BIODIVERSITY
BUILDING
BLOCKS FOR
POLICY

Specification for species occurrence cubes and their production

30/06/2023

Authors: Peter Desmet, Damiano Oldoni, Matthew Blissett, Tim Robertson

Funded by
the European Union

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.

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:	Software specifications
Deliverable n°:	D2.1
WP responsible:	WP2
Nature of the deliverable:	Document, Report
Dissemination level:	Public
Lead partner:	EV INBO
Recommended citation:	Desmet P, Oldoni D, Blissett M, Robertson T (2023). *Specification for species occurrence cubes and their production*. B-cubed project deliverable D2.1.
Due date of deliverable:	Month n°4
Actual submission date:	Month n°4

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final	30 June 2023	Peter Desmet (EV INBO), Damiano Oldoni (EV INBO), Matt Blissett (GBIF), Tim Robertson (GBIF)

Table of contents

Key takeaway messages **4**
Executive summary **4**
Non-technical summary **5**
List of abbreviations **5**
1. Introduction **6**
2. Methodology **7**
3. Cube specification **8**
3.1. Dimensions 8
3.1.1. Taxonomic 8
3.1.2. Temporal 10
3.1.3. Spatial 10
3.1.4. Other 13
3.2. Measures 14
3.2.1. Occurrence count 14
3.2.2. Minimum coordinate uncertainty 14
3.2.3. Minimum temporal uncertainty 15
3.2.4. Sampling bias 15
3.3. Format 18
3.4. Metadata 19
3.5. Findability and storage 20
4. Software specification **21**
4.1. Cube production software 21
4.2. Cube workflow service 23
5. Acknowledgements **24**
6. References **25**

Key takeaway messages

- Effective biodiversity management and policy requires analysis-ready biodiversity data.
- Species occurrence cubes provide such data by grouping species occurrence data along spatial, temporal and/or taxonomic dimensions.
- This document specifies the technical properties of those cubes.
- This document specifies the requirements for software to produce those cubes.
- The software will be implemented as a new service provided by the Global Biodiversity Information Facility (GBIF).

Executive summary

This document presents the specification for “species occurrence cubes”, a format to summarize species occurrence data. It also outlines the requirements for software to produce such cubes and how it can be integrated in services provided by the Global Biodiversity Information Facility (GBIF).

Producing a species occurrence cube will broadly involve the following steps:

1. Search and filter data: a user will be able to restrict a cube to occurrence data of their interest.
2. Define cube dimensions: a user will be able to select from a number of dimensions and categories that determine how occurrence data will be grouped into a cube. For example, taxonomic information can be grouped by family, temporal information by year, and spatial information by a chosen grid reference system, taking into account the spatial uncertainty associated with occurrences.
3. Generate cube: based on the parameter selection by the user, software will process the occurrences into a species occurrence cube, providing measures for each combination of dimensions (e.g. 5 occurrences for species x at year y in grid cell z). The software can also provide reference measures to assess sampling bias.
4. Deposit cube: the user can define in what format and where (cloud storage location) a cube should be deposited. Deposited cubes will be automatically documented with metadata and assigned a Digital Object Identifier (DOI) so they can be reproduced and referenced.

The software developed for this service must be open source and documented, so that users can understand, use, install and contribute to it. It must also be demonstrated to operate on one or more public cloud providers.

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. These data are essential for policy and research, but challenging for users to download and process on a large scale.

This document describes a reporting service that will be built by GBIF to facilitate that process. Rather than downloading individual records, users will be able to download a summary of the data based on their preferences. For example, they can select specific species, years, or other aspects of the data that are relevant to their interests. Based on these preferences, a summary table called a "species occurrence cube" will be generated for download. This cube will contain a condensed representation of the requested information along with record counts.

Users often want to compare the species occurrence cubes with other datasets, such as land cover data sets. To make this comparison easier, the reporting service will enable users to specify how the data are grouped by location. For example, they can choose a known gridding scheme, which divides the study area into smaller sections or grids, allowing for consistent comparisons.

Once a report is generated, it is provided with a Digital Object Identifier (DOI) to ensure that it can be easily and accurately cited and that the report will remain available for others to reuse.

List of abbreviations

API	Application Programming Interface
EBV	Essential Biodiversity Variable
EEA	European Environment Agency
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
GBIF	Global Biodiversity Information Facility
SQL	Structured Query Language

1. Introduction

Climate change, environmental degradation and invasive species represent imminent threats to biodiversity. Effective biodiversity management and policy decisions urgently require access to timely, accurate, and reliable information on biodiversity status, trends, and threats. Unprecedented amounts of biodiversity data are being accumulated from diverse sources, aided by emerging technologies such as automatic sensors, eDNA, and satellite tracking. However, the process of data cleaning, aggregation, and analysis is often time-consuming, convoluted, laborious, and irreproducible. Biodiversity monitoring across large areas and projects faces challenges in evaluating data completeness and sampling bias.

