10 Dipterists Forum Technical Guide

Biodiversity Data Management 5: Digitisation Project 1: Steve Falk

Darwyn P. Sumner

Project managers Darwyn Sumner, David Clements, Steve Falk

Series C, Issue 10 (June 2021), Version 5 (August 2021)

Keywords

Digitisation, Diptera, Steve Falk

Summary

Historical lists of species occurrences by prolific naturalists are commonplace. Observations may have taken place long before the advent of computers and exist as longhand lists. Many of such lists have nowadays been digitised and made publicly available as Open Data by the work of Local Environmental Records Centres and by museums from collections. There remain however several important naturalist records collections of individuals yet to be digitised.

This project addresses the digitisation of the written records of naturalist and author Steven Falk. His material is contained within 13 loose-leaf A4 folders maintained as his personal records over a period of years up to 2014. The number of records is estimated as being around 200,000, around half being records of Hymenoptera and Syrphidae, both groups being the subjects of books written by Falk.

Introduction

Initiated in 2014 by DS & SF, this project began in earnest in 2015 when the SF folders were scanned by Warwickshire LRC and BRC began work on digitising the records. Summaries of progress with the project are to be found in Dipterists Forum Bulletin (Sumner, 2014-2021.)

BRC ceased work on the project in 2015 following the retirement of Val Burton and the pdf scans of the folders were passed to DS in June 2021, thus transferring responsibility for the project back to its originators (see Roy in Sumner, 2021.) The project is therefore now being managed and progressed by Dipterists Forum. BRC continue to act as advisors.

Open Data

There is an agreement that data digitised from Steve Falk's folders will be managed according to **FAIR** principles (findable, accessible, interoperable and reusable) Accordingly **datasets must be submitted to the UK's NBN Atlas as a primary objective** (Dickie et al., 2021). At the time of writing (Version 1, July 2021) precise arrangements have yet to be made. The spreadsheet model detailed here conforms to the Darwin Core requirements of the NBNt team and datasets digitised to this model can thus be submitted directly to them by the project managers.

Secondary objectives are at the discretion of other users but to avoid duplication on GBGs, uploading to iRecord is discouraged. Biological Recording applications and software systems may be used as these are required for analyses, but users of those systems are asked to take care to avoid duplicate uploads to NBN Atlas.

Crowdsourcing

A scoping exercise by DS suggests that a crowdsourcing approach is feasible. It relies on the preparation of spreadsheet datasets (Phase 1) which analyse the folders up to the point of Events (see model). Once these have all been prepared then the files may be utilised to extract records in

Phase 2. Volunteers are essential for Phase 2 and only desirable for Phase 1 work. See Table 2. for the current state of progress with this work.

Techniques and tools

Though many potential tools are available for Biological Recording, none except online systems are available to a wide enough range of potential volunteers. With the exception of spreadsheets each requires either plodding through hand-written sheets one by one or lends itself to highly complex sharing methodologies.

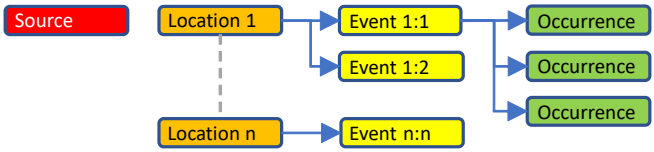
Everyone has a spreadsheet though, its use is frequently the first step before uploading to other systems. Given that the main task in this project is typing in order to transfer from longhand notes in a pdf onto a computer, the spreadsheet is the optimal first step tool. The application used by the author is Excel 2019.

PDF applications are also required. The simple free reader may suffice and though more sophisticated applications permit functions such as bookmarks, notes and highlighting. The nature of handwritten notes (essentially just a picture) means that some of them cannot be used on these files. In particular highlighter functions are generally ruled out (pun intended). In PhantomPDF Pro areas (frre-hand rectangles) can be highlighted. The critical function required from this tool therefore is the ability to display the page number of the pdf document.

