6 Dipterists Forum Technical Guide

Biodiversity Data Management 1: Records collection & verification

Darwyn P. Sumner

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

Series C, Issue 6, Version 2 (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

Summary

A guide to the processes involved in the collation of records submitted in spreadsheet format by recorders through to uploading to the NBN Atlas and GBIF.

Initially intended as personal guide notes, outlining the author's methods for handling incoming datasets for Dipterists Forum's Field Weeks. Audience may now include other Recording Scheme organisers, county recorders, collators of records from expeditions and those wishing to submit non-UK records 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, researchers, biogeographers, phenologists, distribution modellers

Introduction

Specifications for GBG uploads have changed over the years, methods employed to meet these have changed accordingly, and continue to change. The techniques outlined here use sound principles to suggest paths to achieve that objective, tools available may differ.

Principles

Collection, collation, management & dissemination.

These are the four principles of the management of wildlife records. They are detailed in published documents related to Local Environmental Records Centres (Running a Local Records Centre Volume 2, parts 1, 2 & 3) and Recording Schemes (James, 2007;

They are not quite sequential processes, though clearly collection is at the start and dissemination to a publicly accessible format is at the other. Collation and management are intertwined and cover a variety of processes, at points addressing both validation and verification.

A. Collection

Much of the information regarding the initial collection is covered in the notifications and reports (see Bulletins) related to the arrangements for each Field Week.

During the course of the event, a person is designated to act as a combined collector, collator, manager and disseminator.

Though some participants additionally utilise online systems (iRecord) to upload records directly, the method requested is that participants submit records in spreadsheet format to the designated person as an attachment in an email by a specified date.

The following is the process by which these were managed:

A.1. Managing emails

1. Set up an email folder to store incoming emails related to the Event
2. Devise an email filter based upon subject title which will redirect incoming emails to that folder.
3. Use email tagging system to indicate any processes carried out on those messages.

B. Collation

B.1. Managing incoming data

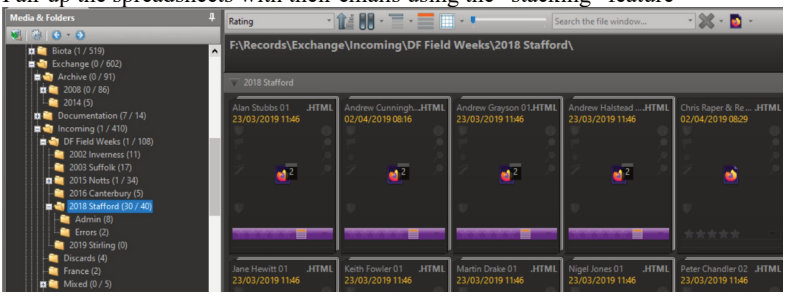
Set up an appropriately named folder to store incoming data (and another for outgoing data). If you have no other document management system then manage through subfolder hierarchy.

Into the appropriate folder place the following:

1. Download the spreadsheet attachments from the email
2. Save the email text as an .html file ([participant name][email sequence].html)

The following illustrates the use of a **digital asset management** system (iMatch) to facilitate management (particularly with respect to archives and backups):

3. Pair up the spreadsheets with their emails using the “stacking” feature



C. Management

C.1. Verification

Contributors providing datasets are presumed to have or have obtained the necessary expertise to ensure that taxa were correctly identified.

C.2. Validation - taxa

The datasets can now be subjected to the following validation checks:

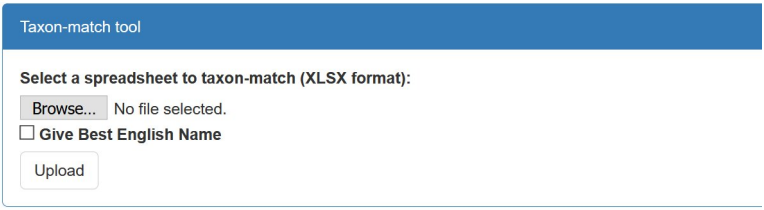
Taxon- match tool

The following methodology was provided by Chris Raper:

1. Create a copy of the spreadsheet. [original name] [copy].xlsx
2. Add a new sheet, naming it “taxon-match”
3. Make “taxon-match” the first sheet
4. Name the sheet containing the original data “raw-data”
5. Copy/paste the species name column from your data into column B on the new sheet
6. If they provided an authority then copy/paste it into column C
7. Copy any record ID number and paste it in column A on the new sheet (not essential but it allows you to quickly look-up the row in the original data)

Note that if an ID column was not present then it will be created.