To address these challenges, the development of tools and infrastructure is crucial for meaningful interpretations and deeper understanding of biodiversity data. Furthermore, a significant delay exists in converting biodiversity data into actionable knowledge. Efforts have been made to reduce this lag through data standardization, rapid mobilisation of biodiversity observations, digitization of collections, and streamlined workflows for data publication. However, delays still occur in the analysis, publication, and dissemination of data.

The B-Cubed project proposes solutions to overcome these challenges. One of those is extending and implementing the intermediary data product “occurrence cube” (Oldoni et al. 2020), which aggregate species occurrence data along spatial, temporal and/or taxonomic dimensions. The idea of creating aggregated biodiversity “data cubes” with taxonomic, spatial and temporal dimensions has also been proposed within the Group on Earth Observations Biodiversity Observation Network (GEOBON) (Kissling et al. 2017) to deliver Essential Biodiversity Variables (EBV). This document specifies the properties of such occurrence cubes. It also documents the requirements for software to produce such cubes and a service to deliver those in a way that is Findable, Accessible, Interoperable and Reusable (FAIR). The software and service will be implemented and provided by the Global Biodiversity Information Facility (GBIF).

By leveraging aggregated occurrence cubes as analysis-ready biodiversity datasets, we aim to enhance comprehension and reduce barriers to accessing and interpreting biodiversity data. Automation of workflows will provide regular and reproducible indicators and models that are open and useful to users. Additionally, the use of cloud computing offers scalability, flexibility, and collaborative opportunities for applying advanced data science techniques anywhere. Finally, close collaboration with stakeholders will inform us of the requirements for tools, increase impact, and facilitate the flow of information from primary data to the decision-making processes.

2. Methodology

The specification in this document are based on the concept of “occurrence cubes” as described in Oldoni et al. (2020). We expanded those to meet the requirements of the B-Cubed project partners and to describe a cube production service to be hosted by GBIF. Feedback was gathered from B-Cubed project partners in the kick-off meeting (March 13-14, 2023), two online calls (April 24 and 27, 2023) and a document open for comments.

Where possible, the specification build on infrastructure and services already provided by GBIF (e.g. occurrence processing, occurrence search, download service, etc.).

The key words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL in this document are to be interpreted as described in RFC 2119.

3. Cube specification

3.1. Dimensions

Dimensions define how occurrences are grouped into a combination of categories, similar to the GROUP BY clause in SQL. A combination of dimension categories is called a “group”, e.g. taxon X, year Y and grid cell Z is a group.

1. A cube MUST have at least one dimension.
2. A cube MUST at maximum have a number of groups that is equal to the number of dimensions multiplied by the number of categories per dimension.
3. Groups without any associated occurrences MUST NOT be included in the cube, to ensure a user won't unwittingly assume this represents a statement of species absence. A cube will therefore typically contain (far) less groups than are theoretically possible.

3.1.1. Taxonomic

The taxonomic dimension groups occurrences into categories using their taxonomic information, i.e. “what was observed?”. Relevant terms are scientificName, kingdom, and terms derived from species matching with the GBIF Backbone Taxonomy (GBIF Secretariat 2022). Grouping is especially useful to lump synonyms and child taxa.

1. This dimension MUST be optional.
2. A number of categories MUST be supported (see Table 1 for details). All of these are existing occurrence properties (example). They are added automatically by the GBIF occurrence processing pipeline, when matching an occurrence to the GBIF Backbone Taxonomy (GBIF Secretariat 2022).
    a. The category speciesKey SHOULD be selected by default.
    b. Note that the category taxonKey is different from the GBIF taxonKey search parameter. The latter lumps synonyms and child taxa, e.g. *Vespa velutina* Lepeletier, 1836 (taxonKey 1311477) includes both the accepted subspecies *Vespa velutina nigrithorax* Buysson, 1905 (taxonKey 6247411) and the synonym *Vespa auraria* Smith, 1852 (taxonKey 1311484). The category taxonKey should only lump occurrences that share the same taxonKey. This SHOULD be communicated clearly to the user.
3. Occurrences that are identified at a higher taxon rank than the selected category MUST NOT be included, e.g. an occurrence identified as genus *Vespa* (taxonKey 1311334) is excluded when using a speciesKey category.
4. Occurrences MUST NOT be assigned to multiple categories.
5. Since the values in the categories are integers that are not self-explanatory, additional columns with the names of the taxa and their higher taxonomy (see Table 2) SHOULD be provided. This MAY be provided in the form of a taxonomic compendium as an additional file (cf. be_species_info.csv in Oldoni et al. 2022).

Table 1: Categories for the taxonomic dimension.

Category	Remarks	Need
kingdomKey	Lumps synonyms and child taxa.	SHOULD
phylumKey	Lumps synonyms and child taxa.	SHOULD
classKey	Lumps synonyms and child taxa.	SHOULD
orderKey	Lumps synonyms and child taxa.	SHOULD
familyKey	Lumps synonyms and child taxa.	MUST
genusKey	Lumps synonyms and child taxa.	SHOULD
speciesKey	Lumps synonyms and child taxa.	MUST
acceptedKey	Lumps synonyms, but not child taxa.	SHOULD
taxonKey	Does not lump synonyms nor child taxa.	MUST

Table 2: Examples of which columns of taxonomic information to include for three different taxonomic dimensions (taxonKey, speciesKey and orderKey).

