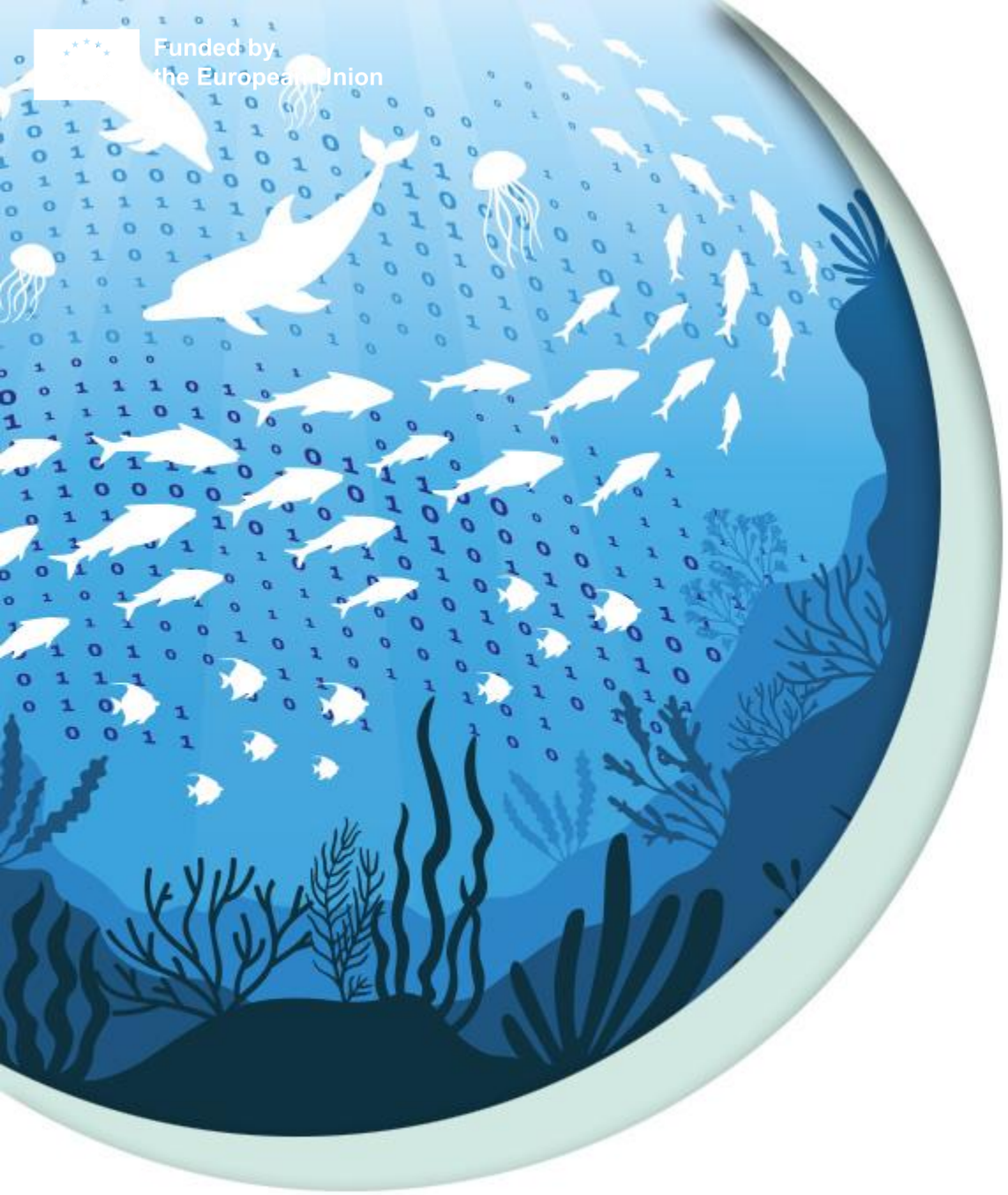


Funded by the European Union

DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

DTO-BioFlow data training workshop:

Automated data publication

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Creating DwC-Archives without IPT

- Creating the data package
  - Multiple libraries available
    - LivingNorwayR
    - DwCA-Writer
    - PyDwCA

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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)

