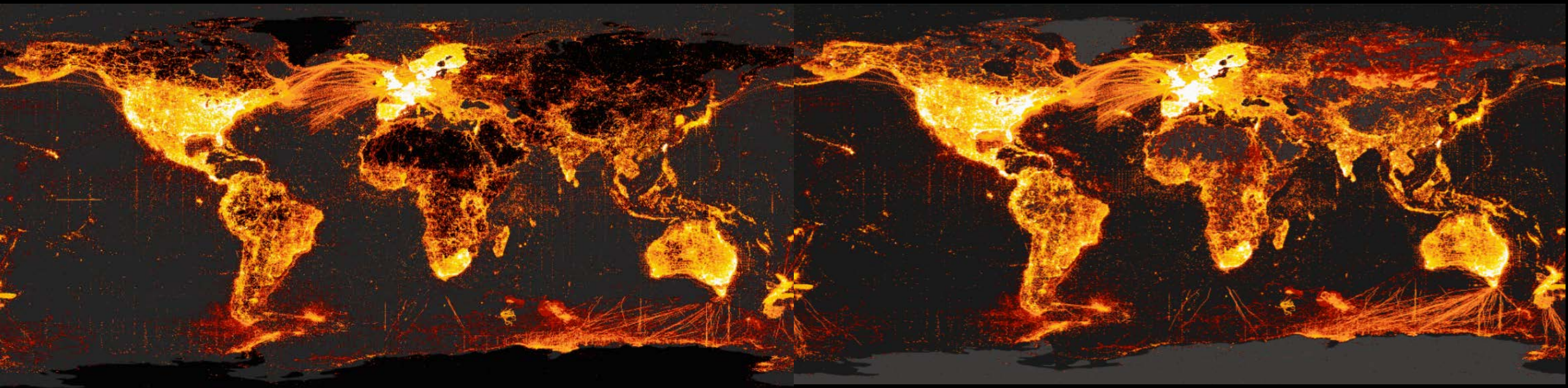
2018 → 2020



981,464,491 → 1,602,089,347

DNA-DERIVED DATA IN GBIF GUIDANCE AND TRAINING

Dmitry Schigel | Scientific officer

GBIF Global Biodiversity Information Facility

October 2020

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

Observations

Digitized specimens

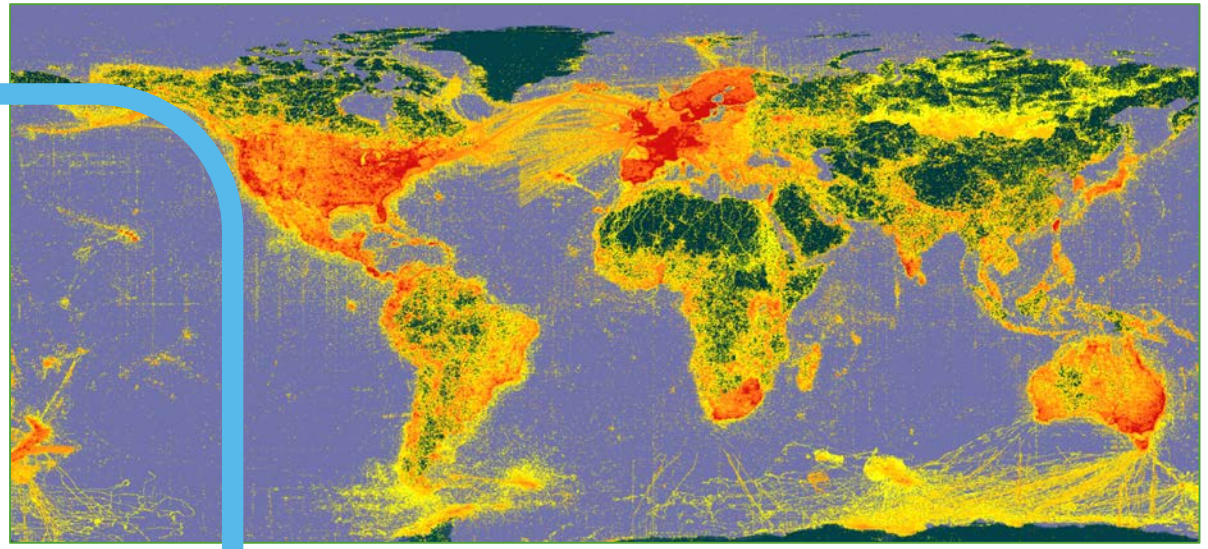
Literature

Data publishing and indexing

Common standards (DwC)