Column	Cube at taxonKey	Cube at speciesKey	Cube at orderKey
kingdomKey	TRUE	TRUE	TRUE
kingdom	TRUE	TRUE	TRUE
phylumKey	TRUE	TRUE	TRUE
phylum	TRUE	TRUE	TRUE
classKey	TRUE	TRUE	TRUE
class	TRUE	TRUE	TRUE
orderKey	TRUE	TRUE	TRUE
order	TRUE	TRUE	TRUE
familyKey	TRUE	TRUE	FALSE
family	TRUE	TRUE	FALSE
genusKey	TRUE	TRUE	FALSE
genus	TRUE	TRUE	FALSE
speciesKey	TRUE	TRUE	FALSE
species	TRUE	TRUE	FALSE
acceptedKey	TRUE	FALSE	FALSE
acceptedScientificName	TRUE	FALSE	FALSE
taxonKey	TRUE	FALSE	FALSE
scientificName	TRUE	FALSE	FALSE
taxonRank	TRUE	TRUE (“SPECIES”)	TRUE
taxonomicStatus	TRUE	TRUE (“ACCEPTED”)	TRUE (“ACCEPTED”)

3.1.2. Temporal

The temporal dimension groups occurrences into categories using their temporal information, i.e. “when was it observed?”. Relevant terms are eventDate, year, day, and month. Grouping is especially useful to reduce the temporal information from a continuum into discrete categories.

1. This dimension MUST be optional.
2. A number of categories MUST be supported (see Table 3 for details). All of these are existing occurrence properties (example), albeit as discrete (year, month, day) not combined (year, yearmonth, yearmonthday) properties. They are added automatically by the GBIF occurrence processing pipeline, when processing the eventDate into year, month, and day.
    a. The category year SHOULD be selected by default.
3. Occurrences that have temporal information that is wider than the selected category SHOULD NOT be included, e.g. an occurrence with date range 2020-12-15/2021-01-15 is excluded when using a year category.
    a. Alternatively, the middle of the date range MAY be used.
4. Occurrences MUST NOT be assigned to multiple categories.

Table 3: Categories for the temporal dimension.

Category	Remarks	Need
year		MUST
yearmonth		SHOULD
yearmonthday (date)		MUST

3.1.3. Spatial

The spatial dimension groups occurrences into categories using their spatial information, i.e. “where was it observed?”. Relevant terms are decimalLatitude, decimalLongitude, geodeticDatum, and coordinateUncertaintyInMeters, as well as a reference grid. Grouping is especially useful to map data to other spatial datasets using the same reference grid and to take into account the coordinate uncertainty.

1. This dimension MUST be optional.
2. Only one spatial dimension MUST be used at a time in a cube.
3. A number of reference grids and cell sizes MUST be supported (see Table 5 for details).
    a. By default, a reference grid SHOULD NOT be selected, so that all options are considered equal.
4. Non-gridded reference datasets SHOULD NOT be supported. Examples include Administrative areas (GADM 2022) and the World Database on Protected Areas (WDPA) (Protected Planet 2012).
    a. Such datasets may not be area-covering and can have overlapping features, leading to misleading results.

   b. Users are advised to make use of such datasets after cube generation. This also allows them more control and flexibility in choosing features of interest and how to combine these with the chosen reference grid.

5. Occurrences SHOULD be considered circles or squares (not points).
   a. Circles MUST be based on the point-radius method (Wieczorek et al. 2004), using the coordinates as the centre and the provided coordinateUncertaintyInMeters as the radius. If not provided, a default coordinateUncertaintyInMeters of 1000m SHOULD be assumed. Users SHOULD be able to specify this value.
   b. Squares SHOULD be based on the provided footprintWKT or MAY be reverse-engineered when the dataset is likely gridded (Waller 2019).
6. A number of grid assignment methods MUST be supported (see Table 4 for detailed needs).
   a. Random grid assignment SHOULD be selected by default.
   b. The seed used for random grid assignment SHOULD be mentioned in the metadata and users SHOULD be able to reuse it to create reproducible results.
   c. Occurrences that have a spatial extent that is wider than the largest grid cell MUST NOT be included when using encompassing grid assignment (they can in random grid assignment).
7. Occurrences that are located beyond the extent of the chosen reference grid MUST NOT be included.
8. Occurrences MUST NOT be assigned to multiple grid cells (i.e. no fuzzy assignment).

Table 4: Grid assignment methods.

Method	Remarks	Need
Random grid assignment	Assigns an occurrence to a random grid cell (of defined size) that overlaps with it. See Oldoni et al. (2020) for details.	MUST
Encompassing grid assignment	Assigns an occurrence to the smallest grid cell size that fully encompasses it. Useful for downscaling approaches (Groom et al. 2018).	SHOULD

Table 5: Reference grids and their cell sizes. Quoted example values are codes for cells encompassing this occurrence in Slovenia at latitude 46.565825 N (46° 33' 56.97" N) and longitude 15.354675 E (15° 21' 16.83" E).