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Automated data publication options

- Connect SQL database to IPT

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

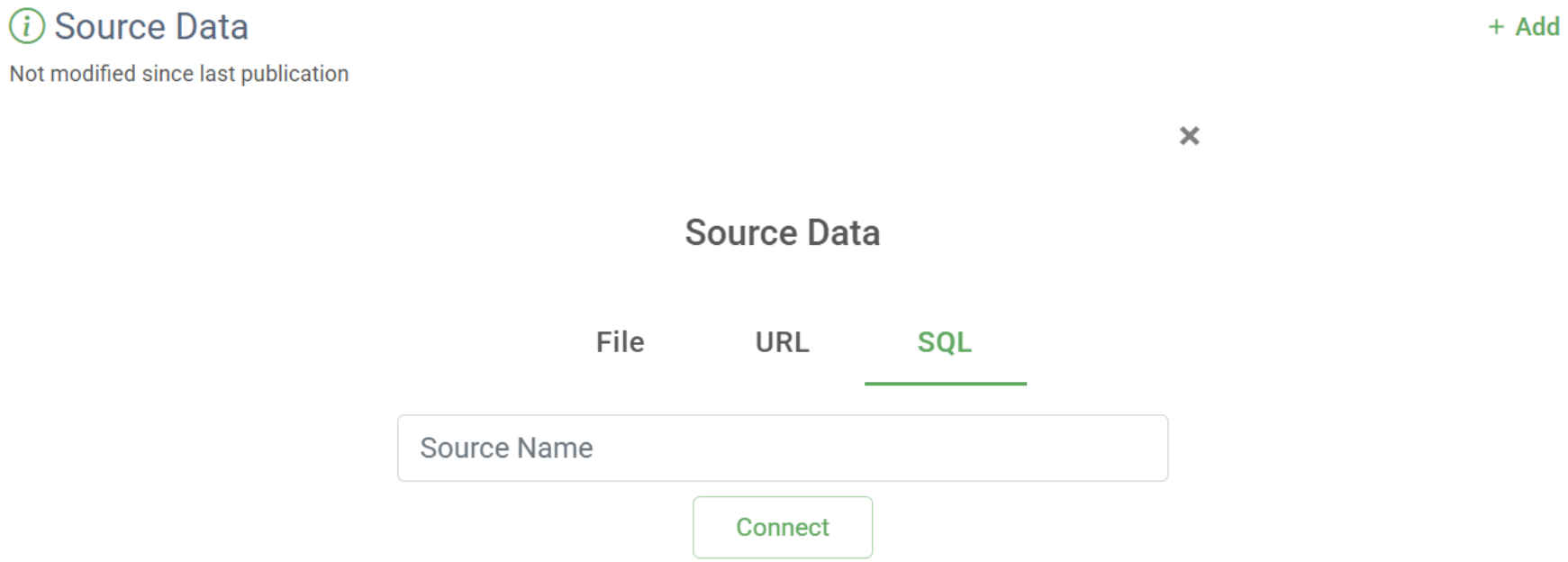
- Host DwC-A online
  - connect to IPT

    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

