The Role of Biodiversity Informatics In the Global Biodiversity Information Facility

Dag Endresen | GBIF Node Manager for Norway

Global Biodiversity Information Facility (GBIF)

GBIF Global Biodiversity Information Facility

Seminar | GC University of Lahore, Pakistan | 18th May 2021

GBIF PROVIDES A BIODIVERSITY DATA INFRASTRUCTURE

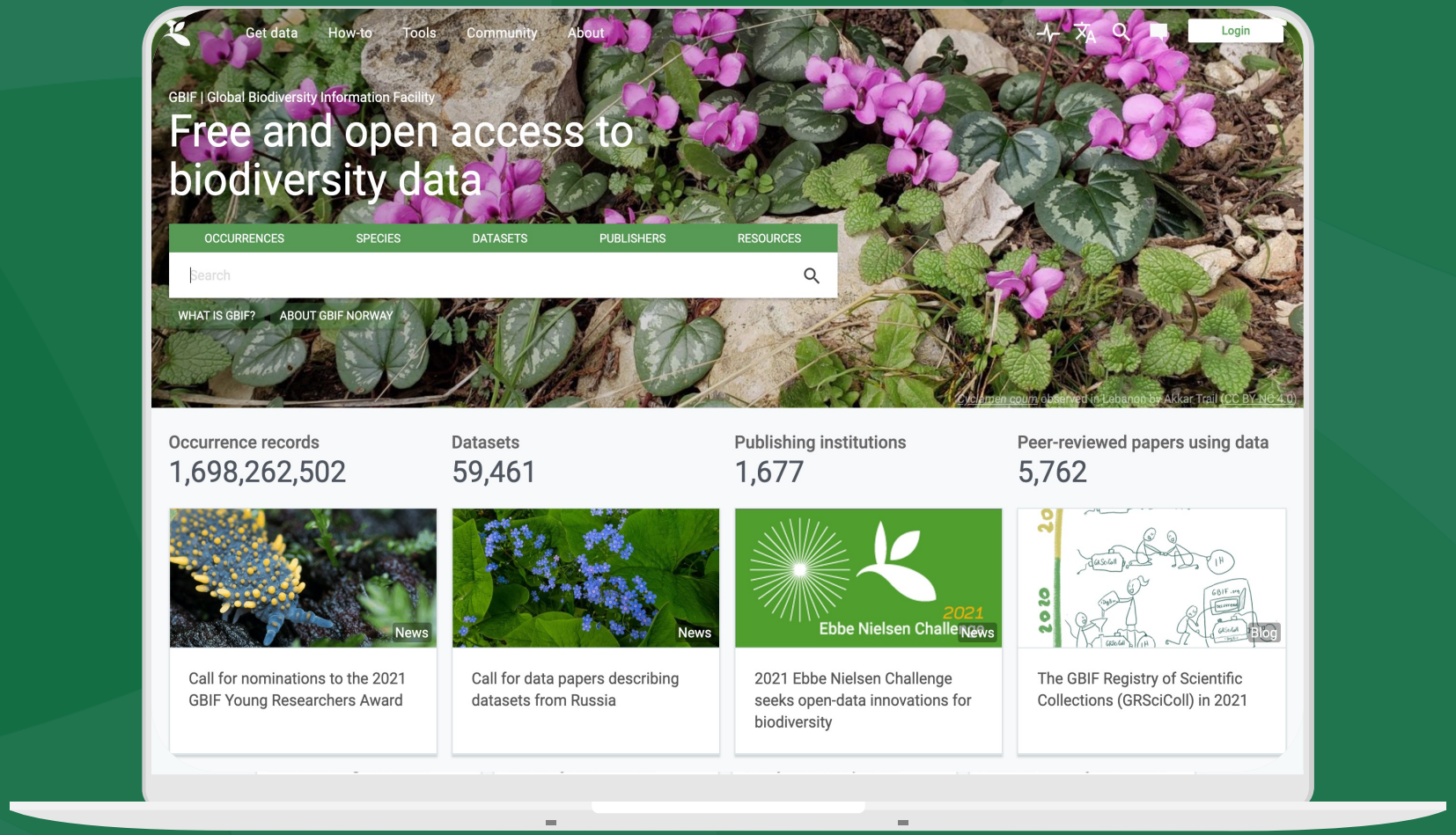
WHAT IS GBIF?

Intergovernmental network and research infrastructure

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

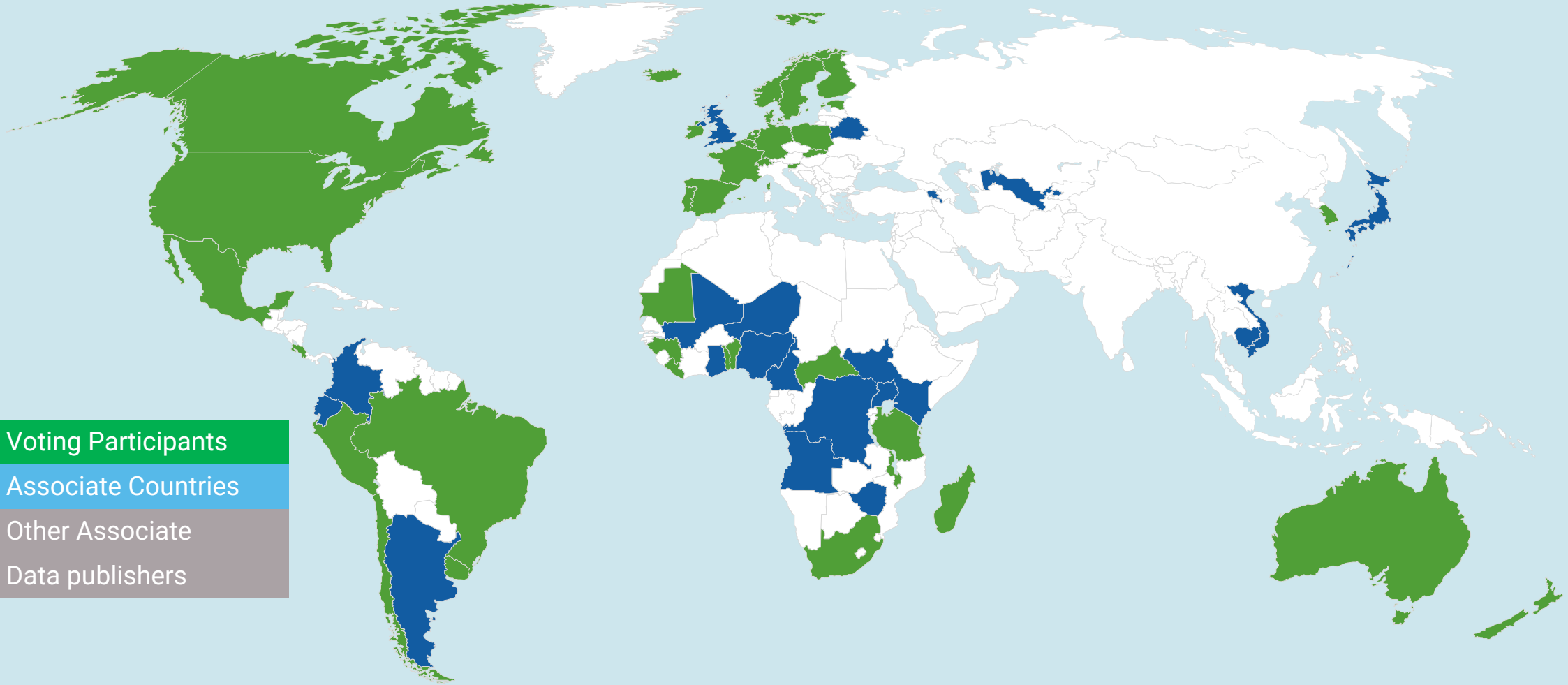
Voluntary collaboration through Memorandum of Understanding (MoU)

