

Global Biodiversity Information Facility

Dag Endresen - GBIF Norway

University of Oslo, Natural History Museum

Illustration: GBIF data portal



GBIF

Global Biodiversity
Information Facility

Open Science course, Hjerkinn | 17th November 2023

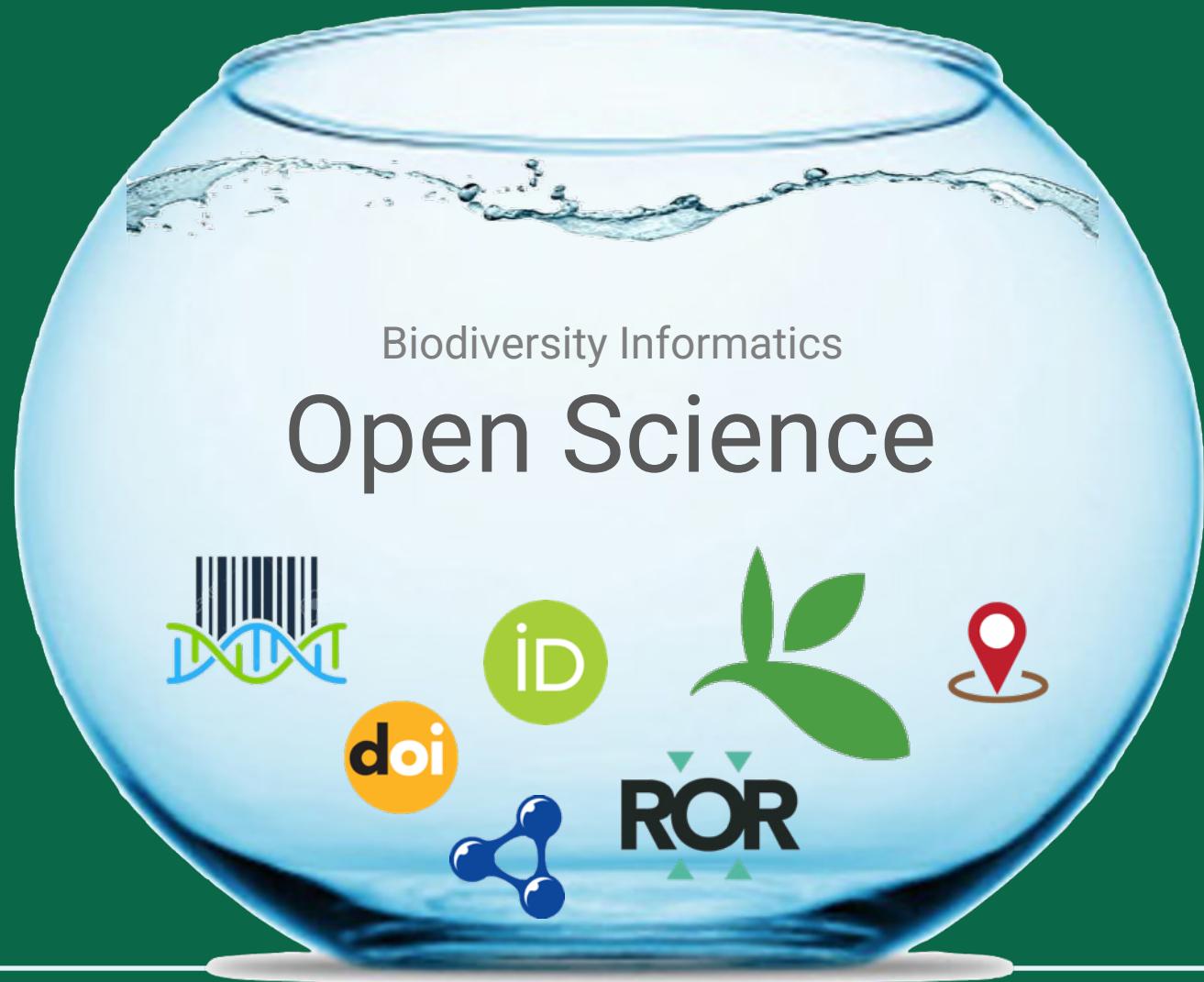
WHY OPEN RESEARCH DATA?

A person wearing a white lab coat and orange gloves is looking through a blue microscope. The background is green and features a grid of binary code (0s and 1s) in light blue.





Traditional
biodiversity science



new possibilities for novel curiosity-driven research





OPPORTUNITIES

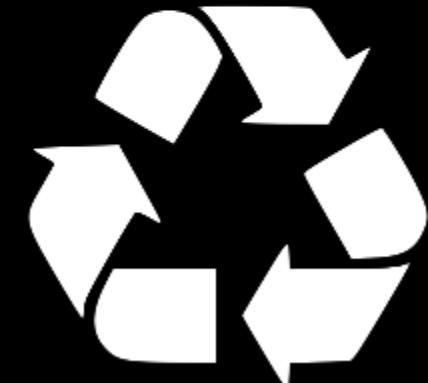
- Enables new research methodologies that were not possible before.
- Skills for open research and open data are in increasing demand!
- Funding opportunities.
- GBIF brings new benefits and opportunities for our museums.

WHY OPEN RESEARCH DATA?

