



BIODT
biodiversitydigitaltwin

School & Hackaton

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

Data Citation using the GBIF Derived Dataset Tool

#CiteTheDOI

Dag Endresen – GBIF Norway, University of Oslo



Funded by
the European Union

Dag Endresen
University of Oslo
GBIF Norway

March 2025



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

- 🔥 GBIF Derived Datasets
- 🔥 Creating Derived Datasets – using the GBIF Tools
- 🔥 Creating Derived Datasets – using the GBIF data API & using R



BIODT

biodiversitydigitaltwin

School & Hackaton

Derived dataset



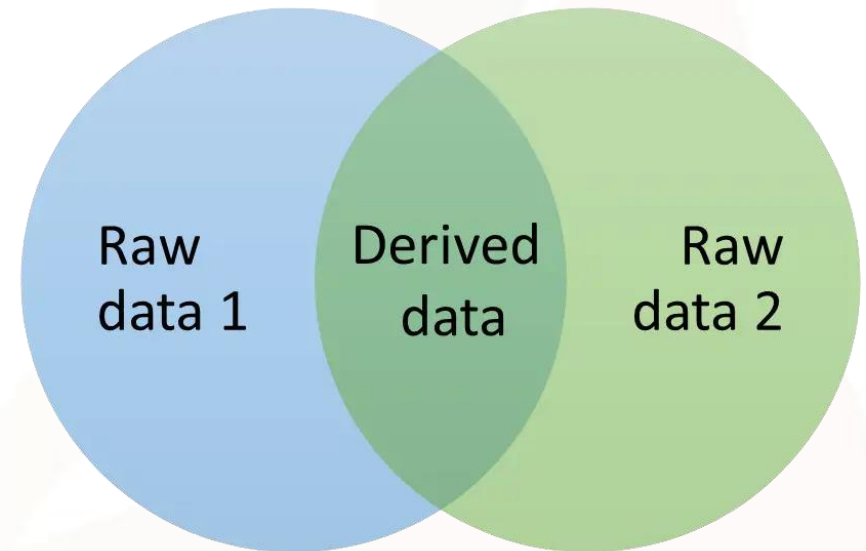
GBIF

Derived Dataset

Create your own citable dataset DOI



- Derived datasets are citable records of GBIF-mediated occurrence data derived either from:
 - 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**)
- When created, a derived dataset is assigned a unique DOI that can be used to cite the data.





BIODT

biodiversitydigitaltwin

School & Hackaton

Derived Dataset – using the GBIF portal Tool

- 🔥 To create a derived dataset, authenticate yourself 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,
- 🔥 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.*

TOOLS

Derived dataset

MY DATASETS REGISTER ABOUT

REGISTER NEW DERIVED DATASET

Title *

The title of your dataset, e.g. "Data used in study abc available at <https://doi.org/xyz>"

Source url *

<https://doi.org/10.5281/zenodo.1234567>

Or enter datasets and occurrence counts here

Dataset key	ad43e954-dd79-4986-a1	Occurrence count	545	
Dataset key	10.15468/inyc6	Occurrence count	253	
Dataset key	03376f38-8456-42c2-91	Occurrence count	74	+

B I H “ ≡ ≡ 🔗 🖼️ 👁️ 📄 ✕ ?

Describe your dataset, how it was obtained, filtered, etc.

DERIVED DATASET | 5 MAY 2021

Records from 3 datasets in derived dataset

DOI 10.21373/dd.enuxdj

[ACCESS DERIVED DATASET](#)

Title: The title of your dataset, e.g. "Data used in study abc available at <https://doi.org/xyz>"

Description: Describe your dataset, how it was obtained, filtered, etc.

Citation: Derived dataset GBIF.org (5 May 2021) Filtered export of GBIF occurrence data <https://doi.org/10.21373/dd.enuxdj>

Involved datasets: 3

Make sure to read the [data user agreement](#) and [citation guidelines](#).

INCLUDES RECORDS FROM 3 DATASETS

Meise Botanic Garden Herbarium (BR)

74

International Barcode of Life project (IBOL)

253

Geographically tagged INSDC sequences

545

- 🔥 Example of a derived dataset endpoint
- 🔥 viewed from the GBIF.org portal

- ✦ For this training session we will use the GBIF UAT Demo Training Server
- ✦ Which is a mirror of the production server.
- ✦ <https://www.gbif-uat.org/derived-dataset/about>
- ✦ Login using your GBIF user account (which you created yesterday).



Image credit: <https://www.vecteezy.com/free-vector/sand-box> Sand Box Vectors by Vecteezy