Remote-sensing

Environmental DNA



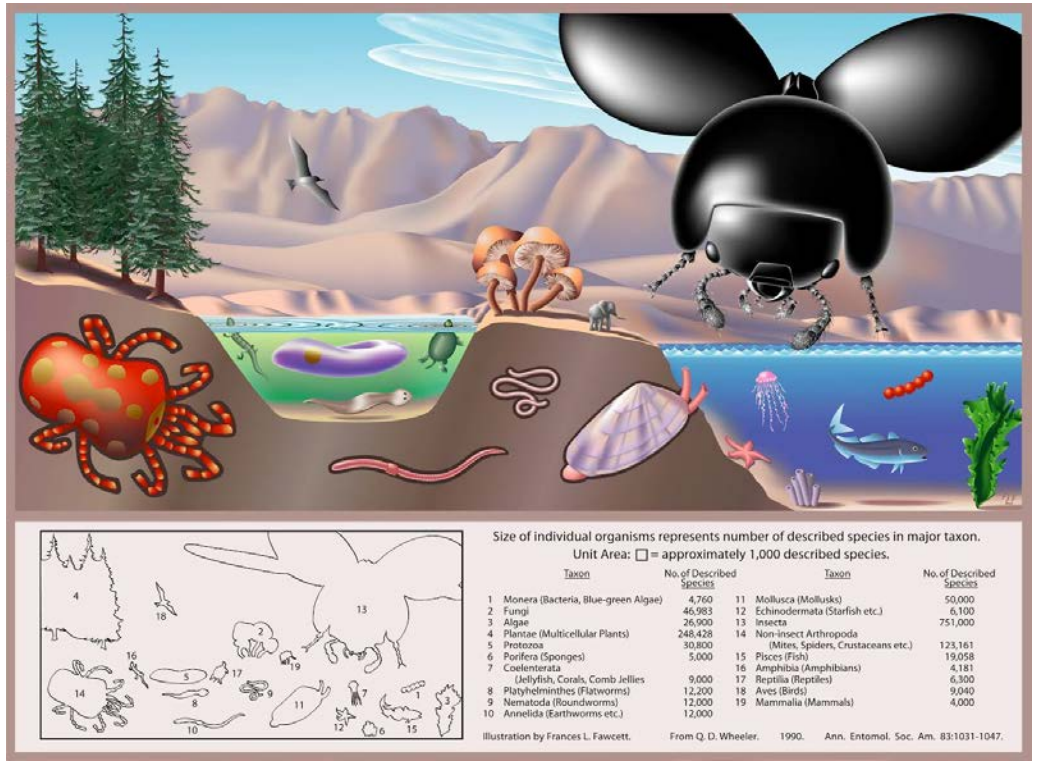
Data discovery and use

https://www.gbif.org/zh/occurrence/search

GBIF

GLOBAL BIODIVERSITY VS. DIGITALLY AVAILABLE DATA

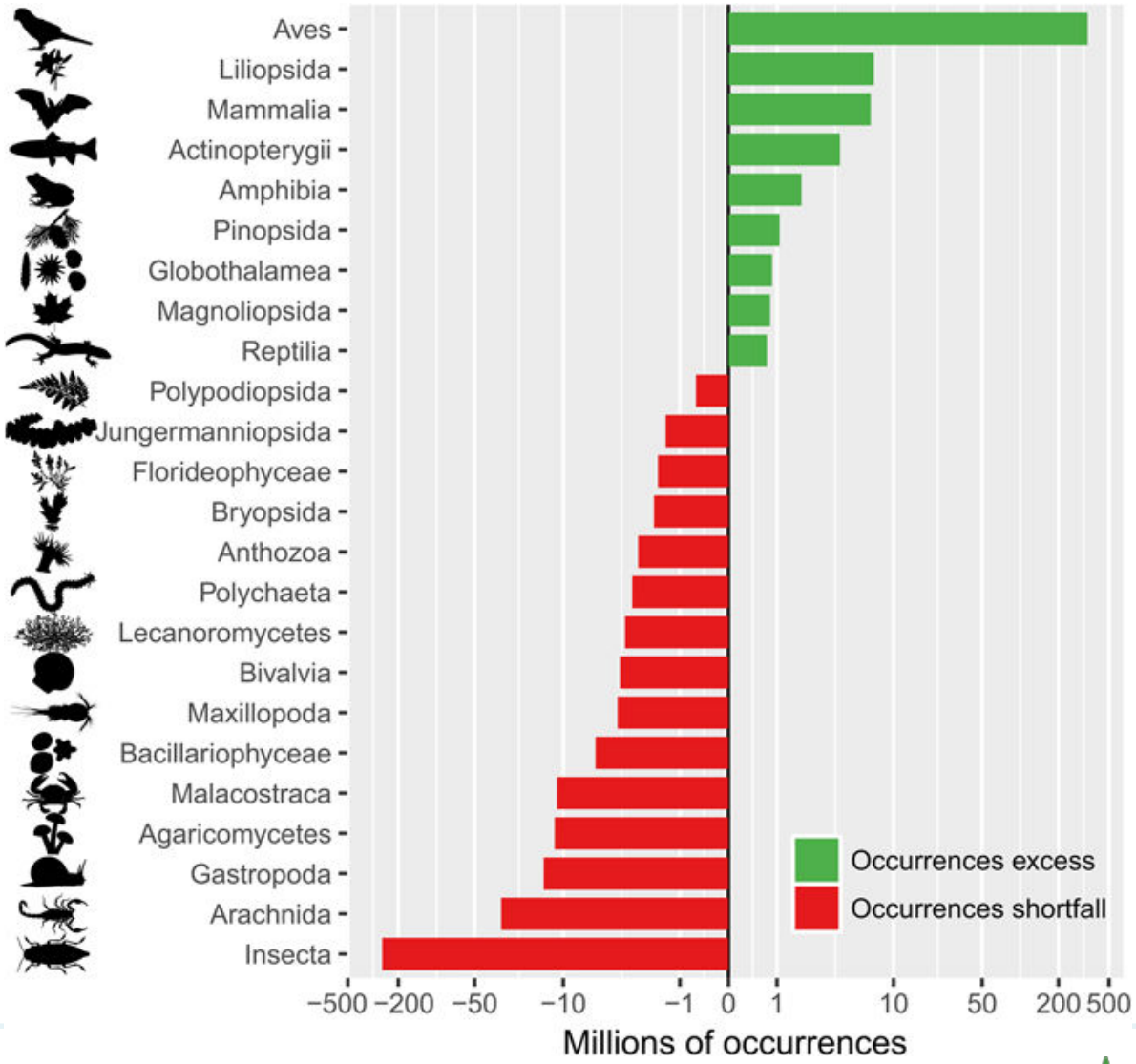
Image: FL Fawcett in Wheller Ann. Entomol. Soc. Am. 1990



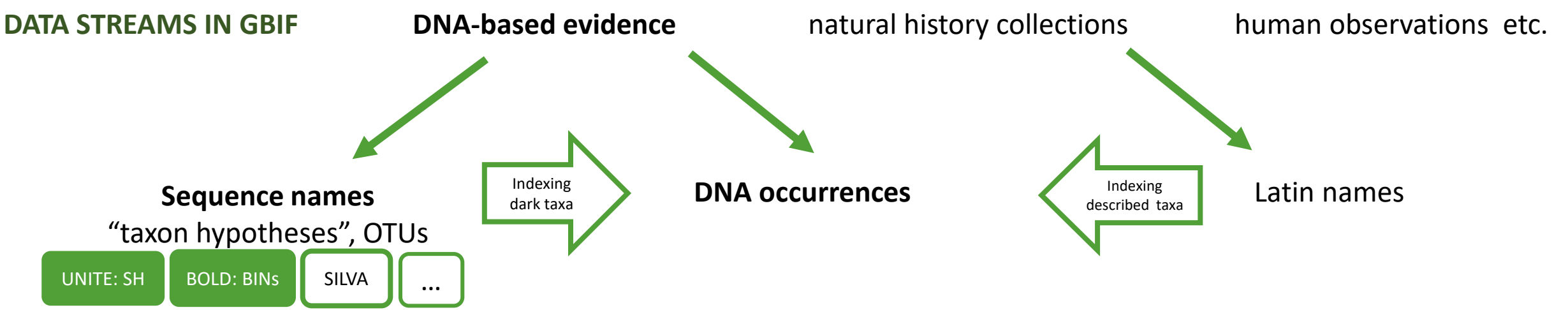
Size of individual organisms represents number of described species in major taxon.
Unit Area: □ = approximately 1,000 described species.