- ❖ We are in the middle of an ongoing **paradigm shift** in scientific practice (*and impact metrics*).
- ❖ Marine science will also need to develop **different approaches**, than they needed in the past – *to remain relevant*.
- ❖ Society is gaining Big Data maturity and will **expect new services** from marine sciences.
- ❖ The open science wave is moving **fast!**



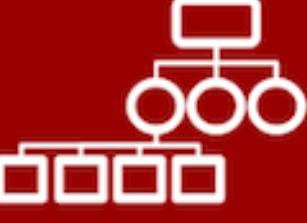
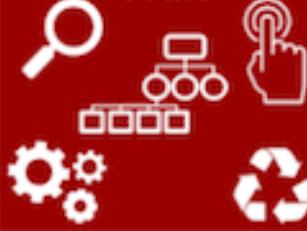
Findable Accessible Interoperable Reusable



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

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

FAIR DATA PRINCIPLES

Findable 	Persistent Identifiers (PIDs) 	Rich metadata 	Indexed data repositories 	PIDs in metadata 
Accessible 	Standard communications protocol 	Open, free protocol 	Authentication, where necessary 	Metadata is always available 
Interoperable 	Vocabularies 	Vocabularies are FAIR 	Linked metadata 	Persistent Identifiers (PIDs) 
Reusable 	Metadata have multiple attributes 	Usage license 	Provenance 	Community standards 

Be

F A I T R

Findable Accessible Interoperable Reusable

and

C A R E

Collective
Benefit

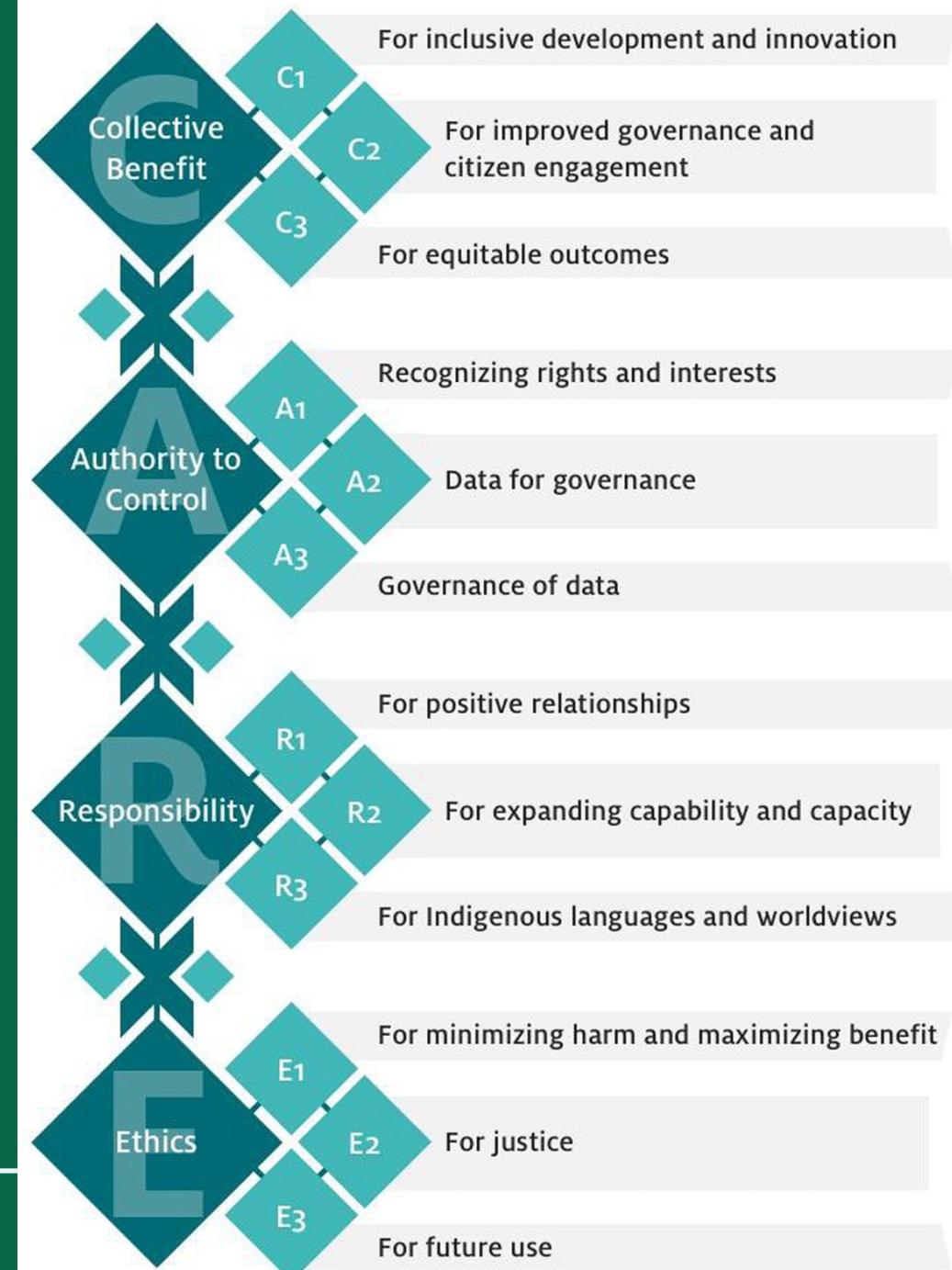
Authority
to Control

Responsibility

Ethics

CARE DATA PRINCIPLES

- CARE is an acronym which stands for:
 - **Collective Benefit**,
 - **Authority to Control**,
 - **Responsibility**,
 - **Ethics**
- CARE was created by the International Indigenous Data Sovereignty Interest Group, a group that is a part of the Research Data Alliance.
- Resolve **Indigenous Peoples' rights** to and interests in their data across the data lifecycle.
- Building on **FAIR** data principles while ensuring indigenous interests.



