Scraping new papers for EGRET: August 5, 2024

Getting Started: copied from ospree howtoscrape2019.tex

You'll need:

1. Excel or other program that makes .xls or .csv files
2. ImageJ download for free from here `https://imagej.net/Welcome`. You'll also need to add the Figure_Calibration.class, which will help for giving x and y calibrations to images.
   (a) To add the the Figure_Calibration.class: In ImageJ go to plugins, select *add plug in*, then navigate to the Figure_Calibration.class file that you downloaded on your computer, click on it and follow through a few clicks to add the plugin.
   (b) If you have some trouble getting the measurements to show up after calibrating, try switching to the pointer and clicking. You might need to set the preferences on your pointer tool to auto-measure.
      i. A clarification on above from Tim Savas (original OPSREE lead data enterer): *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!*

Now here's what to do:

1. Copy the excel file data/egret for git repo (egret/data) and make your own extension using your initials, for example, Jane Doe would write "egret_dmb". This will be the spread sheet you enter your data into and then in the future, someone will save the tab as a csv and merge all of our files together into the master data.
2. Familiarize yourself with each tab:
   (a) **meta_general**: metadata for each sheet
   (b) **source**: list of the paper we are working with. Bibliographic information and notes on usefulness for our purposes. Note the "assigned.to" column, which tells you which figure or table to focus on. You may find upon reviewing the text the paper does not suite out selection criteria, in which case you would change the 'A' (accept) in your accept_reject column to 'R' for reject. This tab should also be updated if a paper is not published in english.
   (c) **data_detailed**: Detailed data for the experiment, with all relevant information filled out.
   (d) **scratch**: For temporary formatting and manipulating data scraped from ImageJ.
   (e) The two most important tabs to fill out are **source** and **data_detailed**.
3. Read your paper and update out the information in the "source" column if needed and fill the "data_detailed tab." Be sure your datasetID and study info agrees with the source tab, if not—figure out what is wrong and fix it.
4. Read the paper to decide if the study is eligible for inclusion in EGRET. These selection criteria include:
   - Seeds are germinated under experimental conditions
   - Not conducted under natural conditions in the field

- Species are not crops
- Experimental treatments are not confounded

5. Take a screen shot of the figure and import into ImageJ, following detailed instructions from Tim's OSPREE data scraping video from git/ospree/notes/howtoscrape/Data Scraping Tutorial.mp4 or the general instructions outlined in the egret wiki. Use the scratch tab to get data into the right format, and then copy into data_detailed.

A few more "how to's", trouble shooting etc

- **For dealing with lats and lons:** Tool to convert to decimal latitude and longitude: `https://andrew.hedges.name/experiments/convert_lat_long/` 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.
- **On entering response times:** If the figure is a time-series or curve of the percent germination over time, the days to germination will be the values from the figures's x-axis. Otherwise the value will be the final day that germination was observed. Remember for time-series to also include the zero point, or to include days along the x-axis that have zero on the y-axis as this is still part of the curve.
- **On dealing with error:** If it's figures records error and it is *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."
- The paper pdf's were saved in a Google drive folder. Unfortunately, the folder went missing in the summer of 2022. We believe what happened was that the folder was made using a student account that was deleted once the owner graduated. Since we were not notified that the folder went missing until the end of the field season, it was too late to request it be temporarily restored and it was never backed up. We reached out to everyone and were able to get many of the pdf's back in a new folder (now owned by Lizzie). We have now re-downloaded the papers into a new folder and backed it up.

Data cleaning and decisions

1. Cleaning code are located in the analysis/cleaning/source folder in the egret repo.
2. Early notes about data scraping can be found at `https://github.com/lizzieinvancouver/egret/issues/4`, with some highlights being:
   - Adding a column for the rare cases where studies start measuring germination day relative to the day of first germination and not the day they start monitoring germination (i.e. day of germination temperatures, imbibing etc)
   - Excluding papers not in english unless the person scraping the data could read the language fluently
   - Decided to just add seed mass as a yes/no for now, we will have to go back and scrape it some day if we want it
   - When dealing with error, enter it as zero if bars are not distinguishable from the points, or as "indistinguishable" if multiple overlapping points have overlapping error bars that can not be scraped with certainty (see issue of a visual example of this).
3. We decided at our first retreat which columns we wanted to clean, the list of columns and who was assigned to clean them is outlined at `https://github.com/lizzieinvancouver/egret/issues/14`, key points from this issue are discussed below for each cleaning script they pertain to. This is also the issue were we list updated column names.

4. JS had 10,000+ empty rows in their csv data file for some reason, which was making the code run slowly. DL manually corrected this in the csv file.

5. The following outlines some of the more complex cleaning decisions and issues:

   - **cleanChemical.R**: Chemical names should be cleaned and standardized, updating names of brand name products to clarify what they are (e.g. include explanations such as "—herbicide"). Names were updated to remove the treatment duration (which was moved to the trt.duration column where it belongs). Also see issue #65.
   - **cleanScarification.R**: We initially thought to move the hot water scarification to the soaking column, but it was confirmed that this was in addition to soaking, so it was left as is. There was also a question about "cold scarification", which had values duplicated in the chilling column (which seemed odd), this is not mentioned in the script, which I think means that is was explicitly called this in the paper. Issue #14 has details about some of the add entries (e.g. partial scarification) that were double checked. Also see issue #41 and #62
   - **cleanCoordinates.R**: For 123 studies there was no specific lat and longs given. We are only interested in source population locations, not where studies were conducted. To infer the coordinates for these studies, we used Google Earth to drop a pin in the centre of the named region (county, city, etc) and get a general value for these studies. We added an additional column to denote when we needed to take these less specific site locations. We also are checking that coordinates seem right by plotting them on a map. Also see issue #21.
   - **cleanGerminationTempDuration.R**: Germination temperatures in 108 studies varied, either as a range or between day and night. These studies with multiple temperatures were manually checked and additional data columns entered for day and night temperatures. If it was not specified what temps were day and night, we assume the warmer temp is day. Also see issue #18.
   - **cleanChillTempDuration.R**: Some studies have alternating day and night temps, so we created unique columns for alternating temps and daylength. Some studies include warm stratification, but this is relatively uncommon. There are 30 studies that are missing chilling temperature. These pdf's were manually checked and notes made to check whether this is true or improperly entered data. It was discovered that in some of these studies, there are no chilling treatments, but there is moderately cold storage temperatures. We discussed whether this storage could function as chilling, but decided that it depended on whether the storage is wet, which would allow the seeds to be imbibed. Also see issue #23, #40 regarding supra-optimal germination temps, or #44 to check papers without chilling for which storage temps may be cold.
   - **cleanPhotoperiod.R**: A total of 113 papers have no photoperiods given. If studies done in complete darkness, photoperiod should be zero, but may have been entered as "NA". Also see issue #42
   - **cleanStorage.R**: Here we created a general column which grouped similar treatments and one that included more diverse treatments. The cleaning of storage is now linked to chilling and whether studies without chilling have moist storage. Also see issue #39.
   - **cleanYearGermination.R**: Over 100 studies do not specify what year the study was conducted. It was discussed and our meeting Aug 2 as to whether it was worth our manually checking the paper. It was decided that it was not, although if we do change our mind, it would be possible for a number of studies to infer it based on the language used in some papers (e.g. seeds were stored for x months after collection"). Also see issue #63, closed for now.
   - **cleanMisc.R**: In this code we are removing redundant data that was reported in both tables and figures of studies, and studies for which we can not use the data (treatments are not properly described). Here we also clean minor columns, such as seed.mass.