Participant nodes, Secretariat in Copenhagen, Denmark



https://www.gbif.org

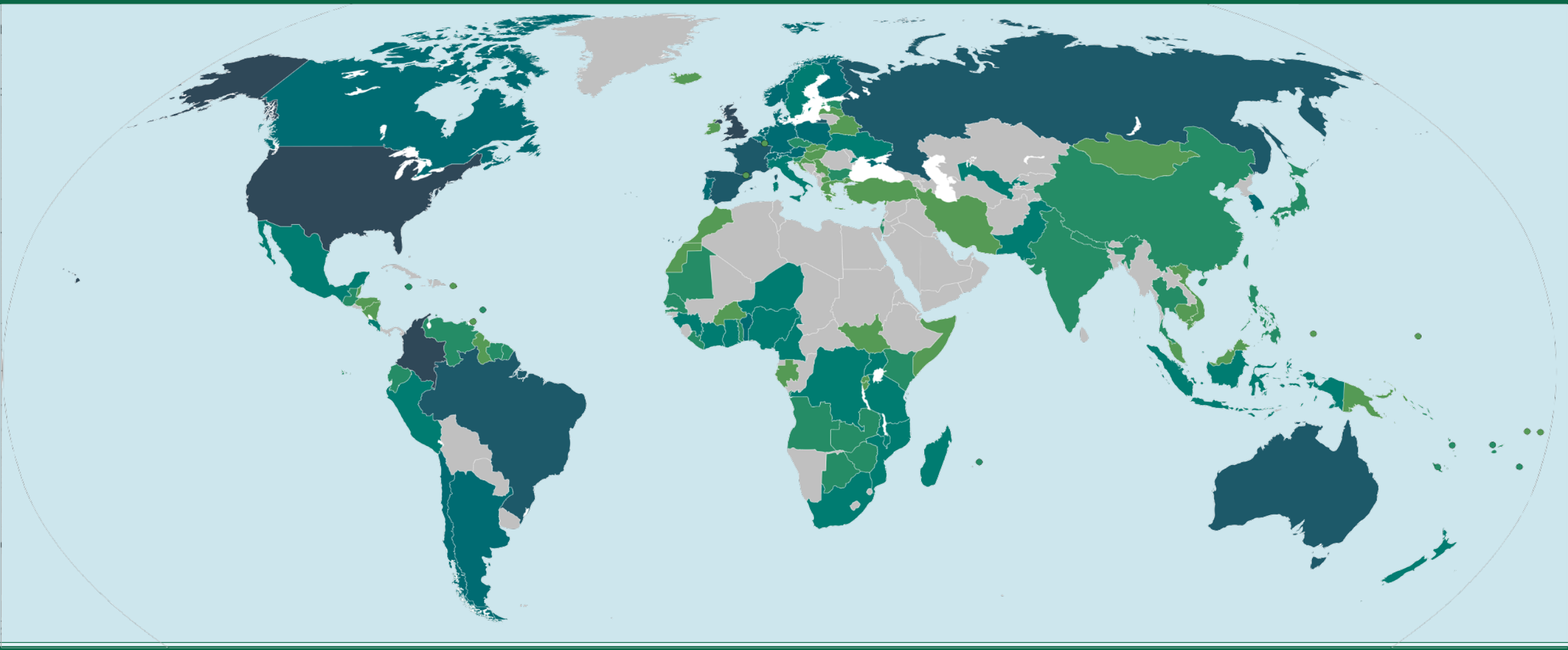
GBIF PARTICIPANT COUNTRIES



https://www.gbif.org/the-gbif-network

GBIF

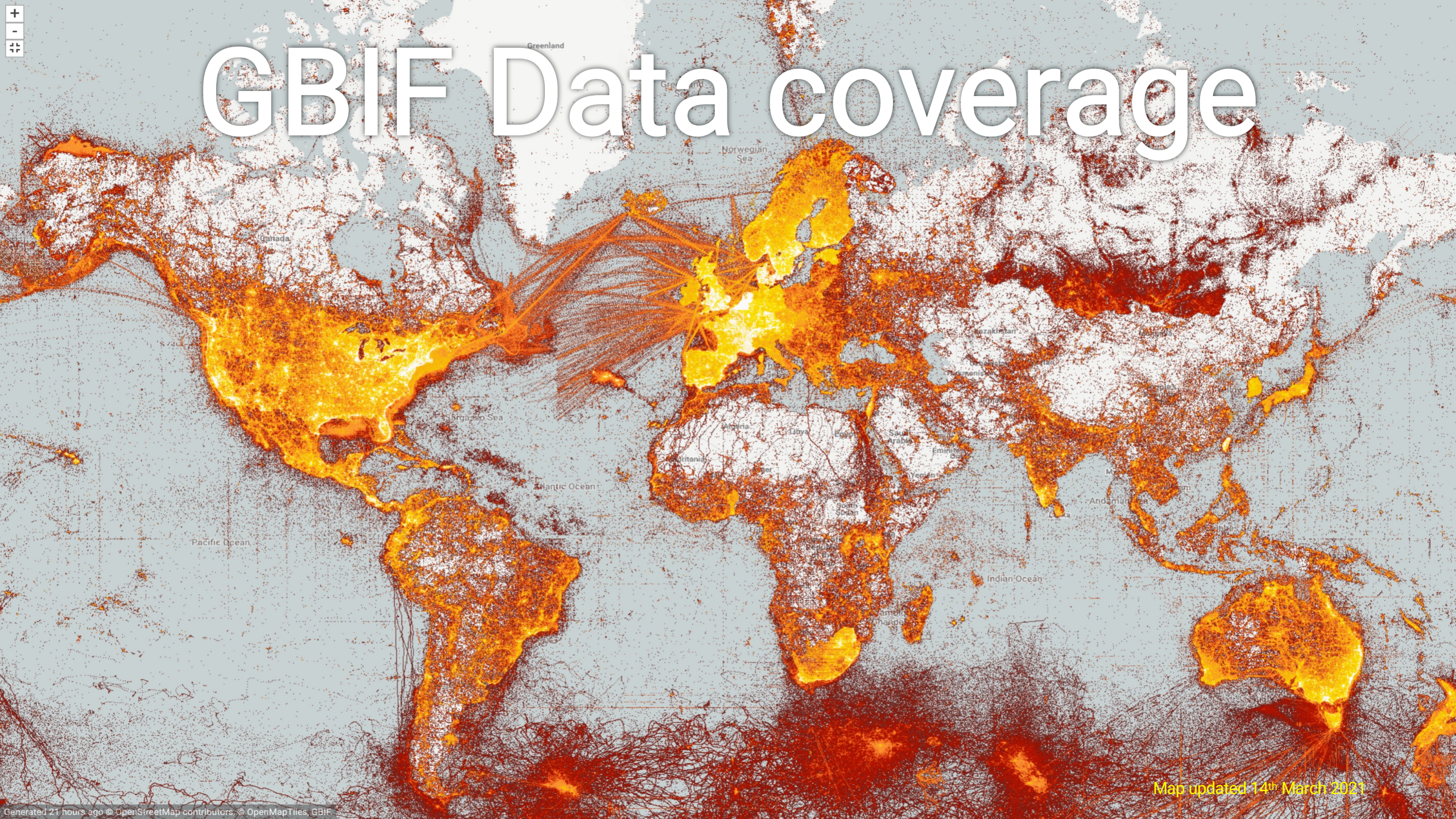
THE GBIF DATA PUBLISHER NETWORK



1677 data publishers; https://www.gbif.org/publisher/search

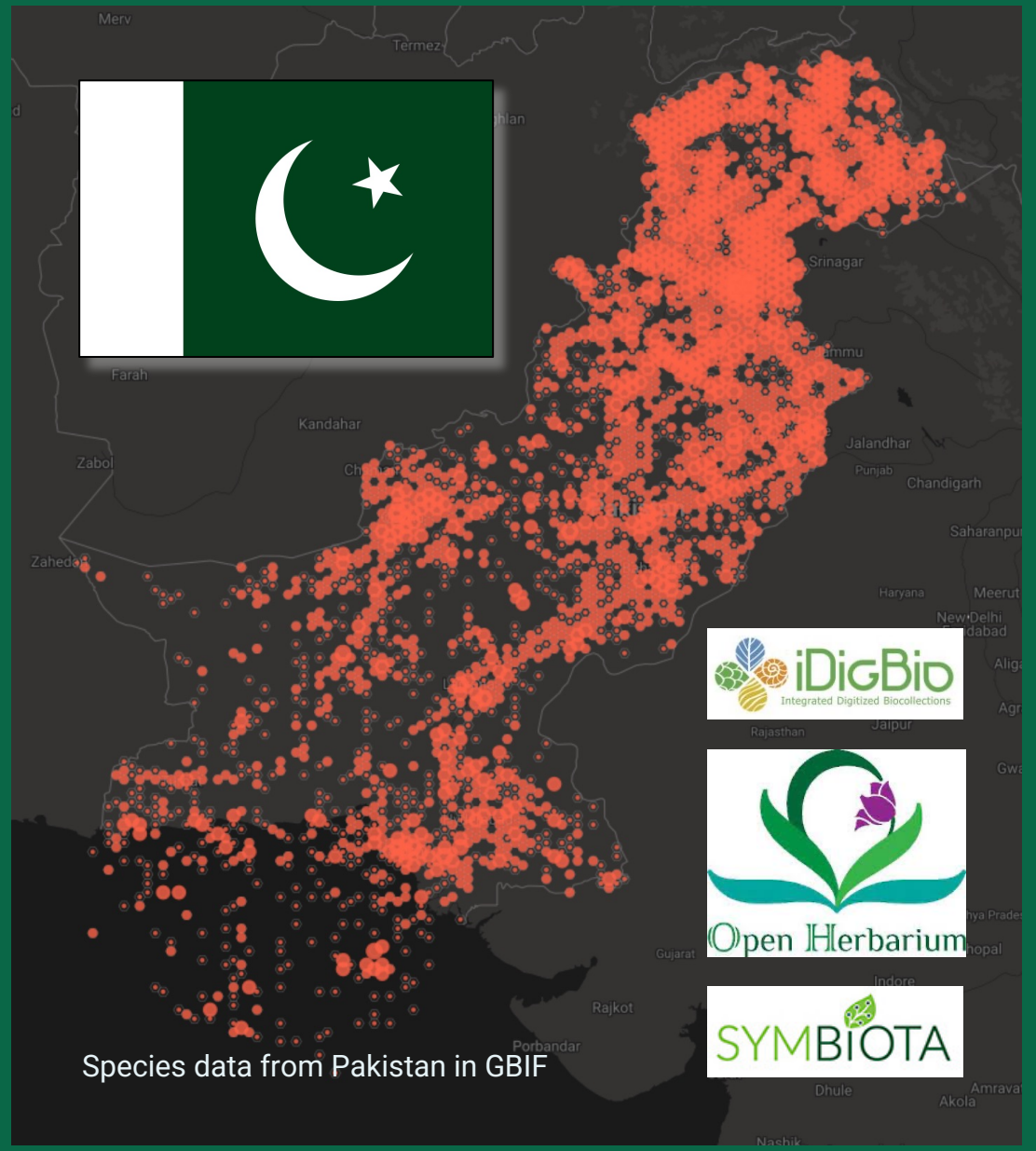
GBIF

GBIF Data coverage



MILESTONES - PAKISTAN AND GBIF

- Pakistan was an associate member of GBIF from August 2001 to 2017.
- March **2006** a GBIF mentor project between Pakistan and **Australia** – installed the **DiGIR** data publishing toolkit, hosted by the Pakistan Museum of Natural History in Islamabad (2 datasets), and the **BioLink** digitisation and collection management database.
- Recent capacity development projects between Pakistan and **USA** (Mary Barkeworth) with datasets hosted from the **OpenHerbarium** (US GBIF node iDigBio) and **Symbiota** collection digitsation and management database (8 datasets). The BIFA program is funded by the Government of **Japan**.



Species data from Pakistan in GBIF

Atlas of Living Australia ala.org.au DIGIR iDigBio Integrated Digitized Biocollections SYMBIOTA GBIF

digitization of collections

Photo: Botany Collection, Algae, Smithsonian National Museum of Natural History Museum, by Chip Clark.

Very few museum specimens are digitized

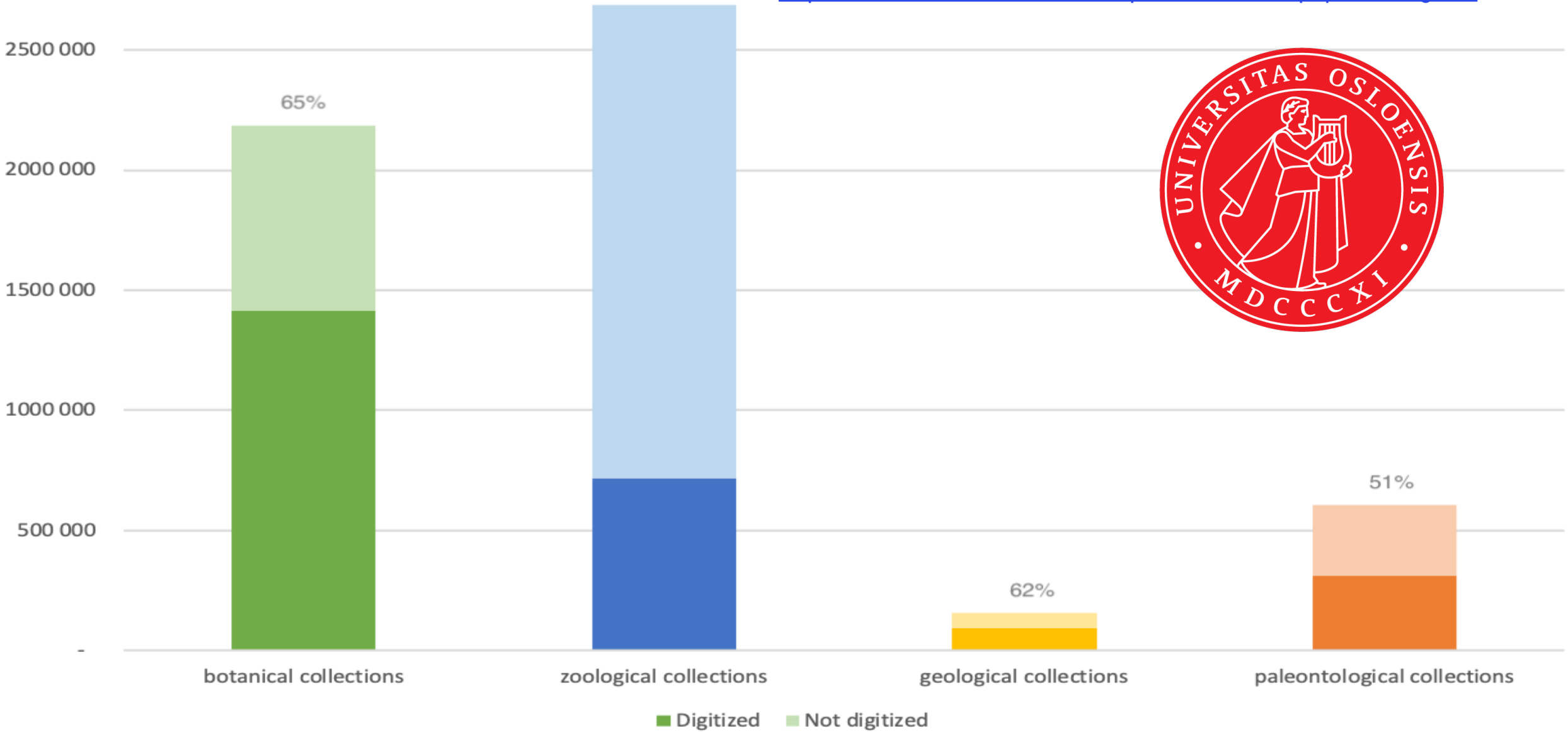
Natural history museum collections worldwide conserve an estimated **1.2 - 3 billion specimens**.
(Ariño 2010; Duckworth *et al*. 1993)

GBIF publishes 1,7 billion records – including **233 million specimens**.

approx. 10% coverage??

