

Funded by
the European Union



DTO-BioFlow

Integration of biodiversity monitoring
data into the Digital Twin Ocean

DTO-BioFlow data training
workshop:

Automated data publication

Levels of automation

- ≡ Level 0: manual
- ≡ Level 1: automated data transformation + manual upload of individual data files to IPT
- ≡ Level 2: automated data transformation and data packaging + manual upload of DwC-A to IPT
- ≡ Level 3: automated data transformation, data packaging and data publication

manual

semi-automated

automated



Levels of automation

- ≡ ~~Level 0: manual~~
- ≡ Level 1: automated data transformation + manual upload of individual data files to IPT
- ≡ Level 2: automated data transformation and data packaging + manual upload of DwC-A to IPT
- ≡ Level 3: automated data transformation, data packaging and data publication

manual

semi-automated

automated



Levels of automation

- ≡ **Level 0: manual**
- ≡ **Level 1: automated data transformation + manual upload of individual data files to IPT**
- ≡ **Level 2: automated data transformation and data packaging + manual upload of DwC-A to IPT**
- ≡ **Level 3: automated data transformation, data packaging and data publication**

manual

semi-automated

automated



Creating DwC-Archives without IPT

≡ Creating the data package

≡ Multiple libraries available

≡ [LivingNorwayR](#)

≡ [DwCA-Writer](#)

≡ [PyDwCA](#)

Creating DwC-Archives without IPT

≡ Creating the data package

≡ Multiple libraries available

≡ Compressing data files+ metadata files

≡ Metadata files schema available ([Documentation](#)):

≡ [Eml.xml](#) -> [Documentation](#)

≡ [Meta.xml](#) -> [Documentation](#)

≡ Libraries to manage xml file

≡ [EML](#)

≡ [Python library for XML](#)

≡ [R library for XML](#)

≡ Example of dataset with metadata files:

≡ [Download the dataset](#)

≡ Extract the .xml files from the DwC-A (.zip file)

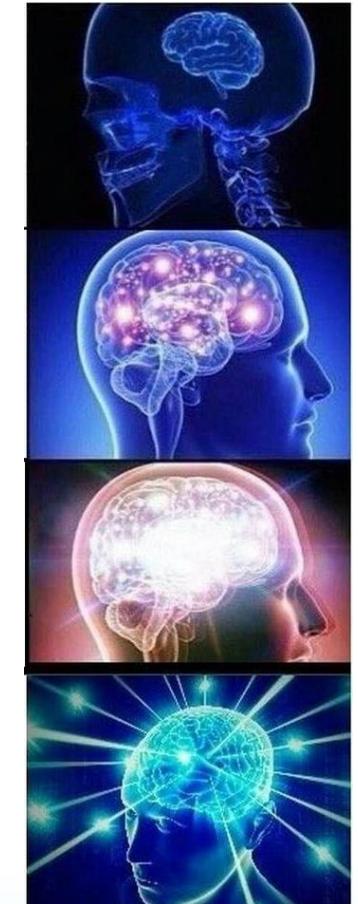
Levels of automation

- ≡ **Level 0: manual**
- ≡ **Level 1: automated data transformation + manual upload of individual data files to IPT**
- ≡ **Level 2: automated data transformation and data packaging + manual upload of DwC-A to IPT**
- ≡ **Level 3: automated data transformation, data packaging and data publication**

manual

semi-automated

automated





Automated data publication options

≡ Connect SQL database to IPT

<https://ipt.gbif.org/manual/en/ipt/latest/manage-resources#source-data>

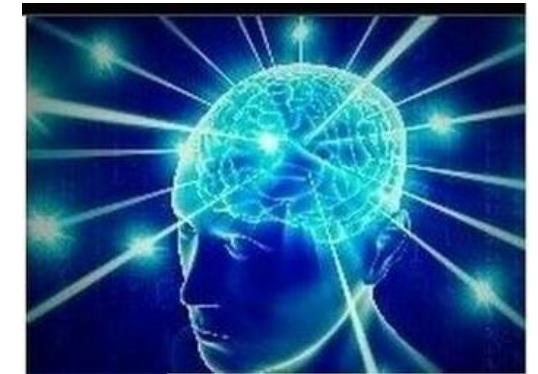
≡ Host DwC-A online

≡ connect to IPT

<https://ipt.gbif.org/manual/en/ipt/latest/manage-resources#source-data>

≡ GBIF Registry API

<https://techdocs.gbif.org/en/data-publishing/register-dataset-api>



Use SQL database

(i) Source Data

Not modified since last publication

+ Add



Source Data

File

URL

SQL

Source Name

Connect



Use SQL database

Manage / Overview / Source

mysql_specimen

Aves Tanzanian collection at the Natural History Museum of Denmark (SNM)