The model

Since the NBNt upgraded to a system which utilises the Darwin Core biological recording data standard (Wieczorek, 2012), some effort has been devoted to the construction of a simple model which would facilitate the submission of records to the NBN Atlas via a DwC spreadsheet emulation. The model has been successfully used to upload a number of Dipterists Forum datasets (https://registry.nbnatlas.org/public/show/dp172) and is detailed in a sequence of papers (see http://www.micropezids.myspecies.info/node/291)

The model is built within a single spreadsheet file using separate sheets with rows linked via the familiar VLOOKUP function to each successive sheet.



Source:

A single row detailing just the folder and the person checking the grid references

Location:

All the locations detailed in each folder, less than 100 of these typically.

Event:

Linking to each location via VLOOKUP in this sheet are added the dates and the page number(s) of the pdf where these are detailed. Typically the number of rows are less than 400.

Occurrence:

In this sheet, again linked to events via VLOOKUP the actual records (species occurrences) are added using a Checklist derived from the UKSI. No name typing involved, just a case of picking from the UKSI list on the Checklist sheet. The number of rows depends upon the group you are working on, small groups such as Micropezids + Tanypezids or Sepsidae will result in <100 per file but medium-sized groups such as Sciomyzids, Empids etc. will find a good deal more whilst Syrphids would be a considerable undertaking

Data quality

The objective of the 2002 NBN trust document: "Improving Wildlife Data Quality" (James, 2002) was to give guidance as to how to do just that.

Two important concepts are defined at the outset in the section "*Guidance on data verification, validation and their application in biological recording*":

> **Data validation**: carrying out standardised, often automated checks on the "completeness", accuracy of transmission and validity of the content of a record.
>
> **Data verification**: ensuring the accuracy of the identification of the things being recorded.

Steve Falk's records are undoubtedly of a high quality. Pains should therefore be taken to ensure that the digitisation of these records achieve an appropriately high standard.

The job may be divided into two phases. The first deals with **Validation** alone but only to the extent of ensuring that the goegraphical locations are properly interpreted.

Where + When

The second phase deals with **Verification** in the main. Here the task, though more laborious, is greatly facilitated by Steve's skills but in view of the old nature of taxon names will require a good deal of taxonomic expertise to digitise into currently accepted names. Some validation is also required here, particularly in regard to ensuring that the collector and determiner are correctly assigned.

What + Who

Phase 1: Validation

The requirement for volunteers to complete this phase is minimal as this has been completed. The task is comprised of the first three of four parts of the simplified Darwin Core model above.

Validation by checking grid references using paper or online maps may require further attention. See the appendices in this document on using Excel and using maps.

Potential volunteers wishing to improve the standard of this work would be most welcome, though Phase 2 work has begun.

Phase 2: Verification

Digitising the entire batch of records would be feasible only if a dedicated full-time contractor were employed. The task of digitising somewhere in the region of 200,000 records is simply too great to be achieved in any kind of reasonable time frame.

Accordingly we adopt a system which permits Diptera Recording Schemes and volunteers focussing on particular Diptera groups to extract records of specific families from the pdf documents.

Each volunteer is provided with a set of Phase 1 spreadsheet files, each one carefully checked for accuracy with respect to grid references and dates. Each file relates to a specific pdf scan of Steve's folders (1 to 12)

Each DwC spreadsheet file contains completed Source, Location and Event sheets. Each has a UKSI Checklist of Diptera species as specified by the Phase 2 task in hand and a single dummy row on the Occurrences sheet - which is where Phase 2 extractors begin.

Technique

Extracting

1. Open the pdf and the spreadsheet matching it.
2. Navigate to page 1 of the pdf and the first row of the Occurrences sheet on the spreadsheet.
3. Ensure that digits 7 & 8 of the *occurrenceID* are changed according to the Recording Scheme codes in the table below. So if you see

SFKF05DFOc000001

and you are dealing with Sepsidae, change this to

SFKF05SeOc000001

this is imperative so as to prevent any likelihood of duplicate GUIs

4. Experiment by changing the UKSI number on that first dummy record (don't change any formulae)
5. Copy the formulae (columns D to AE only) of that first line downwards into new rows - just a few rows at a time (once for each time your taxon group appears on that page). [Select the items in that row, take the cursor pointer to the lower right corner until it is a small black cross then pull downwards]
6. Copy the EventID reference (Column B) down to encompass all those lines ensuring that it remains the same (see tip 1)

7. Drag the occurrenceID entry down so that it increments by 1 (do this throughout - see Unique identifiers)

8. Check that the code in the *catalogueNumber* column (U) matches up with the pdf page

9. Find your taxon in the Checklist, copy the UKSI from its first column and paste into the *taxonNumberID* field (F) in the Occurrence sheet. Name, Family and Rank should now display. If you are unsure then choose a higher rank (Family, Genus) and/or add a note to self in the *taxonRemark* column so that you can return and fix later.

   a. If your pdf editor supports area highlighting (PhantomPDF Pro: Comment | Area Highlight) then highlight the text of the taxa extracted and save the pdf.