DIGITIZED SPECIMENS AT THE UNIVERSITY OF OSLO (APPROX. 47%)

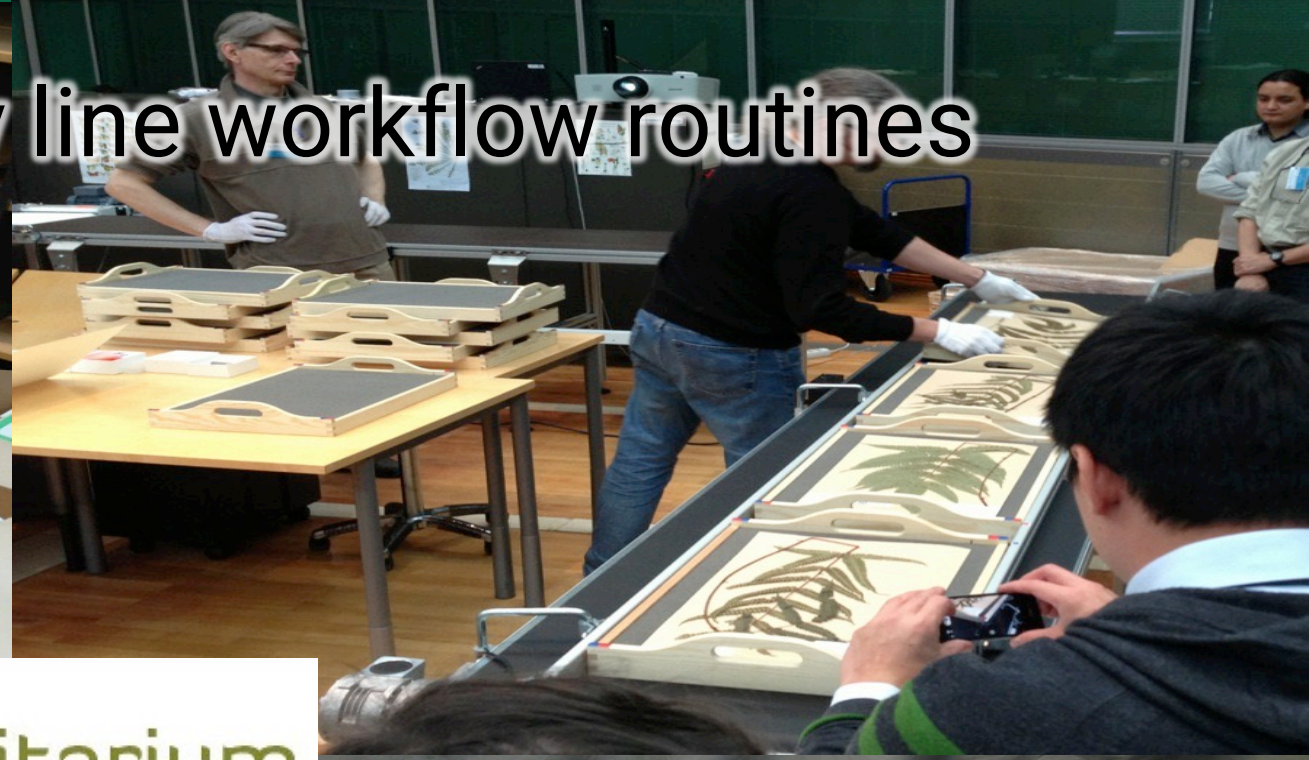
https://wiki.uio.no/nhm/skf/best-practices/index.php/Samlingstall



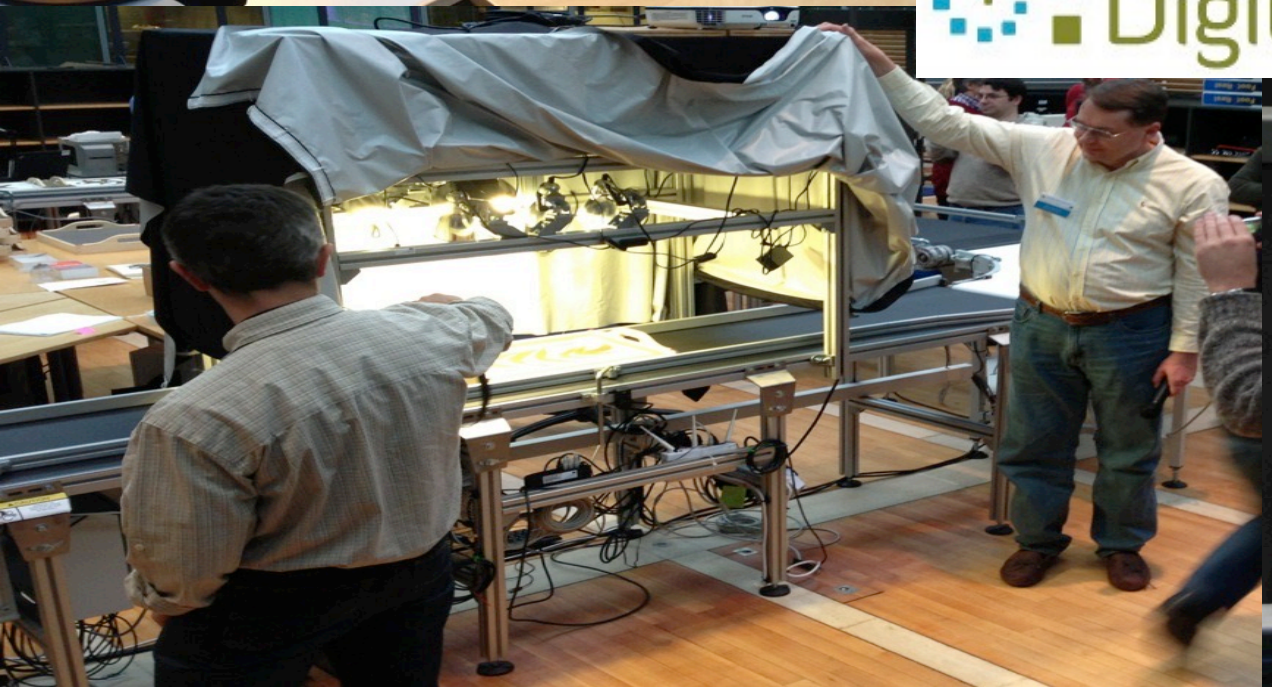
DIGITISATION OF MUSEUM COLLECTIONS AND HERBARIA


Catalog number: O-L-000014
human-friendly
Identifier: urn:uuid:41d9cbb4-4590-4265-8079-ca44d46d27c3
machine-friendly
GBIF.no
Transcription of labels UiO Lichen herbarium (2012)

“Industrial” assembly line workflow routines



Digitarium



http://gbif.no/dugnad/

HERBARIUM O

Herb. Univers. Osloensis

Heterodermia diademata (Tayl.) Awas.

KENYA
Rift Valley Prov. Uasin Gishu Distr.
5 km NW of Timboroa summit, on
moist brush by lake, and fence posts.
0-04 N 35-32 E 2650 m
2/1973 2K19/138 Hildur Krog

TLC (..../..):
zeorin
atranorin
Hildur Krog 19

O-L-185347 Sign up or Sign in! (4 transcriptions)

Scientific name:	ɜrmia diademata (Tayl.) Awas.
Country:	Kenya
Province:	RiftValey Prov.
District:	Uasin Gishu Distr.
Locality:	5 km NW of Timbora summit, on moist brush by lake, and fence posts.
Habitat:	moist brush by lake
Position:	0-04n 35-32e
	Show map...
Elevation:	2650 m
Collector:	Hildur Krog
Collector no:	2K19/138
Date:	Day 2 1973

Skip this record... Next record...

Show help...

GBIF

open data

WHY SHOULD STUDENTS LEARN OPEN SCIENCE?

WHY SHOULD MUSEUMS CARE ABOUT OPEN SCIENCE?

WHY TEACH STUDENTS OPEN SCIENCE?

- We are in the middle of an ongoing **paradigm shift** in scientific practice (and impact metrics).
- The open science wave is moving **fast**!
- Young scientists will need **different skills**, than was needed previously to succeed in academia.
- Researchers will need to develop **different approaches**, than they needed in the past – to remain relevant.
- Society is quickly gaining Big Data maturity and will **expect** new services from biodiversity information and research.

DATA CITATION AS A NEW CURRENCY OF SCIENCE

- **Peer-reviewed scholarly papers** in high impact journals maintain considerable weight for impact metrics.
- A **movement** is under way to **build similar status for** open data, open metadata, open material samples, and other **open scientific research products**…

DECLARATION ON RESEARCH ASSESSMENT

- DORA recognizes **the need to improve how** the outputs of scholarly **research are evaluated**.
- DORA's vision is to advance practical and robust approaches to **research assessment** globally and across all scholarly disciplines.
  - The Declaration on Research Assessment (DORA) was developed in San Francisco in 2012.
  - It has become a worldwide initiative covering all scholarly disciplines and key stakeholders.
  - Covering funders, publishers, professional societies, institutions, and researchers.
  - To date (2021-03-14), 17 085 individuals & 2 169 organizations in 145 countries have signed DORA.
  - Including 12 universities and research institutions from Pakistan

new possibilities for novel curiosity-driven research

Young Researcher

Global Biodiversity Information Facility

GBIF

CAREER OPPORTUNITIES

- Skills for open research and open data are in increasing demand!
- Bring benefits for your career as a (young) researcher.
- Add research datasets to your CV and online researcher profiles.

OPPORTUNITIES

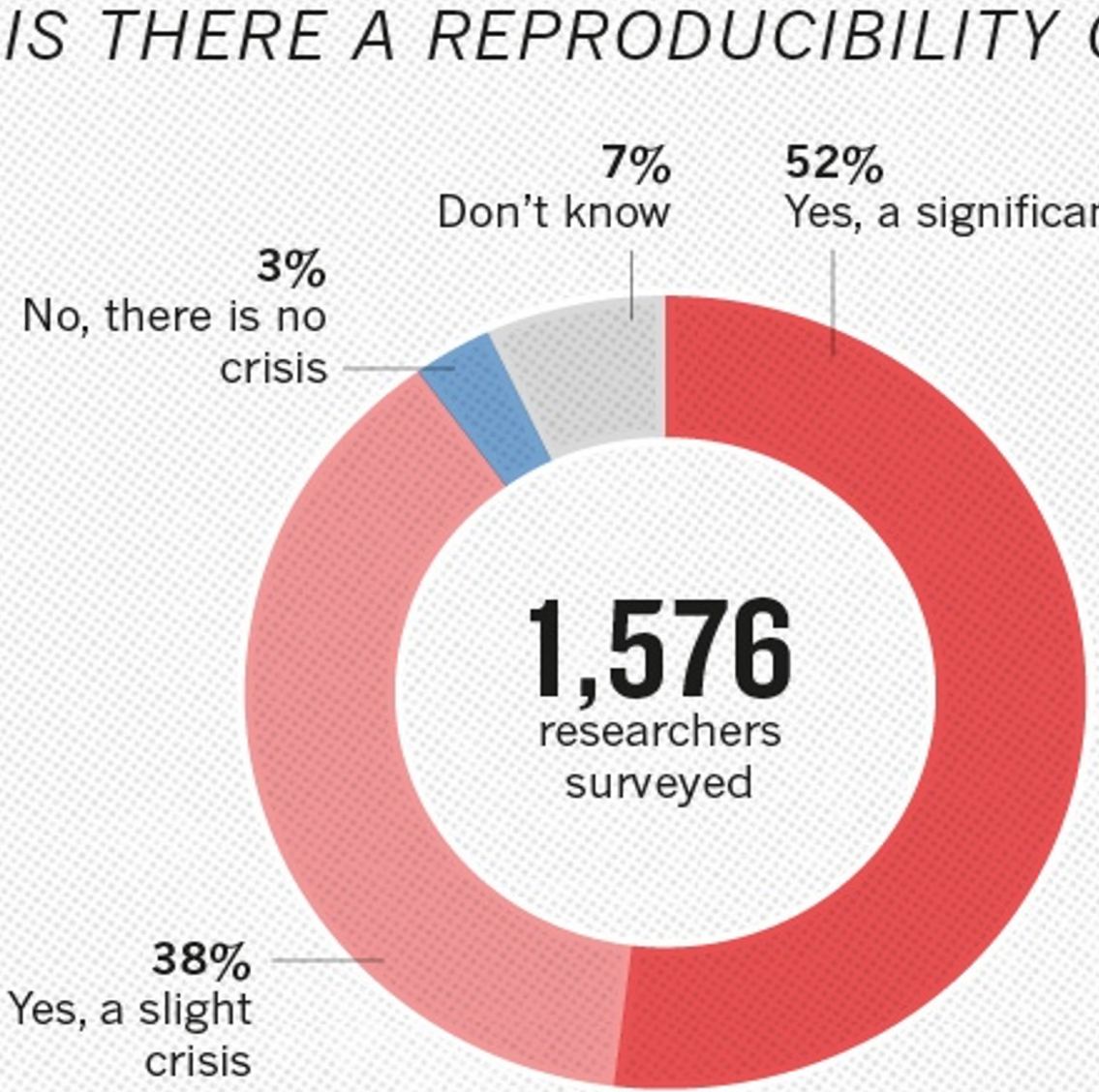
- Enables new research methodologies that were not possible before.
- Scientific citation metrics for collections and specimen data.
- Funding opportunities.


the dark side

REPRODUCIBILITY CRISIS

"***Scientific irreproducibility*** — *the inability to repeat others' experiments and reach the same conclusion*" (Nature 2016)

Baker (2016) 1,500 scientists lift the lid on reproducibility. Nature. doi:10.1038/533452a



Scientific irreproducibility *is a growing concern.*

Open Science solution*:* researchers to share their **methods**, **data**, computer **code** and results in central data repositories.

