GUIDELINES

FOR DATA FORMAT

FOR THE ATLAS OF EUROPEAN MAMMALS 2024

A PROJECT OF THE EUROPEAN MAMMAL FOUNDATION

AND BIODIVERSITY DATA MANAGEMENT

VERSION I

European Mammal Foundation ZOOGDIER VERENIGING DISCOVER THE MAMMALS OF EUROPE nlbif sharing biodiversity data

This document aims,

(1) to share the agreements on data delivery for the compilation of the species presence maps for the second edition of the European Mammal Atlas, and

(2) to provide additional information on data management, data standards, information and links to free for use best practice data management and publication tools/software.

There are many embedded links in the document that will take you to the information source, where more relevant information can be found.

The document combines authored and published information. The sources of the published information are cited.

The communication on the data management will continue also after the end on this project, which may bring changes and development in the document, therefore we consider it as a first version.

It was compiled by Svetlana Miteva, with the contribution of Anthony Mitchell-Jones, Damian McFerran, Andrey Lissovsky and Jeroen Crewels.

These Guidelines are produced within the project of “Second edition of the Atlas of European Mammals in GBIF” (NLBIF Reference: HAF0171123is1), a project of The Habitat Foundation and the European Mammal Foundation, realised in cooperation with the Dutch Mammal Society.

The project “Second edition of the Atlas of European Mammals in GBIF” was possible thanks to the financial help of the NLBIF Foundation. NLBIF supports Dutch organisations and researchers to mobilize and publish biodiversity data.

Dit project/product is tot stand gekomen met een financiële bijdrage van Stichting NLBIF. NLBIF ondersteunt Nederlandse organisaties en onderzoekers bij het open en internationaal gestandaardiseerd online zetten van biodiversiteitsdata.

CONTENT

EMMA2 AGREEMENTS ON THE DATA SUPPLY FORMAT

General agreements on the data standards and protocols to be used for the data comprising the second edition of the European Mammal Atlas (European Mammals on Maps - EMMA2)

All data used for the Atlas of European Mammals will be based on validated observation, which can be traced back to the contributor.

The data delivered by the EMMA 2 National Coordinators will be managed by the Data Coordinator, Damian McFerran.

The National Coordinators will provide the information from the national or regional mammal databases they already have or will develop during this project.

Preferably, the data set from a given country should be supplied in Excel format.

If not possible, then Txt file that is tab delineated.

Each file should contain a fixed number of Tabs. The proposal is that the standards are applicable under Darwin Core for sharing data: see: https://www.gbif.org/darwin-core, and that we standardise the column headings.

Each file should be named, e.g., EMMA2_Russia_(will we get a file each year of a final data set?)

E.g., EMMA2_Russia_2018 or EMMA2_Russia_Final Records (2018-2024).

In advance, best to get a couple of test files/data sets to work through from different countries, so that we can test the mapping system.

In the next month's data sets will be exchanged to check how the process of data conversion from point data in WGS84 to CGRS (UTM) 50km grid will work.

Data format

The contributors should send through the following minimum information in their files:

- Country
- Country Code
- Species Code
- Species Name
- Date Class (1–3)
- Year (allowing us to confirm in correct date)
- Grid Reference. (Full CGRS 50km square details or sent the grid and we will then convert)
- Reference to trace record back to its origin.

Or so:

Country	Country Code	Species Code	Species Latin Name	Species Common Name	Date Class	Year	Grid	UTM Conversion	Date Records Submitted

Metadata

The first Tab should be the Metadata for the records, as shown on the table below.

Title	
Organization Name	
Name of Supplier	
Abstract	
Additional Info	
Licence	
Geographic Description	
Temporal Coverage	
Purpose	
Methods	
Quality Control	
Contact	
Contact Electronic Mail Address	
Contact phone	

See example with filled in table provided

Title	CEDaR_Dragonflies of Ireland
Organization Name	Centre for Environmental Data and Recording (CEDaR)
Abstract	Species records collected through Dragonfly surveys and commissioned work, using various methodology and sampling techniques appropriate to this particular taxon groups and for the purpose of assertaining the status and distribution of species throughout the island of Ireland.
Additional Info	Species
License	CC-BY-NC
Geographic Description	Ireland
Temporal Coverage	All previously recorded species
Purpose	To understand status and distribution of species with view to species/habitat management and production of a Red List
Methods	Various
Quality Control	Verified - Data collected and verified by International experts
Contact	CEDaR, National Museums Northern Ireland, Cultra BT180EU
Contact Electronic Mail Address	cedar.info@nmni.com
Contact phone	028 9039 5255

A requirement for all data sets that are received is to have standard column headings.
In addition, it will be important to provide if there are any Licence requirements that are now being sent through by the contributors.
Information should be provided also on:

- number of individuals that have been recorded
- the source of the records (field, specimen, literature)

For the countries that don't have databases yet and information is spread around, that will be very useful, because then it would be easier to see the data set source. This can be done in additional columns

Data processing

From above, the number of records and species from each 50km square will be generated.

