Selecting and preparing data for the Atlas of European Mammals, 2nd edition.

EMMA2 Steering Group

Version 1, 21 October 2020

Selecting and preparing data for the Atlas of European Mammals, 2nd edition.

This document builds on, and supersedes 'Guidelines For Data Format For The Atlas Of European Mammals 2024 V1'. The Atlas of European Mammals (AEM) is the planned output of the EMMA2 (European Mammals on Maps 2) project.

Qualifying records and the role of the national coordinator

The Atlas of European Mammals (AEM), will be based on verified biological records that meet the basic requirements of species, location, date and observer. Our ambition is that every record (dot) appearing in the Atlas can be traced back to one or more verified biological records, probably recorded at higher resolution than used in the Atlas (a nominally 50km x 50km CGRS[1] cell). These original records should be held in a system that is accessible to *bona fide* researchers, with appropriate safeguards for sensitive records (rare species, vulnerable habitats or other reasons), though we hope that the great majority of records will be freely accessible at a higher resolution than in the Atlas, for example in national or regional atlases. Unpublished, privately-owned or inaccessible records should be avoided, as they will not be available for verification or for further study and violate the principle of FAIR data. If individuals hold large collections of unpublished records, they should be strongly encouraged to make them accessible somewhere, either by adding them to a national or regional database or by publishing them in a scientific paper.

For countries with large databases of records (national or regional), it is likely that each CGRS cell will contain multiple records, at least for common and frequently-recorded species. In other cases, a record in the Atlas may be traceable back to a single observation held in a database in the country of origin.

Responsibility for verifying the accuracy of the data submitted for the AEM rests with the national coordinator. Where records are held in a national or regional database, the task of the coordinator will be to confirm with the owners or managers of the database that a well-defined system is in place to verify records before they are added to it. In other cases, coordinators may wish to verify individual records, particularly for rare species, or check that collections of records meet the requirements of having a defined form of verification. It is the coordinator's decision whether to include records in the dataset for the AEM.

The correct identification of specimens and the validation of records is a crucial step in ensuring that the databases that contribute to the AEM contain only high-quality data. This can be more of a challenge as 'citizen-science' contributions to databases become more common and database managers try to ensure that data contributions are accurate.

For some sympatric species, genetic analysis will be needed. Similarly, acoustic bat species identification should follow established procedures (Ahlen & Baagoe 1999, Rydell et al. 2017, Russo et al. 2017).

[1] Chorological Grid Reference System, based on the 50-km UTM lines and MGRS zone boundaries defined in the technical reports of the U.S. Defence Mapping Agency.

The way in which specimens were identified and the name of the identifier and verifier should normally be included in primary databases, though they will not be needed in the data submission to the AEM.

Species and species names

Taxonomy and nomenclature are in a constant state of flux, so we expect some changes during the lifetime of the EMMA2 project. The taxonomic sub-group of the Steering Group has developed a species list for the area covered by the atlas and will update this as necessary. In many cases, name changes will be the result of nomenclatural changes (e.g. the genus *Myodes* has recently reverted to *Clethrionomys*), so the number of species remains the same. In a few cases, taxonomic studies, particularly using genetics, may reveal new species, often by splitting existing ones, thus increasing the number of species in the atlas (it is currently 270). We are committed to mapping only valid species, so all records should refer to a species on our current list. We will not map *sensu lato* species or pseudo-species. Recognised sub-species will not be mapped separately, but may be referred to in the species accounts.

Although the AEM will not map marine species, we will include Pinnipeds as these spend time hauled-out ashore. For most species, records of seals very close to the coast (within one of the EMMA2 CGRS cells that contains land) could be included, as it seems likely that these animals will also haul-out close by. The ice-breeding species *Cystophora cristata* is problematic, as it is not recorded on land, so we may map this using an indicative area, as in the first Atlas.

A species list spreadsheet, containing the binomial name of the species, will be maintained on the European-mammals.org website and can be used to check the accuracy of data submissions.

Date classes for the atlas

The new atlas will have three data-classes of records, extending the two classes used in the first atlas. These are:

1. **Species present 1999-2023**. Based on data collected since 1st January 1999, modified by published data on extinctions to give the current distribution of the species. Positive records since 1999 are only omitted from the map where more recent detailed surveys have failed to detect the species. We would expect such cases to be rare.

2. **Species present 1970-1998**. Based on data collected between 1st January 1970 and 31st December 1998, modified by published data on extinctions to give the current status of the species. Positive records between 1970 and 1998 are only omitted from the map where more recent detailed surveys have failed to detect the species.

3. **Presence presumed**. Based on data collected before 1970, but where there is no evidence to suggest the species has become extinct. As records in this date-class are now more than 50 years old, we hope that it will rarely be necessary to include such old records.

