BIODT
biodiversitydigitaltwin

School & Hackaton

March 3rd - March 7th
Officine Cantelmo, Lecce IT

Open Science and Data citation

#CiteTheDOI

Dag Endresen – GBIF Norway, University of Oslo

Funded by
the European Union

Dag Endresen
University of Oslo
GBIF Norway

March 2025

GBIF Global Biodiversity Information Facility

BioDT winter school, Lecce, Italy| 4th March 2025

- Open science
- Data citation
- Metadata
- Data papers
- GBIF #CiteTheDOI
- GBIF data API

Why teach students open Science ?

- We are in the middle of an **ongoing paradigm shift** in scientific practice (and impact metrics).
- Young scientists will (already today) need **different skills**, than was needed previously – to succeed in academia.

What is Open Science?

- **Open science** is the movement to make scientific research (*including publications, data, physical samples, and software*) and its dissemination accessible to all levels of an inquiring society, amateur or professional (Woelfle *et al*. 2011).
- **Open science** is transparent and accessible knowledge that is shared and developed through *collaborative networks* (Vicente-Saez *et al*. 2018).
- **Open Science** can be seen as a *continuation of*, rather than a revolution in, practices begun in the 17th century with the advent of the *academic journal* (David 2004).

- Woelfle, M.; Olliaro, P.; Todd, M. H. (2011). "Open science is a research accelerator". Nature Chemistry. 3 (10): 745–748. doi:10.1038/nchem.1149
- Vicente-Saez, Ruben; Martinez-Fuentes, Clara (2018). "Open Science now: A systematic literature review for an integrated definition". Journal of Business Research. 88: 428–436. doi:10.1016/j.jbusres.2017.12.043
- David, P. A. (2004). "Understanding the emergence of 'open science' institutions: Functionalist economics in historical context". Industrial and Corporate Change. 13 (4): 571–589. doi:10.1093/icc/dth023
- https://en.wikipedia.org/wiki/Open_science

Data Citation Principles

FORCE11
The Future of Research Communications and e-Scholarship

1. Data to be legitimate **citable** products of research.
2. Data citations giving **scholarly credit** and attribution.
3. In scholarly literature, whenever claims are based on data, **data should always be cited**.
4. Persistent method for **identification of data**, that is machine actionable, globally unique, universal.
5. Data citation facilitate **access to data** or at least to **metadata**.
6. **Unique identifiers** that persist even beyond the lifespan of the data.
7. Data citation identify and access the specific data that **support verification** of the claim (provenance, time-slice, version).
8. Flexible, but attention to **interoperability** of practices across communities.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

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

Why publish open data?

- Data produced using public funds should be regarded as a common good, and should be made available for inspection, interpretation and re-use by third parties.
- Needless duplication of data-collecting efforts and costs will be reduced.
- Open data increases transparency and overall quality of research.
- Published data can be re-analysed, verified, and improved by others.
- Data publication increase recognition and opportunities for collaboration.
- Published data can be cited and re-used, either alone or in combination with other data.
- Data owners and collection managers can trace data use and citation.
- Data creators, their institutions and funding agencies can be credited.
- Data can be integrated with other datasets across space and time.
- Open data increases potential for interdisciplinary research and re-use in new contexts not envisioned by the data creator.

Penev *et al.* (2017) https://doi.org/10.3897/rio.3.e12431

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

- 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, 22 518 individuals and 3 433 organizations in 166 countries have signed DORA.
- The Research Council of Norway (RCN) signed DORA in May 2018 [link]

Academic policies - NOR-CAM (2021)

Guiding principles for the assessment and evaluation of research(ers) in light of the transition to Open Science. Proposes a more holistic framework for recognition and rewards in academic research assessment.

Academic institutions when recruiting academic and scientific staff; in processes related to promotion from one scientific career stage to the next; career guidance of employees by department and research leaders

Funders when assessing project managers and participants in connection with research applications.

National authorities when evaluating of Norwegian research and education.

https://www.uhr.no/en/news-from-uhr/nor-cam-a-toolbox-for-recognition-and-rewards-in-academic-careers.5780.aspx