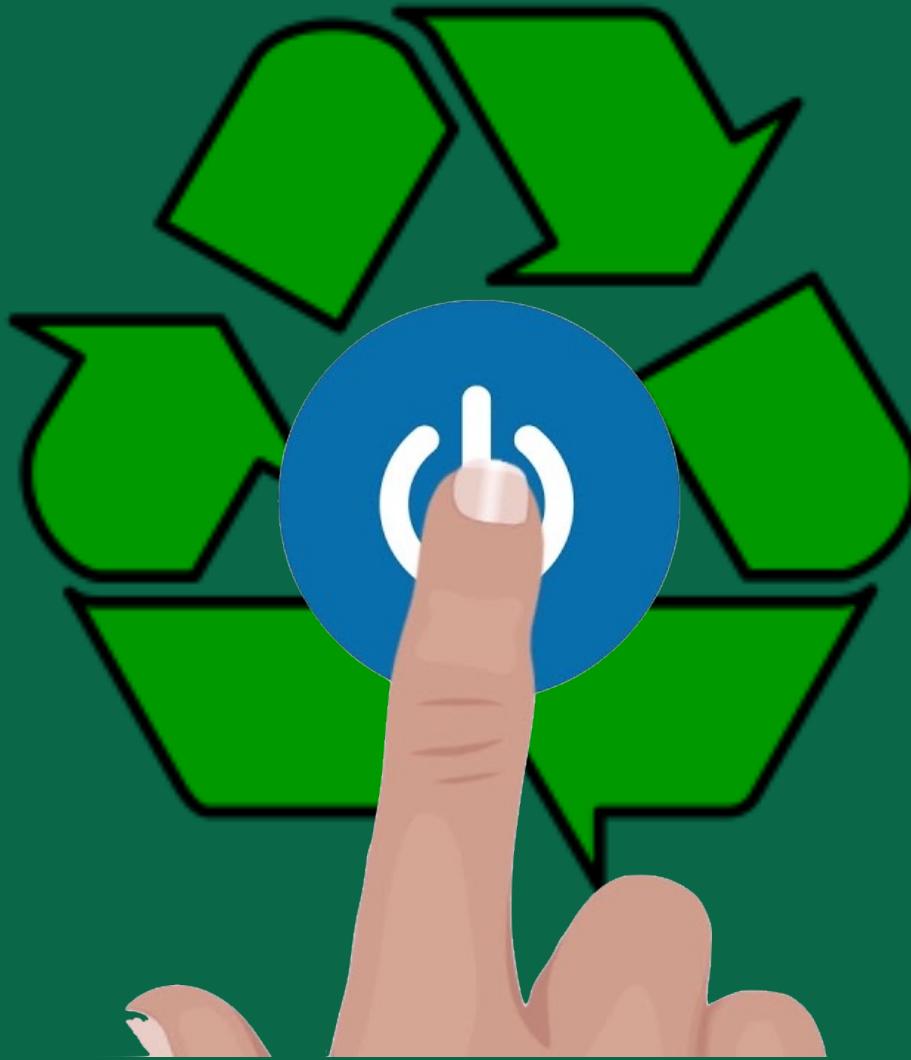
GBIF IS A FAIR & OPEN BIODIVERSITY DATA INFRASTRUCTURE



INCENTIVES 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





DATA CITATION - A NEW CURRENCY OF SCIENCE

- **Peer-reviewed scholarly papers** in high impact journals still 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 access scientific research products...**



DECLARATION ON RESEARCH ASSESSMENT



- DORA recognizes the need to improve the ways in which the outputs of scholarly research are evaluated.
- Worldwide movement covering all scholarly disciplines and all key stakeholders including funders, publishers, professional societies, institutions (universities), and researchers.
- Developed in 2012 in San Francisco
- To date (2023-11-17), 21 157 individuals and 3 031 organizations in 164 countries have signed DORA.
- The Research Council of Norway (RCN) signed DORA in May 2018 [\[link\]](#)