Datasets submitted by each country for the atlas should only include the most recent record for each species/CGRS cell combination, so that each species/CGRS combination should appear only once in each national dataset. Including older data might allow some exploration of changes in distribution over time, but such data are not consistently available across the area of the EMMA2 project and so we have chosen to include only the most recent data.

As with the first edition, the intention is to give a picture of the current distribution of the species, so in a few cases where there have been extensive surveys for a species, coordinators should consider removing records from the map where the surveys have shown with a high degree of probability that the species is no longer present in an area. Where there have not been extensive surveys, so that uncertainty remains, records should be included on the map, as extinction in an area has not been proven.

Where there are two or more records of the same species in the same CGRS cell (where a cell is shared by two or more countries) only the most recent record will appear on the map.

Treatment of migratory or dispersing species and vagrants

Several species of mammals in Europe exhibit long-distance migratory behaviour with predictable two-way seasonal movements, usually between summer and winter ranges. The most migratory bat species are *Pipistrellus nathusii* , *Nyctalus noctula*, *N. leisleri* and *Vespertilio murinus*. For migratory species, all records should be included, as these movements, including stopping-off points, are part of the normal behaviour of the species and the accompanying species account can refer to the seasonal range and movements of the species.

A few non-volant terrestrial species, most notably *Canis lupus* and *Canis aureus* show long-distance dispersal movements, which may perhaps be precursors to the expansion of their range. This seems to be a typical behaviour of these species and individuals can move from their place of birth to appear far from their established range, perhaps being recorded in more than one country. Similarly, some pinnipeds, such as *Odobenus rosmarus* may appear as vagrants, appearing far from their normal range, though with no indication that this is a typical behaviour for the species. Although the seasonal and two-way nature of their movements distinguishes migratory species from dispersing species and vagrants, there is a less clear distinction between dispersers and vagrants.

With the possibility of using colour to indicate the status of records, we are interested to see if we can collect sufficient data to indicate which records are believed to be of dispersing or vagrant individuals. For the national dataset, we now offer the possibility of marking whether a species' record in a particular CGRS cell is, in the opinion of the national coordinator, a record of a dispersing or vagrant individual. This can be indicated by placing a '1' in the column headed 'Vagrant'.

Species not established in the wild

Escaped pets, domestic animals and alien invasive species not established in the wild are excluded from the Atlas and so do not appear on our species list. Also excluded are feral goats, semi-domesticated reindeer herds, and single records of vagrants of species with distributional ranges outside the atlas area.

EMMA2 Data Structure

Data for the EMMA2 project can conveniently be submitted in a spreadsheet, CSV file or Access database (or similar). The data submission should consist of a single table/file, with one record for each species/CGRS cell combination and must contain the following fields (please use the field names in brackets, without the quotation marks):

Species name ('Species')

This should be the binomial name of the species, in accordance with the EMMA2 species list. The species list should be used to check that species names in the data submission are correctly spelled

and that no synonyms are used. This can be done by downloading the EMMA2 species list and using lookup functions to check for errors. Contact the data coordinator if you need help with this.

CGRS cell name ('CGRS')

The CGRS cells used in the Atlas have a 6-figure name, consisting of a two-number UTM zone, a three-letter 100km reference and a figure between 1 and 4 indicating the quadrant of the 100km square, for example 31VDD1. Valid cell names are available for each country in the GIS resources section of our website. We have tried to include all cells which contain any land, but it is possible that some small areas have been excluded because of the resolution of the base map used to generate the lists of cells. If you believe a cell is missing from the list, please contact the data coordinator.

Country code ('Country')

Some CGRS cells are shared by more than one country (a small number of cells are shared by 4 countries), so the country code is needed to ensure that each record is unique using the Species+CGRS+Country combination. We use the ISO-3166 Alpha-2 codes, which can be found here: https://www.iso.org/obp/ui/#search (although we expect you will know your own country code!).

Date class ('Status')

The date class of the record as a single digit, either 1, 2 or 3, as defined above. Please include records with the most recent date-class (1) wherever possible. Do not include more than one date class for each Species+CGRS+Country combination. Where countries share a cell and submit records for the same species with different date-classes, we will plot the most recent date-class on the map.

Year ('Year')

The year of the most recent observation. We will use this to check that the date-class is correct and it may also be used for an analysis using different date-classes.

Vagrant or disperser ('Vagrant')

Place a '1' in this column if you believe that the record of the species for this cell is a dispersing or vagrant individual, otherwise enter zero or leave blank. We are considering ways in which such records can be identified in the Atlas.

