Informatics and data products

Developments and plans

Tim Robertson, Andrea Hahn GBIF Secretariat

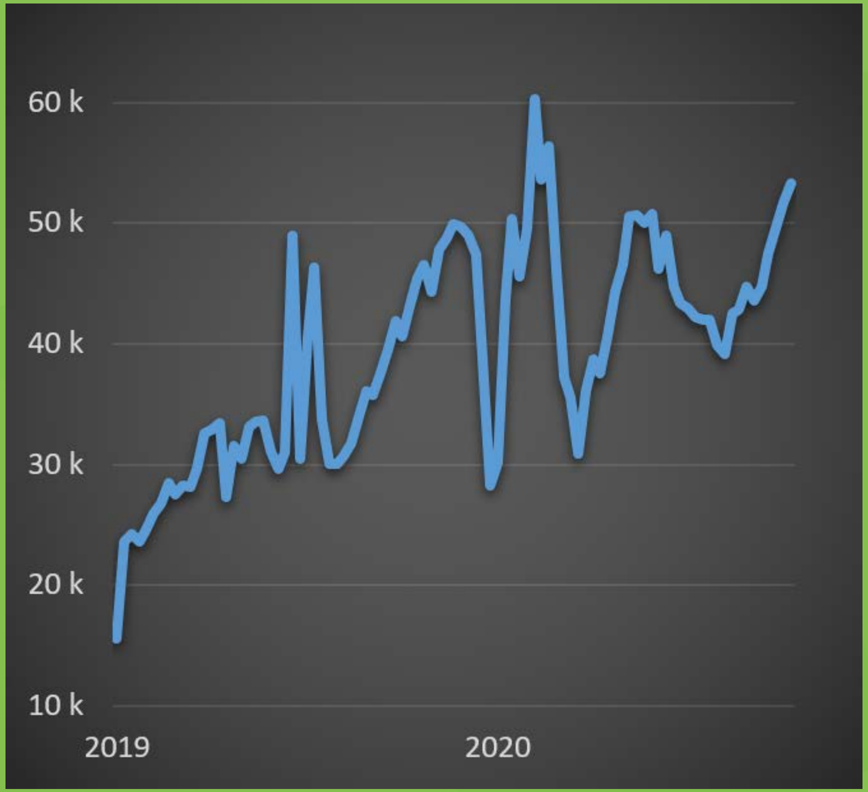
GBIF | Global Biodiversity Information Facility

GB27

Photo of *Aplysina fistularis* by Mike G. Rutherford, via iNaturalist (CC BY-NC 4.0)