A mechanism to retain version controls of the data sets that are received is required, in case errors are found.

Though part of the data will come from publication, which means it will be already verified by the author, validation of unpublished records might be needed. This will require time for the return of revised data sets, therefore the source of the records must be provided.

Currently, countries that have a national database/atlas, generally keep data in a relational database, often very complex.

Individual data collectors are likely to keep data in spreadsheets, and even these people may use one of the freely available recording packages.

We hope, by the end of the project when the time comes to submit data for the atlas all countries will have some sort of national database (or regional ones). It is important to follow a commonly agreed standard. Darwin Core (DwC) is the most popular choice. (See the next chapter with more info on DwC)

Since every participant is pursuing the best practice compatible models, these databases will most probably contain the essential elements of a biological record: species, date, place, recorder, which can always be linked to another databases. All countries that are not familiar and using DwC may consider integrating it in their recording systems.

Contributors

Countries with well-established recording systems (UK, Netherlands, Switzerland etc.) will include data from perhaps hundreds of people. NCs will be preparing a data account for their country, and can list people in that, and/or eventually list all data contributors in the on-line version of the atlas. How the data contributors will be cited/mentioned will be discussed further with the NCs.

Mapping projections

National systems use a variety of national grids to collect map-based observation (e.g. in the UK the Ordnance Survey grid, based on OSGB36 is used). Now most data are collected using GPS (WGS84).

There are apps that convert between the different grids, so data can be collected on phones using WGS84 and submitted using the appropriate national mapping system.

For the atlas, the NCs need to translate the data into the coordinate system based on the **CGRS** (UTM) 50km grid. Various tools are available to help with this, or it can just be done with geographic queries.

Errors of a few metres are possible, but their importance is small for the overall project goals.

ADDITIONAL INFORMATION ON DATA MANAGEMENT

DARWIN CORE – THE STANDARD FOR BIODIVERSITY INFORMATION

The Darwin Core Standard (DwC) offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources. Originally developed by the Biodiversity Information Standards (TDWG) community, Darwin Core is 'an evolving community-developed biodiversity data standard'. It plays a fundamental role in the sharing, use and reuse of open-access biodiversity data and today accounts for vast majority of the hundreds of millions of species occurrence records available through GBIF.org.
In practice, using Darwin Core revolves around a standard file format, the Darwin Core Archive (DwC-A). This compact package (a ZIP file) contains interconnected text files and enables data publishers to share their data using a common terminology. This standardization not only simplifies the process of publishing biodiversity datasets, it also makes it easy for users to discover, search, evaluate and compare datasets as they seek answers to today's data-intensive research and policy questions.

Darwin Core (often abbreviated to DwC) is an extension of Dublin Core for biodiversity informatics. It is meant to provide a stable standard reference for sharing information on biological diversity. The terms described in this standard are a part of a larger set of vocabularies and technical specifications under development and maintained by Biodiversity Information Standards (TDWG) (formerly known as the Taxonomic Databases Working Group (TDWG)).
The Darwin Core is a body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, and samples, and related information. Included in the standard are documents describing how these terms are managed, how the set of terms can be extended for new purposes, and how the terms can be used. The Simple Darwin Core is a specification for one particular way to use the terms and to share data about taxa and their occurrences in a simply-structured way. It is likely what is meant if someone were to suggest "formatting your data according to the Darwin Core".

Each term has a definition and commentaries that are meant to promote the consistent use of the terms across applications and disciplines. Evolving commentaries that discuss, refine, expand, or translate the definitions and examples are referred to through links in the Comments attribute of each term. This approach to documentation allows the standard to adapt to new purposes without disrupting existing applications. There is meant to be a clear separation between the terms defined in the standard and the applications that make use of them. For example, though the data types and constraints are not provided in the term definitions, recommendations are made about how to restrict the values where appropriate. In practice, Darwin Core decouples the definition and semantics of individual terms from application of these terms in different technologies such as XML, RDF or simple CSV text files. Darwin Core provides separate guidelines on how to encode the terms as XML or text files.

