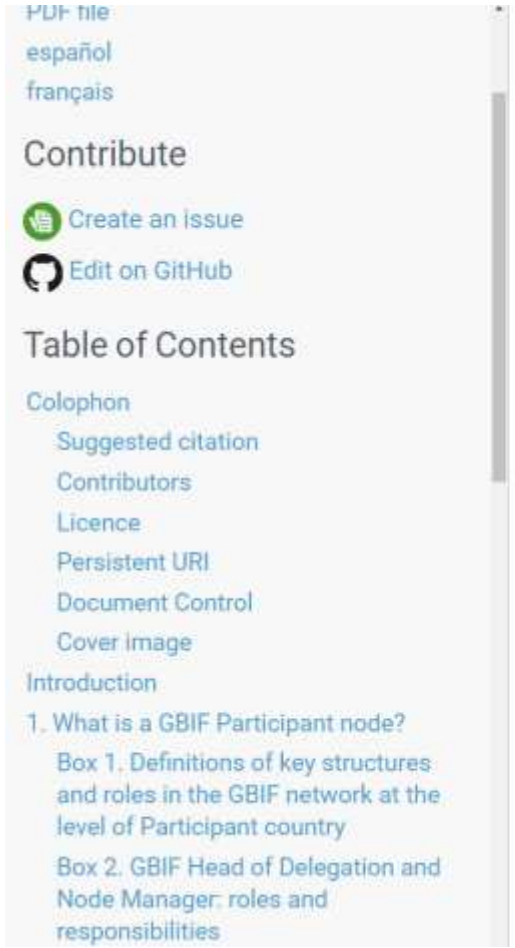
3.2 Mapping recommendations

Examples



TIMELINE AND NEXT STEPS

- Draft closed for edits today
- Copy-editing and formatting
- **Consultation and feedback (July)**
- Online release (autumn)
- Peer-reviewed publication of v 1.0
- Live guide, versioned as needed



NEW HYBRID TRAINING: GBIF, BOLD, UNITE – WITH GBIF NORWAY & BIODATA

Share

Use

New course to pilot in Tbilisi, Georgia in 2020 or 2021
4 days
* * *
Accelerating biodiversity research through
DNA barcodes, collection and observation data

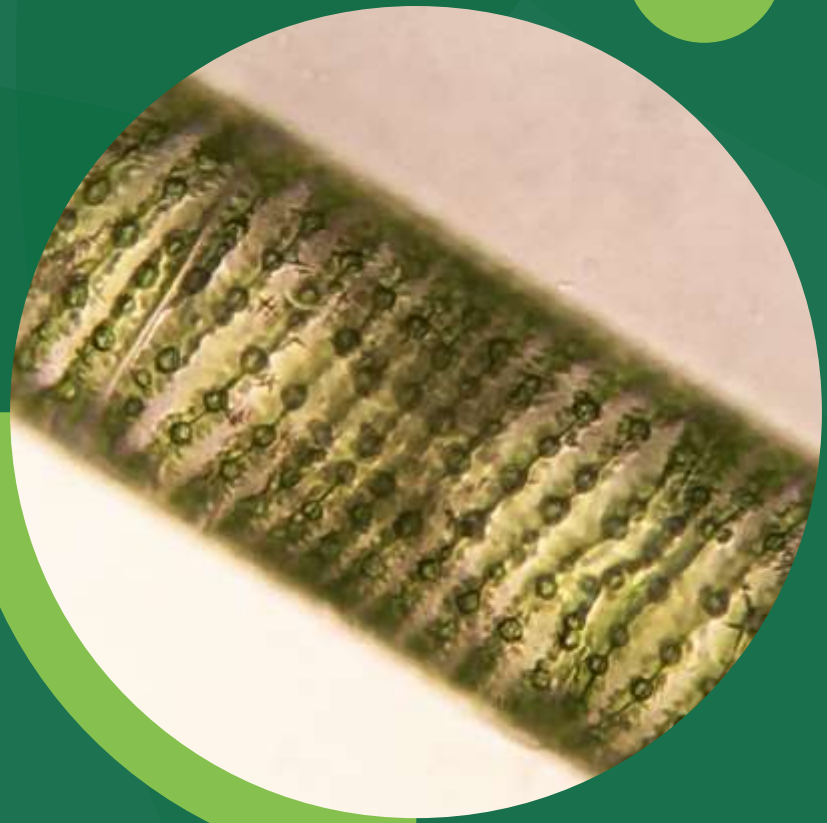
- Publish to and retrieve data from GBIF, BOLD, and UNITE
- Use genetic sequence data as a DNA barcode to identify a species
- Basics of data capture, cleaning, georeferencing, and data citation
- Critically assess data quality and fitness for purpose
- Practice key tools
- FAIR data principles in biodiversity research and collaboration

GBIF

THANK YOU

Dmitry Schigel

dschigel@gbif.org



GBIF