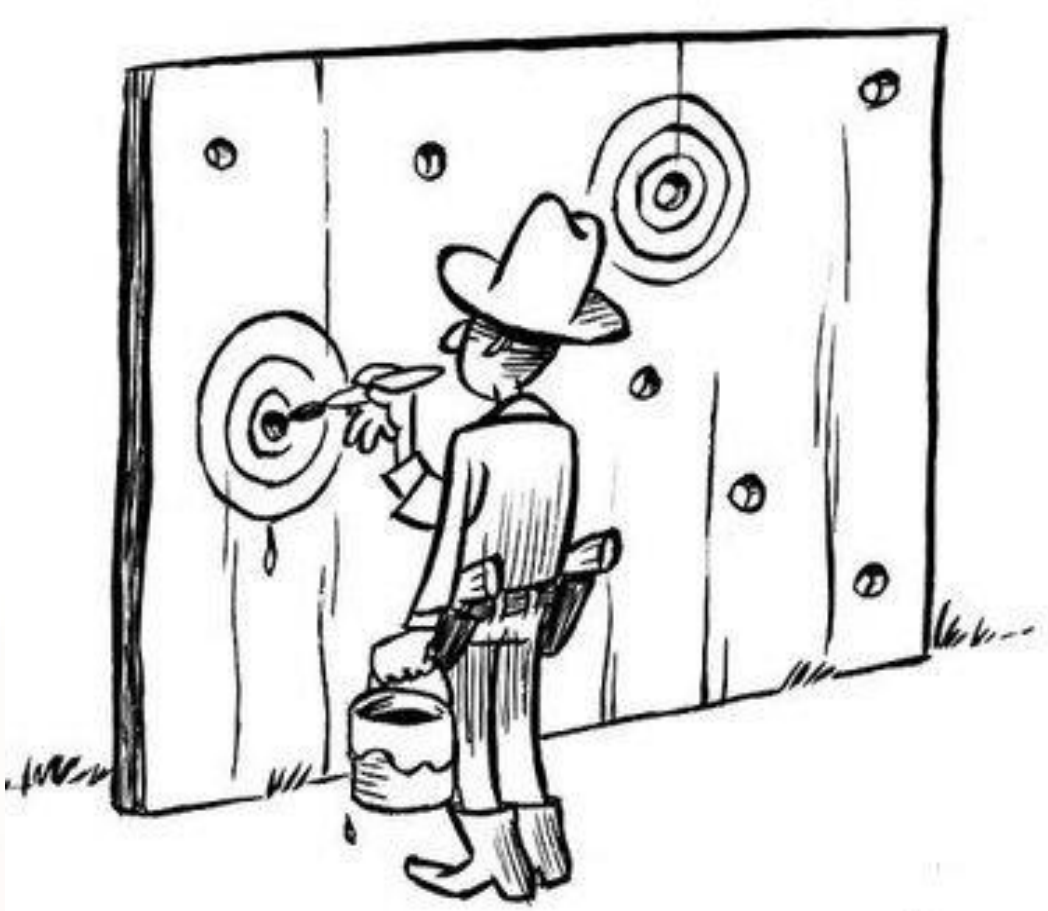
Reproducibility crisis

"***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

Will anybody trust closed science again?

- Recent studies indicates that **p-hacking** [1] is a significant problem – sometimes even without the scientist even being aware of doing so (Ioannidis 2005; Head *et al*. 2015)
- **Pre-registered (open) data** provides a good insurance against suspicion of both data dredging (and plain data falsification).



[1] "p-hacking," (data dredging, data fishing, …) occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant.

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

Meta-
data

Why metadata?

In general, metadata should allow a prospective end user of data to:

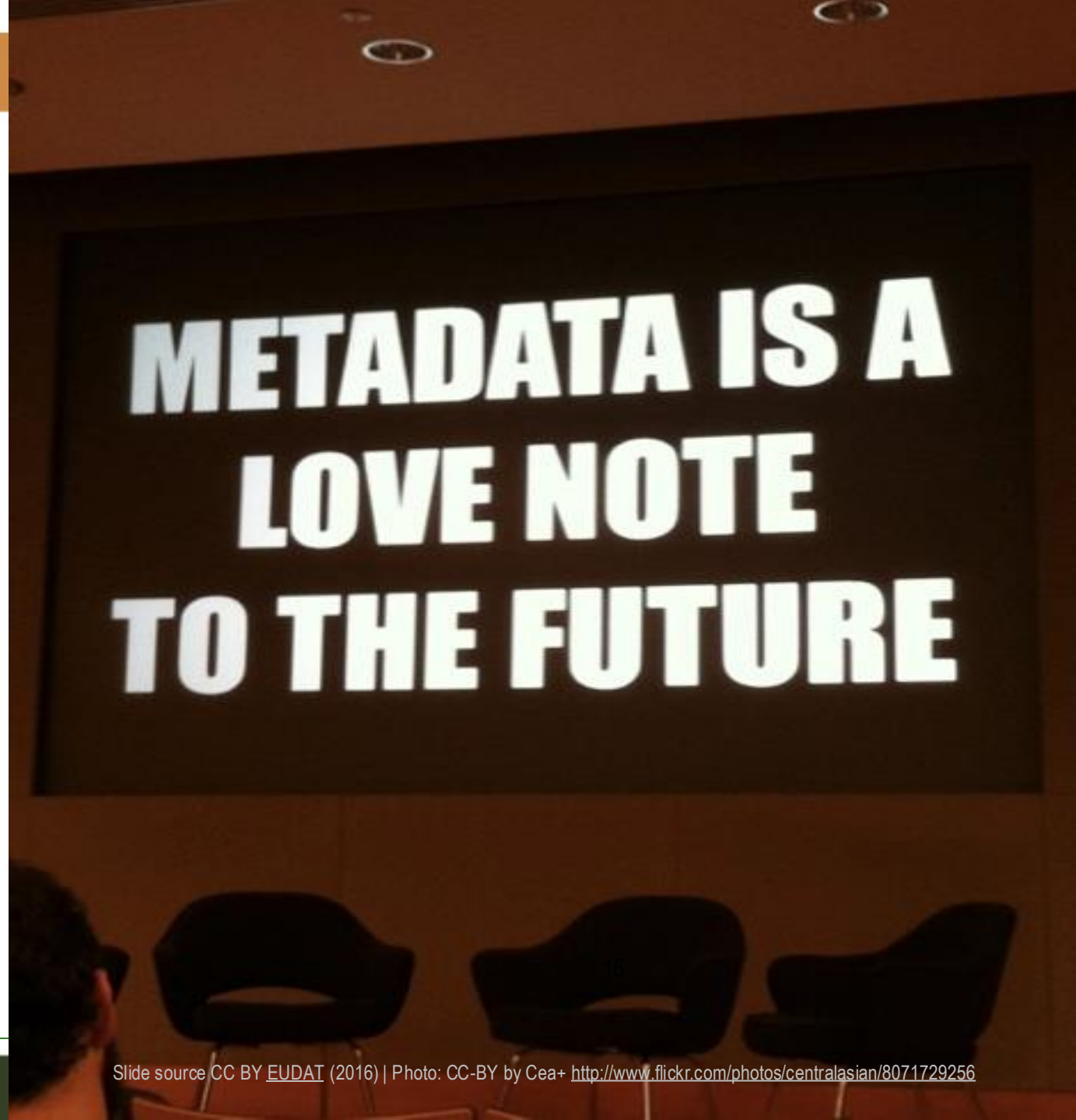
1. identify/discover its existence,
2. learn how to access or acquire the data,
3. understand its fitness-for-use,
4. learn how to transfer (obtain a copy of) the data, and
5. learn how the data should be used.

Photo CC BY-SA Jennifer Fagan-Fry (NOAA) | GBIF Metadata Profile (2011) https://github.com/gbif/ipt/wiki/GMPHowToGuide

BIODT
biodiversitydigitaltwin

What is Metadata?

Metadata, literally "***data about data***" are an essential component of a data management system, describing such aspects as the "what, where, when, who and how" pertaining to a resource.