Grid	Cell sizes	Remarks	Need
EEA reference grid	● 1x1 km (“1kmE4731N2620”) ● 10x10 km (“10kmE473N262”) ● 100x100 km (“100kmE47N26”)	European coverage, used for many reporting purposes. See European Environment Agency (2013) for details.	MUST
Extended Quarter Degree Grid Cells (GDGC)	● 15x15 minutes (“E015N46AD”) ● 30x30 minutes (“E015N46A”) ● 1x1 degrees (“E015N46”)	Worldwide coverage, mostly used in African countries. See Larsen et al. (2009) for details. Cells can be downloaded for a selection of countries (Zenodo 2023) or calculated (Larsen 2021).	MUST
Military Grid Reference System (MGRS)	● 1x1 m (“33TWM2718256978”) ● 10x10 m (“33TWM27185697”) ● 100x100 m (“33TWM271569”) ● 1x1km (“33TWM2756”) ● 10x10 km (“33TWM25”) ● 100x100 km (“33TWM”)	Worldwide coverage, excluding polar regions north of 84°N and south of 80°S. Derived from Universal Transverse Mercator (UTM), but grid codes consist of Grid Zone Designator (33T), 100 km Grid Square ID (WM) and numerical location (Veness 2020).	MUST

3.1.4. Other

Other dimensions could be envisioned to group occurrences.

1. These dimensions MUST be optional.
2. These dimensions MUST be categorical (i.e. controlled vocabularies) or converted to a specified number of quantiles.
3. Occurrences that are not associated with a category MUST be assigned to NOT-SUPPLIED.
4. A number of other categories MAY be supported (see Table 6 for details).
    a. By default, other categories SHOULD NOT be selected.
    b. Note that for some (e.g. establishmentMeans), users are advised to assign these properties after cube production. This also allows them more control and flexibility.
5. Occurrences MUST NOT be assigned to multiple categories.

Table 6: Other categories.

Category	Remarks	Need
Sex		SHOULD
Life stage	Especially important for insects (Radchuk et al. 2013) and invasive species (Wallace et al. 2021).	MAY
Establishment means (derived)	Derived from comparing the occurrence with checklist information (e.g. occurrence is considered “introduced” by checklist x for this species, area and time). This is a spatial dimension, occurrences SHOULD be assigned using one of the methods in Table 4.	MAY
Degree of establishment (derived)	Derived from comparing the occurrence with checklist information (e.g. occurrence is considered “managed” by checklist x for this species, area and time). This is a spatial dimension, occurrences SHOULD be assigned using one of the methods in Table 4.	MAY
IUCN Global Red List Category	Derived from comparing the occurrence with checklist information (e.g. occurrence is considered “vulnerable” by checklist x for this species, area and time). This is a spatial dimension, occurrences SHOULD be assigned using one of the methods in Table 4.	MAY
Trait	More investigation is needed to assess how species trait information (e.g. from Open Traits Network) can be linked to species occurrences.	MAY

3.2. Measures

Measures are the calculated properties per group, similar to aggregate functions (count, sum, average, minimum, etc.) in SQL. Note that a group is a combination of dimension categories (see Section 3.1).

1. The following measures SHOULD be selected by default: occurrence count, minimum coordinate uncertainty.

3.2.1. Occurrence count

1. The occurrence count MUST be included per group.
2. This measure MUST be an integer value expressing the number of occurrences within a group.

The occurrence count provides information on occupancy as well as how many occurrences contributed to the occupancy. Groups with occupancy = FALSE are by definition not present in the cube, see Section 3.1.

3.2.2. Minimum coordinate uncertainty

1. The minimum coordinate uncertainty SHOULD be included per group.
2. This measure MUST be a numeric value expressing the minimum coordinateUncertaintyInMeters associated with an occurrence within a group.

The minimum coordinate uncertainty indicates the minimum spatial extent of occurrences within a group. This is especially useful when using random grid assignment (see Table 4). Consider an example where there are 4 occurrences for taxon X for year Y near grid cell Z (1x1km). Three of those occurrences are coming from a dataset with 10x10km gridded data and have an coordinateUncertaintyInMeters of 7071m. They can be represented as circles that partly or completely include grid cell Z. Due to the random grid assignment method, only one is assigned to grid cell Z, the others to neighbouring grid cells that overlap with their circles. A fourth occurrence is derived from iNaturalist, has an uncertainty of 30m and falls completely within grid cell Z. It is assigned to grid cell Z. The cubed data for XYZ would be:

- year: X
- taxon: Y
- grid: Z
- count: 2
- minimumCoordinateUncertainty: 30

The minimum coordinate uncertainty gives an indication that there was at least one occurrence with a high likelihood of falling completely within grid cell Z. This property can also be used to filter out groups that only contain occurrences that are smeared out over many grid cells (but were randomly assigned to that one). Such groups could be excluded from some spatial analyses at high resolution, but included in temporal analyses.