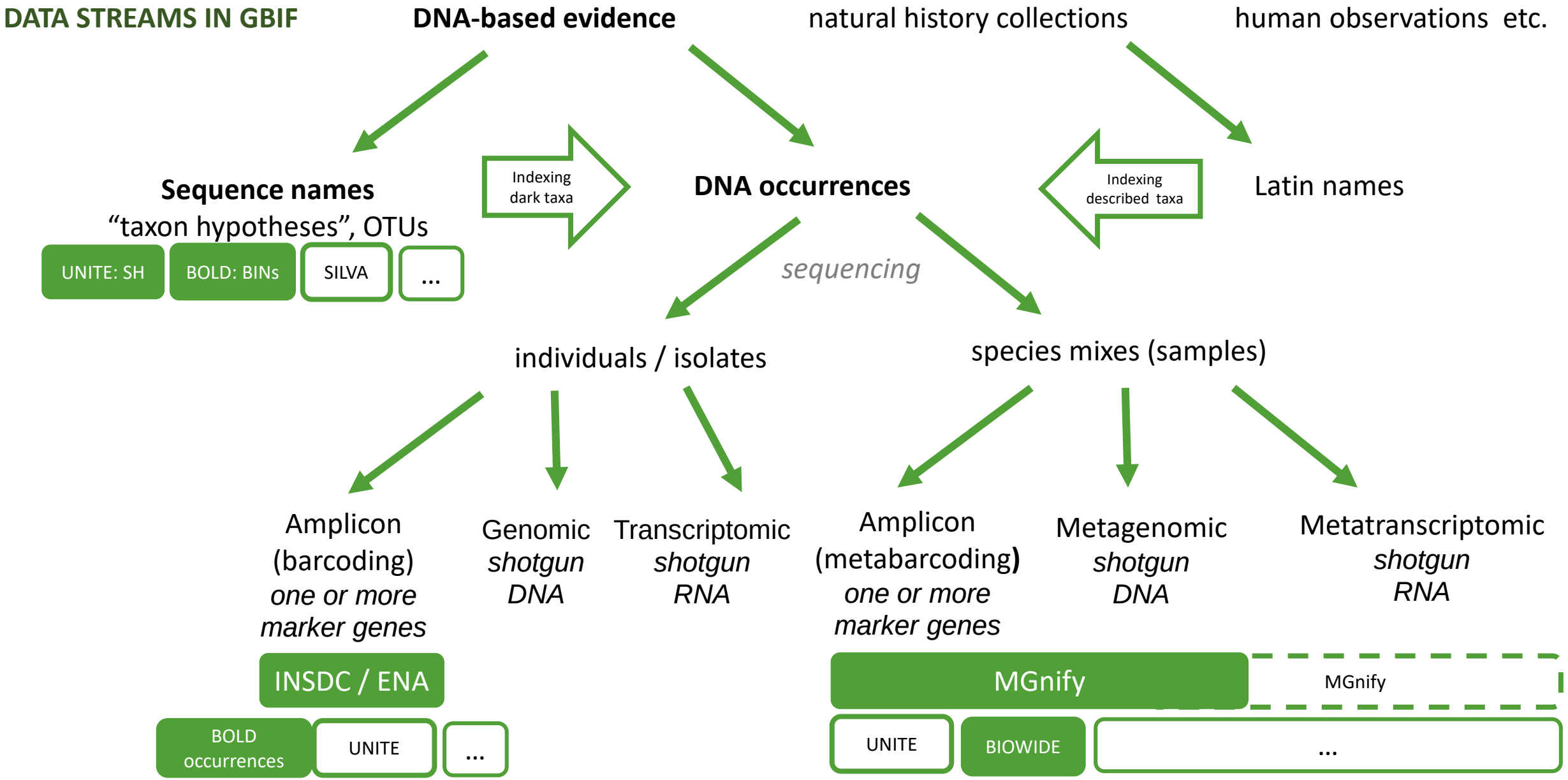
	Taxon	No. of Described Species		Taxon	No. of Described Species
1	Monera (Bacteria, Blue-green Algae)	4,760	11	Mollusca (Molluscs)	50,000
2	Fungi	46,983	12	Echinodermata (Starfish etc.)	6,100
3	Algae	26,900	13	Insecta	751,000
4	Plantae (Multicellular Plants)	248,428	14	Non-insect Arthropoda (Mites, Spiders, Crustaceans etc.)	123,161
5	Protozoa	30,800	15	Pisces (Fish)	19,058
6	Porifera (Sponges)	5,000	16	Amphibia (Amphibians)	4,181
7	Coelenterata (Jellyfish, Corals, Comb Jellies	9,000	17	Reptilia (Reptiles)	6,300
8	Platyhelminthes (Flatworms)	12,200	18	Aves (Birds)	9,040
9	Nematoda (Roundworms)	12,000	19	Mammalia (Mammals)	4,000
10	Annelida (Earthworms etc.)	12,000			

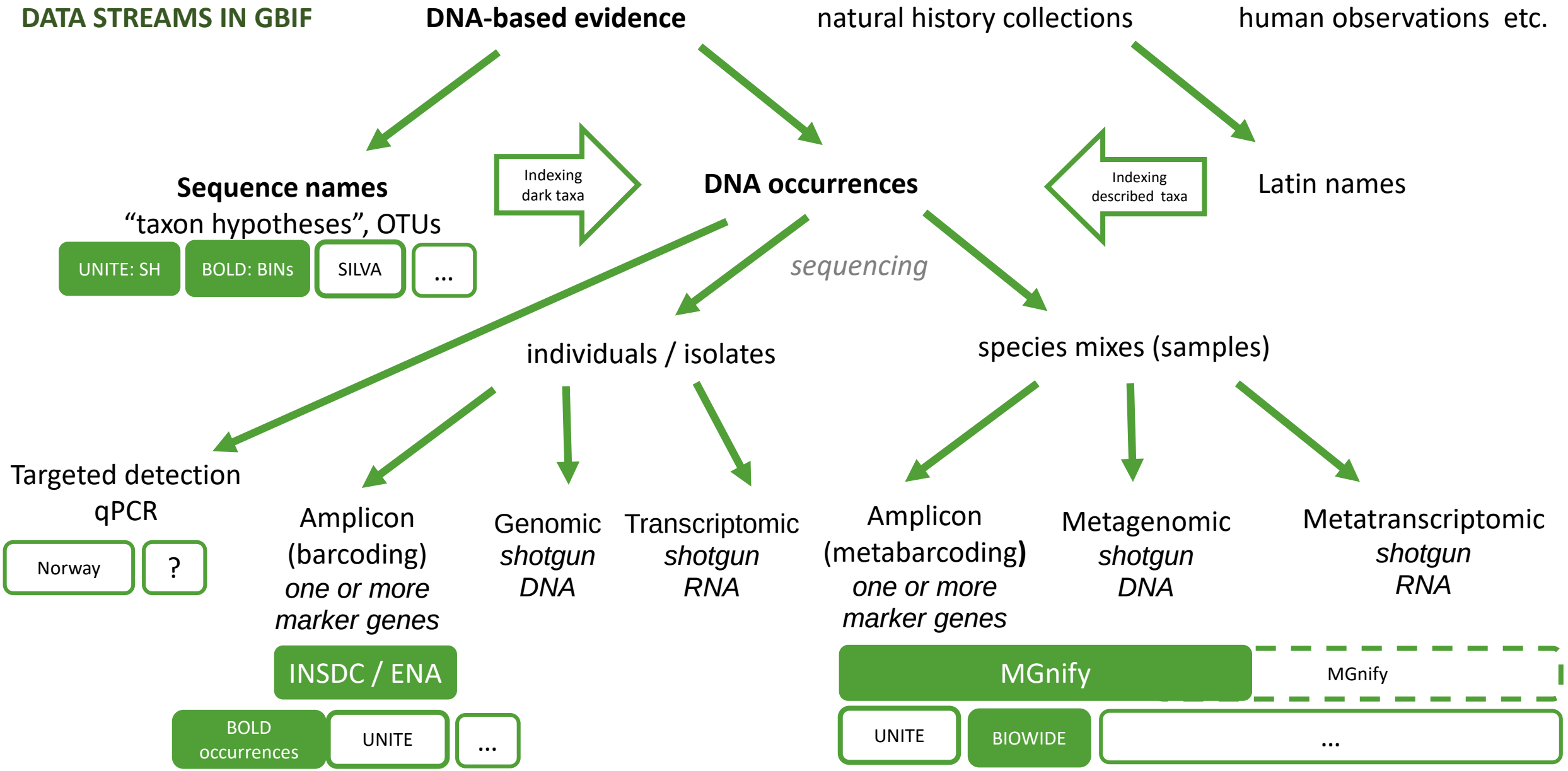
Illustration by Frances L. Fawcett. From Q. D. Wheeler. 1990. Ann. Entomol. Soc. Am. 83:1031-1047.