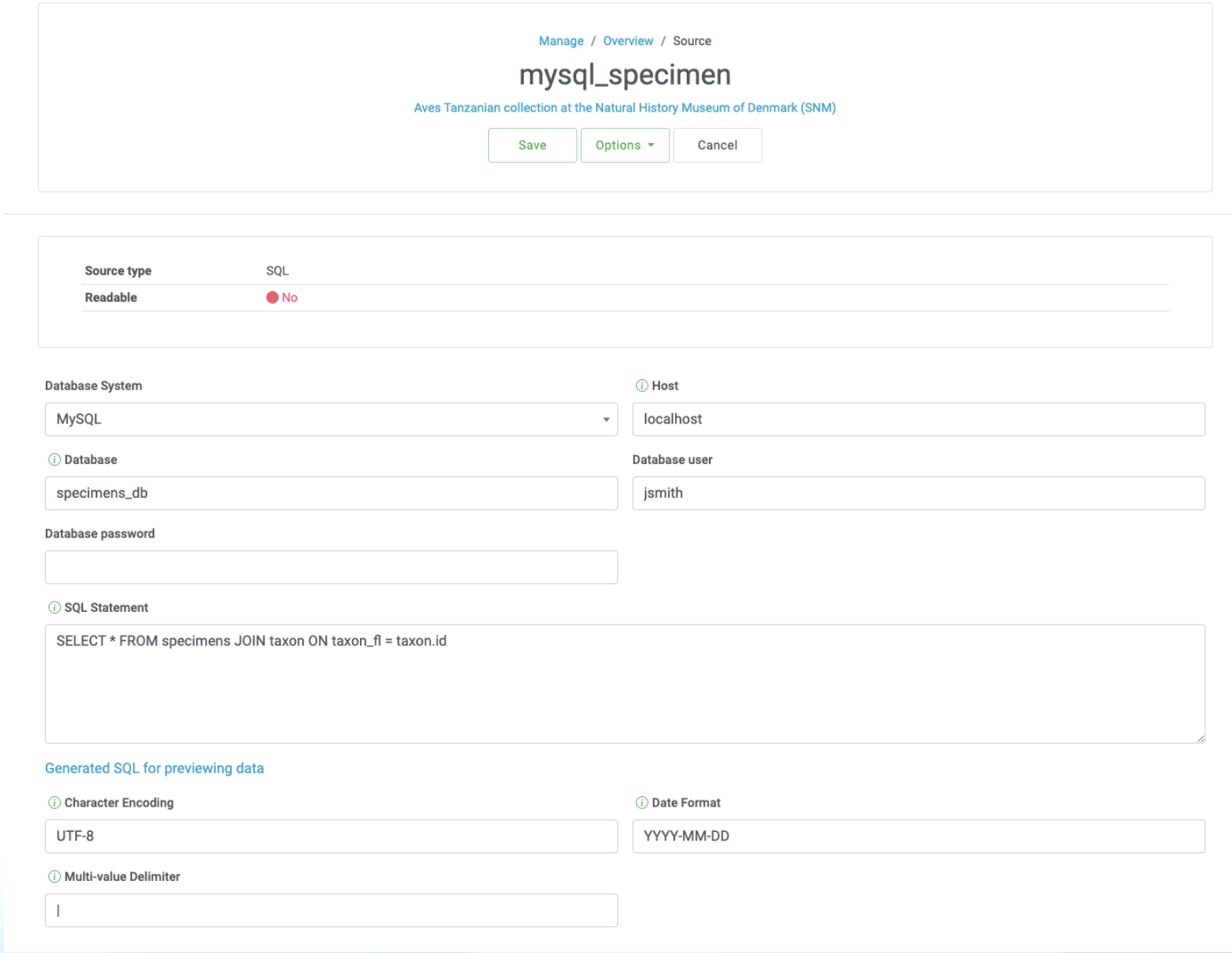
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Use SQL database

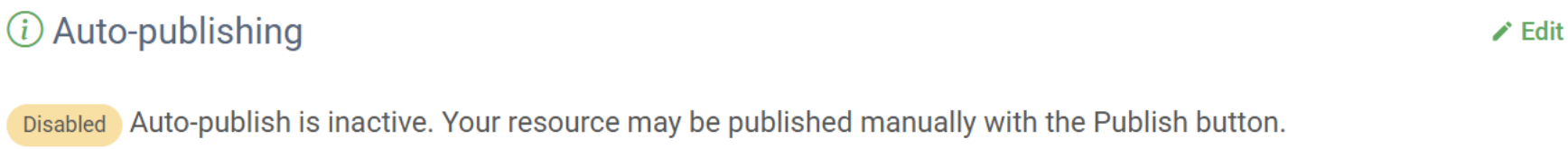


DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Use SQL database



Use SQL database



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:

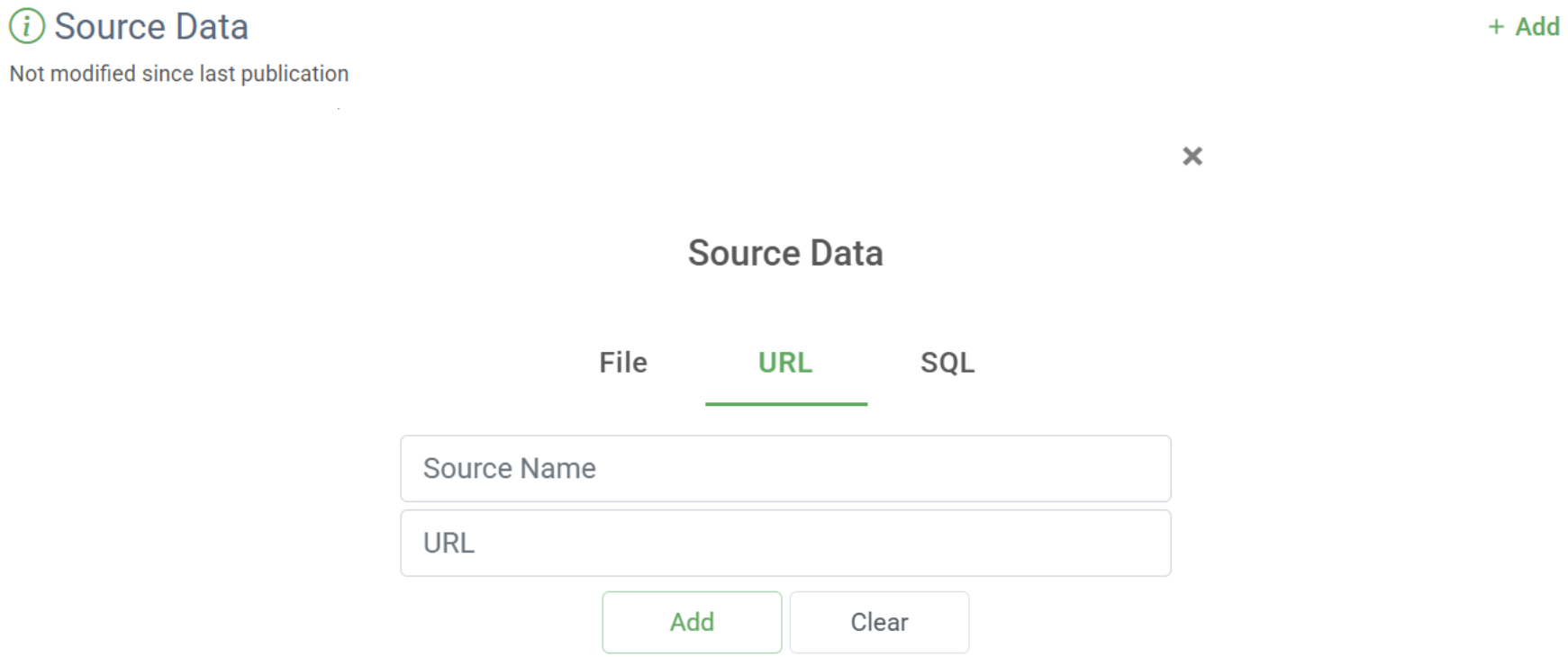
- Microsoft SQL Server
- MySQL
- Oracle
- PostgreSQL
- Sybase

