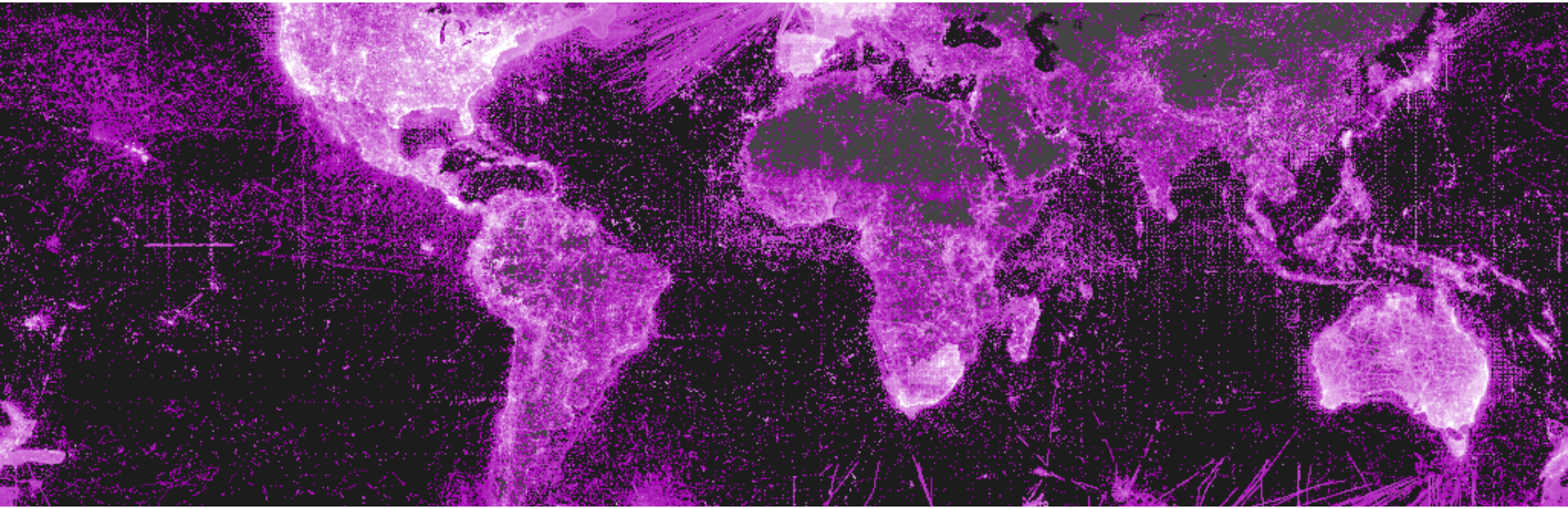
GBIF Global Biodiversity Information Facility

unite community



Shining a light on ‘dark taxa’:

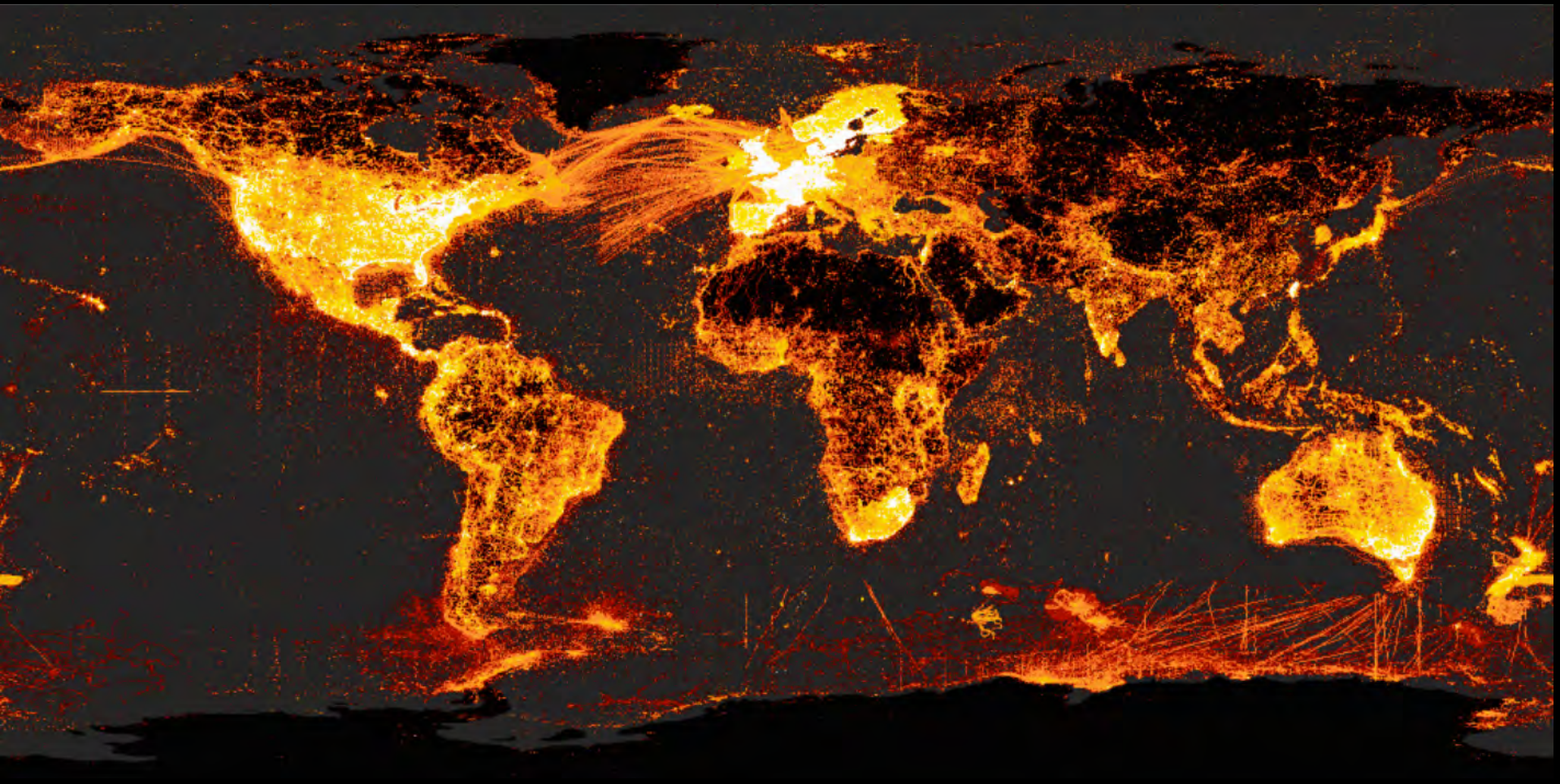
naming and discovering DNA-based biodiversity data

***Urmas Kõljalg**, University of Tartu, Estonia*
***Dmitry Schigel**, GBIF Secretariat*

17 OCTOBER 2018

OUTLINE

- Taxonomic bias in GBIF
- Fungi in GBIF
- UNITE and GBIF
- Latin and non-Latin names
- Going beyond fungi



Animalia
1,170,259
occurrences

Plantae
142,480
occurrences

Fungi
6,661
occurrences

Archaea
0
occurrences

Bacteria
30
occurrences

Chromista
7,394
occurrences

Protozoa
49
occurrences

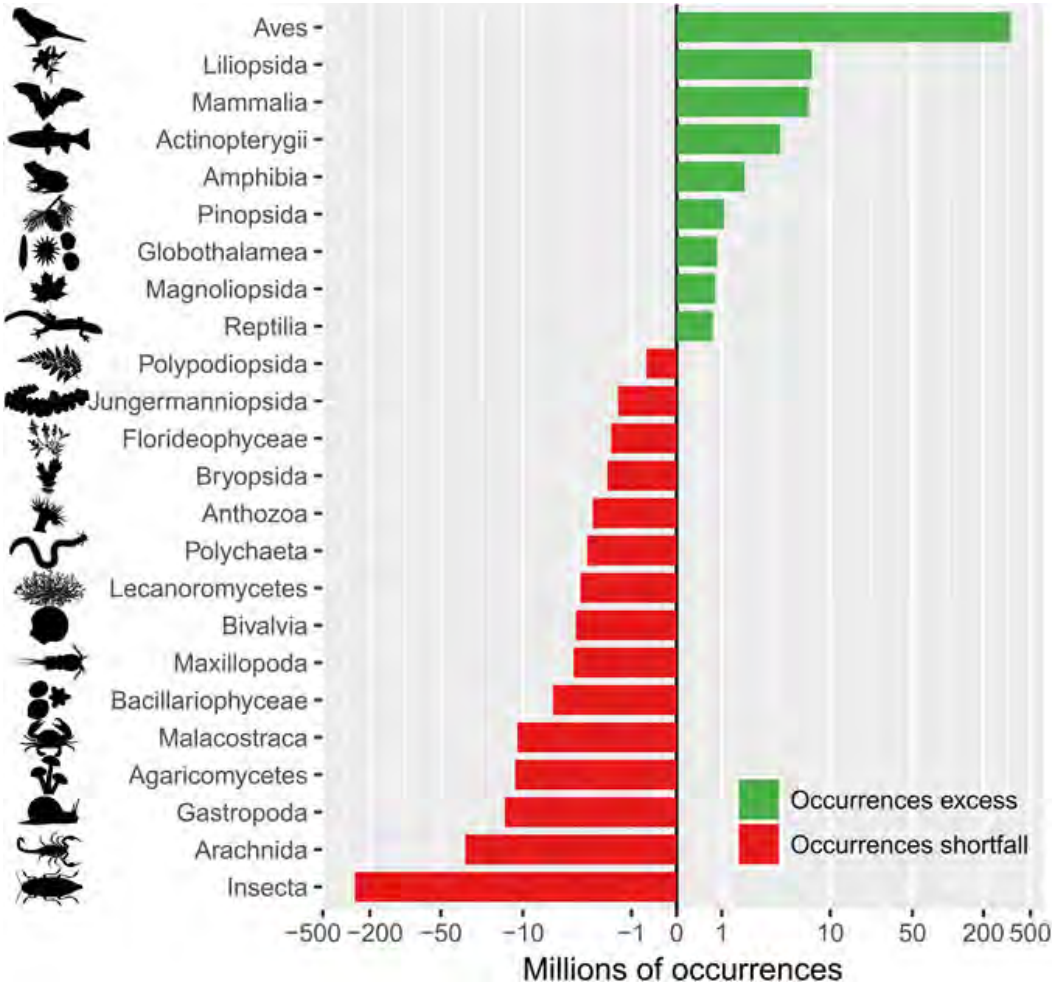
Viruses
0
occurrences

Incertae sedis
1
occurrence

Pica photo by Daniele Colombo via Flickr Creative Commons; the rest are CC0

GBIF

TAXONOMIC BIAS IN GBIF MEDIATED DATA



vertical line at x = 0
'ideal' N occurrences per class, where each class is sampled proportionally to its number of known species

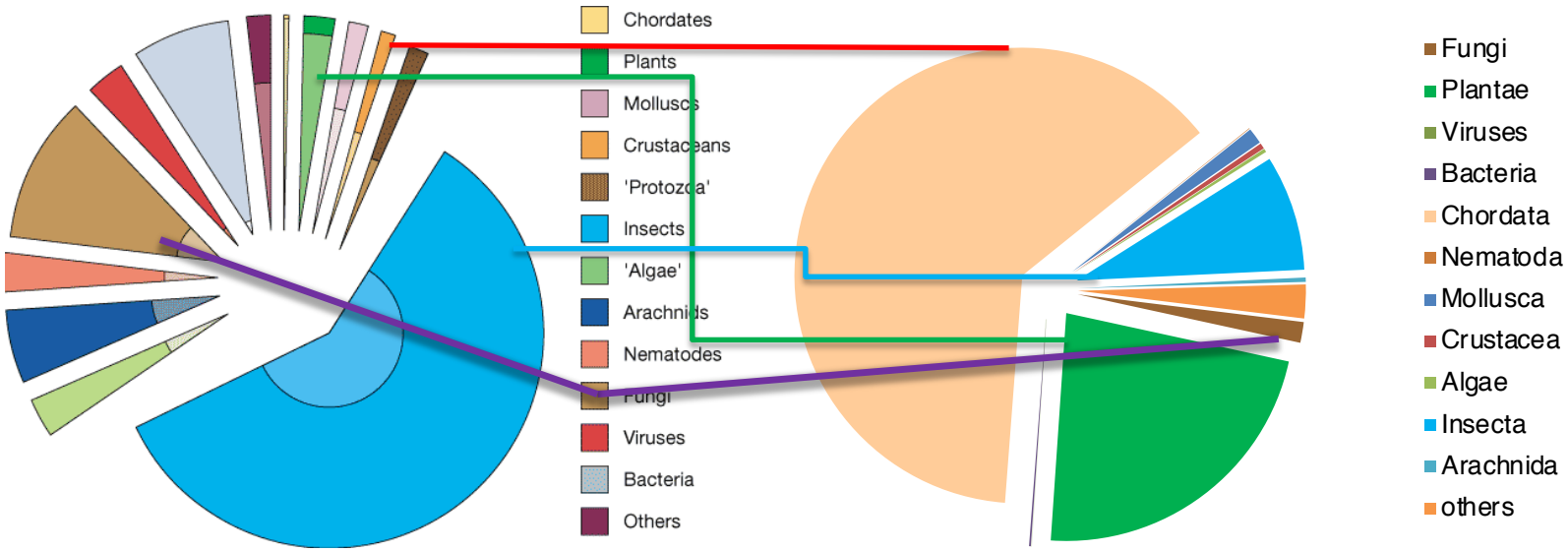
Green and red bars
classes that are over- and under-represented in GBIF compared to 'ideal' sampling