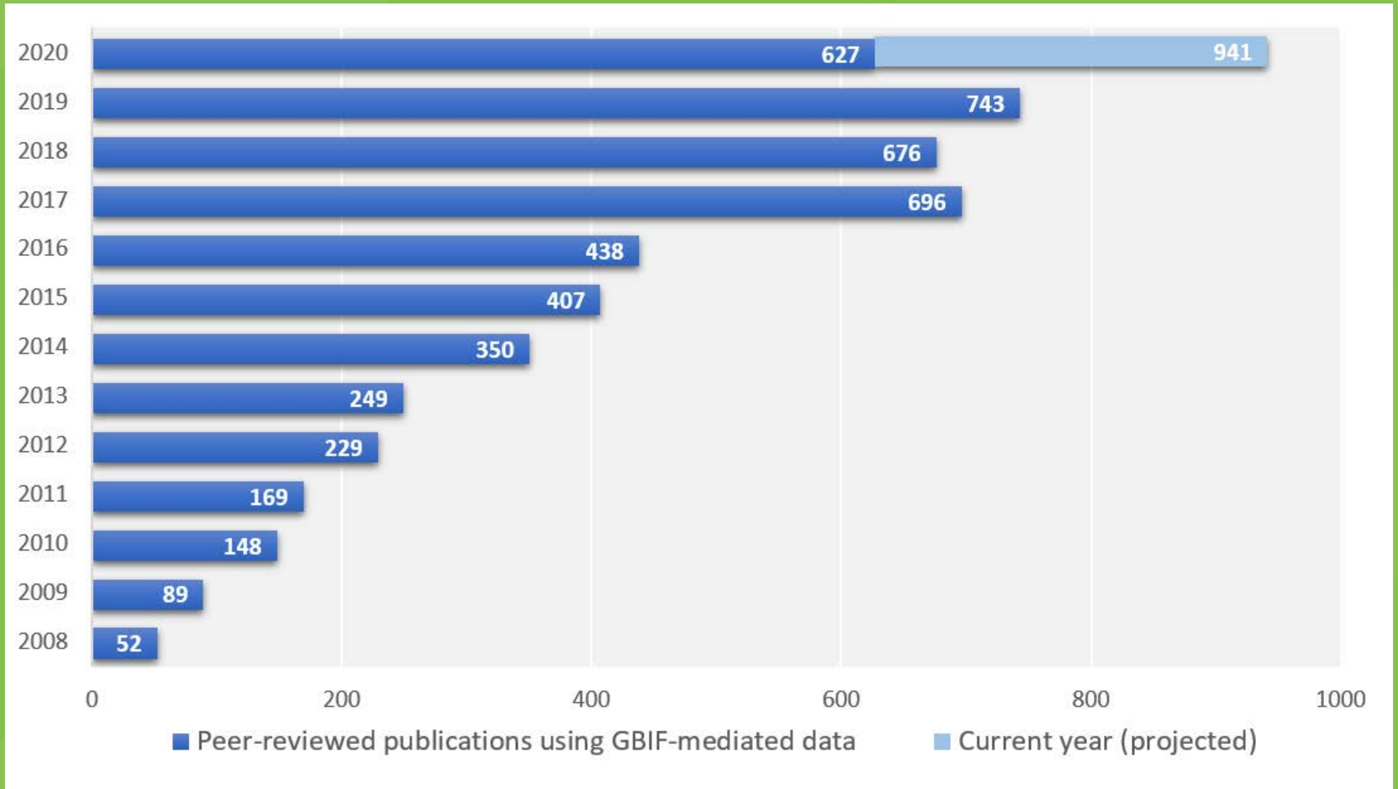
GROWTH AND DATA PUBLISHING

GROWTH IN USE



Users of GBIF.org per week

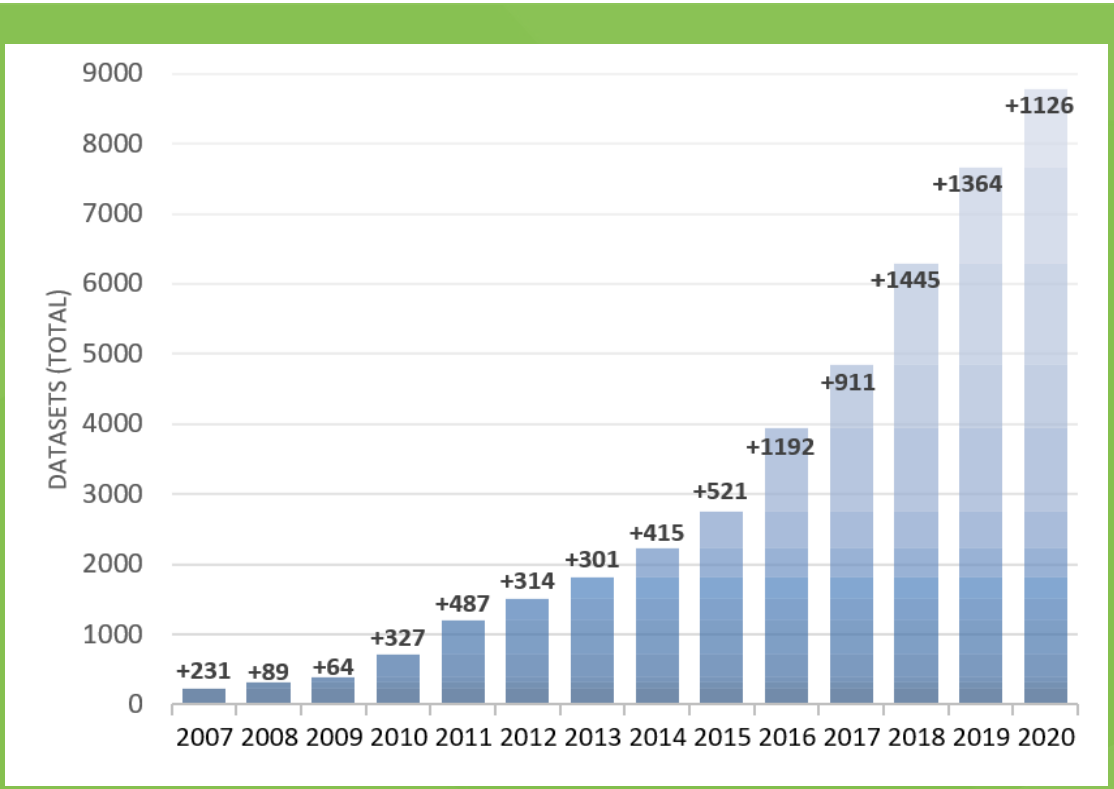
