8 Dipterists Forum Technical Guide

Disseminating Biodiversity Data 3: Recording in Europe

Darwyn P. Sumner

Organiser of the UK Stilt & Stalk Fly Recording Scheme (Micropezids & Tanypezids)

Series C, Issue 8, Version 1 (Mar 2020)

Keywords

Collection, collation, management, dissemination, validation, verification, UKSI, NBN Atlas, GBIF, Global Biodiversity Information Facility, iRecord, iMatch, iNaturalist, Global Biodiversity Gateway, GBG, National Biodiversity Network, Darwin Core, metadata

Summary

A guide to the processes involved in the formatting of records for the purpose of uploading to GBIF.

Audience

Part 1: Recorders, data managers, verifiers, photographers, collectors, surveyors

Part 2: Recording Scheme organisers, county recorders, dataset compilers, validators, verifiers

Part 3: Non-UK recorders, European recorders, researchers, biogeographers, phenologists, distribution modellers

Introduction

Part 3 of this series moves on from UK recording to an examination of methods which might be applied to recording in Europe (and elsewhere in the world) and to submission to GBIF (via the UK's "participating" organisation, NBN).

The methodology for submission to the UK's NBN Atlas is well documented and in frequent use. It utilises Darwin Core which may be simply emulated using a spreadsheet.

Many UK recorders however have records of species occurrences from outside the UK, they have just one recourse to upload those records to the international GBG, GBIF, which is via iNaturalist. That system has a two disadvantages, firstly it requires that an image is available of the recorded taxon and secondly their verification system requires that a second person confirms one's identification; without a collaborator that confirmation rarely occurs.

Though demand for non-UK recording may not be high, there are many individual recorders who wish to upload foreign collections and Recording Schemes which have broadened their remit to study taxa over a wider geographical area. This ranges from Europe to the world as evidenced from various Scratchpad and identification sites (iNaturalist, Diptera.info etc.)

This study arises from one such Recording Scheme (Micropezids & Tanypezids) which

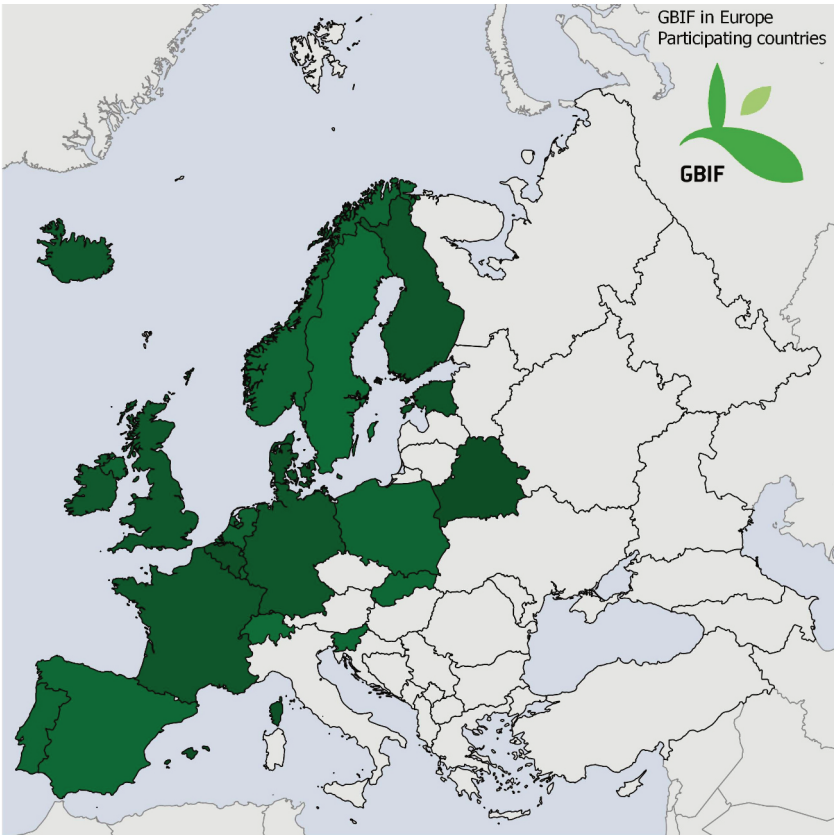
has extended its area of study to Europe and begun to examine European distribution (see http://www.micropezids.myspecies.info/)

In an attempt to be in a position to submit records directly to GBIF, this scheme applied for GBIF endorsement via the UK's NBN and entered into discussions regarding methodology (see https://forums.nbn.org.uk/viewtopic.php?id=7213).

Background

There are mechanisms for the amateur naturalist to make casual records. In the UK these are several, ranging from online systems such as iSpot and iRecord, sharing through biological recording applications such as Recorder 6 and Mapmate to regional and national organisations which have systems to upload to the UK's GBG, the NBN Atlas.

Outside the UK however, the systems may not be so clear to the user. The target GBG is of course the international GBIF but mechanisms within European countries to upload species occurrences there vary from country to country. From systems similar to the UK's to none at all.



GBIF's participating countries. For the current list see https://www.gbif.org/the-gbif-network

GBIF operate through a formal system of endorsing an organisation. Many of these are museums intent on publishing their collection data, though some also support naturalist's endeavours (e.g. Slovenija). Others are systems intended to gather records from a range of contributors, the mechanisms of these will be unclear unless one speaks the particular language. Others use a Bulletin Board/Galerie system to post, verify and upload (Le Monde des Insectes). Austria has recently acquired a system similar to the UK's which was adapted from an Australian system. To use many of them it would be necessary to be able to speak the language.

All this leaves the naturalist who has travelled abroad and collected species occurrences with a confusion of mechanisms (or none) when it comes to publishing their data.

The author, seeking to publish species occurrences to GBIF in support of a recording scheme with a European remit, successfully applied to become a GBIF endorsed organisation, for which support from an existing country-based endorsed organisation was mandatory - the UK's NBN in this case. The NBN have shown interest in helping this scheme upload non-UK datasets to GBIF, they are familiar with the mechanisms as they pass datasets regularly to GBIF. Their last request to this scheme was "*send us a list and we'll see what we can do*".

Which is where this paper originated. That list cannot be the same structure as that used in the UK as once one leaves UK shores the OSGR geospatial coordinate system no longer applies, nor do our lists of taxa. Accordingly this paper follows on from an examination of techniques for collation and dissemination of UK datasets (parts 1 & 2) and compares the structure and requirements of datasets appropriate to European recording in order to arrive at an optimal structure for such a list and a mechanism for submitting to GBIF.

Techniques discussed here mainly involve spreadsheets as this is the optimum format with which to examine and detail the Darwin Core fields and the model proposed. The methods and proposals were tested during the writing of this guide, revisions were made to the text as misconceptions and practicalities were encountered.

Data models & standards

Various data models have been devised or adapted for use in biological recording. These range from systems built around museum record-level standards such as Spectrum, through specifically designed systems which form the core of Recorder 6 (Copp, 2000) and Mapmate to TDWG's Darwin Core which was devised by taxonomists.

They all facilitate the construction of relational databases in which tables containing different classes of information are linked together; each has its strengths and weaknesses and each draws upon published standards.

Researchers wishing to access published occurrences have devised a range of models and analytical solutions. Indeed there is an entire journal devoted to this kind of work (Bioinformatics.) Though their research may be quite complex, they are universally in favour of Darwin Core as a model and in GBIF as a source, though sometines critical.

Darwin Core is a body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data. Darwin Core is maintained by TDWG (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). There is an excellent summary by OBIS (Ocean Biogeographic Information System) though it doesn't deal with land-based taxa.

TDWG also provide free definitive world maps for GIS work.

European recording

1. iNaturalist

This technique works well for non-UK records providing you have both an image and a collaborator to confirm your identifications. It is not really necessary for UK recording unless it is part of a specific team project as we already have good alternative systems in place which gain more attention from UK verifiers. It can upload species occurrences to GBIF fairly quickly under the right circumstances. For those seeking to upload species occurrences from personal overseas collections it would be necessary to photograph each of your specimens. Then wait perhaps a very long time for the material to be identified if the taxa are not popular.

Though there is a spreadsheet (csv) upload facility, this system would be difficult to use for very large numbers of species occurrences. It works well for projects involving invasive species where monitoring involves many countries, an example being the *Drosophila suzukii* project (Chandler, 2017)

Details of the Flickr/iNaturalist methodology, aimed at potential European recorders for a Recording Scheme, are to be found in Micropezid & Tanypezid Newsletter 2

2. GBIF

Resources are available for recording species occurrences etc.. The following details spreadsheet templates which GBIF make available

Remsen, D. P., Döring, M., & Robertson, T. (2011). GBIF Spreadsheet Templates User Guide, version 1.0. (April), 0–3. Retrieved from http://links.gbif.org/dwca-spreadsheet-processor-guide

In the main the systems which GBIF use to collect datasets are aimed at larger institutions rather than individuals or organised schemes. Large museums who wish to record their collections and GBGs such as NBN and the many other European country-based initatives.

The small organiser or individual may consequently find themselves at a disadvantage when wishing to upload non-UK records to GBIF.

Designing a non-UK database

This would have to be a dual-purpose database, suitable for your own research practises (phenology, distribution mapping) and submission of records to GBIF.

To an extent, existing biological recording applications are able to cope, for example Recorder 6 will accept Lat/Long coordinates and does an excellent job of organising Locations - the familiar UK/Vice County hierarchy can be extended to European Country/Province. It lacks the facility to incorporate non-UK taxa and its mapping utility does not extend beyond the UK.

Ideally one's personal database would be in the form of a relational database (e.g. MS Access) but spreadsheets with lookup lists will suffice.

Assembling a practical system from Darwin Core components is like having a huge box of Lego bricks tipped onto the floor and being told to make something out of that.

Issues with recording in Europe

Perhaps accustomed to methods of recording in the UK, many recorders will find some aspects of recording abroad unfamiliar. So as to comply with GBIF requirements and take into account these aspects of recording abroad and develop a personal database that facilitates research, the following considerations should be taken into account:

1. UK species index (UKSI) no longer applies, GBIF uses a different list
2. Geospatial coordinates must be Lat/Long
3. Reference to species occurrences in published papers takes on an important aspect, data from such papers may never have been extracted.
4. Online image postings are good sources & need to be referenced (with links)
5. Links/references to online research sites (Scratchpads) need to be specified in the researcher's database (though not necessarily in the GBIF upload) in order to facilitate research projects such as distribution mapping.

The following examines the above issues in more detail:

A. Species indexes

A.1. UKSI

The UK Species Index is available in a variety of formats. It can be usefully consulted online via the Taxon-match tool and the NBN Record Cleaner. They are also present in Biological Recording applications such as Recorder 6 and Mapmate via continual updates. Checklists are also available through Recording Schemes and websites (e.g. Dipterists Forum) and regional organisations such as Local Environmental Records Centres possess lists which can be adapted for use in lookup tables and the like for keen recorders to use in spreadsheets and other systems.

These are not suitable for overseas species though. Whilst new species arriving in the UK are speedily added to the UKSI, there is no intention to extend this list abroad. The number of species approximately doubles when the range is extended just to Europe.

A.2. GBIF

Unsurprisingly, one has to construct one's own. There's a very good guide to doing

this in the documentation "GBIF Checklist data"

Download the spreadsheet template there and follow GBIF's guide. The fields are readily completed by finding the specific taxon on the GBIF site then adding the taxon name as specified and the taxonID (which can be the url of the taxon's page)

For the casual recorder this needn't be onerous as one would only need to add the taxa recorded. For the Recording Scheme with a remit wider than the UK this is a longer task. If you are a Scratchpad manager you may have a similar list already if you've provided a link to GBIF distribution on your species pages. One cannot simply specify a group and request a download, even in the UK that would be less than straightforward outside biological recording applications.

B. Geospatial coordinates

For recording geospatial coordinates outside the UK it is necessary to specify Latitude and Longitude coordinates. Indeed for a research dataset required for European distribution mapping, all the UK OSGR geospatial coordinates need to be converted to Lat/Long.

There are a number of utilities which can determine Lat/Long and several others which can perform conversions. Their use depend upon the circumstances:

Single location: GPS (if available), Google Earth, iMatch (for image)

Expeditions of multiple sites: GPS, Google Earth kml file list

Single OSGR conversion: UK Grid reference finder ; BGS Coordinate converter ;

Multiple conversions: UK GR Batch converter ;

The Darwin Core downloaded fields are *decimalLatitude* and *decimalLongitude* . This is also the format used in Recorder 6. There are online utilities which will convert from DMS to decimal.

Additional DWC fields:

countryCode

locality

stateProvince

coordinateUncertaintyInMeters

C. Dates

There is a problem in expressing dates in Darwin Core. Their fields are *eventDate* backed up by *day*, *month* & *year*. The problem arises in cases where only imprecise dates are known, a published paper may specify only the year or "VI to VII, 1935" for example. No matter what method is attempted when uploading data, the returned simple downloads invariably cause this to default to 1st January in that year. This has a detrimental effect when constructing phenology charts, one's only recourse is to omit all those with that specific 1st January date.