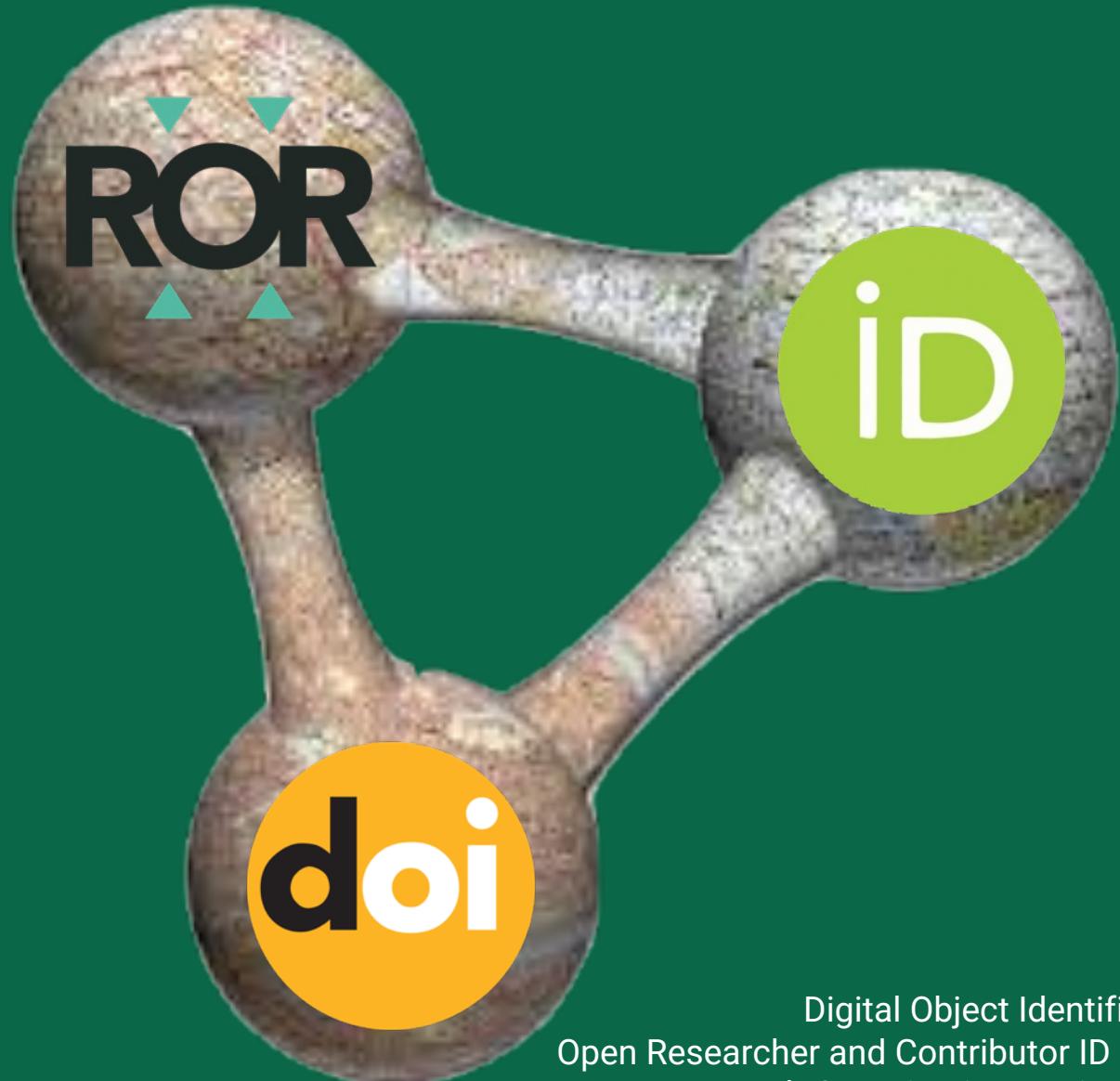
MACHINE-READABILITY REQUIRES PERSISTENT IDENTIFIERS

*The purpose of identifiers is
... to name things
... making it possible to refer to them*



ROR for institutions
ORCID for people
DOI for datasets
GRSciColl UUID for collections

will enable the linking of museum collection specimens to scientific litterature and scientific actors (authors, curators, etc)



Digital Object Identifier (DOI)
Open Researcher and Contributor ID (ORCID)
Research Organisation Registry (ROR)
Universally Unique Identifier (UUID)