~ 170k per month now (+20k over 2019)



status: Aug 2020

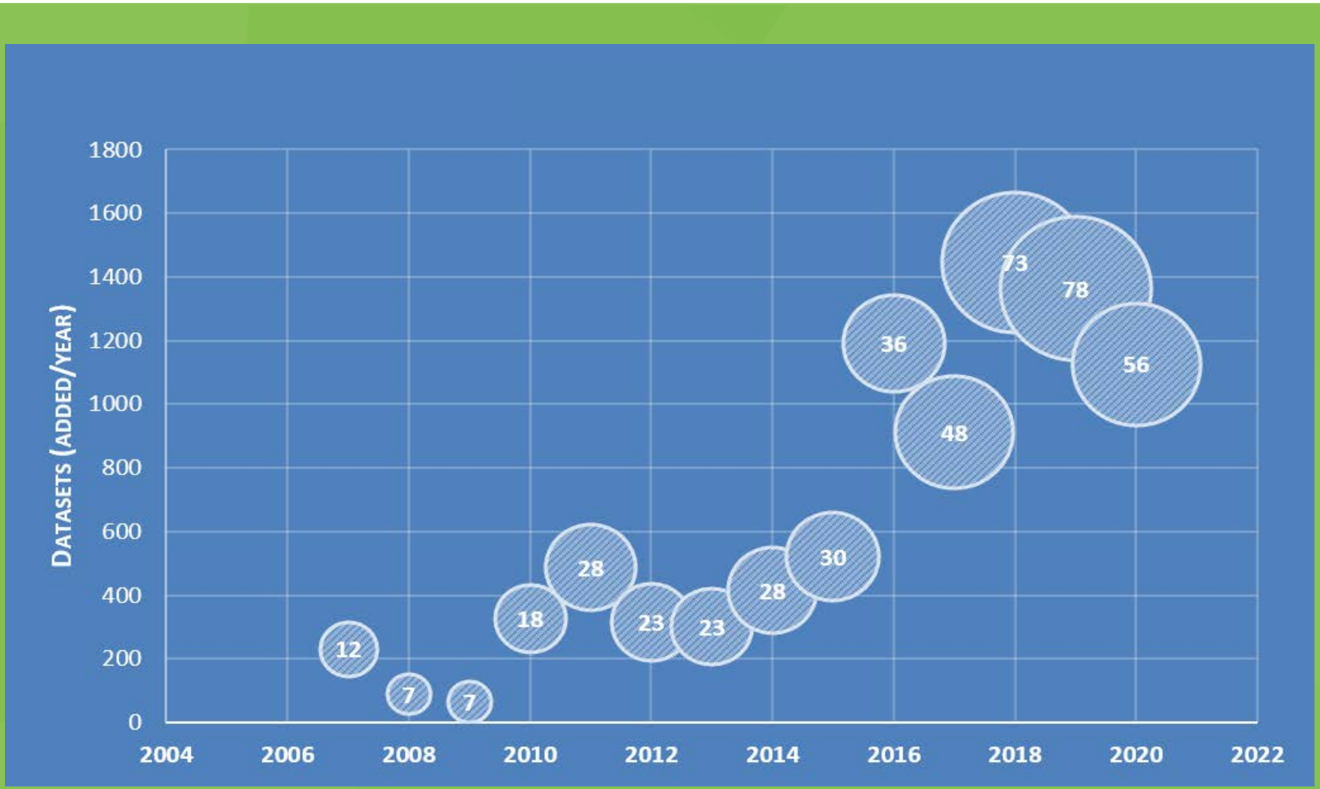
Publications citing use of GBIF-mediated data