Slide source CC BY EUDAT (2016) | Photo: CC-BY by Cea+ http://www.flickr.com/photos/centralasian/8071729256

Data entropy

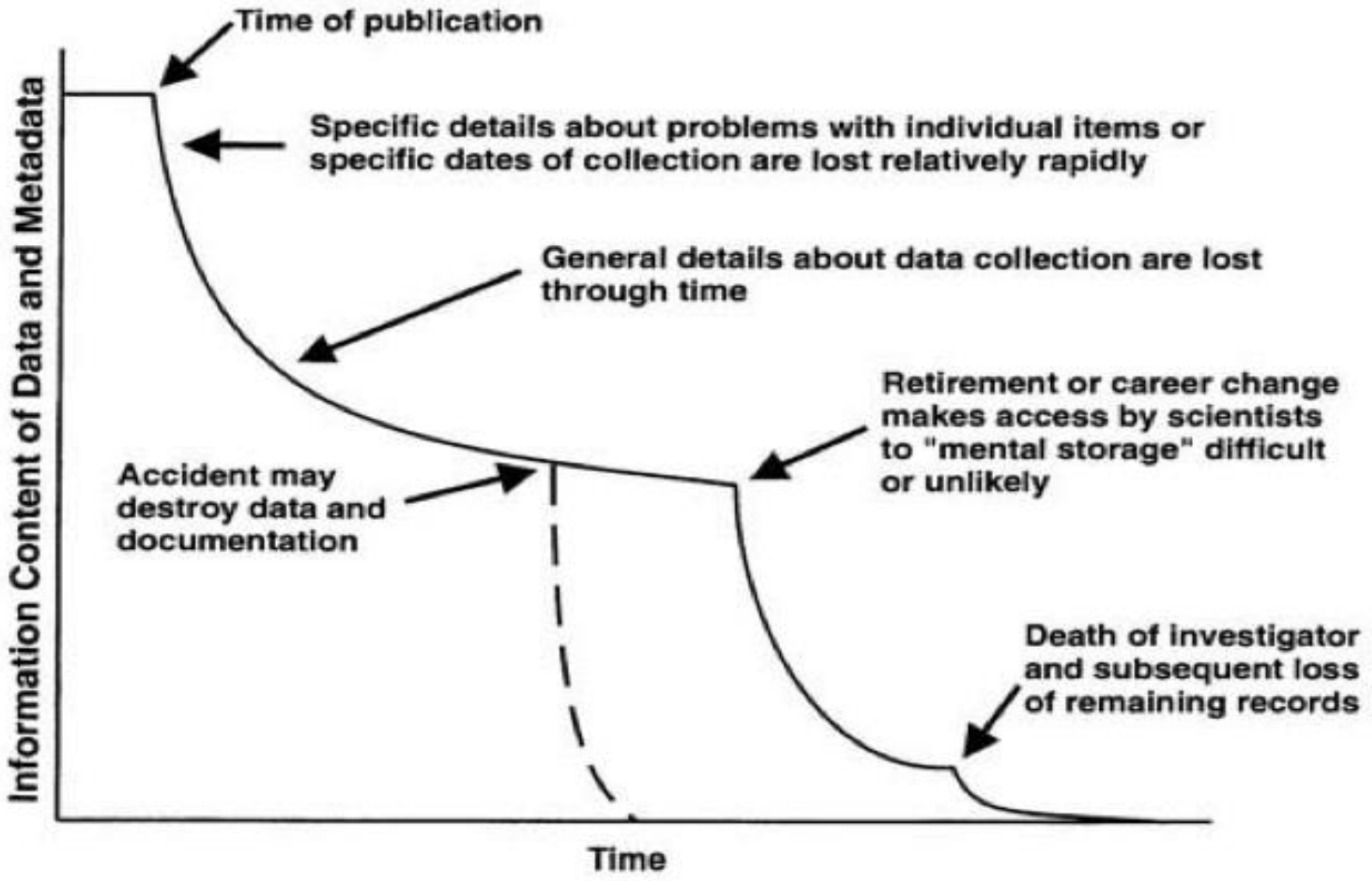
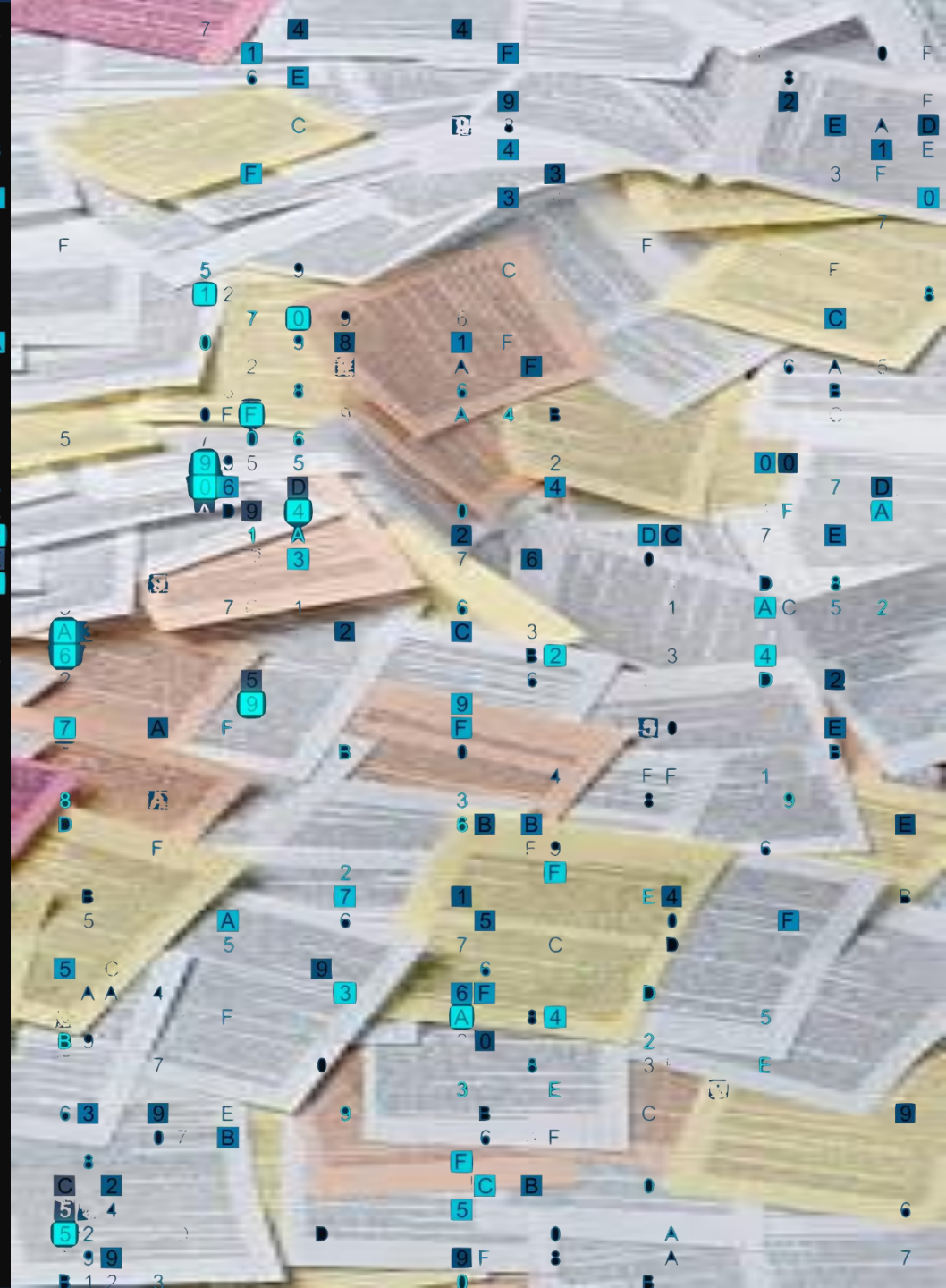