Save Options Cancel

Source type: SQL
Readable: No

Database System: MySQL Host: localhost
Database: specimens_db Database user: jsmith
Database password:

SQL Statement:

```
SELECT * FROM specimens JOIN taxon ON taxon_fl = taxon.id
```

Generated SQL for previewing data

Character Encoding: UTF-8 Date Format: YYYY-MM-DD
Multi-value Delimiter: |

The screenshot displays the DTO-BioFlow application's configuration interface for a MySQL database source. At the top, there are navigation links: 'Manage', 'Overview', and 'Source'. Below this, the source name 'mysql_specimen' is shown, along with a subtitle 'Aves Tanzanian collection at the Natural History Museum of Denmark (SNM)'. There are three buttons at the top right: 'Save', 'Options', and 'Cancel'. The main configuration area starts with 'Source type: SQL' and 'Readable: No'. It then moves to 'Database System' settings, where 'MySQL' is selected as the system and 'localhost' is entered as the host. The 'Database' field contains 'specimens_db' and the 'Database user' field contains 'jsmith'. A 'Database password' field is present but empty. Below these fields is a 'SQL Statement' section containing the query 'SELECT * FROM specimens JOIN taxon ON taxon_fl = taxon.id'. Underneath this, there is a section for generating SQL for previewing data, which includes 'Character Encoding' set to 'UTF-8', 'Date Format' set to 'YYYY-MM-DD', and a 'Multi-value Delimiter' set to '|'. The bottom of the interface features a decorative blue wavy pattern.

Use SQL database

Auto-publishing

 Edit Disabled

Auto-publish is inactive. Your resource may be published manually with the Publish button.



Use SQL database

Supported default databases

The IPT can use database connections to import data from tables or views. Currently the following databases are supported out of the box:

- Microsoft SQL Server
 - MySQL
 - Oracle
 - PostgreSQL
 - Sybase
-
- [Possible to add new JDBC drivers](#)



Use URL

== Two possible flows:

== IPT

== GBIF Registry API

Use URL - IPT

(i) Source Data

Not modified since last publication

+ Add



Source Data

File

URL

SQL

Source Name

URL

Add

Clear



Use URL - IPT

Manage / Overview / Source

occurrence

Aves Tanzanian collection at the Natural History Museum of Denmark (SNM)

Save Options ▾ Cancel

Source type	URL
Readable	<input checked="" type="radio"/> Yes
URL	http://example.org/data/aves/occurrence.txt
Columns	11
Rows	423
Size	84.1 KB
Modified	5 February 2024, 09:25:55

Number of Header Rows: 1

Field Delimiter: ,

Field Quotes: "

Multi-value Delimiter:

Character Encoding: UTF-8

Date Format: YYYY-MM-DD



Use URL - IPT

Auto-publishing

 Edit Disabled

Auto-publish is inactive. Your resource may be published manually with the Publish button.

Use URL - GBIF Registry API

- Webserver, VM, cloud hosting service,...
- Must be accessible to GBIF's servers
- Need a user account on GBIF.org (institution account)
- Contact helpdesk@gbif.org to ask for editor_rights permissions
- Create an account on GBIF-UAT.org for testing
- Testing without account:
 - Test on GBIF-UAT.org
 - Username=ws_client_demo
 - Password=Demo123

Use URL - GBIF Registry API

- Two REST calls (<https://techdocs.gbif.org/en/data-publishing/register-dataset-api>):

- Create the dataset metadata

Mandatory dataset metadata

```
{  
    "publishingOrganizationKey": "0a16da09-7719-40de-8d4f-56a15ed52fb6", ①  
    "installationKey": "92d76df5-3de1-4c89-be03-7a17abad962a", ①  
    "type": "METADATA", ②  
    "title": "Example dataset registration",  
    "description": "The dataset is registered with minimal metadata, which is often sufficient for basic use cases.",  
    "language": "eng",  
    "license": "http://creativecommons.org/publicdomain/zero/1.0/legalcode" ③  
}
```

POST this JSON to GBIF using the Registry API:

```
curl -Ssf --user ws_client_demo:Demo123 -H "Content-Type: application/json"  
  
dataset=$(cat dataset.registration)
```

Use URL - GBIF Registry API

- Two REST calls (<https://techdocs.gbif.org/en/data-publishing/register-dataset-api>):
 1. Create the dataset metadata
 2. Add the endpoints for the data and metadata files

Endpoint definition

```
{  
  "type": "EML", ①  
  "url": "https://techdocs.gbif.org/en/data-publishing/_attachments/test-data-set-1.eml"  
}
```

- type=='DWC_ARCHIVE'

Add this endpoint to the dataset:

```
curl -Ssf --user ws_client_demo:Demo123 -H "Content-Type: application/json"
```