10. Move on to page 2 of the pdf and add the next eventID in the next row

11. Save with your two letter code appended to the filename (e.g. Falk 08aMT.xlsx)

12. Work through the entire pdf a page at a time repeating steps 1. to 8., saving frequently.

13. Repeat for each of the 12 pdfs.

14. Save your final spreadsheet

There are of course a number of ways you might work your way through a pdf file. I did this Phase 2 work on M&T as I developed the Phase 1 files.

With the Phase 1 files ready-prepared and if you have a pdf editor able to highlight areas you could first scroll through the pdf and highlight the taxa for your attention. Then start from the top of the pdf again and fairly rapidly build up your list of occurrences in the spreadsheet.

Tips:

Tip 1. There are a handful of columns you can use for notes, scribblings etc.

The *basisOfRecord* column doesn't get completed until the very end (nearly all "specimen" of course but watch out for Steves photos = "image") so that can be used for any temporary numbering you might wish (e.g. number of your taxa of interest on a page)

The *taxonRemark* column is free text, useful to put the pdf page number here as a marker, even before you've added *eventID* or the *taxonNumberID* so that you can return to that page later and add those. Change it later to a note perhaps if you've got tricky taxon names to resolve

Tip 2.: Because the taxa on each successive pdf page can be similar, you will find yourself copying the UKSI number from previous entries (rather than continually referring to the Checklist sheet) as you work your way down the Occurrence sheet.

To make them easier to locate, fill-colour blocks of taxa UKSIs in the Checklist. For example in the Sciomyzidae I had all the *Tetanocera* brown and all the *Pherbellia* blue; I also finished up with a pretty pattern.

Warning: Take care not to disturb the formulae. Some Data operations may move them and upset the relationships. The Filter command may be relatively safe with care but Sort operations are unadvisable. Best to wait until after you've compiled everything and all the formulae are gone.

Compiling

You've now 9 sets of occurrences in a format which is acceptable for direct submission to NBN Atlas. There's no need to go through iRecord or add to your Biological Recording application. You can do that later but Dipterists Forum is intent upon Open Data - which means we intend to put your compilation on NBN Atlas without much further ado. Or if you've an NBN Atlas dataset for your recording scheme already on NBN Atlas - we'll give you details as to how to append these records to your dataset (or do that for you - with your connivance)

The next steps show you how to meld all those 9 sets of occurrences into a single table in a spreadsheet.

1. Open a new blank copy of Excel
2. Highlight the title row of the Occurrence sheet in the first of your 12 spreadsheets
3. Select Copy (either Ctrl C or Copy on the Home tab)
4. On the new file select cell A1
5. On the Home tab select Paste Values and the last option on that group: *Values and Source formatting* This retains the text and colour coding (but not the column widths - guess you'll have to experiment with that - or just tweak the widths afterwards.)
6. Select cell A2 on your source sheet occurrence tabl and select to the end (Shift Ctrl End) and Copy
7. Position your cursor in cell A2 of the new sheet and select Paste Values (so of course all the links are now gone, you just have the text.)
8. Repeat steps 2-7 but start at the row on the new spreadsheet which is below the last populated row.
9. Continue until you've done all 9 of your extractions.
10. Name your single sheet "Occurrences"
11. Save with a suitable filename
12. Add a second sheet "Sources"
13. In the same fashion copy in each of the single lines from each of the single Phase 1 spreadsheets. This will enable you to track sources in the case of any enquiries.

Now contact us to determine the most appropriate means of uploading to your NBN Atlas dataset.