3.2.3. Minimum temporal uncertainty

1. The minimum temporal uncertainty MAY be included per group.
2. This measure SHOULD be an integer value expressing the minimum temporal range in seconds associated with an occurrence within a group. Examples are provided in Table 7.

The minimum temporal uncertainty indicates the minimum temporal extent of occurrences within a group. This is especially useful to filter out groups that only contain occurrences with broad temporal information.

Table 7: Examples of minimum temporal uncertainty for provided eventDates.

eventDate	minimum temporal uncertainty	Remarks
2021-03-21T15:01:32.456Z	1	Milliseconds are rounded to seconds.
2021-03-21T15:01:32Z	1	
2021-03-21T15:01Z	60	
2021-03-21T15Z	60*60	
2021-03-21	60*60*24	
2021-03-01	60*60*24	For dates at the first day of the month, the minimum temporal uncertainty MAY also be considered 60*60*24*31.
2021-01-01	60*60*24	For dates on the first day of the year, the minimum temporal uncertainty MAY also be considered 60*60*24*365.
2021-03	60*60*24*31	
2021	60*60*24*365	
2021-03-21/2021-03-23	60*60*24*3	

3.2.4. Sampling bias

A species could be well represented for a certain year and grid cell not because it is particularly established there, but because it was observed more (e.g. as result of a bioblitz or because it is a rare species observers seek out). To compensate for this sampling bias, it is important to know the sampling effort. For most cases, direct measures of sampling effort are not available, so one must rely on proxy measures to indicate sampling bias/effort.

An easy metric is the total number of occurrences for a “target group” (Botella et al. 2020, de Beer et al. 2023), a group at a higher taxonomic rank than the focal taxon. To avoid confusion with the term “group” as defined in Section 3.1, we will refer to this as “higher taxon”. For example, the higher taxon for the focal taxon *Vanessa atalanta* could be the genus *Vanessa*, the family *Nymphalidae*, the order *Lepidoptera*, the class *Insecta*, the phylum *Arthropoda* or the kingdom *Animalia*. It allows to calculate a relative occurrence count (i.e. the occurrence count of the focal taxon divided by the occurrence count of the higher taxon). See GBIF Secretariat (2018) for an implementation that makes use of this to show relative observation trends. In addition to the number of occurrences, the number of days the higher taxon was observed and/or the number of observers that observed the higher taxon could also be provided.

1. The target occurrence count SHOULD be included per group to facilitate assessing sampling bias.
2. This measure MUST be an integer value expressing the number of occurrences within a group (see Table 8). Note that by dividing the occurrence count by the target occurrence count, one can calculate a relative count.
3. This measure SHOULD take into account any filters applied to the occurrence data, except for taxonomic filters. For example, for occurrence data filtered on *Vanessa atalanta* (scientificName), human observation (basisOfRecord) and INBO (publisher), a higher taxon at family SHOULD retain the filters basisOfRecord and publisher.
4. This measure SHOULD use the same grid assignment method (see Table 4) as selected for the spatial dimension.
5. This measure SHOULD NOT increase the number of records in the cube. For example, grid cells that are occupied by the higher taxon, but not by the focal taxon, SHOULD NOT be included.
6. The higher taxon rank SHOULD be defined by the user:
    a. It SHOULD either be genus, family, order, class, phylum, kingdom or life (all kingdoms).
    b. The rank MUST be higher than the selected rank for the taxonomic dimension (see Table 1), e.g. only phylum, kingdom or life are valid for a cube at class level (classKey).
    c. family SHOULD be selected by default for cubes with a taxonomic dimension at taxon level (acceptedKey, taxonKey), species level (speciesKey) or genus level (genusKey). The direct higher rank SHOULD be selected by default for other cubes with a higher taxonomic dimension.
    d. It SHOULD NOT be possible to select more than one rank. Note that it is theoretically possible to provide this measure for all (higher) ranks.
    e. If a taxon does not have a parent at the selected rank, its target occurrence count SHOULD be NULL.
7. Other measures than target occurrence count MAY be considered, including:
    a. Number of days observed.
    b. Number of observers (recordedBy). Note that this value is not controlled and can lead to higher numbers than expected.

Table 8: Example of target occurrence counts at genus level for a cube with taxonomic and temporal dimensions.

speciesKey	year	count	genusCount
1311527 (Vespa crabro)	2020	15152	20361
1311527 (Vespa crabro)	2021	15055	20533
1311527 (Vespa crabro)	2022	20655	38641
1311527 (Vespa crabro)	2023	1805	7192
1311477 (Vespa velutina)	2020	3683	20361
1311477 (Vespa velutina)	2021	3825	20533
1311477 (Vespa velutina)	2022	16259	38641
1311477 (Vespa velutina)	2023	5108	7192
1898286 (Vanessa atalanta)	2020	102732	126961
1898286 (Vanessa atalanta)	2021	106411	141924
1898286 (Vanessa atalanta)	2022	76869	125379
1898286 (Vanessa atalanta)	2023	8155	17546

3.3. Format

Since cubes are tabular data, they can be expressed in any format that supports this. It is advised however to choose open formats with broad support.