Insects: <200M occurrences
Birds: >200M occurrences

*Troudet et al. 2017. Scientific Reports **7** 9132 doi:10.1038/s41598-017-09084-6*

GBIF

GLOBAL BIODIVERSITY DATA: 2018

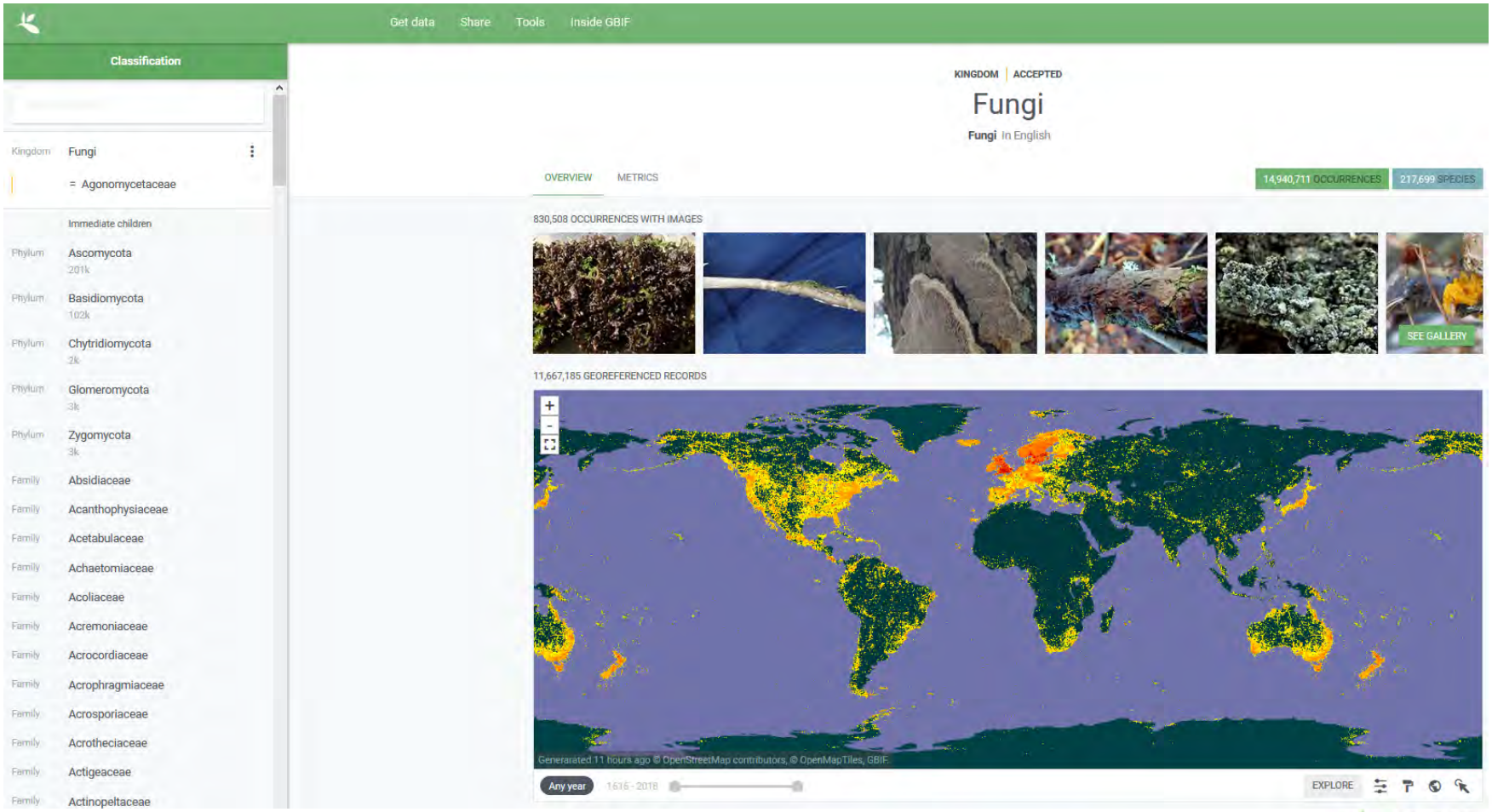


Number of species worldwide:
Purvis & Hector 2000 Nature 405
doi:10.1038/35012221

Number of species occurrences
GBIF.org, 9 October 2018

FUNGI IN GBIF

14,940,711 occurrences:
8M+ observations / 5M+ specimens
11,667,185 georeferenced
830,508 with images
217,699 species



As for 9 October 2018

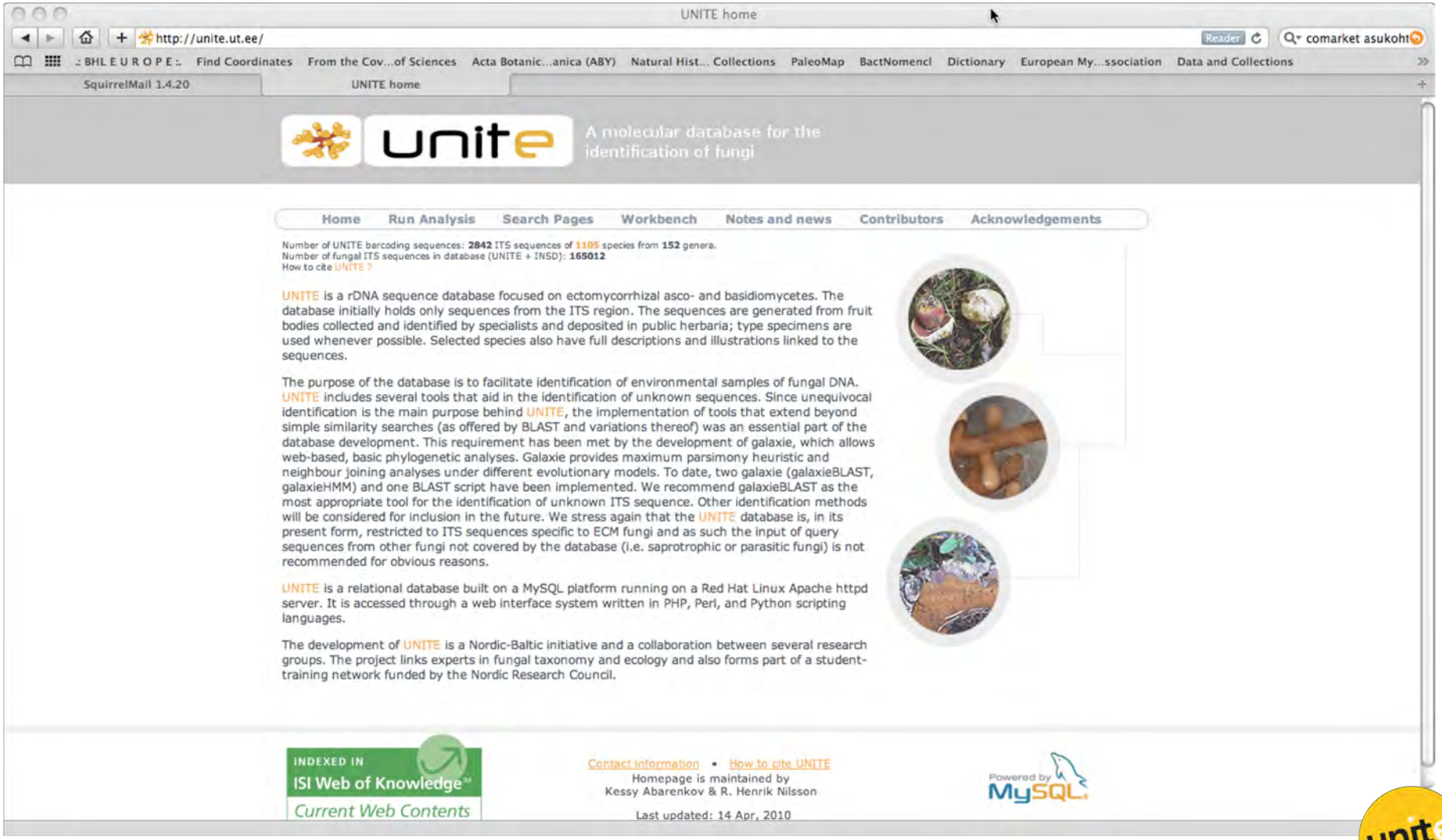
WHY ARE FUNGI UNDERREPRESENTED?

- Fungi are difficult to identify
- Few citizen science portals deal (seriously) with fungi
- Molecular researchers generally don´t publish to GBIF (yet?)
- GBIF portal starts to expose **molecular occurrences named by stable OTU identifiers**

CELEBRATING 15 YEARS IN SERVICE

UNITE 2003 (UNITE ITS SEQUENCES

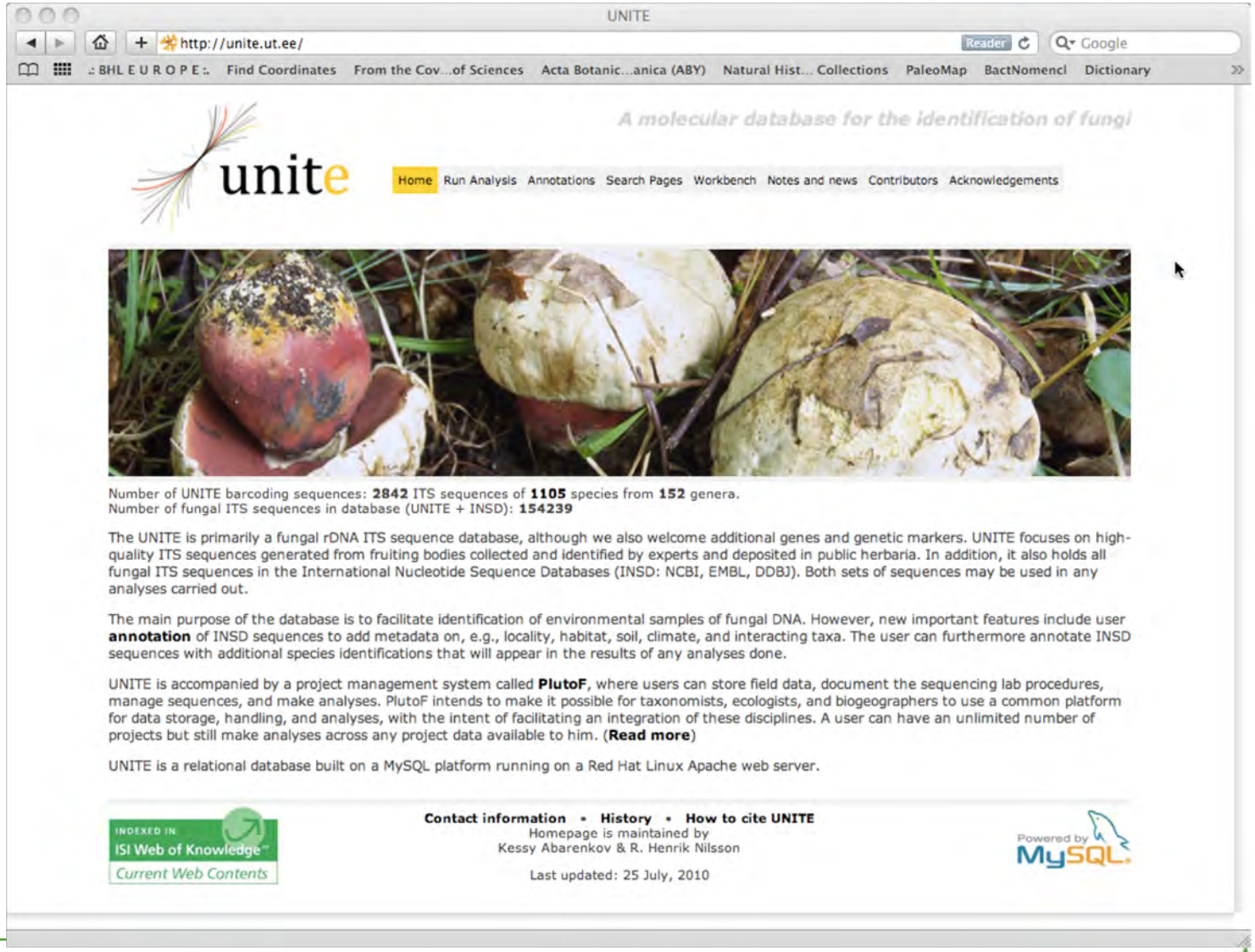
UNITE 2005 (UNITE + INSDC ITS SEQUENCES)