If you found that a breeze and want to try another, then contact us and we'll assign a code if necessary (see table below & "Schemes" sheet in the DwC files.) Other Recording Schemes might be glad of a helping hand. Also consider a non-Recording Scheme group, there are some relatively easy ones such as Bibionidae.

Submitting to NBN Atlas

Though the compilation you have just made is ideally formatted for direct submission to NBN Atlas it is currently unclear how best to deploy these according to FAIR principles.

Options are:

A Simply append them to the appropriate Recording Scheme dataset (e.g https://registry.nbnatlas.org/public/show/dr940)

B Use them to start off a Scheme dataset - for example the Conopidae which hasn't got one yet. A successful upload may incentivise the organiser to process other Scheme data holdings.

C Similarly, append them to the Recording Scheme dataset which is one of a pair. For example the Sciomyzidae has one dataset of iRecord records only and one derived from spreadsheets submitted to the scheme. The latter receives the append.

> The pair system is for the convenience of NBNt who tell me that to update a dataset means taking everything off, adding the new data to that and reloading everything back again. If iRecord records are involved then this process becomes a lot more complex for them.

D Set up an NBN dataset "Steve Falk - miscellaneous Diptera to 2014" (or similar). This would be advisable in any case if non-Recording Scheme records were processed.

Metadata

The NBN datasets must be accompanied by suitable metadata which describes the dataset according to some straightforward guidelines. The author can advise on an acceptable format.

Principles

Checklists

These are the standard UKSI checklists of UK Diptera (or other taxon group) obtained from the UKSI Sandbox and prepared and maintained by Chris Raper who liases closely with Peter Chandler. As this is the formal species index for the UK, these codes and their taxa are also those fully compatible with NBN Atlas.

Note the format of these Checklists. They look a little odd but they evolved from a need to provide a list for European species on a Scratchpad site and also dual purpose to get species name from UKSI code and vice-versa (so UKSI is in 2 columns) and provide the taxon rank and Family name to help on the Occurrence sheet.

These work on the same principle as the links between Locations and Events. In this case the lookup GUI is the formal code used by the UK Species Inventory. These are readily obtained from the UKSI Sandbox at

https://uksi-sandbox.nhm.ac.uk/index.php

The DwC spreadsheets as provided already have a number of checklists to various groups (including dragonflies), they are listed on the Schemes sheet.

If you require others then obtain the list from the UKSI Sandox and **append** them to the bottom of the list on the Checklist sheet, following the same convention.

If you choose to replace rather than append then of course you will break the links on the Occurrence sheet to any Checklist taxa you've replaced.

Beware of extending your list beyond the range of rows specified for the named range "Check" currently set at 2492. Amend that if needed via Excel's Formulas | Name Manager

Provenance

Globally Unique Identifiers (GUIs)

People do swear by these or swear at them when they go wrong (Guralnick, 2015). Basically they comprise a code which has been contrived to make every open data occurrence record unique. This allows each record to be traceable back to its origin - providing provenance in other words.

In addition, if structured nicely, they permit linking across tables (sheets) in a spreadsheet, much in the way that links are created in a structured relational database.

Biological Recording applications such as Recorder 6 use these. Theirs are 16 digits in length, the first 8 being assigned to the user and the rest generated internally by the program. In a database you usually can't see them but in a spreadsheet you've got to pay attention.

Here's the structure we've adopted for use in these spreadsheets:

SFKF05DFSc000001

The first 3 digits are Steve's initials of course.

The next three specify the number of his folder.

Following that, two letters for Dipterists Forum. (always the same in the Phase 1 work, altered when it comes to Phase 2 Occurrences to indicate the particular Recording Scheme - see list below)

Two digits to indicate context within the model (Sc for Source, Lc for Location, Ev for Event and Oc for Occurrence) and to provide links.

Finally we've 6 digits which can increment for each successive record, allowing for nearly a million - plenty for most purposes.

Recording Scheme codes (for GUI digits 7&8)

LB: Soldierflies & Allies

MT: Micropezids & Tanypezids

Se: Sepsidae

Sc: Sciomyzidae

Te: Tephritidae

Co: Conopidae

Ke: Kelp Flies

Te: Tephritids

Sy: Syrphidae

The remainder are to be found in the "Schemes" sheet of the DwC spreadsheets.