This topic is discussed at https://forums.nbn.org.uk/viewtopic.php?id=7056 where it would appear there are additional date fields (startDate & endDate) which are now present in NBN Atlas downloads but not in downloads from GBIF that were submitted by NBN to GBIF.

Time series are also awkward to handle. In the case of malaise traps and the like it may

be prudent to record the occurrence to both the start and end dates.
A *datePrecision* field would be handy for phenology - if there is one.

D. Resources

If you are extracting species occurrence information from published sources such as papers and online postings then it is no longer sufficient merely to specify *recordedBy*, *identifiedBy* and other related fields, we need somehow to specify the source of the information.

D.1. Published papers

There are a number of methods and applications which are able to deal with published papers. None are entirely satisfactory for a taxonomist wishing to link a taxon to a specific paper but it is possible to get close.

Using citation managers

Though Recorder 6 does have something built into the model it is not able to make that taxon/paper link. The most satisfactory way of achieving this is to start with Elsevier's free Mendeley and begin your collection of papers. Straight away you have an organised set of papers and a convenient means of reading them via its built-in pdf reader. To link these to taxa requires some effort though. Adding taxon names as text strings to the Tags field is necessary; subsequently one can "Filter by My Tags" on individual taxa.

For an amateur researcher this is a useful and effective means of managing libraries. Many Open Access papers are readily available.

To incorporate this system into a data model however is more complex. The objective would be to build a table containing at least the citation text and a persistant url. One is able to copy and paste the citation text into one field of a table but the persistant reference is harder to achieve. DOIs or ISBNs are rarely available. One solution to this is the NHM's Scratchpad system. Using BibTex one can transfer publication details (including those taxon tags) from Mendeley to a Scratchpad. This achieves a permanent and persistant url link to a specific paper, for example

ID	Citation	Author	Media	URL
	Roháček, J., & Barták, M. (1990). Micropezidae (Diptera) of Czechoslovakia. Casopis Slezskeho Zemskeho Muzea Opava (A), 39, 97–111.	Roháček	Publication ISSN: 0323-0627	http://micropezids.myspecies.info/node/30

Complex models detailing all aspects of references have been devised (e.g. Copp, 2004) but the above simple table should suffice for the purposes of preparing a dataset for submission to GBIF. Some of the occurrence records one wishes to submit will be present in published papers, accordingly the table is needed in order to indicate "Source".

The Darwin Core fields which map to the above are *bibliographicCitation, recordedBy* & *references* Additionally the Darwin Core field *type* can be used to specify "Text" (as distinct from "Image")

D.2. Referencing online sources

Naturalist photographers post images to a variety of online sites. Some are for the purposes of identification, a few of these have inbuilt systems for passing occurrences to GBIF (iRecord, Le monde des insectes.) Many however are simply for show (e.g. Flickr)

and though the rigour of the four “W”s may not be so good, they can be a rich source for species occurrences.

ID	Description	Author	Media	URL
	European stalk-eyed fly (Sphyracephala europaea)	Nikola Rahmé	Image Flickr	https://www.flickr.com/photos/eurythyrea/5123843851
	Neria octoannulata (Strobl, 1899)	Manuel Sanmartín	Image www.biodiversidadvirtual.org	https://www.biodiversidadvirtual.org/insectarium/Neria-octoannulata-(Strobl-1899)-img694211.html

D.3. Personal collections

D.3.1. Specimens

D.3.2. Photographs

D.4. Museum collections

Downloads

A. Deduced from downloads

A.1. NBN Atlas

A.2. GBIF

There are 3 options for dataset downloads from GBIF: Simple, Darwin Core Archive and Species list, on selecting the green Occurrences button on a species page. My example is *Tanypeza longimana* as I have an iNaturalist record amongst them and it's only 283 records.

Choosing the **Darwin Core Archive** option returns a WinZip file containing files in .xml and .txt formats. To open the list use Excel's Import Wizard on the .txt files (which are tab-delimited.) The occurrence.txt file results in a spreadsheet with columns from A to IE (239 fields) which seems to cover all possible Darwin Core fields.

There may be a simple application which can open this complex of files, it appears to be structured along the lines of a relational database. GBIF's Darwin Core Archive Assistant (at http://tools.gbif.org/dwca-assistant/) may be such an application.

DOWNLOAD pdf and app, p11 deals with spreadsheet methodology

The **Simple** download contains a single .csv file, rename this to [filename].txt so that Excel uses the Import Wizard on opening (otherwise Excel thinks it's already formatted.) This one only runs from Columns A to AX (50 fields). These are as follows:

GBIF Simple download structure

Column	Field	Column	Field
A	gbifID	AD	eventDate
B	datasetKey	AE	day
C	occurrenceID	AF	month
D to K	various taxonomic hierarchy fields	AG	year
L	taxonRank	AH	taxonKey
M	scientificName	AI	speciesKey
N	verbatimScientificName	AJ	basisOfRecord
O	verbatimScientificNameAuthorship	AK	institutionCode
P	countryCode	AL	collectionCode
Q	locality	AM	catalogNumber
R	stateProvince	AN	recordNumber
S	occurrenceStatus	AO	identifiedBy
T	individualCount	AP	dateIdentified
U	publishingOrgKey	AQ	license
V	decimalLatitude	AR	rightsHolder
W	decimalLongitude	AS	recordedBy
X	coordinateUncertaintyInMeters	AT	typeStatus
Y	coordinatePrecision	AU	establishmentMeans
Z	elevation	AV	lastInterpreted
AA	elevationAccuracy	AW	mediaType
AB	depth	AX	issue
AC	depthAccuracy		

The meaning of each of the above fields is detailed in the **Darwin Core quick reference guide**. These are arranged in classes rather than individually detailed terms so hyperlinks to terms in the above table are not really feasible (see scientificName).

A.2.1. Interpretation of colour coding

The fields which we examined in Spreadsheet Method 1 (the UK's system) are marked in RED above (mandatory ones dark red, others pale red.)

Geospatial coordinates are in Lat/Long (presumably when NBN Atlas datasets were submitted to GBIF our OSGB was converted), these are indicated in ORANGE.

If your purpose is simply to examine downloaded records then the columns marked in YELLOW can be safely deleted (except under specialist circumstances) to improve readibility of the spreadsheet.

A.2.2. Usage by organisations

There is a good deal of flexibility in the way that the fields may be used, whilst many are constrained to a particular purpose, few of them are constrained to a particular format.

The following are examples of how iNaturalist and NBN use the fields:

iNaturalist

Very few fields are submitted to iNaturalist, just the image author's name, the geospatial coordinates (lat/long) and the date. Some of this is contained within the image's metadata when the Flickr system is used. Eventually the species name is confirmed by a verifier.

In the downloaded file, GBIF has recorded the following:

- occurrenceID — the url to the record
- species — without author
- scientificName — with author
- countryCode — the two letter ISO code for the country
- locality — not recorded (despite its being in the Flickr post)
- stateProvince — computed by GBIF from the geospatial coordinates
- publishingOrgKey — identifier for iNaturalist as an organisation
- decimalLatitude — provided in Flickr/iNaturalist posting
- decimalLongitude — provided in Flickr/iNaturalist posting
- coordinateUncertaintyInMeters — not recorded
- eventDate — provided in Flickr/iNaturalist posting
- day — calculated from eventDate
- month — calculated from eventDate
- year — calculated from eventDate
- taxonKey — the taxon's code using GBIF's preferred taxonomy
- speciesKey — ditto
- basisOfRecord — the default "HUMAN_OBSERVATION"
- institutionCode — "iNaturalist"
- collectionCode — "Observations"
- catalogNumber — the digits end of the url to the record, evidently a running number
- recordNumber — not assigned
- identifiedBy — not present
- dateIdentified — absent, though this can be found on the iNaturalist posting
- license — "CC-BY-NC" as adopted by iNaturalist for all records
- rightsHolder — "Darwyn Sumner"
- recordedBy — "Darwyn Sumner"
- lastInterpreted — "2020-03-05T21:29:21.199Z" ? code to permit version change
- mediaType — not specified (though other datasets specify "StillImage")
- issue — "GEODETIC_DATUM_ASSUMED_WGS84"

The two missing values are an issue. The precision of the Latitude and Longitude values are known, they were calculated by GPS and thus accurate to 1m but there is no figure for coordinateUncertaintyInMeters, though this level of accuracy cannot be assumed for all iNaturalist postings. This absence makes species distribution modelling awkward, one cannot assess associations with climate, soil type etc. unless this figure is low enough (Mücher, 2010). That figure could be narrowed down if the locality were specified but that isn't recorded either.

The most precise level of GIS mapping achievable from iNaturalist derived records of this nature is to *stateProvince*, for example a map of French départements. This data quality issue has been discussed by several authors, notably Gaiji, 2013 and Otegui, 2016

NBN Atlas

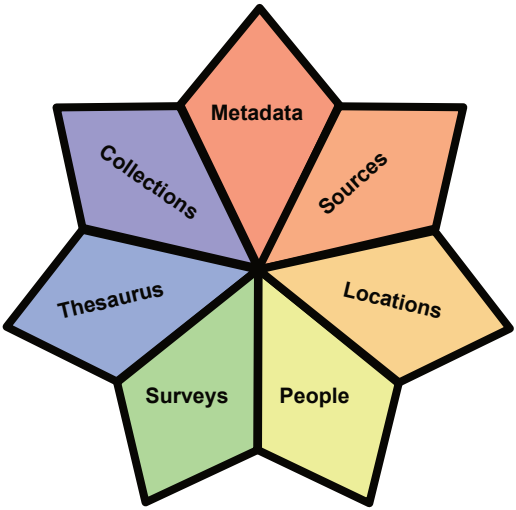
The following example was transferred to GBIF from the NBN Atlas to which it was originally uploaded from a Recording Scheme (DOI https://doi.org/10.15468/mwjnku)

• occurrenceID	a complex GUI
• species	without author
• scientificName	with author
• countryCode	the two letter ISO code for the country
• locality	present only when specified
• stateProvince	computed by GBIF from the geospatial coordinates “England”
• publishingOrgKey	identifier for NBN as an organisation
• decimalLatitude	converted from OSGR
• decimalLongitude	converted from OSGR
• coordinateUncertaintyInMeters	always present (calculated from OSGR)
• eventDate	provided by recorder
• day	calculated from eventDate
• month	calculated from eventDate
• year	calculated from eventDate
• taxonKey	the taxon’s code using GBIF’s preferred taxonomy
• speciesKey	ditto
• basisOfRecord	the default “HUMAN_OBSERVATION”
• institutionCode	“Dipterists Forum”
• collectionCode	The GUI complex identifying both the organisation submitting records to NBN and something else
• catalogNumber	not recorded
• recordNumber	GUI assigned by recorder (Recorder 6 system)
• identifiedBy	not present
• dateIdentified	not present
• license	“CC-BY-NC” as specified by many Recording Schemes
• rightsHolder	not specified
• recordedBy	Recorder name (when known)
• lastInterpreted	“2020-03-05T21:29:21.199Z” ? code to permit version change
• mediaType	not specified (though other datasets specify “StillImage”)
• issue	blank

GBIFis Darwin Core spreadsheet templates

Available at https://www.gbif.org/news/82852/new-darwin-core-spreadsheet-templates-simplify-data-preparation-and-publishing

Data Model



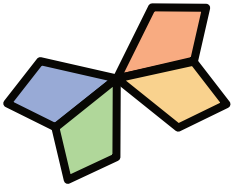
This simplified diagram depicts the **modules** of an idealised Data Model as may be used in a biological recording application, be this a complex application, researcher's stored data or simple spreadsheet system. Copp (2004) details its implementation in earlier versions of Recorder (Recorder 2000) where he added the Collections module. The Metadata module is suggested by the author as an important aspect of archiving.

The key module is the Surveys module which contains all the occurrences. Other modules are more or less independent of that. For example the Thesaurus module, which contains all the dictionaries such as taxa (species dictionary) can be populated independently, as can the People and Locations modules.

The GBIF spreadsheet system specifies this:

Their currency is the flat Survey database, they want the Taxon Dictionary element of the Thesaurus module to check compatibility with theirs.

To manage a personal or research system you require a minimum of this:



Below we discuss how to connect these modules.

Target audience

For this guide we are considering just two types of user. The individual wishing to upload to GBIF records of occurrences from abroad and the Recording Scheme organiser (or other researcher) wishing both to upload to GBIF and maintain a comprehensive dataset in order to support various outputs and analysis relating to taxa in that scheme over a wide geographical range. naturalist & researcher respectively:

Naturalist: The main problem appears to be the route via which GBIF allows such datasets to be submitted. It must be through an accredited organisation. Some countries adhere closely to the principle of dealing only with native records, others may be persuaded to work with non-native data (Spain, UK). A more universal capture system would be desirable, along the lines of the UK's iRecord but effective verification and the personnel required to manage such a system are barriers.

Each dataset may be considered to arise from a single survey. The Locations and Taxon dictionaries may be reused for each survey submitted.