unite community

GBIF

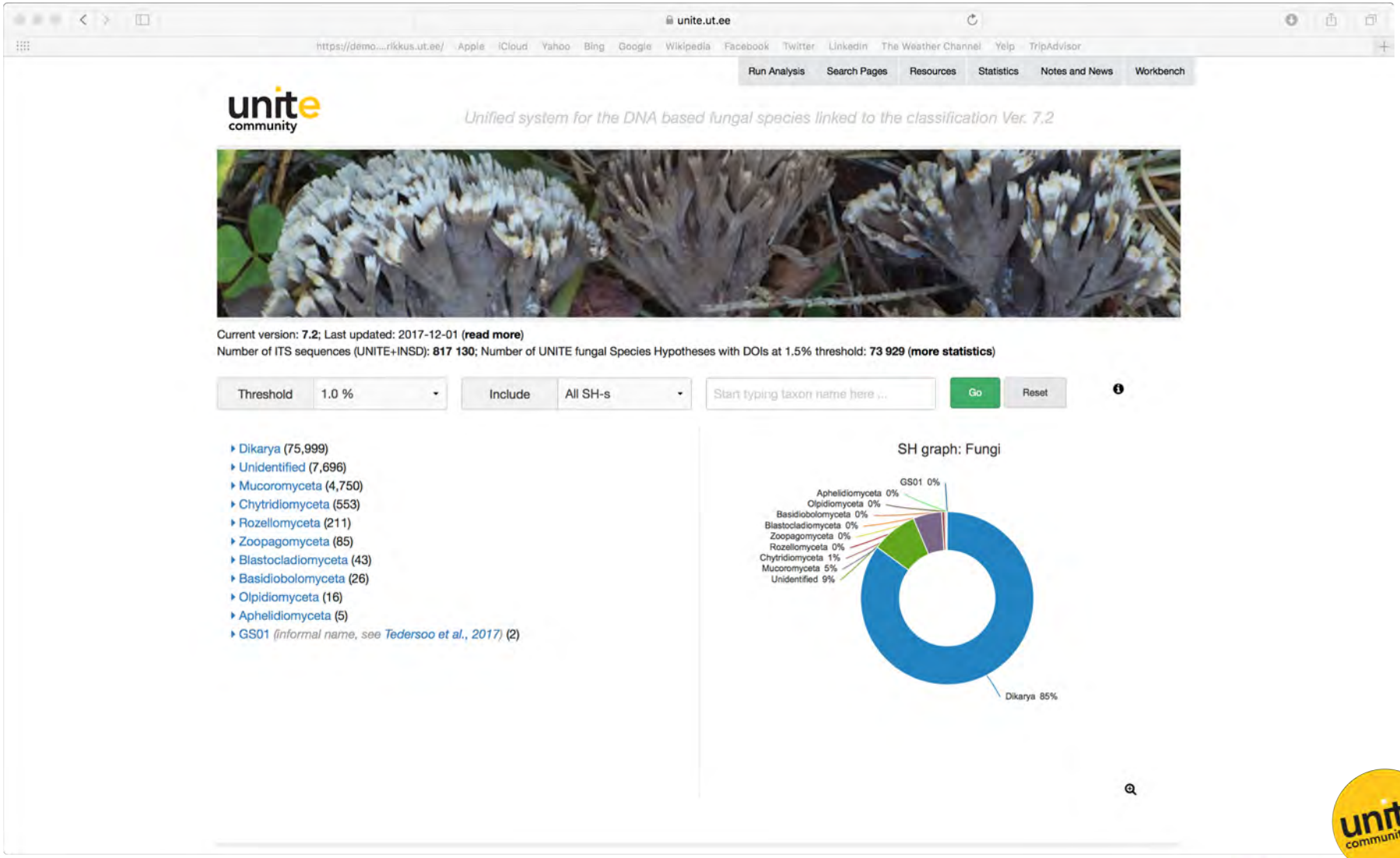
UNITE 2010 (COMMUNICATION SYSTEM FOR SPECIES - NAMES OR NO NAMES AVAILABLE)



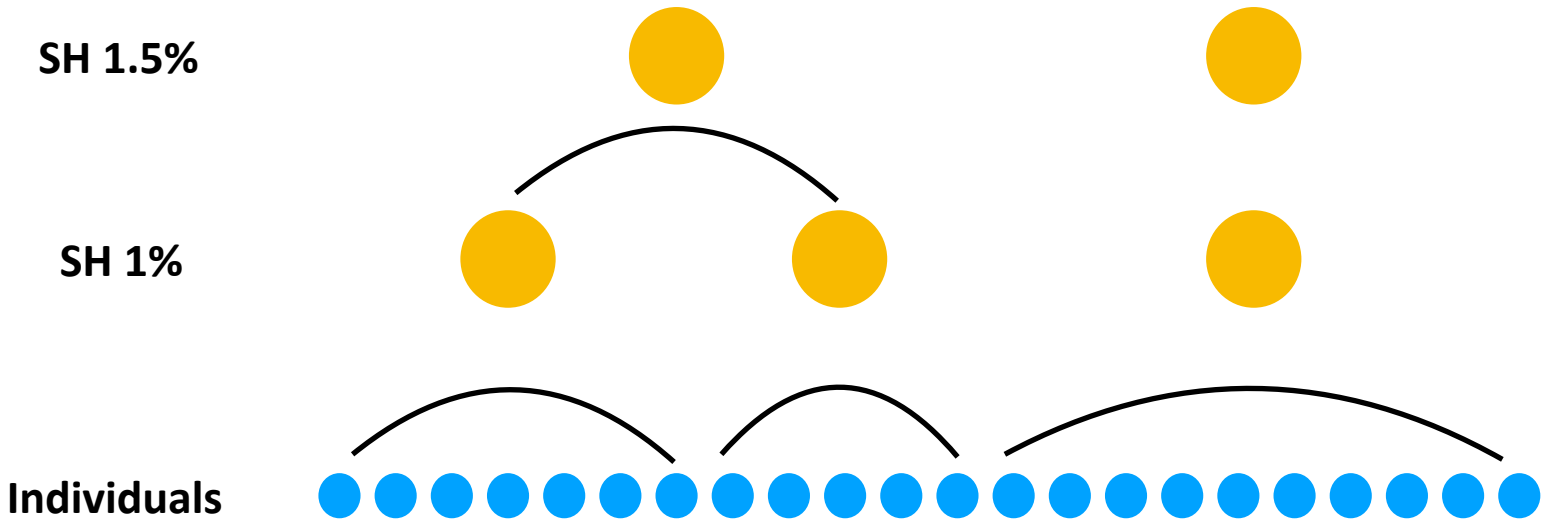
unite community

GBIF

UNITE 2018 (SPECIES HYPOTHESES - IDENTIFICATION AND COMMUNICATION SYSTEM)



UNITE - SPECIES HYPOTHESES (SH)

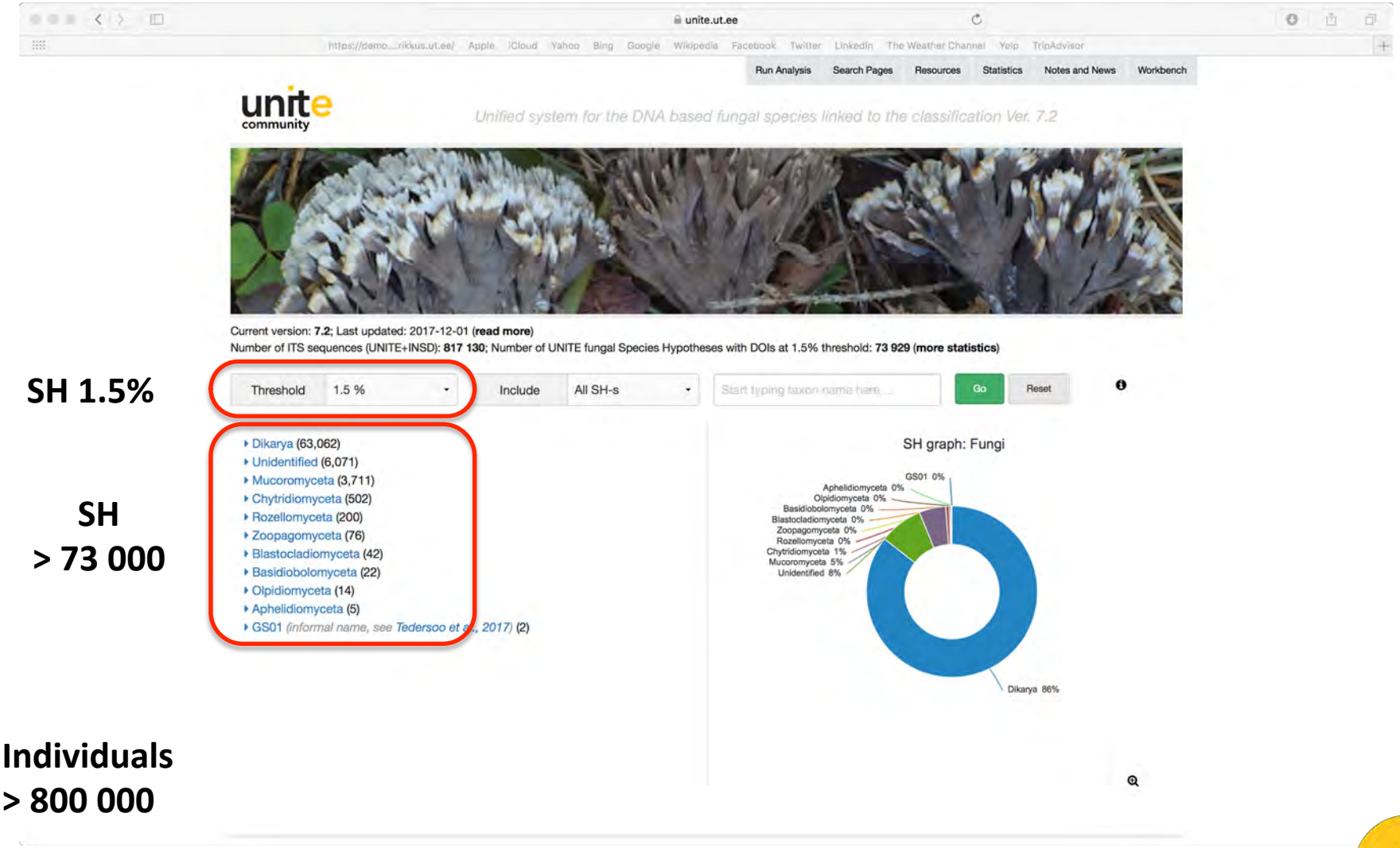


UNITE - SPECIES HYPOTHESES (SH)