We're not entirely constrained even to that list. One could choose a Family not covered by existing Recording Schemes, just as long as a unique two letter code is chosen.

In Phase 2, when the volunteers start to extract records on the "Occurrence" sheet, the first *occurrenceID* must be altered as follows:

SFKF05MTOc000001

For the work on Micropezids & Tanypezids. Or:

SFKF05CoOc000001

For work on the Conopids and other Families within that scheme

This ensures that there are no duplicate GUIs

If you have ever dowloaded datasets from NBN, GBIF, iNaturalist or similar then you will see these GUIs. Spot one of the above structure and you'll know where it came from.

Acknowledgements

Many thanks to

References

Costello, M. J., & Wieczorek, J. (2014). Best practice for biodiversity data management and publication. Biological Conservation, 173, 68–73.

Dickie, I., Kharadi, N., Neupauer, S., Butcher, B., Treweek, J., Judge, J., … Harvey, M. (2021). Mapping the Species Data Pathway : Connecting species data flows in. Geospatial Commission, 44(May), 150.

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

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.pdf

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

Management, B. D. (2020). Technical Guide. 1–12.

Murray-Rust, P. (2008). Open Data in Science. Nature Proceedings. Retrieved from https://www.nature.com/articles/npre.2008.1526.1

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.2017.0391

Robertson, T., Doring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLoS ONE, 9(8). https://doi.org/10.1371/journal.pone.0102623

Sumner, D. P. (2020). Biodiversity Data Management 1: Records collection & verification. Dipterists Forum Report, C, V2(6), 1–12. Retrieved from http://www.micropezids.myspecies.info/node/357

Sumner, D. P. (2020). Biodiversity Data Management 2: Records collation & publishing. Dipterists Forum Report, C, V1(7), 12. Retrieved from http://www.micropezids.myspecies.info/node/357

Sumner, D. P. (2020). Disseminating Biodiversity Data 3: Recording in Europe Darwyn. Dipterists Forum Report, C, V1(8), 1–52. Retrieved from http://www.micropezids.myspecies.info/node/357

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

Using Excel

Tips & tricks

Apologies for telling stuff you already know but mistakes can happen (incremented dates or UKSI numbers) and new revelations made (that Ctrl D trick was new to me)

Copying and extending rows

Incrementing

Any row item which ends in a number will increment when dragged downwards (select a cell, move to lower right until a small black square appears then drag down)

Copying

Select a range of cells in a row and do the same and most of them will not have numbers, several of them have formulae or plain text so this technique usually does what you want it to.

Select a cell containing text and some cells below it in a column, then press Ctrl D. The text gets copied into all your selected cells. A faster method than others, particularly useful when you've to do it on a long column when the value you are copying has scrolled off the top of the screen.

To make a single copy of the cell above, press Ctrl ‘ (the two keys above and below the right hand Shift key on your keyboard)

How the links work

These operate through the various ID codes on each page. So for example you assign *locationID* codes to a list of locations on the Locations sheet and then use that code to repeat pieces of the information onto the Events sheet (where you add more bits of information such as date and page number of the pdf.) The linking mechanism is Excel's VLOOKUP as in

=VLOOKUP(N2,Locat,2,FALSE)

Where N2 is the cell where you bunged the *locationID*, Locat is a predefined range on the Locations sheet and 2 is the column offset on that sheet.

Useful to do it in this way because it means you only have to figure out the Location details once and then everything cascades across to the other sheets. That's of particular value when you come to "improve" a Location entry (e.g. I've used the Woodland Trust website to get a grid reference to Castor Hanglands)

Drop-down lists in Excel

To users of MS Access these might seem a good idea to implement Checklist taxon picking. Excel seems rather unsophisticated in this area in comparison. If you have any successes then let me know, otherwise we're stuck with copying and pasting from a list.

Using pdf editors

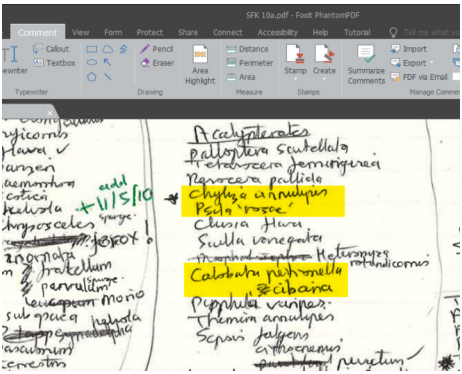
Ideally this application should be a pdf editor rather than a reader.