Illustration from: The Loss of Information about Data (Metadata) Over Time, Michener *et al*, 1997

BIODT
biodiversitydigitaltwin
School & Hackaton

Data papers

Data Paper



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.

- Getting **scholarly recognition** for your datasets.
- Promote and improve the **fitness for reuse** of your research data.

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

Data papers explained

A data paper is a metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal.

Unlike a conventional research article, the primary purpose of a data paper is to **describe data** and the circumstances of their collection, rather than to report hypotheses and conclusions.

GBIF has been working with partners in academic publishing to promote the data paper as a means of bringing **credit and recognition** to all those involved in data publication.

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

Why publish data papers?

Authoring clear, informative **metadata** is an essential step if biodiversity data are going to be discovered and used to inform research and decisions. **This involves extra work, and data publishers need incentives to do it**. In the absence of such incentives, too many datasets are published with poorly-documented metadata or, worse still, no metadata at all.

Data papers help to overcome barriers to **authoring of metadata** by **providing clear acknowledgement of all those involved** in the collection, management, curation and publishing of biodiversity data.

By publishing a data paper

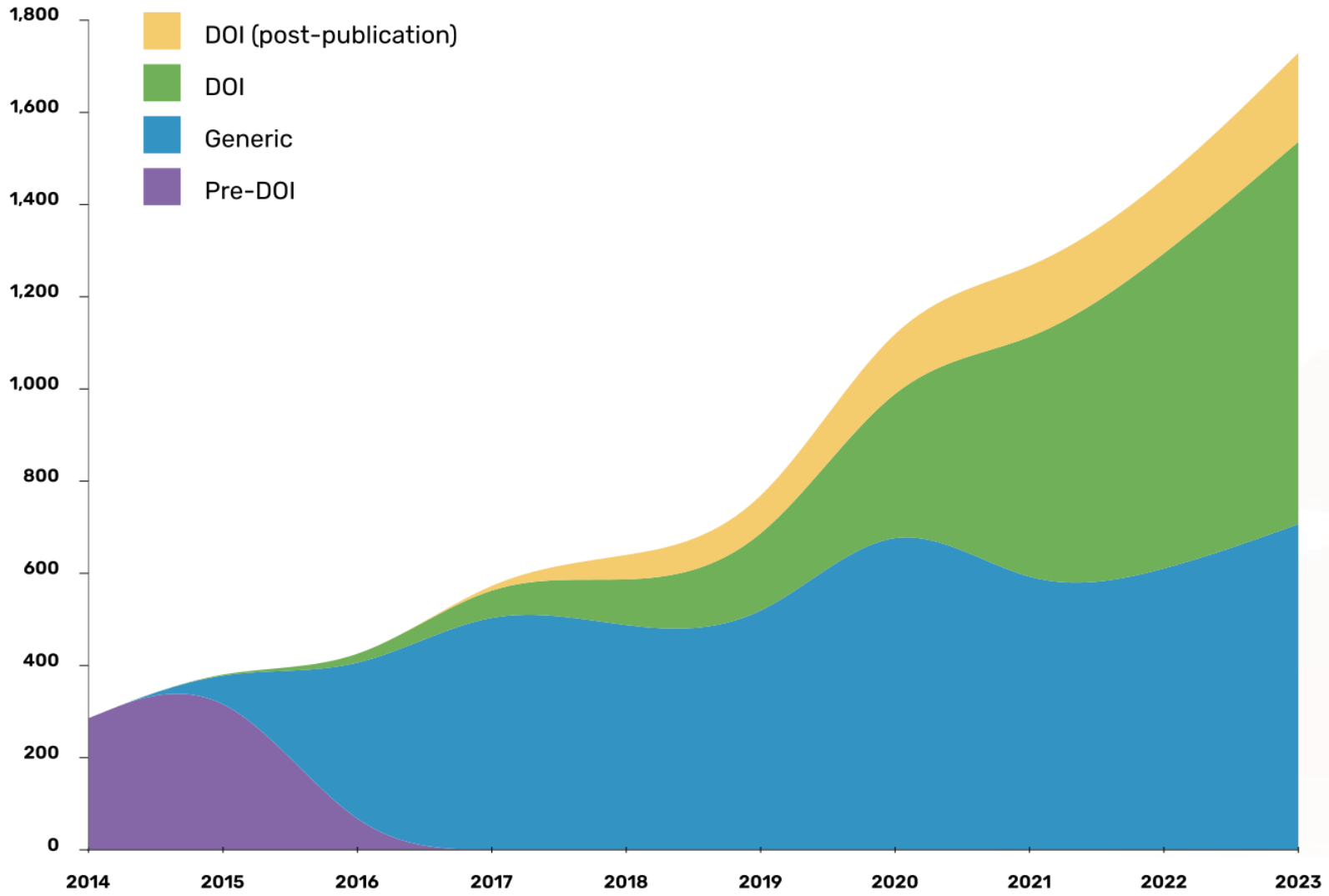
Receive credit through indexing and **citation of the published paper**, in the same way as with any conventional scholarly publication, offering benefits to authors in terms of recognition and career building.

