How to scrape

Project: OSPRÉE - Budburst review - Wolkovich Lab
Version: 1 - Updated on Feb 03 2016 at 19:26 by Tim Savas

Description

Meta-analysis 101!

To to W: drive > WeldShare > Wolkovich Lab > Budburst Review > meta_analysis_labmembers > your folder

In your folder, you'll find a zip file for ImageJ; install this. Add the Figure_Calibration.class, which will help for giving x and y calibrations to images.

You each have an Excel workbook to add data to. Tim and I will manually merge data afterwards.

meta_general - metadata for each sheet

source - list of the paper we are working with. Bibliographic information and notes on usefulness for our purposes.Note the "ToDo" column, which tells you which figure or table to focus on. You may find other figures are better, these were from our initial quick read. Also pay attention to datasetID column, which tells you how you should enter the identifying information for each paper.

study - Details on each experiment within each paper; possibly only one line for a paper, if only one experiment is relevant. This sheet is useful for our overview of what kind of experimental manipulations were done.

data_simple - Aggregated data on each experimental treatment combination.

data_detailed - Detailed data for the experiment, with all relevant information filled out.

- Responses may be pre/post treatment, time, or other. Temporal responses, such as days to 50% budburst, are fairly common. An example of an other type of response would be percent budburst, again fairly common.

scratch - For temporary formatting and manipulating data scraped from ImageJ.

The two most important tabs to fill out are study and data_detailed. We can aggregate data down to get the data_simple version later.

Open your PDF and find the designated figure or table as noted in the source tab of the worksheet. Also note the datasetID column in source; use this as you fill out data_detailed and study.

Take a screen shot and import into ImageJ, following Tim's instructions from lab meeting. Use the scratch tab to get data into the right format, and then copy into data_detailed. Fill out the study tab as best as possible to describe the experimental treatments used in each study within each publication.

For a screen share walk-through on this process, see these tutorials:

Data Scraping_01

Data Scraping_02

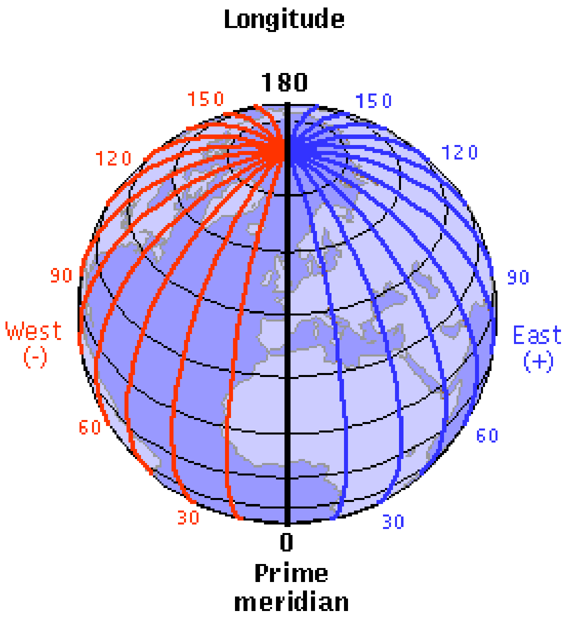
Comments on this notebook

1. Written on Jan 03 2016 at 13:17 by Elizabeth Wolkovich

Notes from Lizzie's first attempt:
To add the the Figure_Calibration.class: In ImageJ go to plugins, navigate to the Figure_Calibration.class, click on it and follow through a few clicks to add the plugin

Also, I had some trouble getting the measurements to show up after calibrating, switching to the pointer and clicking. I think I needed to set the preferences on my pointer tool to auto-measure (I think).

Here's what I am using to convert to decimal latitude and longitude: http://andrew.hedges.name/experiments/convert_lat_long/ (a little easier for me) and remember to add NEGATIVE to your longitude if it's West:



Also, you can check where things are by just typing in lat and long into Google maps.

Finally, I used this: http://www.timeanddate.com/date/duration.html to calculate the number of days between two dates for calculating dormancy. Note that I do NOT include the end date in my calculations.

But here's my question: What goes in 'response' -- it seems to just be 1 if we are doing daystobudburst?

2. Written on Jan 04 2016 at 10:28 by Tim Savas

Also, I had some trouble getting the measurements to show up after calibrating, switching to the pointer and clicking. I think I needed to set the preferences on my pointer tool to auto-measure (I think).

I think you're describing a similar problem I had. After doing the figure calibration and selecting the yellow pointer tool, you start clicking inside the figure but no points appear. The reason for this is that the "rectangle" you drew for the figure calibration is still masking the figure, and until you click out of it, you can't draw points under it. It's hard to see! So to get rid of the invisible rectangle, just click the mouse once outside of its edge.

Side note: After drawing all of your points onto a figure, you can press Command-M to bring up the resulting table of values. I do this in the video tutorial--and whenever I scrape--but didn't describe the key command!