Doable from spreadsheets only - if a little tricky

Researcher: Again with the same issues of finding a "sponsor", the objectives of the researcher are twofold. Firstly to upload collected occurrences to GBIF and secondly to maintain a dataset of all collected records so as to perform various geospatial and other analyses. The researcher is thus both a contributor and a user.

Much has been written from the user perspective regarding publicly available species occurrence datasets ...

The researcher is obliged to carefully manage a wide range of datasets, GBIF downloads for example need to be excluded from resubmission, as do datasets known to be scheduled for GBIF upload such as records downloaded from a countries GBG known to carry out regular uploads.

Modules

The value of the above model and its modules concept is that we can use the definitions to help design our DwC spreadsheets and establish a logical workflow.

A. Sources

The sources module provides a means of tracking the origin of any information in the database, ranging from a whole dataset to individual species occurrences.

Field	Values	Description
sourceID	see GUI	key
type	“StillImage”, “Text”, “PhysicalObject”, “Download”	This will serve to sort source types. **StillImage** will only be referenced once but **Text** may be referenced many times.
bibliographicCitation		Full citation or title of posting
references	URL only	This appears to be the only term available for a url which works for every relevant item in this table.
institutionCode	custodian (author)	Person or organisation custodian of the resource
datasetName	Name of collection of specimens/objects	e.g. NHM (Sumner loan), C.Palmer (Sumner), Sumner
datasetID	Identifier for that collection (not individual records)	[The code assigned to specimen or photograph within the above dataset e.g. “236/18” (specimens) or “ADS_026417” (photographs) belongs in another table - perhaps Occurrence but this has not been implemented in this study]
collectionCode	acquired dataset	use this to name datasets submitted to Recording Scheme or downloads acquired. Include date and/or version to help sort.
modified	date	The most recent date when the resource was changed
language	Russian, French	No DwC terms available
georeferencedBy		People who determined the georeference: “Darwyn Sumner | Andrew Halstead”
georeferencedDate		“2018-06-23”
georeferenceProtocol		Published protocol e.g. “Buchar, 1982”
georeferenceVerificationStatus		Resolution of the georeferencing task (e.g. “low resolution” with respect to a published paper where precise coordinates were difficult to determine from location names alone
georeferenceSources		“Garmin GPS” or “IGN 203 | IGN 204”
georeferenceRemarks		Comment on method used e.g. IGN map, GPS, GoogleEarth, iMatch map, calculated from Buchar grid, country checklist

- There is a need for a comments field, such as *bibliographyComment* where one could record issues regarding a publications, such as the fact that it did not include dates or that it was “Country only”

The fields in **bold** are the only ones from this table which are returned in a simple GBIF download. There’s no means to get a url (*references*) to appear in a simple download. Only iNaturalist manage that because all their records have a url and they can use that as their *occurrenceID*

There are a number of other terms which could be used in the above table, mostly the

concern of museum-level custodians to help them differentiate datasets etc., but they are beyond our scope. Not all of the above fields are transferred to the table which needs to be submitted to GBIF, some are just for the user's convenience. As a spreadsheet the above can be populated completely independently.

Dealing with downloads (e.g. GBIF, NBN Atlas etc.):

- Enter "Download" in *type*
- Enter term such as "GBIF", "NBN", "iRecord" etc. into *institutionCode*
- Enter [Filename] to *datasetName*
- In *collectionCode* enter a term which further identifies the content of that downloaded dataset e.g. "Tanypeza longimana V2" - the version is important to help tou manage later downloads.
- Enter date of download in *modified*
- Enter "Error" or similar in *basisOfRecord*

Accordingly accidental resubmission of records is prevented by *type* (use as filter) and *basisOfRecord*

Use of *datasetName* and *datasetID*

The use of the Sources table varies with the type of material. In the situation where records are being extracted from published papers, each Source record should be a single paper, the *bibliographicCitation*, *reference* and *institutionCode* (=authors) fields deal effectively with these (together with all the blue fields which are used to detail the researcher's work on interpreting the occurrences from each paper. The number of Sources records is likely to be small as each one represents a lot of work. The *datasetName* and *datasetID* fields are redundant in this situation.

In a situation where the Sources table is being used to record online records (e.g. Flickr postings) there are likely to be a lot of records in the Sources table, one per species occurrence. Again *bibliographicCitation*, is the title of the online post, *reference* is the url and *institutionCode* the author's identifier. The only case where *datasetName* would be required is if the Sources record is part of someone's named collection (e.g. filenames of image files)

When the Sources records are being used to record collections then again the number of records should be few, *datasetName* is the name of the collection and *datasetID* its identifier (to another table not present in this account).

The two terms belong to a record of specimens, individual records therein and should not be in the Sources table, it can be placed into the Occurrences table from where it can be linked back to its *datasetName* through the *sourceID* term in the Events table.

B. Locations

This has the potential for a high degree of complexity but for our purposes it is sufficient just to develop a structure which provides sufficient information to suit the desired outputs. See Hill (2010)

Field	Values	Description
locationID	see GUI	key (user-assigned)
location		precise name: “Buron”
verbatimLocality		imprecise directions: “Alania, 10 km SE of Alagir” or code to grid tile: “Buchar 5764”
decimalLatitude		in decimal degrees
decimalLongitude		in decimal degrees
geodeticDatum	“EPSG:4326”	spatial reference system used “WGS84” or “WGS84 Web Mercator” for Google Earth derived coordinates.
coordinateUncertaintyIn Meters	=accuracy	Radius of the smallest circle around the given position containing the whole location. About 30m for 1 second
coordinatePrecision	DwC term (in fractions of a degree)	Good GPS is 0.00001, phones and cameras may be less precise, suggest 0.00005, for a GoogleEarth estimate use 0.0001. Nearest second is 0.000278, minute 0.01667 and degree 1.0. Somewhat academic as it duplicates the purpose behind the above, leave it in the table but don’t populate it unless you discover a need.
higherGeography	DwC term	e.g. “Europe | Pan-Europe”
country	DwC term	TDWG Level 3 (country)
countryCode	DwC term	TDWG Level 3 (country)
stateProvince	DwC term	TDWG Level 4
gridReference	System adopted by Country	UK: facilitates the calculation of *coordinateUncertaintyInMeters*
continent	ISO name = DwC term	TDWG Level 1 (continent) For the most part this can be omitted
continentTDWG1Code	ISO code	TDWG Level 1 (continent)
regionTDWG2Name	ISO name	TDWG Level 2 (region)
regionTDWG2Code	ISO code	TDWG Level 2 (region)
countryTDWG3Code	ISO 3 letter code	TDWG Level 3 (country)
provinceTDWG4Name		TDWG Level 4
provinceTDWG4Code		TDWG Level 4
locationRemarks		It’s no substitute for a proper Biotope dictionary but users may wish to record habitat type here
locationAccordingTo		Publication - could be gazetteer or paper

- As the population of this list may require a good deal of research there is also a need for some kind of flagging system which would indicate progression to completeness.

For GIS purposes it is essential that you are able to match the above table to the fields in TDWG’s published maps. These are described in Brummit (2001) and the SHP files are downloadable from the GitHub link on the TDWG page . Brummit presents the map regions in tiers, Level 1 are continents, Level 2 regions and Level 3 countries. To select, say, Europe from your records it is necessary to include tiers higher than *country* and to split, say, Czechoslovakia into Czech Republic and Slovakia you will need TDWG level 4. Fields indicated in blue are not in GBIF’s “allowable” list so take care to exclude them from

uploads and reassign them from downloads. The *higherGeography* DwC term will help you sort the above table by "Europe", Palaearctic" etc. to some extent but isn't present in the TDWG maps.

> At some points TDWG's Darwin Core terms differ from TDWG's GIS terms, for example for continent they suggest using terms from a controlled vocabulary such as the Getty Thesaurus of Geographic Names and yet they've published their own list in Brummit which is not mentioned in Darwin Core documentatin.

Accuracy & precision.

The *coordinateUncertaintyInMeters* and *coordinatePrecision* terms are a curiosity. Given that **accuracy** is defined as "*the degree of closeness of a measurement to that quantity's true value*" or "*the degree or closeness to which the information on a map matches the values in the real world*" then in geospatial terms one would expect this to be expressed by some sort of distance value. In other words if one travelled to the precise spot indicated by the lat/long coordinates then the actual spot where the specimen was taken is within a radius defined by some sort of **accuracy** term. If I went to Oslo to look for a species and read that this was to an **accuracy** of 30,000m then I wouldn't expect to find it, I'd be in the city centre. If I went to a site that was specified with an **accuracy** of 3m then I would be in the exact spot, I'd be standing next to the malaise trap. The question of accuracy arises because the original recorder did not specify the coordinates to a sufficiently high degree of accuracy, the researcher reading simply "Oslo" from a paper has to make a best estimate of the coordinates and to specify an accuracy figure: circa 30km in this case,.

The **positional accuracy** term used in DwC is *coordinateUncertaintyInMeters* expressed as *the diameter of a circle you would need to draw around the provided coordinates in order to ensure that you had included the original surveyor's exact capture site*. This is all that is needed to filter out inaccurate coordinates when constructing GIS dot maps. In practise when estimating coordinates from papers using Google Earth there is no need to try to be too accurate, figures of 3, 30, 300, 3000 and 30000 should be sufficient, with NULL used when only the country or large region is specified. For situations where grid tiles are specified (e.g. OSGR) the tile size determines the value. The DwC guide also gives the values which may be used based upon coordinates provided in DMS format.

For data models Mucher (2010) indicates that the LANMAP data is to 10km accuracy so data collected to that level (3000m and below) may be used to construct models, note also that Franklin (2009) suggests that as few as 50 occurrences are sufficient to make an informative model.

Dealing with imprecise locations:

One of the desired outputs would be country checklists. Such checklists may be found in published papers or books which give no geospatial coordinates more accurate than just "country". There are also online postings which provide coordinates no better than "province". Though the field *coordinateUncertaintyInMeters* could be calculated, raised to an enormous value, that seems to be an unfeasible approach. The output for GIS work needs to be interrogatable in a simple manner so as to exclude such records from the production of dot maps. Accordingly, assign a value NULL to *coordinateUncertaintyInMeters* and ensure that both *countryCode* and *stateProvince* are completed. Both the latter must map to the Brummit TDWG codes (Levels 3 & 4) which specify map tiles.

Given the above, the term *coordinatePrecision* is of doubtful value. **Precision** is the degree to which such measurements are reproducible or how exact was the description of data. There is no intention of reproducing the measurement and our accuracy figure (above) fulfils practical needs. Ignore *coordinatePrecision* .

Though the GBIF guide squeezes the above into their Events table, it is advisable to deal with Locations separately. Even at the least complex level these may take some scrutiny to work out all the necessary details. The task can become quite complex, for example Roháček, 1990 where authors used traditional grids to provide coordinates. To convert these to Lat/Long it was necessary to construct the Buchar grid in GIS and interpret.

The conversion from DMS to DD can be carried out using a variety of methods:

1. Google Earth. Set the coordinate display to DD, paste the DMS string into the search panel, right-click the resulting pin and select properties. The DD format is displayed.
2. Excel formula, it needs a bit of text string wrangling and if using copied text beware of idiosyncracies such as the use of two single quotes in place of a double quote. Take care to ensure that locations west of the Greenwich meridian, which will be suffixed “W” are entered as negative values, similarly any Latitudes marked “S”.
3. Alternatively there is a macro method described by Geospatiality, (which is susceptible to user idiosyncracies), using macros does mean you have to save the spreadsheet as a macro-enabled file (xlsm).

C. People

Unless you are dealing with large numbers of records from a wide variety of sources, a module which deals with people details is not paramount. The likelihood is that one will be dealing with one’s own records, single online postings by others, papers with a very limited list of collectors or expeditions with a restricted number of surveyors. A very simple table of *code* (for quick entry), *avatar* (username on website postings) and *name* will suffice.

D. Surveys

In Copp, 2004 the contents of this module are stated to comprise Surveys, Survey Events and Samples (the actual species occurrences) In Recorder 6 these must be entered sequentially. Working with spreadsheets however, you can add them in fewer steps.

In Recorder 6 the sequence is as follows:

- Survey
- Event
- Sample
- Occurrence

In OBIS they are termed Classes and are separated out as follows:

- Event
- Occurrence

The GBIF spreadsheets address:

- Sampling Event
- Occurrence

Recorder 6 has the facility to maintain a number of Surveys (there is even a tier higher than this which permits Surveys to be categorised), whilst OBIS and GBIF assume that one is recording just one specific Survey. These latter two also combine Event and Sample into one table.

D.1. Survey

Keeping a table of this is optional, on the assumption that one is either recording one's own overseas collections or maintaining a set of records within a restricted taxonomic group, the most one will need is a group of files related to a single survey. A good Metadata record using the tools detailed below would be of value.