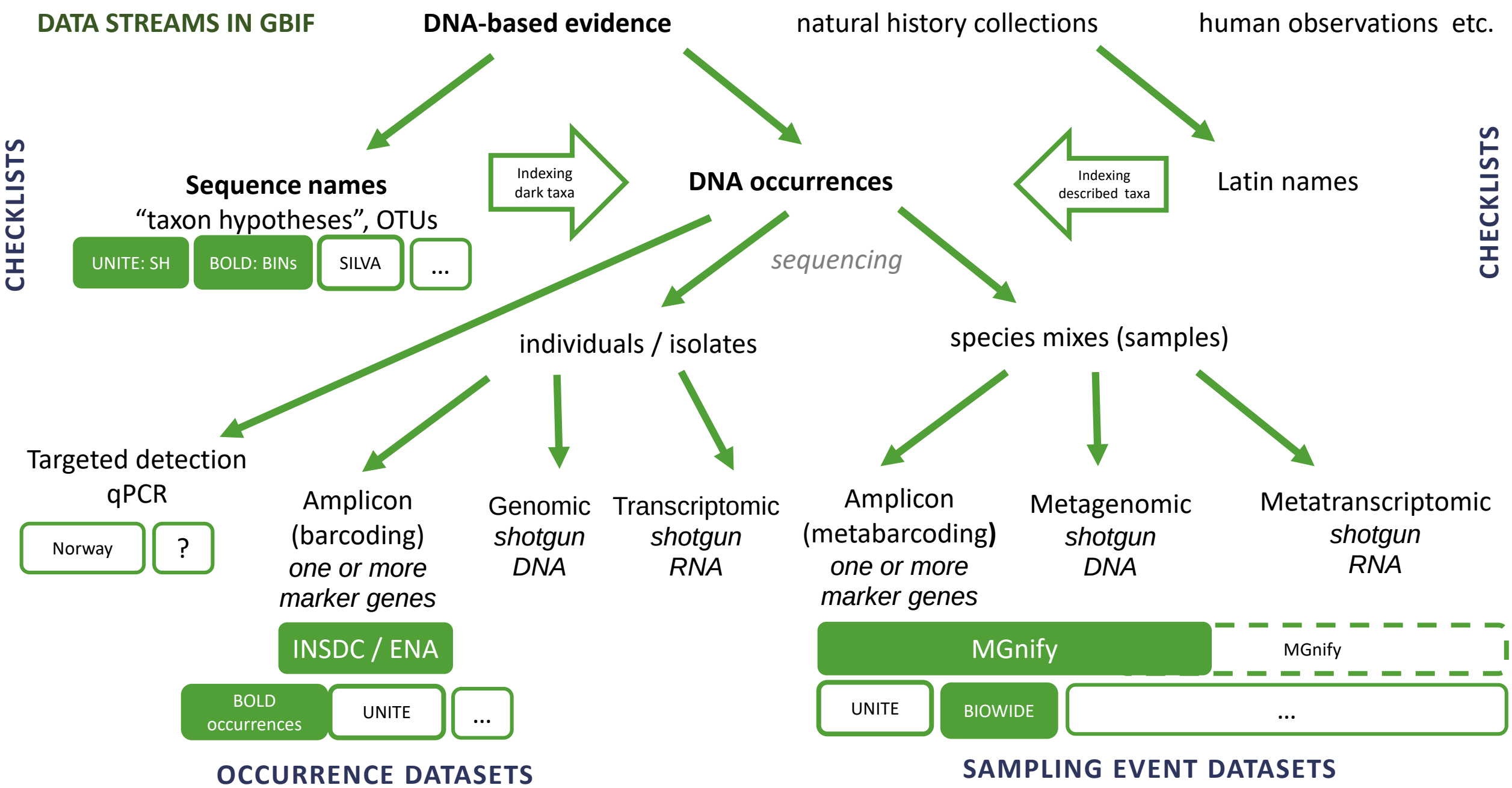


Troudet et al. Nature Scientific Reports 2017

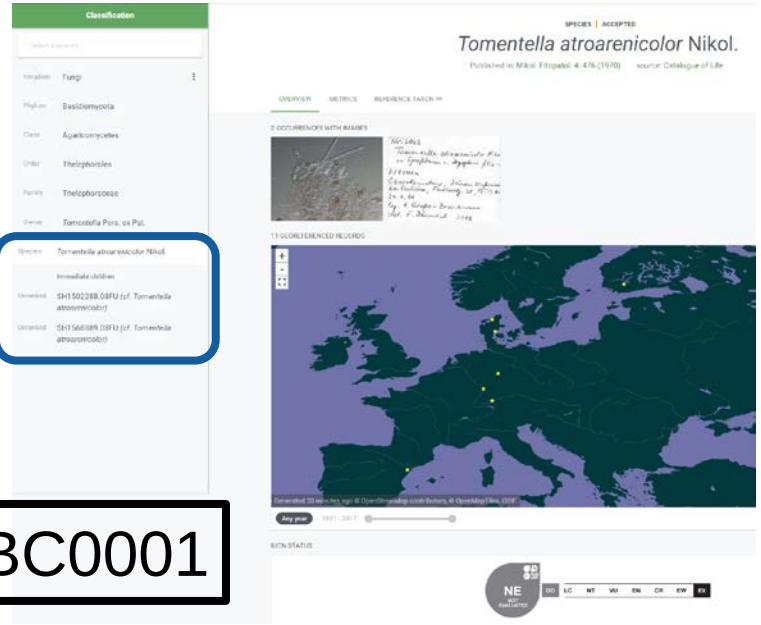

DATA STREAMS IN GBIF
DNA-based evidence
natural history collections
human observations etc.
Sequence names
“taxon hypotheses”, OTUs
UNITE: SH
BOLD: BINs
SILVA
...
Indexing dark taxa
DNA occurrences
Indexing described taxa
Latin names