What goes in 'response' -- it seems to just be 1 if we are doing daystobudburst?

Yes, when "daystobudburst" is the response variable, we've been entering 1 into the "response" column. Meanwhile the "daystobudburst" values scraped from the paper will go into "response time." My original thinking was that temporal data (days to...) should be kept in the response time column, though if any folks have done otherwise, it should be a straightforward fix.

3. Written on Jan 04 2016 at 17:18 by Jehane Samaha

Hello! I've been doing okay on ImageJ-- once I figured out Command-M, everything was smooth!

My questions are more about how to interpret written information in the papers, and which columns call for which pieces of information. Bear with me :)

Side issue: Dan's little yellow sticky notes on the column headings aren't working for me, fyi-- I can't read the full text of the note, just the first few words. Is anyone else having this trouble with excel? I've never had this issue before.

I'm focusing for now on filling out the columns required in data_detailed.

Columns:
Population: I've been treating this as the location of the population, not the study site. I've been approximating coordinates using google earth-- for example, for a cultivar from "Norway" with no other information given. Is it okay to make that estimation? I'm not planning to enter altitude information though unless the paper gives it?
Material: Is this referring to the material that undergoes dormancy, the material that is forced, or the material that is observed? In the paper I'm working on, they induce dormancy in whole plants, then force cuttings indoors, then record bud burst for the uppermost 5 buds. In the spreadsheet, I see "Seedling, cutting, potted sapling", and then I also see "shoots," "apical bud," and many other terms in the example data given in "data detailed."
Other Treatment: what types of treatments should go here (applied to the plant? the cutting?) Also, if there is more than one "other treatment," should we list them with commas in the cells? (I think being clearer on "Material" will help with this one, too).
Columns for Times and Temperatures: These columns seem to be either repetitive or in the wrong order or vague. I'm confused about what the difference between "chilltemp" and "dormancy induction temp" is-- and why does one call for a number while another calls for a descriptor? I have many similar questions. For example, the difference between "chilldays" and "dormancy induction days"-- and for these, I could use a more precise definition of each. For example, is "dormancy induction" days at a certain temperature? And when dormancy is induced at ambient temperatures outside, how should the start date be defined?
Response Columns: More info, please! Tim's answer to Lizzie (above) was very helpful. More like that, for each column, would be great (i.e. what type of data you expect to go into each column). Also, I feel confused as to how to specify what units/ what type of data it is that we are putting in.

Lemme know if you have questions about my questions :)
Thanks!
~Jehane

4. Written on Jan 05 2016 at 11:11 by Ailene Ettinger

Hi all
I have a few questions so far- I also emailed these to you, tim and dan, as i wasn't sure how you prefer to answer these, since two of them are specific to my papers (i.e. should they be included?) and one is more general:

1)One study that I'm looking at (Odium and Colombo 1989) is about budset at the end of the growing season (not leaf-out or flowering, as you mention in the overview document). do you still want this included?
2) Another study (Rinne et al 1997) looks at effects of long-term chilling and short-term freezing on bud burst, but does not manipulate photoperiod. do you still want it included?
3) more general question- for the Length of various treatments- i assume you want this in # of days?
4) One study reports results as both % budburst and days to budburst. Do you only want the days to budburst?
Thank you!
-Ailene

5. Written on Jan 05 2016 at 11:58 by Tim Savas

---

1)One study that I'm looking at (Odium and Colombo 1989) is about budset at the end of the growing season (not leaf-out or flowering, as you mention in the overview document). do you still want this included?

I would say not to include this paper. I'll run it by Dan to see what he thinks.

2) Another study (Rinne et al 1997) looks at effects of long-term chilling and short-term freezing on bud burst, but does not manipulate photoperiod. do you still want it included?

I think we should keep this. I've gone through a paper or two about short-term freezing events, and a good handful more focused on chilling effects than photoperiod. If it's giving you trouble, maybe I can help out at the retreat? :)

3) more general question- for the Length of various treatments- i assume you want this in # of days?

Yes, days is our go-to. Often a figure will be in weeks, in which case I convert it to days. The "response time" column is where these values can go.

4) One study reports results as both % budburst and days to budburst. Do you only want the days to budburst?

Sounds like a good paper! I scrape and record both when they're available. Those units are the best two to find in a paper/figure, so they're worth keeping. If I saw "days to budburst" and "cm elongation above above bud," or something similarly obscure, that's a case where I'd likely just take the days to budburst data. Nothing personal against cm elongation above bud...

6. Written on Jan 05 2016 at 16:04 by Ailene Ettinger

A few more questions:
1) Ruesink98 records the percentage of cuttings with flower buds (but does not record leaf out or budburst). Do you want this study included?

2) If figures show standard error, do you want us to scrape that data as well?
3) how/do you want us to deal with significant digits/rounding? for example, should we round to the nearest day for all our scrubbed data, if
thanks!