1. A number of output formats MUST be supported (see Table 9 for details).
    a. CSV SHOULD be selected by default.
2. A geospatial format MUST only be supported if the cube includes the spatial dimension.

Table 9: Output formats.

Format	Remarks	Need
CSV	Widely used format, including (tab-delimited and compressed) by the GBIF occurrence download service (GBIF Secretariat 2023a). Broad software support.	MUST
EBV NetCDF	Network Common Data Format (netCDF) format adopted by GeoBON to exchange Essential Biodiversity Variables. Can be read by e.g. R package “ebvcube” (Quoss et al. 2021).	MUST
Apache Parquet	Column-oriented data format, optimized for data storage and retrieval. Increasingly used in tools like Google Big Query. Can be read by e.g. R package “arrow” (Richerson et al. 2023).	SHOULD
Apache Avro	Row-oriented data format. Often recommended for long term storage over Apache Parquet, at a cost of performance when reading.	MAY
GeoJSON	See https://geojson.org/	MAY
GeoParquet	See https://geoparquet.org/	MAY
GeoTIFF	See https://www.ogc.org/standard/geotiff/	MAY
HDF5	See https://www.hdfgroup.org/solutions/hdf5/	MAY
JSON	See https://www.json.org/	MAY
PMTiles	See https://protomaps.com/docs/pmtiles	MAY
ZARR	See https://zarr.readthedocs.io/en/stable/	MAY

3.4. Metadata

Metadata documents how a cube was generated and can be cited.

1. Metadata MUST be provided in a machine-readable format such as JSON or XML.
2. Metadata SHOULD make use of DataCite Metadata Schema (DateCite Metadata Working Group 2021). This is currently the case for GBIF occurrence downloads (example).
3. Metadata MUST include the properties in Table 9.
4. Metadata MUST include all the parameters that were used to generate the cube, allowing it to be reproduced.
    a. The parameters MUST be provided in a machine-readable format such as JSON or REST API query parameters.
    b. The parameters MUST include the selected occurrence search filters. This is currently the case for GBIF occurrence downloads (GBIF Secretariat 2023a) (see “description” in this example). Any default values SHOULD also be included.
    c. The parameters MUST include the selected cube properties, such as dimensions, categories, reference grids, default coordinate uncertainty, seed for random grid assignment (see Section 3.1), measures (see Section 3.2) and format (see Section 3.3).
5. Metadata MUST include a stable and unique global identifier, so it can be referenced. This SHOULD be a Digital Object Identifier (DOI).
6. Metadata MUST include the creator, publisher, and creation date of the cube.
7. Metadata MUST include the GBIF-mediated occurrence datasets that contributed to the cube as related identifiers, so these can be credited.
8. Metadata MUST include the licence under which it is deposited.
9. Metadata SHOULD document the columns in the cube. This MAY be expressed using Frictionless Table Schema (Walsh & Pollock 2012).

3.5. Findability and storage

While a cube generated for testing purposes can be ephemeral, downstream use requires cubes to be findable, accessible, persistent and available on (cloud) infrastructure.

1. A cube intended for downstream use MUST be identifiable and findable using a Digital Object Identifier (DOI).
2. A cube intended for downstream use SHOULD be publicly accessible.
3. A cube intended for downstream use SHOULD be deposited on infrastructure that can guarantee its long-term archival (e.g. GBIF, EBV Data Portal, Zenodo). See table 10 for details.
    a. GBIF downloads SHOULD be selected by default.
4. The option SHOULD be offered to make a cube available on the cloud infrastructure where it will be processed. See table 10 for details.
    a. By default, a cloud infrastructure SHOULD NOT be selected.

Table 10: Data storage infrastructures.

Infrastructure	Remarks	Need
GBIF downloads	Infrastructure maintained by GBIF for the long term-archival of occurrence data. See GBIF Secretariat (2023a) for details.	MUST
EBV Data Portal	Infrastructure maintained by GeoBON for the long-term archival of Essential Biodiversity Variables raster datasets, see https://portal.geobon.org/	MUST
Amazon Web Services S3	Commercial cloud infrastructure, see https://aws.amazon.com/s3/	MAY
Google Cloud Storage	Commercial cloud infrastructure, see https://cloud.google.com/storage	MAY
Microsoft Azure Cloud Storage	Commercial cloud infrastructure, see https://azure.microsoft.com/en-us/products/category/storage	MAY

4. Software specification

4.1. Cube production software

This software produces cubes following the specification above.