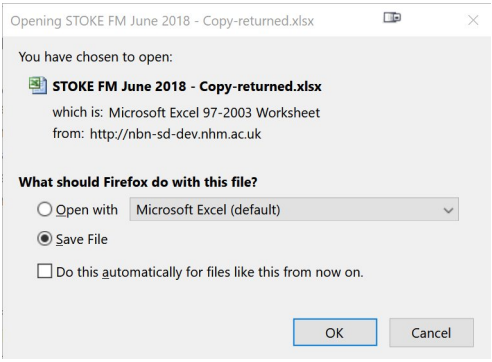
8. Open the utility at http://nbn-sd-dev.nhm.ac.uk/taxonmatch.php



- Pass this script an XLSX format spreadsheet with the following columns:
  - A - Your ID code (optional - used to help you relate the returned information to your records)
  - B - Taxon name (mandatory - e.g. "Tachina fera")
  - C - Taxon authority (optional - e.g. "(Linnaeus, 1761)")
- It will process the names and then return the above PLUS:
  - D - The number of matches found in the UK Species Inventory
  - E - Type of match (exact, name, subgenus, gender or none - see below)
  - F - TAXON_VERSION_KEY (TVK)
  - G - Matched rank
  - H - Matched name
  - I - Matched authority
  - J - Matched attribute
  - K - Language (la=Latin, en=English, gd=Gaelic, cy=Welsh)
  - L - Group name (e.g. bird, flowering plant, insect - butterfly)
  - M - the most appropriate English name
  - If your name is a synonym or we have a better version of your name then we also give:
    - N - Recommended TVK
    - O - Recommended name
    - P - Recommended authority
    - Q - Recommended attribute

9. Browse to your prepared spreadsheet
10. Upload and wait - it will take a while to return the data. There is no progress bar so try it out on shorter lists first.

Eventually the following appears:



10. Save this file and move it to the correct folder.

C.2.1. Interpreting the returned file

The maximum is set to 5000 rows at the moment - if it takes longer than an hour then something is wrong so abort and try a smaller upload.

The "taxon-match" sheet now has rows coloured as follows:

Red	totally unfound names
Purple	possible gender ending differences
Beige	name correct, authority different
Green	perfect match

The main interest are the RED & PURPLE names which require amending.

C.2.2. Managing the returned file

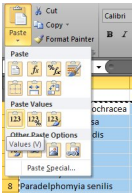
Ideally the contributor will have performed the above check prior to submitting the spreadsheet to the collator. Details were published in Dipterists Forum Bulletin #88 (Autumn 2019.) The collator may now choose either to contact the contributor for clarification regarding any errors or simply exclude those errors. Making verification judgements is not an option for the collator.

Amend the returned spreadsheet as follows in order to exclude errors:

1. Add a new sheet named "amend"
2. Copy row 1 (the header row) from "raw-data" to "amend" (Tip: there's a paste option which preserves the column widths = "keep column widths")
3. Format the date column to date format
4. Use formulae to replicate the first row in "amend" from "raw-data" thus: ='raw-data'!A2 (in cell A2 of "amend")
5. After the last column in "raw-data" add the following header taken from "taxon-match": TAXON_VERSION_KEY (optionally other useful fields may be added but just this one will suffice to sort the good from the bad).
6. Into the first non-header cell in the TAXON_VERSION_KEY column add the formula ='taxon-match'!F2
7. The sheet should now look something like this:

	A	B	C	D	E	F	G	H	I
1	SPECIES	DATE	LOCALITY	VC	GRID REF	COLLECTO	DETERMIN	NOTES	TAXON_VERSION_KEY
2	Austrolimnophila ochracea	23/06/2018	Dimminsc	39	SJ054428	A E Stubbs	A E Stubbs	woodland	NBNSYS0000008140

8. Copy those formulae in row 2 of “amend” down the rows until the end of the records is reached. Note that as you do so, some of the cells in the TAXON_VERSION_KEY column read “0”
9. Highlight the entire dataset from A2 to the last TAXON_VERSION_KEY reading
10. Select Copy (or ctrl C)
11. Select Paste values:



Note that each cell now contains the value rather than the formula. It’s now safe to sort the records on sheet “amend”. **Do not carry out any sorting on the other two sheets at any time**.

12. Select the entire dataset on sheet “amend”, including the headers.
13. Select Data | Sort and ensure that the box “My data has headers” is ticked
14. In the Column | Sort by dropdown box select TAXON_VERSION_KEY and select OK. All the records in which the TAXON_VERSION_KEY could not be resolved have now risen to the top and read “0”
15. Delete all rows in which TAXON_VERSION_KEY = “0” (if this is your own spreadsheet then of course there won’t be any like this after you’ve made corrections and run the utility again)
16. Make a note of how many were deleted, the contributor will surely enquire. Further sorting and deleting (no grid references for example) may be carried out; also replace all the spurious “0”s with blanks
17. Finally, sort the remaining by Date and Locality (or other desired sequence) and Save

C.3. Validation - Record cleaner

Download **NBN Record Cleaner** at https://nbn.org.uk/tools-and-resources/nbn-toolbox/nbn-record-cleaner/ and install it.