The main function required of this tool is that of **area highlighting**.

Simple pdf readers won't allow any kind of editing of course, they'll give you fancy tagging tools on your own system but you can't save them nor send the marked files to others.

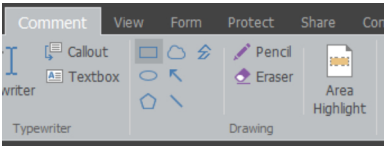
Pdf editors have a variety of commenting tools, two examples are as follows:

Foxit's Phantom PDF



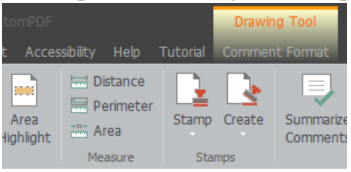
Example of area highlighting in PhantomPDF. Changing colours is laborious, non-persistant and the default yellow is unalterable.

Alternatively the rectangle shape may be used as follows:

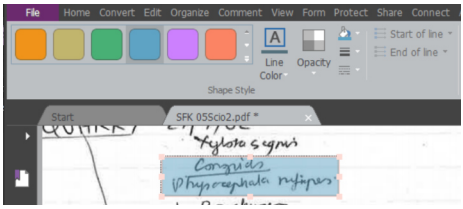


Select the rectangle shape on the Comments tab.

Whereupon the Drawing Tool tab appears.



Selecting this tab gives access to a the drawing tools allowing you to change the outline, colour and opacity of boxes you draw.

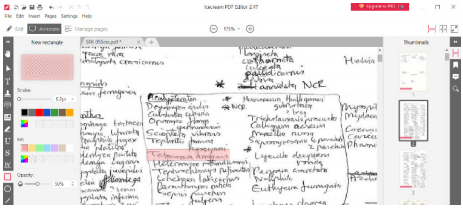


Legibility is not as clear as in highlighting, a balance must be set between that and clarity of the coloured box. This example is set at 50% opacity.

This method has the advantage that the box tool, once set, remains active whereas in area highlighting the button must be pressed each time

After drawing your last box, make a single left click somewhere or the shape is treated as a temporary object and not saved.

Icecream PDF Editor (free version)



Rectangles in Icecream PDF Editor (reviewed in DF Bulletin 92)

You may find others such as **PDFelement** & **PDF Studio**

These techniques of commenting on the pdf are of value in highlighting taxa to be worked upon.

Save your own versions of the pdf, adding your group's two letter code (e.g. "MT") to the filename. This helps with cooperative work on a single group (e.g. a "spotter" marks up the pdf and an "extractor" verifies and enters records into the spreadsheet) - we've started this with the Sciomyzidae. This means too that some helpful volunteer could flag taxa to help out a scheme (e.g. the tiny handful of Oestridae) or simply someone could check your highlighted pdfs for any you missed.

Highlighters of different colours are features in some in some editors but may be laborious to implement in others so it may be least confusing to start off each group with clean un-highlighted pdfs and stick to one colour.

Other commenting tools such as bookmarks may prove to be helpful too.

Using Maps

Online map utilities

OS Grid reference finder

One of the most useful tools to help find grid references of places from named locations (or other data) is the **OS Grid Reference finder**.

https://gridreferencefinder.com/

Here's an example:

1. In the Location box type "Wells-next-the-Sea" and Go
2. But Steve specified "on beach" so it's not the centre of the town as shown
3. Zoom out and navigate to a section of the beach (north to the tip of the East Fleet road)
4. Right click on a portion of the beach ("The Run") to produce the panel. (this doesn't work unless the pointer is still showing from your original search)
5. Copy the 6 figure grid reference from the panel. If that is too precise then reduce it to a 4 figure grid reference, it will still be usable to create 10km distribution maps in the future. Don't be more precise than the originator intended.

Vice county calculator

Not strictly needed for Darwin Core dataset but some researchers might find it of value, it's handy for R6 users.