DATA STREAMS IN GBIF
DNA-based evidence
natural history collections
human observations etc.
Sequence names
"taxon hypotheses", OTUs
UNITE: SH
BOLD: BINs
SILVA
...
Indexing dark taxa
DNA occurrences
Indexing described taxa
Latin names
sequencing
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
INSDC / ENA
BOLD occurrences
UNITE
...
MGnify
MGnify
UNITE
BIOWIDE
...


DATA STREAMS IN GBIF
DNA-based evidence
natural history collections
human observations etc.
Sequence names
"taxon hypotheses", OTUs
UNITE: SH
BOLD: BINs
SILVA
...
Indexing dark taxa
DNA occurrences
Indexing described taxa
Latin names
sequencing
individuals / isolates
species mixes (samples)
Targeted detection qPCR
Norway
?
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
INSDC / ENA
BOLD occurrences
UNITE
...
MGnify
MGnify
UNITE
BIOWIDE
...


DATA STREAMS IN GBIF
DNA-based evidence
natural history collections
human observations etc.
CHECKLISTS
Sequence names
"taxon hypotheses", OTUs
UNITE: SH
BOLD: BINs
SILVA
...
Indexing dark taxa
DNA occurrences
Indexing described taxa
Latin names
CHECKLISTS
sequencing
individuals / isolates
species mixes (samples)
Targeted detection qPCR
Norway
?
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
INSDC / ENA
BOLD occurrences
UNITE
...
MGnify
MGnify
UNITE
BIOWIDE
...
OCCURRENCE DATASETS
SAMPLING EVENT DATASETS

