A global database of soil seed bank richness, density and abundance

A.G. Auffret et al.

Instructional document for data collection.
Created by Alistair G. Auffret and Emma Ladouceur, February-March 2020. Updated as issues were identified during data collection (first half of 2020).

Global seed bank richness

This guide should help you with the data input, and most of the variation in how to get information from papers into the spreadsheet from all the different studies should be covered here. All data entry takes place on the Google sheet. **Do not edit the structure of the spreadsheet or rearrange (e.g. sort) the data.**

1. Read the title.

Skip paper if **clearly** not a community seed bank study (i.e. a study where the aim is to survey all species in the soil) e.g.

- *The viability of Anacamptis laxiflora (Orchidaceae) seeds and the symbiotic germination*

- *Soil seed burial and competition with surrounding plants determine the emergence and development of seedling of an endangered species Horsfieldia hainanensis Merr. in China*

- *Effects of seed age and dormancy-breaking treatments on the viability and germination of the Gulf of Saint Lawrence aster (Symphyotrichum laurentianum)*

or the wrong sort of seed bank e.g.

- *The First Step Is to Bring It Into Our Hands: Wild Seed Conservation, the Stewardship of Species Survival, and Gardening the Anthropocene at the Millennium Seed Bank Partnership*

- *Creative Practices of Care: The Subjectivity, Agency, and Affective Labor of Preparing Seeds for Long-term Banking*

If in doubt, click the URL link. Check that the Authors, Title match the info from the same row as the link.

Note that some papers do not have DOI numbers, and therefore no links based on the DOI. Here, we provide a google scholar search link. Please see if the paper is available there (if there is a result, make sure that it is the same paper), and if not please make a reasonable effort (normal web search, find and browse the journal website, etc.) to locate a copy of the paper.

If no access - Write 'No access' in the appropriate column and leave the 'Human' column blank so that someone else can check. If you take any action on a paper (reading it, entering data, emailing authors), write your initials. If someone else's initials are there, ask them before taking action.

Whether or not you click the link, write your initials in the Human column. Because we are several people working on the spreadsheet at once, it may be a good idea to 'book' a block of 10-50 studies at a time for you to work through during a session. In that case it is very important that you remove your initials from the studies that you don't manage to get through during the session. It's also important not to accidentally copy your initials over someone else's when doing this. Some papers dotted around the spreadsheet have already been booked/inputted because one of the authors is in the group.

2. Read the abstract.

Again, skip the paper if it becomes clear that the community seed bank was not studied. In general we will not enter data from meta-analyses and reviews, but do take a quick look in case the review contains useful data from e.g. grey literature and/or regions with fewer data points that we might want to go back to later.

3. Read the methods.

By now you should know if the community seed bank was studied. If so, add the sampling information to the spreadsheet. This guide gives a general overview of how the information in different studies should be entered on the data, but does not go through each individual column on the spreadsheet. So, when entering data, be sure to read the notes of each column on the spreadsheet (hover your cursor over the different cells of the top row).

Plot location and Habitat

Latitude, Longitude and Location
Add the degrees either in degrees, minutes, seconds and compass direction, or, if decimal degrees are given, add those as they are in the degree column. Ideally, double check the coordinates given on a map before entering into the spreadsheet. At the very least, use your head to make sure that you are in the right hemisphere (and remember that Latitude can be max 90 and Longitude 180), but you can also paste the coordinates into Google maps, which is quite flexible regarding different types of ` and ' for minutes and '' " and `` for seconds. Degree sign is less flexible (so here is one °). Make sure the correct compass direction is in the appropriate column, or for decimal degrees that there is a negative sign where it belongs. Some studies might (confusingly) combine approaches, giving a compass direction and a decimal number, in which case it is very important to think about whether a minus sign is needed or not.

Some examples:

1. Ali's office, Sweden 59°49'5.0"N 17°39'29.7"E
2. Charles Darwin National Park, Australia -12.442269, 130.877048
3. Christ the redeemer, Brazil 22.9509 S, 43.2064 W

	Lat_Deg	Lat_Min	Lat_Sec	Lat_NS	Lon_Deg	Lon_Min	Lon_Sec	Lon_EW
1	59	49	5	N	17	39	29.7	E
2	-12.442269				130.877048			
3	-22.9509				-43.2064			

If coordinates are given for different sites in the same paper, **and the number of seeds and species are given separately per site/type of site (check results section)**, add a new row in the table (*right-click the row number, click add row - check if anyone else is using the spreadsheet, see point 5 later in this doc*), copying all the previous columns (including title, authors etc.) and add separate results for each site. Do not click and drag to copy, because for the DOI and URL columns, this can sometimes incrementally increase the "number" of the, resulting in incorrect DOIs/URLs that might even link to a completely different paper. If coordinates are given separately for different sites, but results are given in total for the whole study, there should only be one row, in which the coordinates should be averaged. Don't forget that minutes and seconds go up to 60, not 100. But, to avoid mistakes, it is best to in these cases convert the different coordinates to decimal degrees (e.g. here: https://geoinfo.sdsu.edu/hightech/LM3/dd1.php), then average that and enter the averaged decimal degrees in the degree columns.