D.2. Event

One sampling event is attached to many occurrences. Lots of moths in a moth trap. As soon as the sampling parameters change, new date, new location, different technique then one has a new sampling event. This is the SamplingEvent spreadsheet from the GBIF downloads. Mandatory fields in red, allowable in pink (see GBIF's Darwin Core Event)

Field	Values	Description
eventID	key	see GUI
parentEventID		used to link to a higher tier of event
eventDate		
samplingProtocol		Method: UV light trap, malaise, sweep net
startDayOfYear	1 to 366	Ordinal day of the year on which the Event started
endDayOfYear	1 to 366	Ordinal day of the year on which the Event ended
year		
month		
day		
habitat		
fieldNotes		
fieldNumber		helps with expedition organiser's administration
eventRemarks		
locationID	key (Locations)	All the non-blue fields in the Locations module Grey ones are optional, they are your record
country, countryCode, stateProvince, locality verbatimLocality, locationAccordingTo decimalLatitude, decimalLongitude geodeticDatum coordinateUncertaintyInMeters coordinatePrecision … etc.		
sourceID	key (Sources)	All fields in the Sources module
type		
bibliographicCitation		
references		
institutionID	institutionCode	
datasetName		
datasetID		
collectionCode	=ownerInstitutionCode	
modified		
license	CC-BY-NC	

Were it not for the fact that the Locations and Sources data have to be squeezed into this table it would look fairly straightforward. The Sources part remains fixed for each expedition or field visit.

Take the example of a typical expedition (Dipterists Forum Field Week), there's one event for each site visited on each day. Mostly everyone has used the same sweep netting method but if there's a photographer then that's another event and if a malaise trap was set up then that's yet another.

The above could be just one line (record) in one table and that's true if it's just you in one place. For an expedition, every time the day or location changes one has to add another record. So with 15 people averaging 4 sites per day over 7 days that's 420 records in this

table. Make that 421 if you set up an Event record to detail the entire expedition and use the *parentEventID* to link all the others to it; that's a fairly rudimentary way of implementing some sort of Survey system. Formal European expeditions can be handled this way too, for example Mark Pollet's expedition to Corsica in 2019 where all sorts of collection methods were used.

D.3. Occurrence

To each of the above Events we now begin to add the occurrences. We've got the "Where" and "When" of the four "W"s in the Events, we just need to add the "What" and the "Who". With the spreadsheet methodology we're reaching the limit of what a single line of a record can cope with, that Events spreadsheet is quite big now.

For each of those Events we need to add one to perhaps hundreds of occurrences, accordingly we begin to move into the realms of relational databases, in other words we simply specify the *eventID* for each occurrence we record. That's how one would suppose that the GBIF dual spreadsheet system would work (but see below), it's the same with more sophisticated models too, though in their case the Locations and Sources modules are also linked in to the Events modules via keys in the manner of relational databases.

The occurrence component lies at the heart of the model. For those who regularly maintain spreadsheet lists, this is their familiar one-by-one list of observations. All the other modules/tables simply connect to this one, lookup lists if you will.

Fit for purpose

At this point the GBIF spreadsheet templates depart from our requirements. The templates are presented as a "guide" but they do not go the whole way.

Compare for example, the structure of our download at (page 11) to the GBIF Occurrence spreadsheet template, there are fields that we expect, to indicate *identifiedBy*, *recordedBy* and other fields associated with sex, stage and so on. Important data which should be included in a list of occurrences. None of theses fields are present in any of the GBIF spreadsheet templates or in the list of additional recommended fields. The following provide terms we need:

- https://dwc.tdwg.org/terms/#identification
- https://dwc.tdwg.org/terms/#occurrence

Also absent is a key to the Events table (*eventID*) so we're unable to link the GBIF spreadsheet ocurrences table to all our work on Events.

Accordingly we have to construct our own as follows:

The following is based on the Occurrence spreadsheet from the GBIF downloads. Added to it are a number of terms related to identification, stage and sex.

Field	Values	Description
occurrenceID	see GUI	key
eventID	key	link to events table
basisOfRecord	Allowed values: PreservedSpecimen, HumanObservation, Occurrence, Stillimage	Use “occurrence” if no other applies.
eventDate		from Events table (via *eventID*)
taxonNumberID	key	from our Taxon dictionary
scientificName		
kingdom		
taxonRank		
decimalLatitude		from Events table (via *eventID*)
decimalLongitude		from Events table (via *eventID*)
geodeticDatum		from Events table (via *eventID*)
countryCode		from Events table (via *eventID*)
individualCount		
organismQuantity		
organismQuantityType		
catalogNumber		specimen or object number (image filename)
recordedBy	DwC term	
identifiedBy	DwC term	
dateIdentified	DwC term	
identificationVerificationStatus	DwC term	verified
identificationRemarks	DwC term	
sex	DwC term	
lifestage	DwC term	
preparations	DwC term	A list (concatenated and separated) of preparations and preservation methods for a specimen
occurrenceRemarks	DwC term	

The mandatory fields are in bold.

E. Thesaurus

These are mostly dictionaries, the only one required being your **Taxon dictionary**. Devise a spreadsheet (or use my Darwin Core template) and populate it yourself. The following structure is suggested:

Field	Values	Description
taxonNumberID	see GUI	key **4516270**
taxonID	DwC term	**GBIF url** It is important to use the full url for this field as this may be required by Scratchpad users to direct readers (via website) to GBIF distribution maps **https://www.gbif.org/species/4516270**
scientificName	DwC term	**Neria cibaria (Linnaeus, 1761)** - must have the author if known
taxonRank	DwC term	**species** - usually
UKSI	NBN term	Include the UK's species index where present, easily generated via Taxon Toolkit
urlNBNAtlas	NBN term	url to NBN Atlas link to this taxon
organismID	DwC term	url to Scratchpad taxon or other favoured site e.g. **http://micropezids.myspecies.info/taxonomy/term/20**
vernacularName	DwC term	**Common Strider** - DwC only allows one, awkward for common names in other languages.
taxonRemarks		your notes, use to indicate which species to refer to should a record be *nomen dubium*
acceptedNameUsage	DwC terms similar in usage to Scratchpad	**Neria cibaria** - the simple binomial
acceptedNameUsageID		
parentNameUsage		**Neria** - consequently that term (the Genus) needs to be in your list as a separate record.
parentNameUsageID		
scientificNameAuthorship	DwC term	**(Linnaeus, 1761)**
sourceID	Sources table	A link back to the Sources table - to record *bibliographicCitation* of the original description
bibliographicCitation	DwC term	Full citation text (best used for taxonomic revisions)
order	DwC term	**Diptera**
family	DwC term	**Micropezidae**
genus	DwC term	**Neria**
subgenus	DwC term	
specificEpithet	DwC term	**cibaria**
taxonomicStatus	DwC term	accepted, sedis incertae, nomen dubium, species inquirenda, synonym etc.

Recording Schemes will already have these in some format, especially if they have previously populated an NHM Scratchpad. Those wishing to record species occurrences to GBIF standards should simply add taxa to this table as they happen across them. The details are easily found, just search GBIF for the taxon then use that URL as the *taxonID*. That url cannot easily be used as a key in a relational database or vlookup though, so a GUI (*taxonNumberID*) which conforms to database usage must be incorporated.

Generate this from the url by deleting the first 28 characters of the url string (https://www.gbif.org/species/1643441 becomes 1643441) There will be other taxon GUIs in other lists but the GBIF ones are relatively simple to find and make the list compatible with GBIF.

Explanation of coloured groupings:

Red are mandatory fields required by GBIF in their spreadsheet. The *taxonID* field permits both the simple number (from GBIF's site) or the GBIF url. However you need both, the numerical ID for links to other tables (e.g. as in relational databases) and the GBIF url for providing the distribution map link on a website. Hence the *taxonNumberID* which can be obtained via spreadsheet fomula from the url.

Blue is a block of fields which allow the production of an hierarchy. It's also present in the spreadsheet one first submits to NHM's Scratchpad site in order to build up their hierarchical checklist on your site. Simply put it consists of assigning an ID to every taxon in your list (Family downwards) no matter what the level, then assigning a *parentID* to each of those, so if it's a species then the *parentID* is its Genus. This is detailed in the Scratchpad documentation. Use your own numbering system for this, no need for elaborate GUIs. It's pretty much all you need to start a Scratchpad off, once you have the taxa loaded up you just add things you know to the pages created,

Green fields can be for the avid collector of the original papers which described the taxon. Perhaps more importantly, if your list deviates from that of GBIF then cite the paper which revised the taxonomy. Just the citation text, the sourceID lets you link back to an item in your Sources table if you wish.

If you are expanding your interest from UK to Europe then expect to double the number of taxa, further afield and that factor rises enormously. There's no need to add taxa which you are not going to use.

As a spreadsheet in its own right it can prove to be rather useful. I would caution against attempts to attempt large groups, my Recording Scheme list has only 120 entries (for 90 species) and covers Europe. Compiling it was quite demanding, three or four times that size would be quite labour intensive although one could cut corners by omitting most of it except the first few terms. A relatively simplified version would be feasible for all UK Recording Schemes and of value to anyone recording or maintaining their own spreadsheets. It's also shareable and thus can facilitate collaboration, to see an example download the European Micropezid & Tanypezid checklist.

It is quite safe to sort this list in Excel, provided all columns are selected as the unique codes are not user-generated. The danger with any lists with user-generated GUIs is that one would lose track of the sequence and be in danger of extending from the wrong final value when adding items.

F. Collections

No complex system is required unless you are a museum or have extensive collections from abroad.

Other modules can cope with collections details though perhaps inelegantly,

G. Metadata

Metadata is structured information about a resource.

In practical terms some sort of Digital Management or Electronic Document Management System will suffice to manage information about your resources, iMatch will help with some of this. Add to this good backup schedules and archive routines. A spreadsheet detailing your work is advisable and Microsoft's OneNote may help you keep a record.

Loss of data richness / Dumbing down data

There are many cases of the loss of data richness, stories abound. There are examples in Part 2 of this guide; spreadsheets submitted as part of an expedition varied from participant to participant but some contained detail fields which did not lend themselves readily to the required NBN structure and so were omitted.

Many other cases have been reported, ranging from the conversion of one database format to another in order to suit a particular range of outputs and services (e.g. Recorder 6 to in-house systems) or idiosyncracies (e.g. converting full OSGR to "tetrads".)

Important data is lost by these processes.

Researchers find this frustrating however, there is much value in that original data and they wish to access it.

Original data should be archived and made accessible.

*find the thread about trends in providing original data - was it Science? publishers are demanding this - esp medical papers - did I do a bulletin item

This is one of the stated purpose of Electronic Document Management Systems, the archiving of original material.

This is a badly neglected area in biological recording, an issue which museums take seriously and address through metadata applications, having devised their own set of standards (see SPECTRUM)

G.1. Metadata structure

Metadata, or structured information, may be applied to just about any level of a collection. From the lowest level such as a specimen or a photograph (e.g. EXIF) up through the hierachy of specimen collections (notes about the donator of the collection) or image collection (e.g. iMatch's "Events" feature) right through to the very top such as details about the museum or the labelling you put on your secure backup hard drive that you keep in the shed for security against fire and theft.

The resource is therefore hierarchical in nature and each level of the hierarchy deserves a label, or many labels. Clearly as one moves up through the hierarchy, metadata cannot be transferred from each member of the lower level and so higher levels must summarise lower levels.

In the case of the above data model the lowest level we need consider is that of Sources, everything below that is pretty much detailed in the lower levels of the hierachy.

Each dataset we assemble comprises a collection of many Sources. That dataset is going to be uploaded to a GBG on a page such as Dipterists Forum NBN data or GBIF and users will wish to know what it's all about, the scope of this particular collection.

G.1.1. Considerations of metadata structures

Formal metadata standards concern the set of descriptors that a particular kind of resource requires in order to describe it adequately, together with rules about the way these descriptors must be expressed

NBN systems

Historically these were derived from a sub-set of Dublin Core, GEMINI, which was a set of geospatial descriptors. Previous incarnations of the UK's Atlas used a cut-down version of the geospatial GEMINI, the current version still retains that geospatial legacy but is more a set of descriptors suitable for a web page.

It focusses on the requirements of the NBN Atlas and some parts of it are unsuitable for scopes broader than the UK.

GBIF systems

Though GBIF have their own system for recording dataset metadata they indicate "*depositors ... must overcome a learning curve in understanding how to map their data to the Darwin Core standard*", thus ruling this out for those without an unfeasibly wide range of skills.

It suggests that there might be DwC terms however - if they could be discovered.

Dublin Core, W3C, ERMS, e-GMS, GEMINI & SPECTRUM

A return to first principles perhaps? The above contain powerful metadata standards and many good ideas.

Actual web page usage

An examination of pages providing descriptions and summaries of other uploaded datasets on web pages is informative. A UK NBN Atlas dataset summary compared to a Swedish GBIF dataset summary.

Personal or organisational usage

Dictated by the need for an organisation to manage and describe their own datasets. Issues such as workflow and "content management" need to be taken into account as do internal processes such as archiving, backup and sources of original material.

Metadata is just labelling, it's not as hard as some would like you to believe. If you've ever labelled a bag of Christmas presents for your family then you've done metadata.