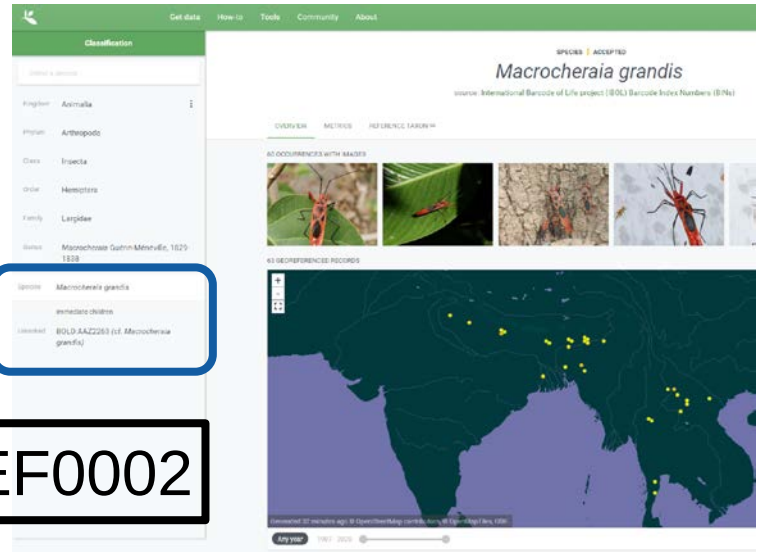
OPERATIONAL TAXONOMIC UNITS



SH ABC0001

OTU = **SH**,
Species hypothesis

GBIF
backbone taxonomy

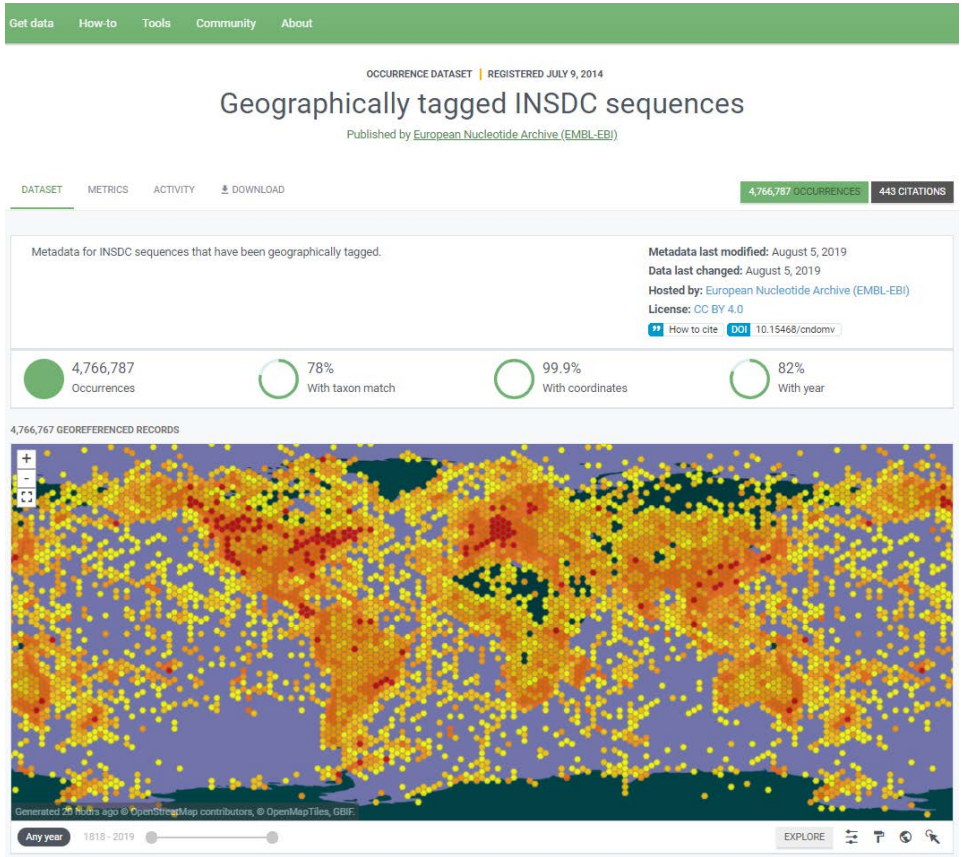


BIN DEF0002

OTU = **BIN**,
Barcode
identification number

INDIVIUDUAL SEQUENCES WITH COORDINATES

European Nucleotide Archive: 4.8M records

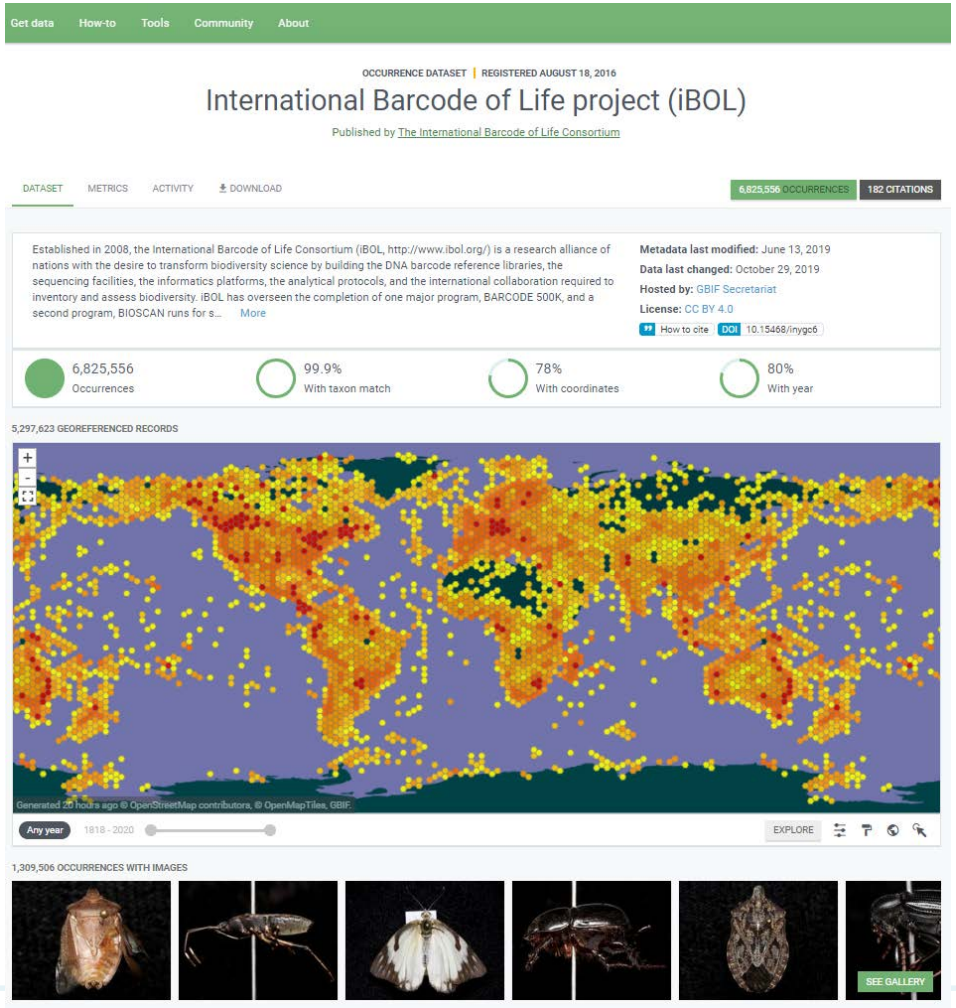


INSDC

ENA

European Nucleotide Archive

International Barcode of Life: 6.8M records



international BARCODE OF LIFE

GBIF

METABARCODING: MGNIFY AND MORE

920 datasets: 20.1M new records

14.8M Bacteria

2.3M Chromista

PUBLISHER | SINCE MARCH 13, 2019

MGnify

ABOUT METRICS HOME PAGE

20,107,151 OCCURRENCES | 920 DATASETS | 41 CITATIONS

Description: MGnify is a free to use analysis platform and database of analysed microbiome sequencing projects, hosted by the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI). Microbiome research typically involves the study of all genomes present within a specific environment, such as soil, seawater or a human body site. Underpinned by dramatically
mainstream. The approach can provide unique insights into the complex p
surroundings, each other, and, in some cases, their host. MGnify offers an
diversity and functional/metabolic potential of environmental samples. Us
held within the database. In addition, users can request analysis of any ap
following publication: EBI Metagenomics in 2017: enriching the analysis o
Mitchell, Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra T
Petra ten Hoopen, Blaise Alako, Clara Amid, Darren J. Wilkinson, Thomas F

ne the taxonomic
ed public datasets
, please refer to the
(2017) Alex L.
Fiona M. I. Hunter,