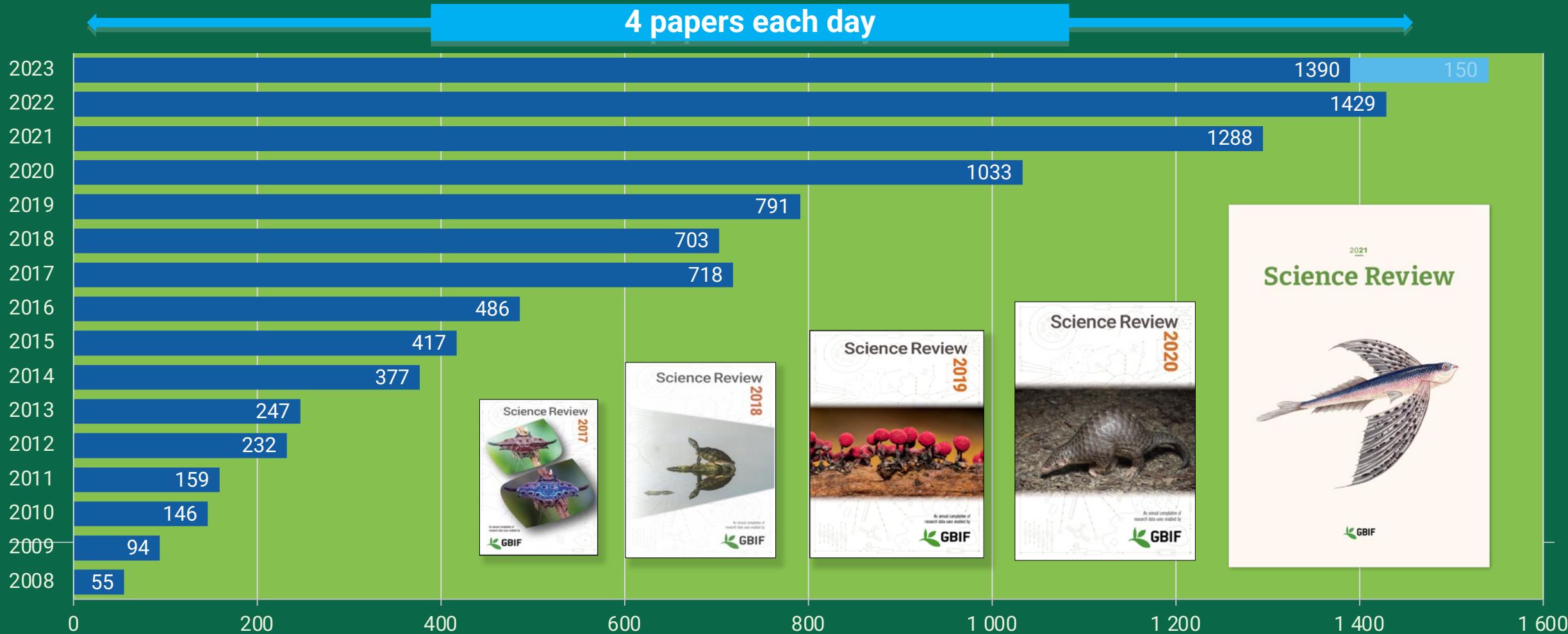
GBIF

#CiteTheDOI

GBIF started issuing DOIs on 3 February 2015



PEER-REVIEWED PUBLICATIONS USING GBIF-MEDIATED DATA

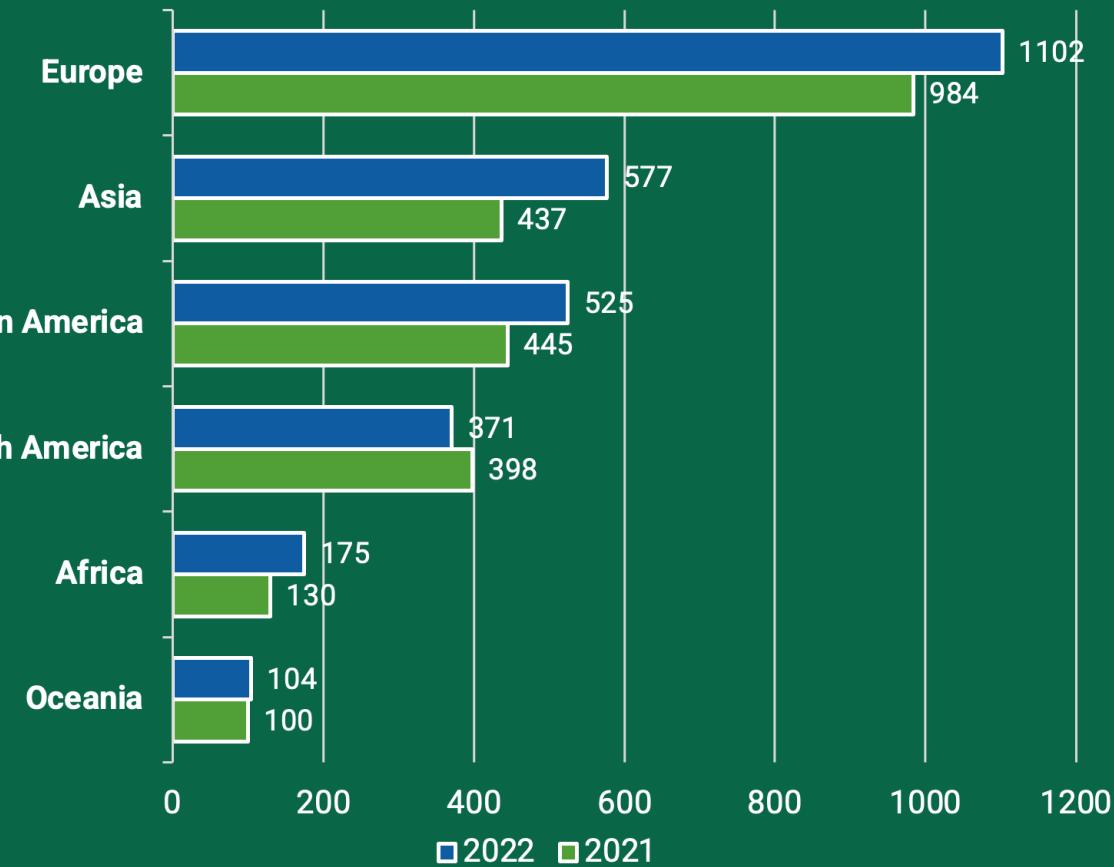


DATA USE IN PEER-REVIEWED JOURNALS

Peer-reviewed uses by country

<i>End of Year</i>		2023	2022	2021
1	United States	306	292	328
2	China	300	300	228
3	Brazil	145	152	147
4	United Kingdom	144	146	137
5	Germany	126	120	131
6	Mexico	105	131	134
7	Spain	95	100	103
8	France	88	75	75
9	Australia	80	78	69
10	India	61	62	35
-	Norway	30	41	37