G.1.2. Proposed format

The following set of metadata elements are examined with respect to their suitability for adequately describing uploaded datasets. The core set of 20 terms is taken from ERMS, a public sector standard (National Archives, UK). Added to this are a further set of 6 terms taken from e-GMS (the UK e-Government ERMS metadata standard). This results in a set of standard metadata elements which are widely acceptable and compatible with standards across a broad range of sectors (well, at least in the UK) This set also happens to cover all the metadata elements from the geospatial GEMINI which happened to be used in the NBN's previous implementation (Gateway), current implementation (Atlas) and Recorder's systems.

Finally, some terms required for our purposes are added.

Element	Purpose
Identifier	unique identifier for an aggregation of records
Title	The title given to the record, folder or class
Subject	Keywords or phrases describing the subject content of the resource
Description	Freetext description of the resource
Creator	The person responsible for the content of the resource
Date	Date submitted
Addressee	The person (or persons) to whom the record was addressed
Type	Type of collection of records
Relation	Identifies instances where a record has a direct relationship with that of another
Aggregation	The unit of measurement used to define where in the information hierarchy any records management action is carried out
Language	The language of the intellectual content of the resource
Location	Physical location
Rights	CC-BY-NC
Disposal	not relevant as the records will never be discarded
Digital Signature	not appropriate
Preservation	Preservation management processes that have been employed to facilitate its survival across technical platforms
Mandate	Purpose for which information is processed
Format	The software format of electronic components forming constituent parts of records
Function	The function of organisation that produced the record[s]
Coverage	The extent or scope of the content of the record
Accessibility	the resource's availability and usability to specific groups
Audience	A category of user for whom the resource is intended
Contributor	An entity responsible for making contributions to the content of the resource
Publisher	An entity responsible for making the resource available
Source	A reference to a resource from which the present resource is derived
Status	The extent to which it has been developed or completed, i.e. is it a first draft, final draft or completed draft? Include version number

Element	Description
Identifier	ADSEMTPb B 2020 V1
Title	Sumner, D.P. 2020. European Micropezids & Tanypezids, Published Papers Part B Occurrence Dataset
Subject	European Micropezids & Tanypezids (Diptera), occurrences extracted from several published papers (Part B)
Description	Diptera: Nerioidea & Diopsoide. Published inventories of several European countries
Creator	Darwyn Sumner
Date	2020-05
Addressee	GBIF
Type	Species occurrences. Diptera: Nerioidea & Diopsoidea
Relation	One of a series of compilations
Aggregation	Published European papers, Part B (31 papers)
Language	English
Location	NHM servers, http://www.micropezids.myspecies.info/node/358
Rights	CC-BY-NC
Preservation	Uploaded to servers of UK Natural History Museum
Mandate	Methods and verification by authors, validation & geographic analyses by D. Sumner
Format	Spreadsheet (Excel 2010) & pdf as zipped files
Function	Recording Scheme (Europe. Diptera: Nerioidea & Diopsoidea)
Coverage	Norway, Estonia, Czech Republic, Slovakia, Switzerland, Portugal, Lithuania, Romania, Netherlands, France, Greece, Italy, Sweden
Accessibility	Open Access
Audience	Researchers, biogeographers, taxonomists
Contributor	European Micropezids & Tanypezids http://www.micropezids.myspecies.info/
Publisher	European Micropezids & Tanypezids
Source	Detailed at http://www.micropezids.myspecies.info/node/358
Status	First draft Version 1

Which appears to work satisfactorily. There is perhaps a little duplication but the purpose of each element is clear (mostly). It omits the temporal coverage of the NBN metadata list but that's a feature of the records which were extracted rather than a particular recording period (e.g. expedition). It substitutes the more rational "**Creator**" for the odd separate "contact" fields. The methods, verifications and validations (in NBN's *qualityControl* are all on a per-paper basis and do not really belong at this high tier of metadata, the general statement in **Mandate** covers this (strictly speaking mandate is "why" it's being done but if you take that as being self-evident then "how" it was done is a fair utilisation of this element.) Researchers will wish to consult the list of papers, it's in the Sources table. Too big a list to be included in a simple summary as above and so the list is best placed on an online page (see **Source** above) or other link to a download

of a simple tabulated list.

The combination of **Location** and **Format** fulfil the requirements of providing access to the raw data, the data deposit. **Function** is equivalent to NBN's *purpose*.

The **Identifier** is your filename (see Globally Unique Identifiers) which is also reflected in the *sources* table and the **Title** can be formatted as your *bibliographicCitation*.

> **Identification** is the code name of the dataset which this metadata is describing e.g. "ADSEMTPb 2020 V1" which should be present in the *collectionCode* of the Sources module. It is constructed from your GUI code + a session indicator + year + version. The session indicator is to allow you to separate your work into batches. Each compendium may be quite time consuming, you are unlikely to cover every conceivable source of records in one session. This allows you to complete one batch, submit it to GBIF and begin work on a second compendium (collectionCode ADSEMTPc C 2021 V1). Note that the GUI changes so as not to duplicate. The Version number allows you to amend elements of a dataset though how this might be reflected in GBIF uploads is unclear.

The above maps well to NBN metadata (see Part 2)

In GBIF metadata we see the following:

- Description (our **Subject**) which can be somewhat wordy in some GBIF datasets. This should all be on one's Scratchpad site (**Contributor**) so no need to repeat it on GBIF.
- Purpose (our **Function**),
- Temporal scope (either nothing or "historic", the occurences are too widely spread over time as to be of any value)
- Geographic scope (**Coverage**)
- Methodology (**Mandate**)
- Additional info (**Description**)
- Contacts (**Creator** + **Contributor**)
- Data description (just languages for metadata and data seemingly, use **Language**)

Somehow a logo might be submitted to GBIF, be sure to include one in the bundle.

The only issues remaining therefore are our own management and record of work carried out. For example DOIs or urls relating to the dataset once it has been uploaded. These are detailed in Part 2

Globally Unique Identifiers

Each of the above tables begins with an identifier. These are fundamental to the construction of relationships between the modules. In simple spreadsheet lookup tables they serve to uniquely identify each item in a list. In relational databases they serve to tie together all the modules and dictionary items.

Though used mostly for internal reference, they are occasionally seen in some form by users of uploaded data and can serve to identify the source of a record.

The NBN system for assigning these is an 8 + 8 coding system where the first 8 digits uniquely identifies the user and the last 8 digits the individual item. The first 8 are assigned by JNCC from a single maintained list - usually when applying for a Recorder 6 license. The last 8 are generated within the application, along the lines of sequential numbers but with systems that incorporate letters too.

Guralnick, 2015 observes: *Biodiversity data is being digitized and made available online at a rapidly increasing rate but current practices typically do not preserve linkages between these data, which impedes interoperation, provenance tracking, and assembly of larger datasets*.

Tthe following propesed system which conforms to the NBN's 16 digit limit is proposed:

- Identifier for organisation - 6 digits — ADSEMT
- Project indicator - 2 digits — Pa
- Context or table identifier - 2 digits — Lc (location) or Oc (occurrence)
- Sequential numbering - 6 digits — 1 million

Six letters specify the individual and organisation carrying out the work.

The next two are available for that organisation's project. For manageability the work has to be carried out on a per small project basis and the inclusion of this not only serves to identify the project (one spreadsheet perhaps) but allows a second project to be begun without reference to the sequence in the previous project, i.e. the second project can begin at number 1 all over again without fear of duplicating GUIs.

Each table has 2 digits to indicate its module's name

Finally, sequential numbers which can extend up to a million, it's unlikely that an organisation would approach that limit, if it does then it can be augmented with letters. Copying coded values like this down a column in a spreadsheet automatically increments the numeric values.

In a downloaded occurrence record one might therefore see:

ADSEMTPaOc000001

Not only fulfilling Guralnick's criteria for interoperation, provenance tracking and assembly of larger datasets but also providing (in the first 8 digits) the basis of a filename and *collectionCode* in the Source module and metadata.

Note that some GUIs may be preordained, such as GBIF's taxa or the UKSI. Others may deviate from this structure in order to facilitate data extraction via codes in a published paper. These latter are all internal to one's database however, not global.

Reference

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., … Page, R. D. M. (2015). Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. Zookeys, 154, 133–154. https://doi.org/10.3897/zookeys.494.9352

Building a model

In order of increasing complexity, the following methods are available:

a. Simple single separate spreadsheets
b. Spreadsheets with additional sheets, each one representing a module
c. Use of Vlookup to cross-reference information across those sheets
d. Use of Excel's Data Model to establish relationships (Excel 2013 or later)
e. Microsoft Access (or e.g. Open Office) relational database
f. SQL and other more advanced methods (not addressed here)

Methods a & b are in common use for those submitting datasets to a recording scheme or organised expedition. They fall short of achieving our full GBIF objective but are very worthwhile. Some more advanced Excel users will also have used method c in order to ensure that they can pick out the correctly spelled taxon name or location details. The remaining methods require more expensive applications or more advanced skills

Our objective is a table that looks pretty much like what we would expect to get back from a GBIF download. Perhaps with a few additions that would help with any jobs that we would like to do (e.g. our own mapping via GIS)

A. Simple spreadsheets

Some of these are of considerable value even if you progress no further than just keeping lists.

The **Sources** table will help track published sources and online image postings (pending ID enquiries of your own perhaps.)

The **Locations** module is invaluable. In this one you can organise locations lists of visits abroad and put together the definitive list of sites from your stored GPS data (e.g. Garmin's Basecamp) or GoogleEarth estimates. If you are analysing a published paper then it's advisable to use a spreadsheet table before transferring to anything else as these investigations can be quite laborius.

The **Taxon dictionary** is invaluable too, if it's a coherent taxonomic group then a careful structured spreadsheet allows you to build a list of taxa over a wider geographic range - and this spreadsheet format is how you develop the list required for a Scratchpad should you ever set one up with NHM. This list works just as well for casual recording, as soon as you find the name of that strange exotic beast from ID sites, look it up in GBIF and add it to the list.

If it's a UK-only list then the Recording Schemes will be able to provide you with the basic spreadsheet which you can then augment using the Taxon Toolkit which will give you the UKSIs with barely any effort.

B. Sheet assemblages

Add the above three as separate sheets in a spreadsheet and you've got a useful starting point for a record of your finds on a "per-expedition" basis. A file such as "France 2014" where you can build in everything you find and deploy your vlookup skills to add records to an Event/Occurrence sheet of your own devising. If this is for your personal use or submission to a Recording Scheme then you need progress no further.

C. Spreadsheet with Cross-referenced sheets

The only means available to progress from B. towards our GBIF objective is via Excel's VLOOKUP

C.1. Prepare all the reference tables

As separate spreadsheets (intially) populate the tables from the above modules with your data. Start with the easiest, just one or two records in **Sources**, all the necessary **Locations** and your Taxon Dictionary (**Checklist**). Make sure each one has a unique key in the first column (see GUI) and that each field conforms to DwC formats. Try to be as comprehensive as possible, you may add or modify them later but bear in mind that you will then be trying to reconcile two versions.

C.2. Assemble the sheets

Start a new spreadsheet of an unambitious nature (e.g. just one small expedition) and name it unambiguously. Add the above tables as separate sheets named Sources, Locations and Checklist. Save it to a dedicated subfolder in your projects folder.

Add a new sheet for Events

C.3. Name the ranges in the sheets

For each sheet select the entire range of data:

1. Shift-End then Shift-Ctrl-Home (those keystrokes save a lot of scrolling)
2. Add a short name to the white Name Box above cell A1, you cannot use the same names as the sheets so use something like "SrcA", "LocA" , "EvtA"& "ChkA"
3. To manage these named ranges use Formulas | Name Manager

C.4. Add Events records

1. Add a sheet "Events" to your spreadsheet . Make this rather "busy" as you may wish to call upon a wide range of fields.

The Events sheet is the one which we wish to populate and our first choice of table should be Sources as this will tend to remain the same for a group of records.

The approach one takes to populating the Events sheet varies according to the material one is working with. A foreign expedition may be straightforward, begin with a single Source and then add an Event for each date+location - the job is done in say 30 records for a week's visit. A more complex source to deal with would be a published list in which the occurrences are listed in taxon sequence, in this case it is necessary to deal partially with the Occurrences first and then the Events

Whatever the type of source information, the transfer of information from the various reference tables to the Events table are carried out using VLOOKUP. A good set of instructions as to how to use this feature is by Excel Easy.

C.5. Add Occurrence records

One or more taxon occurrences belong to each Event.

1. Add the sheet "Occurrences" to your spreadsheet . This can be less "busy" than the Events sheet. The aim is to design it to suit your needs, particularly the requirements of GBIF (see Download). One of those needs might be your own mapping so ensure that you include the *locationID* field.

D. Excel's "Data Model"

The data model features are only present in Excel 2013 and later. Most users have Excel 2010 so this process is not feasible.

A relational database is the proper way to set about linking all the tables in the data model together. Databases such as MS Access are designed for this purpose.