SH 1.5%

SH
> 73 000

Individuals
> 800 000

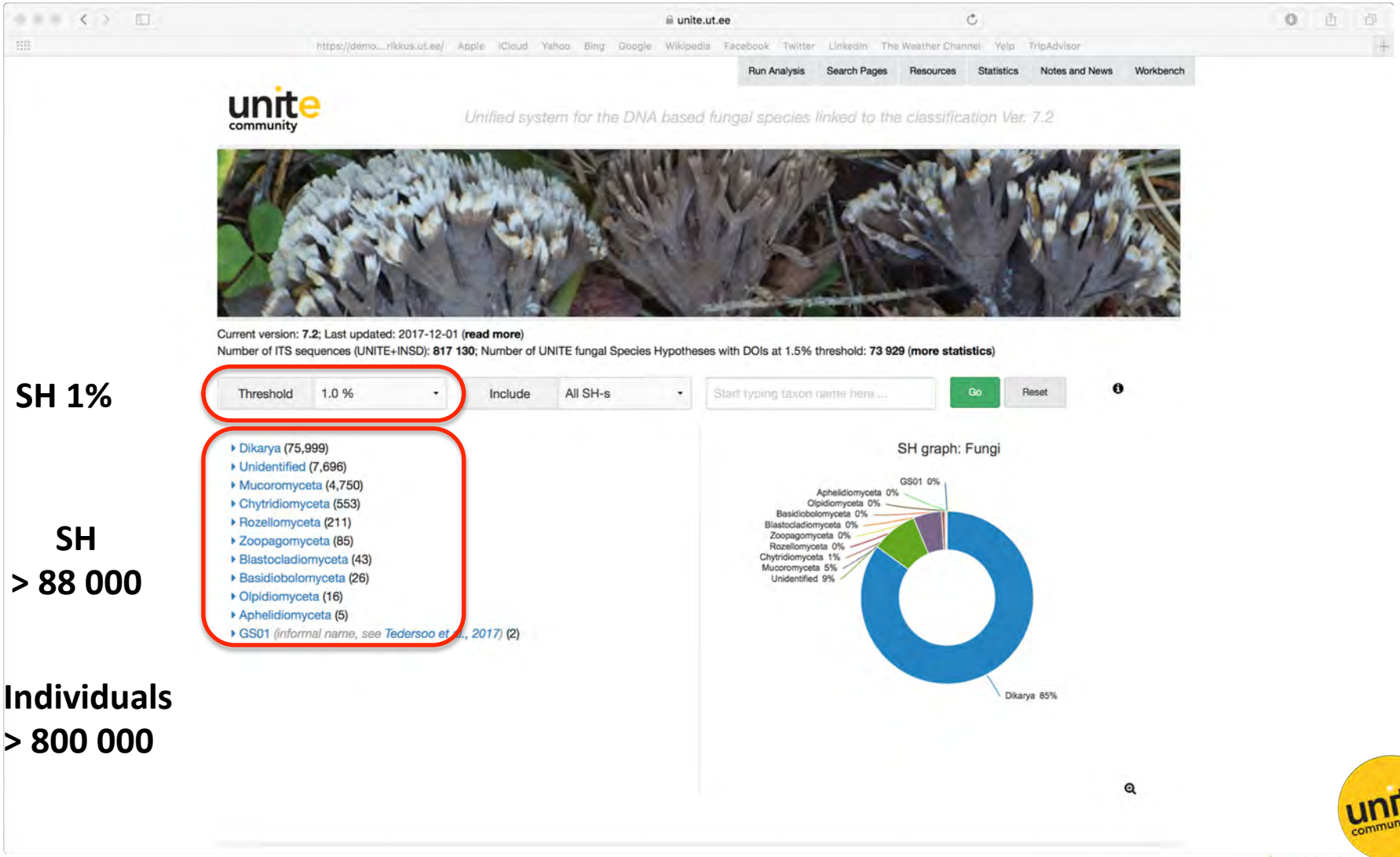


UNITE - SPECIES HYPOTHESES (SH)