Peer-reviewed uses by region



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 (9 November 2021) GBIF Occurrence Download <https://doi.org/10.15468/dl.xxxxxx>

DOI BASED DATA CITATION AT GBIF.ORG -- #CITETHEDOI

paper

J S E Journal of Systematics and Evolution

Research Article | Free Access |

Phylogenomics, biogeography, and evolution of the blue- or white-fruited dogwoods (*Cornus*)—Insights into morphological and ecological niche divergence following intercontinental geographic isolation

Kira Lindelof, Juliete A. Lindo, Wenbin Zhou, Xiang Ji, Qiu-Yun (Jenny) Xiang

First published: 27 August 2020 | <https://doi.org/10.1111/jse.12676> | Citations: 1

SECTIONS PDF TOOLS SHARE

Abstract

The eastern Asian (EA)-eastern North American (ENA) floristic disjunction represents a major pattern of phytogeography of the Northern Hemisphere. Despite 20 years of studies dedicated to identification of taxa that display this disjunct pattern, its origin and evolution remain an open question, especially regarding post-isolation evolution. The blue- or white-fruited dogwoods (BW) are the most species-rich among the four major clades of *Cornus* L., consisting of ~35 species divided into three subgenera (subg. *Yinquaia*, subg. *Mesomora*, and subg. *Kranopsis*). The BW group provides an excellent example of the EA-ENA floristic disjunction for biogeographic study due to its diversity distribution centered in eastern Asia and eastern North America, yet its species relationships and delineation have remained poorly understood. In this study, we combined genome-wide markers from RAD-seq, morphology, fossils, and climate data to understand species relationships, biogeographic history, and ecological niche and morphological evolution. Our phylogenomic analyses with RAxML and MrBayes recovered a strongly supported and well-resolved phylogeny of the BW group with three intercontinental disjunct clades in EA and ENA or Eurasia and North America, of which two are newly identified within subg. *Kranopsis*. These analyses also recovered a

References

Adams DC, Berns CM, Kozak KH, Wiens JJ. 2009. Are rates of species diversification correlated with rates of morphological evolution? *Proceedings Biological Sciences* 276: 2729–2738.

Crossref | PubMed | Web of Science® | Google Scholar

GBIF.org. 2020. GBIF Occurrence Download. Available from <https://doi.org/10.15468/dl.vxxu7>

Google Scholar

GBIF.org. 2020. GBIF Occurrence Download. Available from <https://doi.org/10.15468/dl.vrvfg>

Google Scholar

GBIF.org. 2020. GBIF Occurrence Download. Available from <https://doi.org/10.15468/dl.4utwft>

Google Scholar

GBIF.org. 2020. GBIF Occurrence Download. Available from <https://doi.org/10.15468/dl.yp3my7>

Google Scholar



data citations

High silicon concentrations in grasses are linked to environmental conditions and not associated with C4 photosynthesis

Brightly, W., Hartley, S., Osborne, C., Simpson, K., Strömberg, C. (2019) Global Change Biology, 25, 201–210

The uptake and deposition of silicon (Si) as silica phytoliths in grasses has received little attention, partly because grasses are well known for their high Si...

C4 photosynthesis + grass + grassland + herbivore + phytolith +

Journal article | Peer-reviewed

Data referenced in study DOI 10.15468/dl.heuwwg

Temperature change in subtropical southeastern Africa during the past 790,000 yr

Hanmer, M., Chappell, J., Quick, L., Dupont, L., Johnson, T. (2020) Geology

Across the glacial-interglacial cycles of the late Pleistocene (~700 k.y.), temperature variability at low latitudes is often considerably more variable than precipitation, but reliable assessment of this variability has been difficult compared to changes in precipitation. However, a paucity of quantified temperature re...

Journal article | Open access | Peer-reviewed

Data referenced in study DOI 10.15468/dl.7g3ylt DOI 10.15468/dl.czfuzq DOI 10.15468/dl.hb91rx DOI 10.15468/dl.ln2zvq DOI 10.15468/dl.lpx7qf DOI 10.15468/dl.mv0dln DOI 10.15468/dl.prwbfz DOI 10.15468/dl.ridxz

Landscape Analysis for the Specimen Data Refinement

Walton, S., Livermore, L., Bánki, O., Cubey, R., Drinkwater, R., Englund, M., ... (2020) Methods in Ecology and Evolution

This report reviews the current state-of-the-art applied approaches on automated tools, methods and workflows for extracting information from images of natural history specimens and their labels. We consider the potential for repurposing existing tools, including workflow management systems; and ar...

collections digitisation + data reconciliation + data refinary + digitisation + linked open data + machine learning +

Journal article | Open access | Peer-reviewed

Data referenced in study DOI 10.15468/dl.8pq57z

Phylogenomics, biogeography, and evolution of the blue- or white-fruited dogwoods (*Cornus*)—Insights into morphological and ecological niche divergence

Lindelof, K., Lindo, J., Zhou, W., Ji, X., Xiang, Q. (2020) Journal of Systematics and Evolution

The eastern Asian (EA)-eastern North American (ENA) floristic disjunction represents a major pattern of phytogeography of the Northern Hemisphere. Despite 20 years of studies dedicated to identification of taxa that display this disjunct pattern, its origin and evolution remain an open question, esp...