Please also add some information into the Location column, even if coordinates are given. This will help us identify any errors later on. If coordinates are blank, please give as much detail as possible. You are also welcome to try to find approximate coordinates yourself, using the information given in the paper.

Habitat and Target_Habitat
Add the relevant habitat from the following categories: Arable (incl. other heavily impacted anthropogenic land uses, Paddy field), Grassland (including shrubland, savanna), Forest, Aquatic (Including marine), Wetland (usually clearly described as wetlands, but also including swamp, bog, and e.g. mangroves). Grassland is very broad (i.e. includes shrubland, tundra, desert), but we will be able to gauge the different types according to climatic conditions using WorldClim data, which should be more straightforward and consistent than trying to establish it from the text. Wetlands are different, as climatic conditions can't tell the difference between grasslands and wetlands. Habitats such as 'flood meadows', 'salt marshes' can be difficult to judge. These should be added as wetlands, while 'meadows' and 'wet meadows' would be classed as grasslands. Riparian habitats can also be difficult to place, as some riparian habitats are close to water, but rarely inundated by it. We can add further habitat types later if necessary, let Ali know if you think this is the case.

Because many seed bank studies are related to (potential) restoration of habitats, or degradation of vegetation following land-use change, there is also a 'target habitat' (or ideal habitat) column so that if necessary we can separate core and marginal/degraded habitats for the analysis. Use the same habitat categories as above, or leave blank if it is not clear, or if a pristine or core habitat is being tested. Examples:

Abandoned grassland (that is clearly wooded) would be Habitat: Forest, Target_Habitat: Grassland.

Semi-natural grassland would be Habitat: Grassland, Target_Habitat: <blank>

Tree plantations count as Habitat: Forest, Target_Habitat: Forest.

Arable fields should never have a Target_Habitat. Abandoned or repurposed arable fields should be categorised as the new habitat type, usually in a degraded state. For example Post-agricultural forest would be Habitat: Forest, Target_Habitat: Forest, and Grasslands on former arable fields would be Habitat: Grassland, Target_Habitat: Grassland.

Habitat: Grassland, Target_Habitat: Forest is also quite unusual, and should only be used where grassland is naturally or anthropogenically stable, and not to represent a temporarily open stage during a succession to forest. In many cases the study biome and author language will determine whether it is Habitat: Forest, Target_Habitat: Forest (rainforest degradation, abandonment of slash and burn agriculture) or Habitat: Grassland, Target_Habitat: Grassland - deterioration of traditional, long term native and species rich grasslands. Habitat: Grassland, Target_Habitat: Forest is where 'bad' pasture activities or other stability means that the habitat is currently Grassland but the paper clearly indicates that Forest is the ideal habitat.

Again, if different habitats were studied in the same paper (or if core and degraded habitats were compared), and the number of seeds and species are given separately in the results/tables/appendices, **add a new row for each habitat**, copying all previous rows (including location information). If it is not possible to separate the number of seeds/species from the different habitat types from the text/Figures/Appendices, choose the most appropriate broad habitat type in the study and group all results under that.

Experiment
If the study is an experiment, e.g. restoration or other treatment that has been carried out specifically for the study, put a 1 in that column, but no need to describe further. However, if the experiment involves herbicides or seed sowing, only take data from control plots or plots that don't have those treatments (apart from arable, as herbicide use is pretty standard there). If this is not possible, skip the study. If the study is non-experimental and in a habitat where seed sowing has taken place, or herbicides are sprayed as part of the normal management, then it's fine to keep. Do not mark studies that test different types of germination methods as experimental in this column (unless of course the habitat is also experimentally treated), see below.

Sampling methods

Hold on tight because this has the potential to get confusing. The important thing to realise is that many columns in the spreadsheet will be left blank, but it can be different ones in different studies.

Usually, a study will be based in one or more "sites", in which a number of "samples" are taken, either spread across the site or within "plots'.