Use **Cucaera's Watsonian Vice County calculator** at

https://www.cucaera.co.uk/grid-ref-to-vice-county/

On the Location sheet.

1. In the column to the right of the *gridReference* column (O), add a comma to each row. Best to do this in small batches, omit the comma from your last row of interest.
2. Copy the data from columns N + O (Ctrl C)
3. Paste that into the box at the top of the calculator's site
4. Select all the data from the Results and copy (Ctrl C)
5. Paste to the right of Column O in your spreadsheet
6. It messes with column heights so correct this back to 15
7. Now check carefully that the pasted grid references (Column P) match the original in Column N. This is important as sometimes the calculator returns two values for one grid reference (hence short batches)
8. In the *stateProvince* column ensure that you have a formula which correctly returns the VC

number from column Q (=left(Q2,3) for single digit VCs and =left(Q2,4) for double digit ones.

9. To tidy up, select your work in column M, copy then paste as values. Delete the mess in columns P & Q

Show a map of a bunch of sites

Most quickly done using the OS Batch Convert Tool at

https://gridreferencefinder.com/batchConvert/batchConvert.php

1. Copy selected rows from the *gridReference* column on the Location sheet.
2. Paste that into the first box (1) in the above page
3. At step 5 select Convert
4. Ignore the returned data (6)
5. Use the various options at 7 to view maps

Google Earth Pro

This is of some value because of its huge dictionary of Location names. It doesn't return OS Grid references though, so use it in combination with the OS Grid reference finder if you have no luck with that alone.

Inventory

	Content	Project	pdf	Excel	pages
1	Warwickshire	BRC			225
2	Warwickshire	BRC			293
3	Warwickshire	DF	SFK 03	Falk 03a	130
4	Warwickshire	DF	SFK 04	Falk 04a	127
5	Warwickshire VC38, 33 etc.	DF	SFK 05	Falk 05a	225
6	Cornwall	DF	SFK 06	Falk 06a	93
7	Cambridgeshire & Huntingdonshire	DF	SFK 07	Falk 07a	41
8	Northants to Herefordshire	DF	SFK 08	Falk 08a	131
9	Kent	DF	SFK 09	Falk 09a	118
10	Hampshire	DF	SFK 10	Falk 10a	131
11	duplicate of 6	DF	SFK 11	Falk 11a	91
12	East Sussex	DF	SFK 12	Falk 12a	184
13	Invertebrate Site Register	NE			

- **Folders 1 & 2** were digitised by Val Burton at BRC and presumably exist as records on the BRC servers.
- **Folder 3** was begun by BRC but since we don't know how far they got I propose we start again.
- **Folder 11** is a duplicate scan, the actual scanning job was done twice as evidenced by page 33 in one folder showing a page corner turned and in the other it was straightened. The difference in page numbers is as a result of folder 6 containing 2 blank pages (56 & 91)
- **Folder 13** consists of Invertebrate Site Register forms. The ISR was uploaded to NBN Atlas (https://doi.org/10.15468/7wbiu7) and contains records until the end of 2005.

How to get started

Files may be obtained at

http://www.micropezids.myspecies.info/node/307

They are in the form of ZIP files, each containing both the scan of the folder and its associated DwC format spreadsheet. The latter is already populated with Locations and Events by the author so steps 1 and 2 below may be omitted

1. Construct DwC format spreadsheet files for each folder (except 1, 2, 11 & 13) [DS]
2. Populate the above wrt Locations & Events [DS]
3. Validate all of the above [DS + SF + Volunteers]
4. Use your pdf editor to area highlight taxa of interest in the pdf folder file. Save that as [filename][code] so as not to confuse with the original undefiled copy.
5. Simultaneously record the page number in the *taxonRemark* column on the Occurrence sheet. One row for each taxon you find on that page.
   a. This will help you as you add taxa later. Unless there are loads on every page in which case steps 4 & 5 may not help.
6. Proceed with Verification technique described above.

Revisions

First published on 12th July 2021 at http://micropezids.myspecies.info/node/307
V2: 2021-07-13
V3: 2021-07-14
V4: 2021-07-25
V5: 2021-08-19

Cite as:

Sumner D.P.. (2021) Digitising Project Guide. Dipterists Forum Report, C(10 V5)