Increase the visibility, usability and credibility of the data resources you publish.

Track more effectively the **usage** and **citations** of the data you publish.

BIODT
biodiversitydigitaltwin
School & Hackaton

GBIF Data Citation



#CiteTheDOI

- The number of peer-reviewed journal articles citing GBIF-mediated data has been steadily growing over the past 10 years. This graph illustrates the growth, also indicating the manner in which papers cite GBIF.

GBIF data citation

- The gold standard and recommended practice for data citations includes a persistent identifier (a digital object identifier or DOI), which a narrow majority of GBIF users follow.
- GBIF works with authors, editors and publishers to improve the quality of data citations, enhancing reproducibility of scientific analyses and ensuring credit for data publishers.

BIODT
biodiversitydigitaltwin

Get data Share Tools Inside GBIF Login

OCCURRENCE DATASET | REGISTERED 30 JULY 2012

Vascular Plant Herbarium, Oslo (O)

Published by Natural History Museum, University of Oslo

Reidar Elven

DATASET STATS ACTIVITY DOWNLOAD

855,163 OCCURRENCE 21 CITATIONS

40,738 download events

4 OCCURRENCES FROM THIS DATASET

DOI 10.15468/dl.enncoc
Date: 9 November 2017

Occurrences: 4,801
Involved Datasets: 150

Scientific name Echeveria dc.

RERUN QUERY

4 OCCURRENCES FROM THIS DATASET

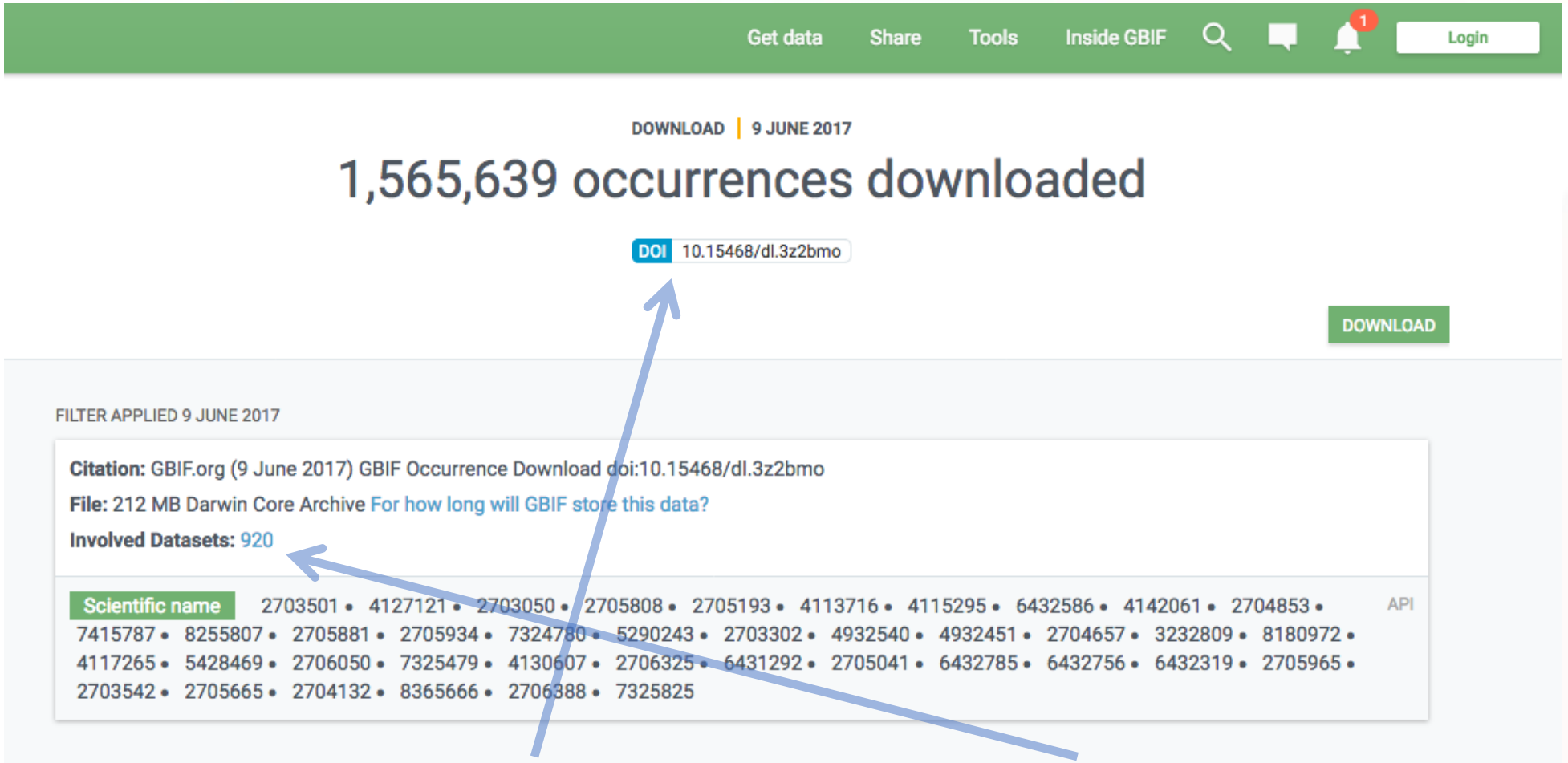
DOI 10.15468/dl.kfv400
Date: 9 November 2017

Occurrences: 4,801
Involved Datasets: 150

Scientific name Echeveria dc.

RERUN QUERY

S&H



Citing the data **download DOI** will resolve to the **dataset DOIs** assigned for each dataset contributing data records to the download set.