Endorsed by: National Biodiversity Network

Administrative contact: Robert Finn

Technical contact: MGnify Helpdesk

Country or area: United Kingdom of Great Britain and Northern Ireland

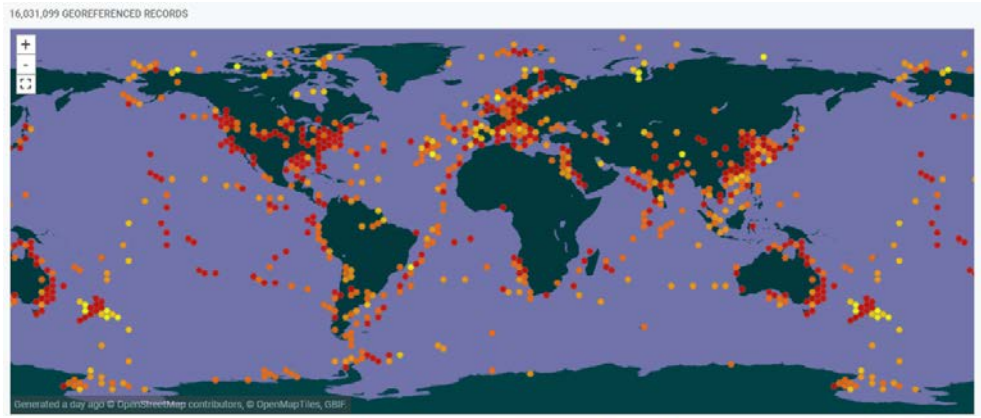
CONTACTS

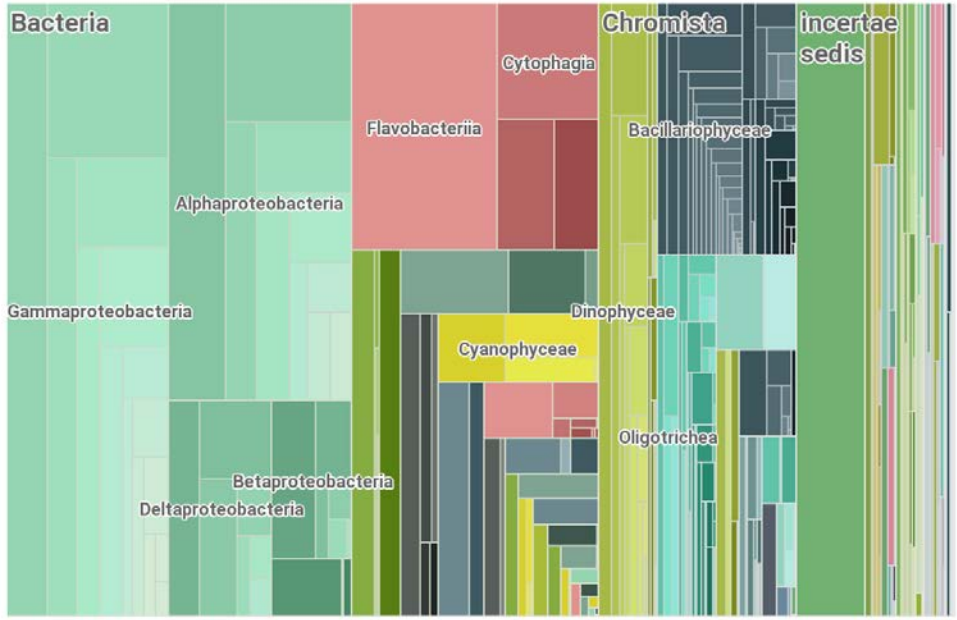
MGnify
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus
Hinxton CB10 1SD
Cambridge

MGnify Helpdesk
Technical point of contact
metagenomics-help@ebi.ac.uk

Robert Finn
Administrative point of contact • Point of contact
rdf@ebi.ac.uk
+441223492679

https://www.gbif.org/publisher/ab733144-7043-4e88-bd4f-fca7bf858880





NEW GUIDE: PUBLISHING DNA-DERIVED DATA THROUGH BIODIVERSITY DISCOVERY PLATFORMS

Mapping and data publishing

Cross-platform

About 40 pages long ”cookbook”

- **Introduction**
- **Categorization**
- **Mapping***
- **Visuals**
- Future prospects
- Resources

**Darwin Core and MIxS based*

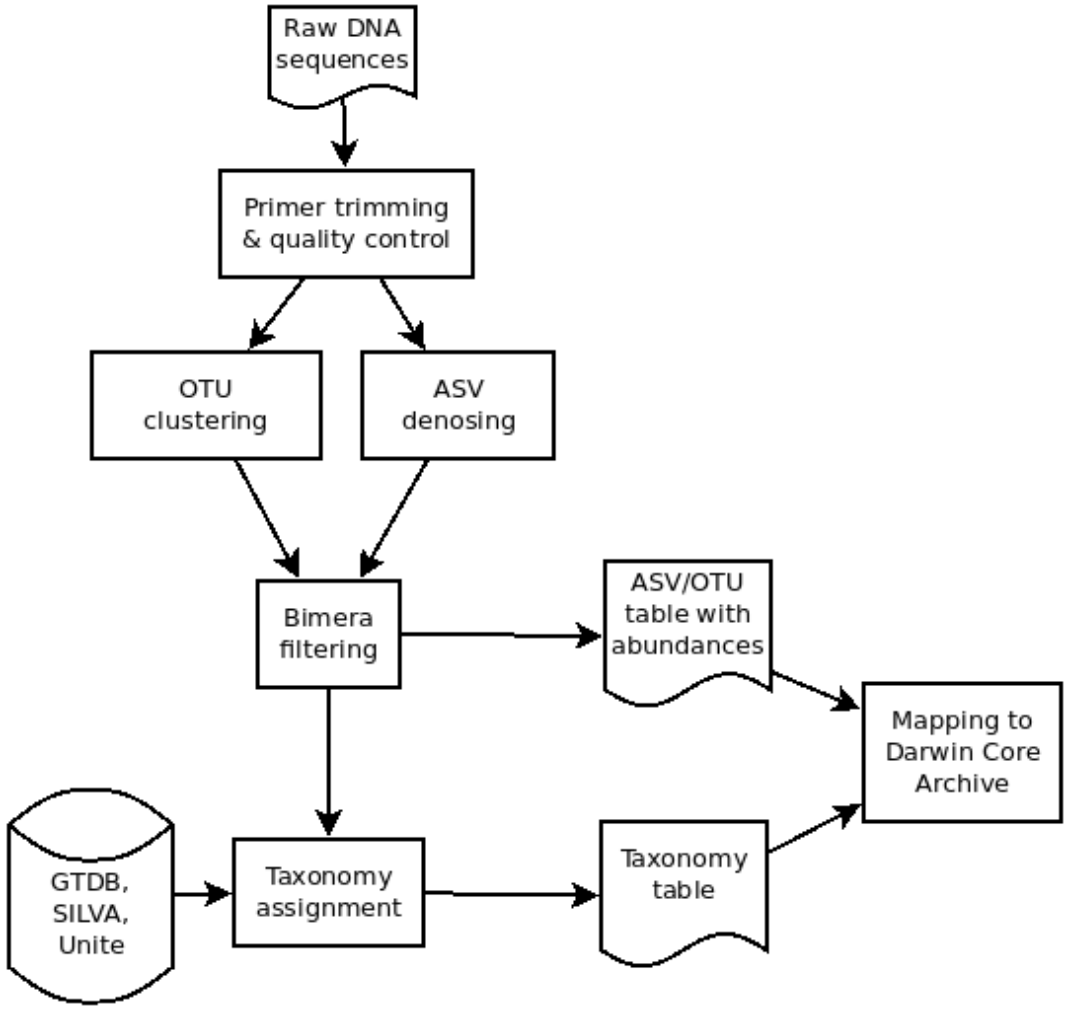
https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en

GBIF

PUBLISHING SEQUENCE-DERIVED DATA: THE “LEARN” SECTION

Introduction

- Rationale
- Audiences
- DNA derived occurrence data
- Biodiversity data
- Processing workflows
- Taxonomy of sequences