IPT (INTEGRATED PUBLISHING TOOLKIT)



not showing: 2385 datasets from UMS PatriNat (2020)

Datasets published through IPT installations, cumulative



status: Sept 2020

newly mobilized IPT datasets per year

circles: number of countries involved

https://tinyurl.com/tdwg-2020-IPT-BPS-report

NAMES INFRASTRUCTURE

In partnership with the Catalogue of Life

MODERNIZING THE CATALOGUE OF LIFE

Complete and completing (Nov 2020)

- New infrastructure deployed to manage data publication and assembly of the taxonomy
  - A new data exchange standard (COL Data Package)
  - An evolution of GBIF ChecklistBank -> COL ChecklistBank
  - A workbench for editors to review and assemble
  - A new suite of APIs and R-based library (rOpenSci)
  - Developed in collaboration, hosted by GBIF
- Deployment of public interfaces
  - New COL website (November)
  - Migration of historical annual checklists (2006-2019)

Upcoming

- Expand the taxonomy with content necessary for GBIF
  - Automate reports on gaps
- Semi-automate assembly (with review)
- Integrate with GBIF.org
  - New backbone end 2020

New website: https://www.catalogue.life/ (temporary URL)
COL ChecklistBank: https://data.catalogue.life/ (temporary URL)

Catalogue of Life GBIF

Communities

Prepare and curate checklists

Data Publishing

COL Repository or other web repositories

ChecklistBank

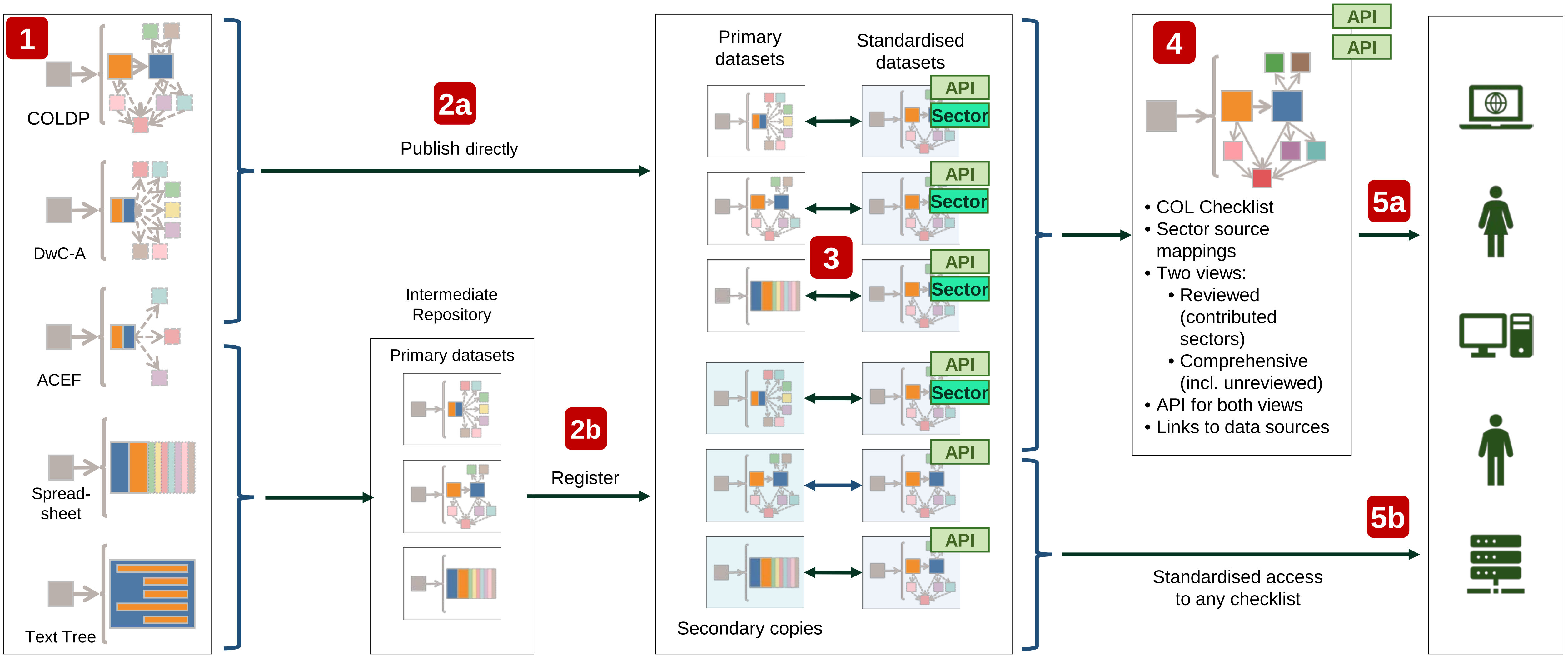
Storage and access to checklists

Workbench

Construct and publish integrated checklist

Users

Web and machine access



- COL Checklist
- Sector source mappings
- Two views:
  - Reviewed (contributed sectors)
  - Comprehensive (incl. unreviewed)
- API for both views
- Links to data sources

Standardised access to any checklist

Catalogue of Life GBIF

GBIF.ORG INFRASTRUCTURE

NEW GBIF.ORG FILTERS

Presence / Absence filter

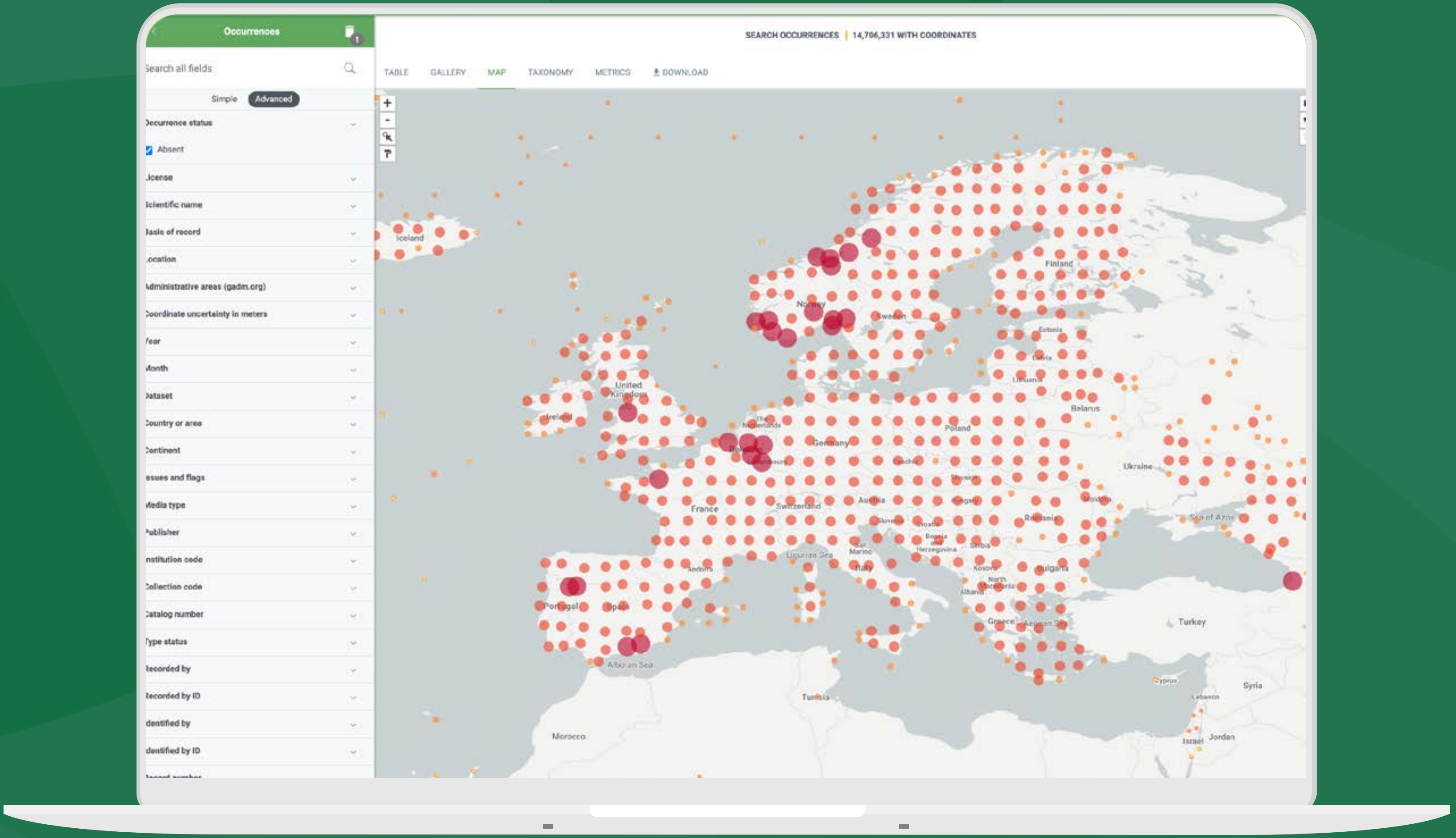
- All records declared as present or absent
- 14,7 million records of absence

Search by person identifiers

- Recorded by or identified by
- Variety of identifier schemas (e.g. from ORCiD, WikiData)
- Working for TDWG Attribution group

Search by Administrative Areas

- Uses 3 levels from the GADM.org database e.g. country / state / county
- Using the API can report counts by e.g. county level



Occurrences

Search all fields

Simple | Advanced

Occurrence status

☑ Absent

License

Scientific name

Basis of record

Location

Administrative areas (gadm.org)