This way, all data publishers contributing data records will be **accredited**!

You receive an email when the download file is ready

Your GBIF data download is ready

Inbox x

downloads@gbif.org 06:29 (0 minutes ago)

to me

Hello dag.endresen,

Your download is available at the following address:
http://api.gbif.org/v1/occurrence/download/request/0070322-160910150852091.zip

When using this dataset please use the following citation:
GBIF.org (16th March 2017) GBIF Occurrence Download http://doi.org/10.15468/dl.uzesiq

Download Information:
DOI: http://doi.org/10.15468/dl.uzesiq (may take some hours before being active)
Creation Date: Thursday, March 16, 2017 5:57:21 AM CET
Records included: 205938 records from 292 published datasets
Data size: 21.2 MB
Download format: DWCA
Filter used: TaxonKey: Hordeum vulgare L.

This download can always be viewed on
http://www.gbif.org/occurrence/download/0070322-160910150852091

For information about download formats please visit
http://www.gbif.org/faq/datause

BIODT
biodiversitydigitaltwin

Get data Share Tools Inside GBIF 1 dag.endresen

PROFILE DOWNLOADS

DOI 10.15468/dl.uzesiq

Scientific name Hordeum vulgare l.

RERUN QUERY

DOI 10.15468/dl.mll0bw
Date: 23 February 2017

Occurrences: 1,642
Involved Datasets: 1

And
- Dataset key: A global database for the distributions of crop wild relatives
- Issue: Taxon match none

DOWNLOAD RERUN QUERY

DOI 10.15468/dl.z7jfht
Date: 27 May 2016

Occurrences: 390,329
Involved Datasets: 39

And
- Scientific name: 2704322 • 5359246 • 2993761 • 2882940 • 5359047 • 8400458 • 2705297 • 5376426 • 2986192 • 2875979 • 2975014 • 3140552 • 3034742 • 2998290 • 2984535 • 2704922 • 3140270 • 5383920 • 3001509 • 2992543 • 2986185 • 2704178 • 3041022 • 5358748 • 5375547 • 5374614 • 8324121 • 3054368 • 7931731 • 5289684 • 8073364 • 7903057 • 7515176 • 2705308 • 6109535 • 3021922 • 2986097 • 8350369 • 2706005 • 2855860 • 5708780 • 4129708 • 3020791 • 3047598 • 2704261 • 5373533 • 2706241 • 3023221 • 5358812 • 7580783 • 2995209 • 2975076 • 8255803 • 3001244 • 3140410 • 2992051 • 3029817 • 3140490 • 2965201 • 2965324 • 2974915 • 2704951 • 3042658 • 8300875 • 5359499 • 3022789 • 3021730 • 5358683 • 2882835 • 8270537 • 2768885 • 2882833 • 2996525 • 8235931 • 8344892 • 3034714 • 5362573 • 2986129 • 2993094 • 2704955
- Country: Sweden
- Has coordinate: True
- Year: After 1970

DOWNLOAD RERUN QUERY

S&H

We store all download files as long as possible.

The download metadata page will always resolve, but the file itself might be removed in the future.

We strive to store all downloads but prioritize downloads that have been cited.