Sampling dimensions
Most studies take a number of soil cores (samples) and pool them (mixing the soil cores together). In these cases, add the diameter of the core and the number of cores per site to the spreadsheet. Sometimes square samples are taken, in which case calculate the plot area (mm^2) and add it to the appropriate column. Add also the depth of the core/sample if given. Even if the top layer of vegetation was removed, as it sometimes is, input the total depth of the core here. Sometimes, instead of giving the dimensions of the soil core, studies give only the volume of each sample. In these cases add this information, otherwise leave the column blank and this will be calculated later. Rarely, individual samples are defined by weight. If so, add that.

Number of samples and sites
What constitutes a site will differ across studies, but generally speaking it is what the author would consider as a replicate. So it could be a grassland, a forest stand, a transect that goes within or across different land use types, etc. etc. Sometimes you will just have to make a judgment. Add the number of sites in the study (or in the relevant row for each study), and add the number of samples per site. You might have to calculate this from the number of

samples per plot and the number of plots per site. It can be useful to have a sheet of scrap paper in front of you so that you can write down the number of sites, plots and samples so that it makes sense to you before you input it to the spreadsheet. If individual cores are split according to depth (e.g. 0-50mm, 50-100mm), this is only one sample as far as we are concerned (because it is only from one area of ground), but authors might consider these separate samples, so be careful in these cases (but generally still report total numbers of seeds and species across the study anyway, although see examples of unusual sampling below). Sometimes, only the total number of samples in the study will be given (for example if different numbers of samples taken per site). If so, add that instead, otherwise leave blank and it will be calculated later. Note that you might have to add new rows per site or group of sites, even if they have the same location and habitat type, depending on how the information is presented in the results. For example if the number of seeds and species are given separately per site and not in total, then it is not possible to know the total number of species found in the study, because of unknown overlap across sites.

Method
"Emergence" is most common (soil extracted and grown later in a greenhouse, but also include garden germination here). Others include "Extraction" (i.e. sorting the sample and identifying seeds under a microscope). Germination in the field ("Field") is sometimes done. In these cases it is important to only include the study if the numbers of seeds and species are known to be from the seed bank (i.e. by actively preventing colonisation from 'above'. Studies that look at total recruitment of disturbed patches that may include both the seed bank and seed rain should not be included.

It happens that authors take a lot of samples, but realise that they don't have time or space to grow all of them (or identify them all under a microscope), or want to use some for looking at other soil properties such as organic matter or chemistry. In these cases, they might take a certain fraction of the samples, a certain volume, or a certain weight. If so, add that information here. If fractions are given or can easily be estimated, add that, because it is consistent regardless of how the samples may have been pooled. However, if a certain volume or weight was taken, this now needs to be calculated so that it represents the row of the spreadsheet. So depending on if the weight or volume of material was taken from the pooled samples at the plot or site level, these now need to be converted to the total amount of material processed in the study (or study-site-habitat combination represented by the row).

Some studies apply different germination methods to their samples. This could e.g. be for wetland sites where some subsamples are grown underwater while others are grown in normal greenhouse conditions to ensure germination of different types of species, or it could be different heat treatments for samples from ecosystems that regenerate after fire. Ideally, such studies should be kept to as few rows as possible in the spreadsheet. For wetlands, it seems that most studies report seeds and species numbers in total, but for more experimental germination studies, results from the different treatments may be given separately. If this is the case, take the results from the treatment that has the most seeds and species, and be sure that the number/volume/fraction of plots in earlier columns represent the samples that match the results that you add.

4. Read the results

Extract the total number of seeds and species from the study (or per habitat, region, site or groups of sites depending on how the sampling and data are presented in the paper). If mentioned at all, this information is usually prominent in the results text or in a table, but check also supplementary materials or linked dataset where available. If seed densities are

given, add this to the appropriate seed_density column (check units), even if total seeds are given as well.

If extracted easily from the text or tables, add the number of positive species (target, specialists, red-listed) and/or negative species (weeds or invasives etc.). This is just a very broad brush for 'good' or 'bad' species that might be something referees are interested in. For 'weeds' in arable fields, leave blank as in theory all non-crop plants in arable fields are weeds.

Here are some examples with a selection of the most important columns

In *study 1*, the authors took samples from three sites, but it is not clear how many samples were taken per site. Perhaps in the methods they write "We took between 7 and 12 samples per site, depending on the size of the field, resulting in a total of 34 samples" in the methods. They had three plots, from which they took two samples (of 10mm diameter) each. Samples were pooled and they used half of the material for identification of seeds under the microscope. They found 43 seeds of 22 different species.