Additional information:

Darwin Core: An Evolving Community-Developed Biodiversity Data Standard
Article Source: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard
Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029715

Darwin Core website: https://dwc.tdwg.org/

GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)

The GBIF Integrated Publishing Toolkit (IPT) is an advanced tool for registering and publishing biodiversity data. The most important function of this tool is to take data from files, in separate files or stored in a database, and to pack them in a standardised zipped file, a so-called DarwinCore Archive (DwC-A). These zipped files can be exchanged quickly and easily online or downloaded directly. GBIF automatically extracts the DarwinCore Archives, unzips them and places the contents in the GBIF data index. The data can then be viewed, filtered, aggregated, downloaded or forwarded via the GBIF portal.

Data in the IPT

The GBIF IPT is used for three types of data:

1. Primary biodiversity data, or "occurrence data". This concerns data about observed or collected individual species.
2. Taxonomic checklists. Checklists are extensive files on the occurrence of certain groups of organisms in a specific area. The emphasis is on the taxonomy. Examples include "Dragonflies of the Netherlands" or "Birds of Northern Europe".
3. Meta-data. The IPT can store a very complete package of meta-data, that is, data about occurrence or checklist data. This meta-data is stored according to a commonly used meta-data standard (Ecological Metadata Language). The meta-data can be exported in a number of different ways. If the package of meta-data is sufficiently complete, this data can even be exported in the format of a so-called "Data Paper". These Data Papers are published in various scientific journals.

In the IPT, the meta-data can be published along with the occurrence or checklist data, or separately. If only the meta-data are online, it is still possible to find the occurrence or checklist data, but it is not possible to download them directly.

DarwinCore Archives (DwC-A)

By way of the IPT, data are stored in DarwinCore Archives (DwC-A). These zipped data packages contain:

1. The primary occurrence or checklist data, standardised according to the DarwinCore data standard, stored in text files.
2. The meta-data stored via the Ecological Metadata Language standard, in an XML file.
3. A Descriptor File that states where which info is located, stored in an XML file.

These DarwinCore Archives receive a unique URL within the IPT, GBIF registers this URL and retrieves the data as soon as the IPT user indicates this should happen. The data are often immediately visible and available in the central GBIF portal.
Additional specific data can also be supplied in so-called DarwinCore Extensions, there are extensions for measurements, non-scientific species names, genetic information, etc.

Other functionalities

The IPT includes RSS feeds, presents usage statistics and has an extensive administrative module that makes it possible for the IPT to be used by multiple users, each with different privileges. Style Sheets make it possible to adapt the IPT to suit the corporate or institutional identity of the data publishing organisation.
The IPT runs on standard web servers on Windows, Linux, Unix and Apple systems.
Additional information:
https://www.nlbif.nl/en/infrastructure/publishing-data/gbif-ipt
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4123864/
https://www.gbif.org/en/tool/81278/ipt-gbif-integrated-publishing-toolkit
https://pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf
https://www.gbif.org/sites/default/files/documents/gbif_best_practice_guide_data_publishing_by_local_governments_en_v1.pdf

LIVING ATLASES

As GBIF nodes, one of our goals is to highlight our publishers and their data. To achieve this, the Atlas of Living Australia (ALA) developed a huge open source platform with several modules re-usable by other organizations. Since 2013, the community around this tool has organized technical workshops to present ALA modules to other institutions that wanted to implement it, to improve already existing national data portals and to learn from each other's achievements.

In order to help new users but also to keep on assisting the experienced ones, we try to arrange at least one workshop per year around specific modules of the platform (e.g. species module, spatial portal, etc.). These meetings are really motivating for new users because they can actually realise that, with some developments, they will be able to have a powerful tool running. And at the same time, these training activities are also very productive for partners with ALA portals already running as they have the opportunity to share doubts and ideas, solve technical issues, get assistance from the ALA developers' team and -in consequence - move forward on the developments of their national data portals. Furthermore, during these technical trainings, we get ideas from other projects and allow the nodes to keep on working significantly on their own.

Thanks to the previous meetings and other engagements arranged around this topic, at least 11 data portals using ALA technology have been released in production since 2014. Other are still under development (some of them are already listed on the new GBIF web page). Katia Cezón from GBIF Spain created a Carto map showing countries with ALA installation or interest in the ALA infrastructure.

On the website, you will find documentation and information about participants and the community but also ALA tools. You will be able to access the materials from past events but also news about future events and different ways to directly talk with members of the community (through HipChat or mailing list).

You will also be able to see the community in action because we are a group of developers that love to work together and improve tools to facilitate a free and open access to biodiversity data.

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data. The authors intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Additional information:

https://living-atlases.gbif.org/

FAIR DATA

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

Additional information:

https://www.go-fair.org/fair-principles/