- Possible to add new JDBC drivers

Use URL

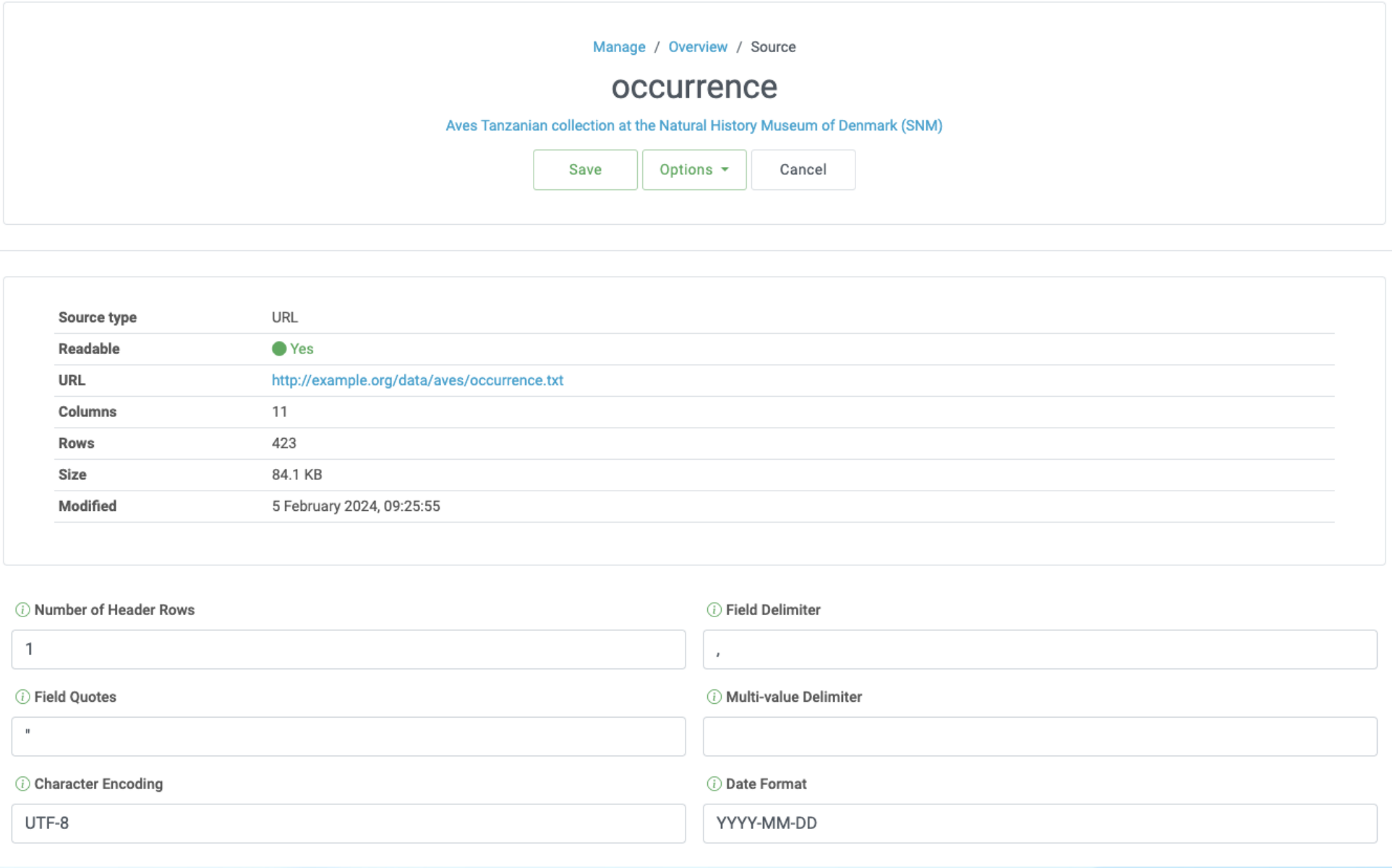
- Two possible flows:
  - IPT
  - GBIF Registry API

Use URL - IPT

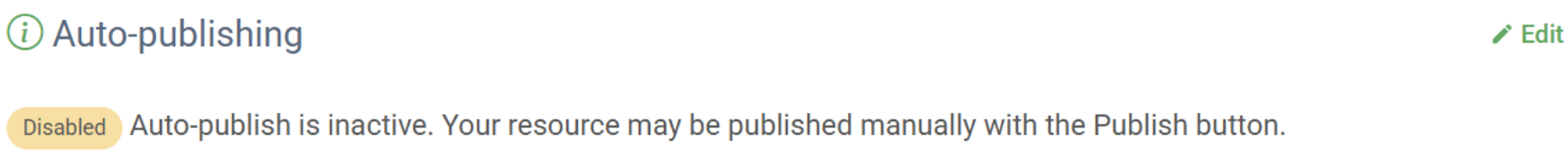

Source Data
Not modified since last publication
+ Add
Source Data
File
URL
SQL
Source Name
URL
Add
Clear

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Use URL - IPT



Use URL - IPT


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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Use URL - GBIF Registry API