Source ('Source')

As stated previously, our intention is that each dot in the Atlas can be traced back to the original observation(s). This field should contain a reference to the source of the data. Please use an alphanumeric reference of the two-letter country code followed by a number. For countries that have a national database and have generated the whole EMMA2 data submission by querying that database, the same reference can be used for each record. In other cases, the reference can be to a regional database or, where data are not held centrally, a literature source. Data coordinators should keep a table of these references (with a copy accompanying the data submission), so they can respond to enquiries about particular records.

Submitter ('Submitter')

Please include your name, so we are clear about who is responsible for the submission.

Date Submitted ('Submitted')

Please include the date of submission, in the form dd/mm/yyyy. This will help keep track of any changes that are made to datasets, as new records are added.

Summary of fields for data submission

The table shows the column headings for the data submission followed by several example rows showing the contents of each field. The source contains a reference to the UK national database.

Species	CGRS	Country	Status	Year	Vagrant	Source	Submitter	Submitted
Erinaceus europaeus	31UCV2	GB	1	2001		GB001	A Jones	02/07/2020
Erinaceus europaeus	31UDT1	GB	2	1995		GB001	A Jones	02/07/2020
Erinaceus europaeus	31UDU2	GB	1	2009		GB001	A Jones	02/07/2020
Sorex araneus	29UPB3	GB	2	1996		GB001	A Jones	02/07/2020
Sorex araneus	29VPC3	GB	1	2004		GB001	A Jones	02/07/2020
Sorex araneus	29VPC4	GB	2	1968		GB001	A Jones	02/07/2020
Sorex araneus	29VPD1	GB	1	2020		GB001	A Jones	02/07/2020
Sorex araneus	29VPD3	GB	1	2011		GB001	A Jones	02/07/2020

References

Ahlen, I. and Baagøe, H.J., 1999. Use of ultrasound detectors for bat studies in Europe: experiences from field identification, surveys, and monitoring. *Acta Chiropterologica*, *1*(2), pp.137-150.

Rydell, J., Nyman, S., Eklöf, J., Jones, G. and Russo, D., 2017. Testing the performances of automated identification of bat echolocation calls: A request for prudence. *Ecological Indicators*, *78*, pp.416-420.

Russo, D., Ancillotto, L. and Jones, G., 2018. Bats are still not birds in the digital era: echolocation call variation and why it matters for bat species identification. *Canadian Journal of Zoology*, *96*(2), pp.63-78.

Appendix 1. Converting data to the CGRS grid.

The CGRS grid used for the Atlas of European Mammals (as well as other European biological atlases) was developed for the Atlas Florae Europaeae (AFE) from the Military Grid Reference System, which in turn is based on the UTM grid. The UTM system divides the world into 60 zones, each 6 degrees in latitude, and the MGRS/CGRS define a consistent way of merging the zones to give a system with global coverage. A more detailed explanation can be found at http://www.luomus.fi/en/utm-mgrs-atlas-florae-europaeae. The cell size used for the atlas is nominally 50km, but because of the need to merge UTM zones, individual cells can range in width from 33 to 62 km.

Location data for biological observations can be stored in a variety of formats and resolutions. Common examples are the lat/long (WGS84) coordinates generated by GPS systems, coordinates on the UTM zonal system or coordinates recorded on a national or regional mapping system (e.g. KFME for central Europe, OSGB for Great Britain). To prepare the dataset for the Atlas of European Mammals, individual observations must be allocated to a single cell in the CGRS.

The approach to allocating observational data to a cell in the CGRS depends on the accuracy of the observation. All geographic locations are a reference to an imaginary square, with the given location at the bottom left-hand (south-west) corner. The size of this square is given by the resolution of the observation, as shown in the table below for some lat/long locations. A similar principle applies to other systems

Quoted location (degrees, decimal minutes)	**Resolution**	**Side of square**
50°N, 10°E	1 degree	111.120km
50°30'N, 10° 30'E	1 minute	1852m
50°30.5N, 10° 30.5'E	0.1 minute	185.2m
50°30.55'N, 10° 30.55'E	0.01 minute	18.52m

When converting observations from lat/long (or any other system) to the CGRS, clearly the lower (coarser) the resolution, the greater the probability of the 'square of uncertainty' intersecting with more than one CGRS cell. At resolutions of 0.01 minute or better ('square of uncertainty' smaller than 18.52m), such uncertainty in converting to a 50km grid can probably be ignored and the observation treated as a point in a GIS. This should cover the great majority of observations in national and regional databases.

In the case of observations with lower resolution, or where greater accuracy is desired, there are several ways to do the conversion.

