BIODIVERSITY BUILDING BLOCKS FOR POLICY

D2.2 Occurrence cube implementation

20/02/2024

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

Funded by the European Union

Prepared under contract from the European Commission
Grant agreement No. 101059592
EU Horizon Europe Research and Innovation Action

Project acronym:	**B3**
Project full title:	**Biodiversity Building Blocks for policy**
Project duration:	01.03.2023 – 31.08.2026 (42 months)
Project coordinator:	Dr. Quentin Groom, Agentschap Plantentuin Meise (MeiseBG)
Call:	HORIZON-CL6-2021-GOVERNANCE-01
Deliverable title:	Occurrence cube implementation
Deliverable n°:	D2.2
WP responsible:	WP2
Nature of the deliverable:	DEM: Demonstrator, pilot, prototype
Dissemination level:	Public
Licence of use:	Creative Commons Attribution 4.0 International
Lead partner:	GBIFS
Recommended citation:	Blissett, M., Robertson T. & Desmet, P (2024). ***Occurrence cube implementation***. B3 project deliverable D2.2.
Due date of deliverable:	Month n°12
Actual submission date:	Month n°12

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final/Draft	20 February 2024	Matthew Blissett (GBIFS), Tim Robertson (GBIFS), Peter Desmet (EV INBO)

Table of contents

Key takeaway messages

- GBIF have developed and released a prototype tool (TRL 6) for producing species occurrence cubes.
- In alignment with the specification (D2.1), the tool provides user defined functions (UDF) which users can add to their SQL to accomplish complex aggregation queries using GBIF or external data.
- The tool can be run on cloud systems, including a test environment using Microsoft Azure (Databricks).
- Documentation on how to use the functions and test environment is provided as well.
- Next steps are to develop the tool further, incorporating all software requirements and user feedback, and releasing it as a GBIF service (D2.3).

Executive summary

GBIF have developed and released a first prototype tool for producing species occurrence cubes. The tool provides a number of user defined functions which users can add to their SQL to accomplish complex aggregation queries on GBIF or external data. The tool is based on the specifications defined in deliverable D2.1 and can be run on cloud systems. Documentation is provided for each function as well as how to use it in a cloud environment. The tool will now be developed further and released as a GBIF service (D2.3).

Non-technical summary

The Global Biodiversity Information Facility (GBIF) provides an increasing amount of occurrence data: data that documents when and where species have been observed. While these data are essential for policy and research, they present a challenge for users attempting large-scale downloads and processing. In response to these challenges, GBIF have now released and documented a first version of a tool aimed at facilitating this process. While this tool currently requires users to have certain technical skills, it serves as a prototype for an upcoming service that will be released later. That service will streamline the download and processing of occurrence data, ensuring better data harmonisation and organisation. Consequently, it will offer an easier user experience and improved functionality, catering to a wide range of users.

List of abbreviations

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

1. Occurrence cube software

The prototype occurrence cube software implements the specification for occurrence cubes and their production (D2.1, Desmet et al. 2023a). In alignment with the specification, SQL has been selected as the query language to generate cubes. The software provides user defined functions (UDF), which users add to their SQL to accomplish complex aggregation queries.

Functions

Three functions provide support for the three gridding schemes required by the specification:

- **eeaCellCode()** — allows random assignment of an occurrence within its coordinate uncertainty to a cell of the EEA Reference Grid (European Environment Agency 2013). This replicates the functionality developed for the TrIAS project, which is currently in use for cube generation. Those cubes can now be generated with the new software.
- **eqdgcCode()** — this allows the same random assignment of an occurrence, but uses the Extended Quarter Degree Grid (Larsen et al. 2009). This grid provides global coverage.
- **mgrsCode()** — this also allows the same random assignment of an occurrence, and uses the Military Grid Reference System (Veness 2020). This grid covers most of the world, excluding the extreme polar regions.

The three functions are documented at https://links.gbif.org/cube-functions.

Metadata

As this is prototype software, metadata are yet not provided in a standardized format and neither is a DOI assigned. The necessary information for the cube metadata can be generated using SQL however. An example query to produce a cube and its metadata is provided in Figure 1 and 2, and covered in more detail in the documentation at https://links.gbif.org/cube-setup.

```
Unset
SELECT
 -- Dimensions
 year,
 eeaCellCode(
   1000, decimalLatitude, decimalLongitude,
   COALESCE(coordinateUncertaintyInMeters, 1000)
 ) AS eeaCellCode,
 speciesKey,
 -- Measurements
 COUNT(*) AS n,
 MIN(
   COALESCE(coordinateUncertaintyInMeters, 1000)
 ) AS minCoordinateUncertaintyInMeters
```

```
FROM
 gbif.occurrence
WHERE occurrenceStatus = 'PRESENT'
 AND countryCode = 'PL'
 AND year >= 2000
 AND kingdom = 'Animalia'
 AND decimalLatitude IS NOT NULL
 AND speciesKey IS NOT NULL
 AND NOT ARRAY_CONTAINS(issue.array_element, 'COUNTRY_COORDINATE_MISMATCH')
 AND month IS NOT NULL
GROUP BY
 year,
 eeaCellCode,
 speciesKey
ORDER BY
 year DESC,
 eeaCellCode ASC,
 speciesKey ASC;
```