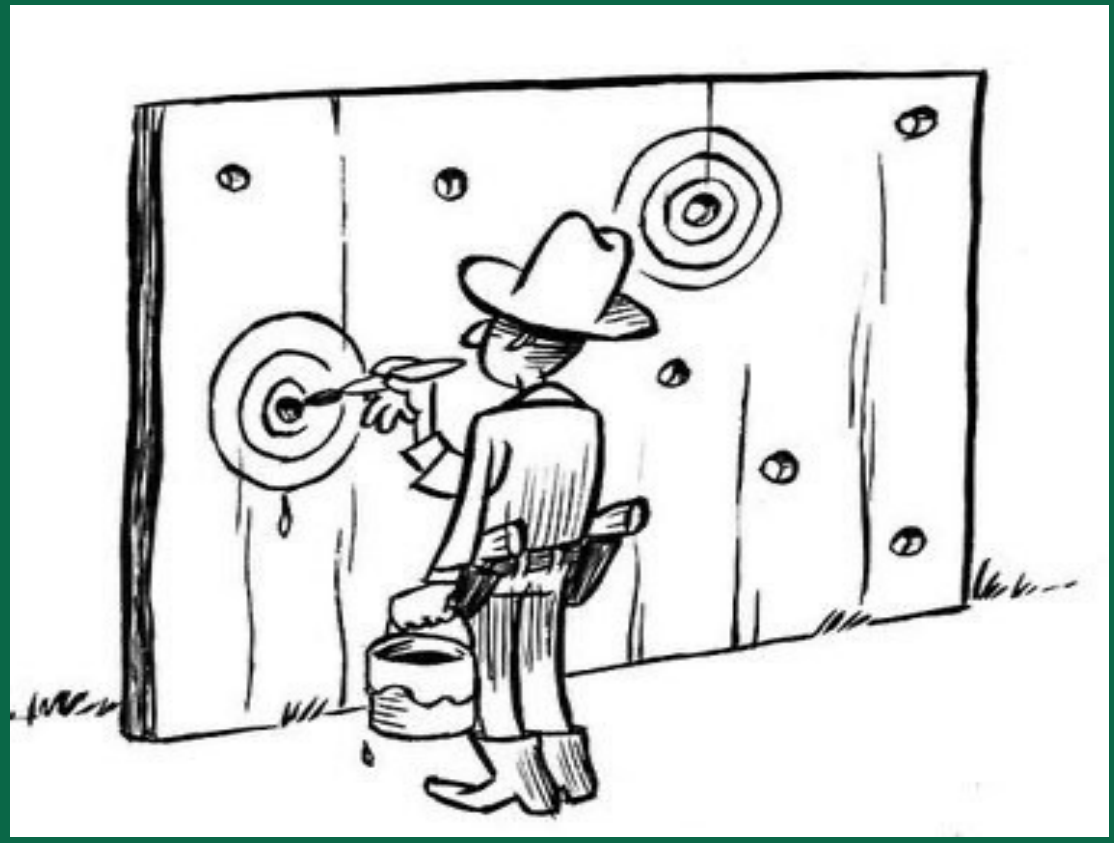
*For physical real-world samples we <u>also</u> need herbarium **specimen** and bio-repositories (museums).*

Baker (2016) 1,500 scientists lift the lid on reproducibility. Nature. doi:10.1038/533452a

GBIF

WILL ANYBODY TRUST CLOSED SCIENCE AGAIN?

- Studies (1,2) indicates that **p-hacking** is a significant problem – sometimes even without the scientist even being aware of doing so.
- **Pre-registered (open) data** provides a good insurance against suspicion of both data dredging (and plain data falsification).
- "p-hacking" = occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant (*data fishing*, …)

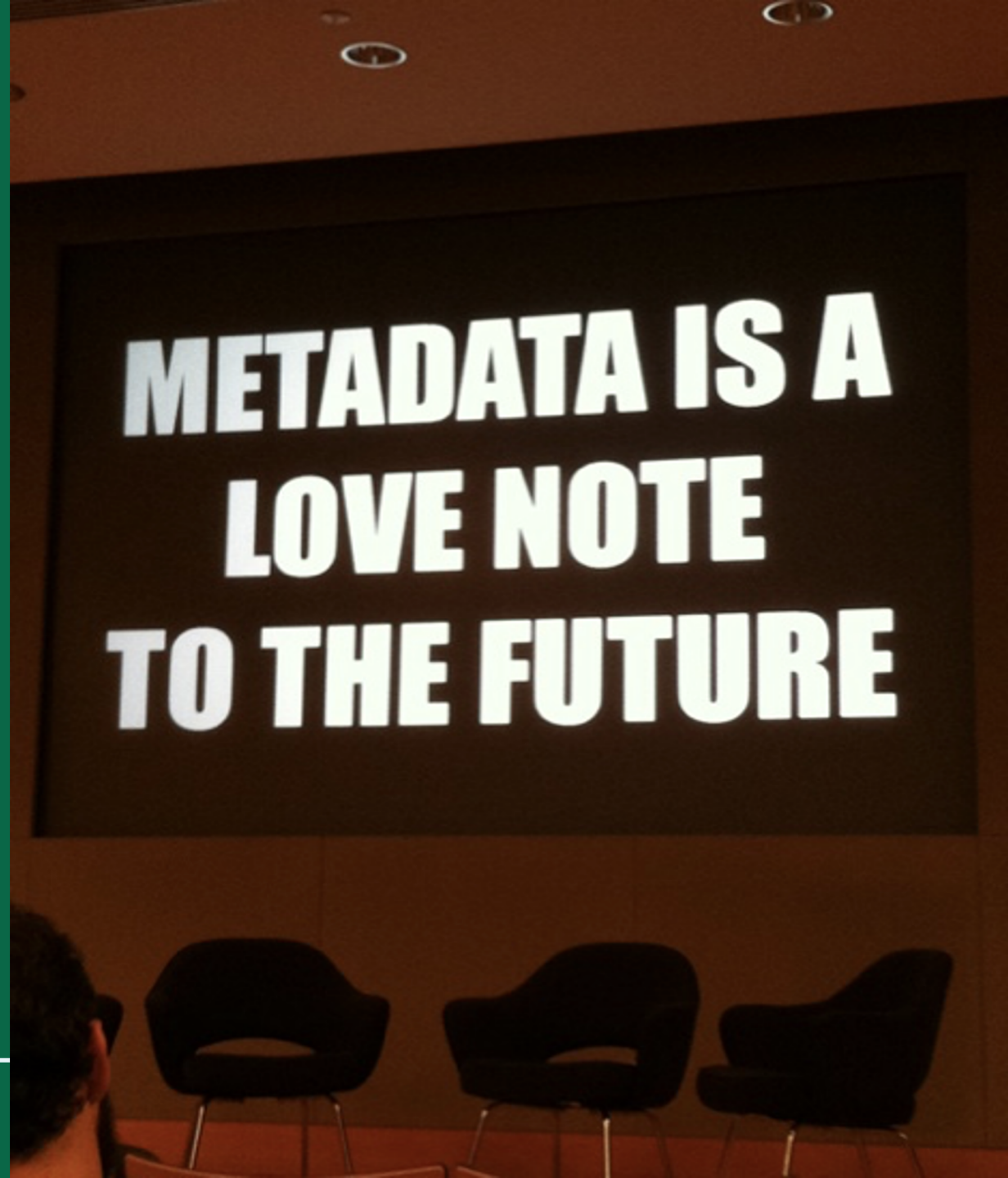


(1) Head *et al*. (2015) The Extent and Consequences of P-Hacking in Science . PLoS Biol. doi:10.1371/journal.pbio.1002106
(2) Ioannidis (2005). "Why Most Published Research Findings Are False". PLoS Medicine. doi:10.1371/journal.pmed.0020124.

reuse of research data

WHAT IS METADATA?

- Metadata, literally means ***“data about data”***.
  - Identify & discover its existence
  - Learn how to access or acquire data
  - Understand its fitness-for-use



DATA MANAGEMENT PLAN (DMP)

- A formal document that describes how research data will be handled during a research project, and after the project is completed.
- Plan data management before the project begins.
- Including costs of data management & archiving.
- Reduce the loss of data.
- DMP saves time in the long run
- Promotes data fitness for reuse.
- Reduce duplication of existing scientific studies.

Illustration CC BY Jørgen Stamp

https://innsida.ntnu.no/wiki/-/wiki/English/Data+management+plan

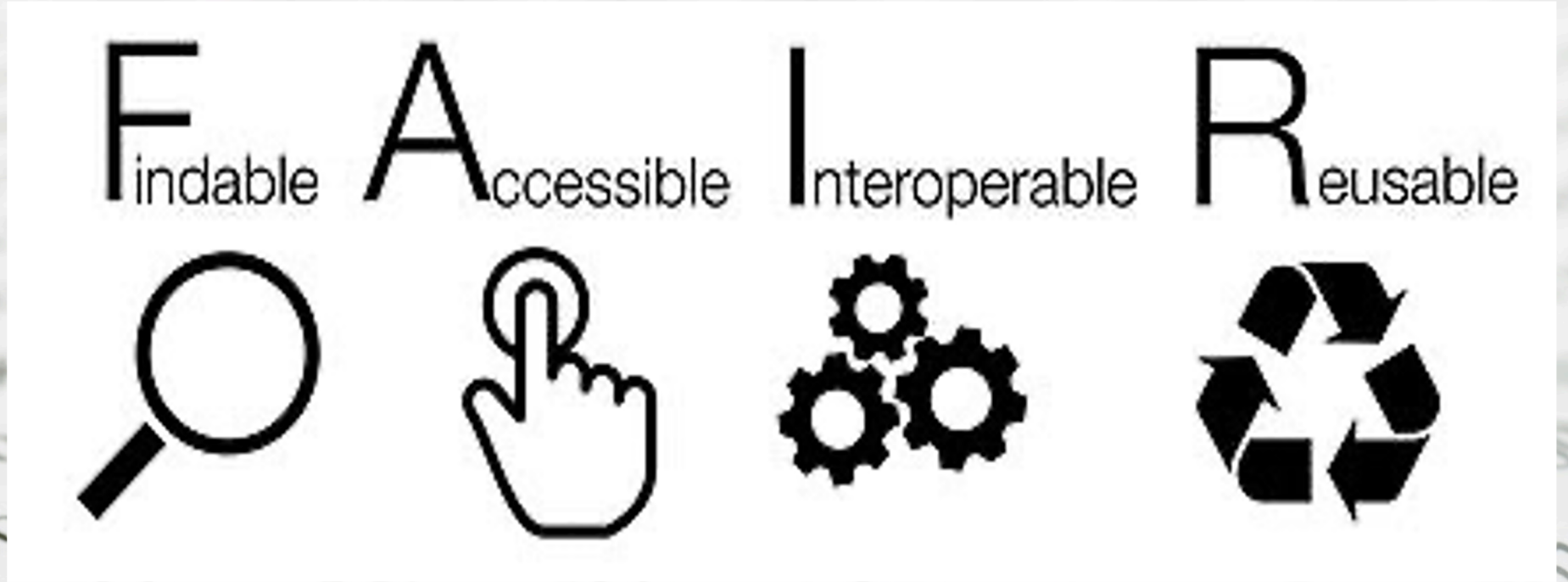
GBIF

WHAT IS A DATA PAPER?

