BIODIVERSITY
BUILDING
BLOCKS FOR
POLICY

M6 First user generated species occurrence cube is available and citable

31/08/2024

Author(s): Matthew Blissett, Tim Robertson, Peter Desmet

Funded by
the European Union

This project receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

Table of contents

Summary

GBIF has developed software (D2.2) to generate occurrence cubes implementing the majority of the specification for occurrence cubes (D2.1). The code is available as open source software in GitHub (https://github.com/gbif/occurrence-cube), providing both version control and issue tracking. This software has additionally been deployed by GBIF as a web service allowing for technical users to generate cube programmatically without the need to manage infrastructure. Cubes generated by the service are provided with a Digital Object Identifier (DOI) allowing for easy citation practices. This milestone describes the status of the software and the deployed service that has resulted in the first public occurrence cube being created (https://doi.org/10.15468/dl.b8ydjq).

List of abbreviations

DOI	Digital Object Identifier
EU	European Union
EEA	European Environment Agency
GBIF	Global Biodiversity Information Facility
HDFS	Hadoop Distributed File System
SQL	Structured Query Language
UDF	User Defined Function

1. Deployment of the occurrence cube software for early adopters

GBIF has developed occurrence cube software (D2.2) that implements the specification for occurrence cubes and their production (D2.1), have tested the software within the project using a public cloud computing environment (M5) and have now integrated the software within an online service providing cube generation.

In alignment with the specification, SQL was selected as the query language to generate cubes. The software provides user defined functions (UDF), which users can include within the SQL statements to simplify some of the complex aggregation queries. Functions are currently available to support three spatial aggregations:

1. The EEA Reference Grid (EEARG) [1]
2. The Extended Quarter Degree Grid Cell (EQDGC) [2]
3. Inverse Snyder Equal-Area Projection (ISEA) Aperture 3 Hexagonal (3H) Discrete Global Grid System (DGGS), ISEA3H [3]

The software has been included within the GBIF download API which is currently available as an experimental feature and documented for early adopters [4].

2. Citing cubes

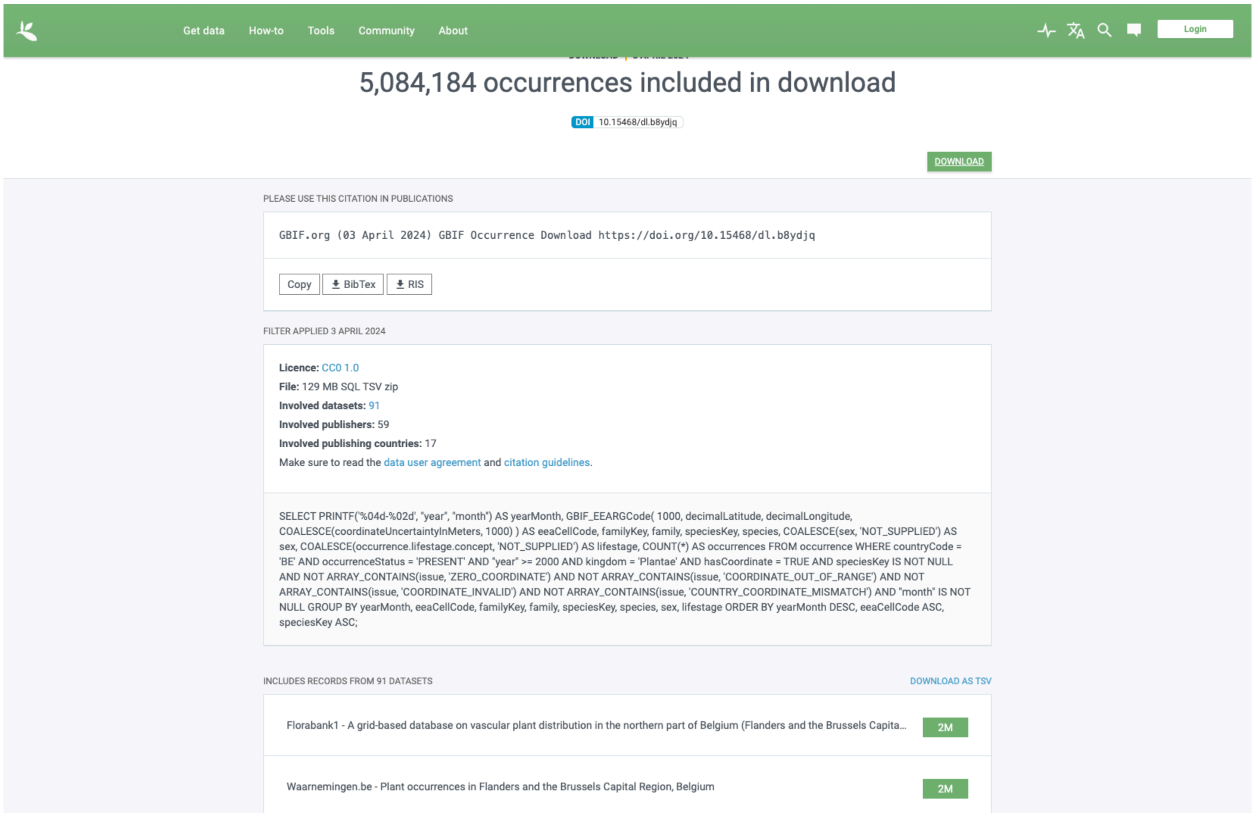
Building on the extensive citation experience within GBIF [5], DOIs are chosen as the mechanism to simplify data citation. Just like other GBIF data downloads, all cubes generated within the GBIF service are issued a DOI through DataCite and come with clear citation guidelines allowing the cube to be cited using the DOI. Additionally, within the cube DOI metadata, a link is created to attribute all the contributing datasets to provide the provenance, and to help raise the visibility of any cube use to the originators of the data. Adopting the mechanism means that any cited use of the cubes themselves will be picked up by the GBIF citation tracking mechanism.