BIODT
biodiversitydigitaltwin
School & Hackaton

#CiteTheDOI

BIODT
biodiversitydigitaltwin

GBIF

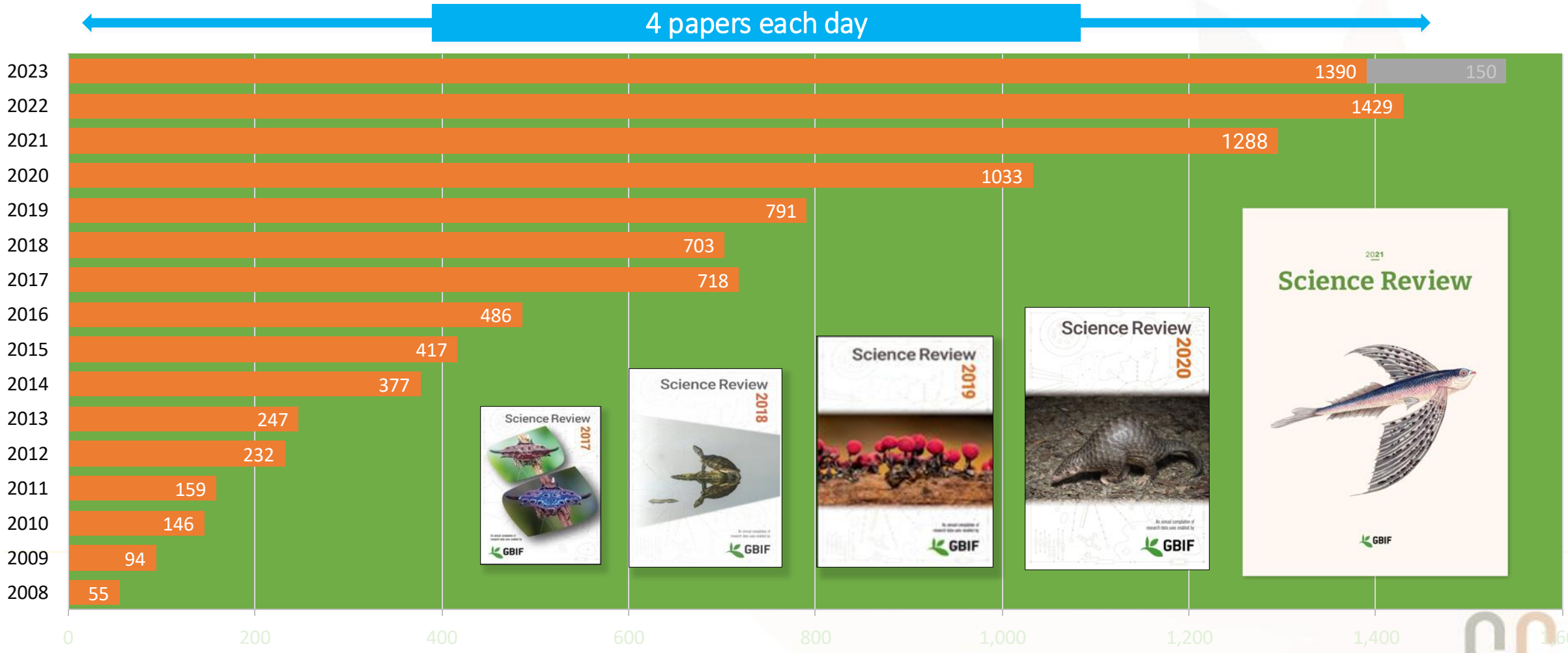
#CiteTheDOI

GBIF started issuing DOIs on 3 February 2015

S&H

Peer-reviewed publications using GBIF-mediated data

Title Here

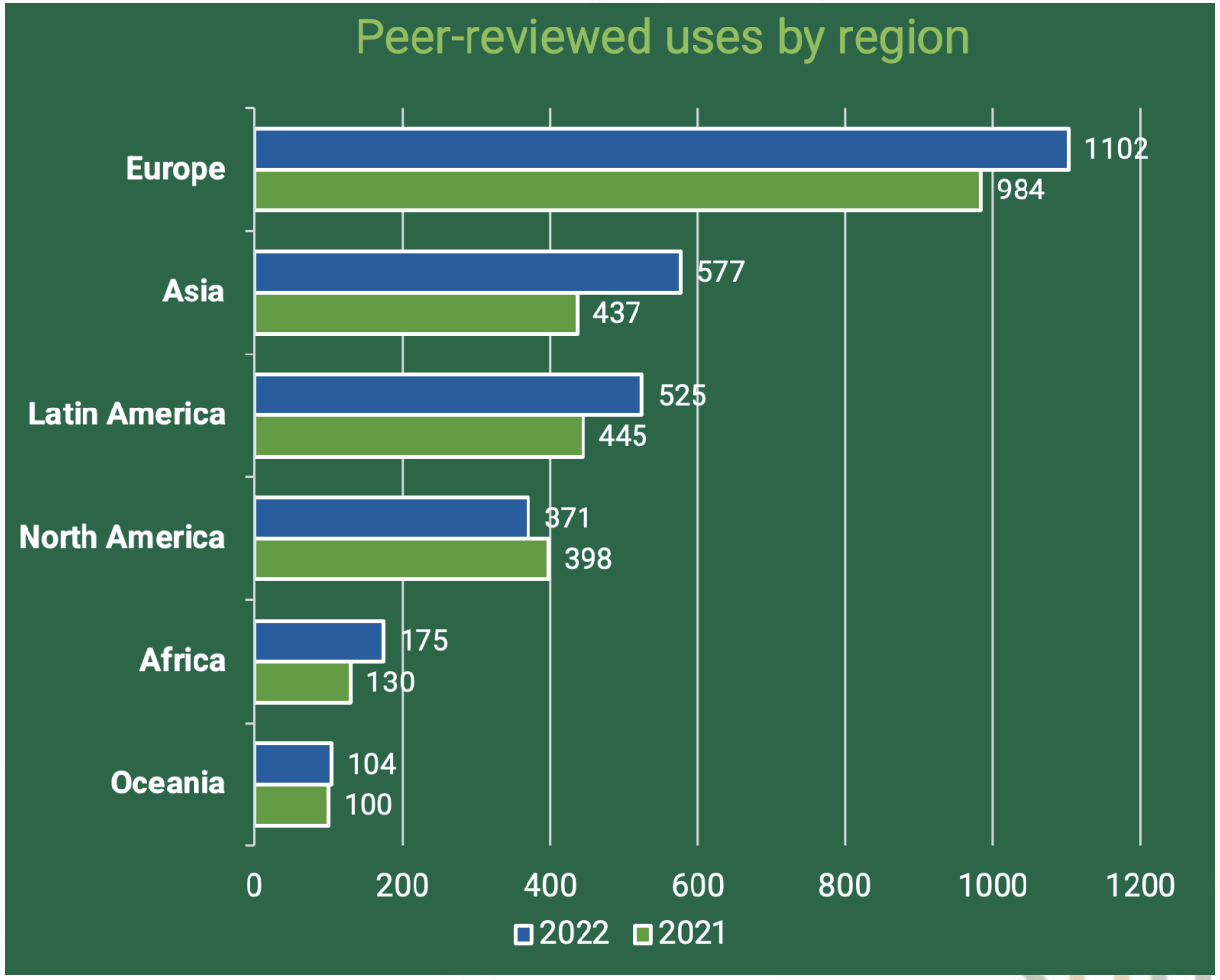


https://www.gbif.org/resource/search?contentType=literature&literatureType=journal&relevance=GBIF_USED&peerReview=true

#CiteTheDOI

Peer-reviewed uses by country

End of Year		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



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

How to cite Data mediated by GBIF?

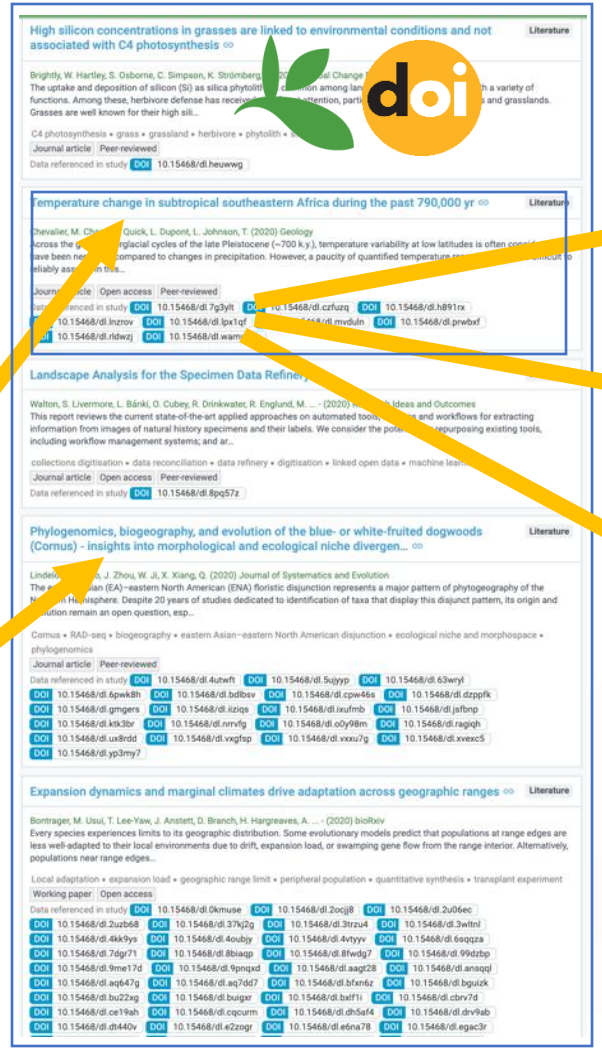
1. **Download data** from GBIF.org
2. and receive recommended citation with a **download DOI**
3. **Cite the DOI** in published research or other work