Download the **User Guide** too.

This tool validates your data – checking the format against a set of built-in rules. This includes spotting bad dates (e.g. 31st February) or spatial references (e.g. TL123) and checking the spelling of items like species and vice county names. You can correct any problems on screen or change the original source and reload before proceeding.

It dates from 2013 so is not as current as the UKSI checker above.

The validation checks that it carries out are sequential, thus if it detects errors concerned with Scientific name then it will not be able to progress to check dates and spatial references.

See https://forums.nbn.org.uk/viewforum.php?id=23 for the latest developments regarding this tool.

Preparing datasets

The formats of datasets will differ according to the chosen method of dissemination (see D. below)

Metadata

Whichever format is chosen, it is advisable to maintain a personal record of metadata describing the particular dataset submitted. This will have the value both of providing metadata descriptions required by the GBG and of keeping your own historic records of work that you have carried out. GBGs do not keep such historic records once a dataset has been updated.

See Appendix 1 for example

D. Dissemination

The methods available for dissemination are varied.

For submission to the UK's NBN Atlas, two methods are available:

1. direct submission of an appropriately formatted dataset, complete with metadata describing that dataset,
2. iRecord methodology.

For submission to GBIF the following are available:

3. iNaturalist methodology (only in cases where images are available)
4. submission of dataset and metadata via endorsing organisation, this method is specifically aimed at non-UK species occurrences

Each of the above methods requires a different means of dataset preparation. Full details of Methods 1. & 4. are to be found in Part 2 of this guide.

Method 3. is detailed in **Stilt & Stalk Fly newsletter 2.** Method 2. was developed by BRC's Martin Harvey and is currently being implemented by Dipterists Forum Secretary Jane Hewitt.

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 (3-10) (Vol. 2). Retrieved from http://www.alerc.org.uk/uploads/7/6/3/3/7633190/running_lrc_vol2_03_to_10.pdf

(NBN). (1999). Running a Local Records Centre V2 (16-21). 2(16–21), 120–164. Retrieved from http://www.alerc.org.uk/uploads/7/6/3/3/7633190/running_lrc_vol2_16_to_21.pdf

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

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

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

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/

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

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

Colwell, R. K. (2004). Biota 2: The Biodiversity Database Manager. University of Connecticut, 2, 862. Retrieved from http://viceroy.eeb.uconn.edu/biota/Biota2Pages/biota2_download.html

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

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

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

GBIF. (2011). Darwin Core Archive Assistant User Data Guide. GBIF, 1.1(January 2011), 1–22.

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

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

Huang, 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

James, T. (2004). Data validation and verification: Report of a Networking Naturalists Seminar. Huntingdon.

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

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

Morris, R. (2012). Developing a Dipterists Forum Database - records from the internet. Bulletin of the Dipterists Forum, 73, 5–6.

Morris, R. (2011). The Future of Dipterists Forum Field Meetings. Bulletin of the Dipterists Forum, 72, 28–29.

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

Morris, R. (2004). Field Week Records. Bulletin of the Dipterists Forum, 58, 7.

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

Perring, F. (1971). The Biological Records Centre - a data centre. Biological Journal of the Linnean Society, 3(3), 237–243. https://doi.org/DOI: 10.1111/j.1095-8312.1971.tb00186.x

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.

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.

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

Rognes, K. (2014). Grossly Inaccurate Biodiversity Data: An Example from Italy Regarding Blowflies (Insecta, Diptera, Calliphoridae). Bulletin of the Museo Civico Di Storia Naturale Di Venezia, (65), 103–120.

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, 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.

Stroud, R. (2016). Improving the flow and quality of Biological Data throughout the National Biodiversity Network. (August).

Sumner, D. P. (2013). Recording Schemes: Made Public? Bulletin of the Dipterists Forum, 75, 5.

Sumner, D. P. (2014). NBN Gateway Chestnuts. Bulletin of the Dipterists Forum, 77, 6.

Sumner, D. P. (2015). The Future of Biological Recording Systems. Bulletin of the Dipterists Forum, 80, 5.

Sumner, D. P. (2005). NBN Gateway: A review. Bulletin of the Dipterists Forum, 59, 10–11.

Sumner, D. P. (2004). Societies & Schemes Steering Group. Bulletin of the Dipterists Forum, 57, 11.

Sumner, D. P. (2006). Field Week Records. Bulletin of the Dipterists Forum, 61, 4.

Sumner, D. P. (2015). Disseminating data: How to share data through the NBN Gateway. Practical Naturalist (Unpublished Manuscript), 58–63.

Sumner, D. P. (2015). Records for the DF Summer Field Meeting 2015. Bulletin of the Dipterists Forum, 80.

Sumner, D. P. (2010). NBN Gateway. Bulletin of the Dipterists Forum, 69, 6–7.