- A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal.
- It takes effort to prepare, curate and describe data.
  - Data papers provide recognition for this effort by means of a scholarly article.
  - Recognition for improving the fitness for reuse for your own datasets.
  - Recognition for making legacy research data available for reuse.

https://www.gbif.org/data-papers

GBIF



*… researchers need to do more than simply post their data on the web for it to be **re-usable**.*

FAIR data is about ***machine-readable*** data


GBIF

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

Observations

Digitized specimens

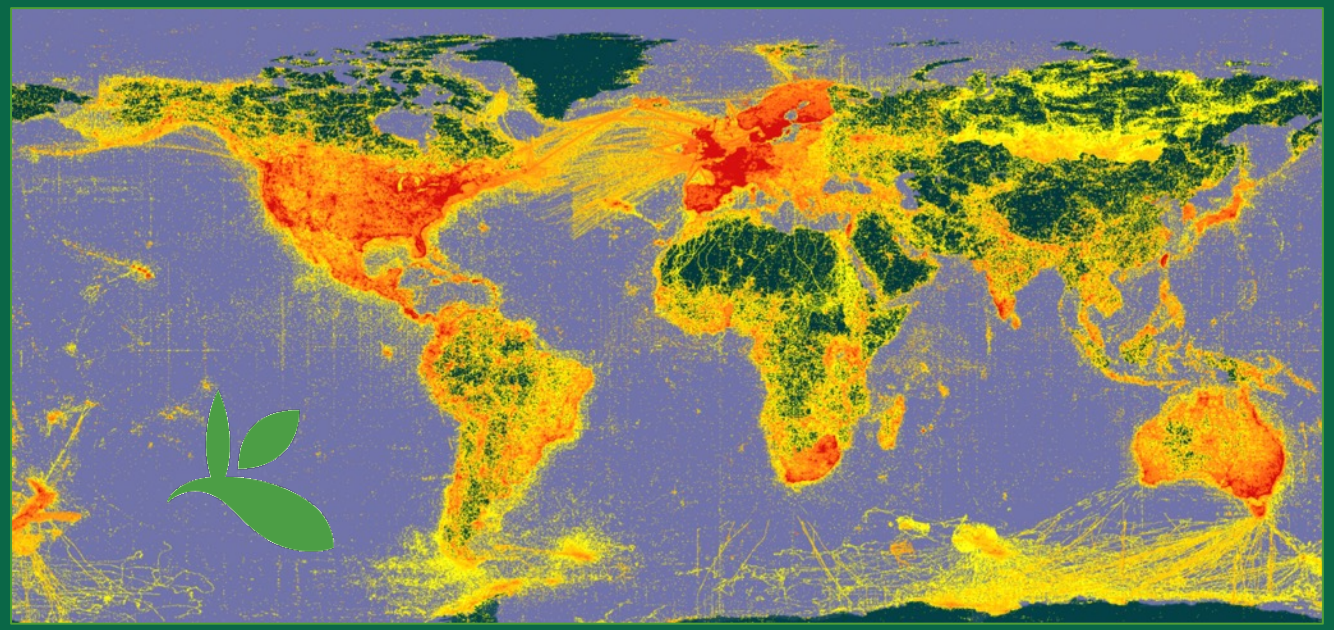
Literature

Data publishing and indexing

Common standards (DwC)

Remote-sensing

Environmental DNA



Data discovery and use

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

GBIF

BY THE NUMBERS | 18 MAY 2021

GBIF

Species occurrence records

1 698 262 502

Datasets

59 461

Country Participants **61**

Organizational Participants **40**

Publishers

1 677

Peer-review papers using data

5 762

Average records downloaded per month (2020)

23.6 billion

BY THE NUMBERS | 15 MAY 2021 - PAKISTAN

Species occurrence records *(published from)*

78 976

Datasets *(published from)*

10

Publishers (from Pakistan)

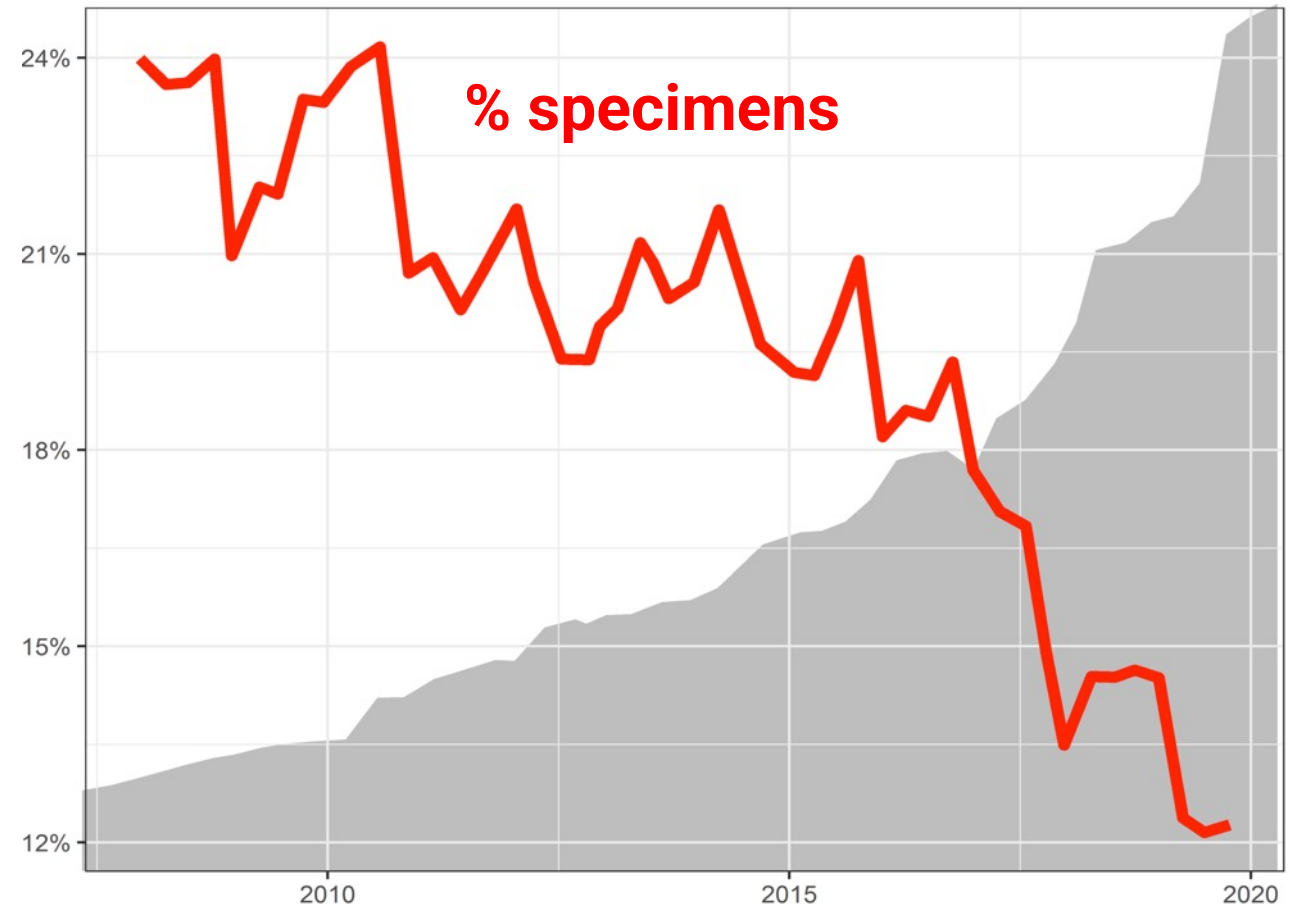
9

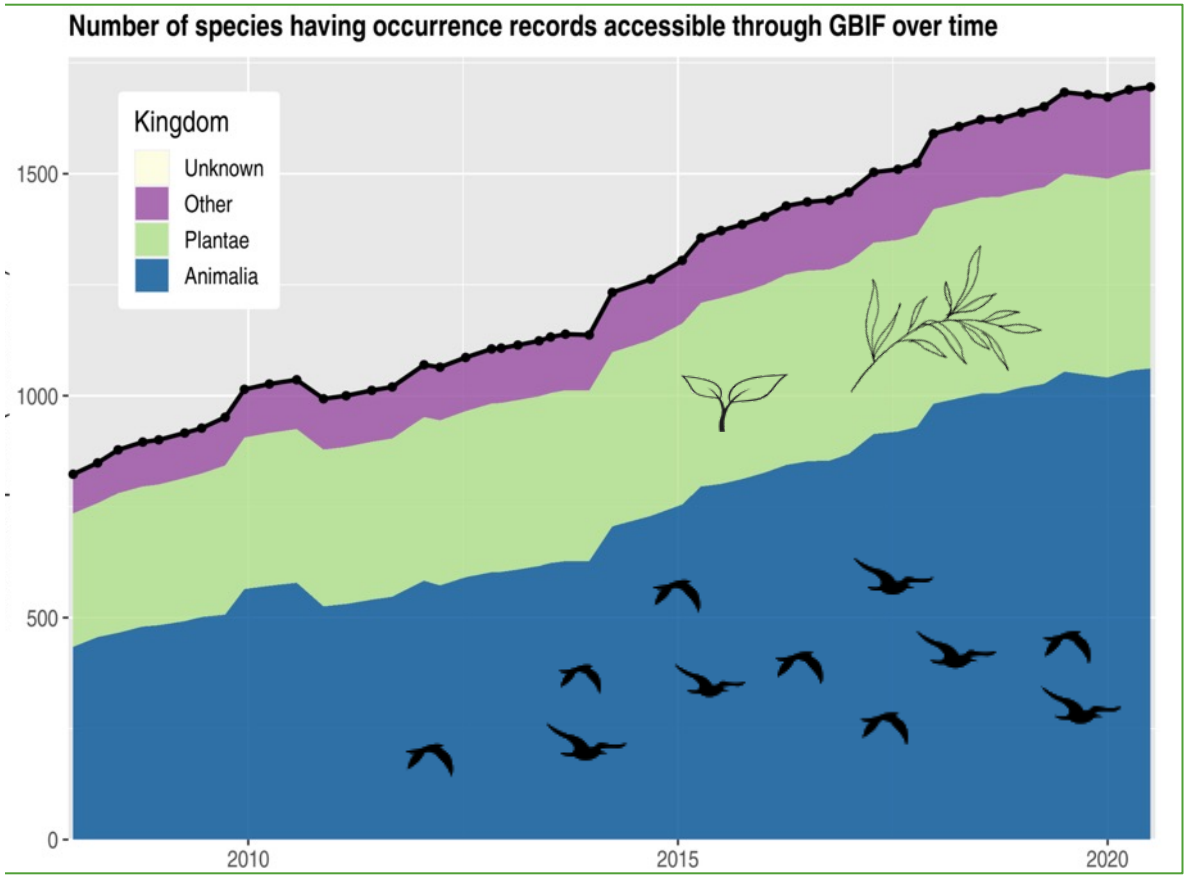
Peer-review papers using data (co-author from Pakistan

18

https://www.gbif.org/country/PK/summary

DATA TRENDS ON GBIF.org





https://www.gbif.org/analytics/global

DATA RICHNESS LEVELS SUPPORTED BY GBIF

Dataset description,
taxonomic/geographic/temporal scope

Dataset metadata

List of taxa
regional or thematic (e.g. invasive, medicinal)

Species checklists

Species occurrences
dates, coordinates, basis of record

Occurrence-only data

Species occurrences and sampling events
dates, coordinates, sampling effort / protocol, abundance

Sampling-event data

https://www.gbif.org/dataset-classes

GBIF

SPECIES OCCURRENCE RECORDS WITH MULTIMEDIA EVIDENCE

Status 15th May 2021

80 million records with taxonomically identified images

- 200 million specimens
- 1 411 million human observations
- 34 million material samples