Cornus + RAD-seq + biogeography + eastern Asian-eastern North American disjunction + ecological niche and morphospace + phylogenomics

Journal article | Peer-reviewed

Data referenced in study DOI 10.15468/dl.4utwft DOI 10.15468/dl.b1wpp DOI 10.15468/dl.63wryf DOI 10.15468/dl.6pwksh DOI 10.15468/dl.bbdby DOI 10.15468/dl.cpw44s DOI 10.15468/dl.dspfk DOI 10.15468/dl.gmgers DOI 10.15468/dl.iizqbs DOI 10.15468/dl.juxfmb DOI 10.15468/dl.jsfbnp DOI 10.15468/dl.ktk3br DOI 10.15468/dl.mnrvg DOI 10.15468/dl.ooy9bm DOI 10.15468/dl.raghj DOI 10.15468/dl.uoxfrd DOI 10.15468/dl.vxgfp DOI 10.15468/dl.vxu07g DOI 10.15468/dl.xvexc5 DOI 10.15468/dl.yp3my7

Expansion dynamics and marginal climates drive adaptation across geographic ranges

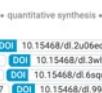
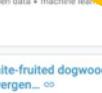
Bontrager, M., Usui, T., Lee-Yau, J., Anstett, D., Branch, H., Hargreaves, A., ... (2020) bioRxiv

Every species experiences limits to its geographic distribution. Some evolutionary models predict that populations at range edges are less well-adapted to their local environments due to drift, expansion load, or swamping gene flow from the range interior. Alternatively, populations near range edges...

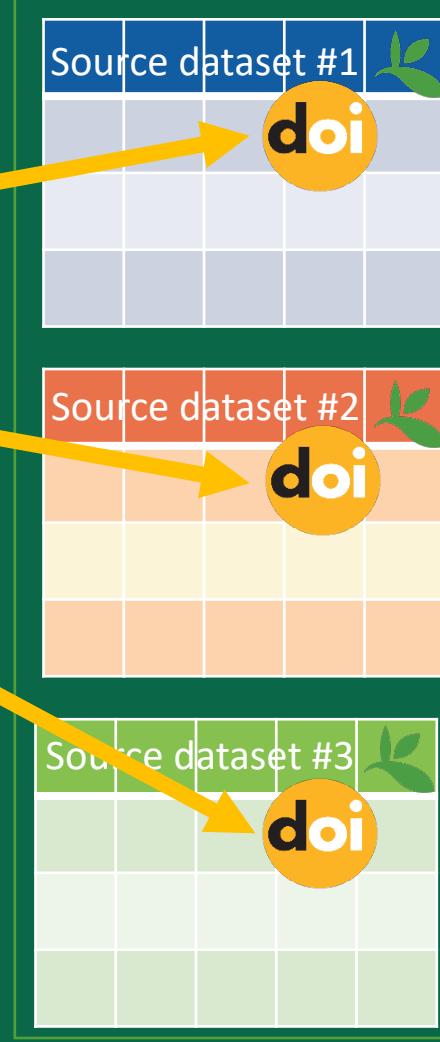
Local adaptation + expansion load + geographic range limit + peripheral population + quantitative synthesis + transplant experiment

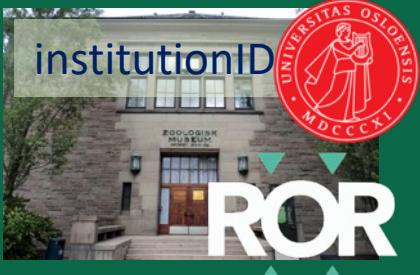
Working paper | Open access

Data referenced in study DOI 10.15468/dl.0kmuse DOI 10.15468/dl.zogic DOI 10.15468/dl.zu0bec DOI 10.15468/dl.2uz668 DOI 10.15468/dl.37k2g DOI 10.15468/dl.3trzu DOI 10.15468/dl.swtml DOI 10.15468/dl.4kk9ys DOI 10.15468/dl.4oubjy DOI 10.15468/dl.4vrryy DOI 10.15468/dl.eeqza DOI 10.15468/dl.7dp771 DOI 10.15468/dl.8baigq DOI 10.15468/dl.kldw7 DOI 10.15468/dl.99zbz DOI 10.15468/dl.9me17 DOI 10.15468/dl.bnpgxd DOI 10.15468/dl.aagz28 DOI 10.15468/dl.ansqql DOI 10.15468/dl.aq747g DOI 10.15468/dl.aq7d7 DOI 10.15468/dl.bfrmf2 DOI 10.15468/dl.bguizk DOI 10.15468/dl.buz2kg DOI 10.15468/dl.buger DOI 10.15468/dl.bfrf1 DOI 10.15468/dl.cbrv7 DOI 10.15468/dl.ce19ah DOI 10.15468/dl.copurn DOI 10.15468/dl.dhsf4 DOI 10.15468/dl.drn9ab DOI 10.15468/dl.dt440v DOI 10.15468/dl.e2zog DOI 10.15468/dl.e6na78 DOI 10.15468/dl.egiacr