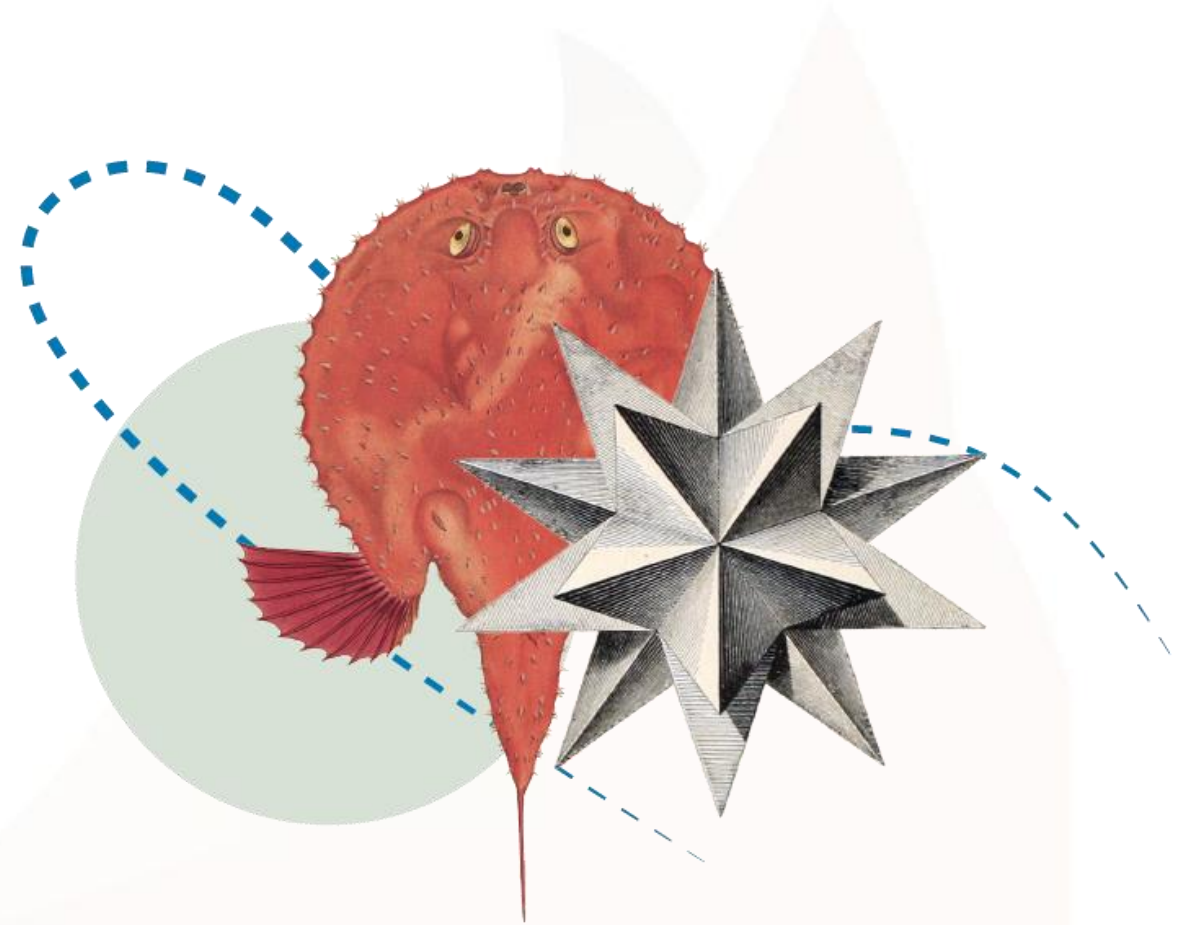
BIODT

biodiversitydigitaltwin

School & Hackaton

Derived dataset – using the GBIF API & R

- 🔥 You can also use the GBIF API or the rgbif R package to create derived datasets.
- 🔥 <https://techdocs.gbif.org/en/openapi/v1/registry-principal-methods#/Derived%20datasets>
- 🔥 https://docs.ropensci.org/rgbif/reference/derived_dataset.html



Request body

```
{
  "originalDownloadDOI": "string",
  "title": "string",
  "description": "string",
  "sourceUrl": "string",
  "registrationDate": "2025-02-25T06:18:55.963Z",
  "relatedDatasets": {
    "additionalProp1": 0,
    "additionalProp2": 0,
    "additionalProp3": 0
  }
}
```

The relatedDatasets map may use GBIF Dataset DOIs or UUIDs as keys, with the values being the number of records to cite from each dataset.

Response

```
{
  "doi": "string",
  "originalDownloadDOI": "string",
  "description": "string",
  "citation": "string",
  "title": "string",
  "sourceUrl": "string",
  "createdBy": "string",
  "modifiedBy": "string",
  "registrationDate": "2025-02-25T06:18:55.964Z",
  "created": "2025-02-25T06:18:55.964Z",
  "modified": "2025-02-25T06:18:55.964Z"
}
```

```
derived_dataset_prep (  
  citation_data = NULL,  
  title = NULL,  
  description = NULL,  
  source_url = NULL,  
  gbif_download_doi = NULL,  
  user = NULL,  
  pwd = NULL,  
  curlopts = list\(\)  
)
```



exercises

- 🔥 Create a Derived Dataset using the GBIF Tool at
 - 🔥 <https://www.gbif-uat.org/derived-dataset>
- 🔥 Use the demo dataset at the training portal (or use your own data)
 - 🔥 <https://training.lifewatch.eu:9001/mod/resource/view.php?id=1820>
 - 🔥 You can use Source URL =
 - 🔥 <https://api.gbif-uat.org/v1/occurrence/download/request/0000110-250227151720179.zip>
- 🔥 Work in groups (same groups as in the morning sessions).
- 🔥 Present your results in plenary, max 5 minutes each group.
- 🔥 Review GBIF documentation and data use course at:
 - 🔥 <https://techdocs.gbif.org/en/tools/#derived-datasets>
 - 🔥 <https://www.gbif.org/derived-dataset> (for actual real-life use)
 - 🔥 <https://www.gbif-uat.org/derived-dataset> (demo, training)

- Using Darwin Core for your dataset
- <https://training.gbif.org/en/data-use/exercise1>
- Programmatic access to GBIF-mediated data
- <https://training.gbif.org/en/data-use/programmatic-access>
- Using rgbif to download, visualize and clean GBIF-mediated data
- <https://training.gbif.org/en/data-use/exercise5>

THANK YOU

GBIF Norway

Dag Endresen | Node Manager

helpdesk@gbif.no



www.gbif.org





BIODT

biodiversitydigitaltwin

School & Hackaton



@BiodiversityDT



BioDT



Funded by
the European Union