Excel does have some other features in the form of Addins which may be of value if you are unwilling to invest in more current versions. It is possible to incorporated Power View sheets and Pivot Tables into Excel using their SQL technology. It doesn't provide much in the way of relational database features but it does provide a means of linking into MSAccess databases and extracting entire datasets with relative ease. Here's how to start:

D.1. Install Excel's Power Pivot feature

This has to be installed from Microsoft at https://www.microsoft.com/en-us/download/details.aspx?id=43348 unless you happen to have a version in which it was included by default. After downloading and running the .msi file, in Excel:

1. Go to File | Options | Add-Ins.
2. In the Manage box, click COM Add-ins | Go
3. Select the item(s) you require and press OK

The PowerPivot feature now occupies its own tab at the top of your Excel screen.

Case studies

During the course of compiling this account the methodology (DwC spreadsheet template) was used on the following:

1. A. A published paper

Comprehensive country-based accounts with sufficient occurrence data to be worth extracting are scarce. My first study (using method C. above) was a 1990 account of 12 taxa in Czechoslovakia (Rohacek & Bartak).

Papers of such a nature from that predigital era were similarly present in the UK and formed the basis of the first digitising efforts in the mid 90s. Fledgling Local Environmental Resources Centres dealt with many such reports (usually only regional in scope) but national Recording Schemes were busy with their spotty maps and BRC were the main digitisers.

The Czechoslovakian paper contained all the essential ingredients for species occurrences except dates. They were based upon museum specimens and current field work. Important material as it forms the backbone, for that taxonomic group, of those countries (now Czech Republic and Slovakia of course) inventory.

a. **Source** module was a single record - the source being that single paper
b. **Locations**: The first hurdle was their use of a country-specific grid, the coordinates provided were in the form of codes to that grid. Several other European countries also record against grids of their own devising (it's not just the UK.) Analysis is a complex GIS process but once the shp file has been developed it should be published in some central repository for others to use (Buchar grid)

   To add the geospatial coordinates to a spreadsheet, they were copied from the GIS table of tiles. The centroid of each tile was determined for each to provide the Lat/Long. Location names, which were provided against each taxon occurrence in the paper were added later.
c. **Checklist** was straightforward, mostly the same taxa as in the UK but since they are records from my Recording Scheme I have all that data from other studies.
d. **Events**: Since the occurrences were presented without dates there was no means of grouping the occurrences into groups of events. Accordingly the Events table has exactly the same number of entries as does the Occurrence table.
e. **Occurrences**: No problems were encounterd here. I avoided any sorting in this and the Events table whilst any formulae were active on the sheets. Names of recorders were implemented via a rudimentary table and vlookup.

A further sheet, devised to facilitate GIS mapping of the occurrences would be advisable as a final check for outliers.

The resulting spreadsheet and the metadata form to it were the first "list" as requested by NBN.

1. B. Other published papers

Make a collection of published papers in your subject area, best done using Elsevier's Mendeley. There are no unique identifiers to such papers (less than one in ten has a DOI) but one can be contructed by uploading details to a Scratchpad where the url will serve as such. Personal communications might be included in this collection. My procedure

was to work through the modules in the following sequence:

a. Source module to cite each paper.

Sort papers in Mendeley, assign candidate papers to a category "georeferenced" then sort in date order, on the assumption that the more recent the paper, the more likely it was to have good geospatial coordinates.

- Country checklists are problematic as they have no geospatial coordinates within country, best to assign these to another dataset (avoid country checklists without supporting data such as Fauna Europaea).

b. Locations may be vague but there is an increasing trend to provide Lat/Long coordinates in more recent papers.

- Beware of large lists, though they may be well arranged in a table, not all of them may be used for your taxa of interest. Grab the location names/codes assigned to the taxa and develop your Location list from those.
- Provincial papers may just provide the location name. Use Google Earth and assign a poor *coordinateUncertaintyInMeters* and a poor *precision*.
- Some papers refer to locations listed in separate yet another published list, which may be quite unintelligible
- Some locations are provided via maps alone, in this case resort to GIS to match sites with Lat/Long as in 1.A. above
- For shortish sets of locations in a single paper, use Google Earth and save each location you find to that countries folder (sort alphabetically either at this point or before you add to your Locations table). Ensure you File | Save | Save My Places. Select the country folder and right-click "Save Place As" ensuring it's a kml file. The kml file will open in Excel, copy the Lat/Long and names you assigned to your Locations table.

c. Events & occurrences should then be straightforward to enter, most papers do not contain many.

Lots of copying and pasting from pdfs involved in this process. Use a spare sheet on the spreadsheet for organising lists or making temporary notes.

Dates: One issue that crops up early on is that of date format. The suggested protocol is that of ISO 8601:2004(E) which is not an easy read. It uses a YMD format rather than the accustomed DMY which 90% of the world uses. The reason for the DwC YMD format (specifically YYYY-MM-DD) is that it is machine readable. This YMD format deals with partial dates well and date ranges (e.g. for traps) are more intelligible, for example

1998-05-28/06-15 versus 28.v.-15.vi.1998 the latter being an ancient format still prevalent in published papers.

The loss of the standard spreadsheet format is a little inconvenient, that format is stored as a number and merely displayed in DMY format. Without it the formulae for working out *startDayOfYear* and *endDayOfYear* is more complex (should you decide that those are necessary). The formal structure of the ISO standard means that Excel's DMY format is readily achieved by formula if it is needed. Darwin Core doesn't like that Excel format however so it is prudent to avoid it. In other words ensure that the field is formatted for text rather than date.

The above does not apply to existing systems which use the Excel DMY format (e.g. UK's recording schemes - see guide part 2)

2. Field Notes | Expeditions

This category covers two types: a) personal trips and collections, b) formal expeditions. One file for each type.

a) In the case of personal material this may be a mixture of specimens, images and observations. Another way of recording field notes.

b) For formal expeditions it would be the responsibility of the organiser to issue blank copies of the spreadsheet and share location details via kml files with other members of the expedition and verifiers of the material. An example of such an expedition is the survey carried out by Marc Pollet (Belgium) on Corsica in 2019. He has to deal with incoming material from several surveyors, obtained by various trapping and other methods and then to collect that material together and send out to various verifiers for identification. It is they who record individual occurrences.

Obtaining an identification may be difficult, it may be necessary to photograph the specimen and post to an online ID site, resulting in long delays. But isn't that the very nature of one's Field Notes and image collections.

a. **Source** module comprised of several records, one per expedition (e.g. France 2008) or per submitted list
b. **Locations** make an initial record of all the locations visited on a particular field trip. If you are collating a list from all participants then all locations are required eventually (kml files), otherwise just your own. If you had a fairly poor record of the coordinates in your field notes then compiling them into the DwC list is an opportunity to check and refine them all.
c. **Checklists** in practise this turns out to be the most demanding of the tasks. Every taxon has to have its GBIF reference and that takes a while (just the first 3 fields initially, tidy up later). For personal use that may not be so bad (just the taxa that you record) but for an expedition recording across a wide range of taxa then this can become immense. There's a strong case for checklists to be developed independently and shared.
d. **Occurrences** presumably you have these already recorded in some format, field notes in a spreadsheet or MSAccess database and used to label specimens. These are readily copied into this DwC format. Mine amounted to around 2000, presumably other's will be around the same region. [Excel tip: =text(d10,"YYYY,MM,DD") converts to the DwC required format]

Outside the users of biological recording applications, personal records are most likely to be kept in some sort of spreadsheet system. Naturalists who take specimens will keep a separate record of those so as to facilitate labelling. Simple observations may just be entered into an online data-gathering system and not kept by the observer. Personal collections are the most likely use of the DwC template.

The template was applied to my field notes. I've now all my collections from abroad accessible in a single spreadsheet, a lot more manageable than the MSAccess tables I had them in previously. Quite an interesting task though, lots of nice hoverflies (codes searched out from GBIF and added to my checklist.) How I'm ever going to be able to send those to GBIF is a conundrum but at least I'll be in a position to do so if I can hunt down an organisation that deals with such records. It was necessary to check geospatial coordinates from IGN maps which were used in the days before GPS.

3. Specimen collections | Correspondents

Collections that are not your own. Loans, gifts or work carried out in a museum. Lists provided by correspondents also belong in this category. Use expeditions (above) for your own material.

Take a blank copy of the spreadsheet file into a museum to work on a collection. As you will be working on a specific taxonomic group, it would be prudent to develop that first.

a. **Checklist:** Material will now be from a wider range of taxa than the above categories. Accordingly items will have to be appended to the taxon dictionary. Likely to be a slow process
b. **Source** module comprises only one record per collection. If it's a museum then keep separate files.

I could not test this out in a museum, perhaps one of my correspondents might do that. I've only experience with one such (UK) collection, many years ago, which involved sourcing the original field note books thus providing the Event data which allowed the pinpointing of poorly labelled specimens from just a cryptic location name and a date.

I was able to test it on donated specimens and submitted lists ...

4. Collected online postings

Records can be made of online postings to identification and other sites. This collection may duplicate records from sites that upload to GBIF. The issue of flagging these so as to avoid duplicates may need to be addressed.

For a Recording Scheme this is an important source of information as it leads to the development of country checklists .

a. **Source** module comprises many records, one per occurrence.
b. **Locations** may be problematic and details may have to be sought from the originator, language may be a barrier to this.

In practise this was fairly straightforward. Sites vary in their intentions, some are clearly set up to record occurrences and have full details (Le mondes des insectes - *French*, MacroID - *Russian*, Ukranian Biodiversity Information Network - *Russian,* iNaturalist), others seem to have the facility to do that but frustratingly users don't add the geospatial coordinates (BiodiversidadVirtual - *Spanish*) others are intended for identification and discussion (Diptera.info) but have no coordinates (though one can enquire) and others are designed simply to register images (Flickr, blogs) where you may be lucky and get some coordinates.

I've clearly only discovered some such sites based on my interests, there is a need for a register of them all somewhere.

A fairly straightforward job extracting the records of all these. The sites are discovered over time by searching for images of one's taxa of interest then homing in on the site that has published them.

As they are all strictly single occurrences the GUIs of the Source, Location, Event and Occurrence tables are all maintained in sequence.

I finished up with around 200 records in this category based on around 90 taxa.

5. Imports

For research purposes datasets need to be imported from a variety of sites, GBIF, NBN Atlas, iRecord, iNaturalist and other sites known to have published to a GBG already.

The objective in obtaining these datasets is simply for one's own research purposes of mapping, phenology, modelling and so on. They are not datasets one would be preparing for a GBIF upload as that is already taken care of. They may be datasets (iRecord) one is preparing for an NBN Atlas upload, this will depend upon the arrangements you have made with BRC in this regard. Some of the following provide systems for download, others not:

One spreadsheet per source as follows:

5.1. iRecord (UK only)

The technique for downloading datasets from iRecord is documented at ???? If you are a verifier then do all that work before continuing.

As a verifier of a particular group, one is able to select that group in the "Download type" dropdown. Filter is "All records" and the records unlimited (all surveys) with Date field set to "Field record date". Since all are required, choose a Start date as early as the system will allow (1st Jan 1999) and today's End date The CSV format is fine for a spreadsheet. Select download and open the file in Excel.

The download is very "data-dense" or "noisy" (47 fields for each occurrence) as this has to support a wide range of services and is not configurable, that problem is left to the end-user.

The first thing to note is that the dataset is upside-down. The most recent records are at the top as indicated by the field "Input on date" (column AL) To work on these in a spreadsheet this sequence has to be reversed so that more recent records can be later appended (at the bottom) - use Data Sort.

5.2. NBN Atlas (UK only)

Data downloads are a primary function of NBN Atlas and are carried out frequently. If you submit records on a regular basis to NBN Atlas then you need to exclude your own submitted records from the download.

5.3. iNaturalist

Though iNaturalist supports complex filters in order to set up Projects, their download system does not support this level of complexity. Downloads (exports) are restricted to single taxonomic groups within single geospatial areas.

Procedure:

1. Set up a Project (minimum requirements that you have 50 IDs or observations)
2. For Species add each group, you may need more than one (e.g. Families, Superfamilies)
3. For Places add Europe, Turkey, Georgia, Armenia & Azerbaijan (these last four ensure coverage of "Pan-Europe")
4. Verify those that you can.
5. On your Home page select Explore
6. Enter a single taxonomic group into the Species field and a single place into the Place field (e.g. Nerioidea + Europe)

The map in your project may indicate no records for the other Places, in which case you need only perform the following task once for each taxonomic group

7. Check that the images shown are correct.
8. Open the Filters box, these require no changes from the default.
9. In the lower right corner of that box select Download
10. This takes you to the iNaturalist Export Observations page which shows you all the settings you just made, ignore all that lot and scroll down to step 4, the blue "Create Export" button
11. Press the button and Save
12. Repeat steps 5 to 11 for each taxonomic group and steps 5 to 12 for each Place which contains any observations

Which results in one or more zipped csv files which will open in Excel.

These are sequenced by *id* (column A) which is a chronological sequence, old to new, based on *created_at* (column H) so these do not need re-sorting as was the case for iRecord.

Each Species|Place combination will require a separate Excel analysis file in order to permit later updates.