1. The software MUST use species occurrence data as its source.
    a. The software MUST accept tabular representations of occurrence data expressed using Darwin Core, including CSV file formats.
    b. The software SHOULD assume occurrence data to be formatted (i.e. have the same fields) as data returned by GBIF in occurrence downloads.
    c. The software MUST NOT assume the GBIF occurrence index to be the source of this data. Users SHOULD be able to provide their own occurrence data (e.g. for testing purposes).
2. The software MUST use parameters by which users can define how a cube is produced.
    a. The parameters MUST include the selected cube properties, such as dimensions, categories, reference grids, default coordinate uncertainty, seed for random grid assignment (see Section 3.1), measures (see Section 3.2) and format (see Section 3.3).
    b. The parameter values MUST be controlled.
    c. The parameters SHOULD use reasonable defaults where relevant (see Section 3).
    d. SQL MAY be considered as the notation format for the parameters.
3. The software MUST be able to use reference grids (see Table 5).
    a. Reference grids MAY be reformatted to optimize processing. This process SHOULD be documented and repeatable to allow updates if necessary.
    b. Representing a reference grid as a formula SHOULD be preferred over storing a reference grid as data.
4. Using the input data and parameters, the software MUST produce the intended cube.
    a. The software MUST support the output formats defined in Section 3.3 or allow downstream services to convert to these formats.
    b. The software MUST return the metadata defined in Section 3.4 or allow downstream service to create this metadata. Note that default parameter values SHOULD also be included in the metadata.
    c. The software SHOULD NOT deposit the cube. This is better reserved for downstream services.
5. Users SHOULD be able to install and use the software, including on cloud processing platforms.
    a. Sufficient technical documentation MUST be provided that documents how the software can be installed.
    b. Sufficient technical documentation MUST be provided that documents how the software may be used on a cloud processing platform.
    c. This MUST be demonstrated on at least one public cloud provider such as Microsoft Azure through a tutorial or recorded demonstration or similar.
6. The software SHOULD be developed using best practices, including:
    a. Source code MUST be version controlled.
    b. The software SHOULD be organized in modular components (functions) to facilitate understanding and code contributions.

- c. The software functions MUST be documented to facilitate understanding and code contributions.
- d. The software MUST include tests to guarantee the intended functionality and prevent breaking changes.

7. The software MUST be released as open source software.
    - a. The software MUST be licensed under an open software licence such as Apache License 2.0.
    - b. The software SHOULD use semantic versioning for releases.
    - c. Source code SHOULD be hosted on GitHub to facilitate collaboration (including code contributions, feature requests, bug reports, etc.).

4.2. Cube workflow service

This service SHOULD embed the cube production software (Section 4.1) into the GBIF occurrence download service (GBIF Secretariat 2023a), allowing users to search for occurrences of interest and download/deposit these as a cube following their specification.

1. The service MUST allow users to **search and filter for occurrences** of interest. Note that the GBIF occurrence search (GBIF Secretariat 2023b) already provides this functionality.
2. The service MAY allow users to **exclude unwanted occurrences** (e.g. occurrences that were flagged). Note that the GBIF occurrence search (GBIF Secretariat 2023b) already provides this functionality through its API, but not at www.gbif.org.
    a. This MAY be implemented as a NOT filter.
3. The service MUST allow users to **define the dimensions** of the cube (see Section 3.1):
    a. The user MUST be able to select what dimensions (controlled list) to include.
    b. The user MUST be able to select what category/categories (controlled list) to use for each dimension.
    c. The user MUST be able to select what reference grid (controlled list, see Table 5) and grid assignment method (controlled list, see Table 4) to use for the spatial dimension.
    d. The user MAY be able to select a default coordinate uncertainty for occurrences that do not have this information.
    e. The user MAY be able to select the seed for random grid assignment.
    f. The service MAY provide information on the cardinality of the selected options, so users have an idea of the number of rows that will be returned in the cube (e.g. year to day "likely to increase the number of rows 360 times").
4. The service MAY allow users to **define the measures** included in the cube (see Section 3.2).
    a. Alternatively, the service MAY return the same measures for all cubes.
5. The service SHOULD allow users to **define the output format** of the cube (see Section 3.3 and Table 9).
    a. Alternatively, the service MAY use the same output format for all cubes, but MUST offer the possibility to create different distributions of a deposited cube in other formats.
6. The service SHOULD allow users to **define a destination** where the cube is deposited (see Section 3.4 and Table 10).
    a. Alternatively, the service MAY use the same destination to deposit all cubes, but MUST offer the possibility to copy a deposited cube to other destinations.
7. Sufficient technical documentation MUST be provided for users to understand and use the service.
8. The service MUST be provided as a REST API and SHOULD be integrated as part of the GBIF occurrence download service (GBIF Secretariat 2023a).
9. Interfaces to GBIF occurrence download API SHOULD be updated to incorporate the new functionality:
    a. The graphical user interface at https://www.gbif.org MUST be updated.
    b. The R package rgbif (Chamberlain et al. 2023a) SHOULD be updated.
    c. The Python package pygbif (Chamberlain et al. 2022b) MAY be updated.

5. Acknowledgements

We would like to thank the following people for providing feedback: Tim Adriaens, Lissa Breugelmans, Miguel Fernandez, Quentin Groom, Cang Hui, Alexis Joly, Sandra MacFadyen, Diego Marcos, Matilde Martini, Ward Langeraert and Andrew Rodrigues.