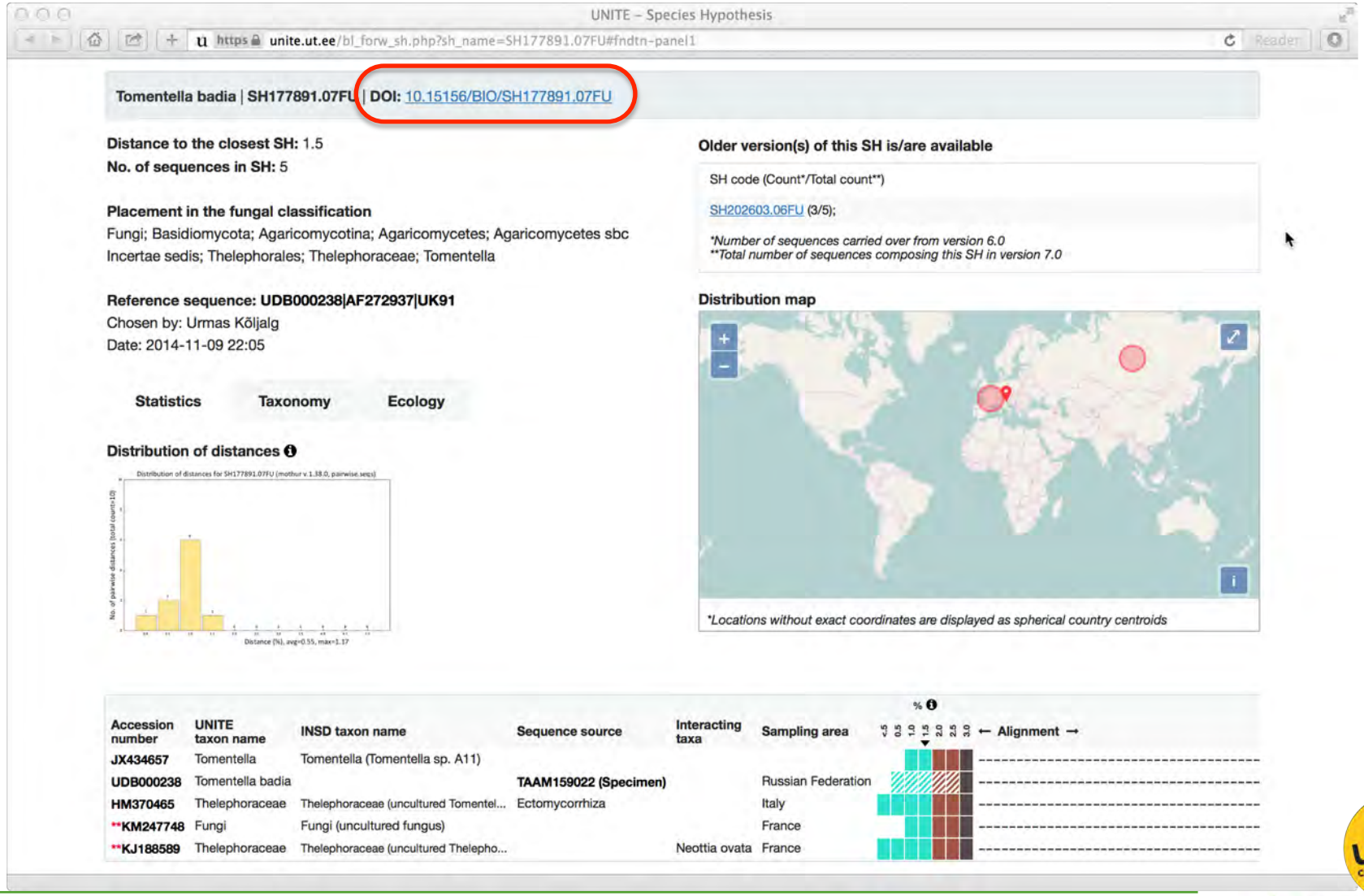
SH 1%

SH
> 88 000

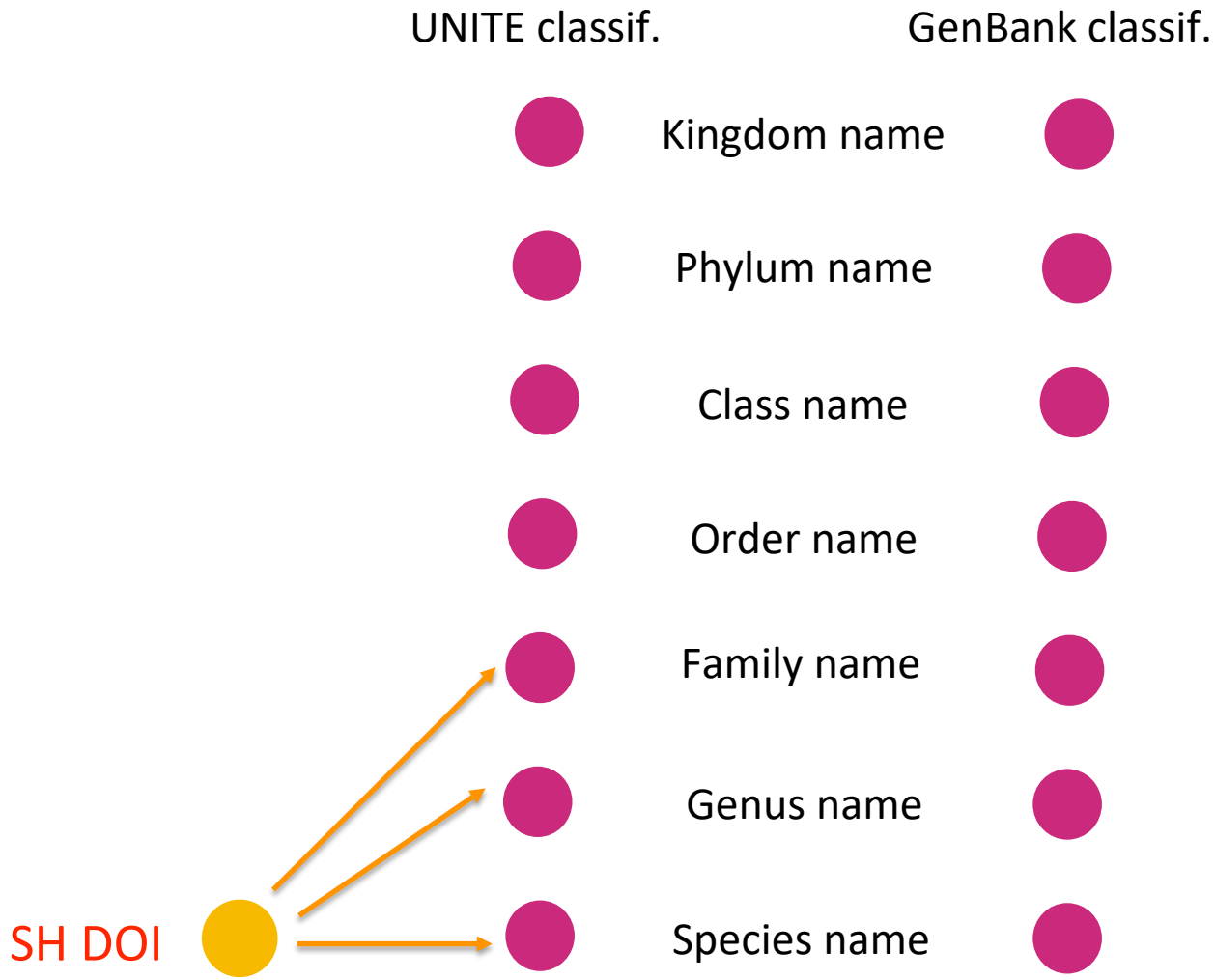
Individuals
> 800 000



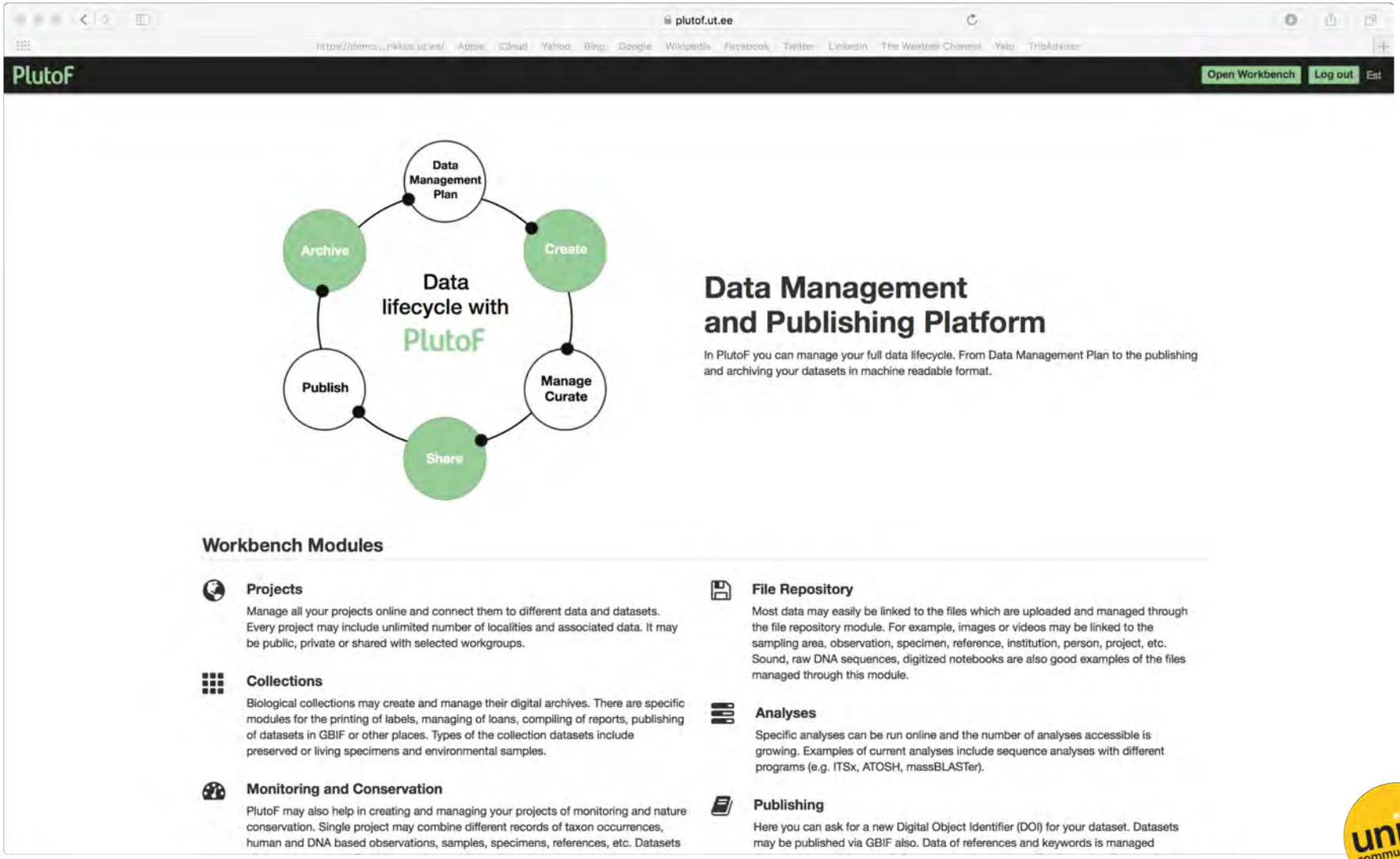
UNITE - EVERY SH HAS DOI (DIGITAL OBJECT IDENTIFIER)



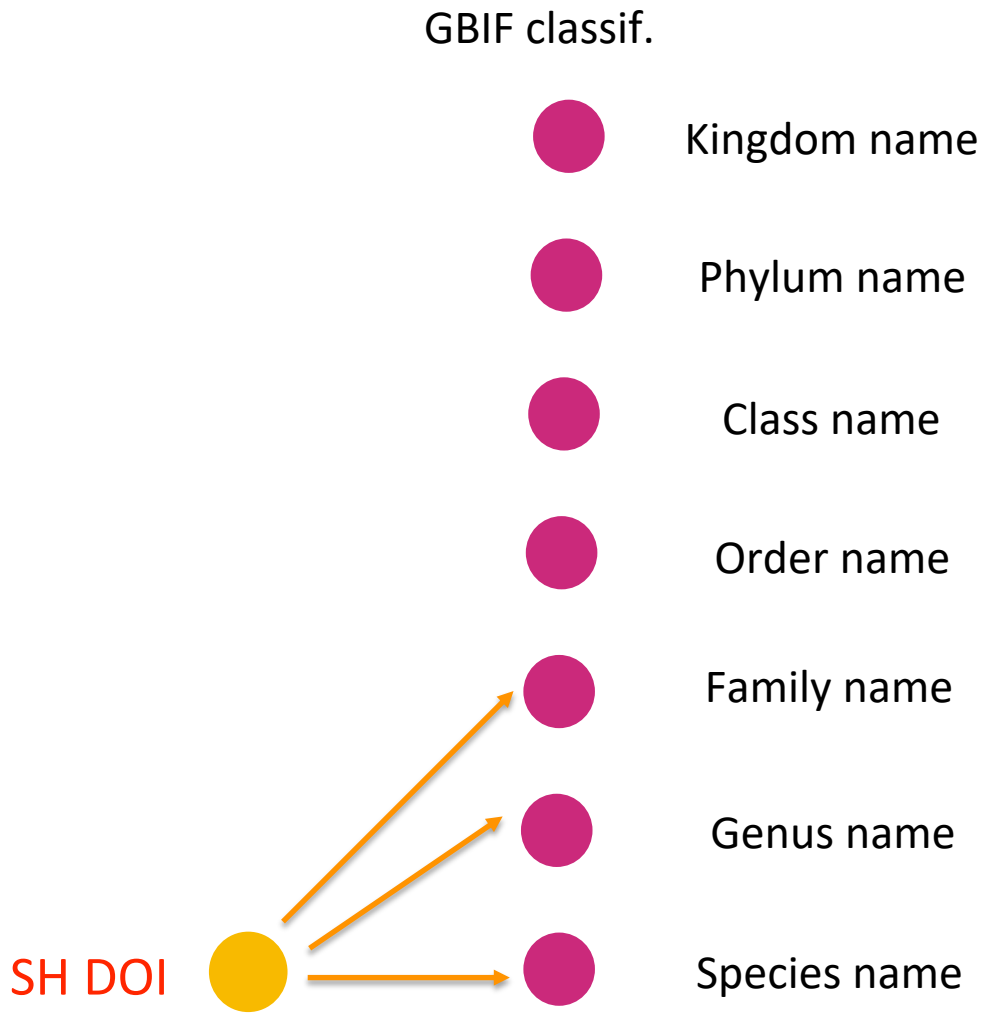
EVERY SH DOI IS CONNECTED TO THE CLASSIFICATION



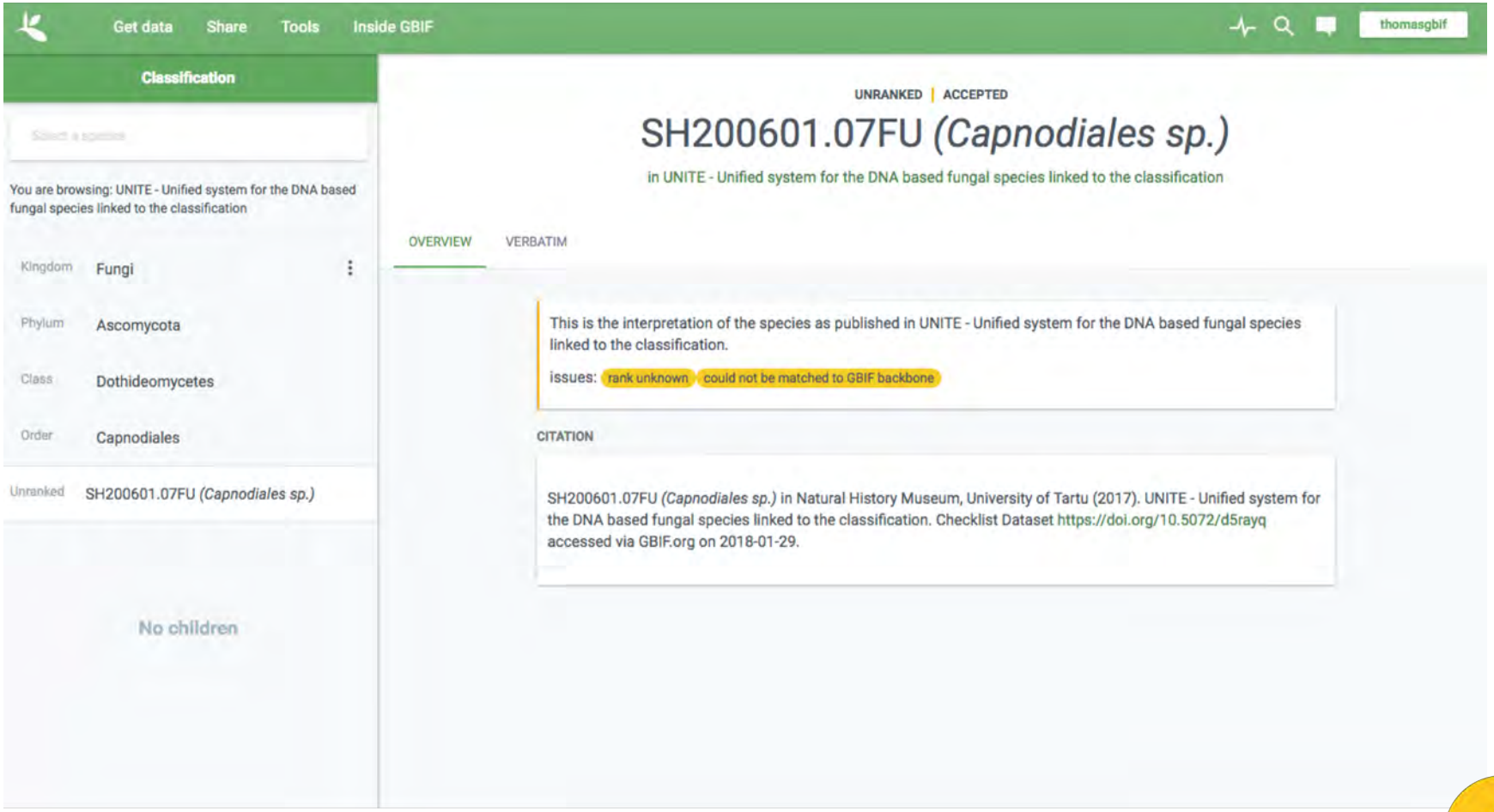
PLUTOF - DATA MANAGEMENT AND PUBLISHING PLATFORM