13. Copy the entire sheet (opened in Excel) and paste into a new sheet in a prepared DwC template Excel file
14. Name the sheet

a. Source module comprises many records, one per occurrence.

b. Locations No *country* or *countryCode* in the download unfortunately. Using Lat/Long coordinates the following methods are available:

a. Javascript and R solutions (would require considerable programming knowledge)

b. Online single records can be determined with https://www.mapdevelopers.com/geocode_tool.php. This also gives the Region (oblast, departement etc.) and an English name which helps with Russian records as the iNaturalist download does not support any non-English alphabets.

c. GIS for larger batches, using TDWG maps to read off the country and region polygons. Somewhat complex and requiring GIS skills

d. Use the record's url to locate the iNaturalist posting then check the details under the map.

e. Scrolling down the location names and picking off the obvious ones (ends in "FR" or "Sweden", contains "Wein")

Choice of method rather depends upon the size of the task, for a couple of hundred records method e. followed by b. looks favourite.

5.4. GBIF

As this is the intended final destination for any exports carried out, some care in interpretation is required in order not to duplicate one's own work.

Procedure

Extracting records from a published paper

The example chosen is Shatalkin & Merz's 2001 paper on Swiss Psilidae. The records were extracted as follows:

1. Paper entered in Mendeley and the pdf opened in that program.
2. In Google Earth set up a folder for Switzerland and ensure that it is highlighted.
3. Set the GE display to DD
4. In Mendeley highlight all the record separators (the authors used a "-") to assist in your navigation
5. Work through all the location names, copying each from the pdf and pasting into the search box in GE. Most of the time GE will find the location name successfully.
6. Set a pin at the chosen spot on the map, paste the location name into the pin's name box and save
7. If all is well then that pin will be appear in the GE Switzerland folder
8. Save My Places (do this frequently, you don't want to forget and lose a long session's worth of data and GE doesn't remind you when you try to exit)
9. Continue through the entire paper, shuffling each recorded GE pin into alphabetical order so as not to duplicate.
10. Ultimately I got around 80 Swiss sites recorded in this way, one was in France but leave that in the Swiss list for now.
11. Save My Places
12. Select the Switzerland folder, right-click and save as kml (note that that file could be shared with a collaborator if you had one)
13. In Excel open the kml file. Copy the three columns, name, lat and long into a new Excel file and save.
14. Copy and append those columns to the Locations table (extend the *locationID* GUIs to cover them all)
15. Add the other data to the Location table, *coordinateUncertaintyInMeters* at perhaps 3000 (*coordinatePrecision* 0.01) which is a cautious estimate. Also add *higherGeography*, *country* & *countryCode* (obtainable just by entering "ISO Switzerland" into internet search engine), not forgetting to spot that French location.
16. Copy the 80 or so *locationID* from the Locations table to the appropriate column in the Events table, extend the *eventID* and other vlookup columns.

Events: There are initially no dates in the Events table, these are to be added as one works one's way through the records. As additional dates are discovered, extend the Events list by copying the locationID to a new Event record, adding a new date. My list of 80 locations produced around 140 events.

Dates: These should be in DwC format (text). A little awkward at first but reassuring that at some future point the Excel date format can be derived by simple formula. The DwC format is also very neat at registering date ranges for malaise traps (a fact that should also be recorded in the Event *samplingProtocol)*

Adding occurrences

1. From the top of the paper (opened in Mendeley) work down the list as follows:
2. In the Occurrence table add the *taxonNumber* by copying the appropriate code from the Checklist table. Add *individualCount*, *sex*, *recordedBy* and *identifiedBy* (=authors of the paper). I used the pipe symbol (|) in the case of mixed sex catches thus - 1|2 m|f
3. Add *basisOfRecord* as "preservedspecimen" since the paper indicates they are all held in museums.
4. Find the *eventID* code (in the Events table - which are arranged alphabetically) to each occurrence and copy that to the *eventID* in the Occurrence table.
5. Copy all the vlookup formulae down the list as you proceed, taking care not to increment numerical values (e.g. *taxonNumber* should be copied rather than extended.) At no point do any sorting, this can only be carried out after all the vlookup formulae are gone from the spreadsheet as a result of Copy | Paste as values
6. Use the highlighter pen in Mendeley to mark your progress.
7. Allow a few days to complete ~200 occurrences.

Publishing to GBIF

The first step is to become an endorsed organisation or be a contributor to such an organisation.

Thanks to the assistance of NBN, the European Micropezids & Tanypezids has become endorsed, it has its own GBIF page for datasets at European Micropezids & Tanypezids, no datasets there yet, figuring out how to do that is the purpose of this guide.

I was offered help by both NBN and GBIF's Data Manager (Jan K. Legind). The approval of this scheme may have something to do with the fact that I have a research site already set up in the form of NHM's Scratchpad. Jan's guess was that I might be sharing data through the Scratchpad infrastructure and though it would be a very useful addition to Scratchpad's tool set it is doubtful that that would be feasible as it would require a considerable amount of liaison amongst developers.

Management tasks

1. Reference material

Download useful background guides, there's this, Hill's paper on geospatial coordinate quality control and the GBIF paper on the templates. Add them to Mendeley.

2. Housekeeping

Assign a folder for your work, some named folder under a "projects" folder is usually a good approach. Keep tabs on this using iMatch or OneNote or other preferred system - beats hunting around for stuff every time.

Implement backup routines.

3. Develop your taxon dictionary

Take the GBIF spreadsheet table and modify it according to your needs (see above). Add all the taxa (including levels up to Family) you will need for your project. Don't get too carried away but bear in mind that this checklist is valuable in its own right for other purposes and is shareable with others. It's probably advisable to make separate dictionaries for each taxonomic group you have an interest in

4. Develop your Locations list

Write a Locations spreadsheet (see above) and add locations on a "per project" basis. Read Hill (2010) for horror stories about geospatial errors and watch or listen to the Dad's Army episode "A question of Reference".

.

Barriers to European recording

Though there are some tools to help facilitate biological recording overseas there is need for many more to help surmount the barriers faced by casual recorders and researchers when attempting to contribute to or utilise Global Biodiversity Gateways.

1. Lack of citation managers with the capacity to assign taxa or provide GUIs
2. Lack of facilities in Recorder 6 to readily add GBIF taxon lists + other shortcomings regarding maps and ease of geospatial coordinate entry.
3. Lack of a desktop database suitable for typical research usage (e.g. supporting the upload of downloads)
4. Absence of online recording systems such as iRecord for overseas use
5. Acute shortage of verifiers
6. Huge backlogs of European data yet to be digitised
7. Lack of adequate documentation regarding pretty much everything, few guides and inadequate support in certain areas.
8. Poor geospatial discipline by posters on online identification sites
9. Inadequate engagement by European countries in GBIF (i.e. non-participating countries)
10. Language barriers

11. Lack of appropriately designed metadata management utilities. Poor implementation of standards
12. Lack of funding (see failed AICHI targets)
13. Unreasonable skills expectation. Construct a Venn diagram of recorder + taxonomist + web designer + database manager + GIS and you'll find perhaps one or two people in the world. The brave few who try to expand their skills a little do not always succeed, support is variable.
14. Poor implementation of Standards throughout (Dublin Core, W3C, ISO, TDWG vs itself, EEA standards)

It is near impossible for European workers to record and research via structured databases. Many resort simply to publishing a sequence of papers on their subject areas resulting in an accumulation of occurrences which add to the backlog of material which needs to be addressed before these can become available for broader research and monitoring.

A multi-lingual application is called for.

References

(NBN). (n.d.). A guide to providing species datasets. Retrieved from http://www.nbn.org.uk/Share-Data/Providing-Data/Data-Provider-Checklist.aspx

(NBN). (1999). Running a Local Records Centre V2 (16-21) (Vol. 2). Retrieved from http://www.alerc.org.uk/uploads/7/6/3/3/7633190/running_lrc_vol2_16_to_21.pdf

(NBN). (1999). Running a Local Records Centre V2 (3-10) (Vol. 2). Retrieved from http://www.alerc.org.uk/uploads/7/6/3/3/7633190/running_lrc_vol2_03_to_10.pdf

Agrin, N., Kline, J., & Ueda, K. (2007). iNaturalist: Final Project Write-up. School of Information, UC Berkeley, 41.

August, T. O. M., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., & Jepson, P. (2015). Emerging technologies for biological recording. Biological Journal of the Linnean Society, 115(3), 731–749.

Austen, G. E., Bindemann, M., Griffiths, R. A., & Roberts, D. L. (2016). Species identification by experts and non-experts: comparing images from field guides. Scientific Reports, 6(August), 33634. https://doi.org/10.1038/srep33634

Ball, S. (2011). NBN Record Cleaner user guide. (February), 1–47.

Ball, S. (2012). NBN Record Cleaner Verification Rules. (August), 1–25. Retrieved from http://data.nbn.org.uk/recordcleaner/documentation/NBNRecordCleanerRuleGuide.pdf

Ball, S., & Rankin, S. (2009). Recorder 6 User Guide. JNCC, 6b(January 2007), 105. Retrieved from http://jncc.defra.gov.uk/page-4595

Barve, V. (2015). Discovering and developing primary biodiversity data from social networking sites. Retrieved from https://kuscholarworks.ku.edu/handle/1808/19011%5Cnhttps://kuscholarworks.ku.edu/bitstream/handle/1808/19011/Barve_ku_0099D_14003_DATA_1.pdf?sequence=1&isAllowed=y

Bell, S., Marzano, M., Cent, J., Kobierska, H., Podjed, D., Vandzinskaite, D., … Mursic, R. (2008). What counts ? Volunteers and their organisations in the recording and monitoring of biodiversity. Biodivers Conserv, 17, 3443–3454. https://doi.org/10.1007/s10531-008-9357-9

Bishop, B. W., & Hank, C. (2016). Data curation profiling of biocollections. Proceedings of the Association for Information Science and Technology, 53(1), 1–9. https://doi.org/10.1002/pra2.2016.14505301046

Brummitt, R. K. (2001). World Geographical Scheme for Recording Plant Distributions Edition 2. In International Working Group on Taxonomic Databases For Plant Sciences (TDWG). https://doi.org/10.1043/0363-6445-28.4.807Bungartz, F. (n.d.). Using IDimager to catalog photographs of Natural History Collections with the Darwin Core XMP metadata standard .

Burnett, J., Copp, C., & Harding, P. (1995). Biological recording in the United Kingdom - present practice and future development (Vol. 1). Retrieved from http://nora.nerc.ac.uk/7869/

Cabinet_Office. (2006). e-Government Metadata Standard Metadata. Retrieved from http://www.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf

Cameron, A. (SNH). (2010). Involving people in biological recording. Retrieved from http://www.snh.org.uk/pdfs/publications/commissioned_reports/382.pdf

Chapman, A.D., Franzier, C.K., Grafton, O., Grant, S., Hobern, D., Lane, M., Wall, J. & Wieczorek, J. (2008). GBIF Training Manual 1: Digitisation of Natural History Collections Data. Retrieved from http://www.gbif.org/resources/2776

Chapman, A. D., & Speers, L. (2005). Uses of Primary species- occurrence data, version 1.0. Report for the Global Biodiversity Information Facility. Retrieved from https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiw4a246ObLAhVBmBoKHcopCvIQFggcMAA&url=http://www.gbif.org/system/files_force/gbif_resource/resource-80545/gibf_uses_data_training_manual_en_v1.pdf?download=1&usg=AFQjCNGqJDuFh8JR

Chapman, D. S., Bell, S., Helfer, S., Roy, D. B., Gardens, R. B., & Tw, S. (2015). Unbiased inference of plant flowering phenology from biological recording data. Biological Journal of the Linnean Society, 115(3), 543–554.

Chavan, V. S., & Ingwersen, P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics, 10 Suppl 1(Lmmc), S2. https://doi.org/10.1186/1471-2105-10-S14-S2

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics, 12(Suppl 15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2

Copp, C. J. T. (2004). Data model documentation Part1.

Copp, C. J. T. (2000). The NBN data model and its implementation in Recorder 2000. Environmental

Information Management, (November), 1–105. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.5946&rep=rep1&type=pdf

D, R., Braak, K., M, D., & Robertson, T. (2011). Darwin Core Archive how-to guide. (April), 21. Retrieved from http://links.gbif.org/gbif_dwc-a_how_to_guide_en_v1

Daume, S., & Galaz, V. (2016). "Anyone Know What Species This Is?" - Twitter Conversations as Embryonic Citizen Science Communities. PloS One, 11(3), e0151387. https://doi.org/10.1371/journal.pone.0151387

E-Government_Unit. (2004). UK GEMINI Standard A Geo-spatial Metadata Interoperability Initiative. Framework, 1–54.

Ellis, R., Pacha, M., & Waterton, C. (2007). Assembling Nature: the social and political lives of biodiversity softwares.

Eynden, A. V. Van Den, Corti, L., Bishop, L., & Horton, L. (2011). Managing and Sharing Data.

Foster, G. N. (2015). Taking the oldest insect recording scheme into the 21st Century. Biological Journal of the Linnean Society, 115(3), 494–504.