In *study 2*, the authors took samples from four sites of the same habitat type. In each site they took four randomly-distributed samples of 30mm diameter. We leave Total_number_samples blank as we can calculate that later. They found 222 seeds representing 36 different species.

In *study 3*, even though sites were within the same study area and have the same habitat type, data were presented in a table separately for each site. For example, samples might have been taken in different years, or they are from restored grasslands along a time-since-restoration gradient but all fit within the same broad categories that we are using. Therefore we add a separate row for each group of sites (so in this case there were 3 sites per group, totalling 9 sites in the study), because without comparing overlap in species lists, we would not know what the total number of species is across the whole study.

Study	Sample diameter mm	Sample depth mm	Number sites	Samples per site	Total number samples	Method	Method volume fraction	Total Seeds	Total Species
Study 1	10	15	3		34	Extraction	0.5	43	22
Study 2	30	10	4	4		Emergence		222	36
Study 3	20	10	3	5		Emergence		97	15
Study 3	20	10	3	5		Emergence		159	10
Study 3	20	10	3	5		Emergence		36	8

Examples of unusual sampling/reporting that has come up so far:

- If authors only give the total sampled area in the site/study rather than the individual cores, use that and write 1 in the samples per site or total samples column, as appropriate.
- Studies that examine the seed bank at multiple depths should be interpreted carefully. If possible, treat the study as if the samples were not separated by depth, i.e. the entire core, representing the sampled area of ground. If there is unknown overlap in species because of the way that the data are presented, add the total

number of seeds for the core, and the **number of species in the uppermost layer**, because that has been consistently shown to be the most species rich. It is also possible to first ask the author for this information (see below).

- Studies that examine the seed bank at different times per year (i.e. take different samples but at the same sites) would be additional Plots_per_site.
- A study with one wetland, but different vegetation types within the wetland, 59 samples per vegetation type. Results given as a total rather than within vegetation types. This was one row, with 5 sites and 295 samples.
- A study where each sample consisted of a top subsample of 50 x 50 x 2.5 cm and then a subsample that continued deeper in the centre of that top layer sample as 10 x 10 x 10 cm. In this case, the results from the subsamples were presented separately, so we take only the results from the top subsample, which has consistent sampling dimensions that are comparable with other studies.

Additional columns

There is also a column for free text comments, if there is anything deemed important about the study that you think we should know about but is not covered in any of the previous columns. There is also an *Action Required* column. This is for the person entering the information for that study to remind themselves if there is something that needs to be added later for any reason. Mostly this is just to show when a request to the author has been made, but can also be to flag potentially valuable reviews or similar. When the required action has been taken, delete it from the cell.

In short, *Comments* won't necessarily be read, but can be useful when trying to understand peculiarities in the data. *Action Required* is for something that (potentially) needs to be done before the database is closed.

5. Adding rows to the spreadsheet

There are a few different situations described above where new rows should be added to the spreadsheet - when more than one set of coordinates are given in the paper (and results are also given separately per coordinates), when different habitat types are studied, and when results (numbers of seeds and species) from different sites or sets of sites are given separately.

To add a row, right-click the row number, click 'add row'.

In Google Sheets, if you add rows to the spreadsheet and there is somebody working on a row below you, their 'box' will remain on the same cell (J:3125), but the other information from that study will have shifted down the number of rows that have been added. **This could cause problems**, so:

- If you are editing and need to enter new rows, check if there is anybody else working on the spreadsheet. If so, write a message in the Google Sheets chat saying that you need to enter some rows and ask if it is okay. If you get a response, then go ahead. If you don't get a response, you can click that person's symbol and see where they are working, and hopefully be able to see if they are active or not before adding the rows.
- If you are editing and others are working on the spreadsheet. Be aware that rows could be added! Keep an eye on the chat function, and when entering data, always make sure that you are entering it under the correct study.

6. Requesting data

If a study looks relevant but important information is not given, please contact the authors - especially if it's from a potentially undersampled region or habitat. Important information means information that without which we cannot calculate seed or species density, i.e. it is not possible to work out the number or dimensions of sampling or the number of species and seeds for a particular row in the spreadsheet. If already contacting an author about such important information, it is of course relevant to also ask e.g. for coordinates of the study in the cases where they only give a study location. If asked for authorship, politely reply that we cannot do that as e.g. it would not be fair on the majority of authors who provide this summary data in their papers. If you engage in email communication with someone, **keep a copy of their emails**. Ideally create a seed bank map project folder for yourself, and save any data shared, and a simple text file of email communications in folders labelled by Author_Journal_Year. More on this below, under ***best data management practices***.

