Make your data accessible

Open Entomology Project

6 October 2016

Introduction

You've been collecting insects all semester, discovering your inner Darwin and sorting specimens from numerous localities. You've also made specimen preparations and labeled each one according to collection best practices. As part of this process you've also established a database of specimens in your collection: a spreadsheet with columns for locality, collector, identifier, *etc*. Each student in the course likely developed the spreadsheet independently, with his/her own column headers. How can you make sure these data are available for research?

Today we'll talk a bit about biodiversity data standards and tools and other resources that facilitate data sharing. We'll also refine your spreadsheets and use a tool that allows your data to be contributed to established biodiversity databases.

Methods and materials

We'll be working at a computer for these exercises, but you can follow-up at home, on your own. Software and datasets required:

- Web browser (*e.g.*, Firefox or Chrome)
- Text editor (TextWrangler for OSX or Notepad++ for Windows)
- Spreadsheet editor (*e.g.*, Google Sheets, Microsoft Excel)
- Your data!

Glossary

We'll be using terms you may not be familiar with. Some of these concepts are provided below, and we'll go over them together. If you have a new one to add raise your hand!

1. Darwin Core
2. GBIF
3. metadata

4. TDWG
5. XML

Your data

Take a look at this label, which probably looks similar to the labels you've created for your specimens:

USA: PA: Centre County:
Pine Grove Mills, 40.730,
-77.884, ± 250m 15.iv.2016
A.R. Deans, sifted litter

Question A: What kinds of data do you see represented? How many kinds are there? List them.

Question B: If you had these data from millions of specimens—all 35+ million insects in the Smithsonian Institution, for example—what kinds of hypotheses could you test? See if you can think of three example research questions.

Question C: Now look at your spreadsheet. How many columns do you have? Is each one a data type? How would you enter the label above into your spreadsheet? What data type(s) is/are in your spreadsheet but not represented in that label above?

Question D: Compare your spreadsheet to your neighbors' spreadsheets. How are they different?

Biodiversity data in research

Biodiversity science is incredibly rich, with respect to the array of research questions and the data types that can be applied to them. We'll discuss a few examples of research that relies on collections data. As we talk, think about the minimum data required for these kinds of questions.

Biodiversity data standards

Hopefully you thought of some compelling research questions for Question B. How can we aggregate data from hundreds (or thousands!) of natural history collections to test your hypotheses? Your spreadsheet almost definitely differed from your neighbors' databases, and you can imagine that a similar scenario exists in the natural history collections world. The Smithsonian would use a different approach than the American Museum of Natural History and the Field Museum.

Fortunately there are established biodiversity data standards and tools that facilitate sharing. We'll look at those developed by TDWG, especially the Darwin Core and associated resources. Open a Web browser and navigate to `http://rs.tdwg.org/dwc/terms/index.htm`. This massive list almost definitely includes the data types you identified in Question A, along with dozens more that may or may not be relevant to your collection.

Find the following terms and read their definitions: *catalogNumber*, *recordedBy*, *eventDate*, *samplingProtocol*, *fieldNotes*, *higherGeography* and *locality*, *decimalLatitude*, *decimalLongitude*, *coordinateUncertaintyInMeters*, *scientificName*. Do any of these sound familiar?

> **Question D:** How many of your spreadsheet column headers are represented in the list above? Based on your reading of these (and maybe your eyes strayed to other terms), how would you change your spreadsheet organization? Or would you?

Prepare your data!

GBIF provides a tool, the Darwin Core Archive Assistant (Global Biodiversity Information Facility, 2011), that will make each of your data sets (spreadsheets)—no matter how different—available more broadly for research. Now that you've thought about Darwin Core fields you're ready to get started:

1. Open a Web browser and navigate to `http://tools.gbif.org/dwca-assistant/`.

2. Our data are primarily "occurrences" (which specimens occurred where), so check that box in the upper left, under "Core".

3. On the upper right of the page you'll see a place to type the name of your spreadsheet file ("Filename:"); make sure it ends with .csv, as we'll be exporting your data as a **c**omma **s**eparated **v**alues file.

4. When we selected "occurrences" the tool auto-populated our list with two required terms (fields), *ID* and *basisOfRecord*. Open a new tab in your browser and navigate back to `http://rs.tdwg.org/dwc/terms/index.htm`; find and read about *basisOfRecord*. We'll discuss *ID* as a group.

5. Now go back to the Darwin Core Archive Assistant. Find the terms under Occurrences (far left) that match the columns in your spreadsheet. Check each one that applies.

6. The terms should have been added to the area in the middle. Are they in the same order as your spreadsheet columns? Item "0" in the list of terms should match column "A" in your spreadsheet, item 1 to column B, *etc*. Note that you can drag the terms up and down in that list. You can also insert a "spacer" for columns in your spreadsheet that do not match any Darwin Core terms. I recommend dragging *ID* to the bottom of your list.

7. Once you have a list of terms that matches your spreadsheet headers it's time to validate! Navigate to the tab that reads "meta.xml". You should see an XML file that describes your database. Click "Save File" and export the file to your desktop. You should also export your spreadsheet as a CSV file. Put both of these files—the CSV and XML files—inside a folder. Now compress that folder into a .zip

8. Go back to the "meta.xml" tab and click "Validate". This should launch the validator in a new browser tab. Find "Upload local archive:", choose your zipped folder and click "Validate". Did you get any warnings? Look at the rendering of your data at the bottom of the page. Do the columns line up? We'll discuss the results and any questions as a group.

More thoughts on biodiversity data

Now that you have an idea of how to create a set of files that contribute to the greater scientific enterprise it's time to think about how to extend and enrich your data sets. We'll discuss some of these issues as a group.

- All of your specimens will be deposited at the Frost Entomological Museum. How do we specify that in your file?
- Do any of you have images or videos? How do we associate them with specimen records?
- How do we associate specimens, for example a parasitoid and its host?
- How do we account for a range of dates, as we might see with a Malaise trap?
- How to explain our approach to georeferencing (*i.e.*, finding a latitude and longitude for each specimen)? Do you know what geodetic datum means?
- Can biodiversity data be copyrighted?

ProTip

GBIF provides structured spreadsheets as Microsoft Excel files: `https://github.com/gbif/ipt/wiki/occurrenceData#templates`. These files can be imported through any of the myriad Integrated Publishing Toolkit (IPT) instances.

Epilogue

This handout is part of an open curriculum. Original files are available free for anyone to download, copy, modify, and improve at the Open Entomology GitHub repository (Open Entomology Project, 2016), which also provides a mechanism for reporting problems and other feedback:
`https://github.com/OpenEntomology/InsectBiodiversityEvolution/issues`

References

Global Biodiversity Information Facility. Darwin Core Archive Assistant User Guide, Version 1.1. `http://tools.gbif.org/dwca-assistant/gbif_dwc-a_asst_en_v1.1.pdf`, 2011. Accessed 5 October 2016.

Open Entomology Project. Insect biodiversity and evolution. `https://github.com/OpenEntomology/InsectBiodiversityEvolution`, 2016. Accessed 19 August 2016.