724 177 audio files

2 845 video files

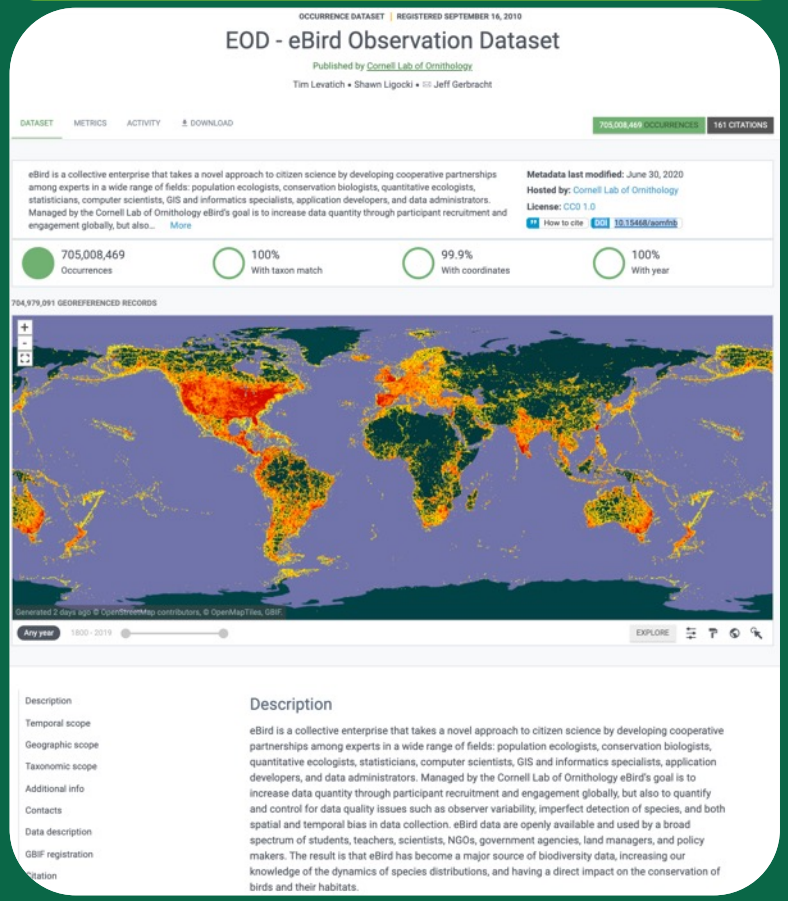
https://www.gbif.org/occurrence/gallery

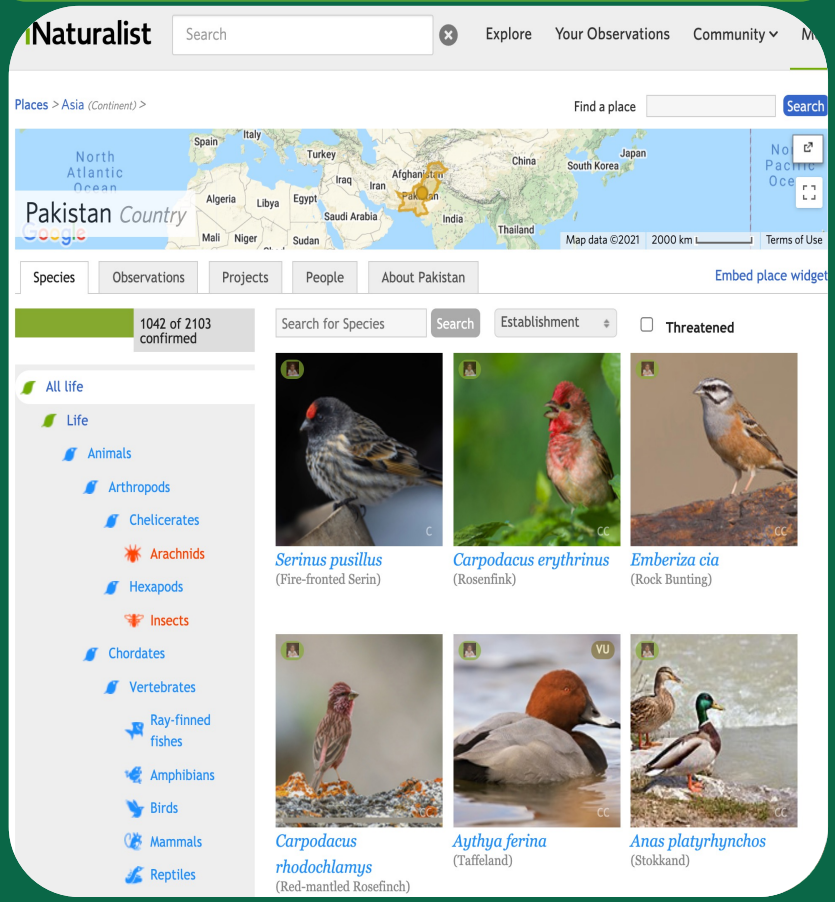
SOURCES OF DATA IN GBIF: CITIZEN SCIENCE OBSERVATIONS

eBird

iNaturalist

ARTSDATABANKEN





GBIF

SOURCES OF DATA IN GBIF: DIGITIZED SPECIMENS FROM MUSEUM COLLECTIONS

OCCURRENCE DATASET | REGISTERED AUGUST 7, 2018

Quaid-i-Azam University Herbarium

Published by Herbarium of Pakistan (ISL) - QAU

DATASET METRICS ACTIVITY DOWNLOAD 36,175 OCCURRENCES 44 CITATIONS

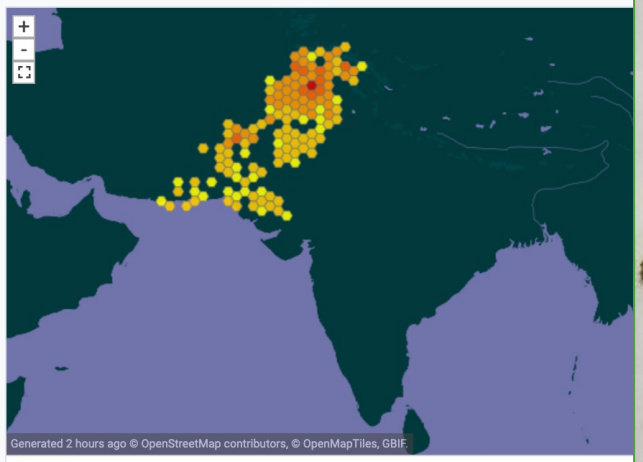
ISL was established in 1974. It is now the largest herbarium in Pakistan having over 180,000 specimens, collected from throughout Pakistan. The oldest specimens were collected over 50 years ago. The herbarium supports and promotes research on a wide range of environmental and biodiversity issues. Its holdings include many important endangered, rare and vulnerable plant species. Currently, the herbarium is being used by projects in taxonomy, molecular systematics, revisions, palynology, diversity... More

Project ID: BIFA3_47
Publication date: November 30, 2020
Metadata last modified: November 30, 2020
Hosted by: iDigBio
License: CC0 1.0

36,175 Occurrences

100% With taxon match

24,284 GEOREFERENCED RECORDS



Any year 1862 - 2018

3 OCCURRENCES WITH IMAGES

OCCURRENCE | 10 JULY 2009

Coenochilus hervillardi Estienne, 2011

Collected in Pakistan

Animalia › Arthropoda › Insecta › Coleoptera › Cetoniidae › *Coenochilus*

DETAILS

Species: Coenochilus hervillardi Estienne, 2011
Location: Pakistan
Basis of record: Preserved specimen
Specimen type: Holotype

Dataset: The Coleoptera collection (EC) of the Muséum national d'Histoire naturelle (MNHN...
Publisher: MNHN - Museum national d'Histoire naturelle
Issues: Occurrence status inferred from individual count

Coordinates missing

This record is published without coordinates, but it includes a textual description of its location.

Location: Pakistan
Locality: Balochistan, forêt du parc national de Ziarat

Creator: Antoine Mantilleri
Record license: http://creativeco...ses/by-nc-nd/4.0/
Identifier: http://mediaphoto...1Qysod78PVnpkJyWH

GBIF

SOURCES OF DATA IN GBIF: TAXONOMIC LITERATURE, OLD AND NEW

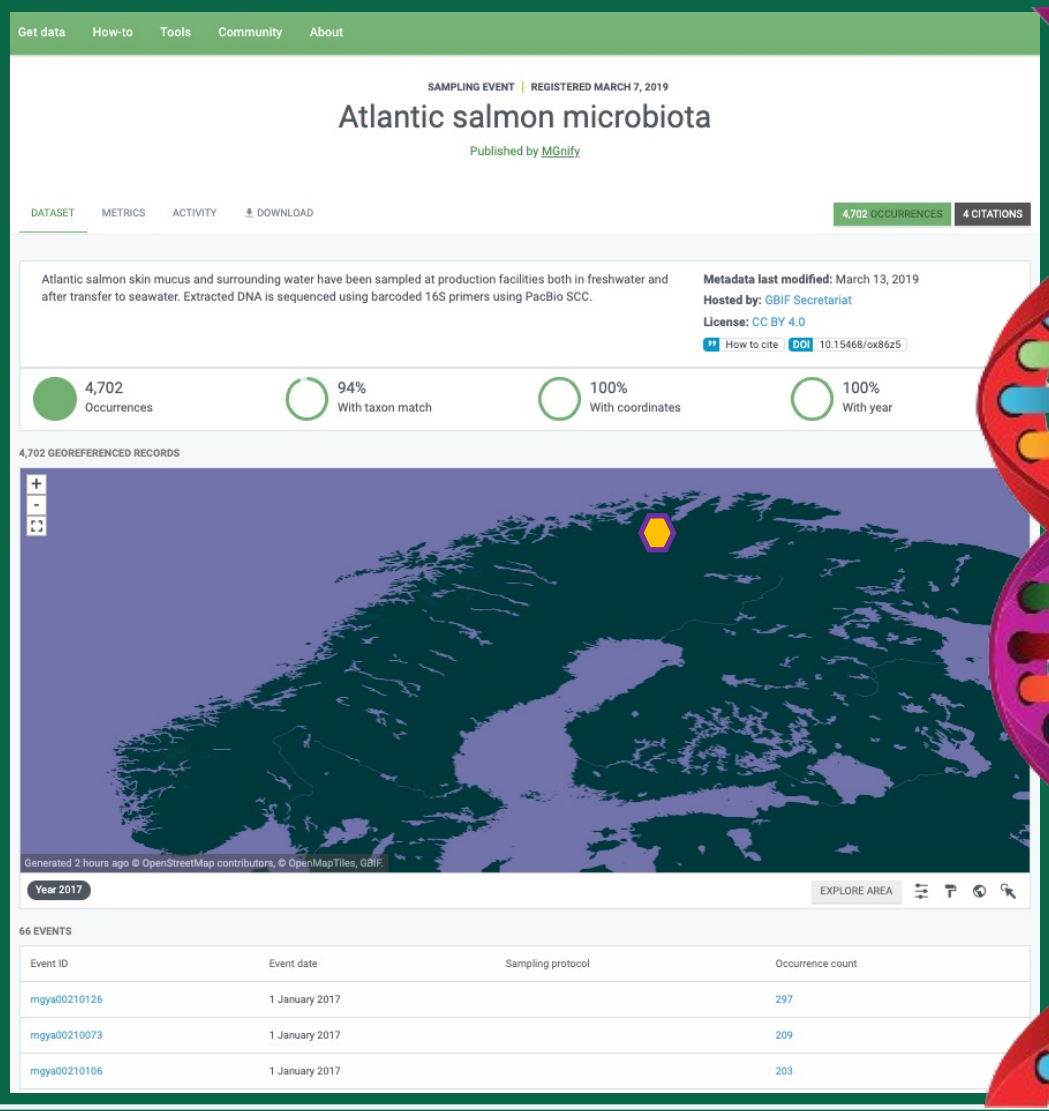
PLAZI
TAKING CARE OF FREEDOM
Data liberation

BHL
Biodiversity Heritage Library

ACS Publications
MOST TRUSTED. MOST CITED. MOST READ.
BioMed Central
The Open Access Publisher
ELSEVIER
NATIONAL PRESS FOUNDATION
npg
nature publishing group
PLOS
aap PSP
stm
Taylor & Francis Group
an informa business
WILEY
Wolters Kluwer