Franklin, J. (2009). Mapping species distributions: spatial inference and prediction. In Landscape Ecology Journal of Vegetation Science. Retrieved from http://www.cambridge.org/ar/academic/subjects/life-sciences/ecology-and-conservation/mapping-species-distributions-spatial-inference-and-prediction

Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content Assessment of the Primary Biodiversity Data Published Through GBIF Network : Status , Challenges and Potentials. Biodiversity Informatics, 8(August 2012), 94–172. https://doi.org/10.17161/bi.v8i2.4124

Goula, M., Sesma, J.-M., & Vivas, L. (2013). Photosharing websites may improve Hemiptera biodiversity knowledge and conservation. ZooKeys, 319(319), 93–105. https://doi.org/10.3897/zookeys.319.4342

Groom, Q., Weatherdon, L., & Geijzendorffer, I. R. (2017). Is citizen science an open science in the case of biodiversity observations? Journal of Applied Ecology, 54(2), 612–617. https://doi.org/10.1111/1365-2664.12767

Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. Biological Conservation, 116(2), 195–203. https://doi.org/10.1016/S0006-3207(03)00190-3

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., … Page, R. D. M. (2015). Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. Zookeys, 154, 133–154. https://doi.org/10.3897/zookeys.494.9352

Heath, J., & Leclercq, J. (1981). Provisional Atlas of the Invertebrates of Europe: Maps 1-27. Retrieved from Istitute for Terrestrial Ecology website: http://nora.nerc.ac.uk/7903/1/Euro_Invert-1-27.pdf

Heberling, J. M., & Isaac, B. L. (2018). iNaturalist as a tool to expand the research value of museum specimens. Applications in Plant Sciences, 6(11), 1–8. https://doi.org/10.1002/aps3.1193

Hill, A. W., Otegui, J., Ariño, A. H., & Guralnick, R. P. (2010). GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network. Gbif, (August), 25. Retrieved from http://www2.gbif.org/GPP-Final.pdfHuang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters, 5(5), 399–406. https://doi.org/10.1111/j.1755-263X.2012.00259.x

Isaac, N. J. B., & Pocock, M. J. O. (2015). Bias and information in biological records. Biological Journal of the Linnean Society, 115(3), 522–531. https://doi.org/10.1111/bij.12517/abstract

Isaac, N. J. B., van Strien, A. J., August, T. a., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution, (ii). https://doi.org/10.1111/2041-210X.12254

James, T. (2007). Running a Biological Recording Scheme or Survey: A handbook to help scheme or society administrators. (September), 61. Retrieved from https://www.fba.org.uk/sites/default/files/Running a Biological Recording Survey or Scheme.pdf

James, T. J. (2011). Improving Wildlife Data Quality. Retrieved from http://www.nbn.org.uk/About/The-Organisation/NBN-Timeline.aspx

JNCC. (2009). Exporting data from Recorder 6 to the NBN Gateway. JNCC, (August 2009). Retrieved from http://jncc.defra.gov.uk/page-4595

Ko, B., Chavan, V., & Remsen, D. (2011). Document Map for Publishing Occurrence Data. Global Biodiversity Information Facility. Retrieved from http://www.gbif.org/resource/80726

Lightfoot, P. (n.d.). NBN Standards for integrated online recording and verification. Retrieved from http://www.nbn.org.uk/nbn_wide/media/Documents/Publications/NBN-Standards-for-Online-

Recording.pdf

Lightfoot, P., Harvey, M., & Breda, J. van. (2015). iRecord: Guidance for verifiers. (June), 1–12. Retrieved from http://www.brc.ac.uk/irecord/sites/default/files/guides/verifiers.pdf

Maes, D., Isaac, N. J. B., Harrower, C. A., Collen, B., van Strien, A. J., & Roy, D. B. (2015). The use of opportunistic data for IUCN Red List assessments. Biological Journal of the Linnean Society, 115(3), 690–706.

Marshall, S. a., & Evenhuis, N. L. (2015). New species without dead bodies: a case for photo-based descriptions, illustrated by a striking new species of Marleyimyia Hesse (Diptera, Bombyliidae) from South Africa. ZooKeys, 525, 117–127. https://doi.org/10.3897/zookeys.525.6143

Mesibov, R. (2018). An audit of some processing effects in aggregated occurrence records. ZooKeys, 751, 129–146. https://doi.org/10.3897/zookeys.751.24791

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., & Cockerill, M. (2011). Towards mainstreaming of biodiversity data publishing : recommendations of the GBIF Data Publishing Framework Task Group. BMC Bioinformatics, 12(Suppl 15), 1–10. Retrieved from http://download.springer.com/static/pdf/643/art:10.1186/1471-2105-12-S15-S1.pdf?originUrl=http://bmcbioinformatics.biomedcentral.com/article/10.1186/1471-2105-12-S15-S1&token2=exp=1458722775~acl=/static/pdf/643/art%3A10.1186%2F1471-2105-12-S15-S1.

Morris, R. (2012). Promoting biological recording - a developing conundrum. British Wildlife, 24(2), 107–110.

Morris, R. K. A., Ball, S. G., & Andrews, I. (2016). Photographic records of Hoverflies collated in 2015 : a preliminary analysis. In (unpublished) UK Hoverfly Recording Scheme. Retrieved from http://www.bacoastal.co.uk/Entomology/2016-Photo-report.pdf

Mücher, C. A., Klijn, J. A., Wascher, D. M., & Schaminée, J. H. J. (2010). A new European Landscape Classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes. Ecological Indicators, 10(1), 87–103. https://doi.org/10.1016/j.ecolind.2009.03.018

National Archives. (2004). Requirements for Electronic Records Management Systems. ReVision. Retrieved from http://www.nationalarchives.gov.uk/documents/metadatafinal.pdf

Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B: Biological Sciences, 374(1763), 2–10. https://doi.org/10.1098/rstb.2017.0391NERC. (2011). UK Taxonomy & Systematics Review Response compiled by NERC on behalf of the community of taxonomy and systematics funders. Retrieved from http://www.nerc.ac.uk/research/programmes/taxonomy/events/pubreview.asp

Otegui, J., & Guralnick, R. P. (2016). The geospatial data quality REST API for primary biodiversity data. Bioinformatics, 32(11), 1755–1757. https://doi.org/10.1093/bioinformatics/btw057

Page, L. M., Macfadden, B. J., Fortes, J. A., Soltis, P. S., & Riccardi, G. (2015). Digitization of Biodiversity Collections Reveals Biggest Data on Biodiversity. 65(9), 841–842. https://doi.org/10.1093/biosci/biv104

Pescott, O. L., Walker, K. J., Pocock, M. J. O., Jitlal, M., Outhwaite, C. L., Cheffings, C. M., … Roy, D. B. (2015). Ecological monitoring with citizen science : the design and implementation of schemes for recording plants in Britain and Ireland. Biological Journal of the Linnean Society, 115(3), 505–521.

Patsatzi, G. M., & Patsatzi, E. (2009). The UK Museum Documentation Standard. Spectrum, (3.2).

Plant, A. (2015). Worldwide Biogeography: Carts, Horses and the Taxonomic Impediment. Bulletin of the Dipterists Forum, 79, 10.

Pocock, M. J. O., Newson, S. E., Henderson, I. G., Peyton, J., Sutherland, W. J., Noble, D. G., … Gifford, C. (2015). Developing and enhancing biodiversity monitoring programmes : a collaborative assessment of priorities. Journal of Applied Ecology, (52), 686–695. https://doi.org/10.1111/1365-2664.12423

Pocock, M. J. O., Roy, H. E., Preston, C. D., & Roy, D. B. (2015). The Biological Records Centre : a pioneer of citizen science. Biological Journal of the Linnean Society, 115(3), 475–493.

Powney, G. D., & Isaac, N. J. B. (2015). Beyond maps : a review of the applications of biological records. Biological Journal of the Linnean Society, 115(3), 532–542.

Pressey, R. L. (2004). Conservation planning and biodiversity: Assembling the best data for the job. Conservation Biology, 18(6), 1677–1681. https://doi.org/10.1111/j.1523-1739.2004.00434.x

Preston, C. D., Roy, D. B., & Roy, H. E. (2012). What have we learnt from 50 years of biological recording? British Wildlife, 24(2), 97–106.

Purse, B. V, & Golding, N. (2015). Tracking the distribution and impacts of diseases with biological records and distribution modelling. Biological Journal of the Linnean Society, 115(3), 664–677.

Redwine, G. (2015). Personal Digital Archiving. Digital Preservation Coalition, (December 2015), 1–37. https://doi.org/10.1081/E-ELIS3-120044332

Remsen, D. P., Döring, M., & Robertson, T. (2011). GBIF Spreadsheet Templates User Guide, version 1.0. (April), 0–3. Retrieved from http://links.gbif.org/dwca-spreadsheet-processor-guide

Rondinini, C., Wilson, K. A., Boitani, L., Grantham, H., & Possingham, H. P. (2006). Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology Letters, (9), 1136–1145. https://doi.org/10.1111/j.1461-0248.2006.00970.x

Roy, D. B., Harding, P. T. (BRC), Preston, C. D., & Roy, H. E. (2014). Celebrating 50 years of the Biological Records Centre. In Biological Records Centre. Retrieved from http://nora.nerc.ac.uk/5213/

Roy, H. E., Preston, C. D., & Roy, D. B. (2015). Fifty years of the Biological Records Centre. Biological Journal of the Linnean Society, 115(3), 469–474.

Roy, H. E., Rorke, S. L., Beckmann, B., Booy, O., Botham, M. S., Brown, P. M. J., … Walker, K. (2015). The contribution of volunteer recorders to our understanding of biological invasions. Biological Journal of the Linnean Society, 115(3), 678–689.

Stewart, A. J. A., Bantock, T. M., Beckmann, B. C., Botham, M. S., Hubble, D., Roy, D. B., … London, N. (2015). The role of ecological interactions in determining species ranges and range changes. Biological Journal of the Linnean Society, 115(3), 647–663.

Sumner, D. P. (2016). Photography: Geotagging. Bulletin of the Dipterists Forum, (81), 20–21.

Sumner, D. P. (2016). A Review of the Status of Diptera Recording in the UK. Dipterists Forum Bulletin, 81, 12–17.

Sutherland, W. J., Roy, D. B., & Amano, T. (2015). An agenda for the future of biological recording for ecological monitoring and citizen science. Biological Journal of the Linnean Society, 115(3), 779–784.

Swanson, A., Kosmala, M., Lintott, C., & Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology, 30(3), 520–531. https://doi.org/10.1111/cobi.12695

Tewksbury, J. J., Anderson, J. G. T., Bakker, J. D., Billo, T. J., Dunwiddie, P. W., Groom, M. J., … Wheeler, T. a. (2014). Natural history's place in science and society. BioScience, 64(4), 300–310. https://doi.org/10.1093/biosci/biu032

Theobald, E. J., Ettinger, A. K., Burgess, H. K., Debey, L. B., Schmidt, N. R., Froehlich, H. E., … Parrish, J. K. (2015). Global change and local solutions : Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation, 181, 236–244. https://doi.org/10.1016/j.biocon.2014.10.021

Tuama, Ó., Eamonn, Braak, K., & Remsen, D. (2011). GBIF Metadata Profile, Reference Guide. (April), 19. Retrieved from http://www.gbif.org/resource/80640

Turnhout, E., & Boonman-berson, S. (2011). Databases , Scaling Practices , and the Globalization of Biodiversity. Ecology and Society, 16(1).

Turnhout, E., Lawrence, A., & Turnhout, S. (2016). Citizen science networks in natural history and the collective validation of biodiversity data: Data Validation in Natural History. Conservation Biology, 30(3), 532–539. https://doi.org/10.1111/cobi.12696

van der Wal, R., Anderson, H., Robinson, A., Sharma, N., Mellish, C., Roberts, S., … Siddharthan, A. (2015). Mapping species distributions: A comparison of skilled naturalist and lay citizen science recording. Ambio, 44 Suppl 4(Suppl 4), 584–600. https://doi.org/10.1007/s13280-015-0709-x

Walisch, T. (2007). Proceedings of the first international Recorder conference. Ferrantia.

Wenger, S. J., & Freeman, M. C. (2008). Estimating Species Occurrence , Abundance , and Detection Probability Using Zero-Inflated Distributions. Ecology, 89(October), 2953–2959.

Wetzel, F. T., Bingham, H. C., Groom, Q., Haase, P., Kõljalg, U., Kuhlmann, M., … Häuser, C. L. (2018). Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. Biological Conservation, 221(January 2017), 78–85. https://doi.org/10.1016/j.biocon.2017.12.024

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. PLoS ONE, 7(1). https://doi.org/10.1371/journal.pone.0029715

Acknowledgements

Many thanks to

Revisions

Version 2: Restructure & redesign to take into account Dipterists Forum's Online Publishing guidelines (Bulletin 88).

Cite as:

Sumner D.P.. (2020). Biodiversity Data Management 3: Recording in Europe. Dipterists Forum Report, C(8 V1), 6?