Sumner, D. P. (2007). NBN Gateway Matters. Bulletin of the Dipterists Forum, 62/63, 18.

Sumner, D. P. (2020). European Micropezids & Tanypezids: Newsletter #2. Dipterists Forum Report: Stilt

& Stalk Fly Recording Scheme, A(2), 1–13.
Sumner, D. P. (2017). Dipterists Forum - Recording Scheme - Stilt & Stalk Flies. Occurrence Dataset. https://doi.org/https://doi.org/10.15468/mwjnku
Sumner, D. P. (2019). European Micropezids & Tanypezids: Newsletter 1. Dipterists Forum Report: Stilt & Stalk Fly Recording Scheme, A(1), 16. Retrieved from http://micropezids.myspecies.info/
Sumner, D. P. (2006). Archiving practices and principles in Biological Records Management r Metadata. Unpublished, (December 2005).
Sumner, D. P., & Ball, S. (2004). Field Week Records 1987 to 2003. Bulletin of the Dipterists Forum, 57, 7.
Sumner, D. P., & Grose, M. (2010). Darwyn Sumner & Matthew Grose. Habitat, 1(06951023), 1–13.
Sumner, D. P., & Grose, M. (2010). Custodianship Responsibilities in Local Records Centres: Archives & Metadata. In ALERC: Reports (Vol. 1).
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.
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
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

Resources

- Running a Local Records Centre Volume 2: 01 to 02
- Running a Local Records Centre Volume 2: 03 to 10
- Running a Local Records Centre Volume 2: 11 to 15

Acknowledgements

Many thanks to

Revisions

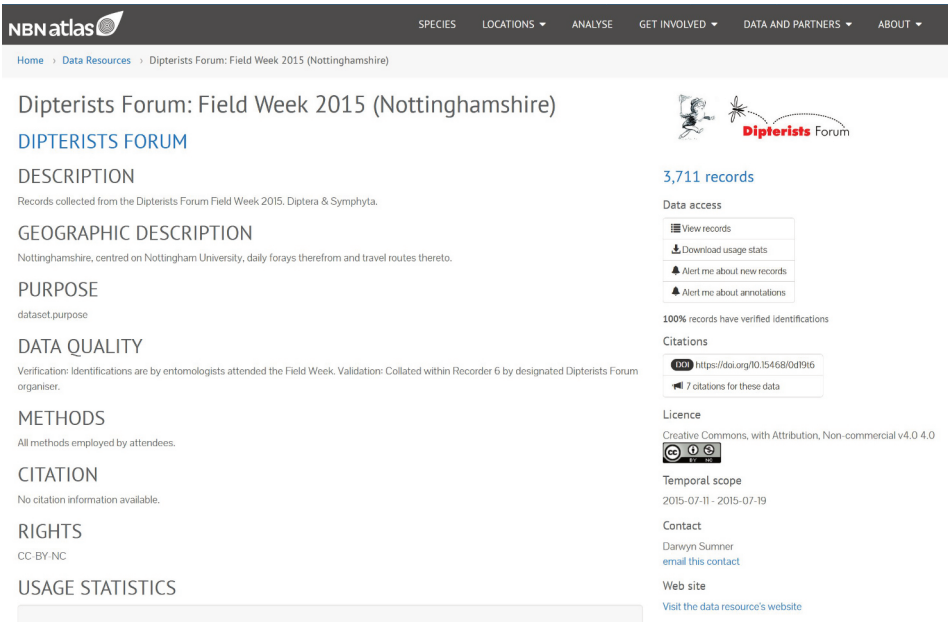
Version 2: Restructure & redesign to take into account Dipterists Forum's Online Publishing guidelines (Bulletin 88).

Cite as:

Sumner D.P.. (2020). Biodiversity Data Management 1: Records collation, Dipterists Forum Report, C(3 V2), 6

Appendix 1: Metadata

The basic set of information is shown on a typical NBN data resources page:



the above information needs to be provided by the person submitting the dataset.

NBN Guidance and a spreadsheet to enter the information are located at https://docs.nbnatlas.org/share-species-occurrence-records-with-the-nbn-atlas/

The current number of fields required is 21, the old Gateway one I got up to 44.

Their spreadsheet format is a good place to start in order to set an additional one up for your own reference. It's only a start though, for your own use it is valuable to add the following fields:

Version	to locate older versions that you submitted
Periodicity	how frequently you propose to update the dataset
Dataset key	[after upload] a unique reference to the dataset
Date uploaded	[after upload]
Number of records	[after upload]
Number of species	[after upload]
DOI	[after upload] see above

Setting up one's own spreadsheet well in advance of the preparations of the datasets is a good practice. When it comes time to submit the data one can simply copy and paste into the NBN metadata spreadsheet.

Use one spreadsheet sheet per dataset.