GBIF

SOURCES OF DATA IN GBIF: DNA SEQUENCE-DERIVED OCCURRENCE DATA



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

GLOBAL BIODIVERSITY VS. DIGITALLY AVAILABLE DATA

1200 mill. animals | 300 m plants | 20 m fungi | 16 m bacteria | 0,04 m virus

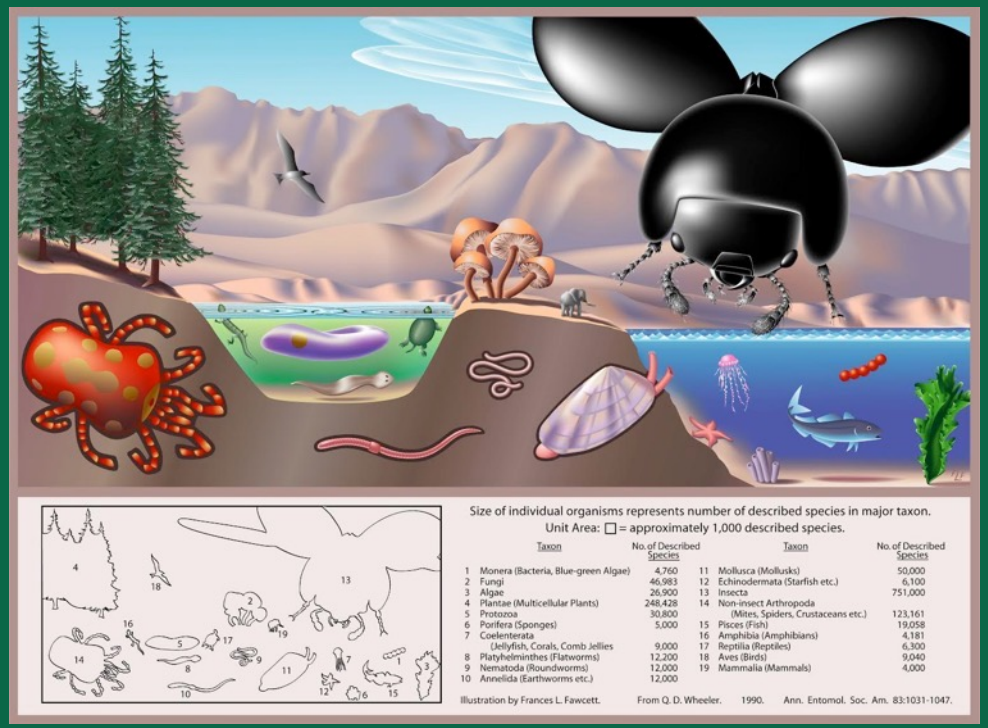
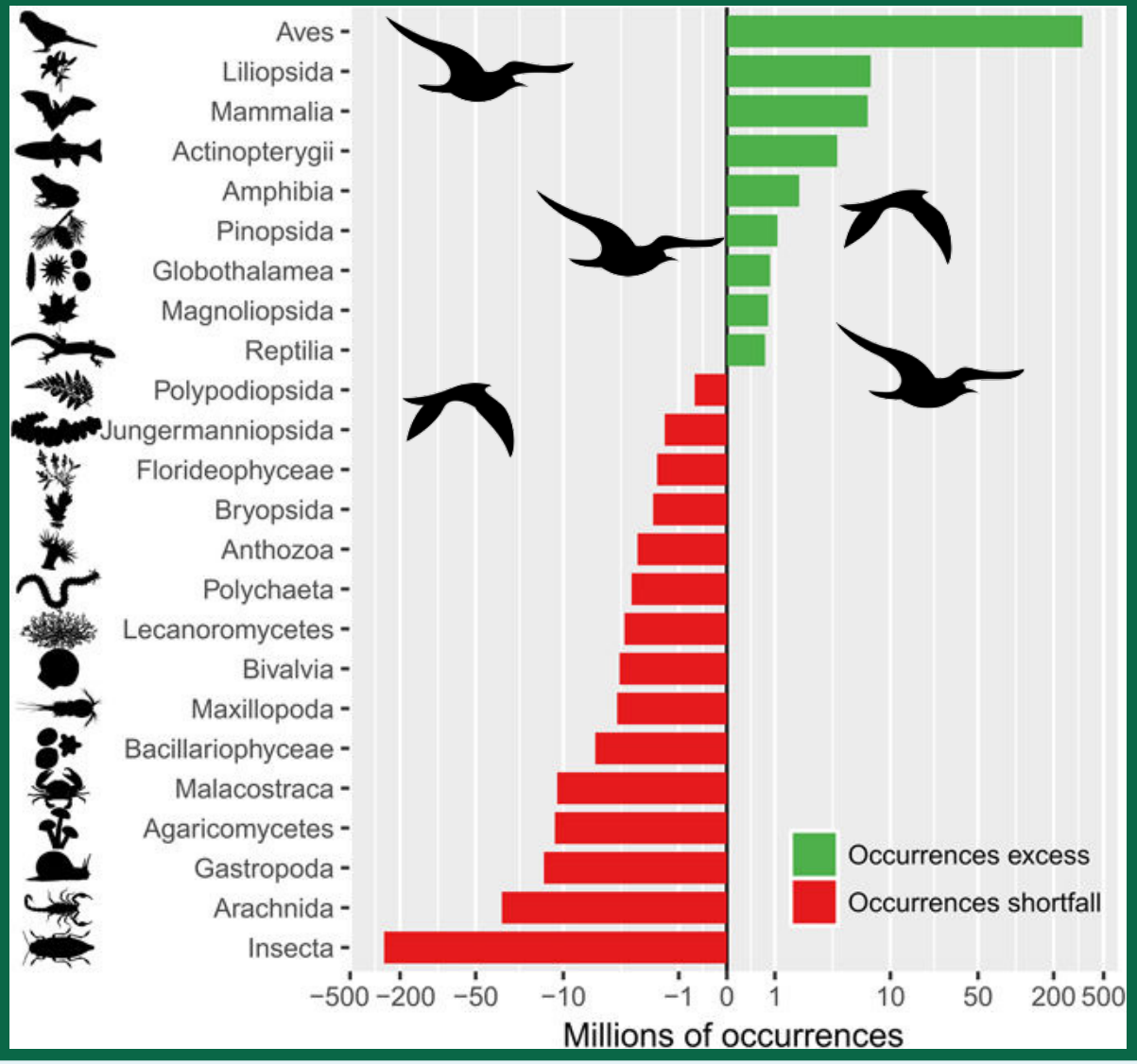


Image: FL Fawcett in Wheller Ann. Entomol. Soc. Am. 1990



Troudet et al. Nature Scientific Reports 2017

Taxonomic bias towards birds and against insects and small organisms

GBIF


GBIF
data publishing

PUBLISH YOUR DATASETS WITH GBIF

- *Step 1*: digitize collections & herbaria
- *Step 2*: register for endorsement in GBIF
- *Step 3*: convert to Darwin Core format
- *Step 4*: publish from national GBIF node

- *Alternative*: publish from regional GBIF cloud data repository - cloud.gbif.org/eca
- *Alternative*: Many citizen science data platforms publish data in GBIF

https://www.gbif.org/publishing-data/

GBIF

GBIF

data use

GBIF: MULTIPLE-PURPOSE DATA PUBLISHING SERVICES

Bio-Collections
& ecology datasets

GBIF

Darwin Core

Environment directive reporting

Invasive Species

IUCN

IUCN RED LIST

Research data portals

elixir

BOLD SYSTEMS

GBIF portal

Convention on Biological Diversity

UNEP

FAO

ipbes
Science and Policy for People and Nature

GBIF

POLICY LINKS

Convention on Biological Diversity

SUSTAINABLE DEVELOPMENT GOALS

COUNTRY PROFILES

Pakistan

Pakistan

Convention

Party since: 1994-10-24

By: Ratification

Cartagena Protocol

Party since: 2009-05-31

By: Ratification

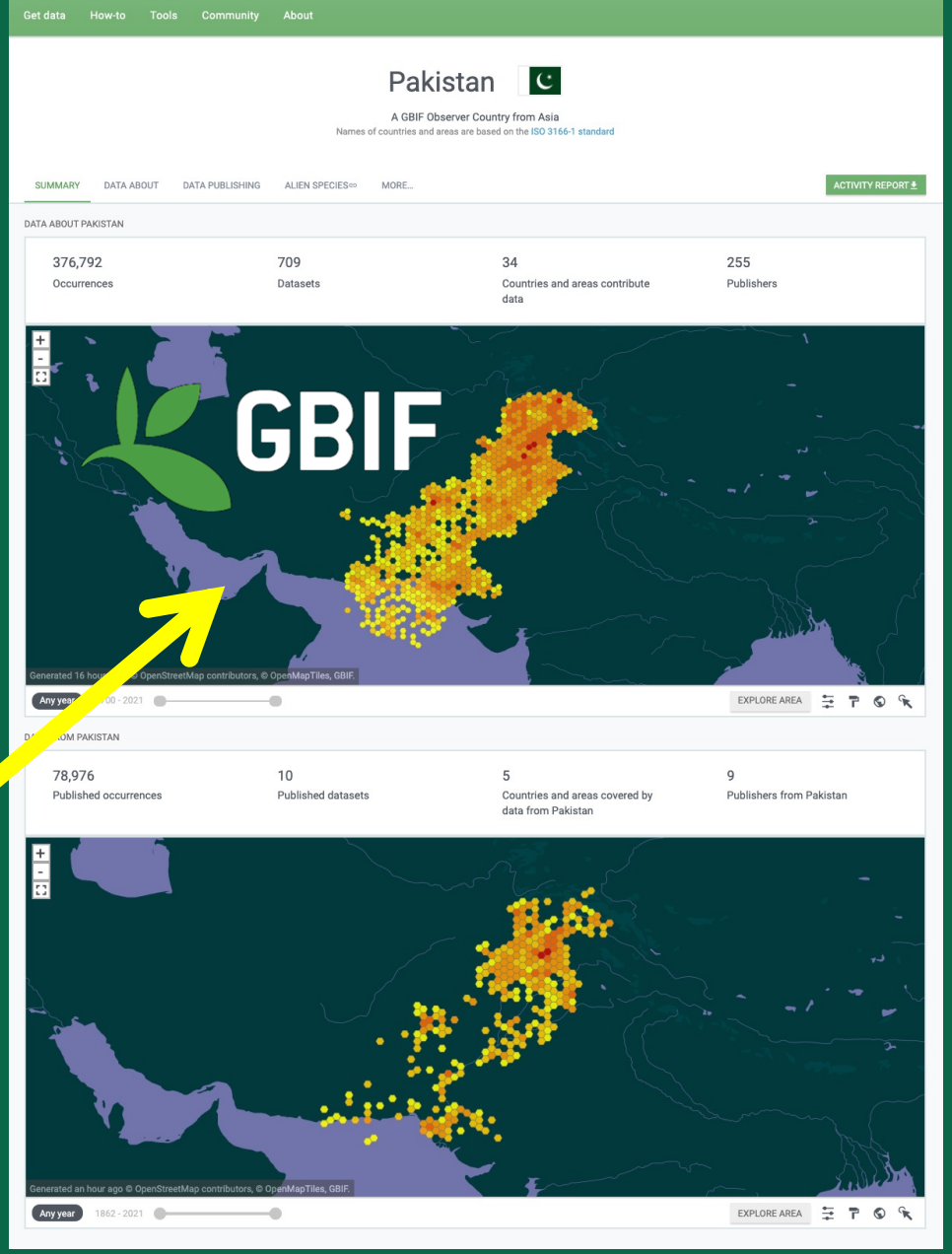
Nagoya Protocol on Access and Benefit-sharing

Party since: 2016-02-21

By: Accession

Further Information

- InforMEA Profile
- UNFCCC Profile
- World Heritage Sites
- GBIF Data
- WTO Statistics



GBIF

POLICY LINKS: AICHI TARGETS

- Trend in invasive alien species introductions (through Global Register of Introduced and Invasive Species)

- Species Protection Index
- Protected Area Representativeness Index

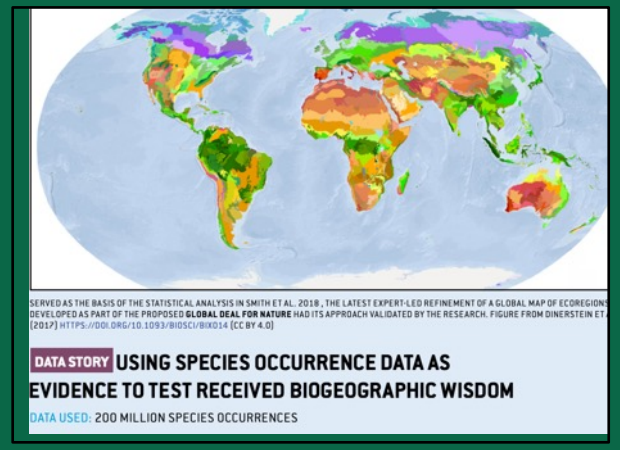
- Comprehensiveness of conservation of socioeconomically/culturally valuable species
- Agrobiodiversity Index
- Crop Wild Relative Index

- Growth in species occurrence records accessible through GBIF
- Species Status Information Index

https://www.cbd.int/cooperation/csp/gbif.shtml | https://www.cbd.int/csp/survey/GBIF.pdf