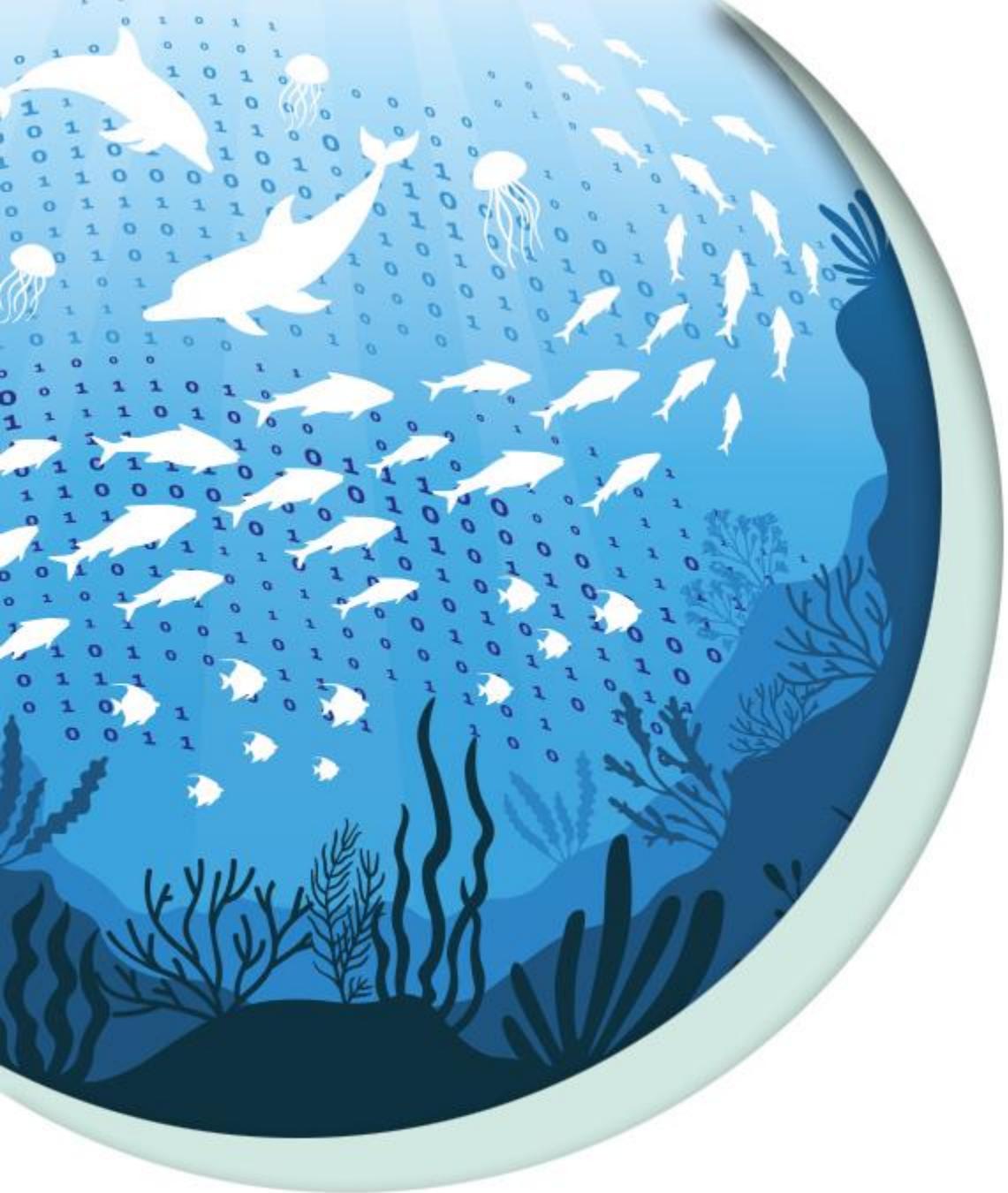


Use URL - GBIF Registry API

- 1-2 minutes for the dataset metadata
- 1-60 minutes for data depending on the size
- Follow the progress :
 - "[Running crawls](#)"
 - "[Running ingestions](#)"
- Use a script to automate the flow (shell, python, R,...)

Relevant sources

- ≡ [GBIF Integrated Publishing Toolkit \(IPT\) User Manual - Source data](#)
- ≡ [GBIF technical documentation - Register dataset with API](#)
- ≡ [GBIF Integrated Publishing Toolkit \(IPT\) User Manual - Supported databases](#)
- ≡ [GBIF Integrated Publishing Toolkit \(IPT\) User Manual – Publishing DwC-A manually](#)



DTO-BioFlow

Integration of biodiversity monitoring
data into the Digital Twin Ocean

THANKS!