Coordinate uncertainty in meters

Year

Month

Dataset

Country or area

Continent

Issues and flags

Media type

Publisher

Institution code

Collection code

Catalog number

Type status

Recorded by

Recorded by ID

Identified by

Identified by ID

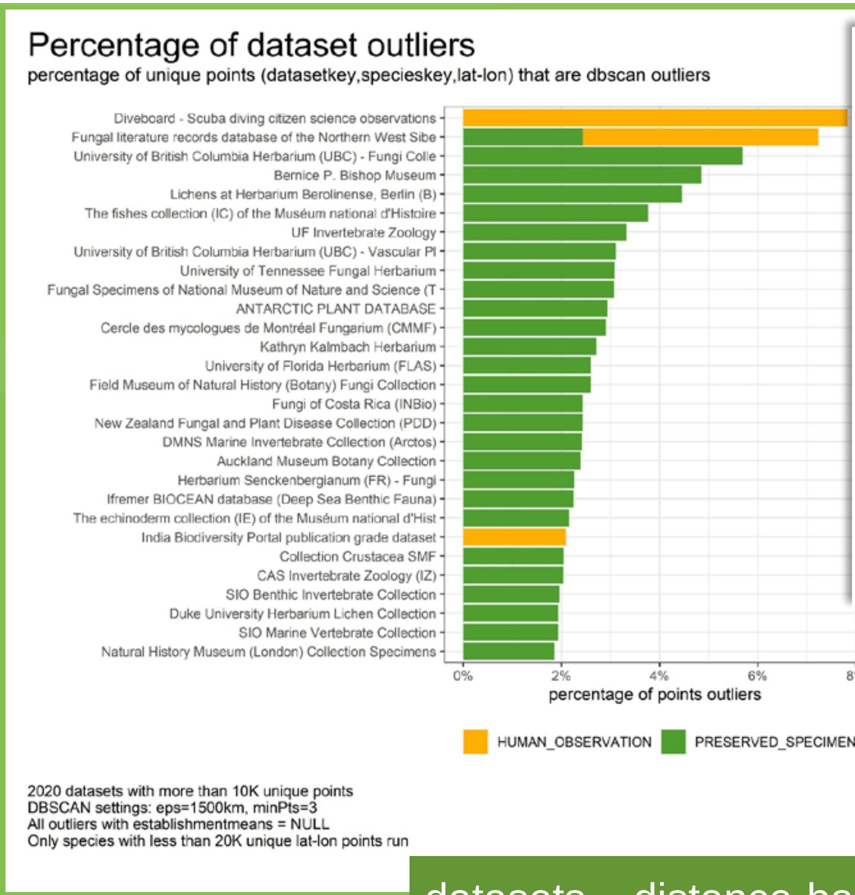
Record number

SEARCH OCCURRENCES | 14,706,331 WITH COORDINATES

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

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

DATA QUALITY - OUTLIER DETECTION

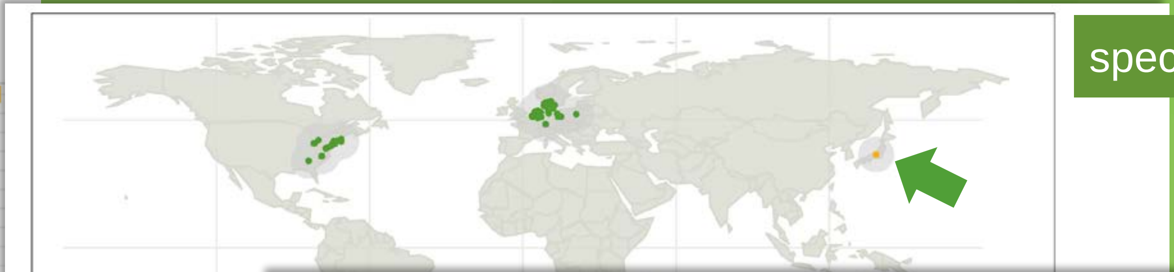


Percentage of dataset outliers

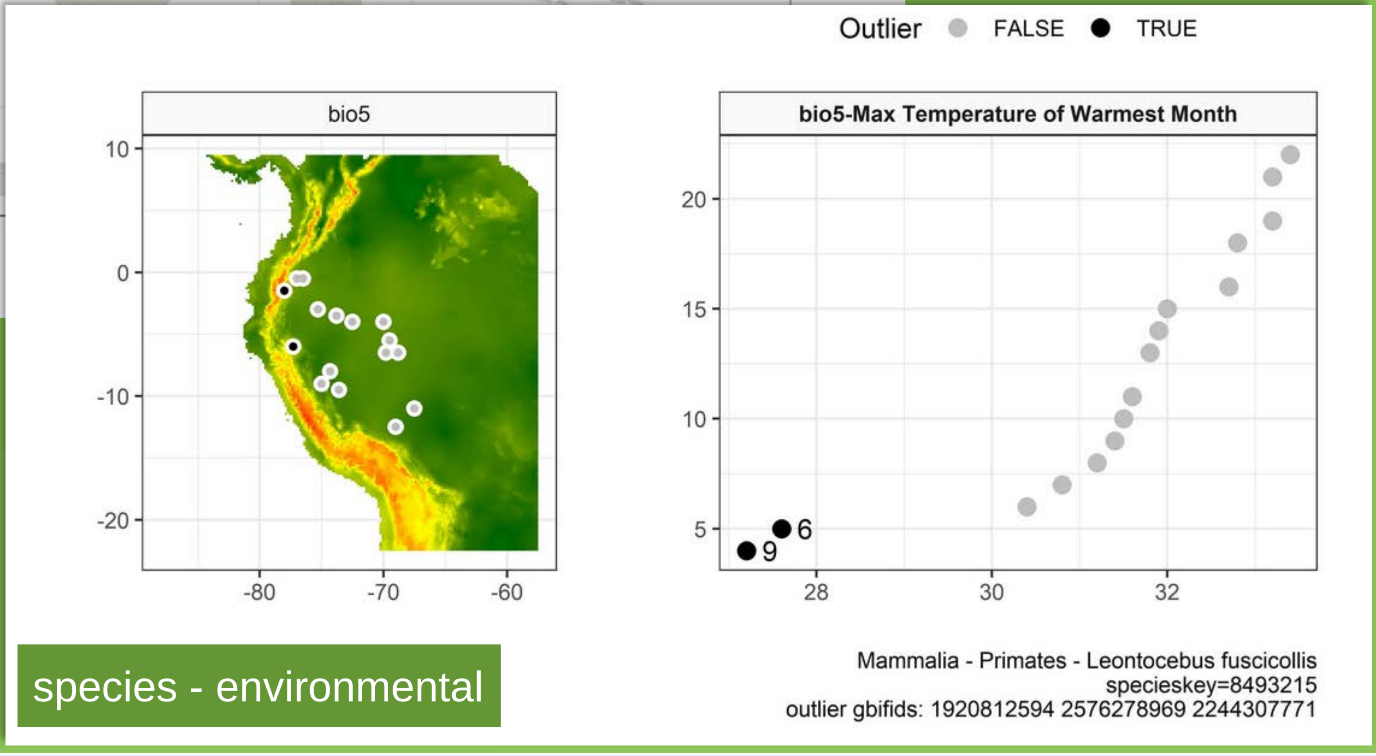
percentage of unique points (datasetkey,specieskey,lat-lon) that are dbscan outliers