GBIF

A DATA RESOURCE TO SUPPORT RESEARCH AND SUSTAINABLE DEVELOPMENT

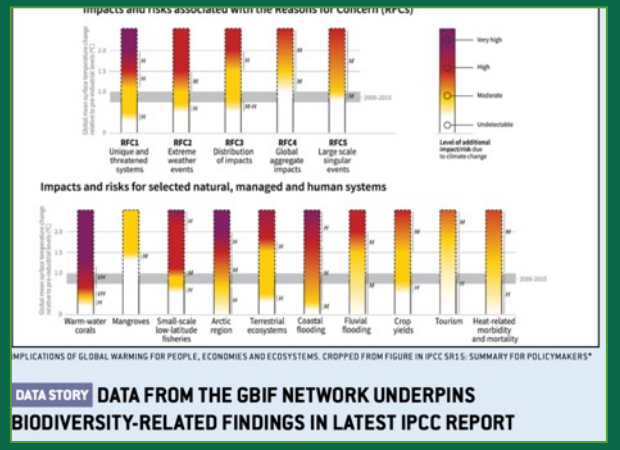


Conservation

- Protected areas
- Threatened species
- Invasive species risk

Food Security

- Crop wild relatives
- *In situ*, *ex situ* conservation of genetic diversity
- Fisheries planning



Climate change

- Modelling impacts on species ranges
- Adaptation strategies
- Mitigation benefits, risks

Human health

- Disease risk based on occurrence of vectors, hosts, reservoirs
- Medicinal plants
- Hazards e.g. snakebite

https://www.gbif.org/science-review

GBIF

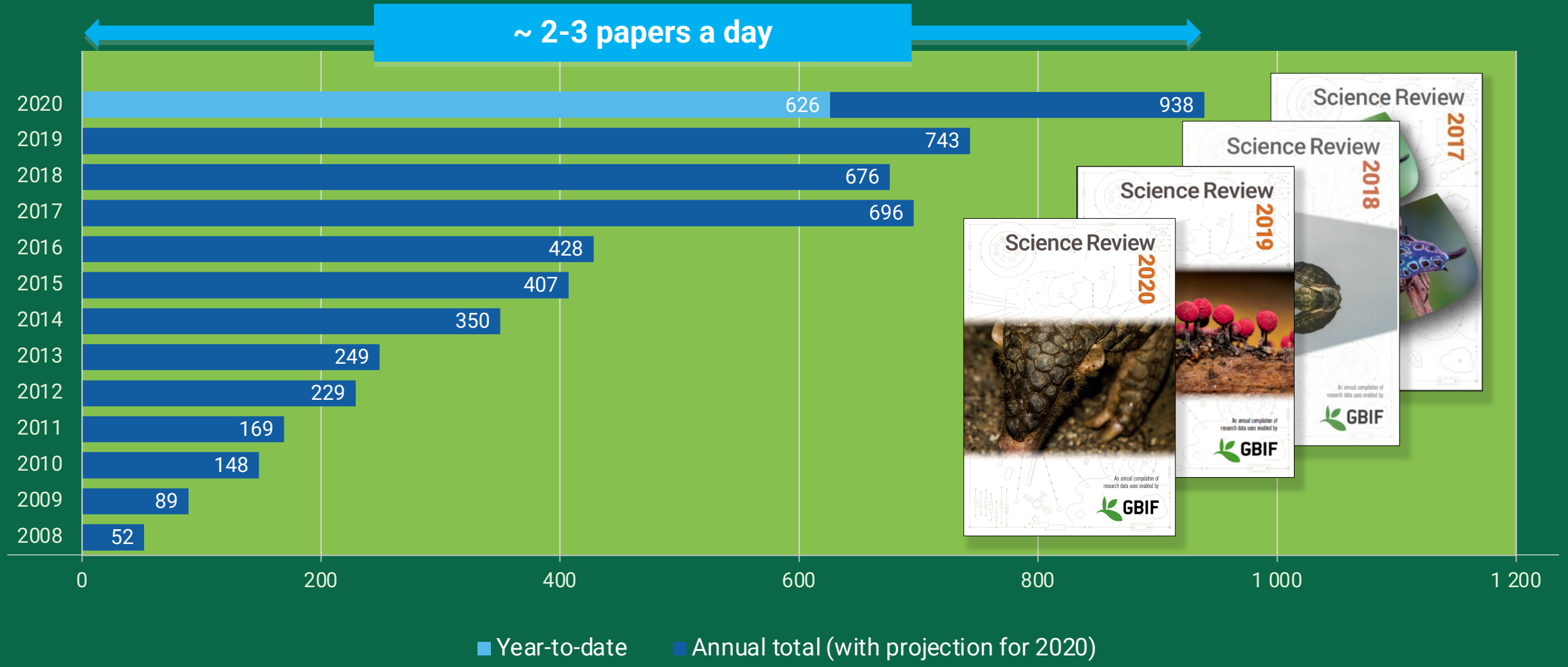
CREDIT FOR DATA REUSE

To incentivize the sharing of useful data, the scientific enterprise needs a well-defined ***system that links individuals with reuse of data sets they generate***

Pierce *et al*. Credit data generators for data reuse, *Nature* 6 June 2019

PEER-REVIEWED PUBLICATIONS USING GBIF-MEDIATED DATA

September 2020



https://www.gbif.org/resource/search?contentType=literature&literatureType=journal&relevance=GBIF_USED&peerReview=true

#CiteTheDOI

GBIF

HOW TO CITE DATA MEDIATED BY GBIF

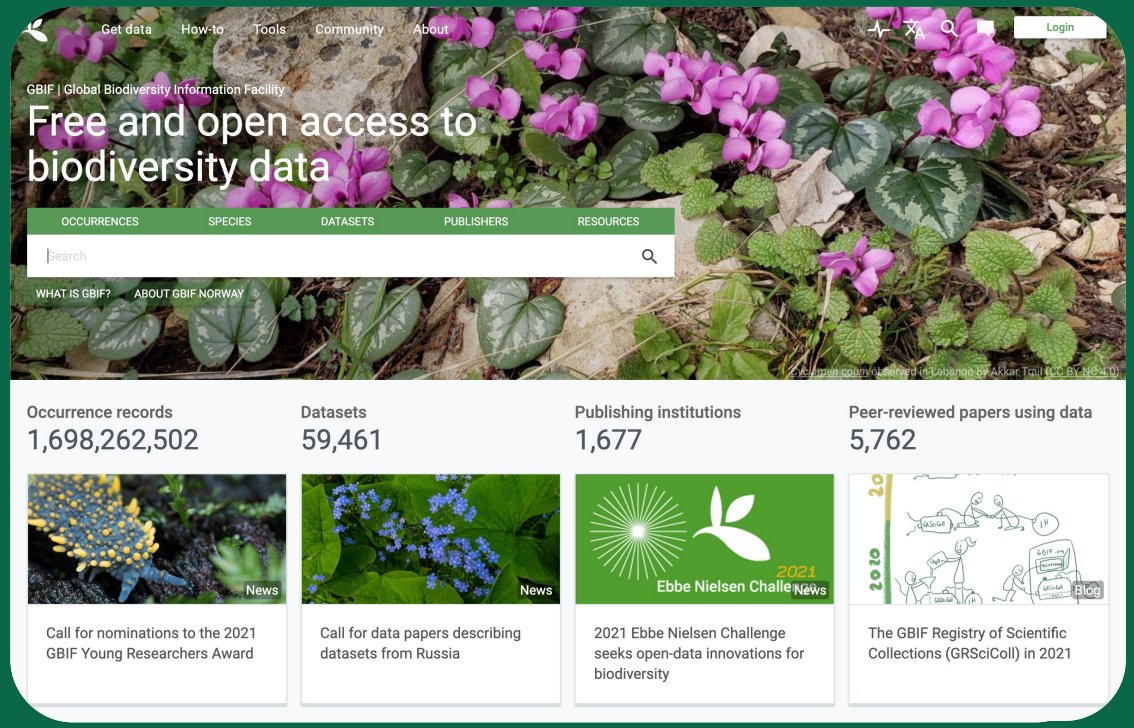
1. **Download data** from GBIF.org
2. and receive recommended citation with a **download DOI**
3. **Cite the DOI** in published research or other work

Example: GBIF.org (12 October 2020) GBIF Occurrence Download **https//doi.org/10.15468/dl.xxxxxx**

WHY CITE DATA?

- **Good academic practice** for transparent and reproducible research
- **Credit institutions** who shared data and supported your research
- Help data publishing institutions to **demonstrate value** of digitization and data publication through research
- Correct citation **encourages data sharing**
- Data accessed through GBIF is free for all – but not free of obligations: see the **user agreement**

THE RESEARCH DATA LIFECYCLE



GBIF.org

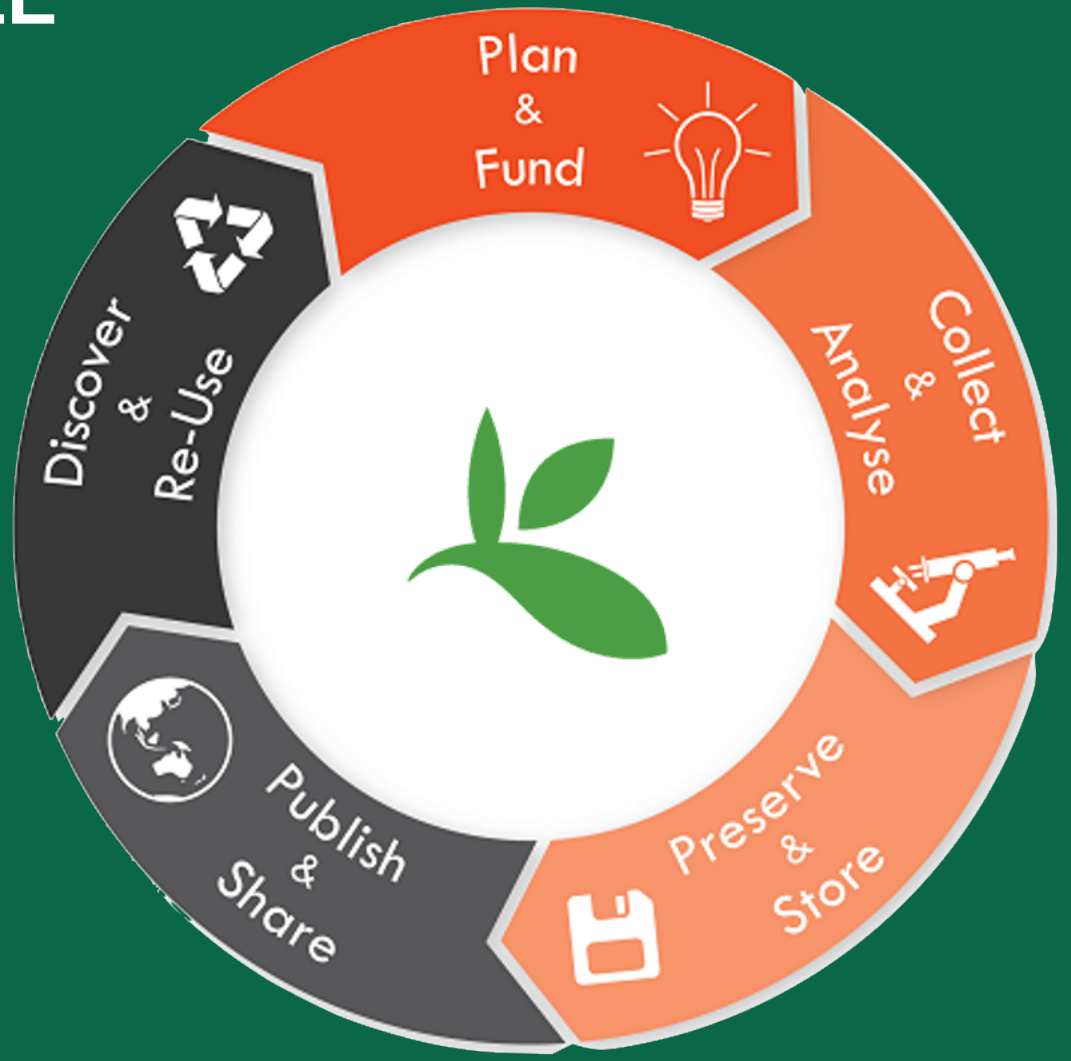


Illustration image credit: Sydney University

GBIF

THANK YOU

Dag Endresen | GBIF Norway

helpdesk@gbif.no