6. References

Botella C, Joly A, Monestiez P, Bonnet P, Munoz F (2020) Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. PLoS One 15:e0232078. https://doi.org/10.1371/journal.pone.0232078

Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K (2023a) rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.7.7.2, https://CRAN.R-project.org/package=rgbif

Chamberlain S, Forkel R, Legind J, Van Hoey S, Desmet P, Noé N (2023b) pygbif: Python client for the GBIF API. Python package version 0.6.3, https://pygbif.readthedocs.io/en/latest/

DataCite Metadata Working Group (2021) DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. https://doi.org/10.14454/3w3z-sa82

de Beer IW, Hui C, Botella C, Richardson DM (2023) Drivers of compositional turnover of narrow-ranged versus widespread naturalised woody plants in South Africa. Frontiers in Ecology and Evolution. 11:1106197. https://doi.org/10.3389/fevo.2023.1106197

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.

GADM (2022) Administrative areas. Accessed via https://gadm.org/ on 2023-06-16.

GBIF Secretariat (2018) Relative observation trends. Accessed via https://www.gbif.org/tools/observation-trends/about on 2023-06-26.

GBIF Secretariat (2022) GBIF Backbone Taxonomy. Checklist dataset https://doi.org/10.15468/39omei accessed via GBIF.org on 2023-06-07.

GBIF Secretariat (2023a) GBIF occurrence download API. Accessed via https://www.gbif.org/developer/occurrence#download on 2023-06-26.

GBIF Secretariat (2023b) GBIF occurrence search. Accessed via https://www.gbif.org/developer/occurrence#search on 2023-06-26.

Groom QJ, Marsh CJ, Gavish Y, Kunin WE. (2018) How to predict fine resolution occupancy from coarse occupancy data. Methods Ecol Evol. 2018; 9: 2273– 2284. https://doi.org/10.1111/2041-210X.13078

Kissling WD, Ahumada JA, Bowser A, Fernandez M, Fernández N, García EA, Guralnick RP, Isaac NJB, Kelling S, Los W, McRae L, Mihoub J-B, Obst M, Santamaria M, Skidmore AK, Williams KJ, Agosti D, Amariles D, Arvanitidis C, Bastin L, De Leo F, Egloff W, Elith J, Hobern D, Martin D, Pereira HM, Pesole G, Peterseil J, Saarenmaa H, Schigel D, Schmeller DS, Segata N, Turak E, Uhlir PF, Wee B, Hardisty AR (2018) Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol Rev, 93: 600-625. https://doi.org/10.1111/brv.12359

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

Larsen R, Holmern T, Prager SD, Maliti H, Røskaft, E. (2009) Using the extended quarter degree grid cell system to unify mapping and sharing of biodiversity data. African Journal of Ecology, 47: 382-392. https://doi.org/10.1111/j.1365-2028.2008.00997.x

Oldoni D, Groom Q, Adriaens T, Davis AJS, Reyserhove L, Strubbe D, Vanderhoeven S, Desmet P (2020) Occurrence cubes: a new paradigm for aggregating species occurrence data. bioRxiv 2020.03.23.983601 https://doi.org/10.1101/2020.03.23.983601

Oldoni D, Groom Q, Adriaens T, Hillaert J, Reyserhove L, Strubbe D, Vanderhoeven S, Desmet P (2022). Occurrence cubes at species level for European countries (Version 20221202) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7389450

Protected Planet (2012) World Database on Protected Areas (WDPA). Accessed via https://www.protectedplanet.net/en/thematic-areas/wdpa?tab=WDPA on 2023-06-15.

Quoss L, Fernandez N, Langer C, Valdez J, Pereira HM (2021) ebvcube: Working with netCDF for Essential Biodiversity Variables. https://cran.r-project.org/package=ebvcube

Radchuk V, Turlure C, Schtickzelle N (2013) Each life stage matters: the importance of assessing the response to climate change over the complete life cycle in butterflies. J Anim Ecol, 82: 275-285. https://doi.org/10.1111/j.1365-2656.2012.02029.x

Richardson N, Cook I, Crane N, Dunnington D, François R, Keane J, Moldovan-Grünfeld D, Ooms J (2023) arrow: Integration to 'Apache' 'Arrow'. https://cran.r-project.org/package=arrow

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.

Walsh P, Pollock R (2012) Table Schema. Version 1. Accessed via https://specs.frictionlessdata.io/table-schema/ on 2023-06-22.

Wallace RD, Bargeron CT, LaForest JH, Carroll RL (2021) The Life Cycle of Invasive Alien Species Occurrence Data. In Invasive Alien Species (eds T. Pullaiah and M.R. Ielmini). https://doi.org/10.1002/9781119607045.ch49

Waller J (2019) Gridded Datasets Update. Accessed via https://data-blog.gbif.org/post/gridded-datasets-update/ on 2023-06-13.

Wieczorek W, Guo G, Hijmans R (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty, International Journal of Geographical Information Science, 18:8, 745-767, https://doi.org/10.1080/13658810412331280211

Zenodo (2023) Quarter Degree Grid Cells community. Accessed via https://zenodo.org/communities/qdgc/ on 2023-06-07.