2020 datasets with more than 10K unique points
DBSCAN settings: eps=1500km, minPts=3
All outliers with establishmentmeans = NULL
Only species with less than 20K unique lat-lon points run

datasets – distance-based



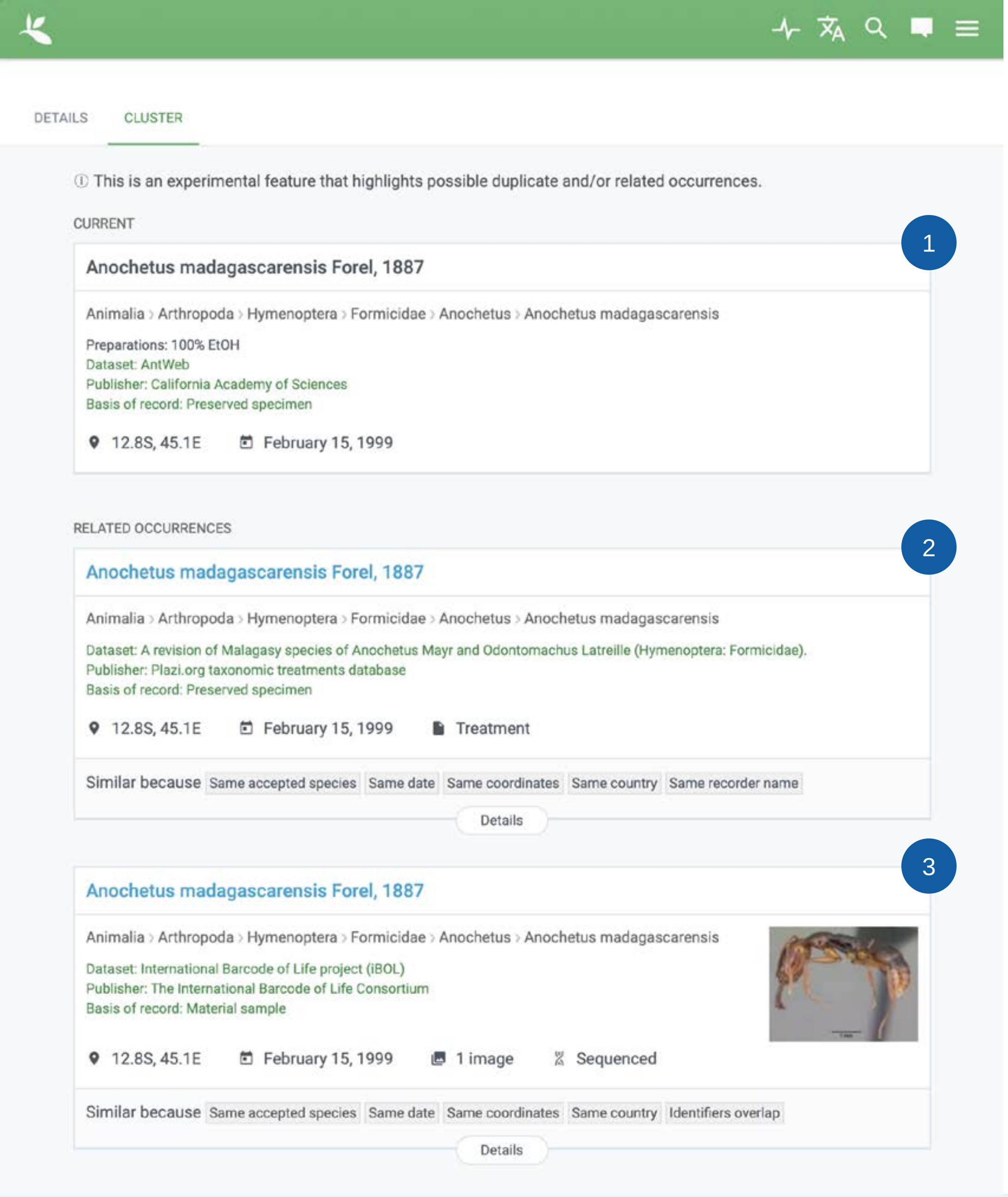
species – distance-based



Mammalia - Primates - Leontocebus fuscicollis
specieskey=8493215
outlier gbifids: 1920812594 2576278969 2244307771

species - environmental

https://data-blog.gbif.org/post/outlier-detection-using-dbscan/



DETAILS CLUSTER

ⓘ This is an experimental feature that highlights possible duplicate and/or related occurrences.