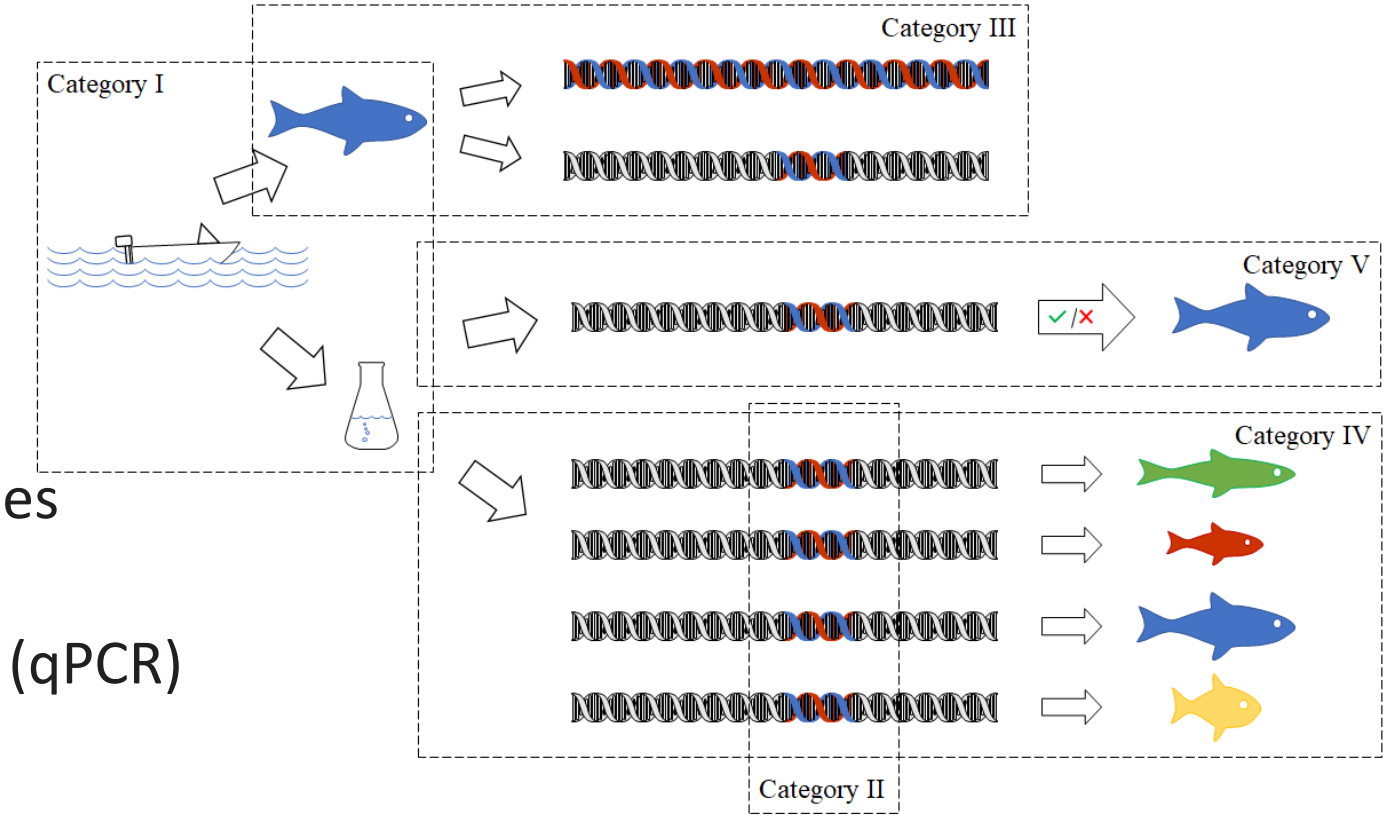
https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en

GBIF

PUBLISHING SEQUENCE-DERIVED DATA: THE "DO" SECTION

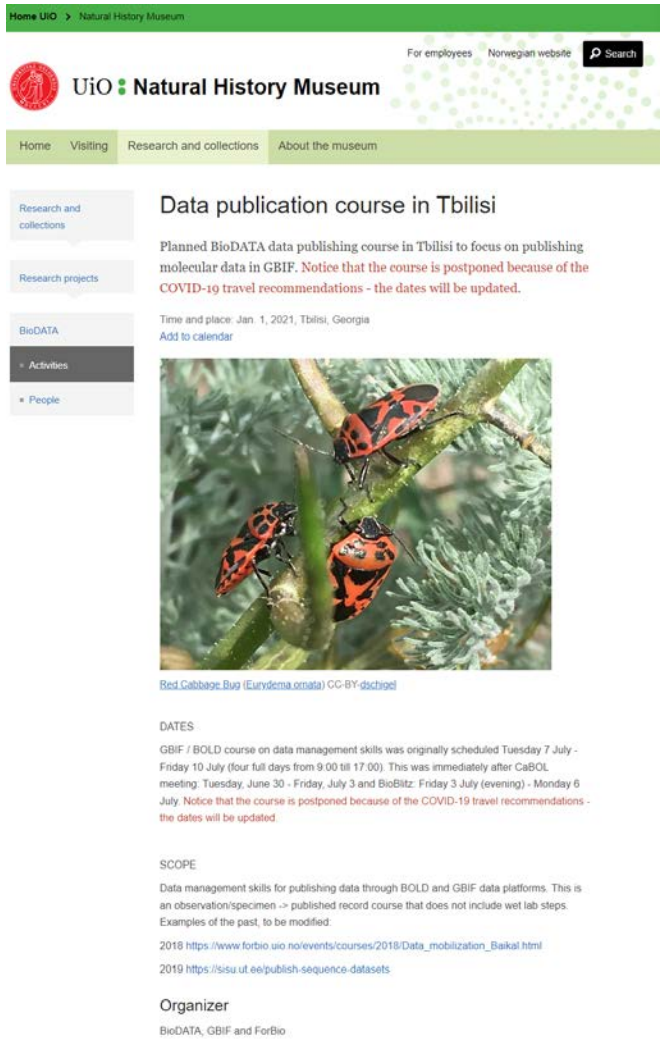
Data packaging and mapping

- Categorisation of data
  - I Sequence-based occurrences
  - II Enriched occurrences
  - III Targeted species detection (qPCR)
  - IV Name references
  - V Metadata-only
- Mapping
- Examples



https://docs.gbif-uat.org/publishing-sequence-derived-data/1.0/en

NEW TRAINING COURSE ON GBIF AND BOLD SKILLS – WITH BIODATA /GBIF NORWAY



Accelerating biodiversity research through DNA barcodes, collection and observation data

Work in progress

Why publish in GBIF?
Barcoding, data and biodiversity
Biodiversity databases in Georgia
Open data as a first-class research citizen
Databases as a research tool
Why do we need to identify species
Concept of DNA barcoding
Data in and data out: recognize and understand your data
Principles of data organization and personal data management
Data structure: standards
Data citation
Data exposure: why and when
Data papers
Barcode reference repositories
Quality control
Use case Bombus and legumes: uncovering pollination mysteries

KEY FACTS

- New GBIF course
- Tbilisi, Georgia
- 2021, 4 days
- Onsite or virtual
- GBIF and BOLD
- publish and use
- DNA-derived FAIR data
- Learn and practice

https://www.nhm.uio.no/english/research/projects/biodata/activities/202007-tbilisi.html

THANK YOU

Dmitry Schigel

dschigel@gbif.org