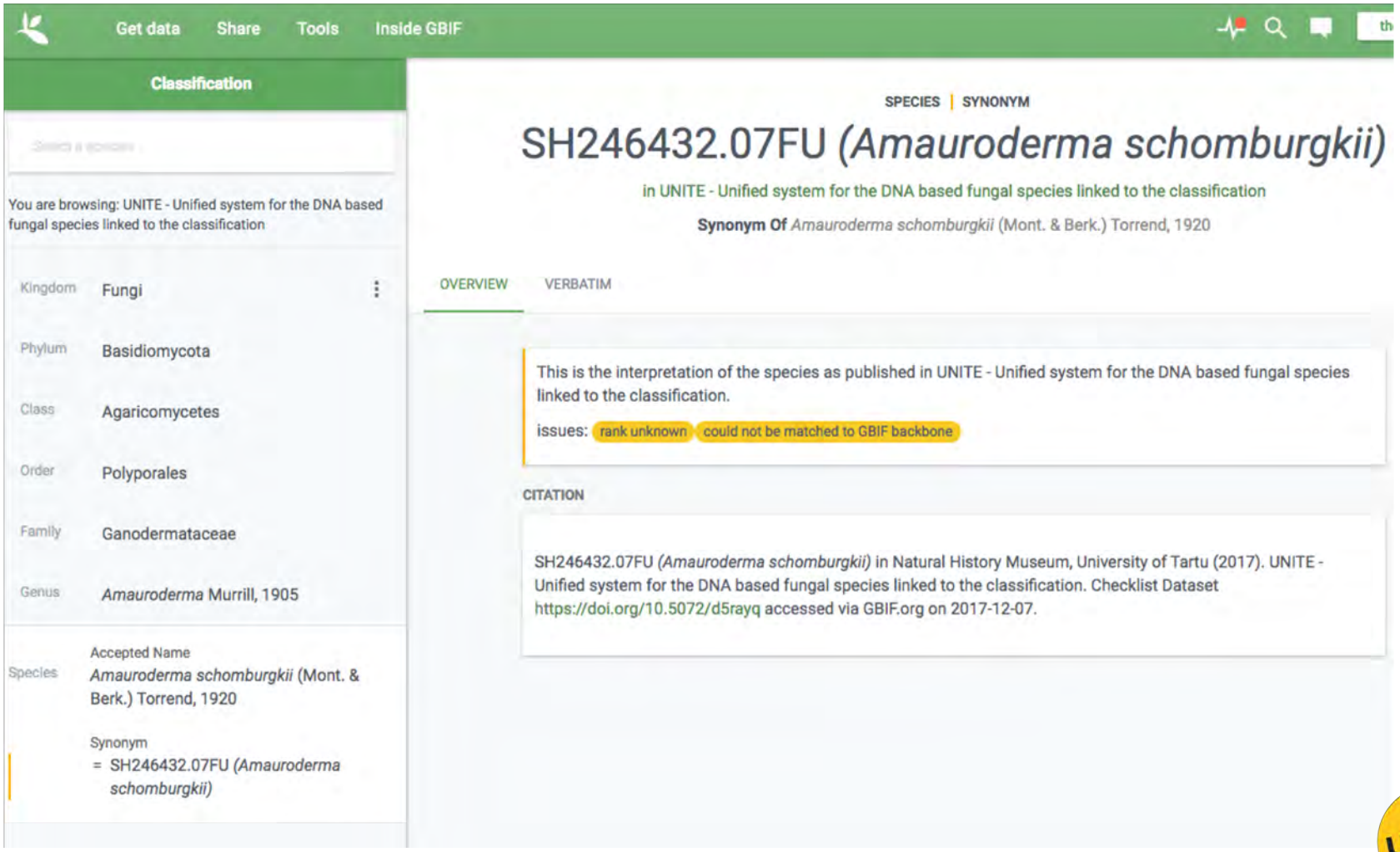
GBIF - EVERY SH DOI IS CONNECTED TO THE CLASSIFICATION



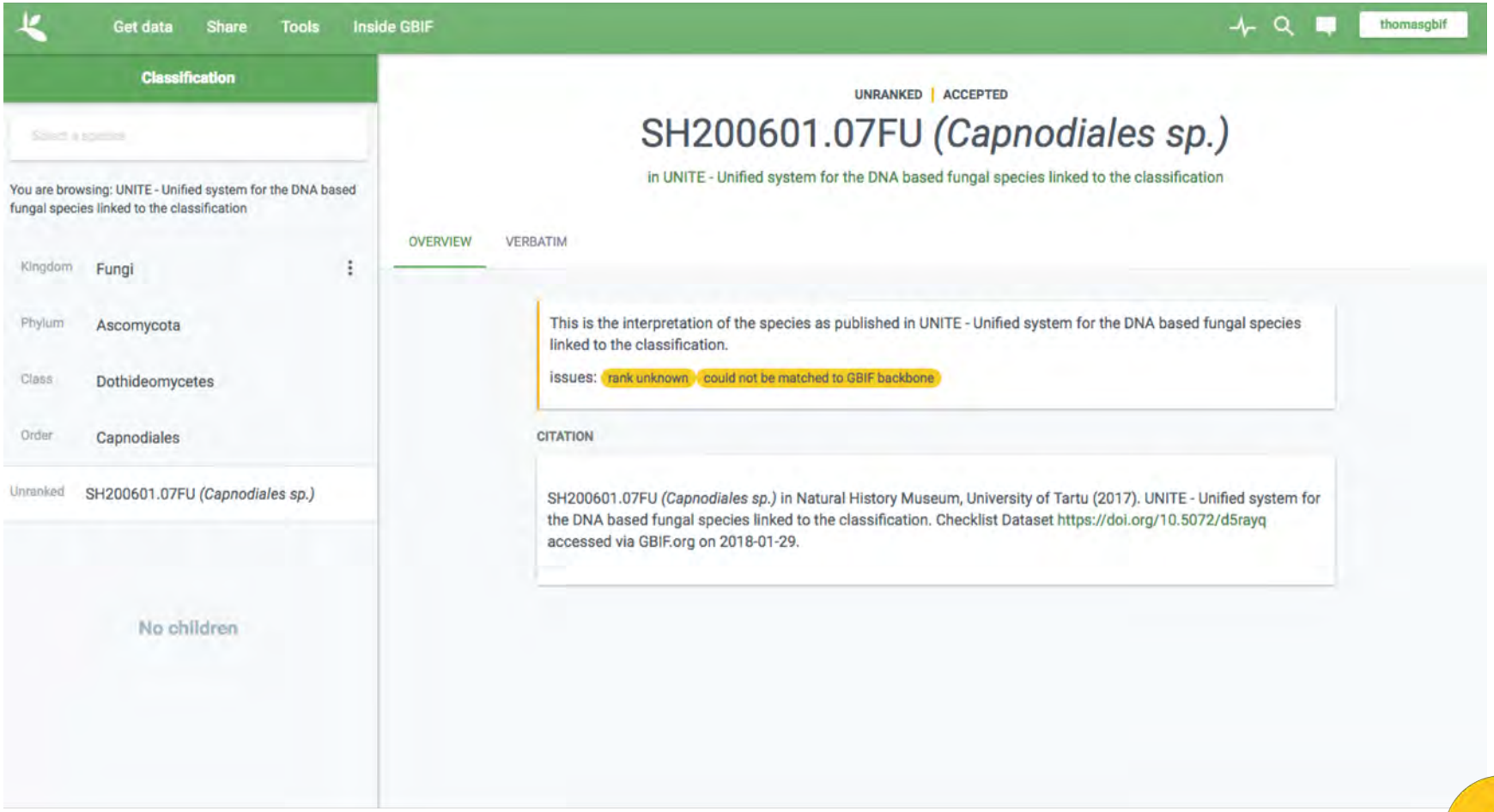
GBIF - EVERY SH DOI IS CONNECTED TO THE CLASSIFICATION



GBIF - EVERY SH DOI IS CONNECTED TO THE CLASSIFICATION



GBIF - EVERY SH DOI IS CONNECTED TO THE CLASSIFICATION



GBIF - NEWS

gbif.org

Get data Share Tools Inside GBIF Login

NEWS | 29 AUGUST 2018

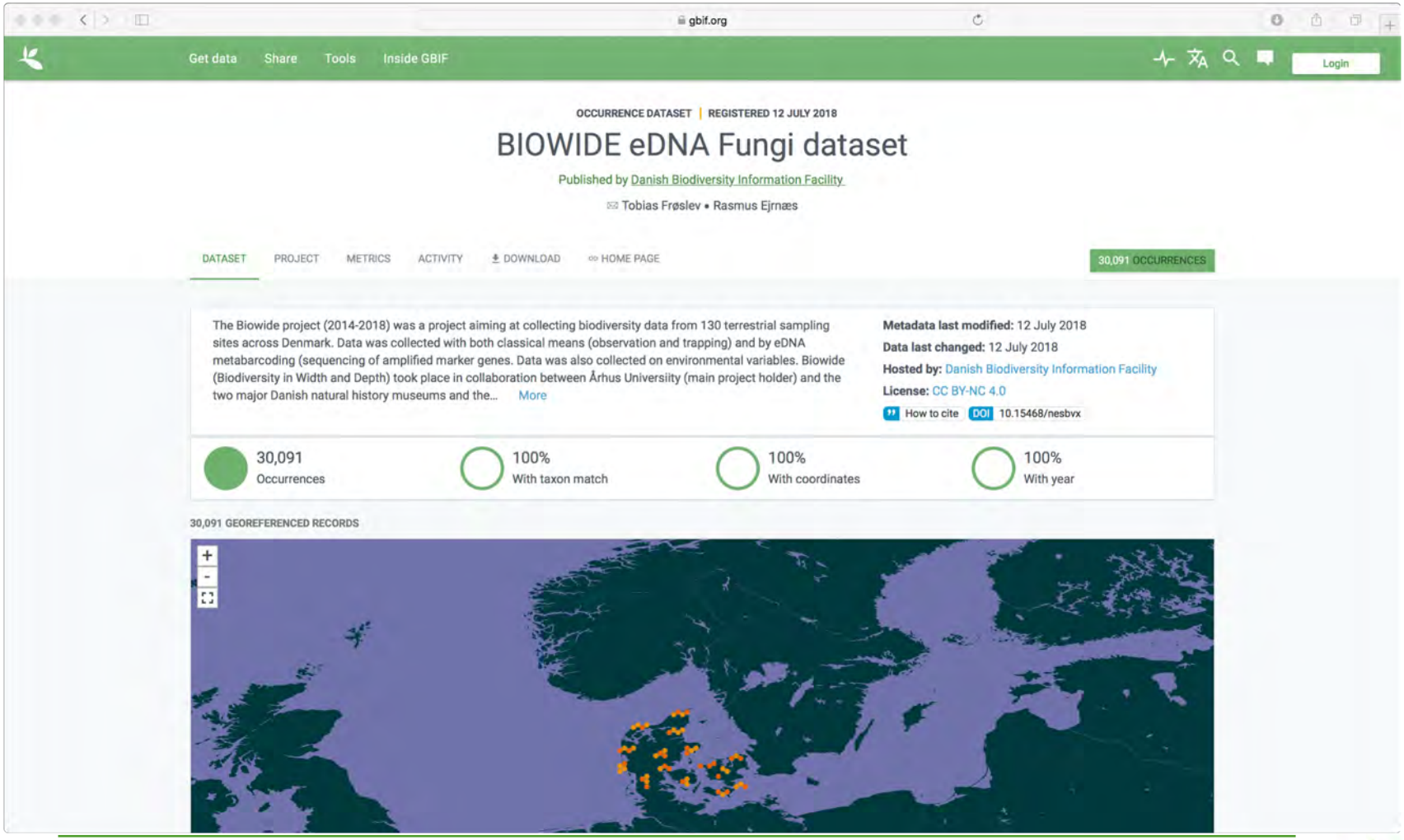
Adding sequence-based identifiers to backbone taxonomy reveals 'dark taxa' fungi

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.

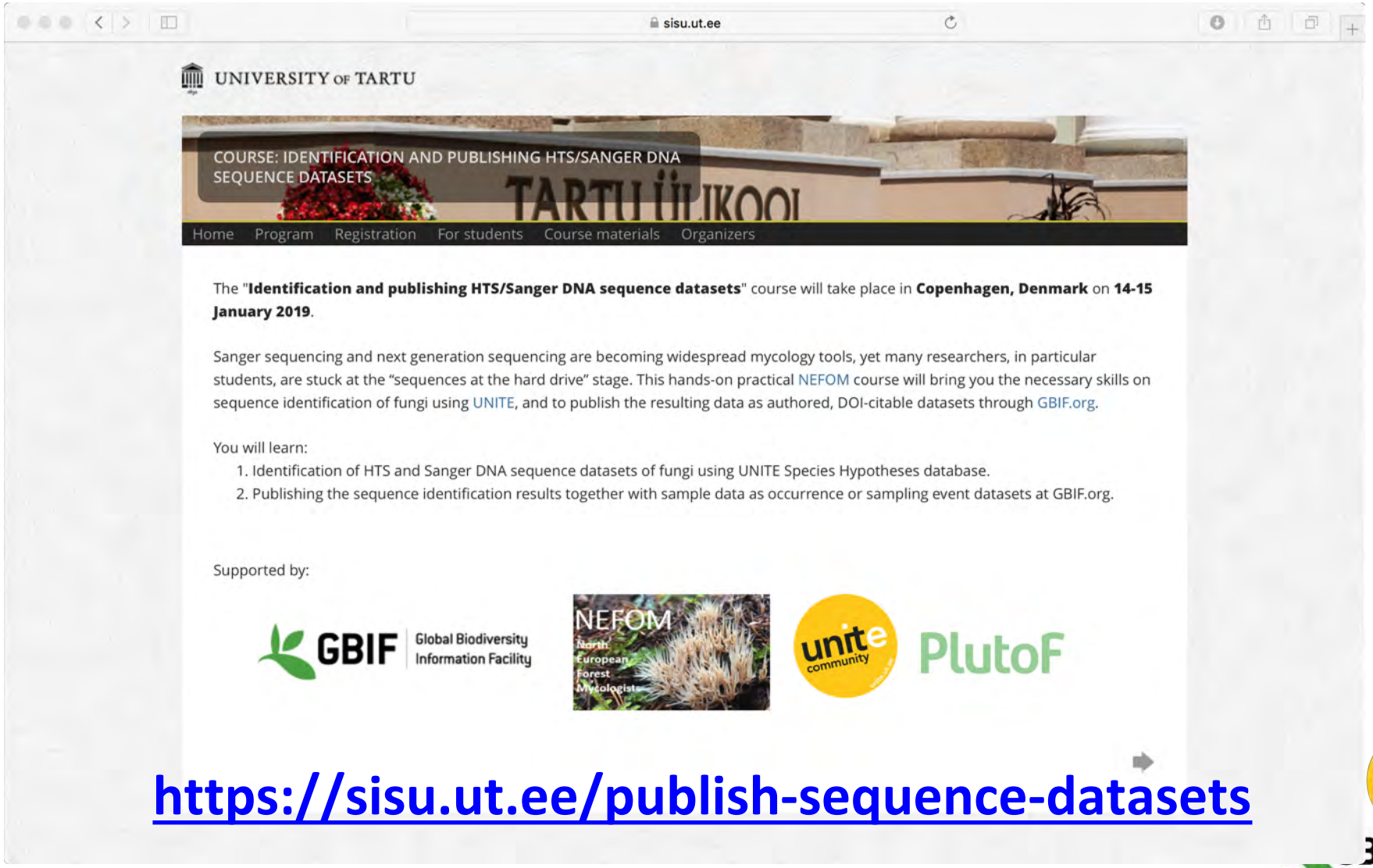
Until a few weeks ago, the GBIF backbone taxonomy fit snugly within a traditional model, classifying and ranking organisms' names using the system that Carl Linnaeus first outlined in *Systema Naturae* in 1735. By combining name-based information from dozens of different authoritative sources like the 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