CURRENT

1

Anochetus madagascarensis Forel, 1887

Animalia › Arthropoda › Hymenoptera › Formicidae › Anochetus › Anochetus madagascarensis

Preparations: 100% EtOH
Dataset: AntWeb
Publisher: California Academy of Sciences
Basis of record: Preserved specimen

12.8S, 45.1E February 15, 1999

RELATED OCCURRENCES

2

Anochetus madagascarensis Forel, 1887

Animalia › Arthropoda › Hymenoptera › Formicidae › Anochetus › Anochetus madagascarensis

Dataset: A revision of Malagasy species of Anochetus Mayr and Odontomachus Latreille (Hymenoptera: Formicidae).
Publisher: Plazi.org taxonomic treatments database
Basis of record: Preserved specimen

12.8S, 45.1E February 15, 1999 Treatment

Similar because Same accepted species Same date Same coordinates Same country Same recorder name

Details

3

Anochetus madagascarensis Forel, 1887

Animalia › Arthropoda › Hymenoptera › Formicidae › Anochetus › Anochetus madagascarensis

Dataset: International Barcode of Life project (iBOL)
Publisher: The International Barcode of Life Consortium
Basis of record: Material sample

12.8S, 45.1E February 15, 1999 1 image Sequenced

Similar because Same accepted species Same date Same coordinates Same country Identifiers overlap

Details

Clustering

- Example shows a specimen record (1), a taxonomic revision (2), a DNA sequence (3)
- Grouped using location, identification, date, local identifiers and type status (so far) and fuzzy matching
- Stored as annotations against the records
- Interface allows easy comparison of the fields on the record (details)
- Purpose
  - Richer model of related information for researchers
  - Transferring information between institutional databases (images, sequences, new identifications) which may reduce effort spent
  - Aid in deduplicating data
  - Known to contribute towards data quality improvements
  - Contibution towards the Extended Digital Specimen / Open Digital specime concept (US and DiSSCo initiatives)

https://www.gbif.org/occurrence/865079977/cluster

GBIF

USING GBIF IN CLOUD ENVIRONMENTS

Growing interest in big data tools

- Google Big Query increasingly being used
- Google / Amazon / Microsoft etc clouds being used by several partners
- Opportunities with science computing clouds like the European Open Science Cloud
- Simplify research, ease integrations with other datasets

New download formats

- Apache Avro format
- Bionomia-specific format to reduce computation on client side
- Services to easily move GBIF downloads onto cloud environments (underway)

GBIF data as public datasets

- Strong interest expressed on the GBIF community forum
- Requires new citation service (underway)
- Exploring possibilities during 2021

https://discourse.gbif.org/t/gbif-exports-as-public-datasets-in-cloud-environments/1835

COLLECTIONS CATALOGUE

COLLECTIONS CATALOGUE GLOBAL CONSULTATION

- Part of the EU Synthesys+ project and run as an initiative under the Alliance
- A virtual workshop held over 2 weeks
- Two two-hour preparatory webinars to introduce the Ideas Paper https://doi.org/10.35035/p93g-te47
  - Available in English, French, Spanish and simplified Chinese
- GBIF community forum https://discourse.gbif.org/g/Collections-catalog
- Daily summaries, in multiple languages
- Outcomes document to be drafted

Trinidad National Herbarium, photo 2019 by Maheva Bagard Laursen (CC BY 2.0)

SOME OUTCOMES

- Many communities need information on collections
- Interest in a core set of common data elements
- Each community has specific needs; may document collections at different levels of granularity

- GBIF can:
  - bring together complementary information
  - promote a consistent approach to the common elements
  - help national nodes showcase their collections and the value they offer
  - Use the DOI-based citation and usage tracking to add value to nodes and to individual collection institutions.

- The catalogue may serve as a mechanism to attract funding; global partnerships to maintain biodiversity collections