- Treat the observation as a point in a GIS. This will be placed at the south-west corner of the square of uncertainty; for example, 50°N will be treated by the GIS as 50.00000°N.
- Add half the resolution to the observation (so 50°N becomes 50.5000°N, 10°E becomes 10.5000°E). This will place the GIS point in the centre of the square of uncertainty.
- Use the AFE spreadsheet (available on our website) to determine the correct CGRS cell and check whether all 4 corners of the square of uncertainty fall within the same CGRS cell. If they do not, you will need to allocate the observation to one CGRS cell – generally the one which includes most of the square of uncertainty.

Two methods for converting lat/long (WGS84) data to the CGRS

1 Using a GIS

Data can be converted by plotting all lat/long observations as points in a layer in the GIS, adding a layer with the CGRS cells (from the GIS resources we supply) and then performing a geographic query to select the cells from the CGRS layer that contain points from the observations layer. An SQL query can then be used to select the most recent observation for each Species+CGRS combination.

If you have a large dataset (> 1 000 000 observations), it might be worth dividing this into two or more smaller datasets to speed up processing.

A detailed example using QGIS:

Begin a new empty project in QGIS, using a WGS84 crs (EPSG:4326)

Add species data from a csv file, with at least the fields Species, Latitude, Longitude, Year. Other fields can be included and can be carried through to the output.

From the toolbox select the algorithm *Vector creation>Create points layer from table*

Set x field to Longitude, y field to Latitude and run the algorithm. This will plot all the points in the map window and produce a table with the default name 'Points from table'.

[Optional] Create a spatial index on this new table using the algorithm *Vector general> Create spatial index*. This takes some time but will speed up subsequent operations.

Add a vector layer to the project, using the shapefile [your country code]_cgrs.shp (or cgrsV2 in some cases) – so for the UK it is GBR_cgrsV2.shp and that's what we'll use for this example. This vector layer must be in the WGS84 crs (EPSG:4326).

Join the tables to add the CGRS cell to each point by selecting *Vector general>Join attributes by location* from the toolbox.

Set:
base layer = 'Points from table'
Join layer = 'GBR_cgrsV2'
Geometric predicate = Within [include 'Intersects' if your data has observations that fall exactly on a CGRS boundary]
Fields to add = CGRSName

This will produce an output table with the default name 'Joined layer'. If there are no errors, this will contain the same number of rows as the input table, with the correct CGRS cell name appended to each row. You can check for errors by sorting the table by CGRS and looking for null values.

The final step is to eliminate duplicates by selecting the most recent record for each species in each CGRS cell. This can be done within QGIS by using a query. From the toolbox select *Vector general> Execute SQL*

Set:
Additional input datasources = Joined layer
SQL query = SELECT Species, CGRSName AS CGRS, max(Year) AS Year from 'Joined layer' GROUP BY Species, CGRSName (the AS commands set the output fieldnames to those needed for the data submission)
Geometry type = No geometry

Running this query should produce an output table with the default name' SQL Output'. This should have far fewer rows than the input layer, as there is only one entry for each species+ CGRS combination. If you have included other fields in the input, you will need to add the fieldnames to the query so they are carried through to the output (and you may also need to specify an aggregate function to specify which value is selected).

The only remaining task is to add the Status column to the output table and set its value. This can be done by creating a new integer column and filling it using an expression like 'CASE WHEN "Year" < 1970 THEN 3 WHEN "Year" <1999 THEN 2 ELSE 1 END'

The final table can then be exported to a CSV file by right-clicking on it in the layers pane and selecting *export>Save features as* then setting the output file format as Comma Separated Value (CSV).

2 Using the Excel spreadsheet

Download the spreadsheet from our website (EMMA2>Documents>EMMA2 GIS resources>Excel CGRS conversion functions); this will only be visible to logged-in users. The file is named afegrid2020.xlsm.

Save and open the spreadsheet and create a new blank tab for your data.

Import your data, which must have Species, Year, Latitude and Longitude fields. Lat and Long should be WGS84. If your import file (csv works well) has column headings, Excel will format the data as a table, which is helpful. There are then two steps to the conversion.

Add a new column and insert the first conversion function *User Defined>LL2GR*. Set LatCell to your Latitude column and LongCell to your Longitude column; set Area to 4, to convert to MGRS references. If your data are formatted as a table, Excel will now fill the column with the formula and you will now have the MGRS grid reference in this column.

Add another new column and insert the conversion function *User Defined>MGRS2CGRS* and set MGRS to the first column you created. The column will be filled with the CGRS cell names and the conversion is complete.

To save the GCRS name column, you will need to copy and paste the values to a new column, save the spreadsheet with a new name, open it and delete all the tabs except the one with your data.

Further work is needed to select unique Species+CGRS combinations. This can easily be done by importing the spreadsheet into Access (or similar) and using an SQL query (as above).