Figure 1: A query for an example cube. It summarizes GBIF occurrence data in Poland, covering all animal records since 2000. The cube uses a species taxonomic dimension, a year temporal dimension and an EEA reference grid spatial dimension. For each group, the count and coordinate uncertainty are measured.

```
Unset
SELECT
 datasetKey,
 license,
 COUNT(*) AS n
FROM
 gbif.occurrence
WHERE occurrenceStatus = 'PRESENT'
 AND countryCode = 'PL'
 AND year >= 2000
 AND kingdom = 'Animalia'
 AND decimalLatitude IS NOT NULL
 AND speciesKey IS NOT NULL
 AND NOT ARRAY_CONTAINS(issue.array_element, 'COUNTRY_COORDINATE_MISMATCH')
GROUP BY
 datasetKey,
 license;
```

Figure 2: Further metadata for the example cube in Figure 1. It lists the datasets that contributed data to the cube, as well as their licence.

2. What requirements are implemented?

The implemented software requirements (D2.1; Desmet et al. 2023a) are listed below. Outstanding “MUST” requirements are recorded in GitHub as issues.

Cube specification (section 3)

- 3.1 Dimensions:
  - Top-level requirements: all implemented.
  - 3.1.1 Taxonomic: All MUST requirements implemented, as well as SHOULD requirements from Table 1.
  - 3.1.2 Temporal: All MUST requirements implemented. Requirement 3 partially implemented.
  - 3.1.3 Spatial: All MUST requirements implemented.
  - 3.1.4 Other: All MUST and SHOULD requirements implemented, some MAY requirements are implemented.
- 3.2 Measures:
  - 3.2.1 Occurrence count: Both MUST requirements implemented.
  - 3.2.2 Minimum coordinate uncertainty: MUST and SHOULD requirements implemented.
  - 3.2.3 Minimum temporal uncertainty: Not yet implemented, although a user may use standard SQL to achieve this.
  - 3.2.4 Sampling bias: Implemented, except the default values.
- 3.3 Format:
  - CSV (MUST): Implemented.
  - EBV NetCDF (MUST): Not implemented, pending work at the hackathon using downstream software to implement this.
  - Other formats (SHOULD, MAY): Not implemented.
- 3.4 Metadata:
  - The software allows some internal metadata (datasets used, licences) to be calculated.
  - External metadata (DataCite schema, DOI assignment) is not implemented, as this requires depositing the cube and will be part of the following deliverable, the cube workflow service.
- 3.5 Findability and storage: Not part of this deliverable.

Cube production software (section 4.1)

- 4.1.1 (source data) is complete.
- 4.1.2 (parameters) is complete, except 2c (reasonable defaults).
- 4.1.3 (reference grids) is complete, except for specifying a random seed (section 3.1.3, 6b).
- 4.1.4 (cube specification) is addressed in section 3, see above.
- 4.1.5 (cloud processing) is complete.
- 4.1.6 (best practices) is complete.
- 4.1.7 (open source) is complete.

3. Cloud environment

The software may be used in a cloud environment, and is capable of using GBIF occurrence data as its source or external data.

A cloud-hosted test environment has been set up by GBIF on Microsoft Azure (M5, Desmet et al. 2023b). It provides access to the occurrence cube software as well as monthly snapshots of GBIF occurrence data, as a queryable table.

Users may also install the software on their own cloud environment following the documentation at https://links.gbif.org/cube-setup.

4. Source code, issues and feedback

The software is developed in a public GitHub repository, including an issue tracker for feedback from users. Releases are deposited on Zenodo, the first release is available at Blissett et al. (2024, https://doi.org/10.5281/zenodo.10607134).

5. Next steps

A first prototype release of the software is now available. In the coming months, it will be extended to meet the entire software specification (D2.1, Desmet et al. 2023a) and improved based on testing and feedback from the partners. Work is already underway to offer the software as a service (D2.3), with documentation at https://links.gbif.org/sql-downloads.

6. References

Blissett M, Robertson T, Desmet P (2024). Occurrence cube functions (cube-0.1.0). Global Biodiversity Information Facility (GBIF). https://doi.org/10.5281/zenodo.10607134

Desmet P, Oldoni D, Blissett M, Robertson T (2023a). Specification for species occurrence cubes and their production. B-cubed project deliverable D2.1. https://b-cubed.eu/storage/app/uploads/public/64d/1f7/5a2/64d1f75a2fc96997998232.pdf

Desmet P, Blissett M, Robertson T (2023b). Software has been tested by partners and feedback is collected. B-cubed project milestone M5.

European Environment Agency (2013). EEA reference grid. Accessed via https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2 on 2023-06-12.

Larsen R (2021). Geocoding and generalisations. Accessed via https://towardsdatascience.com/geocoding-and-generalisations-41fa5652d34c on 2023-06-07.

Microsoft Planetary Computer. https://planetarycomputer.microsoft.com

Veness C (2020). Convert between Latitude/Longitude &

UTM coordinates / MGRS grid references. Accessed via https://www.movable-type.co.uk/scripts/latlong-utm-mgrs.html on 2023-06-15.