GBIF - DANISH BIOWIDE PROJECT UPLOADED FIRST DNA DATASET WHERE SPECIES ARE COMMUNICATED VIA UNITE SH DOI



GBIF: COURSE IN COPENHAGEN JANUARY 2019

IDENTIFICATION AND PUBLISHING HTS/SANGER DNA SEQUENCE DATASETS



https://sisu.ut.ee/publish-sequence-datasets

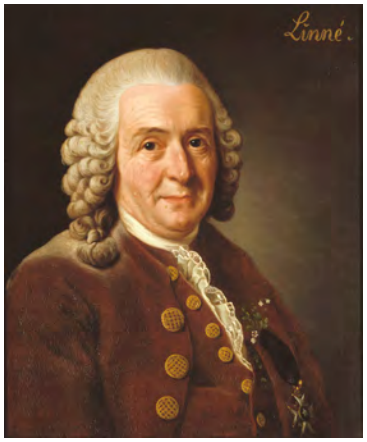
EXISTING DATA MODEL AND DNA EVIDENCE

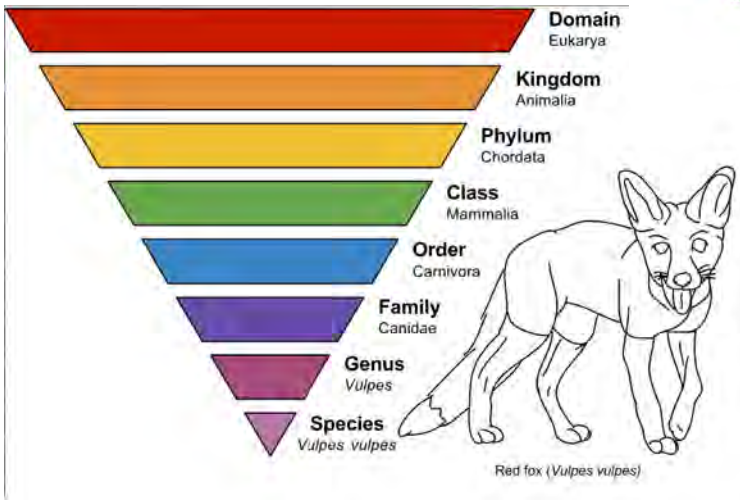

```
{
    "associatedSequences": "KP194575",
    "basisOfRecord": "MATERIAL_SAMPLE"
    "country": "Denmark"
    "dataGeneralizations": "100% identity match with species hypothesis"
    "decimalLatitude": "56.06375636169536"
    "decimalLongitude": "9.7318620682298 27"
    "eventDate": "2014-10-16"
    "eventID": "BIOWIDE-ES066"
    "extensions":
        "http://data.ggbn.org/schemas/ggbn/terms/Amplification":
            "0":
                "consensusSequence": "TAATCTCTCAACTTTTGGTCTTTTGGTTTCTTCTTAACCCAAGGCCTGGAGTTTGGATGGTGGAGGTGTGCTGG
                "marker": "ITS2"
    "footprintWKT": "POLYGON((9.731601299999966 56.0639614,9.731487457013827 56.063607736393486,
    "id": "14193"
    "kingdom": "Fungi"
    "locality": "Odderholm"
    "nameAccordingTo": "http://dx.doi.org/10.15156/BIO/SH221291.07FU"
    "occurrenceID": "14193"
    "preparations": "DNA extract"
    "recordedBy": "Tobias Frøslev"
    "samplingProtocol": "eDNA sampling from soil"
    "scientificName": "SH221291.07FU"
    "taxonConceptID": "SH221291.07FU"
```



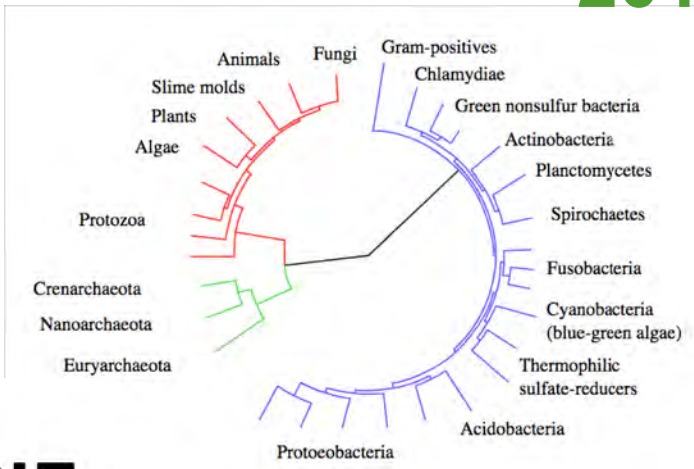
```
    "scientificName": null,
    "sex": null,
    "stateProvince": "Queensland",
    "taxonID": "BOLD:AAF9888"
}
```