Important considerations before emailing :

Many people specialize in seed bank studies, and so some authors have many papers that have come up in our literature search list. Other people may have already emailed them to ask for data from a different study. If you are considering emailing an author, it is recommended to follow some double checks and following steps first;

1. Add your initials to the Human column of the paper you are currently considering.
2. Go to the *Author Search Sheet* tab. In cell A1 it will say e.g.

   REMOVE=QUERY(Data!A:AN, "SELECT* WHERE A CONTAINS 'Auffret, AG'")

   Change the Author name to the first or corresponding author of the paper you are considering, and delete the REMOVE (this is just to prevent potential slow performance of the spreadsheet constantly updating this column). Allow the sheet to populate.

3. See if your author is the first or corresponding author of any of the papers that turn up in the spreadsheet. If some papers have already been processed by others, you should send a message to the team mate(s) in question to ask if they would either consider contacting that person again if they have an existing dialogue (in that case tell them exactly what information is required from what paper), or to let you know if the person e.g. did not answer or was not prepared to share their information.
4. If there are unprocessed papers by the same authors, you should now claim these as yours to process (congratulations!), by adding your initials to the Human column for those papers in the main Data sheet. Check all these papers, filling them out as normal. By doing this, you will see if there are any missing details in those papers that you could ask about at the same time.
5. When you are ready to email, you may also want to use this summary spreadsheet sheet of all their studies. This will show the authors how many times they came up in our search, how organized we are, the effort we have gone to to extract the data we can ourselves, and that the data we need is simple and easy to find and enter. Copying this spreadsheet and attaching it to the email means that they might also be inclined to fill in the missing data themselves, although I would not recommend that you ask them to do that.

Email Template

Feel free to use the following template for requests. (note: on my system, it seems to work better to use 'paste without formatting' into my email program. Otherwise something strange happens when the email is sent and the formatting looks strange to the recipient).

Subject: Information request - seed bank

Dear <AUTHOR>

My colleagues and I are putting together a global review of seed banks. We came across your paper <PAPER NAME>. It appears to fit the criteria for being included in our review, but after reading the paper we were unable to find the following: <DELETE/EDIT AS APPROPRIATE>

- The total number of seeds germinated in the study. <OR SPLIT ACROSS REGIONS/HABITATS/TREATMENTS>

- The number of species identified in the seed bank. <REGIONS/HABITATS/TREATMENTS>

- The total area of ground sampled for the seed bank, or the size and number of soil cores

- Also, if you have coordinates for the study area (degrees and minutes is fine), that would be great - otherwise we can just use the information given in the paper to find an approximate location.

Could you please help? All studies that are included in our review will be cited in an appendix. Together with our eventual paper, we also hope to publish the database in an open-access repository. Therefore, in sending this additional information, please be aware that it will also be included in the published database. This will only include sampling information, the number of species, seeds and/or seed density, along with citation information about your publication. We cannot offer co-authorship on the paper or dataset for this, as it would not be fair for the vast majority of authors who published this summary data in their original papers.

Please let me know if it is not possible for you to easily find this information, or if you choose not to share it with us. Finally, this is a large undertaking and it is also possible that you may already have been contacted in relation to this project. Please tell us if you would prefer not to hear from us again.

Kind regards,
<NAME>

7. Final checks

We have created a few tabs on the spreadsheet to help you check for simple errors. When you are done with your papers (or more regularly, if you prefer), please check the following:

- **Action required check:** Check that you have carried out any actions that you earlier said needed doing.
- **Impossible coordinate check:** Check that none of your papers have impossible coordinates.
- **Invalid habitat check:** Check that none of your papers have habitats that are not in the list

- **My data check:** Check through the papers where you have entered data. You are very welcome to check through all the information again, but at the very least please scan through the rows and columns to see if you spot any clear errors in e.g. coordinates, sampling and results.

8. Best Data Management Practices

Please make a project folder for yourself, and keep the papers/appendices you download and manipulate in a folder named as Author_Journal_Year. Note that this is not really necessary when the information is in the paper, as it is unlikely that we will all lose access to a paper, but if you have to request it, or find the paper via google, researchgate or by contacting the author then please save it. Also if the information is found in supplementaries that are in a separate place (like with Oikos family) or datasets are on e.g. Dryad then please save a copy.
If you engage in email communication with someone, **keep a copy of their emails**. For example, convert to pdf and save to the project folder you created for yourself, or create a document where you copy and paste email communications, including email addresses, dates, and email content. Don't forget to save attachments and give them a name that you can connect to a study.