- Two REST calls (https://techdocs.gbif.org/en/data-publishing/register-dataset-api):
  1. 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
  "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-dat
}
```

- type==‘DWC_ARCHIVE’

Add this endpoint to the dataset:

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

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Use URL - GBIF Registry API

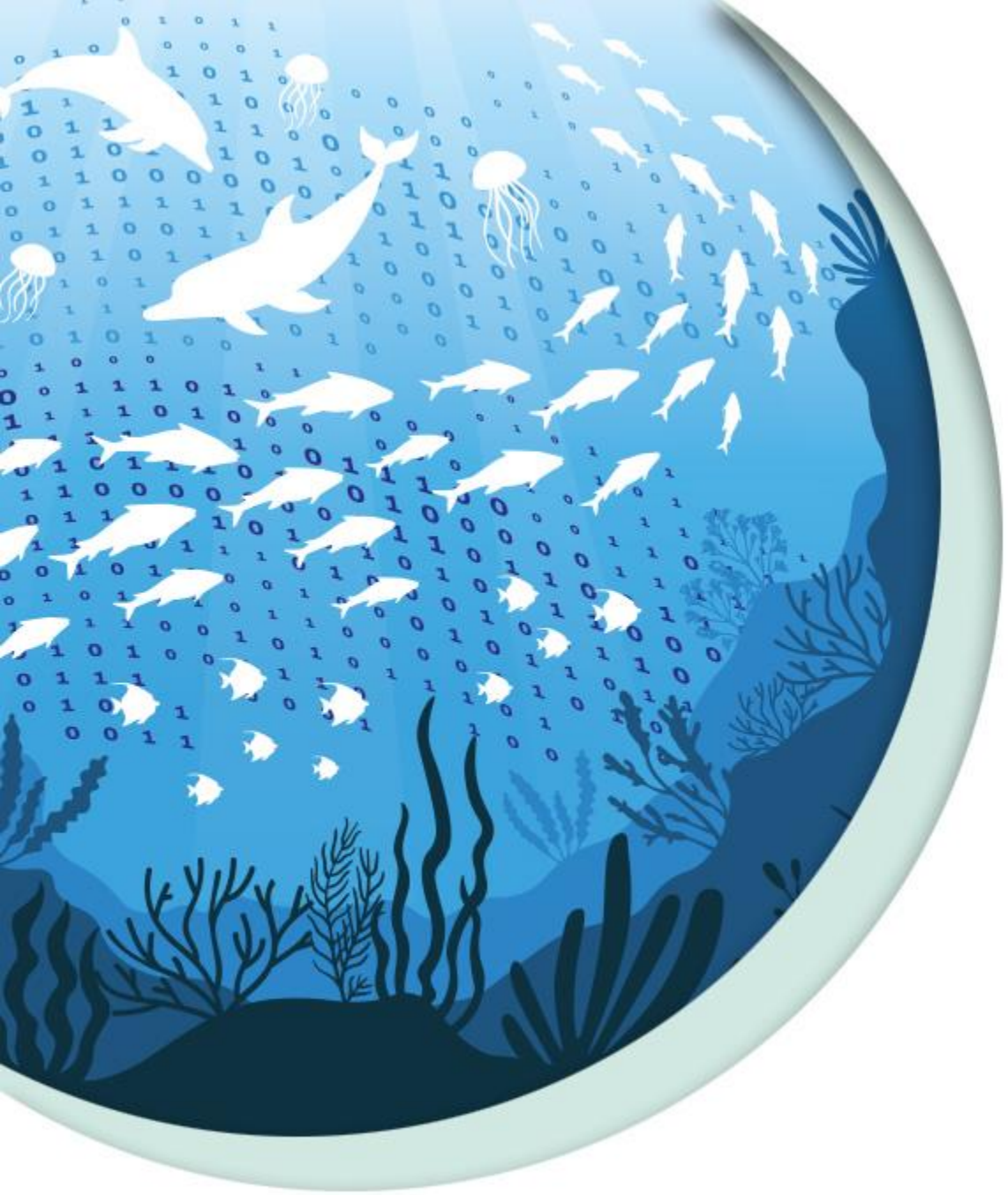
- 1-2 minutes for the dataset metadata
- 1-60 minutes for data depending on the size
- Follow the progress :
  - "Running crawls“
  - "Running ingestions“

- Use a script to automate the flow (shell, python, R,…)

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

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!