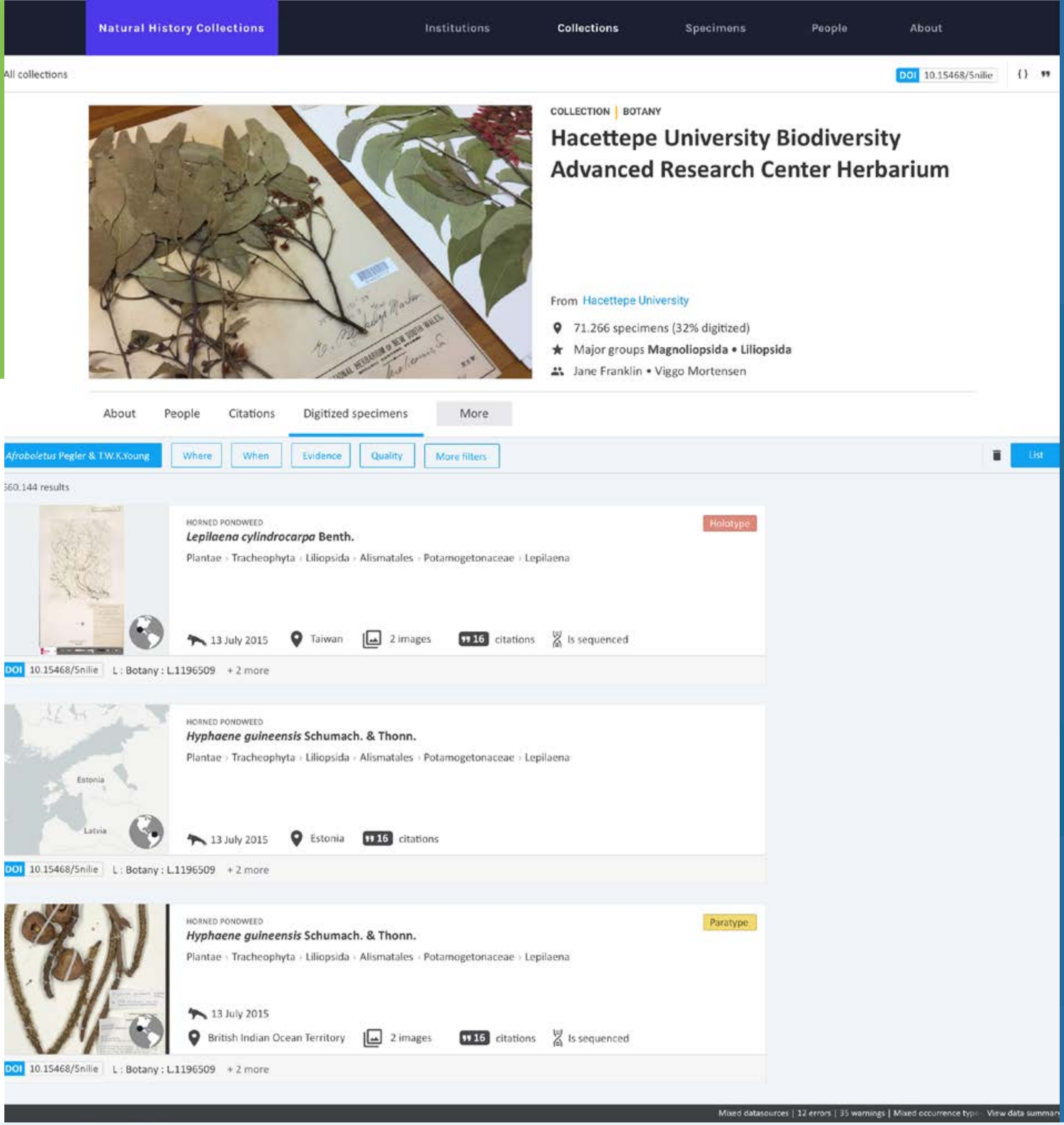
- The GRSciColl Catalogue
  - Holds a basic core suitable to act as a stub for any collection
  - Can act as a first step towards better linking

ENHANCING THE GRSCICOLL COLLECTIONS CATALOGUE

- Monthly synchronisation with Index Herbariorum
- Collaboration with iDigBio
  - Data imported from iDigBio catalogue
  - GRSciColl powering the iDigBio collections catalogue
  - Collaborative editing responsibility
- Early exploration of synchronisation with NCBI BioCollections*
- Linking ~170M specimen-related occurrences to GRSciColl

2021

- Mature the processes around collaborative editing
- Deploy an enhanced collections catalogue portal
- Aim to broaden integrations (Australia)
- Explore a sample profile of the TDWG collections description standard



DOI 10.15468/5nilie

COLLECTION | BOTANY

Hacettepe University Biodiversity Advanced Research Center Herbarium

From Hacettepe University

71.266 specimens (32% digitized)

Major groups **Magnoliopsida • Liliopsida**

Jane Franklin • Viggo Mortensen

560.144 results

HORNED PONDWEED
Lepilaena cylindrocarpa **Benth.**
Plantae › Tracheophyta › Liliopsida › Alismatales › Potamogetonaceae › Lepilaena
Holotype
13 July 2015 · Taiwan · 2 images · 16 citations · Is sequenced
DOI 10.15468/5nilie · L : Botany : L.1196509 · + 2 more

HORNED PONDWEED
Hyphaene guineensis **Schumach. & Thonn.**
Plantae › Tracheophyta › Liliopsida › Alismatales › Potamogetonaceae › Lepilaena
13 July 2015 · Estonia · 16 citations
DOI 10.15468/5nilie · L : Botany : L.1196509 · + 2 more

HORNED PONDWEED
Hyphaene guineensis **Schumach. & Thonn.**
Plantae › Tracheophyta › Liliopsida › Alismatales › Potamogetonaceae › Lepilaena
Paratype
13 July 2015
British Indian Ocean Territory · 2 images · 16 citations · Is sequenced
DOI 10.15468/5nilie · L : Botany : L.1196509 · + 2 more

Mixed datasources | 12 errors | 35 warnings | Mixed occurrence typ... View data summar...

- **NCBI** https://www.ncbi.nlm.nih.gov/biocollections
- **Visual concept of collection catalogue** https://labs.gbif.org/visual-concepts/
- **iDigBio collections powered by GBIF infrastructure** http://beta.idigbio.org/portal/collections **(beta version)**

HOSTED PORTALS

- Introduced at GB26 as part of work to lower technical threshold for participation in GBIF
- Raised interest from many GBIF Participants and partners
- Development work ongoing this year
  - Reworking our APIs (GraphQL-based)
  - Developing the first modules of user interfaces (ReactJS-based)
- Represents the next phase of infrastructure for GBIF
  - Will be consistent with GBIF
  - Will contribute to the envisaged collections catalogue

THANK YOU!

trobertson@gbif.org

ahahn@gbif.org