7. Written on Jan 05 2016 at 16:29 by Tim Savas

---

1) Ruesink98 records the percentage of cuttings with flower buds (but does not record leaf out or budburst). Do you want this study included?

I've just taken a look at this paper. Yes, I would include those data!

2) If figures show standard error, do you want us to scrape that data as well?

Aha, I'm realizing we did not go over this during the meeting. If it's there and *clear enough* to scrape, error can be recorded. Often times the SE bars are in the way of each other or not quite discernible, in which case we've decided to avoid them. But if the bars are clear, record them. Values can go in "resp_error" and just SE in "error type."

3) how/do you want us to deal with significant digits/rounding? for example, should we round to the nearest day for all our scrubbed data, if

Not sure if you're whole message got through but I think I've got it! Sure, when you scrape a figure, "day" values might come out fragmented, e.g. 84.248 days. Before doing any rounding, I would take a look at the methods section and see if they describe the time interval in question--they might say "we took measurements every 10 days," or "we moved samples inside every 30 days to test chilling." I almost always find something along those lines, which then informs how I should alter or round those fragmented "day" values.

On the other hand, you might be dealing with day values not related to any set time interval hiding in the materials and methods section. For that, I would say just to keep the entire number and we'll decide on significant digits to keep soon enough.

Side note to all: I have our papers handy on my computer and can easily pull them up. So feel free to mention an author/year in a question--like 1^-- and I'll likely respond with a better answer :)

8. Written on Jan 07 2016 at 13:35 by Tim Savas

Hello fellow data scrapers,

I've attached a better "meta_general" tab for your spreadsheet. It contains more thorough descriptions of each column in the sheet, specifically data_detailed. I recommend taking a quick look-over.

I'd also like to suggest hiding the following columns. They pertain to a select few papers I've worked on, which you *likely* won't need, or at least in the meantime shouldn't have bloating an otherwise simpler spreadsheet.

Hide: G; O-R; Y; Z-AD

You may have already noticed that the data_detailed sheet is not so intuitively organized. The order of the columns should better reflect the time line of any experiment: dormancy induction, freeze treatment, chill treatment, force treatment. Instead, you'll see that the dormancy induction columns (O-R) are sandwiched between the forcing temperature (M, N) and photoperiod columns (S, T)--which are themselves placed before chilling treatment columns. [scratches brow in puzzlement]

It's too late to hand out a new, well-ordered data_detailed sheet for you to work on, but if anyone is interested, I could make a simple input/output table to help you quickly reorder the columns on your own, for example:

N, force temp night : AD
O, dormancy induction_temp : V
...

Let me know if that would help and I'd be happy to create it, the one caveat being that we should *all* undertake the reordering.

9. Written on Jan 20 2016 at 19:50 by Elizabeth Wolkovich

Update on fields in the xls (especially chilling):

- fieldchill: Was there chilling that occurred in the field that occurred before clipping? ambient (ambient=chilled); OR fieldchamber; OR leave blank if no information (e.g., if seedlings)
- chilldays: chilling days in an experiment OR (if not an experiment) chilling days in field reported OR (if not reported) days since November 1 (or either given in study or southern hemisphere: maybe May 1)
    - Maybe switch to October 1 and April 1...
    - Maybe just give date clipped and calculate (best idea)?

    - Add column for start date if given (sapling brought outside)?
- chilltemp: chilling temperatures (for experiments only)
- chillphotoperiod: chilling photoperiod (for experiments only)
- dormancy (infrequent) -- experiments explicitly on dormancy
- freezing (infrequent) -- experiments explicitly on freezing (that pulled out of chilling and put in freezer)
- number.longdays (infrequent) -- only for Tim and Dan
- response..pre.treatment. and response..post.treatment. (infrequent) -- you'll know if you need this

10. Written on Jan 26 2016 at 13:01 by Tim Savas

One thought after our decision to find weather data for the chilling experiments. On occasion the paper will report their own winter weather data for the site. (How diligent of them.) Heide93 is a good example of this. I've attached a screenshot of their figure. I would image that we should scrape and use this when it's available. I'm just uncertain as to which column these values would go in. Perhaps we add a new one?

11. Written on Jan 26 2016 at 13:20 by Elizabeth Wolkovich

I'd suggest we not scrape weather data until we decide to do it for all studies. We probably will run into issues if we use different weather data sources for different sites as opposed to just pulling data from a common source, where cross-site issues have been at least examined. So for now it just seems worth noting it.

12. Written on Feb 02 2016 at 23:06 by Tim Savas

Hello Dan! I just want to be clear about lat/long data. Which is more valuable to us: the site of the experiment, or the origin of the population sampled? I am currently scraping a paper where both would be valuable: they're testing chilling requirements (good for recording site of experiment) with about ten different populations of a single species (good for recording origin of samples, which they do include in the pub). Hope this question is clear, let me know otherwise :)

13. Written on Feb 03 2016 at 19:26 by Tim Savas

^ Saw those new growing lat/long columns. Thanks!