3. First user-generated cube

Since its launch as an experimental service, early adopters have generated cubes for their research and testing purposes. The first cube was generated on 3rd April 2024, aggregating data across 5 million occurrence records from 91 datasets (see Figure 1). It can be cited using its DOI (https://doi.org/10.15468/dl.b8ydjq). At the time of writing, 178 cubes have been generated by 23 users since.

This constitutes an important step towards deliverable D2.3: a publicly deployed and documented service, scheduled for February 2023. In addition to programmatic access, the service will be accessible through a graphical user interface at gbif.org that will further simplify cube generation.

First user generated species occurrence cube is available and citable



5,084,184 occurrences included in download

DOI 10.15468/dl.b8ydjq

PLEASE USE THIS CITATION IN PUBLICATIONS

```
GBIF.org (03 April 2024) GBIF Occurrence Download https://doi.org/10.15468/dl.b8ydjq
```

FILTER APPLIED 3 APRIL 2024

Licence: CC0 1.0
File: 129 MB SQL TSV zip
Involved datasets: 91
Involved publishers: 59
Involved publishing countries: 17
Make sure to read the data user agreement and citation guidelines.

```
SELECT PRINTF('%04d-%02d', "year", "month") AS yearMonth, GBIF_EEARGCode( 1000, decimalLatitude, decimalLongitude,
COALESCE(coordinateUncertaintyInMeters, 1000) ) AS eeaCellCode, familyKey, family, speciesKey, species, COALESCE(sex, 'NOT_SUPPLIED') AS
sex, COALESCE(occurrence.lifestage.concept, 'NOT_SUPPLIED') AS lifestage, COUNT(*) AS occurrences FROM occurrence WHERE countryCode =
'BE' AND occurrenceStatus = 'PRESENT' AND "year" >= 2000 AND kingdom = 'Plantae' AND hasCoordinate = TRUE AND speciesKey IS NOT NULL
AND NOT ARRAY_CONTAINS(issue, 'ZERO_COORDINATE') AND NOT ARRAY_CONTAINS(issue, 'COORDINATE_OUT_OF_RANGE') AND NOT
ARRAY_CONTAINS(issue, 'COORDINATE_INVALID') AND NOT ARRAY_CONTAINS(issue, 'COUNTRY_COORDINATE_MISMATCH') AND "month" IS NOT
NULL GROUP BY yearMonth, eeaCellCode, familyKey, family, speciesKey, species, sex, lifestage ORDER BY yearMonth DESC, eeaCellCode ASC,
speciesKey ASC;
```

INCLUDES RECORDS FROM 91 DATASETS — DOWNLOAD AS TSV

Florabank1 - A grid-based database on vascular plant distribution in the northern part of Belgium (Flanders and the Brussels Capita... 2M

Waarnemingen.be - Plant occurrences in Flanders and the Brussels Capital Region, Belgium 2M

Figure 1: Screenshot of a generated cube, illustrating the citation guideline and the DOI, the use of SQL for generation, and the first rows documenting the contributing datasets.

4. References

1. https://techdocs.gbif.org/en/data-use/api-sql-download-functions#eea-reference-grid-cell-code-gbif_eeargcode
2. https://techdocs.gbif.org/en/data-use/api-sql-download-functions#extended-quarter-degree-grid-cell-code-gbif_eqdgccode
3. https://techdocs.gbif.org/en/data-use/api-sql-download-functions#isea3h-grid-cell-code-gbif_isea3hcode
4. https://techdocs.gbif.org/en/data-use/api-sql-downloads
5. https://techdocs.gbif.org/en/data-use/citation