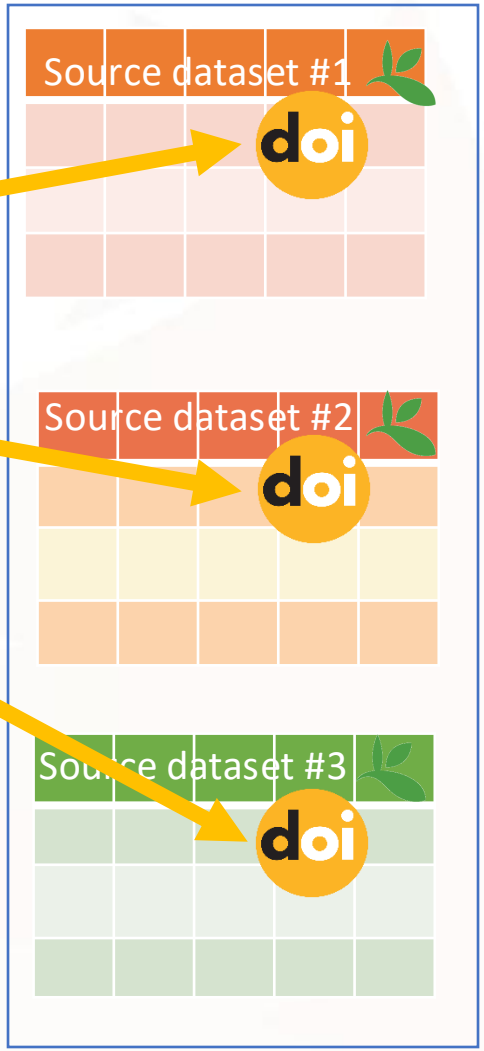
- *Example*: GBIF.org (9 November 2021) GBIF Occurrence Download **https//doi.org/10.15468/dl.xxxxxx**

#CiteTheDOI

https://www.gbif.org/citation-guidelines

DOI BASED DATA CITATION AT GBIF.org -- #CiteTheDOI





#CiteTheDOI

BIODT
biodiversitydigitaltwin

Dataset DOIs

Download DOI

Bibliographic DOI

Publish datasets in GBIF

Source dataset #1 doi

Source dataset #2 doi

Source dataset #3 doi

Filter & download

Add data-DOI to the paper reference list

GBIF download doi

Process & archive

Final state of data

zenodo doi

Analyze & publish

BIODT
biodiversitydigitaltwin

Dataset DOIs

Download DOI

Bibliographic DOI

Publish datasets in GBIF

Source dataset #1 doi

Source dataset #2 doi

Source dataset #3 doi

Filter & download

GBIF download doi

Process & archive

Final state of data

Analyze & publish

Data citation from the published paper

doi

doi

doi

doi

doi

S&H

BIODT
biodiversitydigitaltwin

Dataset DOIs

Download DOI

Bibliographic DOI

Possible with persistent object identifiers

Publish datasets in GBIF

Source dataset #1 doi

UUID

Source dataset #2 doi

UUID

Source dataset #3 doi

UUID

Filter & download

GBIF download doi

UUID

UUID

UUID

UUID

Process & archive

Final state of data

UUID

UUID

UUID doi

Analyze & publish

BIODT
biodiversitydigitaltwin

Dataset DOIs

Download DOI

Bibliographic DOI

*Possible with persistent **external** identifiers*

Source dataset #1

Source dataset #2

Source dataset #3

Publish dataset in GBIF

Filter & download

GBIF download

Analyze & publish

Process & archive

Final state of data

BIODT
biodiversitydigitaltwin
School & Hackaton

BioDT Data Citation Webinars

BIODT
biodiversitydigitaltwin

WEBINAR SERIES

6 November 2023
15:00 CET

Mastering Data Citation:
Insights from GBIF

GBIF

REGISTER NOW

Agenda

- A general introduction of the BioDT project – *Hanna Koivula (CSC)* *Slides*
- Data citation: the GBIF way – *Daniel Noesgaard (GBIF)* *Slides*

BIODT
biodiversitydigitaltwin

WEBINAR SERIES

20 November 2023
15:00 CET

Mastering Data Citation:
Insights from the BioDT Research Infrastructures

DISSCO Distributed System of Scientific Collections
GBIF
LifeWatch ERIC
eLTER

REGISTER NOW

Agenda

- welcome and main remarks from the GBIF data citation webinar - *Rita Giuffrida, Trust-IT*
- the data citation way in DISSCO - *Sharif Islam, DISSCO Slides*
- the data citation way in eLter - *Christoph Wohner, eLter Slides*
- the data citation way in LifeWatch - *Joaquin Lopez Lerida, LifeWatch Slides*

BIODT
biodiversitydigitaltwin

exercises

S&H

Excercises for the Open Science session

- Identify 5 data journals and describe why you would choose one journal over the other for your data paper.
- How is a data paper different from a normal research paper?

- Review GBIF documentation and data use course at:
  - https://techdocs.gbif.org/en/openapi/
  - https://techdocs.gbif.org/en/tools/#derived-datasets
  - https://techdocs.gbif.org/en/data-use/
  - https://docs.gbif.org/course-data-use/en/

THANK YOU

GBIF Norway
Dag Endresen | Node Manager
helpdesk@gbif.no

BIODT

biodiversitydigitaltwin

School & Hackaton

@BiodiversityDT

BioDT

Funded by
the European Union