1753

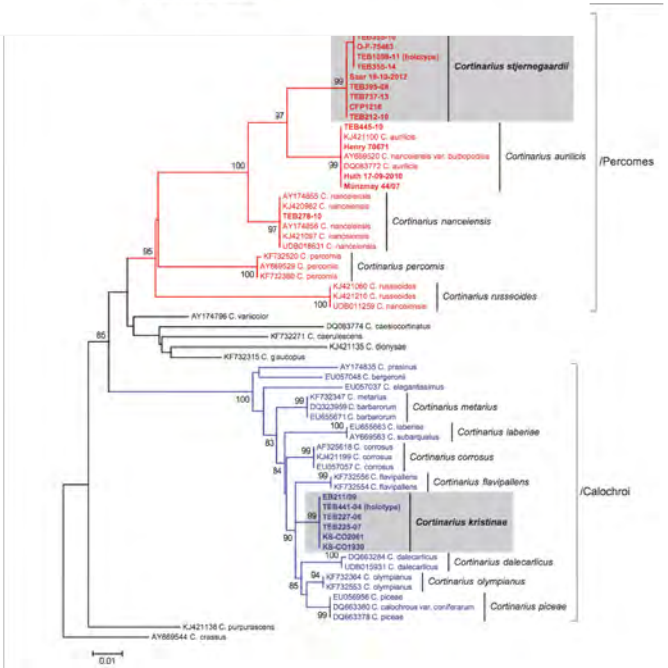




2018

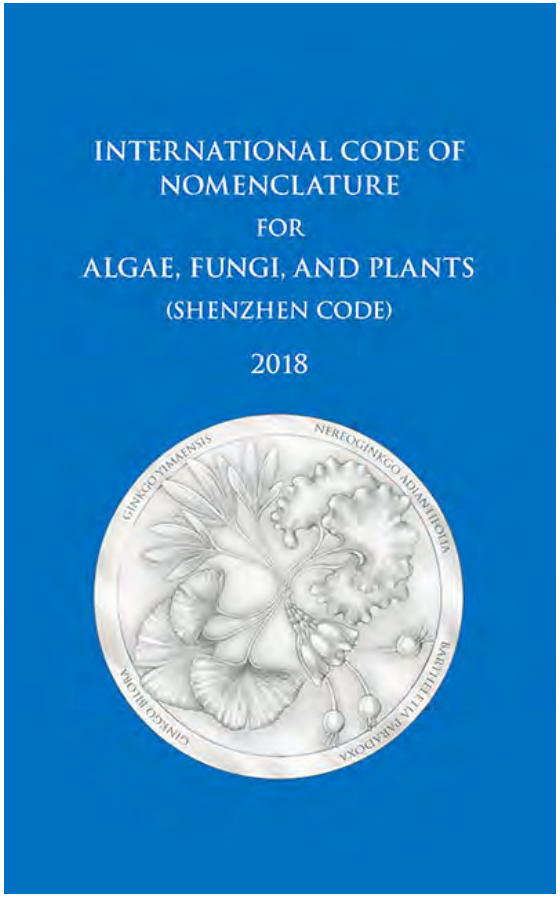


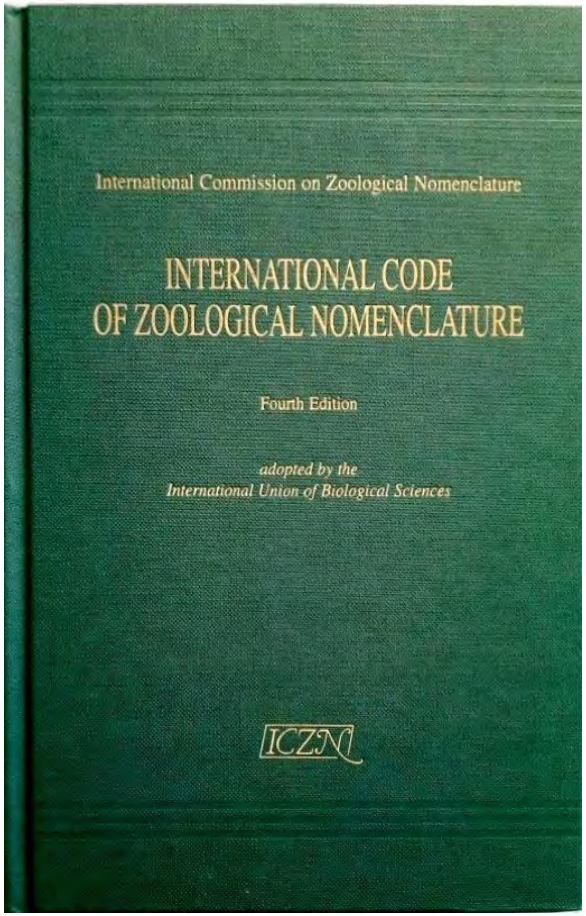
GBIF



GBIF

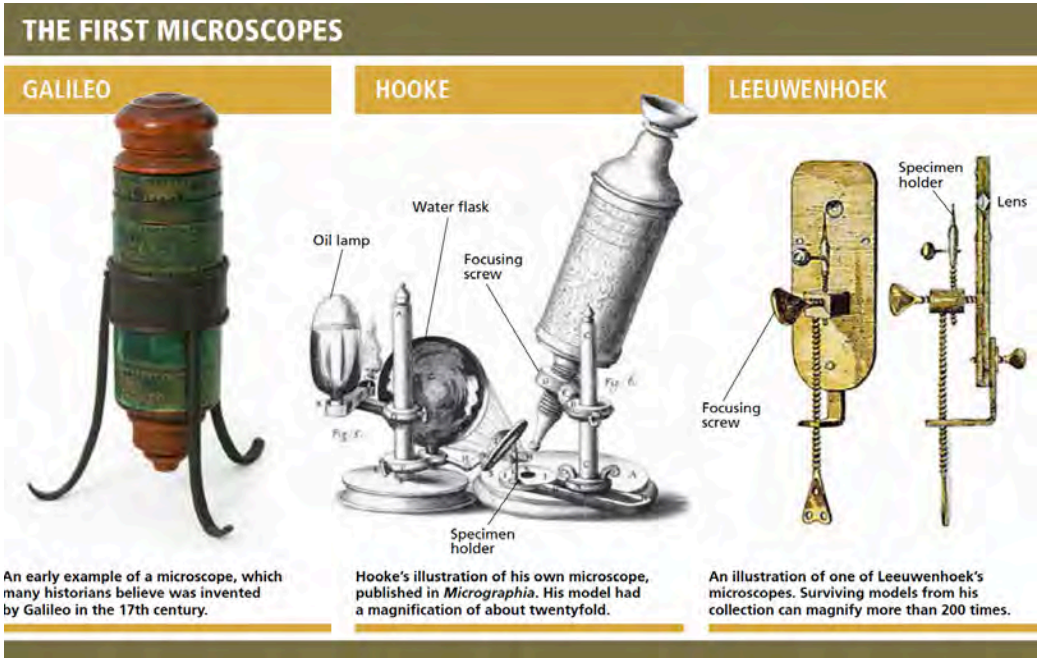
LATIN IS RULED BY THE CODES





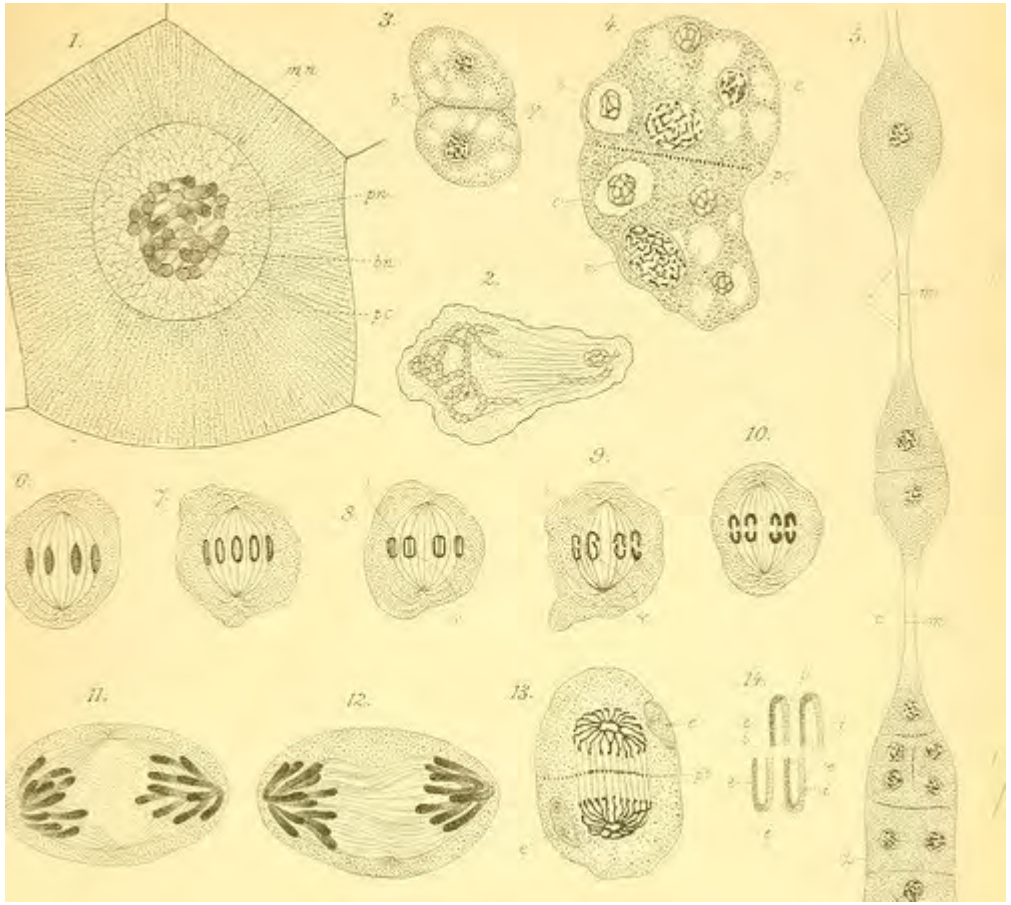
GBIF

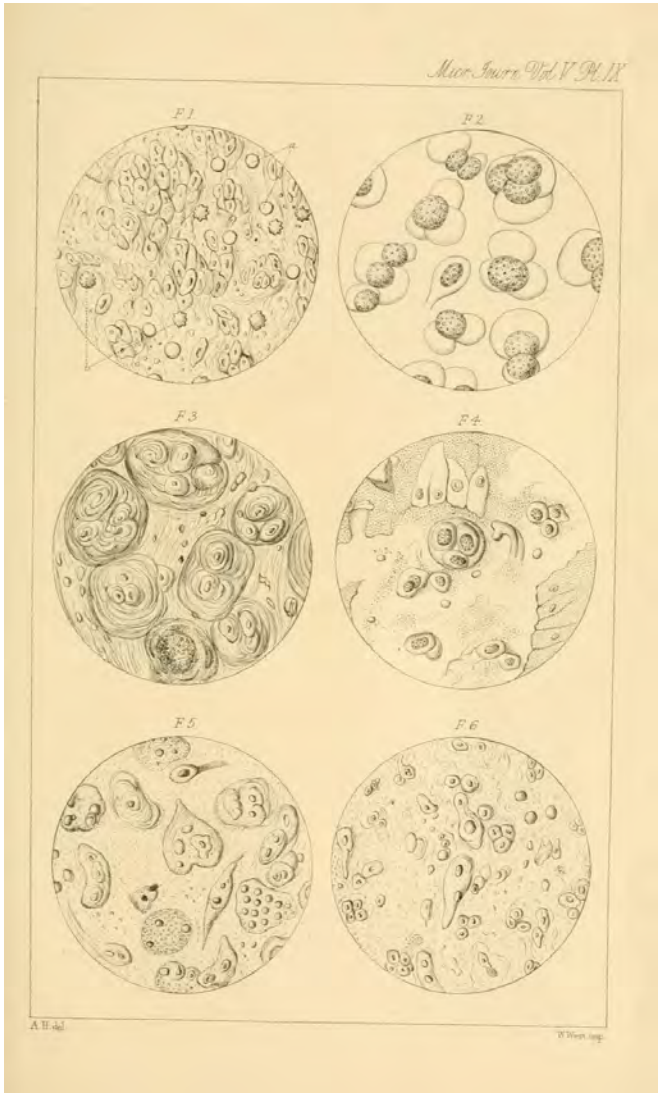
THINGS CHANGE



Wikipedia; Museo Galileo/Institute and Museum of the History of Science; Wellcome Library, London; Granger/NYC; Discover, ABI, Roche

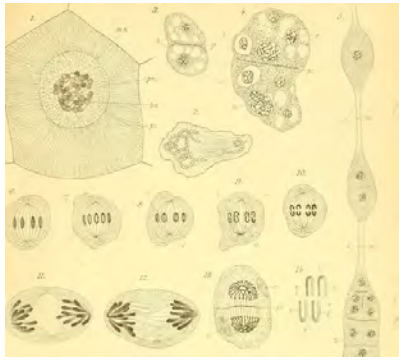
DNA IS NEW (?) MICROSCOPY





Quarterly journal of microscopical science
1853, 1857

MULTIPLE WAYS TO DOCUMENT BIODIVERSITY



Latin: Linnaean nomenclature

OTU codes

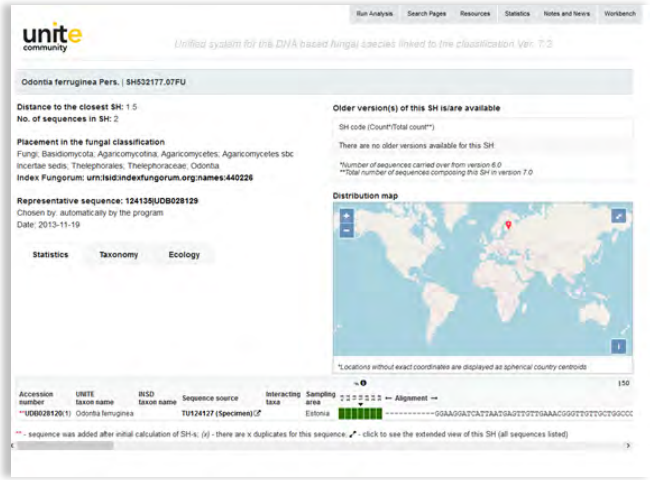
GBIF **Global Biodiversity** Information Facility

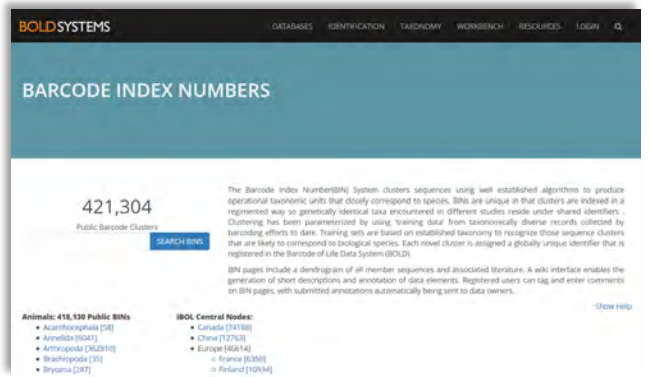
Images: Noah Strycker, Biodiversity Heritage Library, Dmitry Schigel

GBIF

GOING BEOYND FUNGI

BOLD: Trondheim 2019	5.7M
NCBI: EBI and ELIXIR	4Bn
SILVA – bacterial OTUs	6M
R syst – Diatoms	??

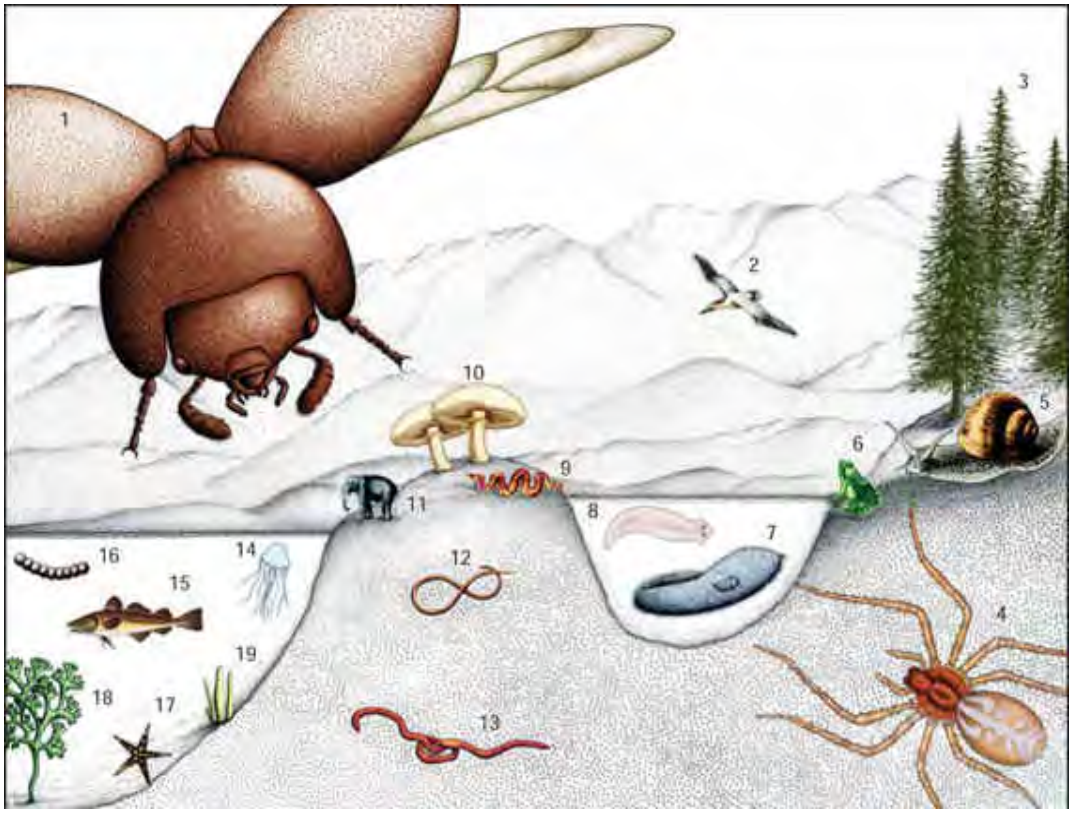




Suggest **key reference libraries of OTUs**:

New names -> indexing dark biodiversity

THANKS



Identification and publishing HTS/Sanger DNA sequence datasets

Copenhagen, Denmark
14–15 January 2019

GBIF Global Biodiversity Information Facility

unite community

PlutoF

https://sisu.ut.ee/publish-sequence-datasets

GBIF