datasets





Dataset DOIs

Source dataset #1

doi

ID



UUID



Source dataset #2

doi

ID



UUID



Source dataset #3

doi

ID



UUID



Download DOI

GBIF download

doi

ID



UUID



Process & archive

Final state of data

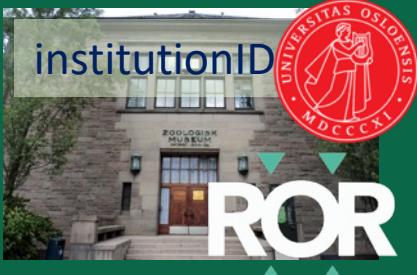
ID

UUID

zenodo doi

Bibliographic DOI





Dataset DOIs

Source dataset #1

doi

iD

ROR

UUID



Source dataset #2

doi

iD

ROR

UUID



Source dataset #3

doi

iD

ROR

UUID



Download DOI

GBIF download

doi

ROR

UUID



Final state of data

zenodo

doi

Process & archive

DOI

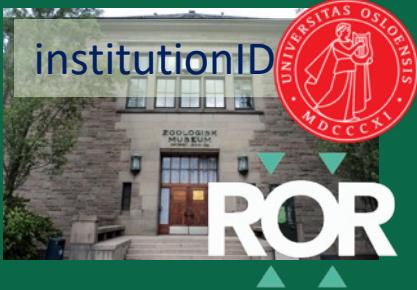
DOI

DOI

DOI

Bibliographic DOI





Dataset DOIs

Source dataset #1	doi
UUID	
UUID	
UUID	
UUID	
Source dataset #2	doi
UUID	
UUID	
UUID	
UUID	
Source dataset #3	doi
UUID	
UUID	
UUID	
UUID	

Download DOI

Possible with persistent object identifiers

GBIF download	doi
UUID	
UUID	
UUID	
UUID	

Analyze & publish

Process & archive

Final state of data
UUID
UUID
UUID
UUID



Dataset DOIs

Source dataset #1
doi

Source dataset #2
doi

Source dataset #3
doi

Download DOI

Possible with persistent **external identifiers**

GBIF download
doi

ROR
WIKIDATA
UUID
ID

Process & archive

Final state of data

zenodo
doi

Analyze & publish

Bibliographic DOI





GBIF

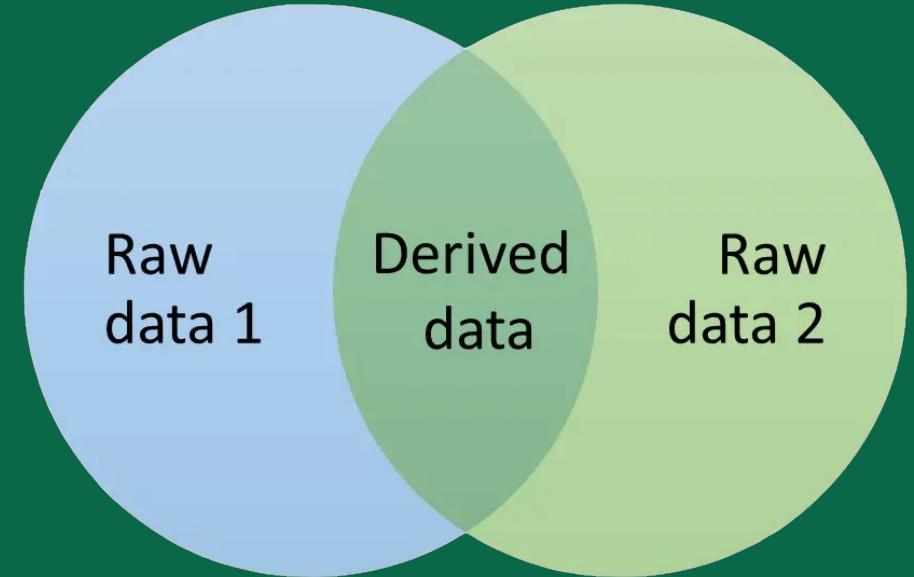
Derived Dataset

Create your own citable dataset DOI



DERIVED DATASET

- a GBIF.org download that has been **filtered/reduced significantly**, or
- data accessed through a cloud service, e.g. Microsoft AI for Earth (Azure), or
- data obtained by any means for which no DOI was assigned, but one is required (e.g. third-party tools accessing the GBIF search API)
 - ... *including rg bif*



HOW TO CREATE A DERIVED DATASET

When created, a derived dataset is assigned a unique DOI that can be used to cite the data.

To create a derived dataset, you will need to authenticate using a GBIF.org account and provide:

- a **title** of the dataset,
- a **list of the GBIF datasets** (by DOI or datasetKey) from which the data originated, ideally with **counts** of how many records each dataset contributed,
- a **persistent URL** of where the extracted dataset can be accessed,
 - [e.g. Zenodo, Dataverse, Dryad, ...]
- a **description** of how the dataset was prepared,
- (*optional*) the GBIF download DOI, if the dataset is derived from an existing download , and
- (*optional*) a date for when the derived dataset should be registered if not immediately .

THANK YOU

GBIF Norway

Dag Endresen | Node Manager

helpdesk@gbif.no



www.gbif.org

